

國立臺灣大學電機資訊學院電信工程學研究所

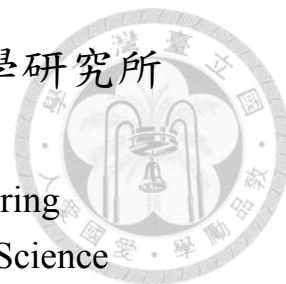
碩士論文

Graduate Institute of Communication Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis



匿名統計推論的理論極限：異質性感測網路與群眾分
包之隱私考量

Fundamental Limits of Anonymous Statistical Inference :
Privacy-Preserving Crowdsourcing and Heterogeneous
Sensor Networks

陳偉寧

Wei-Ning Chen

指導教授：王奕翔博士

Advisor: I-Hsiang Wang, Ph.D.

中華民國 107 年 7 月

July, 2018

國立臺灣大學碩士學位論文
口試委員會審定書

匿名統計推論的理論極限：異質性感測網路與群眾分包
之隱私考量

Fundamental Limits of Anonymous Statistical Inference:
Privacy-Preserving Crowdsourcing and Heterogeneous
Sensor Networks

本論文係陳偉寧君（R05942078）在國立臺灣大學電信工程學研究所完成之碩士學位論文，於民國 107 年 7 月 5 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

王真翔

（簽名）

（指導教授）

陳偉寧

洪榮文

所 長

吳宗霖

（簽名）

誌謝



首先，我要感謝我的指導老師王奕翔教授在碩班兩年、大學專題兩年給我的指導。老師所教的絕不僅止於理論上的知識，還有許多學術、研究上所需的能力，例如構想一個有趣的問題、撰寫期刊會議論文、以及準備並組織報告。在這許多方面，老師會先讓我們自己嘗試，再給予意見讓我們調整修正。如果我在這幾個方面有所長進，那都要感謝老師這幾年所給予的教誨。

再來，我要感謝台科大的林士駿老師及台大電機系的蘇柏青、陳和麟及林茂昭老師，在我的研究成果被 ISIT 會議所接受時願意幫我寫推薦信，讓我的發表得以順利成行。我要特別感謝林士駿老師，知道我有意申請國外 PhD，特別介紹了幾位國際上知名的消息理論教授給我認識，也對於我研究方向提供不少實質上的建議。

我也要感謝清華大學的洪樂文教授與交通大學的陳柏寧教授，除了擔任我的口試委員以外，也給了我的研究成果不少肯定，同時提出許多有趣的方向，希望我可以在往後的研究生涯中朝這些方向探索。

感謝實驗室的眾多成員，特別是與我同屆的宗毅、致伊，許多有趣的問題與解法都是在與大家討論的過程中逐漸成形，同時在研究上遇到的一些數學難題也都可以再遇你們討論過後找到一些可行的方向，希望這樣的實驗室風氣可以一直延續下去。

最後，我要感謝我的家人及秀如，雖然可能不是非常了解我確切的研究在做什麼，仍然願意無怨無悔的支持我，讓我能持續朝自己有興趣的研究方向前進。同時也在我趕會議、學位論文時給了我許多的包容與協助。

摘要

在群眾分包的問題中，平台將許多子問題（例如影像標記）分派給不同的工人，並搜集他們的答案。顧及工人的隱私，搜集的資料通常為匿名的，讓平台在進行最後的決策與估計變得十分困難。在此論文中，我們提出了兩種解決方法，首先是以測試問題推估工人的可靠性，而第二種則是匿名假說檢定。

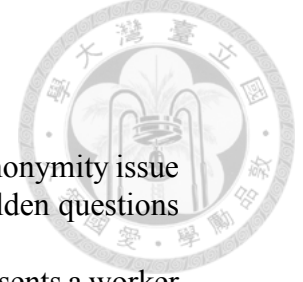
考慮一個大小為 n 的資料庫，其中每筆資料紀錄了一個工人的群組（依據其可靠度所分群），而總共的群組個數為有限的。測試問題讓我們能區分不同群組，也就是說，對於不同類別的工人，他們所標記的答案是不同的。我們可以選擇其中一部份的工人記錄其標記的統計結果（亦即該部分工人當中，每個群組所佔的人數），而我們希望以最少的測試問題 T_n^* 將每個工人確切的類別復原出來。然而，在實際情況中，該統計結果往往是不精確、有雜訊的。我們假定雜訊的大小為 δ_n ，而我們的目標是希望復原錯誤的人數少於 k_n 。在這個問題中，我們主要的貢獻有以下兩點：首先，當 $\delta_n = O(\sqrt{k_n})$ 時，需要大約 $T_n^* = \Theta(n/\log n)$ 數量級的測試問題，可以將工人的類別回復。對於達到上界的驗算法，我們採取“隨機取樣”的想法，將每筆測試問題以二分之一的機率隨機送給每個工人。而下界的部分，則是使用“填裝”的想法構造一個不等式。再來對於 $\delta_n = \Omega(k_n^{(1+\epsilon)/2})$ （其中任意 $\epsilon > 0$ ）的情形，我們證明了至少需要 $T_n^* = \omega(n^p)$ 的測試問題，才足以達成我們的目標。對於這種雜訊強度較高的情形，我們發展了一套新的下界方式，並使用了一些組合最佳化的技巧。

在第二部分，我們考慮匿名假說檢定，在其中不同群組的工人們有著不同的標記能力，因此其答案有著不同的機率分佈。困難之處在於，雖然我們知道每個群組的確切人數，但我們並不知道每筆標記資料是來自哪個群組，因此我們將此問題約化為“多重假說檢定”。在這部分，我們的第一個貢獻是提出了最佳的測試，可寫為在不同假說底下的混合分佈之概似比檢定。第二個貢獻則是刻劃出在樣本數 n 趨近於無限大時，在 Neyman-Pearson setting 底下第二類錯誤趨近於零的速度。該收斂速度的冪次方在機率分布空間之上定義了一個廣義的賦距，同時此結果可以被推廣到 Chernoff regime。

我們的結果量化了匿名性對群眾分包問題所造成的效應，並且此理論工具可應用於許多不同問題，諸如感測網路、物聯網與資訊系統等等。

關鍵字：匿名、隱私、群眾分包、感測網路、資料解碼、假說檢定

Abstract



In this thesis, we propose two treatments to overcome the anonymity issue in privacy-preserving crowdsourcing: group recovery with golden questions and anonymous hypothesis testing.

Consider a data set comprising n items, each of which represents a worker and his or her group index (based on the worker's labeling reliability). The golden questions enable us to *distinguish* disparate group of crowds, that is, workers from different group will respond different answers. By sequentially allocating the golden tasks to a subset of crowds, the fusion center will obtain the histogram of the queried subset. The (unnormalized) histogram, however, is perturbed by some additive noise with magnitude less than δ_n . The goal is to reconstruct the data set such that the Hamming distance between the reconstructed and the actual one is smaller than a tolerance parameter k_n . We are interested in the fundamental limit on the minimum number of queries T_n^* required to recover the n -worker data set within k_n tolerance subject to δ_n noisy perturbation.

We first show that if $\delta_n = O(\sqrt{k_n})$, the minimum query complexity $T_n^* = \Theta(n/\log n)$, where the achievability is based on random sampling, and the converse is based on a packing arguments. On the other hand, if $\delta_n = \Omega(k_n^{(1+\epsilon)/2})$ for some $\epsilon > 0$, we prove that $T_n^* = \omega(n^p)$ for any positive integer p . In other words, no querying methods with $\text{poly}(n)$ query complexity can successfully reconstruct the data set in that regime. This impossibility result is established by a novel combinatorial lower bound on T_n^* .

For the second remedy, anonymous hypothesis testing, each of item in the data set corresponds to a response from a single worker. The fusion center aims to detect a binary parameter by querying the database. The workers are clustered into multiple groups, and hence the responses in different groups follow different distributions under a given hypothesis. The key challenge is the anonymity of the responses – although it knows the exact number of workers n and the distribution of observations in each group, it does not know which group each worker belongs to. It is hence natural to consider it as a composite hypothesis testing problem.

First, we propose an optimal test called *mixture likelihood ratio test*, which is a randomized threshold test based on the ratio of the uniform mixture of all the possible distributions under one hypothesis to that under the other hypothesis. Second, we focus on the Neyman-Pearson setting and specify the error exponent of the worst-case type-II error probability as n tends to infinity, assuming the number of workers in each group is proportional to n . The exponent defines a new divergence between two vector distributions. The results are extended to Chernoff regime by solving a convex optimization problem. Our results elucidate the price of anonymity in anonymous detection.

The developed theories and tools are not restricted to crowdsourcing problem, but can be widely applied to various tasks, such as wireless sensor networks, Internet of Thing, or information retrieval system, etc..

Keywords: Anonymity, Crowdsourcing, Data Decoding, Hypothesis Testing





Contents

誌謝	iii
摘要	iv
Abstract	v
1 Introduction	1
1.1 Privacy-preserving Crowdsourcing	1
1.2 Group Recovery with Golden Questions	3
1.3 Anonymous Hypothesis Testing	5
1.4 Beyond Neyman-Pearson Regime: Anonymous Hypothesis Testing for Bayesian Formulation	8
1.5 Organization of the Thesis	9
2 Background	11
2.1 Basic Information Theory	11
2.1.1 Kullback-Leibler Divergence	11
2.1.2 Method of Types	12
2.2 Concentration Inequalities and Large Deviation	13
2.3 Hypothesis Testing	14
2.3.1 Simple Hypothesis Testing	15
2.3.2 Asymptotic Regime	16
2.3.3 Composite Hypothesis Testing	18



I Group Recovery with Golden Tasks

21

3 Data Extraction with Presence of Noise

23

3.1	Previous Work	23
3.2	Problem Formulation	24
3.2.1	Data Set, Queries, and Responses	25
3.2.2	Criteria of Data Extraction	25
3.3	Main Results	27
3.3.1	Achievability	27
3.3.2	Lower Bounds on Query Complexity	27
3.3.3	Fundamental Limit	28
3.4	Achievability via Randomized Querying	29
3.5	Proof of the Combinatorial Lower Bound	32
3.6	Extension	35

II Anonymous Hypothesis Testing

37

4 Anonymous Hypothesis Testing : Optimal Decision Rules and Type-II Error

Exponents		39
4.1	Previous Work	39
4.2	Problem Formulation	41
4.2.1	Problem Setup	41
4.2.2	Notations	43
4.3	Main Results	44
4.3.1	Main Contributions	45
4.3.2	Numerical Evaluations	47
4.3.3	Distributed Detection with Byzantine Attacks	48
4.4	Proof of Theorem 4.3.1	49
4.5	Proof of Theorem 4.3.2	56
4.6	Extension	62

5	Anonymous Hypothesis Testing : Beyond Neyman-Pearson Regime	65
5.1	Problem Formulation	65
5.2	Main Result	66
5.2.1	Efficient Test	66
5.2.2	Achievable Region: An Information Geometric Perspective . . .	67
5.3	Proof of Theorem 5.2.1	68
5.4	Proof of Theorem 5.2.2	72
6	Conclusions and Future Work	79
6.1	Future Works	80
A	Proof of Lemmas in Chapter 3	83
A.1	proof of Theorem 3.3.2	83
A.2	proof of Theorem 3.3.4	85
A.3	Proof of Claim 3.4.1	87
A.4	Technical Lemmas	88
A.4.1	Lemma A.4.1	88
A.4.2	Proof of Lemma A.3.1	91
A.4.3	Proof of Lemma 3.4.1	92
A.4.4	Proof of Lemma A.2.1	93
A.4.5	Proof of Lemma A.2.2	94
B	Proof of Lemmas in Chapter 5	97
B.1	Proof of Proposition 4.3.1	97
B.2	Proof of Lemma 4.5.1	99
B.3	Proof of Lemma 4.5.2	100
	Bibliography	105





List of Figures

1.1	Crowdsourcing Framwork	2
4.1	Price of anonymity	48
4.2	Comparison between i.i.d. and our setting	49
4.3	Illustration of the auxiliary space	50
5.1	Illustration of achievable region \mathcal{R}	68
5.2	Illustration of $B_r(\cdot)$ and r^*	69
5.3	Relation between \mathcal{A} , \mathcal{A}^c and $B_{r^*}(\mathbf{P}_0)$, $B_{r^*}(\mathbf{P}_1)$	70
5.4	Acceptance region of ϕ_{F^*}	73
5.5	Illustration of L_β and $L_{\beta'}$	76
6.1	Exponents with Partial Information	82





Chapter 1

Introduction

Information retrieval from large-scale data sets plays a crucial role in many fields including data mining, machine learning, Internet of Thing (IoT), etc.. However, as the amounts of data expands exponentially, nearly all of it carries someone's digital fingerprint, which might cause severe disclosure of personal information. One of a common approach to overcome the problem is to restrict the released data from data center being *anonymous*. This motivates us to study the impact of *anonymity* on information retrieval processes. Specifically, in this thesis we focus on the topic of *privacy-preserving crowdsourcing*, studying how anonymity deteriorates the performance of crowdsourcing and how to cope with it. However, we emphasize that the developed theories and tools are not restricted to crowdsourcing problem, but can be widely applied to various fields, such as wireless sensor networks, Internet of Thing, or information retrieval system, etc..

1.1 Privacy-preserving Crowdsourcing

In past decades, the great success on machine learning is attributed to the large-scale well-labeled data sets. Efficiently obtaining accurate labels turns out to be a key to developing machine learning models. Crowdsourcing provides a solution to collect labels by dividing the tasks into numbers of micro-tasks, and then distributes micro-tasks to different crowds. Typically each micro-task is simple, so crowds can solve it rapidly with relatively cheap cost.

However, qualities of each crowd varies greatly as many studies on crowdsourcing suggesting. For example, [25] pointed out that crowds can be tendentious workers who response with certain bias, spammers who always annotate data randomly and independently with the given task, or even adversaries who label data in a malicious way in order to paralyze the system. In general, crowds are quantified into different levels or groups according to their ability, and this quantification may depend on their backgrounds such as credit record, educational level, or the acceptance rate of previous tasks. Because of the heterogeneity among crowds, if we collect the results directly from each worker's answer, it will be very inaccurate and with low utility. A common approach is to allocate a same task to various workers instead of simply one, and then estimate the result according to the the collected answers such as majority votes.

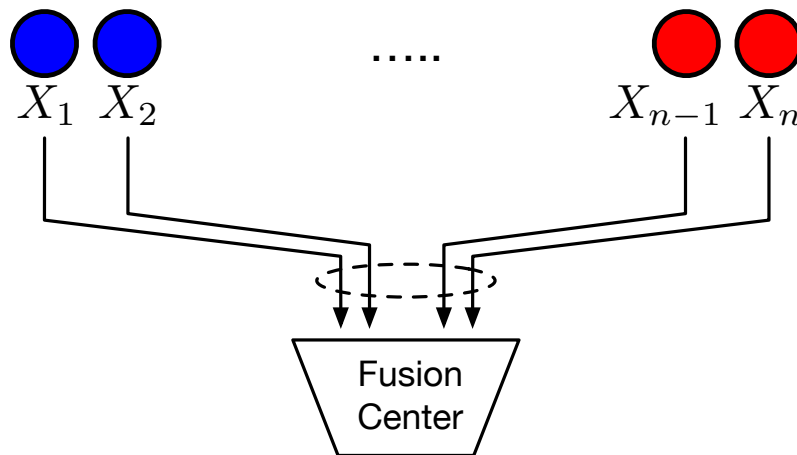


Figure 1.1: Crowdsourcing Framwork

The hardness of this problem lies in that typically the platform (or the fusion center) do not know the group each worker belongs to, although this group information could help requesters estimate true labels more precisely. The reason is that the quantification usually comprises sensitive information such as the crowds' ages, genders, or educational levels, and may severely invade their privacy. Besides, if the number of possible groups is large, then it costs too much communication budget to identify it.

To address the anonymity issue, we propose two treatments. The first one is *estimating the group information with the help of golden tasks*, where golden tasks are tasks with given answers and are used to test the group each worker belongs to. We will elaborate

how golden tasks are employed to efficiently recover the group information in the next subsection. The second treatment is *testing the hypothesis anonymously*. Since the ultimate goal of crowdsourcing is to obtain accurate and reliable labeled data, the group information of each worker is indeed ignorable once we can estimate the true label with small error. Therefore, we seek to design an decision rule to test the hypothesis anonymously with minimum detection error.

1.2 Group Recovery with Golden Questions

The golden tasks (or golden questions) possess the following two characteristics: given the golden tasks, the response from each worker is almost deterministic (with noise small enough, which will be rigorous elaborated in the problem formulation), and the answers from different group of workers are different, that is, these golden questions allow us to *distinguish* disparate group of crowds.

However, the fusion center can only collect the answers in an anonymous query, termed *histogram query*, where each query is a subset of workers, and the response is the histogram (the number of workers belonging to each group) of the corresponding workers. For example, in [19], histogram query (or called *counting query* for binary group) is studied and analyzed as a privacy-preserving database model. Besides, the collected responses may be further perturbed by some noise. We aim to characterizing the fundamental limit on the number of queries (termed *query complexity*) required to recover the group information. In general, we can cast the group recovery problem into a data extraction problem: the group information of each worker can be regarded as a data set, where each item corresponds to a worker, and the value of each item is the group each worker belongs to. The goal is to recover the data set according to the queried answers.

Characterizing the fundamental limit on the number of queries (termed *query complexity*) required to extract the data set is important to both data analysts and data curators. In [43], the fundamental limit on the minimum query complexity to precisely extract the entire n -item data set with noiseless histogram queries is characterized. The optimal query complexity was shown to be $\Theta(n/\log n)$, where n is the size of the data set. Moreover,

an explicit construction of the querying method achieving the optimal query complexity is proposed. However, for the general setting where the goal is to *partially* extract the data set with *noisy* query responses, the characterization of the optimal query complexity remains open.

In this work, we investigate the optimal query complexity T_n^* for partial data extraction with noisy responses to histogram queries. The response from the curator is the actual unnormalized histogram of the queried subset of items, perturbed by an additive noise with maximum magnitude δ_n . The goal of the analyst is to reconstruct the data set partially so that the Hamming distance between the reconstructed and the actual data set is at most k_n .

Our main contribution is characterizing the asymptotic behavior of T_n^* with respect to the size of the data set n and the two parameters k_n, δ_n coupled with n :

- 1) In the regime $\delta_n = O(\sqrt{k_n})$, $T_n^* = \Theta(n/\log n)$, which is the same as the optimal query complexity for perfect reconstruction with noiseless responses to queries [43].
- 2) In the regime $\delta_n = \Omega(k_n^{(1+\epsilon)/2})$ for some $\epsilon > 0$, $T_n^* = \omega(n^p)$ for any positive integer p .

In words, there does not exist querying methods with $\text{Poly}(n)$ query complexity.

For proving the achievability part (upper bound on T_n^*), *randomized* querying is employed. In each query, the items to be included in the queried subset are randomly and uniformly selected. An upper bound on the probability of failure to distinguish two different data sets is then proved, showing that if $\delta_n = O(\sqrt{k_n})$, $\Omega(n/\log n)$ such queries ensure vanishing probability of failure. For proving the converse part (lower bound on T_n^*), we first show that $T_n^* = \Omega(n/\log n)$ based on a packing argument, extending the proof in [43] to general δ_n, k_n . We then develop a novel combinatorial lower bound on T_n^* and show that if $\delta_n = \Omega(k_n^{(1+\epsilon)/2})$ for some $\epsilon > 0$ then no method with polynomial query complexity can reconstruct the data set within Hamming distance of k_n .

Finally, we emphasize that our results for group recovery also apply to various problems, such as pooled data decoding, integer group testing, or coin weighing problem.

1.3 Anonymous Hypothesis Testing

Our second treatment for privacy-preserving crowdsourcing problem is to test hypothesis anonymously. The crowdsourcing problem is directly related to the wireless sensor networks. Each worker can be regarded as a sensor, and observations of sensors under a given hypothesis represents the workers' labeling outputs for a specific tasks. Hence the results from wireless sensor networks directly apply.

For distributed detection in wireless sensor networks [40], when the observations follow identically and independently distributed (i.i.d.) distributions across all sensors, identifying individual sensors is not very important. When the fusion center can fully access the observations, the empirical distribution (types) of the collected observation is a sufficient statistics. When the communication between each sensor and the fusion center is limited, for binary hypothesis testing it is asymptotically optimal to use the same local decision function at all sensors [39]. Hence, anonymity is not a critical issue for the classical (homogeneous) distributed detection problem.

However, when the joint distribution of the workers' responses is *heterogeneous*, that is, the marginal distribution of the response varies across workers, anonymity (unknowing the group information about each worker) may deteriorate the performance of distributed detection, even for binary hypothesis testing. One such example is distributed detection under Byzantine attack [30], where a fixed number of workers are compromised by malicious attackers and report fake responses following certain distributions. Even if the fusion center is aware of the number of compromised workers and the attacking strategy that renders worst-case detection performance (the least favorable distribution as considered in [23, 24, 41]), it is more difficult to detect the hidden parameter when the fusion center does not know which workers are compromised.

In the second part of this thesis, we aim to find the optimal decision rule, and simultaneously quantify the performance loss due to anonymity in heterogeneous distributed detection, with n workers and a single fusion center. Each worker (say worker i , $i \in \{1, \dots, n\}$) has a single random labeling response X_i . The goal of the fusion center is to estimate the hidden parameter $\theta \in \{0, 1\}$ (that is, binary hypothesis testing) from the collected

observations. The distributions of the responses, however, are *heterogeneous* – responses at different workers may follow different sets of distributions. In particular, we assume that these n sensors are clustered into K groups $\{\mathcal{I}_1, \dots, \mathcal{I}_K\}$, and group $\mathcal{I}_k \subseteq \{1, \dots, n\}$ comprises $n\alpha_k$ workers, for $k = 1, \dots, K$. Under hypothesis \mathcal{H}_θ , $\theta \in \{0, 1\}$,

$$X_i \sim P_{\theta;k}, \text{ for } i \in \mathcal{I}_k.$$

Moreover, the workers are *anonymous*, that is, the collected answers at the fusion center is *unordered*. In other words, although the fusion center is fully aware of the *heterogeneity* of its responses, including the set of distributions $\{P_{\theta;k} \mid \theta \in \{0, 1\}, k = 1, \dots, K\}$ and $\{\alpha_k \mid k = 1, \dots, K\}$, it does not know what distribution each individual response will follow.

To address the lack of knowledge about the exact distributions of the observations, we formulate the detection problem as a *composite hypothesis testing* problem, where the vector response of length n follows a product distribution within a finite class of n -letter product distributions under a given parameter θ . The class consists of $\binom{n}{n\alpha_1, \dots, n\alpha_K}$ possible product distributions, each of which follows one of the $\binom{n}{n\alpha_1, \dots, n\alpha_K}$ possible partitions of the workers. The fusion center takes all the possible partitions into consideration when detecting the hidden parameter. We mainly focus on a Neyman-Pearson setting, where the goal is to minimize the worst-case type-II error probability such that the worst-case type-I error probability is not larger than a constant. Towards the end of this part, we also extend our results to a Bayesian setting, where a binary prior distribution is laid on \mathcal{H}_0 and \mathcal{H}_1 .

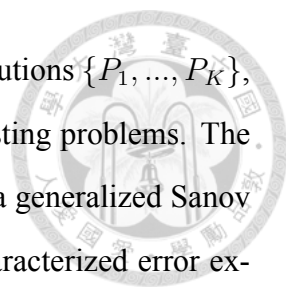
Our main contribution comprises three parts. First, we develop an optimal test, termed *mixture likelihood ratio test* (MLRT), for the anonymous heterogeneous distributed detection problem. MLRT is a randomized threshold test based on the ratio of the uniform mixture of all the possible distributions under hypothesis \mathcal{H}_1 to the uniform mixture of those under \mathcal{H}_0 . To prove the optimality, we first argue that there exists an optimal test that is *symmetric*, that is, it does not depend on the order of observations across the sensors, and thus we only need to consider tests which depend on the histogram of responses. In other words, the histogram of observations contains sufficient information for optimal

detection. Moreover, all possible distributions over the space of observations \mathcal{X}^n under \mathcal{H}_0 (or \mathcal{H}_1) turn out to be the same one over the space of its histogram, so if we test the hypotheses according to the histogram, the original composite hypothesis testing problem boils down to a simple hypothesis testing problem. The one-to-one correspondence between symmetric tests and tests defined on the histogram is the key to derive optimal test. This result extends to M -ary hypothesis testing with heterogeneous observations generated according to hidden latent variables, each of which is associated to a observation, but the decision maker only knows the histogram of the latent variables.

Second, for the case that the alphabet \mathcal{X} is a finite set, we characterize the error exponent of the minimum worst-case type-II error probability as $n \rightarrow \infty$ with $\{\alpha_k \mid k = 1, \dots, K\}$ being fixed. The optimal error exponent turns out to be the minimization of a linear combination of KL divergences with the k -th term being $D(U_k \parallel P_{1;k})$ and α_k being the coefficient, for $k = 1, \dots, K$. The minimization is over all possible distributions U_1, \dots, U_K such that $\sum_{k=1}^K \alpha_k U_k = \sum_{k=1}^K \alpha_k P_{0;k}$. In a simple hypothesis testing problem with i.i.d. responses, a standard approach to derive the type-II error exponent is invoking a strong converse lemma (see, for example, Chapter 12 in [32]) to relate the type-I and type-II error probability, and then applying the large deviation toolkit on the optimal test to single-letterize and find the exponent. In contrast, in our problem, neither can the mixture distributions in the optimal test be decomposed into a product form, nor can the acceptance region be bounded by a large deviation event, making this approach fail to characterize the error exponent. To circumvent the difficulties, we turn to method of types and use bounds on types (empirical distributions) for single-letterization. Intuitively, the exponent measures “how close” the two hypothesis classes \mathcal{H}_0 and \mathcal{H}_1 are from each other, as the role that KL-divergence plays in simple hypothesis testing problems.

For achievability, instead of the optimal MLRT which is difficult to single-letterize, we employ a simpler test that resemble Hoeffding’s test [22]. For the converse, we use an argument based on the method of types. We propose a generalized divergence

$$D_{\alpha_1, \dots, \alpha_K}(P_1, \dots, P_K; Q_1, \dots, Q_K)$$



from a group of distributions $\{Q_1, \dots, Q_K\}$ to another group of distributions $\{P_1, \dots, P_K\}$, which plays a similar role as KL divergence in simple hypothesis testing problems. The key to the characterization of the optimal error exponent is to prove a generalized Sanov Theorem for the composite setting we considered. Based on the characterized error exponent, given the number of bits that a sensor can send to the fusion center, one can also formulate an optimization problem to find the best local decision functions, as in the homogeneous case [39].

As a by-product, we apply our results for $K = 2$ to the distributed detection problem under Byzantine attack and further obtain bounds on the worst-case type-II error exponent. Compared with the worst-case exponent in an alternative Bayesian formulation [30] where the observation of sensors are assumed to be i.i.d. according to a mixture distribution, it is shown that the worst-case exponent in the composite testing formulation is strictly larger. This hints that the conventional approach taken in [30] might be too pessimistic.

In order to be consistent with the previous literatures, we use languages and notations from wireless sensor networks, where the term *sensors* and *observations* represent *crowds* and *responses*, as stated in previous context. After all, the essence of the two problems are indeed identical.

1.4 Beyond Neyman-Pearson Regime: Anonymous Hypothesis Testing for Bayesian Formulation

Finally, we extend our results from the Neyman-Pearson setting to a Bayesian setting (a.k.a. *Chernoff's regime*), minimizing the average probability of error (that is, combining type-I and type-II error). It can be shown that the optimal test is computationally infeasible, since it involves summation over all possible permutations. To overcome the complexity issue, we propose an asymptotically optimal test based on information geometry, which achieves the same error exponent of the average probability of error. We also study the exponent region \mathcal{R} , the collection of all pairs of achievable type-I and type-II error exponents. In particular, we propose a way to parametrize the contour of \mathcal{R} based on

information projection. However, the closed-form expression of \mathcal{R} involves an explicit solution of a convex optimization problem, which remains unsettled.



1.5 Organization of the Thesis

The thesis comprises two parts. The contents of Part I are published in [10], and part of Part II are published in [9] and [12].

In Chapter 2, we set up some basic knowledge, including hypothesis testing, large deviation theory, and other information theoretical bounds, which serve as the main tools for this work. In Chapter 3, we first review some literatures on group testing and pooled-data decoding, summarize their formulation and results, and formally formulate the data extraction (group recovery with golden sample) problem. Fundamental limits on high SNR are established by setting both achievability and converse results; for low SNR regime, we demonstrate the impossibility to recover the group information.

In Chapter 4, we investigate the anonymous hypothesis testing problem. First we revisit previous works related to hypothesis testing, distributed detection, and Byzantine attack, and then we mathematically formulate the anonymous detection problem into a composite hypothesis testing, focusing on Neyman-Pearson regime. As our first contribution in this part, optimal decision rule will be fully characterized. Then, we move on to the asymptotic regime, specify the type-II error exponent, which can be expressed in a generalized divergence between the distributions under each hypothesis. Some properties of exponent will be discussed. We also apply our results under Byzantine attack, and compare with related works. In Chapter 5, we extend our results to Bayesian setting, where an asymptotically optimal, computational efficient test will be given. Then we show that to depict the region of all achievable type-I and type-II error exponents, it suffices to solve the information projection problem with respect to the generalized divergence proposed in previous chapter. However, the closed expression is still an open problem.

Finally, in Chapter 6, we summarize our analysis on privacy-preserving crowdsourcing problem and propose a scheme to overcome the anonymity issue. We closing this thesis by providing some ambitious but interesting directions, such as partial group recover or

optimal crowds clustering.





Chapter 2

Background

In this chapter, we review some useful tools from information theory.

2.1 Basic Information Theory

2.1.1 Kullback-Leibler Divergence

Let P, Q be two probability distributions defined on space \mathcal{X} .

Definition 2.1.1 (Kullback-Leibler Divergence) *The Kullback-Leibler divergence (KL divergence) is defined as*

$$D(P \parallel Q) \triangleq \begin{cases} \mathbb{E}_P \left[\log \frac{P(X)}{Q(X)} \right], & \text{if } P \ll Q \\ +\infty, & \text{else.} \end{cases}$$

Several important properties of KL divergence can be found in [32]. Here we list one which will be frequently used in latter chapters:

Property 2.1.1 (Convexity of KL divergence) *The KL divergence is convex. That is, for all distributions P_1, P_2, Q_1, Q_2 , and for all $\lambda \in [0, 1]$, we have*

$$D(\lambda P_1 + (1 - \lambda)P_2 \parallel \lambda Q_1 + (1 - \lambda)Q_2) \leq \lambda D(P_1 \parallel Q_1) + (1 - \lambda)D(P_2 \parallel Q_2).$$



2.1.2 Method of Types

For a sequence $x^n \in \mathcal{X}^n$, where $\mathcal{X} = \{a_1, a_2, \dots, a_d\}$, its type (empirical distribution) is defined as

$$\Pi_{x^n} = [\pi(a_1|x^n), \pi(a_2|x^n), \dots, \pi(a_d|x^n)],$$

where $\pi(a_i|x^n)$ is the frequency of a_i in the sequence x^n , that is,

$$\pi(a_i|x^n) = \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{\{x_j=a_i\}}.$$

For a given length n , we use \mathcal{P}_n to denote the collection of possible types of length- n sequences. In other words,

$$\mathcal{P}_n \triangleq \left\{ \left[\frac{i_1}{n}, \frac{i_2}{n}, \dots, \frac{i_d}{n} \right] \mid \forall i_1, \dots, i_d \in \mathbb{N} \cup \{0\}, i_1 + i_2 + \dots + i_d = n \right\}.$$

Let $U \in \mathcal{P}_n$ be an n -type. The type class $T_n(U)$ is the set of all length- n sequences with type U ,

$$T_n(U) \triangleq \{x^n \in \mathcal{X}^n \mid \Pi_{x^n} = U\}.$$

Let us introduce some useful lemmas about type.

Lemma 2.1.1 (Cardinality Bound of \mathcal{P}_n)

$$|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}.$$

In words, $|\mathcal{P}_n|$ grows polynomial in n .

Lemma 2.1.2 (Probability of Type Class) Let $P \in \mathcal{P}_n, Q \in \mathcal{P}_{\mathcal{X}}$. Then

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(Q\|P)} \leq Q^{\otimes n}(T_n(P)) \leq 2^{-nD(Q\|P)}.$$

For finite \mathcal{X} , $\mathcal{P}_{\mathcal{X}}$ can be viewed as a subspace in \mathbb{R}^d endowed with Euclidean metric and standard topology. The following theorem, developed by Sanov, depicts the probability of a large deviation event.

Lemma 2.1.3 (Sanov's Theorem) *Let $\Gamma \subseteq \mathcal{P}_{\mathcal{X}}$. Then we have*

$$-\inf_{T \in \text{int } \Gamma} D(T \| Q) \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log Q \{x^n : \Pi_{x^n} \in \Gamma\} \quad (2.1)$$

$$\leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log Q \{x^n : \Pi_{x^n} \in \Gamma\} \quad (2.2)$$

$$\leq -\inf_{T \in \text{cl } \Gamma} D(T \| Q), \quad (2.3)$$

where $\text{int } \Gamma$ and $\text{cl } \Gamma$ respectively denote the interior and the closure of Γ , with respect to the standard topology on \mathbb{R}^d . In particular, if the infimum on the right-hand side is equal to the infimum on the left-hand side in (2.1), we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log Q \{x^n : \Pi_{x^n} \in \Gamma\} = -\inf_{T \in \Gamma} D(T \| Q).$$

Proofs of the lemmas mentioned above can be found in standard information theory textbooks, Chapter 11 in [13] for example. Alternatively, a more rigorous proof of Sanov's theorem Lemma 2.1.3 can be found in [14].

2.2 Concentration Inequalities and Large Deviation

In this section, we introduce some concentration inequalities and basic large deviation theory.

Lemma 2.2.1 (Large Deviation Theory) *Let $X_1, X_2, \dots, X_n \stackrel{i.i.d.}{\sim} P$. Then for any $\gamma \in \mathbb{R}$,*

$$\lim_{n \rightarrow \infty} \frac{1}{n} - \log P \left(\frac{1}{n} \sum_{k=1}^n X_k > \gamma \right) = \inf_{Q: \mathbb{E}_Q X > \gamma} D(Q \| P)$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} - \log P \left(\frac{1}{n} \sum_{k=1}^n X_k \geq \gamma \right) = \inf_{Q: \mathbb{E}_Q X \geq \gamma} D(Q \| P).$$

Proof. See Chapter 14 in [32]. ■

Given a distribution P on $\mathcal{X} \subseteq \mathbb{R}$, and let X be a random variable on \mathcal{X} with probability measure P and Borel σ -field \mathcal{B} . Define the *cumulative generating function (CFG)*

as

$$\psi_X(\lambda) = \log (\mathbb{E}_P e^{\lambda X}) .$$

Sometimes we write $\psi_X(\lambda)$ as $\psi_P(\lambda)$ to emphasize that X are distributed according to the measure P .

Lemma 2.2.2 (Information Projection) *Let*

$$\begin{aligned} A &= \inf \psi'_X = \text{essinf} X \triangleq \sup\{a : X \geq aP - a.s.\} \\ B &= \sup \psi'_X = \text{esssup} X \triangleq \sup\{b : X \leq bP - a.s.\}. \end{aligned}$$

The information projection problem over $\mathcal{E} = \{Q : \mathbb{E}_Q X \geq \gamma\}$ has solution

$$\min_{Q: \mathbb{E}_Q X \geq \gamma} D(Q \| P) = \begin{cases} 0, & \gamma < \mathbb{E}_P X \\ \psi_P^*(\gamma), & \mathbb{E}_P X \leq \gamma < B \\ -\log P(X = B), & \gamma = B \\ +\infty, & \gamma > B, \end{cases}$$

where $\psi_P^*(\lambda)$ is the Lagrange conjugate of $\psi_P(\lambda)$.

Proof. See Chapter 14 in [32]. ■

Lemma 2.2.3 (Chernoff Bound) *Let $X \sim P$. Then the following bound holds for all $\lambda \geq 0$:*

$$P(X \geq \gamma) \leq e^{(\lambda\gamma + \psi_X(\lambda))}.$$

2.3 Hypothesis Testing

In this section, we introduce the classical hypotheses testing problem, which will be frequently used to tackle detection problems in WSN. We summarize some important results in simple hypothesis testing and in asymptotic regime. All the proofs of below lemmas can be found in [32].



2.3.1 Simple Hypothesis Testing



Let X be a random variable taken values in alphabet \mathcal{X} , and P, Q be two probability distributions on \mathcal{X} . Two hypotheses are considered:

$$\begin{cases} \mathcal{H}_0 \text{ (Null hypothesis)} : X \sim P \\ \mathcal{H}_1 \text{ (Alternative hypothesis)} : X \sim Q. \end{cases}$$

We test the hypothesis according to a (randomized) test $\phi : \mathcal{X} \rightarrow [0, 1]$. To evaluate a given test ϕ , probability of two types of error events are considered:

$$\begin{cases} \text{False alarm (type-I error): } P_{\text{FA}}(\phi) \triangleq \mathbb{E}_P[\phi(X)] \\ \text{Miss detection (type-II error): } P_{\text{MD}}(\phi) \triangleq \mathbb{E}_Q[1 - \phi(X)]. \end{cases}$$

Depending on applications, the problem of finding 'best' test function can be formulated in two differently approaches:

1. Neyman-Pearson formulation: In Neyman-Pearson regime, we aim to find a test which minimize type-II error while constraining type-I error not greater than a given constant ϵ :

$$\beta^*(\epsilon) \triangleq \min_{\phi} P_{\text{MD}}(\phi), \text{ subject to } P_{\text{FA}}(\phi) \leq \epsilon.$$

2. Bayesian formulation: In Bayesian's regime, prior on $\mathcal{H}_0, \mathcal{H}_1$ are π_0, π_1 respectively.

Our goal is to design a test to minimize the overall probability of error:

$$P_e^* \triangleq \min_{\phi} \pi_0 P_{\text{FA}}(\phi) + \pi_1 P_{\text{MD}}(\phi).$$

It is well known that the randomized likelihood ratio test (LRT) is optimal for both the Neyman-Pearson and Bayesian formulations. The detailed results can be found in [32].



2.3.2 Asymptotic Regime

In asymptotic regime, we consider $X^n = (X_1, X_2, \dots, X_n) \in \mathcal{X}^n$ as the sample size n tends to infinity. The hypotheses we consider are as below:

$$\begin{cases} \mathcal{H}_0 : X^n \sim P^{\otimes n} \\ \mathcal{H}_1 : X^n \sim Q^{\otimes n}, \end{cases}$$

where we use “ $\otimes n$ ” to denote the i.i.d. extension of distributions. The decision function (a.k.a test) ϕ_n is a mapping from \mathcal{X}^n to $[0, 1]$. Thus the probabilities of two error events are

$$\begin{cases} \text{Type-I error: } P_{\text{FA}}^{(n)}(\phi_n) \triangleq \mathbb{E}_{X^n \sim P^{\otimes n}}[\phi_n(X^n)] \\ \text{Type-II error: } P_{\text{MD}}^{(n)}(\phi_n) \triangleq \mathbb{E}_{X^n \sim Q^{\otimes n}}[1 - \phi_n(X^n)]. \end{cases}$$

Neyman-Pearson formulation and Bayesian formulation are considered respectively. In Section 2.3.1, we see that the optimal test is LRT, and it's not hard to show that under asymptotic regime, the error exponent of LRT decays exponentially in sample size n . For Neyman-Pearson problem in large sample setting (a.k.a. Stein's regime), we aim to find the minimum exponent of type-II error, under the constraint that type-I error less than a threshold ϵ :

$$E(\epsilon) \triangleq \lim_{n \rightarrow \infty} -\frac{1}{n} \log \beta^*(n, \epsilon), \quad (2.4)$$

given that the limit exists, where

$$\beta^*(n, \epsilon) \triangleq \min_{\phi_n} P_{\text{MD}}^{(n)}(\phi_n) \text{ subject to } P_{\text{FA}}^{(n)}(\phi_n) \leq \epsilon.$$

Lemma 2.3.1 (Stein's Lemma) *For all $\epsilon \in (0, 1)$, the optimal type-II error exponent does not depend on ϵ , and is characterized by the KL divergence between P, Q :*

$$E(\epsilon) = D(P \| Q), \quad \forall \epsilon \in (0, 1).$$

Note that Stein's lemma ensures that the limit in (2.3.2) exists and is equal to $D(P \| Q)$.

In the Bayesian problem, the goal is to minimize overall probability of error. Under large sample setting (a.k.a. Chernoff's regime), we aim to characterize the exponent of overall probability of error :

$$F \triangleq - \lim_{n \rightarrow \infty} \frac{1}{n} \log P_e^*(n),$$

where

$$P_e^*(n) \triangleq \min_{\phi_n} \pi_0 P_{FA}^{(n)}(\phi_n) + \pi_1 P_{MD}^{(n)}(\phi_n).$$

Lemma 2.3.2 (Chernoff Information) *The optimal error exponent in Bayesian problem is the Chernoff information between distributions P and Q , which is defined as follows:*

$$CI(P, Q) \triangleq \min_{\lambda \in [0,1]} \log \left(\int_{\mathcal{X}} \left(\frac{Q(x)}{P(x)} \right)^\lambda P(dx) \right)^{(-1)}.$$

Remark 2.3.1 *The optimal error exponent, that is, the Chernoff information $CI(P, Q)$ can be achieved by maximum a posteriori (MAP) test, which is LRT with threshold equal $\log(\pi_0/\pi_1)$:*

$$\phi_n(x^n) = \mathbb{1}_{\{\log(Q(x^n)/P(x^n)) \geq \log(\pi_0/\pi_1)\}}.$$

Though LRT is optimal in simple hypothesis testing problem, sometimes it is hard to analyze, especially when the underlying probability measure is not independent. Therefore, we introduce the following sun-optimal test which also achieves the same error exponent as LRT (but is not optimal in finite-sample regime).

Lemma 2.3.3 (Hoeffding's Test) *For the simple hypothesis testing problem, consider the following test:*

$$\phi(x^n) = \mathbb{1}_{\{D(\Pi_{x^n} \| P) > \epsilon'\}},$$

where ϵ' is chosen such that type-I error is less than ϵ , that is,

$$\mathbb{E}_P [\mathbb{1}_{\{D(\Pi_{x^n} \| P) > \epsilon'\}}] \leq \epsilon.$$

Then, the probability of type-II error decreases exponentially in n , with exponent

$$E(\epsilon) = D(P \| Q).$$

Proof. See [22] ■

2.3.3 Composite Hypothesis Testing

In general, the null or the alternative hypothesis may consist of more than one possible distribution, which forms the *composite hypothesis testing* problem. Finding a test which is *uniformly most powerful* (UMP)¹ is usually impossible except for some special cases. See [29] for more details. Therefore, we focus on *minimax* criterion. Below we formally define the minimax problem for composite hypothesis testing.

Consider the problem for testing the following hypothesis:

$$\begin{cases} \mathcal{H}_0 : X \sim P_\theta, \theta \in \Omega \\ \mathcal{H}_1 : X^n \sim P_{\theta'}, \theta' \in \Omega', \end{cases} \quad (2.5)$$

and we assume Ω and Ω' are disjoint. We aim to find the minimax test:

$$\begin{cases} \phi_{NP}^* \triangleq \arg \min_{\phi} \max_{\theta' \in \Omega'} \mathbb{E}_{P_{\theta'}}[1 - \phi(X)], \text{ subject to } \max_{\theta \in \Omega} \mathbb{E}_{P_\theta}[\phi(X)] \leq \epsilon. \\ \phi_{Bayes}^* \triangleq \arg \min_{\phi} \max_{\theta' \in \Omega', \theta \in \Omega} \pi_0 \mathbb{E}_{P_{\theta'}}[1 - \phi(X)] + \pi_1 \mathbb{E}_{P_\theta}[\phi(X)]. \end{cases}$$

The lemma below demonstrates that the probability of errors under minimax criterion are lower bounded by any reduced simple hypothesis testing problems:

Lemma 2.3.4 (Lower Bound on Composite Hypothesis Testing) *Let $\pi(\theta), \pi'(\theta')$ be two arbitrary prior distributions on Ω, Ω' . Consider the following simple hypothesis testing*

¹ A test ϕ^* is UMP, if $\mathbb{E}_{P_\theta}[\phi] \leq \mathbb{E}_{P_\theta}[\phi^*]$ implies $\mathbb{E}_{Q_{\theta'}}[1 - \phi] \geq \mathbb{E}_{Q_{\theta'}}[1 - \phi^*]$, for any θ, θ' .

reduced from the composite one defined in (2.5):

$$\begin{cases} \tilde{\mathcal{H}}_0 : X^n \sim \int_{\Omega} \pi(\theta) P_{\theta} d\theta \\ \tilde{\mathcal{H}}_1 : X^n \sim \int_{\Omega'} \pi'(\theta') P_{\theta'} d\theta', \end{cases}$$



and denote $\tilde{\beta}^*(\epsilon)$, \tilde{P}_e^* as the type-II and overall probability of error in Neyman-Pearson and Bayesian regime respectively. Let $\beta^*(\epsilon)$, P_e^* be the minimax type-II and minimax overall probability of error in composite setting, that is,

$$\begin{cases} \beta^*(\epsilon) \triangleq \min_{\phi} \max_{\theta' \in \Omega'} \mathbb{E}_{P_{\theta'}}[1 - \phi(X)], \text{ subject to } \max_{\theta \in \Omega} \mathbb{E}_{P_{\theta}}[\phi(X)] \leq \epsilon. \\ P_e^* \triangleq \min_{\phi} \max_{\theta' \in \Omega', \theta \in \Omega} \pi_0 \mathbb{E}_{P_{\theta'}}[1 - \phi(X)] + \pi_1 \mathbb{E}_{P_{\theta}}[\phi(X)]. \end{cases}$$

Then the following bounds hold for all prior $\pi(\theta)$, $\pi'(\theta')$:

$$\beta^*(\epsilon) \geq \tilde{\beta}^*(\epsilon),$$

$$P_e^* \geq \tilde{P}_e^*.$$





Part I

Group Recovery with Golden Tasks





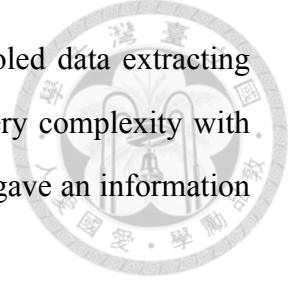
Chapter 3

Data Extraction with Presence of Noise

3.1 Previous Work

As stated in Chapter 1, the golden questions allow us to distinguish different group of crowds, so we can cast the group recovery problem into a categorical data extraction problem: the group information of each worker can be regarded as a data set, where each item corresponds to a worker, and the value of each item is the group each worker belongs to. The goal is to recover the data set according to the queried answers.

Prior work on categorical data extraction with histogram queries for generic alphabet \mathcal{A} was initiated in [43], where the optimal query complexity of exact reconstruction is shown to be $\Theta(n/\log n)$ with noiseless responses, and improved to $\Theta(\frac{k}{\log k} \log \frac{n}{k})$ when the data set is sparse with sparsity level k [42]. Furthermore in [1], upper and lower bounds on the pre-constants in the $n/\log n$ scaling are also proved. In [1, 2], they specified a sharp upper bound on the *rate* of query complexity, and also proposed a computational efficient algorithm (approximate message passing), pointing out there is a gap between information theoretic and algorithmic bounds. Later in [36], converse lower bound on the rate is given, showing that the constant are indeed tight. Besides, [36] also studied the query complexity with presence of *random noise*, and obtained similar behavior like in group testing. However, the considered scenario in [36] is different from us, as they assumed the noise is randomized, while we do not make assumption on the distributions of noise, so it can be in an *adversarial* way.



Independently, in computer science literatures [5, 7, 6], the pooled data extracting problem also termed “coin weighing problem”. [5] studied the query complexity with presence of *erasure error* (which is also different from our setting), gave an information theoretic bound on the query complexity.

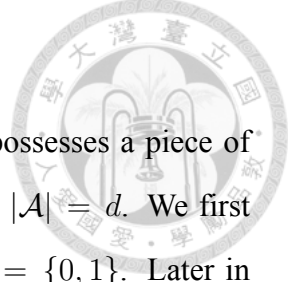
Our problem can also be viewed as generalization of *group testing*[33], where the response is “OR” of the sampled bits while in our setting the response is “SUM” instead. A line of works have taken an information theoretic approach towards group testing problems [4, 3, 37, 27, 35].

Our work is closely related to studies of lower bounds in *data privacy*, where the focus is on deriving conditions on the perturbation level in the response so that no *computationally-efficient* algorithms can reconstruct the private data set from aggregated queries. Binary alphabet ($\mathcal{A} = \{0, 1\}$) is mainly considered in these works. In [19], noisy response to histogram query is proven to be *differential private* with proper perturbation. In [17], it is shown that no algorithm with polynomial running time can reconstruct a constant fraction of the entire data set when the perturbation level $\delta_n = \Omega(\sqrt{n})$. Besides, when $\delta_n = o(\sqrt{n})$, a polynomial-running-time algorithm is given, where the query complexity is $\omega(n)$. In [20], query complexity and running time are improved to n and $\Theta(n \log n)$ respectively.

However, all the reconstruction algorithms [17, 18, 20, 8] aim to recover only a constant fraction of the entire data set ($k_n = \Theta(n)$) with perturbation $\delta_n \approx \sqrt{n}$, and can be viewed as special cases in the regimes considered in this work.

3.2 Problem Formulation

Following [43], we cast the data extraction problem with n items and T_n queries as a linear inverse problem.



3.2.1 Data Set, Queries, and Responses

Consider a data set with n items, labeled from 1 to n . Each item possesses a piece of data which takes value in a finite alphabet $\mathcal{A} = \{a_1, a_2, \dots, a_d\}$ and $|\mathcal{A}| = d$. We first consider the case $d = 2$, and assume without loss of generality $\mathcal{A} = \{0, 1\}$. Later in Section 3.6, it is explained how to extend the results to general d . Let us denote the data set as $\mathbf{x} \in \mathcal{X}$, where \mathcal{X} denotes the collection of all possible realization of data sets. For now, $\mathcal{X} = \{0, 1\}^{n \times 1}$.

To address the partial reconstruction criterion, we use the Hamming distance, formally stated below.

Definition 3.2.1 (Distance between two data sets) *Let $\mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}$ be two data sets with items $[x_1 \dots x_n]^\top$ and $[\tilde{x}_1 \dots \tilde{x}_n]^\top$ respectively. Then, $d_{data}(\mathbf{x}, \tilde{\mathbf{x}}) \triangleq \sum_{j=1}^n \mathbb{1}\{x_j \neq \tilde{x}_j\}$.*

Consider T_n queries, each query being a subset of labels in $[n]$. Let \mathcal{S}_i denote the queried subset in the i -th query. The response to a query is the histogram of the queried subset. We shall use a $T_n \times n$ query matrix $\mathbf{Q} \in \{0, 1\}^{T_n \times n}$ to collectively represent the T_n queries. In particular, $(\mathbf{Q})_{i,j} = 1$ if and only if the j -th item is included in the i -th queried subset. In other words, $(\mathbf{Q})_{i,j} = \mathbb{1}\{j \in \mathcal{S}_i\}$. Hence, the i -th row $\mathbf{q}_i^\top \in \{0, 1\}^{1 \times n}$ represents the queried subset in the i -th query. The responses to the queries can then be represented as the multiplication of the query matrix and the data-set matrix (here, it is an $n \times 1$ matrix). It is not hard to see that the *unnormalized* response to the i -th histogram query $y_i = \mathbf{q}_i^\top \mathbf{x} \in [n]$ and hence $\mathbf{y} = \mathbf{Q}\mathbf{x} \in \{0, 1, \dots, n\}^{T_n \times 1}$.

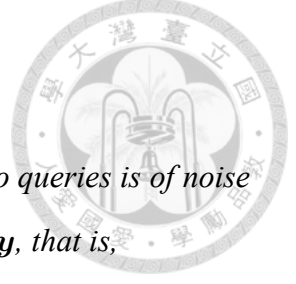
To address the perturbation in the responses, we use the ℓ_∞ norm, formally stated below.

Definition 3.2.2 (Distance between two response) *Suppose $\mathbf{y}, \tilde{\mathbf{y}}$ are the responses to two queries. The distance between them is defined as $d_{response}(\mathbf{y}, \tilde{\mathbf{y}}) \triangleq \max_i |y_i - \tilde{y}_i|$.*

3.2.2 Criteria of Data Extraction

Definition 3.2.3 (Tolerance in Partial Extraction) *The data extraction task is k -tolerable, if the reconstructed data set $\tilde{\mathbf{x}}$ differs from the original one \mathbf{x} by at most k , that is,*

$$d_{data}(\mathbf{x}, \tilde{\mathbf{x}}) \leq k, \forall \mathbf{x} \in \mathcal{X}.$$



Definition 3.2.4 (Noise Level in Perturbed Response) Responses to queries is of noise level δ if the perturbed response $\tilde{\mathbf{y}}$ has distance at most δ to original \mathbf{y} , that is,

$$d_{response}(\mathbf{y}, \tilde{\mathbf{y}}) \leq \delta.$$

The goal of the data analyst is to design the query matrix \mathbf{Q} to extract the data set \mathbf{x} within distance k_n from the δ_n perturbed response $\tilde{\mathbf{y}}$. Formally, \mathbf{Q} has to satisfy the following:

$$\begin{aligned} \forall \mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}, d_{data}(\mathbf{x}, \tilde{\mathbf{x}}) > k_n \\ \implies d_{response}(\mathbf{Q}\mathbf{x}, \mathbf{Q}\tilde{\mathbf{x}}) > 2\delta_n. \end{aligned} \quad (3.1)$$

Definition 3.2.5 (Recoverability) Suppose a query matrix $\mathbf{Q} \in \{0, 1\}^{T_n \times n}$ satisfies (3.1) with respect to tolerance k_n and noise level δ_n , then it is called (T_n, k_n, δ_n) -recoverable.

Definition 3.2.6 (Optimal Query Complexity) $T_n^*(k_n, \delta_n)$ denotes the minimum query complexity for reconstructing a n -item data set with tolerance k_n under noise level δ_n , that is,

- There exists a \mathbf{Q} which is (T_n^*, k_n, δ_n) -recoverable.
- For all $T_n < T_n^*$, there does not exist query matrix \mathbf{Q} which is (T_n, k_n, δ_n) -recoverable.

For a randomized querying method, the query matrix \mathbf{Q} is randomly selected from distribution $P_{\mathbf{Q}}$. To specify the criterion of successfully extracting the data set under randomized querying, let us define the *probability of failure* as follows:

Definition 3.2.7 (Probability of Failure) For a data set $\mathbf{x} \in \mathcal{X}$, the probability of failure $P_f(\mathbf{x}; k_n, \delta_n)$ with respect to the randomly generated query matrix \mathbf{Q} is defined as

$$P_{\mathbf{Q}} \{ \exists \tilde{\mathbf{x}}, d_{data}(\tilde{\mathbf{x}}, \mathbf{x}) > k_n, d_{response}(\mathbf{Q}\tilde{\mathbf{x}}, \mathbf{Q}\mathbf{x}) \leq 2\delta_n \}$$

Definition 3.2.8 ((T_n, k_n, δ_n) -achievable) *Given a sequence of randomly generated query matrices $\{\mathbf{Q}^{(T_n, n)} \mid n \in \mathbb{N}\}$, we say it is (T_n, k_n, δ_n) -achievable, if*

$$\lim_{n \rightarrow \infty} \max_{\mathbf{x} \in \mathcal{X}} P_f(\mathbf{x}; k_n, \delta_n) = 0 \quad (3.2)$$



3.3 Main Results

3.3.1 Achievability

Theorem 3.3.1 (*Achievability of Randomized Querying*)

Suppose one generates the query matrix $\mathbf{Q}_{i,j}^{(T_n, n)}$ according to the following distribution:

$$(\mathbf{Q}^{(T_n, n)})_{i,j} \stackrel{i.i.d.}{\sim} \text{Ber}\left(\frac{1}{2}\right). \quad (3.3)$$

Then, the extraction criterion (3.2) will be satisfied as long as $T_n = \Omega\left(\frac{n}{\log n}\right)$ and one of the following conditions holds:

- 1) $k_n = O(n^\epsilon)$, for some $\epsilon < 1$ and $\delta_n = O(\sqrt{k_n})$
- 2) $\forall \epsilon < 1$, $k_n = \omega(n^\epsilon)$ and $\delta_n = O(k_n^{\frac{1-\epsilon'}{2}})$ for some $\epsilon' > 0$.

Proof. The proof involves finding upper bounds on the probability of failure. Details can be found in Section 3.4. ■

3.3.2 Lower Bounds on Query Complexity

For the converse part, we give two lower bounds in the following two theorems.

Theorem 3.3.2 (*Packing Lower Bound*) *Let $k_n \leq \left(\frac{1-\epsilon}{2}\right)n$ for some $\epsilon > 0$. Then, the following lower bound holds:*

$$T_n^*(k_n, \delta_n) = \Omega\left(\frac{n\left(1 - H_b\left(\frac{1-\epsilon}{2}\right)\right)}{\log(n+1) - \log(4\delta_n + 1)}\right) \quad (3.4)$$

Specifically, when $\delta_n = O(n^{\frac{1-\epsilon'}{2}})$, and ϵ, ϵ' does not depend on n , then (3.4) can be further simplified to

$$T_n^*(k_n, \delta_n) = \Omega(n / \log n).$$

Proof. The successful extraction criterion holds only if for any two data sets $\mathbf{x}, \tilde{\mathbf{x}}$ with distance greater than k_n , the queried output $\mathbf{Q}\mathbf{x}, \mathbf{Q}\tilde{\mathbf{x}}$ differ to each other more than $2\delta_n$, say, $d_{\text{response}}(\mathbf{Q}\mathbf{x}, \mathbf{Q}\tilde{\mathbf{x}}) > 2\delta_n$. Therefore, we cast the problem into a packing problem. The detailed proof is omitted here and can be found in Appendix A of [11] ■

Remark 3.3.1 The condition $k_n \leq (\frac{1-\epsilon}{2})n$ for some $\epsilon > 0$ is reasonable. Let $k_n = n/2$ and consider the following scenario: we simply make a query with $\mathbf{q}_i = [1, \dots, 1]^\top$, and if $\mathbf{q}_i^\top \mathbf{x} > n/2$, we reconstruct \mathbf{x} as $\tilde{\mathbf{x}} = [1, \dots, 1]^\top$, else we say $\tilde{\mathbf{x}} = [0, \dots, 0]^\top$. The reconstruction will succeed with high probability as n grows large enough, by making one query.

The above lower bound is used when the noise level δ_n is relatively small with respect to k_n . Next, we give another lower bound which depends on both k_n and δ_n :

Theorem 3.3.3 (Combinatorial Lower Bound)

$$T_n^*(k_n, \delta_n) \geq \frac{\binom{n}{n/2}}{2 \sum_{\alpha=2\delta_n}^{k_n/2} \binom{k_n/2}{\alpha} \sum_{\delta=2\delta_n}^{\alpha} \binom{k_n/2}{\alpha-\delta} \binom{n-k_n}{n/2-2\alpha+\delta}}. \quad (3.5)$$

This bound is used to prove the impossibility result when δ_n is large with respect to k_n . Detailed proof is given in Section 3.5.

3.3.3 Fundamental Limit

First, Theorem 3.3.1 gives us a sufficient condition for recovering the data set by $\Omega(n / \log n)$ queries. On the other hand, Theorem 3.3.2 states that $T_n = \Omega(n / \log n)$ is also necessary for reconstruction. We combine them into the following corollary:

Corollary 3.3.1 (Fundamental Limit of Query Complexity) Under the one of the following two noise-tolerance conditions



- $k_n = O(n^\epsilon)$ for some $\epsilon < 1$, and $\delta_n = O(\sqrt{k_n})$, or
- $\forall \epsilon < 1$, $k_n = \omega(n^\epsilon)$, and $\exists \epsilon' > 0$, $\delta_n = O(k_n^{(1-\epsilon')/2})$,

the optimal query complexity is

$$T_n^*(k_n, \delta_n) = \Theta\left(\frac{n}{\log n}\right).$$

Next, following Theorem 3.3.3, we give an impossibility result below:

Theorem 3.3.4 (*Impossibility of $\text{Poly}(n)$ Query Complexity*)

If both the following conditions are satisfied:

- $\frac{1}{2}n \geq k_n \geq C_1 n^{\epsilon_1}$
- $\delta_n = \Omega(k_n^{\frac{1+\epsilon_2}{2}})$

where $\epsilon_1, \epsilon_2 \in (0, 1)$, and $C_1 > 0$, then $T_n^*(k_n, \delta_n)$ is $\omega(n^p)$, for all $p \in \mathbb{N}$. In words, there does not exist querying methods with $\text{Poly}(n)$ query complexity that can do the job.

Again, the assumption $\frac{1}{2}n > k_n$ is reasonable due to Remark 3.3.1. To prove this result, we utilize Chernoff bound to derive a lower bound on $T_n^*(k_n, \delta_n)$, and see that it grows exponentially fast with n if δ_n is great enough. The details can be found in Appendix B of [11].

Remark 3.3.2 *Corollary 3.3.1 and Theorem 3.3.4 establish a sharp boundary $\delta_n \approx \sqrt{k_n}$ of partial data extraction under noisy responses to histogram queries. Roughly speaking, if $\delta_n \ll \sqrt{k_n}$, then the sufficient and necessary condition to recover data set is $T_n^* = \Theta(n/\log n)$. On the other hand, if $\delta_n \gg \sqrt{k_n}$, there is no querying method with $\text{Poly}(n)$ query complexity can reconstruct data set successfully.*

3.4 Achievability via Randomized Querying

In this section, we give the proof of Theorem 3.3.1. The proof involves upper bounding the probability of failure. Due to the randomized construction of the querying matrix, each

entry is generated in an i.i.d. fashion. Therefore, we first cast the probability of failure into the central probability of binomial distribution, and then further upper bound it.

Claim 3.4.1 *Under the randomized query defined in (3.3), the probability of failure can be upper bounded by*

$$P_f(\mathbf{x}; k_n, \delta_n) \leq \sum_{t=k_n}^n \binom{n}{t} \Pr(t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n)^{T_n}, \quad (3.6)$$

where $B_t \sim \text{Binomial}(t, 1/2)$.

The proof of the above claim is given in Appendix C in [11].

Continuing the proof of Theorem 3.3.1, the key is to separate the summation of (3.6) into two parts:

$$\underbrace{\sum_{t=k_n}^{k^*} \binom{n}{t} \Pr(t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n)^{T_n}}_{(i)} + \underbrace{\sum_{t=k^*}^n \binom{n}{t} \Pr(t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n)^{T_n}}_{(ii)}. \quad (3.7)$$

Before continuing bounding the probability of failure, we give a lemma to upper bound the central probability of binomial distribution:

Lemma 3.4.1 *Let $B_t \stackrel{iid}{\sim} \text{Binomial}(t, 1/2)$, $\delta_n \in (0, t/16)$ then the following two upper bounds hold:*

$$1) \Pr(t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n) \leq \frac{4\delta_n + 1}{\sqrt{\pi t}}.$$

This bound is used when δ_n is small (with respect to t).

$$2) \Pr(t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n) \leq 1 - \frac{2}{15} e^{-64\delta_n^2/t}.$$

This bound is used when δ_n is large (with respect to t).

The proof can be found in Appendix D in [11].

Now, we are ready for upper bounding (3.7).

For part (i) in (3.7), applying the second bound in Lemma 3.4.1, we have

$$\begin{aligned}
& \sum_{t=k_n}^{k^*} \binom{n}{t} \Pr(t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n)^{T_n} \\
& \leq \sum_{t=1}^{k^*} \binom{n}{t} \Pr(k_n/2 - 2\delta_n \leq B_{k_n} \leq k_n/2 + 2\delta_n)^{T_n} \\
& \leq \sum_{t=1}^{k^*} \binom{n}{t} \left(1 - \frac{2}{15} \exp\left(-\frac{64\delta_n^2}{k_n}\right)\right)^m \\
& \leq \left(1 - \frac{2}{15} \exp\left(-\frac{64\delta_n^2}{k_n}\right)\right)^{T_n} (n+1)^{k^*}
\end{aligned} \tag{3.8}$$

Due to our assumption that $\delta_n = O(\sqrt{k_n})$, $64\delta_n^2/k_n$ is upper bounded by some constant $\eta \geq 0$ for sufficiently large n , and hence

$$\left(1 - \frac{2}{15} \exp\left(-\frac{64\delta_n^2}{k_n}\right)\right) \leq \left(1 - \frac{2}{15} \exp(\eta)\right) =: \xi,$$

for sufficient large n . Note that ξ is a constant which does not depend on n , and is *strictly* less than 1.

Hence (3.8) can be further bounded by $\xi^{T_n} (n+1)^{k^*}$. To get vanishing probability of failure, T_n must satisfy

$$T_n = \Omega\left(\frac{k^* \log n}{\log \xi}\right) = \Omega(k^* \log n), \tag{3.9}$$

since ξ does not depend on n .

For part (ii) in (3.7), we have

$$\begin{aligned}
& \sum_{t=k^*}^n \binom{n}{t} \Pr(t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n)^{T_n} \\
& \leq \sum_{t=0}^n \binom{n}{t} \Pr(k^*/2 - 2\delta_n \leq B_{k^*} \leq k^*/2 + 2\delta_n)^{T_n} \\
& \leq \left(\frac{4\delta_n + 1}{\sqrt{\pi k^*}}\right)^{T_n} 2^{n+1},
\end{aligned} \tag{3.10}$$

where (3.10) is due to Lemma 3.4.1.



To obtain vanishing failure probability,

$$T_n = \Omega \left(\frac{n+1}{\frac{1}{2} \log(\pi k^*) - \log(4\delta_n + 1)} \right). \quad (3.11)$$



Notice that (3.11) requires $\sqrt{\pi k^*} > 4\delta_n + 1$.

In order to choose a proper k^* according to (3.9) and (3.11), we distinguish k_n into two regimes:

1) $k_n = O(n^\epsilon)$, for some $\epsilon \in (0, 1)$:

In this regime, $k_n = O(n^\epsilon)$ and $\delta_n = O(n^{\epsilon/2})$. Hence one can choose k^* such that $k^* = \Theta(n^{\epsilon+\epsilon'})$, where $\epsilon + \epsilon' < 1$.

In this case,

$$(3.9) \implies T_n = \Omega(n^{\epsilon+\epsilon'} \log n)$$

$$(3.11) \implies T_n = \Omega \left(\frac{n}{(\epsilon + \epsilon') \log n/2 - \log \delta_n} \right)$$

2) $k_n = \omega(n^\epsilon)$, for all $\epsilon < 1$:

In this regime, $\delta_n = O(k_n^{(1-\epsilon')/2})$, and therefore we can choose k^* such that $k^* = \Theta(n^{1-\epsilon'})$. In this case,

$$(3.9) \implies T_n = \Omega(n^{1-\epsilon'} \log n)$$

$$(3.11) \implies T_n = \Omega \left(\frac{n}{(1 - \epsilon') \log n/2 - \log \delta_n} \right).$$

The proof is complete by noticing that $T_n = \Omega \left(\frac{n}{\log n} \right)$ is sufficient for the cases in the two regimes.

3.5 Proof of the Combinatorial Lower Bound

In this section, we give the proof of combinatorial lower bound stated in Theorem 3.3.3.

For notational convenience, let us define the right-hand side of (3.5) as τ . Then, the theorem is equivalent to the following statement:

For any $T_n \leq \tau$, $\exists \mathbf{x}, \tilde{\mathbf{x}} \in \{0, 1\}^n$, $\|\mathbf{x} - \tilde{\mathbf{x}}\| > k_n$, such that $|\mathbf{Q}\mathbf{x} - \mathbf{Q}\tilde{\mathbf{x}}| \leq 2\delta_n$.

The main idea of the proof is as follows. Consider a subset S of all confused pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ differing by at least k_n elements. After each query \mathbf{q}_i , one can remove some candidates in S according to the response. If for every single query, the number of removed candidates is at most N , then at least $\frac{|S|}{N}$ queries are needed. We will show that $\tau \leq \frac{|S|}{N}$. Therefore once $T_n \leq \tau$, there exists at least one ambiguous data $\tilde{\mathbf{x}}$, and hence the reconstruction is impossible. Moreover, τ is a lower bound of $T_n^*(k_n, \delta_n)$.

For a data set $\mathbf{x} \in \{0, 1\}^n$, denote an ambiguous data set as $\tilde{\mathbf{x}}$. We focus on the collection of all possible pairs of $(\mathbf{x}, \tilde{\mathbf{x}})$ which have the same one norm, and differs from each other exactly k_n 's element, that is,

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_1 = k_n, \text{ and } \|\mathbf{x}\|_1 = \|\tilde{\mathbf{x}}\|_1.$$

Let $\mathbf{x}, \tilde{\mathbf{x}} \in \{0, 1\}^n$, and define

$$\begin{aligned} S_{k_n} &\triangleq \{(\mathbf{x}, \tilde{\mathbf{x}}) \mid \|\mathbf{x} - \tilde{\mathbf{x}}\|_1 = k_n, \|\mathbf{x}\|_1 = \|\tilde{\mathbf{x}}\|_1\} \\ &= \left\{(\mathbf{x}, \tilde{\mathbf{x}}) \mid \pi(1|\mathbf{x} - \tilde{\mathbf{x}}) = \pi(-1|\mathbf{x} - \tilde{\mathbf{x}}) = \frac{k_n}{2}\right\}, \end{aligned}$$

where we use $\pi(\cdot \mid \mathbf{w})$ to denote the *unnormalized histogram* of vector \mathbf{w} , say, $\pi(x \mid \mathbf{w}) \triangleq (\text{number of } x \text{ in } \mathbf{w})$. Define the collection of all *confusion datasets* after the i -th query :

$$V_i \triangleq \{(\mathbf{x}, \tilde{\mathbf{x}}) \in S_{k_n} \mid |\mathbf{q}_i \cdot (\mathbf{x} - \tilde{\mathbf{x}})| \leq 2\delta_n\}.$$

As long as

$$T_n < \frac{|S_{k_n}|}{\max_{i \in \{1, \dots, T_n\}} |V_i^c|}, \quad (3.12)$$

(with a slight abuse of notation, let $V_i^c = V_i^c \cap S_{k_n}$), we have

$$|S_{k_n}| > T_n \max_{i \in \{1, \dots, T_n\}} |V_i^c| \geq \sum_{i=1}^{T_n} |V_i^c| \geq \left| \bigcup_{i=1}^{T_n} V_i^c \right|$$

due to union bound. Notice that

$$\left| \bigcup_{i=1}^{T_n} V_i^c \right| < |S_{k_n}| \iff \bigcup_{i=1}^{T_n} V_i^c \neq S_{k_n} \iff \bigcap_{i=1}^{T_n} V_i \neq \emptyset,$$



which implies that there exists at least one pair of confusion data sets $(\mathbf{x}, \tilde{\mathbf{x}}) \in S_{k_n}$ after T_n independent queries. To complete the proof, all we need is to the following claim:

Claim 3.5.1

$$\tau \leq \frac{|S_{k_n}|}{\max_{i \in \{1, \dots, T_n\}} |V_i^c|}.$$

Proof. First, we introduce

$$\mathcal{T}_1 \triangleq \{j \mid \tilde{x}_j = 0, x_j = 1\}, \mathcal{T}_2 \triangleq \{j \mid \tilde{x}_j = 1, x_j = 0\}.$$

Note that suppose $(\mathbf{x}, \tilde{\mathbf{x}}) \in S_{k_n}$, then $|\mathcal{T}_1| = |\mathcal{T}_2| = k_n/2$ due to the fact $\|\mathbf{x}\|_1 = \|\tilde{\mathbf{x}}\|_1$, and $\|\mathbf{x} - \tilde{\mathbf{x}}\|_1 = k_n$. Then obviously we have

$$|S_{k_n}| = \binom{n}{k_n/2} \binom{n - k_n/2}{k_n/2} 2^{n-k_n}. \quad (3.13)$$

Let the queried subset corresponding to \mathbf{q} be \mathcal{S} . The *confusion events* $\{|\mathbf{q}\mathbf{x} - \mathbf{q}\tilde{\mathbf{x}}| \leq 2\delta_n\}$ happen if and only if

$$||\mathcal{S} \cap \mathcal{T}_1| - |\mathcal{S} \cap \mathcal{T}_2|| \leq 2\delta_n. \quad (3.14)$$

Therefore, to upper bound $|V_i^c|$, we have

$$\begin{aligned} \max_i |V_i^c| &\leq \max_{\mathbf{q}} |\{(\mathbf{x}, \tilde{\mathbf{x}}) \in S_{k_n} \mid |\mathbf{q} \cdot (\mathbf{x} - \tilde{\mathbf{x}})| > 2\delta_n\}| \\ &= \max_{\mathcal{S} \subseteq [n]} |\{(\mathbf{x}, \tilde{\mathbf{x}}) \in S_{k_n} \mid ||\mathcal{S} \cap \mathcal{T}_1| - |\mathcal{S} \cap \mathcal{T}_2|| > 2\delta_n\}| \end{aligned} \quad (3.15)$$

By symmetry, it is intuitive that the maximum is attained when $|\mathcal{S}| = \frac{n}{2}$ (we also give

a rigorous proof in Appendix, see Lemma D.1), and thus (3.15) is equal to



$$\begin{aligned}
& \left| \{(\mathbf{x}, \tilde{\mathbf{x}}) \in S_{k_n} \mid ||\mathcal{S} \cap \mathcal{T}_1| - |\mathcal{S} \cap \mathcal{T}_2|| > 2\delta_n\} \right| \\
&= \left| \bigcup_{\delta > 2\delta_n} \{(\mathbf{x}, \tilde{\mathbf{x}}) \in S_{k_n} \mid ||\mathcal{S} \cap \mathcal{T}_1| - |\mathcal{S} \cap \mathcal{T}_2|| = \delta\} \right| \\
&= 2^{n-k_n+1} \sum_{\alpha=0}^{k_n/2} \sum_{\delta=2\delta_n}^{\alpha} \binom{n/2}{\alpha} \binom{n/2}{k_n/2 - \alpha} \binom{n/2 - \alpha}{\alpha - \delta} \binom{n/2 - k_n/2 + \alpha}{k_n/2 + \delta - \alpha} \quad (3.16)
\end{aligned}$$

Combining (3.13) and (3.16), we obtain

$$\begin{aligned}
& \frac{|S_{k_n}|}{\max_{i \in \{1, \dots, T_n\}} |V_i^c|} \\
& \geq \frac{\binom{n}{k_n/2} \binom{n-k_n/2}{k_n/2}}{2 \sum_{\alpha=0}^{k_n/2} \sum_{\delta=2\delta_n}^{\alpha} \binom{n/2}{\alpha} \binom{n/2-k_n/2}{k_n/2-\alpha} \binom{n/2-\alpha}{\alpha-\delta} \binom{n/2-k_n/2+\alpha}{k_n/2+\delta-\alpha}} \\
& = \frac{\binom{n}{\frac{n}{2}}}{2 \sum_{\alpha=2\delta_n}^{k_n/2} \binom{k_n/2}{\alpha} \sum_{\delta=2\delta_n}^{\alpha} \binom{k_n/2}{\alpha-\delta} \binom{n-k_n}{n/2-2\alpha+\delta}} = \tau, \quad (3.17)
\end{aligned}$$

where (3.17) is due to direct calculation of binomial coefficient. This proves our claim. ■

3.6 Extension

We close this part by briefly explaining how to extend our results to the general case $|\mathcal{A}| = d$. Following the formulation in [43], the data set can be modeled by a matrix $\mathbf{X} \in \{0, 1\}^{n \times d}$, and the response $\mathbf{Y} \in \{0, 1, \dots, n\}^{T_n \times d}$. To prove the achievability part, we notice that the probability of error $P_f(\mathbf{X}; k_n, \delta_n)$ (w.r.t \mathbf{Q}) is upper bounded by $P_f(\mathbf{x}; k_n, \delta_n)$; here we abuse the notation, denoting $\mathbf{x} \in \{0, 1\}^n$ for some column of \mathbf{X} . Hence, Theorem 3.3.1 also holds for d being constant with respect to n .

On the other hand, suppose \mathbf{X} and $\tilde{\mathbf{X}}$ are two data sets with Hamming distance greater than k_n . Then, there exists some column of $\mathbf{X}, \tilde{\mathbf{X}}$, say $\mathbf{x}, \tilde{\mathbf{x}}$, such that $d_{\text{data}}(\mathbf{x}, \tilde{\mathbf{x}}) \geq k_n/d$. Therefore, the converse results in Theorem 3.3.2, 3.3.3, and 3.3.4 hold for $k'_n = \left(\frac{k_n}{d}\right)$. In particular, as long as d is a constant with respect to n , the asymptotic behavior remains the same.





Part II

Anonymous Hypothesis Testing





Chapter 4

Anonymous Hypothesis Testing : Optimal Decision Rules and Type-II Error Exponents

4.1 Previous Work

As introduced in Chapter 1, in order to be consistent with the previous literatures, we use languages and notations from wireless sensor networks, where the term *sensors* and *observations* represent *crowds* and *responses*, as stated in previous context. After all, the essence of the two problems are indeed identical. In the following we briefly review previous works on distributed detection and wireless sensor networks.

Decentralized detection is a classical topic, and attracts extensive attention in recent years due to its application in wireless sensor networks. See, for example, [39, 40, 38, 41]. Most works in decentralized detection are focused on finding optimal local decision function in both Neyman-Pearson and Bayesian regime. Under some assumptions on the distribution of a given hypothesis, optimal design criteria of local decision function and the decision rule at the fusion center are given. Unlike the anonymous setting considered in our work, the above-mentioned classical works assume fusion centers, as well as the local sensors, have perfect knowledge about the joint distribution, and hence the decision

rules are designed according to it. This is termed an “informed” setting in our paper and is used as a baseline to compare with and see the price of anonymity. On the other hand, in our setting, the fusion center collects observations without knowing the exact index of each one, and thus the problem is formulated into a composite hypothesis testing problem.

Composite hypothesis testing is a long-standing problem in statistics, and is notoriously difficult to find an optimal test. In general, the uniform most powerful (UMP) test does not exist, see, for example, Section 8.3 in [21]. Even if we relax the performance evaluation to the *minimax* regime, the general form of the optimal test is still unknown, except for some special case. For example, [24] considered the case that the composite hypothesis class \mathcal{H}_θ is formed by all ϵ -contaminated distributions of P_θ , that is, $\{(1 - \epsilon)P_\theta + \epsilon Q \mid \forall \text{ possible distributions } Q\}$. Under this structure, Huber showed that a censored version of likelihood ratio test is optimal in the minimax regime. Other works such as [22, 44] followed the idea of Hoeffding’s test [22] and proposed an universal asymptotically optimal test when the null hypothesis is simple. Meanwhile, in our setting, neither the parameter space of the considered distributions is continuous, nor the null hypothesis is simple, making their approaches hard to extend. Another common test for composite hypothesis testing is the *generalized likelihood ratio test* (GLRT). The optimality of GLRT is guaranteed under some circumstances, see, for example, [45]. However, the results in [45] hold only for simple null and composite alternative. In contrast, our result indicates that GLRT is not optimal in our setting.

The concept of Byzantine attack can be traced back to [28] (known as the “Byzantine Generals Problem”), in which reliability of a computer system with malfunctioned components is studied. After that, Byzantine model is developed and generalized by several research areas, especially in communication security. For example, the distributed detection with Byzantine attack is studied under the Neyman-Pearson formulation in [30] and under the Bayesian setting in [26]. In their settings, each sensor is assumed to be compromised with probability α , so the observation turns out to be drawn identically and independently from an mixture distribution, making the hypothesis testing problem simple, and thus Neyman-Pearson lemma can be applied. In contrast, in our work we assume

the number of Byzantine sensors is fixed and is αn , where n is the total number of sensors, and thus the problem falls into a composite hypothesis testing instead of the mixture setting.

This work is presented in part at ISIT 2018. In the conference version [9], upper and lower bounds on the type-II error exponent were given, where the lower bound (achievability) is based on an modified version of Hoeffding's test, and the upper bound (converse) is derived by relaxing the original problem into a simple hypothesis testing. In this journal version, we show that the achievability bound in the conference version is indeed tight, closing the gap between the upper and lower bounds.

The rest of this chapter is organized as follows. In Section 4.2, we formulate the composite hypothesis testing problem for anonymous heterogeneous distributed detection and recap some background knowledge. In Section 4.3, the main results are provided, where the proofs are delegated to Section 4.4 and 4.5.

4.2 Problem Formulation

4.2.1 Problem Setup

Following the description of the setting in Chapter 1, let us formulate the composite hypothesis testing problem. Let $\sigma(i)$ denote the label of the group that sensor i belongs to. This labeling $\sigma(\cdot)$, however, is not revealed to the fusion center. Hence, the fusion center needs to consider all $\binom{n}{n_1, \dots, n_K}$ possible $\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, K\}$ satisfying

$$|\{i \mid \sigma(i) = k\}| = n_k, \forall k = 1, \dots, K, \quad (4.1)$$

and decides whether the hidden θ is 0 or 1. For notational convenience, let ν denote the vector $[n_1 \dots n_K]^\top$, and let $\mathcal{S}_{n, \nu}$ denote the collection of all labelings satisfying (4.1).

Hence, the fusion center is faced with the following *composite* hypothesis testing prob-

lem, where the goal is to infer the parameter θ :

$$\mathcal{H}_\theta : X^n \sim \mathbb{P}_{\theta;\sigma} \triangleq \prod_{i=1}^n P_{\theta;\sigma(i)}, \text{ for some } \sigma \in \mathcal{S}_{n,\nu}.$$



As mentioned in Chapter 1, throughout the paper we consider binary hypothesis testing, that is, $\theta \in \{0, 1\}$.

Let each single observation take values from some measurable space $(\mathcal{X}, \mathcal{F})$, where \mathcal{F} is a σ -algebra on \mathcal{X} . Hence $P_{\theta;k} \in \mathcal{P}_\mathcal{X}$ for all $\theta \in \{0, 1\}$ and $k \in \{1, \dots, K\}$, where $\mathcal{P}_\mathcal{X}$ denotes the collection of all possible distributions over $(\mathcal{X}, \mathcal{F})$. The vector observation x^n is defined on the space $(\mathcal{X}^n, \mathcal{F}^{\otimes n})$, where $\mathcal{F}^{\otimes n}$ is the *tensor product* σ -algebra of \mathcal{F} , that is, the smallest σ -algebra contains the following collection of events:

$$\{\mathcal{E}_1 \times \mathcal{E}_2 \times \dots \times \mathcal{E}_n \mid \mathcal{E}_i \in \mathcal{F}\}.$$

A (randomized) test is a measurable function $\phi : (\mathcal{X}^n, \mathcal{F}^{\otimes n}) \rightarrow ([0, 1], \mathfrak{B})$, where \mathfrak{B} denotes the Borel σ -field on \mathbb{R} . The worst-case type-I and type-II error probabilities of a decision rule ϕ are defined as

$$P_F^{(n)}(\phi) \triangleq \max_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{E}_{\mathbb{P}_{0;\sigma}} [\phi(X^n)] \quad (\text{Type I})$$

$$P_M^{(n)}(\phi) \triangleq \max_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{E}_{\mathbb{P}_{1;\sigma}} [1 - \phi(X^n)] \quad (\text{Type II}).$$

Our focus is on the Neyman-Pearson setting: find a decision rule ϕ satisfying $P_F^{(n)}(\phi) \leq \epsilon$ such that $P_M^{(n)}(\phi)$ is minimized. Let $\beta^{(n)}(\epsilon, \nu)$ denote the minimum type-II error probability.

For the asymptotic regime, we assume that the ratio $\frac{n_k}{n} \rightarrow \alpha_k$ as $n \rightarrow \infty$ for all $k = 1, \dots, K$, and $\sum_{k=1}^K \alpha_k = 1$. We aim to explore if $\beta^{(n)}(\epsilon, \nu)$ decays exponentially fast as $n \rightarrow \infty$, and characterize the corresponding error exponent. For notational convenience, we define upper and lower bounds on the exponent:

$$\overline{E}^*(\epsilon, \alpha) \triangleq \limsup_{n \rightarrow \infty} \left\{ -\frac{1}{n} \log_2 \beta^{(n)}(\epsilon, \nu) \right\},$$

$$\underline{E}^*(\epsilon, \alpha) \triangleq \liminf_{n \rightarrow \infty} \left\{ -\frac{1}{n} \log_2 \beta^{(n)}(\epsilon, \nu) \right\},$$



where in taking the limits, we assume that $\lim_{n \rightarrow \infty} \frac{n_k}{n} = \alpha_k$, for all $k = 1, \dots, K$. If the upper and lower bound match, we simply denote it as $E^*(\epsilon, \alpha)$.

Remark 4.2.1 *The original distributed detection problem [40, 39, 41] involves local decision functions at the sensors to address the limited communication between each sensor and the fusion center. In order to focus on the impact of anonymity, we first absorb them into the distributions $\{P_{\theta;k} : k = 1, \dots, K\}$ because they are symbol-by-symbol maps. Later, we will discuss how to find the best local decision functions according to the characterized error exponent.*

4.2.2 Notations

Let us introduce notations that will be used throughout this paper.

- n denotes the total number of observations, and K denotes the number of groups of sensors.
- $\nu \triangleq [n_1 \dots n_K]^\top$ denotes the number of sensors in the K groups. That is, $n_k \geq 0$, $n_k \in \mathbb{Z}$, and $\sum_{k=1}^K n_k = n$.
- $\alpha \triangleq [\alpha_1 \dots \alpha_K]^\top$ denotes the fraction of each group of sensors in all sensors in the asymptotic regime. That is, $\alpha_k \geq 0$, and $\sum_{k=1}^K \alpha_k = 1$.
- $\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, K\}$ is the labeling function which assigns the index of each sensor to a group. We also denote the collection of indices of sensors in group k as

$$\mathcal{I}_k = \sigma^{-1}(k) \triangleq \{i \mid \sigma(i) = k\}. \quad (4.2)$$

- Let $\mathcal{S}_{n,\nu}$ be the collection of all σ satisfying (4.2). We also use \mathcal{S}_n to denote the collection of length- n permutations:

$$\mathcal{S}_n \triangleq \left\{ \tau : \{1, 2, \dots, n\} \xrightarrow{1-1} \{1, 2, \dots, n\} \right\}.$$

Note that the cardinalities of the two sets are

$$|\mathcal{S}_{n,\nu}| = \binom{n}{n_1, n_2, \dots, n_K}, \quad |\mathcal{S}_n| = n!.$$



- We usually write \mathbf{P}_θ as the vector of $\{P_{\theta;k}\}$:

$$\mathbf{P}_\theta \triangleq \begin{bmatrix} P_{\theta;1} \\ P_{\theta;2} \\ \vdots \\ P_{\theta;K} \end{bmatrix}.$$

4.3 Main Results

As mentioned in Section 4.2, the observations come from the measurable space $(\mathcal{X}^n, \mathcal{F}^{\otimes n})$. Throughout the rest of the paper, we assume that \mathcal{X} is a totally ordered set, and $\mathcal{F}^{\otimes n}$ satisfies the following two assumptions:

1. $\mathcal{F}^{\otimes n}$ contains the following set:

$$\tilde{\mathcal{X}}^n \triangleq \{(x_1, x_2, \dots, x_n) \mid x_1 \geq x_2 \geq \dots \geq x_n\}. \quad (4.3)$$

2. $\mathcal{F}^{\otimes n}$ is closed under permutation. That is, if $\mathcal{A} \in \mathcal{F}^{\otimes n}$, for any length- n permutation $\tau : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$,

$$\mathcal{A}_\tau \triangleq \{(x_{\pi(1)}, \dots, x_{\pi(n)}) \mid (x_1, \dots, x_n) \in \mathcal{A}\} \in \mathcal{F}. \quad (4.4)$$

Remark 4.3.1 We assume that \mathcal{X} is a totally ordered set in order to set the condition such that $\tilde{\mathcal{X}}$ is measurable. The purpose to require $\tilde{\mathcal{X}}$ to be measurable is to preserve the measurability of the ordering map $\Pi(\cdot)$, as later defined in Definition 4.4.1. In general, if \mathcal{X} is not totally ordered, we can still require the collection of representatives in the equivalent classes induced by Π^{-1} to be measurable. However, the regularity assumptions on \mathcal{F}^{\otimes} need to be carefully concerned in that case.

Remark 4.3.2 *The second assumption always holds for tensor σ -fields. The first assumption typically holds too. For example, if \mathcal{X} is finite, we can simply choose \mathcal{F} as the power set $2^{\mathcal{X}}$, and if $\mathcal{X} \subseteq \mathbb{R}$, we can choose \mathcal{F} as the Borel σ -field. In particular, for \mathcal{X} being a finite set, it is straightforward to define a total order over it, and hence it is a totally ordered set. Moreover, the above two assumptions are automatically satisfied.*

4.3.1 Main Contributions

Our first contribution is the characterization of the optimal test:

Theorem 4.3.1 (Optimal Test) *Define the mixture likelihood ratio $\ell(x^n)$:*

$$\ell(x^n) \triangleq \frac{\sum_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{P}_{1;\sigma}(x^n)}{\sum_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{P}_{0;\sigma}(x^n)}. \quad (4.5)$$

Suppose $\mathcal{F}^{\otimes n}$ satisfies the two assumptions (4.3), (4.4). Then an optimal tests $\phi^(x^n)$ takes the following form:*

$$\phi^*(x^n) = \begin{cases} 1, & \text{if } \ell(x^n) > \tau \\ \gamma, & \text{if } \ell(x^n) = \tau \\ 0, & \text{if } \ell(x^n) < \tau. \end{cases} \quad (4.6)$$

That is, for any test ϕ , we have

$$P_F(\phi) \leq P_F(\phi^*) \Rightarrow P_M(\phi) \geq P_M(\phi^*).$$

Remark 4.3.3 *We see that the optimal test, MLRT, is the likelihood ratio test between two uniform mixture distributions*

$$\frac{1}{|\mathcal{S}_{n,\nu}|} \sum_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{P}_{\theta;\sigma}, \quad \theta \in \{0, 1\}.$$

Interestingly, the optimality of MLRT indicates that the widely used decision rule, generalized likelihood ratio test (GLRT), which is defined as the randomized thresholded test

according to the following likelihood ratio

$$\ell_{GLRT}(x^n) \triangleq \frac{\sup_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{P}_{1;\sigma}(x^n)}{\sup_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{P}_{0;\sigma}(x^n)},$$

is strictly sub-optimal in the anonymous hypothesis testing problem.

Sketch of proof. The proof consists of two steps. In the first step, we introduce *symmetric tests* (as later defined in Definition 4.4.2), which do not depend on the order of the observations. Then, we show that among all symmetric tests, (4.6) is optimal. The key is to reduce the original composite hypothesis testing problem into a simple one through the ordering map $\Pi(x^n)$ in Definition 4.4.1, and then apply Neyman-Pearson lemma.

In the second step, we prove that for any test ψ , one can always *symmetrize* it and construct a symmetric one ϕ which is as good as ψ , so (4.6) is optimal among all tests. However, ψ is constructed by assigning values on each equivalence classes introduced by the ordering map $\Pi(\cdot)$, so the measurability of ψ need to be carefully examined. For the detailed proof, please refer to Section 4.4. ■

Our second result specifies the exponent of type-II error in Neyman-Pearson formulation, which does not depend on the type-I error probability ϵ :

Theorem 4.3.2 (Asymptotic Behavior) *Let us consider the case $|\mathcal{X}| < \infty$, The exponent of type-II error probability is characterized as follows.*

$$E^*(\epsilon, \alpha) = \min_{U \in (\mathcal{P}_{\mathcal{X}})^K} \sum_{k=1}^K \alpha_k D(U_k \| P_{1;k}) \quad (4.7)$$

subject to $\alpha^\top U = \alpha^\top P_0$.

Remark 4.3.4 *A standard way to derive the exponent of type-II error probability is to identify the acceptance region (of \mathcal{H}_0) of the optimal test (4.6) as an large-deviation event under \mathcal{H}_1 , and further apply a strong converse lemma to obtain a bound. However, notice that the mixture measure, $\sum_{\sigma} \mathbb{P}_{\theta;\sigma}$, $\theta \in \{0, 1\}$, cannot be factorized into a product form, which makes it hard to single-letterize. Instead, if we add an additional assumption that \mathcal{X} is finite, then we can utilize method of types, such as Sanov's theorem, to circumvent the difficulties.*



Sketch of proof. For the achievability part, we propose a sub-optimal test based on Hoeffding's result [22], in which we accept observations x^n satisfying $D(\Pi_{x^n} \| M_0(\alpha)) \leq \epsilon$ for some threshold ϵ . We apply tools in method of types to bound the type-I and type-II error probabilities, showing that (4.7) is achievable.

For the converse part, given an arbitrary test, we define its acceptance region as \mathcal{A} (if the given test is randomized, we can round the test by 1/2 and make it deterministic, that is, we accept \mathcal{H}_1 if $\phi(x^n) > 1/2$) and consider another high-probability set \mathcal{B} . We analyze the probability of $\mathbb{P}_{1;\sigma} \{\mathcal{A} \cap \mathcal{B}\}$, and show that the exponent cannot be greater than (4.7), which concludes the converse part. For the detailed proof, please refer to Section 4.5. ■

Finally, we give a structural result of the error exponent.

Proposition 4.3.1 *For the case $|\mathcal{X}| < \infty$, the type-II error exponent $E^*(\epsilon, \alpha)$ as characterized in Theorem 4.3.2 only depends on α . Moreover, it is a convex function of α .*

Proof. See Appendix B.1. ■

4.3.2 Numerical Evaluations

To quantify the price of anonymity, note that when the sensors are not anonymous (termed the “informed” setting), it becomes a simple hypothesis testing problem, and the error exponent of the type-II probability of error in the Neyman-Pearson setting is straightforward to derive:

$$E_{\text{Informed}}^*(\epsilon, \alpha) = \sum_{k=1}^K \alpha_k D(P_{0;k} \| P_{1;k}).$$

For ease of illustration, in the following we restrict to the special case of binary alphabet, that is, $|\mathcal{X}| = 2$, and $K = 2$ groups. Let $P_{\theta;1} = \text{Ber}(p_\theta)$ and $P_{\theta;2} = \text{Ber}(q_\theta)$, for $\theta = 0, 1$, where $\text{Ber}(p)$ is the Bernoulli distribution with parameter p . Since there are only two groups, we set $\alpha \equiv \begin{bmatrix} 1 - \alpha & \alpha \end{bmatrix}^\top$. Numerical examples are given in Figure 4.1 to illustrate the price of anonymity versus the mixing parameter α . In general, anonymity may cause significant performance loss. In certain regimes, the type-II error exponent can even be pushed to zero.

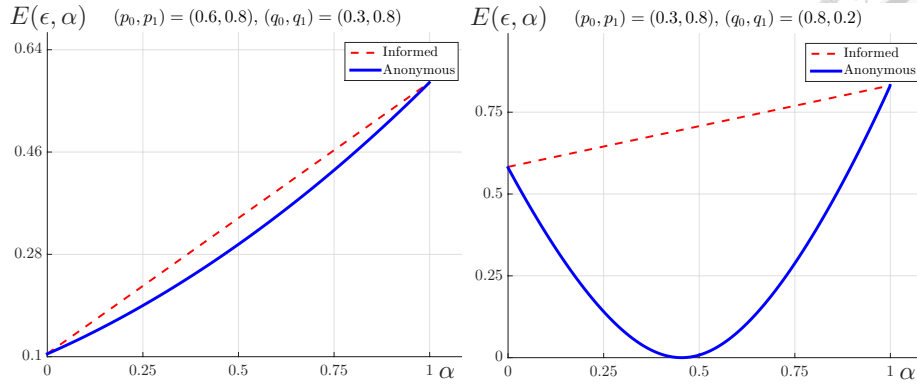


Figure 4.1: Price of anonymity

4.3.3 Distributed Detection with Byzantine Attacks

Let us apply the results to distributed detection under Byzantine attacks, where the sensors are partitioned into two groups. One group consists of $n(1 - \alpha)$ *honest* sensors reporting true i.i.d. observations, while the other consists of $n\alpha$ *Byzantine* sensors reporting fake i.i.d. observations. Here we again neglect the local decision function and assume that each sensor can report its observation to the fusion center. The true observations follow P_θ i.i.d. across honest sensors, while the compromised ones follow Q_θ i.i.d. across Byzantine sensors, for $\theta = 0, 1$. In general, Q_θ is unknown to the fusion center, but in terms of error exponent, one can find the least favorable pair Q_0, Q_1 which minimize the error exponent. Hence, our results can be applied here and arrive the worst-case type-II error exponent as follows:

$$\begin{aligned} \min_{Q_0, Q_1, U, V \in \mathcal{P}_{\mathcal{X}}} & (1 - \alpha)D(U \| P_1) + \alpha D(V \| Q_1) \\ \text{subject to} & (1 - \alpha)U + \alpha V = (1 - \alpha)P_0 + \alpha Q_0. \end{aligned} \quad (4.8)$$

In [30], it assumes that each sensor can be compromised with probability α , and hence it becomes a homogeneous distributed detection problem, where the observation of each sensor follows a mixture distribution $(1 - \alpha)P_\theta + \alpha Q_\theta$ under hypothesis θ , i.i.d. across all sensors. The worst-case exponent of type-II error probability, as derived in [30], is hence

$$\min_{Q_0, Q_1 \in \mathcal{P}_{\mathcal{X}}} D((1 - \alpha)P_0 + \alpha Q_0 \| (1 - \alpha)P_1 + \alpha Q_1). \quad (4.9)$$

We see that the achievable type-II error exponent (4.8) in our setting is always greater than that in the i.i.d. scenario (4.9) (and is *strictly* larger for some α) due to the convexity of KL divergence. This implies the i.i.d. mixture model [30] might be too pessimistic. Figure 4.2 shows a numerical evaluation.

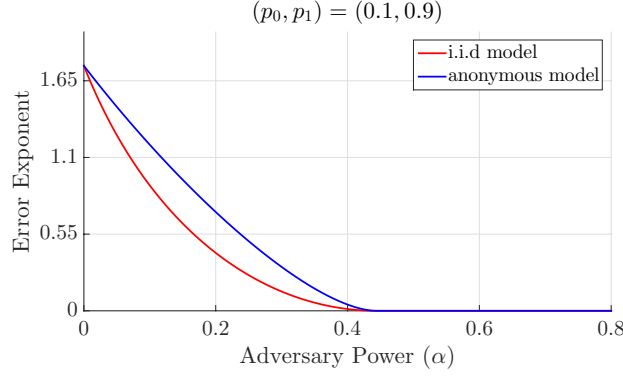


Figure 4.2: Comparison between i.i.d. and our setting

4.4 Proof of Theorem 4.3.1

Before proving Theorem 4.3.1, let us introduce some definitions that help the exposition.

Definition 4.4.1 (Ordering Map) *The ordering map $\Pi(\cdot) : (\mathcal{X}^n, \mathcal{F}^{\otimes n}) \rightarrow (\tilde{\mathcal{X}}^n, \tilde{\mathcal{F}})$, where $\tilde{\mathcal{X}}^n$ is from (4.3) and $\tilde{\mathcal{F}} \triangleq \mathcal{F}^{\otimes n} \cap \tilde{\mathcal{X}}^n$, is defined as follows:*

$$\Pi(x^n) \triangleq (x_{i_1}, x_{i_2}, \dots, x_{i_n}), \text{ such that } x_{i_1} \geq x_{i_2} \geq \dots \geq x_{i_n}.$$

The measurability of Π is easy to check.

Remark 4.4.1 *If $|\mathcal{X}| < \infty$, the mapping Π maps a sample x^n to its type, and the space $\tilde{\mathcal{X}}^n$ is equivalent to $\mathcal{P}_{\mathcal{X}}$.*

Remark 4.4.2 *We will use Π^{-1} to denote the pre-image of Π . That is, for all $\tilde{\mathcal{E}} \subseteq \tilde{\mathcal{X}}^n$,*

$$\Pi^{-1}(\tilde{\mathcal{E}}) \triangleq \{x^n \in \mathcal{X}^n \mid \Pi(x^n) \in \tilde{\mathcal{E}}\}.$$

Notice that the measurability of Π implies for any $\tilde{\mathcal{E}} \in \tilde{\mathcal{F}}$, we have $\Pi^{-1}(\tilde{\mathcal{E}}) \in \mathcal{F}^{\otimes n}$.



Definition 4.4.2 (Symmetric Test) We say a test $\phi(x^n)$ is symmetric, if it is $\sigma(\Pi(X^n))$ -measurable, that is, it can be represented as a composition

$$\phi(x^n) = \tilde{\phi} \circ \Pi(x^n),$$

for some measurable function $\tilde{\phi} : \tilde{\mathcal{X}}^n \rightarrow [0, 1]$. This implies the test ϕ maps a sequence of observations x^n and all its permutations to the same value.

Lemma 4.4.1 Among all symmetric test, $\phi^*(x^n)$, as defined in (4.6), is optimal.

proof of Lemma 4.4.1. To show the optimality of ϕ^* , we first transform the original composite hypothesis testing problem to another one in the auxiliary space $\tilde{\mathcal{X}}^n$ through the ordering mapping $\Pi(\cdot)$, which turns out to be a simple hypothesis testing problem. Hence, applying Neyman-Pearson lemma, we obtain the optimal test. See Figure 4.3 for illustration of the relation between the original space and the auxiliary space.

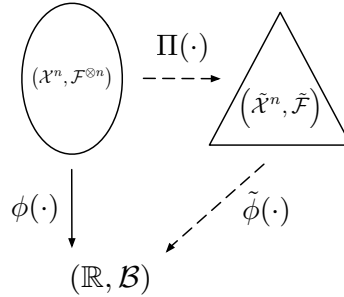


Figure 4.3: Illustration of the auxiliary space

Part 1 First, we claim that for all $\sigma \in \mathcal{S}_{n,\nu}$, the probability measure $\mathbb{P}_{0;\sigma} \circ \Pi^{-1}$, defined on $(\tilde{\mathcal{X}}^n, \tilde{\mathcal{F}})$, does not depend on σ anymore. Thus we can define the probability measure $\tilde{\mathbb{P}}_0 \triangleq \mathbb{P}_{0;\sigma} \circ \Pi^{-1}$, such that for all σ ,

$$(\mathbb{P}_{0;\sigma}, \mathcal{F}^{\otimes n}, \mathcal{X}^n) \xrightarrow{\Pi(\cdot)} (\tilde{\mathbb{P}}_0, \tilde{\mathcal{F}}, \tilde{\mathcal{X}}^n).$$

This claim is quite intuitive, since the labeling σ corresponds to the order of observations, and the ordering map removes the order.

To show this claim, we first observe that for all $\mathcal{E} \in \tilde{\mathcal{F}}$, its pre-image

$$\Pi^{-1}(\mathcal{E}) = \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau, \quad (4.10)$$

where $\mathcal{E}_\tau \triangleq \{(x_{\tau(1)}, \dots, x_{\tau(n)}) \mid (x_1, \dots, x_n) \in \mathcal{E}\}$. Therefore, for any two $\sigma, \sigma' \in \mathcal{S}_{n,\nu}$, we can write $\sigma' = \pi \circ \sigma$ for some $\pi \in \mathcal{S}_n$, and thus have

$$\begin{aligned} \mathbb{P}_{0;\sigma} \circ \Pi^{-1} \{\mathcal{E}\} &= \mathbb{P}_{0;\sigma} \left\{ \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \right\} \stackrel{(a)}{=} \mathbb{P}_{0;\sigma} \left\{ \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_{\tau \circ \pi} \right\} \\ &= \mathbb{P}_{0;\pi \circ \sigma} \left\{ \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \right\} = \mathbb{P}_{0;\sigma'} \circ \Pi^{-1} \{\mathcal{E}\}, \end{aligned}$$

where the equality (a) holds due to the following fact:

$$\forall \pi \in \mathcal{S}_n, \mathcal{S}_n \circ \pi \triangleq \{\tau \circ \pi \mid \tau \in \mathcal{S}_n\} = \mathcal{S}_n.$$

Following the same argument, $\tilde{\mathbb{P}}_1 \triangleq \mathbb{P}_{1;\sigma} \circ \Pi^{-1}$ does not depend on σ either.

Part 2 Second, let us we consider an auxiliary hypothesis testing problem on $\tilde{\mathcal{X}}^n$:

$$\begin{cases} \tilde{\mathcal{H}}_0 : Z \sim \tilde{\mathbb{P}}_0 \\ \tilde{\mathcal{H}}_1 : Z \sim \tilde{\mathbb{P}}_1, \end{cases} \quad (4.11)$$

and let $\tilde{\phi} : \tilde{\mathcal{X}}^n \rightarrow [0, 1]$ be a test with type-I and type-II error probabilities as follows:

$$\begin{cases} P_F(\tilde{\phi}) \triangleq \mathbb{E}_{\tilde{\mathbb{P}}_0} [\tilde{\phi}(Z)] \\ P_M(\tilde{\phi}) \triangleq \mathbb{E}_{\tilde{\mathbb{P}}_1} [1 - \tilde{\phi}(Z)]. \end{cases}$$

We claim that for any symmetric test $\phi(x^n) = \tilde{\phi}(\Pi(x^n))$ as defined in Definition 4.4.2, the following holds:

$$\begin{cases} P_F(\tilde{\phi}) = P_F(\phi) \\ P_M(\tilde{\phi}) = P_M(\phi). \end{cases}$$



To show this, note that a direct calculation gives

$$\begin{aligned}
P_F(\phi) &= \max_{\sigma} \mathbb{E}_{\mathbb{P}_{0;\sigma}} [\phi(X^n)] \\
&= \max_{\sigma} \mathbb{E}_{\mathbb{P}_{0;\sigma}} [\tilde{\phi}(\Pi(X^n))] \\
&= \max_{\sigma} \int \tilde{\phi}(\Pi(x^n)) \mathbb{P}_{0;\sigma}(dx^n) \\
&= \max_{\sigma} \int \tilde{\phi}(z) \mathbb{P}_{0;\sigma}(\Pi^{-1}(dz)) \\
&= \mathbb{E}_{\tilde{\mathbb{P}}_0} [\tilde{\phi}(Z)] = P_F(\tilde{\phi}).
\end{aligned}$$

For the same reason, $P_M(\phi) = P_M(\tilde{\phi})$. Therefore, for any symmetric test on \mathcal{X}^n , the corresponding $\tilde{\phi}$ has exactly the same type-I and type-II error probability. Notice that the auxiliary hypothesis testing problem (4.11) is simple, so by Neyman-Pearson lemma, we have readily seen that the optimal symmetric test on the original problem should be

$$\phi^*(x^n) = \begin{cases} 1, & \text{if } \ell'(x^n) > \tau \\ \gamma, & \text{if } \ell'(x^n) = \tau \\ 0, & \text{if } \ell'(x^n) < \tau, \end{cases}$$

where $\ell'(x^n)$ is defined as

$$\ell'(x^n) = \frac{\tilde{\mathbb{P}}_1(\Pi(x^n))}{\tilde{\mathbb{P}}_0(\Pi(x^n))} = \frac{\mathbb{P}_{1;\sigma} \{ \Pi^{-1}(\Pi(x^n)) \}}{\mathbb{P}_{0;\sigma} \{ \Pi^{-1}(\Pi(x^n)) \}}.$$

Part 3 Finally, we show that $\ell'(x^n)$ is indeed the mixture likelihood ratio $\ell(x^n)$, as defined in (4.5). With a slight abuse of notation, let $\Pi_{x^n} \triangleq \Pi^{-1}(\Pi(x^n)) = \{x_{\tau(1)}, \dots, x_{\tau(n)} \mid \tau \in \mathcal{S}_n\}$. In words, Π_{x^n} is the collection of x^n and all its permutations. We observe that

$$\begin{aligned}
\mathbb{P}_{1;\sigma} \{ \Pi^{-1}(\Pi(x^n)) \} &= \sum_{y^n \in \Pi_{x^n}} \mathbb{P}_{1;\sigma}(y^n) \\
&\stackrel{(a)}{=} \left(\sum_{\tau \in \mathcal{S}_n} \mathbb{P}_{1;\sigma}(\tau(x^n)) \right) c_1(x^n)
\end{aligned}$$

$$\stackrel{(b)}{=} \left(\sum_{\sigma' \in \mathcal{S}_{n,\nu}} \mathbb{P}_{1;\sigma'}(x^n) \right) c_1(x^n) c_2(\sigma).$$



The constant $c_1(x^n)$ in (a) is due to the fact that $x^n = (x_1, \dots, x_n)$ might not be all distinct, so summing over the set $\{\tau(x^n) \mid \tau \in \mathcal{S}_n\}$ may count an element $y^n \in \Pi_{x^n}$ multiple times. Note that if x^n are all distinct, then $c_1(x^n) = 1$. (b) holds because $\mathbb{P}_{1;\sigma}(\tau(x^n)) = \mathbb{P}_{1;\sigma \circ \tau}(x^n)$ and $\mathcal{S}_{n,\nu} = \mathcal{S}_{n,\nu} \circ \tau \triangleq \{\sigma \circ \tau \mid \sigma \in \mathcal{S}_{n,\nu}\}$. Again, the summation counts σ repeatedly, so we normalize by the constant $c_2(\sigma)$. Following the same reason,

$$\mathbb{P}_{0;\sigma} \{ \Pi^{-1}(\Pi(x^n)) \} = \left(\sum_{\sigma' \in \mathcal{S}_{n,\nu}} \mathbb{P}_{0;\sigma'}(x^n) \right) c_1(x^n) c_2(\sigma).$$

Hence,

$$\begin{aligned} \ell'(x^n) &= \frac{\mathbb{P}_{1;\sigma} \{ \Pi^{-1}(\Pi(x^n)) \}}{\mathbb{P}_{0;\sigma} \{ \Pi^{-1}(\Pi(x^n)) \}} \\ &= \frac{\left(\sum_{\sigma' \in \mathcal{S}_{n,\nu}} \mathbb{P}_{1;\sigma'}(x^n) \right) c_1(x^n) c_2(\sigma)}{\left(\sum_{\sigma' \in \mathcal{S}_{n,\nu}} \mathbb{P}_{0;\sigma'}(x^n) \right) c_1(x^n) c_2(\sigma)} \\ &= \frac{\sum_{\sigma} \mathbb{P}_{1;\sigma}(x^n)}{\sum_{\sigma} \mathbb{P}_{0;\sigma}(x^n)} = \ell(x^n), \end{aligned}$$

which establishes the claim.

■

Lemma 4.4.2 For any general (measurable) test $\psi(x^n) : \mathcal{X}^n \rightarrow [0, 1]$, there exists a symmetric test $\phi(x^n)$ whose performance is not worse than ψ . That is,

$$\begin{cases} P_F(\phi) \leq P_F(\psi) \\ P_M(\phi) \leq P_M(\psi). \end{cases} \quad (4.12)$$

proof of Lemma 4.4.2. With a slight abuse of notation, let $\tau(x^n)$ denote the coordinate-permutation function with respect to $\tau \in \mathcal{S}_n$, i.e. $\tau(x^n) = (x_{\tau(1)}, \dots, x_{\tau(n)})$. Then we

construct $\phi(x^n)$ as follows:

$$\phi(x^n) \triangleq \frac{1}{n!} \sum_{\tau \in \mathcal{S}_n} \psi \circ \tau(x^n).$$



We claim the following two facts:

1. $\phi(x^n)$ is symmetric, and thus can be written as $\tilde{\phi} \circ \Pi(x^n)$ for some $\tilde{\mathcal{F}}$ -measurable $\tilde{\phi}$.
2. (4.12) holds for the constructed ϕ .

Part 1 To see that $\phi(x^n) = \tilde{\phi} \circ \Pi(x^n)$, we observe that for any $y^n, z^n \in \Pi^{-1}(\tilde{x}^n)$, there exists a permutation $\pi \in \mathcal{S}_n$ such that $y^n = \pi(z^n)$. Hence it suffices to verify that for all $\pi \in \mathcal{S}_n$, $\phi(x^n) = \phi(\pi(x^n))$.

$$\begin{aligned} \phi(\pi(x^n)) &= \frac{1}{n!} \sum_{\tau \in \mathcal{S}_n} \psi \circ \tau(\pi(x^n)) \\ &= \frac{1}{n!} \sum_{\tau \in \mathcal{S}_n} \psi \circ \tau \circ \pi(x^n) \\ &\stackrel{(a)}{=} \frac{1}{n!} \sum_{\tau' \in \mathcal{S}_n} \psi \circ \tau'(x^n) = \phi(x^n). \end{aligned}$$

The equality (a) holds due to the fact that

$$\mathcal{S}_n \circ \pi \triangleq \{\tau \circ \pi \mid \tau \in \mathcal{S}_n\} = \mathcal{S}_n.$$

Therefore, $\phi(x^n)$ can be decomposed into $\tilde{\phi} \circ \Pi(x^n)$.

Next, we check the measurability of $\tilde{\phi}$. Notice that ϕ is \mathcal{F}^{\otimes} -measurable, since both ψ and τ are measurable. The measurability of τ follows from the τ -permuted closedness assumption of $\mathcal{F}^{\otimes n}$:

$$\forall \mathcal{A} \in \mathcal{F}^{\otimes n}, \mathcal{A}_\tau \triangleq \{\tau(x^n) \mid x^n \in \mathcal{A}\} \in \mathcal{F}^{\otimes n}.$$

Observe that for all Borel-measurable set \mathcal{B} , we have

$$\phi^{-1}\{\mathcal{B}\} = \Pi^{-1}\left\{\tilde{\phi}^{-1}\{\mathcal{B}\}\right\} \in \mathcal{F}^{\otimes n} \Leftrightarrow \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \in \mathcal{F}^{\otimes n},$$



where we use \mathcal{E} to denote event $\tilde{\phi}^{-1}\{\mathcal{B}\}$, and \mathcal{E}_τ to denote the τ -permuted event of \mathcal{E} , as defined in (4.4). Notice here we use the fact given by (4.10). Therefore it suffices to check

$$\forall \mathcal{E} \subseteq \tilde{\mathcal{X}}^n, \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \in \mathcal{F}^{\otimes n} \Rightarrow \mathcal{E} \in \mathcal{F}^{\otimes n} \cap \tilde{\mathcal{X}}^n = \tilde{\mathcal{F}}.$$

We claim that indeed,

$$\left\{ \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \right\} \cap \tilde{\mathcal{X}}^n = \mathcal{E},$$

for every $\mathcal{E} \subseteq \tilde{\mathcal{X}}^n$. This is because

1. Since $\mathcal{E} \subseteq \tilde{\mathcal{X}}^n$, we have $\mathcal{E} = \mathcal{E} \cap \tilde{\mathcal{X}}^n \subseteq \left\{ \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \right\} \cap \tilde{\mathcal{X}}^n$.
2. For any τ and for any $x^n \in \mathcal{E}_\tau \cap \tilde{\mathcal{X}}^n$, $x^n \in \mathcal{E}$. Hence, $\forall \tau \in \mathcal{S}_n$, $\mathcal{E}_\tau \cap \tilde{\mathcal{X}}^n \subseteq \mathcal{E}$, that is, $\left\{ \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \right\} \cap \tilde{\mathcal{X}}^n \subseteq \mathcal{E}$.

Hence,

$$\bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \in \mathcal{F}^{\otimes n} \implies \mathcal{E} = \left\{ \bigcup_{\tau \in \mathcal{S}_n} \mathcal{E}_\tau \right\} \cap \tilde{\mathcal{X}}^n \in \mathcal{F}^{\otimes n} \cap \tilde{\mathcal{X}}^n = \tilde{\mathcal{F}},$$

showing that $\tilde{\phi}$ is $\tilde{\mathcal{F}}$ -measurable.

Part 2 We show that $\phi(x^n)$ cannot be worse than $\psi(x^n)$. Observe that for all $\tau \in \mathcal{S}_n$, we have

$$\begin{aligned} P_F(\psi \circ \tau) &= \max_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{E}_{\mathbb{P}_{0;\sigma}} [\psi(\tau(X^n))] = \max_{\sigma \in \mathcal{S}_{n,\nu}} \mathbb{E}_{\mathbb{P}_{0;\sigma \circ \tau^{-1}}} [\psi(X^n)] \\ &= \max_{\sigma' \in \mathcal{S}_{n,\nu}} \mathbb{E}_{\mathbb{P}_{0;\sigma'}} [\psi(X^n)] = P_F(\psi). \end{aligned}$$

Again, the third equality holds due to the fact

$$\mathcal{S}_{n,\nu} \circ \tau^{-1} \triangleq \{\sigma \circ \tau^{-1} \mid \sigma \in \mathcal{S}_{n,\nu}\} = \mathcal{S}_{n,\nu}.$$

Therefore, we have

$$\begin{aligned} P_F(\phi) &= \max_{\sigma} \mathbb{E}_{\mathbb{P}_{0,\sigma}} \left[\frac{1}{n!} \sum_{\tau \in \mathcal{S}_n} \psi \circ \tau(X^n) \right] \\ &\leq \frac{1}{n!} \sum_{\tau \in \mathcal{S}_n} \max_{\sigma} \mathbb{E}_{\mathbb{P}_{0,\sigma}} [\psi \circ \tau(X^n)] \\ &= \frac{1}{n!} \sum_{\tau \in \mathcal{S}_n} P_F(\psi \circ \tau) = P_F(\psi). \end{aligned}$$



Following the same argument, we obtain $P_M(\phi) \leq P_M(\psi)$, and the proof completes.

■

Finally, the proof of Theorem 4.3.1 directly follows from Lemma 4.4.1 and Lemma 4.4.2.

Proof of Theorem 4.3.1. From Lemma 4.4.2, we only need to consider symmetric tests. From Lemma 4.4.1, we see that the optimal test among all symmetric tests is the mixture likelihood test, as defined in (4.6). This establishes Theorem 4.3.1. ■

Remark 4.4.3 Notice that in the above proof, we do not make use of assumptions on the distribution of X^n , such as independence. Indeed, the proof indicates that for the anonymous composite hypothesis testing problem, under the minimax criterion (i.e. to minimize the worst case error), we should always design tests based on the empirical distribution of X^n (i.e. as a function of $\Pi(x^n)$). This principle also holds for other statistical inference problems, such as M -ary hypothesis testing.

4.5 Proof of Theorem 4.3.2

For the case $|\mathcal{X}| < \infty$, the auxiliary space $\tilde{\mathcal{X}}$ is equivalent to the space of all probability measures on \mathcal{X} , that is, $\mathcal{P}_{\mathcal{X}}$, and the mapping $\Pi(x^n)$ maps a sequence of samples to its type Π_{x^n} . According to Lemma 4.4.2, the optimal test is symmetric, which implies that we only need to consider tests depending on the type. For tests depending only on the empirical distribution, it is natural to view their acceptance region as a collection of empirical distribution, that is, a (measurable) subset of $\mathcal{P}_{\mathcal{X}}$. This motivates us to apply Sanov's theorem. We begin with the following generalization of Sanov's result:

Lemma 4.5.1 (Generalized Sanov Theorem) Let $|\mathcal{X}| < \infty$, and $\Gamma \subseteq \mathcal{P}_{\mathcal{X}}$ be a collection of distributions on \mathcal{X} . Then for all $\sigma \in \mathcal{S}_{n,\nu}$ and $\theta \in \{0, 1\}$, we have

$$- \inf_{\substack{[U_1 \dots U_K]^{\top} \in (\mathcal{P}_{\mathcal{X}})^K \\ \alpha^{\top} U \in \text{int } \Gamma}} \sum_{k=1}^K \alpha_k D(U_k \| P_{\theta;k}) \quad (4.13)$$

$$\leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\theta;\sigma} \{ \Pi_{x^n} \in \Gamma \} \quad (4.14)$$

$$\leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\theta;\sigma} \{ \Pi_{x^n} \in \Gamma \} \quad (4.15)$$

$$\leq - \inf_{\substack{[U_1 \dots U_K]^{\top} \in (\mathcal{P}_{\mathcal{X}})^K \\ \alpha^{\top} U \in \text{cl } \Gamma}} \sum_{k=1}^K \alpha_k D(U_k \| P_{\theta;k}), \quad (4.16)$$

where in taking the limits, we assume that $\lim_{n \rightarrow \infty} \frac{n_k}{n} = \alpha_k$, for all $k = 1, \dots, K$. In particular, if the infimum in the right-hand side is equal to the infimum in the left-hand side, then we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\theta;\sigma} \{ \Pi_{x^n} \in \Gamma \} = - \inf_{\substack{[U_1 \dots U_K]^{\top} \in (\mathcal{P}_{\mathcal{X}})^K \\ \alpha^{\top} U \in \text{cl } \Gamma}} \sum_k \alpha_k D(U_k \| P_{\theta;k}).$$

The proof is a direct extension of Lemma 2.1.3, except that we replace the i.i.d. measure with the product of independent non-identical ones, $\mathbb{P}_{\theta;\sigma}$. For the detailed proof, please refer to Appendix B.2.

Motivated by the generalized Sanov Theorem, we further define the following generalized divergence to measure *how far* from one set of distributions $\mathbf{Q} \triangleq [Q_1 \dots Q_K]^{\top}$ to another set of distributions $\mathbf{P} \triangleq [P_1 \dots P_K]^{\top}$:

Definition 4.5.1 Let $\mathbf{P} = [P_1 \dots P_K]^{\top}$ and $\mathbf{Q} = [Q_1 \dots Q_K]^{\top}$ are both in $(\mathcal{P}_{\mathcal{X}})^K$. Let $\alpha = [\alpha_1 \dots \alpha_K]^{\top}$ be a K -tuple probability vector. Define

$$D_{\alpha}(\mathbf{P}; \mathbf{Q}) \triangleq \inf_{U \in (\mathcal{P}_{\mathcal{X}})^K} \sum_{k=1}^K \alpha_k D(U_k \| Q_k) \quad (4.17)$$

subject to $\alpha^{\top} U = \alpha^{\top} \mathbf{P}$

Thus (4.13) in Lemma 4.5.1 can be rewritten as



$$\begin{aligned}
& - \inf_{\alpha^\top U \in \text{int } \Gamma} D_\alpha(U; \mathbf{P}_\theta) \\
& \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\theta; \sigma} \{ \Pi_{x^n} \in \Gamma \} \\
& \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\theta; \sigma} \{ \Pi_{x^n} \in \Gamma \} \\
& \leq - \inf_{\alpha^\top U \in \text{cl } \Gamma} D_\alpha(U; \mathbf{P}_\theta).
\end{aligned}$$

Also, the result of Theorem 4.3.2, (4.7), is equivalent to the following statement:

$$E^*(\epsilon, \alpha) = D_\alpha(\mathbf{P}_0; \mathbf{P}_1).$$

Remark 4.5.1 *Intuitively, $D_\alpha(\mathbf{P}; \mathbf{Q})$ measures how far between \mathbf{P} and \mathbf{Q} . However, $D_\alpha(\cdot; \cdot)$ is not a divergence, since $D_\alpha(\mathbf{P}; \mathbf{Q}) = 0$ does not always imply $\mathbf{P} = \mathbf{Q}$.*

Notice that for any fixed $\mathbf{Q} \in (\mathcal{P}_\mathcal{X})^K$, $D_\alpha(\mathbf{P}; \mathbf{Q})$ can be regarded as a function of \mathbf{P} . Moreover, this function depends only on the mixture of \mathbf{P} , say, $\alpha^\top \mathbf{P}$. Therefore, for notional convenience, let us use $f_Q(\cdot) : \mathcal{P}_\mathcal{X} \rightarrow \mathbb{R} \cup \{+\infty\}$ to denote this function:

$$\begin{aligned}
f_Q(T) & \triangleq \inf_{U \in (\mathcal{P}_\mathcal{X})^K} \sum_{k=1}^K \alpha_k D(U_k \| Q_k) \\
& \text{subject to } \alpha^\top U = T
\end{aligned}$$

In other words,

$$f_Q(\alpha^\top \mathbf{P}) = D_\alpha(\mathbf{P}; \mathbf{Q}).$$

Before entering the main proof of Theorem 4.3.2, let us introduce some properties of $f_Q(\cdot)$.

Lemma 4.5.2 *Let $\mathbf{Q} \in (\mathcal{P}_\mathcal{X})^K$ and $f_Q(\cdot) : \mathcal{P}_\mathcal{X} \rightarrow \mathbb{R} \cup \{+\infty\}$ be defined as Definition 4.5.1 and above. Then,*

$$1. f_Q(\alpha^\top \mathbf{Q}) = 0$$



2. The collection of all $T \in \mathcal{P}_{\mathcal{X}}$ such that $f_Q(T) < \infty$, denoted as

$$\mathcal{C}_Q \triangleq \{T \in \mathcal{P}_{\mathcal{X}} : f_Q(T) < \infty\},$$

is a compact, convex subset of $\mathcal{P}_{\mathcal{X}}$.

3. $f_Q(T)$ is a convex, continuous function of T on \mathcal{C}_Q (and by the compactness of \mathcal{C}_Q , $f_Q(T)$ is also uniformly continuous).

Proof of Lemma 4.5.2 can be found in Appendix B.3.

proof of Theorem 4.3.2.

Part 1 (Achievability) Let $\delta > 0$ and consider the test :

$$\phi(x^n) \triangleq \mathbb{1}_{\{x^n : D(\Pi_{x^n} \| M_0(\alpha)) > \delta\}}.$$

Denote the acceptance region of ϕ as $\Gamma \triangleq \{T \in \mathcal{P}_{\mathcal{X}} : D(T \| M_0(\alpha)) > \delta\}$. Then the exponent of type-I error probability $P_F(\phi)$ can be bounded by

$$\begin{aligned} & \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{E}_{\mathbb{P}_{0;\sigma}} [\phi(X^n)] \\ &= \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{0;\sigma} \{\Pi_{x^n} \in \Gamma\} \\ &\stackrel{(a)}{\geq} \inf_{T \in \text{cl } \Gamma} f_{P_0}(T) \\ &\stackrel{(b)}{\geq} \delta, \end{aligned}$$

where (a) holds by Lemma 4.5.1, and (b) holds due to the convexity of KL divergence:

$$\begin{aligned} D(T \| M_0(\alpha)) &\leq \min_{U \in (\mathcal{P}_{\mathcal{X}})^K} \sum_{k=1}^K \alpha_k D(U_k \| P_{0;k}) = f_{P_0}(T) \\ &\text{subject to } \alpha^\top U = T \end{aligned}$$

Notice that for any $\delta > 0$, as n large enough, we must have

$$P_F(\phi) < \epsilon.$$

On the other hand, the exponent of type-II error probability $\underline{E}^*(\epsilon, \alpha)$ can be bounded by

$$\begin{aligned} & \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{E}_{\mathbb{P}_{1;\sigma}} [\phi(X^n)] \\ &= \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{1;\sigma} \{X^n : D(\Pi_{X^n} \| M_0(\alpha)) \leq \delta\} \\ &\geq \inf_{T \in cl \Gamma^c} f_{P_1}(T), \end{aligned} \tag{4.18}$$

By Pinsker's inequality (Theorem 6.5 in [32]), we have

$$cl(\Gamma^c) = \{T \in \mathcal{P}_{\mathcal{X}} : D(T \| M_0(\alpha)) \leq \delta\} \subseteq \left\{T \in \mathcal{P}_{\mathcal{X}} : \|T - M_0(\alpha)\|_1 \leq \sqrt{2\delta}\right\} \triangleq B_{\sqrt{2\delta}}(M_0(\alpha)),$$

so (4.18) can be further lower bounded by

$$\inf_{T \in cl \Gamma^c} f_{P_1}(T) \geq \inf_{T \in B_{\sqrt{2\delta}}(M_0(\alpha))} f_{P_1}(T).$$

Also, by the continuity (Lemma 4.5.2) of $f_{P_1}(\cdot)$,

$$\inf_{T \in B_{\sqrt{2\delta}}(M_0(\alpha))} f_{P_1}(T) = f_{P_1}(M_0(\alpha)) + \Delta(\delta),$$

with

$$\lim_{\delta \rightarrow 0} \Delta(\delta) = 0.$$

Finally, since δ can be chosen arbitrarily small, we have

$$\underline{E}^*(\epsilon, \alpha) \geq f_{P_1}(M_0(\alpha)) = D_{\alpha}(P_0; P_1). \tag{4.19}$$

Part 2 (Converse) We have shown that symmetric test is optimal in Lemma 4.4.2. Hence, in the following, it suffices to consider symmetric tests.

For an arbitrary symmetric test $\psi : \mathcal{P}_n \rightarrow [0, 1]$ such that its type-I error probability $P_F(\psi) < \epsilon$, we shall lower bound its type-II error probability as follows. Let $\mathcal{A}^{(n)} \triangleq \{T \in \mathcal{P}_n : \psi(T) \leq 1/2\}$, and recall that

$$\tilde{\mathbb{P}}_0 \triangleq \mathbb{P}_{0;\sigma} \circ \Pi^{-1}$$



is a probability measure independent of σ . Then, we have

$$\begin{aligned} \epsilon > \mathbb{E}_{\tilde{\mathbb{P}}_0} [\psi(T)] &= \sum_{T \in \mathcal{P}_n} \tilde{\mathbb{P}}_0(T) \psi(T) \geq \sum_{T \in (\mathcal{A}^{(n)})^c} \tilde{\mathbb{P}}_0(T) \psi(T) \\ &\stackrel{(a)}{>} \frac{1}{2} \sum_{T \in (\mathcal{A}^{(n)})^c} \tilde{\mathbb{P}}_0(T) = \frac{1}{2} \left(1 - \tilde{\mathbb{P}}_0 \{ \mathcal{A}^{(n)} \} \right), \end{aligned}$$

(a) holds since for all $T \notin \mathcal{A}^{(n)}$, $\psi(T) > 1/2$. In other words, we have

$$\tilde{\mathbb{P}}_0 \{ \mathcal{A}^{(n)} \} > 1 - 2\epsilon.$$

On the other hand, let $\mathcal{B}^{(n)} \triangleq \{T \in \mathcal{P}_n \mid D(T \| M_0(\boldsymbol{\alpha})) \leq \delta\}$. Then, according to the analysis in type-I error probability in the achievability part, we have

$$\tilde{\mathbb{P}}_0 \{ \mathcal{B}^{(n)} \} > 1 - \epsilon.$$

Applying union bound, we see that

$$\tilde{\mathbb{P}}_0 \{ \mathcal{A}^{(n)} \cap \mathcal{B}^{(n)} \} > 1 - 3\epsilon,$$

and hence for $\epsilon < \frac{1}{3}$, $\mathcal{A}^{(n)} \cap \mathcal{B}^{(n)}$ is non-empty.

Let $V_n^* \in \mathcal{A}^{(n)} \cap \mathcal{B}^{(n)}$ and define $\tilde{\mathbb{P}}_1 \triangleq \mathbb{P}_{1;\sigma} \circ \Pi^{-1}$ (which is also independent of σ).

Again we have

$$\begin{aligned} P_F(\psi) &= \mathbb{E}_{\tilde{\mathbb{P}}_1} [1 - \psi(T)] \\ &\geq \sum_{T \in \mathcal{A}^{(n)}} (1 - \psi(T)) \tilde{\mathbb{P}}_1 \{T\} \\ &\geq \frac{1}{2} \tilde{\mathbb{P}}_1 \{V_n^*\}. \end{aligned}$$

We further estimate $\tilde{\mathbb{P}}_1 \{V_n^*\}$ by

$$\tilde{\mathbb{P}}_1 \{V_n^*\} = \mathbb{P}_{1;\sigma} \{T_n(V_n^*)\}$$

$$\begin{aligned}
&= \sum_{\substack{U_k \in \mathcal{P}_{n_k}: \\ \sum_k \alpha_k U_k = V_n^*}} \prod_{k=1}^K P_{1;k}^{\otimes n_k} \{T_{n_k}(U_k)\} \\
&= \sum_{\substack{U_k \in \mathcal{P}_{n_k}: \\ \sum_k \alpha_k U_k = V_n^*}} 2^{-\sum_k n_k D(U_k \| P_{1;k})} \\
&\geq \max_{\substack{U_k \in \mathcal{P}_{n_k}: \\ \sum_k \alpha_k U_k = V_n^*}} 2^{-\sum_k n_k D(U_k \| P_{1;k})} \\
&= 2^{-n \tilde{D}_n},
\end{aligned}$$



where

$$\tilde{D}_n \triangleq \min_{\substack{U_k \in \mathcal{P}_{n_k}: \\ \sum_k \alpha_k U_k = V_n^*}} \left(\sum_k \frac{n_k}{n} D(U_k \| P_{1;k}) \right).$$

Notice that since $V_n^* \in \mathcal{B}^{(n)}$, so we have

$$D(V_n^* \| M_0(\boldsymbol{\alpha})) \leq \delta.$$

Since δ can be chosen arbitrarily small, as $\delta \rightarrow 0$ and $n \rightarrow \infty$ (with $\frac{n_k}{n} \rightarrow \alpha_k$), we have

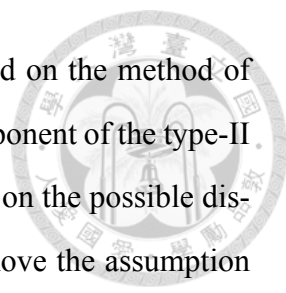
$$\begin{aligned}
\overline{E}^*(\epsilon, \boldsymbol{\alpha}) &\leq \lim_{n \rightarrow \infty} \tilde{D}_n \\
&= \min_{\substack{U_k \in \mathcal{P}_{\mathcal{X}}: \\ \sum_k \alpha_k U_k = M_0(\boldsymbol{\alpha})}} \left(\sum_k \alpha_k D(U_k \| P_{1;k}) \right) \\
&= f_{P_1}(M_0(\boldsymbol{\alpha})) \\
&= D_{\boldsymbol{\alpha}}(\mathbf{P}_0; \mathbf{P}_1),
\end{aligned}$$

which completes the proof.

■

4.6 Extension

Theorem 4.3.1 characterizes the optimal test in the anonymous detection problem, where only a few conditions on the σ -field \mathcal{F} are required. In Theorem 4.3.2, we further assume



the alphabet \mathcal{X} is finite, in order to apply large deviation tools based on the method of types (see Remark 4.3.4 for discussion). However, the optimal exponent of the type-II error probability, given by the result of Theorem 4.3.2, depends only on the possible distributions under \mathcal{H}_θ , and hence it is interesting to see if one can remove the assumption that \mathcal{X} being finite. Recall that in the proof, the main tool we employed is the generalized version of Sanov's theorem (see Lemma 4.5.1), and thus the question turns out to be whether it is possible to prove Lemma 4.5.1 without using method of types. Surprisingly, the answer is yes if \mathcal{X} is a Polish space (a completely separable metrizable topological space). If \mathcal{X} is Polish, the space of all probability measures on \mathcal{X} ($\mathcal{P}_\mathcal{X}$) is also Polish, equipped with weak-topology induced by weak convergence. One can choose, for example, Levy-Prokhorov metric on $\mathcal{P}_\mathcal{X}$. The proof of standard Sanov's Theorem on Polish \mathcal{X} , however, is far more complicated than the case of finite \mathcal{X} , see [16, 15] for detailed proof. Lemma 4.5.1 for Polish \mathcal{X} can be proved with similar techniques. Nevertheless, in order not to digress further from the subject, we only present a proof for finite \mathcal{X} in this paper.





Chapter 5

Anonymous Hypothesis Testing : Beyond Neyman-Pearson Regime

5.1 Problem Formulation

Unlike Neyman-Pearson formulation discussed in Chapter 4, in which we minimize the worst-case type-II error probability, subject to the worst-case type-I error probability not being larger than a constant ϵ . It is natural to extend the result from Chapter 4 to Chernoff's regime, where we aim to minimize the average probability of error:

$$P_e^{(n)}(\phi) \triangleq \pi_0 P_F^{(n)} + \pi_1 P_M^{(n)}.$$

Note that π_0 and π_1 are the prior distributions of \mathcal{H}_0 and \mathcal{H}_1 and do not scale with n . As suggested by Theorem 4.3.1, the optimal test is the mixture likelihood ratio test, so we only need to specify the corresponding threshold τ . However, the mixture likelihood ratio involves summation over $\mathcal{S}_{n,\nu}$, making the computation complexity extremely high. Even for the case $|\mathcal{X}| < \infty$, the computation still takes $\Theta(n^{|\mathcal{X}|})$ operations and thus is difficult to implement. To break the computational barrier, we propose an asymptotically optimal test, based on information projection, which achieves the optimal exponent of the average probability of error. Moreover, the result can be generalized to determine the achievable

exponent region \mathcal{R} , the collection of all achievable pairs of exponents:

$$\mathcal{R} \triangleq \left\{ (E_0, E_1) \mid \text{there exists a test } \phi, \text{ such that } P_F^{(n)}(\phi) \preceq 2^{-nE_0}, P_M^{(n)}(\phi) \preceq 2^{-nE_1} \right\},$$

where a sequence $a_n \preceq 2^{-nE_0}$ means a_n decays to zero at the rate faster than E_0 , that is,

$$-\liminf_{n \rightarrow \infty} \frac{1}{n} \log a_n \geq E_0.$$

5.2 Main Result

5.2.1 Efficient Test

As stated in Theorem 4.3.1, the optimal test $\phi^*(x^n)$ of the anonymous hypothesis testing problem was showed to be the *mixture likelihood ratio test*(4.6):

$$\phi^*(x^n) = \begin{cases} 1, & \text{if } \ell(x^n) > \tau \\ \gamma, & \text{if } \ell(x^n) = \tau \\ 0, & \text{if } \ell(x^n) < \tau \end{cases}$$

where the *mixture likelihood ratio* $\ell'(x^n)$ was defined in (4.5):

$$\ell(x^n) \triangleq \frac{\sum_{\sigma} \mathbb{P}_{1;\sigma}(x^n)}{\sum_{\sigma} \mathbb{P}_{0;\sigma}(x^n)}.$$

However, since the MLR sums over all permutations in $S_{n,\alpha}$, the computational complexity of the optimal test is $\Theta(n^K)$ and thus is obviously impossible to implement. Therefore, we propose an asymptotically optimal test based on convex programing.

Theorem 5.2.1 (Efficient Test) *Recall the function $f_P(T) : \mathcal{P}_{\mathcal{X}} \rightarrow \mathbb{R} \cup \{+\infty\}$ defined in Defintion 4.5.1. Consider the following test based on the function $f_{P_0}(\cdot)$ and $f_{P_1}(\cdot)$:*

$$\phi_{eff}(x^n) \triangleq \begin{cases} 0, & \text{if } f_{P_1}(\Pi_{x^n}) > f_{P_0}(\Pi_{x^n}) \\ 1, & \text{else } f_{P_1}(\Pi_{x^n}) \leq f_{P_0}(\Pi_{x^n}). \end{cases} \quad (5.1)$$

Then ϕ_{eff} is asymptotically optimal in Chernoff's regime. That is, for all priors π_0, π_1 , for all tests ϕ , and for all n large enough,

$$-\frac{1}{n} \log (P_e(\phi)) \leq -\frac{1}{n} \log (P_e(\phi_{\text{eff}})).$$



Remark 5.2.1 From the convexity of KL-divergence and the space $\mathcal{P}_{\mathcal{X}}$, the function $f_{\mathbf{P}}(\cdot)$ is indeed the minimization of a convex function. Hence the proposed test in Theorem 5.2.1 can be computed efficiently.

5.2.2 Achievable Region: An Information Geometric Perspective

To gain more insight on the function $f_{\mathbf{P}}(\cdot)$, we can write it in a "divergence" form, that is, $d(\cdot, \cdot) : \mathcal{P}_{\mathcal{X}} \times \mathcal{P}_{\mathcal{X}} \rightarrow \mathbb{R}$ is defined as

$$d(T, \mathbf{P}) \triangleq f_{\mathbf{P}}(T),$$

measuring the discrepancy between T and \mathbf{P} . As the result of Theorem 4.3.2, we know that $(d(\boldsymbol{\alpha}^\top \mathbf{P}_0, \mathbf{P}_1), 0)$ and $(d(\boldsymbol{\alpha}^\top \mathbf{P}_1, \mathbf{P}_0), 0)$ are two boundary points of \mathcal{R} , where $d(\cdot, \cdot)$ is the pseudo-distance function defined in (4.17). In this section, we aim to characterize \mathcal{R} .

Let us define the function

$$f(\beta) \triangleq \min_{F \in \mathcal{P}_{\mathcal{X}}} \frac{d(F, \mathbf{P}_0)}{\beta} + \frac{d(F, \mathbf{P}_1)}{1 - \beta}, \quad \beta \in [0, 1].$$

Then \mathcal{F} can be characterized by $f(\beta)$:

Theorem 5.2.2 (Achievable Region) \mathcal{R} is the region enclosed by $E_0 = 0$, $E_1 = 0$, and

$$(E_0(\beta), E_1(\beta)) = (\beta^2 (f'(\beta)(1 - \beta) - f(\beta)), (1 - \beta)^2 (\beta f'(\beta) + f(\beta))).$$

See Figure 5.1 for illustration.

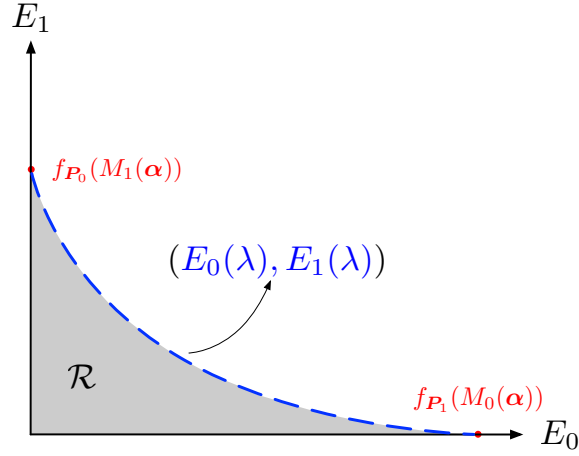


Figure 5.1: Illustration of achievable region \mathcal{R}

5.3 Proof of Theorem 5.2.1

Proof. Let us set some notations. For each $\mathbf{P} \in (\mathcal{P}_{\mathcal{X}})^K$, we use $B_r(\mathbf{P}) \subseteq \mathcal{P}_{\mathcal{X}}$ to denote the r -ball centered at T with respect to $f_{\mathbf{P}}(\cdot)$:

$$B_r(\mathbf{P}) \triangleq \{T \in \mathcal{P}_{\mathcal{X}} \mid f_{\mathbf{P}}(T) < r\}.$$

By the continuity of $f_{\mathbf{P}}(\cdot)$ (from Lemma 4.5.2), $B_r(\mathbf{P})$ is an open set. Then, define the largest packing radius between $\mathbf{P}_0, \mathbf{P}_1$ as follows:

$$r^* \triangleq \sup_r \{B_r(\mathbf{P}_0) \cap B_r(\mathbf{P}_1) = \emptyset\}.$$

See Figure 5.2 for illustration.

The rest of the proof will be organized as follows: we first show that ϕ_{eff} has error exponent at least r^* (the achievability part):

$$-\lim_{n \rightarrow \infty} \frac{1}{n} \log(P_e(\phi_{\text{eff}})) \geq r^*.$$

Then, we will prove that for all tests, the error exponent will be at most r^* (the converse part).

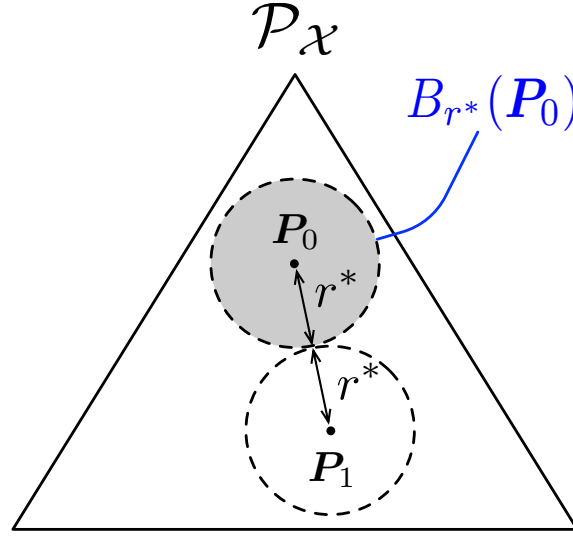


Figure 5.2: Illustration of $B_r(\cdot)$ and r^*

Part 1 (Achievability) Define

$$\mathcal{A} \triangleq \{T \in \mathcal{P}_X \mid f_{P_1}(T) \leq f_{P_0}(T)\},$$

and notice that

$$\begin{cases} P_F^{(n)}(\phi_{eff}) = \mathbb{P}_{0;\sigma} \{\Pi_{x^n} \in \mathcal{A}\} \\ P_M^{(n)}(\phi_{eff}) = \mathbb{P}_{1;\sigma} \{\Pi_{x^n} \in \mathcal{A}^c\}, \end{cases}$$

for any arbitrary σ (recall that ϕ_{eff} depends only on the empirical distribution and therefore is symmetrical, so the error is independent of the choice of a specific σ).

By the generalized Sanov's theorem (Lemma 4.5.1), we see that the exponent of $P_F^{(n)}(\phi_{eff})$ is lower bounded by $\inf_{T \in cl \mathcal{A}} f_{P_0}(T)$. Similarly, the exponent of $P_M^{(n)}(\phi_{eff})$ is lower bounded by $\inf_{T \in cl \mathcal{A}^c} f_{P_1}(T)$. It is not hard to see that indeed,

$$\inf_{T \in cl \mathcal{A}} f_{P_0}(T) = \inf_{T \in \mathcal{A}} f_{P_0}(T), \quad (5.2)$$

and

$$\inf_{T \in cl \mathcal{A}^c} f_{P_1}(T) = \inf_{T \in \mathcal{A}^c} f_{P_1}(T). \quad (5.3)$$

Equation (5.2) holds since \mathcal{A} is a closed set (it is a pre-image of a continuous function from a closed set), so $cl \mathcal{A} = \mathcal{A}$. For the equation (5.3), we notice that \mathcal{A}^c is open, and hence

the infimum of a continuous function on \mathcal{A}^c is actually equal to the infimum on $\text{cl } \mathcal{A}^c$.

Hence, it suffices to show that

$$\inf_{T \in \mathcal{A}} f_{P_0}(T) \geq r^*, \quad \inf_{T \in \mathcal{A}^c} f_{P_1}(T) \geq r^*.$$

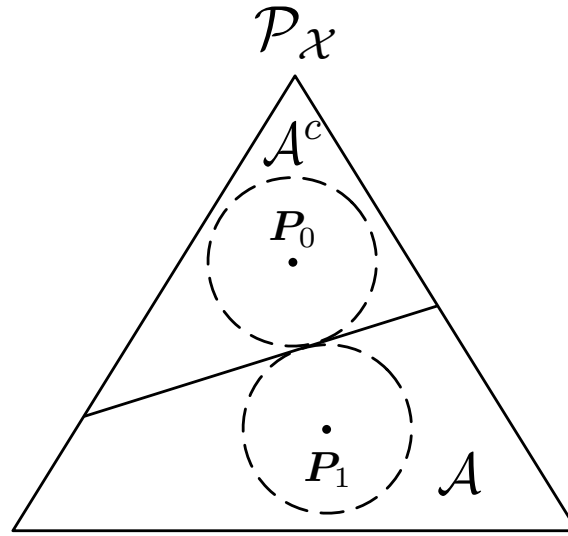


Figure 5.3: Relation between \mathcal{A} , \mathcal{A}^c and $B_{r^*}(P_0)$, $B_{r^*}(P_1)$

It is straightforward to see that \mathcal{A}^c contains $B_{r^*}(P_0)$ and \mathcal{A} contains $B_{r^*}(P_1)$, since we must have

$$1. \quad \forall T \in B_{r^*}(P_0), f_{P_0}(T) < f_{P_1}(T),$$

$$2. \quad \forall T \in B_{r^*}(P_1), f_{P_0}(T) > f_{P_1}(T).$$

Otherwise $B_{r^*}(P_0)$ intersects $B_{r^*}(P_1)$, violating our assumption on r^* . Also notice that \mathcal{A} , \mathcal{A}^c are disjoint, so

$$\mathcal{A}^c \cap B_{r^*}(P_0) = \mathcal{A} \cap B_{r^*}(P_1) = \emptyset,$$

implying that

$$\mathcal{A}^c \subseteq B_{r^*}(P_1)^c, \quad \mathcal{A} \subseteq B_{r^*}(P_0)^c.$$

Therefore, we have

$$\begin{cases} \inf_{T \in \mathcal{A}} f_{P_0}(T) \geq \inf_{T \in B_{r^*}(\mathbf{P}_0)^c} f_{P_0}(T) \geq r^* \\ \inf_{T \in \mathcal{A}^c} f_{P_1}(T) \geq \inf_{T \in B_{r^*}(\mathbf{P}_1)^c} f_{P_1}(T) \geq r^*, \end{cases}$$



proving the achievability part.

Part 2 (Converse) We show that for any test $\phi^{(n)}$, the exponent of the average probability of error greater than r^* leads to contradiction. Suppose the type-I and type-II error exponents of $\phi^{(n)}$ are r_1, r_2 respectively, and $r_1 > r^*, r_2 > r^*$. By Lemma 4.4.2, we only need to consider symmetric tests, that is, tests depend only on the type. Therefore, we can write the acceptance region of $\mathcal{H}_0, \mathcal{H}_1$ as

$$\begin{cases} \mathcal{B}_1^{(n)} = \{\Pi_{x^n} : \phi^{(n)}(x^n) = 1\} \\ \mathcal{B}_0^{(n)} = \{\Pi_{x^n} : \phi^{(n)}(x^n) = 0\}. \end{cases}$$

The exponents of type-I and type-II errors thus are greater than r_1, r_2 respectively, we have

$$\begin{cases} \liminf_{n \rightarrow \infty} \left\{ \min_{T \in \mathcal{B}_1^{(n)}} f_{P_0}(T) \right\} = r_1 > r^* \\ \liminf_{n \rightarrow \infty} \left\{ \min_{T \in \mathcal{B}_0^{(n)}} f_{P_1}(T) \right\} = r_2 > r^*. \end{cases} \quad (5.4)$$

Define $\min\{r_1, r_2\} = \tilde{r}$, and $\delta \triangleq (\tilde{r} - r^*)/2 > 0$. By (5.4), there exists M large enough, such that for all $n > M$,

$$\begin{cases} \min_{T \in \mathcal{B}_1^{(n)}} f_{P_0}(T) > \tilde{r} - \delta > r^* \\ \min_{T \in \mathcal{B}_0^{(n)}} f_{P_1}(T) > \tilde{r} - \delta > r^*. \end{cases}$$

We further define

$$\begin{cases} \mathcal{B}_1 = \bigcup_{n > M} \mathcal{B}_1^{(n)} \\ \mathcal{B}_0 = \bigcup_{n > M} \mathcal{B}_0^{(n)}. \end{cases}$$



We see that

1. $\mathcal{B}_0 \cup \mathcal{B}_1$ are dense in \mathcal{P}_X , since

$$\mathcal{B}_0^{(n)} \cup \mathcal{B}_1^{(n)} = \mathcal{P}_n,$$

and $\bigcup_{n>M} \mathcal{P}_n$ is dense in \mathcal{P}_X . So we have

$$(cl \mathcal{B}_0 \cup cl \mathcal{B}_1)^c = (cl \mathcal{B}_0)^c \cap (cl \mathcal{B}_1)^c = \emptyset. \quad (5.5)$$

2. By construction,

$$\begin{cases} \inf_{T \in \mathcal{B}_1} f_{P_0}(T) = \min_{T \in cl \mathcal{B}_1} f_{P_0}(T) > \tilde{r} - \delta > r^* \\ \inf_{T \in \mathcal{B}_0} f_{P_1}(T) = \min_{T \in cl \mathcal{B}_0} f_{P_1}(T) > \tilde{r} - \delta > r^*. \end{cases} \quad (5.6)$$

From (5.6), we have

$$\begin{cases} B_{(\tilde{r}-\delta)}(\mathbf{P}_0) \subseteq (cl \mathcal{B}_1)^c \\ B_{(\tilde{r}-\delta)}(\mathbf{P}_1) \subseteq (cl \mathcal{B}_0)^c, \end{cases}$$

and by (5.5) $B_{(\tilde{r}-\delta)}(\mathbf{P}_0) \cap B_{(\tilde{r}-\delta)}(\mathbf{P}_1) = \emptyset$. However, this violates our assumption that r^* is the supreme of radius such that the two sets do not overlap. This proves the converse part.

■

5.4 Proof of Theorem 5.2.2

proof of Theorem 5.2.2. Part 1 (Achievability) We first claim that for all $\beta \in [0, 1]$, there exists a test ϕ , such that

$$\frac{E_0(\phi)}{\beta} + \frac{E_1(\phi)}{1-\beta} \geq f(\beta),$$

where $(E_0(\phi), E_1(\phi))$ are the type-I and type-II error exponents respectively:

$$(E_0(\phi), E_1(\phi)) = \left(-\lim_{n \rightarrow \infty} \frac{1}{n} \log P_F^{(n)}(\phi), -\lim_{n \rightarrow \infty} \frac{1}{n} \log P_M^{(n)}(\phi) \right).$$

Define

$$F^* = \arg \min_{F \in \mathcal{P}_X} \frac{d(F, \mathbf{P}_0)}{\beta} + \frac{d(F, \mathbf{P}_1)}{1 - \beta},$$

and consider the test

$$\phi_{F^*}(x^n) = \mathbb{1}_{\{d(\Pi_{x^n}, \mathbf{P}_1) \leq d(F^*, \mathbf{P}_1)\}}.$$

The acceptance region of ϕ_{F^*} is a closed ball centered at \mathbf{P}_1 with radius equal to $d(\Pi_{x^n}, \mathbf{P}_1)$.

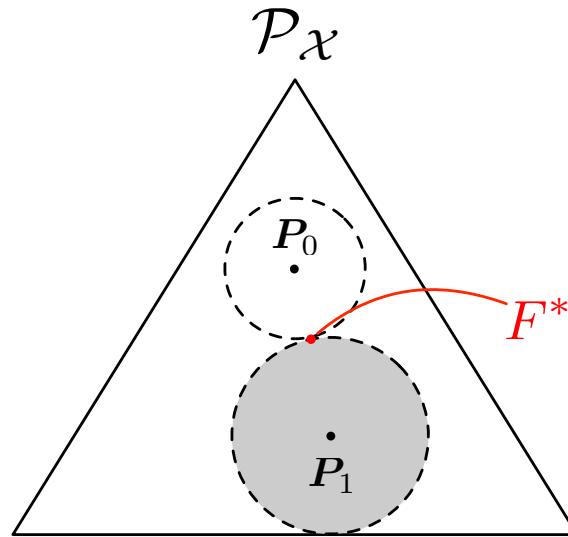


Figure 5.4: Acceptance region of ϕ_{F^*}

By Sanov's theorem,

$$E_1(\phi_{F^*}) = \min_{T \in \mathcal{A}_1^c} d(T, \mathbf{P}_1) = d(F^*, \mathbf{P}_1).$$

To prove the achievability, it suffices to show that

$$E_0(\phi_{F^*}) = \min_{T \in \mathcal{A}_1} d(T, \mathbf{P}_0) = d(F^*, \mathbf{P}_0).$$

Notice that $\min_{T \in \mathcal{A}_1} d(T, \mathbf{P}_0)$ cannot be smaller than $d(F^*, \mathbf{P}_0)$; otherwise there exists

another $\tilde{F} \in \mathcal{A}_1$, such that

$$\frac{d(\tilde{F}, \mathbf{P}_0)}{\beta} + \frac{d(\tilde{F}, \mathbf{P}_1)}{1 - \beta} < \frac{d(F^*, \mathbf{P}_0)}{\beta} + \frac{d(F^*, \mathbf{P}_1)}{1 - \beta},$$

violating the assumption that F^* achieves minimum. Thus we have

$$\min_{T \in \mathcal{A}_1} d(T, \mathbf{P}_0) \geq d(F^*, \mathbf{P}_0).$$

Part 2 (Converse)

For the converse part, following the same idea in Theorem 5.2.1, we show that for all test ϕ ,

$$\frac{E_0(\phi)}{\beta} + \frac{E_1(\phi)}{1 - \beta} \leq f(\beta).$$

We prove by contradiction. Let ϕ^* be a test such that

$$\frac{E_0(\phi^*)}{\beta} + \frac{E_1(\phi^*)}{1 - \beta} > f(\beta),$$

and denote the acceptance region of ϕ^* as \mathcal{B} . Then as Sanov's theorem suggesting, we have

$$\begin{cases} E_0(\phi^*) = \min_{T \in \mathcal{B}} d(T, \mathbf{P}_0), \\ E_1(\phi^*) = \min_{T \in \mathcal{B}^c} d(T, \mathbf{P}_1). \end{cases} \quad (5.7)$$

Assume that the minimum of (5.7) achieves by T_0^*, T_1^* respectively :

$$\begin{cases} T_1^* = \arg \min_{T \in \mathcal{B}} d(T, \mathbf{P}_0), \\ T_0^* = \arg \min_{T \in \mathcal{B}^c} d(T, \mathbf{P}_1), \end{cases}$$

we have

$$\frac{d(T_1^*, \mathbf{P}_0)}{\beta} + \frac{d(T_0^*, \mathbf{P}_1)}{1 - \beta} > f(\beta) = \min_{F \in \mathcal{P}_{\mathcal{X}}} \frac{d(F, \mathbf{P}_0)}{\beta} + \frac{d(F, \mathbf{P}_1)}{1 - \beta}. \quad (5.8)$$

We claim that this cannot happen.



First, notice that T_0^* and T_1^* must appear in the boundary. More precisely, we can find a sequence $T_0^{(i)} \in \mathcal{B}$, and $T_0^{(i)} \rightarrow T_0^*$. To see this, we claim that

$$d((1 - \lambda)T_0^* + \lambda\alpha^\top \mathbf{P}_1, \mathbf{P}_1) < d(T_0^*, \mathbf{P}_1), \forall \lambda \in (0, 1).$$

This can be easily verify:

$$\begin{aligned} d(T_0, \mathbf{P}_1) &= \min_{\mathbf{U} \in (\mathcal{P}_{\mathcal{X}})^K} \sum_{k=1}^K \alpha_k D(U_k \| P_{1;k}) \\ &\quad \text{subject to } \alpha^\top \mathbf{U} = T_0^* \\ &\triangleq \sum_{k=1}^K \alpha_k D(U_k^* \| P_{1;k}) \\ &> \sum_{k=1}^K ((1 - \lambda)\alpha_k D(U_k^* \| P_{1;k}) + \lambda\alpha_k D(P_{1;k} \| P_{1;k})) \\ &\geq \sum_{k=1}^K \alpha_k D((1 - \lambda)U_k^* + \lambda P_{1;k} \| P_{1;k}) \\ &\geq \min_{\mathbf{U} \in (\mathcal{P}_{\mathcal{X}})^K} \sum_{k=1}^K \alpha_k D(U_k \| P_{1;k}) \\ &\quad \text{subject to } \alpha^\top \mathbf{U} = (1 - \lambda)T_0^* + \lambda\alpha^\top \mathbf{P}_1 \end{aligned}$$

This implies for all $\lambda > 0$, $(1 - \lambda)T_0^* + \lambda\alpha^\top \mathbf{P}_1$ is in \mathcal{B} (since T_0^* achieves minimum in \mathcal{B}^c). Therefore, by setting

$$T_0^{(i)} \triangleq (1 - \lambda_i)T_0^* + \lambda_i\alpha^\top \mathbf{P}_1, \lambda_i \rightarrow 0,$$

we find a sequence in \mathcal{B} converging to T_0^* .

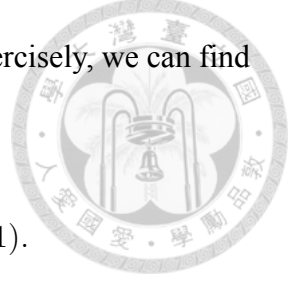
Hence, we can rewrite (5.8) as

$$\frac{d(T_1^*, \mathbf{P}_0)}{\beta} + \frac{d(T_0^*, \mathbf{P}_1)}{1 - \beta} = \lim_{i \rightarrow \infty} \frac{d(T_1^*, \mathbf{P}_0)}{\beta} + \frac{d(T_0^{(i)}, \mathbf{P}_1)}{1 - \beta}.$$

Since both T_1^* and $T_0^{(i)}$ are in \mathcal{B} , we must have

$$\frac{d(T_1^*, \mathbf{P}_0)}{\beta} + \frac{d(T_0^*, \mathbf{P}_1)}{1 - \beta} \leq \min_{F \in \mathcal{P}_{\mathcal{X}}} \frac{d(F, \mathbf{P}_0)}{\beta} + \frac{d(F, \mathbf{P}_1)}{1 - \beta},$$

which completes the converse.



Part 3 (Boundary of \mathcal{R})

Finally, we prove that

$$(\beta^2 (f'(\beta)(1 - \beta) - f(\beta)), (1 - \beta)^2 (\beta f'(\beta) + f(\beta)))$$



characterizes the boundary of \mathcal{R} . Consider two tangential lines

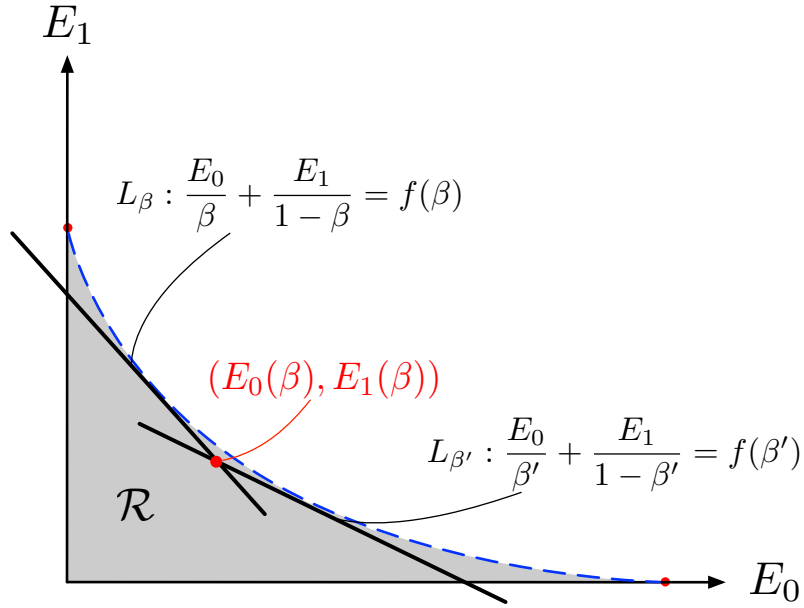


Figure 5.5: Illustration of L_β and $L_{\beta'}$

$$\begin{cases} L_\beta : \frac{E_0}{\beta} + \frac{E_1}{1-\beta} = f(\beta) \\ L_{\beta'} : \frac{E_0}{\beta'} + \frac{E_1}{1-\beta'} = f(\beta'). \end{cases}$$

Denote their intersection as $(E_0(\beta, \beta'), E_1(\beta, \beta'))$:

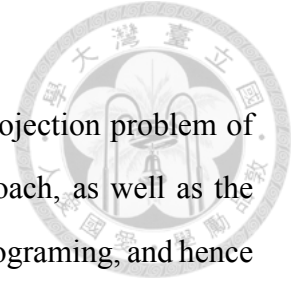
$$\begin{cases} E_1(\beta) = \frac{f(\beta)\beta - f(\beta')\beta'}{\beta/(1-\beta) - \beta'/(1-\beta')} \\ E_0(\beta) = \frac{f(\beta)(1-\beta) - f(\beta')(1-\beta')}{(1-\beta)/\beta - (1-\beta')/\beta'} \end{cases}$$

Taking $\beta' \rightarrow \beta$ and by L'Hospital's rule, we see that

$$\lim_{\beta' \rightarrow \beta} (E_0(\beta, \beta'), E_1(\beta, \beta')) = (\beta^2 (f'(\beta)(1 - \beta) - f(\beta)), (1 - \beta)^2 (\beta f'(\beta) + f(\beta))).$$

■

An alternative to characterize \mathcal{R} is to solving the information projection problem of acceptance region defined by the optimal test. However, this approach, as well as the boundary proposed by Theorem 5.2.2, involve complicated convex programming, and hence the closed-form formula still remains open.







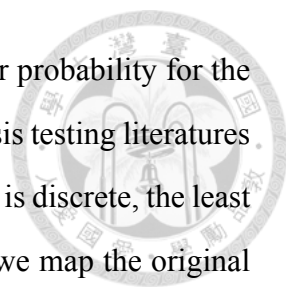
Chapter 6

Conclusions and Future Work

In this thesis, we propose two treatments in order to address the anonymity issue in privacy-preserving crowdsourcing.

In Part I, we study the minimum number of golden tasks (queries) required to recover the group information of the crowds. This problem is equivalent to decoding noisy data via pooling. The problem is cast into a linear inverse problem, where the pooled output takes values in integer, unlike the classic group-testing problem, where the outcomes are the “xor” of the values of the specified subset. We allow the pooled results to be perturbed by some bounded noise with magnitude less than δ_n . Under the noisy setting, the exact recover is usually impossible, and thus we aim to reconstruct the dataset partially, with distortion less than a given threshold n . Inspired by the Shannon’s construction of channel code, we use randomized pooling to obtain an upper bound on query complexity (the minimum number of queries) under small noise regime $\delta = O(\sqrt{\delta_n})$, in which each data will be pooled with probability $1/2$. This randomized pooling scheme takes $O(n/\log n)$ queries, with the same complexity as noiseless setting [43]. On the other hand, we prove the converse lower bound by packing argument, showing that the query complexity is indeed $\Theta(n/\log n)$. Interestingly, with an innovative counting approach, we showed that under the high noise regime $\delta = \Omega(\delta_n^{(1+\epsilon)/2})$, the recovery is impossible.

In Part II, we explore the anonymous detection problem with crowds anonymity, which is also related to the studies of wireless sensor networks. To deal with anonymity, a composite hypothesis testing approach is taken. Focusing on the Neyman-Pearson setting, we



provide an optimal test, and characterize the exponent of type-II error probability for the case that \mathcal{X} is finite. Unlike the settings considered in robust hypothesis testing literatures [23, 24, 41], since the hypothesis classes considered in our framework is discrete, the least favorable distribution might not exist. To circumvent the difficulty, we map the original problem into an auxiliary space by employing the symmetric property of the hypothesis classes, in which the composite hypothesis testing becomes a simple one. Therefore, Neyman-Pearson lemma can be applied to obtain an optimal test, which is a randomized threshold test based on the ratio of the uniform mixture of all the possible distributions under \mathcal{H}_0 to the uniform mixture of those under \mathcal{H}_1 . For the asymptotic regime, we analyze the type-II error exponent using method of types and show that the optimal exponent is the minimization of linear combination of KL-divergences, with the k -th term being $D(U_k \| P_{1;k})$ and α_k being the coefficient, for $k = 1, \dots, K$. The minimization is over all possible distributions U_1, \dots, U_K such that $\sum_{k=1}^K \alpha_k U_k = \sum_{k=1}^K \alpha_k P_{0;k}$. We further extend our result to Chernoff's regime, and indicate that the exponent region can be obtained by solving a convex optimization problem. There are still many open problems. For example, the closed-form expression for the exponents in asymptotic regime, even in Neyman-Pearson formulation, are still unknown. Besides, the solution of information projection is conjectured to have similar form like tilted-distributions, as the classical results in simple hypothesis testing suggested. In addition to hypothesis testing, it is also interesting to investigate other statistical problem such as regression, estimation, or pattern recognition under the anonymous setting.

Though in this thesis we concentrate on the crowdsourcing, the developed theories and tools are not restricted to this problem but widely apply to various fields, especially when the problems involve anonymity or privacy concern.

6.1 Future Works

One of an ambitious and challenging goal is to consider *partial group recovery*. From Figure 4.1, we see that in some cases, the type-II error exponent can be pushed to zero, making reliable detection no longer possible. If each sensor is allowed to transmit a few

bits of information to *partially reveal* their groups, how such partial information can improve the type-II error exponent? Formally speaking, we assume that the total number of groups is K , and each sensor can transmit L bits (with $L < \log K$) through a noiseless channel to the fusion center, providing partial information about the group that it belongs to.

Unsurprisingly, the optimal strategy is the *cluster-and-detect* approach, that is, we first *cluster* the K groups into 2^L super-groups, and each sensor sends L bits to indicate which *super-groups* it belongs to. Inside each super-group, we adopt the optimal anonymous hypothesis testing, and between super-groups, the problem boils down to the equivalent informed hypothesis testing, and hence standard likelihood ratio test can be applied there.

However, the difficulty lies in the clustering step: even the fusion center knows the distribution of each group, the optimal clustering algorithm is indeed a discrete optimization problem and thus NP-hard. When the group number K is large enough, it is intractable to find the optimal clustering. Nevertheless, some suboptimal algorithms suggested by heuristic do demonstrate that this partial information can significantly ameliorate the performance loss caused by anonymity. Below is a numerical example, showing the benefit of partial information.

In the example, we assume there are totally $K = 1024$ (2^{10}) groups, and each group accounts for $1/K$ proportion of total sensors, that is, $\alpha = [\frac{1}{K}, \dots, \frac{1}{K}]^\top$. For the sensors in the k -th group, their observations follow i.i.d. distribution $\text{Ber}(\theta_k)$ under \mathcal{H}_0 , and follow i.i.d. $\text{Ber}(1 - \theta_k)$ under \mathcal{H}_1 , with $\theta_k = \frac{k}{K}$, $k = 1, \dots, K$. Suppose there are L bits available for each sensor to partially inform the fusion center the group it belongs to, then as the clustering-detection algorithm suggests, we first cluster the K groups into 2^L super-groups and then apply anonymous hypothesis testing inside each super-group. As the numerical evaluation in Figure 6.1 illustrates, even with few bits, say, $L = 1$ or 2 , type-II error exponents are significantly improved.

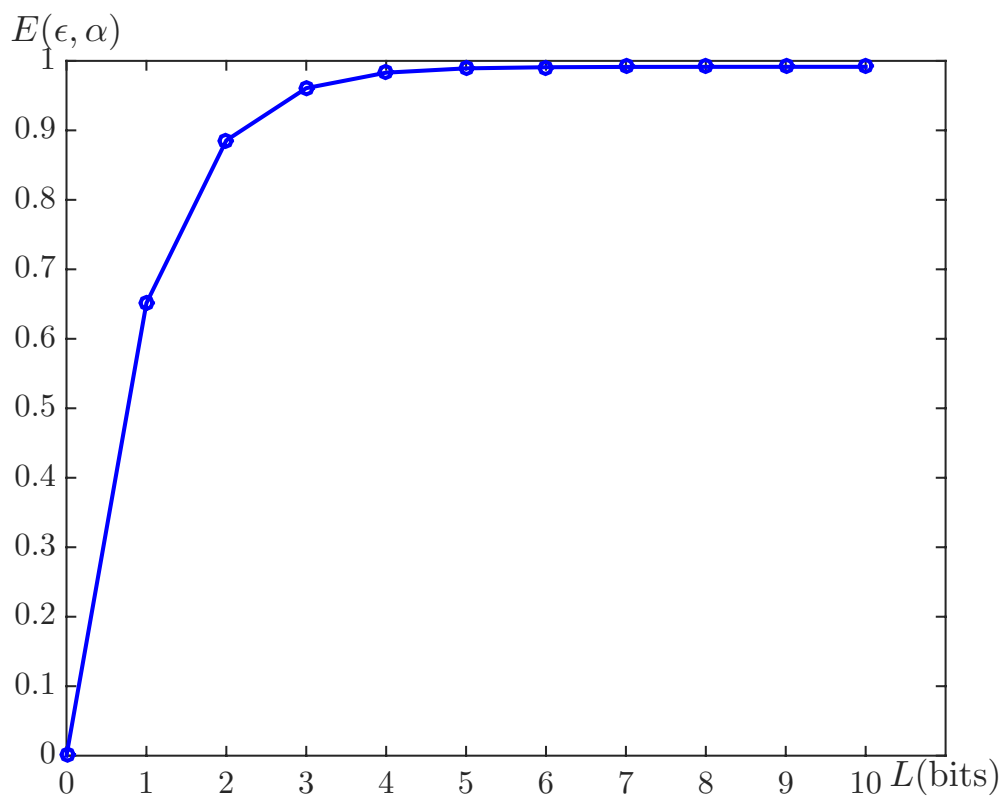


Figure 6.1: Exponents with Partial Information



Appendix A

Proof of Lemmas in Chapter 3

A.1 proof of Theorem 3.3.2

Theorem 3.3.2 (*Packing Lower Bound*) Let $k_n \leq \left(\frac{1-\epsilon}{2}\right)n$ for some $\epsilon > 0$. Then, the following lower bound holds:

$$T_n^*(k_n, \delta_n) = \Omega \left(\frac{n \left(1 - H_b \left(\frac{1-\epsilon}{2}\right)\right)}{\log(n+1) - \log(4\delta_n + 1)} \right) \quad (3.4)$$

Specifically, when $\delta_n = O(n^{\frac{1-\epsilon'}{2}})$, and ϵ, ϵ' does not depend on n , then (3.4) can be further simplified to

$$T_n^*(k_n, \delta_n) = \Omega(n / \log n).$$

We prove Theorem 3.3.2 by packing argument.

Proof. (proof of Theorem 3.3.2)

To reconstruct \mathbf{x} successfully, the cardinality of possible input must less than the cardinality of possible output. The number of possible input (output) turns out to be a packing problem. First, we notice the number of $\tilde{\mathbf{x}}$ with distance to \mathbf{x} less than k_n is

$$|\{\tilde{\mathbf{x}} \in \{0, 1\}^n \mid \|\tilde{\mathbf{x}} - \mathbf{x}\|_1 \leq k_n\}| = \sum_{i=1}^{k_n} \binom{n}{i},$$

thus the total number of all possible pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ with distance greater than k_n to each other

is

$$|\{\mathbf{x}_i \in \{0, 1\}^n \mid |\mathbf{x}_i - \mathbf{x}_j| > k_n, \forall \mathbf{x}_i \neq \mathbf{x}_j\}| = \frac{2^n}{\sum_{i=1}^{k_n} \binom{n}{i}}.$$

On the other hand, the total number of possible outcomes is

$$\frac{|\{0, 1, \dots, n\}^{T_n}|}{(4\delta_n + 1)^{T_n}} = \left(\frac{n+1}{4\delta_n + 1} \right)^{T_n}$$

To guarantee successful reconstruction, the number of possible outputs must greater than the number of possible input, hence we have

$$\frac{2^n}{\sum_{i=1}^{k_n} \binom{n}{i}} \leq \left(\frac{n+1}{4\delta_n + 1} \right)^{T_n} \quad (\text{A.1})$$

Now, we give a bound on summation of binomial coefficient: Suppose $k = pn$, where $p \in (0, 1)$ does not depend on n . Then by Stirling approximation, we have

$$\log \binom{n}{pn} = nH_b(p) + O(\log n),$$

where $H_b(p)$ is the binary entropy function.

$$T_n \geq \frac{n - \log \left(\sum_{i=1}^{k_n} \binom{n}{i} \right)}{\log(n+1) - \log(4\delta_n + 1)} \quad (\text{A.2})$$

$$\geq \frac{n - \log \left(k_n \binom{n}{n(1-\epsilon)/2} \right)}{\log(n+1) - \log(4\delta_n + 1)} \quad (\text{A.3})$$

$$= \frac{n \left(1 - H_b \left(\frac{1-\epsilon}{2} \right) + O\left(\frac{\log n}{n}\right) \right)}{\log(n+1) - \log(4\delta_n + 1)}, \quad (\text{A.4})$$

where (A.3) is due to $\sum_{i=1}^{k_n} \binom{n}{i} \leq k_n \binom{n}{k_n}$ and $k_n \leq \left(\frac{1-\epsilon}{2}\right) n$.

Furthermore, if $\delta_n = O(n^{\frac{1-\epsilon'}{2}})$, and ϵ, ϵ' does not depend on n , then we have

$$T_n = \Omega \left(\frac{n}{\log n} \right).$$

■





A.2 proof of Theorem 3.3.4

Theorem 3.3.4 (*Impossibility of Poly(n) Query Complexity*)

If both the following conditions are satisfied:

- $\frac{1}{2}n \geq k_n \geq C_1 n^{\epsilon_1}$
- $\delta_n = \Omega(k_n^{\frac{1+\epsilon_2}{2}})$

where $\epsilon_1, \epsilon_2 \in (0, 1)$, and $C_1 > 0$, then $T_n^(k_n, \delta_n)$ is $\omega(n^p)$, for all $p \in \mathbb{N}$. In words, there does not exist querying methods with Poly(n) query complexity that can do the job.*

Before we further bounding (3.5), we give two technical lemma:

Lemma A.2.1 *For $n \geq 2$, the following binomial bound holds:*

$$\frac{4^n}{\sqrt{\frac{\pi}{2}(2n+1)}} \leq \binom{2n}{n} \leq \frac{4^n}{\sqrt{\pi n}}$$

Lemma A.2.2 *For $\delta \leq n/2$, the following bound holds:*

$$\sum_{k=n/2-\delta}^{n/2+\delta} \binom{n}{k} \geq 2^n \left(1 - 2 \exp \left(-\frac{\delta^2}{n} \right) \right)$$

Now, we are ready to prove Theorem 3.3.4.

Proof. (proof of Theorem 3.3.4)

We show that as long as $k_n = \Omega(n^{\epsilon})$ and $\delta_n = \Omega\left(k_n^{\frac{1+\epsilon'}{2}}\right)$, the bound given in Theorem 3.3.3 is $\Omega(n)$.

From Theorem 3.3.3, we know that as long as

$$T_n \leq \tau = \frac{\binom{n}{\frac{n}{2}}}{2 \sum_{\alpha=2\delta_n}^{k_n/2} \binom{k_n/2}{\alpha} \sum_{\delta=2\delta_n}^{k_n/2} \binom{k_n/2}{\alpha-\delta} \binom{n-k_n}{n/2-2\alpha+\delta}}, \quad (\text{A.5})$$

the data set cannot be reconstructed successfully.

Applying Lemma A.2.1, we see that

$$1) \quad \binom{n}{n/2} \geq \frac{2^n}{\sqrt{\frac{\pi}{2}(n+1)}}$$

$$2) \binom{n-k_n}{n/2-2\alpha+\delta} \leq \binom{n-k_n}{(n-k_n)/2} \leq \frac{2^{n-k_n}}{\sqrt{\frac{\pi}{2}(n-k_n)}}.$$

Thus (A.5) is lower bounded by

$$\begin{aligned} &\geq \frac{2^{n-1}}{\sqrt{\frac{\pi}{2}(n+1)}} \left\{ \sum_{\alpha=2\delta_n}^{k_n/2} \binom{k_n/2}{\alpha} \sum_{\delta=2\delta_n}^{\alpha} \binom{k_n/2}{\alpha-\delta} \left(\frac{2^{n-k_n}}{\sqrt{\frac{\pi}{2}(n-k_n)}} \right) \right\}^{-1} \\ &= \frac{\sqrt{n-k_n}}{2\sqrt{n+1}} \left\{ 2^{-k_n} \sum_{\alpha=2\delta_n}^{k_n/2} \binom{k_n/2}{\alpha} \sum_{\delta=2\delta_n}^{\alpha} \binom{k_n/2}{\alpha-\delta} \right\}^{-1} \\ &= \frac{\sqrt{n-k_n}}{2\sqrt{n+1}} \left\{ 2^{-k_n} \sum_{\alpha=2\delta_n}^{k_n/2} \binom{k_n/2}{\alpha} \sum_{i=0}^{\alpha-2\delta_n} \binom{k_n/2}{i} \right\}^{-1} \end{aligned} \quad (\text{A.6})$$

Notice that

$$\begin{aligned} &\sum_{\alpha=2\delta_n}^{k_n/2} \binom{k_n/2}{\alpha} \sum_{i=0}^{\alpha-2\delta_n} \binom{k_n/2}{i} \\ &\leq 2^{k_n} - \left(\sum_{j=k_n/4-2\delta_n}^{k_n/4+2\delta_n} \binom{k_n/2}{j} \right)^2 \end{aligned} \quad (\text{A.7})$$

$$\leq 2^{k_n} - \left[2^{k_n/2} \left(1 - 2 \exp \left(-\frac{2\delta_n^2}{k_n} \right) \right) \right]^2 \quad (\text{A.8})$$

$$= 4 \exp \left(-\frac{2\delta_n^2}{k_n} \right) - 4 \exp \left(-\frac{2\delta_n^2}{k_n} \right)^2, \quad (\text{A.9})$$

where (A.7) is due to the observation of summation region, and (A.8) is due to Lemma (A.2.2).

Applying (A.9), we have

$$\begin{aligned} (\text{A.6}) &\geq \frac{\sqrt{n-k_n}}{8\sqrt{n+1}} \left\{ \exp \left(-\frac{2\delta_n^2}{k_n} \right) - \exp \left(-\frac{2\delta_n^2}{k_n} \right)^2 \right\}^{-1} \\ &= \underbrace{\frac{\sqrt{n-k_n}}{8\sqrt{n+1}}}_{(a)} \underbrace{\exp \left(\frac{2\delta_n^2}{k_n} \right)}_{(b)} \underbrace{\left\{ 1 - \exp \left(-\frac{2\delta_n^2}{k_n} \right) \right\}^{-1}}_{(c)} \end{aligned}$$

As long as $n \rightarrow \infty$,

1) (a) $\geq \frac{1}{8\sqrt{2}}$, due to the assumption $\frac{1}{2}n > k_n$



2) $(b) = \omega(n^p)$ for all integer p , due to the fact

$$\begin{aligned} \exp\left(\frac{2\delta_n^2}{k_n}\right) &\geq \exp(k_n^{\epsilon_2/2}) \geq \exp(C_1 n^{\epsilon_1 \epsilon_2/2}) \\ &\geq \exp(p \log n) = n^p \end{aligned}$$



3) $(c) \geq 1$

Combine (a), (b), (c) together, we conclude that as long as T_n polynomial in n , the successful recovery is impossible. ■

A.3 Proof of Claim 3.4.1

First, we use the notation $B_{n_1}^{(1)}, B_{n_2}^{(2)}$ to denote the independent random variables with distribution $\text{Binomial}(n_1, 1/2)$ and $\text{Binomial}(n_2, 1/2)$ respectively. By the definition of probability of failure, $P_f(\mathbf{x}; k_n, \delta_n)$ is

$$P_Q(\exists \tilde{\mathbf{x}}, \|\tilde{\mathbf{x}} - \mathbf{x}\|_1 > k_n, \|\mathbf{Q}\tilde{\mathbf{x}} - \mathbf{Q}\mathbf{x}\|_\infty \leq 2\delta_n) \quad (\text{A.10})$$

$$= P_Q\left(\bigcup_{\tilde{\mathbf{x}} \in B_{k_n}^c(\mathbf{x})} \|\mathbf{Q}\tilde{\mathbf{x}} - \mathbf{Q}\mathbf{x}\|_\infty \leq 2\delta_n\right) \quad (\text{A.11})$$

$$\leq \sum_{\tilde{\mathbf{x}} \in B_{k_n}^c(\mathbf{x})} \Pr(|\mathbf{q}\mathbf{x} - \mathbf{q}\tilde{\mathbf{x}}| \leq 2\delta_n)^{T_n} \quad (\text{A.12})$$

$$= \sum_{t=k_n}^n \sum_{\tilde{\mathbf{x}} \in \partial B_t(\mathbf{x})} \Pr(|\mathbf{q}\mathbf{x} - \mathbf{q}\tilde{\mathbf{x}}| \leq 2\delta_n)^{T_n} \quad (\text{A.13})$$

$$\leq \sum_{t=k_n}^n \binom{n}{t} \max_{t^++t^-=t} \Pr\left(|B_{t^+}^{(1)} - B_{t^-}^{(2)}| \leq 2\delta_n\right)^{T_n} \quad (\text{A.14})$$

Here we use $B_R(\mathbf{x})$ to denote the ball centered at \mathbf{x} with radius R , and use $\partial B_R(\mathbf{x})$ to denote the boundary of $B_R(\mathbf{x})$.

Notice that (A.12) is due to union bound, (A.14) is due to the fact that each \mathbf{q}_i is generated according to $\text{Ber}(1/2)$. To handle (A.14), we give the following lemma:

Lemma A.3.1 For $t_1 + t_2 = T$, T is even, the following fact holds:

$$\begin{aligned} & \Pr \left(\left| B_{t_1}^{(1)} - B_{t_2}^{(2)} \right| \leq \delta \right) \\ & \leq \Pr \left(\left| B_{T/2}^{(1)} - B_{T/2}^{(2)} \right| \leq \delta \right), \end{aligned}$$



where $B_{t_1}^{(1)}, B_{t_2}^{(2)}$ are independent random variables with distribution $\text{Binomial}(n_1, 1/2)$ and $\text{Binomial}(n_2, 1/2)$ respectively.

From lemma A.3.1, we see that the maximum of (A.14) occurs when $t^+ = t^- = t/2$. For simplicity we assume t even, and (A.14) becomes

$$\sum_{t=k_n}^n \binom{n}{t} \Pr \left(\left| B_{t/2}^{(1)} - B_{t/2}^{(2)} \right| \leq 2\delta_n \right)^{T_n} \quad (\text{A.15})$$

$$= \sum_{t=k_n}^n \binom{n}{t} \Pr \left(\left| B_{t/2}^{(1)} - B_{t/2}^{(2)} - \frac{t}{2} \right| \leq 2\delta_n \right)^{T_n} \quad (\text{A.16})$$

$$= \sum_{t=k_n}^n \binom{n}{t} \Pr \left(\left| B_t - \frac{t}{2} \right| \leq 2\delta_n \right)^{T_n} \quad (\text{A.17})$$

$$= \sum_{t=k_n}^n \binom{n}{t} \Pr (t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n)^{T_n}, \quad (\text{A.18})$$

here B_t in (A.16) denotes the random variable with distribution $\text{Binomial}(t, 1/2)$. (A.16) is due to the basic combinatorial fact, and (A.17) is due to the fact that our construction of \mathbf{Q} is independent.

A.4 Technical Lemmas

A.4.1 Lemma A.4.1

Lemma A.4.1 Let $S_{k_n}, V_i, \mathcal{T}_1$ and \mathcal{T}_2 be defined as before. Then

$$\max_{S \subseteq [n]} \left| \left\{ (x, \tilde{x}) \in S_{k_n} \mid \left| |\mathcal{S} \cap \mathcal{T}_1| - |\mathcal{S} \cap \mathcal{T}_2| \right| > 2\delta_n \right\} \right| \quad (\text{A.19})$$

achieves its maximum when $|\mathcal{S}| = \frac{n}{2}$.

Proof. First, let $|S| = s$, $s \in [n]$. Then (A.19) becomes

$$\begin{aligned}
 & 2^{n-k_n+1} \sum_{\alpha=0}^{k_n/2} \sum_{\delta=2\delta_n}^{\alpha} \binom{s}{\alpha} \binom{n-s}{k_n/2-\alpha} \binom{s-\alpha}{\alpha-\delta} \binom{n-s-k_n/2+\alpha}{k_n/2+\delta-\alpha} \\
 &= 2^{n-k_n+1} \binom{n}{k_n/2, k_n/2, n-k_n} \frac{\sum_{\alpha=0}^{k_n/2} \sum_{\delta=2\delta_n}^{\alpha} \binom{k_n/2}{\alpha} \binom{k_n/2}{\alpha-\delta} \binom{n-k_n}{s-2\alpha+\delta}}{\binom{n}{s}} \quad (\text{A.20})
 \end{aligned}$$

Therefore, maximize (A.19) is equivalent to maximize (A.20) over all possible s . After change of variables, (A.20) becomes

$$\begin{aligned}
 & \frac{\sum_{|i-j|>2\delta_n} \binom{k_n/2}{i} \binom{k_n/2}{j} \binom{n-k_n}{s-(i+j)}}{\binom{n}{s}} \\
 &= \frac{\sum_k \sum_{\substack{i+j=k \\ |i-j|>2\delta_n}} \binom{k_n/2}{i} \binom{k_n/2}{j} \binom{n-k_n}{s-k}}{\binom{n}{s}} \\
 &= \sum_k \underbrace{\left(\frac{\sum_{\substack{i+j=k \\ |i-j|>2\delta_n}} \binom{k_n/2}{i} \binom{k_n/2}{j}}{\binom{k_n}{k}} \right)}_{a_k} \underbrace{\left(\frac{\binom{k_n}{k} \binom{n-k_n}{s-k}}{\binom{n}{s}} \right)}_{b_k(s)} \\
 &= \sum_k a_k \cdot b_k(s).
 \end{aligned}$$

One can observe the following facts:

- a_k is symmetric to $k = k_n/2$, that is, $a_k = a_{k_n-k}$, since one can change the variables $(i', j') = (k_n/2 - i, k_n/2 - j)$.
- a_k is maximized as $k = k_n/2$. This can be proved by writing a_k in another form:

$$a_k = \frac{\sum_{\substack{i > (k+2\delta_n)/2 \text{ or } \\ i < (k-\delta_n)/2}} \binom{k}{i} \binom{k_n-k}{k_n/2-i}}{\binom{k_n}{k_n/2}},$$

and it achieves maximum at $k = k_n/2$. Also, a_k is increasing in $[0, k_n/2]$ (and hence decreasing in $[k_n/2, k_n]$).

- For all $s \in [n]$, $\sum_k b_k(s) = 1$.
- For $s \in [0, n/2)$, $b_k(s)$ is maximized at $k^* \in [0, k_n/2]$; for $s \in (n/2, n]$, $b_k(s)$ is

maximized at $k^* \in (k_n/2, k_n]$. Also, $b_k(s)$ is increasing for $k \leq k^*$, and decreasing for $k \geq k^*$.

- $b_k(s)$ is symmetric to $(n/2, k_n/2)$, that is,

$$b_k(s) = b_{k_n-k}(n-s).$$

- $\sum_k a_k \cdot b_k(s)$ is symmetric to $n/2$, that is, $\sum_k a_k \cdot b_k(s) = \sum_k a_k \cdot b_k(n-s)$.

Now, we are ready to show $\sum_k a_k \cdot b_k(s)$ attains its maximum at $x = n/2$. For any $s \in [n]$, we have

$$\sum_k a_k \cdot b_k(s) = \sum_k a_k \cdot \left(\frac{b_k(s) + b_k(n-s)}{2} \right).$$

Consider the equation

$$b_k(n/2) = \left(\frac{b_k(s) + b_k(n-s)}{2} \right).$$

There's exactly a zero at $k = \xi \in [0, n/2]$ for all s , and due to symmetry, there's another zero at $k = k_n - \xi$. Besides, $b_k(n/2) \geq \left(\frac{b_k(s) + b_k(n-s)}{2} \right)$ at $k = k_n/2$. Therefore we conclude that

$$b_k(n/2) \geq \left(\frac{b_k(s) + b_k(n-s)}{2} \right)$$

for $s \in [\xi, k_n - \xi]$, and

$$b_k(n/2) < \left(\frac{b_k(s) + b_k(n-s)}{2} \right)$$

for $s \in [\xi, k_n - \xi]^c$ (With a slight abuse of notation, we denote $[0, k_n] \setminus [\xi, k_n - \xi]$ as $[\xi, k_n - \xi]^c$).

Notice that since

$$\sum_k b_k(s) = \sum_k \left(\frac{b_k(s) + b_k(n-s)}{2} \right) = 1,$$

we have

$$\sum_{k \in [\xi, k_n - \xi]} \left(b_k(n/2) - \left(\frac{b_k(s) + b_k(n-s)}{2} \right) \right)$$



$$= - \sum_{k \in [\xi, k_n - \xi]^c} \left(b_k(n/2) - \left(\frac{b_k(s) + b_k(n-s)}{2} \right) \right). \quad (\text{A.21})$$



Now, consider

$$\begin{aligned} & \sum_k a_k \cdot b_k(n/2) - \sum_k a_k \cdot b_k(s) \\ &= \sum_k a_k \cdot \left(b_k(n/2) - \left(\frac{b_k(s) + b_k(n-s)}{2} \right) \right) \\ &= \sum_{k \in [\xi, k_n - \xi]} a_k \cdot \left(b_k(n/2) - \left(\frac{b_k(s) + b_k(n-s)}{2} \right) \right) \\ &+ \sum_{k \in [\xi, k_n - \xi]^c} a_k \cdot \left(b_k(n/2) - \left(\frac{b_k(s) + b_k(n-s)}{2} \right) \right) \\ &\geq 0. \end{aligned} \quad (\text{A.22})$$

(A.22) is due to the fact that $a_{k_1} \geq a_{k_2}$, for all $k_1 \in [\xi, k_n - \xi]$, $k_2 \in [\xi, k_n - \xi]^c$ and (A.21). Since it holds for all $s \in [n]$, the proof is complete. ■

A.4.2 Proof of Lemma A.3.1

Lemma A.3.1 *For $t_1 + t_2 = T$, T is even, the following fact holds:*

$$\begin{aligned} & \Pr \left(\left| B_{t_1}^{(1)} - B_{t_2}^{(2)} \right| \leq \delta \right) \\ & \leq \Pr \left(\left| B_{T/2}^{(1)} - B_{T/2}^{(2)} \right| \leq \delta \right), \end{aligned}$$

where $B_{t_1}^{(1)}, B_{t_2}^{(2)}$ are independent random variables with distribution $\text{Binomial}(n_1, 1/2)$ and $\text{Binomial}(n_2, 1/2)$ respectively.

Proof. First, note that for $B_t \sim \text{Binomial}(t, \frac{1}{2})$, $t - B_t$ has the same distribution with B_t . Therefore,

$$\begin{aligned} & \Pr \left(\left| B_{t_1}^{(1)} - B_{t_2}^{(2)} \right| \leq \delta \right) \\ &= \Pr \left(|B_{t_1+t_2} - t_2| \leq \delta \right) \\ &= \Pr \left(t_2 - \delta \leq B_T \leq t_2 + \delta \right) \end{aligned}$$

$$\begin{aligned} &\leq \Pr(T/2 - \delta \leq B_T \leq T/2 + \delta) \\ &= \Pr\left(\left|B_{T/2}^{(1)} - B_{T/2}^{(2)}\right| \leq \delta\right). \end{aligned}$$

■



A.4.3 Proof of Lemma 3.4.1

Lemma 3.4.1 *Let $B_t \stackrel{iid}{\sim} \text{Binomial}(t, 1/2)$, $\delta_n \in (0, t/16)$ then the following two upper bounds hold:*

$$1) \Pr(t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n) \leq \frac{4\delta_n + 1}{\sqrt{\pi t}}.$$

This bound is used when δ_n is small (with respect to t).

$$2) \Pr(t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n) \leq 1 - \frac{2}{15}e^{-64\delta_n^2/t}.$$

This bound is used when δ_n is large (with respect to t).

The proof can be found in Appendix D in [11].

Proof.

1) Since $\Pr(B_t = t/2) > \Pr(B_t = k)$, for all $k \in [0, t]$, we have

$$\begin{aligned} \Pr(t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n) &\leq (4\delta_n + 1) \Pr(B_t = t/2) \\ &\leq \frac{4\delta_n + 1}{\sqrt{\pi t}}. \end{aligned}$$

The last inequality is due to Lemma A.2.1.

2) For convenience, let $\delta_n = 2\delta'_n$. This is equivalent to show

$$\Pr(B_t > t/2 + \delta_n) \geq \frac{1}{15} \exp(-16\delta_n^2/t).$$

The proof is first given in [31], which involves some elementary estimates. For the sake of completeness, we state it again.



Write $t = 2m$. We have

$$\begin{aligned}
\Pr(B_t \geq m + \delta_n) &= 2^{-2m} \sum_{j=\delta_n}^m \binom{2m}{m+j} \\
&\geq 2^{-2m} \sum_{j=\delta_n}^{2\delta_n-1} \binom{2m}{m+j} \\
&= 2^{-2m} \sum_{j=\delta_n}^{2\delta_n-1} \binom{2m}{m} \frac{m}{m+j} \cdot \frac{m-1}{m+j-1} \cdots \frac{m-j+1}{m+1} \\
&\geq \frac{1}{\sqrt{m}} \sum_{j=\delta_n}^{2\delta_n-1} \prod_{i=1}^j \left(1 - \frac{j}{m+i}\right) \\
&\geq \frac{1}{\sqrt{m}} \left(1 - \frac{2\delta_n}{m}\right)^{2\delta} \\
&\geq \frac{1}{\sqrt{m}} \cdot \exp(-8\delta_n^2/m).
\end{aligned}$$

For $\delta_n \geq \frac{\sqrt{m}}{4}$, the last expression is at least $\frac{1}{8} \exp\left(\frac{-16\delta_n}{n}\right)$. For $0 \leq \delta_n < \frac{\sqrt{m}}{4}$, we have

$$\Pr(B_t > m + \delta_n) > \Pr(B_t > m + \frac{\sqrt{m}}{4}) \geq \frac{1}{8} \exp(-1/2) > \frac{1}{15}.$$

Thus the claimed bound holds for all $\delta_n \leq m/4$.

■

A.4.4 Proof of Lemma A.2.1

Lemma A.2.1 *For $n \geq 2$, the following binomial bound holds:*

$$\frac{4^n}{\sqrt{\frac{\pi}{2}(2n+1)}} \leq \binom{2n}{n} \leq \frac{4^n}{\sqrt{\pi n}}$$

Proof. First we consider the two expressions:

$$2n \left(\binom{2n}{n} \frac{1}{4^n} \right)^2 = \underbrace{\frac{1}{2} \frac{3}{2} \frac{3}{4} \frac{5}{4} \cdots \frac{2n-1}{2n-2}}_{(1)} \underbrace{\frac{2n-1}{2n}}_{(2)} \quad (\text{A.23})$$

$$= \frac{1}{2} \prod_{j=2}^n \left(1 + \frac{1}{4j(j-1)} \right), \quad (\text{A.24})$$

and

$$(2n+1) \left(\binom{2n}{n} \frac{1}{4^n} \right)^2 = \underbrace{\frac{1}{2} \frac{3}{2} \frac{3}{4} \frac{5}{4} \cdots \frac{2n-1}{2n-2} \frac{2n-1}{2n}}_{(1)} \underbrace{\frac{2n+1}{2n}}_{(3)} \quad (\text{A.25})$$

$$= \prod_{j=1}^n \left(1 - \frac{1}{4j^2} \right). \quad (\text{A.26})$$

By Wallis's formula, (1) converges to $\frac{2}{\pi}$, and (2), (3) converge to 1. Therefore, both (A.24), (A.26) converge to $\frac{2}{\pi}$. Notice that according to the left hand side of two expressions, (A.24) is increasing and (A.26) is decreasing, with the same limit. Therefore we conclude that

$$2n \left(\binom{2n}{n} \frac{1}{4^n} \right)^2 \leq \frac{2}{\pi}, \quad (\text{A.27})$$

and

$$(2n+1) \left(\binom{2n}{n} \frac{1}{4^n} \right)^2 \geq \frac{2}{\pi}. \quad (\text{A.28})$$

Since this holds for $n \geq 2$, the proof is complete. ■

A.4.5 Proof of Lemma A.2.2

Lemma A.2.2 For $\delta \leq n/2$, the following bound holds:

$$\sum_{k=n/2-\delta}^{n/2+\delta} \binom{n}{k} \geq 2^n \left(1 - 2 \exp \left(-\frac{\delta^2}{n} \right) \right)$$

Proof. This is a direct application of Chernoff Bound. Let $X_i \stackrel{i.i.d}{\sim} \text{Ber}(\frac{1}{2})$, $i \in [N]$. Applying Chernoff inequality on $X = \sum_1^N X_i$, we have

$$\Pr(X \geq \mathbb{E}X + \delta) \leq \exp \left(\frac{-\delta^2}{n} \right).$$

Therefore,

$$\begin{aligned}\sum_{k=n/2-\delta}^{n/2+\delta} \binom{n}{k} &\geq 2^n \left(1 - 2 \exp\left(-\frac{2\delta^2}{n}\right)\right) \\ &\geq 2^n \left(1 - 2 \exp\left(-\frac{\delta^2}{n}\right)\right)\end{aligned}$$

■







Appendix B

Proof of Lemmas in Chapter 5

B.1 Proof of Proposition 4.3.1

proof of Proposition 4.3.1. Since the optimal type-II exponent does not depend on ϵ , we denote it as $E^*(\alpha)$ and for simplicity. It suffices to show

$$E^*(\lambda\alpha_1 + (1 - \lambda)\alpha_2) \leq \lambda E^*(\alpha_1) + (1 - \lambda)E^*(\alpha_2), \forall \lambda \in [0, 1].$$

First, let

$$\begin{aligned} E^*(\alpha_1) &= \sum_{k=1}^K \alpha_{1k} D(U_{1k}^* \| P_{1;k}) \\ E^*(\alpha_2) &= \sum_{k=1}^K \alpha_{2k} D(U_{2k}^* \| P_{1;k}) \end{aligned}$$

where $\alpha_1 = [\alpha_{11}, \dots, \alpha_{1K}]^\top$, $\alpha_2 = [\alpha_{21}, \dots, \alpha_{2K}]^\top$, and $U_1^* \triangleq [U_{11}^*, \dots, U_{1K}^*]$, $U_2^* \triangleq [U_{21}^*, \dots, U_{2K}^*]$ are the minimizers of (4.7). Then, by the convexity of KL divergence, we have

$$\begin{aligned} \lambda E^*(\alpha_1) + (1 - \lambda)E^*(\alpha_2) &= \sum_{k=1}^K \lambda \alpha_{1k} D(U_{1k}^* \| P_{1;k}) + (1 - \lambda) \alpha_{2k} D(U_{2k}^* \| P_{1;k}) \\ &\geq \sum_{k=1}^K (\lambda \alpha_{1k} + (1 - \lambda) \alpha_{2k}) D\left(\frac{\lambda \alpha_{1k} U_{1k}^* + (1 - \lambda) \alpha_{2k} U_{2k}^*}{\lambda \alpha_{1k} + (1 - \lambda) \alpha_{2k}} \parallel P_{1;k}\right) \end{aligned} \tag{B.1}$$

Now we claim that $\tilde{\mathbf{U}} \triangleq \left(\frac{\lambda\alpha_{1k}U_{1k}^* + (1-\lambda)\alpha_{2k}U_{2k}^*}{\lambda\alpha_{1k} + (1-\lambda)\alpha_{2k}} \right)_{k=1,\dots,K}$ satisfies

$$(\lambda\alpha_1 + (1-\lambda)\alpha_2)^\top \tilde{\mathbf{U}} = (\lambda\alpha_1 + (1-\lambda)\alpha_2)^\top \mathbf{P}_0, \quad (\text{B.2})$$

and thus

$$\begin{aligned} (\text{B.1}) &= \sum_{k=1}^K (\lambda\alpha_{1k} + (1-\lambda)\alpha_{2k}) D\left(\tilde{U}_k \parallel P_{1;k}\right) \\ &\geq \min_{\substack{\mathbf{U} \in (\mathcal{P}_{\mathcal{X}})^K \\ (\lambda\alpha_1 + (1-\lambda)\alpha_2)^\top \mathbf{U} = (\lambda\alpha_1 + (1-\lambda)\alpha_2)^\top \mathbf{P}_0}} \sum_{k=1}^K (\lambda\alpha_{1k} + (1-\lambda)\alpha_{2k}) D(U_k \parallel P_{1;k}) \\ &= E^*(\lambda\alpha_1 + (1-\lambda)\alpha_2). \end{aligned}$$

To show (B.2), we notice that $\mathbf{U}_1^*, \mathbf{U}_2^*$ satisfy the constraints

$$\alpha_1^\top \mathbf{U}_1^* = \alpha_1^\top \mathbf{P}_0, \quad \alpha_2^\top \mathbf{U}_2^* = \alpha_2^\top \mathbf{P}_0. \quad (\text{B.3})$$

Then we have

$$\begin{aligned} &(\lambda\alpha_1 + (1-\lambda)\alpha_2)^\top \tilde{\mathbf{U}} \\ &= \sum_{k=1}^K (\lambda\alpha_{1k} + (1-\lambda)\alpha_{2k}) \left(\frac{\lambda\alpha_{1k}U_{1k}^* + (1-\lambda)\alpha_{2k}U_{2k}^*}{\lambda\alpha_{1k} + (1-\lambda)\alpha_{2k}} \right) \\ &= \sum_{k=1}^K \lambda\alpha_{1k}U_{1k}^* + (1-\lambda)\alpha_{2k}U_{2k}^* \\ &= \lambda\alpha_1^\top \mathbf{U}_1^* + (1-\lambda)\alpha_2^\top \mathbf{U}_2^* \\ &= (\lambda\alpha_1 + (1-\lambda)\alpha_2)^\top \mathbf{P}_0, \end{aligned}$$

which completes the proof. ■



B.2 Proof of Lemma 4.5.1



proof of Lemma 4.5.1. First, observe that since $\text{int } \Gamma$ is open, the set

$$\tilde{\Gamma} \triangleq \{(U_1, \dots, U_K) \mid \alpha^\top U \in \text{int } \Gamma\} \subset (\mathcal{P}_{\mathcal{X}})^K$$

is open too. This is because the mapping $g(U) = \alpha^\top U$ is continuous, so the pre-image preserves the openness (under standard topology). Therefore, we can find a sequence

$$\left\{ U^{(n)} \in (\mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K}) \cap \tilde{\Gamma} \right\},$$

such that

$$\sum_k \alpha_k D(U_k^{(n)} \parallel P_{\theta;k}) \rightarrow - \inf_{\substack{(U_1, \dots, U_K) \in (\mathcal{P}_{\mathcal{X}})^K \\ \alpha^\top U \in \text{int } \Gamma}} \sum_k \alpha_k D(U_k \parallel P_{\theta;k}),$$

where the limit is taken such that $\frac{n_k}{n} \rightarrow \alpha_k$. So we have

$$\begin{aligned} \mathbb{P}_{\theta;\sigma} \{\Pi_{x^n} \in \Gamma\} &= \sum_{\substack{(U_1, \dots, U_K) \in \mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K} \\ \alpha^\top U \in \Gamma}} \prod_{k=1}^K P_{\theta;k}^{\otimes n_k} \{T_{n_k}(U_k)\} \\ &\geq \sum_{\substack{(U_1, \dots, U_K) \in \mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K} \\ \alpha^\top U \in \text{int } \Gamma}} \prod_{k=1}^K P_{\theta;k}^{\otimes n_k} \{T_{n_k}(U_k)\} \\ &\geq \max_{\substack{(U_1, \dots, U_K) \in \mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K} \\ \alpha^\top U \in \text{int } \Gamma}} \prod_{k=1}^K P_{\theta;k}^{\otimes n_k} \{T_{n_k}(U_k^{(n)})\} \\ &\stackrel{(a)}{\geq} \max_{\substack{(U_1, \dots, U_K) \in \mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K} \\ \alpha^\top U \in \text{int } \Gamma}} \left(\frac{1}{(n_k + 1)^{|\mathcal{X}|}} \right) 2^{\sum_{k=1}^K n_k D(U_k^{(n)} \parallel P_{\theta;k})}, \end{aligned}$$

where inequality (a) holds by Lemma 2.1.2. Thus we have

$$\frac{1}{n} \log \mathbb{P}_{\theta;\sigma} \{\Pi_{x^n} \in \Gamma\} \geq - \min_{\substack{(U_1, \dots, U_K) \in \mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K} \\ \alpha^\top U \in \text{int } \Gamma}} \left(\sum_{k=1}^K \frac{n_k}{n} D(U_k^{(n)} \parallel P_{\theta;k}) + o(1) \right).$$

As $n \rightarrow \infty$ such that $\frac{n_k}{n} \rightarrow \alpha_k$, we see that

$$- \inf_{\substack{(U_1, \dots, U_K) \in (\mathcal{P}_{\mathcal{X}})^K \\ \alpha^\top U \in \text{int } \Gamma}} \sum_k \alpha_k D(U_k \| P_{\theta; k}) \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\theta; \sigma} \{\Pi_{x^n} \in \Gamma\}.$$

On the other hand, for the upper bound, consider

$$\begin{aligned} \mathbb{P}_{\theta; \sigma} \{\Pi_{x^n} \in \Gamma\} &= \sum_{\substack{(U_1, \dots, U_K) \in \mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K} \\ \alpha^\top U \in \Gamma}} \prod_{k=1}^K P_{\theta; k}^{\otimes n_k} \{T_{n_k}(U_k)\} \\ &\stackrel{(a)}{\leq} \sum_{\substack{(U_1, \dots, U_K) \in \mathcal{P}_{n_1} \times \dots \times \mathcal{P}_{n_K} \\ \alpha^\top U \in \Gamma}} 2^{\sum_{k=1}^K D(U_k^{(n)} \| P_{\theta; k})} \\ &\leq \left(\prod_k |\mathcal{P}_{n^k}| \right) 2^{\sum_{k=1}^K n_k D(U_k^{(n)} \| P_{\theta; k})} \\ &\stackrel{(b)}{=} 2^{(\sum_{k=1}^K n_k D(U_k^{(n)} \| P_{\theta; k}) + o(1))}, \end{aligned}$$

where where inequality (a) holds by Lemma 2.1.2, and (b) holds due to the cardinality bound Lemma 2.1.1.

As $n \rightarrow \infty$ and $\frac{n_k}{n} \rightarrow \alpha_k$, we have

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\theta; \sigma} \{\Pi_{x^n} \in \Gamma\} \leq - \inf_{\substack{(U_1, \dots, U_K) \in (\mathcal{P}_{\mathcal{X}})^K \\ \alpha^\top U \in \Gamma}} \sum_k \alpha_k D(U_k \| P_{\theta; k}).$$

Notice that for the case \mathcal{X} finite, the infimum takes over Γ is equal to that one takes in the closure of Γ , since we can use standard topology to find a sequence approaching to the limit point. Thus the proof is complete. ■

B.3 Proof of Lemma 4.5.2

proof of Lemma 4.5.2. Let $\mathbf{Q} \in (\mathcal{P}_{\mathcal{X}})^K$ be a K-tuple of probability measure on \mathcal{X} . We first show that

$$\mathcal{C}_{\mathbf{Q}} \triangleq \{T \in \mathcal{P}_{\mathcal{X}} : f_{\mathbf{Q}}(T) < \infty\}$$

is a compact set.

Part 1 (Compactness) Observe that $f_Q(T) < \infty$ if and only if there exists a $\mathbf{P} = (P_1, \dots, P_K) \in (\mathcal{P}_{\mathcal{X}})^K$, such that

1. $\alpha^\top \mathbf{P} = T$
2. for all $i = 1, \dots, K$, $P_i \ll Q_i$.

Therefore, let us denote

$$\mathcal{M}_Q \triangleq \left\{ \mathbf{P} \in (\mathcal{P}_{\mathcal{X}})^K : P_i \ll Q_i, \forall i = 1, \dots, K \right\} \subseteq (\mathcal{P}_{\mathcal{X}})^K.$$

We claim that \mathcal{M}_Q is a compact set, and thus

$$\mathcal{C}_Q = \{ \alpha^\top \mathbf{P} \mid \mathbf{P} \in \mathcal{M}_Q \}$$

is also compact, since $\alpha^\top \mathbf{P}$ is a linear mapping from $(\mathcal{P}_{\mathcal{X}})^K$ to $\mathcal{P}_{\mathcal{X}}$ so compactness is preserved. To prove the claim, it suffices to show that \mathcal{M}_Q is a closed set, because the boundness is directly followed by the boundness of $(\mathcal{P}_{\mathcal{X}})^K$. It is equivalent to show

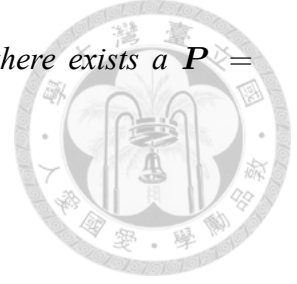
$$\mathcal{M}_Q^C = \left\{ \mathbf{P} \in (\mathcal{P}_{\mathcal{X}})^K : P_i \not\ll Q_i, \text{ for some } i \right\}$$

is open. Notice that

$$\left\{ \mathbf{P} \in (\mathcal{P}_{\mathcal{X}})^K : P_i \not\ll Q_i, \text{ for some } i \right\} = \bigcup_{i=1}^K \left\{ \mathbf{P} \in (\mathcal{P}_{\mathcal{X}})^K : P_i \not\ll Q_i \right\},$$

so it suffices to show $\left\{ \mathbf{P} \in (\mathcal{P}_{\mathcal{X}})^K : P_i \not\ll Q_i \right\}$ is open for all i . Assume $P_i \not\ll Q_i$. Then there must exist some measurable event $\mathcal{E} \subset \mathcal{X}$, such that $Q_i(\mathcal{E}) = 0$, and $P_i(\mathcal{E}) = \epsilon > 0$. Therefore, if \mathcal{X} is finite and thus $\mathcal{P}_{\mathcal{X}}$ equipped with total-variation distance (i.e. one norm), then obviously for any \tilde{Q} such that $\|\tilde{Q} - P_i\| < \frac{\epsilon}{2}$, $\tilde{Q} \not\ll Q_i$. Hence \mathcal{M}_Q^C is open, proving the claim.

Remark B.3.1 If \mathcal{X} is Polish, then $\mathcal{P}_{\mathcal{X}}$ is equipped with Prokhorov's metric, and one can use similar argument to show that \mathcal{M}_Q^C is open.



Next, we show that $f_Q(\cdot)$ is a convex function, so the convexity of \mathcal{C}_Q follows: for all $T_1, T_2 \in \mathcal{C}_Q$,

$$f_Q(\lambda T_1 + (1 - \lambda)T_2) \leq \lambda f_Q(T_1) + (1 - \lambda)f_Q(T_2) < \infty, \quad (\text{B.4})$$

implying $\lambda T_1 + (1 - \lambda)T_2 \in \mathcal{C}_Q$.

Part 2 (Convexity) To show (B.4), we observe

$$\begin{aligned} & \lambda f_Q(T_1) + (1 - \lambda)f_Q(T_2) \\ &= \inf_{\mathbf{U}: \boldsymbol{\alpha}^\top \mathbf{U} = T_1} \lambda \sum_k \alpha_k D(U_k \| P_k) + \inf_{\mathbf{V}: \boldsymbol{\alpha}^\top \mathbf{V} = T_2} (1 - \lambda) \sum_k \alpha_k D(V_k \| P_k) \\ &\stackrel{(a)}{\geq} \inf_{\mathbf{U}, \mathbf{V}: \boldsymbol{\alpha}^\top \mathbf{U} = T_1, \boldsymbol{\alpha}^\top \mathbf{V} = T_2} \sum_k \alpha_k D(\lambda U_k + (1 - \lambda)V_k \| P_k) \\ &\stackrel{(b)}{\geq} \inf_{\mathbf{P}: \boldsymbol{\alpha}^\top \mathbf{P} = \lambda T_1 + (1 - \lambda)T_2} \sum_k \alpha_k D(Q_k \| P_k) \\ &= f_Q(\lambda T_1 + (1 - \lambda)T_2), \end{aligned}$$

where (a) is due to the convexity of KL-divergence, and (b) is because

$$\boldsymbol{\alpha}^\top \mathbf{U} = T_1, \boldsymbol{\alpha}^\top \mathbf{V} = T_2 \Rightarrow \boldsymbol{\alpha}^\top (\lambda \mathbf{U} + (1 - \lambda)\mathbf{V}) = \lambda T_1 + (1 - \lambda)T_2.$$

Therefore, we conclude that $f_Q(\cdot)$ is a convex function and \mathcal{C}_Q is a convex set.

At the final step, we show $f_Q(\cdot)$ is a continuous function on \mathcal{C}_Q . Notice that the convexity of $f_Q(\cdot)$ only guarantees the continuity on the interior of \mathcal{C}_Q , and thus we need to additionally check the boundary points.

Remark B.3.2 Note that in general, the interior of \mathcal{C}_Q may be an empty set since it may lie in a subspace of $\mathcal{P}_\mathcal{X}$. Alternatively, we can define a point \mathbf{P} being interior, if it can be written as

$$\lambda \mathbf{U} + (1 - \lambda)\mathbf{V}, \text{ for some } \lambda \in (0, 1), \text{ and some } \mathbf{V}, \mathbf{U} \in \mathcal{C}_Q.$$

Part 3 (Continuity) First, if the interior of \mathcal{C}_Q is empty, then by the convexity, either \mathcal{C}_Q

is a empty set, or it is a singleton. For both cases, the continuity holds obviously. Hence without losing of generality, we assume that the interior of \mathcal{C}_Q is non-empty, and T_0 is an interior point.

Then for any $T \in \mathcal{C}_Q$, we can construct a sequence $T_n \in \mathcal{C}_Q$, $T_n \rightarrow T$. For example, one can let $T_n = \lambda_n T_0 + (1 - \lambda_n)T$, with $\lambda_n \rightarrow 0$. Let $\mathbf{U}^{(n)} = (U_1^{(n)}, \dots, U_K^{(n)}) \in (\mathcal{P}_{\mathcal{X}})^K$ be a sequence such that

1. $\alpha^\top \mathbf{U}^{(n)} = T_n$

2. $\mathbf{U}^{(n)}$ achieves the infimum of $f_Q(T_n)$:

$$\sum_{k=1}^K \alpha_k D(U_k^{(n)} \| P_k) = \inf_{\mathbf{V}: \alpha^\top \mathbf{V} = T_n} \sum_{k=1}^K \alpha_k D(V_k \| P_k) = f_Q(T_n).$$

Notice that the infimum can always be achieved since $g(\mathbf{V}) \triangleq \sum_{k=1}^K \alpha_k D(V_k \| P_k)$ is a continuous function over the compact set \mathcal{M}_Q .

By construction, $\mathbf{U}^{(n)}$ is a sequence in a compact set \mathcal{M}_Q , and hence by Bolzano-Weierstrass theorem (see Chapter 1 in [34], for example), there exists a convergent subsequence $\mathbf{U}^{(n_i)}$, and let us denote the convergent point

$$\lim_{i \rightarrow \infty} \mathbf{U}^{(n_i)} = \mathbf{U}.$$

Since $\alpha^\top \mathbf{U}^{(n_i)} = T_{n_i}$, and $T_{n_i} \rightarrow T$, we have

$$\alpha^\top \mathbf{U} = T.$$

Notice that the function $f(\mathbf{V}) \triangleq \sum_{k=1}^K \alpha_k D(V_k \| P_k)$ is a continuous function over the compact set \mathcal{M}_Q , we must have

$$\lim_{n \rightarrow \infty} f_Q(T_n) = \lim_{i \rightarrow \infty} f_Q(T_{n_i}) = \sum_{k=1}^K \alpha_k D(U_k \| P_k),$$

and therefore

$$f_{\mathbf{Q}}(T) = \inf_{\mathbf{U}: \mathbf{\alpha}^\top \mathbf{V} = T} \sum_{k=1}^K \alpha_k D(V_k \| P_k) \leq \sum_{k=1}^K \alpha_k D(U_k \| P_k) = \lim_{n \rightarrow \infty} f_{\mathbf{Q}}(T_n).$$

On the other hand, by the convexity of $f_{\mathbf{Q}}(\cdot)$, we must have

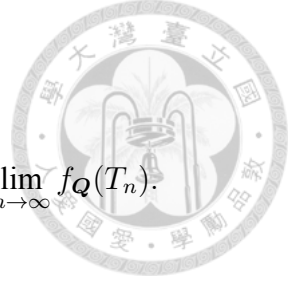
$$f_{\mathbf{Q}}(T) \geq f_{\mathbf{Q}}(T_n), \text{ for all } n \text{ large enough.}$$

Otherwise

$$f_{\mathbf{Q}}(\lambda T_0 + (1 - \lambda)T) > \lambda f_{\mathbf{Q}}(T_0) + (1 - \lambda)f_{\mathbf{Q}}(T),$$

for some λ small enough, which violates the fact that $f_{\mathbf{Q}}(\cdot)$ is a convex function.

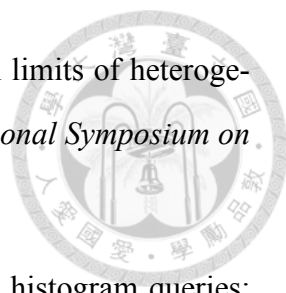
■

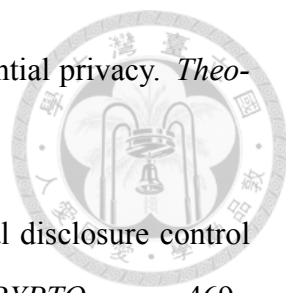


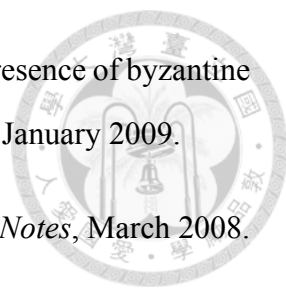


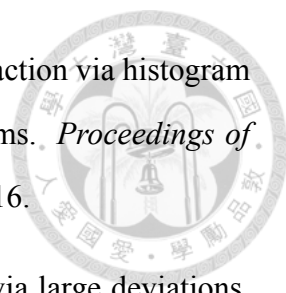
Bibliography

- [1] A. E. Alaoui, A. Ramdas, F. Krzakala, L. Zdeborova, and M. I. Jordan. Decoding from pooled data: Sharp information-theoretic bounds. *arXiv:1611.09981*, 2016.
- [2] A. E. Alaoui, A. Ramdas, F. Krzakala, L. Zdeborova, and M. I. Jordan. Decoding from pooled data: Phase transitions of message passing. *arXiv:1702.02279*, 2017.
- [3] M. Aldridge, L. Baldassini, and O. Johnson. Group testing algorithms: bounds and simulations. *IEEE Transactions on Information Theory*, 60(6):3671–3687, 2014.
- [4] G. K. Atia and V. Saligrama. Boolean compressed sensing and noisy group testing. *IEEE Transactions on Information Theory*, 58(3):1880–1901, 2012.
- [5] N. H. Bshouty. Optimal algorithms for the coin weighing problem with a spring scale. *The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 2009*.
- [6] N. H. Bshouty. On the coin weighing problem with the presence of noise. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 15th International Workshop*, 2012.
- [7] N. H. Bshouty and H. Mazzawi. Algorithms for the coin weighing problems with the presence of noise. *Electronic Colloquium on Computational Complexity*, 2011.
- [8] M. Bun, J. Ullman, and S. Vadhan. Fingerprinting codes and the price of approximate differential privacy. *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, 2014.

- 
- [9] W.-N. Chen, H.-C. Chen, and I.-H. Wang. On the fundamental limits of heterogeneous distributed detection: Price of anonymity. *IEEE International Symposium on Information Theory (ISIT)*, June 2018.
 - [10] W.-N. Chen and I.-H. Wang. Partial data extraction via noisy histogram queries: Information theoretic bounds. *Proceedings of IEEE International Symposium on Information Theory*, 2017.
 - [11] W.-N. Chen and I.-H. Wang. Partial data extraction via noisy histogram queries: Information theoretic bounds. *Manuscript*, January 2017.
<http://homepage.ntu.edu.tw/%7Eihwang/Research/Eprints/isit17nq.pdf>.
 - [12] W.-N. Chen and I.-H. Wang. Anonymous heterogeneous distributed detection: Optimal decision rules, error exponents, and the price of anonymity. *arXiv:1805.03554*, 2018
<https://arxiv.org/abs/1805.03554>.
 - [13] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Number 0471241954. Wiley-Interscience, 2006.
 - [14] I. Csiszár. A simple proof of Sanov’s theorem. *Bull. Braz. Math. Soc. (N.S.)*, 2006.
 - [15] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*, volume Stochastic Modelling and Applied Probability of 38. Springer-Verlag, 2010.
 - [16] F. den Hollander. *Large Deviations*. Number 14 in Fields Institute Monographs. American Mathematical Society, 2000.
 - [17] I. Dinur and K. Nissim. Revealing information while preserving privacy. *Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 202–210, 2003.
 - [18] C. Dwork, F. McSherry, and K. Talwar. The price of privacy and the limits of lp decoding. *Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC)*, pages 85–94, June 2007.

- 
- [19] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, pages 211–407, 2013.
- [20] C. Dwork and S. Yekhanin. New efficient attacks on statistical disclosure control mechanism. *In Proceedings of the Advances in Cryptology-CRYPTO*, pages 469–480, 2008.
- [21] C. George and R. L. Berger. *Statistical inference*. Duxbury, 2002.
- [22] W. Hoeffding. Asymptotically optimal tests for multinomial distributions. *Annals of Mathematical Statistics*, 36(2):369–401, 1965.
- [23] P. J. Huber. A robust version of the probability ratio test. *Annals of Mathematical Statistics*, 36(6):1753–1758, 1965.
- [24] P. J. Huber and V. Strassen. Minimax tests and the neyman-pearson lemma for capacities. *Annals of Statistics*, 1(2):251–263, 1973.
- [25] P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 2010.
- [26] B. Kailkhura, Y. S. Han, S. Brahma, and P. K. Varshney. Asymptotic analysis of distributed Bayesian detection with Byzantine data. *IEEE Signal Processing Letters*, 22, 2015.
- [27] T. Laarhoven. Asymptotics of fingerprinting and group testing: Tight bounds from channel capacities. *IEEE Transactions on Information Forensics and Security*, 10(9):1967–1980, 2015.
- [28] L. Lamport, R. Shostak, and M. Pease. The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4, July 1982.
- [29] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Number 978-0-387-27605-2. Springer-Verlag New York, 2005.

- 
- [30] S. Marano, V. Matta, and L. Tong. Distributed detection in the presence of byzantine attacks. *IEEE Transactions on Signal Processing*, 57(1):16–29, January 2009.
 - [31] J. Matoušek and J. Vondrák. The probabilistic method. *Lecture Notes*, March 2008.
 - [32] Y. Polyanskiy and Y. Wu. Lecture notes on information theory. August 2017.
 - [33] D. Robert. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.
 - [34] H. Royden and P. Fitzpatrick. *Real Analysis*. Pearson, 2010.
 - [35] J. Scarlett and V. Cevher. Phase transitions in group testing. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2016.
 - [36] J. Scarlett and V. Cevher. Phase transitions in the pooled data problem. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.
 - [37] V. Y. Tan and G. K. Atia. Strong impossibility results for sparse signal processing. *Signal Processing Letters, IEEE*, 21(3):260–264, 2014.
 - [38] R. R. Tenney and N. R. Sandell. Detection with distributed sensors. *IEEE Transactions on Aerospace and Electronic Systems*, 1981.
 - [39] J. N. Tsitsiklis. Decentralized detection by a large number of sensors. *Mathematics of Control, Signals and Systems*, 1(2):167–182, 1988.
 - [40] J. N. Tsitsiklis. Decentralized detection. In H. V. Poor and J. B. Thomas, editors, *Advances in Statistical Signal Processing*, volume 2. JAI Press Inc., 1990.
 - [41] V. V. Veeravalli, T. Başar, and H. V. Poor. Minimax robust decentralized detection. *IEEE Transactions on Information Theory*, 40(1):35–40, January 1994.
 - [42] I.-H. Wang, S.-L. Huang, and K.-Y. Lee. Extracting sparse data via histogram queries. *Proceedings of Annual Allerton Conference on Communications, Control, and Computing*, September 2016.

- 
- [43] I.-H. Wang, S.-L. Huang, K.-Y. Lee, and K.-C. Chen. Data extraction via histogram and arithmetic mean queries: Fundamental limits and algorithms. *Proceedings of IEEE International Symposium on Information Theory*, July 2016.
- [44] O. Zeitouni and M. Gutman. On universal hypothesis testing via large deviations. *IEEE Transactions on Information Theory*, 37(2):285–290, March 1991.
- [45] O. Zeitouni, J. Ziv, and N. Merhav. When is the generalized likelihood ratio test optimal? *IEEE Transactions on Information Theory*, 38(5):1597–1602, 1992.