HYPOTHESIS TESTING AND LEARNING WITH SMALL SAMPLES

BY

DAYU HUANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Adjunct Professor Sean P. Meyn, Chair
Professor Richard E. Blahut
Associate Professor Olgica Milenkovic
Professor Venugopal V. Veeravalli

ABSTRACT

Statistical hypothesis testing is a method to make a decision among two or more hypotheses using measurement data. It includes, for instance, deciding whether a system is in its normal state based on sensor measurements, or whether a person is healthy using data from medical tests. We are interested in the situation where the amount of measurement data available is sometimes limited, and the statistical models under the hypotheses have significant uncertainties: for example, a system could have many different abnormal states.

The goal of this thesis is to develop appropriate analysis methods for hypothesis testing problems with a small number of observations and uncertainties regarding the hypotheses. We focus on two problems: a universal hypothesis testing problem and a binary classification problem. In the first problem, only one of the hypotheses has a clearly specified statistical model. In the second problem, the statistical model under either hypothesis is only partially known and training data are available to help learn the model.

For both problems, existing analysis using large deviations has been shown to be a useful tool that leads to asymptotically optimal tests. However, the classical error exponent criterion that forms the foundation of this theory is not applicable for problems where the number of observations is relatively small compared to the number of possible outcomes in each observation (or the size of the observation alphabet). We introduce a new performance criterion based on large deviations analysis that generalizes the classical error exponent. The generalized error exponent characterizes how the probability of error depends on the number of observations and the observation alphabet size. It leads to optimal or near-optimal tests and new insights on some existing tests.

The generalized error exponent analysis, as well as the classical CLT and error exponent analysis, reveals how the size of the alphabet, or more generally the number of features, affects a test's performance. Results from these analyses suggest that quantizing the observation or selecting a subset of features could

help improve a test. We develop an optimization-based algorithm that learns the appropriate features from training data.

To my parents and my fiancée Fangxue Zheng

ACKNOWLEDGMENTS

This thesis would not have been possible without the support of my adviser, Professor Sean Meyn. He introduced me to the beautiful geometric interpretation of large deviations, which inspires the research reported in this thesis. He always encourages me to think both more deeply and more broadly about a problem. His patience and enthusiasm gave me an enormous amount of support.

I wish to express my sincere gratitude to my thesis committee members, Professor Venugopal Veeravalli, Professor Olgica Milenkovic and Professor Richard Blahut. I would also like to thank Professor Pierre Moulin for helpful discussions on the composite hypothesis testing problem.

Two of my close collaborators, Kun Deng and Jayakrishnan Unnikrishnan, deserve acknowledgment for sharing ideas as well as experiences. I would also like to thank many of my colleagues with whom I have had many helpful research discussions. I especially want to thank Wei Dai, Farzad Hassanzadeh, Wei Chen, Ankur Kulkarni, Jianchao Yang, Yu Sun, Yun Li and Anupama Kowli. I learned a lot from them. Life in Urbana-Champaign has been more colorful because of the friends I have met in CSL, on Green Street, or on the basketball court, and I feel grateful for the time spent together.

These acknowledgments would not be complete without mention of my parents and my fiancée Fangxue Zheng. Their support and love made me the person I am. This thesis is dedicated to them.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

In a hypothesis testing problem, we are given a sequence of observations and a few possible hypotheses regarding the observations. The goal is to decide which hypothesis is most likely to be true. The focus of this thesis is on hypothesis testing problems where the number of observations is relatively small compared to the complexity of the hypotheses. We study the universal hypothesis testing and the binary classification problem, in which the complexity is captured by the number of possible observations (or equivalently, the size of the observation alphabet). We are interested in two questions:

1. What are the appropriate tools for analyzing statistical tests when the number of observations is not large compared to the size of the observation alphabet?

2. How can the insights obtained from the analysis be used to design new tests?

1.1 Motivation

Universal hypothesis testing has applications in problems such as detecting abnormal network traffic. The task is to monitor the network data traffic and detect whether it is normal or an attack. A statistical approach to this problem is based on a comparison of the probability distributions of the observations in the two situations – normal behavior, or behavior during an attack [1]. There is significant uncertainty regarding the probability distributions of abnormal traffic data. This can be formulated as the universal hypothesis testing problem in which there are two hypotheses. Under the *null hypothesis*, the traffic is normal and the observations have a specified distribution while the *alternative hypothesis* is that the probability distribution is different from that of the normal traffic. It is usually desirable to detect the anomaly with a small number of observations.

1.2 Related Work

The performance of a test can be characterized by its probability of error. Since the probability of error usually has a complicated formula, a common tool used to gain insights into a test's performance is the asymptotic analysis which approximates the actual probability of error by its limit when some parameters are taken to their asymptotic limits. For the universal hypothesis testing problem, there are three predominant types of asymptotic analysis:

1. Asymptotic consistency analysis: This type of analysis characterizes whether the probability of error decreases to zero in the asymptotic limit. Finer results on the probability of error are obtained in the central limit theorem (CLT) and large deviations analysis.

2. CLT analysis [2]: CLTs are applied to obtain asymptotic approximations of the probability distribution of the test statistic under the hypothesis. To apply CLTs, the threshold in the test should be close to the expectation of the test statistic.

3. Large deviations analysis [3, 4]: The normalized limit of the logarithm of the probability of error is obtained. This limit is called the error exponent. Comparing to the CLT analysis, the difference between the detection threshold and the expectation of the test statistic is larger. The large deviations analysis gives insights that are not available in the CLT analysis, such as the characterization of worst-case distributions. The error exponent has also proven to be useful in test synthesis since it can be used as a surrogate for the probability of error.

When applying asymptotic analysis, it is critical to choose the appropriate asymptotic setting. For example, the Hoeffding test was shown in [4] to be asymptotically optimal for universal hypothesis testing in large deviations analysis under a particular asymptotic setting in which the alphabet size is fixed and the number of observations increases to infinity. The Hoeffding test has a drawback for problems with a large alphabet: For a fixed observation length, its probability of error increases significantly as the alphabet size increases. This drawback can be characterized in refined large deviations [5] or CLT analysis.

For *small sample* problems where the number of observations is significantly smaller than the size of the observation alphabet, the asymptotic setting of [4] is no longer appropriate. The relationship between the number of observations and

the alphabet size is better captured by the high-dimensional model in which both the number of observations and the size of alphabet increase simultaneously. The CLT analysis has been extended to this model [6, 7]. For large deviations analysis, it has been shown in [8] that the classical error exponent is not applicable. In this thesis, we develop generalizations of the error exponent criterion to small sample universal hypothesis testing problems.

The asymptotic analysis suggests methods to improve a test's performance. The CLT and refined large deviations analysis of the Hoeffding test and the related Pearson's chi-square test shows that the alphabet size has a negative impact on the performance, and also shows that the performance could be improved by quantizing the observation alphabet. In this quantization approach, several symbols in the observation alphabet are mapped into one bin, and the test is applied to the new observation alphabet. The optimal choice of bin size has been studied in [9, 10, 11]. The optimal bin size depends on many factors, including the set of alternative distributions.

1.3 Contributions of This Thesis

This thesis contributes to the analysis and design of tests for small sample hypothesis testing problems. The contributions are as follows:

1. We propose a new generalized error exponent criterion for analyzing statistical tests. It is based on identifying the appropriate normalization in large deviations for small sample hypothesis testing problems.

2. The new large deviations analysis is applied to the universal hypothesis testing and binary classification problems. We characterize the best achievable probability of error in each case by showing a pair of upper and lower bounds on the probability of error.

   For the universal hypothesis testing problem, a class of tests including the coincidence-base test is shown to be optimal, while Pearson's chi-square test is shown to be suboptimal.

   For the binary classification problem, our results show how the test and training samples and the size of observation alphabet affect the test's performance. We characterize the region in which there is an asymptotically

consistent test in terms of the number of test and training samples as well as the size of alphabet. The ℓ_2-based test is shown to be suboptimal.

3. The negative impact of a large alphabet size, revealed in asymptotic analysis, can be alleviated by approximating the observations using features that are more relevant for the hypotheses. An example of features is the quantized observation in [9]. The asymptotic approximation obtained in large deviations analysis can be used as the criterion for finding the features. This leads to a feature extraction algorithm proposed in this thesis.

1.4 Outline

The thesis is organized as follows: The preliminaries on universal hypothesis testing and asymptotic analysis are presented in Chapter 2. The generalized error exponent analysis for small sample universal hypothesis testing is presented in Chapter 3. This analysis tool is also applied to the binary classification problem, and the results are presented in Chapter 4. An algorithm for extracting features to alleviate the negative impact of a large alphabet size is described in Chapter 5. Conclusions are provided in Chapter 6 with discussion of future research directions.

CHAPTER 2

PRELIMINARIES

In this chapter, we introduce the universal hypothesis testing problem and describe the large deviations and CLT analysis. We restrict to a model in which observations are i.i.d. (independent and identically distributed).

2.1 Universal Hypothesis Testing

Suppose an i.i.d. sequence $\boldsymbol{Z}_1^n = \{Z_1, \ldots, Z_n\}$ is observed, where $Z_i \in [m] := \{1, 2, \ldots, m\}$. Under the null hypothesis $H0$, Z_i has the distribution π. Under the alternative hypothesis, Z_i has a distribution $\mu \in \Pi_m$, and the exact μ is not known. The set of alternative distribution Π_m is given by

$$\Pi_m := \{\mu : d(\mu, \pi) \geq \varepsilon\} \tag{2.1}$$

where the function d measures the "distance" between two distributions. For example, d can be the *total-variation distance* or the *Kullback-Leibler (KL) divergence*. The Kullback-Leibler divergence for two probability distributions μ, π is defined as

$$D(\mu\|\pi) = \begin{cases} \sum_j \mu_j \log(\mu_j/\pi_j) & \mu \preceq \pi \\ \infty & \text{otherwise} \end{cases}$$

A test $\phi = \{\phi_n\}_{n\geq 1}$ is given by a sequence of binary-valued functions $\phi_n : [m]^n \to \{0, 1\}$. The test decides in favor of $H1$ if $\phi_n(\boldsymbol{Z}_1^n) = 1$, and otherwise in favor of $H0$. The performance of a test is evaluated using the probability of false alarm P_F and probability of missed detection P_M, defined as

$$\begin{aligned} &P_F(\phi_n) = \mathsf{P}_\pi\{\phi_n(\boldsymbol{Z}_1^n) = 1\}, \\ &P_M(\phi_n, \mu) = \mathsf{P}_\mu\{\phi_n(\boldsymbol{Z}_1^n) = 0\}, \quad P_M(\phi_n) = \sup_{\mu\in\Pi_m} P_M(\phi_n, \mu). \end{aligned}$$

The main goal is to design a test with small P_F and P_M.

The foundation of this thesis, as in much of information theory, involves the sequence of empirical distributions denoted,

$$\Gamma_j^n := \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}\{Z_t = j\}, \quad j \in [m], n \geq 0,$$

where the function $\mathbb{I}\{A\}$ is the indicator function that takes value 1 when A holds, and value 0 otherwise. In the i.i.d. setting considered here, it is known that there is no loss of generality in restricting to tests that can be represented by its decision region which is a subset of the space of probability distributions: the test decides in favor of $H1$ if and only if Γ^n falls in to the decision region. Mathematically, a test ϕ takes the following form:

$$\phi_n(\boldsymbol{Z}_1^n) = \mathbb{I}\{S_n \geq s_n + \tau_n\}, \tag{2.2}$$

where S_n is the test statistic, s_n is a normalization term usually chosen as the mean of S_n, and τ_n is the threshold. The test statistics of many important tests are functions of the empirical distribution.

Examples of universal tests include the Hoeffding test, Pearson's chi-square test, and the coincidence-based test designed for the case where n is much smaller than m. These test statistics belong to the class of *separable statistics* (see [6]). A separable statistic is a test statistic of the form

$$S_n = \sum_{j=1}^{m} f_j(n\Gamma_j^n).$$

General theorems on asymptotic distributions and asymptotic moments of separable statistics are available in [6]. Large-deviations analysis for the case $m = O(n)$ is given in [12, 13].

The Hoeffding test can be defined using the Kullback-Leibler divergence. The Hoeffding test statistic is given by

$$S_n^{\mathsf{H}} = D(\Gamma^n \| \pi). \tag{2.3}$$

Pearson's chi-square test statistic after normalization is given by

$$S_n^{\mathsf{P}} = \frac{n}{m}\sum_{j=1}^{m} \frac{(n\Gamma_j^n - n\pi_j)^2}{n\pi_j}. \tag{2.4}$$

Note that Pearson's chi-square statistic is related to the Hoeffding test statistic via the second order Taylor-series approximations for Γ^n close to π.

$$S_n^{\mathsf{H}} = \frac{m}{2n^2} S_n^{\mathsf{P}} + O_n(\|\Gamma^n - \pi\|_1^3).$$

The coincidence-based test, designed for the case where π is the uniform distribution, was introduced in [14]. Its test statistic is given by

$$S_n^* = -\sum_{j=1}^{m} \mathbb{I}\{n\Gamma_j^n = 1\}. \tag{2.5}$$

Applications to continuously-valued observations

Tests designed for finite-valued observations can be applied to solve a universal hypothesis testing problem with continuously valued observations by first partitioning observation space. Suppose the sequence of observations is given by $\boldsymbol{Y}_1^n = \{Y_1, \ldots, Y_n\}$ with $Y_i \in [0, 1]$. We have two hypotheses:

$$H0 : Y_i \sim P, \quad H1 : Y_i \sim Q \in \mathcal{Q} \tag{2.6}$$

where P is absolutely continuous with respect to Lebesgue measure. Consider a partition of the interval $[0, 1]$:

$$\mathbf{A}_n = \{A_1, \ldots, A_m\}.$$

The observation Y_i is mapped to a finite alphabet of size m, and the tests mentioned above such as the Hoeffding test and Pearson's chi-square test are applicable.

The coincidence-based test is applicable if we choose the partition so that

$$P(A_j) = \frac{1}{m}, \tag{2.7}$$

and m is significantly large than n. Related conclusions can be found in Chapter 3 and [14].

2.2 Classical Model of Asymptotic Analysis

2.2.1 Law of large numbers

The law of large numbers is used to show that under the null hypothesis

$$\lim_{n\to\infty} \frac{1}{N_n}(S_n - s_n - \tau_n) \leq 0;$$

and under the alternative hypothesis

$$\lim_{n\to\infty} \frac{1}{N_n}(S_n - s_n - \tau_n) > 0,$$

where N_n is a normalization sequence.

2.2.2 Central limit theorem

In central limit theorem analysis, the threshold is chosen to be close to the expectation of the test statistic so that the central limit theorem can be applied to give asymptotic approximations to the probability of error. Taking the Hoeffding test as an example, the large deviations analysis is applied to the case where the test is given by

$$\mathbb{I}\{D(\Gamma^n \| \pi^0) \geq \tau\},$$

where τ is a fixed threshold. On the other hand, the central limit theorem analysis is applied to the case

$$\mathbb{I}\{D(\Gamma^n \| \pi^0) \geq \frac{1}{n}\tau\}.$$

It can be shown that

$$\lim_{n\to\infty} \mathsf{P}\{D(\Gamma^n \| \pi^0) \geq \frac{1}{n}\tau\} = 1 - F_{\chi^2_{m-1}}(2\tau),$$

where the $F_{\chi^2_{m-1}}$ is the cumulative distribution function of a chi-square random variable with $m-1$ degrees of freedom.

To apply the central limit theorem analysis for both the probability of missed detection and false alarm, the distance between the set of alternative distributions and the null distribution needs to decrease with n so that both error events fall into the region suitable for central limit theorem analysis. For example, the following

alternative distribution could be considered in central limit theorem analysis:

$$\Pi_m := \{\mu : d(\mu, \pi) \geq \frac{1}{\sqrt{n}}\varepsilon\}. \tag{2.8}$$

Unfortunately, the approximation to the probability of missed detection obtained using this method is generally not as good as that for the probability of false alarm.

2.2.3 Large deviations

The classical error exponent criterion is defined as follows:

$$I_F(\phi) := -\limsup_{n\to\infty} \log P_F(\phi_n), \quad I_M(\phi) := -\limsup_{n\to\infty} \log P_M(\phi_n). \tag{2.9}$$

Our use of the term error exponent follows [15]. The error exponent gives the following approximation to the probability of error:

$$P_F(\phi_n) = e^{-nI_F + O(\log(n))}, \quad P_M(\phi_n) = e^{-nI_M + O(\log(n))}. \tag{2.10}$$

The following result in [4] established the asymptotic optimality of the Hoeffding test in terms of the error exponent criterion:

Theorem 2.1. *The Hoeffding test is optimal with respect to the error exponent criterion:*

$$I_M(\phi^{\mathsf{H}}) = \sup\{I_M(\phi) : I_F(\phi) \geq I_F(\phi^{\mathsf{H}})\}$$

In this error exponent analysis, it is not clear whether m affects the performance of the Hoeffding test. In fact, the approximation given by the leading term in (2.10) is very poor for large m. Refined large deviation analysis is needed to better approximate the probability of error when m is large. The result in [16], which generalizes the Bahadur-Rao result (See [5]), can be applied to obtain the following sharp asymptotic approximation:

$$\begin{aligned} P_F(\phi_n^{\mathsf{H}}) &= n^{(m-3)/2} e^{-nI_F}(c_F + O(\frac{1}{n})), \\ P_M(\phi_n^{\mathsf{H}}) &\leq n^{-1/2} e^{-nI_M}(c_M + O(\frac{1}{n})). \end{aligned} \tag{2.11}$$

Comparing the refined large deviations in (2.11) with the classical error exponent result in (2.10), the negative impacts of alphabet size on the performance of

the Hoeffding test are revealed in the refined large deviations.

2.3 High-Dimensional Model

In the classical asymptotic settings surveyed in Section 2.2, m is assumed to be fixed while $n \to \infty$. Analysis in this asymptotic setting is more suitable when n is larger than or comparable to m. For problems where n is much smaller than m, a different asymptotic setting is needed. A popular and useful one is the high-dimensional model, in which we assume

$$n \to \infty, m \to \infty.$$

The dependence between n and m plays a significant role in analysis: the *small sample* case when $n/m \to 0$ has a different nature than the *large sample* case with $n/m \to \infty$. When $n/m \to \infty$, the number of samples per bin increases to infinity, and eventually the underlying probability distribution can be estimated. This does not hold for the *small sample* case where m increases faster than n. In the small sample case, both the large deviations and central limit theorem analysis need a different set of tools. For central limit theorem analysis, an important analysis method for the separable statistics was developed in [6]. For large deviations analysis, the classical method of types no longer applies, and this thesis provides a new tool for this case.

CHAPTER 3

GENERALIZED ERROR EXPONENT FOR SMALL SAMPLE UNIVERSAL HYPOTHESIS TESTING

In this chapter, we develop the generalized error exponent analysis for the small sample universal hypothesis testing problem.

3.1 Problem Statement

We consider a sequence of universal hypothesis testing problems, each with a finite number of outcomes (a finite alphabet). A sequence of i.i.d. observations $\boldsymbol{Z}_1^n$ where $Z_i \in [m] := \{1, 2, \ldots, m\}$ is given. Let $\mathcal{P}_m$ denote the collection of probability mass functions (pmf's) on $[m]$. Let π be the uniform distribution on $[m]$:

$$\pi_j = 1/m \text{ for } 1 \leq j \leq m. \tag{3.1}$$

The set of alternative pmf's is given by

$$\Pi_m := \{\mu \in \mathcal{P}_m : d(\mu, \pi) \geq \varepsilon\}, \tag{3.2}$$

where d is taken to be the total variation distance defined for any pair of pmf's on $[m]$:

$$d_{TV}(\mu, \pi) = \sup_{B \subseteq [m]} \{|\mu(B) - \pi(B)|\}.$$

A test $\phi = \{\phi_n\}_{n \geq 1}$ is given by a sequence of binary-valued functions $\phi_n : [m]^n \to \{0, 1\}$. The test decides in favor of $H0$ if $\phi_n(\boldsymbol{Z}_1^n) = 0$. The test is evaluated using the probabilities of false alarm and missed detection:

$$P_F(\phi_n) = \mathsf{P}_\pi\{\phi_n(\boldsymbol{Z}_1^n) = 1\},$$
$$P_{M,\mu}(\phi_n) = \mathsf{P}_\mu\{\phi_n(\boldsymbol{Z}_1^n) = 0\}, P_M(\phi_n) = \sup_{\mu \in \Pi_m} \mathsf{P}_\mu\{\phi_n(\boldsymbol{Z}_1^n) = 0\}.$$

3.2 Related Work

In this section, we review related results with emphasis on the type of analysis used and the asymptotic settings considered. Many of the results reviewed apply to cases more general than (3.1). The list of references in [7], [8] and [17, Chapter 26, 27 and 34] are good starting points for other related work.

Existing results differ in the asymptotic setting considered, which can be classified into three cases: 1) m is fixed; 2) m is increasing and $m = O(n)$; 3) $n = o(m)$ and $m = o(n^2)$. There is no need to consider the case $n = O(\sqrt{m})$ because the converse result (lower-bounds on probability of error) established in [14] indicates that no asymptotically consistent test exists if $n = O(\sqrt{m})$.

Consider the case where m is fixed.

a) Pearson's chi-square statistic and GLRT statistic are asymptotically distributed as a chi-square distribution whose degree of freedom is $m - 1$. These results and their extensions can be found in [2, 18, 19, 20, 21, 22].

b) The performance of Pearson's chi-square test and GLRT is investigated using a large deviations in [4] . The classical error exponent criterion in (2.9) is used to evaluate a test ϕ. The GLRT is shown to have *optimal* error exponents while Pearson's chi-square test does not.

Next consider the case $m = O(n)$.

a) Pearson's chi-square test and GLRT are both asymptotically consistent (for example, see [7]).

b) Pearson's chi-square statistic and the GLRT statistic both have asymptotically normal distributions. The first work in this line is [23]. Extensions of this result can be found in [24, 25, 26, 27, 11, 28].

c) A lower-bound on the best achievable probability of error in CLT analysis is given by Ermakov in [7]: Under the condition

$$0 < \liminf_{n\to\infty} \frac{\varepsilon}{\sqrt{m}} \leq \limsup_{n\to\infty} \frac{\varepsilon}{\sqrt{m}} < \infty,$$

Pearson's chi-square test is asymptotically minimax. That is, for any test whose P_F is no larger than that of Pearson's chi-square test, its P_M is asymptotically no smaller than that of Pearson's chi-square test. Ermakov's result applies to the range of m satisfying $m = o(n^2)$.

d) An achievability result (a lower-bound on the error exponent) and a complementing converse result (an upper-bound on the error exponent) in the large deviations analysis, have been obtained in [8]: There exists a test for which P_F and P_M both decay *exponentially* fast with respect to n; i.e., I_F and I_M defined in (4.1) are both nonzero, if and only if $m = O(n)$. Other large deviations and moderate-deviations analyses of GLRT and Pearson's chi-square test can be found in [29, 10, 30, 12, 13, 31].

Finally consider the small sample case where $n = o(m)$ and $m = o(n^2)$.

a) Pearson's chi-square test is known to be asymptotically consistent [7]. Two others tests shown to be asymptotically consistent are the test based on counting pairwise-collisions [32] and the coincidence-based test [14]. An approach to extend tests designed for the uniform null distribution to a non-uniform null distribution has been proposed in [33].

b) Ermakov's result on asymptotic minimaxity of Pearson's chi-square test also applies to this case.

c) Results on the asymptotic distribution of Pearson's chi-square statistic and the GLRT statistic have been obtained in [6, 34].

To the best of our knowledge, the proper normalization has not been identified before for the large deviations analysis in this case.[1] We note that the classical error exponent analysis is not suitable.

3.3 Summary of Results

The new large deviations framework proposed here is motivated by and analogous to the classical error exponent (4.1) in the large sample case. While the classical error exponent is defined with the normalization n, for the small sample problem considered here, our main results imply that the following generalized error exponent is best for asymptotic analysis, defined with respect to the normalization

[1] Combining the upper-bounds on probability of error given in [14, 33] with the Chernoff inequality give a loose upper-bound on the probability error and do not yield the proper normalization.

$r(m,n) = n^2/m$:

$$\begin{aligned} J_F(\phi) &:= -\limsup_{n\to\infty} \frac{1}{r(m,n)} \log(P_F(\phi_n)), \\ J_M(\phi) &:= -\limsup_{n\to\infty} \frac{1}{r(m,n)} \log(P_M(\phi_n)). \end{aligned} \tag{3.3}$$

The generalized error exponents give the following approximation to the probabilities of false alarm and missed detection:

$$P_F \approxeq e^{-r(n,m)J_F}, \quad P_M \approxeq e^{-r(n,m)J_m}. \tag{3.4}$$

The generalized error exponent provides new insights that are not available from asymptotic consistency, or CLT analysis. More precisely, the following results are established:

1. The best achievable probability of error, $P_e = \max\{P_F, P_M\}$, decays as $-\log(P_e) = r(n,m)J(1+o(1))$, where $r(n,m) = n^2/m$. This is applicable not only for the case where the set of alternative distributions is defined by the total variation distance, but also for a broad collection of distance or divergence functions.

2. A class of tests based on the separable statistics, including the coincidence-based test, is shown to achieve the *optimal* pair of generalized error exponents J_F and J_M:

$$J_M(\phi^*) = \max\{J_M(\phi) : J_F(\phi) \geq J_F(\phi^*)\}.$$

The *exact* formulae for these generalized error exponents are obtained.

3. The performance of Pearson's chi-square test is asymptotically worse than the optimal test.

Part of the results has been published in [35].

3.4 Generalized Error Exponents

In this section, we describe the main results on the proper normalization for large deviations analysis for the small sample universal hypothesis testing problem. The following assumption is imposed throughout:

Assumption 3.1. *$n = o(m)$ and $m = o(n^2)$.*

To show that the proper normalization to be used in the definition of generalized error exponent is n^2/m, we need to establish:

1. There is a test for which both generalized error exponents are non-zero, and therefore this normalization is not too large.

2. For any test, at least one of the generalized error exponents is finite, and therefore this normalization is not too small.

These are established in Theorem 3.2 and Theorem 3.3. These two theorems characterize the region of (J_F, J_M) that is achievable. This is depicted in Fig. 3.1. The boundary of the achievable region is given by the following pair of functions: For $\tau \in [0, \underline{\kappa}(\varepsilon) - 1]$,

$$\begin{aligned} J_F^*(\tau) &:= \sup_{\theta \geq 0} \{\theta\tau - \tfrac{1}{2}\big(e^{2\theta} - (1+2\theta)\big)\}, \\ J_M^*(\tau) &:= \sup_{\theta \geq 0} \{\theta(\underline{\kappa}(\varepsilon) - 1 - \tau) - \tfrac{1}{2}\big(e^{-2\theta} - (1-2\theta)\big)\underline{\kappa}(\varepsilon)\}, \end{aligned} \tag{3.5}$$

where $\underline{\kappa} : \mathbb{R}_+ \to \mathbb{R}_+$ is the C^1 function,

$$\underline{\kappa}(\varepsilon) = \begin{cases} 1 + 4\varepsilon^2, & \varepsilon < 0.5, \\ 1 + \frac{\varepsilon}{1-\varepsilon}, & \varepsilon \geq 0.5. \end{cases} \tag{3.6}$$

Note that we always have $J_F^*(\tau) < \infty$ and $J_M^*(\tau) < \infty$. For $\tau \in (0, \underline{\kappa}(\varepsilon) - 1)$, we have $J_F^*(\tau) > 0$ and $J_M^*(\tau) > 0$.

Recall the coincidence-based test statistic given in (2.5). The coincidence-based test is given by $\phi^* = \mathbb{I}\{S_n^* \geq \mathsf{E}_\pi[S_n^*] + \tau_n\}$.

Theorem 3.2 (Achievability)**.** *The coincidence-based test ϕ^* achieves the generalized error exponent given in* (3.5)*; i.e., for any $\tau \in [0, \underline{\kappa}(\varepsilon) - 1]$, if the sequence of threshold $\{\tau_n\}$ is chosen so that,*

$$\tau = \lim_{n \to \infty} m\tau_n / n^2, \tag{3.7}$$

then the coincidence-based test has the generalized error exponents:

$$J_F(\phi^*) = J_F^*(\tau), \quad J_F(\phi^*) = J_M^*(\tau). \tag{3.8}$$

Theorem 3.3 (Converse)**.** *Consider any* $\tau \in [0, \kappa(\varepsilon)-1]$*. For any test* ϕ *satisfying*

$$J_F(\phi) \geq J_F^*(\tau),$$

the following upper-bound on the generalized error exponent of missed detection holds:

$$J_M(\phi) \leq J_M^*(\tau).$$

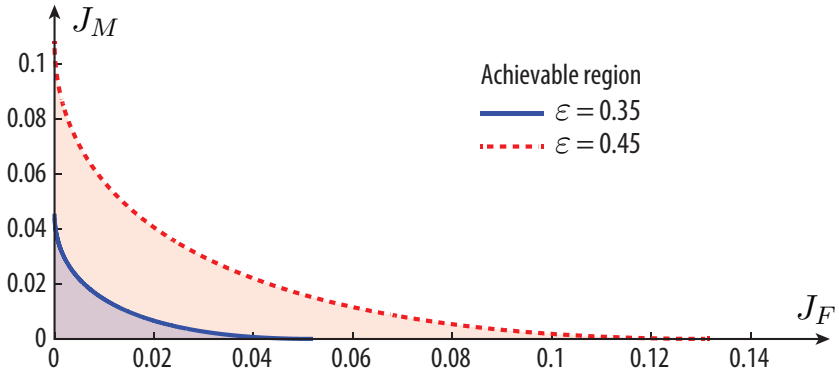

Figure 3.1: Achievable region when $\varepsilon = 0.35$ and $\varepsilon = 0.45$ given by the lower-bound in Theorem 3.2 and upper-bound in Theorem 3.3. The lower and upper bound meet over the entire region.

Theorems 3.2 and 3.3 indicate that the best possible probabilities of false alarm and missed detection decay as in (3.4). We now compare the approximation in (3.4) to the actual empirical performance of the coincidence-based test ϕ^*. The results are shown in Fig. 3.2 for $\varepsilon = 0.35$ and Fig. 3.3 for $\varepsilon = 0.45$. We choose the threshold τ based on (3.8) so that J_F and J_M are the same. The generalized error exponents give estimates of the *slopes* of $\log(P_F)$ and $\log(P_M)$ with respect to $r(n, m)$. It can be observed that the slope from the theoretical approximation by generalized error exponents approximately matches the slope of the simulated value. The remaining difference between the theoretical and the empirical slope in Fig. 3.3 is mainly due to two reasons: First, the threshold chosen is based on the first order approximation and it can be observed in the figure that the slope P_M is slightly smaller while the one for P_F is larger. Second, the generalized error exponent is only the first term in the asymptotic expansion of $\log(P_F)$ and $\log(P_M)$.

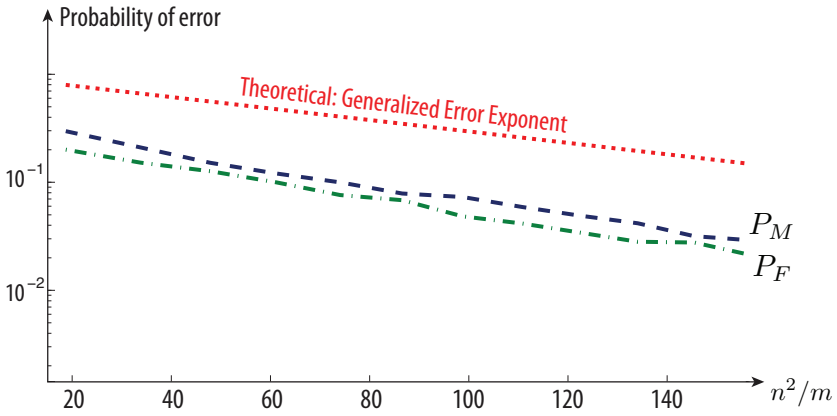

Figure 3.2: Performance of ϕ^* with $\varepsilon = 0.35$.

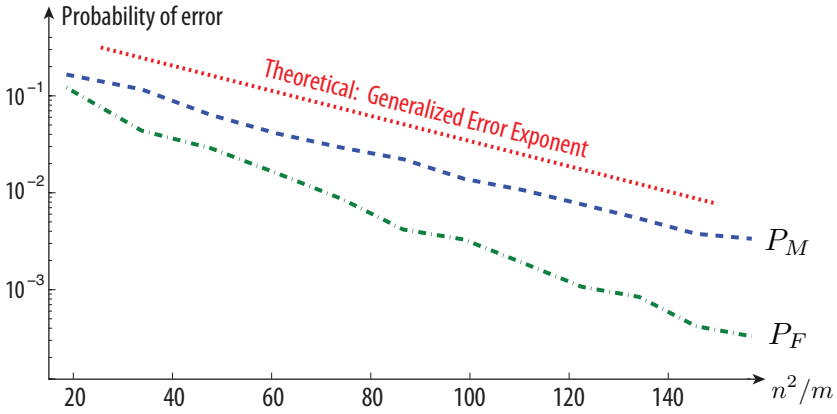

Figure 3.3: Performance of ϕ^* with $\varepsilon = 0.45$.

3.4.1 Rate function and worst-case distributions

Similar to the large deviations for the large sample case, we can define a rate function for the small sample case. This definition holds for a sequence of restricted set of alternative distributions:

$$\mathcal{P}_m^b = \{\mu \in \mathcal{P}_m : \max_j \mu_j \leq c_1 p_j\}, \tag{3.9}$$

where c_1 is a large positive constant satisfying $c_1 \geq \max\{2/(1-\varepsilon), 4\varepsilon\}$. In other words, this set of distributions has a bounded likelihood ratio with respect to π.

Consider the coincidence-based test ϕ^*. The rate function for this test is associated with a sequence of distributions $\boldsymbol{\mu} = \{\mu^{(1)}, \mu^{(2)}, \mu^{(3)}, \ldots\}$ with $\mu^{(n)} \in \mathcal{P}_m^b$ as follows:

$$J_{\boldsymbol{\mu}}(\phi^*, \tau) = -\limsup_{n\to\infty} \frac{m}{n^2} \log(\mathsf{P}_{\mu^{(n)}}\{S_n^* \leq \mathsf{E}_\pi[S_n^*] + \frac{n^2}{m}\tau\}).$$

We show that J is a function of the following quantity:

$$\kappa(\boldsymbol{\mu}) := \liminf_{n} \sum_{j} \frac{(\mu_j^{(n)})^2}{\pi_j}. \tag{3.10}$$

Theorem 3.4.

$$J_{\boldsymbol{\mu}}(\phi^*, \tau) = \sup_{\theta \geq 0} \{\theta(-1-\tau) - \tfrac{1}{2}(e^{-2\theta} - 1)\kappa(\boldsymbol{\mu})\}. \tag{3.11}$$

The rate function can be applied to identify the sequence of worst-case alternative distributions for which the probability of missed detection is asymptotically the largest: Note that $J_{\boldsymbol{\mu}}(\phi^*, \tau)$ is monotonically increasing in $\kappa(\boldsymbol{\mu})$. Therefore, the smaller the quantity $\kappa(\boldsymbol{\mu})$, the larger the probability of missed detection associated with $\boldsymbol{\mu}$. The sequence of distributions achieving the minimum $\kappa(\boldsymbol{\mu})$ is given in the following lemma:

Lemma 3.5. *When π is the uniform distribution, we have*

$$\inf_{\mu \in \Pi_m} \Big(\sum_{j=1}^{k} \frac{\mu_j^2}{\pi_j}\Big) = (1 + \underline{\kappa}(\varepsilon))(1 + o(1)). \tag{3.12}$$

The infimum is achieved by the following bi-uniform distribution:

1. *When $\varepsilon < 0.5$,*

$$\mu_j^* = \begin{cases} \frac{1}{m} + \frac{\varepsilon}{\lfloor m/2 \rfloor}, & j \leq \lfloor m/2 \rfloor, \\ \frac{1}{m} - \frac{\varepsilon}{\lceil m/2 \rceil}, & j > \lfloor m/2 \rfloor. \end{cases} \tag{3.13}$$

2. *When $\varepsilon \geq 0.5$,*

$$\mu_j^* = \begin{cases} \frac{1}{\lfloor m(1-\varepsilon) \rfloor}, & j \leq \lfloor m(1-\varepsilon) \rfloor, \\ 0, & j > \lfloor m(1-\varepsilon) \rfloor. \end{cases} \tag{3.14}$$

Thus, the worst case distributions are identified as bi-uniform distributions.

3.4.2 Sketch of the proof

The large deviations characterization of P_F for the coincidence-based test follows from the following asymptotic approximation of the logarithmic moment generat-

ing function of the test statistic:

$$\log\big(\mathsf{E}_\pi[\exp\{\theta(n-S_n^*)\}]\big) = \tfrac{1}{2}\frac{n^2}{m}\big(m\sum_{j=1}^{m}\pi_j^2\big)(e^{-2\theta}-1) + O(\frac{n^3}{m^2}) + O(1).$$

The characterization of P_M is obtained in a similar way. We will show that the probability of missed detection is dominated by the worst-case distributions given in Lemma 3.5. The details are given in Appendix A.1.

The main idea to prove the converse result is the following: We construct a sequence of events $\{B_{n,\tau,\delta}\}$ so that (i) the probability of these events can be lower-bounded based on the condition on P_F; (ii) the probability of making a missed detection conditioned on these events is lower-bounded. We use the following inequality:

$$\begin{aligned} P_M(\phi_n) &\geq \sup_{\mu\in\Pi_m} \mathsf{P}_\mu\big(\{\phi_n=0\}\cap B_{n,\tau,\delta}\big) \\ &\geq \sup_{\mu\in\Pi_m} \frac{\mu^n}{\pi^n}(\{\phi_n=0\}\cap B_{n,\tau,\delta})\mathsf{P}_\pi(\phi_n=0|B_{n,\tau,\delta})\mathsf{P}_\pi(B_{n,\tau,\delta}). \end{aligned}$$

A lower-bound on the last term follows from the construction of the events. The second term is lower-bounded using the assumption on the probability of false alarm. We construct a collection of distributions so that the first term is always lower-bounded on this set regardless of the test. These distributions are obtained by taking the "dominating" distribution μ^* given in (3.13), and permuting the symbols in $[m]$. Let U_m denote the collection of all subsets of $[m]$ whose cardinality is $\lfloor m/2\rfloor$. For each set $\mathcal{U}\in U_m$, define the distribution $\mu_\mathcal{U}$ as

$$\mu_{\mathcal{U},j} = \begin{cases} \frac{1}{m}+\frac{\varepsilon}{\lfloor m/2\rfloor}, & j\in\mathcal{U}; \\ \frac{1}{m}-\frac{\varepsilon}{\lceil m/2\rceil}, & j\in[m]\setminus\mathcal{U}. \end{cases}$$

Note that $\mu_\mathcal{U}\in\Pi_m$. To prove the converse result, we need to establish a lower-bound on

$$\sup_{\mathcal{U}\in U_m}\frac{\mu_\mathcal{U}^n}{\pi^n}(\{\phi_n=0\}\cap B_{n,\tau,\delta}).$$

The details are given in Appendix A.7.

3.5 Extensions of the Coincidence-Based Test

This section collects together extensions of Section 3.4 in terms of tests and models. We first propose a collection of tests that extend the coincidence-based test, and provide the freedom for fine-tuning the performance for finite samples. We then propose an extension of the coincidence-based test for non-uniform π.

3.5.1 Extending coincidence-based test with high-order statistics

The coincidence-based test uses only the number of symbols that appear in the sequence exactly once. We now add terms to the test statistic that also depend on the number of symbols appearing more than once to create a broader collection of tests. Conditions will be established under which these tests have optimal generalized error exponents. Consider the class of test statistics of the following form:

$$S_n^{*+} = S_n^* + \sum_{l=2}^{\bar{l}} v_l \mathbb{I}\{n\Gamma_j^n = l\}. \tag{3.15}$$

The test is given by

$$\phi^{*+}(\boldsymbol{Z}_1) = \mathbb{I}\{S_n^{*+} - \mathsf{E}_\pi[S_n^{*+}] \geq \tau_n\}.$$

Theorem 3.6. *If $\bar{l} < \infty$, $v_2 = 0$, and $v_l \geq 0$ for all $3 \leq l \leq \bar{l}$, then the test ϕ^{*+} achieves the optimal generalized error exponents given in* (3.5)*.*

The additional terms for $l \geq 3$ in the separable statistic give us ways to fine-tune the test for better finite sample performance. One interesting question is to obtain finer asymptotic approximations of $\log(P_F)$ and $\log(P_M)$ to obtain more insights on how these extensions of the coincidence-based test perform.

For the case with $v_2 \neq 0$, we have the following conjecture:

Conjecture 3.7. *If S_n^{*+} satisfies $\bar{l} < \infty$, $v_2 > -2$, and $v_l \geq 0$ for all $3 \leq l \leq \bar{l}$, then the test is optimal in terms of the generalized error exponent.*

3.5.2 Extensions to non-uniform π

The coincidence-based test (2.5) can be extended to the case where π is not necessarily uniform but the likelihood ratio between the π and the uniform distribution

remains bounded.

Assumption 3.8. *There exists a constant $\eta > 0$ such that $\max_j m\pi_j \leq \eta$ holds for all n.*

The following weighted coincidence-based test is considered:

$$S_n^{\mathsf{W}} = \sum_{j=1}^{m} f_j(n\Gamma_j^n)$$

with

$$f_j(n\Gamma_j^n) = \begin{cases} \frac{1}{2}n^2\pi_j^2, & n\Gamma_j^n = 0, \\ -n\pi_j, & n\Gamma_j^n = 1, \\ 1, & n\Gamma_j^n = 2, \\ 0, & \text{others.} \end{cases} \tag{3.16}$$

The test is given by $\phi_n^{\mathsf{W}} = \mathbb{I}\{S_n^{\mathsf{W}} \geq \tau_n\}$.

The choice of coefficients given in (3.16) ensures $\mathsf{E}_\nu[S_n^{\mathsf{W}}]$ approximates the ℓ_2-distance between ν and π:

Lemma 3.9. *For μ satisfying $\max_j m\pi_j \leq \eta$, the expectation of S_n^{W} is given by:*

$$\mathsf{E}_\nu[S_n^{\mathsf{W}}] = \tfrac{1}{2}\frac{n^2}{m}[m\sum_{j=1}^{m}(\nu_j - \pi_j)^2] + O(\frac{n^3}{m^2}).$$

The proposed test has nonzero generalized error exponents:

Theorem 3.10. *Suppose Assumption 3.1 and Assumption 3.8 hold. For $\tau \in (0, 2\varepsilon^2)$ where τ is defined in* (3.7)*, the test ϕ^{W} has nonzero generalized error exponents:*

$$J_F(\phi^{\mathsf{W}}) > 0, \quad J_M(\phi^{\mathsf{W}}) > 0.$$

3.6 Pearson's Test

In this section, we study the performance of Pearson's chi-square test whose test statistic is given in (2.4). Pearson'chi-square test is given by $\phi^{\mathsf{P}} = \mathbb{I}\{S_n^{\mathsf{P}} \geq \tau_n\}$. We find that this test has a zero generalized error exponent, and therefore its probability of error decays slower than the coincidence-based test.

The test statistic of Pearson's chi-square test is also a separable statistic: $S_n^{\mathsf{P}} = \sum_{j=1}^m f_j(n\Gamma_j^n)$ with $f_j(n\Gamma_j^n) = (n\Gamma_j^n)^2 - \frac{n}{m}$ for uniform π. An important difference between this test statistic and the statistics of the class of optimal tests identified in Theorem 3.6 is that f_j for each the optimal test is bounded, while this is not true in Pearson's chi-square test.

Pearson's chi-square test is also asymptotically consistent in the small sample case:

Proposition 3.11 (Asymptotic consistency)**.** *Under Assumption 3.1, there exists a sequence of thresholds $\{\tau_n\}$, with which the Pearson's chi-square test is asymptotically consistent:*

$$\lim_{n\to\infty} P_F(\phi_n^{\mathsf{P}}) = 0, \quad \lim_{n\to\infty} P_M(\phi_n^{\mathsf{P}}) = 0.$$

The proof of the proposition highlights the connection between Pearson's chi-square test and the coincidence-based test.

Proof of Proposition 3.11. Take $\tau_n = n + \alpha\frac{n^2}{m}$ with $\alpha = \frac{1}{2}\underline{\kappa}(\varepsilon)$. Applying Chebyshev's inequality together with the approximations to the expectation and variance of the test statistic given in the following lemma, we conclude that $\lim_{n\to\infty} P_F(\phi^{\mathsf{P}}) = 0$.

Lemma 3.12. *For any $\nu \in \mathcal{P}_m^b$, the expectations and variance of S_n^{P} is given by*

$$\begin{aligned}
\mathsf{E}_\nu[S_n^{\mathsf{P}}] &= n + \frac{n^2}{m}(m\sum_{j=1}^m(\nu_j - \pi_j)^2) + O(\frac{n^3}{m^2}),\\
\mathrm{var}_\nu[\mathrm{S_n^P}] &= 2\frac{n^2}{m}(m\sum_{j=1}^m \nu_j^2)(1 + o(1)).
\end{aligned}$$

We bound $P_M(\phi_n^{\mathsf{P}})$ by coupling Pearson's chi-square statistic with the coincidence-based test statistic:

$$\begin{aligned}
S_n^{\mathsf{P}} &= \sum_{j=1}^m (n\Gamma_j^n - n\pi_j)^2 = \sum_{j=1}^m (n\Gamma_j^n)^2 - \frac{n^2}{m}\\
&\geq 2\sum_{j=1}^n \mathbb{I}\{n\Gamma_j^n \geq 2\} n\Gamma_j^n + \sum_{j=1}^m \mathbb{I}\{n\Gamma_j^n = 1\} - \frac{n^2}{m}\\
&= 2n + S_n^* - \frac{n^2}{m},
\end{aligned}$$

where the inequality follows from $(n\Gamma_j^n)^2 \geq 2(n\Gamma_j^n)$ when $n\Gamma_j^n > 1$. Thus,

$$\{S_n^* \geq \tau_n - 2n + \frac{n^2}{m}\} \subseteq \{S_n^{\mathsf{P}} \geq \tau_n\}. \tag{3.17}$$

Since $\tau_n - 2n + \frac{n^2}{m} = \mathsf{E}_\pi[S_n^*](1 + o(1)) + \alpha\frac{n^2}{m}$, it follows from Theorem 3.2 that $\lim_{n\to\infty} \sup_{\mu\in\Pi_m} \mathsf{P}_\mu\{S_n^* \leq \tau_n - 2n + \frac{n^2}{m}\} = 0$. Applying (3.17), we obtain $\lim_{n\to\infty} \sup_{\mu\in\Pi_m} \mathsf{P}_\mu\{S_n^{\mathsf{P}} \leq \tau_n\} = 0$, i.e., the worst-case probability of missed detection of Pearson's chi-square test is asymptotically zero. □

However, the probability of false alarm of Pearson's chi-square test decays much slower than the coincidence-based test, as we can show that its generalized error exponent of false alarm is zero:

Theorem 3.13. *Suppose Assumption 3.1 hold. Assume in addition that* $m = o(n^2/\log(n)^2)$. *If the sequence of thresholds are chosen so that*

$$\lim_{n\to\infty} P_M(\phi_n^{\mathsf{P}}) = 0, \tag{3.18}$$

then the generalized error exponent of false alarm is zero, i.e.,

$$J_F(\phi^{\mathsf{P}}) = 0. \tag{3.19}$$

We now compare Pearson's chi-square test and the coincidence-based test. Note that the Pearsons' chi-square test statistic can be written as

$$\begin{aligned} S_n^{\mathsf{P}} = &-\frac{n^2}{m} + \sum_{j=1}^{m} \mathbb{I}\{n\Gamma_j^n = 1\} + \sum_{j=1}^{m} 4\mathbb{I}\{n\Gamma_j^n = 2\} \\ &+ \sum_{l=3}^{\infty}\sum_{j=1}^{m} l^2\mathbb{I}\{n\Gamma_j^n = l\}. \end{aligned} \tag{3.20}$$

The main difference between these two tests is how the coefficients of $\mathbb{I}\{n\Gamma_j^n = l\}$ for $l \geq 3$ are chosen: Remove all the terms corresponding to $l \geq 3$ and consider the resulting test statistic:

$$S_n^{\mathsf{P}_0} = -\frac{n^2}{m} + \sum_{j=1}^{m} \mathbb{I}\{n\Gamma_j^n = 1\} + \sum_{j=1}^{m} 4\mathbb{I}\{n\Gamma_j^n = 2\}.$$

Then we have the following relationship between the three test statistics:

$$\Omega^{\mathsf{P}} := \{S_n^{\mathsf{P}} \le \check{\tau}_n\} \subset \Omega^* := \{S_n^* \le \tau_n\} \subset \Omega^{\mathsf{P0}} := \{S_n^{\mathsf{P}_0} \le \check{\tau}_n\} \tag{3.21}$$

where τ_n and $\check{\tau}_n$ are thresholds, and $\check{\tau}_n = \tau_n + 2n - \frac{n^2}{m}$. This is depicted geometrically in Fig. 3.4.

Note that the region in which Pearson's chi-square test decides in favor of $H1$ is larger than the coincidence-based test, and the probability that samples generating from π fall into this region decays slower than $\exp\{-\alpha n^2/m\}$ for any $\alpha > 0$. This is made precise in the proof of Theorem 3.13. On the other hand, it is not difficult to show that the test associated with ϕ^{P_0} has $J_M = 0$ by considering a sequence of μ whose likelihood ratio with respect to π is unbounded.

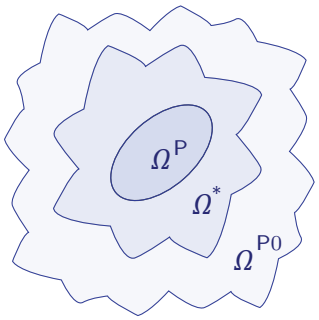

Figure 3.4: Decision regions.

In sum, we have

1. $J_F(\phi^{\mathsf{P}}) = 0, J_M(\phi^{\mathsf{P}}) > 0$;
2. $J_F(\phi^*) > 0, J_M(\phi^*) > 0$;
3. $J_F(\phi^{\mathsf{P}_0}) > 0, J_M(\phi^{\mathsf{P}_0}) = 0$.

Proof of Theorem 3.13. We consider a simpler problem where the set of alternative distributions is given by $\Pi_m \cap \mathcal{P}_m^b$, where $\mathcal{P}_m^b$ is defined in (3.9).

Lemma 3.12 implies $\min_{\mu\in\Pi_m}(\mathsf{E}_\mu[S_n^{\mathsf{P}}] - \mathsf{E}_\pi[S_n^{\mathsf{P}}]) = \frac{n^2}{m}\underline{\kappa}(\varepsilon)(1 + o(1))$.Thus the requirement $P_M(\phi_n^{\mathsf{P}}) \to 0$ imposes an upper-bound on τ_n:

Lemma 3.14. *In order for* (3.18) *to hold, for large enough* n*, we must have*

$$\tau_n \le \bar{\tau}_n := \mathsf{E}_\pi[S_n^{\mathsf{P}}] + \frac{n^2}{m}\underline{\kappa}(\varepsilon) + 2\frac{n}{\sqrt{m}}.$$

Consider the event that the first symbol appears many times:

$$A_n := \{n\Gamma_1^n = \lfloor \frac{n\sqrt{2\underline{\kappa}(\varepsilon)}}{\sqrt{m}} \rfloor\}.$$

In the event A_n, the first summand $f_1(n\Gamma_1^n)$ in the definition of Pearson's chi-square statistic given in (2.4) is $2\frac{n^2}{m}\underline{\kappa}(\varepsilon)$. This drives the value of S_n^{P} above the

threshold τ_n. Thus the probability of false alarm conditioned on this event converges to *one*, as summarized in Lemma 3.15. On the other hand, the probability of A_n does not decay exponentially fast with respect to n^2/m, as summarized in Lemma 3.16.

Lemma 3.15.

$$\mathsf{P}_\pi\{S_n^{\mathsf{P}} \geq \bar{\tau}_n | A_n\} = 1 - o(1).$$

Lemma 3.16.

$$-\lim_{n\to\infty} \frac{m}{n^2} \log(\mathsf{P}_\pi\{A_n\}) = 0.$$

Combining Lemma 3.14, Lemma 3.15 and Lemma 3.16 together, we conclude that

$$J_F(\phi^{\mathsf{P}}) \leq -\liminf_{n\to\infty} \frac{m}{n^2} \log\big(\mathsf{P}_\pi\{S_n^{\mathsf{P}} \geq \bar{\tau}_n | A_n\}\mathsf{P}_\pi\{A_n\}\big) = 0.$$

The proofs of these three lemmas are given in Appendix A.9. □

3.7 Alternative Distributions Based on f-Divergence

The set of alternative distributions studied in previous section is defined using the total variation distance. The generalized error exponent analysis with the same normalization also applies to some other distance or divergence functions, as we will show in Proposition 3.17 and Theorem 3.18. Examples include the KL divergence

$$d^{\mathsf{KL}}(\mu, \pi) = \sum_j \mu_j \log(\mu_j/\pi_j),$$

and Hellinger distance

$$d^{\mathsf{H}}(\mu, \pi) = \sum_j (\sqrt{\mu_j} - \sqrt{\pi_j})^2.$$

Rewrite the definition of set of alternative distributions using a general function d:

$$\Pi_m := \{\mu \in \mathcal{P}_m : d(\mu, \pi) \geq \epsilon\}. \tag{3.22}$$

We now present conditions under which the generalized error exponent analysis applies.

Proposition 3.17. *Suppose the function d satisfies*

1. $(\mu, \pi) \geq \alpha d_{TV}(\mu, \pi)$ *for some* $\alpha > 0$.

2. $\liminf_k \big(\inf\{\sum_j \frac{(\mu_j)^2}{\pi_j} : d(\mu, \pi) \geq \varepsilon, \mu, \pi \in \mathcal{P}_m\}\big) > 0$.

Then n^2/m *is the appropriate normalization for the large deviations analysis for small* $\varepsilon > 0$*: There exists a test* ϕ *such that*

$$J_F(\phi) > 0, J_M(\phi) > 0.$$

There is a constant $0 < \bar{J} < \infty$ *such that for any test* ϕ*, we have*

$$\min\{J_F(\phi), J_M(\phi)\} \leq \bar{J}.$$

We now consider the class of f-divergences, which are defined as

$$d_f(\mu, \pi) = \sum_j \pi_j f(\mu_j/\pi_j), \tag{3.23}$$

where f is a convex function with $f(1) = 0$. For the set of alternative distributions defined in (3.22) with $d = d_f$, we have the following condition:

Theorem 3.18. *Suppose* f *satisfies the following conditions:*

1. *For some* $0 < x < 1$,

$$\tfrac{1}{2}(f(1-x) + f(1+x)) > f(1).$$

2. *There is a constant* $\alpha > 0$ *such that for all* x,

$$f(x) \leq \alpha(x-1)^2.$$

Then n^2/m *is the appropriate normalization for the large deviations analysis for small* $\varepsilon > 0$*: There exists a test* ϕ *such that*

$$J_F(\phi) > 0, J_M(\phi) > 0.$$

There is a constant $0 < \bar{J} < \infty$ *such that for any test* ϕ*, we have*

$$\min\{J_F(\phi), J_M(\phi)\} \leq \bar{J}.$$

Proof of Proposition 3.17. The converse result which gives a lower-bound on the probability of missed detection is proved using a bound on the likelihood ratio between each in the set of bi-uniform distributions similar to those in (3.13) and the null distribution π. The first condition of Proposition 3.17 guarantees that these distributions are still in the set of alternative distributions. Therefore, the converse result holds.

For the achievability result, the critical step is to essentially show that the rate function is positive for any alternative distribution whose likelihood ratio with respect to π is bounded. The second condition of Proposition 3.17 guarantees that κ defined in (3.10) is positive, which in turn ensures a positive rate function. □

Proof of Theorem 3.18. The step is similar to the proof of Proposition 3.17. The first condition of Theorem 3.18 ensures that the collection of bi-uniform distributions used in the proof of the converse result is in the set of alternative distributions. We have for even m, for small enough ε, for μ given by

$$\mu_j = \begin{cases} \frac{1}{m} + \frac{\varepsilon'}{\lfloor m/2 \rfloor}, & j \leq \lfloor m/2 \rfloor, \\ \frac{1}{m} - \frac{\varepsilon'}{\lceil m/2 \rceil}, & j > \lfloor m/2 \rfloor, \end{cases}$$

the following holds:

$$d_f(\mu, \pi) = \tfrac{1}{2} f(1 + 2\varepsilon') + \tfrac{1}{2} f(1 - 2\varepsilon') \geq \varepsilon.$$

The second condition implies that

$$d_f(\mu, \pi) \leq \alpha \sum_j \frac{(\mu_j)^2}{\pi_j}.$$

Thus, the rate function is positive for any alternative distribution whose likelihood ratio with respect to π is bounded. □

3.8 Summary

We have shown that the classical error exponent criterion, which appears in the large deviations analysis for universal hypothesis testing problems with large number of samples, can be extended to the small sample case, provided the normalization is modified to account for both the sample size n and the alphabet size m.

The generalized error exponent offers new insights, which are not available from asymptotic consistency or CLT analysis, such as the optimality of the coincidence-based test and the sub-optimality of Pearson's chi-square test.

We offer a few discussions on directions for future research:

1. The analysis in this chapter is of asymptotic nature. The generalized error exponent gives the leading term in the asymptotic expansion of the probability of error. It is possible to obtain finer approximations with which the approximation error is smaller. This would be valuable when n/m is not very small, and the new term in the approximation can be used to improve the existing test. For example, recall the class of tests that extend the coincidence-based test described in Section 3.5.1. These tests have the same generalized error exponents. Finer approximations can reveal the difference between the test to help identify the best one.

2. The size of alphabet m is used in this and previous work to measure the "complexity" of the alternative hypothesis when the null distribution is uniform. It remains to see how this can be generalized to other cases, where the null distribution is far from uniform or has a countably infinite support and an exponential or polynomial tail. A possible generalization of the size of alphabet is the Rényi entropy of π, which is equal to $\log(m)$ when π is uniform.

3. What is the performance of the test with test statistic $\sum_{j=1}^{m}(n\Gamma_j^n - n\pi_j)^\rho$ for $\rho \in (1, 2)$? For what ρ will this test have a non-zero generalized error exponent? Note that when $\rho = 1$ and π is uniform, this becomes Pearson's chi-square test. For this purpose, it is desirable to establish general large-deviation characterizations of separable statistics for small sample problems, similar to those established for $n \asymp m$ in [12, 13].

4. We have focused on the simple goodness-of-fit problem in this chapter, in which π is fully specified. A natural extension is the composite goodness-of-fit problem in which π is not fully specified but assumed to be in a known set. A similar generalized error exponent concept should exist for the composite case.

CHAPTER 4

GENERALIZED ERROR EXPONENT FOR LARGE ALPHABET CLASSIFICATION

In this chapter, we apply the generalized error exponent analysis developed in the previous chapter to the binary classification problem where the alphabet size is large.

4.1 Background

Consider the following binary classification problem: Two training sequences $\boldsymbol{X} = \{X_1, \ldots, X_N\}$ and $\boldsymbol{Y} = \{Y_1, \ldots, Y_N\}$ generated from two different *unknown* sources are observed. The two sources share the same alphabet $[m] := \{1, \ldots, m\}$. Given a test sequence $\boldsymbol{Z} = \{Z_1, \ldots, Z_n\}$, the test decides whether $\boldsymbol{Z}$ comes from the first source or the second.

The performance of a test is usually assessed by how its probability of classification error depends on N, n, m. Since the exactly formula for the probability of error is usually complicated, asymptotic models and performance criteria are used. For example, the classical error exponent criterion characterizes the exponential rate at which the probability of error decays as N and n increase to infinity. In addition to assessing a particular test's performance, it is desirable to establish fundamental limits on the best achievable performance.

In many applications such as text classification, the number of training and test samples observed, N and n, are much smaller than the size of alphabet m. For example, suppose we want to decide, given two articles written by two different others, which author writes the third article. The number of words appearing in an article is much smaller than the English vocabulary, and the histogram of words is a sparse one [36].

The high-dimensional setting, in which N, n, m all tend to infinity and m is much larger than N, n, is a widely used approach to analyze tests for the small sample problem. A widely used performance criterion is asymptotic consistency:

Given some dependence of N, n on m, does the probability of error decay to zero as m increases to infinity? A fundamental result with respect to this criterion was established in [37]: Assuming that the distribution on all symbols in the alphabet is of order $1/m$, there exists an asymptotic consistent test if and only if $m = o(n^2)$. Note that the result is established only for the case $N = n$.

In most practical scenarios, the number of test samples available is smaller than the number of training samples. It is thus desirable to understand how N and n affect the performance individually. We thus pose the following questions:

1. How fast do N and n need to increase with m in order to have an asymptotic consistent test?

2. Does the probability of error depend on N and n in the same way?

3. If the number of training samples is limited, can the performance be improved by having more test samples?

The goal here is to answer these questions by establishing achievability and converse results on the best achievable probability of classification error. In the classification problem, the classical error exponent analysis has been applied to the case of fixed alphabet in [38] and [39]. It was shown that in order for the probability of error to decay exponentially fast with respect to n, the number of training samples N must grow at least linearly with n. However, in the small sample problem, the classical error exponent concept is not applicable, and we apply the generalized error exponent analysis developed in this thesis.

We identify the appropriate normalization for large deviations analysis, and obtain a generalized error exponent to approximate the probability of error for small number of observations. This analysis yields new insights on the best achievable performance:

1. The numbers of training and test samples N, n have different effects on the performance. This is made precise in Theorem 4.4 and Theorem 4.5.

2. The ℓ_2-norm based test investigated in [37], which compares the ℓ_2 distances between the empirical distribution of the test sequence to those of the two training sequences, is sub-optimal in that it has a zero generalized error exponent, while a weighted coincidence-based test proposed in this chapter has a non-zero generalized error exponent.

The result has been published in [40].

A closely related problem is the problem of testing whether two distributions are close. Achievability and converse results with respect to asymptotic consistency for this problem have been established in [41, 42]. The results in [43] have lead to algorithms for classification and closeness testing [44].

4.2 Problem Statement

Consider the following classification problem: Two training sequences $\boldsymbol{X}$ and $\boldsymbol{Y}$ are generated i.i.d. with marginal distributions π and μ, respectively. Each symbol takes value in $[m] := \{1, 2, \ldots, m\}$. A test sequence $\boldsymbol{Z}$ is observed. The sequence $\boldsymbol{Z}$ is i.i.d. with a marginal distribution π under the null hypothesis $H0$ and with a marginal distribution μ under the alternative hypothesis $H1$. The three sequences $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ are independent.

Denote the set of probability distributions over $[m]$ by $\mathcal{P}_m$. The pair of unknown distributions (π, μ) belongs to the following set $\Pi_m \subseteq \mathcal{P}_m \times \mathcal{P}_m$,

$$\Pi_m = \{(\pi, \mu) : \|\mu - \pi\|_1 \geq \varepsilon, \max_j \pi_j \leq \frac{c_1}{m}, \max_j \mu_j \leq \frac{c_1}{m}\},$$

where c_1 is a large positive constant. The definition of Π_m is essentially the same as the α-large-alphabet source defined in [37], except that we allow the number of training and test samples to be different. While this assumption that all words are rare does not hold for English vocabulary, the insights and tests obtained for rare words will be used to improve the algorithms for the case when there are both frequent and rare words.

In the high-dimensional model, we consider a sequence of classification problems as described above, indexed by m. Thus $\mathcal{P}_m, N, n, p, q, \Pi_m$ all depend on m. Moreover, N, n increase to infinity as m increases.

A test $\phi = \{\phi_m\}_{m\geq 1}$ is a sequence of binary-valued functions with $\phi_m : [m]^N \times [m]^N \times [m]^n \to \{0, 1\}$. It decides in favor of $H1$ if $\phi_m = 1$ and $H0$ otherwise. Use the notation $\mathsf{P}_{(\mu,\pi,\nu)}(A)$ to denote the probability of the event A when $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ have marginal distributions μ, π, ν respectively. The performance of a test ϕ is evaluated using the worst-case average probability of error given by

$$P_e(\phi_m) = \sup_{(\pi,\mu)\in\Pi_m} [\tfrac{1}{2}\mathsf{P}_{(\pi,\mu,\pi)}\{\phi_m = 1\} + \tfrac{1}{2}\mathsf{P}_{(\pi,\mu,\mu)}\{\phi_m = 0\}].$$

It is said to be asymptotically consistent if

$$\lim_{m\to\infty} P_e(\phi_m) = 0.$$

4.3 Region of Asymptotic Consistency

We begin with the asymptotic consistency result.

Theorem 4.1. *There exists an asymptotically consistent test if and only if*

$$m = o(\min\{N^2, Nn\}).$$

Proof. The small sample case where $\max\{N, n\} = o(m)$ is a corollary of the generalized error exponent analysis results given in Theorem 4.4 and Theorem 4.5.

Now consider the case when $m = O(N)$. The only if direction is trivial. For the if direction, when $m = o(N)$, the distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be essentially be estimated with vanishing error since the number of types grows subexponentially in n (see [37, Lemma 3]). When m is linear in N, this problem can be transformed into a small sample problem with alphabet size mb where $b = \lceil \sqrt{\min\{N, n\}} \rceil$: Associate each symbol in $[m]$ with b symbols. Each observation is then randomly mapped to one of the associated symbols. A consistent test for the small sample problem leads to a consistent test for the original problem. □

We depict the region at which asymptotic consistency is achievable in Fig. 4.1. We offer a few remarks:

1. For the case $N = n$, the conclusion of Theorem 4.1 is consistent with the results in [37, Theorem 3 and 4]. Our proof technique is different.

2. The requirements on N and n for asymptotic consistency are different: The first requirement $m = o(N^2)$ needs to be satisfied regardless of how many test samples are available. The second requirement is active only when $n = O(N)$. Therefore, as long as the number of test samples grows linearly with the training samples, further increasing the test samples will not improve the performance in terms of asymptotic consistency.

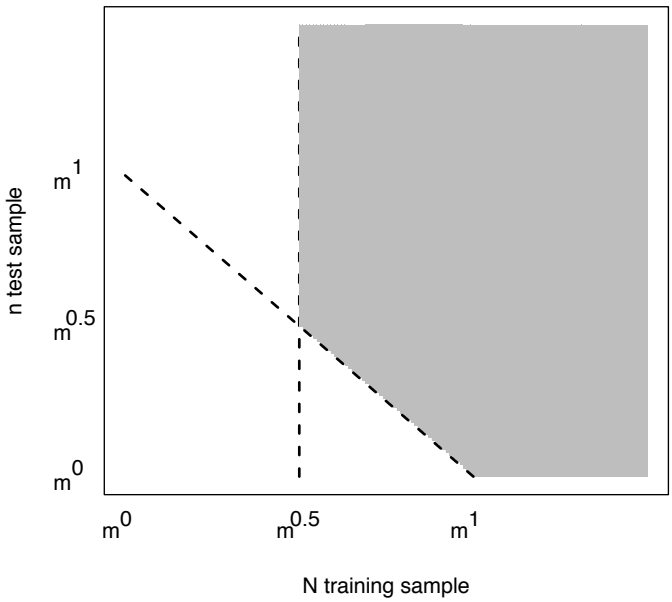

Figure 4.1: Region of asymptotic consistency.

3. On the other hand, increasing the number of *training* samples will always increase the performance. The effect of increasing the training samples is different when $n = o(N)$ and $N = o(n)$.

4.4 Generalized Error Exponent

When m is fixed, the following error exponent criterion has been used to evaluate a test ϕ:

$$I(\phi) := -\limsup_{n\to\infty} \frac{1}{n} \log(P_e(\phi_m)). \tag{4.1}$$

This classical error exponent criterion is no longer applicable in the small sample case where

Assumption 4.2. $N = o(m), n = o(m)$.

One should consider instead the following generalization, defined with respect

to the normalization $r(N, n, m)$:

$$J(\phi) := -\limsup_{n\to\infty} \frac{1}{r(N, n, m)} \log(P_e(\phi_m)). \tag{4.2}$$

The results in Theorem 4.4 and Theorem 4.5 imply that the appropriate normalization is

$$r(N, n, m) = \min\{N^2, Nn\}/m.$$

The generalized error exponent $J(\phi)$ could depend on how N, n increase with m. Note that to have a consistent test, the necessary condition in Theorem 4.1 must be satisfied, as summarized in the assumption below:

Assumption 4.3. $m = o(\min\{N^2, Nn\})$.

This is equivalent to $\lim_{m\to\infty} r(N, n, m) = \infty$.

The following theorems demonstrate that the definition in (4.2) is meaningful:

Theorem 4.4 (Achievability)**.** *Suppose Assumption 4.2 and Assumption 4.3 hold. Then there exists a test ϕ such that*

$$J(\phi) > 0.$$

Theorem 4.5 (Converse)**.** *Suppose Assumption 4.2 holds. There exists a constant $\bar{J}$ such that for any test ϕ,*

$$-\log(P_e(\phi_m)) \le r(N, n, m)\bar{J}.$$

These theorems imply that the best achievable probability of error decays approximately as $P_e = \exp\{-r(N, n, m)J\}$ for some $J > 0$. Note that the probability of error changes exponentially with respect to n only when $n = O(N)$. When $N = o(n)$, the probability of error is mainly determined by the number of training samples. This phenomenon is similar to the case with fixed m, for which results in [38] show that whether $n = O(N)$ holds determines whether the probability of error decreases exponentially in n.

4.5 The ℓ_2-Norm Based Test Is Suboptimal

Let $\Gamma^x, \Gamma^y, \Gamma^z$ be the empirical distributions of $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$. The ℓ_2-norm based test has the following test statistic:

$$F_n := \|\Gamma^z - \Gamma^x\|_2^2 - \|\Gamma^z - \Gamma^y\|_2^2.$$

The test is given by

$$\phi^F = \mathbb{I}\{F_n \geq 0\}.$$

This test was shown in [37] to be asymptotically consistent when $N = n$ and $m = o(N^2)$. We now show, however, this test has zero generalized error exponent:

Theorem 4.6. *Suppose Assumption 4.2 and Assumption 4.3 hold and $N = n$. Assume in addition that $m = o(n^2/\log(n)^2)$. Then*

$$J(\phi^F) = 0.$$

Proof. The sub-optimality of ϕ^F is due to the following reason: For any j, a large variation of the value of $N\Gamma_j^y$ causes a significant change in the value of the statistic F_n. Assume m is even for simplicity of exposition. Let u denote the uniform distribution on $[m]$. Let $q_j = (1+\varepsilon)/m$ for $j \leq m/2$ and $q_j = (1-\varepsilon)/m$ for $j > m/2$. Consider the case where $H1$ is true and the distribution is given by (q, u, u).

Considering the following event where one symbol appears many times:

$$C_n := \{N\Gamma_1^y = \lfloor 4n/\sqrt{m} \rfloor\}. \tag{4.3}$$

We claim that this event is likely to cause an error:

Lemma 4.7.

$$\mathsf{P}_{(q,u,u)}\{\phi^F = 0 | C_n\} = 1 - o(1).$$

On the other hand, the probability of C_n decays slowly:

Lemma 4.8.

$$\mathsf{P}_{(q,u,u)}(C_n) = \exp\{-4(n/\sqrt{m})\log(m)(1 + o(1))\}.$$

Combining these two equality gives the lower-bound

$$\begin{aligned}\log(P_e(\phi^F)) &\geq \log\big(\tfrac{1}{2}\mathsf{P}_{(q,u,u)}(C_n)\mathsf{P}_{(q,u,u)}\{\phi^F = 1|C_n\}\big)\\ &= 34\frac{n}{\sqrt{m}}\log(m)(1+o(1)).\end{aligned}$$

Thus this error decays at most as $nm^{-\frac{1}{2}}\log(m)$, slower than n^2/m. Consequently, $J(\phi^F) = 0$. □

4.6 Weighted Coincidence-Based Test

A nonzero generalized error exponent is achieved by a weighted coincidence-based test. Its construction is inspired by the coincidence-based test for the non-uniform case proposed in Section 3.5.2. Define the test statistic T_n:

$$\begin{aligned}T_n = \sum_j \big[&\frac{1}{N^2}\mathbb{I}\{N\Gamma_j^x = 2, n\Gamma_j^z = 0\} + \frac{1}{n^2}\mathbb{I}\{N\Gamma_j^x = 0, n\Gamma_j^z = 2\}\\ &- \frac{1}{nN}\mathbb{I}\{N\Gamma_j^x = 1, n\Gamma_j^z = 1\} + \frac{1}{nN}\mathbb{I}\{N\Gamma_j^y = 1, n\Gamma_j^z = 1\}\\ &- \frac{1}{n^2}\mathbb{I}\{N\Gamma_j^y = 0, n\Gamma_j^z = 2\} - \frac{1}{N^2}\mathbb{I}\{N\Gamma_j^y = 2, n\Gamma_j^z = 0\}\big].\end{aligned}$$

The test is given by $\phi^T = \mathbb{I}\{T_n \geq 0\}$.

Theorem 4.4 is proved by bounding $P_e(\phi^T)$ via Chernoff:

$$\begin{aligned}\log(\mathsf{P}_{(\pi,\mu,\pi)}\{\phi^T = 1\}) &\leq \inf_\theta \Lambda_{(\pi,\mu,\pi)}(\theta).\\ \log(\mathsf{P}_{(\pi,\mu,\mu)}\{\phi^T = 0\}) &\leq \inf_\theta \Lambda_{(\pi,\mu,\mu)}(\theta).\end{aligned}$$

Here, $\Lambda_{(\pi,\mu,\nu)}(\theta) = \log \mathsf{E}_{(\pi,\mu,\nu)}[\exp(\theta K_n)]$ is the logarithmic moment generating function of K_n. The main step is to obtain an asymptotic approximation to $\Lambda_{(\pi,\mu,\nu)}(\theta)$, given in the following proposition:

Proposition 4.9. *Let $\theta = \min\{N^2, nN\}\gamma$. For γ satisfying $|\gamma| \leq 1$,*

$$\begin{aligned}\Lambda_{(\pi,\mu,\nu)}(\theta) \leq &\min\{N^2, nN\}\big(\gamma[\sum_{j=1}^{m}(\tfrac{1}{2}(\pi_j-\nu_j)^2 - \tfrac{1}{2}(\mu_j-\nu_j)^2)] \\ &\qquad + \gamma^2[\sum_{j=1}^{m}(\pi_j\nu_j + \mu_j\nu_j) + \tfrac{1}{2}(\pi_j^2+\mu_j^2)]\big) \\ &+ O(\frac{\min\{N^2, nN\}\max\{N,n\}}{m^2}) + O(1).\end{aligned}$$

The proof of Proposition 4.9 is given in Appendix B.2.

Proof of Theorem 4.4. Applying Proposition 4.9 with the Chernoff bound for the cases $\nu = \pi$ and $\nu = \mu$, and using Assumption 4.2 and Assumption 4.3, and the facts $\pi_j, \mu_j \leq c_1/m$ and $\sum_{j=1}^{m}(\mu_j - \pi_j)^2 \geq \varepsilon^2/m$, we obtain

$$\begin{aligned}\log(\mathsf{P}_{\pi,\mu,\pi}\{\phi^T = 1\}) &\leq -\frac{\varepsilon^4}{160c_1^2}\frac{\min\{N^2, nN\}}{m}(1+o(1)), \\ \log(\mathsf{P}_{\pi,\mu,\mu}\{\phi^T = 0\}) &\leq -\frac{\varepsilon^4}{160c_1^2}\frac{\min\{N^2, nN\}}{m}(1+o(1)).\end{aligned}$$

The approximation $o(1)$ is uniform over all $(\pi, \mu) \in \Pi_m$. Therefore,

$$J \geq \frac{\varepsilon^4}{160c_1^2}.$$

□

4.7 Proof of the Converse

Proof of Theorem 4.5. ***Step 1***: Establish the upper bound,

$$-\log(P_e(\phi_m)) \leq \bar{J}_1 N^2/m. \tag{4.4}$$

The main idea of the proof is to consider an event under which the observations do not give any information regarding the hypotheses, and lower-bound the probability of such an event.

We now make this precise. Define the event

$$A = \{\text{No symbol in } \boldsymbol{X} \text{ appears more than once;} \\ \text{no symbol in } \boldsymbol{Y} \text{ appears more than once.}\}$$

Assume without loss of generality that m is even. Define a collection of bi-uniform distributions as follows: Let K_m denote the collection of all subsets of $[m]$ whose cardinality is $m/2$. For each set $\omega \in K_m$, define the distribution q^ω as

$$q_j^\omega = \begin{cases} (1+\varepsilon)/m, & j \in \omega; \\ (1-\varepsilon)/m, & j \in [m] \setminus \omega. \end{cases} \tag{4.5}$$

Note that $\|u - q^\omega\|_1 = \varepsilon$, and $(u, q^\omega) \in \Pi_m$ for all ω.

We will use the short-hand notation $\{(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})\} = \{(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})\}$ throughout the chapter.

Our choice of the collection of distributions makes sure that the following result holds:

Lemma 4.10. *For any sequence $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \subseteq A$,*

$$\frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{(u, q^\omega, u)}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{(q^\omega, u, u)}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}).$$

Proof. For any sequence, let φ_i denote the number of symbols appearing i times. The vector $[\varphi_1, \varphi_2, \varphi_3, \ldots]$ is called the *profile* of the sequence [43].

Because of the symmetry of the collection of distributions $\{q^\omega, \omega \in K_m\}$, the symmetry of the uniform distribution u, and the independence among $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$, the value of $\frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{(u, q^\omega, u)}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ only depends on the profiles of $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$. In the event A, the profiles of $\boldsymbol{x}$ and $\boldsymbol{y}$ are fixed, which then leads to the claim of the lemma. □

Lemma 4.10 implies that for any observation $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in A$, it is impossible to tell whether it is more likely to come from the mixture on the left-hand side or

the mixture on the right-hand side. Consequently,

$$\begin{aligned}
P_e(\phi_m) \geq & \frac{1}{4|K_m|}\sum_{\omega}[\mathsf{P}_{(u,q^\omega,u)}\{\phi_m=1\}+\mathsf{P}_{(u,q^\omega,q^\omega)}\{\phi_m=0\}] \\
& + \frac{1}{4|K_m|}\sum_{\omega}[\mathsf{P}_{(q^\omega,u,q^\omega)}\{\phi_m=1\}+\mathsf{P}_{(q^\omega,u,u)}\{\phi_m=0\}] \\
\geq & \frac{1}{4|K_m|}\sum_{\omega}[\Pr_{(u,q^\omega,u)}\{\phi_m=1\}+\Pr_{(q^\omega,u,u)}\{\phi_m=0\})] \\
\geq & \frac{1}{4|K_m|}\sum_{\omega}[\Pr_{(u,q^\omega,u)}(\{\phi_m=1\}\cap A)+\Pr_{(q^\omega,u,u)}(\{\phi_m=0\}\cap A)] \\
= & \frac{1}{4|K_m|}\sum_{\omega}[\Pr_{(u,q^\omega,u)}(\{\phi_m=1\}\cap A)+\Pr_{(u,q^\omega,u)}(\{\phi_m=0\}\cap A)] \\
= & \frac{1}{4|K_m|}\sum_{\omega}\Pr_{(u,q^\omega,u)}(A),
\end{aligned} \tag{4.6}$$

where the first inequality follows from the fact that the maximum is no smaller than the average, and the second last inequality follows from Lemma 4.10. The probability of the event A can be lower-bounded.

Lemma 4.11. *The following approximations holds uniformly for any ω:*

$$\log\big(\Pr_{(u,q^\omega,u)}(A)\big) = -(1+\tfrac{1}{2}\varepsilon^2)\frac{N^2}{m}(1+o(1))+O(1).$$

Proof. It follows from a combinatorial argument that the probability that no symbol appears twice in $\boldsymbol{X}$ when $\boldsymbol{X}$ has marginal distribution u is given by

$$m(m-1)\dots(m-N+1)(1/m)^N = \exp\{-\tfrac{1}{2}\frac{N^2}{m}(1+o(1))\}.$$

Estimating the probability that no symbol appears twice in $\boldsymbol{Y}$ can be done similarly but is more involved. □

The claim (4.4) follows from applying Lemma 4.11 to (4.6), and picking a large enough $\bar{J}$.

Step 2: Establish the second upper-bound

$$-\log(P_e(\phi_m) \leq \bar{J}_2(Nn+n^2)/m. \tag{4.7}$$

We consider the following event:

$$B = \{\text{No symbol in } \boldsymbol{Z} \text{ appears more than once;} \\ \text{no symbol in } \boldsymbol{Z} \text{ has appeared in either } \boldsymbol{X} \text{ or } \boldsymbol{Y}\}.$$

When this event happens, it is impossible (in the worst-case setting) to infer which distribution the test sequence is more likely to be generated from. This is captured by the following lemma:

Lemma 4.12. *Consider any $\boldsymbol{x}, \boldsymbol{y}$. For any two sequences $\boldsymbol{z}$ and $\bar{\boldsymbol{z}}$ such that $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \subseteq B$ and $(\boldsymbol{x}, \boldsymbol{y}, \bar{\boldsymbol{z}}) \subseteq B$, the following holds:*

$$\frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{(u, q^\omega, u)}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{u, q^\omega, u}(\boldsymbol{x}, \boldsymbol{y}, \bar{\boldsymbol{z}}).$$

$$\frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{(u, q^\omega, q^\omega)}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{u, q^\omega, q^\omega}(\boldsymbol{x}, \boldsymbol{y}, \bar{\boldsymbol{z}}).$$

Proof. Since no symbols in $\boldsymbol{z}$ have appeared in $\boldsymbol{x}$ and $\boldsymbol{y}$, due to the symmetry of the collection of distributions $\{q^\omega, \omega \in K_m\}$ and the symmetry of the uniform distribution u, for fixed $\boldsymbol{x}$ and $\boldsymbol{y}$, the value of $\frac{1}{|K_m|}\sum_{\omega \in K_m} \Pr_{(u, q^\omega, q^\omega)}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ only depends on the profile of $\boldsymbol{z}$. It follows from the definition of the event B that the profile of $\boldsymbol{z}$ is the same as the profile of $\bar{\boldsymbol{z}}$. □

The result of Lemma 4.12 can interpreted as follows: In the event B, observing $\boldsymbol{Z}$ does not gives any information since under either hypothesis, each sequence $\boldsymbol{z}$ appears with equal probability.

Consider any $\boldsymbol{x}, \boldsymbol{y}$. Let $D_{\boldsymbol{x},\boldsymbol{y}} = \{\boldsymbol{z} : (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in \{\phi_m = 1\} \cap B\}$ and $D^c_{\boldsymbol{x},\boldsymbol{y}} = \{\boldsymbol{z} : (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in \{\phi_m = 0\} \cap B\}$. Lemma 4.12 implies that the probability of $\{\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}, \phi_m = 1\} \cap B$ only depends on the size of $D_{\boldsymbol{x},\boldsymbol{y}}$, rather than

what sequences the set $D_{\boldsymbol{x},\boldsymbol{y}}$ includes. Consequently,

$$\begin{aligned}
&\frac{1}{|K_m|}\sum_{\omega}\Big[\Pr_{(u,q^\omega,u)}(\{\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y},\phi_m=1\}\cap B)\\
&\qquad\qquad+\Pr_{(u,q^\omega,q^\omega)}(\{\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y},\phi_m=0\}\cap B)\Big]\\
=&\Big[\frac{1}{|K_m|}\sum_{\omega}\Pr_{(u,q^\omega,u)}(\{\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y}\}\cap B)\Big]\frac{|D_{\boldsymbol{x},\boldsymbol{y}}|}{D_{\boldsymbol{x},\boldsymbol{y}}+D^c_{\boldsymbol{x},\boldsymbol{y}}}\\
&+\Big[\frac{1}{|K_m|}\sum_{\omega}\Pr_{(u,q^\omega,q^\omega)}(\{\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y}\}\cap B)\Big]\frac{|D^c_{\boldsymbol{x},\boldsymbol{y}}|}{D_{\boldsymbol{x},\boldsymbol{y}}+D^c_{\boldsymbol{x},\boldsymbol{y}}}\\
\geq&\frac{1}{|K_m|}\min\Big\{\sum_{\omega}\Pr_{(u,q^\omega,u)}(\{\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y}\}\cap B),\\
&\qquad\qquad\sum_{\omega}\Pr_{(u,q^\omega,q^\omega)}(\{\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y}\}\cap B)\Big\},
\end{aligned}\tag{4.8}$$

where the inequality follows from lower-bounding the probability of $\{\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y}\}\cap B$ under (u,q^ω,u) and (u,q^ω,q^ω) by the minimum of these two.

Lemma 4.13. *Let $\bar{J}_2 = 5$. Then the following bounds hold uniformly over all $\omega,\boldsymbol{x},\boldsymbol{y}$:*

$$\log\Big[\frac{\Pr_{(u,q^\omega,u)}(\{\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y}\}\cap B)}{\Pr_{(u,q^\omega,u)}\{\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y}\}}\Big]\geq\bar{J}_2\frac{Nn+n^2}{m}(1+o(1)).$$

$$\log\Big[\frac{\Pr_{(u,q^\omega,q^\omega)}(\{\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y}\}\cap B)}{\Pr_{(u,q^\omega,q^\omega)}\{\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y}\}}\Big]\geq\bar{J}_2\frac{Nn+n^2}{m}(1+o(1)).$$

The proof is similar to that of Lemma 4.11. We omit the details.

Note that the average probability of error is equal to the summation of the left-hand side of (4.8) over all possible $(\boldsymbol{x},\boldsymbol{y})$. Applying Lemma 4.13 to lower-bound the right-hand side of (4.8) leads to the claim.

We now combine (4.4) and (4.7). It is straightforward to verify that

$$\min\{N^2,Nn+n^2\}\leq\min\{N^2,2Nn\}.$$

Taking $\bar{J}=\max\{\bar{J}_1,2\bar{J}_2\}$ leads to the claim of the theorem. □

4.8 Summary

We have studied binary classification when the size of the underlying alphabet m is larger than the number of training samples N and test samples n. We show that there is an asymptotically consistent test if and only if $m = o(\min\{N^2, Nn\})$. Moreover, we characterize the rate of convergence using generalized error exponent: The best achievable probability of error is

$$P_e = \exp\{-J\frac{\min\{N^2, Nn\}}{m}(1 + o(1))\}.$$

The results shed light on the different roles played by the training samples and test samples. We propose a weighted coincidence-based test that achieves $J > 0$, and also show that the known ℓ_2-norm based test has zero generalized error exponent. The above results are established for the case where $N, n = o(m)$, and all symbols are rare, i.e., the probability of observing any symbols is on the same order m^{-1}.

One direction for future research is to relax this assumption that all symbols are rare. One possibility is to consider the rare and frequent symbols separately. This is the approach used in [41] for the problem of testing whether two distributions are close.

CHAPTER 5

APPROXIMATIONS TO THE HOEFFDING TEST

It is clear from the asymptotic analysis surveyed in this thesis that it could be beneficial to quantize the observations depending on the set of alternative distributions. In this chapter, we investigate a feature-based method, which is an alternative to the quantization approach. We then propose a feature extraction algorithm to learn the appropriate features from data.

5.1 Feature-Based Approximations

The performance of the Hoeffding test is significantly affected by the size of alphabet, as demonstrated in (2.11). One way to improve the Hoeffding test is the quantization approach which combines the cells (or bins) together to arrive at a smaller number of cells. This requires some prior knowledge regarding alternative hypothesis which reduces the original set of alternative distributions Π_m in Section 2.1 to a smaller set which we denote by $\mathcal{Q}$ throughout this chapter.

A more flexible method is based on features which are arbitrary functions of the observation. We now introduce the mismatched universal test, a feature-based technique to improve the test performance by incorporating prior information into the test. The mismatched universal test was introduce in [45, 46].

The mismatched test is based on a variational representation of the Kullback-Leibler divergence: The KL divergence can be expressed as the convex dual of the logarithmic moment generating function:

$$D(\nu^1\|\nu^2) = \sup_f \big(\nu^1(f) - \Lambda_{\nu^2}(f)\big), \tag{5.1}$$

where $\nu(f) := \mathsf{E}_\nu[f]$, and the optimization is taken over the space of all real-valued functions on Z where the logarithmic moment generating function is de-

noted

$$\Lambda_{\nu^2}(f) = \log(\nu^2(\exp(f))).$$

When $\nu^1 \preceq \nu^2$, the maximum achieved by log-likelihood ratio function $f = \log(\nu^1/\nu^2)$.

Consider a set of functions denoted by $\mathcal{F}$. The mismatched divergence is a lower bound on the KL divergence obtained by taking the supremum over the smaller set $\mathcal{F}$,

$$D_{\mathcal{F}}^{\text{MM}}(\nu^1\|\nu^2) := \sup_{f\in\mathcal{F}}\big\{\nu^1(f) - \Lambda_{\nu^2}(f)\big\}. \tag{5.2}$$

In this proposal, we focus on the case of linear function class: Let $\{\psi_k, 1 \le k \le \dim_{\mathcal{F}}\}$ denote a set of functions which we call basis functions. We assume that none of these functions is zero everywhere. We denote $\psi = [\psi_1, \ldots, \psi_{\dim_{\mathcal{F}}}]^\top$. The $\dim_{\mathcal{F}}$-dimensional linear function class generated by basis functions $\{\psi_m\}$ is then given by

$$\mathcal{F} = \{f_r := r^\top\psi : r \in \mathbb{R}^{\dim_{\mathcal{F}}}\}. \tag{5.3}$$

We use $\dim_{\mathcal{F}}$ to denote the dimension of the function class. The mismatched universal test is given by

$$\phi^{\text{MM}} = \mathbb{I}\{D_{\mathcal{F}}^{\text{MM}}(\Gamma^n\|\pi) \ge \tau\}.$$

When the function class is chosen as the collection of *all* indicator functions:

$$\psi_i(z) = \mathbb{I}\{z = i\}, 1 \le i \le m,$$

then the mismatched universal test is exactly the same as the Hoeffding test.

The performance of the mismatched test can be characterized using refined large deviations:

Proposition 5.1. *Let*

$$\beta^{\text{MM}}(\tau) = \inf_{\mu\in\mathcal{Q}}\{D^{\text{MM}}(\nu\|\mu) : D^{\text{MM}}(\nu\|\pi) \le \tau\}.$$

Then the probabilities of false alarm and missed detection have the following

approximations:

$$P_F(\phi_n^{\text{MM}}) = n^{(\dim_{\mathcal{F}}-2)/2} e^{-n\tau}(c'_F + O(\frac{1}{n})),$$
$$P_M(\phi_n^{\text{MM}}) \le n^{-1/2} e^{-n\beta^{\text{MM}}(\tau)}(c'_M + O(\frac{1}{n})).$$

This is a direct application of [16]. If we choose the threshold τ so that $\lim_{n\to\infty} P_M(\phi^{\text{MM}}) = c_M > 0$ to maximize the error exponent of P_F, and make a similar choice for the Hoeffding test, then we have the following direct comparison between the performance of these two tests:

The mismatched test

$$\begin{aligned} P_F(\phi^{\text{MM}}) &= n^{(\dim_{\mathcal{F}}-2)/2} \exp\{-n \inf_{\mu\in\mathcal{Q}} D^{\text{MM}}(\mu\|\pi)\}(c'_F + O(\frac{1}{n})), \\ \lim_{n\to\infty} P_M(\phi^{\text{MM}}) &= c_M. \end{aligned} \tag{5.4}$$

The Hoeffding test

$$\begin{aligned} P_F(\phi^{\mathsf{H}}) &= n^{(m-3)/2} \exp\{-n \inf_{\mu\in\mathcal{Q}} D(\mu\|\pi)\}(c_F + O(\frac{1}{n})), \\ \lim_{n\to\infty} P_M(\phi^{\mathsf{H}}) &= c_M. \end{aligned} \tag{5.5}$$

If $\mathcal{F}$ is chosen to be the set of all possible functions, then the mismatched test is the same as the Hoeffding test. The error exponent for the mismatched test is always no larger than that of the Hoeffding test. On the other hand, the order of the polynomial term for the mismatched test could be made smaller than that of the Hoeffding test. For an appropriate choice of $\mathcal{F}$, the mismatched test could have a smaller probability of false alarm.

Ideally, we would like $D(\mu\|\pi) \approx D^{\text{MM}}(\mu\|\pi)$ for all $\mu \in \mathcal{Q}$ where $\mathcal{Q}$ is the set of alternative distributions. When $\mathcal{Q}$ has some special form, it is known how $\mathcal{F}$ should be chosen. For example, the linear function class $\mathcal{F}$ defined in (5.3) satisfies this requirement when $\mathcal{Q}$ is an exponential family:

$$\mathcal{Q} = \{\check{\pi}^r : r \in \mathbb{R}^d\}, \tag{5.6}$$

where the *twisted distribution* $\check{\pi}^r \in \mathcal{P}(\mathsf{Z})$ is defined as

$$\check{\pi}^r := \pi \exp(f_r - \Lambda_\pi(f_r)). \tag{5.7}$$

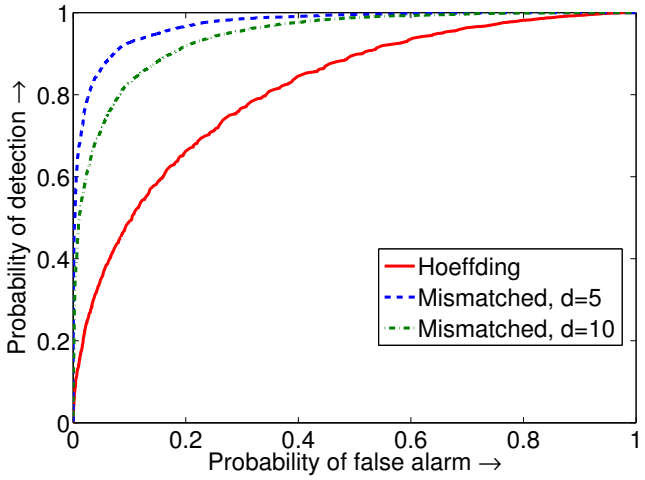

Figure 5.1: Performance of the Hoeffding test and mismatched test. Number of samples $n = 40$. Size of alphabet $m = 39$.

Figure 5.1 plots the performance of the Hoeffding test and mismatched test. The set of alternative distributions is an exponential family given in (5.6) with $d = 5$. We choose the corresponding linear function class. The performance of the mismatched test for one particular distribution in the family is given by the curve with $\dim_{\mathcal{F}} = 5$. We also randomly add another 5 basis functions and the performance is given by the curve $\dim_{\mathcal{F}} = 10$. We observe that the mismatched test with either $\dim_{\mathcal{F}} = 5$ or $\dim_{\mathcal{F}} = 10$ performs better than the Hoeffding test.

5.2 Feature Extraction via Nuclear-Norm Regularized Optimization

The feature extraction algorithm is based on optimizing worst-case mismatched divergence and the dimension of the function class: Given some η as a lower-bound on the worst case $D^{\text{MM}}_{\mathcal{F}}(\mu\|\pi)$, we solve

$$\min_{\mathcal{F}}\{\dim_{\mathcal{F}} : \operatorname*{ess\,inf}_{\mu\in\Theta} D^{\text{MM}}_{\mathcal{F}}(\mu\|\pi) \geq \eta\}. \tag{5.8}$$

where Θ is a probability measure on $\mathcal{Q}$ and the essential supremum of a function h is defined as

$$\operatorname*{ess\,sup}_{\mu\sim\Theta} h = \inf\{\eta : \Theta(\{\mu : h(\mu) \geq \eta\}) = 0\}.$$

An equivalent formulation was proposed in [47].

Let $\{\mu^1, \ldots, \mu^p\} \subset \mathcal{Q}$ be distributions drawn i.i.d. from $\mathcal{Q}$ according to Θ. This corresponds to a supervised learning setting for feature extraction, in which $\{\mu^1, \ldots, \mu^p\}$ are given by the training data. Let $\mathcal{F}^p$ be an optimal solution to the following approximation of (5.8):

$$\min_{\mathcal{F}}\{\dim_{\mathcal{F}} : \min_{\mu\in\{\mu^1,\ldots,\mu^p\}} D^{\text{MM}}_{\mathcal{F}}(\mu\|\pi) \geq \eta\}. \tag{5.9}$$

The features obtained from finite number of alternative distributions are generalizable:

Proposition 5.2. *The following holds with probability one for any small $\epsilon > 0$:*

$$\lim_{p\to\infty} \operatorname*{ess\,inf}_{\mu\sim\Theta} D^{\text{MM}}_{\mathcal{F}^p}(\mu\|\pi) \geq (1-\epsilon)\eta.$$

This proposition follows directly from a result in the Vapnik-Chervonenkis theory [48].

The problem (5.9) can be rewritten as an optimization problem involving the rank function:

$$\begin{aligned} \min \quad & \mathsf{rank}(X) \\ \text{subject to} \quad & \min_{i=1}^{p}\big(\langle \mu^i, X_i\rangle - \log(\langle \pi, e^{X_i}\rangle\big) \geq \eta \end{aligned} \tag{5.10}$$

where the optimization variable X is a $p \times m$ matrix, and X_i is the ith row of X, interpreted as a function on $[m]$. Given an optimizer X^*, we choose $\{\psi_i\}$ to be the set of right singular vectors of X^* corresponding to nonzero singular values. $\psi_i(k) = X^*_{ik}$, $1 \leq k \leq m$, $1 \leq i \leq d$.

The optimization problem (5.10) is not convex. We now apply a heuristic to approximate the solution to (5.10). It uses the nuclear-norm as a surrogate for the rank function. This heuristic was proposed in [49] and [50]. It was later extended to signal processing applications, and is the basis for convex optimization technique for matrix recovery and completion problems (see [51, 52, 53] for theoretical results characterizing the optimality of the solutions, and [54, 55] for

efficient algorithms to solve nuclear norm minimization problems).

The nuclear norm of a matrix is equal to the ℓ_1 norm of its singular values,

$$\|X\|_* = \sum_{i=1}^{\mathsf{rank}(X)} \sigma_i(X).$$

It has been shown in [49] that the nuclear norm is the *convex envelope* of the rank function over the set of matrices whose operator norms are no larger than one; i.e., it is the largest convex function that lower-bounds the rank function.

The nuclear-norm regularized optimization for approximating (5.10) is given as follows:

$$\max \quad \min_{i=1}^{p} \big(\langle \mu^i, X_i \rangle - \log(\langle \pi, e^{X_i} \rangle \big) - t\|X\|_*. \tag{5.11}$$

This optimization problem is convex and can be solved efficiently using *proximal algorithms* [56].

Quadratic approximations

To compute the mismatched test statistic, we need to solve the optimization problem in (5.2). When the function $\mathcal{F}$ is a linear function class, this is a convex optimization problem that can be solved using iterative algorithms such as gradient-descent and Newton method. The computation cost is not too significant when $\dim_{\mathcal{F}}$ is small. A reduction in computation can be obtained by using the quadratic approximation to the mismatched divergence [57]: The logarithmic moment generating function has the following quadratic approximation:

$$\Lambda_\pi(f) \approx \pi(f) + \tfrac{1}{2}\mathrm{var}_\pi(\mathrm{f})$$

where $\mathrm{var}_\pi(\mathrm{f})$ is the variance of f under distribution π. Denote the quadratic approximation to mismatched divergence as:

$$D_{\mathcal{F}}^{\text{PMM}}(\nu\|\pi) = \sup_{f\in\mathcal{F}} \{\nu(f) - \pi(f) - \tfrac{1}{2}\mathrm{var}_\pi(\mathrm{f})\}. \tag{5.12}$$

When $\mathcal{F}$ is a linear function class, the optimization problem in (5.12) becomes a quadratic optimization problem with an explicit solution. Denote the covariance

matrix $\Sigma_{\mathcal{F}}$ by

$$\Sigma_{\mathcal{F}} = \pi(\psi\psi^\top) - \pi(\psi)\pi(\psi^\top).$$

Assumption 5.3. *The covariance matrix $\Sigma_{\mathcal{F}}$ is full rank and thus positive definite.*

We then have

$$D^{\text{PMM}}_{\mathcal{F}}(\nu\|\pi) = \tfrac{1}{2}(\nu(\psi) - \pi(\psi))^\top \Sigma^{-1}_{\mathcal{F}}(\nu(\psi) - \pi(\psi)). \tag{5.13}$$

The mismatched test based on the quadratic approximation is given by

$$\phi^{\text{PMM}}_n = \mathbb{I}\{D^{\text{PMM}}_{\mathcal{F}}(\Gamma^n\|\pi) \geq \tau\}. \tag{5.14}$$

We now derive the feature extraction algorithm for this test. Define the matrix $\bar{A} \in \mathbb{R}^{m,p}$ as follows: The ith row of $\bar{A}$ is given by

$$\bar{A}_{i,j} = \mu^i_j - \pi^i_j.$$

Define the covariance matrix:

$$\Sigma_{i,j} = \mathbb{I}\{i = j\}\pi_i - \pi_i\pi_j.$$

While $\Sigma_{i,j}$ is not invertible, we can take Σ^{-1} to be its pseudo-inverse in the subspace $\{x : x^\top \mathbf{1} = 0\}$.

Consider the following optimization problem with $\bar{X} \in \mathbb{R}^{p,m}$:

$$\begin{array}{ll} \mathsf{min} & \mathsf{rank}(\bar{X}) \\ \mathsf{s.t.} & \bar{A}_i\bar{X}_i^\top - \frac{1}{2}\bar{X}_i\Sigma\bar{X}_i^\top \geq \eta \quad 1 \leq i \leq p. \end{array} \tag{5.15}$$

Let $\bar{X}^*$ denote an optimal solution, and write $r = \mathsf{rank}(\bar{X}^*)$. Let $\bar{X}^* = \bar{U}^*\bar{S}^*(\bar{V}^*)^\top$ be the singular value decomposition of $\bar{X}^*$. Then the selected r features are given by the vector-valued function $(\bar{V}^*)^\top$, and the corresponding function class is given by

$$\mathcal{F}^p = \{f : f(j) = \sum_{k=1}^{r} \theta_k \bar{V}^*_{j,k}, \theta \in \mathbb{R}^r\}. \tag{5.16}$$

The following nuclear norm minimization problem is used to approximate the

solution to (5.15):

$$\begin{array}{ll} \text{min} & \|\bar{X}\Sigma^{\frac{1}{2}}\|_* \\ \text{s.t.} & \bar{A}_i\bar{X}_i^\top - \frac{1}{2}\bar{X}_i\Sigma\bar{X}_i^\top \geq \eta \quad 1 \leq i \leq p. \end{array} \tag{5.17}$$

5.3 Numerical Experiment

We test the performance of the proposed feature extraction algorithm on a real-world data set in the UCI machine learning database [58]. The data set is the traffic count in San Francisco. There are 963 sensors that measure the (real-valued) occupancy rate of different car lanes of the San Francisco bay area freeway over time. Each sensor takes a measurement every 10 minutes. The whole set of data is collected over 15 months. We divide the measurements according to the day and the hour it was taken in. Our task is to infer whether the data comes from a particular time of a day.

We investigate the performance of the quadratic approximation to the mismatched test with the set of linear functions of measurements as the function class. We then apply the feature extraction algorithm to this function class. The output is a function class that is a low-dimensional subspace of the original function class. We then test the performance of the test with the extracted function class.

The input to the feature extraction algorithm is as follows: We took 20 different time slots in a week. For each time slot, we randomly choose 6 weeks. This means that we have $p = 20$ different alternative distributions. In our experiment, we work with measurements from 100 randomly chosen sensors. We expect the performance gain from feature extraction algorithms to be more significant if we instead using measurements from all 963 sensors as a baseline, since most of the sensor measurements will be redundant.

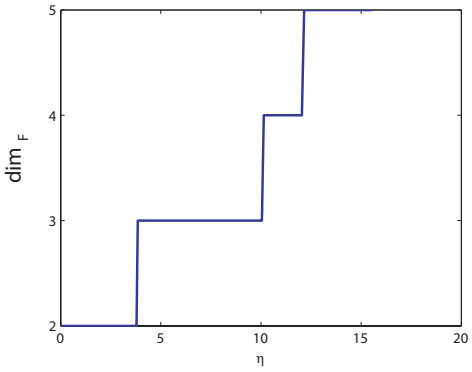

Figure 5.2: Dimension of the function class chosen by the feature extraction algorithm vs. the parameter η in the algorithm.

Figure 5.2 plots the dimension of the function class $\dim_{\mathcal{F}^p}$ chosen by the feature extraction algorithm. It is expected (but not rigorously proved) that $\dim_{\mathcal{F}^p}$ increases with η. We observe that a function class of dimension 5, which is much smaller than 100, gives a good approximation to the worst-case mismatched divergence.

Figure 5.3 plots the performance of the test with the initial features and the features extracted by the algorithm. In the first case shown in the upper figure, we are testing against the set of alternative distributions that is the same as the set used in training. In the second case shown in the lower figure, we are testing against a new set of alternative distributions to study whether the features learned are *generalizable* for the data set. We observe a significant performance gain in the first case. In the second case, we observe performance gains in the regime when the probability of false-alarm is large.

We remark that the performance of the algorithm could depend on the $p = 20$ distributions sampled. We repeat this experiment for training sequences and observe a similar performance gain using feature extraction. Moreover, as p increases, we expect that the dependence on the particular training sequence will become less significant.

5.4 Summary

In this chapter, we have shown the mismatched test could achieve better finite sample performance than the Hoeffding test. The choice of the function class is critical for the performance of the mismatched test. Prior knowledge regarding the set of alternative distributions can be exploited to optimize the function class. We have proposed a feature extraction algorithm based on the nuclear norm optimization.

The feature extraction algorithm is presented in the case of finite observation alphabet. It can be extended to the case where the observation is real-valued, and the goal is to extract a low-dimensional feature space from a initial large number of features. The finite observation alphabet is a special case where the features are indicator functions.

The objective function used in the feature extraction algorithm is based on an approximation to the classical error exponent. When the number of observations is significantly small, the generalized error exponent from the small sample. large

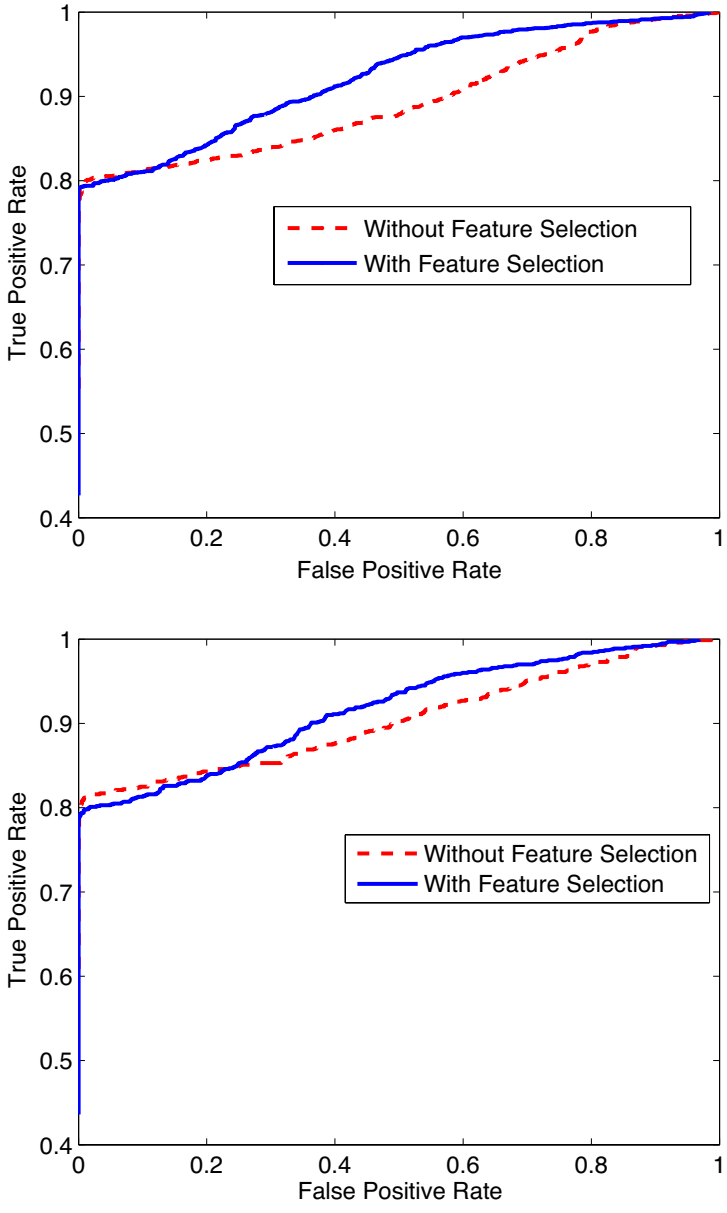

Figure 5.3: Average probability of false alarm and missed detection with / without feature extraction. Above: All the time slots in the test sequences have appeared in the training sequences. Below: Some test sequences have a time slot that has not appeared in the training sequence.

deviations analysis proposed in the thesis might be useful. To apply the generalized error exponent, the main problem to be addressed is to extend the generalized error exponent analysis from the finite alphabet case to the case of finite features.

CHAPTER 6

CONCLUSION

The main contribution of the research surveyed in this thesis is the generalized error exponent analysis framework. In the universal hypothesis testing problem, we show that the error exponent criterion for the large sample problem can be extended to the small sample problem. This requires a new normalization in the asymptotic analysis. The appropriate normalization is a function of both the number of samples and the size of the alphabet, and it holds under general assumptions. In particular, under general conditions on the set of alternative distributions, the coincidence-based test and extensions introduced here have optimal generalized error exponents, while Pearson's chi-square test is suboptimal.

In the binary classification problem, the generalized error exponent analysis is applied to show that the number of training samples and the number of test samples affect the test's performance in different ways. As a corollary, we characterize how the number of training samples and the number of test samples need to increase with the alphabet size for a test to have a vanishing probability of error. The result suggests that given a fixed budget of total training and test samples, it is optimal to choose the number of training samples and number of test samples to be approximately equal.

This thesis leaves many questions for future research:

a) The complexity in the universal hypothesis testing and the binary classification problems studied in this thesis is shown to be captured by the observation alphabet size. For other problems where the set of distributions in a hypothesis takes a different form, the complexity might be captured by other quantities. As a first step in this direction, we are considering extending the generalized error exponent analysis to the situation where the observation alphabet is countably infinite but the probability distributions in the hypotheses are limited to a certain form. For example, for probability distributions over the positive integers, a constraint could be imposed on the "tail" of the probability distribution.

b) The generalized error exponent in its current form is applicable for finite-valued observations. It is valuable to study tests using more flexible features that are not necessarily obtained by quantizing the observations. One interesting direction is to extend the generalized error exponent so that it characterizes how the probability of error depends on the number of samples as well as the number of features. A variation of the generalized error exponent might be used as the objective function in feature extraction algorithms, a role similar to that of the classical error exponent in the method developed in Chapter 5.

c) Topological structure often contains critical information that is easily ignored in statistical methods. Topology is important in other approaches such as the support vector machine. It is likely that current information-theoretic tools can help to create a coherent bridge between topological and statistical approaches to hypothesis testing, such as concepts from lossy source-coding. The mismatched test may be regarded as one step in this direction.

APPENDIX A

PROOF OF RESULTS IN CHAPTER 3

A.1 Proof of Theorem 3.2 and Theorem 3.4

The proof of Theorem 3.2 is based on the Chernoff bound and the Gärtner-Ellis Theorem. To simplify the representation, we work with the follows statistic:

$$\tilde{S}_n^* = \sum_{j=1}^{m} \mathbb{I}\{n\Gamma_j^n = 1\} - n.$$

Its logarithmic moment generating is given by

$$\Lambda_{\nu,\tilde{S}_n^*}(\theta) := \log\big(\mathsf{E}_\nu[\exp\{\theta\tilde{S}_n^*\}]\big). \tag{A.1}$$

Bounds and approximations for $\Lambda_{\nu,\tilde{S}_n^*}$ are first obtained for the restricted set of distributions $\mathcal{P}_m^b$ defined in (3.9).

Proposition A.1. *For any $\nu \in \mathcal{P}_m^b$, the n-sample logarithmic moment generating function for the statistic $\tilde{S}_n^*$ has the following asymptotic expansion:*

$$\Lambda_{\nu,\tilde{S}_n^*}(\theta) = \tfrac{1}{2}\frac{n^2}{m}\big(m\sum_{j=1}^{m}\nu_j^2\big)(e^{-2\theta}-1) + O(\frac{n^3}{m^2}) + O(1). \tag{A.2}$$

The approximation errors $O(\frac{n^3}{m^2})$ and $O(1)$ are uniform over the set $\mathcal{P}_m^b$. The proof of this proposition is given in Appendix A.2.

Applying this proposition together with the Chernoff bound leads to the value of J_F since $\pi \in \mathcal{P}_m^b$. Finding the value of J_M is more involved, because the alternative hypothesis (3.2) is composite. The key step is to identify the sequence of worst-case distributions in Π_m that approximately have the largest probability of missed detection.

So far we have restricted the discussion to alternative distributions in the set

$\mathcal{P}_m^b$. We also need to consider distributions in $\Pi_m \setminus \mathcal{P}_m^b$. For any $\mu \in \Pi_m \setminus \mathcal{P}_m^b$, the set of indices $\mathcal{S}_0 := \{j \in [m] : \mu_j \geq c_1 m^{-1}\}$ is non-empty. Now fix a small constant $\eta > 0$, and consider each index j in $\mathcal{S}_0$ in two separate cases, according to whether $n\mu_j \geq \eta$. Denote

$$\mathcal{T}_\eta(\mu) = \{j : n\mu_j \geq \eta\}, \quad \beta(\mu) = \sum_{j \in \mathcal{T}_\eta(\mu)} \mu_j.$$

Proposition A.2 below addresses the case where $\beta(\mu)$ is large. It implies that the probability of missed detection associated with such a distribution is much smaller than that associated with the dominating distributions: The probability decays exponentially fast with respect to n, which is larger than n^2/m. Proposition A.3 considers the alternate case, and shows that if $\beta(\mu)$ is not large, then a bound similar to that in Proposition A.1 holds.

Proposition A.2. *For all sufficiently small $\eta > 0$, any $\theta \in (0, 0.5]$, and any $\underline{\beta} > 0$, there exists n_0 such that for any $n > n_0$, and any ν satisfying $\beta(\nu) \geq \underline{\beta}$, the following holds:*

$$\Lambda_{\nu, \tilde{S}_n^*}(\theta) \leq -\beta(\nu)\alpha(\theta)n,$$

where $\alpha(\theta) > 0$.

Proposition A.3. *For any $\delta > 0$, $\theta \in (0, 0.5]$, $\overline{\eta} > 0$, there exist $\eta \in (0, \overline{\eta})$, $\overline{\beta} > 0$, and n_0 such that for any $n > n_0$, and any ν satisfying $\beta(\nu) \leq \overline{\beta}$, the following holds:*

$$\Lambda_{\nu, \tilde{S}_n^*}(\theta) \leq \tfrac{1}{2}\frac{n^2}{m}(m \sum_{j \notin \mathcal{T}_\eta(\nu)} \mu_j^2)(e^{-2\theta} - 1)(1 - \delta).$$

The proofs of these two propositions are given in Appendix A.3.

The threshold used in the test involves the term $\mathsf{E}_\pi[S_n^*]$. We also need its asymptotic expansion in the proof of Theorem 3.2:

Lemma A.4. *For any $\nu \in \mathcal{P}_m^b$, the expectation S_n^* have the following asymptotic expansions:*

$$\mathsf{E}_\nu[\tilde{S}_n^*] = -\frac{n^2}{m}\big(m \sum_{j=1}^m \nu_j^2\big) + O(\frac{n^3}{m^2}).$$

The proof of Lemma A.4 is given in Appendix A.10.

Proof of Theorem 3.2. We first prove the lower-bound on the generalized error exponents. Substituting the asymptotic expansion given in Proposition A.1 with

$\nu = \pi$ into the Chernoff bound, we obtain for $\theta > 0$,

$$\begin{aligned}\log \mathsf{P}_\pi(\phi_n^* = 1) =& \log \mathsf{P}_\pi\{\tilde{S}_n^* - \mathsf{E}_\pi[\tilde{S}_n^*] \geq -\tau_n\} \\ \leq& \theta(\mathsf{E}_\pi[\tilde{S}_n^*] - \tau_n) + \Lambda_{\pi, S_n^*}(-\theta) \\ =& -\theta\tau_n + \frac{n^2}{m}\tfrac{1}{2}[e^{2\theta} - (1+2\theta)] + O(\frac{n^3}{m^2}) + O(1).\end{aligned}$$

Normalizing it by $\frac{m}{n^2}$ and taking the limit leads to $J_F(\phi^*) \geq J_F^*(\tau)$.

To obtain the lower-bound on the generalized error exponent of missed detection, we apply Proposition A.2 and Proposition A.3 . We only need to prove it for the case $\tau \in [0, \underline{\kappa}(\varepsilon))$. The case $\tau = \underline{\kappa}(\varepsilon)$ will then follow from a continuity argument.

Take θ_0 to be the maximizer in the optimization problem defining J_M^* (see (3.8)). It is not difficult to see that $\theta_0 > 0$. It follows from Lemma 3.5 that

$$m \sum_{j \notin \mathcal{T}_\eta} \mu_j^2 \geq (1 + \underline{\kappa}(\frac{\varepsilon - \beta(\mu)}{1 - \beta(\mu)}))(1 - \beta(\mu))(1 + o(1)).$$

Thus, for any $\delta > 0$, we can choose η, β_0 small enough so that for any $\mu \in \Pi_m$ satisfying $\beta(\mu) \leq \beta_0$, we have $m \sum_{j \notin \mathcal{T}_\eta} \mu_j^2 \geq (1 + \underline{\kappa}(\varepsilon))(1 - \delta)$. It then follows from Proposition A.3 that for large enough n,

$$\Lambda_{\mu, \tilde{S}_n^*}(\theta_0) \leq \tfrac{1}{2}\frac{n^2}{m}(1 + \underline{\kappa}(\varepsilon))(e^{-2\theta_0} - 1)(1 - \delta)^2 + O(1). \tag{A.3}$$

For μ satisfying $\beta(\mu) \geq \beta_0$, it follows from Proposition A.2 that for large enough n,

$$\Lambda_{\mu, \tilde{S}_n^*}(\theta_0) \leq -\beta_0 \alpha(\theta_0) n. \tag{A.4}$$

We can pick n large enough so that the right-hand side of (A.4) is smaller than the right-hand side of (A.3). Applying the Chernoff bound leads to

$$\begin{aligned}&\log(\sup_{\mu \in \Pi_m} \mathsf{P}_\mu(\phi_n^* = 0)) \\ \leq& -\theta_0(\mathsf{E}_\pi[\tilde{S}_n^*] - \tau_n) + \sup_{\mu \in \Pi_m} \Lambda_{\mu, \tilde{S}_n^*}(\theta_0) \\ \leq& \theta_0(\tau_n - \mathsf{E}_\pi[\tilde{S}_n^*]) + \tfrac{1}{2}\frac{n^2}{m}(1 + \underline{\kappa}(\varepsilon))(e^{-2\theta_0} - 1)(1 - \delta)^2 + O(1).\end{aligned}$$

Consequently,

$$J_M(\phi^*) \geq \theta_0(-1-\tau) - \tfrac{1}{2}(e^{-2\theta_0}-1)(1+\underline{\kappa}(\varepsilon))(1-\delta)^2.$$

This holds for any $\delta > 0$. Thus the claimed lower-bound on the generalized error exponent of missed detection holds.

We now prove the upper-bound on the generalized error exponents. For the upper-bound on $J_M(\phi^*)$, consider the sequence of distributions given in (3.13) and (3.14). Define the limit of the logarithmic moment generating function:

$$\Lambda_1(\theta) := \lim_{n\to\infty} \frac{m}{n^2}\Lambda_{\mu^*,\tilde{S}_n^*}(\theta).$$

By Proposition A.1, the limit exists and is given by the following real-valued continuously differentiable function:

$$\Lambda_1(\theta) = \tfrac{1}{2}(e^{-2\theta}-1)(1+\underline{\kappa}(\varepsilon)).$$

Denote its Fenchel-Legendre transformation $\Lambda_1^*(t) := \sup_\theta[\theta t - \Lambda_1(\theta)]$. Note that $\lim_{n\to\infty} \frac{m}{n^2}(\mathsf{E}_\pi[\tilde{S}_n^*] + \tau_n) = -\tau - 1$. It follows from the Gärtner-Ellis Theorem [59, Theorem 2.3.6] that

$$-\liminf_{n\to\infty} \frac{m}{n^2}\log(\mathsf{P}_{\mu^*}(\phi_n^*=0)) = \inf_{t\geq -\tau-1} \Lambda_1^*(t) = \Lambda_1^*(-\tau-1) = J_M^*(\tau).$$

This leads to the upper-bound on $J_M(\phi^*)$ since $\mu^* \in \Pi_m$.

For the upper-bound on $J_F(\phi^*)$, consider

$$\Lambda_0(\theta) := \lim_{n\to\infty} \frac{m}{n^2}\Lambda_{\pi,\tilde{S}_n^*}(\theta) = \tfrac{1}{2}(e^{-2\theta}-1).$$

Let $\Lambda_0^*(t) = \sup_\theta[\theta t - \Lambda_0(\theta)]$. It follows from the Gärtner-Ellis Theorem that

$$-\liminf_{n\to\infty} \frac{m}{n^2}\log(\mathsf{P}_{\pi}(\phi_n^*=1)) = \inf_{t\leq -\tau-1} \Lambda_0^*(t) = \Lambda_0^*(-\tau-1) = J_F^*(\tau).$$

□

Proof of Theorem 3.4. The proof that the rate function result holds for S_n^* is contained in the proof of Theorem 3.2 except we apply the Gärtner-Ellis Theorem to $\boldsymbol{\mu}$ instead of μ^*. □

A.2 Proof of Proposition A.1

The proof uses the Poissonization technique, and the procedure is applicable for many separable statistics including S_n^*: Let $\{X_j\}$ be a sequence of independent Poisson random variables with parameter $\lambda\nu_j$ for some $\lambda > 0$. Then for any integers $v_1, \ldots, v_m$ satisfying $\sum_{j=1}^m v_j = n$, we have

$$\mathsf{P}\{n\Gamma_j^n = v_j, \text{for all } j\} = \mathsf{P}\{X_j = v_j, \text{for all } j | \sum_{j=1}^m X_j = n\}.$$

Therefore, the moment generating function of a separable statistic $\sum_{j=1}^m f_j(n\Gamma_j^n)$ admits the following representation when it is finite:

$$\mathsf{E}_\nu[\exp\{\theta \sum_{j=1}^m f_j(n\Gamma_j^n)\}] = \mathsf{E}[\exp\{\theta \sum_{j=1}^m f_j(X_j)\} | \sum_{j=1}^m X_j = n].$$

It is related to the moment generating function $A_\lambda(\theta)$ for $\sum_j f_j(X_j)$ as follows:

$$\begin{aligned} A_\lambda(\theta) := \mathsf{E}[\exp\{\theta \sum_{j=1}^m f_j(X_j)\}] &= \sum_{n=0}^\infty \frac{\lambda^n}{n!} e^{-\lambda} \mathsf{E}[\exp\{\theta \sum_{j=1}^m f_j(X_j)\} | \sum_{j=1}^m X_j = n] \\ &= \sum_{n=0}^\infty \frac{\lambda^n}{n!} e^{-\lambda} \mathsf{E}_\nu[\exp\{\theta \sum_{j=1}^m f_j(n\Gamma_j^n)\}]. \end{aligned}$$

Since the variables $\{X_j\}$ are independent, it is easy to obtain the formula for its moment generating function:

$$A_\lambda(\theta) = \prod_{j=1}^m (\sum_{k=0}^\infty \frac{(\lambda\nu_j)^k}{k!} e^{-\lambda\nu_j} e^{\theta f_j(k)}).$$

Since $A_\lambda(\theta)$ is analytic in λ, the moment generating function of $\sum_{j=1}^m f_j(n\Gamma_j^n)$ can be obtained via Cauchy's theorem:

$$\mathsf{E}_\nu[\exp\{\theta \sum_{j=1}^m f_j(n\Gamma_j^n)\}] = \frac{n!}{2\pi i} \oint e^\lambda A_\lambda(\theta) \frac{d\lambda}{\lambda^{n+1}}, \tag{A.5}$$

where the integration is carried out along any closed contour around $\lambda = 0$. These arguments lead to the following lemma:

Lemma A.5. *The n-sample moment generating function of the separable statistic*

$\sum_{j=1}^{m} f_j(n\Gamma_j^n)$ *is given by*

$$\mathsf{E}_\nu[\exp\{\theta \sum_{j=1}^{m} f_j(n\Gamma_j^n)\}] = \frac{n!}{2\pi i} \oint e^\lambda \prod_{j=1}^{m} \big(\sum_{k=0}^{\infty} \frac{(\lambda\nu_j)^k}{k!} e^{-\lambda\nu_j} e^{\theta f_j(k)} \big) \frac{d\lambda}{\lambda^{n+1}}.$$

Proof of Proposition A.1. Applying Lemma A.5 with $f_j(1) = 1$, and $f_j(k) = 0$ for $k \neq 1$, we obtain

$$\mathsf{E}_\nu[\exp\{\theta(\tilde{S}_n^*)\}] = e^{-\theta n} \frac{n!}{2\pi i} \oint g(\lambda) d\lambda \tag{A.6}$$

where

$$g(\lambda) = e^\lambda \prod_{j=1}^{m} (1 - (\lambda\nu_j)e^{-\lambda\nu_j} + (\lambda\nu_j)e^{-\lambda\nu_j} e^\theta) \frac{1}{\lambda^{n+1}}.$$

The rest of the proof is an application of the saddle point method [60]. It consists of two steps: The first step is to pick a particular closed contour around $\lambda = 0$ to carry out the integration. It is desirable to have a contour along which $g(\lambda)$ behaves violently: $g(\lambda)$ is large on a small interval on the contour and significantly smaller along the rest, so that the value of integral can be approximated by the integration over this small interval. Such a contour can be found by identifying a *saddle point* of $g(\lambda)$ at which the derivative of $g(\lambda)$ vanishes, and then pick a contour that goes through the saddle point. The second step is to apply the Laplace method to estimate the integral along the contour.

We now apply the first step of the saddle point method. Note that the derivative of g is given by

$$\frac{d}{d\lambda} g(\lambda) = g(\lambda) [\sum_{j=1}^{m} \frac{\nu_j(e^\theta - 1 + e^{\lambda\nu_j})}{\lambda\nu_j(e^\theta - 1) + e^{\lambda\nu_j}} - \frac{n+1}{\lambda}].$$

To simplify the derivation, we select a point that is close to a saddle point, where the derivative vanishes, defined as the solution to

$$\sum_{j=1}^{m} \frac{\lambda\nu_j(e^\theta - 1 + e^{\lambda\nu_j})}{\lambda\nu_j(e^\theta - 1) + e^{\lambda\nu_j}} = n. \tag{A.7}$$

If λ on the left-hand side was taken to be a saddle point, then the right-hand side would be $n + 1$ instead of n, and we will see this error is negligible for our purposes.

Equation (A.7) has one unique real-valued nonnegative solution, which we denote by λ_0: When restricting λ to $\mathbb{R}$, the left-hand-side is a continuous function of λ. Moreover, its value is 0 when $\lambda = 0$, increases to ∞ when λ increases to ∞, and it is a strictly increasing function on $[0, \infty)$.

We now obtain an asymptotic expansion of λ_0. We first show that $\lambda_0 = O(n)$. It is straightforward to see that

$$\frac{1}{1 + e^{-1}(e^\theta - 1)} \leq \frac{e^\theta - 1 + e^{\lambda\nu_j}}{\lambda\nu_j(e^\theta - 1) + e^{\lambda\nu_j}} \leq e^\theta.$$

Substituting this into (A.7) leads to

$$ne^{-\theta} \leq \lambda_0 \leq n(1 + e^{-1}(e^\theta - 1)). \tag{A.8}$$

It follows from the bound (A.8) and $\nu \in \mathcal{P}_m^b$ that $\lambda_0\nu_j = o(1)$. Thus the denominator of (A.7) satisfies $\lambda_0\nu_j(e^\theta - 1) + e^{\lambda_0\nu_j} = 1 + o(1)$. Substituting this into (A.7) leads to

$$\sum_{j=1}^{m} \lambda_0\nu_j(e^\theta - 1 + e^{\lambda_0\nu_j}) = n(1 + o(1)).$$

Consequently,

$$\lambda_0 = ne^{-\theta}(1 + o(1)).$$

To obtain a refined approximation, let $w := \lambda_0 e^\theta/n - 1$ so that

$$\lambda_0 = ne^{-\theta}(1 + w). \tag{A.9}$$

An approximation for w will be obtained: Since $\lambda_0\nu_j = O(\frac{n}{m})$, we have that the numerator and denominator in the summand of (A.7) satisfy

$$\lambda_0\nu_j(e^\theta - 1 + e^{\lambda_0\nu_j}) = \lambda_0\nu_j(e^\theta + \lambda_0\nu_j + O(\frac{n^2}{m^2})),$$
$$\lambda_0\nu_j(e^\theta - 1) + e^{\lambda_0\nu_j} = 1 + \lambda_0\nu_j e^\theta + O(\frac{n^2}{m^2}).$$

Thus,

$$\sum_{j=1}^{m} \frac{\lambda_0\nu_j(e^\theta - 1 + e^{\lambda_0\nu_j})}{\lambda_0\nu_j(e^\theta - 1) + e^{\lambda_0\nu_j}} = \sum_j [\lambda_0\nu_j e^\theta + \lambda_0^2\nu_j^2(1 - e^{2\theta}) + O(\frac{n^3}{m^3})].$$

Substituting this and (A.9) into (A.7) leads to

$$w + n\sum_j \nu_j^2(1+w)^2(e^{-2\theta}-1) = O(\frac{n^2}{m^3}),$$

which upon solving for w gives

$$w = n\sum_j \nu_j^2(1-e^{-2\theta})(1+O(\frac{n}{m})) = O(\frac{n}{m}). \tag{A.10}$$

The integration in (A.6) is now carried out along the closed contour given by $\lambda = \lambda_0 e^{i\psi} = ne^{-\theta}(1+w)e^{i\psi}$:

$$\begin{aligned} & \mathsf{E}_\nu[\exp\{\theta(\tilde{S}_n^*)\}] \\ = \;& e^{-\theta n}\frac{n!}{2\pi}\int_{-\pi}^{\pi} g(\lambda_0 e^{i\psi})\lambda_0 e^{i\psi} d\psi \\ = \;& \frac{n!}{2\pi}\lambda_0^{-n}e^{-\theta n}\,\mathrm{Re}\big[\int_{-\pi}^{\pi} e^{-in\psi}\prod_{j=1}^{m}(\lambda_0\nu_j(e^\theta-1)e^{i\psi}+e^{\lambda_0\nu_j e^{i\psi}})d\psi\big]. \end{aligned}$$

We now complete the proof Proposition A.1 by applying the second step of the saddle point method: Estimating the integral by the Laplace method. We begin with a rough estimate of the integrand in (A.11). Define

$$h(\psi) := e^{-in\psi}\prod_{j=1}^{m}\big(\lambda_0\nu_j(e^\theta-1)e^{i\psi}+e^{\lambda_0\nu_j e^{i\psi}}\big). \tag{A.11}$$

It follows from $\lambda_0 = n^{-\theta}(1+o(1))$ that

$$\begin{aligned} h(\psi) &= e^{-in\psi}\prod_{j=1}^{m}\big(\lambda_0\nu_j(e^\theta-1)e^{i\psi}+1+\lambda_0\nu_j e^{i\psi}+O(\frac{n^2}{m^2})\big) \\ &= e^{-in\psi}\prod_{j=1}^{m}\big(1+\lambda_0\nu_j e^\theta e^{i\psi}+O(\frac{n^2}{m^2})\big) \\ &= e^{-in\psi}\exp\{\sum_{j=1}^{m}\big(\lambda_0\nu_j e^\theta e^{i\psi}+O(\frac{n^2}{m^2})\big)\} \\ &= e^{-in\psi}e^n\exp\{-n(1-e^{i\psi})+O(\frac{n^2}{m})\}. \end{aligned}$$

For any $\psi \neq 0$, $|h(\psi)|$ is exponentially smaller than the value of $h(\psi)$ at $\psi = 0$. Therefore, the integral in (A.11) can be approximated by integrating over a small

interval around $\psi = 0$. Split the integral in (A.11) into three parts:

$$\begin{aligned} I_1 &= \mathrm{Re}[\int_{-\pi/3}^{\pi/3} h(\psi)d\psi], \\ I_2 &= \mathrm{Re}[\int_{-\pi}^{-\pi/3} h(\psi)d\psi], \\ I_3 &= \mathrm{Re}[\int_{\pi/3}^{\pi} h(\psi)d\psi]. \end{aligned} \tag{A.12}$$

We first estimate I_1. Denote $H(\psi) = \log(h(\psi))$. Simple calculus gives

$$\begin{aligned} H(\psi) &= -in\psi + \sum_{j=1}^{m} \log(\lambda_0\nu_j(e^\theta - 1)e^{i\psi} + \exp\{\lambda_0\nu_j e^{i\psi}\}), \\ H'(\psi) &= -in + i\sum_{j=1}^{m} \frac{\lambda_0\nu_j(e^\theta - 1)e^{i\psi} + \lambda_0\nu_j e^{i\psi}\exp\{\lambda_0\nu_j e^{i\psi}\}}{\lambda_0\nu_j(e^\theta - 1)e^{i\psi} + \exp\{\lambda_0\nu_j e^{i\psi}\}}, \\ H''(\psi) &= -\sum_{j=1}^{m} \exp\{\lambda_0\nu_j e^{i\psi}\} \frac{1}{(\lambda_0\nu_j(e^\theta - 1)e^{i\psi} + \exp\{\lambda_0\nu_j e^{i\psi}\})^2} \\ &\quad \times \big(\lambda_0\nu_j(e^\theta - 1)e^{i\psi}(1 - \lambda_0\nu_j e^{i\psi} + \lambda_0^2\nu_j^2 e^{2i\psi}) \\ &\quad + \lambda_0\nu_j e^{i\psi}\exp\{\lambda_0\nu_j e^{i\psi}\}\big). \end{aligned} \tag{A.13}$$

It is clear that $\mathrm{Im}(H(0)) = 0$. It follows from (A.7) that $H'(0) = 0$. Estimates of $\mathrm{Re}(H(0))$ and $H''(\psi)$ are obtained from substituting (A.9) and (A.10) into the expression of $H(\psi)$ and $H''(\psi)$ and applying asymptotic analysis. In conclusion,

$$\begin{aligned} &\mathrm{Im}(H(0)) = 0, \\ &\mathrm{Re}(H(0)) = n(1 + w) - \tfrac{1}{2}n^2(\sum_{j=1}^{m} \nu_j^2)(1 - e^{-2\theta}) + O(\frac{n^3}{m^2}), \\ &H'(0) = 0, H''(\psi) = -ne^{i\psi} + O(\frac{n^2}{m}). \end{aligned} \tag{A.14}$$

To obtain an upper-bound on I_1, note that for large enough n and for any $\psi \in [-\pi/3, \pi/3]$, we have $\mathrm{Re}(H''(\psi)) \leq -0.4n$. It then follows from the mean value theorem that

$$\mathrm{Re}(H(\psi)) \leq H(0) - 0.2n\psi^2.$$

Consequently, for large enough n and m,

$$I_1 \leq e^{H(0)} \int_{-\pi/3}^{-\pi/3} e^{-0.2\psi^2} d\psi \leq e^{H(0)} \int_{-\infty}^{\infty} e^{-0.2\psi^2} d\psi = e^{H(0)} \frac{\sqrt{\pi}}{\sqrt{0.2n}}. \tag{A.15}$$

To obtain a lower-bound on I_1, note that $\mathrm{Im}(H''(\psi)) = -n\sin(\psi) + O(\frac{n^3}{m^2})$. Applying $|\sin(\psi)| \leq |\psi|$, we have that for large enough n, for any $\psi \in [-\pi/3, \pi/3]$, $|\mathrm{Im}(H''(\psi))| \leq 1.1n|\psi|$. It also follows from (A.14) that $\mathrm{Re}(H''(\psi)) \geq -1.1n$. Applying the mean value theorem, we conclude that there exists some $c > 0$ such that for $\psi \in [-\pi/3, \pi/3]$,

$$\begin{aligned} \mathrm{Re}(H(\psi)) &\geq H(0) - 1.1n\psi^2, \\ |\mathrm{Im}(H(\psi))| &\leq 1.1n|\psi|^3 + c\frac{n^2}{m}\psi^2. \end{aligned}$$

Use the short-hand notation $t_n = 0.1\min\{n^{-1/3}, \sqrt{m}/(\sqrt{c}n)\}$. For $\psi \in [-t_n, t_n]$, we have $\cos(\mathrm{Im}(H(\psi))) \geq 0.5$, and thus $\mathrm{Re}(e^{H(\psi)}) \geq 0.5e^{\mathrm{Re}(H(\psi))}$. The integration for I_1 is further split into three parts:

$$I_1 = \mathrm{Re}[\int_{-\pi/3}^{-t_n} e^{H(\psi)} d\psi] + \mathrm{Re}[\int_{t_n}^{\pi/3} e^{H(\psi)} d\psi] + \mathrm{Re}[\int_{-t_n}^{t_n} e^{H(\psi)} d\psi].$$

The absolute value of the first term is upper-bounded as follows:

$$\begin{aligned} |\int_{-\pi/3}^{-t_n} e^{H(\psi)} d\psi| &\leq e^{H(0)} \int_{-\infty}^{-t_n} e^{-0.4n\psi^2} d\psi \\ &= t_n e^{H(0)} \int_{-\infty}^{-1} e^{-0.4nt_n^2\bar{\psi}^2} d\bar{\psi} \\ &\leq t_n e^{H(0)} \int_{-\infty}^{-1} e^{-0.4nt_n^2|\bar{\psi}|} d\bar{\psi} = e^{H(0)} O(\frac{1}{nt_n}) \\ &= e^{H(0)} o(\frac{1}{\sqrt{n}}). \end{aligned} \tag{A.16}$$

The second term is also bounded in a similar way. The third term is lower-bounded

as follows:

$$\begin{aligned}
&\mathrm{Re}[\int_{-t_n}^{t_n} e^{H(\psi)}d\psi] \\
\geq &\int_{-t_n}^{t_n} 0.5e^{\mathrm{Re}(H(\psi))}d\psi \geq 0.5e^{H(0)}\int_{-t_n}^{t_n} e^{-1.1n\psi^2}d\psi \\
\geq &0.5e^{H(0)}[\int_{-\infty}^{\infty} e^{-1.1n\psi^2}d\psi - 2\int_{-\infty}^{-t_n} e^{-1.1n\psi^2}d\psi] \\
\geq &0.5e^{H(0)}(\frac{\sqrt{\pi}}{\sqrt{1.1n}} + O(\frac{1}{nt_n})) = 0.5e^{H(0)}\frac{\sqrt{\pi}}{\sqrt{1.1n}}(1+o(1)),
\end{aligned}$$

where the last inequality follows from an argument similar to (A.16). Combining these bounds together, we obtain

$$\begin{aligned}
I_1 &\geq \mathrm{Re}[\int_{-t_n}^{t_n} e^{H(\psi)}d\psi] - |\,\mathrm{Re}[\int_{-\pi/3}^{-t_n} e^{H(\psi)}d\psi]| - |\,\mathrm{Re}[\int_{t_n}^{\pi/3} e^{H(\psi)}d\psi]| \\
&\geq e^{H(0)}\frac{0.5\sqrt{\pi}}{\sqrt{1.1n}}(1+o(1)).
\end{aligned}$$

Combing this and (A.15) leads to

$$I_1 = e^{H(0)}\frac{1}{\sqrt{n}}e^{O(1)} = e^{n(1+o(1))}\frac{1}{\sqrt{n}}e^{O(1)}, \tag{A.17}$$

where the last equality follows from the estimate of $H(0)$ given in (A.14) and (A.10).

We now estimate I_2 and I_3. For $\psi \in [-\pi, -\pi/3] \cup [\pi/3, \pi]$, we obtain from (A.12) that $|h(\psi)| \leq \exp\{0.5n + O(\frac{n^2}{m})\}$, which implies $\mathrm{Re}[I_2] + \mathrm{Re}[I_3] = O(e^{0.6n})$.

We are now ready to prove the claim of this proposition. It can be seen from the estimates that I_2 and I_3 are much smaller than I_1. Thus, the integral can be approximated by the estimate of I_1: Substituting (A.17) and (A.14) into (A.11),

we obtain

$$\begin{aligned}
\mathsf{E}_\nu[\exp\{\theta(\tilde{S}_n^*)\}] &= \frac{n!}{2\pi}\lambda_0^{-n}e^{-\theta n}I_1(1+o(1)) \\
&= \frac{n!}{2\pi}\lambda_0^{-n}e^{-\theta n}e^{H(0)}\frac{1}{\sqrt{n}}e^{O(1)}(1+o(1)) \\
&= \frac{n!}{n^n\sqrt{2\pi n}}\big(1+n\sum_j \nu_j^2(1-e^{-2\theta})+O(\frac{n^2}{m^2})\big)^{-n} \\
&\quad \times \exp\{\tfrac{1}{2}n^2(\sum_{j=1}^m \nu_j^2)(1-e^{-2\theta})+O(\frac{n^3}{m^2})\}e^{O(1)}) \\
&= \frac{n!e^n}{n^n\sqrt{2\pi n}}\exp\{-\tfrac{1}{2}n^2(\sum_{j=1}^m \nu_j^2)(1-e^{-2\theta})+O(\frac{n^3}{m^2})\}e^{O(1)}.
\end{aligned}$$

Stirling formula gives $\frac{n!e^n}{n^n 2\pi\sqrt{n}} = 1 + O(\frac{1}{n})$. The claim of the proposition is obtained on taking logarithm on both sides. □

A.3 Proof of Proposition A.2 and Proposition A.3

The proofs of Proposition A.2 and Proposition A.3 use steps similar to those leading to the upper-bound in Proposition A.1. However, the approximation given by (A.9) and (A.10) is no longer valid, so a different approximation is required. The conclusions on the existence and uniqueness of the solution λ_0 and the bounds in (A.8) are still valid, and our proof begins from there.

To simplify the presentation, we use the following notation similar to the small "o" notation: We write $x = o^\eta(1)$ whenever there exists a function $s(\eta)$ that does not depend on θ, n, and ν, such that $|x| \le s(\eta)$ and $\lim_{\eta\to 0} s(\eta) = 0$.

Consider any η and ν. Write $\mathcal{T}_\eta = \mathcal{T}_\eta(\nu)$. For any $j \notin \mathcal{T}_\eta$, we obtain the expansion of the summand in (A.7) via the mean value theorem:

$$\frac{\lambda_0\nu_j(e^\theta - 1 + e^{\lambda_0\nu_j})}{\lambda_0\nu_j(e^\theta-1)+e^{\lambda_0\nu_j}} = \lambda_0\nu_j e^\theta + \lambda_0^2\nu_j^2(1-e^{2\theta})(1+o^\eta(1)).$$

For any $j \in \mathcal{T}_\eta$, the following equality holds:

$$\frac{\lambda_0\nu_j(e^\theta - 1 + e^{\lambda_0\nu_j})}{\lambda_0\nu_j(e^\theta-1)+e^{\lambda_0\nu_j}} = D_j\lambda\nu_j e^\theta,$$

where

$$D_j := \frac{e^{-\theta} + e^{-\lambda_0\nu_j}(1 - e^{-\theta})}{1 + \lambda_0\nu_j e^{-\lambda_0\nu_j}(e^{\theta} - 1)} \geq e^{-2\theta}. \tag{A.18}$$

Substituting these estimates into (A.7) leads to

$$\lambda_0(1 + \sum_{j\in\mathcal{T}_\eta} \nu_j(D_j - 1))e^{\theta} + \lambda_0^2 \sum_{j\notin\mathcal{T}_\eta} \nu_j^2(1 - e^{2\theta})(1 + o^{\eta}(1)) = n. \tag{A.19}$$

Applying $\lambda_0 \sum_{j\notin\mathcal{T}_\eta} \nu_j^2 \leq \eta \sum_{j\notin\mathcal{T}_\eta} \nu_j \leq \eta$ then gives

$$\lambda_0 = \frac{ne^{-\theta}}{1 + \sum_{j\in\mathcal{T}_\eta} \nu_j(D_j - 1)}(1 + o^{\eta}(1)).$$

Introducing a variable w as before,

$$\lambda_0 = \frac{ne^{-\theta}}{1 + \sum_{j\in\mathcal{T}_\eta} \nu_j(D_j - 1)}(1 + w). \tag{A.20}$$

On substituting (A.20) into (A.19), we obtain

$$w = \frac{n\big(\sum_{j\notin\mathcal{T}_\eta} \nu_j^2(1 - e^{-2\theta})\big)}{(1 + \sum_{j\in\mathcal{T}_\eta} \nu_j(D_j - 1))^2}(1 + o^{\eta}(1)) = o^{\eta}(1). \tag{A.21}$$

In the proofs of both propositions, we integrate (A.6) along the closed contour corresponding to $\lambda = \lambda_0 e^{i\psi}$ from $\psi = -\pi$ to $\psi = \pi$, and use the same definition of $h(\psi)$ given in (A.11) and $H(\psi) = \log(h(\psi))$. The integral is given in (A.11) and our task is to estimate it. We now give the details.

Proof of Proposition A.2. We first show that any ψ,

$$\mathrm{Re}(H(\psi)) \leq H(0) = \sum_j [\lambda_0\nu_j + \log\big(1 + \lambda_0\nu_j e^{-\lambda_0\nu_j}(e^{\theta} - 1)\big)]. \tag{A.22}$$

Thus to bound the integral in (A.11), we only need to bound $H(0)$. For $\psi \in [-\frac{1}{2}\pi, \frac{1}{2}\pi]$, the summand in the expression of $\mathrm{Re}(H(\psi))$ given in (A.13) is bounded as follows:

$$\begin{aligned}
&\mathrm{Re}[\log\big(\lambda_0\nu_j(e^{\theta} - 1)e^{i\psi} + e^{\lambda_0\nu_j e^{i\psi}}\big)] \\
&= \mathrm{Re}[\log(e^{\lambda_0\nu_j e^{i\psi}}) + \log\big(1 + \lambda_0\nu_j(e^{\theta} - 1)e^{i\psi}e^{-\lambda_0\nu_j e^{i\psi}}\big)] \\
&\leq \lambda_0\nu_j \cos\psi + \log\big(1 + \lambda_0\nu_j e^{-\lambda_0\nu_j\cos\psi}(e^{\theta} - 1)\big).
\end{aligned} \tag{A.23}$$

The right-hand side is a convex function of $\cos\psi$ for $\psi \in [-\frac{1}{2}\pi, \frac{1}{2}\pi]$. Thus, it achieves its maximum value at $\cos\psi = 1$ or $\cos\psi = 0$. Note that its value at $\cos\psi = 1$ is exactly equal to the summand in $H(0)$. Moreover, we can show that its value at $\cos\psi = 1$ is no smaller than its value at $\cos\psi = 0$:

$$\begin{aligned}
&\lambda_0\nu_j + \log\big(1 + \lambda_0\nu_j(e^\theta - 1)e^{-\lambda_0\nu_j}\big) - \log\big(1 + \lambda_0\nu_j(e^\theta - 1)\big) \\
&= \lambda_0\nu_j + \log\big(\frac{1 + \lambda_0\nu_j(e^\theta - 1)e^{-\lambda_0\nu_j}}{1 + \lambda_0\nu_j(e^\theta - 1)}\big) \\
&\le \lambda_0\nu_j + \log(e^{-\lambda_0\nu_j}) = 0,
\end{aligned}$$

where the inequality follows from $\theta \ge 0$. This leads to (A.22) for $\psi \in [-\frac{1}{2}\pi, \frac{1}{2}\pi]$.

For $\psi \in [-\pi, -\frac{1}{2}\pi] \cup [\frac{1}{2}\pi, \pi]$, we have $|e^{\lambda_0\nu_j e^{i\psi}}| \le 1$. Consequently,

$$|\lambda_0\nu_j(e^\theta - 1)e^{i\psi} + e^{\lambda_0\nu_j e^{i\psi}}| \le 1 + \lambda_0\nu_j(e^\theta - 1),$$

which leads to

$$\mathrm{Re}[\log\big(\lambda_0\nu_j(e^\theta - 1)e^{i\psi} + e^{\lambda_0\nu_j e^{i\psi}}\big)] \le \log\big(1 + \lambda_0\nu_j(e^\theta - 1)\big). \tag{A.24}$$

The right-hand side of (A.24) is equal to the value of the right-hand side of (A.23) at $\cos\psi = 0$, which has been shown in the previous paragraph to be smaller than $H(0)$. This leads to (A.22) for $\psi \in [-\pi, -\frac{1}{2}\pi] \cup [\frac{1}{2}\pi, \pi]$.

We now approximate the right-hand side of (A.22): For $j \notin \mathcal{T}_\eta$, we have

$$\lambda_0\nu_j + \log\big(1 + \lambda_0\nu_j e^{-\lambda_0\nu_j}(e^\theta - 1)\big) = \lambda_0\nu_j^\theta + \tfrac{1}{2}\lambda_0^2\nu_j^2(1 - e^{2\theta})(1 + o^\eta(1)).$$

For $j \in \mathcal{T}_\eta$, we have the inequality

$$\lambda_0\nu_j + \log\big(1 + \lambda_0\nu_j e^{-\lambda_0\nu_j}(e^\theta - 1)\big) \le \lambda_0\nu_j e^\theta + \lambda_0\nu_j(1 - e^{-\lambda_0\nu_j})(1 - e^\theta).$$

Substituting these two estimates, and (A.20), (A.22) into (A.11) leads to

$$
\begin{aligned}
&\mathsf{E}_\nu[\exp\{\theta(\tilde{S}_n^*)\}] \le \frac{n!}{2\pi}\lambda_0^{-n}e^{-\theta n}2\pi\exp\{H(0)\} \qquad \text{(A.25)}\\
&\le n!\lambda_0^{-n}e^{-\theta n}\exp\{\sum_{j\notin\mathcal{T}_\eta}[\lambda_0\nu_j e^\theta + \tfrac{1}{2}\lambda_0^2\nu_j^2(1-e^{2\theta})(1+o^\eta(1))]\}\\
&\quad\times\exp\{\sum_{j\in\mathcal{T}_\eta}[\lambda_0\nu_j e^\theta + \lambda_0\nu_j(1-e^{-\lambda_0\nu_j})(1-e^\theta)]\}\\
&= \frac{n!e^n}{n^n}\big(1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)\big)^n(1+w)^{-n}\\
&\quad\times\exp\{-\frac{\frac{1}{2}n^2\sum_{j\notin\mathcal{T}_\eta}\nu_j^2(1-e^{-2\theta})(1+o^\eta(1))}{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)}\}\\
&\quad\times\exp\{n[\frac{(1+w)+\sum_{j\in\mathcal{T}_\eta}\nu_j(1-e^{-\lambda_0\nu_j})(e^{-\theta}-1)}{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)}]\}\\
&\le \frac{n!e^n}{n^n}\exp\{-n\log(1+w)+\frac{nw}{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)}\}\\
&\quad\times\exp\{-\frac{\frac{1}{2}n^2\sum_{j\notin\mathcal{T}_\eta}\nu_j^2(1-e^{-2\theta})(1+o^\eta(1))}{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)}]\}\\
&\quad\times\exp\{n[\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)-1\\
&\qquad+\frac{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(1-e^{-\lambda_0\nu_j})(e^{-\theta}-1)}{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)}]\}. \qquad \text{(A.26)}
\end{aligned}
$$

We now bound each exponential term on the right-hand side of (A.26). Applying (A.21) and the lower-bound on D_j in (A.18) gives the following bound on the second term:

$$
-\frac{\frac{1}{2}n^2\sum_{j\notin\mathcal{T}_\eta}\nu_j^2(1-e^{-2\theta})}{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)} \le -\tfrac{1}{2}e^{-2\theta}nw(1+o^\eta(1)). \qquad \text{(A.27)}
$$

The first exponential term satisfies

$$
-n\log(1+w)+\frac{nw}{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)} = -nwo^\eta(1), \qquad \text{(A.28)}
$$

which follows from (A.18) and $w = o^\eta(1)$. Combining (A.27) and (A.28) implies that for small enough η, the sum of the first and second term is *negative*.

The last exponential term is bounded in the following lemma:

Lemma A.6.

$$\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)-1+\frac{(1+\sum_{j\in\mathcal{T}_\eta}\nu_j(1-e^{-\lambda_0\nu_j})(e^{-\theta}-1)}{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)}\le-\beta(\nu)\alpha(\theta)\le 0.$$

This lemma is proved in Appendix A.3.1. Applying Lemma A.6 and the conclusion that the sum of the first and second exponential term is negative, we obtain

$$\mathsf{E}_\nu[\exp\{\theta(\tilde{S}_n^*)\}]\le\frac{n!e^n}{\sqrt{2\pi n}n^n}\sqrt{2\pi n}\exp\{-n\beta(\nu)\alpha(\theta)\}.$$

Taking the logarithm on both side and applying Stirling's formula, we obtain

$$\Lambda_{\nu,S_n^*}(\theta)\le-n\beta(\nu)\alpha(\theta)+\tfrac{1}{2}\log(2\pi n)+O(\frac{1}{n}).$$

Since $\beta(\nu)\ge\underline{\beta}$, the second term $\frac{1}{2}\log(2\pi n)$ becomes negligible comparing to the first term for large n. This leads to the claim of the proposition. □

Proof of Proposition A.3. We pick $\overline{\beta}$ so that $\overline{\beta}=o^\eta(1)$. It then follows that

$$\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)=o^\eta(1). \tag{A.29}$$

Substituting this into (A.20) and (A.21) gives

$$\lambda_0=ne^{-\theta}(1+o^\eta(1)),\quad w=n(\sum_{j\notin\mathcal{T}_\eta}\nu_j^2)(1-e^{-2\theta})(1+o^\eta(1)). \tag{A.30}$$

The rest of the proof is similar to the proof of Proposition A.1. Applying (A.22) to $j\in\mathcal{T}_\eta$, we obtain

$$\begin{aligned}|h(\psi)|&\le|e^{-in\psi}\prod_{j\notin\mathcal{T}_\eta}\big(\lambda_0\nu_j(e^\theta-1)e^{i\psi}+e^{\lambda_0\nu_je^{i\psi}}\big)|\\&\quad\times\prod_{j\in\mathcal{T}_\eta}\exp\{\lambda_0\nu_j+\log\big(1+\lambda_0\nu_je^{-\lambda_0\nu_j}(e^\theta-1)\big)\}\\&\le|e^{-in\psi}|\exp\{\big(\sum_{j\notin\mathcal{T}_\eta}\lambda_0\nu_je^\theta\cos\psi(1+o^\eta(1))\big)+\sum_{j\in\mathcal{T}_\eta}\lambda_0\nu_je^\theta\}\\&=e^n\exp\{-n(1-\cos\psi+o^\eta(1))\}.\end{aligned} \tag{A.31}$$

It is clear from (A.31) that the integrand is large at the interval around 0. Thus, we again split the integral in (A.11) into three parts I_1, I_2 and I_3 as in (A.12). We

will show later that I_2 and I_3 are much smaller than I_1.

We first upper-bound I_1. Similar to (A.14), we have

$$\mathrm{Im}(H(0)) = 0, \mathrm{Re}(H'(0)) = 0, \mathrm{Im}(H'(0)) = 0.$$

We now estimate $H''(\psi)$, whose exactly formula is given in (A.13). Consider $j \in \mathcal{T}_\eta$. For $\psi \in [-\pi/3, \pi/3]$, we have the following inequality:

$$|1 + \lambda_0\nu_j(e^\theta - 1)e^{i\psi} \exp\{-\lambda_0\nu_j e^{i\psi}\}| \geq 1,$$

$$|\lambda_0\nu_j(e^\theta - 1)e^{i\psi}(1-\lambda_0\nu_j e^{i\psi}+\lambda_0^2\nu_j^2 e^{2i\psi}) \exp\{-\lambda_0\nu_j e^{i\psi}\}+\lambda_0\nu_j e^{i\psi}| \leq 100\lambda_0\nu_j e^\theta.$$

Substituting these into (A.13), we obtain $|H''(\psi)| \leq 100\overline{\beta}n(1 + o^\eta(1)) = no^\eta(1)$. Substituting this and the estimate (A.30) into the expression of $H''(\psi)$ leads to

$$H''(\psi) = -n(e^{i\psi} + o^\eta(1)).$$

Note that the assumption of the proposition allows us to take very small η. We choose it small enough so that the term $o^\eta(1)$ in the above equation is smaller than 0.05. Then for large enough n, for any $\psi \in [-\pi/3, \pi/3]$, we have $\mathrm{Re}(H''(\psi)) \leq -0.4n$. It follows from the mean value theorem that

$$\mathrm{Re}(H(\psi)) \leq H(0) - 0.2n\psi^2.$$

Consequently, for large enough n and m, we have

$$I_1 \leq e^{H(0)} \int_{-\pi/3}^{-\pi/3} e^{-0.2\psi^2} d\psi \leq e^{H(0)} \int_{-\infty}^{\infty} e^{-0.2\psi^2} d\psi = e^{H(0)} \frac{\sqrt{\pi}}{\sqrt{0.2n}}. \tag{A.32}$$

We now bound the tails I_2 and I_3. For $\psi \in [-\pi, -\pi/3] \cup [\pi/3, \pi]$, we obtain from (A.31) that $|h(\psi)| \leq \exp\{0.5n(1 + o^\eta(1))\}$. Thus, for small enough η, we have

$$\mathrm{Re}[I_2] + \mathrm{Re}[I_3] = O(e^{0.6n}).$$

Substituting the estimate for I_1, I_2 and I_3 into (A.11) gives

$$\mathsf{E}_\nu[\exp\{\theta(\tilde{S}_n^*)\}] \leq \frac{n!}{\sqrt{1.6n\pi}} \lambda_0^{-n} e^{-\theta n} e^{H(0)}(1 + o(1)).$$

Note that the right-hand side is almost the same as the right-hand side of (A.25)

except for the multiplication term $\frac{1}{\sqrt{1.6n\pi}}(1+o(1))$. Thus, we can bound it using the right-hand side of (A.26) after taking into account this additional multiplication term. Applying Lemma A.6 and the estimate (A.28), we obtain

$$\begin{aligned}
&\mathsf{E}_\nu[\exp\{\theta(\tilde{S}_n^*)\}] \\
\leq &\frac{n!e^n}{\sqrt{1.6n\pi}n^n}\exp\{-\frac{\frac{1}{2}n^2\sum_{j\notin\mathcal{T}_\eta}\nu_j^2(1-e^{-2\theta})(1+o^\eta(1))}{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)}\}(1+o^\eta(1)).
\end{aligned}$$

Substituting (A.29) and Stirling's formula into the right-hand side of the above inequality leads to

$$\mathsf{E}_\nu[\exp\{\theta(S_n^*-n)\}] \leq \frac{1}{\sqrt{0.8}}\exp\{-\tfrac{1}{2}n^2(\sum_{j\notin\mathcal{T}_\eta}\nu_j^2)(1-e^{-2\theta})(1+o^\eta(1))\}(1+o(1)).$$

The claim of the proposition is obtained on taking logarithm on both sides. □

A.3.1 Proof of Lemma A.6

Proof of Lemma A.6. The coefficient of n can be bounded as follows:

$$\begin{aligned}
&\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)-1+\frac{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(1-e^{-\lambda_0\nu_j})(e^{-\theta}-1)}{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)} \\
=&\frac{\left(\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)\right)^2+\sum_{j\in\mathcal{T}_\eta}\nu_j(1-e^{-\lambda_0\nu_j})(e^{-\theta}-1)}{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)} \\
\leq&\frac{\left(\sum_{j\in\mathcal{T}_\eta}\nu_j\right)\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)^2}{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)} \\
&+\frac{\sum_{j\in\mathcal{T}_\eta}\nu_j(1-e^{-\lambda_0\nu_j})(e^{-\theta}-1)}{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)} \\
\leq&\frac{\sum_{j\in\mathcal{T}_\eta}\nu_j[(D_j-1)^2+(1-e^{-\lambda_0\nu_j})(e^{-\theta}-1)]}{1+\sum_{j\in\mathcal{T}_\eta}\nu_j(D_j-1)},
\end{aligned}$$

where the first inequality follows from Jessen's inequality and the second follows from $\sum_{j\in\mathcal{T}_\eta}\nu_j\leq 1$.

We now bound the summand in the numerator on the right-hand side of (A.33).

Consider any $j \in \mathcal{T}_\eta$. Let $x = \lambda_0 \nu_j$. Applying the formula of D_j in (A.18) gives

$$\begin{aligned} &(D_j - 1)^2 + (1 - e^{-x})(e^{-\theta} - 1) \\ =&\frac{e^{-x} + e^{-\theta} - e^{-x}e^{-\theta}}{\left(1 + xe^{-x}(e^{\theta} - 1)\right)^2}[(1 - e^{-x})(e^{-\theta} - 1) + \left(xe^{-x}(e^{\theta} - 1)\right)^2]. \end{aligned} \tag{A.33}$$

Let $t(x) = (1 - e^{-x})(e^{-\theta} - 1) + \left(xe^{-x}(e^{\theta} - 1)\right)^2$. For $\theta \in (0, 0.5]$, we can show that $t(x)$ is strictly decreasing on $[0, \infty)$. Note that $j \in \mathcal{T}_\eta$ implies $n\nu_j \geq \eta$, which combined with (A.8) implies $x = \lambda_0 \nu_j \geq \eta e^{-\theta}$. Therefore, $t(x) \leq t(\eta e^{-\theta}) < 0$.

Substituting this into (A.33) and then (A.33), and using the elementary fact that $\frac{e^{-x} + e^{-\theta} - e^{-x}e^{-\theta}}{\left(1 + xe^{-x}(e^{\theta} - 1)\right)^2} \leq e^{-3\theta}$, we obtain $(D_j - 1)^2 + (1 - e^{-x})(e^{-\theta} - 1) \leq -e^{-3\theta} t(\eta e^{-\eta})$. The claim of this lemma follow from combining this with (A.33), and using the fact that $D_j \leq 1$.

□

A.4 Proof of Lemma 3.5

Proof of Lemma 3.5. Let μ^* denote one optimizer in the optimization problem (3.12). The main task is to show that μ^* is the bi-uniform distribution. Let $\mathcal{J}_+ = \{j : \mu_j^* \geq \pi_j\}$, $\mathcal{J}_- = \{j : \mu_j^* < \pi_j\}$. Then μ^* is also the optimizer to the following problem:

$$\begin{aligned} \min \quad & \sum_{j \in \mathcal{J}_+} x_j^2 \\ \text{s.t.} \quad & \sum_{j \in \mathcal{J}_+} x_j = \sum_{j \in \mathcal{J}_+} \mu_j^* \\ & x_j = \mu_j^* \text{ for } j \in \mathcal{J}_- \\ & x_j \geq \pi_j \text{ for } j \in \mathcal{J}_+. \end{aligned}$$

By Jensen's inequality, μ^* must satisfy $\mu_j^* = \mu_{j'}^*$ for all $j, j' \in \mathcal{J}_+$. The same conclusion holds for $j \in \mathcal{J}_-$. Thus, μ^* must be a bi-uniform distribution. □

A.5 Proof of Theorem 3.6

The performance of ϕ^{*+} is analyzed by connecting it to the performance of ϕ^*. We first show that its probability of missed detection is no larger than that of ϕ^*. We then apply a proposition similar to Proposition A.1 to analyze its probability

of false alarm. Consider the statistic $\tilde{S}_n^{*+} = S_n^{*+} - \sum_{l=2}^{\bar{l}} v_l \mathbb{I}\{n\Gamma_j^n = l\}$. Define

$$\Lambda_{\nu,\tilde{S}_n^*}(\theta) := \log\big(\mathsf{E}_\nu[\exp\{\theta(\tilde{S}_n^*)\}]\big). \tag{A.34}$$

Proposition A.7. *For any $\nu \in \mathcal{P}_m^b$, the n-sample logarithmic moment generating function for the statistic $\tilde{S}_n^{*+}$ has the following asymptotic expansion:*

$$\begin{aligned}\Lambda_{\nu,\tilde{S}_n^{*+}}(\theta) =& \frac{n^2}{m}\big(m\sum_{j=1}^m \nu_j^2\big)\{-\theta + \tfrac{1}{2}[e^{-2\theta} - (1-2\theta)]\} \\ &+ O(\frac{n^3}{m^2}) + O(1).\end{aligned} \tag{A.35}$$

Proof of Proposition A.7. The proof follows exactly the same step as that of Proposition A.1. The constraints on S_n^{*+} ensures that the additional term $\sum_{l=2}^{\bar{l}} v_l \mathbb{I}\{n\Gamma_j^n = l\}$ has a negligible effect in terms of general error exponent. We illustrate how the key steps in the approximation apply to the current statistic: Instead of (A.11) in the proof of Proposition A.1, we have

$$\begin{aligned}&\mathsf{E}_\nu[\exp\{\theta(\tilde{S}_n^{*+})\}] \\ =&\frac{n!}{2\pi}\lambda_0^{-n}e^{-\theta n} \\ &\times \mathrm{Re}\big[\int_{-\pi}^{\pi} e^{-in\psi}\prod_{j=1}^m\big(\lambda_0\nu_j(e^\theta-1)e^{i\psi} + e^{\lambda_0\nu_j e^{i\psi}} + \sum_{l=2}^{\bar{l}}\frac{(\lambda_0\nu_j)^l}{l!}(e^{-\theta v_l}-1)\big)d\psi\big].\end{aligned}$$

Define

$$h'(\psi) := e^{-in\psi}\prod_{j=1}^m\big(\lambda_0\nu_j(e^\theta-1)e^{i\psi} + e^{\lambda_0\nu_j e^{i\psi}} + \sum_{l=2}^{\bar{l}}\frac{(\lambda_0\nu_j)^l}{l!}(e^{-\theta v_l}-1)\big).$$

It follows from $\lambda_0 = n^{-\theta}(1+o(1))$ that the last term is negligible when $v_2 = 0$ and $\bar{l} < \infty$.

$$\sum_{l=2}^{\bar{l}}\frac{(\lambda_0\nu_j)^l}{l!}(e^{-\theta v_l}-1) = O(\frac{n^3}{m^3}).$$

Therefore, $h'(\psi)$ has the same asymptotic approximation as that of $h(\psi)$ in (A.11).

$$h'(\psi) = e^{-in\psi}\prod_{j=1}^m\big(\lambda_0\nu_j(e^\theta-1)e^{i\psi} + 1 + \lambda_0\nu_j e^{i\psi} + O(\frac{n^2}{m^2})\big).$$

Therefore, $\Lambda_{\nu,\tilde{S}_n^{*+}}$ has the same asymptotic approximation as that of $\Lambda_{\nu,\tilde{S}_n^*}$ up to an approximation error of $O(\frac{n^3}{m^2})$. □

Proof of Theorem 3.6. Since $v_l \geq 0$, we have

$$\tilde{S}_n^{*+} \leq \tilde{S}_n^*.$$

Thus, for the same sequence of thresholds $\tilde{\tau}_n$, we have

$$\mathsf{P}_\mu\{\tilde{S}_n^{*+} \geq \tilde{\tau}_n\} \leq \mathsf{P}_\mu\{\tilde{S}_n^* \geq \tilde{\tau}_n\}.$$

On the other hand, since $\Lambda_{\nu,\tilde{S}_n^{*+}}$ has the same asymptotic approximation as that of $\Lambda_{\nu,\tilde{S}_n^*}$ up to an approximation error of $O(\frac{n^3}{m^2})$, we have

$$\begin{aligned}
&\log \mathsf{P}_\pi\{\tilde{S}_n^{*+} \leq -\tilde{\tau}_n\} \\
\leq &\theta(-\tau_n) + \Lambda_{\pi,\tilde{S}_n^{*+}}(-\theta) \\
= &-\theta\tau_n + \frac{n^2}{m}\big(\theta + \tfrac{1}{2}[e^{2\theta} - (1+2\theta)]\big) + O(\frac{n^3}{m^2}) + O(1),
\end{aligned}$$

which is the same bound as that for $\log \mathsf{P}_\pi\{\tilde{S}_n^* \leq -\tilde{\tau}_n\}$. □

A.6 Proof of Theorem 3.10

The proof of Theorem 3.10 follows exactly the same steps as those for Theorem 3.2. We use Proposition A.8, Proposition A.9 and Proposition A.10 in place of Proposition A.1, Proposition A.2 and Proposition A.3.

Denote

$$\Lambda_{\nu,S_n^{\mathsf{W}}}(\theta) := \log\big(\mathsf{E}_\nu[\exp\{\theta S_n^{\mathsf{W}}\}]\big).$$

Proposition A.8. *For any $\nu \in \mathcal{P}_m^b$, the n-sample logarithmic moment generating function for the statistic S_n^{W} has the following asymptotic expansion*

$$\begin{aligned}
\Lambda_{\nu,S_n^{\mathsf{W}}}(\theta) =&\tfrac{1}{2}n^2(\sum_{j=1}^m(\pi_j - \nu_j)^2)\theta + \tfrac{1}{2}n^2(\sum_{j=1}^m \nu_j^2)[e^\theta - (1+\theta)] \\
&+ O(\frac{n^3}{m^2}) + O(1).
\end{aligned}$$

Proposition A.9. *For all sufficiently small $\eta > 0$, any $\theta \in [-1, 0)$ and any $\underline{\beta} > 0$. There exists n_0 such that for any $n > n_0$, and any ν satisfying $\beta(\nu) \leq \underline{\beta}$, the following holds:*

$$\Lambda_{\nu, S_n^{\mathsf{W}}}(\theta) \leq -\beta(\mu)\alpha'(\theta)n,$$

where $\alpha'(\theta) > 0$ for $\theta \in [-1, 0)$.

Proposition A.10. *For any $\delta > 0$, $\theta \in [-1, 0)$, $\overline{\eta} > 0$, there exists $\eta \in (0, \overline{\eta})$, $\overline{\beta} > 0$, and n_0 such that for any $n > n_0$, and any ν satisfying $\beta(\mu) \leq \overline{\beta}$, the following holds:*

$$\Lambda_{\nu, S_n^{\mathsf{W}}}(\theta) \leq \frac{n^2}{m}[(m \sum_{j \notin \mathcal{T}_\eta(\nu)} (\pi_j - \nu_j)^2)\theta + \tfrac{1}{2}(m \sum_{j \notin \mathcal{T}_\eta(\nu)} \nu_j^2)(e^\theta - (1 + \theta))](1 - \delta).$$

We only outline the proof for Proposition A.8.

Proof of Proposition A.8. The steps are the same as those in the proof of Proposition A.1. Similar to (A.11), we have $\mathsf{E}_\nu^n[\exp\{\theta S_n^{\mathsf{W}}\}] = \frac{n!}{2\pi}\lambda_0^{-n}\,\mathrm{Re}[\int_{-\pi}^{\pi} h(\psi)d\psi]$, where

$$\begin{aligned}
&h(\psi)\\
=&e^{-in\psi}\prod_{j=1}^{m}[\exp\{\lambda_0\nu_j e^{i\psi}\} + (e^{\frac{1}{2}n^2\pi_j^2\theta} - 1) + \lambda_0 e^{i\psi}\nu_j(e^{-n\pi_j\theta} - 1) + \tfrac{1}{2}\lambda_0^2 e^{2i\psi}\nu_j^2(e^\theta - 1)]\\
=&e^{-in\psi}\exp\{ne^{i\psi} + O(\frac{n^2}{m})\},
\end{aligned}$$

and

$$\lambda_0 = n(1 + w), w = n(\sum_j \nu_j\pi_j\theta - \sum_j \nu_j^2(e^\theta - 1))(1 + O(\frac{n}{m})).$$

Again split the integral into three parts I_1, I_2, I_3 as in (A.12). Similar to that in the proof in Proposition A.1, we will find that the integral can be approximated by I_1, whose estimate is given by $I_1 = e^{H(0)}\frac{1}{\sqrt{n}}e^{O(1)}$, where $H(\psi) = \log(h(\psi))$, $\mathrm{Im}(H(0)) = 0$ and

$$\mathrm{Re}(H(0)) = n(1+w) + \tfrac{1}{2}n^2(\sum_{j=1}^{m}(\pi_j - \nu_j)^2)\theta + \tfrac{1}{2}n^2(\sum_{j=1}^{m}\nu_j^2)(e^\theta - 1 - \theta) + O(\frac{n^3}{m^2}).$$

The rest of the steps are almost the same as those in Proposition A.1. The details are omitted. □

Proof of Theorem 3.10. We first prove the lower-bound on $J_F(\phi^{\mathsf{W}})$. Substituting the asymptotic approximation of $\Lambda_{\pi,S_n^{\mathsf{W}}}(\theta)$ given in Proposition A.8 into the Chernoff bound, we obtain that for $\theta \geq 0$,

$$\begin{aligned}\log \mathsf{P}_\pi(\phi_n^{\mathsf{W}} = 1) \leq & -\theta\tau_n + \Lambda_{\pi,S_n^{\mathsf{W}}}(\theta) \\ & = -\theta\tau_n + n^2(\sum_{j=1}^m \pi_j^2)\tfrac{1}{2}[e^\theta - (1+\theta)] + O(\frac{n^3}{m^2}) + O(1).\end{aligned}$$

Since $m\sum_{j=1}^m \pi_j^2 \leq c_1^2$, which is a consequence of Assumption 3.8, we have

$$J_F(\phi^{\mathsf{W}}) \geq \bar{J}_F(\tau) := \sup_{\theta\geq 0}\{\tfrac{1}{2}\tau\theta - \tfrac{1}{2}c_1^2[e^\theta - (1+\theta)]\} > 0.$$

Lower-bounding $J_M(\phi^{\mathsf{W}})$ requires us to obtain a uniform bound on the probability $\mathsf{P}_\mu(\phi_n = 0)$ over $\mu \in \Pi_m$. This time we apply Proposition A.9 and Proposition A.10. Using an argument similar to the proof in Theorem 3.2, we conclude that for any $\delta > 0$, and $\theta \in (0, 1]$, for large enough n,

$$\begin{aligned}\log \mathsf{P}_\mu(\phi_n^{\mathsf{W}} = 0) \leq & \theta\tau_n + \Lambda_{\mu,S_n^{\mathsf{W}}}(-\theta) \\ & = \theta\tau_n - \frac{n^2}{m}[\tfrac{1}{2}\theta m\sum_{j=1}^m(\mu_j - \pi_j)^2 - \tfrac{1}{2}(m\sum_{j=1}^m \mu_j^2)\big(e^{-\theta} - (1-\theta)\big)](1-\delta).\end{aligned}$$

We need to upper-bound the right-hand side uniformly over all $\mu \in \Pi_m$. Using the inequalities $\mu_j^2 \leq 2\pi_j^2 + 2(\pi_j - \mu_j)^2$ and $e^{-\theta} - (1-\theta) \leq \frac{1}{2}\theta^2$ for $\theta > 0$, we obtain

$$\begin{aligned}&\frac{m}{n^2}\log \mathsf{P}_\mu(\phi_n^{\mathsf{W}} = 0) \\ &\leq \theta\frac{m\tau_n}{n^2} - [\tfrac{1}{2}\theta m\sum_{j=1}^m(\mu_j - \pi_j)^2 - \tfrac{1}{2}\theta^2\big(m\sum_{j=1}^m \pi_j^2 + m\sum_{j=1}^m(\mu_j - \pi_j)^2\big)](1-\delta) + O(1) \\ &= \tfrac{1}{2}\theta[-(m\sum_{j=1}^m(\mu_j - \pi_j)^2)(1-\theta) + \theta(m\sum_{j=1}^m \pi_j^2)](1-\delta) + \theta\frac{m\tau_n}{n^2} + O(1).\end{aligned}$$

Applying $m\sum_{j=1}^m(\mu_j - \pi_j)^2 \geq 4\varepsilon^2$ and $m\sum_{j=1}^m \pi_j^2 \leq c_1^2$ leads to,

$$\frac{m}{n^2}\log[P_M(\phi_n^{\mathsf{W}})] \leq \tfrac{1}{2}\theta[-4\varepsilon^2(1-\theta) + \theta c_1^2](1-\delta) + \frac{m\tau_n}{n^2} + O(1).$$

Taking $\theta = (4\varepsilon^2(1-\delta) - 2\tau)/[(8\varepsilon^2 + 2c_1^2)(1-\delta)]$, and taking the limit on both

sides gives

$$J_M(\phi^{\mathsf{W}}) \geq \frac{1}{4}4\varepsilon^2 \frac{4\varepsilon^2(1-\delta) - 2\tau}{(8\varepsilon^2 + 2c_1^2)(1-\delta)}.$$

Since this holds for all $\delta > 0$, and $2\tau < 4\varepsilon^2$, we conclude that

$$J_M(\phi^{\mathsf{W}}) \geq \bar{J}_M(\tau) := \frac{1}{4}4\varepsilon^2 \frac{2\varepsilon^2 - \tau}{(8\varepsilon^2 + 2c_1^2)(\frac{1}{2} + \tau/(4\varepsilon^2))} > 0.$$

□

A.7 Proof of Theorem 3.3

We first give an outline of the proof: Consider any $\tau \in [0, \underline{\kappa}(\varepsilon)]$. Given $\delta > 0$, a sequence of events $\{B_{n,\tau,\delta}\}_{n\geq 1}$ is constructed so that the following is satisfied:

(i) The probability of the event is close to the probability of false alarm:

$$\limsup_{n\to\infty} -\frac{m}{n^2} \log(\mathsf{P}_\pi(B_{n,\tau,\delta})) \leq J_F^*(\tau) - \delta. \tag{A.36}$$

(ii) For any $\boldsymbol{z}_1^n$ satisfying $\{\boldsymbol{Z}_1^n = \boldsymbol{z}_1^n\} \subseteq B_{n,\tau,\delta}$, the following uniform bound on the likelihood ratio holds:

$$\sup_{\mu\in\Pi_m} \frac{\mu^n}{\pi^n}(\boldsymbol{z}_1^n) \geq \exp\{-\frac{n^2}{m}(J_M^*(\tau) - J_F^*(\tau) + \delta)\}. \tag{A.37}$$

. The upper-bound on J_M is then obtained from the following inequality:

$$\begin{aligned} & P_M(\phi_n) \\ & \geq \sup_{\mu\in\Pi_m} \mathsf{P}_\mu\big(\{\phi_n = 0\} \cap B_{n,\tau,\delta}\big) \\ & \geq \sup_{\mu\in\Pi_m} \frac{\mu^n}{\pi^n}(\{\phi_n = 0\} \cap B_{n,\tau,\delta}) \mathsf{P}_\pi(\phi_n = 0 | B_{n,\tau,\delta}) \mathsf{P}_\pi(B_{n,\tau,\delta}). \end{aligned} \tag{A.38}$$

The first term on the right-hand side is lower-bounded in (A.37). The second term can be shown to satisfy $\mathsf{P}_\pi(\phi_n = 0 | B_{n,\tau,\delta}) = 1 - o(1)$ by combining the inequality

(A.36) together with the assumption $J_F(\phi) \geq J_F^*(\tau)$:

$$\begin{aligned} J_F^*(\tau) &\leq \limsup_{n\to\infty} -\frac{m}{n^2} \log\big(\mathsf{P}_\pi(\phi_n = 1)\big) \\ &\leq \limsup_{n\to\infty} -\frac{m}{n^2} \log\big(\mathsf{P}_\pi(\phi_n = 1|B_{n,\tau,\delta})\mathsf{P}_\pi(B_{n,\tau,\delta})\big) \\ &\leq J_F^*(\tau) - \delta + \limsup_{n\to\infty} -\frac{m}{n^2} \log\big(\mathsf{P}_\pi(\phi_n = 1|B_{n,\tau,\delta})\big). \end{aligned} \tag{A.39}$$

The technique of using uniform lower-bounds on likelihood ratio (LR) to construct lower-bounds of probability of missed detection has been applied in [14, 8]: In this prior work, a uniform bound on LR is obtained *over all possible* $\boldsymbol{z}_1^n$. To prove the tight hardness result as in Theorem 3.3, we require the bound on LR to hold uniformly for the sequences in the event B_n, instead of all sequences. This gives us the freedom to optimize B_n to obtain the tightest bound.

The technique to prove (A.37) has been previously used in providing hardness results for composite and hypothesis testing problems [14, 8, 61]: First, construct a collection of distributions so that for each distribution μ, the likelihood ratio μ/π has a simple expression. Second, show that for all observations $\boldsymbol{z}_1^n := \{z_1, \ldots, z_n\}$ in the event B_n, the *average* of $\mathsf{P}_\mu\{\boldsymbol{Z}_1^n = \boldsymbol{z}_1^n\}/\mathsf{P}_\pi\{\boldsymbol{Z}_1^n = \boldsymbol{z}_1^n\}$ over the collection of distributions can be lower-bounded, which in turn lower-bounds the left-hand side of (A.37). The proof for $\varepsilon < 0.5$ and $\varepsilon \geq 0.5$ uses different constructions of distributions.

Proof of Theorem 3.3. Define the event

$$\begin{aligned} B_{n,\tau,\delta} = \Big\{ &\sum_{j=1}^m \mathbb{I}\{n\Gamma_j^n = 1\} \geq n - (1+\tau+\delta)\frac{n^2}{m}, \\ &\sum_{j=1}^m \mathbb{I}\{n\Gamma_j^n = 2\} \geq \tfrac{1}{2}(1+\tau-\delta)\frac{n^2}{m}\Big\}. \end{aligned} \tag{A.40}$$

The probability of the event $B_{n,\tau,\delta}$ has the following asymptotic approximation:

Lemma A.11. *For $\tau = 0$ and any $\delta > 0$,*

$$\lim_{n\to\infty} \mathsf{P}_\pi(B_{n,\tau,\delta}) = 1. \tag{A.41}$$

For any τ, δ satisfying $\tau > \delta > 0$,

$$\lim_{n\to\infty} -\frac{m}{n^2} \log \mathsf{P}_\pi(B_{n,\tau,\delta}) = J_F^*(\tau - \delta). \tag{A.42}$$

The proofs of all lemmas in this section are given in Appendix A.8.

Returning to the proof of the theorem, consider first the case $\tau > 0$. Consider any $\delta \in (0, \tau)$, and any test ϕ such that $J_F(\phi) \geq J_F^*(\tau)$. Using an argument similar to (A.39), we obtain from $P_F(\phi_n) \geq \mathsf{P}_\pi(B_{n,\tau,\delta})\mathsf{P}_\pi(\phi_n = 1|B_{n,\tau,\delta})$ and (A.42) that

$$\mathsf{P}_\pi(\phi_n = 0|B_{n,\tau,\delta}) = 1 - o(1). \tag{A.43}$$

When $\varepsilon < 0.5$, we use the following construction of distributions: Let U_m denote the collection of all subsets of $[m]$ whose cardinality is $\lfloor m/2 \rfloor$. For each set $\mathcal{U} \in U_m$, define the distribution $\mu_\mathcal{U}$ as

$$\mu_{\mathcal{U},j} = \begin{cases} \frac{1}{m} + \frac{\varepsilon}{\lfloor m/2 \rfloor}, & j \in \mathcal{U}; \\ \frac{1}{m} - \frac{\varepsilon}{\lceil m/2 \rceil}, & j \in [m] \setminus \mathcal{U}. \end{cases}$$

This collection of distributions can be obtained by taking the "dominating" distribution μ^* given in (3.13), and permuting the symbols in the alphabet $[m]$.

Let $\mu_\mathcal{U}^n$ be the n-order product of $\mu_\mathcal{U}$. Define the following mixture distribution:

$$\bar{\mu}^n = \frac{1}{|U_m|} \sum_{\mathcal{U} \in U_m} \mu_\mathcal{U}^n.$$

The LR $\bar{\mu}^n / \pi^n$ can be lower-bounded on $B_{n,\tau,\delta}$:

Lemma A.12. *Suppose $\varepsilon < 0.5$. The following holds for any sequence $\boldsymbol{z}_1^n$ satisfying $\{\boldsymbol{Z}_1^n = \boldsymbol{z}_1^n\} \subseteq B_{n,\tau,\delta}$:*

$$\begin{aligned} &\log\big(\frac{\bar{\mu}^n}{\pi^n}(\boldsymbol{z}_1^n)\big) \\ \geq &-\frac{n^2}{2m}[\underline{\kappa}(\varepsilon) - \log(1 + \underline{\kappa}(\varepsilon))(1 + \tau - \delta)](1 + o(1)) - \frac{n^2}{m} 2\delta \log(1 - 2\varepsilon). \end{aligned}$$

It follows from (A.38) that the probability of missed detection admits the bound,

$$P_M(\phi_n) \geq \frac{\bar{\mu}^n}{\pi^n}(\{\phi_n = 0\} \cap B_{n,\tau,\delta})\mathsf{P}_\pi(\phi_n = 0|B_{n,\tau,\delta})\mathsf{P}_\pi(B_{n,\tau,\delta}). \tag{A.44}$$

Applying (A.43), (A.42), and Lemma A.12 gives a bound on the generalized error

exponent,

$$\begin{aligned} J_M(\phi) \leq & \tfrac{1}{2}[\underline{\kappa}(\varepsilon) - \log(1+\underline{\kappa}(\varepsilon))(1+\tau-\delta) + 4\delta\log(1-2\varepsilon)] + J_F^*(\tau-\delta) \\ = & J_M^*(\tau-\delta) + r_1(\delta), \end{aligned} \tag{A.45}$$

where

$$\begin{aligned} r_1(\delta) = & \tfrac{1}{2}[-\delta\log(1+\kappa(\varepsilon)) + (1+\tau)\log(1-\frac{\delta}{1+\tau}) \\ & - \delta\log(1+\tau-\delta) + \delta + 4\delta\log(1-2\varepsilon)]. \end{aligned}$$

In the derivations above, we have used the following explicit expressions of J_F^* and J_M^*, obtained from solving the optimization problems in their definitions in (3.8):

$$\begin{aligned} J_F^*(\tau) &= \tfrac{1}{2}[-\tau + (1+\tau)\log(1+\tau)], \\ J_M^*(\tau) &= \tfrac{1}{2}[\underline{\kappa}(\varepsilon) - \tau + (1+\tau)\log(\frac{1+\tau}{1+\underline{\kappa}(\varepsilon)})]. \end{aligned}$$

Since the inequality (A.45) holds for any $\delta > 0$, $J_M^*(\tau)$ is continuous in τ, and $r_1(\delta) \to 0$ as $\delta \to 0$, we conclude that $J_M(\phi) \leq J_M^*(\tau)$.

When $\varepsilon \geq 0.5$, we use the following construction of distributions: Let U_m denote the collection of all subsets of $[m]$ whose cardinality is $\lfloor m(1-\varepsilon)\rfloor$. For each $\mathcal{U} \in U_m$, define the distribution

$$\mu_{\mathcal{U},j} = \begin{cases} \frac{1}{\lfloor m(1-\varepsilon)\rfloor}, & j \in \mathcal{U}; \\ 0, & j \in [m] \setminus \mathcal{U}. \end{cases}$$

Consider the mixture $\bar{\mu}^n = \frac{1}{|U_m|}\sum_{\mathcal{U}\in U_m}\mu_{\mathcal{U}}^n$. The following bound on $\bar{\mu}^n/\pi^n$ holds:

Lemma A.13. *Suppose $\varepsilon \geq 0.5$. For any sequence $\boldsymbol{z}_1^n = \{z_1, \ldots, z_n\}$ satisfying $\{\boldsymbol{Z}_1^n = \boldsymbol{z}_1^n\} \subseteq B_{n,\tau,\delta}$, the following holds:*

$$\log\big(\frac{\bar{\mu}^n}{\pi^n}(\boldsymbol{z}_1^n)\big) \geq -\tfrac{1}{2}\frac{n^2}{m}[\underline{\kappa}(\varepsilon) - \log(1+\underline{\kappa}(\varepsilon))(1+\tau-\delta)] + O(\frac{n^3}{m^2}).$$

Again combining (A.44) with (A.43), (A.42), and Lemma A.13, we obtain

$$\begin{aligned} J_M(\phi) &\leq \tfrac{1}{2}[\underline{\kappa}(\varepsilon) - \log(1+\underline{\kappa}(\varepsilon))(1+\tau-\delta)] + J_F^*(\tau-\delta) \\ &= J_M^*(\tau-\delta) + r_2(\delta), \end{aligned} \tag{A.46}$$

where r_2 again vanishes as $\delta \to 0$,

$$r_2(\delta) = \tfrac{1}{2}[-\delta \log(1+\kappa(\varepsilon)) + (1+\tau)\log(1 - \frac{\delta}{1+\tau}) - \delta \log(1+\tau-\delta) + \delta].$$

Since (A.46) holds for any $\delta > 0$ and $J_M^*(\tau)$ is continuous, we have $J_M(\phi) \leq J_M^*(\tau)$.

The proof for the case where $\tau = 0$ is exactly the same as that for the case $\tau > 0$, except (A.41) is used in place of (A.42). We omit the details. □

A.8 Proof of Lemma A.11, Lemma A.12, and Lemma A.13

Proof of Lemma A.11. First consider the case where $\tau = 0$. Applying Theorem 3.2 with τ replaced by δ gives

$$\mathsf{P}_\pi\Big\{ \sum_{j=1}^{m} \mathbb{I}\{n\Gamma_j^n = 1\} \leq n - (1+\delta)\frac{n^2}{m}\Big\} = 1 - o(1). \tag{A.47}$$

The following asymptotic approximations on the expectation and variance of the statistic $\sum_{j=1}^{m} \mathbb{I}\{n\Gamma_j^n = 2\}$ hold:

Lemma A.14.

$$\mathsf{E}_\pi[\sum_{j=1}^{m} \mathbb{I}\{n\Gamma_j^n = 2\}] = \tfrac{1}{2}\frac{n^2}{m}(1+o(1)),$$

$$\mathrm{var}_\pi[\sum_{\mathrm{j}=1}^{\mathrm{m}} \mathbb{I}\{\mathrm{n}\Gamma_{\mathrm{j}}^{\mathrm{n}} = 2\}] = \tfrac{1}{2}\frac{n^2}{m}(1+o(1)).$$

The proof of Lemma A.14 is given in Appendix A.10.

Applying Chebyshev's inequality with the above lemma gives

$$\mathsf{P}_\pi\Big\{\sum_{j=1}^{m} \mathbb{I}\{n\Gamma_j^n = 2\} \leq \tfrac{1}{2}\frac{n^2}{m}(1-\delta)\Big\} = O(\frac{m}{n^2}).$$

The claim of this lemma follows from combining this inequality with (A.47).

Next consider the case where $\tau > 0$. Applying Theorem 3.2 with τ replaced

by $\tau+\delta$, we obtain

$$\lim_{n\to\infty} -\frac{m}{n^2}\log \mathsf{P}_\pi\Big\{\sum_{j=1}^{m}\mathbb{I}\{n\Gamma_j^n=1\}\le n-(1+\tau+\delta)\frac{n^2}{m}\Big\}=J_F^*(\tau+\delta).$$

A large-deviation characterization of $\sum_{j=1}^{m}\mathbb{I}\{n\Gamma_j^n=2\}$ similar to that in Theorem 3.2 can be established:

$$\lim_{n\to\infty} -\frac{m}{n^2}\log \mathsf{P}_\pi\Big\{\sum_{j=1}^{m}\mathbb{I}\{n\Gamma_j^n=2\}\ge \tfrac{1}{2}(1+\tau-\delta)\frac{n^2}{m}\Big\}=J_F^*(\tau-\delta).$$

The proof is similar to that for Theorem 3.2 and is omitted. Note that $J_F^*(\tau+\delta)>J_F^*(\tau-\delta)$. Thus the probability that the first constraint in the definition of $B_{n,\tau,\delta}$ is violated is negligible compared to the probability that the second constraint is satisfied. This leads to the claim of the lemma.

□

Proof of Lemma A.12. For simplicity of exposition we restrict to the case where m is even. Extending the result to the case where m is odd is straightforward. Define

$$\mathcal{S}_1:=\{j\colon j \text{ appears in } \boldsymbol{z}_1^n \text{ exactly } \textit{once}\},$$
$$\mathcal{S}_2:=\{j\colon j \text{ appears in } \boldsymbol{z}_1^n \text{ exactly } \textit{twice}\}.$$

Let $s_1=|\mathcal{S}_1|$, $s_2=|\mathcal{S}_2|$. It follows from $\{\boldsymbol{Z}_1^n=\boldsymbol{z}_1^n\}\subseteq B_{n,\tau,\delta}$ that

$$n\ge s_1\ge n-\frac{n^2}{m}(1+\tau+\delta),\quad s_2\ge \tfrac{1}{2}\frac{n^2}{m}(1+\tau-\delta). \tag{A.48}$$

Consider any set $\mathcal{U}\in U_m$. Let $k_{\mathcal{U},1}=|\mathcal{U}\cap\mathcal{S}_1|$, and $k_{\mathcal{U},2}=|\mathcal{U}\cap\mathcal{S}_2|$. Then

$$\frac{\mu_{\mathcal{U}}^n}{\pi^n}(\boldsymbol{z}_1^n)\ge(1-2\varepsilon)^n\Big(\frac{1+2\varepsilon}{1-2\varepsilon}\Big)^{k_{\mathcal{U},1}+2k_{\mathcal{U},2}}.$$

Consequently,

$$\begin{aligned}\frac{\bar{\mu}^n}{\pi^n}(\boldsymbol{z}_1^n)\ge\frac{1}{|U_m|}\sum_{k_1=1}^{s_1}\sum_{k_2=1}^{s_2}(1-2\varepsilon)^n\Big(\frac{1+2\varepsilon}{1-2\varepsilon}\Big)^{k}\Big(\frac{1+2\varepsilon}{1-2\varepsilon}\Big)^{2k_2}\\ \times|\{\mathcal{U}\in U_m: k_{\mathcal{U},1}=k_1,k_{\mathcal{U},2}=k_2\}|,\end{aligned} \tag{A.49}$$

where $|U_m| = \binom{m}{m/2}$ and

$$|\{\mathcal{U} \in U_m : k_{\mathcal{U},1} = k_1, k_{\mathcal{U},2} = k_2\}| = \binom{s_1}{k_1}\binom{s_2}{k_2}\binom{m-(s_1+s_2)}{m/2-(k_1+k_2)}.$$

The summand on the right-hand side of (A.49) takes its maximum value approximately when $k_1 = \bar{k}_1 := \lceil \frac{1+2\varepsilon}{2} s_1 \rceil$ and $k_2 = \bar{k}_2 := \lceil \frac{1}{2}(1 + \frac{4\varepsilon}{1+4\varepsilon^2}) \rceil$. Applying the Laplace method to approximate the summation leads to the following result:

Lemma A.15.

$$\begin{aligned}\frac{\bar{\mu}^n}{\pi^n}(\boldsymbol{z}_1^n) \geq & e^{O(1)+O(\frac{n^{3/2}}{m})}\sqrt{s_1 s_2}(1-2\varepsilon)^n \left(\frac{1+2\varepsilon}{1-2\varepsilon}\right)^{\bar{k}_1}\left(\frac{1+2\varepsilon}{1-2\varepsilon}\right)^{2\bar{k}_2} \\ & \times \binom{s_1}{\bar{k}_1}\binom{s_2}{\bar{k}_2}\binom{m-(s_1+s_2)}{m/2-(\bar{k}_1+\bar{k}_2)} \Big/ \binom{m}{m/2}.\end{aligned} \tag{A.50}$$

Lemma A.15 is proved at the end of this section.

Stirling's formula gives the following asymptotic approximations to the right-hand side of (A.50):

$$\begin{aligned}
&\binom{s_1}{\bar{k}_1} = \frac{(1+2\varepsilon)^{-\bar{k}_1}(1-2\varepsilon)^{\bar{k}_1 - s_1} 2^{s_1}}{\sqrt{2\pi \bar{k}_1 (s_1 - \bar{k}_1)/s_1}}(1+o(1)), \\
&\binom{s_2}{\bar{k}_2} = \frac{(1+2\varepsilon)^{-2\bar{k}_2}(1-2\varepsilon)^{2(\bar{k}_2 - s_2)}(1+4\varepsilon^2)^{s_2} 2^{s_2}}{\sqrt{2\pi \bar{k}_2 (s_2 - \bar{k}_2)/s_2}}(1+o(1)), \\
&\binom{m-(s_1+s_2)}{m/2-(\bar{k}_1+\bar{k}_2)} = 2^{m-s_1-s_2}\exp\{-\frac{s_1^2(2\varepsilon)^2}{2m}(1+o(1))\}\frac{\sqrt{2}}{\sqrt{\pi m}}(1+o(1)), \\
&\binom{m}{m/2} = \frac{2^m}{\sqrt{2\pi m}}(1+o(1)).
\end{aligned}$$

Substituting the above approximations and the value of $\bar{k}_1$ and $\bar{k}_2$ into (A.50) leads to

$$\frac{\bar{\mu}^n}{\pi^n}(\boldsymbol{z}_1^n) \geq (1-2\varepsilon)^{n-s_1-2s_2}\exp\{-\frac{s_1^2(2\varepsilon)^2}{2m}(1+o(1)) + s_2\log(1+4\varepsilon^2) + O(1) + O(\frac{n^{3/2}}{m})\}.$$

The claim of the lemma follows from substituting (A.48) into the above inequality and using the fact that $\underline{\kappa}(\varepsilon) = 4\varepsilon^2$ when $\varepsilon < 0.5$. □

Proof of Lemma A.13. Let $\mathcal{S} := \{j : j \text{ appears in } \boldsymbol{z}_1^n\}$. Let $s = |\mathcal{S}|$. It follows

from $\{\boldsymbol{Z}_1^n = \boldsymbol{z}_1^n\} \subseteq B_{n,\tau,\delta}$ that

$$n - \tfrac{1}{2}\frac{n^2}{m}(1+\tau+3\delta) \le s \le n - \tfrac{1}{2}\frac{n^2}{m}(1+\tau-\delta). \tag{A.51}$$

The likelihood ratio $\frac{\mu_{\mathcal{U}}^n}{\pi^n}$ has the expression: $\frac{\mu_{\mathcal{U}}^n}{\pi^n}(\boldsymbol{z}_1^n) = (\frac{m}{\lfloor m(1-\varepsilon)\rfloor})^n \mathbb{I}_{\mathcal{S}\subseteq\mathcal{U}}$. Thus,

$$\frac{\bar{\mu}^n}{\pi^n}(\boldsymbol{z}_1^n) = (\frac{m}{\lfloor m(1-\varepsilon)\rfloor})^n (\frac{1}{|U_m|}\sum_{\mathcal{U}\in U_m}\mathbb{I}_{\mathcal{S}\subseteq\mathcal{U}}), \tag{A.52}$$

where

$$\frac{1}{|U_m|}\sum_{\mathcal{U}\in U_m}\mathbb{I}_{\mathcal{S}\subseteq\mathcal{U}} = \frac{\binom{m-s}{\lfloor m(1-\varepsilon)\rfloor - s}}{\binom{m}{\lfloor m(1-\varepsilon)\rfloor}}.$$

Stirling's formula gives

$$\binom{m-s}{\lfloor m(1-\varepsilon)\rfloor - s} = (\frac{\lfloor m(1-\varepsilon)\rfloor}{m})^s \exp\{-\tfrac{1}{2}\frac{s^2}{m}\frac{\varepsilon}{1-\varepsilon} + O(\frac{k^3}{m^2})\}(1+O(\frac{1}{m})).$$

Substituting this into (A.52) leads to

$$\frac{\bar{\mu}^n}{\pi^n}(\boldsymbol{z}_1^n) = (1-\varepsilon)^s \exp\{-\tfrac{1}{2}\frac{s^2}{m}\frac{\varepsilon}{1-\varepsilon} + O(\frac{n^3}{m^2})\}(1+O(\frac{n}{m})).$$

The claim of this lemma follows from applying the inequality (A.51) and the fact that $\underline{\kappa}(\varepsilon) = \frac{\varepsilon}{1-\varepsilon}$ when $\varepsilon \ge 0.5$. □

Proof of Lemma A.15. We only outline the main steps. Denote

$$\begin{aligned} y(\Delta_1,\Delta_2) = (\frac{1+2\varepsilon}{1-2\varepsilon})^{\bar{k}_1+\Delta_1+2(\bar{k}_2+\Delta_2)} &\binom{s_1}{\bar{k}_1+\Delta_1}\binom{s_2}{\bar{k}_2+\Delta_2} \\ &\times \binom{m-(s_1+s_2)}{m/2-(\bar{k}_1+\Delta_1+\bar{k}_2+\Delta_2)}/\binom{m}{m/2}. \end{aligned}$$

It is straightforward to show that

$$\binom{m-(s_1+s_2)}{\frac{m}{2}-(\bar{k}_1+\Delta_1+\bar{k}_2+\Delta_2)}/\binom{m-(s_1+s_2)}{\frac{m}{2}-(\bar{k}_1+\bar{k}_2)} = \exp\{1+O(\frac{(\Delta_1+\Delta_2)(\bar{k}_1+\bar{k}_2)}{m}) + o(1)\}.$$

Let $y_1(\Delta_1) = (\frac{1+2\varepsilon}{1-2\varepsilon})^{\Delta_1}\binom{s_1}{\bar{k}_1+\Delta_1}/\binom{s_1}{\bar{k}_1}$ and $y_2(\Delta_2) = (\frac{1+2\varepsilon}{1-2\varepsilon})^{2\Delta_2}\binom{s_2}{\bar{k}_2+\Delta_2}/\binom{s_2}{\bar{k}_2}$. Note that $y(\bar{k}_1,\bar{k}_2)$ is the largest summand. Keeping only the $\lceil\sqrt{s_1}\rceil\lceil\sqrt{s_2}\rceil$ number of

terms in (A.49) around $\bar{k}_1, \bar{k}_2$, and applying the above approximation, we obtain

$$\begin{aligned}\frac{\bar{\mu}^n}{\pi^n}(\boldsymbol{z}_1^n) &\geq \sum_{\Delta_1=-\lceil\sqrt{s_1}\rceil}^{\lceil\sqrt{s_1}\rceil}\sum_{\Delta_2=-\lceil\sqrt{s_2}\rceil}^{\lceil\sqrt{s_2}\rceil} y(\Delta_1,\Delta_2) \\ &= \Big(\sum_{\Delta_1=-\lceil\sqrt{s_1}\rceil}^{\lceil\sqrt{s_1}\rceil} y_1(\Delta_1)\Big)\Big(\sum_{\Delta_2=-\lceil\sqrt{s_2}\rceil}^{\lceil\sqrt{s_2}\rceil} y_2(\Delta_2)\Big) y(0,0)\exp\{1+O(\frac{n^{\frac{3}{2}}}{m})\}.\end{aligned} \tag{A.53}$$

First estimate $\sum_{\Delta_1=-\lceil\sqrt{s_1}\rceil}^{\lceil\sqrt{s_1}\rceil} y_1(\Delta_1)$. Not that for $\Delta_1 > 0$,

$$\log(y_1(\Delta_1)) = \Delta_1 \log(\frac{1+2\varepsilon}{1-2\varepsilon}) + \sum_{t=1}^{\Delta_1}\log(\frac{s-\bar{k}_1-t}{\bar{k}_1+t}).$$

Approximating the above summation by integrals leads to

$$\log(y_1(\Delta_1)) = -\tfrac{1}{2}(\frac{1}{s_1-\bar{k}_1}+\frac{1}{\bar{k}_1})\Delta_1^2(1+o(1))+O(1).$$

Approximating the summation over Δ_1 using integrals, and applying the above approximation of $y_1(\Delta_1)$ leads to

$$\sum_{\Delta_1=-\lceil\sqrt{s_1}\rceil}^{\lceil\sqrt{s_1}\rceil} y_1(\Delta_1) = e^{O(1)}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(\frac{1}{s_1-\bar{k}_1}+\frac{1}{\bar{k}_1})\Delta_1^2} d\Delta_1 = e^{O(1)}\sqrt{\frac{(s_1-\bar{k}_1)\bar{k}_1}{s_1}} = e^{O(1)}\sqrt{s_1}.$$

A similar approximation for the summation over y_2 holds: $\sum_{\Delta_2=-\lceil\sqrt{s_2}\rceil}^{\lceil\sqrt{s_2}\rceil} y_2(\Delta_2) = e^{O(1)}\sqrt{s_2}$. The claim of this lemma follows from substituting these two approximations into (A.53). □

A.9 Proof of Lemma 3.14, Lemma 3.15 and Lemma 3.16

Proof of Lemma 3.14. Applying Lemma 3.12 to the distribution $\mu^* \in \Pi_m$ given in (3.14) and (3.13) gives $\mathsf{E}_{\mu^*}[S_n^{\mathsf{P}}] = \mathsf{E}_\pi[S_n^{\mathsf{P}}] + \frac{n^2}{m}\kappa(\varepsilon)(1+o(1))$. It follows from Chebyshev's inequality that for $\tau_n > \mathsf{E}_\pi[S_n^{\mathsf{P}}] + \frac{n^2}{m}\underline{\kappa}(\varepsilon)$,

$$\mathsf{P}_\mu^*\{\phi_n^{\mathsf{P}}(\boldsymbol{Z}_1^n)=1\} \leq \frac{\mathrm{var}_{\mu^*}[\mathrm{S}_\mathrm{n}^{\mathsf{P}}]}{(\tau_n - \mathsf{E}_\pi[S_n^{\mathsf{P}}] - \frac{n^2}{m}\underline{\kappa}(\varepsilon))^2}.$$

Thus, in order for $\lim_{n\to\infty} \mathsf{P}^*_\mu\{\phi^{\mathsf{P}}_n(\boldsymbol{Z}^n_1) = 0\} = 0$ to hold, we must have

$$(\tau_n - \mathsf{E}_\pi[S^{\mathsf{P}}_n] - \frac{n^2}{m}\underline{\kappa}(\varepsilon))^2 \le \mathrm{var}_{\mu^*}[\mathrm{S}^{\mathsf{P}}_{\mathrm{n}}](1+\mathrm{o}(1)) = 2\frac{\mathrm{n}^2}{\mathrm{m}}(1+\underline{\kappa}(\varepsilon))(1+\mathrm{o}(1)),$$

where the last equality follows from Lemma 3.12. This leads to the claim of Lemma 3.14. □

Proof of Lemma 3.15. Consider the statistic

$$\bar{S^{\mathsf{P}}_n} = S^{\mathsf{P}}_n - \frac{n}{m}\frac{(n\Gamma^n_1 - n\pi_1)^2}{n\pi_1} = S^{\mathsf{P}}_n - 2\frac{n^2}{m}\underline{\kappa}(\varepsilon) + O(\frac{n}{\sqrt{m}}).$$

The conditional distribution of $\bar{S^{\mathsf{P}}_n}$ in the event A under π is the same as the distribution of $\chi^2_{n'}$ under π', where the number of samples is $n' = n - \lfloor \frac{n\sqrt{2\underline{\kappa}(\varepsilon)}}{\sqrt{m}} \rfloor$ and π' is the uniform distribution over $[m-1]$. It then follows from Lemma 3.12 that

$$\mathsf{E}_\pi[\bar{S^{\mathsf{P}}_n}|A] = \mathsf{E}_{\pi'}[\chi^2_{n'}] = n - \lfloor \frac{n\sqrt{2\underline{\kappa}(\varepsilon)}}{\sqrt{m}} \rfloor + O(\frac{n^2}{m}),$$
$$\mathrm{var}_\pi[\bar{\mathrm{S}^{\mathsf{P}}_{\mathrm{n}}}|\mathrm{A}] = \mathrm{var}_{\pi'}[\chi^2_{\mathrm{n}'}] = 2\frac{\mathrm{n}^2}{\mathrm{m}}(1+\mathrm{o}(1)).$$

It follows from Chebyshev's inequality and Lemma 3.12 that for large enough n,

$$\begin{aligned}
&\mathsf{P}_\pi\{S^{\mathsf{P}}_n \le \mathsf{E}_\pi[S^{\mathsf{P}}_n] + \frac{n^2}{m}\underline{\kappa}(\varepsilon) + 2\frac{n}{\sqrt{m}}|A_n\} \\
&= \mathsf{P}_\pi\{\bar{S^{\mathsf{P}}_n} + 2\frac{n^2}{m}\underline{\kappa}(\varepsilon) \le n + \frac{n^2}{m}\underline{\kappa}(\varepsilon) + 2\frac{n}{\sqrt{m}} + O(\frac{n}{\sqrt{m}})|A_n\} \\
&= \mathsf{P}_\pi\{\bar{S^{\mathsf{P}}_n} \le \mathsf{E}_\pi[\bar{S^{\mathsf{P}}_n}|A] - \frac{n^2}{m}\underline{\kappa}(\varepsilon) + O(\frac{n}{\sqrt{m}})|A_n\} \\
&\le \frac{2\frac{n^2}{m}(1+O(\frac{n}{\sqrt{m}}))}{\left(\frac{n^2}{m}\underline{\kappa}(\varepsilon) + O(\frac{n}{\sqrt{m}})\right)^2} = O(\frac{m}{n^2}).
\end{aligned}$$

□

Proof of Lemma 3.16. A simple combinatorial argument gives

$$\mathsf{P}_\pi\{A_n\} = \binom{n}{\lfloor n\frac{\sqrt{2\underline{\kappa}(\varepsilon)}}{\sqrt{m}} \rfloor} \pi_1^{\lfloor n\frac{\sqrt{2\underline{\kappa}(\varepsilon)}}{\sqrt{m}} \rfloor} (1-\pi_1)^{n - \lfloor n\frac{\sqrt{2\underline{\kappa}(\varepsilon)}}{\sqrt{m}} \rfloor}.$$

Applying Stirling's formula and substituting $\pi_1 = \frac{1}{m}$ leads to

$$\mathsf{P}_\pi\{A_n\} = \exp\{-\tfrac{1}{2}\frac{n\sqrt{2\underline{\kappa}(\varepsilon)}}{\sqrt{m}}\log(m)(1+o(1))\}(1+o(1)). \tag{A.54}$$

Since $m = o(\frac{n^2}{\log(n)^2})$ and $m = o(n^2)$, we have

$$\frac{n\sqrt{2\underline{\kappa}(\varepsilon)}}{\sqrt{m}}\log(m) = \frac{n\sqrt{2\underline{\kappa}(\varepsilon)}}{\sqrt{m}}o(2\log(n)) = o(\frac{n^2}{m}).$$

Substitute this into (A.54) leads to the claim of this lemma. □

A.10 Proof of Lemma A.4, Lemma 3.9, Lemma A.14 and Lemma 3.12

Formulas for the expectation and variance of separable statistics under various conditions have been obtained in [6]. Lemma A.4, Lemma 3.9, Lemma A.14, and Lemma 3.12 follow from the following general results:

Lemma A.16 (Equation 2.11 and Equation 2.20 in [6])**.** *Consider a* symmetric *separable statistic* $\sum_{j=1}^m f(n\Gamma_j^n)$. *Suppose* $|f(x)| \le a_0 e^{a_0 x}$ *for some* $a_0 > 0$. *Moreover,* $f(0) = 0$, $f(2) \neq 2f(1)$. *Then, its variance for* $\nu \in \mathcal{P}_m^b$ *is given by*

$$\mathrm{var}_\nu[\sum_{\mathrm{j}=1}^{\mathrm{m}} \mathrm{f}(\mathrm{n}\Gamma_{\mathrm{j}}^{\mathrm{n}})] = \tfrac{1}{2}\frac{\mathrm{n}^2}{\mathrm{m}}(\mathrm{f}(2)-2\mathrm{f}(1))^2(\mathrm{m}\sum_{\mathrm{j}=1}^{\mathrm{m}}\nu_{\mathrm{j}}^2)(1+\mathrm{o}(1)).$$

Lemma A.16 summarizes two results in [6]. Lemma A.17 extend a result in [6]:

Lemma A.17. *Consider the separable statistic* $\sum_{j=1}^m f_j(n\Gamma_j^n)$. *Suppose we have* $\max_j |f_j(x)| \le a_0 e^{a_0 x}$ *for some* $a_0 > 0$. *Then, its expectation for* $\nu \in \mathcal{P}_m^b$ *is:*

$$\begin{aligned}\mathsf{E}_\nu[\sum_{j=1}^m f_j(n\Gamma_j^n)] = \sum_j f_j(0) + n\sum_{j=1}^m \nu_j(f_j(1)-f_j(0)) \\ + \tfrac{1}{2}\frac{n^2}{m}(m\sum_{j=1}^m \nu_j^2)\big(f_j(0)-2f_j(1)+f_j(2)\big) + O(\frac{n^3}{m^2}).\end{aligned}$$

Proof of Lemma A.17. We have $\nu_j^3\binom{n}{3}|f_j(3)| = O(\frac{n^3}{m^3})$ and

$$\sum_{x=4}^{\infty}\nu_j^x\binom{n}{x}|f_j(x)| \le a_0\sum_{x=4}^{\infty}(\frac{e^{a_0}c_1n}{m})^x \le \frac{a_0}{|\log(e^{a_0}c_1n/m)|}(\frac{e^{a_0}c_1n}{m})^3 = O(\frac{n^3}{m^3}).$$

Consequently,

$$\begin{aligned}
&\mathsf{E}_\nu[\sum_{j=1}^{m} f_j(n\Gamma_j^n)] \\
=&\sum_{j=1}^{m}[f_j(0)(1-\nu_j)^n + f_j(1)n\nu_j(1-\nu_j)^{n-1} + f_j(2)\binom{n}{2}\nu_j^2(1-\nu_j)^{n-2} + O(\frac{n^3}{m^3})] \\
=&\sum_j f_j(0) + n\sum_{j=1}^{m}\nu_j(f_j(1)-f_j(0)) + \frac{n^2}{2}\sum_{j=1}^{m}\nu_j^2(f_j(0)-2f_j(1)+f_j(2)) + O(\frac{n^3}{m^2}).
\end{aligned}$$

□

APPENDIX B

PROOF OF RESULTS IN CHAPTER 4

B.1 Proof of Lemma 4.7 and Lemma 4.8

Proof of Lemma 4.7 . We first obtain asymptotic approximation to $\|\Gamma^z - \Gamma^x\|_2^2$ by applying Lemma A.17.

$$\begin{aligned}
\mathsf{E}\|\Gamma^z - \Gamma^x\|_2^2 =& \sum_{j=1}^{m}[\mathsf{P}\{N\Gamma_j^x = 1, n\Gamma_j^z = 0\} + \mathsf{P}\{N\Gamma_j^x = 0, n\Gamma_j^z = 1\}] \\
&+ \sum_{j=1}^{m}[\mathsf{P}\{N\Gamma_j^x = 2, n\Gamma_j^z = 0\} + \mathsf{P}\{N\Gamma_j^x = 0, n\Gamma_j^z = 2\}] + o(\frac{n^2}{m}) \\
=& 2n + \sum_j n^2(q_j - u_j)^2 + O(\frac{n^3}{m^2}) \\
=& 2n + \frac{n^2}{m}\varepsilon^2(1 + O(\frac{n}{m})).
\end{aligned}$$

Similar to the proof of Lemma 3.15, we have that conditioned on C_n,

$$\begin{aligned}
\mathsf{E}[\sum_{j\neq 1}(\Gamma_j^z - \frac{1}{N}a_j^u)^2|C_n] =& 2n - \lfloor\frac{4n}{\sqrt{m}}\rfloor + \sum_{j\neq 1} n^2(u_j - u_j)^2 + O(\frac{n^2}{m^{3/2}}) \\
=& 2n - \lfloor\frac{4n}{\sqrt{m}}\rfloor + O(\frac{n^2}{m^{3/2}}),
\end{aligned}$$

and

$$\mathsf{E}[(\Gamma_1^z - \Gamma_1^y)^2|C_n] = \frac{16n^2}{m}(1 + o(1)).$$

We can show using Lemma A.17 that that $\|\Gamma^z - \Gamma^x\|_2^2$, $\sum_{j\neq 1}(\Gamma_j^z - \Gamma_j^y)^2$ and $(\Gamma_1^z - \Gamma_1^y)^2$ converge to their expectations asymptotically in probability. Therefore,

conditioned on C_n, we have that with probability $1 - o(1)$

$$\begin{aligned}&\|\Gamma^z - \Gamma^x\|_2^2 - \|\Gamma^z - \Gamma^y\|_2^2\\ =&\frac{n^2}{m}\varepsilon^2(1 + O(\frac{n}{m})) - \frac{16n^2}{m}(1 + o(1)) \leq -\frac{n^2}{m}(1 + o(1)).\end{aligned}$$

□

Proof of Lemma 4.8. A simple combinatorial argument gives

$$\mathsf{P}_{(q,u,q)}\{C_n\} = \binom{n}{\lfloor \frac{4n}{\sqrt{m}} \rfloor} u_1^{\lfloor \frac{4n}{\sqrt{m}} \rfloor}(1 - u_1)^{n - \lfloor \frac{4n}{\sqrt{m}} \rfloor}.$$

Applying Stirling's formula and substituting $u_1 = \frac{1}{m}$ leads to

$$\mathsf{P}_{(q,u,q)}\{C_n\} = \exp\{-\tfrac{1}{2}\frac{4n}{\sqrt{m}}\log(m)(1 + o(1))\}(1 + o(1)). \tag{B.1}$$

□

B.2 Proof of Proposition 4.9

The proof is similar to that for Proposition A.1, which gives asymptotic approximations for the logarithmic moment generating function of the coincidence-based test for the universal hypothesis testing problem. It uses the Poissonization technique. The difference is that there are three independent sequences involved in the test statistic for the classification problem instead of one for the universal hypothesis testing problem. Thus, instead of Lemma A.5, the moment generating function of T_n has the following formula:

$$\begin{aligned}&\mathsf{E}_{(\pi,\mu,\nu)}[\exp\{\theta T_n\}]\\ =&\frac{n!}{2\pi i}\frac{N!}{2\pi i}\frac{N!}{2\pi i}\oint_{\lambda_1}\oint_{\lambda_2}\oint_{\lambda_3} e^{\lambda_1}e^{\lambda_2}e^{\lambda_3}\\ &\times\prod_{j=1}^{m}\Big(\sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\sum_{k_3=0}^{\infty}\frac{(\lambda_1\mu_j)^{k_1}}{k_1!}e^{-\lambda_1\mu_j}\frac{(\lambda_2\pi_j)^{k_2}}{k_2!}e^{-\lambda_2\pi_j}\frac{(\lambda_3\nu_j)^{k_3}}{k_3!}e^{-\lambda_3\nu_j}e^{\theta f_j(k_1,k_2,k_3)}\Big)\\ &\frac{d\lambda_1}{\lambda_1^{N+1}}\frac{d\lambda_2}{\lambda_2^{N+1}}\frac{d\lambda_3}{\lambda_3^{n+1}}.\end{aligned}$$

where $f_j(k_1, k_2, k_3)$ is value of the summand in the definition of T_n corresponding to j and $N\Gamma_j^x = k_1$, $N\Gamma_j^y = k_2$, and $n\Gamma_j^z = k_3$. This leads to

$$
\begin{aligned}
&\mathsf{E}_{(\pi,\mu,\nu)}[\exp\{\theta T_n\}] \\
=&\frac{n!}{2\pi i}\frac{N!}{2\pi i}\frac{N!}{2\pi i}\oint_{\lambda_1}\oint_{\lambda_2}\oint_{\lambda_3} e^{\lambda_1}e^{\lambda_2}e^{\lambda_3} \\
&\times\prod_{j=1}^{m}\big(1+\frac{\lambda_1^2\mu_j^2}{2}\exp\{-\lambda_1\mu_j-\lambda_2\pi_j-\lambda_3\nu_j\}(\exp\{\theta\frac{1}{N^2}\}-1) \\
&\qquad+\frac{\lambda_2^2\pi_j^2}{2}\exp\{-\lambda_1\mu_j-\lambda_2\pi_j-\lambda_3\nu_j\}(\exp\{-\theta\frac{1}{N^2}\}-1) \\
&\qquad+\lambda_1\mu_j\lambda_3\nu_j\exp\{-\lambda_1\mu_j-\lambda_2\pi_j-\lambda_3\nu_j\}(\exp\{-\theta\frac{1}{Nn}\}-1) \\
&\qquad+\lambda_2\pi_j\lambda_3\nu_j\exp\{-\lambda_1\mu_j-\lambda_2\pi_j-\lambda_3\nu_j\}(\exp\{\theta\frac{1}{Nn}\}-1)\big) \\
&\frac{d\lambda_1}{\lambda_1^{N+1}}\frac{d\lambda_2}{\lambda_2^{N+1}}\frac{d\lambda_3}{\lambda_3^{n+1}}.
\end{aligned}
\tag{B.2}
$$

We first find the saddle point and carry out the integration of $\lambda_1, \lambda_2, \lambda_3$ around contours that go through the saddle point. It is straightforward to show that the possible integration is around the contour with $|\lambda_1| = N(1+O(\frac{\max\{N,n\}}{m}))$, $|\lambda_2| = N(1+O(\frac{\max\{N,n\}}{m}))$, $|\lambda_3| = n(1+O(\frac{\max\{N,n\}}{m}))$. Then the summation in (B.2) has the following approximation with $\theta = \min\{N^2, Nn\}\gamma$.

$$
\begin{aligned}
&1+\frac{\lambda_1^2\mu_j^2}{2}\exp\{-\lambda_1\mu_j-\lambda_2\pi_j-\lambda_3\nu_j\}(\exp\{\theta\frac{1}{N^2}\}-1) \\
&\qquad+\frac{\lambda_2^2\pi_j^2}{2}\exp\{-\lambda_1\mu_j-\lambda_2\pi_j-\lambda_3\nu_j\}(\exp\{-\theta\frac{1}{N^2}\}-1) \\
&\qquad+\lambda_1\mu_j\lambda_3\nu_j\exp\{-\lambda_1\mu_j-\lambda_2\pi_j-\lambda_3\nu_j\}(\exp\{-\theta\frac{1}{Nn}\}-1) \\
&\qquad+\lambda_2\pi_j\lambda_3\nu_j\exp\{-\lambda_1\mu_j-\lambda_2\pi_j-\lambda_3\nu_j\}(\exp\{\theta\frac{1}{Nn}\}-1) \\
=&1+\frac{\lambda_1^2\mu_j^2}{2}(\exp\{\theta\frac{1}{N^2}\}-1)+\frac{\lambda_2^2\pi_j^2}{2}(\exp\{-\theta\frac{1}{N^2}\}-1) \\
&+\lambda_1\mu_j\lambda_3\nu_j(\exp\{-\theta\frac{1}{Nn}\}-1)+\lambda_2\pi_j\lambda_3\nu_j(\exp\{\theta\frac{1}{Nn}\}-1) \\
&+O(\frac{\max\{N,n\}\min\{N^2,Nn\}}{m^3}).
\end{aligned}
$$

Applying the fact that for $|x| \leq 0.5$, $e^x - 1 \leq x + x^2$. We obtain for $\lambda_1 = N(1 + O(\frac{\max\{N,n\}}{m}))$, $\lambda_2 = N(1 + O(\frac{\max\{N,n\}}{m}))$, $\lambda_3 = n(1 + O(\frac{\max\{N,n\}}{m}))$,

$$
\begin{aligned}
&1 + \frac{\lambda_1^2\mu_j^2}{2}(\exp\{\theta\frac{1}{N^2}\} - 1) + \frac{\lambda_2^2\pi_j^2}{2}(\exp\{-\theta\frac{1}{N^2}\} - 1) \\
&+ \lambda_1\mu_j\lambda_3\nu_j(\exp\{-\theta\frac{1}{Nn}\} - 1) + \lambda_2\pi_j\lambda_3\nu_j(\exp\{\theta\frac{1}{Nn}\} - 1) \\
&+ O(\frac{\max\{N,n\}\min\{N^2,Nn\}}{m^3}) \\
\leq& 1 + \frac{\lambda_1^2\mu_j^2}{2}\frac{\min\{N^2,Nn\}\gamma}{N^2} - \frac{\lambda_2^2\pi_j^2}{2}\frac{\min\{N^2,Nn\}\gamma}{N^2} \\
&- \lambda_1\mu_j\lambda_3\nu_j\frac{\min\{N^2,Nn\}\gamma}{Nn} + \lambda_2\pi_j\lambda_3\nu_j\frac{\min\{N^2,Nn\}\gamma}{Nn} \\
&+ \gamma^2\tfrac{1}{2}(\mu_j^2\frac{\lambda_1^2}{N^2} + \pi_j^2\frac{\lambda_2^2}{N^2})\frac{\min\{N^2,Nn\}^2}{N^2} \\
&+ \gamma^2(\pi_j\nu_j\frac{\lambda_2\lambda_3}{Nn} + \mu_j\nu_j\frac{\lambda_1\lambda_3}{Nn})\frac{\min\{N^2,Nn\}^2}{Nn} \\
&+ O(\frac{\max\{N,n\}\min\{N^2,Nn\}}{m^3}) \\
\leq& 1 + \frac{\mu_j^2}{2}\min\{N^2,Nn\}\gamma - \frac{\pi_j^2}{2}\min\{N^2,Nn\}\gamma \\
&- \mu_j\nu_j\min\{N^2,Nn\}\gamma + \pi_j\nu_j\min\{N^2,Nn\}\gamma \\
&+ \gamma^2\tfrac{1}{2}(\mu_j^2 + \pi_j^2)\min\{N^2,Nn\} + \gamma^2(\pi_j\nu_j + \mu_j\nu_j)\min\{N^2,Nn\} \\
&+ O(\frac{\max\{N,n\}\min\{N^2,Nn\}}{m^3}).
\end{aligned}
$$

The rest of steps are essentially the same as those in the proof of Proposition A.1. We omit the details.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, pp. 15:1 – 15:58, July 2009.

[2] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, pp. 60 – 62, 1938.

[3] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493 – 507, 1952.

[4] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *The Annals of Mathematical Statistics*, vol. 36, pp. 369 – 401, 1965.

[5] R. R. Bahadur and R. R. Rao, "On deviations of the sample mean," *The Annals of Mathematical Statistics*, vol. 31, no. 4, pp. 1015 – 1027, 1960.

[6] Y. I. Medvedev, "Separable statistics in a polynomial scheme. I," *Theory of Probability and its Applications*, vol. 22, no. 1, pp. 1 – 15, 1977.

[7] M. S. Ermakov, "Asymptotic minimaxity of chi-square tests," *Theory of Probability and its Applications*, vol. 42, pp. 589 – 610, 1998.

[8] A. R. Barron, "Uniformly powerful goodness of fit tests," *The Annals of Statistics*, vol. 17, no. 1, pp. 107 – 124, 1989.

[9] H. B. Mann and A. Wald, "On the choice of the number of class intervals in the application of the chi square test," *The Annals of Mathematical Statistics*, vol. 13, no. 3, pp. 306 – 317, 1942.

[10] W. C. M. Kallenberg, "On moderate and large deviations in multinomial distributions," *The Annals of Statistics*, vol. 13, no. 4, pp. 1554 – 1580, 1985.

[11] J. Oosterhoff, "The choice of cells in chi–square tests," *Statistica Neerlandica*, vol. 39, no. 2, pp. 115 – 128, 1985.

[12] A. F. Ronzhin, "A theorem on large-deviation probabilities for decomposable statistics and its statistical application," *Mathematical Notes*, vol. 36, no. 4, pp. 800 – 807, 1984.

[13] A. V. Kolodzei, "A theorem on probabilities of large deviations for decomposable statistics which do not satisfy the Cramér condition," *Discrete Mathematics & Applications*, vol. 15, no. 3, pp. 255 – 262, 2005.

[14] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Transactions on Information Theory*, vol. 54, no. 10, pp. 4750 – 4755, Oct. 2008.

[15] I. Csiszár and G. Longo, "On the error exponent for source coding and for testing simple statistical hypotheses," *Studia Scientiarum Mathematicarum Hungarica*, vol. 6, pp. 181 – 191, 1971.

[16] M. Iltis, "Sharp asymptotics of large deviations in $\mathbb{R}^d$," *Journal of Theoretical Probability*, vol. 8, no. 3, pp. 501–522, 1995.

[17] A. DasGupta, *Asymptotic Theory of Statistics and Probability*. New York, NY, USA: Springer Verlag, 2008.

[18] A. Wald, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, vol. 54, no. 3, pp. 426 – 482, 1943.

[19] H. Chernoff, "On the distribution of the likelihood ratio," *The Annals of Mathematical Statistics*, vol. 25, no. 3, pp. 573 – 578, 1954.

[20] P. Billingsley, *Statistical inference for Markov processes*. Chicago, IL, USA: University of Chicago Press, 1961.

[21] P. Hall, "Chi squared approximations to the distribution of a sum of independent random variables," *The Annals of Probability*, vol. 11, no. 4, pp. 1028 – 1036, 1983.

[22] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Transactions on Information Theory*, vol. 36, no. 3, pp. 453 – 471, May 1990.

[23] S. K. Tumanyan, "Asymptotic distribution of the χ^2 criterion when the number of observations and number of groups increase simultaneously," *Theory of Probability and its Applications*, vol. 1, pp. 117 – 131, 1956.

[24] G. P. Steck, "Limit theorems for conditional distributions," in *University of California Publications in Statistics*. University of California Press, 1957, vol. 2, pp. 237 – 284.

[25] L. Holst, "Asymptotic normality and efficiency for certain goodness-of-fit tests," *Biometrika*, vol. 59, no. 1, pp. 137 – 145, 1972.

[26] C. Morris, "Central limit theorems for multinomial sums," *The Annals of Statistics*, vol. 3, no. 1, pp. 165 – 188, 1975.

[27] M. P. Quine and J. Robinson, "Normal approximations to sums of scores based on occupancy numbers," *The Annals of Probability*, vol. 12, no. 3, pp. 794 – 804, 1984.

[28] V. M. Kruglov, "The asymptotic behavior of the Pearson statistic," *Theory of Probability and its Applications*, vol. 45, pp. 69 – 92, 2001.

[29] G. Tusnady, "On asymptotically optimal tests," *The Annals of Statistics*, vol. 5, no. 2, pp. 385 – 393, 1977.

[30] M. P. Quine and J. Robinson, "Efficiencies of chi-square and likelihood ratio goodness-of-fit tests," *The Annals of Statistics*, vol. 13, no. 2, pp. 727 – 742, 1985.

[31] S. K. Sirazhdinov, S. A. Mirakhmedov, and S. A. Ismatullaev, "Probabilities of large deviations for randomized divisible statistics in the multinomial scheme," *Theory of Probability and its Applications*, vol. 34, no. 4, pp. 645 – 657, 1989.

[32] O. Goldreich and D. Ron, "On test expansion in bounded-degree graphs," *Electronic Colloquium on Computational Complexity*, 2000, TR00-020.

[33] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White, "Testing random variables for independence and identity," in *Proceedings of 42nd IEEE Symposium on Foundations of Computer Science*, Las Vegas, NV, USA, Oct. 2001, pp. 442 – 451.

[34] Y. I. Medvedev, "Separable statistics in a polynomial scheme. II," *Theory of Probability and its Applications*, vol. 22, no. 3, pp. 607 – 615, 1978.

[35] D. Huang and S. Meyn, "Error exponents for composite hypothesis testing with small samples," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, march 2012, pp. 3261 –3264.

[36] T. Zhang and F. Oles, "Text categorization based on regularized linear classification methods," *Information Retrieval*, vol. 4, pp. 5 – 31, 2001.

[37] B. G. Kelly, T. Tularak, A. B. Wagner, and P. Viswanath, "Universal hypothesis testing in the learning-limited regime," in *Proceedings of 2010 IEEE International Symposium on Information Theory*, Austin, TX, June 2010, pp. 1478 – 1482.

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

[39] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401 – 408, Mar. 1989.

[40] D. Huang and S. Meyn, "Classification with high-dimensional sparse samples," in *Proceedings of 2012 IEEE International Symposium on Information Theory*, July 2012, pp. 2586 –2590.

[41] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing that distributions are close," in *Proceedings of 41st Annual Symposium on Foundations of Computer Science*, Redondo Beach, CA, USA, 2000, pp. 259 – 269.

[42] P. Valiant, "Testing symmetric properties of distributions," in *Proceedings of the 40th Annual ACM symposium on Theory of Computing*. New York, NY, USA: ACM, 2008, pp. 383 – 392.

[43] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1469 – 1481, July 2004.

[44] J. Acharya, H. D. A. Jafarpour, A. Orlitsky, and S. Pan, "Competitive closeness testing," in *Proceedings of 24th Annual Conference on Learning Theory*, Budapest, Hungary, June 2011, pp. 47–68.

[45] D. Huang, J. Unnikrishnan, S. Meyn, V. Veeravalli, and A. Surana, "Statistical SVMs for robust detection, supervised learning, and universal classification," in *IEEE Information Theory Workshop on Networking and Information Theory*, June 2009, pp. 62 – 66.

[46] J. Unnikrishnan, H. Dayu, S. P. Meyn, A. Surana, and V. V. Veeravalli, "Universal and composite hypothesis testing via mismatched divergence," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1587 – 1603, Mar. 2011.

[47] D. Huang and S. Meyn, "Feature extraction for universal hypothesis testing via rank-constrained optimization," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, June 2010, pp. 1618 – 1622.

[48] G. Stengle and J. E. Yukich, "Some new Vapnik-Chervonenkis classes," *The Annals of Statistics*, vol. 17, no. 4, pp. 1441 – 1446, 1989.

[49] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, Stanford University, 2002.

[50] M. Fazel, H. Hindi, and S. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the American Control Conference*, vol. 6, 2001, pp. 4734 – 4739.

[51] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 471, 2010.

[52] E. J. Candes and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717 – 772, 2009.

[53] E. J. Candes and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *Information Theory, IEEE Transactions on*, vol. 56, no. 5, pp. 2053 – 2080, 2010.

[54] J. F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, p. 1956, 2010.

[55] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and bregman iterative methods for matrix rank minimization," *Mathematical Programming*, pp. 1 – 33, 2009.

[56] P. L. Combettes and J. C. Pesquet, "Proximal splitting methods in signal processing," *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185 – 212, 2011.

[57] J. Unnikrishnan, "Decision-making under statistical uncertainty," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2010.

[58] A. Frank and A. Asuncion, "UCI Machine Learning Repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[59] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed., ser. Stochastic Modelling and Applied Probability. New York, NY, USA: Springer-Verlag, 1998.

[60] N. G. D. Bruijn, *Asymptotic Methods in Analysis*. New York, NY, USA: Dover Publications, 1981.

[61] B. Kelly, A. Wagner, T. Tularak, and P. Viswanath, "Classification of homogeneous data with large alphabets," *IEEE Transactions on Information Theory*, 2012, to appear.