




Publicly Accessible Penn Dissertations

2022

Beyond Classical Statistics: Optimality In Transfer Learning And Distributed Learning

Hongji Wei
University of Pennsylvania

Follow this and additional works at: <https://repository.upenn.edu/edissertations>

 Part of the [Computer Sciences Commons](#), [Mathematics Commons](#), and the [Statistics and Probability Commons](#)

Recommended Citation

Wei, Hongji, "Beyond Classical Statistics: Optimality In Transfer Learning And Distributed Learning" (2022). *Publicly Accessible Penn Dissertations*. 5519.
<https://repository.upenn.edu/edissertations/5519>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/5519>
For more information, please contact repository@pobox.upenn.edu.

Beyond Classical Statistics: Optimality In Transfer Learning And Distributed Learning

Abstract

During modern statistical learning practice, statisticians are dealing with increasingly huge, complicated and structured data sets. New opportunities can be found during the learning process with better structured data sets as well as powerful data analytic resources. Also, there are more and more challenges we need to address when dealing with large data sets, due to limitation of computation, communication resources or privacy concerns. Under decision-theoretical framework, statistical optimality should be reconsidered with new type of data or new constraints. Under the framework of minimax theory, this thesis aims to address the following four problems: 1. The first part of this thesis aims to develop an optimality theory for transfer learning for nonparametric classification. An near optimal adaptive classifier is also established. 2. In the second part, we study distributed Gaussian mean estimation with known variance under communication constraints. The exact distributed minimax rate of convergence is derived under three different communication protocols. 3. In the third part, we study distributed Gaussian mean estimation with unknown variance under communication constraints. The results show that the amount of additional communication cost depends on the type of underlying communication protocol. 4. In the fourth part, we investigate the minimax optimality and communication cost of adaptation for distributed nonparametric function estimation under communication constraints.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Statistics

First Advisor

Tony Cai

Keywords

Big Data, Distributed Learning, Machine Learning, Minimax Theory, Transfer Learning

Subject Categories

Computer Sciences | Mathematics | Statistics and Probability

BEYOND CLASSICAL STATISTICS:
OPTIMALITY IN TRANSFER LEARNING AND DISTRIBUTED LEARNING

Hongji Wei

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

T. Tony Cai, Daniel H. Silberberg Professor; Professor of Statistics

Graduate Group Chairperson

Nancy R. Zhang, Ge Li and Ning Zhao Professor; Professor of Statistics

Dissertation Committee

T. Tony Cai, Daniel H. Silberberg Professor; Professor of Statistics

Zongming Ma, Associate Professor of Statistics

Yuxin Chen, Associate Professor of Statistics

BEYOND CLASSICAL STATISTICS:

OPTIMALITY IN TRANSFER LEARNING AND DISTRIBUTED LEARNING

COPYRIGHT

2022

Hongji Wei

ACKNOWLEDGEMENT

First and foremost, I would like to thank my respectful advisor, Tony Cai. Through the past five years, Tony introduced me into a colorful world of theoretical statistics. His guidance and support in every stage of my growth made me become an independent researcher. Tony has broad interest and cutting edge insights in a wide range of statistical problems. I have never had difficulty finding great and interesting problems when I work with him. Tony is also a wonderful mentor. From proving technical theorems, writing papers, to making presentations, he is always very patient and supportive, dedicating so much time teaching me. Tony is a role model. I can always feel his passion for statistics, which becomes an endless source of motivation for me to become a good scholar like him. Tony is a kind friend. He shares interesting news to me. He organizes Thanksgiving parties with students. Those are special and precious memories for me.

I am also deeply thankful to the other members of my dissertation committee, Yuxin Chen and Zongming Ma. Yuxin is definitely an expert in my research area. I am very fortunate to have him in the committee, and appreciate his acceptance of my invitation even before he came to the Wharton Statistics Department. Zongming is not only a teacher but also a friend to us. I have learned a lot from his "killer" asymptotic statistics course. And I still remember those funny moments when we tricked him by drawing a picture of "Zongming fights with Van der Vaart" on his office door.

I feel so fortunate to have been part of the Wharton Statistics Department family. I would like to thank the faculty of the Statistics Department, particularly Weijie Su, Edgar Dobriban, Mark Low, Richard Waterman for their help and guidance in my research, teaching and life. Also, special thanks go to Richard E. Blahut from Electrical & Computer Engineering Department. His information theory class taught me a sharp weapon, without which I could never gain as many fruitful results in my research.

My peers in the department have been a stable source for new ideas and fun. I am thankful

to Linjun Zhang, Yichen Wang, and Hongyang Zhang. They are my senior fellows who have given me a lot of help and inspirations. I also would like to thank my office mates Bo Zhang and Jeff Cai, and other fellow students including Ruijia Wu, Ran Chen, Sheng Gao, Hongming Pu, Hua Wang, Shuxiao Chen, and Mauricio Daros Andrade for their friendship.

Finally, I am deeply grateful to my family. My parents and my grandmother, although they are thousands miles away from me, are always supporting me and encouraging me. I won't be where I am today without their love. Last but not least, I realized it is hard for me to find the words to sufficiently express my gratitude to my dear wife, Tianjiao Wang, for her understanding, caring, dedication and love. She colored my Ph.D life.

This thesis is dedicated to my advisor, my professors, my friends, and my family.

ABSTRACT

BEYOND CLASSICAL STATISTICS:

OPTIMALITY IN TRANSFER LEARNING AND DISTRIBUTED LEARNING

Hongji Wei

T. Tony Cai

During modern statistical learning practice, statisticians are dealing with increasingly huge, complicated and structured data sets. New opportunities can be found during the learning process with better structured data sets as well as powerful data analytic resources. Also, there are more and more challenges we need to address when dealing with large data sets, due to limitation of computation, communication resources or privacy concerns. Under decision-theoretical framework, statistical optimality should be reconsidered with new type of data or new constraints. Under the framework of minimax theory, this thesis aims to address the following four problems:

1. The first part of this thesis aims to develop an optimality theory for transfer learning for nonparametric classification. An near optimal adaptive classifier is also established.
2. In the second part, we study distributed Gaussian mean estimation with known variance under communication constraints. The exact distributed minimax rate of convergence is derived under three different communication protocols.
3. In the third part, we study distributed Gaussian mean estimation with unknown variance under communication constraints. The results show that the amount of additional communication cost depends on the type of underlying communication protocol.
4. In the fourth part, we investigate the minimax optimality and communication cost of adaptation for distributed nonparametric function estimation under communication constraints.

TABLE OF CONTENTS

ACKNOWLEDGEMENT	iii
ABSTRACT	v
LIST OF TABLES	ix
LIST OF ILLUSTRATIONS	x
CHAPTER 1 : INTRODUCTION	1
1.1 Transfer Learning for Nonparametric Classification	1
1.2 Distributed Gaussian Mean Estimation with Known Variance Under Communication Constraints	2
1.3 Distributed Gaussian Mean Estimation with Unknown Variance Under Communication Constraints	3
1.4 Distributed Nonparametric Function Estimation Under Communication Constraints	3
CHAPTER 2 : Transfer Learning for Nonparametric Classification	5
2.1 Introduction	5
2.2 Problem Formulation	11
2.3 Minimax Rate of Convergence	17
2.4 Data-driven Adaptive Classifier	22
2.5 Multiple Source Distributions	28
2.6 Numerical Studies	36
2.7 Application to Crowdsourced Mapping Data	40
2.8 Discussion	42
2.9 Proofs	44

CHAPTER 3 : Distributed Gaussian Mean Estimation with Known Variance Under Communication Constraints	50
3.1 Introduction	50
3.2 Distributed Univariate Gaussian Mean Estimation	60
3.3 Distributed Multivariate Gaussian Mean Estimation	75
3.4 Optimal Distributed Estimation with Sequential and Blackboard Protocols . .	80
3.5 Robustness Against Departures from Gaussianity	82
3.6 Simulation Studies	84
3.7 Discussion	87
3.8 Proofs	89
CHAPTER 4 : Distributed Gaussian Mean Estimation with Unknown Variance Un- der Communication Constraints	94
4.1 Introduction	94
4.2 Problem Formulation	100
4.3 Optimal Adaptive Estimation under the independent protocols	102
4.4 Optimal Adaptive Estimate under Interactive Protocols	108
4.5 Numerical Results	113
4.6 Discussion	114
4.7 Proofs	115
CHAPTER 5 : Distributed Nonparametric Function Estimation Under Communica- tion Constraints	139
5.1 Introduction	139
5.2 Minimax Optimal Rate of Convergence	147
5.3 Adaptive Gaussian sequence estimation	156
5.4 Numerical Studies	168
5.5 Discussion	170
5.6 Proofs	172

APPENDIX 184

BIBLIOGRAPHY 185

LIST OF TABLES

TABLE 4.1	Optimal communication cost for different distributed protocols under adaptive and non-adaptive settings. Adaptive setting: minimal communication cost for rate-optimal adaptive estimator over $\sigma \geq \sigma_0$. Non-adaptive setting: minimal communication cost for rate-optimal estimator with known $\sigma = \sigma_0$	112
TABLE 4.2	MSEs and the communication costs of the three methods. $\sigma_0 = 2^{-12}$, $m = 100$, $\theta = 0.3$. For the two distributed estimators, total communication costs (in bits) are given in the parentheses.	113

LIST OF ILLUSTRATIONS

FIGURE 2.1	Illustration of the relative signal exponent γ . Left panel: feasible region when $\gamma = 0.5$ and $C_\gamma = 0.5$. A pair of distributions (P, Q) has relative signal exponent $\gamma = 0.5$ with $C_\gamma = 0.5$ when $(\eta_P(x), \eta_Q(x))$ falls into the shaded (blue) region for all x in the support. Right panel: feasible region with different choices of γ . Smaller γ implies more information contains in $P_{Y X}$	14
FIGURE 2.2	An illustration of the two-sample weighted K -NN classifier. (X^P, Y^P) are shown by the blue points and (X^Q, Y^Q) are shown by the red points. For each point in the graph, the coordinates represent its two-dimensional covariates X while the number marked inside the point represents its label Y . To classify the black point (x) located in middle of the graph, by calculation we get (say) $k_P = 2$ and $k_Q = 4$. Then we find k_P nearest neighbors from P -data and k_Q nearest neighbors from Q -data. Finally, we calculate their weighted mean to make the final classification.	20
FIGURE 2.3	An illustration of the adaptive procedure given in Algorithm 1. See figure 2.2 for interpretation of the graph. Here we shorthand the threshold $T = (d + 3) \log(n_P + n_Q)$. In each step, we evaluate $r^{(k)}$ and compare it to the threshold R . If $r^{(k)} > T$, then output $\hat{f}^{(k)}$ generated in current step; if $r^{(k)} \leq T$, go to next step and add one more nearest neighbor.	24

FIGURE 2.4	Left: Experiments on non-adaptive methods. We operate the naive K -NN method on only Q -data (dashed line) and our two-sample weighted K -NN classifier on different datasets. The datasets are generated with relative signal exponent $\gamma = 0.7, 0.5, 0.35$ respectively. Right: based on our theory (Theorem 1), the expected ratio of excess risk between the two methods we operate in the experiment.	37
FIGURE 2.5	Left: Experiments on adaptive methods. We operate the naive Lepski method on only Q -data (dashed line) and our adaptive classifier on different datasets. The datasets are generated with relative signal exponent $\gamma = 0.7, 0.5, 0.35$ respectively. Right: based on our theory (Theorem 3), the expected ratio of excess risk between the two methods used in the experiment.	38
FIGURE 2.6	Left: Experiments on transfer learning from multiple source distributions. We apply the naive Lepski method on only Q -data (dashed line) and our adaptive classifier for multiple source distributions. Right: based on our theory (Theorem 5), the expected ratio of excess risk between the two methods we operate in the experiment.	40
FIGURE 2.7	(a) Illustration of the dataset. Each row represents one of a land cover class (farm or forest) and corresponding NDVI values of a pixel from remotely-sensed imagery in 2014-2015. (b) Accuracy of the three methods on the crowdsourced mapping data with different numbers of crowdsourced data involved. Blue: The proposed adaptive classifier. Red: Lepski's method using combined data. Brown: Lepski's method using only crowdsourced data.	41

FIGURE 3.1	(a) Left panel: An illustration of a distributed learning network. Communication between the data servers and the central learner is necessary in order to learn from distributed datasets. (b) Right panel: An illustration of independent distributed protocol. The i -th machine can only transmit a b_i bits transcript to the central machine. The transcript Z_i only depends on observations \tilde{X}_i	53
FIGURE 3.2	The minimax rate of univariate Gaussian mean estimation under communication constraints has 3 phases: localization, refinement and optimal-rate.	55
FIGURE 3.3	An illustration of the Gray functions and Gray codes.	63
FIGURE 3.4	An illustration of MODGAME. The bits in the transcripts are transmitted to the central machine and divided into three types: crude localization bits, finer localization bits, and refinement bits. One then constructs on the central machine a crude interval I_1 , a finer interval I_2 , and the final estimate $\hat{\theta}_D$ step by step.	67
FIGURE 3.5	An illustration for multi-MODGAME. Communication budgets are evenly divided into three parts with each part used for estimating a coordinate of θ by the MODGAME procedure.	77
FIGURE 3.6	Comparisons of the MSEs of MODGAME (red), naive quantization (blue) and sample mean (black). MSEs are plotted on log-scale. In 3.6b and 3.6c, m and σ_n are plotted on log-scale.	85
FIGURE 3.7	Left panel: Comparisons of the MSEs of MODGAME with equal assignment (red), MODGAME with unequal assignment (blue) and sample mean (black). Right panel: Comparisons of the MSEs of multi-MODGAME (red) and sample mean (black). MSEs are plotted on log-scale.	87

FIGURE 5.1	<p>Estimate given by the optimal seq-MODGAME estimator $\hat{\theta}^O$ under the communication constraints. For different choices of total communication budgets $B = 100, 2400, 16000$, we illustrate an example of estimated function \hat{f}^O in each figure. The mean squared error through 1000 trials are also given below each figure.</p>	169
FIGURE 5.2	<p>Estimate given by the local thresholding estimator $\hat{\theta}^A$. Under different choices of ground truth functions f_1, f_2, f_3, we illustrate an example of estimated function \hat{f}^A in each figure. The expected communication cost and their mean squared error through 1000 trials are also given below each figure.</p>	170

CHAPTER 1

INTRODUCTION

During modern statistical learning practice, statisticians are dealing with increasingly huge, complicated and structured data sets. We are stepping into so-called the era of big data. New opportunities can be found during the learning process with better structured data sets as well as powerful data analytic resources. At the mean time, on the other hand, there are more and more challenges we need to address when dealing with large data sets, due to limitation of computation, communication resources or privacy concerns. Under decision-theoretical framework, statistical optimality should be reconsidered with new type of data or new constraints.

This thesis focuses on developing data-driven machine learning algorithms with theoretical guarantees, either to seize more opportunities or to overcome challenges during the modern statistical practice. This thesis consists of the following four chapters.

1.1. Transfer Learning for Nonparametric Classification

Human learners have the natural ability to use knowledge gained in one setting for learning in a different but related setting. This ability to transfer knowledge from one task to another is essential for effective learning. In the first chapter, we study transfer learning in the context of nonparametric classification based on observations from different distributions under the posterior drift model, which is a general framework and arises in many practical problems.

We first establish the minimax rate of convergence and construct a rate-optimal two-sample weighted K -NN classifier. The results characterize precisely the contribution of the observations from the source distribution to the classification task under the target distribution. A data-driven adaptive classifier is then proposed and is shown to simultaneously attain within a logarithmic factor of the optimal rate over a large collection of parameter spaces. Simulation studies and real data applications are carried out where the numerical results further illustrate the theoretical analysis. Extensions to the case of multiple source distributions

are also considered.

This chapter is based on Cai and Wei (2021c), joint work with Tony Cai.

1.2. Distributed Gaussian Mean Estimation with Known Variance Under Communication Constraints

In the conventional statistical decision theoretical framework, the focus is on the centralized setting where all the data are collected together and directly available. The main goal is to develop optimal (estimation, testing, detection, ...) procedures, where optimality is understood with respect to the sample size and parameter space. Communication/computational costs are not part of the consideration.

In the age of big data, communication/computational concerns associated with a statistical procedure are becoming increasingly important in contemporary applications. One of the difficulties for analyzing large datasets is that data are distributed, instead of in a single centralized location.

In this chapter, we study distributed estimation of a Gaussian mean under communication constraints in a decision theoretical framework. Minimax rates of convergence, which characterize the tradeoff between the communication costs and statistical accuracy, are established in both the univariate and multivariate settings. Communication-efficient and statistically optimal procedures are developed. In the univariate case, the optimal rate depends only on the total communication budget, so long as each local machine has at least one bit. However, in the multivariate case, the minimax rate depends on the specific allocations of the communication budgets among the local machines.

Although optimal estimation of a Gaussian mean is relatively simple in the conventional setting, it is quite involved under the communication constraints, both in terms of the optimal procedure design and lower bound argument. The techniques developed in this chapter can be of independent interest. An essential step is the decomposition of the minimax estimation problem into two stages, localization and refinement. This critical decomposition

provides a framework for both the lower bound analysis and optimal procedure design.

This chapter is based on Cai and Wei (2020c), joint work with Tony Cai.

1.3. Distributed Gaussian Mean Estimation with Unknown Variance Under Communication Constraints

In this chapter, we further extend the study in the previous chapter to an adaptive setting. Distributed estimation of a Gaussian mean with unknown variance under communication constraints is studied. Necessary and sufficient communication costs under different types of distributed protocols are derived for any estimator that is adaptively rate-optimal over a range of possible values for the variance. Communication-efficient and statistically optimal procedures are developed.

The analysis reveals an interesting and important distinction among different types of distributed protocols: compared to the independent protocols, interactive protocols such as the sequential and blackboard protocols require less communication costs for rate-optimal adaptive Gaussian mean estimation. The lower bound techniques developed in the present paper are novel and can be of independent interest.

This chapter is based on Cai and Wei (2021d), joint work with Tony Cai.

1.4. Distributed Nonparametric Function Estimation Under Communication Constraints

In this chapter, distributed minimax estimation and distributed adaptive estimation under communication constraints for Gaussian sequence model and white noise model are studied. The minimax rate of convergence for distributed estimation over a given Besov class, which serves as a benchmark for the cost of adaptation, is established. We then quantify the exact communication cost for adaptation and construct an optimally adaptive procedure for distributed estimation over a range of Besov classes.

The results demonstrate significant differences between nonparametric function estimation

in the distributed setting and the conventional centralized setting. For global estimation, adaptation in general cannot be achieved for free in the distributed setting. The new technical tools to obtain the exact characterization for the cost of adaptation can be of independent interest.

This chapter is based on Cai and Wei (2021b), joint work with Tony Cai.

CHAPTER 2

TRANSFER LEARNING FOR NONPARAMETRIC CLASSIFICATION

2.1. Introduction

A key feature of intelligence is the ability to learn from experience. Human learners appear to have the talent to transfer their knowledge gained from one task to another similar but different task. However, in statistical learning, most procedures are designed to solve one single task, or to learn one single distribution based on observations from the same setting. In a wide range of real-world applications, it is important to gain improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned. Transfer learning aims to tackle such a problem. It has attracted increasing attention in machine learning and has been used in many applications. Recent examples include computer vision (Tzeng et al., 2017; Gong et al., 2012), speech recognition (Huang et al., 2013), genre classification (Choi et al., 2017) and also many newly designed algorithms such as Yao and Doretto (2010); Lee et al. (2007). More details about transfer learning can be found in the survey papers (Pan and Yang, 2010; Weiss et al., 2016).

Besides significant successes in applications, much recent focus has also been on the theoretical properties of transfer learning. In many practical situations, there are labeled data available from a distribution P , called the source distribution, while a relatively small quantity of labeled or unlabeled data is drawn from a distribution Q , called the target distribution. They are different but to some extent related distributions. The goal is to make statistical inference under Q . A natural question is: How much information can be transferred from the source distribution P to the target distribution Q , provided a certain level of similarity between the two distributions?

This is quite a general and challenging question. The problem is also known as domain adaptation in the binary classification setting. In domain adaptation, data pairs (X, Y) are drawn from P and Q defined on $\mathbb{R}^d \times \{0, 1\}$. Data from the source distribution P can

be informative about the target distribution Q if the two distributions are similar. Several type of assumptions have been proposed and studied previously in the literature, such as divergence bounds, covariate shift, and posterior drift. The first line of work in the literature measures the similarity by the divergence between P and Q . Generalization bounds are derived on unlabeled testing data from the target distribution Q after training by the data from the source distribution P (Ben-David et al., 2007; Blitzer et al., 2008; Mansour et al., 2009). These bounds are general and can be applied to any two distributions, but for more structured source and target distributions those bounds are not suitable. Another line of work imposes some structural assumptions on P and Q such as covariate shift and posterior drift. Covariate shift assumes that the conditional distributions of Y given X are the same under P and Q , i.e. $P_{Y|X} = Q_{Y|X}$, but the marginal distributions P_X and Q_X can be different. Such a setting typically arises when the same study/survey is carried out in different populations. For example, when constructing a classifier for a certain disease, source data may be generated from clinical studies, but the goal is to classify people drawn from the general public. The task becomes challenging due to the difference between the two populations. Transfer learning under covariate shift has been studied in previous work such as Shimodaira (2000); Sugiyama et al. (2008); Kpotufe and Martinet (2018).

In this chapter, we study transfer learning under the posterior drift model, where it is assumed that $P_X \approx Q_X$ but $P_{Y|X}$ and $Q_{Y|X}$ can highly differ. To be more specific, suppose there are two data generating distributions P and Q on $\Omega \times \{0, 1\}$, where $\Omega \subset [0, 1]^d$. We observe n_P independent and identically distributed (i.i.d.) samples $(X_1^P, Y_1^P), \dots, (X_{n_P}^P, Y_{n_P}^P)$ drawn from a source distribution P , and n_Q i.i.d. samples $(X_1^Q, Y_1^Q), \dots, (X_{n_Q}^Q, Y_{n_Q}^Q)$ drawn from a target distribution Q . The data points from the distributions P and Q are also mutually independent. For each data point (X, Y) , the d -dimensional vector X is regarded as covariates (features) of a certain object, while Y is a (noisy) binary label indicating to which of the two classes this object belongs. The goal is to make classification under the target distribution Q : Given the observed data, construct a classifier $\hat{f} : \Omega \rightarrow \{0, 1\}$ which

minimizes the classification risk under the target distribution Q :

$$R(\hat{f}) \triangleq \mathbb{P}_{(X,Y) \sim Q}(Y \neq \hat{f}(X)).$$

Here $\mathbb{P}_{(X,Y) \sim Q}(\cdot)$ means the probability under the distribution Q .

In binary classification, the regression functions are defined as

$$\eta_P(x) \triangleq P(Y = 1|X = x) \quad \text{and} \quad \eta_Q(x) \triangleq Q(Y = 1|X = x),$$

which can be used to represent the conditional distributions $P_{Y|X}$ and $Q_{Y|X}$. In classification, Y can be regarded as an unknown parameter predicted by X , so from this perspective we refer to P_X and Q_X as the class ‘‘prior’’ probabilities and $\eta_P(x)$ and $\eta_Q(x)$ as the class ‘‘posterior’’ probabilities associated with P and Q respectively (Scott, 2018). We say a ‘‘posterior drift’’ happens when P_X and Q_X have the same support with bounded densities, but $\eta_P(x)$ and $\eta_Q(x)$ are highly different.

Posterior drift is a general framework and arises in many applications, where one collects data from different populations. Here are three examples.

- **Crowdsourcing.** Crowdsourcing is a distributed model for large-scale problem-solving and experimentation such as image classification, video annotation, and translation (Yuen et al., 2011; Karger et al., 2011; Zhang et al., 2014). The tasks are broadcasted to multiple independent workers online in order to collect and aggregate their solutions. In crowdsourcing, many noisy answers/labels are available from a large amount of public workers, while sometimes, more accurate answers/labels may be collected from experienced workers or experts. These expert answers/labels are of higher quality but are relatively few due to the time or budget constraints. One can view this difference in labeling accuracy as a posterior drift. It is desirable to construct a statistical procedure that incorporates both data sets.

- **Concept drift.** Concept drift is a common phenomenon when the underlying distribution of the data changes over time in a streaming environment (Tsymbal, 2004; Gama et al., 2014). One kind of concept drift is called real concept drift where the posterior class probabilities $P(Y|X)$ changes over time. In this situation, posterior drift exists if data are collected at different time. For example, the incidence rate of a certain disease in certain groups may change over time due to the development of treatments and preventive measures.
- **Data corruption.** Data corruption is ubiquitous in applications, where unexpected error on data occurs during storage, transmission or processing (Menon et al., 2015; van Rooyen and Williamson, 2018). In many settings, one receives data of variable quality – perhaps some small amount of clean data, another amount of slightly corrupted data, yet more that is significantly corrupted, and so on (Crammer et al., 2006). Data of variable qualities can be viewed as posterior drift between those data generating distributions, thus better strategies are needed to tackle the problem within the posterior drift framework.

Under the posterior drift model, the main difference between P and Q lies in the regression functions $\eta_P(x)$ and $\eta_Q(x)$. So the relationship between $\eta_P(x)$ and $\eta_Q(x)$, which can be captured by the link function ϕ defined below, is important in characterizing the difficulty of the transfer learning problem. In this work, we propose a new concept called the *relative signal exponent* γ to describe the relationship between $\eta_P(x)$ and $\eta_Q(x)$. Our results show that the relative signal exponent γ plays an important role in the minimax rate of convergence for the excess risk under the posterior drift model.

For conceptual simplicity, we assume $\eta_P(x) = \phi(\eta_Q(x))$ for some strictly increasing link function $\phi(\cdot)$ with $\phi(\frac{1}{2}) = \frac{1}{2}$. Note that this is only a simplified version of our formal model which will be given in Section 2. It is natural to assume ϕ is strictly increasing in the settings where those X that are more likely to be labeled $Y = 1$ under Q are also more likely to be labeled $Y = 1$ under P . The assumption $\phi(\frac{1}{2}) = \frac{1}{2}$ means that those X that are non-

informative under Q are the same under P . Formally, for a given relative signal exponent $\gamma > 0$ and a constant $C_\gamma > 0$, we denote by $\Gamma(\gamma, C_\gamma)$ the collection of all distribution pairs (P, Q) satisfying

$$(\phi(x) - \frac{1}{2})(x - \frac{1}{2}) \geq 0 \quad \text{and} \quad |\phi(x) - \frac{1}{2}| \geq C_\gamma |x - \frac{1}{2}|^\gamma. \quad (2.1)$$

The relative signal exponent is a key parameter in capturing the usefulness of the data from the source distribution P for the task of classification under the target distribution Q . The smaller the relative signal exponent, the more information transferable from P to Q .

In this work we consider transfer learning under the posterior drift model in a nonparametric classification setting. When Q satisfies the margin assumption with the parameter α , defined in Section 2.2, and $\eta_Q(x)$ belongs to the (β, C_β) -Hölder function class, it is shown that, under the regularity conditions, the minimax optimal rate of convergence is given by

$$\inf_{\hat{f}} \max_{(P, Q) \in \Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}) \asymp (n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta(1+\alpha)}{2\beta+d}}, \quad (2.2)$$

where n_P and n_Q are number of data drawn from P and Q respectively, d is the number of features, and Π is the posterior drift regime where the distribution pair (P, Q) belongs to the class $\Gamma(\gamma, C_\gamma)$ with the relative signal exponent γ and satisfies some additional regularity conditions. Here $\mathcal{E}_Q(\hat{f})$ is the excess risk on Q which is defined based on the misclassification error:

$$\mathcal{E}_Q(\hat{f}) = R_Q(\hat{f}) - R_Q(f_Q^*) \quad (2.3)$$

where

$$f_Q^*(x) = \begin{cases} 0 & \text{if } \eta_Q(x) \leq \frac{1}{2} \\ 1 & \text{otherwise} \end{cases} \quad (2.4)$$

is the Bayes classifier under Q . The expectation \mathbb{E}_Z in (2.2) is taken over the random

realizations of all the observed data, namely the set Z , defined as

$$Z \triangleq \{(X_1^P, Y_1^P), \dots, (X_{n_P}^P, Y_{n_P}^P), (X_1^Q, Y_1^Q), \dots, (X_{n_Q}^Q, Y_{n_Q}^Q)\}. \quad (2.5)$$

Note that if one only had observations from the target distribution Q , the minimax rate would be $n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}$. Therefore, the additional term $n_P^{\frac{2\beta+d}{2\gamma\beta+d}}$ in the minimax rate (2.2) quantifies an “effective sample size” for transfer learning from the source distribution P relative to Q , and $\frac{2\beta+d}{2\gamma\beta+d}$ can be viewed as the optimal transfer rate. This result answers one of the main questions in transfer learning: $n_P^{\frac{2\beta+d}{2\gamma\beta+d}}$ is the total amount of information that can be transferred from P to Q , and this quantity depends on the relative signal exponent γ which characterizes the discrepancy between P and Q in posterior drift.

We construct a two-sample weighted K -nearest neighbors (K -NN) classifier and show that it attains the optimal rate given in (2.2). However, this classifier depends on the parameters α , β , and γ , which are typically unknown in practice. In this chapter, we also propose a data-driven classifier \hat{f}_a that automatically adapts to the unknown model parameters α, β and γ , with an additional log term on the excess risk bound:

$$\sup_{(P,Q) \in \Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}_a) \lesssim \left(\left(\frac{n_P}{\log(n_P + n_Q)} \right)^{\frac{2\beta+d}{2\gamma\beta+d}} + \frac{n_Q}{\log(n_P + n_Q)} \right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

This adaptive procedure is essentially different from either the non-adaptive procedure given in this chapter, or any nonparametric classification procedures in the literature. The adaptive classifier is constructed based on the ideas inspired by Lepski’s method for nonparametric regression. The construction begins with a small number of the nearest neighbors, and gradually increases the number of the neighbors used to make the decision. The algorithm terminates once an empirical signal-to-noise ratio reaches a delicately designed threshold. It is shown that the resulting data-driven classifier automatically adapts to a wide collection of parameter spaces.

In some applications, there are data available from multiple source distributions. Intuitively, the samples from all source distributions are helpful to the classification task under the target distribution. We also consider transfer learning in this setting under the posterior drift model. Suppose there are multiple source distributions P_1, \dots, P_m and one target distribution Q , each pair of distributions (P_i, Q) has a relative signal exponent γ_i , $i \in \{1, \dots, m\}$. The minimax optimal rate of convergence is established and the result quantifies precisely the contributions from the data generated by the individual source distributions. An adaptive procedure is constructed and shown to simultaneously attain the optimal rate up to a logarithmic factor over a large class of parameter spaces.

The rest of the chapter is organized as follows. In Section 2.2, after some basic notations and definitions are introduced, the model for transfer learning under the posterior drift model is proposed in a nonparametric classification setting. In Section 2.3, we establish the minimax optimal rate by constructing a minimax optimal procedure with guaranteed upper bound and a matching lower bound. In section 2.4, a data-driven adaptive classifier is proposed and is shown to adaptively attain the optimal rate of convergence, up to a logarithmic factor. Section 2.6 investigates the numerical performance of the data driven procedure. In section 2.7, a real data application is carried out to further illustrate the benefit of our method. Section 2.5 considers transfer learning with multiple source distributions and a brief discussion is given in Section 2.8. For reasons of space, we prove one main result in Section 2.9 and provide the proofs of the other results and some technical lemmas in the Supplementary Material (Cai and Wei, 2019).

2.2. Problem Formulation

We introduce in this section the posterior drift model. We begin with notation and basic definitions.

2.2.1. Notation and definitions

For a distribution G , denote by $G(\cdot)$ and $\mathbb{E}_G(\cdot)$ respectively the probability and expectation under G . Denote by P_X and Q_X the marginal distribution of X under the joint distributions

P and Q for (X, Y) respectively. Let $\text{supp}(\cdot)$ denote the support of a probability distribution. Throughout the chapter we write $\|\cdot\|$ to denote the Euclidean norm. We use $\mathbb{I}_{\{\cdot\}}$ to denote the indicator function taking values in $\{0, 1\}$. We define $a \vee b = \max(a, b)$, $a \wedge b = \min(a, b)$, and $\lfloor a \rfloor$ be the maximum integer that is not larger than a . We denote by $B(x, r)$ a Euclidean ball centered at x with radius r . We write $\lambda(\cdot)$ to denote Lebesgue measure of a set in a Euclidean space. We denote by C or c some generic constants not depending on n_P or n_Q that may vary from place to place.

2.2.2. Posterior drift in nonparametric classification

For two distributions P and Q for a random pair (X, Y) taking values in $[0, 1]^d \times \{0, 1\}$, we observe two independent random samples, $(X_1^P, Y_1^P), \dots, (X_{n_P}^P, Y_{n_P}^P) \stackrel{\text{i.i.d.}}{\sim} P$ and $(X_1^Q, Y_1^Q), \dots, (X_{n_Q}^Q, Y_{n_Q}^Q) \stackrel{\text{i.i.d.}}{\sim} Q$. We shall use P -data and Q -data to refer to the data sets drawn from the distributions P and Q respectively. We consider the transfer learning problem when there is a posterior drift between P and Q . In the posterior drift model, the covariates/features X are drawn from distributions having the same support with bounded densities, but the response/label Y has different conditional distributions given X between P and Q . The readers should notice that the model we introduced in Section 2.1 is a special case within the model we will introduce in this section.

The regression functions have been defined informally in the introduction, now we give a precise definition. Let

$$\eta_P(x) = \begin{cases} P(Y = 1|X = x) & \text{if } x \in \text{supp}(P_X) \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

$$\eta_Q(x) = \begin{cases} Q(Y = 1|X = x) & \text{if } x \in \text{supp}(Q_X) \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

denote the corresponding regression functions of P and Q . Besides the previous definition (2.4) of Bayes classifier under the target distribution Q , we can similarly define the Bayes

classifier for the source distribution P as:

$$f_P^*(x) = \begin{cases} 0 & \text{if } \eta_P(x) \leq \frac{1}{2} \\ 1 & \text{otherwise} \end{cases} .$$

Now assume (X^P, Y^P) is a data pair drawn from the distribution P . From the definition, given $X^P = x$, Y^P is more likely to be equal to 1 if $f_P^*(x) = 1$ whereas Y^P is more likely to be equal to 0 if $f_P^*(x) = 0$. It is similar for the distribution Q . Thus informally one can regard $f_P^*(x)$ ($f_Q^*(x)$) as the true label at the covariate value x under the distribution P (Q).

In transfer learning, although the observed data are drawn from two or more different distributions, these distributions are usually related to each other so that all of them are useful for learning the intrinsic true labels. For instance, in a crowdsourcing survey, although accuracy varies among different workers, their answers should be no worse than random guessing. It is reasonable to assume that the answer is correct with probability at least $\frac{1}{2}$. This means we may reasonably assume that, given the same covariate x , the “true labels” under the distributions P and Q are the same. That is

$$f^*(x) \triangleq f_P^*(x) = f_Q^*(x) \quad \forall x \in \text{supp}(P_X),$$

which is equivalent to

$$(\eta_P(x) - \frac{1}{2})(\eta_Q(x) - \frac{1}{2}) \geq 0.$$

The definitions and assumptions introduced so far treat the P -data and Q -data symmetrically and interchangeably. But in real applications, usually the two data sets are treated differently. We call P the source distribution and Q the target distribution. The goal is to transfer the knowledge gained from the P -data together with the information contained in the Q -data for constructing an optimal classifier under the target distribution Q .

Intuitively it is clear that the amount of information that can be transferred from the P -data

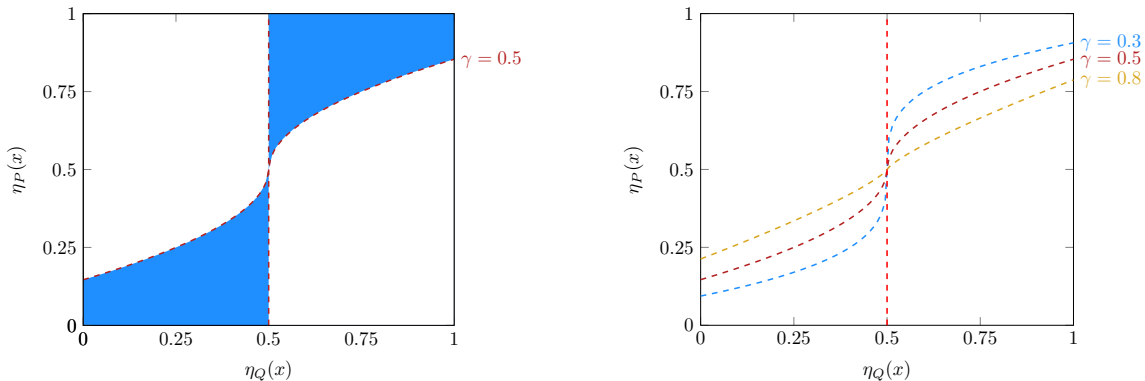


Figure 2.1: Illustration of the relative signal exponent γ . Left panel: feasible region when $\gamma = 0.5$ and $C_\gamma = 0.5$. A pair of distributions (P, Q) has relative signal exponent $\gamma = 0.5$ with $C_\gamma = 0.5$ when $(\eta_P(x), \eta_Q(x))$ falls into the shaded (blue) region for all x in the support. Right panel: feasible region with different choices of γ . Smaller γ implies more information contains in $P_{Y|X}$.

for the inference under Q depends on the similarity between the distributions P and Q . In this chapter, we quantify the similarity by the **relative signal exponent** of P with respect to Q .

Definition 1 (Relative Signal Exponent). The class $\Gamma(\gamma, C_\gamma)$ with relative signal exponent $\gamma \in (0, \infty)$ and a constant $C_\gamma \in (0, \infty)$ is defined as the set of distribution pairs (P, Q) , both supported on $\mathbb{R}^d \times \{0, 1\}$, satisfying $\forall x \in \text{supp}(P_X) \cup \text{supp}(Q_X)$,

$$(\eta_P(x) - \frac{1}{2})(\eta_Q(x) - \frac{1}{2}) \geq 0 \quad (2.6)$$

$$|\eta_P(x) - \frac{1}{2}| \geq C_\gamma |\eta_Q(x) - \frac{1}{2}|^\gamma. \quad (2.7)$$

Remark 1. The relative signal exponent γ indicates the signal strength of the P -data relative to the Q -data. Note that $|\eta_Q(x) - \frac{1}{2}|$ is always bounded by $1/2$. So generally speaking, the smaller γ is, the larger the difference between $\eta_P(x)$ and $\frac{1}{2}$, which means the P -data is more informative about $f^*(x)$ and consequently more information can be transferred from the P -data to the Q -data.

One can see that the above definition of relative signal exponent implies when $|\eta_Q(x) - \frac{1}{2}|$ is large, then $|\eta_P(x) - \frac{1}{2}|$ should be relatively large. This is intuitively true in a wide range of real applications. Taking again the crowdsourcing surveys as an example. If one crowd of workers can answer a question correctly with a larger probability, then for another crowd of workers the accuracy of their answers is also usually larger because this question is likely to be easier.

In addition to the relative signal exponent γ , we also need to define a smoothness parameter of η_Q and characterize its behavior near $1/2$:

Definition 2 (Smoothness). The (β, C_β) -Hölder class of functions ($0 < \beta \leq 1$), denoted by $\mathcal{H}(\beta, C_\beta)$, is defined as the set of functions $g : \mathbb{R}^d \rightarrow \mathbb{R}$ satisfying, for any $x_1, x_2 \in \mathbb{R}^d$,

$$|g(x_1) - g(x_2)| \leq C_\beta \|x_1 - x_2\|^\beta.$$

Definition 3 (Margin Assumption). The margin class $\mathcal{M}(\alpha, C_\alpha)$ with $\alpha \geq 0, C_\alpha > 0$ is defined as the set of distributions Q satisfying

$$Q_X(|\eta_Q(X) - \frac{1}{2}| < t) \leq C_\alpha t^\alpha.$$

In this chapter we consider the nonparametric classification problem when $\eta_Q(x)$ belongs to a (β, C_β) -Hölder class and Q belongs to a margin class $\mathcal{M}(\alpha, C_\alpha)$. When $Q \in \mathcal{M}(\alpha, C_\alpha)$, we also say that Q satisfies the margin assumption with the parameter α .

Remark 2. In the main part of our discussion, we focus on the case with $0 < \beta \leq 1$, i.e. η belongs to a Hölder function class with smoothness less than or equal to 1. Generally it is possible to consider more general classes where the smoothness parameter can be larger than 1. The discussion on the model and methods associated with the general smoothness parameter $\beta > 1$ will be deferred to the discussion section.

The margin assumption was first introduced in Tsybakov (2004); Audibert and Tsybakov (2007) to characterize the convergence rate in nonparametric classification. The margin assumption put a constraint on the mass around $\eta_Q(x) \approx \frac{1}{2}$ so that with large probability $\eta_Q(x)$ is either $\frac{1}{2}$ or far from $\frac{1}{2}$. Generally, if an underlying distribution satisfies the margin assumption, then a more accurate classification can be guaranteed.

Another definition is about density constraints on the marginal distributions P_X and Q_X .

Definition 4 (Common Support and Strong Density Assumption). (P_X, Q_X) is said to have common support and satisfy strong density assumption with parameter $\mu = (\mu_-, \mu_+)$, $c_\mu > 0$, $r_\mu > 0$ if both P_X and Q_X are absolutely continuous with respect to the Lebesgue measure on \mathbb{R}^d , and

$$\Omega \triangleq \text{supp}(P_X) = \text{supp}(Q_X)$$

$$\lambda[\Omega \cap B(x, r)] \geq c_\mu \lambda[B(x, r)] \quad \forall 0 < r \leq r_\mu, \forall x \in \Omega$$

$$\mu_- < \frac{dP_X}{d\lambda}(x) < \mu_+ \quad \forall x \in \Omega$$

$$\mu_- < \frac{dQ_X}{d\lambda}(x) < \mu_+ \quad \forall x \in \Omega.$$

Define $\mathcal{S}(\mu, c_\mu, r_\mu)$ to be the set of the marginal densities pairs (P_X, Q_X) that have common support and satisfy the strong density assumption with parameter μ, c_μ, r_μ .

Remark 3. The strong density assumption was first introduced in Audibert and Tsybakov (2007). In this chapter we focus on the scenario that the marginal densities of P_X and Q_X have regular support and are bounded from below and above on the support.

Moreover, note that when Q_X satisfies the strong density assumption, in the regime $\alpha\beta > d$, there is no distribution Q such that the regression function η_Q crosses $\frac{1}{2}$ in the interior of the support Q_X (Audibert and Tsybakov, 2007). So this regime only contains the trivial cases for classification. Therefore, we further assume $\alpha\beta \leq d$ in the following discussion.

Given a classifier $\hat{f} : \mathbb{R}^d \rightarrow \{0, 1\}$, the excess risk on Q of the classifier \hat{f} , defined in equation (2.3), has a dual representation (Gyorfi, 1978)

$$\mathcal{E}_Q(\hat{f}) = 2\mathbb{E}_{(X,Y) \sim Q}(|\eta_Q(X) - \frac{1}{2}\mathbb{I}_{\{\hat{f}(X) \neq f_Q^*(X)\}}|). \quad (2.8)$$

A major goal in transfer learning is to construct an empirical decision rule \hat{f} incorporating both the P -data and Q -data, so that the excess risk on Q is minimized. It is interesting to understand when the minimax rate in the transfer learning setting is faster than the optimal rate where only the Q -data is used to construct the decision rule.

Putting the above definitions together, in this chapter we consider the posterior drift non-parametric parameter space:

$$\begin{aligned} \Pi(\alpha, C_\alpha, \beta, C_\beta, \gamma, C_\gamma, \mu, c_\mu, r_\mu) = \{ & (P, Q) : (P, Q) \in \Gamma(\gamma, C_\gamma), Q \in \mathcal{M}(\alpha, C_\alpha), \\ & \eta_Q \in \mathcal{H}(\beta, C_\beta), (P_X, Q_X) \in \mathcal{S}(\mu, c_\mu, r_\mu)\}. \end{aligned}$$

In the rest of this chapter, we will use the shorthand $\Pi(\alpha, \beta, \gamma, \mu)$ or Π if there is no confusion. The space $\Pi(\alpha, \beta, \gamma, \mu)$ is also called the posterior drift regime with $(\alpha, \beta, \gamma, \mu)$.

2.3. Minimax Rate of Convergence

In this section, we establish the minimax rate of convergence for the excess risk on Q for transfer learning under the posterior drift model and propose an optimal procedure using the two-sample weighted K -NN classifier.

The K -NN method has attracted much attention (Cover and Hart, 1967; Gyorf, 1978; Gadat et al., 2016) due to its massive practical success and appealing theoretical properties. In the conventional setting where one only has access to the Q -data and there is no P -data, with a suitable choice of the neighborhood size k , the K -NN classifier can achieve the minimax rate of convergence for the excess risk on Q (Gadat et al., 2016). The K -NN classifier is generated in two steps:

Step 1: For any given x to be classified, one can estimate $\eta_Q(x)$ by taking the empirical mean of the response variables (Y) according to its k nearest covariates (X). Formally, define $X_{(i)}^Q(x)$ be the i -th nearest covariates to x among $X_1^Q, \dots, X_{n_Q}^Q$ and $Y_{(i)}^Q(x)$ is its corresponding response (label). The estimate $\hat{\eta}_Q(x)$ is given by

$$\hat{\eta}_Q(x) = \frac{1}{k} \sum_{i=1}^k Y_{(i)}^Q(x).$$

Step 2: The class label for x is estimated by the plug-in rule:

$$\hat{f}(x) = \mathbb{I}_{\{\hat{\eta}_Q(x) > \frac{1}{2}\}}.$$

In transfer learning, one also has access to the P -data in addition to the Q -data, the P -data can be used to help the classification task under the target distribution Q and should be taken into consideration. To accommodate the existing K -NN methods, we should take the empirical mean of not only the k -nearest response variables from the Q -data, but also some nearest response variables from the P -data. In addition, when taking the average, data from the different distributions should have different weights because the signal strength varies between the two distributions. To make the classification at $x \in [0, 1]^d$, a new strategy called the two-sample weighted K -NN classifier is summarized as follows:

Step 1: Define $X_{(i)}^P(x)$ to be the i -th nearest covariates to x among $X_1^P, \dots, X_{n_P}^P$ and $Y_{(i)}^P(x)$ is its corresponding response. $X_{(i)}^Q(x)$ and $Y_{(i)}^Q(x)$ can be defined likewise. Construct the two-sample weighted K -NN estimator

$$\hat{\eta}_{NN}(x) = \frac{w_P \sum_{i=1}^{k_P} Y_{(i)}^P(x) + w_Q \sum_{i=1}^{k_Q} Y_{(i)}^Q(x)}{w_P k_P + w_Q k_Q}$$

where the number of neighbors k_P and k_Q and the weights w_P and w_Q will be specified later.

Step 2: The class label for x is estimated by the plug-in rule:

$$\hat{f}_{NN}(x) = \mathbb{I}_{\{\hat{\eta}_{NN}(x) > \frac{1}{2}\}}.$$

The final decision rule $\hat{f}_{NN}(x)$, which is generated by both the P -data and Q -data, is called the *two-sample weighted K -NN classifier*.

The performance of the two-sample weighted K -NN classifier $\hat{f}_{NN}(x)$ clearly depends on the choice of (k_P, k_Q, w_P, w_Q) . The next theorem gives a set of choices of (k_P, k_Q, w_P, w_Q) and a provable upper bound on the excess risk, which gives a guarantee for the performance of the two-sample weighted K -NN classifier with these specific choices of (k_P, k_Q, w_P, w_Q) .

Theorem 1 (Upper Bound). *Let \hat{f}_{NN} be the two-sample weighted K -NN classifier with $w_Q = (n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta}{2\beta+d}}$, $w_P = (n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\gamma\beta}{2\beta+d}}$, $k_Q = \lfloor n_Q(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{d}{2\beta+d}} \rfloor$, and $k_P = \lfloor n_P(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{d}{2\beta+d}} \rfloor$. Then*

$$\sup_{(P,Q) \in \Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}_{NN}) \leq C(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta(1+\alpha)}{2\beta+d}}$$

for some constant $C > 0$ not depending on n_P or n_Q .

The following lower bound result shows that the two-sample weighted K -NN classifier \hat{f}_{NN} given in Theorem 1 is in fact rate optimal.

Theorem 2 (Lower Bound). *There exists a constant $c > 0$ not depending on n_P or n_Q such that*

$$\inf_{\hat{f}} \sup_{(P,Q) \in \Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}) \geq c(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

The proof of Theorem 1 will be given in Section 2.9, which is based on the general techniques for proving K -NN methods, for instance, see Gadat et al. (2016); Samworth (2012). In the literature of classical nonparametric classification problem, the focus was mainly on bias-variance trade-off. Under posterior drift model, we further extend the general techniques

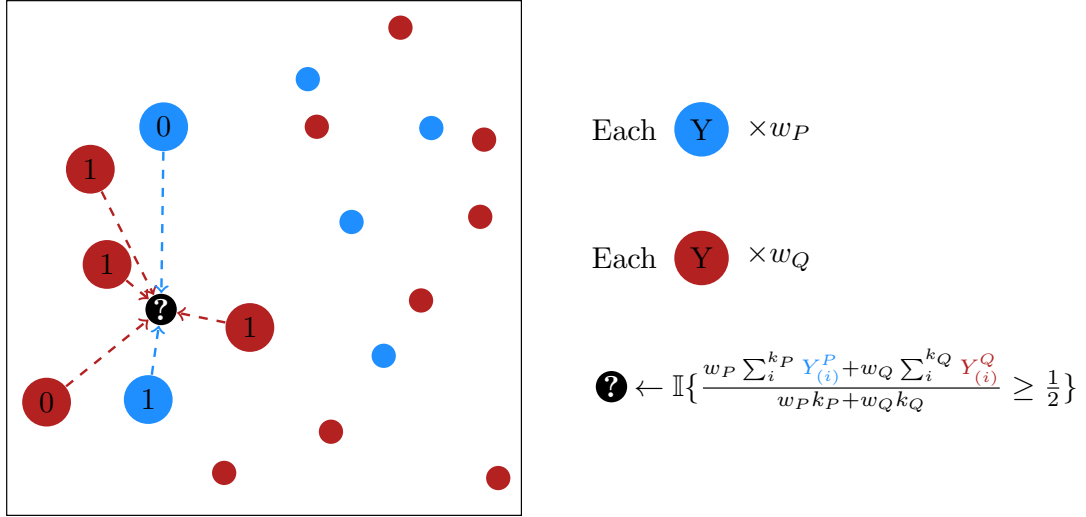


Figure 2.2: An illustration of the two-sample weighted K -NN classifier. (X^P, Y^P) are shown by the blue points and (X^Q, Y^Q) are shown by the red points. For each point in the graph, the coordinates represent its two-dimensional covariates X while the number marked inside the point represents its label Y . To classify the black point (x) located in middle of the graph, by calculation we get (say) $k_P = 2$ and $k_Q = 4$. Then we find k_P nearest neighbors from P -data and k_Q nearest neighbors from Q -data. Finally, we calculate their weighted mean to make the final classification.

to the two-sample setting, where the weights and the number of neighbors are carefully selected to make the best combination of information. The proof of Theorem 2 is given in the supplementary material (Cai and Wei, 2019), using the same general scheme as in (Audibert and Tsybakov, 2007; Kpotufe and Martinet, 2018). Theorems 1 and 2 together establish the minimax rate of convergence for transfer learning under the posterior drift model,

$$\inf_{\hat{f}} \sup_{(P,Q) \in \Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}) \asymp (n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta(1+\alpha)}{2\beta+d}}. \quad (2.9)$$

We make a few remarks on the minimax rate of convergence.

- Based on the minimax rate given in (2.9), it is easy to see that, in terms of the classification accuracy, the contribution from the P -data is substantial when $n_P^{\frac{2\beta+d}{2\gamma\beta+d}} \gg n_Q$, and the contribution is not significant otherwise.

- Comparing the convergence rates (2.9) with (2.10), the minimax rate for transfer learning under the posterior drift model is the same as if one had a sample of size $n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q$ from the distribution Q in the conventional setting. Therefore, one can intuitively view $n_P^{\frac{2\beta+d}{2\gamma\beta+d}}$ as the “effective sample size” of the P -data for the classification task under Q . The exponent $\frac{2\beta+d}{2\gamma\beta+d}$ here can be regarded as the *transfer rate*. The smaller the relative signal exponent γ is, the larger $\frac{2\beta+d}{2\gamma\beta+d}$ is, and more information is transferred from the P -data. This transfer rate provides a quantitative answer to the question posed in the introduction: How much information can be transferred from the source distribution P to the target distribution Q ? It is also interesting to note that, when $\gamma < 1$, $\frac{2\beta+d}{2\gamma\beta+d} > 1$, which implies that in this case an observation from P is more valuable than an observation from Q for the classification problem.
- In the transfer learning literature, much attention has been on an interesting special case where there is no data from the target distribution Q at all, i.e., $n_Q = 0$ (Mansour et al., 2009; Blitzer et al., 2008). This case arises when a classifier has been trained based on the data drawn from the source distribution P , and one wishes to generalize the classifier to unlabeled testing data drawn from the target distribution Q . Our results show that generalization is possible in the posterior drift framework and the optimal rate of convergence is

$$\inf_{\hat{f}} \sup_{(P,Q) \in \Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}) \asymp n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}}.$$

- It is worth noting that in the conventional setting with access to the Q -data only, the minimax rate, which is given in Audibert and Tsybakov (2007), would be

$$\inf_{\hat{f}} \sup_{(P,Q) \in \Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}) \asymp n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}, \quad (2.10)$$

which is a special case of (2.9) with $n_P = 0$. This rate can be achieved by the K -NN classifier given above with the choice of $k \asymp n_Q^{\frac{2\beta}{2\beta+d}}$.

2.4. Data-driven Adaptive Classifier

In the previous section, we have established the minimax optimal rate over the parameter space $\Pi(\alpha, \beta, \gamma, \mu)$ for transfer learning under the posterior drift model. This rate can be achieved by the two-sample weighted K -NN classifier given in Theorem 1. A major drawback of this classifier is that it requires the prior knowledge of β and γ , which is typically unavailable in practice. An interesting and practically important question is whether it is possible to construct a data-driven adaptive decision rules that can achieve the same rate of convergence, while automatically adapt to a wide collection of the parameter spaces $\Pi(\alpha, \beta, \gamma, \mu)$.

In nonparametric regression, Lepski's method (Lepski, 1991, 1992, 1993) is a well known approach for the construction of a data driven estimator that adapts to the unknown smoothness parameter β by screening from a small bandwidth to larger bandwidths with delicately designed stopping rules. This procedure can be used for nonparametric classification in the conventional setting where only Q -data is available and only adaptation to one smoothness parameter β is needed. For readers' convenience we include this construction in Section 2.9. The transfer learning setting is more challenging: we need to adapt to both parameters β and γ . In this section, we modify Lepski's method in our context and introduce a new stopping rule and show that the resulting classifier adapts to all unknown parameters.

Now we develop a data-driven procedure to make classification at a specific point $x \in [0, 1]^d$. The construction combines all data points from the P -data and the Q -data together and finds nearest neighbors among all the data. Denote by $X_{(i)}(x)$ the i -th nearest data point to x in the combined set $\{X_1^P, \dots, X_{n_P}^P\} \cup \{X_1^Q, \dots, X_{n_Q}^Q\}$. Similar to Lepski's method, we begin with a small number of nearest neighbors, and gradually increase the number of neighbors used to make the decision. One more nearest neighbor is added in each step. At the k -th step, there are k nearest neighbors $X_{(1)}(x), \dots, X_{(k)}(x)$ among all the points in the combined set $\{X_1^P, \dots, X_{n_P}^P\} \cup \{X_1^Q, \dots, X_{n_Q}^Q\}$. Suppose among these k nearest neighbors there are $k_P^{(k)}$ points from the P -data and $k_Q^{(k)}$ points from the Q -data. Heuristically, given these k nearest

neighbors, one can obtain a weighted K -NN estimate as

$$\hat{\eta}^{(k)}(x, w_P, w_Q) = \frac{w_P \sum_{i=1}^{k_P^{(k)}} Y_{(i)}^P(x) + w_Q \sum_{i=1}^{k_Q^{(k)}} Y_{(i)}^Q(x)}{w_P k_P^{(k)} + w_Q k_Q^{(k)}}.$$

If β and γ are known, one can calculate the optimal choice of the weights w_P and w_Q for a two-sample weighted K -NN classifier. To construct an adaptive procedure, we need to find a data driven method for choosing the weights w_P and w_Q . Define the ‘‘variance’’ of $\hat{\eta}^{(k)}(x, w_P, w_Q)$ as

$$v^{(k)}(w_P, w_Q) = \frac{w_P^2 k_P^{(k)} + w_Q^2 k_Q^{(k)}}{(w_P k_P^{(k)} + w_Q k_Q^{(k)})^2}.$$

For a given k , we call the maximum value of the ratio between $(\hat{\eta}^{(k)}(x, w_P, w_Q) - \frac{1}{2})^2$ and the ‘‘variance’’ $v^{(k)}(w_P, w_Q)$ as the signal-to-noise ratio index $\hat{r}^{(k)}$:

$$\hat{r}^{(k)} = \max_{w_P, w_Q} \frac{(\hat{\eta}^{(k)}(x, w_P, w_Q) - \frac{1}{2})^2}{v^{(k)}(w_P, w_Q)}.$$

The algorithm is terminated when $\hat{r}^{(k)} > (d+3) \log(n_P + n_Q)$, and the corresponding w_P and w_Q are chosen as the maximizers of $\frac{(\hat{\eta}^{(k)}(x, w_P, w_Q) - \frac{1}{2})^2}{v^{(k)}(w_P, w_Q)}$. If the algorithm doesn’t terminate at any step, the optimal k can be alternatively chosen by the maximizer of $\hat{r}^{(k)}$. That is, we choose $k = k^*$ with

$$k^* = \begin{cases} \min\{k : \hat{r}^{(k)} > (d+3) \log(n_P + n_Q)\} & \text{if } \max_k \hat{r}^{(k)} > (d+3) \log(n_P + n_Q) \\ \operatorname{argmax}_k \hat{r}^{(k)} & \text{otherwise} \end{cases} \quad (2.11)$$

and choose $(w_P, w_Q) = (w_P^*, w_Q^*)$ with

$$(w_P^*, w_Q^*) = \operatorname{argmax}_{(w_P, w_Q)} \frac{(\hat{\eta}^{(k^*)}(x, w_P, w_Q) - \frac{1}{2})^2}{v^{(k^*)}(w_P, w_Q)}.$$

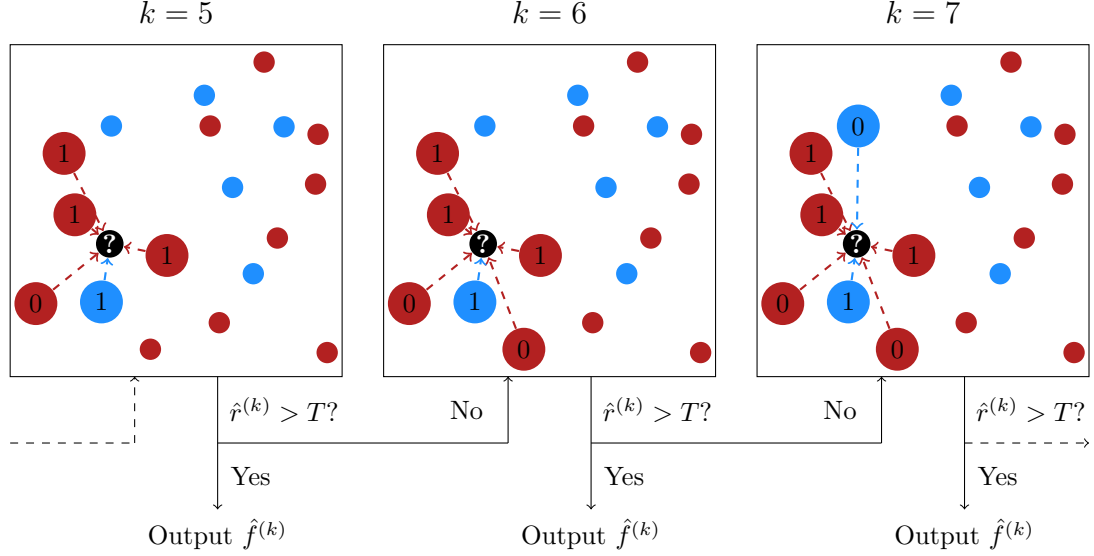


Figure 2.3: An illustration of the adaptive procedure given in Algorithm 1. See figure 2.2 for interpretation of the graph. Here we shorthand the threshold $T = (d + 3) \log(n_P + n_Q)$. In each step, we evaluate $r^{(k)}$ and compare it to the threshold R . If $r^{(k)} > T$, then output $\hat{f}^{(k)}$ generated in current step; if $r^{(k)} \leq T$, go to next step and add one more nearest neighbor.

The data driven adaptive classifier is then defined as

$$\hat{f}_a(x) = \mathbb{I}_{\{\hat{\eta}^{(k^*)}(x, w_P^*, w_Q^*) \geq \frac{1}{2}\}}.$$

Remark 4. The choice of $(d+3) \log(n_P + n_Q)$ as the threshold in the stopping rule (2.11) is important and requires some explanation. Roughly speaking, this is due to the fact that the maximum fluctuation of $\hat{\eta}^{(k)}(x, w_P, w_Q)$ is bounded by $\sqrt{(d+3) \log(n_P + n_Q) v^{(k)}(w_P, w_Q)}$ with high probability, which will be shown in Lemma 5 with a suitable change of parameter (stated in the supplementary material (Cai and Wei, 2019)). Thus, when $\hat{r}^{(k)} > (d+3) \log(n_P + n_Q)$, $\hat{\eta}^{(k)}(x, w_P, w_Q) > \frac{1}{2}$ indicates $\mathbb{E}\hat{\eta}^{(k)}(x, w_P, w_Q) > \frac{1}{2}$, which suggests $f^*(x) = 1$, and vice versa.

The procedure is summarized in Algorithm 1 where the above procedure is simplified by using the actual closed form expression for $\hat{r}^{(k)}$ and $\hat{f}_a(x)$.

We investigate the theoretical properties of this data-driven classifier \hat{f}_a in terms of both global and local adaptivity. The theoretical analysis shows that the proposed classifier is, both globally and locally, adaptive to the unknown smoothness and relative signal exponent.

2.4.1. Global adaptivity

Note that \hat{f}_a is a data-driven classifier. The following theorem gives an upper bound for the excess risk under Q :

Theorem 3. *Let $n = n_P + n_Q$. There exists a constant $C > 0$ not depending on n_P or n_Q such that*

$$\sup_{(P,Q) \in \Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}_a) \leq C \left(\left(\frac{n_P}{\log n} \right)^{\frac{2\beta+d}{2\gamma\beta+d}} + \frac{n_Q}{\log n} \right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}. \quad (2.12)$$

The proof of Theorem 3 is given in the supplementary material (Cai and Wei, 2019).

Comparing the rate of convergence in (2.12) for the adaptive classifier \hat{f}_a with the minimax rate (2.9), the data driven classifier \hat{f}_a simultaneously achieves within a logarithmic factor of the minimax optimal rate over a large collection of parameter spaces.

Remark 5. If only the Q -data is available and Lepski's method is applied, then the following upper bound on the excess risk under Q holds:

$$\sup_{(P,Q) \in \Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}_L) \leq C \cdot \left(\frac{n_Q}{\log n_Q} \right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}. \quad (2.13)$$

One can verify that by setting $n_P = 0$, our new adaptive procedure is exactly equivalent to Lepski's method (Algorithm 3), while the rates of convergence for the two methods also coincide.

2.4.2. Local adaptivity

In practice, one might be focused on classifying a given observation x_0 and thus especially interested in the accuracy of a classifier at a specific point x_0 . Interestingly, the weights w_P and w_Q , the number k of neighbors of the proposed classifier $\hat{f}_a(x)$ are all locally selected and

calculated based on samples in a neighborhood of x . It is of practical interest to investigate the local adaptivity of the proposed classifier.

In order to study the local behavior of the classifier \hat{f}_a at a given point x_0 , we need to extend the definitions for the posterior drift model to their local versions. First, we define the local excess risk on Q at a point x_0 :

Definition 5. For any $x_0 \in \Omega$ and a classifier $\hat{f} : \Omega \rightarrow \{0, 1\}$, define the classification risk at x_0 on distribution Q for \hat{f} as:

$$R(\hat{f}, x_0) = \mathbb{P}_{(X,Y) \sim Q}(Y \neq \hat{f}(x_0) | X = x_0).$$

Further, define the local excess risk at x_0 on distribution Q for \hat{f} as

$$\mathcal{E}_Q(\hat{f}, x_0) = R(\hat{f}, x_0) - R(f_Q^*, x_0).$$

Next, we give a formal definition for local smoothness $\beta_0 = \beta(x_0)$ and local relative signal exponent $\gamma_0 = \gamma(x)$:

Definition 6. A function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ has local Hölder smoothness β_0 ($0 < \beta_0 \leq 1$) at point $x_0 \in \mathbb{R}^d$ if there exists $r > 0$ and $C_\beta > 0$ such that for any $x' \in B(x_0, r)$,

$$|g(x') - g(x_0)| \leq C_\beta \|x' - x_0\|^\beta.$$

Definition 7. A pair of distributions (P, Q) , both supported on $\Omega \times \{0, 1\}$, are defined to have local relative signal exponent γ_0 at a point $x_0 \in \Omega$, if there exists $r > 0$ and $C_\gamma > 0$ such that for any $x \in B(x_0, r)$,

$$(\eta_P(x) - \frac{1}{2})(\eta_Q(x) - \frac{1}{2}) \geq 0$$

$$|\eta_P(x) - \frac{1}{2}| \geq C_\gamma |\eta_Q(x) - \frac{1}{2}|^\gamma.$$

The definitions of local smoothness and local relative signal exponent are similar to their global versions, except we only consider in a small neighborhood of x_0 . Based on the above definitions, the local adaptivity of \hat{f}_a at x_0 is characterized as follows:

Theorem 4. *Suppose the distributions (P, Q) are both supported on $\Omega \times \{0, 1\}$ and a point $x_0 \in \Omega$. Suppose the following holds.*

1. (P, Q) have local relative signal exponent γ_0 at x_0 ;
2. η_Q has local Hölder smoothness β_0 at x_0 ;
3. $(P_X, Q_X) \in \mathcal{S}(\mu, c_\mu, r_\mu)$, i.e. P_X and Q_X have common support and satisfy the strong density assumption.

Let $n = n_P + n_Q$. There exists a constant $C > 0$ such that

$$\mathbb{E}_Z \mathcal{E}_Q(\hat{f}_a, x_0) \leq C \left(\left(\frac{n_P}{\log n} \right)^{\frac{2\beta_0+d}{2\gamma_0\beta_0+d}} + \frac{n_Q}{\log n} \right)^{-\frac{\beta_0}{2\beta_0+d}} \quad (2.14)$$

The proof of Theorem 4 is provided in the supplementary material.

Remark 6. Under the same setting as in Theorem 4, when β_0 and γ_0 are known, the local minimax rate of convergence is

$$\inf_{\hat{f}} \sup_{(P, Q)} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}_a, x_0) \asymp (n_P^{\frac{2\beta_0+d}{2\gamma_0\beta_0+d}} + n_Q)^{-\frac{\beta_0}{2\beta_0+d}}$$

where the supremum is taken over all distribution pairs (P, Q) satisfying conditions 1,2,3 stated in Theorem 4. This minimax rate can be achieved by the same construction as the minimax classifier in Section 2.3 (using local parameters β_0, γ_0 instead of global parameters β, γ). As a result, Theorem 4 shows that \hat{f}_a also achieves within a logarithmic factor of the

local minimax optimal rate. In other words, \hat{f}_a adapts to local smoothness and local signal relative exponent.

Remark 7. For simplicity, this chapter focuses on the posterior drift model, which is somewhat restrictive since the relation between P and Q is described by a signal parameter γ . However, because \hat{f}_a is adaptive to the local signal relative exponent, it can make nearly optimal classification under heterogeneity where γ varies. In other words, \hat{f}_a works optimally even when P is stronger than Q in some places and weaker than Q elsewhere.

Remark 8. Note that there is also a dual representation of $\mathbb{E}_Z \mathcal{E}_Q(\hat{f}_a, x_0)$:

$$\mathbb{E}_Z \mathcal{E}_Q(\hat{f}_a, x_0) = 2|\eta_Q(x_0) - \frac{1}{2}| \mathbb{P}_Z \left(\hat{f}_a(x_0) \neq f_Q^*(x_0) \right)$$

Theorem 4 can be interpreted as follows. For any point x_0 , the classifier \hat{f}_a performs well (i.e. the accuracy of \hat{f}_a is bounded away from 1/2) when

$$|\eta_Q(x_0) - \frac{1}{2}| \geq C \left(\left(\frac{n_P}{\log n} \right)^{\frac{2\beta_0+d}{2\gamma_0\beta_0+d}} + \frac{n_Q}{\log n} \right)^{-\frac{\beta_0}{2\beta_0+d}}$$

for some constant $C > 0$. Other than the sample sizes n_P and n_Q , the rate only depends on the local smoothness β_0 and local relative signal exponent γ_0 . Also, it is optimal up to a logarithmic factor. The result thus shows that \hat{f}_a is adaptive to the local smoothness and local relative signal exponent.

2.5. Multiple Source Distributions

We have so far focused on transfer learning with one source distribution P and one target distribution Q . In practice, data may be generated from more than one source distribution. In this section, we generalize our methods to treat transfer learning in the setting where multiple source distributions are available.

We consider a model where there are several source distributions with different relative

signal exponents with respect to the target distribution Q . Suppose there are n_{P_1}, \dots, n_{P_m} , and n_Q i.i.d data points generated from the source distributions P_1, \dots, P_m , and the target distribution Q respectively,

$$\begin{aligned} (X_1^{P_1}, Y_1^{P_1}), \dots, (X_{n_{P_1}}^{P_1}, Y_{n_{P_1}}^{P_1}) &\stackrel{\text{i.i.d.}}{\sim} P_1 \\ &\vdots \\ (X_1^{P_m}, Y_1^{P_m}), \dots, (X_{n_{P_m}}^{P_m}, Y_{n_{P_m}}^{P_m}) &\stackrel{\text{i.i.d.}}{\sim} P_m \\ (X_1^Q, Y_1^Q), \dots, (X_{n_Q}^Q, Y_{n_Q}^Q) &\stackrel{\text{i.i.d.}}{\sim} Q \end{aligned}$$

and all the samples are independent. The goal is to make classification under the target distribution Q . Similar as before, it is intuitively clear that how useful the data from the source distributions $P_i, i \in [m]$, to the classification task under Q depends on the relationship between each P_i and Q . The definition of the relative signal exponent needs to be extended to accommodate the multiple source distributions. It is natural to consider the situation where each source distribution P_i and the target distribution Q have a relative signal exponent. This motivates the following definition of the vectorized relative signal exponent when there are multiple source distributions.

Definition 8. Suppose the distributions P_1, \dots, P_m , and Q are supported on $\mathbb{R}^d \times \{0, 1\}$. Define the class $\Gamma(\boldsymbol{\gamma}, C_\boldsymbol{\gamma})$ with the relative signal exponent $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_m) \in \mathbb{R}_+^m$ and constants $C_\boldsymbol{\gamma} = (C_1, \dots, C_m) \in \mathbb{R}_+^m$, is the set of distribution tuples (P_1, \dots, P_m, Q) that satisfy, for each $i \in [m]$, (P_i, Q) belongs to the class $\Gamma(\gamma_i, C_i)$ with the relative signal exponent γ_i .

Similar as in Section 2.2, adding the regularity conditions on Q including the smoothness, margin assumption and strong density assumption, we define the parameter space in the

multiple source distributions setting as follows:

$$\begin{aligned} \Pi(\alpha, C_\alpha, \beta, C_\beta, \gamma, C_\gamma, \mu, c_\mu, r_\mu) = \{ & (P_1, \dots, P_m, Q) : (P_1, \dots, P_m, Q) \in \Gamma(\gamma, C_\gamma), \\ & Q \in \mathcal{M}(\alpha, C_\alpha), \eta_Q \in \mathcal{H}(\beta, C_\beta), (P_{i,X}, Q_X) \in \mathcal{S}(\mu, c_\mu, r_\mu) \text{ for all } i \in [m]\}. \end{aligned}$$

We will simply denote $\Pi(\alpha, C_\alpha, \beta, C_\beta, \gamma, C_\gamma, \mu, c_\mu, r_\mu)$ by Π or $\Pi(\alpha, \beta, \gamma, \mu)$ if there is no confusion.

In this section we establish the minimax optimal rate of convergence and propose an adaptive classifier which simultaneously achieves the optimal rate of convergence within a logarithmic factor over a wide collection of the parameter spaces. The proofs are similar to those for Theorems 1, 2 and 3 in the one source distribution setting. For reasons of space, we omit the proofs.

2.5.1. Minimax rate of convergence

We begin with the construction of a minimax rate optimal classifier \hat{f}_{NN} in the case of multiple source distributions. The classifier is an extension of the two-sample weighted K -NN classifier given in Section 2.3. It incorporates the information contained in the data drawn from the source distributions P_i , $i \in [m]$, as well as the data drawn from the target distribution Q . The detailed steps are as follows.

Step 1: Compute the weights w_{P_1}, \dots, w_{P_m} , and w_Q by

$$\begin{aligned} w_{P_i} &= (n_Q + \sum_{i=1}^m n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}})^{-\frac{\gamma_i\beta}{2\beta+d}}, \quad \text{for all } i \in [m], \\ w_Q &= (n_Q + \sum_{i=1}^m n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}})^{-\frac{\beta}{2\beta+d}}. \end{aligned}$$

Compute the numbers of neighbors $k_{P_1}, \dots, k_{P_m}, k_Q$ by

$$k_{P_i} = \lfloor n_{P_i} (n_Q + \sum_{i=1}^m n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}})^{-\frac{d}{2\beta+d}} \rfloor, \quad \text{for all } i \in [m]$$

$$k_Q = \lfloor n_Q (n_Q + \sum_{i=1}^m n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}})^{-\frac{d}{2\beta+d}} \rfloor.$$

Step 2: Define $X_{(j)}^{P_i}(x)$ to be the j -th nearest data point to x among $X_1^{P_i}, \dots, X_{n_{P_i}}^{P_i}$ and $Y_{(j)}^{P_i}(x)$ is its corresponding response (label). Likewise, let $X_{(j)}^Q(x)$ be the j -th data point to x among $X_1^Q, \dots, X_{n_Q}^Q$ and $Y_{(j)}^Q(x)$ is its corresponding response (label). Define the weighted K -NN estimator

$$\hat{\eta}_{NN}(x) = \frac{w_Q \sum_{j=1}^{k_Q} Y_{(j)}^Q(x) + \sum_{i=1}^m \left(w_{P_i} \sum_{j=1}^{k_{P_i}} Y_{(j)}^{P_i}(x) \right)}{w_Q k_Q + \sum_{i=1}^m w_{P_i} k_{P_i}}.$$

This estimator takes weighted average of k_{P_i} nearest neighbors from the data points drawn from P_i , each with weight w_{P_i} , and k_Q nearest neighbors from the data points drawn from Q , each with weight w_Q .

Step 3: The final classifier is then defined as

$$\hat{f}_{NN}(x) = \mathbb{I}_{\{\hat{\eta}_{NN}(x) > \frac{1}{2}\}}.$$

We now analyze the theoretical properties of the classifier \hat{f}_{NN} . Theorem 5 gives an upper bound for the excess risk $\mathcal{E}_Q(\hat{f}_{NN})$, while Theorem 6 provides a matching lower bound on the excess risk. These two theorems together establish the minimax rate of convergence and show that \hat{f}_{NN} attains the optimal rate. In the following theorems, the expectation \mathbb{E} is taken over random realization of all data drawn from source and target distributions.

Theorem 5 (Upper Bound). *There exists a constant $C > 0$ not depending on n_P or n_Q ,*

such that

$$\sup_{(P_1, \dots, P_m, Q) \in \Pi(\alpha, \beta, \gamma, \mu)} \mathbb{E} \mathcal{E}_Q(\hat{f}_{NN}) \leq C(n_Q + \sum_{i=1}^m n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}})^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Theorem 6 (Lower Bound). *There exists a constant $c > 0$ not depending on n_P or n_Q , such that*

$$\inf_{\hat{f}} \sup_{(P_1, \dots, P_m, Q) \in \Pi(\alpha, \beta, \gamma, \mu)} \mathbb{E} \mathcal{E}_Q(\hat{f}) \geq c(n_Q + \sum_{i=1}^m n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}})^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Theorems 5 and 6 together yield the minimax optimal rate for transfer learning with multiple source distributions:

$$\inf_{\hat{f}} \sup_{(P_1, \dots, P_m, Q) \in \Pi(\alpha, \beta, \gamma, \mu)} \mathbb{E} \mathcal{E}_Q(\hat{f}) \asymp (n_Q + \sum_{i=1}^m n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}})^{-\frac{\beta(1+\alpha)}{2\beta+d}}. \quad (2.15)$$

As discussed in Section 2.3, here $n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}}$ can be viewed as the effective sample size for data drawn from the source distribution P_i when the information in this sample is transferred to help the classification task under the target distribution Q . Even when there are multiple source distributions, the transfer rate associated with P_i remains to be $\frac{2\beta+d}{2\gamma_i\beta+d}$, which is not affected by the presence of the data drawn from the other source distributions.

2.5.2. Adaptive classifier

Again, the minimax classifier is not practical as it depends on the parameters γ and μ which are typically unknown. It is desirable to construct a data driven classifier that does not rely on the knowledge of the model parameters. A similar adaptive data-driven classifier can be developed. The detailed steps are summarized in Algorithm 2.

It is clear from the construction that the classifier \hat{f}_a is a data-driven decision rule. Theorem 7 below provides a theoretical guarantee for the excess risk of \hat{f}_a under the target distribution Q . In view of the optimal rate given in (2.15), Theorem 7 shows that \hat{f}_a is adaptively nearly optimal over a wide range of parameter spaces.

Theorem 7. Let $n = n_Q + \sum_{i=1}^m n_{P_i}$. There exists a constant $C > 0$ such that for $\Pi = \Pi(\alpha, \beta, \gamma, \mu)$,

$$\sup_{(P_1, \dots, P_m, Q) \in \Pi} \mathbb{E} \mathcal{E}_Q(\hat{f}_a) \leq C \cdot \left(\frac{n_Q}{\log n} + \sum_{i=1}^m \left(\frac{n_{P_i}}{\log n} \right)^{\frac{2\beta+d}{2\gamma_i\beta+d}} \right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Algorithm 1 The Data Driven Procedure

Input: $x \in \text{supp}(Q_X)$

for $k = 1, \dots, (n_P + n_Q - 1), (n_P + n_Q)$ **do**

Find k nearest covariates to x among all covariates in data $\{X_1^P, X_2^P, \dots, X_{n_P}^P\} \cup \{X_1^Q, X_2^Q, \dots, X_{n_Q}^Q\}$. Suppose among those k nearest neighbors $X_{(1)}(x), X_{(2)}(x), \dots, X_{(k)}(x)$ there are $k_P^{(k)}$ covariates from P -data and $k_Q^{(k)}$ covariates from Q -data.

Compute $k_P^{(k)}$ nearest neighbor estimate in P -data (If $k_P^{(k)} = 0$, set $\hat{\eta}_P^{(k)} \leftarrow \frac{1}{2}$)

$$\hat{\eta}_P^{(k)} \leftarrow \frac{1}{k_P^{(k)}} \sum_{i=1}^{k_P^{(k)}} Y_{(i)}^P(x)$$

and $k_Q^{(k)}$ nearest neighbor estimate in Q -data (If $k_Q^{(k)} = 0$, set $\hat{\eta}_Q^{(k)} \leftarrow \frac{1}{2}$)

$$\hat{\eta}_Q^{(k)} \leftarrow \frac{1}{k_Q^{(k)}} \sum_{i=1}^{k_Q^{(k)}} Y_{(i)}^Q(x)$$

Let $\hat{r}^{(k)}$ be the signal-to-noise ratio index calculated by

$$\hat{r}^{(k)} \leftarrow \begin{cases} k_P^{(k)} \left(\hat{\eta}_P^{(k)} - \frac{1}{2} \right)^2 + k_Q^{(k)} \left(\hat{\eta}_Q^{(k)} - \frac{1}{2} \right)^2 & \text{if } \text{sign}(\hat{\eta}_P^{(k)} - \frac{1}{2}) = \text{sign}(\hat{\eta}_Q^{(k)} - \frac{1}{2}) \\ \max \left(k_P^{(k)} \left(\hat{\eta}_P^{(k)} - \frac{1}{2} \right)^2, k_Q^{(k)} \left(\hat{\eta}_Q^{(k)} - \frac{1}{2} \right)^2 \right) & \text{if } \text{sign}(\hat{\eta}_P^{(k)} - \frac{1}{2}) \neq \text{sign}(\hat{\eta}_Q^{(k)} - \frac{1}{2}) \end{cases}$$

Define the intermediate classifier by

$$\hat{f}^{(k)}(x) \leftarrow \mathbb{I}_{\{\sqrt{k_P^{(k)}}(\hat{\eta}_P^{(k)} - \frac{1}{2}) + \sqrt{k_Q^{(k)}}(\hat{\eta}_Q^{(k)} - \frac{1}{2}) \geq 0\}}$$

if $\hat{r}^{(k)}(x) > (d + 3) \log(n_P + n_Q)$ **then**

Stop and output $\hat{f}_a(x) \leftarrow \hat{f}^{(k)}(x)$

Output $\hat{f}_a(x) \leftarrow \hat{f}^{(k_m)}(x)$ where $k_m = \arg \max_k \hat{r}^{(k)}$

Algorithm 2 The Data Driven Classifier

Input: $x \in \text{supp}(Q_X)$

for $k = 1, \dots, (n_Q + \sum_{i=1}^m n_{P_i} - 1), (n_Q + \sum_{i=1}^m n_{P_i})$ **do**

Find k nearest neighbors $X_{(1)}(x), \dots, X_{(k)}(x)$ to x among all the covariates $\{X_j^Q : j \in [n_Q]\} \cup \bigcup_{i=1}^m \{X_j^{P_i} : j \in [n_{P_i}]\}$. Suppose $k_{P_i}^{(k)}$ of them are from the distribution P_i , $i = 1, \dots, m$, and $k_Q^{(k)}$ of them are from Q . That is, the k nearest neighbors are partitioned into $m + 1$ parts according to which distribution they are drawn from.

For each $i \in [m]$, Compute the K -NN estimate for η_{P_i} (If $k_{P_i}^{(k)} = 0$, set $\hat{\eta}_{P_i}^{(k)} \leftarrow \frac{1}{2}$)

$$\hat{\eta}_{P_i}^{(k)}(x) \leftarrow \frac{1}{k_{P_i}^{(k)}} \sum_{j=1}^{k_{P_i}^{(k)}} Y_{(j)}^{P_i}(x)$$

and nearest neighbor estimate for η_Q (If $k_Q^{(k)} = 0$, set $\hat{\eta}_Q^{(k)} \leftarrow \frac{1}{2}$)

$$\hat{\eta}_Q^{(k)} \leftarrow \frac{1}{k_Q^{(k)}} \sum_{i=1}^{k_Q^{(k)}} Y_{(i)}^Q(x).$$

Compute the positive signal-to-noise index

$$\hat{r}_+^{(k)} \leftarrow \mathbb{I}_{\{\eta_Q^{(k)} \geq \frac{1}{2}\}} k_Q^{(k)} \left(\eta_Q^{(k)} - \frac{1}{2} \right)^2 + \sum_{i=1}^m \mathbb{I}_{\{\eta_{P_i}^{(k)} \geq \frac{1}{2}\}} k_{P_i}^{(k)} \left(\eta_{P_i}^{(k)} - \frac{1}{2} \right)^2$$

and negative signal-to-noise index

$$\hat{r}_-^{(k)} \leftarrow \mathbb{I}_{\{\eta_Q^{(k)} < \frac{1}{2}\}} k_Q^{(k)} \left(\eta_Q^{(k)} - \frac{1}{2} \right)^2 + \sum_{i=1}^m \mathbb{I}_{\{\eta_{P_i}^{(k)} < \frac{1}{2}\}} k_{P_i}^{(k)} \left(\eta_{P_i}^{(k)} - \frac{1}{2} \right)^2.$$

Let $\hat{r}^{(k)}$ be the signal-to-noise ratio index calculated by

$$\hat{r}^{(k)} \leftarrow \max \left\{ \hat{r}_+^{(k)}, \hat{r}_-^{(k)} \right\}.$$

Define the classifier

$$\hat{f}^{(k)}(x) \leftarrow \mathbb{I}_{\{\hat{r}_+^{(k)} \geq \hat{r}_-^{(k)}\}}.$$

if $\hat{r}^{(k)} > (d + 3) \log(n_Q + \sum_{i=1}^m n_{P_i})$ **then**

Stop and output $\hat{f}_a(x) \leftarrow \hat{f}^{(k)}(x)$.

Output $\hat{f}_a(x) \leftarrow \hat{f}^{(k_m)}(x)$ where $k_m = \text{argmax}_k \hat{r}^{(k)}$.

2.6. Numerical Studies

In this section, we carry out simulation studies to further illustrate the performance of the adaptive transfer learning procedure. Numerical comparisons with the existing methods are given. The simulation results are consistent with the theoretical predictions.

For all simulation experiments in this section, the data is generated under the posterior drift model with $d = 2$. The distributions (P, Q) used to generate data is specified as following:

1. Marginal distributions: $P_X = Q_X$ are both uniform distribution on the square $\Omega = [-1, 1]^2$.
2. Regression functions: η_Q and η_P are defined as

$$\eta_Q(x) = 0.5 + p \operatorname{sign}(x_1) (|x_1| \max\{0, 1 - |x_2|\})^\beta$$

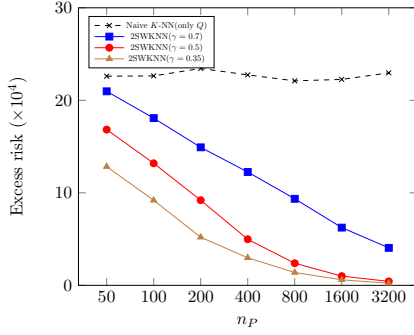
and

$$\eta_P(x) = 0.5 + p \operatorname{sign}(x_1) (|x_1| \max\{0, 1 - |x_2|\})^{\gamma\beta}$$

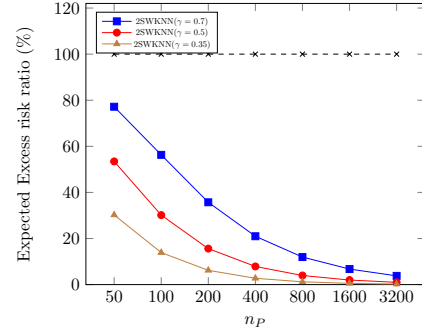
where $x = (x_1, x_2) \in [-1, 1]^2$, p, β and γ are parameters that may vary in different simulation studies.

According to the above construction, both η_P and η_Q take the maximum values at $(1, 0)$ and the minimum values at $(-1, 0)$, and equal to 0.5 when $x_1 = 0$. it can be easily verified that $\eta_Q \in \mathcal{H}(\beta, C_\beta)$ with some $C_\beta > 0$, $(P, Q) \in \Gamma(\gamma, 1)$, Q satisfies the margin assumption with $\alpha = 0.99/\beta$, and P_X and Q_X have the common support and bounded densities.

In the following experiments, we focus on evaluating the average excess risk at a random test sample x drawn uniformly from the square $\Omega = [-1, 1]^2$, given n_P data generated from P and n_Q data generated from Q .



(a) Experimental results



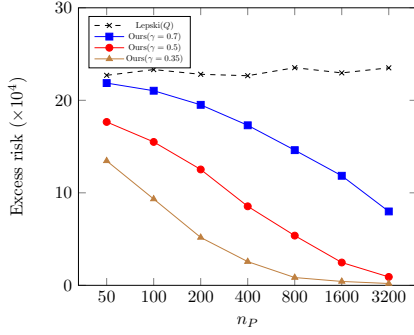
(b) Theoretical prediction

Figure 2.4: Left: Experiments on non-adaptive methods. We operate the naive K -NN method on only Q -data (dashed line) and our two-sample weighted K -NN classifier on different datasets. The datasets are generated with relative signal exponent $\gamma = 0.7, 0.5, 0.35$ respectively. Right: based on our theory (Theorem 1), the expected ratio of excess risk between the two methods we operate in the experiment.

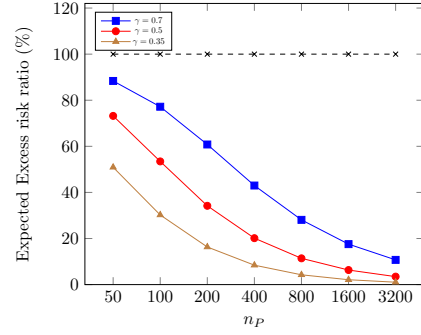
2.6.1. Minimax non-adaptive classifier

For this particular distribution pair (P, Q) , theoretically, the minimax rate of convergence for the excess risk can be achieved via the two-sample weighted K -NN classifier when we are able to make use of model parameters β, γ . In the following simulation, we fix $p = 0.03$, $n_Q = 1000$ and $\beta = 1$. By comparing the proposed non-adaptive classifier with a naive K -NN classifier on just the Q -data, we evaluate the improvement on the excess risk under different values of γ and n_P .

During the experiment, we generated datasets with choices of the relative signal exponent $\gamma \in \{0.7, 0.5, 0.35\}$ and number of P -data n_P varying from 50 to 3200. The excess risk of the two-sample weighted K -NN classifier and the naive K -NN method are illustrated in Figure 2.4a. Meanwhile, a planer plot is given in Figure 2.4b to illustrate the expected ratio of the excess risk between the two methods based on our developed theory (Theorem 1). One can clearly see how the transfer rates play a role in the experiments with different relative signal exponent γ . The empirical performance and our theoretical prediction are matched to some extent.



(a) Experimental results



(b) Theoretical prediction

Figure 2.5: Left: Experiments on adaptive methods. We operate the naive Lepski method on only Q -data (dashed line) and our adaptive classifier on different datasets. The datasets are generated with relative signal exponent $\gamma = 0.7, 0.5, 0.35$ respectively. Right: based on our theory (Theorem 3), the expected ratio of excess risk between the two methods used in the experiment.

2.6.2. Adaptive classifier

We also compare the proposed adaptive classifier with the existing methods to see whether its numerical performance matches its theoretical guarantees. Lepski’s method is a good competitor as it is also adaptive to the smoothness parameter β . Following a similar routine as in the previous experiments, we compare the excess risk between our proposed classifier and Lepski’s method applying only the Q -data to evaluate the improvement we may gain empirically.

Fix $p = 0.03$ and $\beta = 1$, we generated $n_Q = 1000$ data from the target distribution Q , and $n_P \in \{50, 100, 200, 400, 800, 1600, 3200\}$ data from the source distribution P with different choices of relative signal exponent $\gamma \in \{0.7, 0.5, 0.35\}$. Results of the numerical experiments are shown in Figure 2.5a. A figure of the expected improvement on excess risk, calculated according to Theorem 3, is also available in Figure 2.5b. In both figures, the curve looks like a reversed "S" shape when γ is large, whereas a curve of exponential decrease appears when γ is small. Therefore, it is justified that the simulation results are consistent with the theoretical predictions.

2.6.3. Multiple source distributions

Other than involving only a single source distribution during the previous numerical studies, it is also worthwhile to see whether we can gain desired improvement as our theory predicts when there are multiple source distributions. We only illustrate in this subsection the performance of our adaptive classifier applying to multiple source distributions (Algorithm 2).

Different from the previous simulation studies, in this subsection we generate data from three different source distributions P_1, P_2, P_3 and one target distribution Q . In a similar vein, the distributions (P_1, P_2, P_3, Q) are specified as following:

1. Marginal distributions: we set $P_{1,X} = P_{2,X} = P_{3,X} = Q_X$ to be all uniformly distributed on the square area $\Omega = [-1, 1]^2$.
2. Regression functions: we set η_Q and $\eta_{P_1}, \eta_{P_2}, \eta_{P_3}$ as

$$\eta_Q(x) = 0.5 + p \operatorname{sign}(x_1) (|x_1| \max\{0, 1 - |x_2|\})^\beta$$

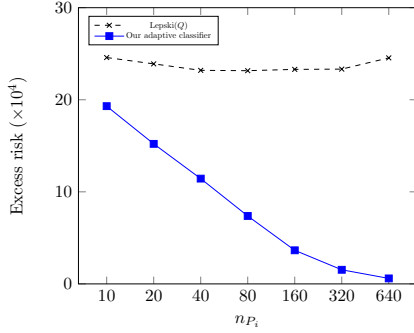
and

$$\eta_{P_i}(x) = 0.5 + p \operatorname{sign}(x_1) (|x_1| \max\{0, 1 - |x_2|\})^{\gamma_i \beta} \quad i = 1, 2, 3$$

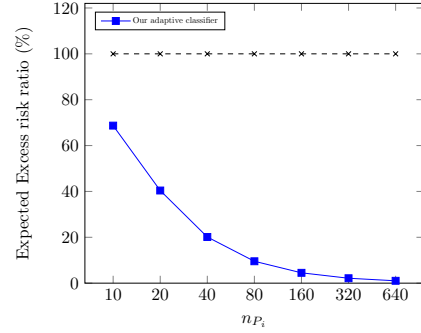
where $x = (x_1, x_2) \in [-1, 1]^2$, p, β and $\gamma_1, \gamma_2, \gamma_3$ are parameters that will be specified later.

In the simulation, we fix $p = 0.03, \beta = 1$ and $\gamma_1 = 0.35, \gamma_2 = 0.5, \gamma_3 = 0.7$, and we always set $n_{P_1} = n_{P_2} = n_{P_3}$. We compare the average excess risk of the two classifiers: our proposed adaptive classifier and the Lepski's procedure with only Q -data involved. By varying number of data drawing from source distributions, we can clearly see an improvement when applying transfer learning methods.

The excess risk of the two methods during the experiments are illustrated in Figure 2.6a.



(a) Experimental results



(b) Theoretical prediction

Figure 2.6: Left: Experiments on transfer learning from multiple source distributions. We apply the naive Lepski method on only Q -data (dashed line) and our adaptive classifier for multiple source distributions. Right: based on our theory (Theorem 5), the expected ratio of excess risk between the two methods we operate in the experiment.

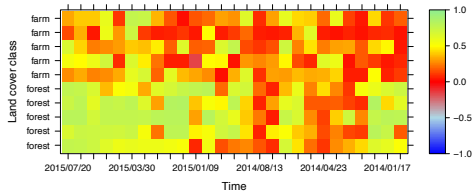
Also, we calculate the expected ratio between the two methods according to the theory we developed in Theorem 5. Again, the empirical performance and our theoretical prediction are similar to some extent.

For reasons of space, additional simulation results on different choices of β are given in the supplementary material (Cai and Wei, 2019).

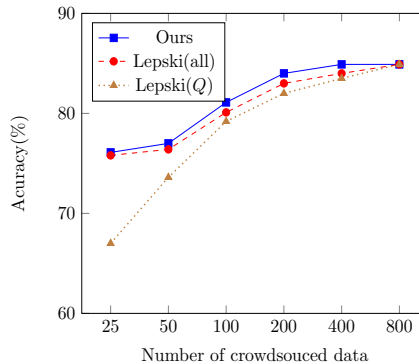
2.7. Application to Crowdsourced Mapping Data

To illustrate the proposed adaptive classifier, we consider in this section an application based on the crowdsourced mapping data (Johnson and Iizuka, 2016). Land use/land cover maps derived from remotely-sensed imagery are important for geographic studies. This dataset contains Landsat time-series satellite imagery information on given pixels and their corresponding land cover class labels (farm, forest, water, etc.) obtained from multiple sources. The goal is to make classification of land cover classes based on NDVI (normalized difference vegetation index) values of those remotely-sensed imagery from the years 2014-2015. In this chapter we focus on classification of two specific classes: farm and forest.

Within this dataset, there are two kinds of label sources, given the NVDI values of the images: 1) crowdsourced georeferenced polygons with land cover labels obtained from OpenStreetMap; 2) accurately labeled data by experts. Although crowdsourced data are massive,



(a) Illustration of the dataset.



(b) Accuracy of different classifiers.

Figure 2.7: (a) Illustration of the dataset. Each row represents one of a land cover class (farm or forest) and corresponding NDVI values of a pixel from remotely-sensed imagery in 2014-2015. (b) Accuracy of the three methods on the crowdsourced mapping data with different numbers of crowdsourced data involved. Blue: The proposed adaptive classifier. Red: Lepski’s method using combined data. Brown: Lepski’s method using only crowdsourced data.

free and public, the labels contain various types of errors due to user mislabels or outdated images. Whereas the expert labels are almost accurate, but they are usually too expensive to obtain a large volume. The challenge is to accurately combine the information contained in the two datasets to minimize the classification error.

As in Section 2.6.2, we apply three methods to make the classification: (1) our proposed adaptive procedure; (2) Lepski’s method with all data involved where we do not distinguish data from different sources; (3) Lepski’s method with only the crowdsourced data. We use $n_P = 50$ accurately labeled data, and change the number of involved crowdsourced data from $n_Q = 25$ to $n_Q = 800$. We use other 166 accurately labeled data to evaluate the classification accuracy of the three methods mentioned above.

Figure 2.7b shows the accuracy of the three methods with different numbers of crowdsourced data involved. As more and more crowdsourced data are used, the amount of information contained in the crowdsourced data gradually increases, and the relative contribution from the accurately labeled data gradually decreases. The proposed adaptive classifier consistently outperforms the naive Lepski’s method, especially when the number of the crowdsourced data is between 100 and 400, because in these cases the adaptive classifier can

significantly increase the accuracy by better leveraging the information gained from both distributions.

2.8. Discussion

We studied in this chapter transfer learning under the posterior drift model and established the minimax rate of convergence. The optimal rate quantifies precisely the amount of information in the P -data that can be transferred to help classification under the target distribution Q . A delicately designed data-driven adaptive classifier was also constructed and shown to be, both globally and locally, adaptive to the unknown smoothness and relative signal exponent. It is simultaneously within a log factor of the optimal rate over a large collection of parameter spaces.

The results and techniques developed in this chapter serve as a starting point for the theoretical analysis of other transfer learning problems. For example, in addition to classification, it is also of significant interest to characterize the relationship between the source distribution and the target distribution, so that the data from the source distribution P can help in other statistical problems under the target distribution Q . Examples include regression, hypothesis testing, and construction of confidence sets. We will investigate these transfer learning problems in the future.

Within the posterior drift framework of this chapter, some of the technical assumptions can be relaxed to a certain extent. For the smoothness parameter β , we focused on the case $0 < \beta \leq 1$. It is possible to consider more general classes where β can be larger than 1, with strengthened relative signal exponent assumptions on the higher order derivatives of $\eta_P(x)$ and $\eta_Q(x)$. When $\beta > 2$, the problem might be solved with a carefully designed weighted K -NN classifier, as was introduced in Samworth (2012). Construction of such a weighted K -NN method is involved and we leave it as future work. For the marginal distributions P_X and Q_X , other than the strong density assumption, there are also weaker regularity conditions introduced in the literature. See, for example, Gadat et al. (2016); Kpotufe and Martinet (2018). Similar results on the minimax rate of convergence can be established under these

different regularity conditions. The minimax and adaptive procedures should also be suitably modified.

Also, in complementary work, Kpotufe and Martinet (2018) studied K -NN classifiers for transfer learning in the *covariate shift* framework where the marginal distributions P_X and Q_X are allowed to differ significantly. It is interesting to consider nonparametric classification under both *covariate shift* and *posterior drift*. In such a setting, besides the relative signal exponent γ , one also assumes (P, Q) have *transfer-exponent* $\tau \geq 0$ such that

$$\forall x, r \in (0, \Delta_{\mathcal{X}}], P_X(B(x, r)) \geq Q_X(B(x, r)) \cdot C_\tau \left(\frac{r}{\Delta_{\mathcal{X}}} \right)^\tau,$$

and Q_X is (C_d, d) -doubling, as is defined in Definitions 3 and 6 in Kpotufe and Martinet (2018). The detailed analysis appears to be quite involved, we only make some conjectures here based on our preliminary calculations and leave the rigorous proofs and further investigations for future work. Our calculations indicate that the optimal rate of convergence for the excess risk on Q under both covariate shift (transfer-exponent τ) and posterior drift (relative signal exponent γ) should be

$$\inf_{\hat{f}} \sup_{(P, Q)} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}) \asymp \left(n_P^{\frac{2\beta+d}{2\gamma\beta+\tau+d}} + n_Q \right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

An additional transfer-exponent τ appears in the denominator of the transfer rate $\frac{2\beta+d}{2\gamma\beta+\tau+d}$. The above optimal rate can be achieved by two-sample weighted K -NN classifier (proposed in our work) with proper choices of w_P, w_Q, k_P and k_Q . In addition, our proposed classifier \hat{f}_a should be nearly optimal adaptive classifier (up to a logarithmic term) in a sense that

$$\sup_{(P, Q)} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}) \lesssim \left(\left(\frac{n_P}{\log n} \right)^{\frac{2\beta+d}{2\gamma\beta+\tau+d}} + \frac{n_Q}{\log n} \right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}$$

where $n = n_P + n_Q$.

2.9. Proofs

We prove Theorem 1 in this section and leave the proofs of other theorems and additional technical lemmas in the supplementary material (Cai and Wei, 2019). For readers' convenience, we begin by stating Lepski's method for nonparametric classification in the conventional setting where there are only the Q -data.

2.9.1. Lepski's method

Algorithm 3 is a version of Lepski's method in nonparametric classification. We state the algorithm here for reference.

Algorithm 3 Lepski's method (Lepski and Spokoiny, 1997)

Input: n labeled samples $(X_i, Y_i) \in \mathbb{R}^d \times \{0, 1\}$, $i \in [n]$, and a point $x \in \mathbb{R}^d$ to be classified.

Set $\eta_0^- \leftarrow -\infty$ and $\eta_0^+ \leftarrow +\infty$.

for $k = 1, \dots, (n_P + n_Q - 1), (n_P + n_Q)$ **do**

Find k nearest neighbor estimates $\hat{\eta}_k(x) = \frac{1}{k} \sum_{i=1}^k Y_{(i)}$, where $Y_{(i)}$ denote the label to i -th nearest covariates to x .

Set $\eta_k^- \leftarrow \eta_{k-1}^- \vee (\hat{\eta}_k(x) - \sqrt{\frac{d+3}{k}} \log n)$.

Set $\eta_k^+ \leftarrow \eta_{k-1}^+ \wedge (\hat{\eta}_k(x) + \sqrt{\frac{d+3}{k}} \log n)$.

if $\eta_k^- > \frac{1}{2}$ or $\eta_k^+ < \frac{1}{2}$ **then**

Stop and output $\hat{f}_L(x) \leftarrow \mathbb{I}_{\{\hat{\eta}_k(x) \geq \frac{1}{2}\}}$.

Output $\hat{f}_L(x) \leftarrow \mathbb{I}_{\{\hat{\eta}_n(x) \geq \frac{1}{2}\}}$.

2.9.2. Proof of Theorem 1

First we define some new notations for convenience. In the proof, we use $\zeta_Q(x) = |\eta_Q(x) - \frac{1}{2}|$ and $\zeta_P(x) = |\eta_P(x) - \frac{1}{2}|$ to denote the signal strength. Let $\bar{Y}_{(1:k_Q)}^Q(x) := \frac{1}{k_Q} \sum_{i=1}^{k_Q} Y_{(i)}^Q(x)$ and $\bar{Y}_{(1:k_P)}^P(x) := \frac{1}{k_P} \sum_{i=1}^{k_P} Y_{(i)}^P(x)$ denote the average of k_Q nearest neighbors in Q -data and k_P nearest neighbors in P -data respectively. We will sometime omit x in the notations such as $X_{(i)}^Q(x), X_{(i)}^P(x)$ if there is no confusion in the context. We also use the shorthand $X_{1:n_Q}^Q$ to denote the whole set of the Q -data covariates $\{X_1^Q, \dots, X_{n_Q}^Q\}$. And similarly $X_{1:n_P}^P$ denotes $\{X_1^P, \dots, X_{n_P}^P\}$. We define $\mathbb{E}_{Y|X}(\cdot) = \mathbb{E}(\cdot | X_{1:n_Q}^Q, X_{1:n_P}^P)$ to denote the expectation conditional on the covariates of all data, and \mathbb{E} is the expectation taken over random realization of all

data (the same as \mathbb{E}_Z we defined before). Also, in following proofs we always assume $(P, Q) \in \Pi(\alpha, \beta, \gamma, \mu)$ so we will not state this assumption again in the lemmas.

Before proving the theorem, we first state three useful lemmas. The first lemma 1 provides a high probability uniform bound on the distance between any point and its k -th nearest neighbor.

Lemma 1 (*K-NN Distance Bound*). *There exists a constant $C_D > 0$ such that, with probability at least $1 - C_D \frac{n_Q}{k_Q} \exp(-\frac{k_Q}{6})$, for all $x \in \Omega$,*

$$\|X_{(k_Q)}^Q(x) - x\| \leq C_D \left(\frac{k_Q}{n_Q}\right)^{\frac{1}{d}}. \quad (2.16)$$

And with probability at least $1 - C_D \frac{n_P}{k_P} \exp(-\frac{k_P}{6})$, for all $x \in \Omega$,

$$\|X_{(k_P)}^P(x) - x\| \leq C_D \left(\frac{k_P}{n_P}\right)^{\frac{1}{d}}. \quad (2.17)$$

Let E_Q denote the event that Inequality (2.16) holds for all $x \in \Omega$ and let E_P denotes (2.17) holds for all $x \in \Omega$. It follows from Lemma 1 that

$$\mathbb{P}(E_Q) \geq 1 - C_D \frac{n_Q}{k_Q} \exp(-\frac{k_Q}{6}) \quad \text{and} \quad \mathbb{P}(E_P) \geq 1 - C_D \frac{n_P}{k_P} \exp(-\frac{k_P}{6}).$$

Lemma 2 points out that when the signal is sufficiently strong, bias of $\bar{Y}^Q(x)$ and $\bar{Y}^P(x)$ will not be too large to overwhelm the signal.

Lemma 2 (*Bias Bound*). *There exist constants $c_b, C_b > 0$ such that:*

If a point $x \in \Omega$ satisfies $\zeta_Q(x) \geq 2C_b \|X_{(k_Q)}^Q(x) - x\|^\beta$, then we have

$$\mathbb{E}_{Y|X}(\bar{Y}_{(1:k_Q)}^Q(x)) - \frac{1}{2} \geq c_b \zeta_Q(x) \text{ if } f^*(x) = 1, \quad (2.18)$$

$$\mathbb{E}_{Y|X}(\bar{Y}_{(1:k_Q)}^Q(x)) - \frac{1}{2} \leq -c_b \zeta_Q(x) \text{ if } f^*(x) = 0. \quad (2.19)$$

If a point $x \in \Omega$ satisfies $\zeta_Q(x) \geq 2C_\beta \|X_{(k_P)}^P(x) - x\|^\beta$, then we have

$$\mathbb{E}_{Y|X}(\bar{Y}_{(1:k_P)}^P(x)) - \frac{1}{2} \geq c_b \zeta_Q(x)^\gamma \text{ if } f^*(x) = 1, \quad (2.20)$$

$$\mathbb{E}_{Y|X}(\bar{Y}_{(1:k_P)}^P(x)) - \frac{1}{2} \leq -c_b \zeta_Q(x)^\gamma \text{ if } f^*(x) = 0. \quad (2.21)$$

Hence, if a point $x \in \Omega$ satisfies $\zeta_Q(x) \geq C_b (\max\{\frac{k_Q}{n_Q}, \frac{k_P}{n_P}\})^{\frac{\beta}{d}}$, then

- Under the event E_Q , we have

$$\mathbb{E}_{Y|X}(\bar{Y}_{(1:k_Q)}^Q(x)) - \frac{1}{2} \geq c_b \zeta_Q(x) \text{ if } f^*(x) = 1, \quad (2.22)$$

$$\mathbb{E}_{Y|X}(\bar{Y}_{(1:k_Q)}^Q(x)) - \frac{1}{2} \leq -c_b \zeta_Q(x) \text{ if } f^*(x) = 0. \quad (2.23)$$

- Under the event E_P , we have

$$\mathbb{E}_{Y|X}(\bar{Y}_{(1:k_P)}^P(x)) - \frac{1}{2} \geq c_b \zeta_Q(x)^\gamma \text{ if } f^*(x) = 1, \quad (2.24)$$

$$\mathbb{E}_{Y|X}(\bar{Y}_{(1:k_P)}^P(x)) - \frac{1}{2} \leq -c_b \zeta_Q(x)^\gamma \text{ if } f^*(x) = 0. \quad (2.25)$$

Lemma 3 gives a bound on the probability of misclassification at certain covariates x .

Lemma 3 (Misclassification Bound). *Let C_b and c_b be the constants defined in Lemma 2.*

If $\zeta_Q(x) \geq C_b (\max\{\frac{k_Q}{n_Q}, \frac{k_P}{n_P}\})^{\frac{\beta}{d}}$, then

- Under the event E_Q , we have

$$\mathbb{P}_{Y|X}(\hat{f}_{NN}(x) \neq f_Q^*(x)) \leq \exp\left(-2 \frac{[(c_b w_Q k_Q \zeta_Q(x) - w_P k_P) \vee 0]^2}{k_P w_P^2 + k_Q w_Q^2}\right).$$

- Under the event E_P , we have

$$\mathbb{P}_{Y|X}(\hat{f}_{NN}(x) \neq f_Q^*(x)) \leq \exp\left(-2 \frac{[(c_b w_P k_P \zeta_Q(x)^\gamma - w_Q k_Q) \vee 0]^2}{k_P w_P^2 + k_Q w_Q^2}\right).$$

- Under the event $E_P \cap \mathbb{E}_Q$, we have

$$\mathbb{P}_{Y|X}(\hat{f}_{NN}(x) \neq f_Q^*(x)) \leq \exp\left(-2c_b^2 \frac{(w_P k_P \zeta_Q(x)^\gamma + w_Q k_Q \zeta_Q(x))^2}{k_P w_P^2 + k_Q w_Q^2}\right).$$

Given the three lemmas above, the remain proof generally follows that of Lemma 3.1 in Audibert and Tsybakov (2007). Let $\delta = (n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta}{2\beta+d}}$. When w_P, w_Q, k_P, k_Q are given as in Theorem 1, we have

$$w_Q = \delta, w_P = \delta^\gamma, k_Q = \lfloor n_Q \delta^{\frac{d}{\beta}} \rfloor, k_P = \lfloor n_P \delta^{\frac{d}{\beta}} \rfloor. \quad (2.26)$$

We will approximate $k_Q = n_Q \delta^{\frac{d}{\beta}}$ and $k_P = n_P \delta^{\frac{d}{\beta}}$ in the following proof because one can easily show such an approximation only result sin changing the constant factor in the upper bound.

The following lemma gives a bound for the local misclassification risk when the parameters in the weighted K-NN estimator are properly chosen.

Lemma 4. *Using w_P, w_Q, k_P, k_Q defined in theorem 1 to construct a weighted K-NN estimator \hat{f}_{NN} . Then there exist constants $c_1, C_1 > 0$ such that, with probability at least $1 - 2(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta(1+\alpha)}{2\beta+d}}$, for all x we have*

$$\mathbb{P}_{Y|X}(\hat{f}_{NN}(x) \neq f_Q^*(x)) \leq C_1 \exp\left(-c_1 \left(\frac{\zeta_Q(x)}{\delta}\right)^{1 \wedge \gamma}\right). \quad (2.27)$$

Let E_0 be the event that inequality (2.27) holds for all x . Lemma 4 implies

$$\mathbb{P}(E_0) \geq 1 - 2(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Consider the disjoint sets $\mathcal{A}_j \subset \Omega, j = 0, 1, 2, \dots$ defined as

$$\begin{aligned}\mathcal{A}_0 &:= \{x \in \Omega : 0 < \zeta_Q(x) \leq \delta\}, \\ \mathcal{A}_j &:= \{x \in \Omega : 2^{j-1}\delta < \zeta_Q(x) \leq 2^j\delta\} \text{ for } j \geq 1.\end{aligned}$$

Note that by the margin assumption, for all j ,

$$Q_X(A_j) \leq Q_X(|\eta_Q - \frac{1}{2}| \leq 2^j\delta) \leq C_\alpha 2^{\alpha j} \delta^\alpha.$$

Based on the partition A_0, A_1, \dots and the dual representation of $\mathcal{E}_Q(\hat{f})$ shown in (2.8), we have a decomposition of $\mathcal{E}_Q(\hat{f}_{NN})$:

$$\begin{aligned}\mathcal{E}_Q(\hat{f}_{NN}) &= 2\mathbb{E}_{X \sim Q_X}(|\eta_Q(X) - \frac{1}{2}| \mathbb{I}_{\{\hat{f}_{NN}(X) \neq f_Q^*(X)\}}) \\ &= 2 \sum_{j=0}^{\infty} \mathbb{E}_{X \sim Q_X}(\zeta_Q(X) \mathbb{I}_{\{\hat{f}_{NN}(X) \neq f_Q^*(X)\}} \mathbb{I}_{\{X \in A_j\}}).\end{aligned}$$

For $j = 0$, $\mathbb{E}_{X \sim Q_X}(\zeta_Q(X) \mathbb{I}_{\{\hat{f}_{NN}(X) \neq f_Q^*(X)\}} \mathbb{I}_{\{X \in A_0\}}) \leq \delta \cdot Q_X(A_0) \leq C_\alpha \delta^{\alpha+1}$.

Under the event E_0 , $2^{j-1}\delta < \zeta(x) \leq 2^j\delta$ for $x \in A_j$ and $j > 1$. Inequality (2.27) now yields

$$\begin{aligned}\mathbb{E}_{Y|X} \mathbb{E}_{X \sim Q_X}(\zeta_Q(X) \mathbb{I}_{\{\hat{f}_{NN}(X) \neq f_Q^*(X)\}} \mathbb{I}_{\{X \in A_j\}}) \\ &= \mathbb{E}_{X \sim Q_X}(\zeta_Q(X) \mathbb{P}_{Y|X}(\hat{f}_{NN}(X) \neq f_Q^*(X)) \mathbb{I}_{\{X \in A_j\}}) \\ &\leq 2^j \delta \cdot C_1 \exp(-c_1 \cdot 2^{(j-1) \cdot (1 \wedge \gamma)}) \cdot Q_X(A_j) \\ &\leq C_\alpha C_1 [2^{(1+\alpha)j} \exp(-c_1 \cdot 2^{(j-1) \cdot (1 \wedge \gamma)})] \delta^{\alpha+1}.\end{aligned}$$

Combining these summands together yields

$$\begin{aligned}
\mathbb{E}_{Y|X} \mathcal{E}_Q(\hat{f}_{NN}) &= 2 \sum_{j=0}^{\infty} \mathbb{E}_{Y|X} \mathbb{E}_{X \sim Q_X} (\zeta_Q(X) \mathbb{I}_{\{\hat{f}_{NN}(X) \neq f_Q^*(X)\}} \mathbb{I}_{\{X \in A_j\}}) \\
&\leq 2C_\alpha \left(1 + C_1 \sum_{j=0}^{\infty} [2^{(1+\alpha)j} \exp(-c_1 \cdot 2^{(k-1) \cdot (1 \wedge \gamma)})] \right) \delta^{1+\alpha} \\
&\leq C \delta^{1+\alpha}.
\end{aligned}$$

where the last step follows from the fact that $\sum_{j=0}^{\infty} [2^{(1+\alpha)j} \exp(-c_1 \cdot 2^{(k-1) \cdot (1 \wedge \gamma)})]$ converges when $\gamma > 0$. Finally, it follows from Lemma 4 that

$$\mathbb{P}(E_0^c) \leq 2(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Applying the trivial bound $\mathcal{E}_Q(\hat{f}_{NN}) \leq 1$ when E_0^c occurs, we have

$$\begin{aligned}
\mathbb{E} \mathcal{E}_Q(\hat{f}_{NN}) &= \mathbb{E}(\mathbb{E}_{Y|X} \mathcal{E}_Q(\hat{f}_{NN})) \\
&\leq \mathbb{E}(\mathbb{E}_{Y|X} \mathcal{E}_Q(\hat{f}_{NN}) | E_0) \mathbb{P}(E_0) + \mathbb{E}(\mathbb{E}_{Y|X} \mathcal{E}_Q(\hat{f}_{NN}) | E_0^c) \mathbb{P}(E_0^c) \\
&\leq C(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta(1+\alpha)}{2\beta+d}} \cdot 1 + 1 \cdot 2(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta(1+\alpha)}{2\beta+d}} \\
&= (C+2)(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.
\end{aligned}$$

□

CHAPTER 3

DISTRIBUTED GAUSSIAN MEAN ESTIMATION WITH KNOWN VARIANCE UNDER COMMUNICATION CONSTRAINTS

3.1. Introduction

In the conventional statistical decision theoretical framework, the focus is on the centralized setting where all the data are collected together and directly available. The main goal is to develop optimal (estimation, testing, detection, ...) procedures, where optimality is understood with respect to the sample size and parameter space. Communication/computational costs are not part of the consideration.

In the age of big data, communication/computational concerns associated with a statistical procedure are becoming increasingly important in contemporary applications. One of the difficulties for analyzing large datasets is that data are distributed, instead of in a single centralized location. This setting arises naturally in many statistical practices.

- **Large datasets.** When the datasets are too large to be stored on a single computer or data center, it is necessary to divide the whole dataset into multiple computers or data centers, each assigned a smaller subset of the full dataset. Such is the case for a wide range of applications.
- **Privacy and security.** Privacy and security concerns can also cause the decentralization of the datasets. For example, medical and financial institutions often collect datasets that contain sensitive and valuable information. For privacy and security reasons, the data cannot be released to a third party for a centralized analysis and need to be stored in different and secure places while performing data analysis.

Distributed learning, which aims to learn from distributed datasets, has attracted much recent attention. For example, Google AI proposed a machine learning setting called "Federated Learning" (McMahan and Ramage, 2017), which develops a high-quality centralized

model while the training data remain distributed over a large number of clients. Figure 3.1a provides a simple illustration of a distributed learning network. In addition to advances on architecture design for distributed learning in practice, there is also an increasing amount of literature on distributed learning theories, including Jordan et al. (2019), Battey et al. (2018), Dobriban and Sheng (2018), and Fan et al. (2019) in statistics, computer science, and information theory communities. Several distributed learning procedures with some theoretical properties have been developed in recent works. However, they do not impose any communication constraints on the proposed procedures thus fail to characterize the relationship between the communication costs and statistical accuracy. Indeed, in a decision theoretical framework, if no communication constraints are imposed, one can always output the original data from the local machines to the central machine and treat the problem same as in the conventional centralized setting.

For large-scale data analysis, communications between machines can be slow and expensive and limitation on bandwidth and communication sometimes becomes the main bottleneck on statistical efficiency. It is therefore necessary to take communication constraints into consideration when constructing statistical procedures. When the communication budget is limited, the algorithm must carefully “compress” the information contained in the data as efficiently as possible, leading to a trade-off between communication costs and statistical accuracy. The precisely quantification of this trade-off is an important and challenging problem.

Estimation of a Gaussian mean occupies a central position in parametric statistical inference. In this chapter we consider distributed Gaussian mean estimation under the communication constraints in both the univariate and multivariate settings. Although optimal estimation of a Gaussian mean is a relatively simple problem in the conventional setting, this problem is quite involved under the communication constraints, both in terms of the construction of the rate optimal distributed estimator and the lower bound argument. Optimal distributed estimation of a Gaussian mean also serves as a starting point for investigating other more

complicated statistical problems in distributed learning including distributed nonparametric function estimation, distributed high-dimensional linear regression, and distributed large-scale multiple testing.

3.1.1. Problem formulation

We begin by giving a formal definition of **transcript**, **distributed estimator**, and **independent distributed protocol**. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a parametric family of distributions supported on space \mathcal{X} , where $\theta \in \Theta \subseteq \mathbb{R}^d$ is the parameter of interest. Suppose there are m local machines and a central machine, where each local machine contains n i.i.d observations and the central machine produces the final estimator of θ under the communication constraints between the local and central machines. More precisely, suppose we observe i.i.d. random samples drawn from a distribution $P_\theta \in \mathcal{P}$:

$$X_{i,j} \stackrel{\text{i.i.d.}}{\sim} P_\theta, \quad i = 1, \dots, m; j = 1, \dots, n$$

where the i -th local machine has access to $X_{i,1}, X_{i,2}, \dots, X_{i,n}$ only. We denote $\tilde{X}_i = (X_{i,1}, X_{i,2}, \dots, X_{i,n})$ as the set of data on the i -th local machine.

For $i = 1, \dots, m$, let $b_i \geq 1$ be a positive integer and the i -th local machine can only transmit b_i bits to the central machine. That is, the observation \tilde{X}_i on the i -th local machine needs to be processed to a binary string of length b_i by a (possibly random) function $\Pi_i : \mathcal{X}^n \rightarrow \{0, 1\}^{b_i}$. The resulting string $Z_i \triangleq \Pi_i(\tilde{X}_i)$, which is called the **transcript** from the i -th machine, is then transmitted to the central machine. Finally, a **distributed estimator** $\hat{\theta}$ is constructed on the central machine based on the transcripts Z_1, Z_2, \dots, Z_m ,

$$\hat{\theta} = \hat{\theta}(Z_1, Z_2, \dots, Z_m).$$

The above scheme to obtain a distributed estimator $\hat{\theta}$ is called an **independent distributed protocol**, or independent protocol.

In addition to the independent protocol, there are other more general and interactive dis-

tributed protocols including the sequential protocol and blackboard protocol, which are two popular communication protocols considered in the literature (Zhang et al., 2013a; Barnes et al., 2019b). We shall first focus on the independent protocol, then introduce the sequential and blackboard protocols and establish optimality results for these two types of distributed protocols in Section 3.4.

The class of independent distributed protocols with communication budgets b_1, b_2, \dots, b_m is defined as

$$\mathcal{A}_{ind}(b_1, b_2, \dots, b_m) = \{(\hat{\theta}, \Pi_1, \Pi_2, \dots, \Pi_m) : \Pi_i : \mathcal{X}^n \rightarrow \{0, 1\}^{b_i}, i = 1, 2, \dots, m, \\ \hat{\theta} = \hat{\theta}(\Pi_1(\tilde{X}_1), \dots, \Pi_m(\tilde{X}_m))\}.$$

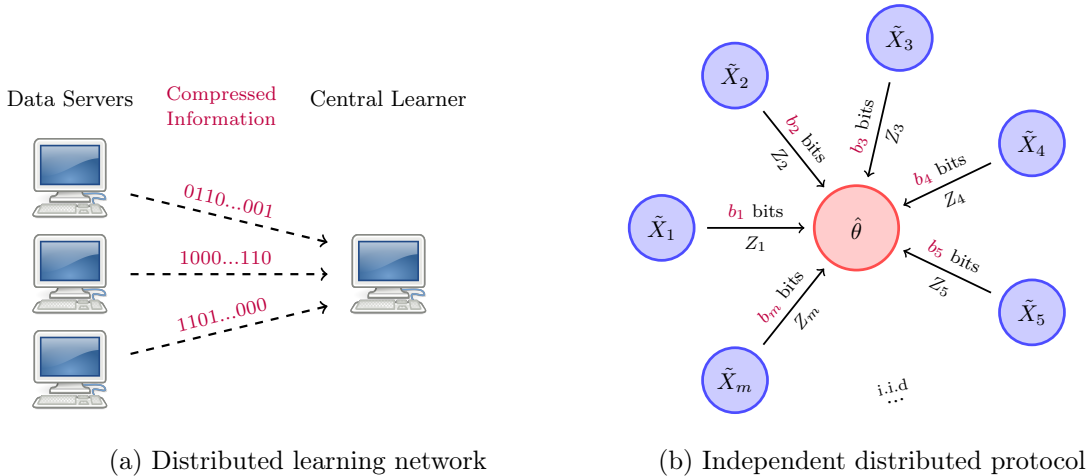


Figure 3.1: (a) Left panel: An illustration of a distributed learning network. Communication between the data servers and the central learner is necessary in order to learn from distributed datasets. (b) Right panel: An illustration of independent distributed protocol. The i -th machine can only transmit a b_i bits transcript to the central machine. The transcript Z_i only depends on observations \tilde{X}_i .

We use $b_{1:m}$ as a shorthand for (b_1, b_2, \dots, b_m) and denote $\hat{\theta} \in \mathcal{A}_{ind}(b_{1:m})$ for $(\hat{\theta}, \Pi_1, \dots, \Pi_m) \in \mathcal{A}_{ind}(b_{1:m})$. We shall always assume $b_i \geq 1$ for all $i = 1, 2, \dots, m$, i.e. each local machine can transmit at least one bit to the central machine. Otherwise, if no communication is allowed from any of the local machines, one can just exclude those local machines and treat the problem as if there are fewer local machines available. Figure 3.1b gives a simple illustration

for the distributed protocols.

As usual, the estimation accuracy of a distributed estimator $\hat{\theta}$ is measured by the mean squared error (MSE), $\mathbb{E}_{P_\theta} \|\hat{\theta} - \theta\|_2^2$, where the expectation is taken over the randomness in both the data and construction of the transcripts and estimator. As in the conventional decision theoretical framework, a quantity of particular interest in distributed learning is the minimax risk for the distributed protocols

$$\inf_{\hat{\theta} \in \mathcal{A}_{ind}(b_{1:m})} \sup_{P_\theta \in \mathcal{P}} \mathbb{E}_{P_\theta} \|\hat{\theta} - \theta\|_2^2,$$

which characterizes the difficulty of the distributed learning problem under the communication constraints $b_{1:m}$. As mentioned earlier, in a rigorous decision theoretical formulation of distributed learning, the communication constraints are essential. Without the constraints, one can always output the original data from the local machines to the central machine and the problem is then reduced to the usual centralized setting.

3.1.2. Distributed estimation of a univariate Gaussian mean

We first consider distributed estimation of a univariate Gaussian mean under the communication constraints $b_{1:m}$, where $P_\theta = N(\theta, \sigma^2)$ with $\theta \in [0, 1]$ and the variance σ^2 known. Set $\sigma_n = \sigma/\sqrt{n}$. Note that by a sufficiency argument, one can estimate θ based on the sample means $X_i \triangleq \frac{1}{n} \sum_{j=1}^n X_{i,j}$ on the local machines, and the problem is the same as if one only observes $X_i \sim N(\theta, \sigma_n^2)$ on the i -th machine, for $i = 1, \dots, m$.

Our analysis in Section 3.2 establishes the following minimax rate of convergence for distributed univariate Gaussian mean estimation under the communication constraints $b_{1:m}$,

$$\inf_{\hat{\theta} \in \mathcal{A}_{ind}(b_{1:m})} \sup_{\theta \in [0,1]} \mathbb{E}(\hat{\theta} - \theta)^2 \asymp \begin{cases} 2^{-2B} & \text{if } B < \log_2 \frac{1}{\sigma_n} + 2 \\ \frac{\sigma_n^2}{(B - \log_2 \frac{1}{\sigma_n})} & \text{if } \log_2 \frac{1}{\sigma_n} + 2 \leq B < \log_2 \frac{1}{\sigma_n} + m, \\ \min \left\{ \frac{\sigma_n^2}{m}, 1 \right\} & \text{if } B \geq \log_2 \frac{1}{\sigma_n} + m \end{cases} \quad (3.1)$$

where $B = \sum_{i=1}^m b_i$ is the total communication budget, and $a \asymp b$ denotes $cb \leq a \leq Cb$ for some constants $c, C > 0$. The same optimal rate of convergence holds for the class of sequential protocols and blackboard protocols.

The above minimax rate characterizes the trade-off between the communication costs and statistical accuracy for univariate Gaussian mean estimation. An illustration of the minimax rate is shown in Figure 3.2.

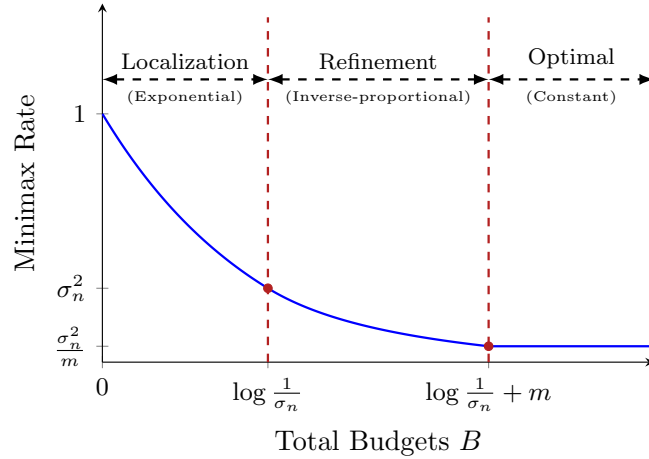


Figure 3.2: The minimax rate of univariate Gaussian mean estimation under communication constraints has 3 phases: localization, refinement and optimal-rate.

The minimax rate (3.1) is interesting in several aspects. First, the optimal rate of convergence only depends on the total communication budget $B = \sum_{i=1}^m b_i$, but not the specific allocation of the communication budgets among the m local machines, as long as each machine has at least one bit. Second, the rate of convergence has three different phases:

1. Localization phase. When $B < \log_2 \frac{1}{\sigma_n} + 2$, as a function of B , the minimax risk decreases fast at an exponential rate. In this phase, having more communication budget is very beneficial in terms of improving the estimation accuracy.
2. Refinement phase. When $\log_2 \frac{1}{\sigma_n} + 2 \leq B < \log_2 \frac{1}{\sigma_n} + m$, as a function of B , the minimax risk decreases relatively slowly and is inverse-proportional to the total communication budget B .

3. Optimal-rate phase. When $B \geq \log_2 \frac{1}{\sigma_n} + m$, the minimax rate does not depend on B , and is the same as in the centralized setting where all the data are combined (Bickel, 1981b).

An essential technique for solving this problem is the decomposition of the minimax estimation problem into two steps, *localization* and *refinement*. This critical decomposition provides a framework for both the lower bound analysis and optimal procedure design. In the lower bound analysis, the statistical error is decomposed into “localization error” and “refinement error”. It is shown that one of these two terms is inevitably large under the communication constraints. In our optimal procedure called MODGAME, bits of the transcripts are divided into three types: crude localization bits, finer localization bits, and refinement bits. They compress the local data in a way that both the localization and refinement errors can be optimally reduced. Further technical details and discussion are presented in Section 3.2. Furthermore, it will be shown that MODGAME is also robust against departures from Gaussianity. See Section 3.5 for a detailed discussion.

3.1.3. Distributed estimation of a multivariate Gaussian mean

We then consider the multivariate case under the communication constraints $b_{1:m}$, where $P_\theta = N_d(\theta, \sigma^2 I_d)$ with $\theta \in [0, 1]^d$ and the noise level σ is known. As in the univariate case, by a sufficiency argument, it is equivalent to consider distributed estimation where each local machine only observes a local sample mean vector $X_i \sim N_d(\theta, \sigma_n^2 I_d)$, with $\sigma_n = \sigma/\sqrt{n}$. The goal is to optimally estimate the mean vector θ under the squared error loss.

The construction and the analysis given in Section 3.3 show that the minimax rate of convergence in this case is given by

$$\inf_{\hat{\theta} \in \mathcal{A}_{ind}(b_{1:m})} \sup_{\theta \in [0,1]^d} \mathbb{E} \|\hat{\theta} - \theta\|_2^2 \asymp \begin{cases} 2^{-2B/d} d & \text{if } B/d < \log_2 \frac{1}{\sigma_n} + 2 \\ \frac{d\sigma_n^2}{(B/d - \log_2 \frac{1}{\sigma_n})} & \text{if } \log_2 \frac{1}{\sigma_n} + 2 \leq B/d < \log_2 \frac{1}{\sigma_n} + \max\{m', 2\} \\ d \min \left\{ \frac{\sigma_n^2}{m'}, 1 \right\} & \text{if } B/d \geq \log_2 \frac{1}{\sigma_n} + \max\{m', 2\} \end{cases} \quad (3.2)$$

where $B = \sum_{i=1}^m b_i$ is the total communication budgets and $m' = \sum_{i=1}^m \min \left\{ \frac{b_i}{d}, 1 \right\}$ is the

“effective sample size”. The same optimal rate of convergence holds true for the class of sequential protocols or blackboard protocols.

The minimax rate in the multivariate case (3.2) is an extension of its univariate counterpart (3.1), but it also has its distinct features, both in terms of the estimation procedure and lower bound argument. Intuitively, the total communication budgets B are evenly divided into d parts so that roughly B/d bits can be used to estimate each coordinate. Because there are d coordinates, the risk is multiplied by d . The effective sample size m' is a special and interesting quantity in multivariate Gaussian mean estimation. This quantity suggests that even when the total communication budgets are sufficient, the rate of convergence must be larger than the benchmark $d \min \left\{ \frac{\sigma_n^2}{m'}, 1 \right\}$. There is a gap between the distributed optimal rate and centralized optimal rate if $m' \ll m$. See Section 3.3 for further technical details and discussion.

3.1.4. Related literature

The study on how the communication constraints compromise the estimation accuracy in the distributed settings has a long history. Dating back to 1980’s, Zhang and Berger (1988) proposed an asymptotically unbiased distributed estimator and calculated its variance. In recent years, there has been emerging literature focusing on the theoretical properties of distributed estimation under the communication constraints. Among them, distributed Gaussian mean estimation has been intensively studied. We divide the discussion into two parts – lower bound and upper bound.

Lower bound: Zhang et al. (2013a) introduced general technical tools to derive lower bounds for several distributed estimation problems. Specifically, for d -dimensional Gaussian mean estimation with independent protocols, the lower bound is of order $\frac{\sigma_n^2 d^2}{(\sum_{i=1}^m b_i \wedge d) \log m}$. Garg et al. (2014) studied distributed estimation of the mean of a high-dimensional Gaussian distribution. A lower bound of order $\min \left\{ \frac{\sigma_n^2 d^2}{B}, d \right\}$ is established for the mean squared error of any independent protocol. Braverman et al. (2016) applied a strong data processing inequality to obtain lower bounds for distributed estimation with blackboard protocols. A

lower bound for sparse Gaussian mean estimation is derived. Han et al. (2018); Barnes et al. (2019b) proposed non-information theoretic approaches to obtain lower bounds for distributed estimation. In the case of Gaussian mean estimation, it was shown in Barnes et al. (2019b) that a lower bound of order $\sigma_n^2 \max\{\frac{d^2}{B}, \frac{d}{m}\}$ holds for any independent, sequential or blackboard protocols.

Upper bound: Garg et al. (2014) proposed a blackboard distributed protocol with the communication cost $O(md)$ which estimates the mean vector up to a squared loss of $O(\frac{d\sigma_n^2}{m})$. Braverman et al. (2016) introduced an independent distributed protocol for Gaussian mean estimation. If $\log(md/\sigma_n) = o(m)$, the protocol achieves the mean squared error $O(\frac{\sigma_n^2 d}{\alpha m})$ with the communication cost $C = \alpha dm$.

In summary, the known minimax rate for distributed Gaussian mean estimation is $\frac{\sigma_n^2 d^2}{B}$ when $\log(md/\sigma_n) = o(m)$. However, when n is large such that $\log(\sigma_n)/m$ is bounded away from zero, the optimal rate is still unknown.

In addition to the above closely related literature, Szabó and van Zanten (2018); Zhu and Lafferty (2018) considered distributed nonparametric regression with Gaussian noise and derived an optimal rate of convergence up to a logarithmic factor. The optimal rate is divided into three phases, namely insufficient regime, intermediate regime, and sufficient regime. Current best results for distributed nonparametric regression also suffer from a logarithmic gap, which in our opinion is due to the incomplete understanding of distributed Gaussian mean estimation with a small variance. Other related results can be found in the literature, see, for example, Zhang et al. (2013b); Shamir (2014); Diakonikolas et al. (2017); Han et al. (2018); Lee et al. (2017); Kipnis and Duchi (2019); Hadar and Shayevitz (2019); Szabó and van Zanten (2019, 2020).

3.1.5. Our contribution

Although the interplay between communication costs and statistical accuracy has drawn increasing recent attention, to the best of our knowledge, this chapter is the first work to

establish a sharp minimax rate for distributed Gaussian mean estimation that holds for all values of the parameters d, m, σ_n and in all communication budget regimes for three communication protocols – independent, sequential, and blackboard. Two rate-optimal estimation procedures – MODGAME for the univariate case and multi-MODGAME for the multivariate case – are developed and are shown to be robust against departures from Gaussianity.

In particular, the unified minimax rate applies to the case $\sigma_n < 1$. In comparison, when $\sigma_n < 1$, the previous results are not sharp even in the high communication budget regime (i.e. refinement phase and optimal-rate phase). See Remarks 5 and 6 for detailed comparison with previous results. This is an important case that arises in many statistical applications including distributed nonparametric regression and sparse signal recovery. Establishing a sharp and complete minimax rate is not only important for distributed Gaussian mean estimation itself, but also fundamental for solving these related problems.

This chapter also develops a key technique – the decomposition of the minimax estimation problem into two steps, *localization* and *refinement*. We provide a general framework and techniques to study the optimal trade-off between the localization and refinement errors. This is reflected in both the construction of the MODGAME procedure and in the lower bound argument. In contrast, the previous literature focused exclusively on the refinement error, and failed to consider the localization error. As a result, the existing results are sharp only when the communication costs for localization are negligible. We believe the technique for understanding the interplay between the localization and refinement errors is of independent interest as it can be used to solve other distributed estimation problems.

3.1.6. Organization of the chapter

We finish this section with notation and definitions that will be used in the rest of the chapter. Section 3.2 studies distributed estimation of a univariate Gaussian mean under communication constraints with independent protocols and Section 3.3 considers the multivariate case. Section 3.4 introduces sequential and blackboard protocols and extends the optimality results to these two types of communication protocols. Section 3.5 considers the

robustness of the proposed procedures against departures from Gaussianity. The numerical performance of the proposed distributed estimators is investigated in Section 3.6 and further research directions are discussed in Section 3.7. For reasons of space, we prove the lower bound for the univariate case in Section 3.8 and defer the proofs of the other main results and the technical lemmas to the Supplementary Material (Cai and Wei, 2020a).

3.1.7. Notation and definitions

For any $a \in \mathbb{R}$, let $\lfloor a \rfloor$ denote the floor function (the largest integer not larger than a). Unless otherwise stated, we shorthand $\log a$ as the base 2 logarithmic of a . For any $a, b \in \mathbb{R}$, let $a \wedge b \triangleq \min\{a, b\}$ and $a \vee b \triangleq \max\{a, b\}$. For any vector a , we will use $a^{(k)}$ to denote the k -th coordinate of a , and denote by $\|a\| \triangleq \sqrt{\sum_k (a^{(k)})^2}$ its l_2 norm. For any set S , let $S^k \triangleq S \times S \times \dots \times S$ be the Cartesian product of k copies of S . Let $\mathbb{I}_{\{\cdot\}}$ denote the indicator function taking values in $\{0, 1\}$.

For any discrete random variables X, Y supported on \mathcal{X}, \mathcal{Y} , the entropy $H(X)$, conditional entropy $H(X|Y)$, and mutual information $I(X; Y)$ are defined as

$$\begin{aligned} H(X) &\triangleq - \sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \log \mathbb{P}(X = x), \\ H(X|Y) &\triangleq - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y) \log \mathbb{P}(X = x|Y = y), \\ I(X; Y) &\triangleq \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y) \log \frac{\mathbb{P}(X = x|Y = y)}{\mathbb{P}(X = x)}. \end{aligned}$$

3.2. Distributed Univariate Gaussian Mean Estimation

In this section we consider distributed estimation of a univariate Gaussian mean, where one observes on m local machines i.i.d. random samples:

$$X_i \stackrel{\text{i.i.d.}}{\sim} N(\theta, \sigma_n^2), \quad i = 1, \dots, m,$$

under the constraints that the i -th machine has access to X_i only and can transmit b_i bits only to the central machine. We denote by $\mathcal{P}_{\sigma_n}^1$ the Gaussian location family

$$\mathcal{P}_{\sigma_n}^1 = \{N(\theta, \sigma_n^2) : \theta \in [0, 1]\},$$

where $\theta \in [0, 1]$ is the mean parameter of interest and the variance σ_n^2 is known. For given communication budgets $b_{1:m}$ with $b_i \geq 1$ for $i = 1, \dots, m$, the goal is to optimally estimate the mean θ under the squared error loss. A particularly interesting quantity is the minimax risk under the communication constraints, i.e., the minimax risk for the independent distributed protocol $\mathcal{A}_{ind}(b_{1:m})$:

$$R_1(b_{1:m}) = \inf_{\hat{\theta} \in \mathcal{A}_{ind}(b_{1:m})} \sup_{\theta \in [0, 1]} \mathbb{E}(\hat{\theta} - \theta)^2,$$

which characterizes the difficulty of the estimation problem with independent protocols under the communication constraints. We first focus on the independent protocols. Same results for sequential and blackboard protocols will be established in Section 3.4.

We first introduce an estimation procedure and provide an upper bound for its performance and then establish a matching lower bound on the minimax risk. The upper and lower bounds together establish the minimax rate of convergence and the optimality of the proposed estimator.

3.2.1. Estimation procedure - MODGAME

We begin with the construction of an estimation procedure under the communication constraints and provide a theoretical analysis of the proposed procedure. The procedure, called MODGAME (Minimax Optimal Distributed GAussian Mean Estimation), is a deterministic procedure that generates a distributed estimator $\hat{\theta}_D$ under the distributed protocol $\mathcal{A}_{ind}(b_{1:m})$. We divide the discussion into two cases: $\sigma_n < 1$ and $\sigma_n \geq 1$.

MODGAME procedure when $\sigma_n < 1$

When $\sigma_n < 1$, MODGAME consists of two steps: localization and refinement. Roughly speaking, the first step utilizes $\log \frac{1}{\sigma_n} + o(B - \log \frac{1}{\sigma_n})$ bits, out of the total budget $B = \sum_{i=1}^m b_i$ bits, for localization to roughly locate where θ is, up to $O(\sigma_n)$ error. Building on the location information, the remaining $B - \log \frac{1}{\sigma_n}$ bits are used for refinement to further increase the accuracy of the estimator. Detailed theoretical analysis will show that the optimality of the final estimator.

Before describing the MODGAME procedure in detail, we define several useful functions that will be used to generate the transcripts. For any interval $[L, R]$, let $\tau_{[L,R]} : \mathbb{R} \rightarrow [L, R]$ be the truncation function defined by

$$\tau_{[L,R]}(x) = \begin{cases} L & \text{if } x \leq L \\ x & \text{if } L < x < R \\ R & \text{if } x \geq R \end{cases} . \quad (3.3)$$

For any integer $k \geq 0$, denote $g_k : \mathbb{R} \rightarrow \{0, 1\}$ be the k -th Gray function defined by

$$g_k(x) \triangleq \begin{cases} 0 & \text{if } \lfloor 2^k \tau_{[0,1]}(x) \rfloor \bmod 4 = 0 \text{ or } 3 \\ 1 & \text{if } \lfloor 2^k \tau_{[0,1]}(x) \rfloor \bmod 4 = 1 \text{ or } 2. \end{cases}$$

Similarly we denote by $\bar{g}_k : \mathbb{R} \rightarrow \{0, 1\}$ the k -th conjugate Gray function defined by

$$\bar{g}_k(x) \triangleq \begin{cases} 0 & \text{if } \lfloor 2^k \tau_{[0,1]}(x) \rfloor \bmod 4 = 0 \text{ or } 1 \\ 1 & \text{if } \lfloor 2^k \tau_{[0,1]}(x) \rfloor \bmod 4 = 2 \text{ or } 3. \end{cases}$$

To unify the notation we set $g_k(x) \equiv \bar{g}_k(x) \equiv 0$ if $k < 0$.

It is worth mentioning that these Gray functions mimic the behavior of the Gray codes (for

reference see Savage (1997)). Fix $K \geq 1$, if we treat $(g_1(x), g_2(x), \dots, g_K(x))$ as a string of code for any source $x \in [0, 1]$, then those x within the interval $[2^{-K}(s-1), 2^{-K}s]$ where s is a integer will match the same code. Moreover, the code for adjacent intervals only differs by one bit, which is also a key feature for the Gray codes. This key feature guarantees the robustness of the Gray codes. Such robustness makes the Gray functions very useful for distributed estimation. An example for $K = 3$ is shown in Figure 3.3 to better illustrate the behavior of the Gray functions.

Along with the figure, we also provide a simple example to show why the Gray codes are robust to stochastic errors. Suppose X_1, X_2 , and X_3 are three i.i.d random variables with mean $1/4 + \epsilon$ and a small variance that is slightly larger than ϵ^2 . The goal is to estimate their mean by one-bit measurement of each observation. By using the Gray codes, $(g_1(X_1), g_2(X_2), g_3(X_3))$ is equal to (001) or (011) with large probability, whose decoded interval $(1/8, 1/4)$ or $(1/4, 3/8)$ is close to $1/4$. As a contrast, if one uses the binary codes, the result will be unstable due to the stochastic error of X_2 . In the MODGAME procedure, the Gray codes are used to help crudely “locate” the final estimator $\hat{\theta}_D$ to an interval of length $O(\sigma_n)$.

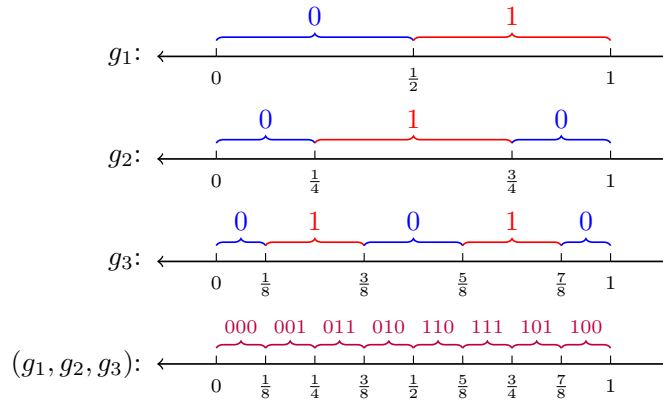


Figure 3.3: An illustration of the Gray functions and Gray codes.

Define the refinement function $h(x) : \mathbb{R} \rightarrow \{0, 1\}$ and the conjugate refinement function

$\bar{h}(x) : \mathbb{R} \rightarrow \{0, 1\}$ by

$$h(x) \triangleq \lfloor 2^{\lfloor \log \frac{1}{\sigma_n} \rfloor - 7} x \rfloor \bmod 2 \quad \text{and} \quad \bar{h}(x) \triangleq \lfloor 2^{\lfloor \log \frac{1}{\sigma_n} \rfloor - 7} x - \frac{1}{2} \rfloor \bmod 2. \quad (3.4)$$

For any function f , define the convolution function

$$\Phi_f(x) \triangleq \mathbb{E}_{X \sim N(x, \sigma_n^2)} f(X) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{(y-x)^2}{2\sigma_n^2}} f(y) dy.$$

The above refinement functions and convolution function are used to accurately estimate the mean of the Gaussian observations. In the MODGAME procedure, the central machine collects one-bit measurements of some observations, say $h(X_1), h(X_2), \dots, h(X_u)$. By definition, the mean of those one-bit measurements is exactly $\Phi_h(\theta)$. Note that $\Phi_h(x)$ is a periodic wave-shape function, therefore after locating θ to a short interval of length $O(\sigma_n)$ during the preliminary steps, the central machine obtains a good estimate for θ by solving estimating equation $\Phi_h(\theta) = u^{-1} \sum_{i=1}^u h(X_i)$. A similar communication strategy is also adopted in Braverman et al. (2016).

For any $K \geq 1$, let $\text{Dec}_K(y_1, y_2, \dots, y_K) : \{0, 1\}^K \rightarrow 2^{[0,1]}$ be the decoding function defined by

$$\text{Dec}_K(y_1, y_2, \dots, y_K) \triangleq \{x \in [0, 1] : g_k(x) = y_k \quad \text{for } k = 1, 2, \dots, K\}.$$

Last, we define the distance between a point $x \in \mathbb{R}$ and a set $S \subseteq \mathbb{R}$ as

$$d(x, S) \triangleq \min_{y \in S} |x - y|.$$

We are now ready to introduce the MODGAME procedure in detail. Again, we divide into three cases.

Case 1: $B < \log \frac{1}{\sigma_n} + 2$. In this case, the output is the values of the first B localization bits

from local machines, where the k -th localization bit is defined as the value of the function $g_k(\cdot)$ evaluated on the local sample. The procedure can be described as follows.

Step 1: *Generate transcripts on local machines.* Define $s_0 = 0$ and $s_i = \sum_{j=1}^i b_j$ for $i = 1, \dots, m$. On the i -th machine, the transcript Z_i is concatenated by the $(s_{i-1} + 1)$ -th, $(s_{i-1} + 2)$ -th, ..., $(s_{i-1} + b_i)$ -th Gray functions evaluated at X_i . That is,

$$Z_i = (U_{s_{i-1}+1}, U_{s_{i-1}+2}, \dots, U_{s_{i-1}+b_i}),$$

where $U_{s_{i-1}+k} \triangleq g_{s_{i-1}+k}(X_i)$ for $k = 1, 2, \dots, b_i$.

Step 2: *Construct distributed estimator $\hat{\theta}_D$.* Now we collect the bits U_1, U_2, \dots, U_B from the transcripts Z_1, Z_2, \dots, Z_m . Note that U_k is the k -th Gray function evaluate at a random sample drawn from $N(\theta, \sigma_n^2)$, one may reasonably "guess" that $U_k \approx g_k(\theta)$. By this intuition, we set $\hat{\theta}_D$ to be the minimum number in the interval $\text{Dec}_B(U_1, U_2, \dots, U_B)$, i.e.

$$\hat{\theta}_D = \min\{x : x \in \text{Dec}_B(U_1, U_2, \dots, U_B)\}.$$

Case 2: $\log \frac{1}{\sigma_n} + 2 \leq B \leq \log \frac{1}{\sigma_n} + m$. Let

$$u \triangleq \max \left\{ s \in \mathbb{N} : \lfloor \log s \rfloor^2 + 2s \leq B - \lfloor \log \frac{1}{\sigma_n} \rfloor \right\}, \quad (3.5)$$

and define finer localization functions:

$$\begin{aligned} f_1(x) &\triangleq g_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 2}(x), \\ f_2(x) &\triangleq \bar{g}_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 2}(x), \\ f_k(x) &\triangleq g_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 4 + k}(x) \text{ for } k \geq 3. \end{aligned} \quad (3.6)$$

In this case the total communication budget is divided into 3 parts: crude localization bits (roughly $\lfloor \log \frac{1}{\sigma_n} \rfloor$ bits), finer localization bits ($\lfloor \log u \rfloor^2$ bits), and refinement bits ($2u$ bits).

The crude localization bits are the values of the functions $g_1(\cdot), g_2(\cdot), \dots, g_{\lfloor \log \frac{1}{\sigma_n} \rfloor}(\cdot)$, each evaluated on a local sample. We denote those resulting binary bits by $U_1, U_2, \dots, U_{\lfloor \log \frac{1}{\sigma_n} \rfloor}$. The finer localization bits are the values of the functions $f_1(\cdot), f_2(\cdot), \dots, f_{\lfloor \log u \rfloor}(\cdot)$, each function is evaluated on $\lfloor \log u \rfloor$ different local samples. The function values of $f_k(\cdot)$ are denoted by $W_{k,1}, W_{k,2}, \dots, W_{k, \lfloor \log u \rfloor}$. The refinement bits are the values of the function $h(\cdot)$, evaluated on u local samples; and the values of the function $\bar{h}(\cdot)$, evaluated on u different local samples. The resulting binary bits are denoted by V_1, V_2, \dots, V_n and $\bar{V}_1, \bar{V}_2, \dots, \bar{V}_n$ respectively.

These three types of bits are assigned to local machines by the following way: (1) Among all m machines, there are $\lfloor \log u \rfloor^2$ local machines who will output transcript consisting of 1 finer localization bit and $b_i - 1$ crude localization bits. (2) Among all m machines, there are $2u$ local machines who will output transcript consist of 1 refinement bit and $b_i - 1$ crude localization bits. (3) The remain $m - (\lfloor \log u \rfloor^2 + 2u)$ machines will output transcript consist of b_i crude localization bits. The above assignment is feasible because

$$\lfloor \log u \rfloor^2 + 2u \leq B - \lfloor \log \frac{1}{\sigma_n} \rfloor \leq m.$$

It is worth mentioning that every finer localization bits and every refinement bits are assigned to different machines. Intuitively, this is because we need these bits to be independent so that we can gain enough information for estimation. See Figure 3.4 for an overview of the MODGAME procedure. The procedure can be summarized as follows:

Step 1: *Generate transcripts on local machines.* Define $s_i = \sum_{j=1}^i (b_j - \mathbb{I}_{\{j \leq \lfloor \log u \rfloor^2 + 2u\}})$ and $s_0 = 0$. On the i -th machine:

- If $(j - 1)\lfloor \log u \rfloor + 1 \leq i \leq j\lfloor \log u \rfloor$ for some integer $1 \leq j \leq \lfloor \log u \rfloor$, output

$$Z_i = (U_{s_{i-1}+1}, U_{s_{i-1}+2}, \dots, U_{s_{i-1}+b_i-1}, W_{j, i-(j-1)\lfloor \log u \rfloor});$$

(If $b_i = 1$, just output $Z_i = W_{j, i-(j-1)\lfloor \log u \rfloor}$.)

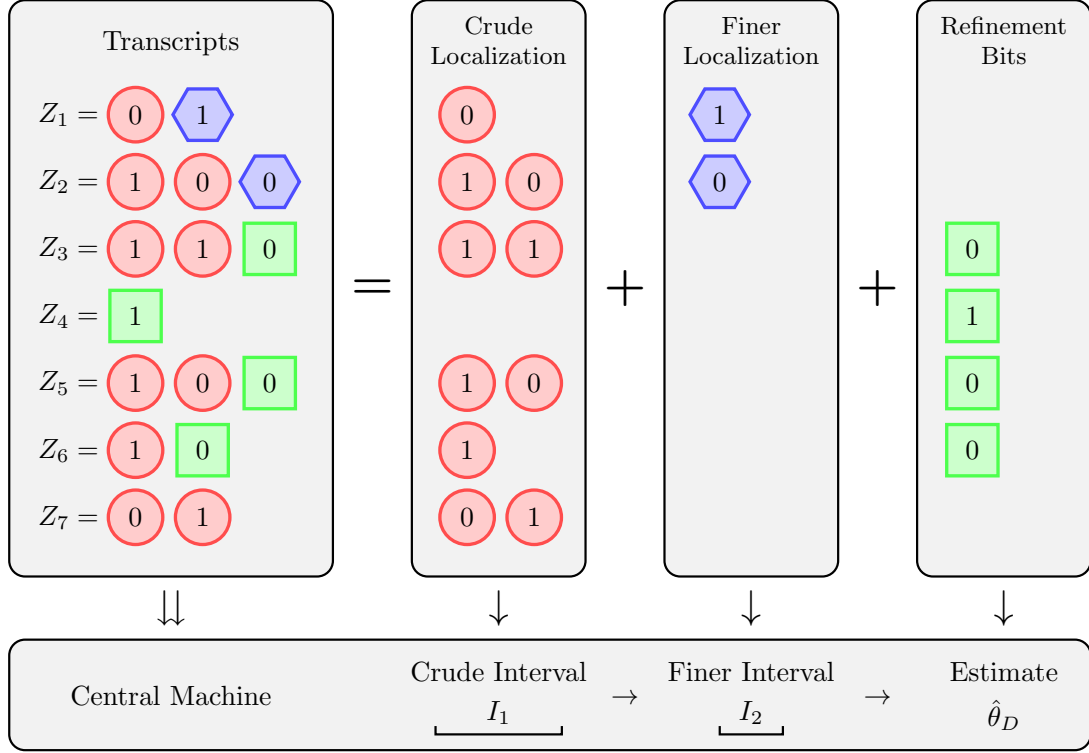


Figure 3.4: An illustration of MODGAME. The bits in the transcripts are transmitted to the central machine and divided into three types: crude localization bits, finer localization bits, and refinement bits. One then constructs on the central machine a crude interval I_1 , a finer interval I_2 , and the final estimate $\hat{\theta}_D$ step by step.

- If $\lfloor \log u \rfloor^2 + 1 \leq i \leq \lfloor \log u \rfloor^2 + u$, output

$$Z_i = (U_{s_{i-1}+1}, U_{s_{i-1}+2}, \dots, U_{s_{i-1}+b_i-1}, V_{i-\lfloor \log u \rfloor^2});$$

(If $b_i = 1$, just output $Z_i = V_{i-\lfloor \log u \rfloor^2}$.)

- If $\lfloor \log u \rfloor^2 + u + 1 \leq i \leq \lfloor \log u \rfloor^2 + 2u$, output

$$Z_i = (U_{s_{i-1}+1}, U_{s_{i-1}+2}, \dots, U_{s_{i-1}+b_i-1}, \bar{V}_{i-(\lfloor \log u \rfloor^2+u)});$$

(If $b_i = 1$, just output $Z_i = \bar{V}_{i-(\lfloor \log u \rfloor^2+u)}$.)

- If $i \geq \lfloor \log u \rfloor^2 + 2u + 1$, output

$$Z_i = (U_{s_{i-1}+1}, U_{s_{i-1}+2}, \dots, U_{s_{i-1}+b_i}).$$

where the above binary bits are calculated by

$$\begin{aligned} U_{s_{i-1}+k} &\triangleq g_{s_{i-1}+k}(X_i) \quad \text{for } i = 1, 2, \dots, m; \quad k = 1, 2, \dots, b_i. \\ W_{j, i-(j-1)\lfloor \log u \rfloor} &\triangleq f_j(X_i) \quad \text{for } j = 1, 2, \dots, \lfloor \log u \rfloor - 1; \\ &\quad i = (j-1)\lfloor \log u \rfloor + 1, (j-1)\lfloor \log u \rfloor + 2, \dots, j\lfloor \log u \rfloor. \\ V_{i-\lfloor \log u \rfloor^2} &\triangleq h(X_i) \quad \text{for } i = \lfloor \log u \rfloor^2 + 1, \lfloor \log u \rfloor^2 + 2, \dots, \lfloor \log u \rfloor^2 + u. \\ \bar{V}_{i-(\lfloor \log u \rfloor^2+u)} &\triangleq \bar{h}(X_i) \quad \text{for } i = \lfloor \log u \rfloor^2 + u + 1, \dots, \lfloor \log u \rfloor^2 + 2u. \end{aligned}$$

Step 2: Construct distributed estimator $\hat{\theta}_D$. From transcripts Z_1, Z_2, \dots, Z_m , we can collect

- crude localization bits $U_1, U_2, \dots, U_{\lfloor \log \frac{1}{\sigma_n} \rfloor}$;
- finer localization bits $W_{1,1}, W_{1,2}, \dots, W_{\lfloor \log u \rfloor, \lfloor \log u \rfloor}$;
- refinement bits V_1, V_2, \dots, V_u and $\bar{V}_1, \bar{V}_2, \dots, \bar{V}_u$.

Step 2.1: First, we use crude localization bits $U_1, U_2, \dots, U_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 3}$ to roughly locate θ . The "crude interval" I_1 will be obtained in this step.

- If $\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor \leq 3$, just set $I_1 = I'_1 = [0, 1]$.
- If $\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor \geq 4$, let

$$I'_1 \triangleq \text{Dec}_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 3}(U_1, U_2, \dots, U_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 3}). \quad (3.7)$$

Then we further stretch I'_1 to a larger interval I_1 so that I_1 will double the length of I'_1 :

$$I_1 \triangleq \left\{ x : d(x, I'_1) \leq 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 2)} \right\}. \quad (3.8)$$

Step 2.2: Then, we use finer localization bits to locate θ to a smaller interval of length roughly $O(\sigma_n)$. A "finer interval" I_2 will be generated in this step. For any $1 \leq k \leq \lfloor \log u \rfloor$, let

$$W_k = \mathbb{I}_{\{\sum_{j=1}^{\lfloor \log u \rfloor} W_{k,j} \geq \frac{1}{2} \lfloor \log u \rfloor\}}$$

be the majority voting summary statistic for $W_{k,1}, W_{k,2}, \dots, W_{k, \lfloor \log u \rfloor}$.

(a) If $\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor \leq 3$, and $\lfloor \log \frac{1}{\sigma_n} \rfloor \leq 4$, let

$$I_2 = I'_2 = [0, 1].$$

(b) If $\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor \leq 3$, and $\lfloor \log \frac{1}{\sigma_n} \rfloor \geq 5$, let

$$I'_2 \triangleq \text{Dec}_{\lfloor \log \frac{1}{\sigma_n} \rfloor - 4}(W_{\lfloor \log u \rfloor - \lfloor \log \frac{1}{\sigma_n} \rfloor + 5}, W_{\lfloor \log u \rfloor - \lfloor \log \frac{1}{\sigma_n} \rfloor + 6}, \dots, W_{\lfloor \log u \rfloor}). \quad (3.9)$$

Then we further stretch I'_2 to a larger interval I_2 so that I_2 will double the length of I'_2 :

$$I_2 \triangleq \left\{ x : d(x, I'_2) \leq 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 3)} \right\}.$$

(c) If $\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor \geq 4$, let

$$I'_2 \triangleq \{x \in I_1 : f_k(x) = W_k \text{ for all } k = 1, 2, \dots, \lfloor \log u \rfloor\}. \quad (3.10)$$

Lemma 7 in the Supplementary Material Cai and Wei (2020a) shows I'_2 is an interval. Then we further stretch I'_2 to a larger interval I_2 so that I_2 will double the length of I'_2 :

$$I_2 \triangleq \left\{ x : d(x, I'_2) \leq 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 3)} \right\}.$$

Step 2.3: Finally, we use refinement bits V_1, V_2, \dots, V_u and $\bar{V}_1, \bar{V}_2, \dots, \bar{V}_u$ to get an accurate

estimate $\hat{\theta}_D$. Lemma 8 in the Supplementary Material Cai and Wei (2020a) shows that one of the following two conditions must hold:

$$I_2 \subseteq \left[\left(2j - \frac{3}{4}\right) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}, \left(2j + \frac{3}{4}\right) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)} \right] \text{ for some } j \in \mathbb{Z}$$

or

$$I_2 \subseteq \left[\left(2j + \frac{1}{4}\right) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}, \left(2j + \frac{7}{4}\right) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)} \right] \text{ for some } j \in \mathbb{Z}.$$

So we can divide the procedure into the following two cases.

(a) If $I_2 \subseteq \left[\left(2j - \frac{3}{4}\right) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}, \left(2j + \frac{3}{4}\right) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)} \right]$ for some $j \in \mathbb{Z}$. Then $\Phi_h(x)$ is a strictly monotone function on I_2 (proved in Lemma 8 in the Supplementary Material Cai and Wei (2020a)). Denote

$$L_I \triangleq \inf_{x \in I_2} \Phi_h(x) \quad \text{and} \quad R_I \triangleq \sup_{x \in I_2} \Phi_h(x).$$

By monotonicity, Φ_h is invertible on I_2 . Let $\Phi_h^{-1} : [L_I, R_I] \rightarrow I_2$ be the inverse of Φ_h , the distributed estimator $\hat{\theta}_D$ is given by

$$\hat{\theta}_D = \Phi_h^{-1} \left(\tau_{[L_I, R_I]} \left(\frac{1}{u} \sum_{j=1}^u V_j \right) \right) \quad (3.11)$$

where $\tau_{[L_I, R_I]}$ is the truncation function defined in (3.3).

(b) Otherwise, we have $I_2 \subseteq \left[\left(2j + \frac{1}{4}\right) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}, \left(2j + \frac{7}{4}\right) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)} \right]$ for some $j \in \mathbb{Z}$. In this case $\Phi_{\bar{h}}(x)$ is a strictly monotone function on I_2 (proved in Lemma 8 in the Supplementary Material Cai and Wei (2020a)). Denote

$$\bar{L}_I \triangleq \inf_{x \in I_2} \Phi_{\bar{h}}(x) \quad \text{and} \quad \bar{R}_I \triangleq \sup_{x \in I_2} \Phi_{\bar{h}}(x).$$

By monotonicity, $\Phi_{\bar{h}}$ is invertible on I_2 . Let $\Phi_{\bar{h}}^{-1} : [\bar{L}_I, \bar{R}_I] \rightarrow I_2$ be the inverse of $\Phi_{\bar{h}}$, the distributed estimator $\hat{\theta}_D$ is given by

$$\hat{\theta}_D = \Phi_{\bar{h}}^{-1} \left(\tau_{[\bar{L}_I, \bar{R}_I]} \left(\frac{1}{u} \sum_{j=1}^u \bar{V}_j \right) \right) \quad (3.12)$$

where $\tau_{[\bar{L}_I, \bar{R}_I]}$ is the truncation function defined in (3.3).

Case 3: $B > \log \frac{1}{\sigma_n} + m$. We only need to use part of the total communication budget B as if we deal with the case $B = \lfloor \log \frac{1}{\sigma_n} \rfloor + m$. To be precise, we can always easily find b'_1, b'_2, \dots, b'_m so that $1 \leq b'_i \leq b_i$ for $i = 1, 2, \dots, m$ and

$$\sum_{i=1}^m b'_i = \lfloor \log \frac{1}{\sigma_n} \rfloor + m.$$

Then we can implement the procedure introduced in Case 2 where we let the i -th machine only output a transcript of length b'_i .

MODGAME procedure when $\sigma_n \geq 1$

When $\sigma_n \geq 1$, each machine only need to output a one-bit measurement to achieve the global optimal rate as if there are no communication constraints. Some related results are available in Kipnis and Duchi (2019). The following procedure is based on the setting when $b_i = 1$ for all $i = 1, \dots, m$. If $b_i > 1$ for some i , then one can simply discard all remain bits so that only one bit is sent by each machine.

Here is the MODGAME procedure when $\sigma_n \geq 1$:

Step 1. The i -th machine outputs

$$Z_i = \begin{cases} 0 & \text{if } X_i < 0 \\ 1 & \text{if } X_i \geq 0 \end{cases}.$$

Step 2. The central machine collects Z_1, Z_2, \dots, Z_m and estimates θ by

$$\hat{\theta}_D = \tau_{[0,1]} \left(\sigma_n \Phi^{-1} \left(\frac{1}{m} \sum_{i=1}^m Z_i \right) \right)$$

where τ is the truncation function defined in (3.3) and Φ is the cumulative distribution function of a standard normal, $\Phi(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$. Here Φ^{-1} is the inverse of Φ and we extend it by defining $\Phi^{-1}(0) = -\infty$ and $\Phi^{-1}(1) = \infty$.

3.2.2. Theoretical properties of the MODGAME procedure

Section 3.2.1 gives a detailed construction of the MODGAME procedure, which clearly satisfies the communication constraints by construction. The following result provides a theoretical guarantee for the statistical performance of MODGAME.

Theorem 8. *For given communication budgets $b_{1:m}$ with $b_i \geq 1$ for $i = 1, \dots, m$, let $B = \sum_{i=1}^m b_i$ and let $\hat{\theta}_D$ be the MODGAME estimate. Then there exists a constant $C > 0$ such that*

$$\sup_{\theta \in [0,1]} \mathbb{E}(\hat{\theta}_D - \theta)^2 \leq \begin{cases} C \cdot 2^{-2B} & \text{if } B < \log \frac{1}{\sigma_n} + 2 \\ C \cdot \frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B < \log \frac{1}{\sigma_n} + m \cdot \\ C \cdot \left(\frac{\sigma_n^2}{m} \wedge 1 \right) & \text{if } B \geq \log \frac{1}{\sigma_n} + m \end{cases} \quad (3.13)$$

An interesting and somewhat surprising feature of the upper bound is that it depends on the communication constraints $b_{1:m}$ only through the total budget $B = \sum_{i=1}^m b_i$, not the specific value of $b_{1:m}$, so long as each machine can transmit at least one bit. The proof of Theorem 8 is provided in the Supplementary Material (Cai and Wei, 2020a).

3.2.3. Lower bound analysis and discussions

Section 3.2.1 gives a detailed construction of the MODGAME procedure and Theorem 8 provides a theoretical guarantee for the estimator. We shall now prove that MODGAME is indeed rate optimal among all estimators satisfying the communication constraints by

showing that the upper bound in Equation (3.13) cannot be improved. More specifically, the following lower bound provides a fundamental limit on the estimation accuracy under the communication constraints.

Theorem 9. *Suppose $b_i \geq 1$ for all $i = 1, 2, \dots, m$. Let $B = \sum_{i=1}^m b_i$. Then there exists a constant $c > 0$ such that*

$$R_1(b_{1:m}) \geq \begin{cases} c \cdot 2^{-2B} & \text{if } B < \log \frac{1}{\sigma_n} + 2 \\ c \cdot \frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B < \log \frac{1}{\sigma_n} + m \\ c \cdot \left(\frac{\sigma_n^2}{m} \wedge 1 \right) & \text{if } B \geq \log \frac{1}{\sigma_n} + m. \end{cases}$$

The key novelty in the lower bound analysis is the decomposition of the statistical risk into *localization* error and *refinement* error based on a delicate construction of the following candidate set G_δ :

$$G_\delta \triangleq \left\{ \theta_{u,v} = \sigma_n u + \delta v : u = 0, 1, 2, \dots, \left(\lfloor \frac{1}{\sigma_n} \rfloor - 1 \right), v = 0, 1 \right\},$$

where δ is a precision parameter that will be specified later. By assigning θ a uniform prior on the candidate set G_δ , estimation of θ can be decomposed into estimation of u and v . One can view estimation of u as the *localization* step and estimation of v as the *refinement* step. The following lemma is a key technical tool.

Lemma 5. *Let $0 < \sigma_n < 1$ and let u be uniformly distributed on $\{0, 1, \dots, \lfloor \frac{1}{\sigma_n} \rfloor - 1\}$ and v be uniformly distributed on $\{0, 1\}$. Let u and v be independent and let $\theta = \theta_{u,v} = \sigma_n u + \delta v$ where $0 < \delta < \frac{\sigma_n}{8}$. Then for all $\hat{\theta} \in \mathcal{A}_{ind}(b_{1:m})$,*

$$I(\hat{\theta}; u) + \frac{\sigma_n^2}{64\delta^2} I(\hat{\theta}; v) \leq B. \quad (3.14)$$

Remark 1. The proof of Lemma 5 mainly relies on the strong data processing inequality

(Lemma 14 in Cai and Wei (2020a)). The strong data processing inequality was originally developed in information theory, for reference see Raginsky (2016). Zhang et al. (2013a) and Braverman et al. (2016) applied this technical tool to obtain lower bounds for distributed mean estimation. However, their lower bounds are not sharp when σ_n is very small, due to the fact that the focus was on bounding the refinement error using the strong data processing inequality, but failed to bound the localization error.

Lemma 5 suggests that under the communication constraints $b_{1:m}$, there is an unavoidable trade-off between the mutual information $I(\hat{\theta}; u)$ and $I(\hat{\theta}; v)$. So one of the above two quantities must be "small". When $I(\hat{\theta}; u)$ (or $I(\hat{\theta}; v)$) is smaller than a certain threshold, it can be shown that the estimator $\hat{\theta}$ cannot accurately estimate u (or v), which means the *localization* error (or the *refinement* error) is large. Given that one of *localization* error and *refinement* error must be larger than a certain value, the desired lower bound follows. A detailed proof of Theorem 9 is given in Section 3.8.

Minimax rate of convergence. Theorems 8 and 9 together yield a sharp minimax rate for distributed univariate Gaussian mean estimation with independent protocols:

$$\inf_{\hat{\theta} \in \mathcal{A}_{ind}(b_{1:m})} \sup_{\theta \in [0,1]} \mathbb{E}(\hat{\theta} - \theta)^2 \asymp \begin{cases} 2^{-2B} & \text{if } B < \log \frac{1}{\sigma_n} + 2 \\ \frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B < \log \frac{1}{\sigma_n} + m \\ \frac{\sigma_n^2}{m} \wedge 1 & \text{if } B \geq \log \frac{1}{\sigma_n} + m \end{cases} \quad (3.15)$$

The results also show that MODGAME is rate optimal.

The minimax rate only depends on the total communication budgets $B = \sum_{i=1}^m b_i$. As long as each transcript contains at least one bit, how these communication budgets are allocated to local machines does not affect the minimax rate. This surprising phenomenon is due to the symmetry among the local machines since samples on different machines are independent and identically distributed.

Remark 2. Figure 3.2 gives an illustration for the minimax rate (3.15), which is divided into three phases: localization, refinement, and optimal-rate. The minimax risk decreases quickly in the localization phase, when the communication constraints are extremely severe; then it decreases slower in the refinement phase, when there are more communication budgets; finally the minimax rate coincides with the centralized optimal rate (Bickel, 1981b) and stays the same, when there are sufficient communication budgets. The value for each additional bit decreases as more bits are allowed.

In the localization phase, the risk is reduced to as small as $O(\sigma_n^2)$, which can be achieved by using the sample on only ONE machine and there is no need to “communicate” with multiple machines. In the refinement phase, the risk is further reduced to $O(\sigma_n^2/m)$. However, one must aggregate information from all machines in order to achieve this rate.

Remark 3. If the central machine itself also has an observation, or equivalently if one of the local machines serves as the central machine, then the communication constraints can be viewed as one of b_i is equal to infinity. This setting is considered in some related literature, for instance, see Jordan et al. (2019). Then according to Theorem 8, MODGAME always achieves the centralized rate $\frac{\sigma_n^2}{m} \wedge 1$, as long as at least one bit is allowed to communicate with each local machine.

Remark 4. Our analysis on the minimax rate can be generalized to the l_r loss for any $r \geq 1$, with suitable modifications on both the lower bound analysis and optimal procedure.

3.3. Distributed Multivariate Gaussian Mean Estimation

We turn in this section to distributed estimation of a multivariate Gaussian mean under the communication constraints. Similar to the univariate case, suppose we observe on m local machines i.i.d. random samples:

$$X_i \stackrel{\text{i.i.d.}}{\sim} N_d(\theta, \sigma_n^2 I_d), \quad i = 1, \dots, m,$$

where the i -th machine has access to X_i only. Here we consider the multivariate Gaussian location family

$$\mathcal{P}_{\sigma_n}^d = \left\{ N_d(\theta, \sigma_n^2 I_d) : \theta \in [0, 1]^d \right\},$$

where $\theta \in [0, 1]^d$ is the mean vector of interest and the noise level σ_n is known. Under the constraints on the communication budgets $b_{1:m}$ with $b_i \geq 1$ for $i = 1, \dots, m$, the goal is to optimally estimate the mean vector θ under the squared error loss. We are interested in the minimax risk for the distributed protocol $\mathcal{A}_{ind}(b_{1:m})$:

$$R_d(b_{1:m}) = \inf_{\hat{\theta} \in \mathcal{A}_{ind}(b_{1:m})} \sup_{\theta \in [0, 1]^d} \mathbb{E} \|\hat{\theta} - \theta\|^2.$$

Another goal is to develop a rate-optimal estimator that satisfies the communication constraints. The multivariate case is similar to the univariate setting, but it also has some distinct features, both in terms of the estimation procedure and the lower bound argument.

3.3.1. Lower bound analysis

We first obtain the minimax lower bound which is instrumental in establishing the optimal rate of convergence. The following lower bound on the minimax risk shows a fundamental limit on the estimation accuracy when there are communication constraints. In view of the upper bound to be given in Section 3.3.2 that is achieved by a generalization of the MODGAME procedure, the lower bound is rate optimal.

Theorem 10. *Suppose $b_i \geq 1$ for all $i = 1, 2, \dots, m$. Let $B = \sum_{i=1}^m b_i$ and $m' = \frac{1}{d} \sum_{i=1}^m (b_i \wedge d)$, then there exists a constant $c > 0$ such that*

$$R_d(b_{1:m}) \geq \begin{cases} c \cdot 2^{-2B/d} d & \text{if } B/d < \log \frac{1}{\sigma_n} + 2 \\ c \cdot \frac{d\sigma_n^2}{(B/d - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B/d < \log \frac{1}{\sigma_n} + (m' \vee 2) \cdot \\ c \cdot d \left(\frac{\sigma_n^2}{m'} \wedge 1 \right) & \text{if } B/d \geq \log \frac{1}{\sigma_n} + (m' \vee 2) \end{cases}$$

A detailed proof of Theorem 10 is given in the Supplementary Material (Cai and Wei, 2020a).

Remark 5. In the earlier work including Garg et al. (2014); Barnes et al. (2019b), a lower bound for distributed Gaussian mean estimation has been established as $\Omega(\frac{\sigma_n^2 d^2}{B})$, where B is the total communication cost. This lower bound is sharp for $\sigma_n \geq 1$. However, when $\sigma_n < 1$, by showing that the additional and exact $\log(1/\sigma_n)$ localization bits are necessary for estimating a Gaussian mean, the lower bound can be improved to $\Omega(\min\{\frac{\sigma_n^2 d^2}{B-d \log(1/\sigma_n)}, \sigma_n^2 d\})$. The improvement is significant when $\log(1/\sigma_n)/m$ is bounded away from 0.

3.3.2. Optimal procedure

We now construct an estimator of the mean vector under the communication constraints. Roughly speaking, the procedure, called multi-MODGAME, first divides the communication budgets evenly into d parts and then each part of communication budgets will be used to estimate one coordinate of θ . Our analysis shows that multi-MODGAME achieves the mini-max optimal rate under the communication constraints. The construction of the distributed estimator $\hat{\theta}_D$ is divided into three steps.

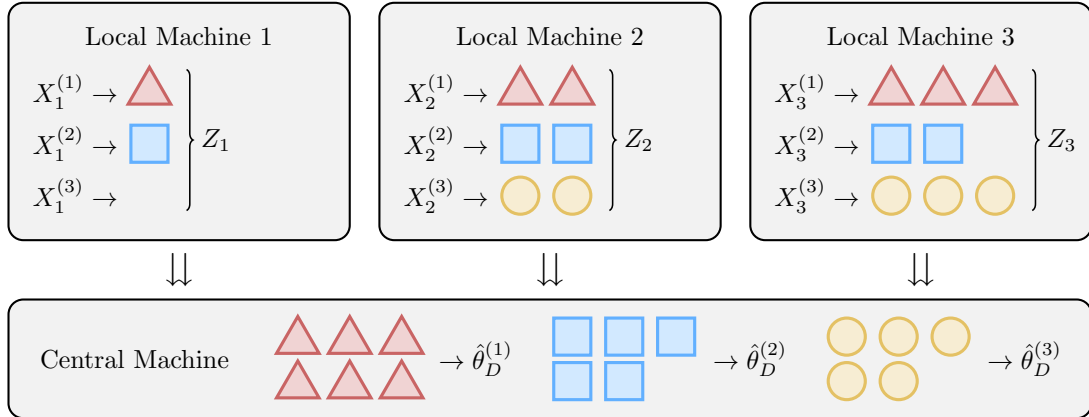


Figure 3.5: An illustration for multi-MODGAME. Communication budgets are evenly divided into three parts with each part used for estimating a coordinate of θ by the MODGAME procedure.

Step 1: *Assign communication budgets.* In this step we will calculate $b_i^{(k)}$ ($i = 1, 2, \dots, m; k = 1, 2, \dots, d$) so that

$$b_i = b_i^{(1)} + b_i^{(2)} + \dots + b_i^{(d)} \quad \text{for all } i = 1, 2, \dots, m.$$

where $b_i^{(k)}$ is the number of bits within the transcript Z_i which is associated with estimation

of $\hat{\theta}^{(k)}$.

Without loss of generality we assume $b_1 \leq b_2 \leq \dots \leq b_m$, which can always be achieved by permuting the indices of the machines. Write $1, 2, 3, \dots, d$ repeatedly to form a sequence:

$$Q \triangleq 1, 2, 3, \dots, d, 1, 2, 3, \dots, d, 1, 2, 3, \dots$$

The sequence Q is then divided into subsequences of lengths b_1, b_2, \dots, b_m . Let Q_1 be the subsequence of Q from index 1 to index b_1 ; let Q_2 be the next subsequence from index $b_1 + 1$ to $b_1 + b_2$; ... let Q_m be the subsequence from index $\sum_{i=1}^{m-1} b_i + 1$ to $\sum_{i=1}^m b_i$. For each $1 \leq k \leq d$, let $b_i^{(k)}$ be the number of occurrence of k within Q_i . To be more precise, $b_i^{(k)}$ can be calculated by

$$b_i^{(k)} = \left\lfloor \frac{\sum_{j=1}^i b_j - k}{d} \right\rfloor - \left\lfloor \frac{\sum_{j=1}^{i-1} b_j - k}{d} \right\rfloor.$$

Step 2: *Generate transcripts on local machines.* On the i -th machine, the transcript Z_i is concatenated by short transcripts $Z_i^{(1)}, Z_i^{(2)}, \dots, Z_i^{(d)}$, where the length of $Z_i^{(k)}$ is $b_i^{(k)}$ for $k = 1, 2, \dots, d$. Note that the k -th coordinate of the observations on each machine, $X_1^{(k)}, X_2^{(k)}, \dots, X_m^{(k)}$, can be viewed as i.i.d univariate Gaussian variables with mean $\theta^{(k)}$ and variance σ_n^2 . For $1 \leq k \leq d$, the transcripts $Z_1^{(k)}, Z_2^{(k)}, \dots, Z_m^{(k)}$ can be generated the same way as if we implement MODGAME to estimate $\theta^{(k)}$ from observations $X_1^{(k)}, X_2^{(k)}, \dots, X_m^{(k)}$, within the communication budgets $b_1^{(k)}, b_2^{(k)}, \dots, b_m^{(k)}$. Some machines may be assigned zero communication budget, if that happens those machines are ignored and the procedure is implemented as if there are fewer machines.

Step 3: *Construct distributed estimator $\hat{\theta}_D$.* We have collected $Z_i^{(k)}$ ($i = 1, 2, \dots, m; k = 1, 2, \dots, d$) from the m local machines. For $1 \leq k \leq d$, as in MODGAME, one can use $Z_1^{(k)}, Z_2^{(k)}, \dots, Z_m^{(k)}$ to obtain an estimate for $\hat{\theta}^{(k)}$:

$$\hat{\theta}_D^{(k)} = \hat{\theta}_D^{(k)} \left(Z_1^{(k)}, Z_2^{(k)}, \dots, Z_m^{(k)} \right).$$

The final multi-MODGAME estimator $\hat{\theta}_D$ of the mean vector θ is just the vector consisting of the estimates for the d coordinates:

$$\hat{\theta}_D \triangleq \left(\hat{\theta}_D^{(1)}, \hat{\theta}_D^{(2)}, \dots, \hat{\theta}_D^{(d)} \right).$$

The following result provides a theoretical guarantee for multi-MODGAME.

Theorem 11. *Let $B = \sum_{i=1}^m b_i$ and $m' = \frac{1}{d} \sum_{i=1}^m (b_i \wedge d)$. Then there exists a constant $C > 0$ such that*

$$\sup_{\theta \in [0,1]^d} \mathbb{E} \|\hat{\theta}_D - \theta\|^2 \leq \begin{cases} C \cdot 2^{-2B/d} & \text{if } B/d < \log \frac{1}{\sigma_n} + 2 \\ C \cdot \frac{d\sigma_n^2}{(B/d - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B/d < \log \frac{1}{\sigma_n} + (m' \vee 2) \\ C \cdot d \left(\frac{\sigma_n^2}{m'} \wedge 1 \right) & \text{if } B/d \geq \log \frac{1}{\sigma_n} + (m' \vee 2). \end{cases} \quad (3.16)$$

Remark 6. Compared to the state-of-art results in the literature including Braverman et al. (2016), the multi-MODGAME procedure is more communication-efficient and more flexible in communication budget allocation. To be specific, the algorithm proposed in Braverman et al. (2016) achieves the mean squared error $O(\frac{\sigma_n^2 d}{\alpha m})$ with the total communication cost of order $\alpha m d + d \log^2(\alpha m d / \sigma_n)$. In comparison, to achieve the same statistical performance, MODGAME only needs $\alpha m d + d \log(1/\sigma_n)$ bits. The difference could be significant when $\sigma_n \ll 1$.

Moreover, multi-MODGAME achieves the optimal statistical performance in the distributed setting with any pre-specified communication budget allocation (b_1, b_2, \dots, b_m) . That is, the constraint is imposed on each individual local machine. In comparison, the protocol in Braverman et al. (2016) assigns the total communication budget by the algorithm thus in a way solves a simpler "total communication constrained" problem.

The lower and upper bounds given Theorems 10 and 11 together establish the minimax rate

for distributed multivariate Gaussian mean estimation:

$$\inf_{\hat{\theta} \in \mathcal{A}_{ind}(b_{1:m})} \sup_{\theta \in [0,1]^d} \mathbb{E} \|\hat{\theta} - \theta\|^2 \asymp \begin{cases} 2^{-2B/d} d & \text{if } B/d < \log \frac{1}{\sigma_n} + 2 \\ \frac{d\sigma_n^2}{(B/d - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B/d < \log \frac{1}{\sigma_n} + (m' \vee 2) \\ d \left(\frac{\sigma_n^2}{m'} \wedge 1 \right) & \text{if } B/d \geq \log \frac{1}{\sigma_n} + (m' \vee 2) \end{cases} \quad (3.17)$$

where $B = \sum_{i=1}^m b_i$ is the total communication budget and $m' = \frac{1}{d} \sum_{i=1}^m (b_i \wedge d)$ is the “effective sample size”. In particular, the minimax rate (3.15) for the univariate case is an special case for the above minimax rate (3.17) with $d = 1$.

Remark 7. Different from the univariate case, in the multivariate case the minimax rate depends on not only the total communication budget B , but also the effective sample size m' . How the communication budgets assigned to individual local machines affects the difficulty of the estimation problem. If the communication budgets are tight on some machines, then one may have $m' \ll m$, which means the centralized minimax rate cannot be achieved even if the total communication budget B is sufficiently large.

Remark 8. This chapter focuses on the unit hypercube $[0, 1]^d$ as the parameter space. A similar analysis can be applied to other “regular” shape constraints, such as a ball or a simplex, and the minimax rate depends on the constraint.

3.4. Optimal Distributed Estimation with Sequential and Blackboard Protocols

Independent protocols are considered as a “non-interactive” communication strategy, where each machine can only access its own samples. However, feedback could be helpful in the learning process. There are other more general and interactive communication protocols considered in the literature, including the sequential protocols and blackboard protocols (Zhang et al., 2013a; Barnes et al., 2019b).

- **Sequential protocols.** The local machines sequentially send transcripts to the next

local machine, and finally the central machine collects all the transcripts. The transcript Z_i sent by the i -th local machine, which is at most b_i bits, can depend on the previous transcripts Z_1, Z_2, \dots, Z_{i-1} .

- **Blackboard protocols.** The local machines communicate via a publicly shown blackboard. When a local machine writes a message on the blackboard, all other local machines can see the content. Finally, the central machine collects all the information and outputs the final estimate. The total length of the messages written by the i -th local machine is at most b_i bits.

As for distributed estimation with the independent protocols, it is interesting to establish the optimal rates of convergence for the sequential protocols and blackboard protocols. This is also related to a question of both theoretical and practical interest: is feedback useful for distributed Gaussian mean estimation?

Note that any independent protocol can be viewed as a sequential protocol (by ignoring messages provided by the previous machines). Similarly, any sequential protocol can be implemented as a blackboard protocol. Therefore, the upper bounds (3.13) for the proposed MODGAME procedure and (3.16) for multi-MODGAME still hold over the class of sequential protocols and blackboard protocols. The question is: Can these bounds be improved by using more sophisticated algorithms?

The answer is no. The following theorem provides a lower bound for d -dimensional distributed Gaussian mean estimation with blackboard protocols. We denote by $\mathcal{A}_{seq}(b_{1:m})$ and $\mathcal{A}_{BB}(b_{1:m})$ the class of sequential protocols and blackboard protocols respectively.

Theorem 12. *Suppose $b_i \geq 1$ for all $i = 1, 2, \dots, m$. Let $B = \sum_{i=1}^m b_i$ and $m' = \frac{1}{d} \sum_{i=1}^m (b_i \wedge$*

d), then there exists a constant $c > 0$ such that

$$\inf_{\hat{\theta} \in \mathcal{A}_{BB}(b_{1:m})} \sup_{\theta \in [0,1]^d} \mathbb{E} \|\hat{\theta} - \theta\|^2 \geq \begin{cases} c \cdot 2^{-2B/d} d & \text{if } B/d < \log \frac{1}{\sigma_n} + 2 \\ c \cdot \frac{d\sigma_n^2}{(B/d - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B/d < \log \frac{1}{\sigma_n} + (m' \vee 2) \cdot \\ c \cdot d \left(\frac{\sigma_n^2}{m'} \wedge 1 \right) & \text{if } B/d \geq \log \frac{1}{\sigma_n} + (m' \vee 2) \end{cases}$$

The proof of Theorem 12 is also based on the localization-refinement error decomposition. A sketch of the proof is given in the Supplementary material Cai and Wei (2020a). Theorem 12 and the upper bound given in (3.16) together show that the optimal rate of convergence is the same and MODGAME and multi-MODGAME are rate-optimal for the three classes of communication protocols— independent, sequential, and blackboard. To some extent, the results imply that feedback is not necessary to achieve communication-efficiency for distributed Gaussian mean estimation.

3.5. Robustness Against Departures from Gaussianity

We have so far focused exclusively on the Gaussian location families. Both the optimal distributed procedures and lower bound arguments are established under the assumption of Gaussian observations. We consider in this section robustness of the proposed MODGAME and multi-MODGAME procedures against departures from Gaussianity.

Even if the i.i.d observations $X_{i,j}, i = 1, 2, \dots, m, j = 1, 2, \dots, n$ are drawn from a non-Gaussian distribution, after taking the sample mean on each local machine, according to the central limit theorem, the distribution of these sample means is close to a Gaussian distribution when n is large. Thus intuitively the proposed procedures should still work even when the original observations are nongaussian.

For simplicity we focus on the one-dimensional estimation problem. The multivariate case can be considered as a direct generalization to the univariate case. Let P_θ be a location family where θ is the mean, and its variance is σ^2 . Denote \bar{P}_θ^n as the distribution of the mean of n i.i.d copies drawn from P_θ . If on each local machine we can access to n i.i.d

observations $X_{i,1}, X_{i,2}, \dots, X_{i,n} \sim P_\theta$, then each machine can take the local sample mean $X_i \triangleq \sum_{j=1}^n X_{i,j} \sim \bar{P}_\theta^n$. Even though \bar{P}_θ^n is a non-Gaussian distribution, the MODGAME procedure can take X_i as inputs to generate a final estimate.

Recall that MODGAME is divided into three steps: crude localization step, finer localization step, and refinement step. During the first two steps, in order to obtain the desired statistical guarantee for the “confidence interval” I_2 , we only need sub-Gaussian tail condition for X_i . During the refinement step, the key is to use Φ_h or $\Phi_{\bar{h}}$ to generate estimates from the one-bit measurements. If X_i is not drawn from a Gaussian distribution, there is additional bias that could be controlled under certain conditions.

Let $TV(\cdot, \cdot)$ denote the total variation distance between two probability distributions. A random variable X (or a distribution P where $X \sim P$) is called v -subgaussian if $\mathbb{E} \exp(s(X - \mathbb{E}X)) \leq \exp(\frac{v^2 s^2}{2})$, $\forall s \in \mathbb{R}$. The following theorem shows that when the total variation distance between the distribution \bar{P}_θ^n of the local sample mean and the Gaussian distribution $N(\theta, \sigma_n^2)$ is sufficiently small, MODGAME has the same theoretical guarantee as in the Gaussian case. This implies that MODGAME is robust against departures from the Gaussian distribution.

Theorem 13. *If \bar{P}_θ^n is a $D\sigma_n$ -subgaussian distribution and $TV(\bar{P}_\theta^n, N(\theta, \sigma_n^2)) \leq \frac{D}{\sqrt{m}}$ for some $D > 0$. Then there exists a constant $C > 0$ such that*

$$\sup_{\theta \in [0,1]} \mathbb{E}(\hat{\theta} - \theta)^2 \leq C \cdot \begin{cases} 2^{-2B} & \text{if } B < \log \frac{1}{\sigma_n} + 2 \\ \frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B < \log \frac{1}{\sigma_n} + m \\ \left(\frac{\sigma_n^2}{m} \wedge 1\right) & \text{if } B \geq \log \frac{1}{\sigma_n} + m \end{cases} \quad (3.18)$$

where $\hat{\theta}$ is the output of the MODGAME procedure and $B = \sum_{i=1}^m b_i$ is the total communication cost.

A sketch of the proof is given in the Supplementary Material Cai and Wei (2020a). Note that $X_i \sim \bar{P}_\theta^n$ is the mean of i.i.d observations in the i th local machine. The L_1 Berry-Esseen

bound (e.g. (Chen et al., 2010, Corollary 4.2)) suggests $TV(\bar{P}_\theta^n, N(\theta, \sigma_n^2)) \leq \frac{\mathbb{E}(|X_1 - \theta|/\sigma)^3}{2\sqrt{n}}$. If X_1 is a $D\sigma$ -subgaussian distribution, then $\mathbb{E}(|X_1 - \theta|/\sigma)^3$ is bounded by a constant (depending on D). Hence the following corollary holds.

Corollary 14. *If P_θ is a $D\sigma$ -subgaussian distribution, and $m \leq Dn$ for some $D > 0$. Then there exist a constant $C > 0$ such that*

$$\sup_{\theta \in [0,1]} \mathbb{E}(\hat{\theta} - \theta)^2 \leq C \cdot \begin{cases} 2^{-2B} & \text{if } B < \log \frac{1}{\sigma_n} + 2 \\ \frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B < \log \frac{1}{\sigma_n} + m \\ \left(\frac{\sigma_n^2}{m} \wedge 1\right) & \text{if } B \geq \log \frac{1}{\sigma_n} + m \end{cases} \quad (3.19)$$

where $\hat{\theta}$ is the output of the MODGAME procedure. $B = \sum_{i=1}^m b_i$ is the total communication cost.

Corollary 14 shows that, if n/m is asymptotically bounded away from 0, then MODGMAE achieves the same statistical performance as in the Gaussian case as long as the observations are drawn from a subgaussian distribution.

3.6. Simulation Studies

It is clear by construction that MODGAME and multi-MODGAME satisfy the communication constraints and are easy to implement. We investigate in this section their numerical performance through simulation studies. Comparisons with the existing methods are given and the results are consistent with the theory. In this section, we implement a slightly modified version of MODGAME procedure, where each local machine output three refinement bits instead of one. This slightly modified MODGAME procedure has better numerical performance and also has the same theoretical guarantee as what is stated in Section 3.2.

We first consider MODGAME for estimating a univariate Gaussian mean. In this case, we set $d = 1$ and $b_1 = b_2 = \dots = b_m = b$, i.e. the communication budgets for all machines are equal, and compare the empirical MSEs of MODGAME, naive quantization (see e.g.

Zhang et al. (2013a)), and sample mean. For naive quantization, each machine projects its observation to $[0, 1]$ and quantizes it to precision 2^{-b} . The quantized observation is sent to the central machine and the central machine uses their average as the final estimate. The sample mean is the efficient estimate when there are no communication constraints, which can be viewed as a benchmark for any distributed Gaussian mean estimation procedure.

First, we fix $m = 100$, $\sigma_n = 2^{-8}$ and assign the communication budget for each machine b from 1 to 7. The MSEs of the three estimators are shown in Figure 3.6a, which shows that MODGAME makes better use of the communication resources in comparison to naive quantization. It can be seen from the figure, MODGAME outperforms naive quantization when the communication constraints are extremely severe. As the communication budgets increases, naive quantization can nearly achieve the optimal MSE, meanwhile MODGAME still performs very well.

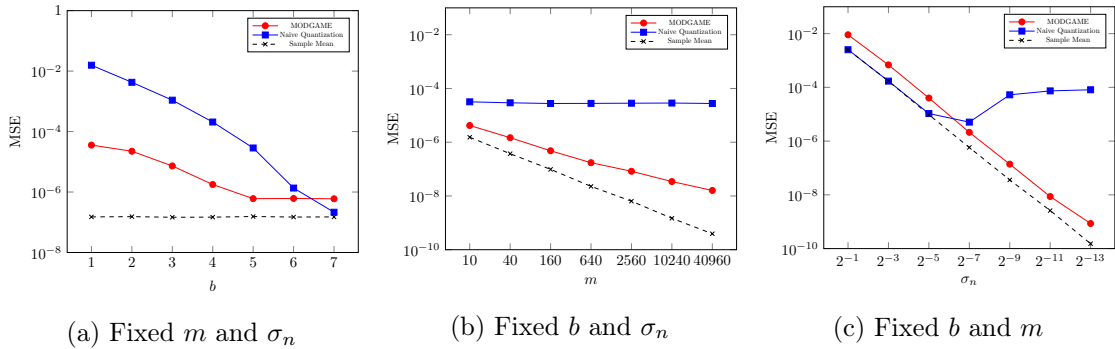


Figure 3.6: Comparisons of the MSEs of MODGAME (red), naive quantization (blue) and sample mean (black). MSEs are plotted on log-scale. In 3.6b and 3.6c, m and σ_n are plotted on log-scale.

In the second setting, we fix $\sigma_n = 2^{-8}$, $b = 5$ and vary the number of machines m from 10 to 40960. Figure 3.6b plots the MSEs of the three methods. The MSE of MODGAME decreases as number of machine increases and outperforms naive quantization; the MSE of naive quantization remains constant as the quantization error plays a dominant role in the MSE.

Finally, we fix $b = 5$, $m = 100$ and vary the standard deviation σ_n from 2^{-1} to 2^{-13} . Figure

3.6c shows the MSEs of the three estimators. It can be seen that MODGAME is robust for all choices of σ_n . The difference between the MSE of MODGAME and the optimal MSE for non-distributed sample mean is small. For naive quantization, it is as good as the optimal non-distributed sample mean when σ_n is large. However, as seen in the previous experiment, when σ_n is small, the MSE of naive quantization is dominated by the quantization error and is much larger than the MSE of MODGAME. In all three settings, it can be seen clearly that the MSE of MODGAME decreases as the communication budgets increases. This is consistent with the theoretical results established in Section 3.2 and demonstrates the tradeoff between the communication costs and statistical accuracy.

Besides, to demonstrate that the performance of the MODGAME procedure only depends on total communication budget B , we implement another simulation. We fix $m = 6$, $\sigma_n = 2^{-12}$ and assign the total communication budgets B from 18 to 36. We compare the performance of the MODGAME procedure with different communication allocation. That is, in one simulation we assign $b_i = 3$ bits to each local machine except one, and that one machine are assigned $B - 3(m - 1)$ bits. In another simulation we assign equal communication budget $b_i = B/m$ to each machine. As a benchmark, we also implement non-distributed sample mean estimator. Figure 3.7a shows the MSEs of the above three methods. It is shown clearly that how communication budgets are assigned to local machines doesn't affect the performance of the MODGAME procedure, which is consistent with our theory.

We now turn to multi-MODGAME. Different values of the dimension d yield similar phenomena. We use $d = 50$ here for illustration. When d is larger than the number of bits that is allowed to communicate on each machine, naive quantization is not valid as it is unclear how to quantize the d coordinates of the observed vector. As a comparison, it can be seen in the following experiments that multi-MODGAME still performs well even if d is large and the communication budgets are tight.

Same as before, we set $b_1 = b_2 = \dots = b_m = b$, i.e. the communication budgets for all machines are equal. We set $d = 50$, $\sigma_n = 2^{-8}$, $m = 25$ and assign the communication

budgets b for each machine from 2 to 21. The MSEs of different methods are shown in Figure 3.7b. A phase transition at $b = 10$ can be clearly seen. When $b \leq 10$, the MSE decreases quickly at an exponential rate. When $b > 10$, the decrease becomes relatively slow. This phenomenon is consistent with the theoretical prediction that different phases appear in the convergence rate for multi-MODGAME (Theorem 11).

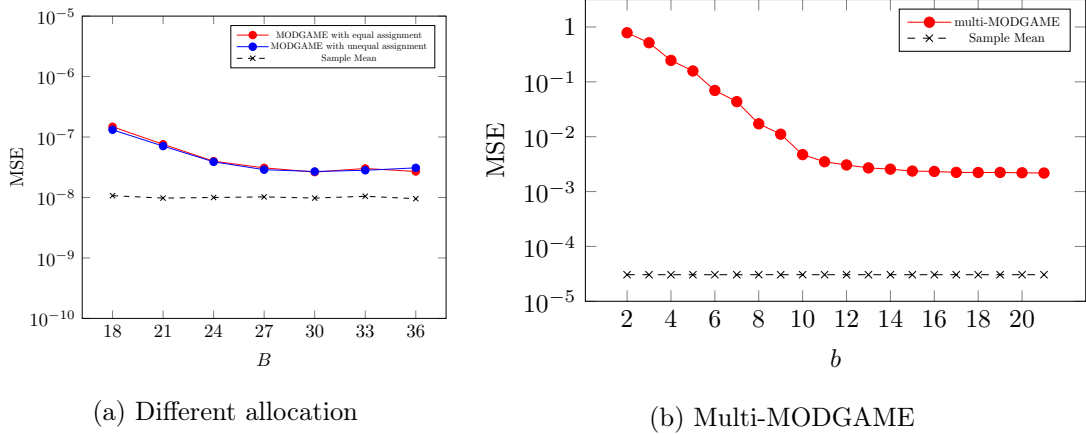


Figure 3.7: Left panel: Comparisons of the MSEs of MODGAME with equal assignment (red), MODGAME with unequal assignment (blue) and sample mean (black). Right panel: Comparisons of the MSEs of multi-MODGAME (red) and sample mean (black). MSEs are plotted on log-scale.

3.7. Discussion

We established in this chapter a sharp and complete minimax rate that holds for all values of the parameters d, m, n, σ in all communication budget regimes under the independent, sequential, and blackboard protocols. A key technique is the decomposition of the minimax estimation problem into two steps, *localization* and *refinement*, which appears in both the lower bound analysis and optimal procedure design. The optimality results and techniques developed can be useful for solving other problems such as distributed nonparametric function estimation and distributed sparse signal recovery.

In spite of these optimality results, there are still several open problems on distributed Gaussian mean estimation. For example, an interesting problem is the optimal estimation of the mean θ when the variance σ^2 is unknown. The lack of knowledge of σ^2 requires additional communication efforts for optimally estimating θ . When there are more than one

sample available on each local machine, a natural approach is to estimate σ^2 on each local machine and then use MODGAME to estimate θ . It would be interesting to investigate the performance of such an estimator. Other than estimating the mean θ , distributed estimation of the variance σ^2 is also an interesting and important problem. When there are multiple samples on each local machine, the local estimate of σ^2 can be viewed as an observation drawn from a scaled χ^2 distribution. The problem then becomes a distributed χ^2 estimation problem and it might be solved by using a similar approach to the one used in this chapter. We leave these for future work.

Optimal estimation of the mean of a multivariate Gaussian distribution with a general (known) covariance matrix is another interesting problem. A naive approach is to ignore the dependency and apply MODGAME to estimate the coordinates individually, this is arguably not communication efficient in general. For instance, if the correlation between certain coordinates is large, it may be possible to save a significant amount of communication budget by utilizing the information from one coordinate to help estimate the other. Another approach is to use multi-MODGAME after orthogonalization. More specifically, consider the Gaussian location family with a general non-singular covariance matrix Σ . Let $\lambda_{\min} > 0$ be the smallest eigenvalue of Σ . For $X \sim N_d(\theta, \Sigma)$, $\lambda_{\min}^{1/2}(d\Sigma)^{-1/2}X \sim N_d\left(\lambda_{\min}^{1/2}(d\Sigma)^{-1/2}\theta, \frac{\lambda_{\min}}{d}I_d\right)$. Note that $\lambda_{\min}^{1/2}(d\Sigma)^{-1/2}\theta \in [0, 1]^d$ for any $\theta \in [0, 1]^d$, therefore one can apply multi-MODGAME to estimate $\lambda_{\min}^{1/2}(d\Sigma)^{-1/2}\theta$, then transform it back to get an estimate for θ . However, this is generally not rate-optimal. A systematic study is needed for this problem. Another related and more challenging problem is optimal distributed estimation of the covariance matrix Σ .

This chapter arguably considered one of the simplest settings for optimal distributed estimation under the communication constraints, but as can be seen in the chapter, both the construction of the rate optimal estimators and the theoretical analysis are already quite involved for such a seemingly simple problem. As we deepen our understanding on distributed learning under the communication constraints, we hope to extent this line of work to investigate other statistical problems in distributed settings, including nonparametric function

estimation, high-dimensional linear regression, and large-scale multiple testing. For Gaussian mean estimation, as we showed in this chapter, the optimal rates of convergence under the three different communication protocols – independent, sequential, and blackboard – are the same. In some more complicated problems, feedback might be useful in improving estimation accuracy and the optimal rates will thus be different under these three classes of communication protocols. It is interesting to understand fully when and to what extent feedback helps in terms of improving statistical accuracy.

3.8. Proofs

In this section we prove Theorem 9 for the univariate case. For reasons of space, Theorems 8, 10, 11, 12, 13 and the technical lemmas are proved in the Supplementary Material (Cai and Wei, 2020a).

We prove separately the three cases in Theorem 9: $B < \log \frac{1}{\sigma_n} + 2$, $\log \frac{1}{\sigma_n} + 2 \leq B < \log \frac{1}{\sigma_n} + m$, and $B \geq \log \frac{1}{\sigma_n} + m$. We first focus on the most important case $\log \frac{1}{\sigma_n} + 2 \leq B < \log \frac{1}{\sigma_n} + m$. New technical tools are developed in the proof. The other two cases are relatively easy.

Case 1: $\log \frac{1}{\sigma_n} + 2 \leq B < \log \frac{1}{\sigma_n} + m$. Note that $b_i \geq 1$ for all $i = 1, 2, \dots, m$ implies that $B = \sum_{i=1}^m b_i \geq m$. Therefore in this case we must have $\sigma_n < 1$.

Let $0 < \delta < \frac{1}{8}\sigma_n$ be a parameter to be specified later. Define a grid of candidate values of θ as

$$G_\delta \triangleq \left\{ \theta_{u,v} = \sigma_n u + \delta v : u = 0, 1, 2, \dots, \left(\lfloor \frac{1}{\sigma_n} \rfloor - 1 \right), v = 0, 1 \right\}. \quad (3.20)$$

Let $\mathbb{U}(G_\delta)$ be a uniform prior of θ on G_δ . Note that $G_\delta \subset [0, 1]$, so the minimax risk is lower bounded by the Bayesian risk:

$$\inf_{\hat{\theta} \in \mathcal{A}(b_{1:m})} \sup_{\theta \in [0,1]} (\hat{\theta} - \theta)^2 \geq \inf_{\hat{\theta} \in \mathcal{A}(b_{1:m})} \mathbb{E}_{\theta \sim \mathbb{U}(G_\delta)} (\hat{\theta} - \theta)^2. \quad (3.21)$$

For any estimator $\hat{\theta} \in \mathcal{A}(b_{1:m})$, the rounded estimator $\hat{\theta}' \triangleq \operatorname{argmin}_{\tilde{\theta} \in G_\delta} |\tilde{\theta} - \hat{\theta}|$ always satisfy

$(\hat{\theta} - \theta)^2 \geq \frac{1}{4}(\hat{\theta}' - \theta)^2$ for all $\theta \in G_\delta$. Note that $\hat{\theta}'$ also belongs to the protocol class $\mathcal{A}(b_{1:m})$, and only takes value in G_δ , this implies

$$\inf_{\hat{\theta} \in \mathcal{A}(b_{1:m})} \mathbb{E}_{\theta \sim \mathbb{U}(G_\delta)} (\hat{\theta} - \theta)^2 \geq \frac{1}{4} \inf_{\hat{\theta} \in \mathcal{A}(b_{1:m}) \cap G_\delta} \mathbb{E}_{\theta \sim \mathbb{U}(G_\delta)} (\hat{\theta} - \theta)^2, \quad (3.22)$$

where $\mathcal{A}(b_{1:m}) \cap G_\delta$ is a shorthand for $\mathcal{A}(b_{1:m}) \cap \{\hat{\theta} : \hat{\theta} \text{ only takes value in } G_\delta\}$.

Now we have $\hat{\theta}, \theta \in G_\delta$ thus they can be reparametrized by $\hat{\theta} = \theta_{\hat{u}, \hat{v}}$ and $\theta = \theta_{u, v}$. It is easy to verify the inequality

$$(\hat{\theta}_{\hat{u}, \hat{v}} - \theta_{u, v})^2 \geq \max \left\{ \frac{\sigma_n^2}{4} (\hat{u} - u)^2, \delta^2 \mathbb{I}_{\{\hat{v} \neq v\}} \right\}.$$

Hence

$$\inf_{\hat{\theta} \in \mathcal{A}(b_{1:m}) \cap G_\delta} \mathbb{E}_{\theta \sim \mathbb{U}(G_\delta)} (\hat{\theta} - \theta)^2 \geq \inf_{\theta_{\hat{u}, \hat{v}} \in \mathcal{A}(b_{1:m}) \cap G_\delta} \mathbb{E}_{\theta_{u, v} \sim \mathbb{U}(G_\delta)} \max \left\{ \frac{\sigma_n^2}{4} (\hat{u} - u)^2, \delta^2 \mathbb{I}_{\{\hat{v} \neq v\}} \right\}. \quad (3.23)$$

Putting together (3.21), (3.22), and (3.23), we have

$$\begin{aligned} \inf_{\hat{\theta} \in \mathcal{A}(b_{1:m})} \sup_{\theta \in [0, 1]} (\hat{\theta} - \theta)^2 &\geq \frac{1}{4} \inf_{\theta_{\hat{u}, \hat{v}} \in \mathcal{A}(b_{1:m}) \cap G_\delta} \mathbb{E}_{\theta_{u, v} \sim \mathbb{U}(G_\delta)} \max \left\{ \frac{\sigma_n^2}{4} (\hat{u} - u)^2, \delta^2 \mathbb{I}_{\{\hat{v} \neq v\}} \right\} \\ &\geq \inf_{\theta_{\hat{u}, \hat{v}} \in \mathcal{A}(b_{1:m}) \cap G_\delta} \max \left\{ \frac{\sigma_n^2}{16} \mathbb{E}_{\theta_{u, v} \sim \mathbb{U}(G_\delta)} (\hat{u} - u)^2, \frac{\delta^2}{4} \mathbb{P}_{\theta_{u, v} \sim \mathbb{U}(G_\delta)} (\hat{v} \neq v) \right\}. \end{aligned} \quad (3.24)$$

Therefore, by assigning a prior $\theta \sim \mathbb{U}(G_\delta)$, we have successfully decomposed the estimation problem of θ into estimation problems of u and v . We can view estimation of u as "localization" step and estimation of v as "refinement" step, so (3.24) essentially has decomposed the statistical risk into localization error and refinement error. To lower bound the right hand side of (3.24), we show that under communication constraints, one cannot simultaneously estimate both u and v accurately, i.e. the localization and refinement errors cannot be both too small. Lemma 5, which shows that for any distributed estimator $\hat{\theta}$, there is unavoidable trade-off between the mutual information $I(\hat{\theta}; u)$ and $I(\hat{\theta}; v)$, is a key step.

We set $\delta = \frac{\sigma_n}{\sqrt{256(B+1-\log(\lfloor \frac{1}{\sigma_n} \rfloor))}}$, and assign the uniform prior $\mathbb{U}(G_\delta)$ to the parameter $\theta = \theta_{u,v}$. One can easily verify $\delta < \frac{1}{8}\sigma_n$, and u, v are independent random variables where u is uniform distributed on $\{0, 1, \dots, \lfloor \frac{1}{\sigma_n} \rfloor - 1\}$, and v is uniform distributed on $\{0, 1\}$. Therefore, we can apply Lemma 5 to get inequality (3.14). From the inequality (3.14) we can further get, for any $\hat{\theta} \in \mathcal{A}(b_{1:m}) \cap G_\delta$, one of the following two inequalities

$$I(\hat{\theta}; u) \leq \log(\lfloor \frac{1}{\sigma_n} \rfloor) - 1 \quad \text{or} \quad I(\hat{\theta}; v) \leq \frac{64\delta^2}{\sigma_n^2} \left(B + 1 - \log(\lfloor \frac{1}{\sigma_n} \rfloor) \right)$$

must hold. We show that either of the above bounds on the mutual information will result in a large statistical risk.

Case 1.1: $I(\hat{\theta}; u) \leq \log(\lfloor \frac{1}{\sigma_n} \rfloor) - 1$. Note that \hat{u} is a function on $\hat{\theta}$, thus by data processing inequality, $I(\hat{u}; u) \leq I(\hat{\theta}; u) \leq \log(\lfloor \frac{1}{\sigma_n} \rfloor) - 1$. Note that u is uniform distributed on $\{0, 1, \dots, \lfloor \frac{1}{\sigma_n} \rfloor - 1\}$, thus $H(u) = \log(\lfloor \frac{1}{\sigma_n} \rfloor)$. We have

$$H(u|\hat{u}) = H(u) - I(\hat{u}; u) \geq 1. \tag{3.25}$$

The following lemma shows that large conditional entropy will result in large L_2 distance between two integer-valued random variables.

Lemma 6. *Suppose A, D are two integer-valued random variables. If $H(A|D) \geq \frac{1}{2}$, then there exist a constant $c_2 > 0$ such that*

$$\mathbb{E}(A - D)^2 \geq c_2.$$

Given (3.25) and the fact that \hat{u}, u are integer valued, Lemma 6 yields

$$\mathbb{E}_{\theta_{u,v} \sim \mathbb{U}(G_\delta)} (\hat{u} - u)^2 \geq c_2. \tag{3.26}$$

Case 1.2: $I(\hat{\theta}; v) \leq \frac{\delta^2}{c_1 \sigma_n^2} (B + 1 - \log(\lfloor \frac{1}{\sigma_n} \rfloor))$. By the strong data processing inequality, plug in $\delta = \frac{\sigma_n}{\sqrt{256(B+1-\log(\lfloor \frac{1}{\sigma_n} \rfloor))}}$ we have $I(\hat{v}; v) \leq I(\hat{\theta}; v) \leq \frac{1}{4}$, so $H(v|\hat{v}) = H(v) - I(\hat{v}; v) \geq \frac{3}{4}$. It follows from Lemma 6 that

$$\mathbb{P}_{\theta_{u,v} \sim \mathbb{U}(G_\delta)}(\hat{v} \neq v) = \mathbb{E}_{\theta_{u,v} \sim \mathbb{U}(G_\delta)}(\hat{v} - v)^2 \geq c_2. \quad (3.27)$$

Combine (3.26) for Case 1.1 and (3.27) for Case 1.2 together, we have for any $\hat{\theta} \in \mathcal{A}(b_{1:m}) \cap G_\delta$,

$$\begin{aligned} & \max \left\{ \frac{\sigma_n^2}{16} \mathbb{E}_{\theta_{u,v} \sim \mathbb{U}(G_\delta)}(\hat{u} - u)^2, \frac{\delta^2}{4} \mathbb{P}_{\theta_{u,v} \sim \mathbb{U}(G_\delta)}(\hat{v} \neq v) \right\} \\ & \geq c_2 \min \left\{ \frac{\sigma_n^2}{16}, \frac{\delta^2}{4} \right\} = \frac{c_2 \sigma_n^2}{1024(B+1-\log(\lfloor \frac{1}{\sigma_n} \rfloor))} \geq \frac{c_2}{2048} \cdot \frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})}. \end{aligned} \quad (3.28)$$

The minimax lower bound follows by combining (3.24) and (3.28),

$$\inf_{\hat{\theta} \in \mathcal{A}(b_{1:m})} \sup_{\theta \in [0,1]} (\hat{\theta} - \theta)^2 \geq \frac{c_2}{2048} \cdot \frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})}.$$

Case 2: $B < \log \frac{1}{\sigma_n} + 2$. Let $S = 2^{B+1}$ and $K_S \triangleq \{\frac{i}{S} : i = 0, 1, \dots, S-1\}$. Denote by $\mathbb{U}(K_S)$ the uniform distribution on K_S . For the same reason as in (3.21) and (3.22) we have

$$\begin{aligned} \inf_{\hat{\theta} \in \mathcal{A}(b_{1:m})} \sup_{\theta \in [0,1]} (\hat{\theta} - \theta)^2 & \geq \inf_{\hat{\theta} \in \mathcal{A}(b_{1:m})} \mathbb{E}_{\theta \sim \mathbb{U}(K_S)}(\hat{\theta} - \theta)^2 \geq \frac{1}{4} \inf_{\hat{\theta} \in \mathcal{A}(b_{1:m}) \cap K_S} \mathbb{E}_{\theta \sim \mathbb{U}(K_S)}(\hat{\theta} - \theta)^2 \\ & = \frac{1}{4S^2} \inf_{\hat{\theta} \in \mathcal{A}(b_{1:m}) \cap K_S} \mathbb{E}_{\theta \sim \mathbb{U}(K_S)}(S\hat{\theta} - S\theta)^2. \end{aligned} \quad (3.29)$$

The parameter θ can be treated as a random variable drawn from $\mathbb{U}(K_S)$. Note that by the data processing inequality, for any $\hat{\theta} \in \mathcal{A}(b_{1:m})$,

$$I(\hat{\theta}; \theta) = I(\hat{\theta}(Z_1, Z_2, \dots, Z_m); \theta) \leq I(Z_1, Z_2, \dots, Z_m; \theta) \leq \sum_{i=1}^m H(Z_i) \leq B.$$

By $\theta \sim \mathbb{U}(K_S)$ we have $H(\theta|\hat{\theta}) = H(\theta) - I(\hat{\theta}; \theta) \geq \log S - B \geq 1$. Note that when $\theta \sim \mathbb{U}(K_S)$, for any $\hat{\theta} \in \mathcal{A}(b_{1:m}) \cap K_S$, $S\hat{\theta}$ and $S\theta$ both take value in $\{0, 1, 2, \dots, S-1\}$. Also we have $H(S\theta|S\hat{\theta}) = H(\theta|\hat{\theta}) \geq 1$. Therefore, Lemma 6 yields that $\mathbb{E}_{\theta \sim \mathbb{U}(K_S)}(S\hat{\theta} - S\theta)^2 \geq c_2$. We thus conclude that

$$\frac{1}{4S^2} \inf_{\hat{\theta} \in \mathcal{A}(b_{1:m}) \cap K_S} \mathbb{E}_{\theta \sim \mathbb{U}(K_S)} (S\hat{\theta} - S\theta)^2 \geq \frac{c_2}{4 \cdot 2^{2(B+1)}} = \frac{c_2}{16} \cdot 2^{-2B}.$$

The desired lower bound follows by plugging into (3.29).

Case 3: $B \geq \log \frac{1}{\sigma_n} + m$. The minimax risk for distributed protocols is always lower bounded by the minimax risk with no communication constraints:

$$\inf_{\hat{\theta} \in \mathcal{A}(b_{1:m})} \sup_{\theta \in [0,1]} (\hat{\theta} - \theta)^2 \geq \inf_{\hat{\theta}} \sup_{\theta \in [0,1]} (\hat{\theta} - \theta)^2 \asymp \frac{\sigma_n^2}{m} \wedge 1.$$

which is given in Bickel (1981b). □

CHAPTER 4

DISTRIBUTED GAUSSIAN MEAN ESTIMATION WITH UNKNOWN VARIANCE UNDER COMMUNICATION CONSTRAINTS

4.1. Introduction

Distributed statistical analysis is becoming increasingly important and challenging, as distributed data sets naturally arise in a range of applications due to size constraints, security concerns, or privacy considerations. For large-scale data analysis, communication costs can be expensive and become the main bottleneck in the learning process. When communication resources are limited, it is important to understand the interplay between the communication constraints and statistical accuracy in order to construct optimal estimation and inference procedures under the communication constraints.

Significant recent effort has been made to gain fundamental understanding of distributed estimation. For example, Zhang et al. (2013a); Garg et al. (2014); Braverman et al. (2016); Han et al. (2018); Barnes et al. (2019b) developed lower bound techniques for distributed parametric estimation. Zhu and Lafferty (2018); Szabó and van Zanten (2018, 2020); Cai and Wei (2020c, 2021b); Szabó et al. (2020) considered information-theoretical limits under communication constraints for various distributed problems, such as Gaussian mean estimation, linear regression, nonparametric regression and testing. Optimality results have been established under different communication constraints. Besides theoretical analysis, progress has also been made on developing practical methodologies for distributed estimation. See, for example, Kleiner et al. (2014); Deisenroth and Ng (2015); Lee et al. (2017); Diakonikolas et al. (2017); Jordan et al. (2019); Battey et al. (2018); Fan et al. (2019).

In this chapter we study distributed adaptive Gaussian mean estimation with unknown variance in a decision-theoretical framework. This is a basic yet fundamental distributed estimation problem. Gaussian mean estimation with known variance has been intensively studied in the distributed setting. See, for example, Garg et al. (2014); Braverman et al.

(2016); Barnes et al. (2019b); Cai and Wei (2020c). The optimality results in these papers were established in the non-adaptive setting where the variance of Gaussian observations is known a priori, and the estimation procedures and statistical lower bound arguments critically depend on the knowledge of variance. In a wide range of statistical applications, the variance of the observations is unknown and the procedures and results developed in the aforementioned papers are no longer applicable. Adaptive Gaussian mean estimation with unknown variance is technically challenging, and differs significantly from the non-adaptive setting. Understanding distributed adaptive Gaussian mean estimation with unknown variance also provides insight into other related statistical problems including distributed density estimation and distributed nonparametric regression with random design.

The primary goal of this chapter is to precisely characterize the minimal communication costs for adaptive Gaussian mean estimation without prior knowledge of variance under different types of distributed protocols, and construct communication-efficient estimators. Our analysis shows that the case of unknown variance differs significantly from the case when σ^2 is known. In particular, in sharp contrast to the known variance case, the behaviors of adaptive Gaussian mean estimation with unknown variance are very different under the independent and interactive protocols.

4.1.1. Distributed estimation framework and distributed protocols

We begin by introducing a general framework for distributed estimation by giving a formal definition of transcript, distributed estimator, and distributed protocols. Let $\mathcal{P} = \{P_{\theta, \xi} : \theta \in \Theta, \xi \in \Xi\}$ be a parametric family of distributions supported on space \mathcal{X} , where $\theta \in \Theta$ is the parameter of interest and $\xi \in \Xi$ are nuisance parameters. Suppose there are m local machines and a central machine, where the local machines contain the observations and each local machine has access only to data in that machine, and the central machine produces the final estimator of θ under the communication constraints between the local and central machines. More precisely, suppose we observe i.i.d. random samples drawn from a

distribution $P_{\theta,\xi} \in \mathcal{P}$:

$$X_i \stackrel{\text{i.i.d.}}{\sim} P_{\theta,\xi}, \quad i = 1, \dots, m,$$

where the i -th local machine has access to X_i only.

On each machine, because of limited communication budget, the observation X_i on the i -th local machine needs to be processed to a uniquely decodable binary string Z_i . The resulting string Z_i , which is called the **transcript** from the i -th machine, is transmitted to the central machine. Finally, after all transcripts Z_1, \dots, Z_m are generated, a **distributed estimator** $\hat{\theta}$ is constructed on the central machine based on the transcripts Z_1, \dots, Z_m ,

$$\hat{\theta} = \hat{\theta}(Z_1, \dots, Z_m).$$

The rules and constraints related to how transcripts can be constructed, which is called **distributed protocol**, has a lot of different variety. we are primarily interested in three different types of distributed protocols: independent protocol, sequential protocol, and black-board protocols:

- **Independent protocol.** The local machines simultaneously generate transcripts and then send them to the central machine. The i -th transcript only depends on the observation X_i on the i -th machine, so it can be expressed by $Z_i = \Pi_i(X_i)$ with some (possibly random) function Π_i . Let $|Z_i|_l$ denote the length of transcript Z_i . The class of independent protocols with total communication cost B is defined as

$$\mathcal{A}_{ind}(B) = \{\hat{\theta} : \hat{\theta} = \hat{\theta}(Z_1, \dots, Z_m), Z_i = \Pi_i(X_i), i = 1, \dots, m, \sum_{i=1}^m |Z_i|_l \leq B\}.$$

- **Sequential protocol.** The local machines sequentially send transcripts to the next local machine, and finally the central machine collects all the transcripts. The transcript Z_i sent by the i -th local machine depends on local observation X_i and the previous

transcripts Z_1, \dots, Z_{i-1} , which can be written as

$$Z_i = \Pi_i(X_i, Z_1, \dots, Z_{i-1})$$

where Π_i is a (possibly random) function. The class of sequential protocols with total communication cost B is defined as

$$\mathcal{A}_{seq}(B) = \{\hat{\theta} : \hat{\theta} = \hat{\theta}(Z_1, \dots, Z_m), Z_i = \Pi_i(X_i, Z_1, \dots, Z_{i-1}), i = 1, \dots, m, \sum_{i=1}^m |Z_i|_l \leq B\}.$$

- **Blackboard protocol.** The local machines communicate via a publicly shown blackboard. When a local machine writes a message on the blackboard, all other local machines can see the content. Finally, the central machine collects all the information and outputs the final estimate. The total length of the messages written by all local machines is at most B bits. Similarly, we denote the class of blackboard protocols with total communication cost B as $\mathcal{A}_{bb}(B)$, where the estimator is obtained by a blackboard protocol with total communication cost $\sum_{i=1}^m |Z_i|_l \leq B$. It is clear by definitions that the sequential protocols can be considered as a subset of the blackboard protocols.

Independent protocols are considered as **non-interactive** whereas sequential and blackboard protocols are considered as **interactive protocols**. See Kushilevitz (1997); Barnes et al. (2019a) for further discussion on these communication protocols.

4.1.2. Main results and our contribution

If a distributed Gaussian mean estimator achieves the same mean squared error as the optimal centralized estimator (up to a constant factor) over a range of possible value of the variance, we call it rate-optimal adaptive Gaussian mean estimator. This chapter first establishes the lower bounds for the communication costs of rate-optimal adaptive Gaussian mean estimators under the independent, sequential or blackboard protocols respectively. The lower bounds serve as a benchmark for the communication-efficiency of any rate-optimal

adaptive Gaussian mean estimator. We then develop estimation algorithms that use the minimal communication cost to achieve the statistical optimal rate of convergence. With the matching upper and lower bounds, we derive the necessary and sufficient communication costs for rate-optimal adaptive Gaussian mean estimators under the independent, sequential or blackboard protocols respectively.

The results exhibit interesting new phenomena. First, the behavior of adaptive Gaussian mean estimation with unknown variance differs significantly from the distributed estimation problem with known variance. Compared to the non-adaptive minimax rate in the case of known variance established in Cai and Wei (2020c), there is a cost of adaptation in communication budget for Gaussian mean estimation under the independent protocols, whereas no additional communication budget is needed for adaptation under the interactive protocols. Moreover, it is somewhat surprising that the minimal communication cost for distributed adaptive Gaussian mean estimation under the non-interactive and interactive protocols are different. To the best of our knowledge, this is the first example in statistical distributed estimation showing that interactions could help with estimation.

The technical tools developed in this chapter to prove the main theorems are novel and can be of independent interest. Most of the existing lower bound techniques are universal for all types of distributed protocols, and also lack the ability to study adaptation over nuisance parameters. The proof of the lower bound under the independent protocols (Theorem 15) are dedicated for adaptive estimation under the independent protocols with a non-information theoretic approach.

4.1.3. Related Literature

As mentioned earlier, distributed Gaussian mean estimation has been intensively studied in the setting of known variance. Zhang et al. (2013a); Garg et al. (2014) analyze the distributed estimation problems under the independent protocols. Braverman et al. (2016) applied a strong data processing inequality to obtain lower bounds under the blackboard protocols. Kipnis and Duchi (2017) considers distributed estimation with one-bit measure-

ments under the independent and sequential protocols. Han et al. (2018); Barnes et al. (2019b) proposed non-information theoretic approaches to obtain lower bounds for distributed estimation. Cai and Wei (2020c) established a sharp minimax rate of convergence for distributed Gaussian mean estimation with known variance under the independent, sequential, and blackboard protocols. In particular, the results show that the optimal rates are the same under the three protocols when σ^2 is known.

The behavior of estimation problems under various types of distributed protocols has been studied in two different settings. One common setting is that i.i.d. data are distributed over different machines. For example, Braverman et al. (2016); Barnes et al. (2019b) developed unified approach to establish lower bounds for distributed estimation in this setting under independent, sequential, and blackboard protocols. More recently, Acharya et al. (2020) proposed private-coin protocol and public-coin protocol and show that they have different behavior in a distributed Gaussian signal detection problem. Another setting is that data are drawn from different distributions on different local machines. Various two-sample estimation and testing problems have been considered in this setting. Xiang and Kim (2013); Liu (2021) showed that in independence testing problem and two-sample joint density estimation problem, interactions between local machines improve statistical accuracy and communication-efficiency, compared to the classical one-shot communication approaches.

An emerging topic in distributed estimation is the interplay between communication constraints and adaptation. The focus so far has been mainly on adaptive nonparametric function estimation with unknown smoothness in the distributed setting. Szabó and van Zanten (2020); Cai and Wei (2021b) showed that additional communication budget is required in order to achieve adaptation in distributed nonparametric function estimation under the independent protocols. This is in sharp contrast to the classical centralized setting where global adaptation can be achieved for free over a wide range of smoothness classes (Donoho and Johnstone, 1995; Johnstone, 2017).

4.1.4. Organization of the chapter

We finish this section with notation and definitions. We first formulate the problem in Section 4.2. Then we derive the minimal communication cost for rate-optimal adaptive Gaussian mean estimation under the independent protocols in Section 4.3 and establish the minimal communication cost for rate-optimal adaptive Gaussian mean estimation under the sequential and blackboard protocols in Section 4.4. The numerical performance of the proposed distributed estimators is investigated in Section 4.5. Further research directions are discussed in Section 4.6 and the proofs of main theorems and lemmas are provided in Section 4.7.

4.1.5. Notation and definitions

For any $a \in \mathbb{R}$, let $\lfloor a \rfloor$ denote the floor function (the largest integer not larger than a), and $\lceil a \rceil$ denote the ceiling function (the smallest integer not smaller than a). Unless otherwise stated, we shorthand $\log a$ as the logarithm to the base 2 of a . For any $a, b \in \mathbb{R}$, let $a \wedge b \triangleq \min\{a, b\}$ and $a \vee b \triangleq \max\{a, b\}$. We use $a = O(b)$ or equivalently $b = \Omega(a)$ to denote there exist a constant $C > 0$ such that $a \leq Cb$, and we use $a \asymp b$ to denote $a = O(b)$ while $b = O(a)$. We use $\tau_{[a,b]}(x)$ to denote the truncation function, which is the projection of x onto $[a, b]$. Define the density of a Gaussian distribution with mean 0 and standard deviation σ as

$$\phi_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}.$$

and the tail probability of a standard Gaussian distribution with mean 0 and standard deviation 1 as

$$\Phi(x) = \mathbb{P}(N(0, 1) > x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy.$$

4.2. Problem Formulation

In this section, we formulate the statistical problem of distributed Gaussian mean estimation with unknown variance σ^2 . Suppose there are m local machines, on the i -th machine there

is an i.i.d. normal observation:

$$X_i \stackrel{\text{i.i.d.}}{\sim} N(\theta, \sigma^2).$$

The goal is to optimally estimate $\theta \in [0, 1]$ with unknown σ^2 under a certain distributed protocol with a total communication budget B . In other words, the distributed estimator needs to be adaptive to the unknown variance σ^2 .

In the conventional centralized setting, the minimax risk of restricted Gaussian mean estimation is given in Bickel (1981a):

$$\inf_{\hat{\theta}} \sup_{\theta \in [0,1]} \mathbb{E}(\hat{\theta} - \theta)^2 = \frac{\sigma^2}{m} - 4\pi^2 \frac{\sigma^4}{m^2} + o(\sigma^2) \asymp \frac{\sigma^2}{m} \wedge 1.$$

The above quantity serves as a benchmark for the Gaussian mean estimation problem. For a given $\sigma_0 > 0$, we call distributed estimator $\hat{\theta}$ a *rate-optimal adaptive estimator* if there exists a constant $C > 0$, not depending on σ, σ_0 or m , such that for any $\sigma \geq \sigma_0$, we have

$$\sup_{\theta \in [0,1]} \mathbb{E}(\hat{\theta} - \theta)^2 \leq C \left(\frac{\sigma^2}{m} \wedge 1 \right).$$

Such distributed estimators are considered as statistically optimal and adaptive as they achieve the optimal rate of convergence in the centralized-setting over a wide range of σ . Let $\mathcal{P}_{\sigma_0} = \{P_{\theta, \sigma} = N(\theta, \sigma^2) : \theta \in [0, 1], \sigma \in [\sigma_0, \infty)\}$ be the Gaussian location family with unknown variance. The distributed estimation problem of θ is considered with the nuisance parameter σ .

Setting a lower bound $\sigma \geq \sigma_0$ is necessary. This is due to the fact that no distributed estimator with a finite total communication cost B is able to achieve the optimal rate of convergence over all $\sigma > 0$. With total communication cost B , the mean squared error of any distributed estimator is at least of order 2^{-2B} due to discretization error, however, the optimal rate of convergence for Gaussian mean estimation is of order $\frac{\sigma^2}{m}$. As a result, when σ is extremely small, any distributed estimator cannot attain optimal rate of convergence.

Therefore, there is no distributed estimator with finite communication cost that can be rate-optimal adaptive with all possible positive real number σ . A lower bound on σ is needed here to make the problem well-formulated. With smaller lower bound σ_0 , the distributed estimator needs more communication cost in order to be adaptive over the range $\sigma \geq \sigma_0$, and the estimating procedure would be also different. In the real data application, people needs to choose σ_0 as a priori, depending on prior knowledge on the dataset or the communication budget. See also Remark 13 for further discussion on σ_0 .

Throughout this chapter, we assume $0 < \sigma_0 \leq \frac{1}{2}$. When $\sigma_0 > \frac{1}{2}$, the solution to the problem is essentially identical to the case $\sigma_0 = \frac{1}{2}$. See Remark 11 for further explanation.

4.3. Optimal Adaptive Estimation under the independent protocols

We consider in this section adaptive distributed estimation under the independent protocols. We begin by establishing a lower bound for the minimax relative efficiency under the independent protocols with a given communication budget. A rate-optimal adaptive distributed estimator is then constructed. It is shown that the proposed estimator achieves the minimum communication cost among all rate-optimal adaptive estimators, as is shown by the matching lower bound.

4.3.1. Lower bound analysis

It is difficult to directly derive the minimal communication cost for rate-optimal adaptive estimators. In our analysis, we first analyze the statistical performance of the estimators in the class $\mathcal{A}_{ind}(B)$. Then we argue that only when the communication budget B is larger than a certain value, a distributed estimator in $\mathcal{A}_{ind}(B)$ can possibly be a rate-optimal adaptive estimator. This leads to a lower bound for the communication cost among the rate-optimal adaptive estimators.

We use the relative efficiency as a measure for the statistical performance when we derive

the lower bound. The relative efficiency for an estimator $\hat{\theta}$ is defined as

$$r(\hat{\theta}, \theta, \sigma) = \left(\frac{\sigma^2}{m} \wedge 1 \right)^{-1} \mathbb{E}(\hat{\theta} - \theta)^2$$

which indicates the gap between the mean squared error of the estimator $\hat{\theta}$ and the optimal rate of convergence when data are drawn from $N(\theta, \sigma^2)$.

We consider the minimax relative efficiency under the total communication constraint B :

$$R_{ind}(\sigma_0, B) = \inf_{\hat{\theta} \in \mathcal{A}_{ind}(B)} \sup_{\theta \in [0,1], \sigma \geq \sigma_0} r(\hat{\theta}, \theta, \sigma).$$

The quantity $R_{ind}(\sigma_0, B)$ is a benchmark for the limit of estimation accuracy under the independent protocols with the total communication constraint B , when σ^2 is unknown.

The relative efficiency is closely related to rate-optimal adaptive estimators. According to the definition, $\hat{\theta}$ is a rate-optimal adaptive estimator over $\sigma \geq \sigma_0$, if and only if the maximum relative efficiency for the estimator $\hat{\theta}$ is bounded by some constant C , i.e.

$$\sup_{\theta \in [0,1], \sigma \geq \sigma_0} \left(\frac{\sigma^2}{m} \wedge 1 \right)^{-1} \mathbb{E}(\hat{\theta} - \theta)^2 \leq C.$$

Remark 9. As a contrast, the conventional distributed minimax risk

$$\inf_{\hat{\theta} \in \mathcal{A}_{ind}(B)} \sup_{\theta \in [0,1], \sigma \geq \sigma_0} \mathbb{E}(\hat{\theta} - \theta)^2$$

is not a good proxy to study because the estimation problem becomes more difficult when σ^2 is large. When σ is sufficiently large, say, $\sigma > \sqrt{m}$ this minimax mean squared risk is bounded away from zero according to centralized minimax rate given in Bickel (1981a).

The following theorem provides a lower bound on the minimax relative efficiency for estimators in $\mathcal{A}_{ind}(B)$.

Theorem 15. *If $B > \frac{1}{\frac{\sigma_0}{m}}$, there exists a constant $c > 0$, not depending on σ_0, σ, θ or m , such that*

$$R_{ind}(\sigma_0, B) \geq c \sqrt{\frac{m \log \frac{1}{\sigma_0}}{B}} \wedge 1.$$

The techniques used to prove Theorem 15 are novel and can be of independent interest. Roughly speaking, our goal is to prove that there must exist $\sigma \geq \sigma_0$ and θ, δ such that the central machine cannot tell whether the data are drawn from $N(\theta - \delta, \sigma^2)$ or $N(\theta + \delta, \sigma^2)$ by only looking at these transcripts. To accomplish this goal, we give an upper bound on the integrated squared Hellinger distances over different choices of θ and σ :

$$I = \sum_{i=1}^m \sum_{j=0}^{J-1} \int_{\lambda\sigma_j}^{1-\lambda\sigma_j} H^2(Z_i|X_i \sim N(\theta - \lambda\sigma_j, \sigma_j^2); Z_i|X_i \sim N(\theta + \lambda\sigma_j, \sigma_j^2)) d\theta$$

where $\sigma_1, \sigma_2, \dots, \sigma_J$ are carefully chosen different levels of σ , λ is a tuning constant factor. $H^2(Z_i|X_i \sim N(\theta - \lambda\sigma_j, \sigma_j^2); Z_i|X_i \sim N(\theta + \lambda\sigma_j, \sigma_j^2))$ denotes the squared Hellinger distances between distribution of Z_i if $X_i \sim N(\theta - \lambda\sigma_j, \sigma_j^2)$ and distribution of Z_i if $X_i \sim N(\theta + \lambda\sigma_j, \sigma_j^2)$. If I is proved to be small, then there must exist some θ and σ_j such that

$$\sum_{i=1}^m H^2(Z_i|X_i \sim N(\theta - \lambda\sigma_j, \sigma_j^2); Z_i|X_i \sim N(\theta + \lambda\sigma_j, \sigma_j^2))$$

is small, and then we can conclude that the central machine does not have enough information to distinguish whether the data are drawn from $N(\theta - \lambda\sigma_j, \sigma_j^2)$ or $N(\theta + \lambda\sigma_j, \sigma_j^2)$. This will give a lower bound on the relative efficiency $R_{ind}(\sigma_0, B)$. The above technique can be summarized into the following lemma:

Lemma 7. *Let $J > 0$ be an integer. Let $\lambda > 0$, $0 < \sigma_0 < \sigma_1 < \dots < \sigma_{J-1}$ satisfy $\lambda\sigma_{J-1} < \frac{1}{6}$.*

If for any distributed estimator $\hat{\theta} \in \mathcal{A}_{ind}(B)$, we have

$$I = \sum_{i=1}^m \sum_{j=0}^{J-1} \int_{\lambda\sigma_j}^{1-\lambda\sigma_j} H^2(Z_i|X_i \sim N(\theta - \lambda\sigma_j, \sigma_j^2); Z_i|X_i \sim N(\theta + \lambda\sigma_j, \sigma_j^2)) d\theta \leq \frac{J}{2},$$

then there exists a constant $c > 0$ such that

$$R_{ind}(\sigma_0, B) \geq c\lambda^2 m.$$

Theorem 15 gives a lower bound on the relative efficiency for all distributed estimators from $\mathcal{A}_{ind}(B)$. Note that a rate-optimal adaptive estimator should have bounded relative efficiency, the following Corollary 4.3.1 can be directly derived from Theorem 15.

Corollary 4.3.1. *If an estimator $\hat{\theta} \in \mathcal{A}_{ind}(B)$ is a rate-optimal adaptive estimator, that is, there exists a constant $C > 0$ such that*

$$\mathbb{E}(\hat{\theta} - \theta)^2 \leq C \left(\frac{\sigma^2}{m} \wedge 1 \right) \quad \text{for all } \sigma \geq \sigma_0.$$

Then there exists a constant $c > 0$ (which only depends on C) such that

$$B \geq cm \log \frac{1}{\sigma_0}.$$

The above corollary states that the minimum communication cost needed for a rate-optimal adaptive estimator is of order $m \log \frac{1}{\sigma_0}$.

4.3.2. Optimal estimator under the independent protocols - $\hat{\theta}_q$

We now construct a communication efficient rate-optimal adaptive estimator under the independent protocol. The optimal estimator $\hat{\theta}_q$ makes use of $m \log \frac{3}{\sigma_0}$ total communication budget to achieve the centralized optimal rate of convergence for all $\sigma \geq \sigma_0$.

The estimator $\hat{\theta}_q$ can be constructed by the following steps.

Step 1: Generating transcripts. Let $d = 2^{\lceil \log_2 \sigma_0 \rceil}$. Let S_d denote the following grid of interval d between $-1 - d$ and 2 :

$$S_d = \{-1 - d, -1, -1 + d, -1 + 2d, \dots, 2 - d, 2\}.$$

Let Z_i be the quantized version of X_i and then truncate in $[-1, 2]$. That is,

$$Z_i = \begin{cases} -1 - d & \text{if } X_i \leq -1 \\ 2 & \text{if } X_i \geq 2 \\ \max\{z \in S_d : z \leq X_i\} & \text{if } -1 < X_i < 2 \end{cases}$$

In the third case when $-1 < X_i < 2$, Z_i is the maximum number in S_d that is less than or equal to X_i . Since Z_i has only $3/d + 2$ possible values, it can be encoded using at most $\log\left(\frac{3}{d} + 2\right) \leq \log\left(\frac{6}{\sigma_0} + 2\right)$ bits.

Step 2: Estimation. The central machine receives the transcripts Z_1, \dots, Z_m from the local machines. Let $Z_{(1)} \leq \dots \leq Z_{(m)}$ be the order statistics of Z_1, \dots, Z_m . First, we calculate $\hat{\sigma}$ by

$$\hat{\sigma} = \begin{cases} \sigma_0 & \text{if } Z_{(\lceil 0.84m \rceil)} - Z_{(\lfloor 0.16m \rfloor)} < \sigma_0 \\ Z_{(\lceil 0.84m \rceil)} - Z_{(\lfloor 0.16m \rfloor)} & \text{if } \sigma_0 \leq Z_{(\lceil 0.84m \rceil)} - Z_{(\lfloor 0.16m \rfloor)} \leq 1 \\ 1 & \text{if } Z_{(\lceil 0.84m \rceil)} - Z_{(\lfloor 0.16m \rfloor)} > 1 \end{cases}$$

Then, let $\tilde{\sigma} = \min\{2^{-k} : 1 \geq 2^{-k} \geq \hat{\sigma}, k \text{ is an integer}\}$, i.e. the minimum number that is power of 2 and also larger than $\hat{\sigma}$. Let $L = \max\{k\tilde{\sigma} : k\tilde{\sigma} \leq Z_{(\lfloor 0.16m \rfloor)}, k \text{ is an integer}\}$, i.e. the largest multiple of $\tilde{\sigma}$ that is less than or equal to $Z_{(\lfloor 0.16m \rfloor)}$. Similarly we define $R = \max\{k\tilde{\sigma} : k\tilde{\sigma} \geq Z_{(\lceil 0.84m \rceil)}, k \text{ is an integer}\}$, i.e. the smallest multiple of $\tilde{\sigma}$ that is larger than or equal to $Z_{(\lceil 0.84m \rceil)}$. Let $\hat{p}_L = \frac{1}{m} \sum_{i=1}^m \mathbb{I}_{\{Z_i < L\}}$ be the proportion of transcripts that is less than L , and $\hat{p}_R = \frac{1}{m} \sum_{i=1}^m \mathbb{I}_{\{Z_i \geq R\}}$ be the proportion of transcripts that is larger than or equal to R ,

Finally, recall that $\Phi(\cdot)$ denotes the tail probability of a standard Gaussian variable, let $(\hat{\theta}_q, \hat{\sigma}_q)$ be the solution to the equations:

$$\Phi\left(\frac{\hat{\theta}_q - L}{\hat{\sigma}_q}\right) = \hat{p}_L \vee \frac{1}{m},$$

$$\Phi\left(\frac{R - \hat{\theta}_q}{\hat{\sigma}_q}\right) = \hat{p}_R \vee \frac{1}{m}.$$

The above equation always has one unique solution where $\hat{\theta}_q \in [L, R]$, we take this $\hat{\theta}_q$ as the final estimate.

It is easy to verify that the above estimator $\hat{\theta}_q \in \mathcal{A}_{ind}\left(m \log\left(\frac{1}{\sigma_0} + 5\right)\right)$. The next theorem establishes an upper bound on its mean squared error, showing that the estimator is rate-optimal adaptive over $\sigma \geq \sigma_0$.

Theorem 16. *There exists a constant $C > 0$, not depending on σ_0, σ, θ or m , such that*

$$\sup_{\theta \in [0, 1], \sigma \geq \sigma_0} \mathbb{E}(\hat{\theta}_q - \theta)^2 \leq C \left(\frac{\sigma^2}{m} \wedge 1\right).$$

Remark 10. The construction of the estimator $\hat{\theta}_q$ is involved. A more straightforward and simpler estimator is the quantization-then-average estimator proposed in Zhang et al. (2013a). However, it can be shown that the quantization-then-average estimator is not even consistent when each local machine has only limited communication budget, because the quantization bias ($\mathbb{E}Z_i - \theta$) is not exactly zero if one just rounds the observations to a certain precision on the local machines. As a result, when the number of machines $m \rightarrow \infty$, the estimation error will not converge to zero. Therefore, a more sophisticated procedure such as $\hat{\theta}_q$ is necessary to achieve the optimal rate of convergence with the communication constraint.

Remark 11. The above estimator $\hat{\theta}_q$ is designed under the assumption that $0 < \sigma_0 \leq \frac{1}{2}$. When $\sigma_0 > \frac{1}{2}$, we can use the estimator for the case $\sigma_0 = \frac{1}{2}$, which is rate-optimal adaptive estimator over $\sigma \geq \sigma_0$. The total communication cost is of order m , which cannot be further reduced because each machine needs to transmit at least one bit in order to involve its observation into the estimation procedure. The choice of $\frac{1}{2}$ is for convenience; it can be changed to any positive number and all the results hold with minor modifications.

Remark 12. Corollary 4.3.1 and Theorem 16 together show that the necessary and sufficient communication cost for a rate-optimal adaptive estimator is of order $m \log \frac{1}{\sigma_0}$ bits. The order of communication cost of the estimator $\hat{\theta}_q$ cannot be further reduced. Compared to the minimax rate of convergence for non-adaptive Gaussian mean estimation established in the previous complementary work Cai and Wei (2020c), the communication cost for adaptive Gaussian mean estimation is larger, so there is a cost of adaptation under the independent protocols.

Remark 13. The construction of adaptive estimator $\hat{\theta}_q$ requires knowledge of the lower bound σ_0 for unknown σ , which seems unnatural. However, as Theorem 15 suggests, if one lets $\sigma_0 \rightarrow 0$, the required communication cost for a distributed estimator to achieve the optimal rate of convergence will go to infinity. Therefore, there is no rate-optimal adaptive estimator for all $\sigma > 0$ without a lower bound on σ . A similar phenomenon also appears in the construction of adaptive confidence ball in nonparametric regression. If one assumes the smoothness $\beta \geq \beta_0$, then it is possible to be adaptive from β_0 to $2\beta_0$. If one does not assume any lower bound for the smoothness, then no adaptation is possible. See Theorem 4 and the discussion thereafter in Cai and Low (2006).

4.4. Optimal Adaptive Estimate under Interactive Protocols

In the previous section we show that an order of $m \log \frac{1}{\sigma_0}$ bits are necessary and sufficient for an adaptive estimator to achieve its optimal statistical performance under the independent protocols with $\sigma \geq \sigma_0$. However, under the sequential protocols or blackboard protocols, it may require less communication cost to achieve the same statistical performance, because the local machines can “communicate” with each other to some extent. This leads to an interesting question: do we still need $m \log \frac{1}{\sigma_0}$ bits to achieve the optimal rate of convergence over all $\sigma \geq \sigma_0$ under the sequential or blackboard protocols?

We consider in this section distributed estimation under two types of interactive protocols, namely the sequential protocols and the blackboard protocols. We first construct a distributed estimator under the sequential protocols that is statistical optimal for all $\sigma \geq \sigma_0$.

A matching lower bound is then established to show that the communication cost of the proposed estimator cannot be further improved for all distributed estimators under the blackboard protocols. Recall that the sequential protocols are a subset of the blackboard protocols, we obtain the sufficient and necessary communication cost for the statistical optimal estimators under the interactive protocols. The results show an interesting phenomenon. Compared to the independent protocols, under the sequential protocols or the blackboard protocols, it requires less communication cost for the rate-optimal adaptive estimation. So feedback and information sharing are helpful in distributed Gaussian mean estimation with unknown variance.

4.4.1. Optimal estimator under the sequential protocols

In the following procedure we assume $m \geq 12$. The case of $m \leq 11$ is relatively simple. For example, when $m \leq 11$, the problem can be solved by only looking at the first local machine and outputs its best approximation up to σ_0 precision. The estimation process can be divided into three steps:

Step 1: Preliminary estimation of θ and σ . For the first 11 local machines, the i -th machine ($i = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11$) outputs

$$Z_i = \lfloor \tau_{[-1,2]}(X_i + 1)/\sigma_0 \rfloor.$$

There are at most $\lfloor \frac{3}{\sigma_0} \rfloor + 1$ possible outputs for each local machine, so each transcript Z_i ($i = 1, 2, \dots, 11$) can be encoded by no larger than $\log \frac{3}{\sigma_0} + 1$ bits.

On the 12-th and later local machines, based on Z_1, Z_2, \dots, Z_{11} , each machine can calculate a preliminary estimate of θ and σ by

$$\hat{\theta}_{11} = \sigma_0 Z_{11},$$

$$\hat{\sigma} = \sigma_0 \max \left\{ 1, \left(\frac{1}{10} \sum_{i=1}^{10} \left(Z_i - \frac{1}{10} \sum_{i=1}^{10} Z_i \right)^2 \right)^{1/2} \right\}.$$

Step 2: One-bit passing. Starting at the 12-th local machine, on the i -th local machine, we output

$$Z_i = \text{sign}(X_i - \hat{\theta}_{i-1}),$$

and update the current state $\hat{\theta}$ by

$$\hat{\theta}_i = \hat{\theta}_{i-1} + \hat{\sigma} \gamma_i Z_i$$

where $\gamma_i = i^{-2/3}$.

Step 3: Final estimation of θ . On the central machine, because we have access to Z_1, Z_2, \dots, Z_m , thus we can calculate $\hat{\theta}_i$ accordingly for all $i = 11, \dots, m$. The final estimator of the mean θ is given by

$$\hat{\theta}_{sq} = \frac{1}{m-10} \sum_{i=11}^m \hat{\theta}_i.$$

Since each of first 11 local machines outputs at most $\log \frac{3}{\sigma_0} + 1$ bits, and the later local machines only output 1 bit per machine, it is easy to verify that the above proposed estimator $\hat{\theta}_{sq} \in \mathcal{A}_{sq}(11 \log \frac{3}{\sigma_0} + m)$. The following theorem gives an upper bound on its mean squared error for all $\sigma \geq \sigma_0$

Theorem 17. *The estimator $\hat{\theta}_{sq} \in \mathcal{A}_{sq}(11 \log \frac{3}{\sigma_0} + m)$ and satisfies*

$$\mathbb{E}(\hat{\theta}_{sq} - \theta)^2 \leq C \left(\frac{\sigma^2}{m} \wedge 1 \right),$$

where C is a universal constant not depending on σ_0, σ, θ or m .

That is, the proposed sequential protocol estimator $\hat{\theta}_{sq}$ is rate-optimal for all $\sigma \geq \sigma_0$, whose total communication cost is $\log \frac{3}{\sigma_0} + m$ bits.

Remark 14. The one-bit passing step of the above estimator $\hat{\theta}_{sq}$ is established in light of the previous work Kipnis and Duchi (2017), where the goal is to construct an estimator using one-bit measurements from local machines. Their proposed estimator was shown to

be asymptotically normal. However, the finite sample mean squared error of their estimator was not guaranteed, as the finite sample performance is significantly influenced by the initial position $\hat{\theta}_{11}$ and the initial step size $\hat{\sigma}$.

We introduce in this chapter the preliminary estimates $\hat{\theta}_{11}$ and $\hat{\sigma}$, which can be obtained at a small amount of communication cost, as an approximation for the optimal initial position and initial step size. This warm start initialization is the key to obtain finite sample bound in Theorem 17. The hardcoded number “11” in the procedure can be set to any larger constants, but not smaller ones. Due to a technical reason we require the preliminary estimate $\hat{\sigma}$ to have bounded -5 order moment, i.e. $\mathbb{E}\hat{\sigma}^{-5} < \infty$.

Remark 15. The proof of Theorem 17 extends the techniques developed in the previous seminal work Polyak (1990) on stochastic approximation. Polyak (1990) developed upper bounds for stochastic approximation with averaging. The additional difficulty to prove Theorem 17, compared to the previous work, is to control the uncertainty brought to the estimator $\hat{\theta}_{sq}$ from the random initialization $\hat{\theta}_{11}$ and $\hat{\sigma}$. Much more careful calculation is needed here.

4.4.2. Lower bound under interactive protocols

The above proposed estimator $\hat{\theta}_{sq}$ achieves the optimal rate of convergence for all $\sigma \geq \sigma_0$ with communication cost $(11 \log \frac{3}{\sigma_0} + m)$ bits. The next theorem is a direct corollary derived from Theorem 5 in Cai and Wei (2020c). The lower bound argument shows that the communication cost for $\hat{\theta}_{sq}$ cannot be improved.

Theorem 18. *For any $\hat{\theta} \in \mathcal{A}_{bb}(B)$, if $\hat{\theta}$ is rate-optimal when $\sigma = \sigma_0$, i.e. there is a constant $C > 0$ such that*

$$\sup_{\theta \in [0,1]} \mathbb{E}(\hat{\theta} - \theta)^2 \leq C \left(\frac{\sigma_0^2}{m} \wedge 1 \right).$$

Then there exists a constant $c > 0$ (depends on C) such that

$$B \geq c \left(\log \frac{1}{\sigma_0} + m \right).$$

The above theorem establishes a lower bound on the communication cost for any distributed estimator under the blackboard protocols that achieves optimal rate of convergence when $\sigma = \sigma_0$. The same lower bound also holds for any estimator that achieves the optimal rate of convergence for $\sigma \geq \sigma_0$, as those estimators are feedback-free under more strict conditions. Recall that the sequential protocols are a subset of the blackboard protocols. Therefore, the lower bound in Corollary 18, together with the proposed adaptive estimator $\hat{\theta}_{sq}$, shows that order $\log \frac{1}{\sigma_0} + m$ communication cost is necessary and sufficient for rate-optimal adaptive estimation under the interactive protocols, including the sequential and blackboard protocols.

Remark 16. Recall that for any rate-optimal adaptive estimator under the independent protocols, the minimal communication cost is of order $m \log \frac{1}{\sigma_0}$, which is larger than that for a rate-optimal adaptive estimator under the interactive protocols. Feedback and information sharing are necessary to improve communication-efficiency in adaptive Gaussian mean estimation.

Remark 17. The lower bound on communication cost in Corollary 18 holds for the non-adaptive case when $\sigma = \sigma_0$ is known in advance. Since the adaptive estimator $\hat{\theta}_{sq}$ is constructed with no more communication cost than the non-adaptive case, there is no cost of adaptation for Gaussian mean estimation under the two types of the interactive protocols. In contrast, under the independent protocols, as more communication cost is needed to establish a rate-optimal adaptive estimator, there is a cost of adaptation for Gaussian mean estimation.

Table 4.1: Optimal communication cost for different distributed protocols under adaptive and non-adaptive settings. Adaptive setting: minimal communication cost for rate-optimal adaptive estimator over $\sigma \geq \sigma_0$. Non-adaptive setting: minimal communication cost for rate-optimal estimator with known $\sigma = \sigma_0$.

Protocol	adaptive estimator	non-adaptive estimator
Independent	$O(m \log \frac{1}{\sigma_0})$	$O(m + \log \frac{1}{\sigma_0})$
Sequential	$O(m + \log \frac{1}{\sigma_0})$	$O(m + \log \frac{1}{\sigma_0})$
Blackboard	$O(m + \log \frac{1}{\sigma_0})$	$O(m + \log \frac{1}{\sigma_0})$

4.5. Numerical Results

The proposed adaptive estimators under independent protocol and under interactive protocols are easy to implement. In this section, we conduct simulation studies to investigate the numerical performance of these two estimators. The numerical results show that the proposed estimators are practically useful, having high statistical accuracy while only requiring a small amount of communication cost.

We consider in the simulation study a setting where $\sigma_0 = 2^{-12}$, i.e. we know a priori $\sigma \geq \sigma_0 = 2^{-12}$. Assume there are $m = 100$ machines, where each machine has access to a univariate normal variable $X \sim N(\theta, \sigma^2)$, with $\theta = 0.3$ and choices of $\sigma \in \{2^{-2}, 2^{-4}, 2^{-6}, 2^{-8}, 2^{-10}, 2^{-12}\}$. We compare the following three estimators: the classical sample-mean estimator (under the centralized setting), the adaptive estimator under the independent protocol, and the adaptive estimator under the sequential protocol. The average mean squared errors (MSEs) of the three different estimators over 100 simulation runs, and the communication costs (in bits) of the two distributed estimators are given in Table 4.2.

Table 4.2: MSEs and the communication costs of the three methods. $\sigma_0 = 2^{-12}$, $m = 100$, $\theta = 0.3$. For the two distributed estimators, total communication costs (in bits) are given in the parentheses.

σ	Sample-mean	Independent Protocol	Sequential Protocol
2^{-2}	6.17×10^{-4}	$4.14 \times 10^{-3}(1500)$	$2.12 \times 10^{-3}(266)$
2^{-4}	4.04×10^{-5}	$1.45 \times 10^{-4}(1500)$	$1.28 \times 10^{-4}(266)$
2^{-6}	2.14×10^{-6}	$9.02 \times 10^{-6}(1500)$	$8.31 \times 10^{-6}(266)$
2^{-8}	1.46×10^{-7}	$5.23 \times 10^{-7}(1500)$	$4.85 \times 10^{-7}(266)$
2^{-10}	8.59×10^{-9}	$5.00 \times 10^{-8}(1500)$	$2.66 \times 10^{-8}(266)$
2^{-12}	5.68×10^{-10}	$5.10 \times 10^{-9}(1500)$	$2.47 \times 10^{-9}(266)$

The numerical results shown in Table 4.2 are interesting and consistent with the theoretical analysis given earlier. The adaptive estimator under the sequential protocol uniformly outperforms the one under the independent protocol, in terms of both the MSE and communication cost. This shows the clear advantage of the sequential protocol over the independent protocol. Comparing with the classical centralized sample-mean estimator, the MSEs of the adaptive estimator under the sequential protocol and under the independent protocol are

respectively within a factor of 4 and a factor of 10 times of the corresponding MSEs of the sample mean. This is consistent with the theoretical results that the statistical accuracy of both distributed estimators is within a constant factor of the centralized optimality. Indeed, the simulation results show that the actual constant gaps are relatively small. In particular, it is interesting to see that the adaptive estimator under the sequential protocol achieves such a good performance with only 266 bits. Considering their low communication costs, we find the proposed adaptive estimators could be practically useful in real distributed estimation applications.

4.6. Discussion

We studied in this chapter the problem of distributed adaptive Gaussian mean estimation with unknown σ . In the conventional centralized setting, Gaussian mean estimation with unknown σ is arguably one of the most basic and fundamental problems in classical statistics. As seen in this chapter, the theoretical analysis is rich and difficult in the distributed setting.

The insights gained from the analysis can be used to solve other related problems where the variance is unknown. One such problem is nonparametric regression with random design. As pointed out in Cai and Wei (2021b), despite being asymptotically equivalent in the centralized setting, the problem of distributed nonparametric regression with random design is significantly different from that with fixed design. For example, when wavelet methods are used, the empirical wavelet coefficients in this case have unknown variance due to the unknown design distribution and the techniques developed in this chapter can potentially be used to construct a wavelet estimator in that problem. More discussion on the connections and differences among various distributed nonparametric function estimation problems can be found in Cai and Wei (2021b).

In this chapter, the focus is on the optimal estimation of the mean θ . A closely related problem is statistical inference for the mean including the construction of optimal confidence intervals for θ . This involves optimal estimation of the variance σ^2 in the same setting, which is a challenging problem by itself. We leave the inference problem for future work.

The results in this chapter reveal an interesting phenomenon: the communication costs required under different types of distributed protocols can be substantially different. This is in sharp contrast to Gaussian mean estimation with known variance. It is interesting to investigate further the differences among various types of distributed protocols for other distributed statistical problems. It is technically challenging to develop a general optimality theory under different types of communication constraints. More generally, it is of significant interest to understand the interplay between communication cost, statistical accuracy, adaptation, and different types of distributed protocols for a wide range of problems. This is an important topic in data science that is wide open and merits further study.

4.7. Proofs

We prove the main results in this section. Throughout this section, L_x^1 denotes the L^1 function space with respect to the x variable and \mathbb{I}_Ω denotes the indicator function taking values in $\{0, 1\}$. We use shorthand $a \lesssim b$ to denote there exists a universal constant $C > 0$ such that $a \leq Cb$. With slight abuse of notation, we define ϕ be the standard Gaussian density, ϕ_σ be the density of $N(0, \sigma^2)$, and $\phi_{\theta, \sigma}$ be the density of $N(\theta, \sigma^2)$.

4.7.1. Proof of Theorem 15

We first define several quantities. They play important roles to establish the proof.

Let P, Q be two distributions that are absolutely continuous with respect to a Lebesgue measure on the measurable space \mathcal{Z} . p, q are the density functions of P, Q respectively. Define squared Hellinger distance $H^2(P, Q)$ as

$$H^2(P, Q) \triangleq \frac{1}{2} \int_{\mathcal{Z}} (\sqrt{p} - \sqrt{q})^2 dx.$$

Define total variation distance $TV(P, Q)$ as

$$TV(P, Q) \triangleq \frac{1}{2} \int_{\mathcal{Z}} |p - q| dx.$$

Let \mathcal{Z} be a finite set, $h : \mathbb{R} \rightarrow \mathcal{Z}$ a random function, and $f, g \in L^1(\mathbb{R})$ are non-negative functions. Define “generalized squared Hellinger distance” for Z be

$$H^2(h; f, g) \triangleq \frac{1}{2} \sum_{z \in \mathcal{Z}} \left(\sqrt{\int_{-\infty}^{\infty} f(x) \mathbb{P}(h(x) = z) dx} - \sqrt{\int_{-\infty}^{\infty} g(x) \mathbb{P}(h(x) = z) dx} \right)^2.$$

Note that when f, g are densities, $H^2(h; f, g)$ is exactly the squared Hellinger distance between distribution of $h(X)$ when $X \sim f$, and distribution of $h(X)$ when $X \sim g$. This is why we call this quantity generalized squared Hellinger distance.

Similarly, we define “generalized total variation distance” as

$$TV(h; f, g) \triangleq \frac{1}{2} \sum_{z \in \mathcal{Z}} \left| \int_{-\infty}^{\infty} f(x) \mathbb{P}(h(x) = z) dx - \int_{-\infty}^{\infty} g(x) \mathbb{P}(h(x) = z) dx \right|.$$

Also when f, g are densities, $TV(h; f, g)$ is exactly the total variation distance between distribution of $h(X)$ when $X \sim f$, and distribution of $h(X)$ when $X \sim g$.

The following lemma provides two basic but useful inequalities for $H^2(h; f, g)$ and $TV(h; f, g)$.

Lemma 8. *For any random function $h : \mathbb{R} \rightarrow \mathcal{Z}$, the following two inequalities hold:*

- (a) *Sub-additivity of $H^2(h; f, g)$: if $f(x, s), g(x, s) \in L_x^1(\mathbb{R})$ are non-negative functions for each $s \in (s_l, s_r)$, and $\int_{s_l}^{s_r} f(x, s) ds, \int_{s_l}^{s_r} g(x, s) ds \in L_x^1(\mathbb{R})$. Then we have*

$$H^2(h; \int_{s_l}^{s_r} f(\cdot, s) ds, \int_{s_l}^{s_r} g(\cdot, s) ds) \leq \int_{s_l}^{s_r} H^2(h; f(\cdot, s), g(\cdot, s)) ds. \quad (4.1)$$

- (b) *Bound between $TV(h; f, g)$ and $H^2(h; f, g)$: if f and g have the same support (i.e. $\{x : f(x) > 0\} = \{x : g(x) > 0\}$) and there exist $M \geq 1$ such that $1/M \leq f(x)/g(x) \leq M$ for all $x \in \{x : g(x) > 0\}$. Then we have*

$$H^2(h; f, g) \leq \frac{\sqrt{M} - 1}{\sqrt{M} + 1} TV(h; f, g). \quad (4.2)$$

Besides, we define $\phi_{\theta,\sigma}$ as the density function of $N(\theta, \sigma^2)$, i.e.

$$\phi_{\theta,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}.$$

Now we move back to the main proof. Let $\lambda = c_\lambda \left(\frac{\log \frac{3}{mB}}{mB}\right)^{1/4}$ where c_λ is a positive constant that will be specified later. Let J be the maximum integer such that $2^{-J} \geq \sigma_0$. Let $\sigma_j = 2^j \sigma_0, j = 1, 2, \dots, J-1$.

We are interested in the following *integrated squared Hellinger distances*:

$$I \triangleq \sum_{i=1}^m \sum_{j=0}^{J-1} \int_{\lambda\sigma_j}^{1-\lambda\sigma_j} H^2(\Pi_i; \phi_{(\theta-\lambda\sigma_j),\sigma_j}, \phi_{(\theta+\lambda\sigma_j),\sigma_j}) d\theta. \quad (4.3)$$

The following subsection is dedicated to show that under proper choice of the constant c_λ , we have $I \leq \frac{1}{2}m$.

Bound integrated integrated squared Hellinger distances I

We first "slice" $\phi_{(\theta-\lambda\sigma_j),\sigma_j}$ and $\phi_{(\theta+\lambda\sigma_j),\sigma_j}$ in (4.3) into pieces so that we can apply Lemma 8(a) to give an upper bound for I . Let

$$s^* = \sup_{x \in \mathbb{R}} |\phi_{-\lambda,1}(x) - \phi_{\lambda,1}(x)|,$$

$$A(s) = \{x : |\phi_{-\lambda,1}(x) - \phi_{\lambda,1}(x)| \geq s\}, \quad 0 < s < s^*,$$

$$x_s = \sup A(s),$$

$$f(x, s) = \mathbb{I}_{\{x \in A(s)\}} \frac{\phi_{-\lambda,1}(x)}{|\phi_{-\lambda,1}(x) - \phi_{\lambda,1}(x)|}, \quad x \neq 0, 0 < s < s^*,$$

$$g(x, s) = \mathbb{I}_{\{x \in A(s)\}} \frac{\phi_{\lambda,1}(x)}{|\phi_{-\lambda,1}(x) - \phi_{\lambda,1}(x)|}, \quad x \neq 0, 0 < s < s^*.$$

When $x = 0$ we set $f(0, s) = g(0, s) = \phi_{\lambda,1}(0)/s^*$. It is easy to verify that

$$\phi_{-\lambda,1}(x) = \int_0^{s^*} f(x, s) ds \quad \text{and} \quad \phi_{\lambda,1}(x) = \int_0^{s^*} g(x, s) ds.$$

The reason why we design the function f and g is for a good property: $g(x, s) - f(x, s) = \mathbb{I}_{\{x \in A(s)\}} \text{sign}(x)$, which is a compact supported piecewise function only taking values in $\{-1, 0, 1\}$. By this way we "discretize" the problem and is able to adopt combinatoric techniques (in Lemma 15).

Note that $\phi_{(\theta-\lambda\sigma_j),\sigma_j}(x) = \frac{1}{\sigma_j} \phi_{-\lambda,1}((x-\theta)/\sigma_j)$ and $\phi_{(\theta+\lambda\sigma_j),\sigma_j}(x) = \frac{1}{\sigma_j} \phi_{\lambda,1}((x-\theta)/\sigma_j)$, so we have

$$\begin{aligned} \phi_{(\theta-\lambda\sigma_j),\sigma_j}(x) &= \int_0^{s^*} \frac{1}{\sigma_j} f((x-\theta)/\sigma_j, s) ds, \\ \phi_{(\theta+\lambda\sigma_j),\sigma_j}(x) &= \int_0^{s^*} \frac{1}{\sigma_j} g((x-\theta)/\sigma_j, s) ds. \end{aligned}$$

The above equations and Lemma 8(a) implies

$$I = \sum_{i=1}^m \sum_{j=0}^{J-1} \int_{\lambda\sigma_j}^{1-\lambda\sigma_j} H^2(\Pi_i; \phi_{(\theta-\lambda\sigma_j),\sigma_j}, \phi_{(\theta+\lambda\sigma_j),\sigma_j}) d\theta \quad (4.4)$$

$$\leq \sum_{i=1}^m \sum_{j=0}^{J-1} \int_{\lambda\sigma_j}^{1-\lambda\sigma_j} d\theta \int_0^{s^*} H^2 \left(\Pi_i(x); \frac{1}{\sigma_j} f((x-\theta)/\sigma_j, s), \frac{1}{\sigma_j} g((x-\theta)/\sigma_j, s) \right) ds \quad (4.5)$$

$$= \sum_{i=1}^m \sum_{j=0}^{J-1} \frac{1}{\sigma_j} \int_{\lambda\sigma_j}^{1-\lambda\sigma_j} d\theta \int_0^{s^*} H^2(\Pi_i(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)) ds. \quad (4.6)$$

Note that $f(x, s)$ and $g(x, s)$ both are supported on $A(s)$ and when $x \in A(s)$,

$$f(x, s)/g(x, s) = \phi_{-\lambda,1}(x)/\phi_{\lambda,1}(x) = e^{2\lambda x} \in [e^{-2\lambda x_s}, e^{2\lambda x_s}].$$

Apply Lemma 8(b), we have

$$\begin{aligned} & H^2(\Pi_i(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)) \\ & \leq \frac{e^{\lambda x_s} - 1}{e^{\lambda x_s} + 1} TV(\Pi_i(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)). \end{aligned}$$

Substitute into (4.4) and apply Fubini's theorem, we get

$$I \leq \int_0^{s^*} ds \frac{e^{\lambda x_s} - 1}{e^{\lambda x_s} + 1} \sum_{i=1}^m \sum_{j=0}^{J-1} \frac{1}{\sigma_j} \int_{\lambda \sigma_j}^{1-\lambda \sigma_j} TV(\Pi_i(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)) d\theta. \quad (4.7)$$

The following lemma bridges the partial total variation distances and communication costs, which is crucial to our proof.

Lemma 9. *If $\Pi : \mathbb{R} \rightarrow \{0, 1\}^b$ takes value in $\{0, 1\}^b$, then there exist a constant $C_1 > 0$ such that*

$$\sum_{j=0}^{J-1} \frac{1}{\sigma_j} \int_{\lambda \sigma_j}^{1-\lambda \sigma_j} TV(\Pi(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)) d\theta \leq C_1 x_s (1 + x_s) \sqrt{J(b \wedge J)}.$$

Another Lemma gives an upper bound on the integral by analysis.

Lemma 10. *If $\lambda \leq 1/6$ then there exists a constant $C_2 > 0$ such that*

$$\int_0^{s^*} \frac{e^{\lambda x_s} - 1}{e^{\lambda x_s} + 1} x_s (1 + x_s) ds \leq C_2 \lambda^2.$$

Apply Lemma 9 and 10 on (4.7), we have

$$I \leq C_1 C_2 \sqrt{J} \lambda^2 \sum_{i=1}^n \sqrt{(b_i \wedge J)}.$$

Jensen's inequality implies that $\sum_{i=1}^m \sqrt{b_i} \leq m \sqrt{1/m \sum_{i=1}^m b_i} = \sqrt{mB}$, therefore we have

$$I \leq C_1 C_2 \sqrt{mJB} \lambda^2.$$

Recall that $\lambda = c_\lambda \left(\frac{\log \frac{3}{\sigma_0}}{mB} \right)^{1/4}$. Note that $\log \frac{3}{\sigma_0} \leq 2J$, so when c_λ is a sufficiently small constant such that $0 < c_\lambda < \frac{1}{\sqrt{8C_1 C_2}}$, we have

$$I \leq \frac{J}{2}.$$

Recall the definition of I in (4.3):

$$I = \sum_{j=0}^{J-1} \int_{\lambda\sigma_j}^{1-\lambda\sigma_j} \sum_{i=1}^m H^2(\Pi_i; \phi_{(\theta-\lambda\sigma_j), \sigma_j}, \phi_{(\theta+\lambda\sigma_j), \sigma_j}) d\theta.$$

The above upper bound $I \leq J/2$ holds for any distributed estimator $\hat{\theta}$. Note that we have $B > \frac{1}{m}$ thus $\lambda\sigma_{J-1} < 1/6$ if we set $c_\lambda < 1/6$. Apply Lemma 7, we can conclude the desired lower bound:

$$R_{ind}(\sigma_0, B) \geq c\lambda^2 m \geq c_1 c_\lambda^2 \sqrt{\frac{m \log \frac{3}{\sigma_0}}{B}}. \quad \square$$

4.7.2. Proof of Theorem 16

For simplicity of notations we define $Z_{(-)} = Z_{(\lfloor 0.16m \rfloor)}$ and $Z_{(+)} = Z_{(\lceil 0.84m \rceil)}$. Before we proceed to the proof, we give a lemma showing large deviation bounds on $Z_{(-)}$ and $Z_{(+)}$. These bounds can be directly derived using Gaussian tail bounds so we omit the proof.

Lemma 11. *There exists universal constants $C, c > 0$ such that for any $k \geq 2$, we have*

$$\mathbb{P}(Z_{(-)} < \theta - k\sigma) \leq C \exp(-ck^2 m),$$

$$\mathbb{P}(Z_{(-)} > \theta - \sigma/2) \leq C \exp(-cm),$$

$$\mathbb{P}(Z_{(+)} > \theta + k\sigma) \leq C \exp(-ck^2m),$$

$$\mathbb{P}(Z_{(+)} < \theta + \sigma/2) \leq C \exp(-cm).$$

We first define several events:

$$E_1 = \{\theta \notin [Z_{(-)}, Z_{(+)}]\},$$

$$E_2 = \{\theta \in [Z_{(-)}, Z_{(+)}], \hat{\sigma} \notin [\min\{1, \frac{1}{2}\sigma\}, 4\sigma]\},$$

$$E_3 = (E_1 \cup E_2)^c = \{\theta \in [Z_{(-)}, Z_{(+)}], \min\{1, \frac{1}{2}\sigma\} \leq \hat{\sigma} \leq 4\sigma\}.$$

Note that we have

$$\mathbb{E}(\hat{\theta}_q - \theta)^2 = \sum_{k=1}^3 \mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_k\}}.$$

Therefore, the proof can be divided into showing $\mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_k\}} \leq C_k \frac{\sigma^2}{m}$ with some universal constant C_k respectively for $k = 1, 2, 3$.

1. Bound on $\mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_1\}}$.

Under E_1 , we have either $E_{11} = \{Z_{(-)} > \theta\}$ or $E_{12} = \{Z_{(+)} < \theta\}$ happens.

Define $E_{11,k} = \{Z_{(-)} > \theta, \theta + k\sigma < Z_{(+)} \leq \theta + (k+1)\sigma\}$. Under $E_{11,k}$, note that we have $Z_{(+)} - Z_{(-)} \leq (k+1)\sigma$, this implies $\hat{\sigma} \leq (k+1)\sigma$, then $\bar{\sigma} \leq 2(k+1)\sigma$, thus $R \leq \theta + 3(k+1)\sigma$. Note that the final estimate $\hat{\theta}_q$ must lie in the interval $[L, R]$, So we have $|\hat{\theta}_q - \theta| \leq 3(k+1)\sigma$ under event $E_{11,k}$.

Apply Lemma 11, when $k = 0, 1$ we have $\mathbb{P}(E_{11,k}) \leq \mathbb{P}(E_{11}) \leq C \exp(-cm)$. When $k \geq 2$ we have $\mathbb{P}(E_{11,k}) \leq \mathbb{P}(Z_{(+)} > \theta + k\sigma) \leq C \exp(-ck^2m)$. Therefore, note that $E_{11} = \bigcup_{k=0}^{\infty} E_{11,k}$,

we have

$$\begin{aligned}
\mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_{11}\}} &\leq \sum_{k=0}^{\infty} \mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_{11,k}\}} \\
&\leq (6\sigma)^2 \cdot 2C \exp(-cm) + \sum_{k=2}^{\infty} (3(k+1)\sigma)^2 \cdot C \exp(-ck^2m) \\
&\leq C' \frac{\sigma^2}{m}.
\end{aligned}$$

with some universal constant $C' > 0$.

By a symmetric argument we can also prove that $\mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_{12}\}} \leq C' \frac{\sigma^2}{m}$. Therefore, we conclude that

$$\mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_1\}} \leq \mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_{11}\}} + \mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_{12}\}} \leq 2C' \frac{\sigma^2}{m}.$$

2. Bound on $\mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_2\}}$.

Let $E_{21} = \{\theta \in [Z_{(-)}, Z_{(+)}], \hat{\sigma} < \min\{1, \frac{1}{2}\sigma\}\}$ and $E_{22,k} = \{\theta \in [Z_{(-)}, Z_{(+)}], k\sigma < \hat{\sigma} \leq (k+1)\sigma\}$ ($k \geq 4$).

Under E_{21} we have $|\hat{\theta} - \theta| < \frac{3}{2}\sigma$, under $E_{22,k}$ we have $|\hat{\theta} - \theta| < 3(k+1)\sigma$. Moreover, we have the probability bounds

$$\mathbb{P}(E_{21}) \leq \mathbb{P}(Z_{(+)} < \theta + \sigma/2) \leq C \exp(-ck^2m),$$

$$\mathbb{P}(E_{22,k}) \leq \mathbb{P}(Z_{(+)} > \theta + \frac{k}{2}\sigma) + \mathbb{P}(Z_{(-)} < \theta - \frac{k}{2}\sigma) \leq 2C \exp(-ck^2m/4).$$

Note that $E_2 = E_{21} \cup \bigcup_{k=4}^{\infty} E_{22,k}$, we have

$$\begin{aligned} \mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_2\}} &\leq \mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_{21}\}} + \sum_{k=4}^{\infty} \mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_{22,k}\}} \\ &\leq \left(\frac{3}{2}\sigma\right)^2 \cdot C \exp(-ck^2m) + \sum_{k=4}^{\infty} (3(k+1)\sigma)^2 \cdot 2C \exp(-ck^2m/4) \\ &\leq C'' \frac{\sigma^2}{m} \end{aligned}$$

with some universal constant $C'' > 0$.

3. Bound on $\mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_3\}}$.

Under event E_3 , because we have $\min\{1, \frac{1}{2}\sigma\} \leq \hat{\sigma} \leq 4\sigma$, also note that $\hat{\sigma} \leq 1$ almost surely, so there are at most 5 possible values of $\tilde{\sigma}$, whose range is between $\min\{1, \frac{1}{2}\sigma\}$ to $\min\{1, 8\sigma\}$ (recall that $\tilde{\sigma}$ is chosen only from powers of 2).

For each possible value of $\tilde{\sigma}$, the length of the interval $[L, R]$ is either $\tilde{\sigma}$ or $2\tilde{\sigma}$. Recall we have requirements that L, R are multiples of $\tilde{\sigma}$ and event E_3 suggest $\theta \in [L, R]$, so there are at most 5 possible values of (L, R) pairs for each possible value of $\tilde{\sigma}$. Putting together, we can conclude that under event E_3 , the possible values of the pair (L, R) is at most 25. We use $(L_1, R_1), (L_2, R_2), \dots, (L_{25}, R_{25})$ to denote these 25 possible values of the (L, R) pair. Thus we have the following decomposition:

$$\mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_3\}} = \sum_{k=1}^{25} \mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_3, (L, R) = (L_k, R_k)\}} \leq \sum_{k=1}^{25} \mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{(L, R) = (L_k, R_k)\}}. \quad (4.8)$$

For each possible pair (L_k, R_k) ($k = 1, 2, \dots, 25$), we have $R_k - L_k \leq 2\tilde{\sigma} \leq 16\sigma$. Define the function $F_k : (-\infty, \infty) \times (0, \infty) \rightarrow (0, 1) \times (0, 1)$ as

$$F_k(t, s) = \begin{pmatrix} \Phi\left(\frac{t-L_k}{s}\right) \\ \Phi\left(\frac{R_k-t}{s}\right) \end{pmatrix}.$$

When $(L, R) = (L_k, R_k)$, we have

$$\mathbb{E}\|F_k(\hat{\theta}_{sq}, \hat{\sigma}_{sq}) - F_k(\theta, \sigma)\|^2 = \mathbb{E}(\max\{\hat{p}_L, \frac{1}{m}\} - \mathbb{P}(X_1 < L))^2 + \mathbb{E}(\max\{\hat{p}_R, \frac{1}{m}\} - \mathbb{P}(X_1 > R))^2 \leq \frac{4}{m} \quad (4.9)$$

where the last inequality is due to mp_L is binomial distributed with mean $m\mathbb{P}(X_1 < L)$, and mp_R is binomial distributed with mean $m\mathbb{P}(X_1 > R)$.

Note that $R_k - \theta \leq R_k - L_k \leq 16\sigma$, therefore $\frac{t-L_k}{\sigma} < 16$ and $\frac{R_k-t}{\sigma} < 16$ for $t \in [L_k, R_k]$.

Then it is easy to prove that there exists a constant $c' = |\frac{d\Phi(x)}{dx}|_{x=16}| > 0$ such that for any $t, \in [L_k, R_k]$,

$$\begin{aligned} \left| \Phi\left(\frac{t-L_k}{\sigma}\right) - \Phi\left(\frac{\theta-L_k}{\sigma}\right) \right| &\geq c' \frac{|t-\theta|}{\sigma}; \\ \left| \Phi\left(\frac{R_k-t}{\sigma}\right) - \Phi\left(\frac{R_k-\theta}{\sigma}\right) \right| &\geq c' \frac{|t-\theta|}{\sigma}. \end{aligned}$$

Besides, note that for any $t \in [L, R]$ and $s > 0$, at least one of the following inequalities holds:

$$\begin{aligned} \left| \Phi\left(\frac{t-L_k}{s}\right) - \Phi\left(\frac{\theta-L_k}{\sigma}\right) \right| &> \left| \Phi\left(\frac{t-L_k}{\sigma}\right) - \Phi\left(\frac{\theta-L_k}{\sigma}\right) \right|; \\ \left| \Phi\left(\frac{R_k-t}{s}\right) - \Phi\left(\frac{R_k-\theta}{\sigma}\right) \right| &\geq \left| \Phi\left(\frac{R_k-t}{\sigma}\right) - \Phi\left(\frac{R_k-\theta}{\sigma}\right) \right|. \end{aligned}$$

Combine the two observations above, we have

$$\|F_k(\hat{\theta}_{sq}, \hat{\sigma}_{sq}) - F_k(\theta, \sigma)\|^2 \geq \frac{(c')^2}{\sigma^2} (\hat{\theta}_{sq} - \theta)^2.$$

Substitute above inequality into (4.9), when $(L, R) = (L_k, R_k)$ we have

$$\mathbb{E}(\hat{\theta}_{sq} - \theta)^2 \leq \frac{4}{(c')^2} \frac{\sigma^2}{m}.$$

Substitute the above inequality into (4.8) we obtained the desired bound

$$\mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_3\}} \leq \sum_{k=1}^{25} \mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{(L,R)=(L_k,R_k)\}} \leq \frac{100}{(c')^2} \frac{\sigma^2}{m}. \quad \square$$

4.7.3. Proof of Theorem 17

The proof of Theorem 17 will be carried out by several stages. Throughout the proof, we define $\delta_k = \frac{\hat{\theta}_k - \theta}{\sigma}$, $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is standard Gaussian density. $\Phi(x) = \mathbb{P}(X > x)$ where $X \sim N(0, 1)$, and $\Lambda(x) = 1 - 2\Phi(x)$. We also define $\mu_k = \frac{1}{k-10}$.

We first give a lemma that will be useful in the proof.

Lemma 12. *Let $\{A_k\}_{k=0}^\infty$ be a positive sequence, and $\{b_k\}_{k=0}^\infty, \{d_k\}_{k=0}^\infty$ be two decreasing positive sequences that satisfy*

$$A_k \leq (1 - \alpha b_k) A_{k-1} + \beta b_k d_k, \quad k = 1, 2, \dots$$

where $\alpha, \beta > 0$. If there exists $K > 0$ such that

$$\frac{d_{k-1}}{d_k} \leq 1 + \frac{\alpha}{2} b_k \text{ for all } k \geq K.$$

Then we have for all $k \geq 0$,

$$A_k \leq \left(\frac{A_0 + \beta \sum_{i=1}^K b_i d_i}{d_K} + \frac{2\beta}{\alpha} \right) d_k. \quad (4.10)$$

Then we provide several claims and show the proof to each claim directly after their statement.

Claim 1. *There exists a constant $C_1 > 0$ (doesn't depend on θ, σ or $\hat{\sigma}$) such that for all $k \geq 11$, we have*

$$\mathbb{E}[(\hat{\theta}_k - \theta)^2 | \hat{\sigma}] \leq C_1 ((\hat{\sigma}/\sigma)^{-2} + (\hat{\sigma}/\sigma)^2) \sigma^2 \gamma_k. \quad (4.11)$$

PROOF OF CLAIM 1. Define the Lyapunov function

$$L(x) = \begin{cases} x^2 & \text{if } -2 < x < 2 \\ 4e^{|x|/2-1} & \text{if } |x| \geq 2 \end{cases}.$$

Note that $x^2 \lesssim L(x)$, therefore to prove Claim 1, it suffices to show that

$$\mathbb{E} \left[L \left(\frac{\hat{\theta}_k - \theta}{\sigma} \right) \middle| \hat{\sigma} \right] \lesssim ((\hat{\sigma}/\sigma)^{-2} + (\hat{\sigma}/\sigma)^2) \gamma_k. \quad (4.12)$$

We have the following lemma.

Lemma 13. (a) *If $\gamma_k \hat{\sigma}/\sigma \geq 2$, we have*

$$\mathbb{E}[L(\delta_k)|\hat{\sigma}] \leq 11e^{\gamma_k \hat{\sigma}/(2\sigma)}. \quad (4.13)$$

(b) *If $\gamma_k \hat{\sigma}/\sigma \leq 2$ and $k \geq 12$, we have*

$$\mathbb{E}[L(\delta_k)|\hat{\sigma}] \leq (1 - 0.25\gamma_k \hat{\sigma}/\sigma) \mathbb{E}[L(\delta_{k-1})|\hat{\sigma}] + (\gamma_k \hat{\sigma}/\sigma)^2. \quad (4.14)$$

Case 1: When $\gamma_k \hat{\sigma}/\sigma \geq 2$. Consider surrogate function

$$\tilde{L}(x) = \begin{cases} 4 & \text{if } 0 \leq y \leq 4 \\ 4e^{\sqrt{x}/2-1} & \text{if } x \geq 4 \end{cases}.$$

Note that $\tilde{L}(\delta_k^2) \leq L(\delta_k) + 4$ and $\tilde{L}(x)$ is convex, apply Lemma 13(a) and Jensen's inequality we have

$$\tilde{L}(\mathbb{E}[\delta_k^2|\hat{\sigma}]) \leq \mathbb{E}[\tilde{L}(\delta_k^2)|\hat{\sigma}] \leq \mathbb{E}[L(\delta_k)|\hat{\sigma}] + 4 \leq 11e^{\gamma_k \hat{\sigma}/(2\sigma)} + 4$$

which suggests that

$$\mathbb{E}[\delta_k^2|\hat{\sigma}] \lesssim (\gamma_k \hat{\sigma}/\sigma)^2 \leq (\hat{\sigma}/\sigma)^2 \gamma_k.$$

Case 2: When $\gamma_k \hat{\sigma}/\sigma < 2$. Let k_0 be the largest k such that $\gamma_k \hat{\sigma}/\sigma \geq 2$ (if there is no such k , set $k_0 = 0$). Given Lemma 13(b), we can apply Lemma 12 with $A_i = \mathbb{E}[L(\delta_{k_0+i})|\hat{\sigma}]$, $b_i = d_i = \gamma_{k_0+i} \hat{\sigma}/\sigma$, $\alpha = 0.25$, $\beta = 1$, and $K = \lceil 8^3(\hat{\sigma}/\sigma)^{-3} \rceil - k_0$. This is a valid K value because

$$d_{i-1}/d_i = (1 - 1/k)^{-2/3} \leq 1 + 1/k \leq 1 + \frac{k^{-2/3} \hat{\sigma}}{8\sigma} = 1 + \frac{\alpha}{2} b_k \text{ when } i \geq 8^3(\hat{\sigma}/\sigma)^{-3}$$

where $k = k_0 + i$.

Also note that we have $\gamma_{k_0} \hat{\sigma}/\sigma \leq 4$ due to the definition of k_0 , thus $A_0 \leq 11e^2$ according to Lemma 13(a). And note that $\sum_{i=1}^K b_i d_i < \sum_{i=1}^{\infty} b_i d_i = (\hat{\sigma}/\sigma)^2 \sum_{i=1+k_0}^{\infty} \gamma_i^2 < \infty$. Therefore, apply Lemma 12, we have

$$\mathbb{E}[L(\delta_k)|\hat{\sigma}] \lesssim \left(\frac{11e^2 + (\hat{\sigma}/\sigma)^2 \sum_{i=1+k_0}^{\infty} \gamma_i^2}{(\hat{\sigma}/\sigma)^3} + 8 \right) d_k \lesssim ((\hat{\sigma}/\sigma)^{-3} + 1) \hat{\sigma}/\sigma \gamma_k.$$

Combine the two cases above, we prove the desired bound (4.11).

Claim 2. *There exists a constant $C_2 > 0$ (doesn't depend on θ, σ or $\hat{\sigma}$) such that for all $k \geq 11$, we have*

$$\mathbb{E}[(\hat{\theta}_k - \theta)^4|\hat{\sigma}] \leq C_2 ((\hat{\sigma}/\sigma)^{-4} + (\hat{\sigma}/\sigma)^4) \sigma^4 \gamma_k^2. \quad (4.15)$$

PROOF OF CLAIM 2. The proof is very similar to Claim 1. We will omit some details in the proof. Re-define the Lyapunov function

$$L(x) = \begin{cases} x^4 & \text{if } -2 < x < 2 \\ 16e^{|x|/2-1} & \text{if } |x| \geq 2 \end{cases}.$$

We have the following lemma.

Lemma 14. (a) If $\gamma_k \hat{\sigma} / \sigma \geq 2$, we have

$$\mathbb{E}[L(\delta_k) | \hat{\sigma}] \leq 44e^{\gamma_k \hat{\sigma} / (2\sigma)}. \quad (4.16)$$

(b) If $\gamma_k \hat{\sigma} / \sigma \leq 2$ and $k \geq 12$, we have

$$\mathbb{E}[L(\delta_k) | \hat{\sigma}] \leq (1 - 0.25\gamma_k \hat{\sigma} / \sigma) \mathbb{E}[L(\delta_{k-1}) | \hat{\sigma}, \delta_{k-1}] + (6C_1 + 1)((\hat{\sigma} / \sigma)^4 + 1)\gamma_k^3. \quad (4.17)$$

Case 1: When $\gamma_k \hat{\sigma} / \sigma \geq 2$. Similarly we can conclude that

$$\mathbb{E}[\delta_k^4 | \hat{\sigma}] \lesssim (\gamma_k \hat{\sigma} / \sigma)^4 \leq (\hat{\sigma} / \sigma)^4 \gamma_k^2.$$

Case 2: When $\gamma_k \hat{\sigma} / \sigma < 2$. Let $i = k - k_0$ where k_0 is defined as in the proof of Claim 1. Given Lemma 14(b), we can apply Lemma 12 with $b_i = \gamma_k \hat{\sigma} / \sigma$, $d_i = \gamma_k^2$, $\alpha = 0.25$, $\beta = (6C_1 + 1)((\hat{\sigma} / \sigma)^{-1} + (\hat{\sigma} / \sigma)^3)$, and $K = \lceil 16^3 (\hat{\sigma} / \sigma)^{-3} \rceil - k_0$, then we have

$$\mathbb{E}[L(\delta_k) | \hat{\sigma}] \lesssim \left(\frac{44e^2 + (\hat{\sigma} / \sigma)^4 + 1}{(\hat{\sigma} / \sigma)^4} + 8((\hat{\sigma} / \sigma)^{-1} + (\hat{\sigma} / \sigma)^3) \right) \gamma_k^2 \lesssim ((\hat{\sigma} / \sigma)^{-4} + (\hat{\sigma} / \sigma)^3) \gamma_k^2.$$

Combine the two cases above, and note that $\mathbb{E}[(\hat{\theta}_k - \theta)^4] \lesssim \mathbb{E}[L(\delta_k) | \hat{\sigma}]$, we can conclude (4.15).

Claim 3. Let $\bar{\theta}_k = \frac{1}{k} \sum_{i=1}^k \hat{\theta}_i$. There exists a constant $C_3 > 0$ (doesn't depend on θ, σ or $\hat{\sigma}$) such that

$$\mathbb{E}[(\bar{\theta}_k - \theta)^2 | \hat{\sigma}] \leq C_3 ((\hat{\sigma} / \sigma)^{-2} + (\hat{\sigma} / \sigma)^2) \sigma^2 \gamma_k.$$

PROOF OF CLAIM 3. Let $\mu_k = \frac{1}{k-10}$, note that

$$\bar{\theta}_{k+1} - \theta = (1 - \mu_k)(\bar{\theta}_k - \theta) + \mu_k(\hat{\theta}_{k+1} - \theta).$$

This implies

$$\mathbb{E}[(\bar{\theta}_{k+1} - \theta)^2 | \hat{\sigma}]^{1/2} \leq (1 - \mu_k) \mathbb{E}[(\bar{\theta}_k - \theta)^2 | \hat{\sigma}]^{1/2} + \mu_k \mathbb{E}[(\hat{\theta}_i - \theta)^2 | \hat{\sigma}]^{1/2}. \quad (4.18)$$

From the above inequality we can show

$$\mathbb{E}[(\bar{\theta}_k - \theta)^2 | \hat{\sigma}]^{1/2} \leq 3 (C_1 ((\hat{\sigma}/\sigma)^{-2} + (\hat{\sigma}/\sigma)^2) \sigma^2 \gamma_k)^{1/2} \quad (4.19)$$

holds for all $k \geq 1$ by induction, which suggest Claim 3 holds with $C_3 = 9C_1$. The induction is concluded by:

1. From Claim 1 we have

$$\mathbb{E}[(\bar{\theta}_{11} - \theta)^2 | \hat{\sigma}]^{1/2} = E[(\hat{\theta}_{11} - \theta)^2 | \hat{\sigma}]^{1/2} \leq (C_1 ((\hat{\sigma}/\sigma)^{-2} + (\hat{\sigma}/\sigma)^2) \sigma^2 \gamma_{11})^{1/2},$$

therefore (4.19) holds when $k = 11$.

2. If (4.19) holds for k , from (4.18) and Claim 1 we have

$$\mathbb{E}[(\bar{\theta}_{k+1} - \theta)^2 | \hat{\sigma}]^{1/2} \leq \left(3(1 - \mu_k) \sqrt{\frac{\gamma_k}{\gamma_{k+1}}} + \mu_k \right) (C_1 ((\hat{\sigma}/\sigma)^{-2} + (\hat{\sigma}/\sigma)^2) \sigma^2 \gamma_{k+1})^{1/2}.$$

Note that $\sqrt{\frac{\gamma_k}{\gamma_{k+1}}} < (1 - \mu_k)^{-1/3}$ and $3(1 - \mu_k)^{2/3} + \mu_k \leq 3$ for all k . So we have (4.19) holds for $k + 1$.

Claim 4. Let $\bar{\theta}_k = \frac{1}{k} \sum_{i=1}^k \hat{\theta}_i$. There exists a constant $C_4 > 0$ (doesn't depend on θ, σ or $\hat{\sigma}$)

such that

$$\mathbb{E}[(\bar{\theta}_k - \theta)(\hat{\theta}_k - \theta)|\hat{\sigma}] \leq C_4 ((\hat{\sigma}/\sigma)^{-5} + (\hat{\sigma}/\sigma)^3) \sigma^2 \mu_k. \quad (4.20)$$

PROOF OF CLAIM 4. We have

$$\mathbb{E}[(\bar{\theta}_{k+1} - \theta)(\hat{\theta}_{k+1} - \theta)|\hat{\sigma}, \hat{\theta}_k] = (1 - \mu_k) \mathbb{E}[(\bar{\theta}_k - \theta)(\hat{\theta}_{k+1} - \theta)|\hat{\sigma}, \hat{\theta}_k] + \mu_k \mathbb{E}[(\hat{\theta}_{k+1} - \theta)^2|\hat{\sigma}, \hat{\theta}_k].$$

Take expectation with respect to $\hat{\theta}_k$ we have

$$\begin{aligned} & \mathbb{E}[(\bar{\theta}_{k+1} - \theta)(\hat{\theta}_{k+1} - \theta)|\hat{\sigma}] \\ &= (1 - \mu_k) \mathbb{E}[(\bar{\theta}_k - \theta)(\hat{\theta}_k - \theta - \hat{\sigma} \gamma_k \Lambda(\delta_k))|\hat{\sigma}] + \mu_k \mathbb{E}[(\hat{\theta}_{k+1} - \theta)^2|\hat{\sigma}] \\ &= (1 - \mu_k)(1 - 2\hat{\sigma}/\sigma \gamma_k \phi(0)) \mathbb{E}[(\bar{\theta}_k - \theta)(\hat{\theta}_k - \theta)|\hat{\sigma}] + (1 - \mu_k) \hat{\sigma} \gamma_k \mathbb{E}[(\bar{\theta}_k - \theta)(2\phi(0)\delta_k - \Lambda(\delta_k))|\hat{\sigma}] \\ & \quad + \mu_k \mathbb{E}[(\hat{\theta}_{k+1} - \theta)^2|\hat{\sigma}]. \end{aligned}$$

Note that $2\phi(0)\delta_k - \Lambda(\delta_k) \lesssim \delta_k^2$. Cauchy-Schwartz inequality suggests

$$\begin{aligned} \mathbb{E}[(\bar{\theta}_k - \theta)(2\phi(0)\delta_k - \Lambda(\delta_k))|\hat{\sigma}]^2 &\leq \mathbb{E}[(\bar{\theta}_k - \theta)^2|\hat{\sigma}] \mathbb{E}[(2\phi(0)\delta_k - \Lambda(\delta_k))^2|\hat{\sigma}] \\ &\lesssim \mathbb{E}[(\bar{\theta}_k - \theta)^2|\hat{\sigma}] \mathbb{E}[\delta_k^4|\hat{\sigma}] \\ &\lesssim ((\hat{\sigma}/\sigma)^{-2} + (\hat{\sigma}/\sigma)^2) \sigma^2 \gamma_k \cdot ((\hat{\sigma}/\sigma)^{-4} + (\hat{\sigma}/\sigma)^4) \gamma_k^2 \\ &\lesssim ((\hat{\sigma}/\sigma)^{-6} + (\hat{\sigma}/\sigma)^6) \sigma^2 \gamma_k^3. \end{aligned}$$

Thus we have

$$\begin{aligned}
& \mathbb{E}[(\bar{\theta}_{k+1} - \theta)(\hat{\theta}_{k+1} - \theta)|\hat{\sigma}] \\
& \leq (1 - \mu_k)(1 - 2\hat{\sigma}/\sigma\gamma_k\phi(0))\mathbb{E}[(\bar{\theta}_k - \theta)(\hat{\theta}_k - \theta)|\hat{\sigma}] \\
& \quad + C'_4(1 - \mu_k) \left((\hat{\sigma}/\sigma)^{-3} + (\hat{\sigma}/\sigma)^3 \right) \hat{\sigma}\sigma\gamma_k^{5/2} \\
& \quad + \mu_k C_1 \left((\hat{\sigma}/\sigma)^{-2} + (\hat{\sigma}/\sigma)^2 \right) \sigma^2\gamma_k \\
& \leq (1 - 2\hat{\sigma}/\sigma\gamma_k\phi(0))\mathbb{E}[(\bar{\theta}_k - \theta)(\hat{\theta}_k - \theta)|\hat{\sigma}] + (C'_4 + C_1) \left((\hat{\sigma}/\sigma)^{-2} + (\hat{\sigma}/\sigma)^4 \right) \sigma^2\mu_k\gamma_k.
\end{aligned}$$

with some constant $C'_4 > 0$. The last inequality is due to the fact that $\gamma_k^{5/2} \leq \mu_k\gamma_k$.

Now we apply Lemma 12 with $b_k = \hat{\sigma}/\sigma\gamma_k$, $d_k = \mu_k$, $\alpha = 2\phi(0)$,

$\beta = (C'_4 + C_1) \left((\hat{\sigma}/\sigma)^{-3} + (\hat{\sigma}/\sigma)^3 \right) \sigma^2$, and $K = (\phi(0)\hat{\sigma}/\sigma)^{-3}$. We have

$$\mathbb{E}[(\bar{\theta}_k - \theta)(\hat{\theta}_k - \theta)|\hat{\sigma}] \lesssim \left(\frac{\sigma^2 + \left((\hat{\sigma}/\sigma)^{-2} + (\hat{\sigma}/\sigma)^4 \right) \sigma^2}{(\hat{\sigma}/\sigma)^3} + \left((\hat{\sigma}/\sigma)^{-3} + (\hat{\sigma}/\sigma)^3 \right) \sigma^2 \right) \mu_k.$$

Then we can conclude (4.20).

Claim 5. *There exists a constant $C_5 > 0$ such that*

$$\mathbb{E}[(\bar{\theta}_k - \theta)^2|\hat{\sigma}] \leq C_5 \left((\hat{\sigma}/\sigma)^{-5} + (\hat{\sigma}/\sigma)^3 \right) \sigma^2\mu_k. \tag{4.21}$$

PROOF OF CLAIM 5. Note that we have

$$\begin{aligned}
& E[(\bar{\theta}_{k+1} - \theta)^2 | \hat{\sigma}] \\
&= (1 - \mu_k)^2 E[(\bar{\theta}_k - \theta)^2 | \hat{\sigma}] + 2\mu_k(1 - \mu_k) E[(\bar{\theta}_k - \theta)(\hat{\theta}_{k+1} - \theta) | \hat{\sigma}] + \mu_k^2 E[(\hat{\theta}_{k+1} - \theta)^2 | \hat{\sigma}] \\
&\leq (1 - 2\mu_k + \mu_k^2) E[(\bar{\theta}_k - \theta)^2 | \hat{\sigma}] \\
&\quad + 2C_4 ((\hat{\sigma}/\sigma)^{-5} + (\hat{\sigma}/\sigma)^3) \sigma^2 \mu_k^2 \\
&\quad + C_3 ((\hat{\sigma}/\sigma)^{-2} + (\hat{\sigma}/\sigma)^2) \sigma^2 \gamma_k \mu_k^2 \\
&\leq (1 - 2\mu_k + \mu_k^2) E[(\bar{\theta}_k - \theta)^2 | \hat{\sigma}] + (2C_4 + C_3) ((\hat{\sigma}/\sigma)^{-5} + (\hat{\sigma}/\sigma)^3) \sigma^2 \mu_k^2
\end{aligned}$$

which implies there exists a constant $C_5 > 0$ such that

$$E[(\bar{\theta}_{k+1} - \theta)^2 | \hat{\sigma}] \leq C_5 ((\hat{\sigma}/\sigma)^{-5} + (\hat{\sigma}/\sigma)^3) \sigma^2 \mu_k.$$

PROOF OF THE THEOREM. Now we are ready to prove the theorem. Take expectation on (4.21) with respect to $\hat{\sigma}$, we have

$$\mathbb{E}[(\bar{\theta}_k - \theta)^2] \leq C_5 \mathbb{E}[(\hat{\sigma}/\sigma)^{-5} + (\hat{\sigma}/\sigma)^3] \sigma^2 \mu_k.$$

Note that $\hat{\sigma}$ is the empirical estimate of σ over 10 observations. Thus we have $\mathbb{E}((\hat{\sigma}/\sigma)^{-5}) < \infty$ and $\mathbb{E}(\hat{\sigma}/\sigma)^3 < \infty$, therefore there exists a constant $C_6 > 0$ such that

$$\mathbb{E}[(\bar{\theta}_k - \theta)^2] \leq C_6 \sigma^2 \mu_k.$$

Substitute $k = m$ into the above equation, also note that $(\hat{\theta}_{sq} - \theta)^2 \leq (\bar{\theta}_k - \theta)^2$ and $(\hat{\theta}_{sq} - \theta)^2 \leq 1$, we conclude the theorem. \square

4.7.4. Proof of Lemma 7

Note that the length of each integral interval $(\lambda\sigma_j, 1 - \lambda\sigma_j)$ is at least $2/3$. Therefore, for any distributed estimator $\hat{\theta}$, there exists $0 \leq j^* \leq J - 1$ and $\theta^* \in [0, 1]$ such that

$$\sum_{i=1}^m H^2(Z_i|X_i \sim N(\theta^* - \lambda\sigma_{j^*}, \sigma_{j^*}^2), Z_i|X_i \sim N(\theta^* + \lambda\sigma_{j^*}, \sigma_{j^*}^2)) \leq \frac{3}{4}$$

where $Z_i|X_i \sim P$ denotes the distribution of $\Pi_i(X_i)$ when $X_i \sim P$, H^2 denotes the squared Hellinger distances.

Note that $Z_i, i = 1, 2, \dots, m$ are independent, by sub-additivity of squared Hellinger distances for product measures, we have

$$H^2((Z_i)_{i=1}^m |_{X_i \sim N(\theta^* - \lambda\sigma_{j^*}, \sigma_{j^*}^2)}, \text{ for } i=1, 2, \dots, m, (Z_i)_{i=1}^m |_{X_i \sim N(\theta^* + \lambda\sigma_{j^*}, \sigma_{j^*}^2)}, \text{ for } i=1, 2, \dots, m) \leq \frac{3}{4}$$

where $(Z_i)_{i=1}^m$ is a shorthand for (Z_1, Z_2, \dots, Z_m)

Note that the distributed estimator $\hat{\theta}$ is a (possibly random) function of $(Z_i)_{i=1}^m$. Given that the squared Hellinger distance between the distribution of $(Z_i)_{i=1}^m$ under $N(\theta^* - \lambda\sigma_{j^*}, \sigma_{j^*}^2)$ and $N(\theta^* + \lambda\sigma_{j^*}, \sigma_{j^*}^2)$ are bounded by $3/4$, which means we cannot “distinguish” whether the data are drawn from $N(\theta^* - \lambda\sigma_{j^*}, \sigma_{j^*}^2)$ or $N(\theta^* + \lambda\sigma_{j^*}, \sigma_{j^*}^2)$ by looking at those transcripts, we can apply Le Cam’s method to conclude a minimax lower bound: when $\sigma = \sigma_{j^*}$, there exists a constant $c_1 > 0$ such that

$$\sup_{\theta \in \{\theta^* - \lambda\sigma_{j^*}, \theta^* + \lambda\sigma_{j^*}\}} \mathbb{E}(\hat{\theta} - \theta)^2 \geq c_1 \lambda^2 \sigma^{*2},$$

which is equivalent to

$$\sup_{\theta \in \{\theta^* - \lambda\sigma_{j^*}, \theta^* + \lambda\sigma_{j^*}\}} \left(\frac{\sigma^{*2}}{m} \right)^{-1} \mathbb{E}(\hat{\theta} - \theta)^2 \geq c_1 \lambda^2 m.$$

Thus we can conclude that

$$R_{ind}(\sigma_0, B) \geq c\lambda^2 m. \quad \square$$

4.7.5. Proof of Lemma 8

PROOF OF (4.1). For any $z \in \mathcal{Z}$, define

$$F_z(s) = \int_{-\infty}^{\infty} f(x, s) \mathbb{P}(h(x) = z) dx,$$

$$G_z(s) = \int_{-\infty}^{\infty} g(x, s) \mathbb{P}(h(x) = z) dx.$$

By definition, we have

$$\begin{aligned} H^2(h; \int_{s_l}^{s_r} f(\cdot, s) ds, \int_{s_l}^{s_r} g(\cdot, s) ds) &= \frac{1}{2} \sum_{z \in \mathcal{Z}} \left(\sqrt{\int_{s_l}^{s_r} F_z(s) ds} - \sqrt{\int_{s_l}^{s_r} G_z(s) ds} \right)^2 \\ &= \frac{1}{2} \sum_{z \in \mathcal{Z}} \left(\int_{s_l}^{s_r} F_z(s) ds + \int_{s_l}^{s_r} G_z(s) ds - 2 \sqrt{\int_{s_l}^{s_r} F_z(s) ds \int_{s_l}^{s_r} G_z(s) ds} \right), \end{aligned}$$

$$\begin{aligned} \int_{s_l}^{s_r} H^2(h; f(\cdot, s), g(\cdot, s)) ds &= \frac{1}{2} \sum_{z \in \mathcal{Z}} \int_{s_l}^{s_r} \left(\sqrt{F_z(s)} - \sqrt{G_z(s)} \right)^2 ds \\ &= \frac{1}{2} \sum_{z \in \mathcal{Z}} \left(\int_{s_l}^{s_r} F_z(s) ds + \int_{s_l}^{s_r} G_z(s) ds - 2 \int_{s_l}^{s_r} \sqrt{F_z(s) G_z(s)} ds \right). \end{aligned}$$

Therefore, from Cauchy-Schwartz inequality

$$\sqrt{\int_{s_l}^{s_r} F_z(s) ds \int_{s_l}^{s_r} G_z(s) ds} \geq \int_{s_l}^{s_r} \sqrt{F_z(s) G_z(s)} ds,$$

we can conclude (4.1).

PROOF OF (4.2). Define

$$F_z = \int_{-\infty}^{\infty} f(x) \mathbb{P}(h(x) = z) dx,$$

$$G_z = \int_{-\infty}^{\infty} g(x)\mathbb{P}(h(x) = z)dx.$$

$1/M \leq f(x)/g(x) \leq M$ for all $x \in \{x : g(x) > 0\}$ implies that $1/M \leq F_z/G_z \leq M$ when $F(z) > 0$ or $G(z) > 0$. This suggests that

$$\left| \frac{\sqrt{F_z} - \sqrt{G_z}}{\sqrt{F_z} + \sqrt{G_z}} \right| \leq \frac{\sqrt{M} - 1}{\sqrt{M} + 1}.$$

By definition we have

$$H^2(h; f, g) = \frac{1}{2} \sum_{z \in \mathcal{Z}} (\sqrt{F_z} - \sqrt{G_z})^2 = \frac{1}{2} \sum_{z \in \mathcal{Z}} \left| \frac{\sqrt{F_z} - \sqrt{G_z}}{\sqrt{F_z} + \sqrt{G_z}} \right| |\sqrt{F_z} - \sqrt{G_z}| \leq \frac{\sqrt{M} - 1}{\sqrt{M} + 1} TV(h; f, g). \quad \square$$

4.7.6. Proof of Lemma 9

First, note that by definition we have

$$TV(\Pi(x); f((x - \theta)/\sigma_j, s), g((x - \theta)/\sigma_j, s)) \leq x_s \sigma_j.$$

So we have

$$\sum_{j=0}^{J-1} \frac{1}{\sigma_j} \int_{\lambda \sigma_j}^{1 - \lambda \sigma_j} TV(\Pi(x); f((x - \theta)/\sigma_j, s), g((x - \theta)/\sigma_j, s)) d\theta \leq x_s J.$$

Therefore, it only remains to prove

$$\sum_{j=0}^{J-1} \frac{1}{\sigma_j} \int_{\lambda \sigma_j}^{1 - \lambda \sigma_j} TV(\Pi(x); f((x - \theta)/\sigma_j, s), g((x - \theta)/\sigma_j, s)) d\theta \leq C_1 x_s (1 + x_s) \sqrt{Jb}.$$

The next technical lemma is the key to prove Lemma 9.

Lemma 15. Let $\{a_k\}, k = 1, 2, \dots, 2^J$ be a non-negative sequence such that

$$0 \leq a_k \leq 1, k = 1, 2, \dots, 2^J.$$

Then there exists a constant $C_3 > 0$ such that

$$\sum_{j=1}^J \sum_{l=1}^{2^{J-j}} \left| \sum_{k=(l-1)2^j+1}^{(l-1)2^j+2^{j-1}} a_k - \sum_{k=(l-1)2^j+2^{j-1}+1}^{l \cdot 2^j} a_k \right| \leq C_3 2^J \sqrt{J} \int_0^w \sqrt{-\log t} dt$$

where $w = 2^{-J} \sum_{k=1}^{2^J} a_k$ is the mean of the sequence.

Let $x'_s = \inf_{x \in A(s)} |x|$. For any real number $\theta, z \in \{0, 1\}^b$ and $k \in [2^J]$, let $a_k(\theta, z) = \int_{\theta+(k-1)x_s\sigma_0}^{\theta+kx_s\sigma_0} \mathbb{P}(\Pi(x) = z) dx$ and $a'_k(\theta, z) = \int_{\theta+(k-1)x'_s\sigma_0}^{\theta+kx'_s\sigma_0} \mathbb{P}(\Pi(x) = z) dx$. Note that it is easy to check $A(s) = [-x_s, -x'_s] \cup [x'_s, x_s]$, so we have

$$\begin{aligned} & TV(\Pi(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)) \\ &= \frac{1}{2} \sum_{z \in \{0,1\}^b} \left| \int_{\theta-\sigma_j x_s}^{\theta-\sigma_j x'_s} \mathbb{P}(\Pi(x) = z) dx - \int_{\theta+\sigma_j x'_s}^{\theta+\sigma_j x_s} \mathbb{P}(\Pi(x) = z) dx \right| \\ &\leq \frac{1}{2} \sum_{z \in \{0,1\}^b} \left| \int_{\theta-\sigma_j x_s}^{\theta} \mathbb{P}(\Pi(x) = z) dx - \int_{\theta}^{\theta+\sigma_j x_s} \mathbb{P}(\Pi(x) = z) dx \right| \\ &\quad + \frac{1}{2} \sum_{z \in \{0,1\}^b} \left| \int_{\theta-\sigma_j x'_s}^{\theta} \mathbb{P}(\Pi(x) = z) dx - \int_{\theta}^{\theta+\sigma_j x'_s} \mathbb{P}(\Pi(x) = z) dx \right| \\ &= \frac{1}{2^{J-j}} \sum_{z \in \{0,1\}^b} \sum_{r=1}^{2^{J-j-1}} \left| \sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j} a_k(\theta - \sigma_j x_s - 2^{j+1}(r-1)x_s\sigma_0, z) \right. \\ &\quad \left. - \sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r} a_k(\theta - \sigma_j x_s - 2^{j+1}(r-1)x_s\sigma_0, z) \right| \\ &\quad + \frac{1}{2^{J-j}} \sum_{z \in \{0,1\}^b} \sum_{r=1}^{2^{J-j-1}} \left| \sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j} a'_k(\theta - \sigma_j x'_s - 2^{j+1}(r-1)x'_s\sigma_0, z) \right. \\ &\quad \left. - \sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r} a'_k(\theta - \sigma_j x'_s - 2^{j+1}(r-1)x'_s\sigma_0, z) \right|. \end{aligned}$$

Substitute the above inequality and rewrite the integral variable, also recall that $\sigma_j = 2^j \sigma_0$,

we have

$$\begin{aligned}
& \sum_{j=0}^{J-1} \frac{1}{\sigma_j} \int_{\lambda \sigma_j}^{1-\lambda \sigma_j} TV(\Pi(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)) d\theta \\
& \leq \frac{1}{2^J \sigma_0} \sum_{j=0}^{J-1} \sum_{z \in \{0,1\}^b} \sum_{r=1}^{2^{J-j-1}} \int_{\lambda \sigma_j - \sigma_j x_s - 2^{j+1}(r-1)x_s \sigma_0}^{1-\lambda \sigma_j - \sigma_j x_s - 2^{j+1}(r-1)x_s \sigma_0} \\
& \quad \left| \sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j} a_k(\theta, z) - \sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r} a_k(\theta, z) \right| d\theta \\
& \quad + \frac{1}{2^J \sigma_0} \sum_{j=0}^{J-1} \sum_{z \in \{0,1\}^b} \sum_{r=1}^{2^{J-j-1}} \int_{\lambda \sigma_j - \sigma_j x'_s - 2^{j+1}(r-1)x'_s \sigma_0}^{1-\lambda \sigma_j - \sigma_j x'_s - 2^{j+1}(r-1)x'_s \sigma_0} \\
& \quad \left| \sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j} a'_k(\theta, z) - \sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r} a'_k(\theta, z) \right| d\theta \\
& \leq \frac{1}{2^J \sigma_0} \sum_{j=0}^{J-1} \sum_{z \in \{0,1\}^b} \sum_{r=1}^{2^{J-j-1}} \int_{-x_s}^1 \left| \sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j} a_k(\theta, z) - \sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r} a_k(\theta, z) \right| d\theta \\
& \quad + \frac{1}{2^J \sigma_0} \sum_{j=0}^{J-1} \sum_{z \in \{0,1\}^b} \sum_{r=1}^{2^{J-j-1}} \int_{-x'_s}^1 \left| \sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j} a'_k(\theta, z) - \sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r} a'_k(\theta, z) \right| d\theta \\
& = \frac{x_s \sigma_0}{2^J \sigma_0} \int_{-x_s}^1 d\theta \sum_{z \in \{0,1\}^b} \sum_{j=0}^{J-1} \sum_{r=1}^{2^{J-j-1}} \left| \sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j} \frac{a_k(\theta, z)}{x_s \sigma_0} - \sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r} \frac{a_k(\theta, z)}{x_s \sigma_0} \right| \\
& \quad + \frac{x'_s \sigma_0}{2^J \sigma_0} \int_{-x'_s}^1 d\theta \sum_{z \in \{0,1\}^b} \sum_{j=0}^{J-1} \sum_{r=1}^{2^{J-j-1}} \left| \sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j} \frac{a'_k(\theta, z)}{x'_s \sigma_0} - \sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r} \frac{a'_k(\theta, z)}{x'_s \sigma_0} \right|. \tag{4.22}
\end{aligned}$$

Define $w(\theta, z) \triangleq 2^{-J} \sum_{k=1}^{2^J} a_k(\theta, z)/(x_s \sigma_0) = \frac{1}{2^J x_s \sigma_0} \int_{\theta}^{\theta+2^J x_s \sigma_0} \mathbb{P}(\Pi(x) = z)$. Note that $a_k(\theta, z)/(x_s \sigma_0) \in [0, 1]$, apply Lemma 15 gives

$$\begin{aligned}
& \sum_{z \in \{0,1\}^b} \sum_{j=0}^{J-1} \sum_{r=1}^{2^{J-j-1}} \left| \sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j} a_k(\theta, z)/(x_s \sigma_0) - \sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r} a_k(\theta, z)/(x_s \sigma_0) \right| \\
& \leq C_3 2^J \sqrt{J} \sum_{z \in \{0,1\}^b} \int_0^{w(\theta, z)} \sqrt{-\log t} dt.
\end{aligned}$$

Then note that $\int_0^w \sqrt{-\log t} dt$ is a concave function of w and $\sum_{z \in \{0,1\}^b} w(\theta, z) = 1$, we can apply Jensen's inequality to get

$$\sum_{z \in \{0,1\}^b} \int_0^{w(\theta, z)} \sqrt{-\log t} dt \leq 2^b \int_0^{2^{-b}} \sqrt{-\log t} dt.$$

It is not difficult to prove that there exists a constant $C_{1,1}$ such that

$$\int_0^{2^{-b}} \sqrt{-\log t} dt \leq C_{1,1} 2^{-b} \sqrt{b}.$$

Combine the three inequalities above we can conclude

$$\begin{aligned} & \frac{x_s \sigma_0}{2^J \sigma_0} \int_{-x_s}^1 d\theta \sum_{z \in \{0,1\}^b} \sum_{j=0}^{J-1} 2^{J-j-1} \sum_{r=1}^{2^{j+1}(r-1)+2^j} \left| \frac{a_k(\theta, z)}{x_s \sigma_0} - \sum_{k=2^{j+1}(r-1)+2^{j+1}}^{2^{j+1}r} \frac{a_k(\theta, z)}{x_s \sigma_0} \right| \\ & \leq C_3 C_{1,1} x_s \int_{-x_s}^1 \sqrt{Jb} d\theta = C_3 C_{1,1} x_s (1 + x_s) \sqrt{Jb}. \end{aligned}$$

By a similar argument we also have

$$\begin{aligned} & \frac{x'_s \sigma_0}{2^J \sigma_0} \int_{-x'_s}^1 d\theta \sum_{z \in \{0,1\}^b} \sum_{j=0}^{J-1} 2^{J-j-1} \sum_{r=1}^{2^{j+1}(r-1)+2^j} \left| \frac{a_k(\theta, z)}{x'_s \sigma_0} - \sum_{k=2^{j+1}(r-1)+2^{j+1}}^{2^{j+1}r} \frac{a_k(\theta, z)}{x'_s \sigma_0} \right| \\ & \leq C_3 C_{1,1} x'_s (1 + x'_s) \sqrt{Jb}. \end{aligned}$$

Substitute the above two inequalities into (4.22) and note that $x'_s \leq x_s$, we conclude Lemma 9. □

CHAPTER 5

DISTRIBUTED NONPARAMETRIC FUNCTION ESTIMATION UNDER COMMUNICATION CONSTRAINTS

5.1. Introduction

Distributed statistical estimation and inference are becoming increasingly important as in many applications data can be necessarily distributed at different locations due to the size constraint or privacy and security concerns. Such a setting arises in a range of medical, financial, and business applications. With distributed data, separate statistical analyses need to be carried out at individual sites and then the results are transmitted to and aggregated at a central location in order to make the final statistical decision. For large-scale data analysis, communication costs can be expensive and become the main bottleneck in statistical practice. It is important to understand the interplay between communication constraints and statistical accuracy, as well as how to design optimal estimation and inference procedures under communication constraints.

There has been an increasing amount of recent literature on distributed estimation when the communication budget is limited. For example, Zhang et al. (2013a); Garg et al. (2014); Braverman et al. (2016); Han et al. (2018); Zhu and Lafferty (2018); Szabó and van Zanten (2018); Barnes et al. (2019b); Cai and Wei (2020c); Szabó and van Zanten (2020) considered information-theoretical limits under communication constraints for various distributed estimation problems, such as Gaussian mean estimation, linear regression and nonparametric regression. Optimality results have been established under different communication constraints. Besides theoretical analysis, progress has also been made on developing practical methodologies for distributed estimation. See, for example, Kleiner et al. (2014); Deisenroth and Ng (2015); Lee et al. (2017); Diakonikolas et al. (2017); Jordan et al. (2019); Battey et al. (2018); Fan et al. (2019). Further literature review is given in Section 5.1.4.

In this chapter we study distributed minimax and distributed adaptive nonparametric esti-

mation under communication constraints in a decision theoretical framework. In the conventional non-distributed settings, adaptation has been a central goal for nonparametric function estimation. It is well-known that adaptive estimation can be achieved for free under a range of global losses such as the integrated squared error over a wide collection of Besov classes (Donoho and Johnstone, 1995; Johnstone, 2017). Indeed, it is possible to adaptively achieve superefficiency for free (Cai, 2008). However, in the distributed settings, adaptation becomes more difficult and involved due to the additional communication constraints. A rate-optimal adaptive algorithm needs to perform well statistically while efficiently compressing the information from the local machines to the central learner. Intuitively, the difficulty arises from the fact that only limited amount of information can be transmitted and information that is critical for estimation over one function class might not be essential for estimation over another. In such a setting, it is easy to imagine that achieving adaptation over a collection of function classes requires more communication budget than what is needed for a given function class in the minimax setting.

The primary goal of this chapter is to precisely characterize the communication cost of adaptation for distributed nonparametric function estimation. We first establish the minimax rate of convergence for distributed estimation over a given Besov class, which serves as a benchmark for the cost of adaptation when the smoothness parameters are unknown. We then quantify the exact cost of adaptation and construct an optimally adaptive procedure for distributed nonparametric estimation over a range of Besov classes.

5.1.1. Distributed estimation framework

We begin by introducing a general framework for distributed estimation by giving a formal definition of transcript, distributed estimator, and independent distributed protocol. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a parametric family of distributions supported on space \mathcal{X} , where $\theta \in \Theta$ is the parameter of interest. Suppose there are m local machines and a central machine, where the local machines contain the observations and each local machine has access only to data in that machine, and the central machine produces the final estimator

of θ under the communication constraints between the local and central machines. More precisely, suppose we observe i.i.d. random samples drawn from a distribution $P_\theta \in \mathcal{P}$:

$$X_i \stackrel{\text{i.i.d.}}{\sim} P_\theta, \quad i = 1, \dots, m,$$

where the i -th local machine has access to X_i only.

On each machine, because of limited communication budget, the observation X_i on the i -th local machine needs to be processed to a uniquely decodable string Z_i by a (possibly random) function $Z_i : \mathcal{X} \rightarrow \bigcup_{b=1}^{\infty} \{0, 1\}^b$. The resulting string $Z_i = Z_i(X_i)$, which is called the **transcript** from the i -th machine, is then transmitted to the central machine. Here we denote the length of transcript Z_i as $|Z_i|_l$, which indicates the communication cost for sending this transcript. Finally, a **distributed estimator** $\hat{\theta}$ is constructed on the central machine based on the transcripts Z_1, Z_2, \dots, Z_m ,

$$\hat{\theta} = \hat{\theta}(Z_1, Z_2, \dots, Z_m).$$

The above scheme to obtain a distributed estimator $\hat{\theta}$ is called an **independent distributed protocol**. Within an independent distributed protocol, the transcripts from each local machine only depend on its local observations and no information is exchanged between the local machines. There are also other types of distributed protocols with more interactive communication schemes in the literature (Zhang et al., 2013a). In the present work we focus on independent distributed protocol. Define $L(\hat{\theta}) \triangleq \sum_{i=1}^m |Z_i|_l$ as the total communication cost for distributed estimator $\hat{\theta}$. The class of distributed protocols with total communication budgets B can be defined as

$$\mathcal{A}_T(B) = \{(\hat{\theta}, Z_1, Z_2, \dots, Z_m) : L(\hat{\theta}) \leq B\}.$$

The above classes of distributed protocol imposes uniform hard upper bounds on the length

of transcripts, that is, the (total) length of transcripts are constrained to be less than a certain value given any possible observations. It is sometimes worthwhile to consider transcripts with variable length in order to gain possible adaptation to the data. In such settings, we introduce a class of distributed protocols with the expected total communication budgets for the family \mathcal{P} :

$$\mathcal{A}_E(B, \Theta) = \{(\hat{\theta}, Z_1, Z_2, \dots, Z_m) : \sup_{\theta \in \Theta} \mathbb{E}_{P_\theta} L(\hat{\theta}) \leq B\} \quad (5.1)$$

where the expected total communication cost is uniformly bounded by B under any data generating distribution $P_\theta \in \mathcal{P}$.

As usual, the estimation accuracy of a distributed estimator $\hat{\theta}$ is measured by the mean squared error (MSE), $\mathbb{E}_{P_\theta} \|\hat{\theta} - \theta\|_2^2$, where the expectation is taken over the randomness in both the data and construction of the transcripts and estimator. As in the conventional decision theoretical framework, a quantity of particular interest in distributed learning is the minimax risk for the distributed protocols

$$\inf_{\hat{\theta} \in \mathcal{A}_E(B, \Theta)} \sup_{P_\theta \in \mathcal{P}} \mathbb{E}_{P_\theta} \|\hat{\theta} - \theta\|_2^2,$$

which characterizes the difficulty of the distributed learning problem under the expected total communication constraints $\mathcal{A}_E(B, \Theta)$. Similarly $\mathcal{A}_E(B, \Theta)$ can be replaced by other class of distributed protocols to illustrate minimax risk under other kind of communication constraints. In a rigorous decision theoretical formulation of distributed learning, the communication constraints are essential. Without the constraints, one can always output the original data from the local machines to the central machine and the problem is then reduced to the usual centralized setting.

5.1.2. Distributed estimation

We consider distributed minimax and adaptive estimation for the Gaussian sequence model and white noise model. For the white noise model, the goal is to recover the unknown

function based on the noisy observations collected on m machines, where on the i -th machine, for $1 \leq i \leq m$, one observes a Gaussian process,

$$dY_i(t) = f(t)dt + \frac{\epsilon}{\sqrt{n}}dW_i(t) \quad t \in [0, 1], i = 1, 2, \dots, m. \quad (5.2)$$

Here $\frac{\epsilon}{\sqrt{n}}$ is the noise level and $W_i(t), i = 1, 2, \dots, m$ are independent standard Wiener process. The i -th machine has access to $Y_i(t)$ only. The goal is to recover the unknown function f based on the distributed observed processes $Y_1(t), Y_2(t), \dots, Y_m(t)$.

In the conventional centralized setting, wavelet methods (Donoho and Johnstone, 1994; Hall et al., 1999; Cai, 1999) have been shown to be a powerful tool for nonparametric function estimation as it decomposes a function into a structured wavelet series and a nonparametric estimation problem is then transformed into a Gaussian sequence estimation problem. Motivated by the equivalence between the white noise model and Gaussian sequence model, we begin by focusing on the following distributed Gaussian sequence estimation problem. Suppose there are m machines, on i -th machine we have i.i.d Gaussian observations

$$X_{i,jk} = \theta_{jk} + \sigma z_{i,jk}, \quad j = 0, 1, 2, \dots; k = 1, 2, \dots, 2^j \quad (5.3)$$

where $z_{i,jk} \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$ for $i = 1, 2, \dots, m; j = 0, 1, \dots; k = 1, 2, \dots, 2^j$, the noise level σ known. The i -th machine can only access to $X_i \triangleq (X_{i,jk})_{j \geq 0, k=1, 2, \dots, 2^j}$ only. The goal is to estimate $\theta \triangleq (\theta_{i,jk})_{j \geq 0, k=1, 2, \dots, 2^j}$ under the mean-squared error

$$R(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_2^2 = \sum_{j=0}^{\infty} \sum_{k=1}^{2^j} (\hat{\theta}_{jk} - \theta_{jk})^2.$$

We consider estimation over a collection of Besov classes $\mathcal{B}_{p,q}^\alpha(M)$ with $\alpha, p, q, M > 0$, where $\mathcal{B}_{p,q}^\alpha(M)$ is defined as the set of sequences θ satisfying $|\theta|_{b_{p,q}^\alpha} \leq M$ with the Besov sequence

seminorm $|\theta|_{b_{p,q}^\alpha}$ given by

$$|\theta|_{b_{p,q}^\alpha} \triangleq \left(\sum_{j=0}^{\infty} \left(2^{js} \left(\sum_{k=1}^{2^j} |\theta_{jk}|^p \right)^{1/p} \right)^q \right)^{1/q}. \quad (5.4)$$

Here $s = \alpha + 1/2 - 1/p > 0$ and $1 \leq p, q \leq \infty$, with the obvious replacement of the corresponding ℓ_p or ℓ_q norms to ℓ_∞ norms when $p, q = \infty$. The Besov sequence norm $|\theta|_{b_{p,q}^\alpha}$ is equivalent to the Besov function norm on the original function f ; see, for example, Meyer (1992). Therefore, the distributed Gaussian sequence model (5.3) is equivalent to the white noise model (5.2). In the classical centralized setting, the Gaussian sequence model is also known to be a good proxy to study estimation of a function under the nonparametric regression model.

5.1.3. Main contributions

For estimation under the Gaussian sequence model (5.3) with communication constraints, a distributed estimation procedure, called seq-MODGAME, is proposed, and its rate of convergence under the communication constraints is derived. A matching lower bound is established to show that the seq-MODGAME procedure is optimal. The upper and lower bounds together yield the sharp optimal rate of convergence for estimation over a Besov class $\mathcal{B}_{p,q}^\alpha(M)$:

$$\mathcal{R}_E(B, \mathcal{B}_{p,q}^\alpha(M)) \triangleq \inf_{\hat{\theta} \in \mathcal{A}_E(B, \mathcal{B}_{p,q}^\alpha(M))} \sup_{\theta \in \mathcal{B}_{p,q}^\alpha(M)} \|\hat{\theta} - \theta\|_2^2.$$

where $\mathcal{A}_E(B, \mathcal{B}_{p,q}^\alpha(M))$ is the set of distributed protocols under the expected total communication constraints defined in (5.1). The same optimal rate holds for the white noise model. To the best of our knowledge, this is the first exact minimax rate of convergence for the distributed nonparametric function estimation. In comparison, the existing results have at least a logarithmic gap in the upper and lower bounds and are for more specialized parameter spaces such as a Hölder or Sobolev class.

We then quantify the exact communication cost for adaptation and construct an optimally

adaptive procedure for distributed estimation over a range of Besov classes. Our analysis shows interesting phenomena. In the classical non-distributed setting, it is well known that adaptation can be achieved for free in terms of global risk measures such as the mean integrated squared error over a wide collection of Besov classes. See, for example, Donoho and Johnstone (1995); Johnstone (2017). However, in the distributed setting, our results show that there are unavoidable additional communication costs for any adaptive procedure over a collection of Besov classes. Specifically, the results provide a sharp characterization for the communication costs for adaptation, where it is shown that $O(m^3)$ total additional bits are necessary and sufficient to achieve the adaptation over a wide collection of Besov classes. In addition, a local thresholding procedure is constructed and is shown to be the most communication-efficient among all adaptive distributed estimators. Our newly proposed local thresholding procedure requires no prior knowledge on the range of the smoothness parameters, and is able to automatically achieve statistical adaptation over a wide collection of Besov classes $\mathcal{B}_{p,q}^\alpha(M)$ with $p \geq 2$ at the guaranteed minimum communication cost. The analysis on adaptive estimation makes significant improvement over existing results. The new technical tools used to obtain the exact characterization for the cost of adaptation can be of independent interest.

5.1.4. Related literature

Distributed nonparametric function estimation has been investigated in the recent literature. Zhu and Lafferty (2018) studied distributed minimax rate of convergence for the white noise model over the Sobolev classes with a logarithmic gap between the upper and lower bounds. Szabó and van Zanten (2018) derived distributed minimax rate for nonparametric regression under the integrated squared error and supnorm error losses over the Hölder classes and Sobolev classes, also with a logarithmic gap between the upper and lower bounds. The paper also showed that adaptation is possible within the range $\alpha \in [\alpha_{\min}, \alpha_{\max})$ where α_{\min} depends on the given communication budget.

Szabó and van Zanten (2020) considered a two-point adaptation problem for distributed nonparametric estimation and showed that two-point adaptation is impossible when the smoothness indices of the two function classes are both larger than a certain threshold. It also proposed an adaptive distributed protocol that achieves statistical adaptation over a range of Sobolev classes with the smoothness indices below a certain threshold, while at the same time transmitting the minimal number of bits, up to a logarithmic gap. Szabó and van Zanten (2020) provided a clear solution when two-point adaptation can be achieved without additional communication cost. However, it is not clear whether adaptation is possible with additional communication budgets under the same settings. In comparison, we provide a more general lower bound for the communication cost for adaptive distributed estimators over a collection of Besov classes and construct an estimator that is adaptive over a wider range of parameter spaces at the guaranteed minimum communication cost.

5.1.5. Organization of the chapter

We finish this section with notation, definitions, and some assumptions that will be used in the rest of the chapter. Section 5.2 establishes the optimal rate of convergence for distributed Gaussian sequence estimation and Section 5.3 characterizes the communication cost of adaptation and introduces adaptive distributed procedures. The numerical performance of the proposed distributed estimators is investigated in Section 5.4 and further research directions are discussed in Section 5.5. For reasons of space, we only prove lower bounds for communication cost of adaptive estimators in Section 5.6 and defer the proofs of other main results and the technical lemmas to the supplementary material Cai and Wei (2020b).

5.1.6. Notation, definitions, and assumptions

For simplicity, in later sections we denote $n_j = 2^j$ be the number of coefficients at the j -th resolution level. For any positive integers n, N , let $[n] \triangleq \{1, 2, \dots, n\}$ and $n \bmod N$ be the remainder of n divided by N . For any $a \in \mathbb{R}$, let $\lfloor a \rfloor$ denote the floor function (the largest integer not larger than a). Unless otherwise stated, we shorthand $\log a$ as the base 2 logarithmic of a . For any $a, b \in \mathbb{R}$, let $a \wedge b \triangleq \min\{a, b\}$ and $a \vee b \triangleq \max\{a, b\}$. We

use $a = O(b)$ or equivalently $b = \Omega(a)$ to denote there exist a constant $C > 0$ such that $a \leq Cb$, and we use $a \asymp b$ to denote $a = O(b)$ while $b = O(a)$. For any vector a , denote by $\|a\| \triangleq \sqrt{\sum_k (a^{(k)})^2}$ its l_2 norm. For any finite set S , let $\text{card}(S)$ denote the cardinality of S . Define the density of a Gaussian distribution with mean 0 and standard deviation σ as

$$\phi_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}.$$

Throughout the chapter, we shall assume $s = \alpha + 1/2 - 1/p > 0$. This condition is necessary for the estimation problem to be well-formulated. When $s \leq 0$, the closure of the Besov ball $\mathcal{B}_{p,q}^\alpha(M)$ is not compact and the compactness of the closure of the parameter space is a necessary condition for consistent estimation under the homoskedastic Gaussian sequence model. See Ibragimov and Khasminskii (1997) and Johnstone (2017, Theorem 5.7). Moreover, we assume $M \geq \sigma$. Otherwise estimation over the Besov ball $\mathcal{B}_{p,q}^\alpha(M)$ is trivial as the simple estimator $\hat{\theta} = 0$ is optimal.

5.2. Minimax Optimal Rate of Convergence

In this section we study the minimax rate of convergence for estimating the mean of a Gaussian sequence $\theta \in \mathcal{B}_{p,q}^\alpha(M)$ under the expected total communication constraint:

$$\inf_{\hat{\theta} \in \mathcal{A}_E(B, \mathcal{B}_{p,q}^\alpha(M))} \sup_{\theta \in \mathcal{B}_{p,q}^\alpha(M)} \mathbb{E} \|\hat{\theta} - \theta\|^2 \tag{5.5}$$

where we assume the parameters α, p, q, M are known in an oracle setting.

If there is no communication constraint, or equivalently we are in a centralized setting, Donoho and Johnstone (1998) pointed out the minimax rate of convergence over Besov classes is

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathcal{B}_{p,q}^\alpha(M)} \mathbb{E} \|\hat{\theta} - \theta\|^2 \asymp M^{\frac{2}{2\alpha+1}} \left(\frac{\sigma^2}{m} \right)^{\frac{2\alpha}{2\alpha+1}}.$$

However, when the communication constraints take effect, there will be a loss of statistical accuracy thus the optimal rate of convergence (5.5) will further depend on the expected

total communication cost B .

We first introduce a distributed estimation procedure satisfying the communication-constraint and provide an upper bound for its statistical performance. A matching lower bound on its minimax risk is then established. The upper and lower bounds together unveil a sharp minimax rate of convergence and the optimality of the proposed estimator.

5.2.1. Optimal procedure

We begin with the construction of an estimation procedure under the communication constraints and provide a theoretical analysis of the proposed procedure. The construction of the following procedure, called *seq-MODGAME*, is inspired by the MODGAME procedure proposed in Cai and Wei (2020c) for distributed Gaussian mean estimation. However, unlike the simple Gaussian mean estimation problem considered in Cai and Wei (2020c), the magnitude of each coordinate of θ is not known a priori because within Besov space $\mathcal{B}_{p,q}^\alpha(M)$, the constraint on the Besov norm (5.4) is imposed on the whole vector, but not individual entries. Therefore, to estimate a mean vector $\theta \in \mathcal{B}_{p,q}^\alpha(M)$ under Gaussian sequence model (5.3), one needs a more sophisticated quantization strategy than the MODGAME procedure proposed in Cai and Wei (2020c).

We first define several useful functions and quantities. Define localization encoding function $g : \mathbb{Z} \rightarrow \bigcup_{k=1}^{\infty} \{0, 1\}^k$ by the following rule:

- $g(0) = \text{“0”}$.
- When x is a positive integer, let k be the length of its binary representation, and define $g(x)$ to be a string starting with “1”, followed by k zeros and then followed by the binary representation of x . For example, $g(1) = \text{“101”}$ and $g(8) = \text{“100001000”}$.
- When x is a negative integer, let k be the length of the binary representation of $-x$, and define $g(x)$ to be a string starting with “11”, followed by $k - 1$ zeros and then followed by the binary representation of $-x$. For example, $g(-1) = \text{“111”}$ and $g(-8) = \text{“110001000”}$.

The function $g(x)$, as an encoding mechanism, has two main properties. First, it is a prefix code, thus uniquely decodable (Blahut and Blahut, 1987). We denote g^{-1} as its inverse function (decoding function). Second, the length of $g(x)$ is guaranteed to be no larger than $2\log(|x|+1)+3$, which means that its length is adaptive to the magnitude of x . We will see that $g(x)$ plays an important role in the construction of the transcripts with variable length under the communication constraints.

As in the conventional centralized setting, we estimate the coordinates of the vector $\theta = (\theta_{j,k}) \in \mathcal{B}_{p,q}^\alpha(M)$ from its noisy observation up to a certain resolution level j_{\max} and truncate all the components above j_{\max} to zero. Note that when the communication budget is insufficient, the estimation accuracy in the distributed setting is not as good as in the centralized setting. So we first decide the maximal resolution level j_{\max} , and precision parameter δ according to communication budget B and other model parameters. At those resolution levels lower than j_{\max} , we estimate each entry in an optimal way so that the stochastic error is roughly $O(\delta)$. At those higher resolution levels, we just truncate all entries to zero. The advanced communication strategy used in the procedure is the key to the optimality results.

We are now ready to introduce the seq-MODGAME procedure in detail. It is divided into two cases: $B < \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}}$ and $B \geq \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}}$.

Case 1: $B < \left(\frac{\Lambda_0 M}{\sigma}\right)^{\frac{2}{2\alpha+1}}$.

Let δ be a precision parameter calculated by

$$\delta \triangleq \Lambda_0 M B^{-(\alpha+1/2)}$$

where $\Lambda_0 > 0$ is a large tuning parameter. Let j_{\max} be the maximal resolution level, defined as

$$j_{\max} \triangleq \max \left\{ j : M \cdot 2^{-j(\alpha+1/2)} \geq \delta \right\}.$$

In this case, only one local machine is needed to sent transcripts to the central machine.

First step: Generate the transcripts on the first local machine. On the first local machine (who can access to data X_1), the output transcript Z_1 is the collection of the "crude localization" strings $Z_{1,jk}$, $0 \leq j \leq j_{\max}, k \in [n_j]$ where $Z_{1,jk}$ is defined as

$$Z_{1,jk} = g(\lfloor X_{1,jk}/\delta \rfloor).$$

Second step: Generate the distributed estimator $\hat{\theta}^O$ on the central machine. The central machine can receive $Z_{1,jk}$, $0 \leq j \leq j_{\max}, k \in [n_j]$ from the first local machine. The final estimate $\hat{\theta}^O$ is given by

$$\hat{\theta}_{jk}^O = g^{-1}(Z_{1,jk}) \cdot \delta \text{ if } 0 \leq j \leq j_{\max}, k \in [n_j]$$

$$\hat{\theta}_{jk}^O = 0 \text{ if } j > j_{\max}, k \in [n_j]$$

Case 2: $B \geq \left(\frac{\Lambda_0 M}{\sigma}\right)^{\frac{2}{2\alpha+1}}$

Let u be a parameter and δ be the precision parameter. They are calculated by

$$u \triangleq \left((\Lambda_0 M / \sigma)^{-\frac{1}{\alpha+1}} B^{\frac{2\alpha+1}{2\alpha+2}} \right) \wedge m, \quad \delta = \sigma / \sqrt{u},$$

and let j_{\max} be the maximal resolution level, defined as

$$j_{\max} \triangleq \max \left\{ j : M \cdot 2^{-j(\alpha+1/2)} \geq \delta \right\}.$$

In this case, with the help of communication strategy introduced in Cai and Wei (2020c), each entry of θ at lower resolution levels can be estimated in the most communication-efficient way so that their estimation errors is roughly of order δ .

First step: Generate the transcripts on the local machines.

1. On the first machine (which has access to data X_1), the output transcript Z_1 is the collection of the “crude localization” strings $Z_{1,jk}$, $0 \leq j \leq j_{\max}, k \in [n_j]$ where $Z_{1,jk}$ is defined as

$$Z_{1,jk} = g(\lfloor X_{1,jk}/\sigma \rfloor);$$

2. On the i -th machine where $2 \leq i \leq 1 + \lfloor \log^2 u \rfloor$, the output transcript Z_i is the collection of the “finer localization” strings $Z_{i,jk}$, $0 \leq j \leq j_{\max}, k \in [n_j]$ where $Z_{i,jk}$ is defined as

$$Z_{i,jk} = g(\lfloor X_{i,jk}/\sigma \rfloor \bmod \lfloor \log u \rfloor);$$

3. On the i -th machine where $2 + \lfloor \log^2 u \rfloor \leq i \leq u$ the output transcript Z_i is the collection of the ‘refinement’ strings $Z_{i,jk}$, $0 \leq j \leq j_{\max}, k \in [n_j]$ where $Z_{i,jk}$ is defined as

$$Z_{i,jk} = \lfloor X_{i,jk}/\sigma \rfloor \bmod 8.$$

4. On the i -th machine where $u < i \leq m$, the local machine does not output anything.

Second step: Generate the distributed estimator $\hat{\theta}$ on the central machine. The central machine receives the transcripts Z_1, Z_2, \dots, Z_u from the local machines. Note that the code words in Z_1, Z_2, \dots, Z_u are all uniquely decodable, thus those transcripts can be decomposed into short strings $Z_{i,jk}$ for $i \in [u], j \in J_i, k \in [n_j]$.

The final estimator $\hat{\theta}^O$ is constructed as follows.

- For each $0 \leq j \leq j_{\max}, k \in [n_j]$:
 1. Because $g(x)$ is uniquely decodable, from $Z_{1,jk} = g(\lfloor X_{1,jk}/\sigma \rfloor)$ one can recover the value of $\lfloor X_{1,jk}/\sigma \rfloor$. Let I_{jk}^a be an left-closed-right-open interval of length u

defined as

$$I_{jk}^a \triangleq \begin{cases} \left[\lfloor X_{i,jk}/\sigma \rfloor - \frac{\lfloor \log u \rfloor - 1}{2}, \lfloor X_{i,jk}/\sigma \rfloor + \frac{\lfloor \log u \rfloor + 1}{2} \right) & \text{if } \lfloor \log u \rfloor \text{ is an odd number} \\ \left[\lfloor X_{i,jk}/\sigma \rfloor - \frac{\lfloor \log u \rfloor}{2}, \lfloor X_{i,jk}/\sigma \rfloor + \frac{\lfloor \log u \rfloor}{2} \right) & \text{if } \lfloor \log u \rfloor \text{ is an even number} \end{cases}$$

2. Denote $z_{ik}^b \triangleq \operatorname{argmax}_{z'} \sum_{i=2}^{\lfloor \log^2 u \rfloor + 1} \mathbb{I}_{\{Z_{i,jk}=z'\}}$ be the mode statistic among the $Z_{i,jk}$, $2 \leq i \leq \lfloor \log^2 u \rfloor + 1$. Note that the length of I_{jk}^a is $\lfloor \log u \rfloor$, so there will be exactly one integer $x_{jk}^b \in I_{jk}^a$ that satisfies

$$x_{jk}^b \bmod \lfloor \log u \rfloor = g^{-1}(z_{ik}^b).$$

Let I_{jk}^b be an interval of length 3 defined by

$$I_{jk}^b \triangleq [x_{jk}^b - 1, x_{jk}^b + 1].$$

3. Let p^h be the proportion of those refinement strings whose value is equal to $g(x_{jk}^b - 2 \bmod 8)$:

$$p^h \triangleq \frac{1}{u - 1 - \lfloor \log^2 u \rfloor} \sum_{i=\lfloor \log^2 u \rfloor + 2}^u \mathbb{I}_{\{Z_{i,jk}=g(x_{jk}^b - 2 \bmod 8)\}}$$

Define a function

$$h_{jk}(y) \triangleq \sum_{l=-\infty}^{\infty} \int_{x_{jk}^b - 2 + 8l}^{x_{jk}^b - 1 + 8l} \phi_1(x - y) dx$$

It is easy to see that $h_{jk}(y)$ is a strictly decreasing function on I_{jk}^b . Let $h_{jk}^{-1}(y)$ be the inverse function of $h_{jk}(y)$ which maps $h_{jk}(I_{jk}^b)$ to I_{jk}^b . The estimate is

calculated by

$$\hat{\theta}_{jk}^O = \begin{cases} (x_{jk}^b + 1)\sigma & \text{if } p^h \leq h_{jk}(x_{jk}^b + 1) \\ h_{jk}^{-1}(p^h)\sigma & \text{if } h_{jk}(x_{jk}^b + 1) < p^h < h_{jk}(x_{jk}^b - 1) \\ (x_{jk}^b - 1)\sigma & \text{if } p^h \geq h_{jk}(x_{jk}^b - 1) \end{cases}$$

- For each $j \geq j_{\max} + 1, k \in [n_j]$, set

$$\hat{\theta}_{jk}^O = 0.$$

The following theorem provides the theoretical guarantee for the communication cost of $\hat{\theta}^O$, as well as an upper bound for its statistical performance.

Theorem 19. *If Λ_0 is set to be a sufficient large constant such that $\Lambda_0 > (24\alpha + 64)^{\alpha+1/2}$, then the estimator $\hat{\theta}^O \in \mathcal{A}_E(B, \mathcal{B}_{p,q}^\alpha(M))$ and there exists a constant $C > 0$ such that*

$$\sup_{\theta \in \mathcal{B}_{p,q}^\alpha(M)} \mathbb{E} \|\hat{\theta}^O - \theta\|^2 \leq C \cdot \begin{cases} M^2 B^{-2\alpha} & \text{if } B < \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} \\ M^{\frac{2}{\alpha+1}} \left(\frac{\sigma^2}{B}\right)^{\frac{\alpha}{\alpha+1}} & \text{if } \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} \leq B < \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}} \\ M^{\frac{2}{2\alpha+1}} \left(\frac{\sigma^2}{m}\right)^{\frac{2\alpha}{2\alpha+1}} & \text{if } B \geq \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}} \end{cases} \quad (5.6)$$

for all $2 \leq p \leq \infty, 0 < q \leq \infty, \alpha > 0, M > 0$

Remark 9. The proposed distributed estimator $\hat{\theta}^O$ satisfies expected total communication constraint, which is weaker than other types of constraint considered in the literature. The reason we work on this type of communication constraint is to illustrate the main idea and omit unnecessary complication when presenting the estimator. With suitable modification, the estimator can be made to satisfy other kinds of communication constraint, say, a fixed/hard total communication constraint or an equally assigned communication constraint on each single local machines.

For example, the following proposition provides a quick look on how $\hat{\theta}^O$ satisfies fixed/hard total communication constraint with high probability.

Proposition 5.2.1. *With probability at least $1 - \exp(-B/18)$, we have*

$$L(\hat{\theta}^O) < 2B.$$

That is, the proposed estimator $\hat{\theta}^O$ satisfies the total communication constraint $2B$ with high probability. Note that the additional factor on the communication constraint doesn't affect the rate of convergence given in Theorem 19, therefore the estimator is still rate-optimal.

5.2.2. Lower bound analysis

Section 5.2.1 gives a detailed construction of the seq-MODGAME procedure for distributed Gaussian sequence estimation and provides a theoretical guarantee for the estimator in Theorem 19. In this section we shall show that the estimator $\hat{\theta}^O$ is indeed rate optimal among all estimators satisfying the total communication constraints by proving that the upper bound in Equation (5.6) cannot be improved. The following theorem gives a lower bound on the minimax risk under the expected total communication constraints.

Theorem 20. *There exists a constant $c > 0$ such that*

$$\mathcal{R}_E(B, \mathcal{B}_{p,q}^\alpha(M)) \geq c \cdot \begin{cases} M^2 B^{-2\alpha} & \text{if } B < \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} \\ M^{\frac{2}{\alpha+1}} \left(\frac{\sigma^2}{B}\right)^{\frac{\alpha}{\alpha+1}} & \text{if } \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} \leq B < \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}} \\ M^{\frac{2}{2\alpha+1}} \left(\frac{\sigma^2}{m}\right)^{\frac{2\alpha}{2\alpha+1}} & \text{if } B \geq \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}} \end{cases} \quad (5.7)$$

for all $0 < p \leq \infty, 0 < q \leq \infty, \alpha > 0, M > 0$

The lower bound given in Theorem 20 is proved by constructing simultaneous tests $\theta_{jk} = 0$ vs $\theta_{jk} = \delta$ for all $j \leq J, k = 1, 2, \dots, 2^j$, with pre-specified choices of δ and J . Then by strong data processing inequalities, we can prove that at least a proportion of entries cannot be accurately estimated. The detailed proof is deferred to the supplementary material

Cai and Wei (2020a).

Theorems 19 and 20 together establish the minimax rate for distributed Gaussian sequence estimation:

$$\mathcal{R}_E(B, \mathcal{B}_{p,q}^\alpha(M)) \asymp \begin{cases} M^2 B^{-2\alpha} & \text{if } B < \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} \\ M^{\frac{2}{\alpha+1}} \left(\frac{\sigma^2}{B}\right)^{\frac{\alpha}{\alpha+1}} & \text{if } \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} \leq B < \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}} \\ M^{\frac{2}{2\alpha+1}} \left(\frac{\sigma^2}{m}\right)^{\frac{2\alpha}{2\alpha+1}} & \text{if } B \geq \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}} \end{cases} \quad (5.8)$$

where $2 \leq p \leq \infty, q \leq \infty, \alpha > 0, M > 0$. The results also show that the distributed estimator $\hat{\theta}^O$ proposed in Section 5.2.1 is rate optimal under the total communication constraints.

The theorem also suggests that in order to achieve the centralized rate of convergence, which is of order $M^{\frac{2}{2\alpha+1}} \left(\frac{\sigma^2}{m}\right)^{\frac{2\alpha}{2\alpha+1}}$, a communication cost of order $\left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}}$ is sufficient and necessary.

Remark 10. Similar as the optimal rate of convergence for distributed univariate Gaussian mean estimation Cai and Wei (2020c), the minimax rate (5.8) can be divided into three phases: localization ($B < \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}}$), refinement ($\left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} \leq B < \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}}$), and optimal-rate ($B \geq \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}}$). The minimax rate decreases quickly in the localization phase, when the communication constraints are extremely severe; then it decreases slower in the refinement phase, when there are more communication budgets; finally the minimax rate coincides with the centralized optimal rate (Donoho and Johnstone, 1998) and stays the same, when there are sufficient communication budgets. The value for each additional bit decreases as more bits are allowed.

Remark 11. As mentioned in the introduction, distributed minimax estimation was considered in Zhu and Lafferty (2018) for the Hölder classes and in Szabó and van Zanten (2020) for the Sobolev classes. These two types of function classes are special cases of the Besov classes with the Hölder class being $B_{\infty,\infty}^\alpha$ and Sobolev class being $B_{2,\infty}^\alpha$. Furthermore, in both Zhu and Lafferty (2018) and Szabó and van Zanten (2020), the existing upper bound

and lower bound are sub-optimal (with a poly-logarithmic gap to the optimal rate of convergence (5.8)). In contrast, the minimax rate given in (5.8) is sharp for a wide collection of Besov spaces.

5.3. Adaptive Gaussian sequence estimation

The minimax rate of convergence established in Section 5.2 provides an important benchmark for the evaluation of the performance of distributed Gaussian sequence estimators. However, the estimator $\hat{\theta}^{\mathcal{O}}$, in spite of its statistical optimality and communication efficiency, requires explicit knowledge of the smoothness parameters which are typically unknown in practice. The optimal seq-MODGAME procedure proposed in Section 5.2 highly depends on the prior knowledge on the parameter space $\mathcal{B}_{p,q}^{\alpha}(M)$ so that local machines efficiently transmit useful information when the communication budget is limited. It is evident from the construction and theoretical analysis that the estimator $\hat{\theta}^{\mathcal{O}}$ designed for one Besov class $B_{p,q}^{\alpha}(M)$ with a given smoothness parameter α would perform poorly over another Besov class $B_{p,q}^{\alpha'}(M)$ with a different smoothness parameter α' . Therefore, the estimator $\hat{\theta}^{\mathcal{O}}$ is not practical for real applications because the model parameters are typically unavailable.

This naturally leads to the important question of adaptive distributed estimation: Is it possible to construct a single distributed estimator, satisfying the communication constraints and not depending on the smoothness parameters, that achieves the optimal rate of convergence simultaneously over a wide collection of Besov classes $B_{p,q}^{\alpha}(M)$? In the conventional centralized setting, the answer is affirmative. That is, one can achieve adaptation for free for estimating a Gaussian sequence over a collection of Besov classes $B_{p,q}^{\alpha}(M)$ under the mean squared error.

Adaptive estimation in the centralized setting has been a major goal in the classical non-parametric function estimation literature. In particular, wavelet thresholding is well known to be a powerful technique to achieve adaptivity. For example, Donoho and Johnstone (1995); Abramovich et al. (2006) proposed adaptive term-by-term thresholding methods and Cai (1999); Cai and Zhou (2009) introduced data-driven block thresholding procedures to

achieve optimal rate of convergence over a wide collection of Besov spaces. In contrast, little has been understood on how to construct a communication-efficient adaptive estimator for most distributed estimation problems, including but not limited to distributed Gaussian sequence estimation. It is interesting and practically important to investigate the interplay between communication constraints and adaptation for distributed estimation problems.

In this section we address the following questions: how to construct a data-driven distributed estimation procedure that can achieve the centralized optimal rate with communication cost as small as possible? Can adaptation be achieved for free? If not, what is the cost of adaptation?

It was shown in Section 5.2 that, for distributed estimation over the Besov class $\mathcal{B}_{p,q}^\alpha(M)$, one needs at least $\Omega\left(\left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}}\right)$ total bits to communicate in order to achieve the centralized optimal rate $O\left(M^{\frac{2}{2\alpha+1}} \left(\frac{\sigma^2}{m}\right)^{\frac{2\alpha}{2\alpha+1}}\right)$. It is tempting to consider the question: Is there a distributed estimator with a total communication budget $O\left(\left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}}\right)$ that adaptively achieves the centralized optimal rate over a wide collection of Besov classes $\theta \in \mathcal{B}_{p,q}^\alpha(M)$?

To rigorously formulate this problem, let $\tilde{S} \subset (0, \infty) \times (0, \infty) \times (0, \infty] \times (0, \infty]$ be a collection of Besov parameter combinations (α, M, p, q) , and $\tilde{C}(\cdot)$ is a function $(0, \infty) \rightarrow (0, \infty)$. Let $\mathcal{G}(\tilde{S}, \tilde{C})$ be the set of adaptive distributed estimators that achieve the centralized optimal rate of convergence over Besov classes $\mathcal{B}_{p,q}^\alpha(M)$ for all $(\alpha, M, p, q) \in \tilde{S}$. To be precise, $\mathcal{G}(\tilde{S}, \tilde{C})$ is the collection of distributed estimators $\hat{\theta}$ who satisfy the following property: for any $(\alpha, M, p, q) \in \tilde{S}$,

$$\sup_{\theta \in \mathcal{B}_{p,q}^\alpha(M)} \mathbb{E} \|\hat{\theta} - \theta\|^2 \leq \tilde{C}(\alpha) M^{\frac{2}{2\alpha+1}} \left(\frac{\sigma^2}{m}\right)^{\frac{2\alpha}{2\alpha+1}}$$

Estimators in $\mathcal{G}(\tilde{S}, \tilde{C})$ are called statistically-optimal adaptive estimators over parameter set \tilde{S} . We are interested in the minimum expected communication cost among all statistically-

optimal adaptive estimators:

$$Q(\tilde{S}, \tilde{C}, \mathcal{B}_{p,q}^\alpha(M)) \triangleq \inf_{\hat{\theta} \in \mathcal{G}(\tilde{S}, \tilde{C})} \sup_{\theta \in \mathcal{B}_{p,q}^\alpha(M)} \mathbb{E}_\theta L(\hat{\theta})$$

The above quantity, which is called *the minimax communication cost for statistically-optimal adaptive estimators*, serves as a benchmark for the communication-efficiency of estimators in $\mathcal{G}(\tilde{S}, \tilde{C})$. For any statistically-optimal adaptive estimators, its expected communication cost is at least $Q(\tilde{S}, \tilde{C}, \mathcal{B}_{p,q}^\alpha(M))$ when estimating a function in $\mathcal{B}_{p,q}^\alpha(M)$. The analysis of the minimax communication cost $Q(\tilde{S}, \tilde{C}, \mathcal{B}_{p,q}^\alpha(M))$ is divided into two steps: upper bound and lower bound. We first propose in Section 5.3.1 an adaptive distributed estimator $\hat{\theta}^A$ which can achieve the centralized optimal rate of convergence when $2 \leq p \leq \infty$, and provide a upper bound on the expected communication cost. We then derive in Section 5.3.2 a lower bound for the rate of convergence of $Q(\tilde{S}_0, \tilde{C}, \mathcal{B}_{p,q}^\alpha(M))$ where \tilde{S}_0 is collection of all Besov class parameters with $p \geq 2$. The lower bound provides a fundamental limit on the communication cost for a statistically-optimal adaptive estimator, while it matches the upper bound for $\hat{\theta}^A$ on the expected communication cost. Therefore, the proposed distributed estimator $\hat{\theta}^A$ is shown to be the most communication-efficient one among all statistically-optimal adaptive estimators over a wide range of Besov classes.

5.3.1. Optimal adaptive procedure by local thresholding

In order to establish an upper bound on $Q(\tilde{S}, \tilde{C}, \mathcal{B}_{p,q}^\alpha(M))$, we first construct a statistically-optimal adaptive distributed procedure which simultaneously achieves the optimal rate of convergence over a wide collection of Besov classes, while the rate of convergence for its expected communication cost matches that of the minimax lower bound given in Section 5.3.2.

Wavelet thresholding methods have been shown to be a powerful tool for adaptive nonparametric function estimation problems in the conventional centralized settings. Estimators derived from data-driven thresholding rules can automatically adapt to a wide collection

of Besov spaces. See Donoho and Johnstone (1995); Abramovich et al. (2006); Cai (1999); Cai and Zhou (2009); Johnstone (2017) and the references therein. However, in the distributed settings, due to the communication constraints, it is typically impossible to estimate individual coordinates accurately by thresholding them all together on the central machine. In such a setting, it is unclear how to optimally threshold on each local machine and efficiently transmit the information to the central machine with minimal communication cost such that a final aggregated estimator is statistically-optimal adaptive. Indeed, it is unclear if this goal is even achievable.

Fortunately, the answer is affirmative. The following “local thresholding” procedure is proposed for adaptive distributed Gaussian sequence estimation. We should emphasize that here “local thresholding” referred to the fact that the thresholding step is carried out on individual local machines, not on the central machine. The meaning is different from that in the conventional wavelet estimation literature in the centralized setting. The general strategy can be summarized as follows. On each local machine, we first select “significant resolution levels” by certain thresholding rule. Only information about the significant resolution levels is transmitted to the central machine, where an estimation subroutine called “ada-MODGAME” is applied to generate good estimates for individual coordinates based on the transcripts collected from the local machines. These estimates will be further processed to yield a final estimate $\hat{\theta}^A$.

Now we are ready to introduce the local thresholding procedure in detail. Let $g : \mathbb{Z} \rightarrow \bigcup_{k=1}^{\infty} \{0, 1\}^k$ denote the localization encoding function defined in Section 5.2.1. The estimation procedure is divided into two steps, with the subroutine *ada-MODGAME* in the second step of the procedure.

First step: Generate the transcripts on the local machines by thresholding. For $1 \leq i \leq m$, on the i -th machine:

1. Define the set of "significant resolution levels" on the i -th machine by

$$J_i = \{0, 1, 2, \dots, (\lfloor 2 \log m \rfloor)\} \cup \{j \geq \lfloor 2 \log m \rfloor + 1 : \sum_{k=1}^{n_j} X_{i,jk}^2 \geq n_j \sigma^2 (1 + \frac{\Lambda_1}{m})\},$$

where $\Lambda_1 > 0$ is a prespecified parameter. Only those coordinates at the resolution levels in the set J_i are processed as part of the transcript outputs from the i -th machine. All the resolution levels that are not in J_i are considered to be "locally thresholded", because the signal strength on those resolution levels is weak.

2. If $i = 1$, the output transcript Z_1 is the collection of the "crude localization" strings $Z_{1,jk}$, $j \in J_1, k \in [n_j]$ where $Z_{1,jk}$ is defined as

$$Z_{1,jk} = g(\lfloor X_{1,jk}/\sigma \rfloor);$$

If $2 \leq i \leq 1 + \lfloor \log^2 m \rfloor$, the output transcript Z_i is the collection of the "finer localization" strings $Z_{i,jk}$, $j \in J_i, k \in [n_j]$ where $Z_{i,jk}$ is defined as

$$Z_{i,jk} = g(\lfloor X_{i,jk}/\sigma \rfloor \bmod \lfloor \log m \rfloor);$$

If $i \geq 2 + \lfloor \log^2 m \rfloor$ the output transcript Z_i is the collection of the "refinement" strings $Z_{i,jk}$, $j \in J_i, k \in [n_j]$ where $Z_{i,jk}$ is defined as

$$Z_{i,jk} = \lfloor X_{i,jk}/\sigma \rfloor \bmod 8.$$

Second step: Generate the distributed estimator $\hat{\theta}$ on the central machine. The central machine receives the transcripts Z_1, Z_2, \dots, Z_m from the local machines. Note that the code words in Z_1, Z_2, \dots, Z_m are all uniquely decodable, thus the central machine is able to recover short strings $Z_{i,jk}$ for $i \in [m], j \in J_i, k \in [n_j]$. Also, note that the total number of short strings from the i -th machine is $\sum_{j \in J_i} 2^j$, so from the binary representation of the

total number of short strings from the i -th machine, one can recover significant resolution level J_i .

To wrap up, from those transcripts that the central machine receives

- significant resolution levels on the local machines J_1, J_2, \dots, J_m .
- short strings $Z_{i,jk}$ for $i \in [m], j \in J_i, k \in [n_j]$.

Let \hat{J} be defined as

$$\hat{J} \triangleq \left\{ j : j \in J_1; \sum_{i=2}^{1+\lfloor \log^2 m \rfloor} \mathbb{I}_{\{j \in J_i\}} \geq \frac{\lfloor \log^2 m \rfloor}{2}; \sum_{i=2+\lfloor \log^2 m \rfloor}^m \mathbb{I}_{\{j \in J_i\}} \geq \frac{m-1-\lfloor \log^2 m \rfloor}{2} \right\}$$

Intuitively, \hat{J} is the set of resolution levels that are significant on most local machines. The resolution levels within \hat{J} will be estimated whereas those not in \hat{J} will be zero out (thresholded).

The final estimator $\hat{\theta}^A$ is constructed as follows: For $j = 1, 2, \dots$,

- If $j \notin \hat{J}$, let

$$\hat{\theta}_{jk}^A = 0 \text{ for all } k \in [n_j].$$

- If $j \leq \lfloor 2 \log m \rfloor$, let $S_j = [m]$ and

$$(\hat{\theta}_{j1}^*, \hat{\theta}_{j2}^*, \dots, \hat{\theta}_{jn_j}^*) = \hat{f}_{\text{ada}}(S_j, \{Z_{i,jk} : i \in S_j, k \in [n_j]\})$$

be the output of the subroutine "ada-MODGAME". Then apply the thresholding rule to get the final estimate

$$(\hat{\theta}_{j1}^A, \hat{\theta}_{j2}^A, \dots, \hat{\theta}_{jn_j}^A) = \begin{cases} (\hat{\theta}_{j1}^*, \hat{\theta}_{j2}^*, \dots, \hat{\theta}_{jn_j}^*) & \text{if } \sum_{k=1}^{n_j} (\hat{\theta}_{jk}^*)^2 \geq \Lambda_2 \frac{n_j \sigma^2}{m} \\ (0, 0, 0, \dots, 0) & \text{otherwise} \end{cases}$$

where $\Lambda_1 > 0$ is a prespecified parameter.

- If $j \geq \lfloor 2 \log m \rfloor + 1$ and $j \in \hat{J}$, define $S_j = \{i \in [m] : j \in J_i\}$, and let

$$(\hat{\theta}_{j1}^A, \hat{\theta}_{j2}^A, \dots, \hat{\theta}_{jn_j}^A) = \hat{f}_{\text{ada}}(S_j, \{Z_{i,jk} : i \in S_j, k \in [n_j]\})$$

be the output of the subroutine “ada-MODGAME”.

Subroutine: ada-MODGAME

Input: $\sigma, m, j, n_j, S_j, \{Z_{i,jk} : i \in S_j, k \in [n_j]\}$.

For each $k \in [n_j]$, do following steps:

1. Because $g(x)$ is uniquely decodable, from $Z_{1,jk} = g(\lfloor X_{i,jk}/\sigma \rfloor)$ one can recover the value of $\lfloor X_{i,jk}/\sigma \rfloor$. Let I_{jk}^a be a left-closed-right-open interval of length m defined as

$$I_{jk}^a \triangleq \begin{cases} \left[\lfloor X_{i,jk}/\sigma \rfloor - \frac{\lfloor \log m \rfloor - 1}{2}, \lfloor X_{i,jk}/\sigma \rfloor + \frac{\lfloor \log m \rfloor + 1}{2} \right) & \text{if } \lfloor \log m \rfloor \text{ is an odd number} \\ \left[\lfloor X_{i,jk}/\sigma \rfloor - \frac{\lfloor \log m \rfloor}{2}, \lfloor X_{i,jk}/\sigma \rfloor + \frac{\lfloor \log m \rfloor}{2} \right) & \text{if } \lfloor \log m \rfloor \text{ is an even number} \end{cases}$$

2. Let $S_j^b \triangleq S_j \cap \{i : 2 \leq i \leq \lfloor \log^2 m \rfloor + 1\}$ be the set of machines that output the finer localization strings. Let $z_{ik}^b \triangleq \operatorname{argmax}_{z'} \sum_{i \in S_j^b} \mathbb{I}_{\{Z_{i,jk}=z'\}}$ be the mode statistic among $Z_{i,jk}, i \in S_j^b$. Note that the length of I_{jk}^a is $\lfloor \log m \rfloor$, so there will be exactly one integer $x_{jk}^b \in I_{jk}^a$ satisfying

$$x_{jk}^b \bmod \lfloor \log m \rfloor = g^{-1}(z_{ik}^b).$$

Let I_{jk}^b be an interval of length 3 defined by

$$I_{jk}^b \triangleq [x_{jk}^b - 1, x_{jk}^b + 1].$$

3. Let $S_j^h \triangleq S_j \cap \{i : i \geq \lfloor \log^2 m \rfloor + 2\}$ be the set of machines that output the refinement

strings. Let p^h be the proportion of those refinement strings whose value is equal to $g(x_{jk}^b - 2 \bmod 8)$:

$$p^h \triangleq \text{card}(S_j^h)^{-1} \sum_{i \in S_j^h} (\mathbb{I}_{\{Z_{i,jk} = g(x_{jk}^b - 2 \bmod 8)\}})$$

Define a function

$$h_{jk}(y) \triangleq \sum_{l=-\infty}^{\infty} \int_{x_{jk}^b - 2 + 8l}^{x_{jk}^b - 1 + 8l} \phi_1(x - y) dx.$$

It is easy to see that $h_{jk}(y)$ is a strictly decreasing function on I_{jk}^b . Let $h_{jk}^{-1}(y)$ be the inverse function of $h_{jk}(y)$ which maps $h_{jk}(I_{jk}^b)$ to I_{jk}^b . Finally the estimate can be calculated by

$$\hat{\theta}_{jk}^* = \begin{cases} (x_{jk}^b + 1)\sigma & \text{if } p^h \leq h_{jk}(x_{jk}^b + 1) \\ h_{jk}^{-1}(p^h)\sigma & \text{if } h_{jk}(x_{jk}^b + 1) < p^h < h_{jk}(x_{jk}^b - 1) \\ (x_{jk}^b - 1)\sigma & \text{if } p^h \geq h_{jk}(x_{jk}^b - 1) \end{cases} .$$

Output: $\hat{\theta}_{jk}^*$ for $k \in [n_j]$.

We have given above a detailed construction of the local thresholding estimator $\hat{\theta}^A$. The following theorem provides a theoretical guarantee for the statistical performance and communication cost of the proposed procedure over the Besov classes $\mathcal{B}_{p,q}^\alpha(M)$ with $\alpha > 0$, $M \geq \sigma$, $1 < q \leq \infty$, and $2 \leq p \leq \infty$.

Theorem 21 (Upper Bound for the Communication Cost). *If $\Lambda_1 > 10$ and Λ_2 is chosen sufficiently large, there exists a constant $C > 0$ such that, the local thresholding estimator $\hat{\theta}^A$ is adaptively rate-optimal, i.e.*

$$\sup_{\theta \in \mathcal{B}_{p,q}^\alpha(M)} \mathbb{E} \|\hat{\theta}^A - \theta\|^2 \leq CM^{\frac{2}{2\alpha+1}} \left(\frac{\sigma^2}{m} \right)^{\frac{2\alpha}{2\alpha+1}}$$

and we also have

$$\sup_{\theta \in \mathcal{B}_{p,q}^\alpha(M)} \mathbb{E}_\theta L(\hat{\theta}^A) \leq C \left(m^3 + \left(\frac{M}{\sigma} \right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}} \right)$$

for all $\alpha > 0$, $M \geq \sigma$, $1 < q \leq \infty$, and $2 \leq p \leq \infty$.

Remark 12. The proof of Theorem 21 is involved due to the fact that, after thresholding on the local machines, the conditional distribution of the observations given that their resolution level is selected into the significant set J_i is no longer Gaussian. Lemma 7 (from the supplementary material Cai and Wei (2020b)) is the key to the proof, which shows that the ada-MODGAME subroutine is robust even if the additive noise is slightly different from Gaussian distribution.

Remark 13. One of the merits of the local thresholding estimator $\hat{\theta}^A$ is its "communication-adaptivity", which means the communication cost of the estimation procedure is also adaptive to the smoothness of the underlying function. Compared to the two-point adaptive procedure proposed in the previous work Szabó and van Zanten (2020) which is able to achieve adaptation with smoothness less than certain threshold, our newly proposed local thresholding procedure requires no prior knowledge on the range of the smoothness parameters, and is able to achieve statistical adaptation over a wide collection of Besov classes. The user can apply local thresholding procedure to obtain adaptation over the Besov classes $\mathcal{B}_{p,q}^\alpha(M)$ as long as $p \geq 2$ with guaranteed minimum communication cost.

5.3.2. Lower bound analysis

In this subsection, we are going to obtain a lower bound for the minimax communication cost for statistically-optimal adaptive estimators, which is instrumental in establishing the optimal rate of convergence. Before we establish a lower bound for the minimax communication cost $Q(\tilde{\mathcal{S}}, \tilde{\mathcal{C}}, \mathcal{B}_{p,q}^\alpha(M))$, we first state the following theorem, which gives a lower bound for the communication cost when the estimator achieves statistical-optimal rate of convergence in two different Besov classes.

Theorem 22 (Lower bound for communication cost for two-point adaptation). *For any distributed estimator $\hat{\theta}$, let $\mathcal{B}_{p_1, q_1}^{\alpha_1}(M_1)$ and $\mathcal{B}_{p_2, q_2}^{\alpha_2}(M_2)$ be two different Besov classes. If there exists a constant $C > 0$ such that $M_1 \leq C\sigma m^{2\alpha_1 + \frac{1}{2}}$, and*

$$\sup_{\theta \in \mathcal{B}_{p_l, q_l}^{\alpha_l}(M_l)} \mathbb{E} \|\hat{\theta} - \theta\|^2 \leq CM_l^{\frac{2}{2\alpha_l + 1}} \left(\frac{\sigma^2}{m} \right)^{\frac{2\alpha_l}{2\alpha_l + 1}} \text{ for } l = 1, 2. \quad (5.9)$$

Then there exists a constant $c > 0$ (depending on C) such that

$$\sup_{\theta \in \mathcal{B}_{p_2, q_2}^{\alpha_2}(M_2)} \mathbb{E} L(\hat{\theta}) \geq c \left(\left(\frac{M_1}{\sigma} \right)^{\frac{2}{2\alpha_1 + 1}} m^{\frac{2\alpha_1 + 2}{2\alpha_1 + 1}} + \left(\frac{M_2}{\sigma} \right)^{\frac{2}{2\alpha_2 + 1}} m^{\frac{2\alpha_2 + 2}{2\alpha_2 + 1}} \right).$$

Remark 14. If one sets $\sigma = \sqrt{m/n}$, $M_1 = M_2 = 1$ and $\alpha_2 > \alpha_1 > \frac{\log n}{4 \log m} - \frac{1}{2}$, the above Theorem 22 recovers the result of Theorem 2.4 in Szabó and van Zanten (2020) which shows that two-point adaptation is impossible without additional communication cost when $m^{4\alpha + 2} \gg n$. Comparing with the previous result, the result given in Theorem 22 here is stronger because we prove the lower bound for the communication cost $\sup_{\theta \in \mathcal{B}_{p_2, q_2}^{\alpha_2}(M_2)} \mathbb{E} L(\hat{\theta})$ under the only assumption that $\hat{\theta}$ is adaptive. In particular, no upper bound is imposed on $\sup_{\theta \in \mathcal{B}_{p_1, q_1}^{\alpha_1}(M_1)} \mathbb{E}_\theta L(\hat{\theta})$, which is in fact necessary to obtain Theorem 2.4 in Szabó and van Zanten (2020).

The above Theorem 22 only considers two-point adaptation between two specific Besov classes. However, in real data application, we are more interested in developing estimators that are able to adapt to a wide range of parameter spaces, such as our adaptive estimator $\hat{\theta}^A$. It is necessary to extend the above Theorem 22 to a general lower bound on $Q(\tilde{S}, \tilde{C}, \mathcal{B}_{p, q}^\alpha(M))$.

We define $\tilde{S}_0 = \{(\alpha, M, p, q) : \alpha > 0, M \geq \sigma, 2 \leq p \leq \infty, 1 < q \leq \infty\}$ a wide collection of Besov class parameters. The following lower bound on $Q(\tilde{S}_0, \tilde{C}, \mathcal{B}_{p, q}^\alpha(M))$ shows a fundamental limit on the communication cost of statistically-optimal estimators over Besov classes $\mathcal{B}_{p, q}^\alpha(M)$ where $(\alpha, p, q, M) \in \tilde{S}_0$. In view of the upper bound to be given in Section 5.3.1 that is achieved by the adaptive distributed estimator $\hat{\theta}^A$, the lower bound is rate optimal.

Theorem 23 (Lower bound for the communication cost over Besov ball collection \tilde{S}_0). *For any $\tilde{C} : (0, \infty) \rightarrow (0, \infty)$ and $(\alpha, M, p, q) \in \tilde{S}_0$, there exists a constant $c > 0$ such that*

$$Q(\tilde{S}_0, \tilde{C}, \mathcal{B}_{p,q}^\alpha(M)) \geq c \left(m^3 + \left(\frac{M}{\sigma} \right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}} \right). \quad (5.10)$$

Remark 15. The lower bound in Theorem 23 shows that, if a distributed estimator adaptively achieves the optimal rate of convergence over the all Besov classes where $p \geq 2$, the minimum required expected communication cost for estimating functions in $\mathcal{B}_{p,q}^\alpha(M)$ is of order $m^3 + \left(\frac{M}{\sigma} \right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}}$. The additional communication cost, which is of order m^3 and not depending on the values of α, M, p, q and σ , is required and necessary for constructing an adaptive estimator. When $m \gtrsim \left(\frac{M}{\sigma} \right)^{\frac{2}{4\alpha+1}}$, the cost of adaptation is significant.

Remark 16. Although in Theorem 23 we provide a lower bound on $Q(\tilde{S}, \tilde{C}, \mathcal{B}_{p,q}^\alpha(M))$ where $\tilde{S} = \tilde{S}_0$, the same lower bound also holds when \tilde{S} is other sufficiently large Besov ball collections. With the help from Theorem 22, we are able to establish lower bounds for other Besov ball collection \tilde{S} .

Remark 17. The techniques used to prove Theorems 22 and 23 can be of independent interest. Roughly speaking, if the algorithm aims to perform well on both $B_{p_1,q_1}^{\alpha_1}(M_1)$ and $B_{p_2,q_2}^{\alpha_2}(M_2)$ where $\alpha_1 < \alpha_2$, since we cannot tell whether each local sample is drawn from $B_{p_1,q_1}^{\alpha_1}(M_1)$ or $B_{p_2,q_2}^{\alpha_2}(M_2)$ on the local machines, the algorithm needs to transmit more bits than non-adaptive estimation for $B_{p_2,q_2}^{\alpha_2}(M_2)$, because it also needs to estimate well in $B_{p_1,q_1}^{\alpha_1}(M_1)$. More specifically, we prove that the local machines cannot “distinguish” samples that is drawn from a null model ($\theta = \vec{0}$) or drawn from a mixture of models with θ having m^2 non-zero elements. If the observations are truly drawn from the mixture, the minimum communication cost required to achieve the statistical optimal rate of convergence is of order m^3 . Thus one can further show that the minimax communication cost is at least $\Omega(m^3)$ even if $\theta = \vec{0}$. This is a key step in the argument for establishing Theorems 22 and 23.

A similar technique was also used in Szabó and van Zanten (2020). But a finer analysis

is needed here, especially for the key Claim 3 where we first prove a *conditional strong data processing inequality* and use it to establish a stronger result without unnecessary assumptions.

Lemma 16 (Conditional strong data processing inequality). *For $t > 0$ and $k \in \mathbb{Z}_+$, let θ be a random vector uniformly distributed on the set $\{-t\sigma, t\sigma\}^k$ and let $X \sim N(\theta, \sigma^2 I_k)$. Let $D \subseteq \mathbb{R}^k$ be a k -dimensional region such that the event $X \in D$ is independent with θ and let Z be a random variable such that $\theta \rightarrow X \rightarrow Z$ forms a Markov chain. Then*

$$I(Z; \theta | X \in D) \mathbb{P}(X \in D) \leq 256t^2 (H(Z | X \in D) \mathbb{P}(X \in D) + H(\{X \in D\})),$$

where $I(\cdot; \cdot | \cdot)$, $H(\cdot)$, and $H(\cdot | \cdot)$ denote conditional mutual information, entropy, and conditional entropy respectively.

The definitions of the conditional mutual information $I(\cdot; \cdot | \cdot)$, entropy $H(\cdot)$, and conditional entropy $H(\cdot | \cdot)$ are given in Section 5.6.1. Note that the classical strong data processing inequality for the Gaussian channels serves as a special case if we set $D = \mathbb{R}^k$. The above inequality is the key to the proof of Theorem 22. We omit the proof of Lemma 16 since it is similar to the proof of Claim 3 in the proof of Theorem 22.

The upper and lower bounds given in Theorems 21 and 23 together establish the minimax rate of communication cost for statistically-optimal adaptive estimators:

$$Q(\tilde{S}_0, \tilde{C}, \mathcal{B}_{p,q}^\alpha(M)) \asymp m^3 + \left(\frac{M}{\sigma}\right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}} \quad (5.11)$$

where \tilde{C} is large enough and recall that $\tilde{S}_0 = \{(\alpha, M, p, q) : \alpha > 0, M > \sigma, 2 \leq p \leq \infty, 1 < q \leq \infty\}$. The minimax rate (5.11) also implies that $\hat{\theta}^A$ is the optimal adaptive distributed estimator with respect to both statistical performance and communication cost.

1. The estimator $\hat{\theta}^A$ simultaneously achieves the centralized optimal rate over the Besov classes $\mathcal{B}_{p,q}^\alpha(M)$ for all $\alpha > 0, M \geq \sigma, 1 < q \leq \infty$, and $2 \leq p \leq \infty$. There is no

statistical cost of adaptation in terms of the rate of convergence.

2. Among all the statistically-optimal adaptive estimators, the expected communication cost for $\hat{\theta}^A$ is rate-optimal over the Besov classes $\mathcal{B}_{p,q}^\alpha(M)$ for all $\alpha > 0, M \geq \sigma, 1 < q \leq \infty$, and $2 \leq p \leq \infty$.

Remark 18. Compared with the minimum communication cost $(\frac{M}{\sigma})^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}}$ for achieving the optimal rate of convergence in the minimax setting in (5.8), an additional communication cost of order m^3 bits is needed to achieve the adaptation over a collection of Besov classes. The term m^3 can be viewed as the communication cost of adaptation. This interplay between communication and statistical adaptation in the distributed setting is an interesting phenomenon: It costs more bits to communicate in order to achieve adaptivity. In contrast, statistical adaptation can be achieved for free in the centralized setting (Donoho and Johnstone, 1995; Johnstone, 2017).

5.4. Numerical Studies

The proposed seq-MODGAME estimator $\hat{\theta}^O$ and the adaptive local thresholding estimator $\hat{\theta}^A$ are easily to implement. In this section, we conduct simulation studies to investigate the numerical performance of these two estimators in various settings.

5.4.1. The seq-MODGAME estimator $\hat{\theta}^O$

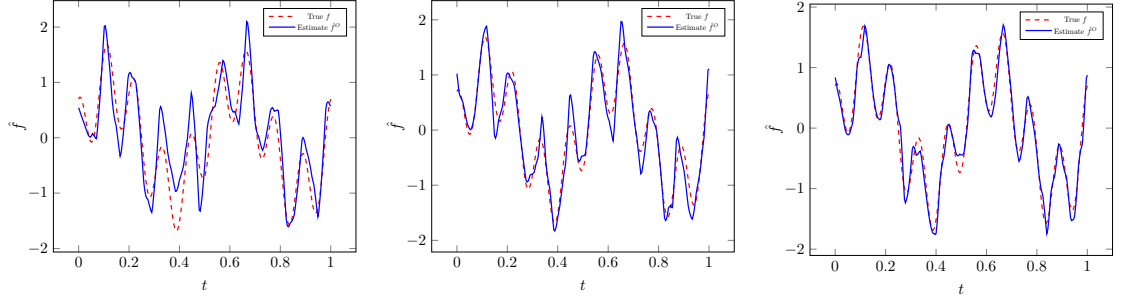
We first study the seq-MODGAME estimator $\hat{\theta}^O$ proposed in Section 5.2. We generate i.i.d data according to the distributed Gaussian sequence model (5.3) on $m = 100$ different virtual machines, where the mean vector θ is the wavelet coefficients of certain specified underlying function. The underlying function f is chosen as

$$f(t) = \sin(4\pi t) + 0.7 \cos(18\pi t) \quad t \in [0, 1]$$

and the noise level $\sigma = 1/16$.

We apply the optimal seq-MODGAME estimator $\hat{\theta}^O$ to estimate wavelet coefficients of f

given their noisy observations on virtual machines. Afterwards, we transform estimated wavelet coefficients back to estimated smooth functions \hat{f}^O . The results are shown in Figure 5.1. As more and more bits are allowed to communicate, the mean squared error are decreasing so that the estimate is becoming more and more accurate.



(a) $B = 100$, $\text{MSE} = 31.47$ (b) $B = 2400$, $\text{MSE} = 12.54$ (c) $B = 16000$, $\text{MSE} = 5.10$

Figure 5.1: Estimate given by the optimal seq-MODGAME estimator $\hat{\theta}^O$ under the communication constraints. For different choices of total communication budgets $B = 100, 2400, 16000$, we illustrate an example of estimated function \hat{f}^O in each figure. The mean squared error through 1000 trials are also given below each figure.

5.4.2. The local thresholding estimator $\hat{\theta}^A$

Similar to the setting in Section 5.4.1, we generate i.i.d data according to the distributed Gaussian sequence model (5.3) and set $m = 100, \sigma = 1/16$. However, in this simulation study we work on three different choices for the underlying functions $f = f_1, f_2$ or f_3 :

$$f_1(t) = 1.5 \sin(4\pi t) \quad t \in [0, 1];$$

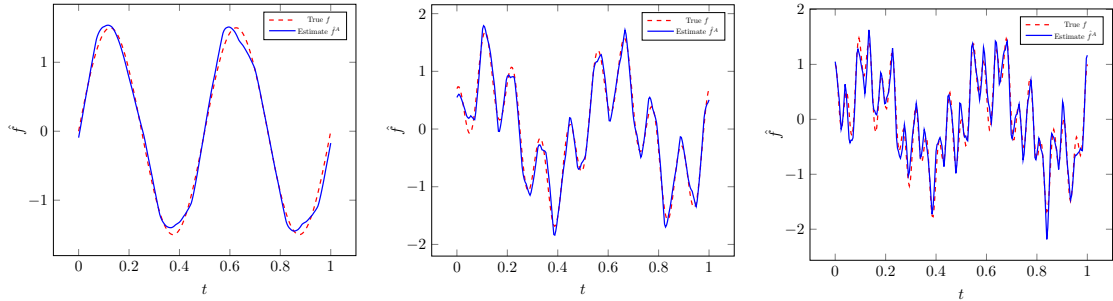
$$f_2(t) = \sin(4\pi t) + 0.7 \cos(18\pi t) \quad t \in [0, 1];$$

$$f_3(t) = 0.8 \sin(4\pi t) + 0.5 \cos(18\pi t) + 0.5 \cos(44\pi t) \quad t \in [0, 1].$$

The three functions given above are designed to have different smoothness. f_1 is the smoothest function among the three functions whereas f_3 is the most “wiggly” one. We expect to see a data-driven estimator can adapt to their smoothness automatically during the estimation.

Similarly, given random distributed data generated by adding noise to the wavelet coefficients

of f_1, f_2 and f_3 respectively, we apply the local thresholding estimator $\hat{\theta}^A$ to estimate the wavelet coefficients. The estimated smooth functions \hat{f}^A are obtained by reversed discrete wavelet transform on the estimated wavelet coefficients. The results are shown in Figure 5.2. It can be clearly seen from simulation that, when the underlying function are relatively smooth, the local thresholding estimator requires less communication cost while achieves better statistical accuracy. The numerical results are consistent with the theory, which shows the local thresholding estimator can adapt to the smoothness of the underlying function.



(a) $f_1 : \mathbb{E}L = 3330, \text{MSE} = 2.06$ (b) $f_2 : \mathbb{E}L = 8083, \text{MSE} = 5.03$ (c) $f_3 : \mathbb{E}L = 15862, \text{MSE} = 8.9$

Figure 5.2: Estimate given by the local thresholding estimator $\hat{\theta}^A$. Under different choices of ground truth functions f_1, f_2, f_3 , we illustrate an example of estimated function \hat{f}^A in each figure. The expected communication cost and their mean squared error through 1000 trials are also given below each figure.

5.5. Discussion

In this chapter, both distributed minimax and distributed adaptive estimation under the communication constraints were studied for the Gaussian sequence model and white noise model. Optimal minimax rate of convergence is established and the cost of adaptation is characterized. In addition, a data-driven adaptive distributed estimator with theoretical guarantees is constructed. Several technical tools and the formulation for the study of the interplay between adaptation and communication cost can be of independent interest.

Distributed nonparametric function estimation is still very much a new area with a range of interesting open problems. One such problem is the construction of an adaptive distributed procedure for Gaussian sequence estimation under a fixed communication constraint. It is notable that the communication cost for the local thresholding procedure $\hat{\theta}^A$ is related to

the smoothness of the underlying function. When the communication budget is tight, there is not enough budget to implement the local thresholding procedure. Therefore, it will be useful to have an estimator whose communication cost is controlled, while its estimation accuracy is adaptive to the smoothness of underlying function.

In this chapter, we focused on estimation over the Besov classes with $p \geq 2$. Another direction is the study of distributed Gaussian sequence estimation over the Besov classes with $p < 2$. Similar to the centralized setting, the case $p < 2$ is very different from the case $p \geq 2$ in the distributed setting. The techniques developed in this chapter are not sufficient for the case $p < 2$ and we leave this case for future work.

Besides the white noise model considered in this chapter, it is also interesting to study other related nonparametric function estimation problems, including nonparametric density estimation, nonparametric regression with fixed design, and nonparametric regression with random design, which have all been well studied in the centralized setting. In particular, it is shown that these three models are asymptotic equivalent to the white noise model (Nussbaum, 1996; Brown and Low, 1996; Brown et al., 2002, 2004) in the centralized setting under mild regularity conditions when the smoothness parameter $\alpha > \frac{1}{2}$. Practically, for example, by applying the root-unroot algorithm to the binned data (Brown et al., 2010), the density estimation problem can essentially be turned into the problem of nonparametric regression with fixed design. However, in the distributed settings, these four problems may exhibit different asymptotic behaviors due to the communication constraints. In the distributed setting, nonparametric density estimation, nonparametric regression with fixed design, and nonparametric regression with random design merit careful and separate investigations. We leave them for future work.

Broadly speaking, virtually any problem studied in the classical centralized setting has its counterpart in the distributed setting. Examples include minimax and adaptive estimation of linear and quadratic functionals as well as hypothesis testing under these nonparametric function models. It is challenging to develop a general optimality theory and construct statis-

tically optimal distributed procedures under the communication constraints, New technical tools for both the lower bound and upper bound analyses are needed.

5.6. Proofs

We prove Theorems 22 and 23 in this section. For reasons of space, the proofs of the other theorems, propositions and additional technical lemmas are given in the supplementary material (Cai and Wei, 2020a).

5.6.1. Notation and definitions

For any finite S , denote $\mathbb{U}(A)$ be a uniform distribution on S . For any a, b , let $a \lesssim b$ denote there exists a universal constant $C > 0$ such that $a \leq Cb$, whereas $a \gtrsim b$ denotes there exists a universal constant $c > 0$ such that $a \geq cb$. For any discrete random variables X, Y supported on \mathcal{X}, \mathcal{Y} , the entropy $H(X)$, conditional entropy $H(X|Y)$, and mutual information $I(X; Y)$ are defined as

$$\begin{aligned} H(X) &\triangleq - \sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \log \mathbb{P}(X = x), \\ H(X|Y) &\triangleq - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y) \log \mathbb{P}(X = x|Y = y), \\ I(X; Y) &\triangleq \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y) \log \frac{\mathbb{P}(X = x|Y = y)}{\mathbb{P}(X = x)}. \end{aligned}$$

5.6.2. Proof of Theorem 22

It follows from Theorem 20 that for any estimator $\hat{\theta}$ satisfying $\sup_{\theta \in \mathcal{B}_{p,q}^\alpha(M)} \mathbb{E}_\theta L(\hat{\theta}) \leq B$, we have

$$\sup_{\theta \in \mathcal{B}_{p,q}^\alpha(M)} \mathbb{E} \|\hat{\theta} - \theta\|^2 \geq cM^{\frac{2}{\alpha+1}} \left(\frac{\sigma^2}{B} \right)^{\frac{\alpha}{\alpha+1}}$$

for some constant $c > 0$. By the assumption,

$$\sup_{\theta \in \mathcal{B}_{p_2, q_2}^{\alpha_2}(M_2)} \mathbb{E} \|\hat{\theta} - \theta\|^2 \leq CM_2^{\frac{2}{2\alpha_2+1}} \left(\frac{\sigma^2}{m} \right)^{\frac{2\alpha_2}{2\alpha_2+1}}.$$

So it follows that

$$\sup_{\theta \in \mathcal{B}_{p_2, q_2}^{\alpha_2}(M_2)} \mathbb{E}_\theta L(\hat{\theta}) \gtrsim \left(\frac{M_2}{\sigma} \right)^{\frac{2}{2\alpha_2+1}} m^{\frac{2\alpha_2+2}{2\alpha_2+1}}.$$

To prove Theorem 22, it now suffices to show

$$\sup_{\theta \in \mathcal{B}_{p_2, q_2}^{\alpha_2}(M_2)} \mathbb{E}_\theta L(\hat{\theta}) \gtrsim \left(\frac{M_1}{\sigma} \right)^{\frac{2}{2\alpha_1+1}} m^{\frac{2\alpha_1+2}{2\alpha_1+1}}.$$

The remain part of the proof aims to prove the above inequality.

Define the constant λ (only depends on C) and variable u as follows:

$$\lambda = \max\{10, 32\sqrt{C}\},$$

$$u = \left(\frac{M_1}{\sigma} \right)^{\frac{2}{2\alpha_1+1}} m^{\frac{1}{2\alpha_1+1}}.$$

Define the set of sequences

$$S_{m,u} \triangleq \left\{ \left(\tau_1 \frac{\lambda\sigma}{\sqrt{m}}, \tau_2 \frac{\lambda\sigma}{\sqrt{m}}, \dots, \tau_u \frac{\lambda\sigma}{\sqrt{m}}, 0, 0, \dots \right) : \tau_1, \tau_2, \dots, \tau_u \in \{-1, +1\} \right\}.$$

Since for any $\theta \in S_{m,u}$ and $p_1, q_1 < \infty$ we have

$$\begin{aligned} |\theta|_{b_{p_1, q_1}^{\alpha_1}} &= \left(\sum_{j=0}^{\infty} \left(2^{j(\alpha_1+1/2-1/p_1)} \left(\sum_{k=1}^{2^j} |\theta_{jk}|^{p_1} \right)^{1/p_1} \right)^{q_1} \right)^{1/q_1} \\ &\leq \left(\sum_{j=0}^{[\log u]+1} 2^{jq_1(\alpha_1+1/2)} \right)^{1/q_1} \frac{\lambda\sigma}{\sqrt{m}} \\ &\leq \left(\frac{(2u)^{q_1(\alpha_1+1/2)}}{1-2^{-q_1(\alpha_1+1/2)}} \right)^{1/q_1} \frac{\lambda\sigma}{\sqrt{m}} \\ &= \lambda \left(\frac{2^{q_1(\alpha_1+1/2)}}{1-2^{-q_1(\alpha_1+1/2)}} \right)^{1/q_1} u^{\alpha_1+1/2} \frac{\sigma}{\sqrt{m}} \leq M_1. \end{aligned}$$

When $p_1 = \infty$ or $q_1 = \infty$, the above inequality also holds by similar argument. Therefore we have $S_{m,u} \subset \mathcal{B}_{p_1, q_1}^{\alpha_1}(M_1)$.

Since we have assumed

$$\sup_{\theta \in \mathcal{B}_{p_1, q_1}^{\alpha_1}(M_1)} \mathbb{E} \|\hat{\theta} - \theta\|^2 \leq CM_1^{\frac{2}{2\alpha_1+1}} \left(\frac{\sigma^2}{m} \right)^{\frac{2\alpha_1}{2\alpha_1+1}}.$$

Note that the maximum risk is lower bounded by the Bayesian risk, assign to θ a uniform prior $\theta \sim \mathbb{U}(S_{m,u})$, then we have

$$\mathbb{E}_{\theta \sim \mathbb{U}(S_m)} \|\hat{\theta} - \theta\|^2 \leq CM_1^{\frac{2}{2\alpha_1+1}} \left(\frac{\sigma^2}{m} \right)^{\frac{2\alpha_1}{2\alpha_1+1}} = Cu \frac{\sigma^2}{m}.$$

In the following proof, we are going to provide several claims and prove each claim accordingly. Let Q_0 denote the probability law of X_1 when $\theta = (0, 0, 0, \dots, 0, \dots)$. Let Q_m denote the probability law of X_1 when $\theta \sim \mathbb{U}(S_{m,u})$. Note that there are multiple distributions we need to consider, we shorthand the probability, expectation, entropy and mutual information when $\theta = (0, 0, 0, \dots, 0, \dots)$ as $\mathbb{P}_0, \mathbb{E}_0, H_0$ and I_0 respectively. Similarly we use shorthands $\mathbb{P}_m, \mathbb{E}_m, H_m$ and I_m to denote those quantities when $\theta \sim \mathbb{U}(S_{m,u})$.

Claim 6. *We have $I_m(\hat{\theta}, \theta) \geq \frac{15}{16}u$.*

PROOF OF CLAIM 6: Define $\hat{\theta}^* \triangleq \mathcal{P}_{S_{m,u}}(\hat{\theta})$ be the nearest point in $S_{m,u}$ to $\hat{\theta}$. Then we have

$$\mathbb{E}_m \|\hat{\theta}^* - \theta\|^2 \leq 4\mathbb{E}_m \|\hat{\theta} - \theta\|^2 \leq 4Cu \frac{\sigma^2}{m}. \quad (5.12)$$

Note that $\hat{\theta}^* \in S_m$ thus we can reparametrize $\hat{\theta}^*$ to

$$\hat{\theta}^* = \left(\hat{\tau}_1 \frac{\lambda\sigma}{\sqrt{m}}, \hat{\tau}_2 \frac{\lambda\sigma}{\sqrt{m}}, \dots, \hat{\tau}_{m^2} \frac{\lambda\sigma}{\sqrt{m}}, 0, 0, \dots \right) \quad \text{where } \hat{\tau}_1, \hat{\tau}_2, \dots, \hat{\tau}_{m^2} \in \{-1, +1\}$$

Then we can simplify (5.12) to

$$\mathbb{E}_m \sum_{k=1}^{m^2} (\hat{\tau}_k - \tau_k)^2 \leq 4C\lambda^{-2}u. \quad (5.13)$$

Recall that $\lambda = \max\{10, 32\sqrt{C}\}$. Substitute into (5.13) we have

$$\mathbb{E}_m \sum_{k=1}^{m^2} (\hat{\tau}_k - \tau_k)^2 \leq \frac{1}{256}u.$$

Apply Fano's inequality, we can conclude

$$\sum_{k=1}^{m^2} H_m(\tau_k | \hat{\tau}_k) \leq \frac{1}{16}u.$$

The following lemma is instrumental to establish later results:

Lemma 17. *If A is a random variable and Y_1, Y_2, \dots, Y_d are independent random variables, then*

$$I(A; (Y_1, Y_2, \dots, Y_d)) \geq \sum_{k=1}^d I(A; Y_k).$$

Note that $\tau_1, \tau_2, \dots, \tau_k$ are i.i.d Rademacher variables, apply Lemma 17 we have

$$\begin{aligned} I_m(\hat{\theta}^*; \theta) &= I_m(\hat{\theta}^*; (\tau_1, \tau_2, \dots, \tau_{m^2})) \geq \sum_{k=1}^{m^2} I_m(\hat{\theta}^*; \tau_k) \geq \sum_{k=1}^{m^2} I_m(\hat{\tau}_k; \tau_k) \\ &= \sum_{k=1}^{m^2} H_m(\tau_k) - H_m(\tau_k | \hat{\tau}_k) \geq \frac{15}{16}m^2. \end{aligned}$$

The second inequality above is due to data processing inequality applied to the fact $\hat{\tau}_k$ only depends on $\hat{\theta}^*$. Finally the claim can be concluded by data processing inequality $I_m(\hat{\theta}; \theta) \geq I_m(\hat{\theta}^*; \theta)$.

Claim 7. Let $\delta > 0$ be a parameter that will be specified later. For any $\delta > 0$, there exist a constant $C_3 > 0$ (depending on C, α_1 and δ) such that

$$\mathbb{P}_m \left(\frac{dQ_m}{dQ_0}(X_1) > C_3 \right) \leq \delta, \quad (5.14)$$

$$I_m \left(X_1; \theta \middle| \frac{dQ_m}{dQ_0}(X_1) > C_3 \right) \mathbb{P}_m \left(\frac{dQ_m}{dQ_0}(X_1) > C_3 \right) \leq \frac{u}{2m}. \quad (5.15)$$

PROOF OF CLAIM 7: We first prove (5.14) holds with large enough constant C_3 . Let $X_{1,k}$ denote the k -th coordinate of X_1 . Note that

$$\frac{dQ_m}{dQ_0}(X_1) = \prod_{k=1}^u \left(e^{-\frac{\lambda^2}{2m}} \cdot \frac{e^{-\frac{\lambda X_{1,k}}{\sqrt{m}\sigma}} + e^{\frac{\lambda X_{1,k}}{\sqrt{m}\sigma}}}{2} \right).$$

Using the basic inequality $\ln\left(\frac{e^t + e^{-t}}{2}\right) \leq \frac{t^2}{2}$, we have

$$\begin{aligned} \mathbb{P}_m \left(\frac{dQ_m}{dQ_0}(X_1) > C_3 \right) &= \mathbb{P}_m \left(\ln \frac{dQ_m}{dQ_0}(X_1) > \ln C_3 \right) \\ &= \mathbb{P}_m \left(\sum_{k=1}^u \left(\ln \left(\frac{e^{-\frac{\lambda X_{1,k}}{\sqrt{m}\sigma}} + e^{\frac{\lambda X_{1,k}}{\sqrt{m}\sigma}}}{2} \right) - \frac{\lambda^2}{2m} \right) > \ln C_3 \right) \\ &\leq \mathbb{P}_m \left(\sum_{k=1}^u \frac{\lambda^2}{2m\sigma^2} (X_{1,k}^2 - \sigma^2) > \ln C_3 \right) \\ &= \mathbb{P}_m \left(\sum_{k=1}^u \frac{\lambda^2}{2m\sigma^2} \left(X_{1,k}^2 - \sigma^2 - \frac{\lambda^2\sigma^2}{m} \right) > \ln C_3 - \frac{\lambda^4 u}{2m^2} \right). \end{aligned}$$

Note that $\sum_{k=1}^u \frac{\lambda^2}{2m\sigma^2} \left(X_{1,k}^2 - \sigma^2 - \frac{\lambda^2\sigma^2}{m} \right)$ has mean 0 and variance at most $(1 + \lambda^2)\lambda^4 u/m^2$.

Note that we have assumed $M_1 \leq C\sigma m^{2\alpha_1 + \frac{1}{2}}$, this implies $u \leq C^{\frac{2}{2\alpha_1 + 1}} m^2$. So by Chebyshev's inequality, as long as

$$\ln C_3 \geq \frac{C^{\frac{2}{2\alpha_1 + 1}} \lambda^4}{2} + \sqrt{C^{\frac{2}{2\alpha_1 + 1}} (1 + \lambda^2) \lambda^4 / \delta},$$

we have

$$\mathbb{P}_m \left(\frac{dQ_m}{dQ_0}(X_1) > C_3 \right) \leq \delta.$$

We now prove the second inequality (5.15). Note that when $\theta \sim \mathbb{U}(S_{m,u})$, θ and event $\{\frac{dQ_m}{dQ_0}(X_1) > C_3\}$ are independent (due to symmetry of $S_{m,u}$). Define

$$\theta_a = \left(\frac{\lambda\sigma}{\sqrt{m}}, \frac{\lambda\sigma}{\sqrt{m}}, \dots, \frac{\lambda\sigma}{\sqrt{m}}, 0, 0, \dots, 0 \right) \in S_m.$$

By symmetry, it is easy to show that

$$\begin{aligned} I_m \left(X_1; \theta \middle| \frac{dQ_m}{dQ_0}(X_1) > C_3 \right) & \mathbb{P}_m \left(\frac{dQ_m}{dQ_0}(X_1) > C_3 \right) \\ & = \int_{\frac{dQ_m}{dQ_0}(X_1) > C_3} p(x_1 | \theta = \theta_a) \log \frac{p(x_1 | \theta = \theta_a)}{q_m(x_1)} dx_1 \end{aligned}$$

where $p(x_1 | \theta = \theta_a)$ denote the density of x_1 when $\theta = \theta_0$, and $q_m(x_1)$ denote the density of law Q_m .

Further, note that we have following decomposition for $p(x_1 | \theta = \theta_a)$ and $q_m(x_1)$:

$$\begin{aligned} p(x_1 | \theta = \theta_a) & = \prod_{i=1}^u \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_{1,k} - \frac{\lambda\sigma}{\sqrt{m}})^2}{2\sigma^2}}, \\ q_m(x_1 | \theta = \theta_a) & = \prod_{i=1}^u \frac{1}{2\sqrt{2\pi}\sigma} \left(e^{-\frac{(x_{1,k} - \frac{\lambda\sigma}{\sqrt{m}})^2}{2\sigma^2}} + e^{-\frac{(x_{1,k} + \frac{\lambda\sigma}{\sqrt{m}})^2}{2\sigma^2}} \right). \end{aligned}$$

So we can get

$$\begin{aligned}
& \int_{\frac{dQ_m}{dQ_0}(X_1) > C_3} p(x_1 | \theta = \theta_a) \log \frac{p(x_1 | \theta = \theta_a)}{q_m(x_1)} dx_1 \\
&= u \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - \frac{\lambda\sigma}{\sqrt{m}})^2}{2\sigma^2}} \cdot \log \left(\frac{2}{1 + \exp(-\frac{2\lambda y}{\sqrt{m}\sigma})} \right) \mathbb{P}_m \left(\frac{dQ_m}{dQ_0}(X_1) > C_3 \middle| x_{1,1} = y \right) dy \\
&\leq u \int_{y \in [-2\lambda\sqrt{m}\sigma, 2\lambda\sqrt{m}\sigma]} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - \frac{\lambda\sigma}{\sqrt{m}})^2}{2\sigma^2}} \cdot \log \left(\frac{2}{1 + \exp(-\frac{2\lambda y}{\sqrt{m}\sigma})} \right) \\
&\quad \cdot \mathbb{P}_m \left(\frac{dQ_m}{dQ_0}(X_1) > C_3 \middle| x_{1,1} = y \right) dy \\
&\quad + u \int_{y \notin [-2\lambda\sqrt{m}\sigma, 2\lambda\sqrt{m}\sigma]} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - \frac{\lambda\sigma}{\sqrt{m}})^2}{2\sigma^2}} \cdot \log \left(\frac{2}{1 + \exp(-\frac{2\lambda y}{\sqrt{m}\sigma})} \right) \\
&\quad \cdot \mathbb{P}_m \left(\frac{dQ_m}{dQ_0}(X_1) > C_3 \middle| x_{1,1} = y \right) dy. \tag{5.16}
\end{aligned}$$

Now we bound the first term of the right hand side in (5.16). It can be shown that when C_3 is a large enough constant, we could get

$$\mathbb{P}_m \left(\frac{dQ_m}{dQ_0}(X_1) > C_3 \middle| x_{1,1} = 2\lambda\sqrt{m}\sigma \right) \leq \frac{\ln 2}{4\lambda^2},$$

(we omit the proof here because it is similar to the proof of (5.14).)

Thus it is easy to show

$$\begin{aligned}
& \int_{y \in [-2\lambda\sqrt{m}\sigma, 2\lambda\sqrt{m}\sigma]} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - \frac{\lambda\sigma}{\sqrt{m}})^2}{2\sigma^2}} \cdot \log \left(\frac{2}{1 + \exp(-\frac{2\lambda y}{\sqrt{m}\sigma})} \right) \\
&\quad \cdot \mathbb{P}_m \left(\frac{dQ_m}{dQ_0}(X_1) > C_3 \middle| x_{1,1} = y \right) dy \\
&\leq \frac{\ln 2}{4\lambda^2} \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - \frac{\lambda\sigma}{\sqrt{m}})^2}{2\sigma^2}} \cdot \log \left(\frac{2}{1 + \exp(-\frac{2\lambda y}{\sqrt{m}\sigma})} \right) dy \\
&\leq \frac{\ln 2}{4\lambda^2} \cdot \frac{1}{\ln 2} \left(\frac{\lambda}{\sqrt{m}} \right)^2 = \frac{1}{4m}.
\end{aligned}$$

where the second inequality is due to the entropy bound given in Michalowicz et al. (2008).

Next we are going to bound the second term of the right hand side in (5.16). Because $\lambda \geq 10$, it is easy to show

$$\begin{aligned} & \int_{y \notin [-2\lambda\sqrt{m}\sigma, 2\lambda\sqrt{m}\sigma]} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - \frac{\lambda\sigma}{\sqrt{m}})^2}{2\sigma^2}} \cdot \log \left(\frac{2}{1 + \exp(-\frac{2\lambda y}{\sqrt{m}\sigma})} \right) \\ & \cdot \mathbb{P}_m \left(\frac{dQ_m}{dQ_0}(X_1) > C_3 \mid x_{1,1} = y \right) dy \\ & \leq \log 2 \cdot \int_{y \notin [-2\lambda\sqrt{m}\sigma, 2\lambda\sqrt{m}\sigma]} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - \frac{\lambda\sigma}{\sqrt{m}})^2}{2\sigma^2}} dy \leq 2 \exp \left(-\frac{(2\lambda\sqrt{m} - \frac{\lambda}{\sqrt{m}})^2}{2} \right) < \frac{1}{4m}. \end{aligned}$$

Apply the above two bounds to (5.16) we can get

$$\int_{\frac{dQ_m}{dQ_0}(X_1) > C_3} p(x_1 | \theta = \theta_a) \log \frac{p(x_1 | \theta = \theta_a)}{q_m(x_1)} dx_1 \leq \frac{u}{2m}$$

when C_3 is a large enough constant. This directly implies inequality (5.15).

Denote the set $R = \{x \in \mathbb{R}^\infty : \frac{dQ_m}{dQ_0}(x) \leq C_3\}$ and random variable $W_i = \mathbb{I}_{\{X_i \in R\}}$.

Claim 8. *For each $i = 1, 2, \dots, m$, we have*

$$I_m(Z_i; \theta | X_i \in R) \mathbb{P}_m(X_i \in R) \leq \frac{256\lambda^2}{m} (\mathbb{E}_m(L_i \mathbb{I}_{\{X_i \in R\}}) + H_m(W_i)).$$

PROOF OF CLAIM 8: Let \tilde{Z}_i defined as

$$\tilde{Z}_i \triangleq \begin{cases} Z_i & \text{if } X \in R \\ \star & \text{if } X \notin R \end{cases}$$

where \star is a unique symbol which is different with any 0-1 string.

The following lemma is instrumental to establishing later results.

Lemma 18 (Multidimensional strong data processing inequality).

Suppose $T = (T^{(1)}, T^{(2)}, \dots, T^{(d)})$ be a collection of random variables where each entry is

an i.i.d Bernoulli random variable with mean $\frac{1}{2}$. Let μ_0 be a d -dimensional vector and $\Delta > 0$ be a positive real number. Let X be a d -dimensional Gaussian random variable where $X^{(1)}, X^{(2)}, \dots, X^{(d)}$ are independent with distribution

$$X^{(k)} \sim N(\mu_0^{(k)} + T^{(k)} \Delta, \sigma^2).$$

Let Z be a discrete random variable such that $T \rightarrow X \rightarrow Z$ is a Markov chain, i.e. $Z \perp T | X$. Then the following multidimensional strong data processing inequality holds:

$$I(T; Z) \leq 64 \left(\frac{\Delta}{\sigma} \right)^2 I(X; Z). \quad (5.17)$$

Lemma 18 has been proved in Cai and Wei (2020c). For sake of completeness, we provide its proof in the present supplementary material.

Apply Lemma 18 on Markov chain $\theta \rightarrow X_i \rightarrow \tilde{Z}_i$ where $\theta \sim \mathbb{U}(S_m)$, we have

$$I_m(\theta; \tilde{Z}_i) \leq \frac{256\lambda^2}{m} I_m(\tilde{Z}_i; X_i).$$

Note that $W_i \perp \theta$ when $\theta \sim \mathbb{U}(S_{m,u})$, and W_i is determined given \tilde{Z} , we have

$$\begin{aligned} I_m(\theta; \tilde{Z}_i) &= I_m(\theta; (\tilde{Z}_i, W_i)) = I_m(\theta; \tilde{Z}_i | W_i) + I_m(\theta; W_i) \\ &= I_m(\theta; \tilde{Z}_i | X_i \in R) \mathbb{P}_m(X_i \in R) + I_m(\theta; \tilde{Z}_i | X_i \notin R) \mathbb{P}_m(X_i \notin R) + I_m(\theta; W_i) \\ &= I_m(\theta; \tilde{Z}_i | X_i \in R) \mathbb{P}_m(X_i \in R). \end{aligned}$$

For similar reasons, we have

$$\begin{aligned}
I_m(\tilde{Z}_i; X_i) &\leq H_m(\tilde{Z}_i) = H_m(\tilde{Z}_i, W_i) = H_m(\tilde{Z}_i|W_i) + H_m(W_i) \\
&= H_m(\tilde{Z}_i|X_i \in R)\mathbb{P}(X_i \in R) + H_m(\tilde{Z}_i|X_i \notin R)\mathbb{P}(X_i \notin R) + H_m(W_i) \\
&= H_m(Z_i|X_i \in R)\mathbb{P}(X_i \in R) + H_m(W_i) \\
&\leq \mathbb{E}_m(L_i|X_i \in R)\mathbb{P}(X_i \in R) + H_m(W_i)
\end{aligned}$$

where the latter inequality is due to Shannon's source coding theorem (Shannon, 1948).

Combining the above three formulas yields the desired inequality.

PROOF OF THE MAIN THEOREM:

Note that the region R is "symmetric" where $x \in R$ is equivalent to $|x| \in R$ ($|x|$ is entry-wise absolute value). So $\mathbb{P}(X \in R|\theta)$ is invariant for all $\theta \in S_m$, therefore $W_i \perp \theta$ when $\theta \sim \mathbb{U}(S_m)$. Based on this, for each $i = 1, 2, \dots, m$ we have

$$\begin{aligned}
I_m(Z_i; \theta) &\leq I_m((Z_i, W); \theta) = I_m(Z_i; \theta|W) + I_m(W; \theta) \\
&= I_m(Z_i; \theta|W) \\
&= I_m(Z_i; \theta|X_i \in R)\mathbb{P}_m(X_i \in R) + I_m(Z_i; \theta|X_i \notin R)\mathbb{P}_m(X_i \notin R) \quad (5.18) \\
&\leq I_m(Z_i; \theta|X_i \in R)\mathbb{P}_m(X_i \in R) + I_m(X_i; \theta|X_i \notin R)\mathbb{P}_m(X_i \notin R) \\
&\leq \frac{256\lambda^2}{m} (\mathbb{E}_m(L_i\mathbb{I}_{\{X_i \in R\}}) + H_m(W_i)) + \frac{u}{2m}
\end{aligned}$$

where the second inequality is due to data processing inequality and the last inequality is derived from Claim 7 and Claim 8.

Taking summation over (5.18), we have

$$\frac{256\lambda^2}{m} \left(mH_m(W_1) + \sum_{i=1}^m \mathbb{E}_m(L_i\mathbb{I}_{\{X_i \in R\}}) \right) + \frac{u}{2} \geq \sum_{i=1}^m I_m(Z_i; \theta) \geq I_m(\hat{\theta}; \theta) \geq \frac{15}{16}u \quad (5.19)$$

where the last inequality is due to Claim 6.

Note that for each $i = 1, 2, \dots, m$, we have

$$\mathbb{E}_m(L_i \mathbb{I}_{\{X_i \in R\}}) = \mathbb{E}_0(L_i \mathbb{I}_{\{X_i \in R\}} \frac{dQ_m}{dQ_0}(X_i)) \leq C_3 \mathbb{E}_0(L_i \mathbb{I}_{\{X_i \in R\}}) \leq C_3 \mathbb{E}_0(L_i).$$

Substitute the above inequality into (5.19) we can get

$$\mathbb{E}_0(L) = \sum_{i=1}^m \mathbb{E}_0(L_i) \geq \frac{1}{C_3} \left(\frac{7}{4096\lambda^2} mu - mH_m(W_1) \right).$$

Note that $H_m(W_1) \leq -\delta \log \delta - (1 - \delta) \log(1 - \delta)$. We can always set δ to a sufficient small constant so that $H_m(W_1) \leq \frac{7}{2048\lambda^2}$. Note that $u \geq 1$, then we can conclude that

$$\mathbb{E}_0(L) \geq \frac{7}{8192C_3\lambda^2} mu.$$

Finally, for any $\alpha, p, M > 0$, given the fact that $(0, 0, 0, \dots, 0, \dots) \in \mathcal{B}_{p,q}^\alpha(M)$, we have

$$\sup_{\theta \in \mathcal{B}_{p,q}^\alpha(M)} \mathbb{E}_\theta L \geq \mathbb{E}_0(L) \geq \frac{7}{8192C_3\lambda^2} mu \gtrsim \left(\frac{M_1}{\sigma} \right)^{\frac{2}{2\alpha_1+1}} m^{\frac{2\alpha_1+2}{2\alpha_1+1}}.$$

5.6.3. Proof of Theorem 23

This theorem can be viewed as an extension of Theorem 22. Note that there exists

$(\alpha_0, M_0, p_0, q_0) \in \tilde{\mathcal{S}}_0$ such that

$$M_0 = \sigma m^{2\alpha_0 + \frac{1}{2}}. \tag{5.20}$$

Note that for any $\hat{\theta} \in \mathcal{G}(\tilde{\mathcal{S}}, C(\cdot))$ and $(\alpha, M, p, q) \in \tilde{\mathcal{S}}$, we have

$$\sup_{\theta \in \mathcal{B}_{p,q}^\alpha(M)} \mathbb{E} \|\hat{\theta} - \theta\|^2 \leq \tilde{C}(\alpha) M^{\frac{2}{2\alpha+1}} \left(\frac{\sigma^2}{m} \right)^{\frac{2\alpha}{2\alpha+1}},$$

$$\sup_{\theta \in \mathcal{B}_{p_0, q_0}^{\alpha_0}(M)} \mathbb{E} \|\hat{\theta} - \theta\|^2 \leq \tilde{C}(\alpha_0) M_0^{\frac{2}{2\alpha_0+1}} \left(\frac{\sigma^2}{m} \right)^{\frac{2\alpha_0}{2\alpha_0+1}}.$$

Based on above two inequalities and (5.20), apply Theorem 22, then apply (5.20) again, we can conclude

$$\sup_{\theta \in \mathcal{B}_{p, q}^{\alpha}(M)} \mathbb{E} L(\hat{\theta}) \gtrsim \left(\frac{M_0}{\sigma} \right)^{\frac{2}{2\alpha_0+1}} m^{\frac{2\alpha_0+2}{2\alpha_0+1}} + \left(\frac{M}{\sigma} \right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}} = m^3 + \left(\frac{M}{\sigma} \right)^{\frac{2}{2\alpha+1}} m^{\frac{2\alpha+2}{2\alpha+1}}.$$

APPENDIX

SUPPLEMENTARY MATERIALS

Due to the limit of the space, we put additional proofs into supplementary materials. Please refer the supplements of Chapter 2-5 to the following documents.

1. **Supplement1 Transfer Learning.pdf.** This document is a copy of Cai and Wei (2019), serves as supplement to Chapter 2: Transfer Learning for Nonparametric Classification.
2. **Supplement2 Distributed Gaussian.pdf.** This document is a copy of Cai and Wei (2020a), serves as supplement to Chapter 3: Distributed Gaussian Mean Estimation with Known Variance under Communication Constraints.
3. **Supplement3 Distributed Adaptive Gaussian.pdf.** This document is a copy of Cai and Wei (2021a), serves as supplement to Chapter 4: Distributed Gaussian Mean Estimation with Unknown Variance under Communication Constraints.
4. **Supplement4 Distributed Nonparametric.pdf.** This document is a copy of Cai and Wei (2020b), serves as supplement to Chapter 5: Distributed Nonparametric Function Estimation under Communication Constraints.

BIBLIOGRAPHY

- Felix Abramovich, Yoav Benjamini, David L Donoho, and Iain M Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2): 584–653, 2006.
- Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Distributed signal detection under communication constraints. In *Conference on Learning Theory*, pages 41–63. PMLR, 2020.
- Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.
- Leighton Pate Barnes, Yanjun Han, and Ayfer Özgür. Fisher information for distributed estimation under a blackboard communication protocol. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2704–2708. IEEE, 2019a.
- Leighton Pate Barnes, Yanjun Han, and Ayfer Özgür. Learning distributions from their samples under communication constraints. *CoRR*, abs/1902.02890, 2019b. URL <http://arxiv.org/abs/1902.02890>.
- Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *Annals of statistics*, 46(3):1352, 2018.
- Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 137–144, 2007.
- P. J. Bickel. Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.*, 9(6):1301–1309, 11 1981a. doi: 10.1214/aos/1176345646. URL <https://doi.org/10.1214/aos/1176345646>.
- PJ Bickel. Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *The Annals of Statistics*, 9(6):1301–1309, 1981b.
- Richard E Blahut and Richard E Blahut. *Principles and practice of information theory*, volume 1. Addison-Wesley Reading, MA, 1987.
- John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 129–136, 2008.
- Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and David P Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Com-*

- puting, pages 1011–1020. ACM, 2016.
- Lawrence Brown, T. Tony Cai, Ren Zhang, Linda Zhao, and Harrison Zhou. The root–unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probability Theory and Related Fields*, 146(3-4):401, 2010.
- Lawrence D Brown and Mark G Low. Asymptotic equivalence of nonparametric regression and white noise. *The Annals of statistics*, 24(6):2384–2398, 1996.
- Lawrence D Brown, T Tony Cai, Mark G Low, and Cun-Hui Zhang. Asymptotic equivalence theory for nonparametric regression with random design. *The Annals of statistics*, 30(3): 688–707, 2002.
- Lawrence D Brown, Andrew V Carter, Mark G Low, and Cun-Hui Zhang. Equivalence theory for density estimation, poisson processes and gaussian white noise with drift. *The Annals of Statistics*, 32(5):2074–2097, 2004.
- T Tony Cai. Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *The Annals of Statistics*, 27(3):898–924, 1999.
- T. Tony Cai. On information pooling, adaptability and superefficiency in nonparametric function estimation. *Journal of Multivariate Analysis*, 99:421–436, 2008.
- T. Tony Cai and M. G. Low. Adaptive confidence balls. *The Annals of Statistics*, 34:202–228, 2006.
- T Tony Cai and Hongji Wei. Supplement to “Transfer learning for nonparametric classification: Minimax rate and adaptive classifier”. 2019.
- T Tony Cai and Hongji Wei. Supplement to “Distributed Gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms”. 2020a.
- T Tony Cai and Hongji Wei. Supplement to “Distributed nonparametric regression: Optimal rate of convergence and cost of adaptation”. 2020b.
- T Tony Cai and Hongji Wei. Distributed gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms. *arXiv preprint arXiv:2001.08877*, 2020c.
- T Tony Cai and Hongji Wei. Supplement to “Distributed adaptive gaussian mean estimation with unknown variance: Interactive protocol helps adaptation”. 2021a.
- T Tony Cai and Hongji Wei. Distributed nonparametric function estimation: Optimal rate of convergence and cost of adaptation. *The Annals of Statistics*, to appear, 2021b.

- T Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128, 2021c.
- T Tony Cai and Hongji Wei. Distributed adaptive gaussian mean estimation with unknown variance: Interactive protocol helps adaptation. *The Annals of Statistics*, to appear, 2021d.
- T. Tony Cai and Harrison H Zhou. A data-driven block thresholding approach to wavelet estimation. *The Annals of Statistics*, 37:569–595, 2009.
- Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein's method*. Springer Science & Business Media, 2010.
- Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Transfer learning for music classification and regression tasks. *arXiv preprint arXiv:1703.09179*, 2017.
- Thomas M Cover and Peter E Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.
- Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from data of variable quality. In *Advances in Neural Information Processing Systems*, pages 219–226, 2006.
- Marc Deisenroth and Jun Wei Ng. Distributed Gaussian processes. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1481–1490, 07–09 Jul 2015. URL <http://proceedings.mlr.press/v37/deisenroth15.html>.
- Ilias Diakonikolas, Elena Grigorescu, Jerry Li, Abhiram Natarajan, Krzysztof Onak, and Ludwig Schmidt. Communication-efficient distributed learning of discrete distributions. In *Advances in Neural Information Processing Systems*, pages 6391–6401, 2017.
- Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *arXiv preprint arXiv:1810.00412*, 2018.
- David L Donoho and Iain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.
- David L Donoho and Iain M Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.
- David L Donoho and Iain M Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921, 1998.
- Jianqing Fan, Dong Wang, Kaizheng Wang, and Ziwei Zhu. Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47(6):3009–3031, 2019.

- Sabastien Gadat, Thierry Klein, and Clément Marteau. Classification in general finite dimensional spaces with the k -nearest neighbor rule. *The Annals of Statistics*, 44(3): 982–1009, 2016.
- João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):44, 2014.
- Ankit Garg, Tengyu Ma, and Huy Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2014.
- Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073. IEEE, 2012.
- László Györfi. On the rate of convergence of nearest neighbor rules (corresp.). *IEEE Transactions on Information Theory*, 24(4):509–512, 1978.
- Uri Hadar and Ofer Shayevitz. Distributed estimation of gaussian correlations. *IEEE Transactions on Information Theory*, 65(9):5323–5338, 2019.
- Peter Hall, Gérard Kerkycharian, and Dominique Picard. On the minimax optimality of block thresholded wavelet estimators. *Statistica Sinica*, 9(1):33–49, 1999.
- Y. Han, P. Mukherjee, A. Ozgur, and T. Weissman. Distributed statistical estimation of high-dimensional and nonparametric distributions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 506–510, 2018.
- Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. *arXiv preprint arXiv:1802.08417*, 2018.
- Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7304–7308. IEEE, 2013.
- I Ibragimov and R Khasminskii. Some estimation problems in infinite dimensional gaussian white noise. In *Festschrift for Lucien Le Cam*, pages 259–274. Springer, 1997.
- Brian A Johnson and Kotaro Iizuka. Integrating openstreetmap crowdsourced data and landsat time-series imagery for rapid land use/land cover (lulc) mapping: Case study of the laguna de bay area of the philippines. *Applied Geography*, 67:140–149, 2016.
- Iain M Johnstone. *Gaussian Estimation: Sequence and Wavelet Models*. Manuscript, 2017.

- Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.
- David R Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal crowdsourcing using low-rank matrix approximations. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 284–291. IEEE, 2011.
- Alon Kipnis and John C. Duchi. Mean estimation from adaptive one-bit measurements. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1000–1007, 2017. doi: 10.1109/ALLERTON.2017.8262847.
- Alon Kipnis and John C Duchi. Mean estimation from one-bit measurements. *arXiv preprint arXiv:1901.03403*, 2019.
- Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.
- Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. *arXiv preprint arXiv:1803.01833*, 2018.
- Eyal Kushilevitz. Communication complexity. In *Advances in Computers*, volume 44, pages 331–360. Elsevier, 1997.
- Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1):115–144, 2017.
- Su-In Lee, Vassil Chatalbashev, David Vickrey, and Daphne Koller. Learning a meta-level prior for feature relevance from multiple related tasks. In *Proceedings of the 24th international conference on Machine learning*, pages 489–496. ACM, 2007.
- O V Lepski. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35:454–466, 1991.
- O V Lepski. Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates. *Theory of Probability & Its Applications*, 36:682–697, 1992.
- O V Lepski. Asymptotically minimax adaptive estimation. II: Schemes without optimal adaptation: Adaptive estimators. *Theory of Probability & Its Applications*, 37:433–448, 1993.
- Oleg V Lepski and Vladimir G Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546, 1997.
- Jingbo Liu. A few interactions improve distributed nonparametric estimation, optimally,

2021.

- Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of The 22nd Annual Conference on Learning Theory (COLT 2009)*, Montréal, Canada, 2009. URL <http://www.cs.nyu.edu/~mohri/postscript/nadap.pdf>.
- Brendan McMahan and Daniel Ramage. Federated learning: Collaborative machine learning without centralized training data. *Google Research Blog*, 3, 2017.
- Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pages 125–134, 2015.
- Yves Meyer. *Wavelets and operators*, volume 1. Cambridge university press, 1992.
- Joseph Michalowicz, Jonathan Nichols, and Frank Bucholtz. Calculation of differential entropy for a mixed gaussian distribution. *Entropy*, 10(3):200–206, 2008.
- Michael Nussbaum. Asymptotic equivalence of density estimation and gaussian white noise. *The Annals of Statistics*, 24(6):2399–2430, 1996.
- Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.
- Boris Polyak. New stochastic approximation type procedures. *Avtomatica i Telemekhanika*, 7:98–107, 01 1990.
- Maxim Raginsky. Strong data processing inequalities and ϕ -sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016.
- Richard J Samworth. Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763, 2012.
- Carla Savage. A survey of combinatorial gray codes. *SIAM review*, 39(4):605–629, 1997.
- Clayton Scott. A generalized neyman-pearson criterion for optimal domain adaptation. *arXiv preprint arXiv:1810.01545*, 2018.
- Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In *Advances in Neural Information Processing Systems*, pages 163–171, 2014.
- Claude Elwood Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

- Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.
- Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, pages 1433–1440, 2008.
- Botond Szabó and Harry van Zanten. Adaptive distributed methods under communication constraints. *arXiv preprint arXiv:1804.00864*, 2018.
- Botond Szabó and Harry van Zanten. An asymptotic analysis of distributed nonparametric methods. *Journal of Machine Learning Research*, 20(87):1–30, 2019.
- Botond Szabó and Harry van Zanten. Distributed function estimation: Adaptation using minimal communication. *arXiv preprint arXiv:2003.12838*, 2020.
- Botond Szabó, Lasse Vuursteen, and Harry van Zanten. Optimal distributed testing in high-dimensional gaussian models. *CoRR*, abs/2012.04957, 2020. URL <https://arxiv.org/abs/2012.04957>.
- Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.
- Alexey Tsymbal. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2):58, 2004.
- Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.
- Brendan van Rooyen and Robert C. Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018. URL <http://jmlr.org/papers/v18/16-315.html>.
- Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.
- Yu Xiang and Young-Han Kim. Interactive hypothesis testing against independence. In *2013 IEEE International Symposium on Information Theory*, pages 2840–2844. IEEE, 2013.
- Yi Yao and Gianfranco Doretto. Boosting for transfer learning with multiple sources. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1855–1862. IEEE, 2010.

- Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. A survey of crowdsourcing systems. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 766–773. IEEE, 2011.
- Yuchen Zhang, John Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pages 2328–2336, 2013a.
- Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, 14(1):3321–3363, 2013b.
- Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems*, pages 1260–1268, 2014.
- Zhen Zhang and Toby Berger. Estimation via compressed information. *IEEE transactions on Information theory*, 34(2):198–211, 1988.
- Yuancheng Zhu and John Lafferty. Distributed nonparametric regression under communication constraints. *arXiv preprint arXiv:1803.01302*, 2018.