ROBUST STATISTICAL INFERENCE THROUGH THE LENS OF OPTIMIZATION

A Dissertation
Presented to
The Academic Faculty

By

Liyan Xie

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology

August 2021

ROBUST STATISTICAL INFERENCE THROUGH THE LENS OF OPTIMIZATION

Thesis committee:

Dr. Yao Xie, Advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Jianjun Shi
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. George V. Moustakides
Department of Electrical and Computer Engineering
University of Patras

Dr. Jeff Wu
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Arkadi Nemirovski
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: May 19, 2021

To my parents.

ACKNOWLEDGMENTS

My deepest gratitude goes to my advisor, Prof. Yao Xie, for guiding me to the research field of statistical signal processing, especially change-point detection and its applications in various domains, and for her constant support during my Ph.D. study. I have been very fortunate to work with her on exciting research topics. It is hard to imagine this thesis without her guidance and support. I am constantly amazed by her passion for research and her insights and vision into both practical problems and theoretical results. She can always find the perfect balance between theory and practice, which greatly enriched my research experience during my Ph.D. study. I cannot overstate my appreciation for her constant encouragement whenever I have encountered any difficulty in research and life.

I would like to thank all committee members for their time and effort serving in my thesis committee and for their help in the preparation of this work – Prof. George Moustakides, Prof. Arkadi Nemirovski, Prof. Jeff Wu, and Prof. Jianjun Shi. I would like to thank Prof. George Moustakides for his support and guidance in my research on change-point detection problems and during my job hunting. As an expert in change-point detection, he has taught me how to approach a problem from the theoretical side and also how to implement the algorithms efficiently. Moreover, I am also amazed by his insights and vision into machine learning problems and his innovative ideas of combining machine learning with change-point detection in certain contexts. I would like to thank Prof. Arkadi Nemirovski for his guidance in optimization-related research. He is always willing to explain the research problems and his ideas with great patience whenever I have a question. I also would like to thank him for his support in my fellowship applications and job hunting process. I would like to thank Prof. Jeff Wu for encouraging me a lot to pursue the academic career and for providing me valuable suggestions for my career path. I still remember that I felt much more confident about choosing the academia path after talking to him before my thesis proposal. I would like to thank Prof. Jianjun Shi for his guidance and insights for my

research, especially possible applications of change detection algorithms. After my thesis proposal, Prof. Jianjun Shi has discussed with me a variety of possible improvements and research topics about my research proposal, which has provided me a lot of potential future directions and exciting problems to think about in my future career. I have benefited a lot from my communication with all my thesis committee members, and I am very fortunate to have them on my committee.

I would like to thank the H. Milton Stewart School of Industrial and Systems Engineering. The students, faculties, and staffs here have provided me with a wonderful study and work environment. Many thanks go to Santanu S. Dey, Alan Erera, Amanda Ford, Dawn Strickland, and Sandra Bryant-Turner for the assistance in a lot of administrative procedures. I also want to thank the Transdisciplinary Research Institute for Advancing Data Science (TRIAD) and the Algorithms & Randomness Center at Georgia Tech for providing financial support through the IDEaS-TRIAD and ARC fellowship.

I would like to thank all my colleagues and collaborators who made this work possible. Special thanks to Prof. Rui Gao at UT-Austin for his guidance on distributionally robust optimization ever since when he was still at Georgia Tech. Special thanks to Prof. Wenzhan Song at UGA for his guidance on signal processing for seismic sensors and in health care. Special thanks to all my colleagues in Yao's group, including Yang Cao, Junzhuo Chen, Shuang Li, Haoyun Wang, Jie Wang, Song Wei, Chen Xu, Minghe Zhang, Rui Zhang, Shixiang Zhu, etc. I have either collaborated with them or have learned a lot from them. I would like to thank all my friends, and I'm glad we could accompany and support each other on this journey. Special thanks go to Shanshan Cao, Jialei Chen, Zhehui Chen, Ana María Estrada Gómez, Junqi Hu, Arvind Krishna, Feng Liu, Tyler Perini, Guanyi Wang, Yujia Xie, Chuanping Yu, Ruizhi Zhang, Wanrong Zhang, Yujie Zhao, to name only a few. I would also like to say thanks to Minshuo for his constant love and support.

Last but not least, I want to thank my parents for their unconditional love and care during my pursuit of a Ph.D. To them I dedicate this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Robust statistical inference is an important and fundamental problem in modern data science. Many classical works in sequential analysis are designed for the case when we have full knowledge of the underlying data-generating distributions, such as the well-known Neyman-Pearson lemma for hypothesis testing and the cumulative sum (CUSUM) algorithm for sequential change-point detection. However, there are many cases when we do not have sufficient prior knowledge about the true distributions. In such cases, we need robust statistical methods that can guarantee the worst-case performance. Moreover, we also need algorithms that can be implemented efficiently in the online setting where data comes sequentially.

Such kind of problem is frequently seen and is widely applicable to a variety of applications. For example, in health care applications, we might do not have much information about the anomaly data (such as a new disease), and we would like to develop a method that can detect anomaly pattern quickly from data; In sensor network modeling such as social networks and seismic sensors, the goal is to detect any structural or correlation changes among sensors as quickly as possible.

This thesis tackles the robust statistical inference from three aspects. Chapter 3 and 4 study the sequential change-point detection problem with unknown distributions, from both the parametric side and non-parametric sides. Chapter 5 studies a data-driven setting of the robust hypothesis testing problem when the only information we have is data. Chapter 6 studies the spatio-temporal modeling of event data over networks. Useful preliminaries and important background information are summarized in Chapter 2 and all the proofs are delegated to the Appendices.

CHAPTER 1
INTRODUCTION AND BACKGROUND

Statistical signal processing and hypothesis testing are fundamental problems in modern data science and engineering applications. The development of modern data acquisition enables us to have an increasing amount of data that can be accessed with high speed and high resolution. This has brought new opportunities for us to build more reliable machine learning models to perform estimation, prediction, and decision-making. However, this also poses several new challenges, as explained below.

(1) *High-dimensionality.* Sequential data in modern applications is usually high dimensional. For example, in sensor networks, the Long Beach 3D seismic array consists of approximately 5300 seismic sensors that record data continuously for seismic activity detection and analysis; and in multi-stage manufacturing processes [175], we usually have a huge amount of dependent data sequences as well. We need more efficient online algorithms to deal with a large amount of high-dimensional data and to detect the anomaly pattern in an online manner. Usually, changes in high-dimensional time series exhibit low-dimensional structures in the form of sparsity, low-rankness, and subset structures, which can be exploited to enhance the capability to detect weak signals quickly and efficiently.

(2) *Data uncertainties.* In certain cases, we might have enough training data from one category but very limited data from another category, causing certain biases in our decision-making. For example, in health care applications, we usually have enough data for normal people or known diseases/medicine, but much fewer samples for new diseases/patients. We need robust data-driven hypothesis testing/classification algorithms that can achieve the optimal worst-case performance even when the true data-generating distribution deviates from our estimate.

(3) *Complex data distributions.* Modern sequential data is more involved in nature. It

is much more challenging to come up with a simple parametric form to describe the data distribution, as commonly done in the traditional setting. Therefore, the non-parametric methods have been studied and extended a lot recently. Moreover, in modern applications, sequential data could have complex spatial and temporal dependencies, for instance, induced by the network structure [159, 81, 15]. For example, in social networks, dependencies are usually due to interaction and information diffusion [121]. We need a general modeling framework for spatio-temporal event data and efficient algorithms with strong theoretical guarantees.

This thesis mainly focuses on developing new theories and algorithms for three research problems in the area of statistical inference, aiming to tackle the above challenges. The first problem we study is sequential (quickest) change detection. We consider a subspace change for high-dimensional data sequences, which is a fundamental problem since subspace structure is commonly used for modeling high-dimensional data. We also consider a non-parametric setting that can be useful when the data distributions cannot be represented by simple parametric families, and the weighted ℓ_2 divergence is proposed to detect the change. The second problem we study is data-driven robust hypothesis testing when the true date-generating distributions are all unknown and we only have access to a limited number of training samples. The third problem is parameter recovery for spatio-temporal models, with potential applications in modeling crime events and COVID-19 cases.

The structure of this thesis is organized as follows. Chapter 1 introduces the background and motivation for each topic as explained below. Chapter 2 reviews some preliminary and fundamental results in sequential change detection, distributionally robust optimization, and variational inequalities. Those basics can help readers understand the problem set-up and proof strategies in the Appendices.

In Chapter 3, we consider the online monitoring of multivariate streaming data for changes that are characterized by an unknown subspace structure manifested in the covariance matrix. In particular, we consider the covariance structure changes from an identity

matrix to an unknown spiked covariance model. We assume the post-change distribution is unknown, and propose two detection procedures: the Largest-Eigenvalue Shewhart chart and the Subspace-CUSUM detection procedure. We present theoretical approximations to the average run length and the expected detection delay for the Largest-Eigenvalue Shewhart chart, as well as the asymptotic optimality analysis for the Subspace-CUSUM procedure. The performance of the proposed methods is illustrated using simulation and real data for human gesture detection and seismic event detection.

In Chapter 4, we present a new non-parametric statistic, called the weighed ℓ_2 divergence, based on empirical distributions for sequential change detection. We start by constructing the weighed ℓ_2 divergence as a fundamental building block for two-sample tests and change detection. The proposed statistic is proved to attain the optimal sample complexity in the offline setting. We then study the sequential change detection using the weighed ℓ_2 divergence and characterize the fundamental performance metrics, including the average run length and the expected detection delay. We also present practical algorithms to find the optimal projection to handle high-dimensional data and the optimal weights, which is critical to quick detection since, in such settings, there are not many post-change samples. Simulation results and real data examples are provided to validate the good performance of the proposed method.

In Chapter 5, we consider a data-driven robust hypothesis test where the optimal test will minimize the worst-case performance regarding distributions close to the empirical distributions with respect to the Wasserstein distance. This leads to a new non-parametric hypothesis testing framework based on distributionally robust optimization, which is more robust when there are limited samples for one or both hypotheses. Such a scenario often arises from applications such as health care, online change-point detection, and anomaly detection. We study the computational and statistical properties of the proposed test by presenting a tractable convex reformulation of the original infinite-dimensional variational problem exploiting Wasserstein's properties and characterizing the optimal radius for the

uncertainty sets to control the generalization error. We also demonstrate the good performance of our method on synthetic and real data.

In Chapter 6, we introduce a new general modeling approach for multivariate discrete event data with categorical interacting marks, which we refer to as marked Bernoulli processes. In the proposed model, the probability of an event of a specific category to take place in a location may be influenced by past events at this and other locations. We do not restrict interactions to be positive or decaying over time as it is commonly adopted, allowing us to capture an arbitrary shape of influence from historical events, locations, and events of different categories. In our modeling, prior knowledge is incorporated by allowing general convex constraints on model parameters. We develop two parameter estimation procedures utilizing the constrained least square and maximum likelihood estimation, which can be solved as convex problems based on variational inequalities. We discuss different applications of our approach and illustrate the performance of proposed recovery routines on synthetic examples and real-world data.

1.1 Sequential Subspace Change Detection

Detecting the change from high-dimensional streaming data is a fundamental problem in various applications such as video surveillance [191], sensor networks [214], wearable sensors [188], and seismic events detection [122]. In many scenarios, the change happens to the *covariance* structure and can be represented as a *low-rank subspace* to capture the similarity of signal waveforms observed at multiple sensors. In Chapter 3, we consider the fundamental problem of detecting such a change in the covariance matrix that shifts from an identity matrix to a spiked covariance model [97]. Different from the offline hypothesis test considered in [22], we assume a sequential setting, where the goal is to detect such a change as quickly as possible after it occurs.

A formal description of the problem is as follows. Assume a sequence of multivariate observations $x_1, x_2, \ldots, x_t, \ldots$, where $x_t \in \mathbb{R}^k$ and k is the data dimension. At a cer-

tain time τ, the distribution of the observation changes. In particular, we are interested in structural changes that happen to the covariance matrix, which we describe below: (1) the *emerging subspace*, meaning the change is a subspace emerging from a noisy background and thus the covariance matrix changes from an identity matrix to a spiked covariance matrix; (2) the *switching subspace*, meaning that the signals are along with different subspaces before and after the change, resulting the covariance matrix to change from one spiked covariance matrix to another. The emerging subspace problem can arise, for instance, from weak signal detection for seismic sensor arrays [188], and the switching subspace detection can arise from monitoring principal component analysis (PCA) for streaming data [14]. The switching subspace problem, as we will show, can be reduced to the emerging subspace problem; therefore, we focus on the emerging subspace problem.

The main contribution of this work is two-fold. From the methodology perspective, we propose two sequential detection procedures: the Largest-Eigenvalue Shewhart chart and the Subspace-CUSUM procedure. The Largest-Eigenvalue Shewhart chart computes the largest eigenvalue of the sample covariance matrix over a sliding window and detects a change when the statistic exceeds the threshold. The Subspace-CUSUM is derived based on the likelihood ratio following the approach of classical CUSUM [140], but instead of assuming the parameters are fully specified, we estimate the parameters and plug-in, which is analogous to the generalized likelihood ratio (GLR) statistic [108]. From the theoretical perspective, we provide a theoretical analysis of the proposed procedures, which facilitates efficient calibration of the parameters. We consider two commonly used metrics: the *average run length* (ARL) and the *expected detection delay* (EDD). Theoretical approximations can help us determine the threshold in the detection procedure efficiently. Moreover, building on Anderson's results for the distribution of eigenvectors [7], we provide theoretical guidelines on how to choose the parameters involved in the Subspace-CUSUM procedure. Through proper parameter optimization, we prove that the resulted procedure is first-order asymptotically optimal in the sense that the ratio of its expected detection delay with the

corresponding of the optimum CUSUM (that has complete knowledge of the pre- and post-change statistics) tends to one as the average run length tends to infinity.

The proposed detection procedures are computationally efficient since they only require computing the leading eigenvalue and eigenvector of the sample covariance matrix, respectively. They are widely applicable to real data whenever there is a low-rank subspace change. For example, we have demonstrated its use in human activity detection from wearable sensors data and seismic event detection.

Related Work

In change-point detection and industrial quality control, commonly used methods include Shewhart chart, cumulative sum (CUSUM), generalized likelihood ratio (GLR) types of detection procedures, etc.

Shewhart charts can be viewed as scan statistics over time. A change is detected when the process is out of control, i.e., the detection statistic falls out of the control limit. A commonly used Shewhart chart for multivariate observations is the Hotelling's T^2 control chart [88], which can detect both mean and covariance shifts and the control limits are set through chi-square distributions. Modified T^2 charts based on principal component analysis are considered in [92, 93]. The U^2 multivariate control chart in [168] considers detecting the mean shift in a known subspace. Those work does not consider the largest eigenvalue as a detection statistic.

While Shewhart charts use the current subgroup samples to compute the detection statistic, the CUSUM procedure utilizes all past samples and updates the detection statistic recursively based on the log-likelihood ratio [140]. Multivariate CUSUM procedure for detecting mean shift has been developed in [147] and a more recent work [28] presents CUSUM based on projected data. In classic CUSUM, the pre-change and post-change distributions are specified completely. The proposed Subspace-CUSUM procedure here is not a typical CUSUM since we estimate the unknown subspace after the change.

Usually, the post-change distributions or their parameters are unknown and hard to pre-specify. One solution is to set the post-change parameter to represent the "smallest possible change" of interest. However, when there is a parameter mismatch, the CUSUM procedure suffers from a performance loss. The generalized likelihood ratio (GLR) procedure is introduced to handle unknown post-change distributions [108]. The Subspace-CUSUM procedure here is different from the GLR procedure since we do not estimate the full log-likelihood function; instead, we only estimate the subspace and introduce an additional parameter to control the performance.

Covariance shift detection has been considered in the past literature using various detection statistics. A multivariate CUSUM based on likelihood functions of multivariate Gaussian is studied in [86] considering a specific setting where the covariance changes from Σ to $c\Sigma$ for a constant c. The determinant of the sample covariance matrix was used in [3] and [4] to detect the covariance change. [36] considers a CUSUM chart for monitoring covariance shift using the projection pursuit [90] and likelihood ratio, with simulation studies on the performance of the proposed methods. Offline change detection of covariance change is studied in [40] using the Schwarz information criterion [171], where the change-point location must satisfy certain regularity condition to ensure the existence of the maximum likelihood estimator. Recently, [203] uses the wide binary segmentation through independent projection (WBSIP) to recover the change-points for the covariance matrix in the offline setting. [11] uses the distance between empirical precision matrices to detect abrupt changes in covariance for the offline case. Classical approaches usually consider the general setting, and here we are interested in detecting the structural change, i.e., spiked covariance matrix.

Recent work has also considered other types of structured covariance changes. The detection of a shift in an off-diagonal sub-matrix of the covariance matrix is studied in [9] using likelihood ratios. The detection of switching subspaces is studied in [95] based on a CUSUM type procedure, but they only estimate the pre-change subspace using historical

data and assume the post-change subspace is known, this is different from our work since we also estimate the post-change subspace. [221] develops an offline modeling framework for multivariate functional data based on sparse subspace clustering.

The Largest-Eigenvalue Shewhart Chart is related to [22], which studies the sparse principal component test based on sparse eigenvalue statistics. The largest eigenvalue statistic is shown to be asymptotically minimax optimal in [22] for detecting whether there exists a sparse and low-rank component. A natural sequential version of this idea is to use a sliding window and estimate the largest eigenvalue of the corresponding sample covariance matrix. However, this sequential version does not enjoy any form of (asymptotic) optimality.

1.2 Sequential Change Detection by Weighted ℓ_2 Divergence

Many classic results and procedures for sequential change detection have been developed, see [151, 17, 196]. However, many widely used methods assume a parametric form of the distributions before and after the change. For high-dimensional data, such parametric methods can be difficult to implement in certain scenarios since the post-change distribution is typically unknown and complicated. Recently, there have been many interests in developing non-parametric change detection procedures for high-dimensional streaming data.

We focus on a type of *distribution-free* methods based on empirical distributions. Compared with parametric methods, such non-parametric tests are more flexible and can be more applicable for various real-world situations. They tend to perform better when (i) the data does not follow a parametric distribution or (ii) we do not have enough historical samples to estimate the underlying distribution reliably. However, one particular challenge is to establish performance guarantees and improve the sample efficiency of the non-parametric test statistic [160].

In Chapter 4, we develop a new data-driven distribution-free sequential change detection procedure based on the *weighted* ℓ_2 divergence between empirical distributions as the

test statistic, which is related to the idea of *testing closeness* between two distributions [19]. More specifically, we start by considering the problem of testing closeness between two discrete distributions from samples observed. Suppose we are given two independent sample sets $x_1^1, \ldots, x_{n_1}^1 \overset{\text{iid}}{\sim} p$ and $x_1^2, \ldots, x_{n_2}^2 \overset{\text{iid}}{\sim} q$, where p and q are discrete distributions defined on the finite observation space and they can be both unknown. Our goal is to design a *test* which, given these two sample sets, claims whether $p = q$ or there is a significant difference between p and q. We use the ℓ_2 norm $\|p - q\|_2$ to characterize the difference between two distributions. Note that the ℓ_1 norm is also commonly used in literature.

We propose a new type of test by considering a family of distance-based divergence between empirical distributions of the two sets of observations. More specifically, the proposed test rejects $p = q$ whenever the distance-based divergence between empirical distributions is larger than a data-dependent (random) threshold ℓ. We introduce "weights" that are design parameters, which can be particularly important in achieving good performance in practice when we do not have a large number of samples. We show the optimality of the proposed procedure in achieving the theoretical lower bound of the sample complexity required for a low-risk test that meets the specifications. Moreover, we develop practical optimization procedures for selecting the optimal weights and the low-dimensional projections for high-dimensional data. Finally, we extend the proposed test to sequential change detection and characterize theoretical performances in both offline and online settings.

The proposed non-parametric test can fit into many potential applications, such as wetland and dryland classification [51], sensor network monitoring of cascades failures [41], spatial crime rate change detection [224], personalize healthcare [45], etc.

Related Work

There is a long history of studying similar problems in both statistics and computer science. In statistics, a two-sample test is a fundamental problem in which one aims to decide if two sets of observations are drawn from the same distribution [117], with a wide range

of applications [113]. Available approaches to the two-sample test can be largely divided into two categories: parametric and non-parametric. The parametric approach assumes that the data distribution belongs to certain parametric families, but the parameters can be unknown [49]. The non-parametric setting does not impose any assumption on the underlying distribution and therefore is widely applicable to real scenarios.

Classical approaches focus on the so-called "goodness-of-fit" test to decide whether the observations follow a pre-specified distribution. Non-parametric goodness-of-fit tests can be generalized for two-sample (and multi-sample) tests; in this case, the focus is the asymptotic analysis when the sample size goes to infinity. For instance, the Kolmogorov-Smirnov test [187], and the Anderson-Darling test [79] focus on univariate distributions and compute divergences between the empirical cumulative distributions of two (and multi) samples. The Wilcoxon-Mann-Whitney test [127, 102] is based on the data ranks and is also limited to univariate distributions. Van der Waerden tests are based on asymptotic approximation using quantiles of the standard Gaussian distribution [48, 170]. The nearest neighbors test for multivariate data is based on the proportion of neighbors belonging to the same sample [169].

There is much work aimed at extending univariate tests to the multivariate setting. A distribution-free generalization of the Smirnov two-sample test is proposed in [24] by conditioning on the empirical distribution functions. Wald-Wolfowitz run test and Smirnov two-sample test are generalized to multivariate setting using minimal spanning trees in [65]. A class of distribution-free multivariate tests based on nearest neighbors is studied in [25, 87, 169], and a multivariate k-sample test based on Euclidean distance between sample elements is proposed in [192]. Some recent work includes methods based on maximum mean discrepancy (MMD) [75] and the Wasserstein distance [161]. In particular, the ℓ_2 test enables us to draw a conclusion directly based on comparing empirical distributions. Compared with existing methods such as the MMD test, which requires a huge gram matrix when the sample size is large, the ℓ_2 test enables us to choose weights flexibly to better

serve the testing task.

Another line of research in theoretical computer science deals with *closeness testing*. It is first studied in [19, 20], in which the testing algorithm with sub-linear sample complexity is presented; the lower bound to the sample complexity is given in [198]; a test that meets the optimal sample complexity is proposed in [37]; see [165] and [32] for recent surveys. The ℓ_2 case has also been studied in [19, 73, 37], and optimal algorithms are given. Many variants of closeness testing have also been studied recently. In [1], sublinear algorithms are provided for generalized closeness testing. In [23], the closeness testing is studied under the case where sample sizes are unequal for two distributions. In [52], a nearly optimal algorithm for closeness testing for discrete histograms is given. In [2], the problem is studied from a differentially private setting.

Outstanding early contributions of sequential change detection mainly focus on parametric methods [140, 141, 176, 125] and is well-summarized in recent books [109, 196]. Recently, there have been growing interests in the non-parametric hypothesis test used in change detection problems. In [29], the "QuantTree" framework is proposed to define the bins in high-dimensional cases recursively, and the resulted histograms are used for change detection. In [39], a sequential change detection procedure using nearest neighbors is proposed. In the seminal work [114], a binning strategy is developed to discretize the sample space to construct the detection statistic to approximate the well-known generalized likelihood ratio test. The binned detection statistic's asymptotic properties are studied, and it is shown to be asymptotically optimal when the pre-and post-change distributions are discrete. Note that here we do not rely on likelihood ratios and assume the pre- and post-change distributions are unknown, and all we have are some possible "training data."

1.3 Robust Hypothesis Testing with Wasserstein Uncertainty Sets

Hypothesis testing is a fundamental problem in statistics and an essential building block for machine learning problems such as classification and anomaly detection. The goal

of hypothesis testing is to find a decision rule to discriminate between two hypotheses given new data while achieving a small probability of errors. However, the exact optimal test is difficult to obtain when the underlying distributions are unknown. This issue is particularly challenging when the number of samples is limited, and we cannot obtain accurate estimations of the distributions. The limited sample scenario (for one or both hypotheses) commonly arises in many real-world applications such as medical imaging diagnosis [8], online change-point detection [151], and online anomaly detection [38].

For hypothesis testing, the well-known Neyman-Pearson Lemma [138] establishes that the likelihood ratio gives the optimal test for two simple hypotheses. This requires to specify *a priori* two true distribution functions P_1 and P_2 for the two hypotheses, which, however, are usually unknown in practice. When the assumed distributions deviate from true distributions, the likelihood ratio test may experience a significant performance loss.

Typically there are "training" samples available for both hypotheses. A commonly used approach is the generalized likelihood ratio test (GLRT) [206], which assumes parametric forms for the distributions and estimates parameters using data and plug into the likelihood ratio statistic. Another popular method is the density ratio estimation [190]. However, in many scenarios, the training samples for one or both hypotheses can be small. For instance, we tend to have a small sample size for patients in healthcare applications. In limited-sample scenarios, it can be challenging to estimate parameters for GLRT (especially in the high dimensional case) or to estimate density ratios accurately. Without reliable estimation of the underlying distributions, various forms of robust hypothesis testing [89, 91, 119, 77] have been developed by considering different "uncertainty sets". Huber's seminar work [89] sets the uncertainty set as the ϵ-contamination sets that contain distributions close to a nominal distribution defined by total-variation distance. In [91], the optimal tests are characterized under majorization conditions, which, however, are intractable in general. Thus, there remains a computational challenge to find the optimal test, especially when the data is *multi-dimensional*. This has become a significant obstacle in applying robust

hypothesis tests in practice.

We consider a setting where the sample size is small. When there are limited samples, the empirical distribution may have "holes" in the sample space: places where we do not have samples yet, but there is a non-negligible probability for the data to occur, as illustrated in Figure 1.1. Thus, we may not want to restrict the true distribution to be on the same support of the empirical distribution. However, many commonly used distance divergences for probability distributions, such as Kullback-Leibler divergence, are defined for distributions with common support. Thus, in our setting, it can be restrictive if we were to construct uncertainty sets using the Kullback-Leibler divergence (e.g., [119] and [77]). Similarly, total-variational norm-induced uncertainty sets will have this issue since they encourage distributions with the same support as the empirical distribution. This motivates us to consider an uncertainty set formed by the Wasserstein distance. It measures the distance between distributions using optimal transport metric, which is more suitable for distributions without common support.

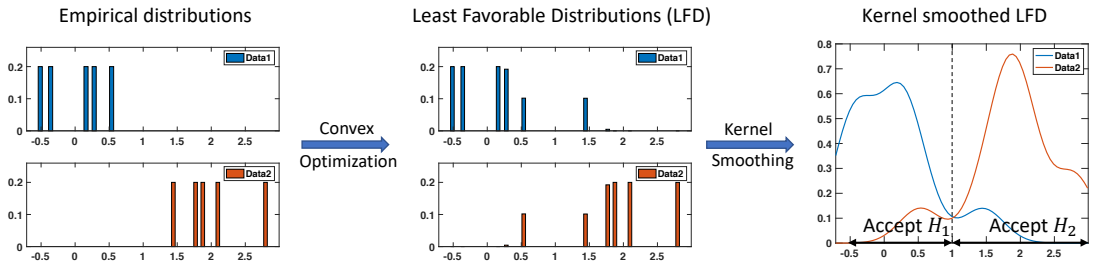


Figure 1.1: Left: Empirical distributions of two sets of training samples (5 samples each), generated from $\mathcal{N}(0,1)$ and $\mathcal{N}(2,1.2)$, respectively. Middle: Least Favorable Distributions (LFD) solve from Lemma 5.2 with radius equal to 0.1. Right: Kernel smoothed versions of LFD (with kernel bandwidth $h = 0.3$).

In Chapter 5, we present a new non-parametric minimax hypothesis test assuming the distributions under each hypothesis belong to two disjoint "uncertainty sets" constructed using the Wasserstein distance. Specifically, the uncertainty sets contain all distributions close to the empirical distributions formed by the training samples in Wasserstein distance.

This approach is more robust in small sample sizes when we cannot estimate the true distributions accurately.

A notable feature of our approach is the computational tractability and explicit characterization of the optimal test. The optimal test is based on a pair of least favorable distributions (LFD) from the uncertainty sets, which is a reminiscence of Huber's robust test. However, here the optimal test form is different, and our LFDs are computationally tractable in general. An outstanding challenge in finding the minimax test is that we face an infinite-dimensional optimization problem (finding the saddle point for optimal test and LFDs), which is hard to solve in general. To tackle the challenge, we make a connection to recent advances in distributionally robust optimization. In particular, we decouple the original minimax problem into two sub-problems using strong duality, which enables us first to find the optimal test for a given pair of distribution P_1 and P_2, and then find the LFDs P_1^* and P_2^* by solving a finite-dimensional convex optimization problem. We further characterize the general optimal test and extend the test to the "batch" setting containing multiple test samples. We also characterize the optimal radii of the uncertainty sets, which is an important question that affects the optimal test's generalization property. Finally, we show our method's good performance using simulated and real data, and demonstrate its applicability for sequential human activity detection.

Related Work

Robust hypothesis testing has been developed under the minimax framework by considering various forms of "uncertainty sets". Seminal work by Huber [89] considers the ϵ-contamination sets that contain distributions close to a nominal distribution defined by total-variation distance. Huber and Strassen later generalized the results in [91] based on the observation that the ϵ-contamination sets can be described using the so-called alternating capacities. It is claimed that under this capacity assumption, there is a representative pair (namely the LFDs) such that the Neyman-Pearson test between this pair is minimax

optimal. Although Huber provides an explicit characterization of the robust hypothesis test in the form of a truncated likelihood ratio, the "capacities" condition is required to obtain the optimality result; the LFDs are difficult to obtain in general. Our result is consistent with [89] in that our robust test also depends on the least favorable distributions, but we find the LFDs from data by solving a tractable optimization problem.

More recently, [119] and [77] consider uncertainty sets induced by Kullback-Leibler (KL) divergence in the one-dimensional setting without specifying parametric forms; the optimal test is obtained using the strong duality of problem induced by the KL divergence. Aiming to develop a computationally efficient procedure, [72, 33] consider a convex optimization framework for hypothesis testing, assuming parametric forms for the distributions and the parameters under the null and the alternative hypothese belong to convex sets. We consider a new way to construct uncertainty sets using Wasserstein metrics and empirical distributions to achieve distributional robustness. Using Wasserstein metric to achieve robustness is a popular technique and has been applied to many areas, including computer vision [166, 118, 155], generative adversarial networks [10, 78], and two-sample test [161].

Our work is also closely related to the distributional robust optimization (DRO) framework [58, 68, 27, 185, 172]. However, we cannot obtain our results using a simple extension of the existing DRO framework. In particular, we need new techniques to obtain tractable reformulation in our problem. Existing DRO problems typically involve only one class of empirical samples. In contrast, our problem involves two classes and we cannot rely on existing strong duality results in DRO [27, 58, 68]. Besides, we provide new insights regarding our solution's structural properties that are different from those that occurred in other DRO problems. Another line of work in DRO aims to characterize the size of uncertainty sets. The robust Wasserstein profile inference (RWPI) is one tool to provide the asymptotic distribution of the sufficient radius. It is first proposed and applied in the finite-dimensional case in [26] and is generalized to the infinite-dimensional case in [178]. The non-asymptotic concentration bound for the uncertainty set size is given in [67].

1.4 Convex Parameter Recovery for Interacting Marked Processes

Discrete events are a type of sequential data, where each data point is a tuple consisting of event time, location, and possibly category. Such event data is ubiquitous in modern applications, such as police data [130], electronic health records [96], and social network data [189, 105]. In modeling discrete events, we are particularly interested in estimating the interactions of events, such as triggering or inhibiting effects of past events on future events. For example, in crime event modeling, the triggering effect has been empirically verified; when a crime event happens, it makes the future events more likely to happen in the neighborhood. Similar empirical observations have been made for other applications such as in biological neural networks, social networks [223, 121], financial networks [56], and spatio-temporal epidemiological processes [103].

A popular model for capturing *interactions* between discrete events is the so-called Hawkes processes [84, 83, 85, 162]. The Hawkes process is a type of mutually-exciting non-homogeneous point process with intensity function consisting of a deterministic part and a stochastic part depending on the past event. The stochastic part of the intensity function can capture the interactions of past events and the current event, and it may be parameterized in different ways. In a certain sense, Hawkes processes may be viewed as a point process analog to classical autoregression in time series analysis. Hawkes process has received a lot of attention since it is quite general and can conveniently model interactions. For instance, in a network Hawkes process, interactions between nodes are modeled using a directed weighted graph in which direction and magnitude of edges indicate direction and strength of influence of one node on another. Along this line, there are various generalizations that allow for other types of point process modeling, where different "link" functions are considered, such as self-correcting process, reactive process, and specialized process (see [162] for an overview).

Estimating the *interactions* of the past events and the current event is a fundamental

problem for Bernoulli processes since it reveals the underlying temporal and spatial structures and allows for predicting future events. There has been much prior work in estimating model parameters, assuming that interactions are shift-invariant and captured through *kernel functions*. Furthermore, various simplifying assumptions are typically made for the kernel functions, e.g., that the spatio-temporal interactions are decoupled (e.g., [223]), implying that the interaction kernel function is a product of the interaction over time and interaction over locations and can be estimated separately. It is often assumed that the temporal kernel function decays exponentially over time with an unknown decay rate [121], or it is completely specified [80]; thus, the problem focus is on estimating spatial interaction between locations. It is also commonly assumed that the interactions are positive, i.e., the interaction triggers rather than inhibit future events [220]. Such simplification, however, may impede capturing complex interaction effects between events. For instance, negative interaction or inhibition is well known to play a major role in neuronal connectivity [55]. The study of more complex modeling of spatial aspects, especially jointly with discrete marks, is still in infancy.

In Chapter 6, we present a general computational framework for estimating marked spatio-temporal processes with categorical marks. Motivated by Hawkes processes, we consider a model of a discrete-time process on a finite spatio-temporal grid, which we refer to as Bernoulli processes. A brief description of the proposed modeling is as follows. At each time t a site k of the grid of the *M-state Bernoulli process* can be in one of $M + 1$ states – a ground state, in which "nothing happens," or an event state if an event of one of M given types at every (discrete) time instant t takes place at the site. We assume that the probability distribution of the events at each location at time t is a (linear or nonlinear) function on the process history – past events at different sites at times from $t - d$ to $t - 1$, d being the memory depth parameter of the process. For instance, each site of a 1-state linear (vanilla) Bernoulli process can be in one of two states – 0 (no event) or 1 (event takes place).

From the point of view of time series, this process can be seen as a vector autoregressive process with observations at sites of the grid at time t being Bernoulli random variables with conditional expectation given the process history being a linear combination, with coefficients which are unknown process parameters, of states of the process sites at times $t-d$ to $t-1$. This model can be seen as a natural simplification of the continuous-time Hawkes process where spatio-temporal cells are so small that one can ignore the chances for two or more events occurring in a cell.

A notable feature of this model is that prior information on the structure of interactions is represented by general convex constraints on the parameters, allowing for very general types of structures of interactions. Here convexity is assumed for the sake of computational tractability. For instance, we can relax the nonnegativity restrictions on interaction parameters and/or avoid assumptions of monotone or exponential time decay of interactions commonly used in the literature; when the situation has a "network component" allowing to assume that interacting sites are pairs neighboring nodes in a known graph, we can incorporate this information, for instance, by restricting the interaction coefficients for non-neighboring pairs of sites to be zero.

Related Work

The considered model is related to information diffusion processes over continuous time, for example, nonlinear Hawkes model [42], self-exciting processes over networks (see [162] for an overview), information diffusion networks [74], and multivariate stationary Hawkes processes [55]. Compared to these well-known models, time and space discretization leading to the spatio-temporal Bernoulli process is a considerable simplification that, nonetheless, leads to practical estimation routines that can be used in "real world" scenarios.

Various approaches to parametric and nonparametric estimation of spatio-temporal processes have been proposed in the literature. A line of work [63, 220, 131] consider non-

parametric Hawkes process estimation based on the Expectation-Maximization (EM) algorithms and the Kernel method. Least-square estimates for link functions of continuous-time multivariate stationary Hawkes process are studied in [55]. There is also much work [129, 57, 148] considering the estimation in the Bayesian framework. In particular, [153] considers estimation in a Bernoulli model similar to the one we promote in this work using the Bayesian approach and impose prior distributions on parameters. Several authors consider the problem of sparse model estimation for point processes [82].

Our approach to processing the estimation problem is based on convex optimization, which leads to computationally efficient procedures. We consider two classes of recovery procedures based on the Least Squares (LS) (which, in hindsight, is resembling but not identical to what is done in [55]) and Maximum Likelihood (ML) estimation. We cast estimation into convex optimization using Variational Inequality (VI) formulation of the corresponding statistical problems, which allows us to provide interpretable performance bounds and confidence intervals for the estimates and leads to computationally efficient numerical algorithms when processing large data sets.

The main contribution is in proposing models for the marked interacting processes which allow for simple "computation-friendly" statistical analysis utilizing *variational inequalities with monotone operators*. To the best of our knowledge, except for [98], we do not know other examples of using this approach in the statistical literature. The importance of this approach becomes clear in the case of nonlinear link function. Assuming the link to be a monotone vector field, our variational inequality-based approach reduces the estimation problem to an efficiently solvable problem with convex structure. In contrast, the ML and the LS estimation in the "nonlinear monotone link" case of our model result, in general, lead to solving nonconvex optimization problems and thus are not computationally friendly. It is also worth mentioning that concentration inequalities for martingales are not standard: the corresponding bounds are expressed using observable characteristics, which leads to confidence sets for the estimates expressed in terms of observations.

CHAPTER 2
PRELIMINARIES

This chapter presents some preliminaries and useful background information in the filed of sequential change detection, distributionally robust optimization, and variational inequalities with monotone operators. The standard problem formulations are introduced and classical results are reviewed.

2.1 Sequential Change Detection

The efficient detection of abrupt changes in the statistical behavior of streaming data, referred to as sequential (quickest) change detection, is an important and fundamental research topic in statistics, signal processing, and information theory. Such problems have been studied under the theoretical framework [151, 196, 199], and has a wide range of applications, such as power networks [43], internet traffic [111], cyber-physical systems [139], sensor networks [156], social networks [157, 143], epidemic detection [16], scientific imaging [154], genomic signal processing [174], seismology [6], video surveillance [116], and wireless communications [106]. This section presents the basics of sequential change detection and some related generalizations and extensions. Some detailed reviews on this topic can be found in [213, 199, 109], etc.

2.1.1 Problem Definition

In the sequential change detection problem, the aim is to detect a possible change in the data generating distribution of a sequence of observations $\{X_n, n = 1, 2, \ldots\}$. The initial distribution of the observations is the one corresponding to normal system operation. At some unknown time ν (referred to as the *change-point*), due to some event, the distribution of the random observations changes. The goal is to detect the change as quickly as possi-

ble, subject to false-alarm constraints. We start by assuming that the observations X_n are independent and identically distributed (i.i.d.) with probability density function (pdf) f_0 before the change-point ($n \leq \nu$) and pdf f_1 after the change-point ($n > \nu$), respectively. The generalization to non-i.i.d. settings is summarized in some recent books [196, 194].

A central problem when designing sequential change detection procedures is about the *tradeoff* between false-alarm and detection delay. The goal in sequential change detection theory is to find detection procedures that have guaranteed optimality properties in terms of this tradeoff.

2.1.2 Mathematical Preliminaries

Sequential change detection is closely related to the problem of statistical hypothesis testing, in which observations, whose distribution depends on the hypothesis, are used to decide which of the hypotheses is true. For the special case of binary hypothesis testing, we have two hypotheses, the *null* hypothesis and the *alternative* hypothesis. The classic Neyman-Pearson Lemma [138] establishes the form of the optimal test for this problem. In particular, consider the case of a single observation X, and suppose the pdf of X under the null and alternative hypotheses are f_0 and f_1, respectively. Then, the test that minimizes the false negative error (Type-II error), under the constraint of the false-positive error (Type-I error), is to compare the *likelihood ratio* $f_1(X)/f_0(X)$ to a threshold to decide which hypothesis is true. The likelihood ratio test is also optimal under other criteria such as Bayesian and minimax [133]. As we will see, the likelihood ratio also plays a key role in the development of sequential change detection algorithms.

The goal of sequential change detection is to design a *stopping time* on the observation sequence at which it is declared that a change has occurred. A stopping time is formally defined as follows:

Definition 2.1 (Stopping time)**.** A stopping time with respect to a random sequence $\{X_n, n = 1, 2, \ldots\}$ is a random variable τ such that for each n, the event $\{\tau = n\} \in \sigma(X_1, \ldots, X_n)$,

where $\sigma(X_1, \ldots, X_n)$ denotes the σ-algebra generated by $(X_1, \ldots, X_n)$. Equivalently, the event $\{\tau = n\}$ is a function of only $X_1, \ldots, X_n$.

The main results on stopping times that are most useful for sequential change detection problems include Doob's Optional Stopping Theorem [46] and Wald's Identity [180].

A quantity that plays an important role in the performance of sequential change detection algorithms is the Kullback-Leibler (KL) divergence between two distributions.

Definition 2.2. (KL Divergence). The KL divergence between two pdfs f_1 and f_0 is defined as $D(f_1 \| f_0) = \int f_1(x) \ \log(f_1(x)/f_0(x)) \ dx$.

Note that $D(f_1 \| f_0) \geq 0$ with equality if and only if $f_1 = f_0$ almost surely. It is usually assumed that $0 < D(f_1 \| f_0) < \infty$.

Define the log-likelihood ratio for an observation X:

$$\ell(X) := \log \frac{f_1(X)}{f_0(X)}. \tag{2.1}$$

A fundamental property of the log-likelihood ratio, which is useful for constructing sequential change detection algorithms, is that before the change $n \leq \nu$, the expected value of $\ell(X_n)$ is equal to $-D(f_0||f_1) < 0$; and after the change, $n > \nu$, the expected value of $\ell(X_n)$ is equal to $D(f_1||f_0) > 0$. As will be seen later, the KL divergence between the pre- and post-change distributions is an important quantity that characterizes the tradeoff between the average detection delay and the false-alarm rate.

2.1.3 Common Sequential Change Detection Procedures

We now present two commonly used sequential change detection procedures, the CUSUM and GLR procedure, which will be used in subsequent chapters.

The CUSUM procedure was first introduced by Page [140]. Although the CUSUM procedure was developed heuristically, it was later shown in [125, 134, 164, 107] that it has very strong optimality properties, which we will discuss further in Section 2.1.4.

The CUSUM procedure utilizes the properties of the cumulative log-likelihood ratio sequence:

$$S_n = \sum_{k=1}^{n} \ell(X_k).$$

Before the change occurs, the statistic has a negative drift because the expected value of $\ell(X_k)$ before the change is negative. After the change, it has a positive drift because the expected value of $\ell(X_k)$ after the change is positive. Thus, S_n roughly attains its minimum at the change-point ν. The CUSUM procedure is then constructed to detect this change in the drift of S_n. Specifically, the exceedance of S_n with respect to its past minimum is taken and compared with a threshold $b > 0$:

$$\tau_{\mathrm{c}} = \inf \left\{ n \geq 1 : W_n = \left(S_n - \min_{0 \leq k \leq n} S_k \right) \geq b \right\}. \tag{2.2}$$

The CUSUM statistic can also be rewritten as:

$$W_n = S_n - \min_{0 \leq k \leq n} S_k = \max_{0 \leq k \leq n} \sum_{i=k+1}^{n} \ell(X_i) = \max_{1 \leq k \leq n+1} \sum_{i=k}^{n} \ell(X_i). \tag{2.3}$$

Note that the maximization over all possible $\nu = k$ corresponds to plugging in a maximum likelihood estimate of the unknown change-point location in the log-likelihood ratio of the observations to form the CUSUM statistic. It can be shown that W_n can be computed recursively:

$$W_n = (W_{n-1} + \ell(X_n))^+, \quad W_0 = 0,$$

where $(x)^+ = \max\{x, 0\}$. This recursion enables the efficient online implementation of the CUSUM procedure in practice. In many cases, a slightly different recursion is also frequently used:

$$W_n = (W_{n-1})^+ + \ell(X_n), \quad W_0 = 0. \tag{2.4}$$

The CUSUM procedure requires full knowledge of pre- and post-change distributions

to obtain the log-likelihood ratio $\ell(X)$ used in computing the test statistics. In practice, the post-change distribution f_1 may be unknown. In the parametric setting, the post-change distribution can be parametrized using f_θ, where $\theta \in \Theta$ is the unknown parameter. A commonly used method for the situation here, which corresponds to the problem of composite hypothesis testing, is the generalized likelihood ratio (GLR) approach. In the GLR approach, a supremum over $\theta \in \Theta$ is taken in constructing the test statistic. In particular, the test statistic for the GLR-CUSUM algorithm is given by:

$$W_n^{\mathrm{G}} = \max_{1 \le k \le n+1} \sup_{\theta \in \Theta} \sum_{i=k}^{n} \ell_\theta(X_i), \tag{2.5}$$

where $\ell_\theta(X) = \log(f_\theta(X)/f_0(X))$. Performance analyses of the GLR-CUSUM algorithm for one-parameter exponential families can be found in [125, 126]. A major drawback of the GLR approach is that the corresponding GLR statistic (e.g., the one given in (2.5)) cannot be computed recursively in time, except in some special cases (e.g., when the parameter set Θ has finite cardinality). To reduce the computational cost, a window-limited GLR approach was developed in [207] and generalized in [107, 110]. Window-limited versions of the GLR algorithm can be shown to be asymptotically optimal in certain cases if the window size is carefully chosen.

2.1.4 Optimality Results

We now briefly summarize optimality results in the existing literature for the above procedures. We only review the minimax (non-Bayesian) setting, where we do not assume a prior distribution for the change-point ν. The Bayesian setting is not discussed here, more details can be found in [176, 196, 213].

A fundamental problem in sequential change detection is to optimize the tradeoff between the false-alarm rate and the average detection delay. Controlling the false-alarm rate is commonly achieved by setting an appropriate threshold on a test statistic such as the

one in (2.2). But the threshold also affects the average detection delay. A larger threshold incurs fewer false alarms and leads to a larger detection delay, and vice versa.

In minimax settings, the change-point is assumed to be a *deterministic* unknown variable. In this case, the average run length (ARL) to false alarm is generally used as a performance measure for false alarms:

$$\mathsf{ARL}(\tau) = \mathbb{E}_\infty[\tau], \tag{2.6}$$

where $\mathbb{P}_\infty$ is the probability measure on the sequence of observations when the change never occurs, and $\mathbb{E}_\infty$ is the corresponding expectation. Denote the set of stopping times that satisfy a constraint γ on the ARL by $\mathcal{D}_\gamma := \{\tau : \mathsf{ARL}(\tau) \geq \gamma\}$.

Finding a uniformly powerful test that minimizes the delay over all possible values of the change-point ν, subject to a ARL constraint, is generally intractable. Therefore, it is more tractable to pose the problem in the so-called minimax setting. There are two essential measures of the detection delay in the minimax setting, due to Lorden [125] and Pollak [149], respectively.

Lorden considers the supremum of the average detection delay conditioned on the worst possible realizations. In particular, Lorden defines:

$$\mathsf{WADD}(\tau) = \sup_{n\geq 0} \text{ess sup}\, \mathbb{E}_n\left[(\tau - n)^+ | X_1, \ldots, X_n\right], \quad \text{(Lorden)} \tag{2.7}$$

where $\mathbb{P}_n$ denotes the probability measure on the observations when the change-point $\gamma = n$, and $\mathbb{E}_n$ denotes the corresponding expectation. We then have the following Lorden's formulation:

$$\text{minimize } \mathsf{WADD}(\tau) \text{ subject to } \mathsf{ARL}(\tau) \geq \gamma. \tag{2.8}$$

For the i.i.d. setting, Lorden showed that Page's CUSUM procedure given in (2.2) is asymptotically optimal as $\gamma \to \infty$. It was later shown in [134] and [164] that the slight

modification of the CUSUM procedure in (2.4) is exactly optimal for (2.8) for all $\gamma > 0$.

Although the CUSUM procedure is exactly optimal under Lorden's formulation, WADD is a pessimistic measure of detection delay since it considers the worst-case pre-change samples. An alternative measure of detection delay was suggested by Pollak [149]:

$$\mathsf{CADD}(\tau) = \sup_{n \geq 0} \mathbb{E}_n[\tau - n | \tau \geq n], \quad \text{(Pollak)} \tag{2.9}$$

for all stopping times τ for which the expectation is well-defined. It is easy to see that for any stopping time τ, $\mathsf{WADD}(\tau) \geq \mathsf{CADD}(\tau)$, and therefore, Pollak's formulation is less pessimistic. On the other hand, Pollak's measure applies in the case where the change imposing mechanism uses data that are independent from the observations, while Lorden's measure applies when the change imposing mechanism can use data dependent with the observations. In terms of this perspective, the Lorden's performance measure can be viewed as less pessimistic since the corresponding limitation of its applicability is less obvious. We will use the Lorden's measure in remaining chapters.

In general, it may be challenging to exactly solve the problem in (2.8) and the corresponding problem defined using CADD in (2.9). For this reason, asymptotically optimal solutions for the above problems are often investigated in the literature. Specifically, a stopping time τ is said to be *first-order* asymptotically optimal if it satisfies:

$$\frac{\mathsf{CADD}(\tau)}{\inf_{\tau \in \mathcal{D}_\gamma} \mathsf{CADD}(\tau)} \to 1, \quad \text{as } \gamma \to \infty;$$

the notions can also be defined similarly for the problem in (2.8) defined using WADD.

Pollak's formulation has been studied for the i.i.d. data in [149] and [195]. The first-order asymptotic optimality for Lorden's formulation can also be extended to Pollak's formulation. To show this, Lorden in [125] established a universal lower bound for WADD and Lai in [107] proved the lower bound to CADD:

Theorem 2.1 (Lower bound for CADD [107])**.** *As* $\gamma \to 0$*,*

$$\inf_{\tau \in \mathcal{D}_\gamma} \mathsf{CADD}(\tau) \geq \frac{\log \gamma}{D(f_1||f_0)}(1 + o(1)).$$

It can be shown that the CUSUM procedure with a threshold $b = |\log \gamma|$ is first-order asymptotically optimum for both Lorden's and Pollak's formulations. In particular, as $\gamma \to \infty$,

$$\mathsf{CADD}(\tau_{\mathrm{c}}) = \mathsf{WADD}(\tau_{\mathrm{c}}) \sim \frac{\log \gamma}{D(f_1||f_0)},$$

where $\sim$ means the ratio of the quantities on its two sides approaches 1 as $\gamma \to \infty$.

Moreover, an alternative detection method called the SRP algorithm (Pollak's version of the Shiryaev-Roberts algorithm that starts from a quasi-stationary distribution of the Shiryaev-Roberts statistic) was proved to be third-order asymptotically optimal in [149], namely the correspoonding stopping time τ satisfies $\mathsf{CADD}(\tau) - \inf_{\tau \in \mathcal{D}_\gamma} \mathsf{CADD}(\tau) = o(1)$ as $\gamma \to \infty$. It was later shown by [150] that it is not strictly optimum.

Remark 2.1 (Evaluating the performance metrics)**.** In the definition of the WADD metric (2.7) and the CADD metric (2.9), it appears that we need to consider the supremum over all possible past observations and all possible change-points. However, we can actually show that for the CUSUM procedure, and some other algorithms, that the supremum over all possible change-points in WADD and CADD is achieved at time $n = 0$, i.e., the change happens before we take observations:

$$\mathsf{CADD}(\tau_{\mathrm{c}}) = \mathsf{WADD}(\tau_{\mathrm{c}}) = \mathbb{E}_0 \left[\tau_{\mathrm{c}}\right].$$

Therefore, the CADD and the WADD can be conveniently evaluated by setting $\gamma = 0$, without "taking the supremum". In such cases, we can also use the term expected detection delay (EDD) instead to denote $\mathbb{E}_0 [\tau_{\mathrm{c}}]$, and it is equivalent to the worst-case detection delay WADD in Lorden's definition and the CADD in Pollak's definition.

2.2 Distributionally Robust Optimization

For many modern data-driven decision-making problems, we typically solve an optimization problem to find the optimal decision variable. The performance of the decision is usually affected by uncertainties for the underlying data-generating distribution. Classical optimization problems for a fixed distribution might perform poorly in practice when the fixed distribution deviates significantly from the true distribution. In order to learn a decision from limited training samples that will generalize well to unseen test samples, the distributionally robust optimization (DRO) framework is commonly used and has gain much attention recently [50, 71, 158, 53, 205]. We focus on the distributionally robust optimization problem based on Wasserstein distances in the following [68, 26, 58].

2.2.1 Problem Definition

We start by considering the single-state stochastic program where the goal is to find a decision variable $x \in \mathbb{R}^d$ which minimizes the expected risk $\mathbb{E}_{\xi\sim P}[\Psi(x,\xi)]$, where the expectation is taken with respect to a random variable ξ taking values in the set $\Omega \subset \mathbb{R}^m$ following a distribution P and Ψ is the loss function. Let (Ω, c) be a metric space Ω with metric c. The space of Borel probability measures on Ω is denoted by $\mathscr{P}(\Omega)$ and $P \in \mathscr{P}(\Omega)$. When the true underlying distribution P of the random variable ξ is unknown, a possible solution is to use the distributionally robust optimization as an alternative.

In distributionally robust optimization problems, we typically construct an ambiguity (uncertainty) set containing all possible distributions of the random variable ξ, denoted as $\mathcal{P} \subset \mathscr{P}(\Omega)$. The goal is to find a decision variable x which minimizes the *worst-case* expected risk defined as the supreme expected risk over the ambiguity set $\mathcal{P}$:

$$\sup_{P\in\mathcal{P}} \mathbb{E}_{\xi\sim P}[\Psi(x,\xi)].$$

Then the problem of solving the optimal robust decision variable x can be written as:

$$\inf_{x} \sup_{P \in \mathcal{P}} \mathbb{E}_{\xi \sim P}[\Psi(x, \xi)]. \tag{2.10}$$

There are multiple ways to construct the ambiguity set $\mathcal{P}$. We focus on the case when we have a nominal distribution P_0 and the ambiguity set $\mathcal{P}$ is the collection of probabilities that are close to the nominal distribution P_0 with respect to certain divergence measures:

$$\mathcal{P} = \{P \in \mathscr{P}(\Omega) : D(P, P_0) \leq r\},$$

where $r \geq 0$ is the so-called radius parameter that controls the size of the ambiguity set. If r is set to zero, then the ambiguity set is a singleton that only contains the nominal distribution P_0. Some commonly used divergence measures $D(\cdot, \cdot)$ include the Kullback-Leibler divergence [77, 119], Total-Variation distance [89, 91], Wasserstein metric [68, 58, 69], etc. In the following subsection, we focus on the data-driven distributionally robust optimization using Wasserstein metric and review some useful results.

2.2.2 Data-driven Formulation and Strong Duality

In particular, in the data-driven case, we may not know the true date-generating distribution P for the random variable ξ exactly and the only information we have is historical samples of the random variable ξ. Suppose we have a set of training samples $\{\xi_1, \ldots, \xi_n\}$ that are i.i.d. sampled from the unknown distribution P, then we can estimate the distribution P using the empirical distribution defined as follows:

$$\hat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i},$$

where δ_ξ denotes the Dirac point mass concentrated on ξ for each $\xi \in \mathbb{R}^m$, i.e., $\delta_\xi(A) = \mathbb{1}_{\{\xi \in A\}}$ for any Borel measurable set and $\mathbb{1}_{\{.\}}$ is the indicator function.

In the following, we first define the Wasserstein metric and then construct the ambiguity set $\mathcal{P}$ based on the Wasserstein metric.

Definition 2.3. (Wasserstein metric). The *Wasserstein metric* of order p for two given distributions $P, Q \in \mathscr{P}(\Omega)$ is defined as:

$$\mathsf{W}_p(P,Q) := \left(\min_{\gamma \in \Gamma(P,Q)} \left\{ \mathbb{E}_{(\omega,\omega')\sim\gamma} \left[c^p(\omega,\omega') \right] \right\} \right)^{1/p},$$

where $c(\cdot,\cdot) : \Omega \times \Omega \to \mathbb{R}_+$ is a metric on Ω, and $\Gamma(P,Q)$ is the collection of all Borel probability measures on $\Omega \times \Omega$ with marginal distributions P and Q.

When both distributions P and Q are discrete measures, the Wasserstein metric is also known as finding the optimal transport plan that maps P to Q [201]. The well-known Kantorovich duality establishes the equivalent dual form for the Wasserstein metric that is commonly used in practice.

Theorem 2.2 (Kantorovich Duality, [201])**.** *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish probability spaces and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous cost function. Then we have the duality*

$$\min_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\gamma(x,y) = \sup_{\substack{(\phi,\psi)\in L^1(\mu)\times L^1(\nu) \\ \phi(x)+\psi(y)\le c(x,y),\ \forall x,y}} \left(\int_{\mathcal{X}} \phi(x) d\mu + \int_{\mathcal{Y}} \psi(y) d\nu \right),$$

where $\gamma \in \Pi(\mu,\nu)$ denotes the joint distribution on $\mathcal{X} \times \mathcal{Y}$, with marginal distributions μ and ν, respectively.

Note that when μ is a discrete distribution on $\{x_1, \ldots, x_m\}$, then the function $\phi \in L^1(\mu)$ can be viewed as a vector $\eta := [\phi(x_1); \ldots; \phi(x_m)] \in \mathbb{R}^m$. And the above dual formulation will be reduced to:

$$\min_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\gamma(x,y) = \sup_{\substack{\eta\in\mathbb{R}^m, \psi\in L^1(\nu) \\ \eta_i+\psi(y)\le c(x_i,y) \\ \forall 1\le i\le m, \forall y}} \left(\sum_{i=1}^{m} \eta_i \mu(x_i) + \int_{\mathcal{Y}} \psi(y) d\nu \right),$$

this is what we will use frequently in Chapter 5.

Based on the definition of the empirical distribution and the Wasserstein metric, the ambiguity set can be constructed as all distributions whose Wasserstein distance of order p with the nominal distribution $\hat{P}_n$ is at most a given radius r:

$$\mathcal{P}_p(\hat{P}_n, r) = \{P \in \mathscr{P}(\Omega) : \mathsf{W}_p(P, \hat{P}_n) \leq r\}. \tag{2.11}$$

We have the following strong duality result for solving the inner supreme in the Wasserstein distributionally robust optimization problem (2.10) when the ambiguity set is constructed from Wasserstein metric as in (2.11).

Theorem 2.3 (Theorem 1, [68])**.** *Suppose Ω is the sample space for the random variable ξ. Consider any $\nu \in \mathscr{P}(\Omega)$ and $\Psi \in L^1(\nu)$. Let $p \in [1, \infty)$ and $\theta > 0$. Then we have:*

$$\sup_{\mu \in \mathcal{P}_p(\nu, \theta)} \int_\Omega \Psi(\xi)\mu(d\xi) = \inf_{\lambda \geq 0} \left\{ \lambda\theta^p - \int \inf_{\xi \in \Omega} [\lambda c^p(\zeta, \xi) - \Psi(\xi)]\nu(d\zeta) \right\},$$

where $\mathcal{P}_p(\nu, \theta)$ is the ambiguity set induced by Wasserstein metric:

$$\mathcal{P}_p(\nu, \theta) = \{\mu \in \mathcal{P}(\Omega) : \mathsf{W}_p(\mu, \nu) \leq \theta\}.$$

Due to the strong duality results, the optimal decisions can often be computed by solving tractable convex optimization problems which can be computed efficently through off-the-shelf optimization software.

2.3 Variational Inequalities with Monotone Operators

Variational inequality (VI) with monotone operators is the principal computational tool for optimization problems, and is a general method for solving convex optimization problems. The structural assumptions to be imposed on the variational inequality formulations are weaker than those resulting in convex maximum likelihood based problems and their

sample average approximations. For example, it can be shown that using a least-square estimation approach for parameter recovery in generalized linear models (GLM) sometimes will lead to a non-convex optimization problem. However, using VI approach for parameter estimation will lead to a convex program. The solution to the VI can be found in a computationally efficient way (more details can be found in [98]).

We start with the related preliminaries. A vector field $F : \mathcal{X} \to \mathbb{R}^N$ defined on a nonempty convex subset $\mathcal{X}$ of $\mathbb{R}^N$ is called *monotone*, if

$$\langle F(x) - F(y), x - y \rangle \geq 0, \ \forall x, y \in \mathcal{X}.$$

When $N = 1$, monotonicity means that the scalar function F is nondecreasing on $\mathcal{X}$. A basic example of a multivariate monotone vector field is the subgradient of a convex function $f : \mathcal{X} \to \mathbb{R}$, i.e., $F(x) = \partial f(x)$ be defined as the set of all subgradients of f at x. Since f is convex, for any $x, x' \in \mathcal{X}$ and $f'(x) \in \partial f(x), f'(x') \in \partial f(x')$, we have that $(f'(x))^\top (x - x') \geq f(x) - f(x') \geq (f'(x'))^\top (x - x')$, thus $(f'(x) - f'(x'))^\top (x - x') \geq 0$.

We say that $\alpha \geq 0$ is a *modulus of strong monotonicity* of vector field F, when

$$\langle F(x) - F(y), x - y \rangle \geq \alpha \|x - y\|_2^2, \ \forall x, y \in \mathcal{X};$$

when $\alpha > 0$, F is called *strongly monotone.*

A pair $(\mathcal{X}, F)$ comprised of nonempty convex domain $\mathcal{X}$ and monotone vector field F on this domain gives rise to *variational inequality* $\mathrm{VI}(F, \mathcal{X})$. A *weak solution* to this VI is a point $\bar{x} \in \mathcal{X}$ such that

$$\langle F(x), x - \bar{x} \rangle \geq 0, \ \forall x \in \mathcal{X}.$$

A *strong solution* to this VI is a point $\bar{x} \in \mathcal{X}$ such that

$$\langle F(\bar{x}), x - \bar{x} \rangle \geq 0, \ \forall x \in \mathcal{X}.$$

Whenever F is strongly monotone, weak solution, if exists, is unique. Every strong solution is a weak one; when F is continuous on $\mathcal{X}$, the inverse is also true. When $\mathcal{X}$ is a convex compact set, $\mathrm{VI}(F, \mathcal{X})$ always has weak solutions. When F is the gradient field of a continuously differentiable convex function f on $\mathcal{X}$, the weak and the strong solutions to $\mathrm{VI}(F, \mathcal{X})$ are exactly the minimizers of f on $\mathcal{X}$.

Finally, we should stress that variational inequalities with monotone operators are the most general "problems with convex structure;" under mild computability assumptions, that can be efficiently solved to a high accuracy, see [98] for the stochastic algorithms and convergence analysis.

CHAPTER 3

SEQUENTIAL SUBSPACE CHANGE DETECTION

This chapter presents the work on sequential subspace change-point detection. This work is mainly summarized in [212, 209, 211]. Section 3.1 introduces the problem set-up for sequential subspace change-point detection and shows a unified framework for the emerging and switching subspace problems. Section 3.2 presents the proposed two sequential change detection procedures: the Largest-Eigenvalue Shewhart chart and Subspace-CUSUM procedure. Section 3.3 presents theoretical approximations and bounds for the average run length and the expected detection delay of the Largest-Eigenvalue Shewhart chart, as well as asymptotic optimality of the Subspace-CUSUM procedure. Section 3.4 contains simulation studies to demonstrate the performance of the proposed algorithms in different settings. Section 3.5 shows two real data examples using human gesture detection and seismic event detection.

3.1 Problem Setup

We first introduce the spiked covariance model considered in [97], which assumes that a small number of directions explain most of the variance. For simplicity, we consider the spiked covariance model of rank-one. The results can be generalized to the case where rank is greater than one using similar ideas. In particular, the rank-one spiked covariance matrix is given by

$$\Sigma = \sigma^2 I_k + \theta u u^\top,$$

where I_k denotes an identity matrix of size k; θ is the signal strength; $u \in \mathbb{R}^k$ represents a basis for the subspace with unit norm $\|u\| = 1$; σ^2 is the noise variance, which will be considered known since it can be estimated from historical data. It is possible to consider σ^2

unknown as well and provide estimates of this parameter along with the necessary estimates of u. However, to simplify our presentation, we decide to consider σ^2 known. The Signal-to-Noise Ratio (SNR) is defined as $\rho = \theta/\sigma^2$.

Formally, the *emerging* subspace problem can be cast as follows:

$$\begin{aligned} x_t &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_k), \qquad\qquad t = 1, 2, \ldots, \tau, \\ x_t &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_k + \theta u u^\top), \, t = \tau + 1, \tau + 2, \ldots \end{aligned} \tag{3.1}$$

where τ is the unknown change-point that we would like to detect from data that are acquired sequentially. Similarly, the *switching* subspace problem can be formulated as follows

$$\begin{aligned} x_t &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_k + \theta u_1 u_1^\top), \, t = 1, 2, \ldots, \tau, \\ x_t &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_k + \theta u_2 u_2^\top), \, t = \tau + 1, \tau + 2, \ldots \end{aligned} \tag{3.2}$$

where $u_1, u_2 \in \mathbb{R}^k$ represent bases for the subspaces before and after the change, with $\|u_1\| = \|u_2\| = 1$ and u_1 is considered known. In both settings, our goal is to detect the change as quickly as possible, subject to the constraint that false detections occurring before the true change-point are very rare.

The switching subspace problem in (3.2) can be reduced into the emerging subspace problem in (3.1) by a simple data projection. In detail, we can select any orthonormal matrix $Q \in \mathbb{R}^{(k-1)\times k}$ such that

$$Q u_1 = 0, \quad Q Q^\top = I_{k-1},$$

which means that all rows of Q are orthogonal to u_1, and they are orthogonal to each other and have unit norm. Then, we project each observation x_t using the projection matrix Q onto a $k-1$ dimensional space and obtain a new sequence:

$$y_t = Q x_t \in \mathbb{R}^{k-1}, t = 1, 2, \ldots.$$

Then y_t is a zero-mean random vector with covariance matrix $\sigma^2 I_{k-1}$ before the change and $\sigma^2 I_{k-1} + \theta Q u_2 u_2^\top Q^\top$ after the change. Let $u = Qu_2 / \|Qu_2\|$, and

$$\tilde{\theta} = \theta \|Qu_2\|^2 = \theta[1 - (u_1^\top u_2)^2].$$

Thus, problem in (3.2) can be reduced to the following

$$\begin{aligned} y_t &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_{k-1}), \qquad\qquad t = 1, 2, \ldots, \tau, \\ y_t &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_{k-1} + \tilde{\theta} u u^\top), \, t = \tau + 1, \tau + 2, \ldots \end{aligned} \tag{3.3}$$

Note that this way the switching subspace problem is reduced into the emerging subspace problem, where the new signal strength $\tilde{\theta}$ depends on the angle between u_1 and u_2, which is consistent with our intuition.

We would like to emphasize that by projecting the data onto a lower-dimensional space, we lose information, suggesting that the two versions of the problem are not exactly equivalent. Indeed, the optimum detector for the transformed data in (3.3) and the one for the original data in (3.2) do not coincide. This can be easily verified by computing the corresponding CUSUM tests and their optimum performance. Despite this difference, it is clear that with the proposed approach, we put both problems under the same framework, offering computationally simple methods to solve the original problem in (3.2). Consequently, in the following analysis, we focus solely on problem in (3.1).

3.2 Detection Procedures

We propose two online methods: the Largest-Eigenvalue Shewhart chart and the Subspace-CUSUM procedure. The notations are standard. Denote by $\mathbb{P}_\tau$ and $\mathbb{E}_\tau$ the probability and expectation induced when there is a change-point at the time τ. Under this definition, $\mathbb{P}_\infty$ and $\mathbb{E}_\infty$ is the probability and the expectation under the nominal regime (change never happens) while $\mathbb{P}_0$ and $\mathbb{E}_0$ the probability and expectation under the alternative regime

(change happens before we take any data).

3.2.1 Largest-Eigenvalue Shewhart Chart

Motivated by the test statistic in [22], we consider a Shewhart chart by computing the largest eigenvalue of the sample covariance matrix repeatedly over a sliding window. Assume the window length is w. For each time $t > 0$, the *un-normalized* sample covariance matrix using the available samples is given by

$$\hat{\Sigma}_{t,\min\{t,w\}} = \begin{cases} x_1 x_1^\top + \cdots + x_t x_t^\top, \text{ for } t < w; \\ x_{t-w+1} x_{t-w+1}^\top + \cdots + x_t x_t^\top, \text{ for } t \geq w. \end{cases} \tag{3.4}$$

We note that for $t = 1$ the matrix contains a single outer product and as time progresses the number of outer products increases linearly until it reaches w. After this point, namely for $t \geq w$, the number of outer products remains equal to w.

Let $\lambda_{\max}(X)$ denote the largest eigenvalue of a symmetric matrix X. We define the *Largest-Eigenvalue Shewhart chart*, as the one that stops according to the following rule:

$$T_{\mathrm{E}} = \inf\{t > 0 : \lambda_{\max}(\hat{\Sigma}_{t,\min\{t,w\}}) \geq b\}, \tag{3.5}$$

where $b > 0$ is a constant threshold selected to meet a suitable false alarm constraint. We need to emphasize that we do *not* divide by $\min\{t, w\}$ when forming the un-normalized sample covariance matrix. As we will explain in Section 3.3.1, it is better for T_{E} to always divide by w instead of $\min\{t, w\}$. Consequently, we can omit the normalization constant w from our detection statistics by absorbing it into the threshold.

3.2.2 Subspace-CUSUM Procedure

The CUSUM procedure [140, 180] is the most popular sequential test for change detection. When the observations are independent and identically distributed before and after

the change, CUSUM is known to be exactly optimum [134] in the sense that it solves a very well defined constrained optimization problem introduced in [125], see Chapter 2.1 for useful preliminaries. However, the CUSUM procedure can only be applied when we have exact knowledge of the pre- and post-change distributions. For our problem, this requires complete specification of all parameters, namely the subspace u, noise power σ^2, and SNR ρ. In this section, we first derive the exact CUSUM statistic in our setting and then present the proposed Subspace-CUSUM procedure.

To derive the CUSUM procedure, let $f_\infty(\cdot), f_0(\cdot)$ denote the pre- and post-change pdf of the observations. Then recursive CUSUM statistics as defined in (2.4) is as follows:

$$S_t = (S_{t-1})^+ + \log \frac{f_0(x_t)}{f_\infty(x_t)}, \tag{3.6}$$

and the CUSUM stopping time in turn is defined as

$$T_{\mathrm{C}} = \inf\{t > 0 : S_t \geq b\}, \tag{3.7}$$

where $b > 0$ is a threshold selected to meet a suitable false alarm constraint.

For our problem of interest (3.1), we can derive that

$$\begin{aligned}
\log \frac{f_0(x_t)}{f_\infty(x_t)} &= \log \left[\frac{[(2\pi)^k |\sigma^2 I_k + \theta u u^\top|]^{-1/2}}{[(2\pi)^k \sigma^{2k}]^{-1/2}} \times \frac{\exp\{-x_t^\top (\sigma^2 I_k + \theta u u^\top)^{-1} x_t/2\}}{\exp\{-x_t^\top x_t/(2\sigma^2)\}} \right] \\
&= \log \left[\left| I_k + \rho u u^\top \right|^{-\frac{1}{2}} \exp\left\{ \frac{1}{2} \frac{\theta}{\theta + \sigma^2} \frac{(u^\top x_t)^2}{\sigma^2} \right\} \right] \\
&= \frac{\rho}{2\sigma^2(1+\rho)} \left\{ (u^\top x_t)^2 - \sigma^2 \left(1 + \frac{1}{\rho} \right) \log(1+\rho) \right\}.
\end{aligned}$$

The second equality is due to the matrix inversion lemma [208] that allows us to write

$$(\sigma^2 I_k + \theta u u^\top)^{-1} = \frac{1}{\sigma^2} I_k - \frac{\theta}{\theta + \sigma^2} \frac{u u^\top}{\sigma^2},$$

which, after substitution into the equation, yields the desired result. Note that the multi-

plicative factor $\rho/[2\sigma^2(1+\rho)]$ is positive, so we can omit it from the log-likelihood ratio when forming the CUSUM statistic in (3.6). This leads to

$$S_t = (S_{t-1})^+ + (u^\top x_t)^2 - \sigma^2 \left(1 + \frac{1}{\rho}\right) \log(1+\rho). \tag{3.8}$$

Remark 3.1. We can show that the increment term in (3.8), i.e.,

$$(u^\top x_t)^2 - \sigma^2 \left(1 + \frac{1}{\rho}\right) \log(1+\rho),$$

has the following property: its expected value is negative under the pre-change and positive under the post-change probability measure. The proof relies on a simple argument based on Jensen's inequality [167]. Due to this property, before the change, the CUSUM statistics S_t will oscillate near 0 while it will exhibit, on average, a positive drift after the occurrence of the change forcing it, eventually, to hit or exceed the threshold.

Usually, the subspace u and SNR ρ are unknown. In this case it is impossible to form the exact CUSUM statistic depicted in (3.8). One option is to estimate the unknown parameters and substitute them back into the likelihood function. Here we propose to estimate only u and introduce a new *drift* parameter d which plays the same role as $\sigma^2(1+1/\rho)\log(1+\rho)$, this leads to the following Subspace-CUSUM update

$$\mathcal{S}_t = (\mathcal{S}_{t-1})^+ + (\hat{u}_t^\top x_t)^2 - d, \quad t \geq 1 \tag{3.9}$$

and $\mathcal{S}_0 = 0$. To apply (3.9), we need to specify d and of course provide the estimate $\hat{u}_t$. Regarding the latter we simply use the *unit-norm* eigenvector corresponding to the largest eigenvalue of the un-normalized sample covariance matrix $\hat{\Sigma}_{t+w,w}$ depicted in (3.4). We denote the estimator of u as $\hat{u}_t$ because at time t the estimate will rely on the data $x_{t+1}, \ldots, x_{t+w}$ that are in the "future" of t. Practically, this is always possible by properly

delaying our data by w samples. Stopping occurs similarly to CUSUM, that is

$$T_{\mathrm{SC}} = \inf\{t > 0 : \mathcal{S}_t \geq b\}. \tag{3.10}$$

Of course, in order to be fair, at the time of stopping we must make the appropriate correction, namely if $\mathcal{S}_t$ exceeds the threshold at t for the first time, then the actual stopping takes place at $t + w$. The reason we use estimates based on "future" data is to make x_t and $\hat{u}_t$ *independent* which in turn will help us decide what is the appropriate choice for the drift constant d in Section 3.3.3.

Remark 3.2. An alternative possibility is to use the generalized likelihood ratio (GLR) statistic, where both ρ and u are estimated for each possible change location κ. The GLR statistic is

$$\max_{\kappa<t} \left\{ -\frac{t-\kappa}{2} \log(1+\hat{\rho}_{\kappa,t}) + \frac{1}{2\sigma^2} \frac{\hat{\rho}_{\kappa,t}}{1+\hat{\rho}_{\kappa,t}} \sum_{i=\kappa+1}^{t} (\hat{u}_{\kappa,t}^{\top} x_i)^2 \right\},$$

where $\hat{\rho}_{\kappa,t}$, $\hat{u}_{\kappa,t}$ are estimated from samples $\{x_i\}_{i=\kappa+1}^{t}$. However, this computation is more intensive since there is no recursive implementation for the GLR statistic, furthermore it requires growing memory. There are finite memory versions [110], however are equally complicated in their implementation. Therefore, we do not consider the GLR statistic here.

3.3 Theoretical Analysis

To fairly compare the detection procedures discussed in the previous section, we need to calibrate them properly. The calibration process must be consistent with the performance measure we are interested in. Recall from Chapter 2.1 that for a given stopping time T we measure false alarms through the *average run length* (ARL) expressed with $\mathbb{E}_{\infty}[T]$. For the detection capability of T we use the *worst-case average detection delay* (WADD) defined in (2.7), which considers the worst possible data before the change and the worst possible

change-point.

In this section, we first discuss the scenarios that lead to the worst-case detection delay for the proposed procedures. Then we characterize the ARL and WADD of the Largest-Eigenvalue Shewhart chart. The theoretical characterization of ARL is very important because it can serve as a guideline on how to choose the threshold b used in the detection procedure. Without theoretical analysis, people usually use Monte Carlo simulation to set the threshold, which can be time-consuming when the problem structure is complicated. Therefore a theoretical way to choose the threshold can be beneficial, especially for online change-point detection where computational efficiency is of great importance. We will also provide performance estimates for the Subspace-CUSUM test. This will allow for the optimum design of the two parameters w, d and for demonstrating that the resulting detector is asymptotically optimum.

3.3.1 Worst-Case Average Detection Delay

We now consider scenarios that lead to the worst-case detection delay. For the Largest-Eigenvalue Shewhart chart, assume $1 \leq t - w + 1 \leq \tau < t$. Since for the detection we use $\lambda_{\max}(\hat{\Sigma}_{t,w})$ and compare it to a threshold, it is clear that the worst-case data before τ are the ones that will make $\lambda_{\max}(\hat{\Sigma}_{t,w})$ as small as possible. We observe that

$$\begin{aligned}\lambda_{\max}(\hat{\Sigma}_{t,w}) &= \lambda_{\max}(x_{t-w+1}x_{t-w+1}^{\top} + \cdots + x_{\tau}x_{\tau}^{\top} + \cdots + x_t x_t^{\top}) \\ &\geq \lambda_{\max}(x_{\tau+1}x_{\tau+1}^{\top} + \cdots + x_t x_t^{\top}) = \lambda_{\max}(\hat{\Sigma}_{t,t-\tau}),\end{aligned}$$

which corresponds to the data $x_{t-w+1}, \ldots, x_{\tau}$, before the change, being all equal to zero. In fact, the worst-case scenario at any time instant τ is equivalent to forgetting all data before and including τ and restarting the procedure from $\tau + 1$ using up to w outer products in the un-normalized sample covariance matrix, exactly as we do when we start at time 0. Due to stationarity, this suggests that we can limit ourselves to the case $\tau = 0$ and compute $\mathbb{E}_0[T_{\mathrm{E}}]$ and this will constitute the worst-case average detection delay. Furthermore, the fact that

in the beginning we do not normalize with the number of outer products, is beneficial for T_{E} since it improves its ARL.

We should emphasize that if we do not force the data before the change to become zero and use simulations to evaluate the detector with a change occurring at some time different from 0, then it is possible to arrive at misleading conclusions. Indeed, it is not uncommon for this test to appear outperforming the exact CUSUM test for low ARL values. Of course this is impossible since the exact CUSUM is optimum for *any* ARL in the sense that it minimizes the WADD depicted in (2.7) [134].

Let us now consider the worst-case scenario for the Subspace-CUSUM procedure. We observe that

$$\mathcal{S}_t = (\mathcal{S}_{t-1})^+ + (\hat{u}_t^\top x_t)^2 - d \geq 0 + (\hat{u}_t^\top x_t)^2 - d,$$

suggesting that when $\mathcal{S}_t$ restarts this is the worst it can happen for the detection delay. Therefore, the well-known property of the worst-case scenario in the exact CUSUM carries over to Subspace-CUSUM. Again, because of stationarity, this allows us to fix the change-point time at $\tau = 0$. Of course, as mentioned before, because $\hat{u}_t$ uses data coming from the future of t, if our detector stops at some time t (namely when for the first time we experience $\mathcal{S}_t \geq b$) then the *actual* time of stopping must be corrected to $t + w$. A similar correction is not necessary for CUSUM because this test has the exact information for all parameters.

Threshold b is chosen so that the ARL meets a pre-specified value. In practice, b is usually determined by simulation. More specifically, by simulating multiple streams of data from pre-change distribution, we can obtain the ARL for different thresholds. Therefore the threshold can be determined by the simulation results.

A very convenient tool in accelerating the estimation of ARL (which is usually large) is the usage of the following formula that connects the ARL of CUSUM to the average of the

sequential probability ratio test (SPRT) stopping time [180]:

$$\mathbb{E}_\infty[T_{\mathrm{C}}] = \frac{\mathbb{E}_\infty[T_{\mathrm{SPRT}}]}{\mathbb{P}_\infty(S_{T_{\mathrm{SPRT}}} \geqslant b)}, \tag{3.11}$$

where the SPRT stopping time is defined as

$$T_{\mathrm{SPRT}} = \inf\{t > 0 : S_t \notin [0, b]\}.$$

The validity of (3.11) relies on the CUSUM property that, after each restart, S_t is independent of the data before the time of the restart. Unfortunately, this key characteristic is no longer valid in the proposed Subspace-CUSUM scheme since $\hat{u}_t$ uses data from the future of t. We could, however, argue that this dependence is weak. Indeed, as we will see in Lemma A.2, each $\hat{u}_t$ is equal to u plus some small random perturbation (estimation error with the power of the order of $1/w$), with these perturbations being practically independent in time. As we observed with numerous simulations, estimating the ARL directly and through (3.11) (with S_t replaced by $\mathcal{S}_t$), results in almost indistinguishable values even for moderate window sizes w. This suggests that we can use (3.11) to estimate the ARL of the Subspace-CUSUM as well. As we mentioned, in the final result, we need to add w to account for the future data used by the estimate $\hat{u}_t$.

3.3.2 Analysis of Largest-Eigenvalue Shewhart Chart

In this section, we first introduce some connection with random matrix theory, which are the building blocks for the theoretical derivation. Then we provide the approximation to ARL as a function of threshold b after taking into account the *temporal correlation* between detection statistics. The comparison with simulation results shows the high accuracy of our results.

The study of ARL requires the understanding of the property of the largest eigenvalue under the null hypothesis, i.e., the samples are i.i.d. Gaussian random vectors with zero-

mean and identity covariance matrix. In [97], the Tracy-Widom law [197] was used to characterize the distribution of the largest eigenvalue. Define the center constant $\mu_{w,k}$ and scaling constant $\sigma_{w,k}$:

$$\begin{aligned}\mu_{w,k} &= \left(\sqrt{w-1}+\sqrt{k}\right)^2, \\ \sigma_{w,k} &= \left(\sqrt{w-1}+\sqrt{k}\right)\left(\frac{1}{\sqrt{w-1}}+\frac{1}{\sqrt{k}}\right)^{1/3}.\end{aligned} \tag{3.12}$$

If $k/w \to \gamma < 1$, then the centered and scaled largest eigenvalue converges in distribution to a random variable W_1 with the so-called Tracy-Widom law of order one [97]:

$$\frac{\lambda_{\max}(\hat{\Sigma}_w) - \mu_{w,k}}{\sigma_{w,k}} \to W_1. \tag{3.13}$$

The Tracy-Widom law can be described in terms of a partial differential equation and the Airy function, and its tail can be computed numerically (using for example the R-package RMTstat).

Remark 3.3 (Connection with random matrix theory)**.** There has been an extensive literature on the distribution of the largest eigenvalue of the sample covariance matrix, see, e.g., [97, 219, 13, 94]. The so-called *bulk* [54] results are typically used for eigenvalue distributions. It treats a continuum of eigenvalues, and the *extremes*, which are the (first few) largest and smallest eigenvalues. Assume there are w samples which are k-dimensional Gaussian random vectors with zero-mean and identity covariance matrix. Let $\hat{\Sigma}_w = \sum_{i=1}^{w} x_i x_i^\top$ denote the un-normalized sample covariance matrix. If $k/w \to \gamma > 0$, the largest eigenvalue of the sample covariance matrix converges to $w(1+\sqrt{\gamma})^2$ almost surely [70]. Here we use the Tracy-Widom law to characterize its limiting distribution and tail probabilities.

If we ignore the temporal correlation of the largest eigenvalues produced by the sliding window, we can obtain a simple approximation for the ARL. If we call $p = \mathbb{P}_\infty(\lambda_{\max}(\hat{\Sigma}_{t,w}) > b)$ for $t \geq w$ then the probability to stop at t is geometric and it is easy to see that the

ARL can be expressed as $1/p$. We note that to obtain this result, we must assume that $\mathbb{P}_\infty(\lambda_{\max}(\hat{\Sigma}_{t,w}) > b) = p$ for $t < w$ as well, which is clearly not true. Since for $t < w$ the un-normalized sample covariance has less than w terms, the corresponding probability is smaller than p. This suggests that $1/p$ is a lower bound to the ARL while $w + 1/p$ an upper bound. If $w \ll 1/p$, then approximating the ARL with $1/p$ is quite acceptable. We can use the Tracy-Widom law to obtain an asymptotic expression relating the ARL with the threshold b. The desired formula is depicted in the following theorem.

Theorem 3.1 (Approximation of ARL by Ignoring Temporal Correlation)**.** *For any* $0 < p \ll 1$ *we have* $\mathbb{E}_\infty[T_{\mathrm{E}}] \approx 1/p$*, if we select*

$$b = \sigma_{w,k} b_p + \mu_{w,k}, \tag{3.14}$$

where b_p *denotes the* p*-upper-percentage point of* W_1 *namely* $\mathbb{P}(W_1 \geq b_p) = p$*.*

Now we aim to capture the temporal correlation between detection statistics due to overlapping time windows. We leverage a proof technique developed in [179], which can obtain satisfactory approximation for the tail probability of the maximum of a random field.

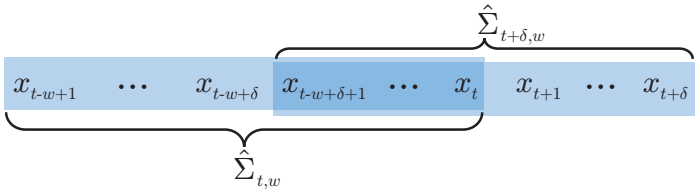


Figure 3.1: Illustration of the temporal correlation between largest eigenvalues, $\delta \in \mathbb{Z}^+$.

Figure 3.1 illustrates the overlap of two sample covariance matrices and provides necessary notation. For each $\hat{\Sigma}_{t,w}$, define $Z_t = \lambda_{\max}(\hat{\Sigma}_{t,w})$. We note that for any given $M > 0$,

$$\mathbb{P}_\infty(T_{\mathrm{E}} \leq M) = \mathbb{P}_\infty\left(\max_{1 \leq t \leq M} Z_t \geq b\right),$$

which is the max over a set of correlated variables $\{Z_t\}_{t=1}^{M}$. Capturing the temporal dependence of $\{Z_t\}$ is challenging. Below, we assume the dimension k and the window size w are fixed, and consider local covariance structure of the detection statistic when they only non-overlap at a small shift δ relative to the window size, i.e., δ/w is small. By leveraging the properties of the local approximation, we can obtain an asymptotic approximation using the localization theorem [179]. Define a special function $\nu(\cdot)$ which is closely related to the Laplace transform of the overshoot over the boundary of a random walk [183]:

$$\nu(x) \approx \frac{\frac{2}{x}[\Phi\big(\frac{x}{2}\big) - 0.5]}{\frac{x}{2}\Phi\big(\frac{x}{2}\big) + \phi\big(\frac{x}{2}\big)}, \tag{3.15}$$

where $\phi(x)$ and $\Phi(x)$ are the probability density function (pdf) and cumulative distribution function (cdf) of the standard normal distribution $\mathcal{N}(0,1)$. Then we have the following results.

Theorem 3.2 (ARL of Largest-Eigenvalue Shewhart Chart)**.** *For large values of b we can write*

$$\mathbb{E}_\infty[T_{\mathrm{E}}] = \left[b'\phi(b')\beta_{k,w}\nu(b'\sqrt{2\beta_{k,w}/w})/w\right]^{-1}\big(1 + o(1)\big), \tag{3.16}$$

where

$$\beta_{k,w} = 1 + \frac{\left(1 + c_1 k^{-\frac{1}{6}}/\sqrt{w}\right)\left(2 + c_1 k^{-\frac{1}{6}}/\sqrt{w}\right)}{c_2^2 k^{-\frac{1}{3}}/w}, \quad b' = \frac{b - (\mu_{w,k} + \sigma_{w,k}c_1)}{\sigma_{w,k}c_2},$$

with $c_1 = \mathbb{E}[W_1] = -1.21$ and $c_2 = \sqrt{\mathrm{Var}(W_1)} = 1.27$.

We perform simulations to verify the accuracy of the threshold values obtained without and with considering the temporal correlation (Theorem 3.1 and Theorem 3.2, respectively). The results are shown in Table 3.1. Compared with the thresholds obtained from Monte Carlo simulation, we find that the threshold in (3.16), when temporal correlation is taken into account, is more accurate than its counterpart obtained by using the Tracy-Widom law in (3.14).

Table 3.1: Comparison of the threshold b obtained from simulations and using the approximations ignoring the correlation in (3.14), and considering the correlation in (3.16). Window length $w = 200$, dimension $k = 10$. The numbers shown are b/w. Approximations that are closer to simulation values are indicated in boldface.

Target ARL	5k	10k	20k	30k	40k	50k
Simulation	1.633	1.661	1.688	1.702	1.713	1.722
Approx (3.14)	1.738	1.763	1.787	1.800	1.809	1.816
Approx (3.16)	**1.699**	**1.713**	**1.727**	**1.735**	**1.740**	**1.744**

We now focus on the detection performance and present a tight lower bound for the EDD of the Largest-Eigenvalue Shewhart chart. The results are based on a known result for CUSUM [180] and requires the derivation of the Kullback-Leibler divergence for our problem.

Theorem 3.3. *For large values of* b *we have*

$$\mathbb{E}_0[T_{\mathrm{E}}] \geq 2\frac{b' + e^{-b'} - 1}{\rho - \log(1+\rho)}\big(1 + o(1)\big), \tag{3.17}$$

where

$$b' = \frac{1}{2\sigma^2(1+\rho)}\left[b\rho - (1+\rho)\sigma^2 \log(1+\rho)\right].$$

Consistent with intuition, in Theorem 3.3, the right-hand-side of (3.17) is indeed a decreasing function of the SNR ρ. Comparing the lower bound in Theorem 3.3 with simulated average delay, as shown in Figure 3.2, we can show that in the regime of small detection delay (which is the main regime of interest), the lower bound serves as a reasonably good approximation.

3.3.3 Analysis of Subspace-CUSUM Procedure

In this section, we focus on how to set the drift parameter d for Subspace-CUSUM procedure and the proof of its asymptotic optimality. The drift d is an important parameter for the Subspace-CUSUM to achieve desired properties of change-point detection algorithms.

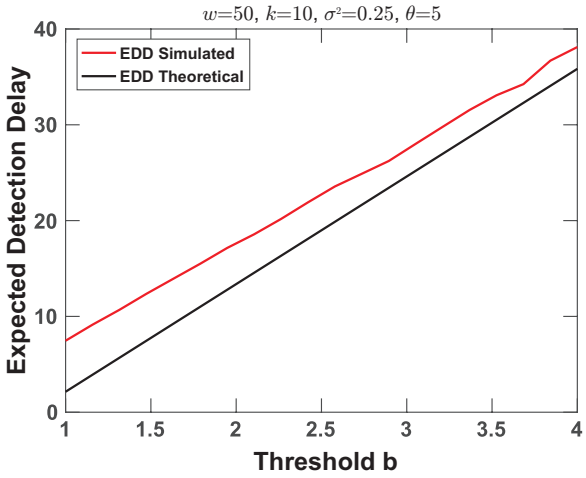


Figure 3.2: Simulated EDD and lower bound as a function of the threshold b.

For the drift parameter d we need the following inequalities to be true:

$$\mathbb{E}_\infty[(\hat{u}_t^\top x_t)^2] < d < \mathbb{E}_0[(\hat{u}_t^\top x_t)^2]. \tag{3.18}$$

With (3.18), we can guarantee that $\mathcal{S}_t$ mimics the behavior of the exact CUSUM statistic S_t mentioned in Remark 3.1, namely, it exhibits a negative drift before and a positive after the change. As we mentioned, the main advantage of using $\hat{\Sigma}_{t+w,w}$ is that it provides estimates $\hat{u}_t$ that are *independent* from x_t. This independence property allows for the straightforward computation of the two expectations in (3.18) and contributes towards the proper selection of d. However, for this computation to be possible, especially under the alternative regime, it is necessary to be able to describe the statistical behavior of our estimate $\hat{u}_t$. We will assume that the window size w is sufficiently large so that Central Limit Theorem (CLT) type approximations are possible for $\hat{u}_t$ and we will consider that $\hat{u}_t$ is actually Gaussian with mean u (the correct vector) and (error) covariance matrix that can be specified, analytically, of being size $1/w$ [7, 142].

Lemma 3.1. *Adopting the Gaussian approximation for $\hat{u}_t$ we have the following two mean*

values under the pre- and post-change regime:

$$\mathbb{E}_\infty[(\hat{u}_t^\top x_t)^2] = \sigma^2, \quad \mathbb{E}_0[(\hat{u}_t^\top x_t)^2] \ = \sigma^2(1+\rho)\left[1 - \frac{k-1}{w\rho}\right]. \tag{3.19}$$

Lemma 3.1 also suggests that the window size w and the drift d must satisfy $\sigma^2 < d < \sigma^2(1+\rho)\left(1 - \frac{k-1}{w\rho}\right)$. Necessary condition for this to be true is that $w > (k-1)(1+\rho)/\rho^2$. Actually this constraint is required for the Gaussian approximation to make sense. But in order for the approximation to be efficient we, in fact, need w to be significantly larger than the lower bound. We can see that when the SNR is high ($\rho \gg 1$) then with relatively small window size we can obtain efficient estimates. When on the other hand SNR is low ($\rho \ll 1$) then far larger window sizes are necessary to guarantee validity of the Gaussian approximation.

Consider now the case where ρ is *unknown* but exceeds some pre-set minimal SNR $\rho_{\min}$ of interest. From the above derivation, given the worst-case SNR and an estimation for the noise variance $\hat{\sigma}^2$, we can give a lower bound for $\mathbb{E}_0[(\hat{u}_t^\top x_t)^2]$. Consequently, the drift d can be anything between $\hat{\sigma}^2$ and $\hat{\sigma}^2(1+\rho_{\min})(1-(k-1)/(w\rho_{\min}))$ where, we observe, that the latter quantity exceeds $\hat{\sigma}^2$ when $w > (k-1)(1+\rho_{\min})/\rho_{\min}^2$. Below, for simplicity, for d we use the average of the two bounds. It is worthwhile mentioning that the lower and upper bound in Lemma 3.1 are derived based on the assumption that the window size w is large enough.

Remark 3.4 (Monte Carlo simulation to choose the threshold)**.** Alternatively, and in particular when w does not satisfy $w \gg k$, we can estimate $\mathbb{E}_0[(\hat{u}_t^\top x_t)^2]$ by Monto Carlo simulation. This method requires: (i) estimating the noise level $\hat{\sigma}^2$, which can be obtained from training data without a change-point; (ii) the pre-set worst-case SNR $\rho_{\min}$; (iii) a unit norm vector u_0 that is generated randomly. Under the nominal regime we have $\mathbb{E}_\infty[(\hat{u}_t^\top x_t)^2] = \hat{\sigma}^2$. Under the alternative $\mathbb{E}_0[(\hat{u}_t^\top x_t)^2]$ depends only on the SNR ρ as shown in (3.19). We can therefore simulate the worst-case scenario $\rho_{\min}$ using the randomly gen-

erated vector u_0 by generating samples from the distribution $\mathcal{N}(0, \hat{\sigma}^2 I_k + \rho_{\min} u_0 u_0^\top)$.

Even though the average of the update in (3.9) does not depend on true subspace u, the computation of the test statistic $\mathcal{S}_t$ in (3.9) requires the estimate $\hat{u}_t$ of the eigenvector. This can be accomplished by applying the singular value decomposition (SVD) (or the power method [128]) on the un-normalized sample covariance matrix $\hat{\Sigma}_{t+w,w}$.

We then provide performance estimates for the proposed Subspace-CUSUM test. This will allow for the optimum design of the two parameters w, d and for demonstrating that the resulting detector is asymptotically optimum.

From [196, Pages 396–397] we have that the exact CUSUM test has the following performance

$$\mathbb{E}_\infty[T_{\mathrm{C}}] = \frac{e^b}{\mathsf{K}}\big(1 + o(1)\big), \quad \mathbb{E}_0[T_{\mathrm{C}}] = \frac{b}{\mathsf{I}_0}\big(1 + o(1)\big), \tag{3.20}$$

where b is the constant threshold; K is of the order of a constant with its exact value being unimportant for the asymptotic analysis; finally I_0 is the Kullback-Leibler information number $\mathsf{I}_0 = \mathbb{E}_0[\log(f_0(x)/f_\infty(x))]$. We recall that the worst-case average detection delay in CUSUM is equal to $\mathbb{E}_0[T_{\mathrm{C}}]$, as detailed in Chapter 2.1. This is the reason we consider the computation of this quantity. If now, we impose the constraint that the ARL is equal to $\gamma > 1$ and for the asymptotic analysis that $\gamma \to \infty$, then we can compute the threshold b that can achieve this false alarm performance namely $b = (\log \gamma)\big(1 + o(1)\big)$. Substituting this value of the threshold in EDD we obtain

$$\mathbb{E}_0[T_{\mathrm{C}}] = \frac{\log \gamma}{\mathsf{I}_0}\big(1 + o(1)\big). \tag{3.21}$$

Applying this formula in our problem we end up with the following optimum performance

$$\mathbb{E}_0[T_{\mathrm{C}}] = \frac{2 \log \gamma}{\rho - \log(1 + \rho)}\big(1 + o(1)\big). \tag{3.22}$$

For the performance computation of Subspace-CUSUM, since the increment $(\hat{u}_t^\top x)^2 - d$

in (3.9) is not a log-likelihood, we cannot use (3.21) directly. To compute the ARL of T_{SC} we first find $\delta_\infty > 0$ from the solution of the equation:

$$\mathbb{E}_\infty[e^{\delta_\infty[(\hat{u}_t^\top x_t)^2 - d]}] = 1 \tag{3.23}$$

and then we note that $\delta_\infty[(\hat{u}_t^\top x)^2 - d]$ is the log-likelihood ratio between the two pdfs $\tilde{f}_0 = \exp\{\delta_\infty[(\hat{u}_t^\top x)^2 - d]\} f_\infty$ and f_∞. This allows us to compute the threshold b asymptotically as $b = (\log\gamma)\big(1 + o(1)\big)/\delta_\infty$ after assuming that $w = o(\log\gamma)$. Similarly we can find a $\delta_0 > 0$ and define $\tilde{f}_\infty = \exp\{-\delta_0[(\hat{u}_t^\top x_t)^2 - d]\} f_0$ so that $\delta_0[(\hat{u}_t^\top x_t)^2 - d]$ is the log-likelihood ratio between f_0 and $\tilde{f}_\infty$ leading to $\mathbb{E}_0[T_{\mathrm{SC}}] = b\big(1 + o(1)\big)/(\mathbb{E}_0[(\hat{u}_t^\top x_t)^2] - d)$ with the dependence on δ_0 being in the $o(1)$ term. Substituting b we obtain

$$\mathbb{E}_0[T_{\mathrm{SC}}] = \frac{\log\gamma}{\delta_\infty\big(\mathbb{E}_0[(\hat{u}_t^\top x_t)^2] - d\big)}\big(1 + o(1)\big) + w, \tag{3.24}$$

where the last term w is added because we use data from the future of t as we explained before. Parameter δ_∞, defined in (3.23), is directly related to d. We show in the Appendix that d can be expressed in terms of δ_∞ as follows

$$d = -\frac{1}{2\delta_\infty}\log(1 - 2\sigma^2\delta_\infty). \tag{3.25}$$

By (3.25), we obtain the following expression for the EDD:

$$\mathbb{E}_0[T_{\mathrm{SC}}] = \frac{\log\gamma(1+o(1))}{\sigma^2\delta_\infty(1+\rho)\big(1-\frac{k-1}{w\rho}\big)+\frac{1}{2}\log(1-2\sigma^2\delta_\infty)} + w. \tag{3.26}$$

Note that in the previous equation we have two parameters δ_∞ and w and the goal is to select them so as to minimize the EDD. Therefore if we first fix the window size w we can minimize over δ_∞ (which is equivalent to minimizing with respect to the drift d). We observe that the denominator is a concave function of δ_∞ therefore it exhibits a single

maximum. The optimum δ_∞ can be computed by taking the derivative and equating to 0 which leads to a particular δ_∞. Substituting this optimal value we obtain the following minimum EDD:

$$\mathbb{E}_0[T_{\mathrm{SC}}] = \frac{2\log\gamma(1+o(1))}{(1+\rho)\left(1-\frac{k-1}{w\rho}\right)-1-\log\left[(1+\rho)\left(1-\frac{k-1}{w\rho}\right)\right]} + w. \tag{3.27}$$

Equation (3.27) involves only the target ARL level γ and the window size w. If we keep w constant it is easy to verify that the ratio of the EDD of the proposed scheme over the EDD of the optimum CUSUM tends, as $\gamma \to \infty$, to a quantity which is strictly greater than 1. In order to make this ratio tend to 1 and therefore establish asymptotic optimality we need to select the window size w as a function of γ. Actually we can perform this selection optimally by minimizing (3.27) with respect to w for given γ. The following proposition identifies the optimum window size.

Proposition 3.1. *For each* ARL *level γ, the optimal window size that minimizes the corresponding* EDD *is given by*

$$w^* = \sqrt{\log\gamma} \cdot \frac{\sqrt{2(k-1)}}{\rho - \log(1+\rho)}\big(1+o(1)\big),$$

resulting in an optimal drift

$$d^* = \frac{\sigma^2(1+\rho)\left(1-\frac{k-1}{w^*\rho}\right)}{(1+\rho)\left(1-\frac{k-1}{w^*\rho}\right)-1} \log\left[(1+\rho)\left(1-\frac{k-1}{w^*\rho}\right)\right].$$

Using these optimal parameter values it is straightforward to establish that the corresponding Subspace-CUSUM is first-order asymptotically optimum. This is summarized in our next theorem.

Theorem 3.4. *As the ARL level $\gamma \to \infty$, the corresponding* EDD *of the Subspace-CUSUM*

procedure T_{SC} with the two parameters d and w optimized as above satisfies

$$\lim_{\gamma\to\infty} \frac{\mathbb{E}_0[T_{\mathrm{SC}}]}{\mathbb{E}_0[T_{\mathrm{C}}]} = 1. \tag{3.28}$$

Proof. As we pointed out, the proof is straightforward. Indeed if we substitute the optimum d and w and then take the ratio with respect to the optimum CUSUM performance depicted in (3.22) we obtain

$$\frac{\mathbb{E}_0[T_{\mathrm{SC}}]}{\mathbb{E}_0[T_{\mathrm{C}}]} = 1 + \sqrt{\frac{k-1}{2\log\gamma}} + o(1) \to 1,$$

which proves the desired limit. Even though the ratio tends to 1, we note that $\mathbb{E}_0[T_{\mathrm{SC}}] - \mathbb{E}_0[T_{\mathrm{C}}] = \Theta(\sqrt{\log\gamma}) \to \infty$. This is corroborated by our simulations (see Figure 3.4, red curve). □

3.4 Simulation Study

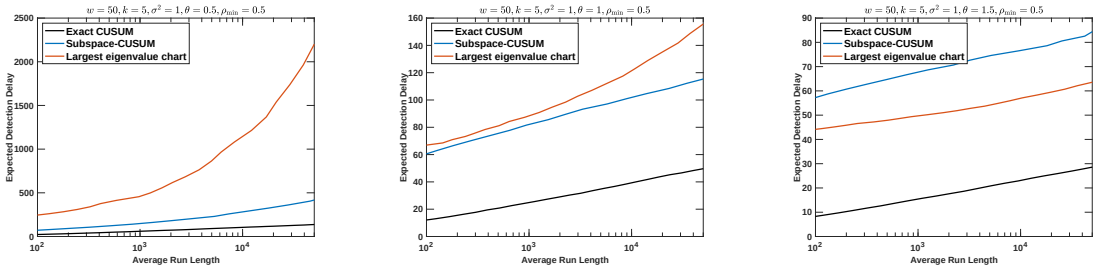
In this section, numerical results are presented to compare the proposed detection procedures. The tests are first applied to synthetic data, where the performance of the Subspace-CUSUM and Largest-Eigenvalue Shewhart chart are compared against the CUSUM optimum performance. Then the performance of Subspace-CUSUM is optimized by selecting the most appropriate window size.

3.4.1 Performance Comparison

Simulation studies are performed to compare the Largest-Eigenvalue Shewhart chart and the Subspace CUSUM procedure. The exact CUSUM procedure with all parameters known is chosen as the baseline and gives the minimal detection delay to all detection procedures.

Figure 3.3 depicts the EDD-ARL curves for parameter values $k = 5$, $\sigma^2 = 1$, $w = 50$ and three different levels of signal strength (SNR): $\theta = 0.5$, $\theta = 1$, and $\theta = 1.5$. For fair comparison, the SNR lower bound is set to be a constant $\rho_{\min} = 0.5$ in all three scenarios.

The threshold for each procedure is determined using the pre-set lower bound $\rho_{\min}$ as



(a) $\theta = 0.5$ (b) $\theta = 1$ (c) $\theta = 1.5$

Figure 3.3: Comparison of Subspace-CUSUM and the Largest-Eigenvalue Shewhart chart, fixed window size $w = 50$. Baseline: Exact CUSUM (optimal).

discussed in Remark 3.4. In Figure 3.3, the black line corresponds to the exact CUSUM procedure, which is clearly the best and it lies below the other curves. Subspace-CUSUM has much smaller EDD than the Largest-Eigenvalue Shewhart chart, and the difference increases with increasing ARL for SNR $\theta = 0.5$ and $\theta = 1$. However, when the signal is stronger ($\theta = 1.5$), the Largest-Eigenvalue Shewhart chart outperforms the Subspace-CUSUM as shown in Figure 3.3 (c). This is consistent with previous research findings in [137] that Shewhart charts are more efficient when detecting strong signals, while the CUSUM-type charts can detect weak signals more quickly due to its cumulative structure.

3.4.2 Optimal Window Size

We also consider the EDD-ARL curve when w is optimized to minimize the detection delay at every ARL value. We first compute the EDD for window sizes $w = 1, 2, \ldots, 50$, given each ARL value. Then we plot in Figure 3.5 (a) the lower envelope of EDDs corresponding to the optimal EDD achieved by varying w. We also plot the optimal value of w as a function of ARL in Figure 3.5 (b). The comparison of the Largest-Eigenvalue Shewhart chart and the Subspace-CUSUM procedure with optimal window size w is summarized in Figure 3.4. Even though the best EDD of the Subspace-CUSUM is diverging from the performance enjoyed by CUSUM, this divergence is slower than the increase of the optimum CUSUM EDD, as shown in Theorem 3.4.

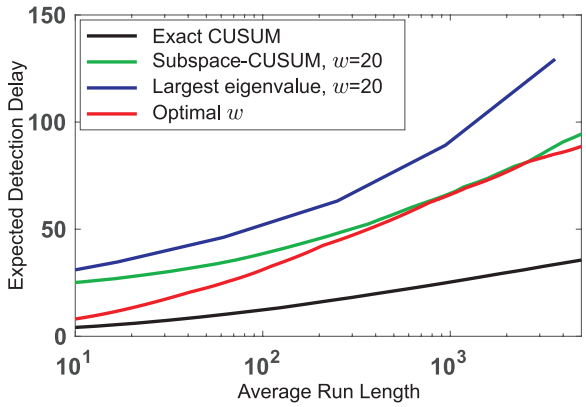
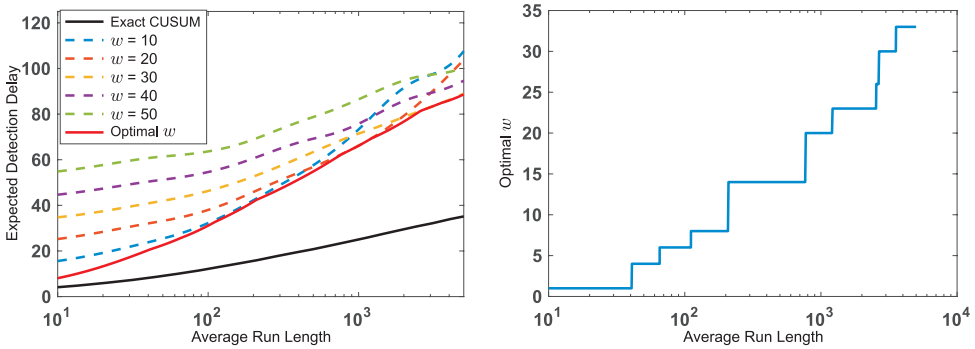


Figure 3.4: Comparison of the largest eigenvalue procedure and CUSUM procedures.



(a) Minimal EDD (b): Optimal window size

Figure 3.5: (a): Minimal EDD vs ARL among window sizes w from 1 to 50; (b): Corresponding optimal window size w.

3.5 Real Data Examples

In this section, we show how to apply the proposed methods to real data problems and demonstrate the performance using two real datasets: one is the human gesture detection dataset, the other is a seismic dataset. It is worth mentioning that the subspace model formulation is a fundamental problem in high dimensional problems, and the proposed methods are widely applicable to a variety of applications.

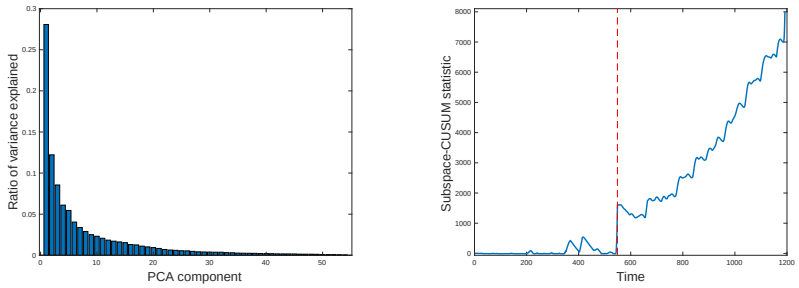
3.5.1 Human Gesture Detection

We apply the proposed method to the sequential posture detection problem using a real dataset: the "Microsoft Research Cambridge-12 Kinect gesture" dataset [61]. The cross-correlation structure of such multivariate functional data may change over time due to the posture change. [221] studies the same dataset from the dynamic subspace learning perspective in the offline setting, our goal is to detect the change-point from sequential observations. This dataset contains 18 sensors. At each time t, each sensor records the coordinates in the three-dimensional cartesian coordinate system. Therefore there are 54 attributes in total.

We select a subsequence with a posture change from "bow" to "throw", and we use the first 250 training samples to estimate the subspace before the change. Figure 3.6 (a) shows the eigenvalues resulted from the principal component analysis (PCA). We select r leading eigenvectors of the sample covariance matrix as our estimate of the pre-change subspace. For example, when $r = 1$, we estimate the pre-change subspace to be a rank one space characterized by the leading eigenvector of the sample covariance matrix of training samples. Then we normalize the observations by multiplying them with a matrix Q that is orthogonal to the pre-change subspace, as discussed in Section 3.1. This enables us to approximate the covariance of pre-change observations by an *identity matrix*. Then we apply the proposed Subspace-CUSUM procedure to detect the change.

The detection statistic is shown in Figure 3.6 (b,c) for different r values, the detection statistic both increase significantly at the true change-point time (indicated by the red dash line). It also shows that the proposed test performs well not only when $r = 1$, but also for $r > 1$ cases, which means that although we focus on the rank one case in the previous theoretical discussion, the propose method can be widely used in many problems that involves such low-rank change.

We also compare the proposed method with Hotelling's T^2 control chart [88]. We use the same training data to estimate the pre-change mean $\bar{\mu}$ and covariance matrix $\bar{\Sigma}$, and then



(a) Eigenvalues (b) $r = 1$

(c) $r = 10$ (d) Hotelling T^2

Figure 3.6: (a): PCA Eigenvalues; (b,c): Subspace-CUSUM statistic over time; (d): Hotelling's T^2 statistic. True change-point indicated by red line.

construct the Hotelling's T^2 statistics $(x_t - \bar{\mu})^\top \bar{\Sigma}^{-1}(x_t - \bar{\mu})$. As shown in Figure 3.6 (d), the detection statistic has a much larger vibration then the Subspace-CUSUM procedure and the performance is sensitive to the estimation of $\bar{\mu}$ and $\bar{\Sigma}$.

3.5.2 Seismic Event Detection

In this example, we consider a seismic signal detection problem. The goal is to detect micro-earthquakes and tremor-like signals, which are weak signals caused by minor subsurface changes in the earth. The tremor signal may be seen at a subset of sensors, and the affected sensors observe a similar waveform corrupted by noise. The tremor signals are not earthquakes, but they are useful for geophysical study and prediction of potential earthquakes. Usually, the tremor signals are challenging to detect using an individual sensor's data; therefore, network detection methods have been developed, which mainly uses covariance information of the data for detection [122]. We will show that network-based detection can be cast as a subspace detection problem.

Assume that we have N sensors. At an unknown onset, the tremor signal may affect all sensors. Let $s(t)$ be the unknown signal waveform , then the signal observed at sensors can be represented as

$$x_i(t) = u_i s(t-\tau) + w_i(t),\ i = 1, 2, \ldots, n, \tag{3.29}$$

where $w_i(t) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ denotes the random noise, $u_i > 0$ is the unknown *deterministic* magnitude of the signal, and τ is the unknown change-point or the time when the seismic event happens. Here the waveform function $s(t)$ is assumed to be causal, i.e., $s(t) = 0, \forall t < 0$. Moreover, we suppose the signal waveform at time time follows a zero-mean normal distribution with time-varying variance (vibration), i.e., $s(t) \sim \mathcal{N}(0, \sigma_t^2)$.

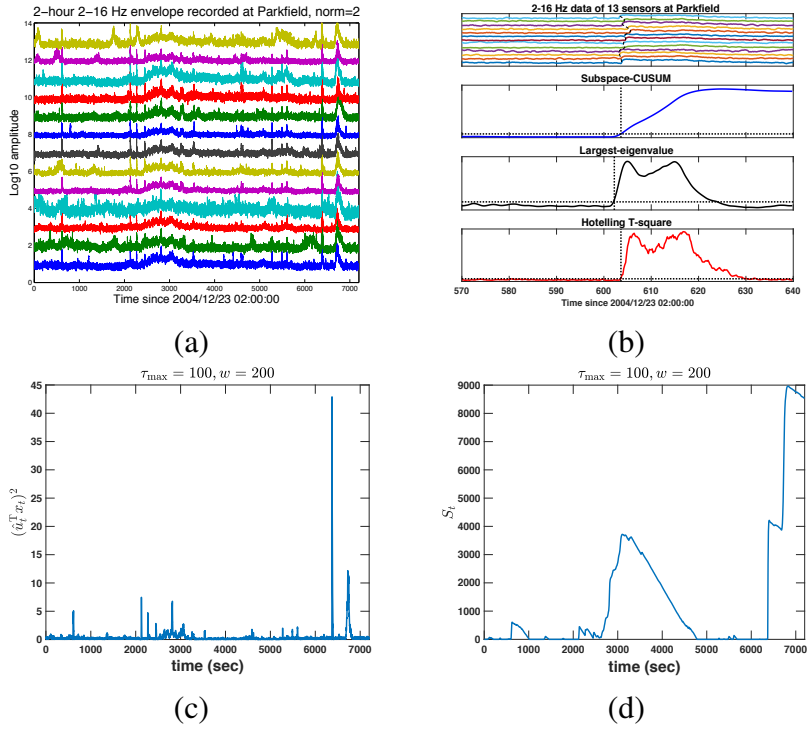


Figure 3.7: (a) the raw data; (b) comparison of different detection procedures; (c) increment term; (d) Subspace-CUSUM statistic.

Denote the observation vector $X(t)$ and magnitude u as

$$X(t) := [x_1(t), \ldots, x_n(t)]^\top,\ u := [u_1, \ldots, u_n]^\top.$$

Following (3.29), we can formulate the problem as follows

$$\begin{aligned} X(t) &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_n), \qquad\qquad t = 1, 2, \ldots, \tau, \\ X(t) &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_n + \sigma_t^2 u u^\top), \, t = \tau + 1, \tau + 2, \ldots. \end{aligned} \tag{3.30}$$

We apply the proposed methods to a real seismic dataset recorded at Parkfield, California from 2am to 4am on 12/23/2004. The raw data contains records at 13 seismic sensors that simultaneously records a continuous stream of data. The frequency of the raw data is 250Hz. In this example, we set the window size $w = 200$, which corresponds to a $0.8s$ time window. For each procedure, we use the data within the first 600s to find the threshold by controlling the false alarm rate.

We apply the proposed Largest-Eigenvalue Shewhart chart and the Subspace-CUSUM procedure. We further compare them with the classic Hotelling's T^2 procedure based on the estimated sample mean and sample covariance. The results are shown in Figure 3.7 (b). Using the detection statistics in Figure 3.7 (c,d), we find three main events at the 615, 2127, and 6371 seconds, as well as some continuous vibration during 2500$\sim$3200 seconds. By comparing the detection results with the true seismic event catalog that can be found online at the Northern California Earthquake Data Center, we found that our findings match the three true events at 594, 2124 and 6369 seconds, along with a tremor catalog around 2500$\sim$3180 seconds. The comparison shows that all detection delays are within 20 seconds. Both the Largest-Eigenvalue Shewhart chart and the Subspace-CUSUM procedure work for this dataset effectively.

3.6 Conclusion and Discussions

We study two online detection procedures for detecting the emergence of a spiked covariance model: the Largest-Eigenvalue Shewhart chart and the Subspace-CUSUM procedure. For Subspace-CUSUM, we perform a simultaneous estimate of the required subspace in parallel with its sequential detection. We avoid estimating all unknown parameters by fol-

lowing a worst-case analysis with respect to the subspace power. We are able to derive theoretical expressions for the ARL and an interesting lower bound for the EDD of the Largest-Eigenvalue Shewhart chart. In particular, we are able to handle the correlations resulted from the usage of a sliding window, which is an issue that is not present in the off-line version of the same procedure. For the comparisons of the two proposed detection procedures, we discuss how it is necessary to calibrate each detector so that comparisons are fair. Comparisons are performed using simulated data and real data about human gesture detection and seismic event detection.

CHAPTER 4

SEQUENTIAL CHANGE DETECTION BY WEIGHTED ℓ_2 DIVERGENCE

This chapter studies the non-parametric sequential change detection based on weighted ℓ_2 divergence. This work is mainly summarized in [210]. Section 4.1 introduces the problem set-up and proposes the weighted ℓ_2 test and shows its optimality in ℓ_2 sense. Section 4.2 studies sequential change detection using the proposed statistic. Section 4.3 discusses different aspects to optimize the parameters involved in the proposed method. Section 4.4 and Section 4.5 demonstrate the performance of the proposed detection procedure using simulation and real-data studies.

4.1 Problem Setup and Weighted ℓ_2 Divergence Test

We consider the problem of testing closeness between two discrete distributions from samples observed. Let Ω be an n-element observation space, identified with $\{1, ..., n\}$. Probability distributions on Ω are denoted as vector $p \in \Delta_n = \{p \in \mathbb{R}^n : p \geq 0, \sum_i p_i = 1\}$, where p_i is the probability mass of the i-th element in Ω. Given two independent sample sets $X^1 = \{x_1^1, \ldots, x_{n_1}^1\}$ and $X^2 = \{x_1^2, \ldots, x_{n_2}^2\}$, where $x_1^1, \ldots, x_{n_1}^1 \overset{\text{iid}}{\sim} p$, and $x_1^2, \ldots, x_{n_2}^2 \overset{\text{iid}}{\sim} q$, our goal is to design a *test* which, given observations X^1 and X^2, claims one of the following hypotheses:

$$H_0 : p = q, \quad H_1 : \|p - q\|_2 \geq \epsilon \|p\|_2,$$

where $\| \cdot \|_2$ is the ℓ_2 norm on $\mathbb{R}^n$ and $\epsilon > 0$ is a parameter that represents the relative difference of magnitude. Note that the alternative hypothesis H_1 considered here is slightly different but closely related to the traditional setting where H_1 is defined as $\|p - q\|_2 \geq \epsilon$.

Define the type-I risk of a test as the probability of rejecting hypothesis H_0 when it is

true, i.e., the probability of claiming $\|p-q\|_2 \geq \epsilon\|p\|_2$ when $p=q$. The type-II risk is the probability of claiming $p=q$ when $\|p-q\|_2 \geq \epsilon\|p\|_2$. We aim at building a test which, given $0<\alpha,\beta<1/2$, has the type-I risk at most α (which we call *at level* α), and the type-II risk at most β (*of power* $1-\beta$); and we aim to meet these specifications with sample sizes n_1 and n_2 as small as possible.

We focus on a family of distance-based divergence between empirical distributions of the two sets of observations. More specifically, we consider tests that reject the null hypothesis H_0 (and accept the alternative H_1) when

$$D(X^1, X^2) > \ell,$$

where $D(\cdot,\cdot)$ is a proxy for the weighted ℓ_2 divergence between distributions p and q underlying X^1 and X^2, and ℓ is a data-dependent (random) threshold.

The motivation for considering the ℓ_2 divergence for the non-parametric test is twofold. First, the ℓ_2 divergence-based test has a certain (near) optimality that we will show in Section 4.1.3. Second, the ℓ_2 divergence is more *robust* compared to other divergences such as χ^2-divergence, which is commonly used when ℓ_1 separation between distributions is of interest. The χ^2-divergence becomes numerically difficult to evaluate when there are "small" p_i (meaning some atoms have small probability), while the ℓ_2 distance remains bounded in such cases. Similar argument holds for the α– and β–divergences [44, 163, 18, 47, 5] and detection statistic for robust change detection [222]. Moreover, here we focus on a new *weighted* ℓ_2 divergence, which emphasizes atoms that contribute most to $\|p-q\|_2$.

Our goal in this section is to develop a test statistic, the weighted ℓ_2 divergence, used as the basic building block of the change detection procedure. We aim to construct a test with the following properties. When applied to two independent sets of size N, i.i.d. samples $\{x_1^1, ..., x_N^1\}$ and $\{x_1^2, ..., x_N^2\}$ drawn from unknown distributions $p, q \in \Delta_n$, the test

(i) rejects the null hypothesis with probability at most a given α under $H_0 : p = q$;

(ii) accepts the null hypothesis with probability at most a given β when there is a relative difference "of magnitude at least a given $\epsilon > 0$," i.e., under $H_1 : \|p - q\|_2 \geq \epsilon \|p\|_2$.

We want to meet these reliability specifications with as small sample size N as possible.

4.1.1 Test Statistic

The main ingredient of weighted ℓ_2 divergence test is the *individual test* built as follows. Let us fix "weights" $\sigma_i \geq 0$, $i = 1, ..., n$, and let $\Sigma = \text{Diag}\{\sigma_1, ..., \sigma_n\}$ be a diagonal matrix with diagonal entries being $\sigma_1, \ldots, \sigma_n$. Given $\{x_1^1, ..., x_N^1\}$ and $\{x_1^2, ..., x_N^2\}$, we divide them into two *consecutive* (left) parts E, E', of cardinality L each, and (right) parts F, F', of cardinality R each, respectively. Note that the cardinality L and R are at most $N/2$ and can be less than $N/2$ if we do not use all N samples. Set

$$\gamma = \frac{R}{L+R}, \bar{\gamma} = 1 - \gamma = \frac{L}{L+R}, M = \frac{2LR}{L+R} = 2\gamma L = 2\bar{\gamma} R. \tag{4.1}$$

Let $\omega, \omega', \zeta, \zeta' \in \Delta_n$ be the empirical distributions of observations in sets E, E', F, F', and χ be the weighted ℓ_2 test statistics defined as

$$\chi = (\omega - \zeta)^\top \Sigma (\omega' - \zeta') = \sum_{i=1}^{n} \sigma_i (\omega_i - \zeta_i)(\omega_i' - \zeta_i'). \tag{4.2}$$

The *weighted ℓ_2 divergence test* $\mathcal{T}$ claims a change if and only if

$$|\chi| > \ell,$$

where ℓ is the threshold. The following lemma summarizes the properties of $\mathcal{T}$:

Proposition 4.1 (Test Properties)**.** *Let $\mathcal{T}$ be the weighted ℓ_2 divergence test applied to a pair of samples drawn from distributions $p, q \in \Delta_n$, and let the threshold ℓ satisfy*

$$\ell \geq 2\sqrt{2}\theta M^{-1} \sqrt{\textstyle\sum_i \sigma_i^2 p_i^2}, \tag{4.3}$$

for some $\theta \geq 1$. Then

1. *Risk: The type-I risk of $\mathcal{T}$ is at most $1/\theta^2$;*

2. *Power: Under the assumption*

$$\sum_i \sigma_i(p_i - q_i)^2 > \ell + 2\sqrt{2}\theta \Big[M^{-1/2}\sqrt{\sum_i \sigma_i^2(p_i - q_i)^2(\gamma p_i + \bar{\gamma} q_i)} + M^{-1}\sqrt{\gamma \sum_i \sigma_i^2 p_i^2 + \bar{\gamma}\sum_i \sigma_i^2 q_i^2}\Big], \tag{4.4}$$

the power of $\mathcal{T}$ is at least $1 - 3/\theta^2$.

For simplicity, in the rest of this section we assume that $\sigma_i = 1$, $1 \leq i \leq n$, so the left hand side of (4.4) reduces to $\|p - q\|_2^2$. In Section 4.3.1, we will discuss how to utilize the non-uniform weights.

4.1.2 Special Case: ℓ_2 Test With Uniform Weights

The individual test $\mathcal{T}$ in the previous section has two drawbacks: (i) to control the type-I risk, the threshold ℓ in (4.3) specifying $\mathcal{T}$ must be chosen with respect to the magnitude $\|p\|_2$ which is typically *unknown*; (ii) to achieve a small type-I risk of $\mathcal{T}$ we need to set a large θ, thus resulting in poor power of the test. This section will show that we can reduce these limitations by "moderately" increasing the sample sizes. To simplify the notation, from now on, we use the fixed value $\theta = 3$ (i.e., the type-I risk is at most $1/9$ and the power is at least $2/3$), and use $M = L = R$ as a special case of the definition in (4.1).

The testing procedure will be as follows. We first give the Algorithm 1 to specify the threshold ℓ that satisfies the condition (4.3) with high probability and then introduce the testing procedure.

When the nominal distribution p is *unknown*, we perform a *training-step* – use part of the first set of observations to build, with desired reliability $1 - \delta$, a tight upper bound ϱ (the output of Algorithm 1) on the squared norm $\|p\|_2^2$ of the unknown distribution p such

that

$$\mathbb{P}\left[\|p\|_2^2 \leq \varrho \leq 3\|p\|_2^2\right] \geq 1-\delta, \tag{4.5}$$

where the probability is taken with respect to the observations sampled from distribution p.

The training-step is organized in Algorithm 1, where the input parameter S is defined as

$$S := \min\left\{S \in \mathbb{N} : \sum_{k=S}^{2S} \binom{2S}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{2S-k} \leq \frac{\delta}{\lceil \log_2(n) \rceil}\right\}. \tag{4.6}$$

The definition in (4.6) has an intuitive explanation: S is the smallest number such that in $2S$ independent tosses of a coin, with probability of getting a head in each toss being $\leq 1/3$, the probability of getting at least S heads does not exceed δ/m, where $m = \lceil \log_2(n) \rceil$.

Algorithm 1: Training-step to estimate a tight upper bound on $\|p\|_2^2$.

Input: Samples $X^1 := \{x_1, \ldots, x_N\}$; Reliability $1-\delta$; $m = \lceil \log_2(n) \rceil$; S in (4.6);
Output: A tight upper bound ϱ on $\|p\|_2^2$ satisfying the condition (4.5);
for $i = 1, \ldots, m$ **do**
  $\rho_i = 2^{-i/2}$;
  Set $P_i \in \mathbb{R}_+$ as the solution to

$$3\left[2^{7/4} P_i^{-1/2} \rho_i^{3/2} + 2P_i^{-1}\rho_i\right] = \frac{1}{3}\rho_i^2;$$

  Set $Q_i = \lceil P_i \rceil$;
  Use $4S$ consecutive segments, of cardinality Q_i each, of the sample X^1 to build $2S$ pairs $\{(\xi_s, \xi_s'), s = 1, \ldots, 2S\}$ of empirical distributions;
  Set $\theta_s = \xi_s^\top \xi_s'$ for $s = 1, \ldots, 2S$;
  Set Θ_i as the median of $\theta_1, ..., \theta_{2S}$;
  if $\Theta_i \geq 2\rho_i^2/3$ *or* $i = m$ *or* $N = |X^1| < 4SQ_{i+1}$ *(running out of sample)* **then**
    Terminate.
  end
end
Return $\varrho = \Theta_i + \rho_i^2/3$.

Properties of the training-step in Algorithm 1 can be summarized as follows:

Proposition 4.2 (Bounding $\|p\|_2^2$)**.** *Let $\rho_i = 2^{-i/2}$ and $i(p)$ be the smallest $i \leq m$ such that $\rho_i \leq \|p\|_2$ (note that $i(p)$ is well defined due to $\rho_m \leq n^{-1/2}$). Assume that the size of the first group of sample X^1 is at least $4SQ_{i(p)}$. Then the probability for the training-step to*

terminate in the first $i(p)$ stages and to output ϱ satisfying the condition (4.5) *is at least $1-\delta$, where δ is the reliability tolerance specifying the training-step. Besides this, the number of observations utilized in a successful training-step is at most*

$$4SQ_{i(p)} = O(1)\ln(\ln(n)/\delta)/\|p\|_2. \tag{4.7}$$

After ϱ is built, we use the part of the first sample X^1 *not* used in the training-step and the entire second sample X^2 to run $K = K(\alpha, \beta)$ individual tests to make a decision. Here $\alpha < 1/2$ and $\beta < 1/2$ are pre-specified upper bounds on the type-I and type-II risks of the testing problem, and $K(\alpha, \beta)$ is the smallest integer such that the probability of getting at least $K/2$ heads in K independent tosses of a coin is

(i) $\leq \alpha$, when the probability of getting head in a single toss is $\leq 1/9$,

(ii) $\geq 1-\beta$, when the probability of getting head in a single toss is $\geq 2/3$.

It is easy to check that $K \leq O(1)[\ln(1/\alpha) + \ln(1/\beta)]$.

The k-th individual test is applied to two $2M$-long segments of observations taken first from the sample X^1 (and these are non-overlapping with the training-step observations), and second from X^2, with non-overlapping segments of observations used in different individual tests. Here the positive integer M, same as the reliability tolerances δ, α, β, is a parameter of our construction, and the threshold ℓ for individual tests is chosen as

$$\ell = 6\sqrt{2}M^{-1}\sqrt{\varrho}. \tag{4.8}$$

After running K individual tests, we claim H_1 if and only if the number of tests where H_1 is claimed is at least $K/2$. The properties of the resulting ℓ_2 test are presented as follows:

Theorem 4.1 (Sample Complexity)**.** *Consider the ℓ_2 test above with design parameters δ, $\alpha, \beta \in (0, 1/2)$ and M. Then for properly selected absolute constants $O(1)$, the following*

holds true. Let p, q be the true distributions from which X^1 and X^2 are sampled, and let the size N of X^1, X^2 satisfies

$$N \geq O(1)\left[\ln(\ln(n)/\delta)/\|p\|_2 + [\ln(1/\alpha) + \ln(1/\beta)]M\right]. \tag{4.9}$$

Then

1. *The probability for the training-step in Algorithm 1 to be successful is at least $1 - \delta$, and when it happens there are enough observations to carry out $K = K(\alpha, \beta)$ subsequent individual tests.*

2. *Under the condition that the training-step is successful:*

   (a) The type-I risk (claiming H_1 when $p = q$) is at most α;

   (b) For every $\epsilon > 0$, with positive integer M satisfying

$$M \geq O(1)\frac{1}{\epsilon^2\|p\|_2}, \tag{4.10}$$

the type-II risk (claiming H_0 when $\|p - q\|_2 \geq \epsilon\|p\|_2$) is at most β.

4.1.3 Near-Optimality of ℓ_2 Divergence Test

From the above analysis, when testing a difference of magnitude $\|p-q\|_2 \geq \epsilon\|p\|_2$, reliable detection is guaranteed when the size N of samples X^1 and X^2 is at least $O(n^{1/2}\epsilon^{-2})$ (due to the fact that $\|p\|_2 \geq n^{-1/2}$), with just logarithmic in the reliability parameters factors hidden in $O(\cdot)$. We will show that the $O(n^{1/2})$ sample size is the best rate can achieve unless additional *a priori* information on p and q is available.

Proposition 4.3 (Optimality)**.** *Given cardinality n of the set Ω and sample size N. For i.i.d. N-observation samples X^1 and X^2, suppose there exists a low-risk test that can detect reliably any difference of magnitude $\|p - q\|_2 \geq \epsilon\|p\|_2$ for $0 < \epsilon < 1/2$ such that*

1. *for every distribution p, the type-I risk is at most a given $\alpha < 1/2$, and*

2. *for every distributions p, q satisfying $\|p - q\|_2 \geq \epsilon\|p\|_2$, the type-II risk is at most a given $\beta < 1/4$.*

Then $N \geq O(1)\sqrt{n}$, with a positive absolute constant $O(1)$ that depends on α, β, ϵ.

4.1.4 Illustrating Example: Quasi-Uniform Distribution

Now we present an illustrative example using "quasi-uniform" distributions. Assume that the nominal distribution p and the alternative distribution q are *quasi-uniform*, i.e., there exists a known constant κ satisfying $2 \leq \kappa \leq n$ such that $\|p\|_\infty \leq \kappa/n$ and $\|q\|_\infty \leq \kappa/n$. Since $\|x\|_2 \leq \sqrt{\|x\|_1\|x\|_\infty}$, we have $\max[\|p\|_2, \|q\|_2] \leq \sqrt{\kappa/n}$, and hence the threshold

$$\ell = 6\sqrt{2}M^{-1}\sqrt{\kappa/n} \tag{4.11}$$

satisfies the condition (4.3) with $\theta = 3$ (recall that we are in the case of uniform weights $\sigma_i \equiv 1$). With this choice of ℓ, the right hand side of (4.4) is at most $6\sqrt{2}[2M^{-1}\sqrt{\kappa/n} + M^{-1/2}\sqrt{\kappa/n}\|p - q\|_2]$. To ensure the validity of (4.4) with $\theta = 3$, it suffices to have

$$\|p - q\|_2^2 \geq 6\sqrt{2}\left[2M^{-1}\sqrt{\kappa/n} + M^{-1/2}\sqrt{\kappa/n}\|p - q\|_2\right],$$

which holds when

$$\|p - q\|_2^2 \geq O(1)M^{-1}\sqrt{\kappa/n}, \tag{4.12}$$

with properly selected moderate absolute constant $O(1)$. For quasi-uniform distributions, $\|p - q\|_2$ is no larger than $2\sqrt{\kappa/n}$. Therefore, for $\|p - q\|_2 \geq \lambda n^{-1/2}$ with some $\lambda \in (0, 2\sqrt{\kappa}]$, the sample size M should satisfy

$$M \geq O(1)\frac{\sqrt{\kappa n}}{\lambda^2}$$

in order to ensure (4.12). We see that in the case of $L = R$, given $\alpha \ll 1$, $\beta \ll 1$, the sample size of

$$O(1)[\ln(1/\alpha) + \ln(1/\beta)]\frac{\sqrt{\kappa n}}{\lambda^2}$$

ensures that for the ℓ_2 test with the threshold in (4.11), its type-I risk and type-II risk are upper bounded by α and β, respectively.

In the following, we provide numerical examples to validate the optimality results in Proposition 4.3. Suppose the support size n is even and set $L = R = M$ for simplicity. The experiment set-up is described as the following two steps:

(i) Draw two $n/2$-element subsets independently, Ω_1 and Ω_2, of Ω from the uniform distribution on the family of all subsets of Ω of cardinality $n/2$.

(ii) The samples X^1 are i.i.d. drawn from the uniform distribution on Ω_1, denoted as p; and the second group of samples X^2 are i.i.d. drawn from the uniform distribution on Ω_2, denoted by q.

Therefore we have $\max[\|p\|_2, \|q\|_2] \leq \sqrt{2/n}$, implying that we can set the threshold as

$$\ell = 12M^{-1}n^{-1/2}.$$

In all simulations, the individual test was applied. We perform the simulation for various n and M values. The power is shown in Figure 4.1, averaged over 1000 trials. The results show that for magnitude $\|p - q\|_2 = O(1/\sqrt{n})$, at least $O(\sqrt{n})$ samples are required in order to detect the difference between p and q with high probability.

4.2 Change Detection Procedures Based on ℓ_2 Test

In this section, we construct the change detection procedure based on the proposed weighted ℓ_2 divergence test. Change detection is an important instance of the sequential hypothesis

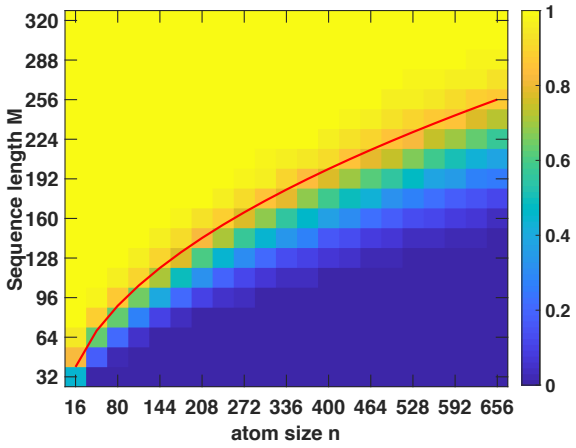


Figure 4.1: Validation of the theoretical $O(\sqrt{n})$ bound by plotting the empirical test power of "quasi-uniform" in Section 4.1.4, averaged over 1000 random trials. The Type-I risk is controlled to be less than 10^{-3}. The theoretical lower bound to sample complexity $O(\sqrt{n})$ is shown in red line, which match the empirical phase-transition "watershed".

test, but it has unique characteristics that require a separate study due to different performance metrics considered. Since we do not know the change location, we have to perform scanning when forming the detection statistic. We discuss two settings: the offline scenario where we have fixed samples and the online setting where the data come sequentially.

4.2.1 Offline Change Detection by "Scan" Statistic

In the offline setting, we observe samples $X^T = \{x_1, ..., x_T\}$ on a time horizon $t = 1, ..., T$, with x_t's taking values in an n-element set $\Omega = \{1, \ldots, n\}$. Assume there exists time $K \in \{1, ..., T\}$ such that for $t \leq K$, x_t are i.i.d. drawn from some *pre-change* distribution p, and for $t \geq K + 1$, x_t are i.i.d. drawn from the *post-change* distribution q. Our goal is to design a test which, based on the samples X^T, decides on the null hypothesis $K = T$ ("no change") versus the alternative $K < T$ ("change"). Meanwhile, we want to control the probability of false alarm to be at most a given $\alpha > 0$, and under this restriction to make the probability of successfully detecting the change as large as possible, at least when K and $T - K$ both are moderately large and q "significantly differs" from p.

We use the proposed test in Section 4.1 to construct a scan statistic for change detection.

Given T, we select a collection of *bases* B_j, $1 \leq j \leq J$. A base B is a segment of $\{1, ..., T\}$ partitioned into three *consecutive* parts: *pre-change* part B_{lf}, middle part B_{md}, and *post-change* part B_{rg}; the last instant in B_{lf} is the first instant in B_{md}, and the first instant in B_{rg} is by 1 larger than the last instant in B_{md}. For example: $B_{\text{lf}} = \{1, ..., 10\}$, $B_{\text{md}} = \{10, 11\}, B_{\text{rg}} = \{12, ..., 20\}$. We associate with base B an *individual test* $\mathcal{T}_B$ which operates with observations $\{x_t, t \in B_{\text{lf}} \cup B_{\text{rg}}\}$ only. This test aims at deciding on two hypotheses: (1) "No change:" there is no change on B, that is, either K is less than the first, or larger than or equal to the last time instant from B; (2) "Change:" the change point K belongs to the middle set B_{md}.

Given $\alpha > 0$ and a base B, we call individual test $\mathcal{T}_B$ associated with this base α-*feasible*, if the probability of false alarm for $\mathcal{T}_B$ is at most α, meaning that whenever there is no change on the base B of the test, the probability for the test to claim change is at most α. Our "overall" test $\mathcal{T}$ works as follows: we equip bases B_j, $1 \leq j \leqslant J$, with tolerances $\alpha_j > 0$ such that $\sum_{j=1}^{J} \alpha_j = \alpha$, and then associate with each base B_j with a α_j-feasible individual test $\mathcal{T}_{B_j}$ (as given by the ℓ_2 test in Section 4.1.1). Given observations X^T, we perform one by one the individual tests in some fixed order, until either (i) the current individual test claims change; and when it happens, the overall test claims change and terminates, or (ii) all J individual tests are performed and no one of them claimed change; in this case, the overall test claims no change and terminates.

Proposition 4.4 (False Alarm Rate for Offline Change Detection)**.** *With the outlined structure of the overall test and under condition $\sum_j \alpha_j = \alpha$, the probability of false alarms for $\mathcal{T}$ (of claiming change when $K = T$) is at most α.*

4.2.2 Online Change Detection

In practice, the online detection of abrupt changes is often of more interest. Instead of giving a fixed duration of samples in the offline setting, the observations arrive sequentially for online detection tasks. The goal is to detect the change as quickly as possible, under the

constraint that the false alarm rate is under control.

The proposed detection procedure based on ℓ_2 test is illustrated in Figure 4.2. Given a sequence $\{x_t, t = 1, 2, \ldots\}$, as each time t, we search over all possible change-points $k < t$. In particular, we form two sequences before k and two sequences between $[k, t]$ with the length all equal to $M_{t,k} = \lceil (t-k)/2 \rceil$; their corresponding empirical distributions are denoted as $\xi_{t,k}$, $\xi'_{t,k}$, and $\eta_{t,k}$, $\eta'_{t,k}$. The detection statistic $\chi_{t,k}$ is formed as:

$$\chi_{t,k} = M_{t,k}(\xi_{t,k} - \eta_{t,k})^\top \Sigma (\xi'_{t,k} - \eta'_{t,k}). \tag{4.13}$$

We note that the multiplicative term $M_{t,k}$ can be viewed as a *scaling* parameter (which is proportional to the standard deviation of the test statistic) such that the variance of $\chi_{t,k}$ is of a constant order as $t - k$ increases.

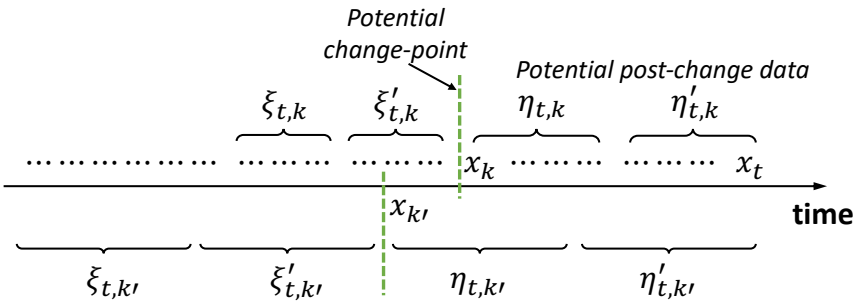


Figure 4.2: Illustration of the sequential change detection procedure.

The online change-point detection procedure is given by a stopping time

$$\mathcal{T} := \inf\{t : \max_{0 \le k \le t} \chi_{t,k} \ge b\}, \tag{4.14}$$

where $b > 0$ is a pre-specified threshold that needs to be determined by controlling the false alarm rate. An intuitive interpretation of $\mathcal{T}$ is that at each time t, we search over all possible change-pints $k < t$, and raise alarm if the maximum statistic exceeds the threshold.

Remark 4.1 (Window-Limited Procedure)**.** In practice, we can also adopt a *window-limited*

version of $\mathcal{T}$ as follows to improve the computational efficiency:

$$\mathcal{T}' = \inf\{t : \max_{m_0 \le t-k \le m_1} \chi_{t,k} \ge b\}, \tag{4.15}$$

where m_0 and m_1 are the lower and upper bounds of the window size that we would like to scan for the possible changes. Usually m_0 can be set such that the resulted sequences are long enough to have meaningful empirical distributions for constructing the detection statistic in (4.13). For practical considerations, we usually require the window size m_1 to be at least the expected detection delay, as discussed in [107] where the original theoretical study of the window-limited test was proposed.

Remark 4.2 (Comparison with the Binning Approach [114])**.** We note that the binning approach in [114] also considers discretized space and scans over all possible change-points when approximating log-likelihood ratio statistics. Compared with [114] which assumes the pre-change distribution is known, the detection procedure (4.14) and its window-limited version (4.15) do not need the prior knowledge of the pre-change distribution. We did not use the log-likelihood ratio statistic here but scan over all neighboring time windows directly to detect any significant difference in empirical distributions.

4.2.3 Theoretical Analysis

Now we characterize the two fundamental performance metrics for sequential change detection, namely the average run length (ARL) and the expected detection delay (EDD). We cannot use the previous method in Proposition 4.1 to determine the threshold because the bound is too conservative and will be intractable when the ARL is large. Here we present an asymptotic theoretical approximation.

To compute the ARL, we need to quantify the distribution of $\mathcal{T}$ when data are sampled from the same distribution p. Intuitively, the detection statistic $\chi_{t,k}$ is small when the samples are from the same distribution. A relatively standard result is that when the threshold

b tends to infinity, the stopping time's asymptotic distribution is approximately exponential when there is no change. This is proven true in various scenarios [181, 184, 216]. The main idea is to show that the number of boundary cross events for detection statistics over disjoint intervals converges to Poisson random variable in the total variation norm; the result can be established by invoking the Poisson limit theorem for dependent samples. Detailed proofs by adapting those techniques into the specific ℓ_2 test setting are left for future work. Under such approximation, we have

$$\mathbb{P}_\infty(\mathcal{T}' \geq m) = \mathbb{P}_\infty\left(\max_{1\leq t\leq m}\max_{m_0\leq t-k\leq m_1}\chi_{t,k} \geq b\right) \approx e^{-\lambda m},$$

where $\mathbb{P}_\infty$ is the probability measure when the change-point equals to ∞, i.e., the change never happens; and $\mathbb{E}_\infty$ denotes the corresponding expectation under this probability measure. Therefore, we only need to compute the probability $\mathbb{P}_\infty(\mathcal{T}' \geq m)$ and find the parameter λ, then the expectation of $\mathcal{T}'$ equals $1/\lambda$. We adopt the change-of-measure transformation [217, 214, 179] and characterize the local properties of a random field. We first quantify the correlation between $\chi_{t,k}$ ane $\chi_{\tau,s}$ in order to find the probability $\mathbb{P}_\infty(\mathcal{T}' \geq m)$ theoretically.

Proposition 4.5 (Temporal Correlation of Sequential Detection Statistics)**.** *Suppose all samples are i.i.d. drawn from the same distribution p, denote $M = t - k = \tau - s$, then the correlation between $\chi_{t,k}$ and $\chi_{\tau,s}$ is*

$$\mathrm{Corr}(\chi_{t,k}, \chi_{\tau,s}) = 1 - \frac{2}{M}|t-\tau| - \frac{2}{M}|k-s| + \frac{(t-\tau)^2 + (s-k)^2}{M^2}.$$

Based on the correlation result, we have the following Theorem characterizing the ARL of the proposed ℓ_2 sequential detection procedure. The main idea is to use a linear approximation for the correlation between detection statistics $\chi_{t,k}$ and $\chi_{\tau,s}$. Then the behavior of the detection procedure can be related to a random field. By leveraging the localization theorem [179], we can obtain an asymptotic approximation for ARL when the threshold b is

large enough (in the asymptotic sense). Define a special function $v(\cdot)$ which is closely related to the Laplace transform of the overshoot over the boundary of a random walk [183]:

$$v(x) \approx \frac{\frac{2}{x}[\Phi\left(\frac{x}{2}\right) - 0.5]}{\frac{x}{2}\Phi\left(\frac{x}{2}\right) + \phi\left(\frac{x}{2}\right)}, \tag{4.16}$$

where $\phi(x)$ and $\Phi(x)$ are the probability density function and cumulative density function of the standard Gaussian distribution; this is the same function as used in Chapter 3 (see (3.15)). For simplicity, we denote the variance of $\chi_{t,k}$ as

$$\sigma_p^2 := \mathrm{Var}[\chi_{t,k}] = 4\left[\sum_{i=1}^n \sigma_i^2 p_i^2 (1-p_i)^2 + \sum_{i \neq j} \sigma_i \sigma_j p_i^2 p_j^2\right]. \tag{4.17}$$

Theorem 4.2 (ARL Approximation)**.** *For large values of threshold $b \to \infty$, the* ARL *of the test $\mathcal{T}'$ can be approximated as*

$$\mathbb{E}_\infty[\mathcal{T}'] = \frac{1}{2} b^{-1} e^{b^2/(2\sigma_p^2)} [2\pi\sigma_p^2]^{1/2} \Big/ \int_{[4b^2/(m_1\sigma_p^2)]^{1/2}}^{[4b^2/(m_0\sigma_p^2)]^{1/2}} y\nu^2(y)dy(1+o(1)). \tag{4.18}$$

The main contribution of Theorem 4.2 is to provide a theoretical method to set the threshold that can avoid the Monte Carlo simulation, which could be time-consuming, especially when ARL is large. Although there is no close-form analytical solution for b, when we let the right-hand side of (4.18) equals to a specific ARL value (a target lower bound for ARL), we can numerically compute the right-hand side of (4.18) for any given threshold value b. Then we search over a grid to find the corresponding threshold values. Table 4.1 validates the approximation's good accuracy by comparing the threshold obtained from (4.18) and compares it with that obtained by the Monte Carlo simulation. In detail, we generate 2000 independent trials of data from nominal distribution p and perform the detection procedure $\mathcal{T}'$ for each trial; the ARL for each threshold b is estimated by the average stopping time over 2000 trials. In Table 4.1, we report the threshold obtained through Monte Carlo simulation (as a proxy for the ground-truth) and on the approximation (4.18), for a

range of ARL values. The ARL values in Table 4.1 correspond to the lower bound of an ARL; since ARL will increase when increasing the threshold. So if we have a good approximation, this can help us to calibrate the threshold and control the false alarm rate. The results in Table 4.1 indicate that the approximation is reasonably accurate since the relative error is around 10% for all specified ARL values. It is worth mentioning that ARL is very sensitive to the choice of threshold, making it challenging to estimate the threshold with high precision. However, the EDD is not that sensitive to the choice of the threshold, which means that a small difference in the threshold will not significantly change EDD.

Table 4.1: Comparison of the threshold b obtained from simulations and the approximation (4.18). Scanning window $m_0 = 10, m_1 = 50$, support size $n = 20$, nominal distribution p is uniform.

ARL	5k	10k	20k	30k	40k	50k
Simulation	2.0000	2.1127	2.2141	2.2857	2.3333	2.3750
Theoretical	1.8002	1.8762	1.9487	1.9897	2.0183	2.0398

After the change occurs, we are interested in the expected detection delay, i.e., the expected number of additional samples to detect the change. There are a variety of definitions for the detection delay [125, 149, 144, 196]. To simplify the study of EDD, it is customary to consider a specific definition $\mathbb{E}_0[\mathcal{T}']$, which is the expected stopping time when the change happens at time 0 and only depends on the underlying distributions p, q. It is not always true that $\mathbb{E}_0[\mathcal{T}']$ is equivalent to the standard worst-case EDD in literature [125, 149]. However, since $\mathbb{E}_0[\mathcal{T}']$ is certainly of interest and is reasonably easy to approximate, we consider it as a surrogate here. We adopt the convention that there are certain pre-change samples $\{x_{-1}, x_{-2}, \ldots\}$ available before time 0, which can be regarded as reference samples.

Note that for any $t > 0$ and $k = 0$, the sequences $\xi_{t,0}$ and $\xi'_{t,0}$ come from the pre-change distribution p since they belong to the reference sequence $\{x_{-1}, x_{-2}, \ldots\}$, and the sequences $\eta_{t,0}$ and $\eta'_{t,0}$ are from the post-change distribution q. Therefore, the expectation

of the detection statistic $\chi_{t,k}$ is $\mathbb{E}[\chi_{t,k}] = \lceil (t-k)/2 \rceil (p-q)^\top \Sigma (p-q)$, which determines the asymptotic growth rate of the detection statistic after the change. Using Wald's identity [180], we are able to obtain a first-order approximation for the detection delay, provided that the maximum window size m_1 is large enough compared to the EDD.

Theorem 4.3 (EDD Approximation)**.** *Suppose $b \to \infty$, with other parameters held fixed. If the window size m_1 is sufficiently large and greater than $2b/[(p-q)^\top \Sigma (p-q)]$, then the expected detection delay*

$$\mathbb{E}_0[\mathcal{T}'] = \frac{b(1+o(1))}{(p-q)^\top \Sigma (p-q)/2}. \tag{4.19}$$

Remark 4.3 (Optimize weights to minimize EDD)**.** From the EDD approximation in (4.19), it is obvious that we can minimize EDD by optimizing over the weights matrix Σ. In particular, the EDD can be minimized when we can find the weights Σ such that the weighted ℓ_2 divergence between p and q is maximized. This is consistent with the subsequent discussion in Section 4.3.1. In particular, when we have certain prior information about the distributions p and q, we could apply the optimization-based method in Section 4.3.1 to find the optimal weights to reduce the detection delay.

4.3 Optimized Weights and Projection of High-Dimensional Data

This section discusses setting optimal weights that adapt to the closeness at different elements in Ω, given some *a prior* information on p and q. In addition, we tackle the data high-dimensionality by adopting the Wasserstein-based principal differences analysis [135] to find the optimal projection.

4.3.1 Optimize Weights for ℓ_2 Test

So far, we primarily focused on the case with uniform weights $\sigma_i \equiv 1$. In this section, we will discuss how to further improve performance by choosing the optimal weights. In the

simplest case, when we know in advance (or can infer from additional "training" samples) that the distribution shift $p \to q$ (nearly) does not affect probabilities with indexes from some *known* set I, we can set $\sigma_i = 0$ for $i \in I$ and $\sigma_i = 1$ for $i \notin I$. This will keep the magnitude $\sum_i \sigma_i (p_i - q_i)^2$ on the left hand side of (4.4), as compared to uniform weights, intact, but will reduce the right hand side of (4.4).

A framework to optimize over σ_i's is as follows. Assume that we know distributions p, q belong to a set $\mathcal{X} \subset \Delta_n$, which is defined by a set of quadratic constraints:

$$\mathcal{X} = \{p \in \Delta_n : p^\top Q_k p \leq 1,\ k = 1, \ldots, K\}, \tag{4.20}$$

where $Q_k \in \mathbb{R}^{n \times n}$ are positive semi-definite ($Q_k \succeq 0$).

A natural way to measure "magnitude of difference" is to use $\|p - q\|_2$ (the case using $\|p - q\|_1$ can also be similarly defined and solved). Assume we want to select $\sigma = [\sigma_1, ..., \sigma_n] \geq 0$ to make reliable detection of difference $\|p - q\|_2 \geq \rho$, for some given $\rho > 0$. To achieve this, we can impose a fixed upper bound on the right hand side in (4.4) when $p = q \in \mathcal{X}$, i.e., to require σ to satisfy

$$g_*(\sigma) := \max_{p \in \mathcal{X}} \sum_i \sigma_i^2 p_i^2 \leq a$$

with some given a, and to maximize under this constraint the quantity

$$f_*(\sigma) := \min_{p,q} \left\{ \textstyle\sum_i \sigma_i (p_i - q_i)^2 : p, q \in \mathcal{X}, \|p - q\|_2 \geq \rho \right\}.$$

For any σ that satisfies $g_*(\sigma) \leq a$, the associated test which claims H_1 when the statistics (defined in (4.2)) $|\chi| > 2\sqrt{2}\theta M^{-1}\sqrt{a}$ is with type-I risk at most $1/\theta^2$. At the same time, large $f_*(\sigma)$ is in favor of good detection of distribution shift of magnitude $\|p - q\|_2 \geq \rho$. By the homogeneity in σ, we can set $a = 1$ without loss of generality.

In general, both g_* and f_* are difficult to compute. Therefore, we replace the problem

$$\max_{\sigma \geq 0} \{f_*(\sigma) : g_*(\sigma) \leq 1\}$$

with its *safe tractable approximation*:

$$\max_{\sigma \geq 0} \{f(\sigma) : g(\sigma) \leq 1\}, \tag{4.21}$$

where f is a concave efficiently computable *lower* bound on f_*, and g is a convex efficiently computable *upper* bound on g_*.

To build g, note that when $p \in \mathcal{X}$, the matrix $P = pp^T \in \mathbb{R}^{n \times n}$ is positive semi-definite ($P \succeq 0$), non-negative in each entry ($P \geq 0$), $\sum_{i,j=1}^n P_{ij} = 1$, and $\text{Tr}(PQ_k) \leq 1$, $k \leq K$, by (4.20). Consequently, the function

$$g(\sigma) := \max \left\{ \text{Tr}(\Sigma^2 P) : P \succeq 0, P \geq 0, \sum_{i,j=1}^n P_{ij} = 1, \text{Tr}(PQ_k) \leq 1,\ k \leq K \right\},$$

with $\Sigma := \text{Diag}\{\sigma_1, ..., \sigma_n\}$, is an efficiently computable convex upper bound on g_*. Similarly, to build f, observe that the matrix $S = (p-q)(p-q)^T$ stemming from $p, q \in \mathcal{X}$ with $\|p - q\|_2 \geq \rho$ belongs to the convex set

$$\mathcal{S} = \left\{ S : S \succeq 0, \sum_{i,j=1}^n |S_{ij}| \leq 4, \sum_{i,j=1}^n S_{ij} = 0, \text{Tr}(S) \geq \rho^2, \text{Tr}(SQ_k) \leq 4, k \leq K \right\}.$$

Therefore,

$$f_*(\sigma) \geq f(\sigma) := \min_{S \in \mathcal{S}} \text{Tr}(\Sigma S),$$

and the function $f(\sigma)$ is concave and efficiently computable.

To implement the problem in (4.21) efficiently, we derive the tractable dual formulation in the following. Note that these constraints can be greatly simplified if Q_k are *diagonal* matrices, especially for the high dimensional case.

Proposition 4.6 (Dual Reformulation)**.** *The dual formulation of the optimization problem* (4.21) *is*

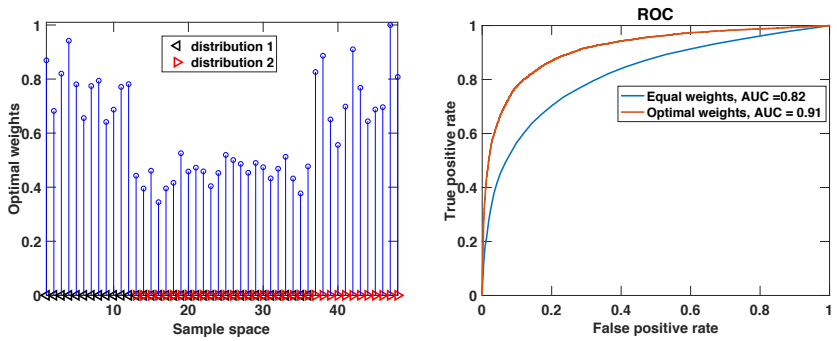
$$
\begin{aligned}
\max\ & \lambda\rho^2 - 4\textstyle\sum_k x_k - 4\xi \\
\textit{s.t.}\ & \lambda \geq 0, P \succcurlyeq 0, \xi \geq 0, x_k \geq 0, U \geq 0, W \geq 0, \Lambda \succcurlyeq 0, V \geq 0, \mu_k \geq 0, 1 \leq k \leq K, \\
& \textstyle\sum_k x_k Q_k + U - W - P - rJ - \lambda I_n \succcurlyeq -\Sigma, \\
& U_{ij} + W_{ij} \leq \xi, \quad 1 \leq i \leq n, 1 \leq j \leq n, \\
& \textstyle\sum_k \mu_k - \nu \leq 1, \quad -\Lambda - V + \sum_k \mu_k Q_k - \nu J \succcurlyeq \Sigma^2.
\end{aligned}
$$

where $\Sigma = \text{Diag}\{\sigma_1, ..., \sigma_n\}$ *and* $J \in \mathbb{R}^{n\times n}$ *is a matrix with all elements equal to* 1.

We present an illustrative simulation example to show the benefits of optimizing weights σ. The experimental set-up is as follows. Consider the sample space $\Omega = \{1, \ldots, n\}$ with $n = 48$. The distributions p and q are set as uniform distributions on the subset $\Omega_1 = \{1, \ldots, 3n/4\}$ and $\Omega_2 = \{n/4 + 1, \ldots, n\}$, respectively. The common support of p and q consists of $n/2$ elements. We first use training data to estimate the matrix Q in our formulation. Specifically, we sample 32 observations from each distribution and compute the empirical distribution of all observations. This process is repeated for $m = 50$ trials, and the resulting Q is solved from the following optimization problem

$$
\begin{aligned}
\text{min} \quad & \log\det A^{-1} \\
\text{s.t.} \quad & \|Ap_i\|_2 \leq 1,\ i = 1, \ldots, m,
\end{aligned}
$$

where $A = Q^{1/2}$ and p_i is the empirical distribution in the i-th trial. The volume of the ellipsoid defined with Q is proportional to $\det A^{-1}$. Thus the solution to the above optimization problem is the minimum ellipsoid that contains the m empirical distributions [30]. The optimal weights are shown in Figure 4.3 (a). Moreover, we compare the ROC curve of the test under equal weights $\sigma_i = 1$ and optimal weights in Figure 4.3 (b), averaged over 10,000 trials. The result shows the benefits of using optimized weights.



(a) (b)

Figure 4.3: Illustration of optimal weights on a simulated example. (a): Optimal weights; (b): The ROC curves under optimal weights and equal weights.

4.3.2 Optimal Projection for High-Dimensional Data

Assume that the two distributions p and q, rather than discrete distributions on a given finite set, are continuous distributions on $\mathbb{R}^d$. In this situation, we may try to convert observations $x \in \mathbb{R}^d$ into observations $f(x)$ taking values in a *finite* set and apply the proposed test to the transformed observations.

One way to build f is to project observations x onto one-dimensional subspace and then split the range of the projection into bins. We propose to select this subspace using, when available, "training sample" $x_1, ..., x_{2T}$, with the first T observations drawn, independently of each other, from the nominal distribution p, and the last T observations drawn independently of each other and of $x_1, ..., x_T$, from the distribution q. A natural selection of the one-dimensional subspace can be as follows. Denote by e the unit vector spanning the subspace. Let us look at the sample empirical distributions of the projections of the observations $x_1, ..., x_{2T}$ on e, and try to find unit vector e for which the Wasserstein distance between the distributions of the first half and the second half of the projections is as large

as possible [135]. The distance above is, up to factor $1/T$, the quantity

$$\begin{aligned}\phi(e) = \min_{\omega_{ij},\, 1\le i,j\le 2T} \Bigg\{ & \sum_{i,j} |e^\top (x_i - x_j)| \omega_{ij} : \omega_{ij} \ge 0, 1 \le i, j \le 2T; \\ & \sum_j \omega_{ij} = \begin{cases} 1, & i \le T \\ 0, & i > T \end{cases}; \sum_i \omega_{ij} = \begin{cases} 0, & j \le T \\ 1, & j > T \end{cases} \Bigg\} \\ = \max_\lambda \big\{ & \textstyle\sum_{i=1}^T \lambda_i - \sum_{i=T+1}^{2T} \lambda_i : \lambda_i - \lambda_j \le |e^\top (x_i - x_j)|, 1 \le i, j \le 2T \big\} \\ = \Phi(E[e]), & \end{aligned}$$

where

$$\begin{aligned} E[e] &= ee^\top, \\ \Phi(E) &= \max_\lambda \Big\{ \sum_{i=1}^T \lambda_i - \sum_{i=T+1}^{2T} \lambda_i : \lambda_i - \lambda_j \le \sqrt{[x_i - x_j]^\top E [x_i - x_j]}, 1 \le i, j \le 2T \Big\}. \end{aligned}$$

Note that function $\Phi(E)$ is concave and the goal is to maximize $\Phi(E)$ over positive semi-definite rank-one matrices $E = ee^\top$ with trace 1. An efficiently solvable convex relaxation after relaxing the rank-one constraint is:

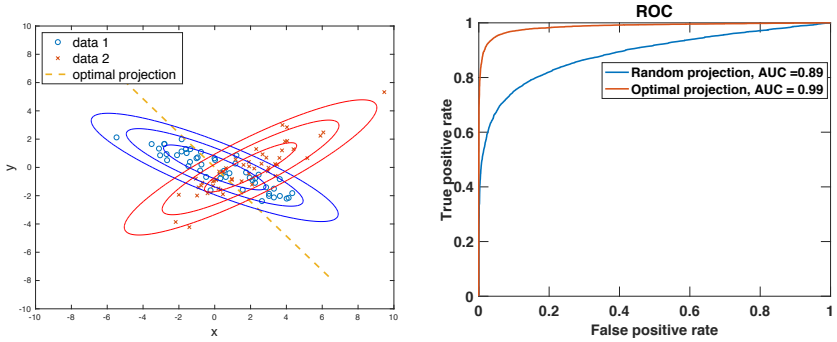
$$\max_{E,\lambda} \left\{ \sum_{i=1}^T \lambda_i - \sum_{i=T+1}^{2T} \lambda_i : \lambda_i - \lambda_j \le \sqrt{[x_i - x_j]^\top E [x_i - x_j]}, \forall i, j;\ E \succeq 0, \mathrm{Tr}(E) = 1 \right\}.$$

After the optimal solution E_* to the problem is found, we can use standard methods to obtain a reasonably good e, e.g., take e as the leading eigenvector of E_*.

Here we present a simple numerical illustration for the optimal project. Consider the two-dimensional Gaussian distributions with same mean value and different covariance structures. More specifically, let the data X_1 to be sampled i.i.d. from $\mathcal{N}(\mu_1, \Sigma_1)$ and data X_2 to be sampled from $\mathcal{N}(\mu_2, \Sigma_2)$, where

$$\mu_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top,\ \mu_2 = \begin{bmatrix} 2 & 0 \end{bmatrix}^\top,\ \Sigma_1 = \begin{bmatrix} 5.03 & -2.41 \\ -2.41 & 1.55 \end{bmatrix},\ \Sigma_2 = \begin{bmatrix} 5.50 & 3.30 \\ 3.30 & 2.53 \end{bmatrix}. \tag{4.22}$$

Figure 4.4 shows the optimal projection obtained from 50 training samples from each distribution (which can be seen to optimally "separate" the two distributions), and the ROC curve averaged over 10,000 trials that demonstrates the performance gain of the optimal projection.



(a) (b)

Figure 4.4: Illustration of optimal projection on simulated data. (a): Optimal projection for two training sets; (b): The ROC curves for optimal projection and random projection.

4.4 Numerical Examples

In this section, we perform some simulations to validate the performance of the ℓ_2 test and compare with two benchmarks: (i) the classical parametric Hotelling's T^2 test [88]; and (ii) the non-parametric maximum mean discrepancy (MMD) test [75]. More specifically, we study the test power of the two-sample test for Gaussian distributions under various dimensions. Moreover, we show the performance in change detection by studying the detection power in the offline case and the expected detection delay in the online case, respectively.

We first introduce briefly the two benchmark procedures.

Hotelling's T^2 statistic: The Hotelling's T^2 statistic is a classical parametric test designed utilizing the mean and covariance structures of data, and thus it can detect both the mean and covariance shifts [88]. Given two set of samples $\{x_1, \ldots, x_{n_1}\}$ and $\{y_1, \ldots, y_{n_2}\}$,

the Hotelling's T^2 statistic is defined as

$$t^2 = \frac{n_1 n_2}{(n_1 + n_2)} (\bar{x} - \bar{y})^\top \widehat{\Sigma}^{-1} (\bar{x} - \bar{y}), \tag{4.23}$$

where $\bar{x}$ and $\bar{y}$ are the sample mean and $\widehat{\Sigma}$ is the pooled covariance matrix estimate.

MMD statistic: The MMD test is a non-parametric benchmark for two-sample test and change detection [75, 120]. Given a class of functions $\mathcal{F}$ and two distributions p and q, the MMD distance between p and q is defined as $\mathrm{MMD}_{\mathcal{F}}(p, q) = \sup_{f \in \mathcal{F}} (\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{y \sim q}[f(y)])$. For MMD in reproducing kernel Hilbert spaces (RKHS), given samples $\{x_1, \ldots, x_{n_1}\}$ and $\{y_1, \ldots, y_{n_2}\}$, an unbiased estimate of squared MMD distance is given by

$$\begin{aligned} \mathrm{MMD}_u^2 = & \frac{1}{n_1(n_1-1)} \sum_{i=1}^{n_1} \sum_{j \neq i} k(x_i, x_j) + \frac{1}{n_2(n_2-1)} \sum_{i=1}^{n_2} \sum_{j \neq i} k(y_i, y_j) \\ & - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} k(x_i, y_j), \end{aligned} \tag{4.24}$$

where $k(\cdot, \cdot)$ is the kernel function associated with RKHS.

4.4.1 Two-Sample Test

Following a similar setup as in [75], we investigate the performance of various tests as a function of the dimension d of the sample space $\mathbb{R}^d$, when both p and q are Gaussian distributions. We consider values of d up to 256. The type-I risk for all tests is set as $\alpha = 0.05$. The sample size is chosen as $n_1 = n_2 = 100$, and results are averaged over 500 independent trials. In the first case, the distributions p, q have different means and the same variance. More specifically, $p = \mathcal{N}(0, I_d)$ and $q = \mathcal{N}(\mu \mathbf{1}/\sqrt{d}, I_d)$ with $\mu = 0.8$. Note that the division of each element of the mean vector by $\sqrt{d}$ makes the difficulty of the hypothesis testing similar across all d values. In the second case, the distributions p, q have

the same means but different variance. More specifically, $p = \mathcal{N}(0, I_d)$ and $q = \mathcal{N}(0, \Sigma)$ with $\Sigma = \text{Diag}\{0.25, 1, \ldots, 1\}$, i.e., we only scale the first diagonal entry in the covariance matrix to make the hypothesis testing problem challenging to perform.

The test power for different methods is shown in Figure 4.5. The test power drops when the dimension increases, which is consistent with the results in [160]. Hotelling's T^2 test performs good in low dimensions, but its performance degrades quickly when we consider higher dimensional problems. The MMD test is comparable to ℓ_2 test in low dimensions, but the ℓ_2 test tends to outperform the MMD test in high dimensions. The reason can be that by projecting to one-dimensional spaces using a good projection, the power of ℓ_2 test tends to decrease slower compared to Hotelling's T^2 and MMD tests.

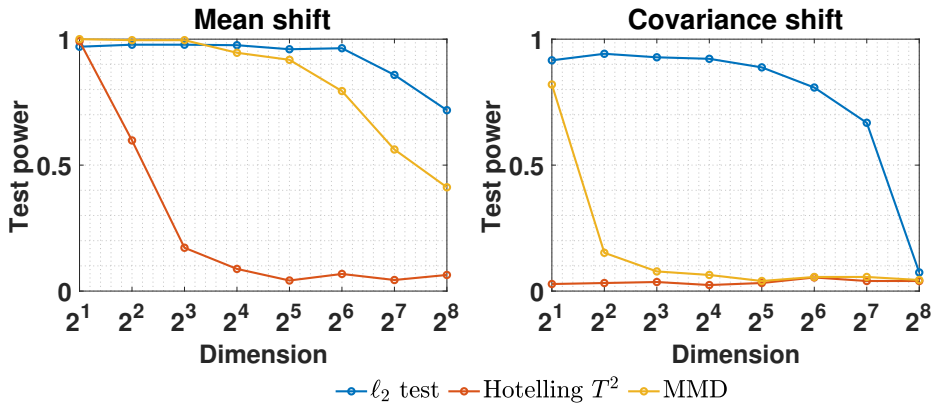


Figure 4.5: Comparison of test power of the proposed test versus classic Hotelling's T^2 statistic and the MMD statistic, when performing a two-sample test on two Gaussian distributions, with significance level $\alpha = 0.05$. (Left) Gaussian distributions having the same variance and different means; (Right) Gaussian distributions having same mean and different variances.

4.4.2 Offline Change Detection

As an extension and application of the proposed ℓ_2 test, we investigate the performance for the *offline* change detection and compare the detection power, i.e., the probability of successfully detecting the change when there is a change.

Assume we have sample $x_1, \ldots, x_T$ with a fixed time horizon $T = 200$, when there is a change, we set the change-point $K = 100$. The ℓ_2 detection statistic at each time t is

$M\chi$ with χ defined in (4.2) (here $M = 2LR/(L+R)$ is the normalizing constant). To avoid the segment being too short, we compute the detection statistics for time instances $t \in [w, T-w]$ with $w = 20$, and then take the maximum. Similarly, the Hotelling's T^2 statistic at each time t is computed using (4.23) by treating data before t as one sample and after t as another sample; the MMD statistic is computed in a similar way from (4.24). We claim there is a change-point when the maximum of the detection statistics within window $t \in [w, T-w]$ exceeds the threshold. The thresholds for different methods will be chosen by Monte Carlo simulation to control the false alarm rate.

We consider the following cases (distribution changes) in the numerical experiments.

Case 1 (Discrete distributions). The support size is $n = 10$, distribution shifts from $p = \mathbf{1}/10$ (uniform) to $q = [1/30, 2/30, \ldots, 5/30, 5/30, \ldots, 2/30, 1/30]$ (non-uniform).

Case 2 (Gaussian mean and covariance shift). The distribution shifts from two-dimensional Gaussian $\mathcal{N}(0, I_2)$ to $\mathcal{N}\left([0.5 \;\; 0]^\top, [1 \;\; 0.7]^\top [1 \;\; 0.7] + [-1 \;\; 0.4]^\top [-1, \;\; 0.4]\right)$.

Case 3 (Gaussian to Gaussian mixture). The distribution shifts from $\mathcal{N}(0, I_{20})$ to the Gaussian mixture $0.8\mathcal{N}(0, I_{20}) + 0.2\mathcal{N}(0, 0.1 I_{20})$.

Case 4 (Gaussian to Laplace). The distribution shifts from standard Gaussian $\mathcal{N}(0, 1)$ to Laplace distribution with zero mean and standard deviation 0.8.

The detection power is averaged over 500 repetitions and is reported in Table 4.2. It shows that the proposed ℓ_2 test outperforms the classic Hotelling's T^2 and MMD tests, especially when the distribution change is difficult to detect (such as Case 3 and Case 4, where pre- and post-change distributions are close). For Case 2 detecting mean and covariance shifts, the MMD test performs slightly better. A possible explanation is that the MMD metric can capture the difference between pre- and post-change Gaussian distributions well in a fairly low-dimensional setting.

Table 4.2: Detection power in offline change detection. The sequence of length is 200. Thresholds for all methods are calibrated so that the significance level is $\alpha = 0.10$ and $\alpha = 0.25$. Averaged over 500 trials.

	$\alpha = 0.10$				$\alpha = 0.25$			
	Case 1	Case 2	Case 3	Case 4	Case 1	Case 2	Case 3	Case 4
ℓ_2 test	**0.52**	0.85	**0.18**	**0.56**	**0.70**	0.90	**0.35**	**0.71**
MMD	0.32	**0.90**	0.16	0.43	0.60	**0.95**	0.34	0.69
Hotelling's T^2	0.07	0.23	0.09	0.06	0.20	0.23	0.23	0.23

4.4.3 Online Change Detection

We further investigate the performance for *online* change detection and compare the average detection delay, i.e., the number of samples it takes to detect the change after the change happens. More specifically, the detection delay is the difference between the stopping time and the true change-point.

Assume we have samples $\{x_t, t = 1, 2, \ldots\}$ that are available sequentially. We adopt the convention that there are pre-change samples available as $\{\ldots, x_{-2}, x_{-1}, x_0\}$, which are referred as historical data and can be used during the detection procedure. Consider the window-limited ℓ_2 detection procedure defined in (4.15) with parameter $m_0 = 20$ and $m_1 = 100$. The Hotelling's T^2 detection statistic at each time t is constructed as $(\bar{x}_t - \hat{\mu})^T \widehat{\Sigma}^{-1} (\bar{x}_t - \hat{\mu})$ where $\bar{x}_t$ is the average of samples within window $[t - m_0 + 1, t]$, and $\hat{\mu}, \widehat{\Sigma}$ are estimated from historical data. The MMD statistic is constructed in the same way as in [120] with block size $B_0 = 20$ and number of blocks $N = 5$. We will claim change and stop the detection procedures when the detection statistic exceeds the threshold; the thresholds for different methods are chosen by Monte Carlo simulation to control the average run length.

We consider the following four cases, which are modified slightly from the offline case. We have increased slightly the signal-to-noise ratio in certain cases to increase the detectability in the online setting.

Case 1 (Discrete distributions). The support size is $n = 10$, distribution shifts from $p =$

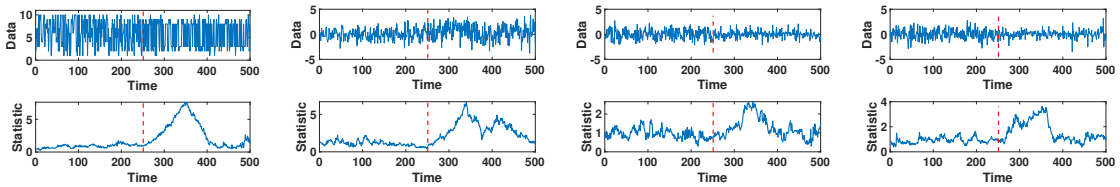
$\mathbf{1}/10$ (uniform) to $q = [0.04, 0.14, 0.32, 0, 0, 0, 0, 0.32, 0.14, 0.04]$ (non-uniform).

Case 2 (Gaussian mean and covariance shift). The distribution shifts from two-dimensional Gaussian $\mathcal{N}(0, I_2)$ to $\mathcal{N}\left([0.5 \;\; 0]^\top, [1 \;\; 0.7]^\top [1 \;\; 0.7] + [-1 \;\; 0.4]^\top [-1 \;\; 0.4]\right)$.

Case 3 (Gaussian to Gaussian mixture). The distribution shifts from $\mathcal{N}(0, I_{20})$ to the Gaussian mixture $0.4\mathcal{N}(0, I_{20}) + 0.6\mathcal{N}(0, 0.1 I_{20})$.

Case 4 (Gaussian to Laplace). The distribution shifts from standard Gaussian $\mathcal{N}(0, 1)$ to Laplace distribution with zero mean and standard deviation 0.7.

The evolution paths of detection statistics for all cases are given in Figure 4.6. To simulate EDD, we let the change occur at the first time instant of the testing data. The detection delay is averaged over 500 repetitions and reported in Table 4.3.



Case 1 Case 2 Case 3 Case 4

Figure 4.6: Illustration of online change detection using the ℓ_2 divergence under four simulated cases explained in Section 4.4.3. For each case, the upper plot shows the raw data and the bottom plot shows the evolution path of the ℓ_2 detection statistic, with true change-point indicated in red dash lines.

Table 4.3: Comparison of EDD for online change detection using the proposed statistic, the MMD, and the Hotelling's T^2 statistic. The parameter is $n = 10$, $m_0 = 20$, $m_1 = 100$ and thresholds for all methods are calibrated so that $\mathsf{ARL} = 500$. The dashed line indicates the method fails to detect the change (i.e., the delay is larger than the time horizon).

	Case 1	Case 2	Case 3	Case 4
ℓ_2 test	**20.34**	89.66	**69.23**	**92.49**
MMD	258.02	47.72	—	394.91
Hotelling's T^2	406.42	**36.79**	—	370.61

4.5 Real Data Study: Online Gesture Change Detection

In this section, we apply our method to the sequential gesture detection problem using a real dataset: the Microsoft Research Cambridge-12 (MSRC-12) Kinect gesture dataset [61]. This dataset consists of sequences of human skeletal body part movements (represented as body part locations) collected from 30 people performing 12 gestures. There are 18 sensors in total, and each sensor records the coordinates in the three-dimensional Cartesian coordinate system at each time. Therefore there are 54 attributes, denoted by $y_t \in \mathbb{R}^{54}$, $t = 1, 2, \ldots, T$. The goal is to detect the transition of gestures from the sequences of sensor observations.

We apply the proposed online change detection procedure defined in (4.15) to the MSRC-12 dataset, and the detailed scheme is outlined as follows. We first preprocess the data by removing the frames that the person is standing still or with little movements. Then we select a *unit-norm* vector $u \in \mathbb{R}^{54}$ and project data into this direction to obtain a univariate sequence: $x_t = u^\top y_t$. The projection vector u is found by finding the optimal projection to maximize the Wasserstein distance described in Section 4.3.2. Then we discretize the univariate sequence into n bins. At each time t, we construct the detection statistic $\max_{m_0 \le t-k \le m_1} \chi_{t,k}$ as illustrated in Figure 4.2.

The parameters are set as $m_0 = 20$, $m_1 = 300$ for the detection procedure. The detection statistics are shown in Figure 4.7, with the true change indicated by red dash lines. We also compare the ℓ_2 test with tests based on Hotelling's T^2 statistic, ℓ_1 distance, and KL divergence. For the ℓ_1 distance based approach, we build the test statistic as for time t and potential change-point $k < t$ as the ℓ_1 distance between empirical distributions of samples before and after k, then the detection statistic is computed by maximizing over all potential change-points k in $m_0 \le t - k \le m_1$. The KL divergence based approach is built similarly by replacing the ℓ_1 distance with KL divergence (see [210] for detailed implementations).

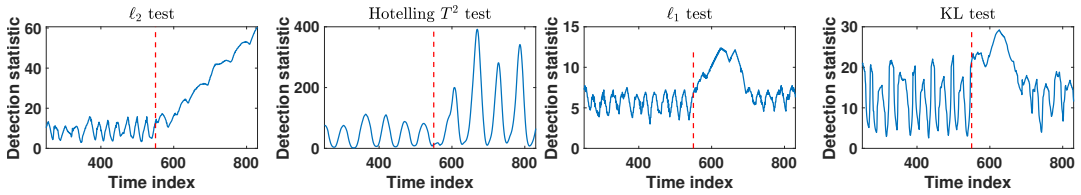


Figure 4.7: Real-data example using online gesture change detection. Comparison of detection statistics (under uniform weights) for "bow" to "throw", for the proposed procedure, the Hotelling's T^2 test, ℓ_1 test, and the KL test. Red dash lines indicate the true change-point (hand-labeled).

From the results in Figure 4.7, we can observe that our sequential detection procedure based on the ℓ_2 divergence can detect the change right after it happens. This is because the detection statistic before the change has a smaller variance, which indicates that we can set the threshold to be reasonably low for quicker detection. Moreover, there is a clear linear increasing trend after the change, enabling quick and reliable detection. In contrast, Hotelling's T^2 statistic does not have the desired online change detection behavior. The detection statistic is noisy before the change and does not have a consistent positive shift after the change; the KL test is even worse in this regard. The ℓ_1 divergence-based test has a similar behavior as the ℓ_2 divergence. However, the ℓ_1 divergence has smaller "signal-to-noise" ratio in that the variance between the change is larger, and post-change distribution drift seems to be smaller.

4.6 Conclusion and Discussions

We have presented a new non-parametric change detection procedure based on the optimal weighted ℓ_2 divergence. We study the optimality and various theoretical properties of the weighted ℓ_2 divergence for the offline and online change-point detection. We also study the practical aspects, including calibration threshold using training data, optimizing weights, and finding an optimal projection for high-dimensional data. We demonstrate the good performance of the proposed method using simulated and real-data for human gesture detection.

CHAPTER 5

ROBUST HYPOTHESIS TESTING WITH WASSERSTEIN UNCERTAINTY SETS

This chapter studies the data-driven robust hypothesis testing problem with Wasserstein uncertainty sets. This work is partially summarized in [69]. Section 5.1 introduces the problem formulation of data-driven robust hypothesis testing. Section 5.2 presents the optimal test. Section 5.3 discusses the choice of the optimal radii of the uncertainty sets. Section 5.4 demonstrates our robust tests' good performance using both synthetic and real data.

5.1 Problem Setup and Wasserstein Minimax Test

Let $\Omega \subset \mathbb{R}^d$ be the sample space, where d is the data dimension. Denote $\mathscr{P}(\Omega)$ as the set of Borel probability measures on Ω. Given $P_1, P_2 \in \mathscr{P}(\Omega)$, the *simple hypothesis test* decides whether a given test sample ω is from P_1 or P_2. In many practical situations, P_1, P_2 are not exactly known, but instead we have access to n_1 and n_2 training samples for each hypothesis. Denote the two sets of training samples as $\widehat{\Omega}_k = \{\widehat{\omega}_k^1, \ldots, \widehat{\omega}_k^{n_k}\}$, $k = 1, 2$, and define empirical distributions constructed using training data sets as

$$Q_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \delta_{\widehat{\omega}_k^i}.$$

Here δ_ω denotes the Dirac point mass concentrated on $\omega \in \Omega$.

To capture the distributional uncertainty, we consider *composite hypothesis test* of the form:

$$\begin{aligned} H_0 \ : \quad & \omega \sim P_1, \quad P_1 \in \mathcal{P}_1; \\ H_1 \ : \quad & \omega \sim P_2, \quad P_2 \in \mathcal{P}_2, \end{aligned}$$

where $\mathcal{P}_1, \mathcal{P}_2$ are collections of relevant probability distributions. In particular, we will consider them to be Wasserstein uncertainty sets. Below let us describe our problem setup.

5.1.1 Randomized Test

We consider the set of all *randomized tests* defined as follows [99].

Definition 5.1 (Randomized test)**.** Given hypotheses H_0, H_1, a randomized test is any Borel measurable function $\pi : \Omega \to [0, 1]$ which, for any observation $\omega \in \Omega$, accepts the hypothesis H_0 with probability $\pi(\omega)$ and H_1 with probability $1 - \pi(\omega)$.

In the randomized test, the decision to accept a hypothesis can be a random selection based on the function $\pi(\omega)$. Thus, the usual deterministic test (e.g., considered in [69]) is a special case by setting $\pi(\omega) \in \{0, 1\}$ and the randomized test is more general.

For a simple hypothesis test with hypotheses P_1 and P_2, we define the *risk* of a randomized test π as the summation of Type-I and Type-II errors:

$$\Phi(\pi; P_1, P_2) := \mathbb{E}_{P_1}[1 - \pi(\omega)] + \mathbb{E}_{P_2}[\pi(\omega)], \tag{5.1}$$

where $\mathbb{E}_P$ denotes the expectation with respect to the random variable ω that follows distribution P. Here we consider equal weights on the Type-I and Type-II errors; other weighted combinations can be addressed similarly.

5.1.2 Wasserstein Minimax Formulation

The minimax hypothesis test finds the optimal test that minimizes the *worst-case risk* over all possible distributions in the composite hypotheses:

$$\inf_{\pi} \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \Phi(\pi; P_1, P_2). \tag{5.2}$$

The resulting worst-case solution P_1^*, P_2^* are called the *least favorable distributions* (LFDs) in the classical robust hypothesis testing literature [89, 91].

We consider uncertainty sets based on the *Wasserstein metric* of order 1, defined as:

$$\mathsf{W}(P,Q) := \min_{\gamma\in\Gamma(P,Q)} \left\{\mathbb{E}_{(\omega,\omega')\sim\gamma}\left[c(\omega,\omega')\right]\right\},$$

where $c(\cdot,\cdot) : \Omega\times\Omega \to \mathbb{R}_+$ is a metric on Ω, and $\Gamma(P,Q)$ is the collection of all Borel probability measures on $\Omega\times\Omega$ with marginal distributions P and Q. Define the Wasserstein uncertainty sets $\mathcal{P}_1, \mathcal{P}_2$ as Wasserstein balls centering at two empirical distributions:

$$\mathcal{P}_k := \{P_k \in \mathscr{P}(\Omega) :\ \mathsf{W}(P_k, Q_k) \le \theta_k\}, \quad k = 1,2, \tag{5.3}$$

where $\theta_1, \theta_2 > 0$ specify the radii of the uncertainty sets. See Chapter 2.2 for useful background information and preliminaries.

Remark 5.1 (Comparison with Huber's censored likelihood ratio test)**.** Huber's seminal paper [89] considered a deterministic minimax test with uncertainty sets referred to as ϵ-contamination sets:

$$\mathcal{P}_k = \{(1-\epsilon_k)p_k + \epsilon_k f_k,\ f_k \in \mathscr{P}(\Omega)\},$$

where $\epsilon_k \in (0,1)$, p_k is the nominal density function, and f_k is the density that can be viewed as the perturbation, $k = 1,2$. Huber proved that the optimal test in this setting is a censored version of the likelihood ratio test, with censoring thresholds c', c'', and the LFDs are given by:

$$q_1(x) = \begin{cases} (1-\epsilon_1)p_1(x) & p_2(x)/p_1(x) < c'' \\ \frac{1}{c''}(1-\epsilon_1)p_2(x) & p_2(x)/p_1(x) \ge c'' \end{cases};$$

$$q_2(x) = \begin{cases} (1-\epsilon_2)p_2(x) & p_2(x)/p_1(x) > c' \\ c'(1-\epsilon_2)p_1(x) & p_2(x)/p_1(x) \le c' \end{cases}.$$

Huber assumed the exact knowledge of the nominal distributions p_1 and p_2. This is different from our setting, where we only have limited samples from each hypothesis. A simple

observation is that if we set p_k to be the empirical distribution, then the ratio $p_2(x)/p_1(x)$ will be ∞ on $\widehat{\Omega}_2 \setminus \widehat{\Omega}_1$ and 0 on $\widehat{\Omega}_1 \setminus \widehat{\Omega}_2$. In such a case, the LFDs proposed by Huber are degenerate

$$q_1(x) = \begin{cases} (1-\epsilon_1)/n_1 & x \in \widehat{\Omega}_1 \\ \epsilon_1/n_2 & x \in \widehat{\Omega}_2 \end{cases}; \quad q_2(x) = \begin{cases} (1-\epsilon_2)/n_2 & x \in \widehat{\Omega}_2 \\ \epsilon_2/n_1 & x \in \widehat{\Omega}_1 \end{cases},$$

which do not lead to any meaningful test.

5.2 Tractable Convex Reformulation and Optimal Test

The saddle point problem (5.2) for the Wasserstein minimax test is an *infinite-dimensional* variational problem, which in the original form does not amend to any tractable solution. In this section, we derive a *finite-dimensional* convex reformulation for finding the optimal test.

At the core of our analysis is the following strong duality result, which means we can exchange the order of infimum and supremum in our problem:

$$\inf_{\pi} \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \Phi(\pi; P_1, P_2) = \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \inf_{\pi} \Phi(\pi; P_1, P_2). \tag{5.4}$$

This is essential in leading to closed-form expression for the optimal test and convex reformulation in solving the LFDs. Our proof strategy is as follows. First, in Section 5.2.1, we derive a closed-form expression of the optimal test for the simple hypothesis problem $\inf_\pi \Phi(\pi; P_1, P_2)$. Next in Section 5.2.2, we develop a convex reformulation of the $\sup \inf$ problem on the right-hand side of (5.4), whose optimal solution gives the LFDs. Then in Section 5.2.3, we construct the optimal minimax test for the original formulation (left-hand side of (5.4)) leveraging strong duality. Note that here we cannot directly rely on existing tools such as Sion's minimax theorem [186], because (i) the space of all randomized tests is not endowed with a linear topological structure and, (ii) Wasserstein ball is not compact

in the space $\mathscr{P}(\Omega)$ since Ω may not be compact.

5.2.1 Optimal Test for Simple Hypothesis Test

Let us start by considering the simple hypothesis test for given $P_1, P_2 \in \mathscr{P}(\Omega)$, the inner minimization in the right-hand-side of (5.4):

$$\inf_{\pi} \ \Phi(\pi; P_1, P_2). \tag{5.5}$$

Define the total variation distance between two distributions P_1 and P_2 as $\mathsf{TV}(P_1, P_2) := (1/2) \int_\Omega |dP_1(\omega) - dP_2(\omega)|$. The following Lemma gives a closed-form expression for the optimal test, which resembles a randomized version of the Neyman-Pearson Lemma. The proof is provided in the Appendix.

Lemma 5.1. *Let* $p_1(\omega) := \frac{dP_1}{d(P_1+P_2)}(\omega)$. *The test*

$$\pi(\omega) = \begin{cases} 1, & \text{if } p_1(\omega) > 1/2, \\ 0, & \text{if } p_1(\omega) < 1/2, \\ \text{any real number in } [0,1], & \text{otherwise,} \end{cases}$$

is optimal for (5.5) *with the risk*

$$\psi(P_1, P_2) := \int_\Omega \min\{p_1(\omega), 1 - p_1(\omega)\}\, d(P_1 + P_2)(\omega) = 1 - \mathsf{TV}(P_1, P_2). \tag{5.6}$$

Lemma 5.1 shows that the optimal test for the simple hypothesis takes a similar form as the likelihood ratio test that accepts the hypothesis with a higher likelihood and breaks the tie arbitrarily. An important observation from the lemma is that the risk only depends on the *common* support of the two distributions, defined as $\Omega_0(P_1, P_2) := \big\{\omega \in \Omega : 0 < p_1(\omega) < 1\big\}$, on which P_1 and P_2 are absolutely continuous with respect to each other. In particular, if the supports of P_1, P_2 have measure-zero overlap, then $\inf_\pi \Phi(\pi; P_1, P_2)$ equals to zero

— the optimal test for two non-overlapping distributions P_1, P_2 has zero risk.

5.2.2 Least Favorable Distributions

Now we continue with finding the LFDs given the form of the optimal test in Lemma 5.1, which corresponds to the remaining supermum part of the right-hand side of (5.4):

$$\sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \psi(P_1, P_2). \tag{5.7}$$

Note that from the definition of ψ in (5.6), the risk associated with the optimal test, the problem of finding LFDs admits a clear statistical interpretation: the LFDs correspond to a pair of distributions in the uncertainty sets that are closest to each other in the total variation distance.

To tackle the infinite-dimensional variational problem (5.7), let us first discuss some structural properties of the LFDs that will lead to a finite-dimensional convex reformulation. Consider a toy example where $Q_1 = \delta_{\widehat{\omega}_1}, Q_2 = \delta_{\widehat{\omega}_2}$, i.e., there is only one sample in each training data set. The goal of solving LFDs can be understood as moving part of the probability mass on $\widehat{\omega}_1$ and $\widehat{\omega}_2$ to other places such that the objective function $\psi(P_1, P_2)$ is maximized. Note that, to find the LFDs, we need to (i) move the probability mass such that P_1 and P_2 overlap as much as possible, since the objective value $\psi(P_1, P_2)$ depends only on the common support; (ii) then if we were to move p_k from $\widehat{\omega}_k$ to a common point $\omega \in \Omega$, $k = 1, 2$, in the least favorable way, then we solve $\min_{\omega \in \Omega}[p_1 c(\omega, \widehat{\omega}_1) + p_2 c(\omega, \widehat{\omega}_2)]$ by the definition of the Wasserstein metric. From the triangle inequality satisfied by the metric $c(\cdot, \cdot)$, we need ω to be on the linear segment connecting $\widehat{\omega}_1$ and $\widehat{\omega}_2$ and in fact, it has to be one of the endpoints $\widehat{\omega}_1$ or $\widehat{\omega}_2$. More generally, one can generalize this argument, and there exist LFDs supported on the empirical observations.

The following lemma shows that the LFDs can be solved via a finite-dimensional convex optimization problem. For simplicity, define the total number of observations

$n := n_1 + n_2$ and the union of observations from both hypotheses:

$$\widehat{\Omega} := \widehat{\Omega}_1 \cup \widehat{\Omega}_2.$$

Without causing confusions, we re-label the samples in $\widehat{\Omega}$ as $\{\widehat{\omega}^1, \ldots, \widehat{\omega}^n\}$.

Lemma 5.2 (LFDs)**.** *The LFD problem in* (5.7) *can be reformulated as the following finite-dimensional convex program:*

$$\begin{aligned} \max_{\substack{p_1,p_2\in\mathbb{R}_+^n \\ \gamma_1,\gamma_2\in\mathbb{R}_+^{n\times n}}} \quad & \sum_{l=1}^{n} \min\left\{p_1^l, p_2^l\right\} \\ \text{subject to} \quad & \sum_{l=1}^{n}\sum_{m=1}^{n} \gamma_{k,l,m} c(\widehat{\omega}^l, \widehat{\omega}^m) \le \theta_k,\ k=1,2; \\ & \sum_{m=1}^{n} \gamma_{k,l,m} = Q_k^l,\ 1 \le l \le n, k = 1,2; \\ & \sum_{l=1}^{n} \gamma_{k,l,m} = p_k^m,\ 1 \le m \le n,\ k=1,2. \end{aligned} \tag{5.8}$$

Above, the decision variables γ_k are square matrices that can be viewed as a joint distribution on $\widehat{\Omega} \times \widehat{\Omega}$ with marginals specified by Q_k and candidate LFDs p_k. The lm-th entry of γ_k is specified by $\gamma_{k,l,m}$ and the l-th entry of p_k (respectively, Q_k) is specified by p_k^l (respectively, Q_k^l). In the following, we will denote (P_1^*, P_2^*) as the LFDs solved from (5.8). Note that Lemma 5.2 simplifies the LFD problem (5.7) from infinite-dimensional to finite-dimensional, using the fact that there exist LFDs supported on a finite set $\widehat{\Omega} \subset \Omega$ due to our analysis. We also comment that the complexity of solving the LFDs in (5.8) is *independent* of the dimension of the data, once the pairwise distances $c(\widehat{\omega}^l, \widehat{\omega}^m)$ are calculated and given as input parameters of the convex program.

5.2.3 General Optimal Test

Thus far, we have found one of the LFDs defined on the discrete set of training samples $\widehat{\Omega}$ by solving the original minimax problem's dual form. However, it may be common in practice that the given test sample is different from all training samples. In this case, the current optimal test in Lemma 5.1 is not well-defined. Moreover, this optimal test is *not* uniquely defined when there is a tie between the likelihood of samples under two hypotheses. To overcome these limitations, we will establish an optimal test that is well-defined anywhere in the observation space Ω.

Our main result is the following theorem which specifies the general form of the optimal test π^* and LFDs (P_1^*, P_2^*) to the saddle point problem (5.2), whose proof is given in the Appendix.

Theorem 5.1 (General Optimal Test)**.** *Let $\widehat{\mathcal{P}}_k := \mathcal{P}_k \cap \mathscr{P}(\widehat{\Omega})$, $k = 1, 2$, and let (P_1^*, P_2^*) be the LFDs solved from* (5.8)*. The optimal test over the whole observation space $\pi^* : \Omega \to [0, 1]$ is given by*

(i) On the support of training samples $\omega \in \widehat{\Omega}$, $\pi^*(\omega) = \widehat{\pi}_m^*$ for $\omega = \widehat{\omega}^m$, $m = 1, \ldots, n$, where $\widehat{\pi}_m^* \in [0, 1]$, $m = 1, \ldots, n$, is the solution to the following system of linear equations

$$\begin{aligned} \sum_{m=1}^{n} (1 - \widehat{\pi}_m) P_1^*(\widehat{\omega}^m) &= \min_{\lambda_1 \geq 0} \left\{ \lambda_1 \theta_1 + \frac{1}{n_1} \sum_{l=1}^{n} \max_{1 \leq m \leq n} \{1 - \widehat{\pi}_m - \lambda_1 c(\widehat{\omega}^l, \widehat{\omega}^m)\} \right\}, \\ \sum_{m=1}^{n} \widehat{\pi}_m P_2^*(\widehat{\omega}^m) &= \min_{\lambda_2 \geq 0} \left\{ \lambda_2 \theta_2 + \frac{1}{n_2} \sum_{l=1}^{n} \max_{1 \leq m \leq n} \{\widehat{\pi}_m - \lambda_2 c(\widehat{\omega}^l, \widehat{\omega}^m)\} \right\}; \end{aligned} \tag{5.9}$$

the solution is guaranteed to exist.

(ii) Off the support of training samples $\omega \in \Omega \setminus \widehat{\Omega}$, $\pi^*(\omega) \in [\ell(\omega), u(\omega)]$, *where*

$$\begin{aligned} \ell(\omega) &= \max\left\{ \max_{i=1,\ldots,n_1} \min_{\widehat{\omega}\in\widehat{\Omega}} \left\{ \pi^*(\widehat{\omega}) + \lambda_1^* c(\widehat{\omega}, \widehat{\omega}_1^i) - \lambda_1^* c(\omega, \widehat{\omega}_1^i) \right\}, 0 \right\}, \\ u(\omega) &= \min\left\{ \min_{j=1,\ldots,n_2} \max_{\widehat{\omega}\in\widehat{\Omega}} \{ \pi^*(\widehat{\omega}_2^j) - \lambda_2^* c(\widehat{\omega}, \widehat{\omega}_2^j) + \lambda_2^* c(\omega, \widehat{\omega}_2^j) \}, 1 \right\}, \end{aligned} \tag{5.10}$$

λ_k^*, $k = 1, 2$ *are the minimizers to the* inf *problems on the right hand side of* (5.9)*, and it is guaranteed that* $u(\omega) \geq \ell(\omega)$, $\forall \omega \in \Omega \setminus \widehat{\Omega}$.

To illustrate Theorem 5.1, let us consider a toy example as shown in Figure 5.1. Suppose the training samples for hypothesis H_0 is $\widehat{\omega}_1 = -2$ and for hypothesis H_1 are $\widehat{\omega}_2 = 1$ and $\widehat{\omega}_3 = 3$. Then, the two empirical distributions Q_1 is a point mass on $\widehat{\omega}_1 = -2$ and Q_2 is a discrete distribution that $\widehat{\omega}_2 = 1$ and $\widehat{\omega}_3 = 3$ occur with equal probability 1/2. By setting the radii of the uncertainty sets $\theta_1 = \theta_2 = 1$, the LFDs solution to (5.8) becomes $P_1^*(\widehat{\omega}_1) = 0.69$, $P_1^*(\widehat{\omega}_2) = 0.28$, $P_1^*(\widehat{\omega}_3) = 0.03$, and $P_2^*(\widehat{\omega}_1) = 0.29$, $P_2^*(\widehat{\omega}_2) = 0.28$, $P_2^*(\widehat{\omega}_3) = 0.43$. Notice that there is a tie at the point $\widehat{\omega}_2$. Now we will invoke Theorem 5.1 to break this tie. According to (5.9), the general optimal test $\pi^*(\widehat{\omega}_i)$, $i = 1, 2, 3$ needs to satisfy:

$$1 - \pi^*(\widehat{\omega}_1) - \lambda_1^* c(\widehat{\omega}_1, \widehat{\omega}_1) = 1 - \pi^*(\widehat{\omega}_2) - \lambda_1^* c(\widehat{\omega}_1, \widehat{\omega}_2) = 1 - \pi^*(\widehat{\omega}_3) - \lambda_1^* c(\widehat{\omega}_1, \widehat{\omega}_3).$$

Therefore, we can set $\pi^*(\widehat{\omega}_2) = 1 - c(\widehat{\omega}_1, \widehat{\omega}_2)/c(\widehat{\omega}_1, \widehat{\omega}_3) = 0.4$. This means that the optimal test at $\widehat{\omega}_2$ should accept the hypothesis H_0 with probability 0.4 (note that the tie is not broken arbitrarily). As a comparison, consider a different case where $\widehat{\omega}_2 = 2$ while everything else is kept the same. It can be verified that there is still a tie at $\widehat{\omega}_2$. However, this time we have $\pi^*(\widehat{\omega}_2) = 1 - c(\widehat{\omega}_1, \widehat{\omega}_2)/c(\widehat{\omega}_1, \widehat{\omega}_3) = 0.2$, meaning that the optimal test at $\widehat{\omega}_2$ should accept the hypothesis H_0 with probability 0.2. We note that in this simple experiment, the chance of accepting H_0 decreases if we move $\widehat{\omega}_2$ away from $\widehat{\omega}_1$, which is

consistent with our intuition as illustrated in Figure 5.1. Moreover, we also plot the upper and lower bounds $u(\omega)$ and $\ell(\omega)$, as defined in (5.10), showing the range of the optimal test off the support of training samples. This example also demonstrates the advantage of using Wasserstein metrics in defining the uncertainty sets: the optimal test will directly reflect the data geometry.

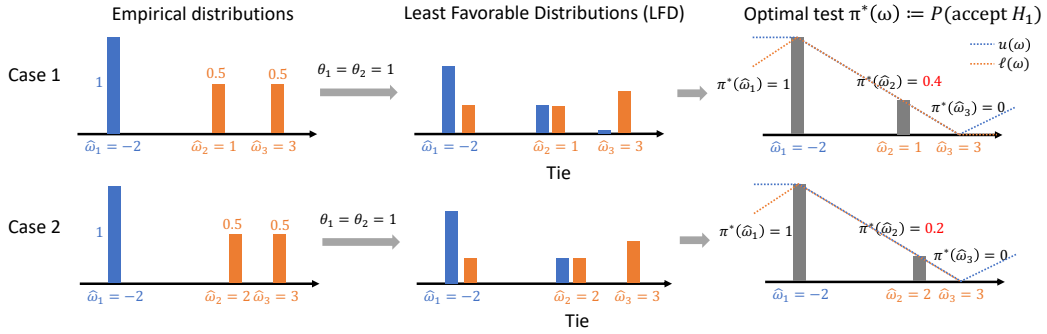


Figure 5.1: A toy example illustrating the optimal test depends on the training data configuration. In these two cases, there are three samples, and only $\widehat{\omega}_2$ is different, which takes values 1 and 2, respectively. Note that the optimal test $\pi^*(\widehat{\omega}_2)$ will change when the gap between empirical samples are different. We also illustrate the upper and lower bounds $u(\omega)$ and $\ell(\omega)$ from (5.10).

5.2.4 Extension to Whole Space via Kernel Smoothing

We observe that for samples ω off the empirical support, it is possible to have $u(\omega)$ strictly larger than $\ell(\omega)$ with $u(\cdot)$ and $\ell(\cdot)$ given in (5.10). In such cases, there are infinite choices for $\pi^*(\omega)$ according to Theorem 5.1. In this subsection, we describe a specific choice for $\pi^*(\omega)$ under such situation by kernel smoothing. As a natural strategy, we may use kernel smoothing to extend LFDs solved from (5.8) to the whole space. This can be done by convolving the discrete LFDs with a kernel function $G_h : \mathbb{R}^d \to \mathbb{R}$ parameterized by a (bandwidth) parameter h:

$$P_k^h(\omega) := \sum_{l=1}^{n} P_k^*(\widehat{\omega}^l) G_h(\omega - \widehat{\omega}^l), \ k = 1, 2, \ \forall \omega \in \Omega. \tag{5.11}$$

There can be various choices of kernel functions. For instance, given normalized data, we can use the product of one-dimensional kernel function $g : \mathbb{R} \to \mathbb{R}$ with bandwidth $h > 0$:

$$G_h(x) = \frac{1}{h^d} \prod g\left(\frac{x_i}{h}\right), \ x \in \mathbb{R}^d.$$

An example of the kernel-smoothed LFDs is shown in Figure 1.1. Through convolution, we can obtain the kernel-smoothed LFDs and the corresponding test π_h^* that is defined as the optimal test for the simple hypothesis under (P_1^h, P_2^h) as specified in Lemma 5.1. To ensure the risk after kernel-smoothing is comparable to that of the general optimal test π^*, we truncate the resulted π_h^* such that (5.10) is satisfied after truncation. After such a procedure, the test based on the kernel-smoothed LFDs will achieve a good performance as validated by the numerical experiments in Section 5.4.

5.2.5 Test with Batch Samples

Testing using a batch of samples is important in practice, as one test sample may not achieve sufficient power. We can construct a test for a batch of samples by assembling the optimal test for each individual sample. Assume m i.i.d. test samples ω_1, ω_2, ..., ω_m. Consider a *batch test* based on the "majority rule" with the acceptance region for H_0 defined as $\mathbb{A} := \{(\omega_1, \omega_2, \ldots, \omega_m) : \pi^m(\omega_1, \omega_2, \ldots, \omega_m) \geq 1/2\}$, where

$$\pi^m(\omega_1, \omega_2, \ldots, \omega_m) = \frac{1}{m} \sum_{i=1}^{m} \pi^*(\omega_i),$$

can be viewed as the fraction of votes in favor of hypothesis H_0 (due to Lemma 5.1). We can bound the risk of such a majority rule batch test as follows.

Proposition 5.1 (Risk for batch test)**.** *The risk of the test $\pi^m(\omega_1, \ldots, \omega_m)$ is upper bounded by*

$$\max\left\{\sup_{P_1 \in \mathcal{P}_1} \mathbb{P}_{P_1}[\mathbb{A}], \sup_{P_2 \in \mathcal{P}_2} \mathbb{P}_{P_2}[\mathbb{A}^c]\right\} \leq \sum_{m/2 \leq i \leq m} \binom{m}{i} (\epsilon^*)^i (1-\epsilon^*)^{m-i},$$

where

$$\epsilon^* = \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \Phi(\pi^*; P_1, P_2),$$

is the worst-case risk of the optimal randomized test and $\mathbb{A}$ is the acceptance region for H_0. Thus, when $\epsilon^* < 1/2$, the above probability tends to 0 exponentially fast as the batch size $m \to \infty$.

5.3 Optimal Radii

In this section, we discuss how to choose the radii θ_1, θ_2, which is critical to the performance of the robust optimal test. There is clearly a tradeoff: when the radius is too small, the optimal test is not robust and does not generalize to new test data; while the radius is too large, the solution may be too conservative, causing performance degradation. We expect sample sizes n_k to play a major role in determining the optimal radii and thus in the following we emphasize by denoting the radii as θ_{k,n_k} and the empirical distributions as Q_{k,n_k}. It should also be remembered that the uncertainty sets $\mathcal{P}_k(\theta_{k,n_k})$, $k = 1, 2$, also depend on the sample sizes. We will show that the theoretically optimal radii $\theta_{k,n_k} \leq \mathcal{O}(n_k^{-1/d})$, $k = 1, 2$ for all distributions on the sample space $\Omega \subset \mathbb{R}^d$. The order $\mathcal{O}(n_k^{-1/d})$ is the worst-case scenario that be achieved. In particular, we would like to point out that the worst-case radii coincides with the common strategy for establishing theoretical guarantees by requiring the uncertainty sets to contain true distributions with high probability. This is because the empirical Wasserstein metric has poor concentration properties [31] and will lead to radii choices $\mathcal{O}(n_k^{-1/d})$ as well.

To characterize the optimal radii, we adopt an alternative technique inspired by using empirical likelihood to study distributionally robust optimization [112, 26]. The strategy is as follows. We define the optimal test for true distributions as the *oracle*; however, since the true distributions are unknown, such an oracle test is also unknown. Our goal is to select the smallest radii that are sufficiently large such that the oracle test is achievable

with high probability. To this end, we define a set $\mathcal{S}$ that contains all pairs of distributions that lead to the oracle test. Then within the set $\mathcal{S}$, we find the closest distribution to the empirical distributions with respect to the Wasserstein metric; intuitively, this gives rise to the tightest radii that meet our goal. Then we discuss the asymptotic performance of such distance and thus the radii; we can show that such a choice will lead to the oracle test with high probability.

To state the precise result, we start with two definitions. Let P_1°, P_2° be the underlying true distributions of the hypotheses H_0 and H_1 respectively. Define the *oracle* test π° corresponding to the true distributions as specified by Lemma 5.1. Also define the set of optimal tests for resolving simple hypothesis for all pairs of distributions in our uncertainty sets with radii θ_{1,n_1} and θ_{2,n_2} (using Lemma 5.1):

$$\Pi(\theta_{1,n_1}, \theta_{2,n_2}) := \{\pi : \exists P_1 \in \mathcal{P}_1(\theta_{1,n_1}), P_2 \in \mathcal{P}_2(\theta_{2,n_2}), \text{ s.t. } \pi \in \arg\min_{\pi'} \Phi(\pi'; P_1, P_2)\}.$$

We are interested in finding the radii such that the set is likely to include the oracle test $\pi^\circ \in \Pi(\theta_{1,n_1}, \theta_{2,n_2})$ following the similar idea of [112, 26]. To achieve this goal, we introduce a set $\mathcal{S}$ that contains all possible pairs of distributions giving rise to the oracle test π°:

$$\mathcal{S} := \{(P_1, P_2) : \pi^\circ \in \arg\min_{\pi:\Omega\to[0,1]} \Phi(\pi; P_1, P_2)\}.$$

Note that $\mathcal{S}$ is guaranteed to be non-empty since it at least contains the true distribution $\{P_1^\circ, P_2^\circ\}$. Then consider within $\mathcal{S}$, the distributions that are closest to the empirical distributions and define the so-called *profile function* to capture the notion of "distance to the empirical distributions" within the set:

$$\mathsf{F}_{n_1,n_2} := \inf_{\{P_1,P_2\}\in\mathcal{S}} \max_{k=1,2}\{\mathsf{W}(P_k, Q_{k,n_k})\}; \tag{5.12}$$

here the subscript indicates its dependence on the sample sizes n_1 and n_2. Clearly if the

radii $\theta_{1,n_1}, \theta_{2,n_2} \geq \mathsf{F}_{n_1,n_2}$, then the intersection $(\mathcal{P}_1(\theta_{1,n_1}) \times \mathcal{P}_2(\theta_{2,n_2})) \cap \mathcal{S}$ is nonempty, and thus $\pi^\circ \in \Pi(\theta_{1,n_1}, \theta_{2,n_2})$, as illustrated in Figure 5.2.

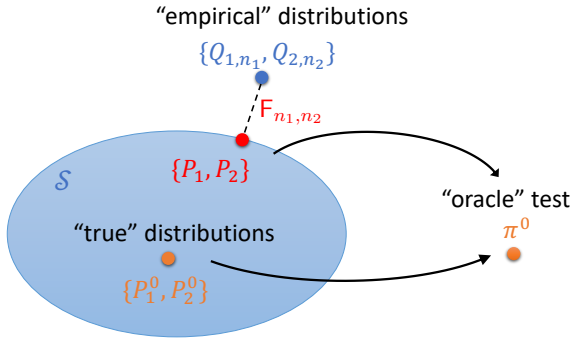


Figure 5.2: An illustration of the profile function. The set $\mathcal{S}$ contains all pairs of distributions $\{P_1, P_2\}$ such that the oracle test is optimal; F_{n_1,n_2} denotes the minimal distance from the empirical distribution to the set $\mathcal{S}$.

We first derive an equivalent dual representation of the profile function F_{n_1,n_2}. We partition the sample space Ω as

$$\Omega_1^\circ := \{\omega \in \Omega : dP_1^\circ(\omega) \geq dP_2^\circ(\omega)\}, \quad \Omega_2^\circ := \{\omega \in \Omega : dP_1^\circ(\omega) < dP_2^\circ(\omega)\}.$$

Thereby the oracle test π° accepts hypothesis H_0 on set Ω_1° and accept hypothesis H_1 on set Ω_2°; and the boundary between Ω_1° and Ω_2° corresponds to the decision boundary of the oracle test π°. Denote by $\mathcal{B}_+(\Omega)$ ($\mathrm{Lip}(\Omega)$) the set of bounded and non-negative (respectively, 1-Lipschitz continuous) functions on Ω. Define the function class

$$\mathcal{A} := \Big\{\boldsymbol{\alpha} = \boldsymbol{\alpha}_2 \mathbb{1}_{\Omega_2^\circ} - \boldsymbol{\alpha}_1 \mathbb{1}_{\Omega_1^\circ} : \ \boldsymbol{\alpha}_k \in \mathcal{B}_+(\Omega_k^\circ) \cap \mathrm{Lip}(\Omega_k^\circ),\ \boldsymbol{\alpha}(\omega_k^\circ) = 0,\ k = 1, 2\Big\}, \quad (5.13)$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function and $\omega_k^\circ \in \Omega_k^\circ$, $k = 1, 2$. Thus for each function $\boldsymbol{\alpha} \in \mathcal{A}$, the positive part is on Ω_2° and the negative part is on Ω_1°, and all functions in $\mathcal{A}$ coincide on $\omega_1^\circ, \omega_2^\circ$. We have the following lemma, whose proof is given in Appendix.

Lemma 5.3. *The profile function* F_{n_1,n_2} *defined in* (5.12) *equals*

$$\mathsf{F}_{n_1,n_2} = \sup_{\substack{\lambda_1,\lambda_2\geq 0,\lambda_1+\lambda_2\leq 1\\ \boldsymbol{\alpha}\in\mathcal{A}}} \Bigg\{ \mathbb{E}_{\widehat{\omega}_1\sim Q_{1,n_1}}\Big[\inf_{\omega\in\Omega}\big\{\lambda_1 c(\omega,\widehat{\omega}_1)+\boldsymbol{\alpha}(\omega)\big\}\Big] + \mathbb{E}_{\widehat{\omega}_2\sim Q_{2,n_2}}\Big[\inf_{\omega\in\Omega}\big\{\lambda_2 c(\omega,\widehat{\omega}_2)-\boldsymbol{\alpha}(\omega)\big\}\Big]\Bigg\}.$$

We note that the objective function, denoted as $F_N(\lambda_1,\lambda_2,\boldsymbol{\alpha})$, of the above supreme problem can be decoupled into two terms: $F_N(\lambda_1,\lambda_2,\boldsymbol{\alpha}) = E_N(\lambda_1,\lambda_2,\boldsymbol{\alpha}) + G_N(\boldsymbol{\alpha})$, where

$$\begin{aligned} E_N(\lambda_1,\lambda_2,\boldsymbol{\alpha}) &:= \frac{1}{n_1(N)}\sum_{i=1}^{n_1(N)}\inf_{\omega\in\Omega}\{\lambda_1 c(\omega,\widehat{\omega}_1^i)+\boldsymbol{\alpha}(\omega)-\boldsymbol{\alpha}(\widehat{\omega}_1^i)\} \\ &\quad + \frac{1}{n_2(N)}\sum_{j=1}^{n_2(N)}\inf_{\omega\in\Omega}\{\lambda_2 c(\omega,\widehat{\omega}_2^j)-(\boldsymbol{\alpha}(\omega)-\boldsymbol{\alpha}(\widehat{\omega}_2^j))\}, \\ G_N(\boldsymbol{\alpha}) &:= \frac{1}{n_1(N)}\sum_{i=1}^{n_1(N)}\boldsymbol{\alpha}(\widehat{\omega}_1^i) - \frac{1}{n_2(N)}\sum_{j=1}^{n_2(N)}\boldsymbol{\alpha}(\widehat{\omega}_2^j). \end{aligned}$$

It follows that $E_N(\lambda_1,\lambda_2,\boldsymbol{\alpha})\leq 0$ since the $\inf$ value is non-positive by taking $\omega=\widehat{\omega}_1^i$ and $\omega=\widehat{\omega}_2^j$, respectively, whence $F_N(\lambda_1,\lambda_2,\boldsymbol{\alpha})\leq G_N(\boldsymbol{\alpha})$ and

$$\mathsf{F}_{n_1(N),n_2(N)} \leq \sup_{\boldsymbol{\alpha}\in\mathcal{A}} G_N(\boldsymbol{\alpha}).$$

Based on the definition of the set $\mathcal{A}$, we observe a close-form solution for $\sup_{\boldsymbol{\alpha}\in\mathcal{A}} G_N(\boldsymbol{\alpha})$ as follows. By definition of $\mathcal{A}$, $\boldsymbol{\alpha}(\widehat{\omega}_1^i)\leq 0$ for $\widehat{\omega}_1^i\in\Omega_1^\circ$ and $\boldsymbol{\alpha}(\widehat{\omega}_2^j)\geq 0$ for $\widehat{\omega}_2^j\in\Omega_2^\circ$. Therefore, to maximize $G_N(\boldsymbol{\alpha})$, we should let $\boldsymbol{\alpha}(\widehat{\omega}_1^i)=0$ for $\widehat{\omega}_1^i\in\Omega_1^\circ$ and $\boldsymbol{\alpha}(\widehat{\omega}_2^j)=0$ for $\widehat{\omega}_2^j\in\Omega_2^\circ$. In addition, since $\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2$ are 1-Lipschitz, we have $\boldsymbol{\alpha}(\widehat{\omega}_1^i)\leq \min_{j:\,\widehat{\omega}_2^j\in\Omega_2^\circ} c(\widehat{\omega}_1^i,\widehat{\omega}_2^j)$ for $\widehat{\omega}_1^i\in\Omega_2^\circ$ and $\boldsymbol{\alpha}(\widehat{\omega}_2^j)\geq -\min_{i:\,\widehat{\omega}_1^i\in\Omega_1^\circ} c(\widehat{\omega}_2^j,\widehat{\omega}_1^i)$ for $\widehat{\omega}_2^j\in\Omega_1^\circ$. Hence we have

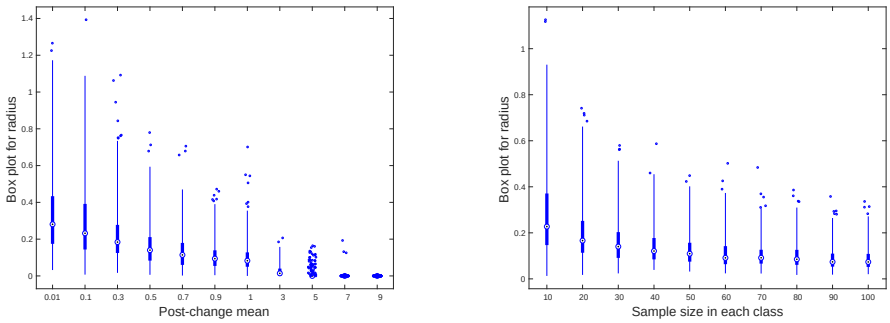
$$\sup_{\boldsymbol{\alpha}\in\mathcal{A}} G_N(\boldsymbol{\alpha}) = \frac{1}{n_1}\sum_{i:\,\widehat{\omega}_1^i\in\Omega_2^\circ}\min_{j:\,\widehat{\omega}_2^j\in\Omega_2^\circ} c(\widehat{\omega}_1^i,\widehat{\omega}_2^j) + \frac{1}{n_2}\sum_{j:\,\widehat{\omega}_2^j\in\Omega_1^\circ}\min_{i:\,\widehat{\omega}_1^i\in\Omega_1^\circ} c(\widehat{\omega}_2^j,\widehat{\omega}_1^i). \tag{5.14}$$

To ease the exposition, we consider a balanced sample size regime by assuming $n_1 := n_1(N)$ and $n_2 := n_2(N)$ such that $\lim_{N\to\infty} n_1(N)/n_2(N) = 1$, although our results can be generalized to the setting where the sample size ratio converges to a positive constant. Under this regime, we denote the sample sizes using a single parameter N. Assume f_1 and f_2 are the density functions for the true distributions P_1° and P_2°, respectively. The right-hand side of (5.14) consists of minimum distances between samples in two sets, whose limiting distribution can be computed theoretically in similar works [59, 145, 202, 146]. Here we restate the result in [59] to provide explicitly the constant factor within order $\mathcal{O}(N^{-1/d})$. Under mild conditions on the support and data-generating distributions, in particular, if the sets Ω_1°, Ω_2° are compact convex sets and the densities f_k are continuous on Ω and has bounded partial derivatives on Ω, and $f_k(\omega) > 0$, for all $\omega \in \Omega$, $k = 1, 2$, then [59] has shown that as $N \to \infty$, the right-hand side of (5.14) converges to

$$\frac{\Gamma(1+1/d)}{V_d^{1/d} N^{1/d}} \left(\int_{\Omega_2^\circ} \frac{f_1(x)}{[f_2(x)]^{1/d}} dx + \int_{\Omega_1^\circ} \frac{f_2(x)}{[f_1(x)]^{1/d}} dx \right), \tag{5.15}$$

where $V_d = \pi^{d/2}/\Gamma(1+d/2)$ is the volume of the unit ball in $\mathbb{R}^d$, $\Gamma(x) = \int_0^\infty z^{x-1} e^{-z} dz$ is the Gamma function. This shows that the order of the limiting upper bound for the profile function is $\mathcal{O}(n^{-1/d})$. Compared with the same order obtained from the convergence of empirical measures in Wasserstein distance [62], the constant term in (5.15) shows the explicit dependence on the underlying data-generating distributions and the relation between two density functions.

We remark that although we adopt a similar principle as used in [26, 178] by considering the profile function F_{n_1,n_2}, the proof in our case is technically much more challenging, because: (1) the uncertainty set here involves the empirical samples from two classes, while there is only one uncertainty set in [26, 178]; (2) the introduced variable $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ are functions in the continuous samples space, not the finite-dimensional vector as in [26], thus the optimality condition is not a simple first-order condition but involves inequalities due to the



(a) (b)

Figure 5.3: Consider a simulated example with the null distribution $\mathcal{N}(0,1)$ and the alternative distribution $(\mu,1)$. We illustrate the dual profile function F_{n_1,n_2} as a function of (a) the mean shift μ and (b) the sample size n, which are consistent with our theory.

variational principle; (3) the inequality constraint here differs from the equality constraints in [26, 178], resulting in an additional supreme involved in the constraint for solving F_{n_1,n_2}. Thus, we develop quite different analytical techniques here.

5.4 Numerical Experiments

In this section, we present several numerical experiments to demonstrate the good performance of our method.

5.4.1 Synthetic Data: Testing Gaussian Mixtures

Assume data are 100-dimensional and the samples under two hypotheses are generated from Gaussian mixture models (GMM) following the distributions $0.5\mathcal{N}(0.4e, I_{100}) + 0.5\mathcal{N}(-0.4e, I_{100})$ and $0.5\mathcal{N}(0.4f, I_{100}) + 0.5\mathcal{N}(-0.4f, I_{100})$, respectively. Here $e \in \mathbb{R}^{100}$ is a vector with all entries equal to 1, and $f \in \mathbb{R}^{100}$ is a vector with the first 50 entries equal to 1 and remaining 50 entries equal to -1. Consider a setting with a small number of training samples $n_1 = n_2 = 10$, and then test on 1000 new samples from each mixture model. The radius of the uncertainty set and the kernel bandwidth are determined by cross-validation.

Table 5.1: GMM data, 100-dimensional, comparisons averaged over 500 trials.

# observation (m)	Ours	GMM	Logistic	Kernel SVM	3-layer NN
1	**0.2145**	0.2588	0.4925	0.3564	0.4164
2	**0.2157**	0.2597	0.4927	0.3581	0.4164
3	**0.1331**	0.1755	0.4905	0.3122	0.3796
4	**0.1329**	0.1762	0.4905	0.3129	0.3808
5	**0.0937**	0.1310	0.4888	0.2877	0.3575
6	**0.0938**	0.1315	0.4881	0.2893	0.3570
7	**0.0715**	0.1034	0.4880	0.2727	0.3399
8	**0.0715**	0.1038	0.4876	0.2745	0.3401
9	**0.0579**	0.0850	0.4873	0.2634	0.3264
10	**0.0578**	0.0851	0.4874	0.2641	0.3267

We compare the performance of the proposed approach with several commonly used classifiers. They are comparable since binary classifiers can be used for deciding hypotheses, although they are designed with different targets. The competitors include the Gaussian Mixture Model (GMM), logistic regression, kernel support vector machine (SVM) with radial basis function (RBF) kernel, and a three-layer perceptron [66] to illustrate the performance of neural networks. The results are summarized in Table 5.1, where the first column corresponds to the single observation scheme, while other columns are results using multiple observations, with the number of observations m varying from 2 to 10. We use the majority rule for GMM, logistic regression, kernel SVM, and three-layer neural networks (NN) for testing batch samples. Note that there are over 2500 parameters in the neural network model with two hidden layers (50 nodes in each layer), which is challenging to learn when the training data size is small. Moreover, given only ten samples per class, estimating the underlying Gaussian mixture model is unrealistic, so that any parametric methods will suffer. The results demonstrate that when there is a small sample size, our minimax test outperforms other methods.

5.4.2 Real Data: MNIST Handwritten Digits Classification

We also compare the performance using MNIST handwritten digits dataset [115]. The full dataset contains 70,000 images, from which we randomly select five training images from each class. We solve the optimal randomized test from (5.8) with the radii parameters chosen by cross-validation. For the batch test setting, we divide test images from the same class into batches, each consisting of m images. The decision for each batch is made using the majority rule for the optimal test in Section 5.2.5, as well as for logistic regression and SVM. We repeat this process to 500 randomly selected batches, and the average misclassification rates are reported in Table 5.2. The results show that our method significantly outperforms logistic regression and SVM. Moreover, the performance gain is higher in the batch test setting: the errors decay quickly as m increases. Note that the neural network-based deep learning model is not appropriate for this setting since the data-size is too small to train the model.

Table 5.2: MNIST data, comparisons averages over 500 trials.

# observation (m)	Ours	Logistic	SVM
1	**0.3572**	0.3729	0.3674
2	**0.3631**	0.3797	0.3712
3	**0.2772**	0.2897	0.2840
4	**0.2122**	0.2239	0.2169
5	**0.1786**	0.1882	0.1827
6	**0.1540**	0.1643	0.1588
7	**0.1347**	0.1446	0.1391
8	**0.1185**	0.1276	0.1222
9	**0.1063**	0.1160	0.1119
10	**0.0960**	0.1057	0.1010

5.4.3 Application: Human Activity Detection

In this subsection, we apply the optimal test for human activity detection from sequential data, using a dataset released by the Wireless Sensor Data Mining Lab [124, 204, 104]. In this dataset, 225 users were asked to perform specific activities, including walking, jogging,

stairs, sitting, standing, and lying down; the data were recorded using accelerometers. Our goal is to detect the change of activity in real-time from sequential observations. Since it is difficult to build precise parametric models for distributions of various activities, traditional parametric change-point detection methods do not work well. We compare the proposed method with a standard nonparametric multivariate sequential change-point detection procedure based on the Hotelling's T^2 statistic [132]. The raw data consists of sequences of observations for one person; each sequence may contain more than one change-points, and the time duration for each activity is also different. For this experiment, we only consider two types of transitions of activities: walking to jogging and jogging to walking. We extract 360 sequences of length 100 such that each sequence only contains one change-point.

We construct a change-point detection procedure using our optimal test as follows. Denote the data sequence as $\{\omega_t, t = 1, 2, \ldots\}$. At any possible change-point time t, we treat samples in time windows $[t - w, t - 1]$ and $[t + 1, t + w]$ as two groups of training data and find the LFDs $\{P_1^*, P_2^*\}$ by solving the convex problem in Equation (5.8). Then we calculate the detection statistic as $P_2^*(\omega_t) - P_1^*(\omega_t)$, inspired by the optimal detector in Lemma 5.1. We couple this test statistic with the CUSUM-type recursion [140], which can accumulate change and detects small deviations quickly. The recursive detection statistic is defined as $S_t = \max\{0, S_{t-1} + P_2^*(\omega_t) - P_1^*(\omega_t)\}$, with $S_0 = 0$. A change is detected when S_t exceeds a pre-specified threshold for the first time. Such scheme is similar to the combination of convex optimization solution and change-point detection procedure [34]. In the experiment, we set the window size $w = 10$ and choose the same radii for uncertainty sets using cross-validation. The Hotelling's T^2 procedure is constructed similarly. Using historical samples, we estimate the nominal (pre-change) mean $\widehat{\mu}$ and covariance $\widehat{\Sigma}$. The Hotelling's T^2 statistics at time t is defined as $(\omega_t - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\omega_t - \widehat{\mu})$ and the Hotelling procedure uses a CUSUM-type recursion: $H_t = \max\{0, H_{t-1} + (\omega_t - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\omega_t - \widehat{\mu})\}$.

We compare the expected detection delay (EDD) versus Type-I error. Here EDD is defined as the average number of samples that a procedure needs before detects a change

after it has occurred, which is a commonly used metric for sequential change-point detection [214]. The Type-I error corresponds to the probability of detecting a change when there is no change. We consider a range of thresholds such that the corresponding Type-I error is from 0.05 to 0.35. The results in Figure 5.4 show that our test significantly outperforms Hotelling's T^2 procedure in detecting the change quicker under the same Type-I error.

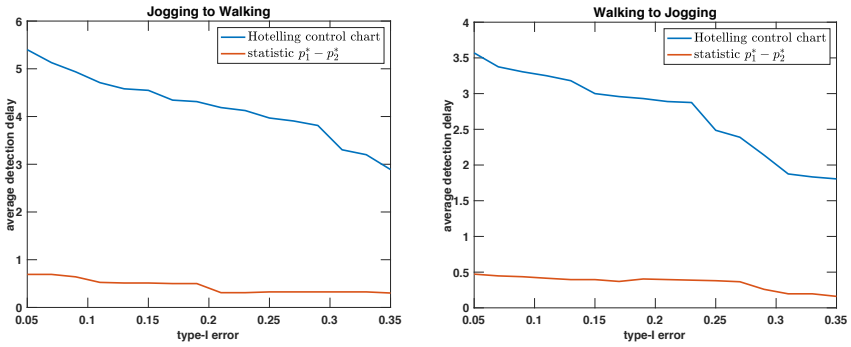


Figure 5.4: Comparison of the Expected Detection Delay (EDD) of our test with the Hotelling's T^2 procedure for detecting two type of activity transitions: jogging to walking (left) and walking to jogging (right).

5.5 Conclusion and Discussions

We present a new approach for robust hypothesis testing when there are limited "training samples" for each hypothesis. We formulate the problem as a minimax hypothesis testing problem to decide between two disjoint sets of distributions centered around empirical distributions in Wasserstein metrics. This formulation, although statistically sound – can be treated as a "data-driven" version of Huber's robust hypothesis test, is computationally challenging since it involves an infinite-dimensional optimization problem. Thus, we present a computationally efficient framework for solving the minimax test, revealing the optimal test's statistical meaning. Moreover, we characterize the optimal radius's asymptotic distribution and shed light on the optimal test's generalization property. Furthermore, we discuss how to extend the minimax test from empirical support to the whole space and

use it for the "batch" test settings and demonstrate its good performance on simulated and real data.

The method can be kernelized to handle more complex data structures (e.g., the observations are not real-valued). The kernelization can be conveniently done by replacing the metric $c(\cdot,\cdot)$ used in solving the optimal test (5.8) with other distances metrics between features after kernel transformation. Take the Euclidean norm as an example. Given a kernel function $\mathcal{K}(\cdot,\cdot)$ that measures similarity between any pair of data, the pairwise norm $c(\omega^l,\omega^m) = \left\|\omega^l - \omega^m\right\|$ in (5.8) can be replaced with the kernel version distance $\mathcal{K}(\omega^l,\omega^m)$. Moreover, this means that the proposed framework can be combined with feature selection and neural networks.

CHAPTER 6
CONVEX PARAMETER RECOVERY FOR INTERACTING MARKED PROCESSES

This chapter studies the parameter recovery for spatio-temporal marked processes. This work is mainly summarized in [100]. Section 6.1 discusses the parameter estimation of the single-state network Bernoulli process. Section 6.2 presents the performance guarantee in terms of the recovery error and the entry-wise confidence intervals. Section 6.3 and Section 6.4 discuss extensions to the multi-state Bernoulli process and non-linear link functions, respectively. Section 6.5 discusses the properties of the Maximum Likelihood estimate of parameters of the general Bernoulli process. Section 6.6 illustrates the application of the proposed approach using various simulation examples and a "real-world" data analysis of crime events in Atlanta.

6.1 Spatio-Temporal Bernoulli Process and Parameter Estimation

We consider the spatio-temporal Bernoulli process with discrete-time over discrete locations. Specifically, we assume that the discrete-time and location grid we deal with is fine enough so that we can neglect the possibility for more than one event to occur in a cell of the grid. We will model the interactions of these events in the grid.

6.1.1 Single-State Model

Define a *spatio-temporal Bernoulli process with memory depth* d as follows. We observe on discrete time horizon $\{t : -d+1 \leq t \leq N\}$ random process as follows. At time t we observe Boolean vector $\omega_t \in \mathbb{R}^K$ with entries $\omega_{tk} \in \{0, 1\}$, $1 \leq k \leq K$. Here $\omega_{tk} = 1$ and $\omega_{tk} = 0$ mean, respectively, that at time t in location k an event took/did not take place. We

set:

$$\omega^t = \{\omega_{sk}, -d+1 \le s \le t, 1 \le k \le K\} \in \mathbb{R}^{(t+d)\times K},$$
$$\omega_\tau^t = \{\omega_{sk}, \tau \le s \le t, 1 \le k \le K\} \in \mathbb{R}^{(t-\tau+1)\times K}.$$

In other words, ω^t denotes all observations (at all locations) until current time t, and ω_τ^t contains observations on time horizon from τ to t.

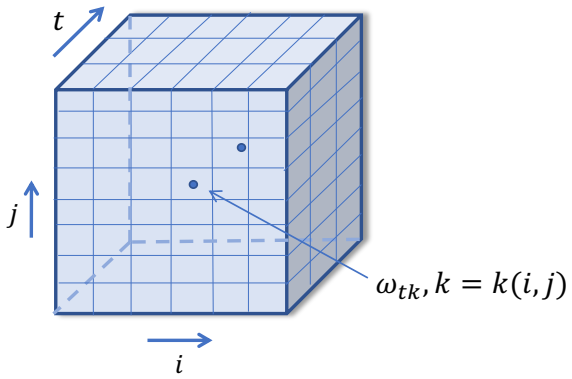


Figure 6.1: Illustration of the discretized process. Observation ω_{tk}, $k = k(i,j)$ is at the location of a $3d$ spatio-temporal grid.

We assume that for $t \ge 1$ the conditional probability of the event $\omega_{tk} = 1$, given the history ω^{t-1}, is specified as

$$\beta_k + \sum_{s=1}^{d}\sum_{\ell=1}^{K} \beta_{k\ell}^s \omega_{(t-s)\ell},\ 1 \le k \le K, \tag{6.1}$$

where $\beta = \{\beta_k, \beta_{k\ell}^s : 1 \le s \le d, 1 \le k, \ell \le K\}$ is a collection of coefficients. Here

- β_k corresponds to the *baseline intensity* at the k-th location (i.e., the intrinsic probability for an event to happen at a location without the exogenous influence, also called the birthrate);

- $\beta_{k\ell}^s$ captures the magnitude of the *influence* of an event that occurs at time $t-s$ at the ℓ-th location on chances for an event to happen at time t in the k-th location; so the sum in (6.1) represents the cumulative influence of past events at the k-th location.

Throughout this chapter, d is considered as a pre-specified parameter of the procedure. In reality, it can be tuned by verifying the "predictive abilities" of models with different d's for a given dataset. Since the probability of occurrence is between 0 and 1, we require the coefficients to satisfy

$$\begin{aligned} 0 &\leq \beta_k + \textstyle\sum_{s=1}^{d} \sum_{\ell=1}^{K} \min\left[\beta_{k\ell}^s, 0\right], \ \ \forall\, k \leq K, \\ 1 &\geq \beta_k + \textstyle\sum_{s=1}^{d} \sum_{\ell=1}^{K} \max\left[\beta_{k\ell}^s, 0\right], \ \ \forall\, k \leq K. \end{aligned} \tag{6.2}$$

Note that constraints in (6.2) allow some of the coefficients $\beta_{k\ell}^s$ to be negative, permitting the corresponding model to capture the inhibitive effect of past events. Our goal is to recover the collection of parameters β using a set of observations ω^N. Instead of using the classical loss function approach, we reduce the estimation problem to another problem with a convex structure, a variational inequality (VI) with monotone operators, see Chapter 2.3 for preliminaries.

6.1.2 Variational Inequality for Least Squares (LS) Estimation

Let $\kappa = K + dK^2$; we arrange all reals from the collection β in (6.1) into a column vector (still denoted as β):

$$\beta = [\beta_1, \ldots, \beta_K, \beta_{11}^1, \ldots, \beta_{11}^d, \beta_{1K}^1, \ldots, \beta_{1K}^d, \ldots, \beta_{KK}^1, \ldots, \beta_{KK}^d]^\top \in \mathbb{R}^\kappa.$$

Note that constraints (6.2) above state that β must reside in the polyhedral set $\mathcal{B}$ given by explicit polyhedral representation. Assume that we are given a convex compact set $\mathcal{X} \subset \mathcal{B}$ such that $\beta \in \mathcal{X}$; thus $\mathcal{X}$ means the domain of β. Clearly, the inclusion $\mathcal{X} \subset \mathcal{B}$ is a must in our model, but one can cut $\mathcal{X}$ off $\mathcal{B}$ by additional constraints reflecting additional *a priori* information on model's parameters. Our model says that for $t \geq 1$, the conditional

expectation of ω_t given ω^{t-1} is $\eta^\top(\omega_{t-d}^{t-1})\beta$:

$$\mathbb{P}_{\omega^{t-1}}\{\omega_t = 1\} = \eta^\top(\omega_{t-d}^{t-1})\beta,$$

for a known to us function $\eta(\cdot)$ which is defined on the set of all zero-one arrays $\omega_{t-d}^{t-1} \in \{0,1\}^{d\times K}$ and takes values in the matrix space $\mathbb{R}^{\kappa\times K}$:

$$\eta^\top(\omega_{t-d}^{t-1}) = \left[I_K, I_K \otimes \mathrm{vec}(\omega_{t-d}^{t-1})^T\right] \in \mathbb{R}^{K\times\kappa}, \tag{6.3}$$

where I_K is a $K\times K$ identity matrix, $\otimes$ denotes the standard Kronecker product, and $\mathrm{vec}(\cdot)$ vectorizes a matrix by stacking all columns. Note that the matrix $\eta(\omega_{t-d}^{t-1})$ is Boolean and has at most one nonzero entry in every row. Indeed, (6.1) says that a particular entry in β, β_k or $\beta_{k\ell}^s$, affects at most one entry in $\eta^\top(\omega_{t-d}^{t-1})\beta$, namely, the k-th entry, implying that each column of $\eta^\top(\cdot)$ has at most one nonzero entry.

Consider a vector field $F : \mathcal{X} \to \mathbb{R}^\kappa$, defined as

$$F(x) = \frac{1}{N}\mathbb{E}_{\omega^N}\left\{\sum_{t=1}^{N}[\eta(\omega_{t-d}^{t-1})\eta^\top(\omega_{t-d}^{t-1})x - \eta(\omega_{t-d}^{t-1})\omega_t]\right\} : \mathcal{X} \to \mathbb{R}^\kappa,$$

where $\mathbb{E}_{\omega^N}$ denotes expectation taken with respect to the distribution of ω^N (notation $\mathbb{E}_{\omega^t}$ is similarly defined). Below, all expectations and probabilities are conditional given a specific realization of the initial fragment ω_{-d+1}^0 of observations.

Observe that we have

$$\langle F(x) - F(y), x - y\rangle = \frac{1}{N}\sum_{t=1}^{N}\mathbb{E}_{\omega^N}\left\{(x-y)^\top\eta(\omega_{t-d}^{t-1})\eta^\top(\omega_{t-d}^{t-1})(x-y)\right\} \geq 0, \forall x, y \in \mathcal{X}.$$

Thus, the vector field F is *monotone* (see Chapter 2.3 for details). Moreover, we have

$F(\beta) = 0$, since

$$\begin{aligned} F(\beta) &= \tfrac{1}{N}\mathbb{E}_{\omega^N}\left\{\textstyle\sum_{t=1}^{N}\eta(\omega_{t-d}^{t-1})[\eta^\top(\omega_{t-d}^{t-1})\beta - \omega_t]\right\} \\ &= \tfrac{1}{N}\textstyle\sum_{t=1}^{N}\mathbb{E}_{\omega^t}\left\{\eta(\omega_{t-d}^{t-1})[\eta^\top(\omega_{t-d}^{t-1})\beta - \omega_t]\right\} \\ &= \tfrac{1}{N}\textstyle\sum_{t=1}^{N}\mathbb{E}_{\omega^{t-1}}\left\{\eta(\omega_{t-d}^{t-1})\left[\eta^\top(\omega_{t-d}^{t-1})\beta - \mathbb{E}_{|\omega^{t-1}}\{\omega_t\}\right]\right\} \\ &= \tfrac{1}{N}\textstyle\sum_{t=1}^{N}\mathbb{E}_{\omega^{t-1}}\left\{\eta(\omega_{t-d}^{t-1})[\eta^\top(\omega_{t-d}^{t-1})\beta - \eta^\top(\omega_{t-d}^{t-1})\beta]\right\} = 0, \end{aligned}$$

where $\mathbb{E}_{|\omega^{t-1}}$ denotes the conditional expectation given ω^{t-1}. Therefore, $\beta \in \mathcal{X}$ is a zero of the monotone operator F and therefore it is a solution to the variational inequality $\mathrm{VI}[F, \mathcal{X}]$.

Now consider the empirical version

$$F_{\omega^N}(x) = \underbrace{\left[\frac{1}{N}\sum\nolimits_{t=1}^{N}\eta(\omega_{t-d}^{t-1})\eta^\top(\omega_{t-d}^{t-1})\right]}_{A[\omega^N]} x - \underbrace{\frac{1}{N}\sum\nolimits_{t=1}^{N}\eta(\omega_{t-d}^{t-1})\omega_t}_{a[\omega^N]} \tag{6.4}$$

of vector field F. Note that $F_{\omega^N}(x)$ is monotone and affine, and its expected value is $F(x)$ at every point x.

We propose to use, as an estimate of β, a weak solution to the Sample Average Approximation of $\mathrm{VI}[F, \mathcal{X}]$, i.e., the variational inequality:

$$\text{find } z \in \mathcal{X} : \langle F_{\omega^N}(w), w - z\rangle \geq 0, \quad \forall w \in \mathcal{X}. \qquad \mathrm{VI}[F_{\omega^N}, \mathcal{X}]$$

The monotone vector field $F_{\omega^N}(\cdot)$ is continuous (even affine), so that weak solutions to $\mathrm{VI}[F_{\omega^N}, \mathcal{X}]$ are exactly the same as strong solutions (defined in Chapter 2.3). Moreover, the empirical vector field $F_{\omega^N}(x)$ is just the gradient field of the convex quadratic function

$$\Psi_{\omega^N}(x) = \frac{1}{2N}\sum_{t=1}^{N}\|\eta^\top(\omega_{t-d}^{t-1})x - \omega_t\|_2^2, \tag{6.5}$$

so that weak (same as strong) solutions to $\mathrm{VI}[F_{\omega^N}, \mathcal{X}]$ are just minimizers of this function on $\mathcal{X}$. In other words, our estimate based on solving variational inequality is an optimal

solution to the Least Squares (LS) formulation: the constrained optimization problem

$$\min_{x\in\mathcal{X}} \Psi_{\omega^N}(x) \tag{6.6}$$

with a convex quadratic objective. Problem (6.6), the same as a general variational inequality with a monotone operator, can be routinely and efficiently solved by convex optimization algorithms.

6.2 Toward Performance Guarantees

Our objective in this section is to construct *non-asymptotic* confidence sets for parameter estimates built in the previous section. Utilizing concentration inequalities for martingales, we can express these sets in terms of the process observations in the spirit of results of [101, 123, 82].

Observe that the vector of true parameters β underlying our observations not only solves variational inequality $\mathrm{VI}[F,\mathcal{X}]$, but also solves the variational inequality $\mathrm{VI}[\overline{F}_{\omega^N},\mathcal{X}]$, where

$$\overline{F}_{\omega^N}(x) = A[\omega^N]x - \underbrace{\frac{1}{N}\sum\nolimits_{t=1}^{N}\eta(\omega_{t-d}^{t-1})\eta^\top(\omega_{t-d}^{t-1})\beta}_{\overline{a}[\omega^N]}$$

with $A[\omega^N]$ defined in (6.4).

In fact, β is just a root of $\overline{F}_{\omega^N}(x)$: $\overline{F}_{\omega^N}(\beta) = 0$. Moreover, the monotone affine operators $F_{\omega^N}(x)$ and $\overline{F}_{\omega^N}(x)$ differ only in the value of constant term: in $F_{\omega^N}(x)$ this term is $a[\omega^N]$, and in $\overline{F}_{\omega^N}(x)$ this term is $\overline{a}[\omega^N]$. Thus, equivalently, β is the minimizer on $\mathcal{X}$ of the quadratic form

$$\overline{\Psi}_{\omega^N}(x) := \frac{1}{2N}\sum_{t=1}^{N}\|\eta^\top(\omega_{t-d}^{t-1})x - \eta^\top(\omega_{t-d}^{t-1})\beta\|_2^2,$$

and the functions Ψ in (6.5) and $\overline{\Psi}$ above differ only in the constant terms (which do not

affect the results of minimization) and in the linear terms. Moreover, the difference of the vectors of coefficients of linear terms is given by (due to $\overline{F}_{\omega^N}(\beta) = 0$):

$$\Delta_F := F_{\omega^N}(\beta) - \overline{F}_{\omega^N}(\beta) = F_{\omega^N}(\beta) = \overline{a}[\omega^N] - a[\omega^N] = \frac{1}{N}\sum_{t=1}^{N} \underbrace{\eta(\omega_{t-d}^{t-1})[\eta^\top(\omega_{t-d}^{t-1})\beta - \omega_t]}_{\xi_t}. \tag{6.7}$$

Note that this is the same as the difference of constant terms in $F_{\omega^N}(\cdot)$ and $\overline{F}_{\omega^N}(\cdot)$. Since the conditional expectation of ω_t given ω^{t-1} is $\eta^\top(\omega_{t-d}^{t-1})\beta$, we have $\mathbb{E}_{|\omega^{t-1}}[\xi_t] = 0$. Thus, ξ_t is a *martingale-difference*.

Concentration bounds for $F_{\omega^N}(\beta)$ can be obtained by applying general Bernstein-type inequalities for martingales.

Lemma 6.1. *For all $\epsilon \in (0,1)$ vector $F_{\omega^N}(\beta) = \Delta_F$ in (6.7) satisfies*

$$\mathbb{P}_{\omega^N}\left\{\|F_{\omega^N}(\beta)\|_\infty \geq \sqrt{\frac{\ln(2\kappa/\epsilon)}{2N}} + \frac{\ln(2\kappa/\epsilon)}{3N}\right\} \leq \epsilon. \tag{6.8}$$

Proof. Both ω_t and $\eta^\top(\omega_{t-d}^{t-1})\beta$ are vectors with nonnegative entries not exceeding 1, we have $\|\eta^\top(\omega_{t-d}^{t-1})\beta - \omega_t\|_\infty \leq 1$. Besides this, $\eta(\omega_{t-d}^{t-1})$ is a Boolean matrix with at most one nonzero in every row, whence $\|\eta^\top(\omega_{t-d}^{t-1})z\|_\infty \leq \|z\|_\infty$ for all z, thus $\|\xi_t\|_\infty \leq 1$. Furthermore, the conditional variance of components of ω_t is bounded by $1/4$, so, applying the Azuma-Hoeffding inequality [12] to components $(F_{\omega^N}(\beta))_k$, $k = 1, ..., \kappa$, of $F_{\omega^N}(\beta)$ we conclude that

$$\mathbb{P}_{\omega^N}\left\{|(F_{\omega^N}(\beta))_k| \geq \sqrt{\frac{x}{2N}} + \frac{x}{3N}\right\} \leq 2\exp\{-x\},\ \forall 1 \leq k \leq \kappa,\ x \geq 0.$$

The latter bound results in (6.8) by application of the total probability formula. □

A somewhat finer analysis allows to establish more precise data-driven deviation bounds for components of $F_{\omega^N}(\beta)$.

Lemma 6.2. *For all $y > 1$, entries $(F_{\omega^N}(\beta))_k$, $k = 1, ..., \kappa$, of $F_{\omega^N}(\beta)$ satisfy, with proba-*

bility at least $1 - 2e\big(y\big[\ln((y-1)N\big) + 2\big] + 2\big)e^{-y}$*,*

$$a[\omega^N]_k - \overline{\psi}(a[\omega^N]_k, N; y) \le (F_{\omega^N}(\beta))_k \le a[\omega^N]_k - \underline{\psi}(a[\omega^N]_k, N; y), \tag{6.9}$$

where $a[\omega^N]_k$ *is the* k*-th component of* $a[\omega^N]$ *as in* (6.4) *and lower and upper functions* $\underline{\psi}(\cdot)$*,* $\overline{\psi}(\cdot)$ *are defined in relation* (D.4) *in the Appendix.*

We are about to extract from this Lemma upper bounds on the accuracy of recovered coefficients. Recall that our estimate $\widehat{\beta} := \widehat{\beta}(\omega^N)$ solves the variational inequality $\mathrm{VI}[F_{\omega^N}, \mathcal{X}]$ with $F_{\omega^N}(x) = A[\omega^N]x - a[\omega^N]$ in (6.4). Note that $A[\omega^N]$ is positive semidefinite (we write $A \succeq 0$, and we write $A \succ 0$ for positive definite A). Given $A \in \mathbb{R}^{\kappa\times\kappa}$, $A \succeq 0$, and $p \in [1, \infty]$, define the "condition number":

$$\theta_p[A] := \max\left\{\theta \ge 0 : g^\top A g \ge \theta \|g\|_p^2, \ \ \forall g \in \mathbb{R}^\kappa\right\}. \tag{6.10}$$

Observe that $\theta_p[A] > 0$ whenever $A \succ 0$, and that for $p, p' \in [1, \infty]$ one has

$$g^\top A g \ge \frac{1}{2}\left\{\theta_p[A]\|g\|_p^2 + \theta_{p'}[A]\|g\|_{p'}^2\right\} \ge \sqrt{\theta_p[A]\theta_{p'}[A]}\|g\|_p\|g\|_{p'}. \tag{6.11}$$

The following result is immediate:

Theorem 6.1 (Bounding ℓ_p estimation error)**.** *For every* $p \in [1, \infty]$ *and every* ω^N *one has*

$$\|\widehat{\beta}(\omega^N) - \beta\|_p \le \|F_{\omega^N}(\beta)\|_\infty / \sqrt{\theta_p[A[\omega^N]]\theta_1[A[\omega^N]]}. \tag{6.12}$$

As a result, for every $\epsilon \in (0, 1)$*, the probability of the event*

$$\|\widehat{\beta}(\omega^N) - \beta\|_p \le \big(\theta_p[A[\omega^N]]\theta_1[A[\omega^N]]\big)^{-1}\left(\sqrt{\frac{\ln(2\kappa/\epsilon)}{2N}} + \frac{\ln(2\kappa/\epsilon)}{3N}\right), \ \forall p \in [1, \infty] \tag{6.13}$$

is at least $1 - \epsilon$*.*

Proof. Let us fix ω^N and set $\widehat{\beta} = \widehat{\beta}[\omega^N]$, $A = A[\omega^N]$. Since $F_{\omega^N}(\cdot)$ is continuous and $\widehat{\beta}$ is a weak solution to $\mathrm{VI}[F_{\omega^N}, \mathcal{X}]$, $\widehat{\beta}$ is also a strong solution: $\langle F_{\omega^N}(\widehat{\beta}), z - \widehat{\beta}\rangle \geq 0$ for all $z \in \mathcal{X}$; in particular, $\langle F_{\omega^N}(\widehat{\beta}), \beta - \widehat{\beta}\rangle \geq 0$. On the other hand, $F_{\omega^N}(\widehat{\beta}) = F(\beta) - A(\beta - \widehat{\beta})$. As a result, $0 \leq \langle F_{\omega^N}(\widehat{\beta}), \beta - \widehat{\beta}\rangle = \langle F_{\omega^N}(\beta) - A(\beta - \widehat{\beta}), \beta - \widehat{\beta}\rangle$, whence

$$(\beta - \widehat{\beta})^\top A(\beta - \widehat{\beta}) \leq \langle F_{\omega^N}(\beta), \beta - \widehat{\beta}\rangle \leq \|F_{\omega^N}(\beta)\|_\infty \|\beta - \widehat{\beta}\|_1. \tag{6.14}$$

Setting $p' = 1$ in (6.11), we obtain

$$(\beta - \widehat{\beta})^\top A(\beta - \widehat{\beta}) \geq \sqrt{\theta_1[A]\theta_p[A]}\|\beta - \widehat{\beta}\|_1 \|\beta - \widehat{\beta}\|_p.$$

This combines with (6.14) to imply (6.12); then (6.12) together with (6.8) imply (6.13). □

Remark 6.1 (Evaluating the condition number)**.** To assess the upper bound (6.13) one needs to compute "condition numbers" $\theta_p[A]$ of a positive definite matrix A. The computation is easy when $p = 2$, in which case $\theta_2[A]$ is the minimal eigenvalue of A, and when $p = \infty$:

$$\theta_\infty[A] = \min_{1 \leq i \leq \kappa} \left\{x^\top A x : \|x\|_\infty \leq 1, x_i = 1\right\}$$

is the minimum of κ efficiently computable quantities. In general, $\theta_1[A]$ is difficult to compute, but this quantity admits an efficiently computable tight within the factor $\pi/2$ lower bound. Specifically, for a symmetric positive definite A, $\min_z\{z^\top A z : \|z\|_1 = 1\}$ is the largest $r > 0$ such that the ellipsoid $\{z : z^\top A z \leq r\}$ is contained in the unit $\|\cdot\|_1$-ball, or, passing to polars, the largest r such that the ellipsoid $y^\top A^{-1} y \leq r^{-1}$ contains the unit $\|\cdot\|_\infty$-ball. Because of this, the definition of $\theta_1[A]$ in (6.10) is equivalent to $\theta_1[A] = \left[\max_{\|x\|_\infty \leq 1} x^\top A^{-1} x\right]^{-1}$. It remains to note that when Q is a symmetric positive semidefinite $\kappa \times \kappa$ matrix, the efficiently computable by semidefinite relaxation upper

bound on $\max_{\|x\|_\infty \le 1} x^\top Q x$, given by

$$\min_{\lambda} \left\{ \sum_i \lambda_i : \lambda_i \ge 0,\ \forall i;\ \mathbf{Diag}\{\lambda_1, ..., \lambda_\kappa\} \succeq Q \right\},$$

is tight within the factor $\pi/2$ [136].

Under favorable circumstances, we can expect that for large N the minimal eigenvalue of $A[\omega^N]$ will be of the order of one with overwhelming probability implying that the lengths of the confidence intervals (6.16) go to 0 as $N \to \infty$ at the rate $O(1/\sqrt{N})$. Note, however, that inter-dependence of the "regressors" $\eta(\omega_{t-d}^{t-1})$ across t makes it difficult to prove something along these lines.

We can use concentration bounds of Lemmas 6.1 and 6.2 to build confidence intervals for linear functionals of β. For instance, inequality (6.9) of Lemma 6.2 leads to the following estimation procedure of the linear form $e(\beta) = e^\top \beta$, $e \in \mathbb{R}^\kappa$. Given $y > 1$, consider the pair of optimization problems

$$\begin{aligned}
\underline{e}[\omega^N, y] &= \min_x \left\{ e^\top x : x \in \mathcal{X}, \underline{\psi}(a[\omega^N]_k, N; y) \le (A[\omega^N]x)_k \le \overline{\psi}(a[\omega^N]_k, N; y), \forall k \right\}, \\
\overline{e}[\omega^N, y] &= \max_x \left\{ e^\top x : x \in \mathcal{X}, \underline{\psi}(a[\omega^N]_k, N; y) \le (A[\omega^N]x)_k \le \overline{\psi}(a[\omega^N]_k, N; y), \forall k \right\},
\end{aligned} \tag{6.15}$$

where $\underline{\psi}(\cdot)$ and $\overline{\psi}(\cdot)$ are defined as in (D.4) of the appendix. These problems clearly are convex, so $\underline{e}[\omega^N, y]$ and $\overline{e}[\omega^N, y]$ are efficiently computable. Immediately, we have the following:

Lemma 6.3. *Given $y > 1$, the probability of the event*

$$\underline{e}[\omega^N, y] \le e^\top \beta \le \overline{e}[\omega^N, y],\ \ \forall e, \tag{6.16}$$

is at least $1 - 2\kappa e\big(y\big[\ln((y-1)N\big) + 2\big] + 2\big)e^{-y}$.

Indeed, when events

$$a[\omega^N]_k - \overline{\psi}(a[\omega^N]_k, N; y) \leq F_{\omega^N}(\beta)_k \leq a[\omega^N]_k - \underline{\psi}(a[\omega^N]_k, N; y),\ k = 1, ..., \kappa,$$

take place, β is a feasible solution to optimization problems in (6.15). Due to Lemma 6.2, this implies that (6.16) takes place with probability at least $1 - 2\kappa e\big(y\big[\ln((y-1)N\big) + 2\big] + 2\big)e^{-y}$.

6.3 Multi-State Spatio-Temporal Processes

In this section, we consider the multi-state spatio-temporal process in which an event outcome contains additional information about its category [131]. So far, we considered the case where at every time instant t every location k maybe be either in the state $\omega_{tk} = 0$ ("no event"), or $\omega_{tk} = 1$ ("event"). We are now extending the model by allowing the state of a location at a given time instant to take $M \geq 2$ "nontrivial" values on the top of the zero value "no event." In other words, observation of the multi-state Bernoulli process is categorical — we can either observe no event or observe one of M possible event outcomes.

We define *M-state spatio-temporal process with memory depth d* as follows. We observe a random process on time horizon $\{t : -d + 1 \leq t \leq N\}$, observation at time t being $\omega_t = \{\omega_{tk} \in \{0, 1, \ldots, M\}, 1 \leq k \leq K\}$. For every $t \geq 1$, the conditional, $\omega^{t-1} = (\omega_{-d+1}, \omega_{-d+2}, \ldots, \omega_{t-1})$ given, distribution of ω_{tk} is to be of category p, $1 \leq p \leq M$, is given by

$$\mathbb{P}_{\omega^{t-1}}\{\omega_{tk} = p\} = \beta_k(p) + \sum_{s=1}^{d}\sum_{\ell=1}^{K} \beta^s_{k\ell}(p, \omega_{(t-s)\ell}), \tag{6.17}$$

and the probability for ω_{tk} to take value 0 (no event or "ground event") is the complemen-

tary probability

$$\mathbb{P}_{\omega^{t-1}}\{\omega_{tk}=0\}=1-\sum_{p=1}^{M}\left[\beta_k(p)+\sum_{s=1}^{d}\sum_{\ell=1}^{K}\beta^s_{k\ell}(p,\omega_{(t-s)\ell})\right].$$

In other words, $\beta^s_{k\ell}(p,q)$ is the contribution of the location ℓ in state $q\in\{0,1,\dots,M\}$ at time $t-s$ to the probability for the location k to be in state $p\in\{1,\dots,M\}$ at time t, and $\beta_k(p)$, $p\in\{1,\dots,M\}$ is the "endogenous" component of the probability of the latter event.

Of course, for this description to make sense, the β-parameters should guarantee that for every ω^{t-1}, that is, for every collection $\{\omega_{\tau\ell}\in\{0,1,\dots,M\}:\tau<t,1\le\ell\le K\}$, the prescribed by (6.17) probabilities are nonnegative and their sum over $p=1,\dots,M$ is ≤ 1. Thus, the β-parameters should satisfy the system of constraints

$$\begin{array}{l}0\le\beta_k(p)+\sum_{s=1}^{d}\sum_{\ell=1}^{K}\min\limits_{0\le q\le M}\beta^s_{k\ell}(p,q),\ 1\le p\le M,\ 1\le k\le K,\\ 1\ge\sum_{p=1}^{M}\beta_k(p)+\sum_{s=1}^{d}\sum_{\ell=1}^{K}\max\limits_{0\le q\le M}\sum_{p=1}^{M}\beta^s_{k\ell}(p,q),\ 1\le k\le K.\end{array}\tag{6.18}$$

The solution set $\mathcal{B}$ of this system is a polyhedral set given by explicit polyhedral representations. We are given convex compact set $\mathcal{X}$ in the space of parameters $\beta=\{\beta_k,\beta^s_{k\ell}(p,q),1\le s\le d,1\le k,\ell\le K,1\le p\le M,0\le q\le M\}$ such that $\mathcal{X}$ contains the true parameter β of the process we are observing, and $\mathcal{X}$ is contained in the polytope $\mathcal{B}$ given by constraints (6.18).

We arrange the collection of β-parameters associated with a M-state spatio-temporal process with memory depth d into a column vector (still denoted β) and denote by κ the dimension of β. In general, $\kappa=KM+dK^2M^2$. However, depending on application, it could make sense to postulate that some of the components of β are zeros, thus reducing the actual dimension of β; for example, we could assume that $\beta_{k\ell}(\cdot,\cdot)=0$ for some "definitely non-interacting" pairs k,ℓ of locations.

Note that (6.17) says that the M-dimensional vector of conditional probabilities for ω_{tk}

to take values $p \in \{1, \ldots, M\}$ given ω^{t-1} is $[\eta_k^\top(\omega_{t-d}^{t-1})\beta]_p$ with known to us function $\eta_k(\cdot)$ defined on the set of arrays $\omega_{t-d}^{t-1} \in \{0, 1, \ldots, M\}^{d\times K}$ and taking values in the space of $\kappa \times M$ matrices. Note that the value of ω_{tk} is the index of the category, and does not mean magnitude. Same as above, $\eta_k(\omega_{d-1}^{t-1})$ is a Boolean matrix.

To proceed, for $0 \leq q \leq M$, let $\chi_q \in \mathbb{R}^M$ be defined as follows: $\chi_0 = 0 \in \mathbb{R}^M$, and χ_q, $1 \leq q \leq M$, is the q-th vector of the standard basis in $\mathbb{R}^M$. In particular, the state ω_{tk} can be encoded by vector $\bar{\omega}_{tk} = \chi_{\omega_{tk}}$, and the state of our process at time t — by the block vector $\overline{\omega}_t \in \mathbb{R}^{MK}$ with blocks $\bar{\omega}_{tk} \in \mathbb{R}^M$, $k = 1, ..., K$. In other words: the k-th block in $\overline{\omega}_t$ is an M-dimensional vector which is the p-th basic orth of $\mathbb{R}^M$ when $\omega_{tk} = p \geq 1$, and is the zero vector when $\omega_{tk} = 0$. Arranging $\kappa \times M$ matrices $\eta_k(\cdot)$ into a matrix

$$\eta(\cdot) = [\eta_1(\cdot), ..., \eta_K(\cdot)] \in \{0, 1\}^{\kappa\times MK},$$

we obtain

$$\mathbb{E}_{|\omega^{t-1}}\{\overline{\omega}_t\} = \eta^\top(\omega_{t-d}^{t-1})\beta \in \mathbb{R}^{MK},$$

where $\mathbb{E}_{|\omega^{t-1}}$ is the conditional expectation given ω^{t-1}. Note that similarly to Section 6.1.1, (6.17) says that every particular entry in β, $\beta_k(p)$ or $\beta_{k\ell}^s(p, q)$, affects at most one of the entries in the block vector $[\eta_1^\top(\omega_{t-d}^{t-1})\beta; ...; \eta_K^\top(\omega_{t-d}^{t-1})\beta]$ specifically, the p-th entry of the k-th block, so that the Boolean matrix $\eta(\omega_{t-d}^{t-1})$ has at most one nonzero entry in every row.

Note that the spatio-temporal Bernoulli process with memory depth d, as defined in Section 6.1.1, is a special case of M-state ($M = 1$) spatio-temporal process with memory depth d, the case where state 0 at a location contributes nothing to probability of state 1 in another location at a later time, that is, $\beta_{k\ell}^s(1, 0) = 0$ for all s, k, ℓ.

As an illustration, consider a spatio-temporal model of crime events of different types, e.g., burglary and robbery, in a geographic area of interest. We split the area into K non-overlapping cells, which will be our locations. Selecting the time step in such a way that we can ignore the chances for two or more crime events to occur in the same spatio-temporal

cell, we can model the history of crime events in the area as a $M = 2$-state spatio-temporal process, with additional to (6.18) convex restrictions on the vector of parameters β expressing our *a priori* information on the probability $\beta_k(p)$ of a "newborn" crime event of category p to occur at time instant t at location k and on the contribution $\beta^s_{k\ell}(p,q)$ of a crime event of category q in spatio-temporal cell $\{t-s,\ell\}$ to the probability of crime event of category p, $p \geq 1$, to happen in the spatio-temporal cell $\{t,k\}$.

The problem of estimating parameters β of the M-state spatio-temporal process from observations of this process can be processed exactly as in the case of the single state spatio-temporal Bernoulli process. Specifically, observations ω^N give rise to two monotone and affine vector fields on $\mathcal{X}$, the first observable and the second unobservable:

$$\begin{aligned} F_{\omega^N}(x) &= \underbrace{\left[\frac{1}{N}\sum_{t=1}^{N}\eta(\omega_{t-d}^{t-1})\eta^\top(\omega_{t-d}^{t-1})\right]}_{A[\omega^N]}x - \underbrace{\frac{1}{N}\sum_{t=1}^{N}\eta(\omega_{t-d}^{t-1})\overline{\omega}_t}_{a[\omega^N]}, \\ \overline{F}_{\omega^N}(x) &= A[\omega^N]x - A[\omega^N]\beta. \end{aligned} \tag{6.19}$$

The two fields differ only in constant term, β is a root of the second field, and the difference of constant terms, same as the vector $F_{\omega^N}(\beta)$ due to $\overline{F}_{\omega^N}(\beta) = 0$, are zero-mean satisfying, for exactly the same reasons as in Section 6.2, concentration bounds (6.8) and (6.9) of Lemmas 6.1 and 6.2. To recover β from observations, we may use the Least Squares (LS) estimate obtained by solving variational inequality $\mathrm{VI}[F_{\omega^N},\mathcal{X}]$ with the just defined F_{ω^N}, or, which is the same, by solving

$$\min_{x\in\mathcal{X}}\left\{\Psi_{\omega^N}(x) := \frac{1}{2N}\sum_{t=1}^{N}\|\eta^\top(\omega_{t-d}^{t-1})x - \overline{\omega}_t\|_2^2\right\}. \tag{6.20}$$

Note that (6.8) and (6.9), by the same argument as in Section 6.2, imply the validity in our present situation of Theorem 6.1 and Lemma 6.3.

6.4 Nonlinear Link Function

So far, our discussion focused on "linear" link functions, where past events contribute *additively* to the probability of a specific event in a given spatio-temporal cell. We now consider the case of non-linear link functions. This generalizes our model to allow more complex spatio-temporal interactions.

We first consider the single-state process. Let $\phi(\cdot) : D \to \mathbb{R}^K$ be a continuous *monotone* vector field defined on a closed convex domain $D \subset \mathbb{R}^K$ such that

$$y \in D \Rightarrow 0 \leq \phi(y) \leq [1; \ldots; 1].$$

For example, we may consider "sigmoid field" $\phi(u) = [\phi_1(u); ...; \phi_K(u)]$ with

$$[\phi(u)]_k = \frac{\exp\{u_k\}}{1 + \exp\{u_k\}}, \quad k \leq K,\ D = \mathbb{R}^K.$$

Given positive integer N, we define a *spatio-temporal Bernoulli process with memory depth d and link function ϕ* as a random process with realizations $\{\omega_{tk} \in \{0, 1\}, k \leq K, -d+1 \leq t \leq N\}$ in the same way it was done in Section 6.1.1 with assumptions of Section 6.1.1 replaced as follows. We assume a given a convex compact set $\mathcal{X} \subset \mathbb{R}^\kappa$ such that the vector of parameters β underlying the observed process belongs to $\mathcal{X}$ and every $\beta \in \mathcal{X}$ satisfies

$$\eta^\top(\omega_{t-d}^{t-1})\beta \in D, \quad \forall 1 \leq t \leq N \tag{6.21}$$

with given functions $\eta(\omega_{t-d}^{t-1})$ taking values in the space of $\kappa \times K$ matrices; and the conditional expectation of $\omega_t \in \{0, 1\}^K$ given ω^{t-1} is $\phi(\eta^\top(\omega_{t-d}^{t-1})\beta)$.

Let us set

$$\begin{array}{rl} F(x) = & \frac{1}{N}\mathbb{E}_{\omega^N}\left\{\sum_{t=1}^N\left[\eta(\omega_{t-d}^{t-1})\phi\left(\eta^\top(\omega_{t-d}^{t-1})x\right)-\eta(\omega_{t-d}^{t-1})\omega_t\right]\right\}:\mathcal{X}\to\mathbb{R}^\kappa,\\ F_{\omega^N}(x) = & \underbrace{\frac{1}{N}\sum_{t=1}^N\eta(\omega_{t-d}^{t-1})\phi\left(\eta^\top(\omega_{t-d}^{t-1})x\right)}_{A_{\omega^N}(x)}-\underbrace{\frac{1}{N}\sum_{t=1}^N\eta(\omega_{t-d}^{t-1})\omega_t}_{a[\omega^N]}:\mathcal{X}\to\mathbb{R}^\kappa,\\ \overline{F}_{\omega^N}(x) = & A_{\omega^N}(x)-\underbrace{\frac{1}{N}\sum_{t=1}^N\eta(\omega_{t-d}^{t-1})\phi\left(\eta^\top(\omega_{t-d}^{t-1})\beta\right)}_{\overline{a}[\omega^N]}:\mathcal{X}\to\mathbb{R}^\kappa. \end{array} \tag{6.22}$$

We are now essentially in the situation of Section 6.1.2 (where we considered the special case $\phi(z)\equiv z$ of our present situation). Specifically, $F(\cdot)$ is a monotone (albeit not affine) vector field on $\mathcal{X}$, $F(\beta)=0$. The empirical version $F_{\omega^N}(x)$, for every $x\in\mathcal{X}$, is a monotone on $\mathcal{X}$ vector field which is an unbiased estimate of $F(x)$. Besides this, $\overline{F}_{\omega^N}(x)$ is a monotone on $\mathcal{X}$ vector field, and the true vector of parameters β underlying our observations solves the variational inequality $\mathrm{VI}[\overline{F}_{\omega^N},\mathcal{X}]$ (is a root of $\overline{F}_{\omega^N}$). These observations suggest estimating β by weak solution to the variational inequality $\mathrm{VI}[F_{\omega^N},\mathcal{X}]$.

Note that, same as above, vector fields F_{ω^N} and $\overline{F}_{\omega^N}$ differ only in the constant terms, and this difference is nothing but $F_{\omega^N}(\beta)$ due to $\overline{F}_{\omega^N}(\beta)=0$; moreover $\xi_t=\eta(\omega_{t-d}^{t-1})(\omega_t-\phi(\eta^\top(\omega_{t-d}^{t-1})\beta))$ is a martingale difference. Though deviation probabilities for $F_{\omega^N}(\beta)$ do not obey the same bound as in the case of $\phi(z)\equiv z$ (since the matrices $\eta(\omega_{t-d}^{t-1})$ now not necessarily are Boolean with at most one nonzero in a row), the reasoning which led us to (6.8) demonstrates that the vector $F_{\omega^N}(\beta)$ in our present situation does obey the bound

$$\mathbb{P}_{\omega^N}\left\{\|F_{\omega^N}(\beta)\|_\infty\geq\Theta\left[\sqrt{\frac{\ln(2\kappa/\epsilon)}{2N}}+\frac{\ln(2\kappa/\epsilon)}{3N}\right]\right\}\leq\epsilon,\ \forall\epsilon\in(0,1), \tag{6.23}$$

where Θ is the maximum, over all possible ω_{d-1}^{t-1}, of the $\|\cdot\|_1$-norm of rows of $\eta(\omega_{t-d}^{t-1})$. Note that in the situation of this section, our $O(1/\sqrt{N})$ exponential bounds on large deviations of $F_{\omega^N}(\beta)$ from zero, while being good news, do *not* result in easy-to-compute on-line upper-risk bounds and confidence intervals for linear functions of β. Indeed, in order to adjust to our present situation Theorem 6.1, we need to replace the condition numbers $\theta_p[\cdot]$

with constants of strong monotonicity of the vector field $F_{\omega^N}(\cdot)$ on $\mathcal{X}$. On the other hand, to adopt the result of Lemma 6.3 in the present setting, we need to replace the quantities $\overline{e}$ and $\underline{e}$, see (6.15), with the maximum (resp., minimum) of the linear form $e^\top x$ over the set $\{x \in \mathcal{X} : \|F_{\omega^N}(x)\|_\infty \leq \delta\}$. Both these tasks for a *nonlinear* operator $F_{\omega^N}(\cdot)$ seem to be problematic.

The construction in the previous paragraph can be extended to M-state processes. Below, with a slight abuse of notation, we redefine notation for the multi-state processes.

Let us identify two-dimensional $K \times M$ array $\{a_{k\ell} : 1 \leq k \leq K, 1 \leq \ell \leq M\}$ with KM-dimensional block vector with K blocks $[a_{k1}; a_{k2}; \ldots; a_{kM}]$, $1 \leq k \leq K$, of dimension M each. With this convention, a parametric $K \times M$ array $\psi(z) = \{\psi_{kp}(z) \in \mathbb{R} : k \leq K, 1 \leq p \leq M\}$ depending on KM-dimensional vector z of parameters becomes a vector field on $\mathbb{R}^{KM}$. Assume that we are given an array $\phi(\cdot) = \{\phi_{kp}(\cdot) \in \mathbb{R} : k \leq K, 1 \leq p \leq M\}$ of the outlined structure such that vector field $\phi(\cdot)$ is continuous and monotone on a closed convex domain $D \subset \mathbb{R}^{KM}$, and for all $y \in D$

$$0 \leq \phi_{kp}(y) \leq 1,\ 1 \leq p \leq M, 1 \leq k \leq K \ \ \& \ \ \sum_{p=1}^{M} \phi_{kp}(y) \leq 1, \quad 1 \leq k \leq K. \tag{6.24}$$

We assume that the conditional probability for location k at time t to be in state $p \in \{1, \ldots, M\}$ (i.e., to have $\omega_{tk} = p$) given ω^{t-1} is $\phi_{kp}(\eta^\top(\omega_{t-d}^{t-1})\beta)$ for some vector of parameters $\beta \in \mathbb{R}^\kappa$ and known to us function $\eta(\cdot)$ taking values in the space of $\kappa \times KM$ matrices and such that $\eta^\top(\omega_{d-1}^{t-1})\beta \in D$ whenever $\omega_{\tau k} \in \{0, 1, ..., M\}$ for all τ and k. As a result, the conditional probability to have $\omega_{tk} = 0$ is $1 - \sum_{p=1}^{M} \phi_{kp}(\eta^\top(\omega_{t-d}^{t-1})\beta)$.

In addition, we assume that we are given a convex compact set $\mathcal{X} \subset \mathbb{R}^\kappa$ such that $\beta \in \mathcal{X}$ and for all such β and for all $\{\omega_{\tau k} \in \{0, 1, ..., M\}, \forall \tau, k\}$, $\eta^\top(\omega_{t-d}^{t-1})\beta \in D$. Same as in Section 6.3, we encode the collection $\{\omega_{tk} : 1 \leq k \leq K\}$ of locations' states at time t by block vector $\overline{\omega}_t$ with K blocks of dimension M each, with the k-th block equal to the ω_{tk}-th vector of the standard basis in $\mathbb{R}^M$ when $\omega_{tk} > 0$ and equal to 0 when $\omega_{tk} = 0$. We

clearly have

$$\mathbb{E}_{|\omega^{t-1}}\left\{\overline{\omega}_t\right\} = \phi(\eta^\top(\omega_{t-d}^{t-1})\beta).$$

Setting

$$\begin{array}{rl} F(x) = \frac{1}{N}\mathbb{E}_{\omega^N}\left\{\sum_{t=1}^N\left[\eta(\omega_{t-d}^{t-1})\phi\left(\eta^\top(\omega_{t-d}^{t-1})x\right) - \eta(\omega_{t-d}^{t-1})\overline{\omega}_t\right]\right\} : \mathcal{X} \to \mathbb{R}^\kappa, \\ F_{\omega^N}(x) = \underbrace{\frac{1}{N}\sum_{t=1}^N \eta(\omega_{t-d}^{t-1})\phi\left(\eta^\top(\omega_{t-d}^{t-1})x\right)}_{A_{\omega^N}(x)} - \underbrace{\frac{1}{N}\sum_{t=1}^N \eta(\omega_{t-d}^{t-1})\overline{\omega}_t}_{a[\omega^N]} : \mathcal{X} \to \mathbb{R}^\kappa \\ \overline{F}_{\omega^N}(x) = A_{\omega^N}(x) - \underbrace{\frac{1}{N}\sum_{t=1}^N \eta(\omega_{t-d}^{t-1})\phi\left(\eta^\top(\omega_{t-d}^{t-1})\beta\right)}_{\overline{a}[\omega^N]} : \mathcal{X} \to \mathbb{R}^\kappa, \end{array} \tag{6.25}$$

(cf. equation (6.22)), we can repeat word by word the previous comment for single-state process with nonlinear links.

6.5 Maximum Likelihood (ML) Estimate

In the previous sections, we have discussed the Least Squares estimate of the parameter vector β. Now, we consider commonly used in statistics alternative approach based on the Maximum Likelihood (ML) estimation. ML estimate is obtained by maximizing over $\beta \in \mathcal{X}$ the conditional likelihood of what we have observed, the condition being the actually observed values of ω_{tk} for $-d+1 \le t \le 0$ and $1 \le k \le K$. In this section, we study the properties of the ML estimate and show that its calculation reduces to a convex optimization problem.

6.5.1 ML Estimation: Case of Linear Link Function

We start by considering the single-state model. Assume, in addition to what has been already assumed, that for every t random variables ω_{tk} are conditionally independent across

k given ω^{t-1}. Then the negative log-likelihood, conditioned by the value of ω^0, is given by

$$L(\beta) = \tfrac{1}{N}\sum_{t=1}^{N}\sum_{k=1}^{K}\Big[-\omega_{tk}\ln\Big(\beta_k + \sum_{s=1}^{d}\sum_{\ell=1}^{K}\beta^s_{k\ell}\omega_{(t-s)\ell}\Big) - (1-\omega_{tk})\ln\Big(1-\beta_k-\sum_{s=1}^{d}\sum_{\ell=1}^{K}\beta^s_{k\ell}\omega_{(t-s)\ell}\Big)\Big].$$

Note that $L(\beta)$ is a convex function, so the ML estimate in our model reduces to the convex program:

$$\min_{x\in\mathcal{X}} L(x). \tag{6.26}$$

For the multi-state model, assume that states ω_{tk} at locations k at time t are conditionally independent across $k \le K$ given ω^{t-1}. Then the ML estimate is given by minimizing, over $\beta \in \mathcal{X}$, the conditional negative log-likelihood of collection ω^N of observations (the condition being the initial segment ω^0 of the observation). The objective in this minimization problem is the convex function

$$L_{\omega^N}(\beta) = -\frac{1}{N}\sum_{t=1}^{N}\sum_{k=1}^{K}\psi_{tk}(\beta,\omega^N),$$

where

$$\psi_{tk}(\beta,\omega^N) = \begin{cases} \ln\left([\eta_k^\top(\omega_{t-d}^{t-1})\beta]_{\omega_{tk}}\right), & \omega_{tk}\in\{1,\dots,M\},\\ \ln\Big(1-\sum_{j=1}^{M}[\eta_k^\top(\omega_{t-d}^{t-1})\beta]_j\Big), & \omega_{tk}=0. \end{cases} \tag{6.27}$$

We are about to show that the ML estimate has a structure similar to the LS estimator that we have dealt within Section 6.1, and obeys bounds similar to (6.23). Given a small positive tolerance ϱ, consider M-state spatio-temporal process with K locations and vector of parameters $\beta \in \mathbb{R}^\kappa$, as defined in Section 6.3, restricted to reside in the polyhedral set B_ϱ

cut off $\mathbb{R}^\kappa$ by "ϱ-strengthened" version of constraints (6.18), specifically, the constraints

$$\begin{array}{c} \varrho \le \beta_k(p) + \sum_{s=1}^d \sum_{\ell=1}^K \min\limits_{0\le q\le M} \beta^s_{k\ell}(p,q),\ 1 \le p \le M\ , 1 \le k \le K, \\ 1-\varrho \ge \sum_{p=1}^{M-1} \beta_k(p) + \sum_{s=1}^d \sum_{\ell=1}^K \max\limits_{0\le q\le M} \sum_{p=1}^M \beta^s_{k\ell}(p,q),\ 1 \le k \le K. \end{array} \tag{6.28}$$

The purpose of strengthening the constraints on β is to make the maximum likelihood, to be defined below, continuously differentiable on the given parameter domain.

In what follows, we treat vectors from $\mathbb{R}^{KM}$ as block vectors with K blocks of dimension M each. For such a vector z, $[z]_{kp}$ stands for the p-th entry in the k-th block of z. Let

$$Z_0 = \left\{ \omega \in \mathbb{R}^{MK} : \omega \ge 0, \sum_{p=1}^M [\omega]_{kp} \le 1,\ \forall k \le K \right\}.$$

Similarly, for a small positive tolerance ϱ, define

$$Z_\varrho = \left\{ z \in \mathbb{R}^{MK} : [z]_{kp} \ge \varrho,\ \forall k,p,\ \sum_{p=1}^M [z]_{kp} \le 1-\varrho,\ \forall k \right\} \subset Z_0.$$

We associate with a vector $w \in Z_0$ the convex function $\mathcal{L}_w : Z_\varrho \to \mathbb{R}$,

$$\mathcal{L}_w(z) := -\sum_{k=1}^K \left[\sum_{p=1}^M [w]_{kp} \ln([z]_{kp}) + [1-\sum_{p=1}^M [w]_{kp}] \ln(1-\sum_{p=1}^M [z]_{kp}) \right]. \tag{6.29}$$

From now on, assume that we are given a convex compact set $\mathcal{X} \subset B_\varrho$ known to contain the true vector β of parameters. Then the problem of minimizing the negative log-likelihood becomes

$$\min_{x\in\mathcal{X}} \left\{ L_{\omega^N}(x) = \frac{1}{N} \sum_{t=1}^N \mathcal{L}_{\overline{\omega}_t}(\eta^\top(\omega_{t-d}^{t-1})x) \right\}, \tag{6.30}$$

where $\overline{\omega}_t = \overline{\omega}_t(\omega^t)$ encodes, as explained in Section 6.3, the observations at time t, and $\eta(\omega_{t-d}^{t-1})$ are as defined in Section 6.3.

Note that by construction, $\overline{\omega}_t$ belongs to Z_0. Moreover, by construction, we have $\eta^\top(\omega_{t-d}^{t-1})x \in Z_\varrho$ whenever $x \in B_\varrho$ and $\omega_{tk} \in \{0,1,...,M\}$ for all t and k. Now, min-

imizers of $L_{\omega^N}(x)$ over $x \in \mathcal{X}$ are exactly the solutions of the variational inequality stemming from $\mathcal{X}$ and the monotone and smooth vector field (the smoothness property is due to $L_{\omega^N}(x)$ being convex and smooth on $\mathcal{X}$):

$$F_{\omega^N}(x) = \nabla_x L_{\omega^N}(x) = \frac{1}{N}\sum_{t=1}^{N} \eta(\omega_{t-d}^{t-1})\theta(\eta^\top(\omega_{t-d}^{t-1})x, \overline{\omega}_t(\omega^t))$$

with

$$\theta(z,\omega) = \nabla_z \mathcal{L}_w(z) = -\sum_{k=1}^{K}\left[\sum_{p=1}^{M}\frac{[w]_{kp}}{[z]_{kp}}e^{kp} - \frac{1-\sum_{p=1}^{M}[w]_{kp}}{1-\sum_{p=1}^{M}[z]_{kp}}\sum_{p=1}^{M}e^{kp}\right],\ [w \in Z_0]$$

where $e^{kp} \in \mathbb{R}^{KM}$ is the block-vector with the p-th vector of the standard basis in $\mathbb{R}^M$ as the k-th block and all other blocks equal to 0.

Note that we clearly have

$$w \in Z_\varrho \Rightarrow \theta(w,w) = 0. \tag{6.31}$$

Let us show that $F_{\omega^N}(\beta)$ is "typically small": its magnitude obeys the large deviation bounds similar to (6.8) and (6.23). Indeed, let us set $\overline{z}_t(\omega^{t-1}) = \eta^\top(\omega_{t-d}^{t-1})\beta$, so that $\overline{z}_t \in Z_\varrho$ due to $\beta \in B_\varrho$. Invoking (6.31) with $w = \overline{z}_t(\omega^{t-1})$, we have

$$F_{\omega^N}(\beta) = \frac{1}{N}\sum_{t=1}^{N}\underbrace{\eta(\omega_{t-d}^{t-1})\vartheta_t[\omega^t]}_{\xi_t},$$

where

$$\vartheta_t[\omega^t] = -\sum_{k=1}^{K}\left[\sum_{p=1}^{M}\frac{[\overline{\omega}_t(\omega^t)]_{kp} - [\overline{z}_t(\omega^{t-1})]_{kp}}{[\overline{z}_t(\omega^{t-1})]_{kp}}e^{kp} + \frac{\sum_{p=1}^{M}[[\overline{z}_t]_{kp} - [\overline{\omega}_t(\omega^t)]_{kp}]}{1-\sum_{p=1}^{M}[\overline{z}_t(\omega^{t-1})]_{kp}}\sum_{p=1}^{M}e^{kp}\right].$$

Since the conditional expectation of $[\overline{\omega}_t(\omega^t)]_{kp}$ given ω^{t-1} equals $[\overline{z}_t(\omega^{t-1})]_{kp}$ the conditional expectation of ξ_t given ω^{t-1} is zero. Besides this, random vectors ξ_t take their values

in a bounded set (of size depending on ϱ). As a result, $\|F_{\omega^N}(\beta)\|_\infty$ admits bound on probabilities of large deviations of the form (6.23), with properly selected (and depending on ϱ) factor Θ. However, for the reasons presented in Section 6.4, extracting from this bound meaningful conclusions on the accuracy of the ML estimate is a difficult task, and it remains an open problem.

Remark 6.2 (Decomposition of LS and ML estimation)**.** In the models we have considered, the optimization problems (6.6), (6.20), (6.26), and (6.30), we aim to solve when building the LS and the ML estimates under mild assumptions are decomposable (in spite of the fact that the observations are dependent). Indeed, vector

$$\beta = \{\beta_{kp}, \beta^s_{k\ell}(p,q),\ 1 \le k,\ell \le K,\ 1 \le p \le M,\ 0 \le q \le M,\ 1 \le s \le d\}$$

of the model parameters can be split into K subvectors

$$\beta^k = \{\beta_{kp}, \beta^s_{k\ell}(p,q),\ 1 \le \ell \le K,\ 1 \le p \le M,\ 0 \le q \le M,\ 1 \le s \le d\},\quad k = 1, ..., K.$$

It is immediately seen that the objectives to be minimized in the problems in question are sums of K terms, with the k-th term depending only on x^k. As a result, if the domain $\mathcal{X}$ summarizing our a priori information on β is decomposable: $\mathcal{X} = \{x : x^k \in \mathcal{X}_k, 1 \le k \le K\}$, the optimization problems yielding the LS and the ML estimates are collections of K uncoupled convex optimization problems in variables x^k. Moreover, under favorable circumstances optimization problem (6.20) admits even finer decomposition. Namely, splitting β^k into subvectors

$$\beta^{kp} = \{\beta_{kp}, \beta^s_{k\ell}(p,q),\ 1 \le \ell \le K,\ 1 \le s \le d,\ 0 \le q \le M\},$$

it is easily seen that the objective in (6.20) is the sum, over $k \le K$ and $p \le M$, of functions $\Psi^{kp}_{\omega^N}(x^{kp})$. As a result, when $\mathcal{X} = \{x : x^{kp} \in \mathcal{X}_{kp}, 1 \le k \le K, 1 \le p \le M\}$, (6.20)

is a collection of KM uncoupled convex problems $\min_{x^{kp}\in\mathcal{X}_{kp}} \Psi^{kp}_{\omega^N}(x^{kp})$. The outlined decompositions may be used to accelerate the solution process.

6.5.2 ML Estimate: General Link Functions

Let us now derive ML estimate for the case of nonlinear link function considered in Section 6.4. In this situation, we strengthen constraints (6.24) on D to

$$y \in D \Rightarrow \varrho \le \phi_{kp}(y),\ \sum_{p=1}^{M} \phi_{kp}(y) \le 1-\varrho, \quad 1 \le k \le K,\ 1 \le p \le M,$$

with some $\varrho > 0$. Assuming that ω_{tk}'s are conditionally independent across k given ω^{t-1}, computing ML estimate for the general link-function reduces to solving problem (6.30) with $\mathcal{L}_w(z) : D \to \mathbb{R}$, $w \in Z_0$, given by

$$\mathcal{L}_w(z) = -\sum_{k=1}^{K}\left[\sum_{p=1}^{M}[w]_{kp}\ln(\phi_{kp}(z)) + [1-\sum_{p=1}^{M}[w]_{kp}]\ln(1-\sum_{p=1}^{M}\phi_{kp}(z))\right].$$

Assuming ϕ continuously differentiable on D and $\mathcal{L}_w(\cdot)$ convex on D, we can repeat, with straightforward modifications, everything that was said above (that is, in the special case of $\phi(z) \equiv z$), including exponential bounds on probabilities of large deviations of $F_{\omega^N}(\beta)$. However, in general, beyond the case of affine $\phi_{kp}(\cdot)$, function $\mathcal{L}_w(\cdot)$ becomes nonconvex. This is due to the fact that convexity on D of functions

$$-\ln(\phi_{kp}(\cdot)),\ \ -\ln\Big(1-\sum_{p}\phi_{kp}(\cdot)\Big)$$

is a rare commodity. As a special case, convexity of these functions does take place in the case logistic link function

$$\phi_{kp}(z) = \frac{\exp\{a_{kp}(z)\}}{\sum_{q=0}^{M}\exp\{a_{kq}(z)\}}$$

with functions $a_{kq}(z)$, $0 \leq q \leq M$ that are affine in z.

6.6 Numerical Experiments

6.6.1 Experiments with Simulated Data

This section presents the results of several simulation experiments illustrating applications of the proposed Bernoulli process models. We compare performances of Least Squares (LS) and Maximum Likelihood (ML) estimates in terms of ℓ_1, ℓ_2, and ℓ_∞ norms of the error of parameter vector recovery. We assume that d (or a reasonable upper bound of it) is known in our simulation examples. The bracket percentage inside the table below shows the norm of the error relative to the norm of the corresponding true parameter vector.

Single state spatio-temporal processes

First, consider a single state setting with memory depth $d = 8$ and the number of locations $K = 8$. The true parameter values are selected randomly from the set $\mathcal{X}_0$ as follows:

- $\beta_k \geq 0$, $\beta^s_{kl} \geq 0$; and $\beta_k + \sum_{s=1}^{d} \sum_{\ell=1}^{K} \beta^s_{k\ell} \leq 1$, $\forall k$;
- $\beta^s_{k\ell} = 0$ when $|k - \ell| > 1$ (interactions are local);
- For every $1 \leq k, \ell \leq K$, $\beta^s_{k\ell}$ is a non-increasing convex function of s.

Above, the convexity of a function $f(s)$ in $s \in G = \{1, \ldots, d\}$ means that the function is the restriction of a convex function on the segment $[1, d]$ onto the grid G or, which is the same, that $f(s-1) - 2f(s) + f(s+1) \geq 0, s = 2, 3, \ldots, d-1$. This translates into the constraint $\beta^{s-1}_{k,\ell} - 2\beta^s_{k,\ell} + \beta^{s+1}_{k,\ell} \geq 0, s = 2, 3, \ldots, d-1, \forall k, \ell$. Note that we have imposed additional constraints than (6.2). More specifically, we assume interacting pairs should be adjacent in the sense that $|k - \ell| \leq 1$, and that $\beta^s_{k,\ell}$ is monotone and a convex-function in s.

We report the performance of the LS estimate (obtained by solving $\mathrm{VI}[F_{\omega^N}, \mathcal{X}]$) and the ML estimate (obtained by solving (6.26)). To have a fair comparison, we do not introduce any additional constraints on the interaction coefficients in our estimation procedure

(meaning that the LS and ML estimates do not have any prior knowledge about $\mathcal{X}_0$ and their assumed admissible set $\mathcal{X}$ is much larger than $\mathcal{X}_0$). Utilizing the Matlab implementation [215] of the EM algorithm, we also compute estimations of parameters of the commonly used model of Hawkes process with exponential temporal kernel (see, e.g., [223]). The latter is equivalent to assuming that $\beta^s_{k\ell} = a_{k\ell}\tau e^{-\tau s}$, $s = 1, 2, \ldots$, where $\tau > 0$ is the decay rate parameter and $a_{k\ell} > 0$ represents the interactions between two locations.

Figure 6.2 shows the recovered interaction coefficients using various methods with $N = 10,000$ observations, for a single (randomly generated) instance. The associated error metrics are presented in Table 6.1. The confidence intervals in Figure 6.3 are computed according to (6.15) by letting e be standard basis vectors in $\mathbb{R}^\kappa$ and restricting the parameter space to $\mathcal{X}$. We also repeat the experiment 100 times (each time generate a new random true parameters), and the average errors are reported in Table 6.2. The experiments show that ML and LS estimates exhibit similar performance (ML outperforming slightly the LS estimates), and both of them outperform the recovery by EM algorithm based on the exponential kernel, which may be due to a more flexible parameterization of our model.

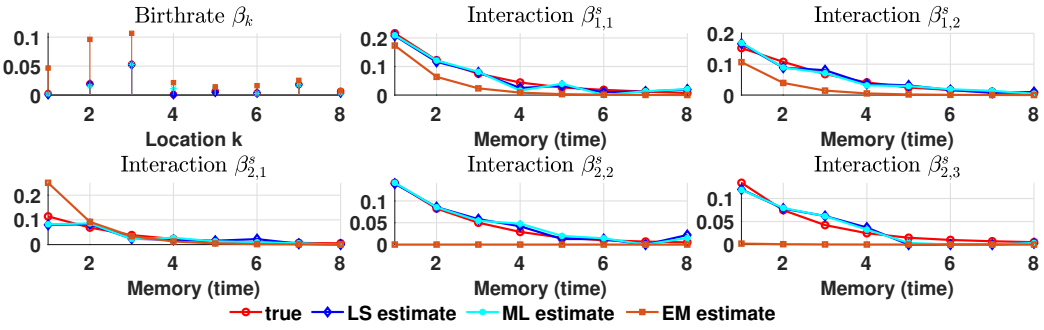


Figure 6.2: Single-state process: estimates for baseline intensity β_k and interactions parameters $\beta^s_{k\ell}$ for one random instance.

Multi-state spatio-temporal processes

Now consider a multi-state spatio-temporal Bernoulli process with the number of states $M = 2$. Here the possible states $p = 0$ represents no event, $p = 1, 2$ represent the event of category 1 and 2, respectively. We assume memory depth $d = 8$ and the number of

Table 6.1: Single-state process: error of ML, LS and EM estimation for the one instance shown in Figure 6.2.

Estimate	ℓ_1 error	ℓ_2 error	ℓ_∞ error
ML	1.7150 (22.57%)	0.1534 (17.67%)	0.0342 (13.64%)
LS	1.8849 (24.80%)	0.1714 (19.73%)	0.0372 (14.84%)
EM (exponential kernel)	6.3127 (83.06%)	0.6413 (73.83%)	0.2105 (83.97%)

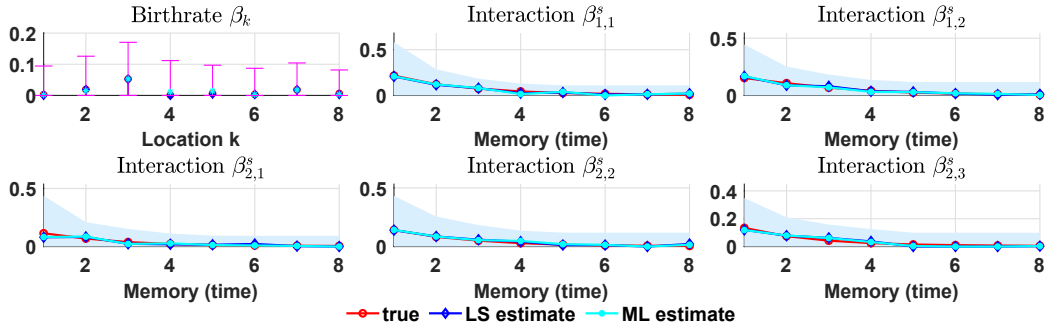


Figure 6.3: Computed 90% confidence intervals corresponding to Figure 6.2.

Table 6.2: Single-state process: error of ML, LS and EM estimation averaged over 100 trials.

Estimate	ℓ_1 error	ℓ_2 error	ℓ_∞ error
ML	1.1482 (15.11%)	0.1112 (12.60%)	0.0336 (11.87%)
LS	1.9776 (26.02%)	0.1831 (20.72%)	0.0472 (16.62%)
EM (exponential kernel)	6.4725 (85.16%)	0.6695 (75.72%)	0.2209 (75.17%)

locations $K = 10$. The true parameters are randomly generated from the set $\mathcal{X}_0$ specified by (again we impose additional constraints as in Section 6.6.1):

- $\beta_k(p) \geq 0$, $\beta^s_{kl}(p,q) \geq 0$; $\sum_{p=1}^M \beta_k(p) + \sum_{s=1}^d \sum_{\ell=1}^K \max_{0\leq q\leq M} \sum_{p=1}^M \beta^s_{k\ell}(p,q) \leq 1, \forall k \leq K$;
- $\beta^s_{k\ell}(p,q) = 0$ when $|k-\ell| > 1$, $\forall p, q$ (interactions are local);
- For every $1 \leq k, \ell \leq K$ and $1 \leq p \leq M, 0 \leq q \leq M$, $\beta^s_{k\ell}(p,q)$ is a non-increasing convex function of s.

Furthermore, we consider two scenarios, with additional constraints on the parameters

- Scenario 1: events can only trigger future events of the same category, i.e., $\beta^s_{k\ell}(p,q) \equiv$

$0, q \neq p$;

- Scenario 2: events of category $q = 0, \ldots, M$, only trigger events with category $p \leq q$. This can happen, for example, when modeling earthquakes aftershocks: events are marked using M categories according to their magnitudes: $u_1 < \ldots < u_M$. Set $u_0 = 0$ and treat the event "no earthquake" as "earthquake of magnitude 0." Then each earthquake can trigger "aftershocks" with the same or smaller magnitudes.

We generate a synthetic data sequence of length $N = 20,000$. For a single (randomly generated) instance, recovery of baseline and interaction parameters are presented in Figure 6.4. The associated recovery errors of the LS estimate (solution to (6.20)) and the ML estimate (solution to (6.30)) are reported in Table 6.3. In addition, we also report the recovery errors separately for (i) the baseline intensity vector (referred to as "birthrates") $\beta_{\text{birth}} = \{\beta_k(p), k \leq K, 1 \leq p \leq M\} \in \mathbb{R}^{KM \times 1}$; and (ii) the vector of interactions between different locations $\beta_{\text{inter}} = \{\beta^s_{k\ell}(p, q)\} \in \mathbb{R}^{dK^2M(M+1) \times 1}$. As shown in Table 6.3, the ℓ_1 recovery error for estimating birthrate is smaller than that for the interaction parameters. Thus, the recovery error for β is dominated by the error for interaction parameters. This could be explained because the magnitude of the baseline intensity is higher than the influence parameters (which is usually needed to have stationary processes).

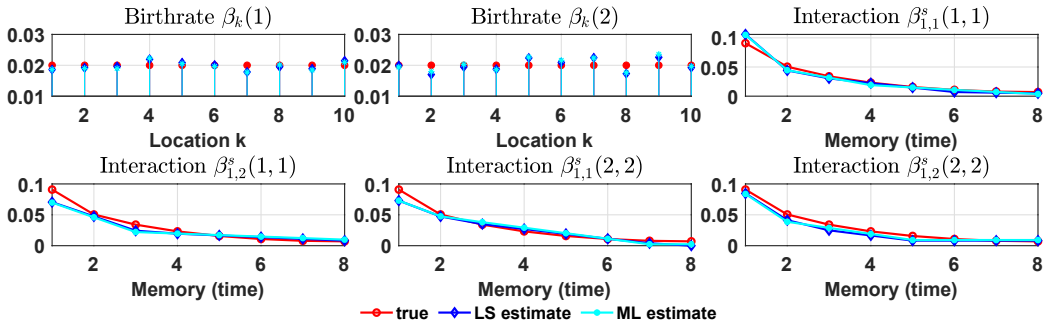


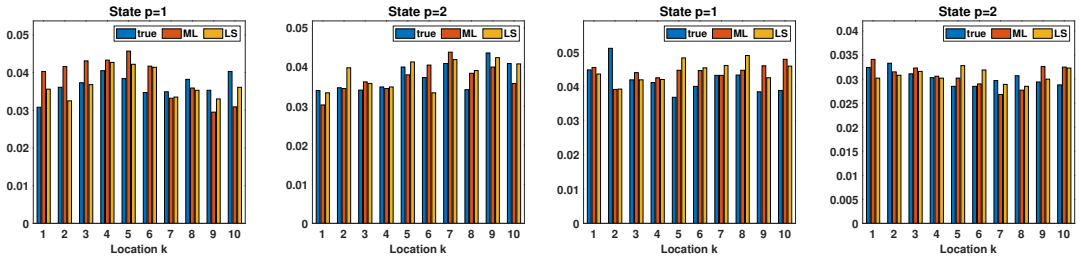
Figure 6.4: Multi-state process: examples of LS and ML estimates for baseline intensity $\beta_k(p)$ and interactions parameters $\beta^s_{k\ell}(p, q)$.

Finally, to assess the predictive capability of our model, we did the following experiment. Generate one sequence of discrete events, with length $N = 20,000$, using randomly

Table 6.3: Multi-state process recovery: norms of recovery error for LS estimate $\hat{\beta}_{LS}$ and ML estimate $\hat{\beta}_{\mathrm{ML}}$.

Estimate	Scenario 1		Scenario 2	
	ℓ_1 error	ℓ_2 error	ℓ_1 error	ℓ_2 error
$\hat{\beta}_{\mathrm{ML}}$	0.3524 (4.7%)	0.0532 (2.5%)	1.0179 (13.6%)	0.1146 (5.9%)
$\hat{\beta}_{\mathrm{LS}}$	0.4947 (6.6%)	0.0744 (3.4%)	1.0854 (14.5%)	0.1230 (6.3%)
$\hat{\beta}_{\mathrm{ML,\ birth}}$	0.0106 (2.7%)	0.0028 (3.1%)	0.0226 (5.7%)	0.0060 (6.7%)
$\hat{\beta}_{\mathrm{LS,\ birth}}$	0.0160 (4.0%)	0.0044 (5.0%)	0.0237 (5.9%)	0.0066 (7.4%)
$\hat{\beta}_{\mathrm{ML,\ inter}}$	0.3419 (4.8%)	0.0531 (2.5%)	0.9952 (14.0%)	0.1144 (5.9%)
$\hat{\beta}_{\mathrm{LS,\ inter}}$	0.4786 (6.7%)	0.0743 (3.4%)	1.0617 (15.0%)	0.1228 (6.3%)

selected parameters. We divide the sequence in half: use half for "training" and the other half for "testing". In particular, we (1) use the first half of the sequence for estimating the Bernoulli process model parameter, (2) use the "trained" model to generate a new "synthetic" sequence of length $N/2$, and (3) compare the "synthetic" sequence with the "test" sequence, in terms of the frequency of events, for each category, at each location. The results in Figure 6.5 show that the synthetic sequence has a reasonably good match with the testing sequence, based on the LS and the ML estimates.



Scenario 1 Scenario 2

Figure 6.5: Multi-state process: experiment to compare the frequency of events from a synthetic sequence (generated using models estimated from training sequence using LS and ML estimates) with that from the testing sequence.

Sparse network recovery with negative and non-monotone interactions

In the last synthetic example, we consider an example to recover a network with "non-conventional" interactions: non-monotonic temporal interactions and negative interactions.

Consider a sparse, directed, and non-planar graph (meaning that this cannot be embedded on a two-dimensional Euclidean space and, thus, this does not correspond to discretized space) with $K = 8$ nodes. The interaction functions are illustrated in Figure 6.6.

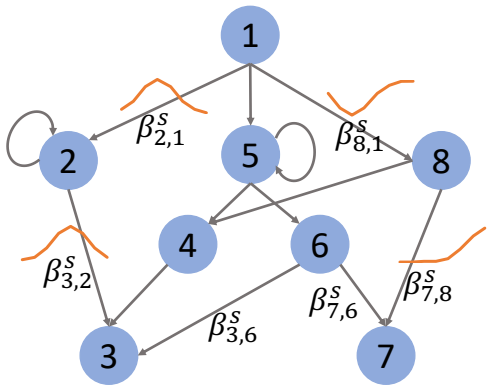


Figure 6.6: Sparse non-planar graph with non-monotonic and negative interaction. Note that the interaction $1 \to 8$ is negative.

The baseline intensities are all positive at all 8 nodes. The directed edge (arrows) means there is a one-directional "influence" from one node to its neighbor, e.g., $1 \to 5$. The self-edges, e.g., $2 \to 2$ and $5 \to 5$, denote that these nodes have a self-exciting effect: events happen at the node will trigger future events at itself. The true parameters of the model are generated as follows.

- Baseline parameters values at all locations are drawn independently from a uniform distribution on $[0, 0.2]$;
- For each *directed* edge $\ell \to k$, the interaction $\beta^s_{k,\ell}$ is given by $\beta^s_{k\ell} = 0.05e^{-0.25(s-\tau_{k\ell})^2}$, $s \geq 0$, and the peak $\tau_{k\ell}$ is randomly chosen from $\{1, \ldots, d\}$, except for one edge $1 \to 8$, whose interaction function is set to be negative: $\beta^s_{8,1} = -0.05e^{-0.25(s-\tau_{8,1})^2}$.

In our implementation, we consider two scenarios: (1) the graph structure is *unknown*: we do *not* impose sparsity constraints while obtaining the LS and ML estimates; (2) the graph structure is *known*, and then we impose the sparsity constraints by setting the interactions to be 0 when there is no edge; this illustrate the scenario when we have some prior information

about the network structure. We report recovery errors for the two scenarios in Table 6.4 and compare the recovery of interaction parameters under scenario (1) with the true values in Figure 6.7.

Table 6.4: Sparse network recovery with non-conventional interactions: errors of LS and ML estimates $\hat{\beta}_{\text{LS}}$, $\hat{\beta}_{\text{ML}}$.

Estimate	Unknown Graph			Known Graph		
	ℓ_1 error	ℓ_2 error	ℓ_∞ error	ℓ_1 error	ℓ_2 error	ℓ_∞ error
$\hat{\beta}_{\text{ML}}$	1.7694 (58.71%)	0.1128 (24.65%)	0.0224 (13.79%)	0.4715 (15.64%)	0.0593 (12.95%)	0.0173 (10.68%)
$\hat{\beta}_{\text{LS}}$	1.8757 (62.23%)	0.1166 (25.48%)	0.0211 (13.01%)	0.4773 (15.84%)	0.0606 (13.23%)	0.0204 (12.58%)
$\hat{\beta}_{\text{ML, birth}}$	0.0367 (3.84%)	0.0162 (4.42%)	0.0111 (6.84%)	0.0126 (1.32%)	0.0068 (1.85%)	0.0061 (3.75%)
$\hat{\beta}_{\text{LS, birth}}$	0.0378 (3.95%)	0.0172 (4.69%)	0.0129 (7.94%)	0.0126 (1.32%)	0.0069 (1.89%)	0.0061 (3.75%)
$\hat{\beta}_{\text{ML, inter}}$	1.7327 (84.20%)	0.1117 (40.69%)	0.0224 (44.73%)	0.4589 (22.30%)	0.0589 (21.46%)	0.0173 (34.65%)
$\hat{\beta}_{\text{LS, inter}}$	1.8379 (89.31%)	0.1153 (42.02%)	0.0211 (42.19%)	0.4648 (22.58%)	0.0602 (21.92%)	0.0204 (40.81%)

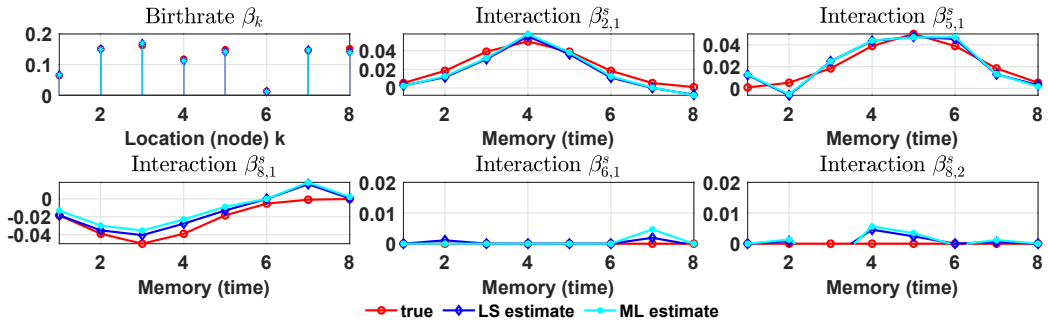


Figure 6.7: Sparse network identification when graph is unknown: examples of LS and ML estimates of baseline intensity and vectors of interaction parameters; interactions $\beta_{6,1}$ and $\beta_{8,2}$ correspond to edges $1 \to 6$ and $2 \to 8$ which do *not* exist in the graph in Figure 6.6.

From the experiment results, we observe that both the LS and ML estimates match closely with the true parameters, even when the underlying graph structure is unknown. The comparison in Table 6.4 shows a significant improvement in the estimation error when the graph structure is known *a priori*. This is consistent with our previous remark that knowing the network structure allows for a better choice of the feasible region resulting in reduced estimation error. Moreover, by examining the histogram of the maximum interaction between each pair, i.e., $\{\max_{s=1}^{d} |\beta^s_{k,\ell}|, 1 \leq k, \ell \leq K\}$ as shown in Figure 6.8, we observe that we can indeed accurately recover the support of the graph: the estimates of the edges with non-zero interactions, are completely separable from the estimates of the edges with zero interactions. This indicates that we can apply an appropriate threshold (in this

case, e.g., 0.03) to recover precisely the unknown graph structure completely. This example also shows that even when prior information about the spare structure of the underlying network is not available, LS and ML estimates can recover the underlying network reasonably well, which opens possibilities of applying the proposed approach to perform casual inference [193] using discrete-event data.

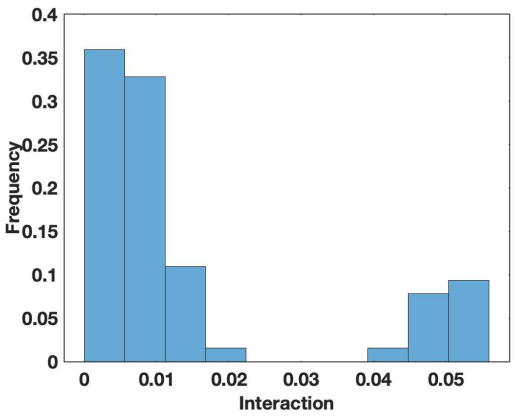


Figure 6.8: Sparse network support recovery: histogram of the recovered interaction parameters $\{\max_{s=1}^{d} |\beta_{k,\ell}^{s}|, 1 \leq k, \ell \leq K\}$. Edges with non-zero interactions can be perfectly separated from those with zero interactions.

6.6.2 Real Data Studies: Crime in Atlanta

Finally, we study a real crime dataset in Atlanta to demonstrate the promise of our methods to recover interesting structures from real-data. We consider two categories of crime incidents, "burglary" and "robbery". These incidents were reported to the Atlanta Police Department from January 1, 2015, to September 19, 2017. The dataset contains 47,245 "burglary" and 3,739 "robbery" incidents. As mentioned in the introduction, it is believed that crime incidents are related and have "self-exciting" patterns: once crime incidence happens, it triggers similar crimes more likely to happen in the neighborhood in the near future [177]. Here, we model the data using a multi-state Bernoulli process with two states: no event ($p = 0$), burglary ($p = 1$), and robbery ($p = 2$).

We extract crime events around the Atlanta downtown area, as shown in Figure 6.9,

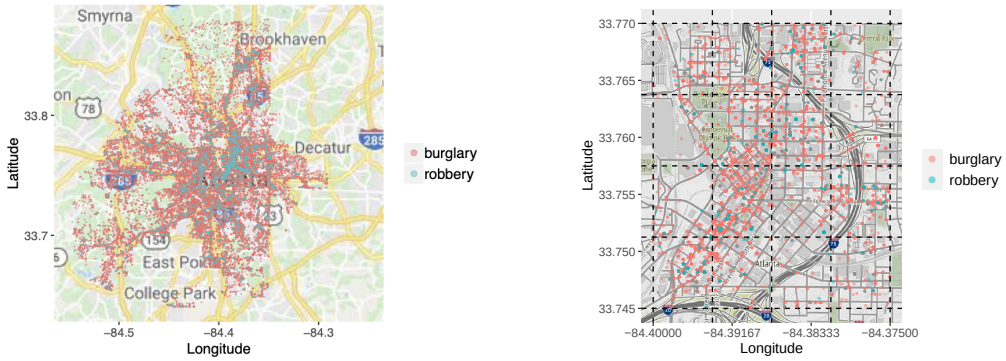


Figure 6.9: Raw data map: burglary and robbery incidents in Atlanta. Left: the full map; Right: zoom-in around downtown Atlanta.

which contains 6031 "burglary" events and 454 "robbery" events. The whole time horizon (788 days) is split into discrete time intervals with a duration of four hours. The memory depth d is set to 6 in this example (each sample is 1 hour). This is a "reasonable" value retrieved from observations by verifying the predictions of frequencies of the burglary and robbery incidents in various spatial cells. The downtown region is divided uniformly into 16 sub-regions.

We compute the LS estimates of the parameters $\{\beta_k(p), \beta^s_{k,l}(p,q)\}$, in two different ways to set up the constraints: in the first setup, we do not impose additional constraints on the parameters apart from "basic" constraints (6.18); in the second setup, we impose constraints to only consider temporal interaction function, $\beta^s_{k\ell}$, with monotonic and convex "shapes". Such constraints are routinely imposed when estimating parameters of Hawkes model, see, e.g., [162]. The estimated parameters are shown in Figure 6.10. In the figure, the size of the red dot in each region is proportional to the magnitude of the estimated birthrate $\beta_k(p), k = 1, \ldots, K$, for Burglary/Robbery, respectively; the width of the arrow is proportional to the magnitude of the interaction $\beta^s_{k,l}(p,q)$ between locations. It is interesting to notice that our model recovers the dynamic of the interactions and how they change over time. There also seem to be strong interactions between burglary and robbery at different locations.

To validate the model, we take the two-year duration of data, divide the sequence in

half, use the first half of the sequence to estimate a multi-state Bernoulli process model, generate a synthetic sequence, and compare with the second half of the sequence reserved for testing. We compare the frequency of Burglary and Robbery events across all locations, for the synthetic and testing sequence. The results are shown in Table 6.5. The results look to be a reasonably good match, considering that the crime events are relatively rare and with highly complex (and unknown) dynamics: predicting their frequency in the first place is a highly challenging task and an essential research task of criminology.

We also note that the prediction for burglary seems to be better since the frequencies from the synthetic sequence are very close, and the relative error is smaller. This is expected since the number of burglary cases is much larger than the number of robbery cases in our dataset, and the frequency of robbery cases is very small (typically below 0.01, as shown in Table 6.5). The experiment serves as a sanity check and shows that for challenging and noisy real-world datasets, there could be a certain truth to the proposed methods.

Table 6.5: Crime event model recovery: frequency of Burglary and Robbery events at each location.

Locations	Burglary			Robbery		
	True	With constr	Without constr	True	With constr	Without constr
1	0.1499	**0.1707**	0.1766	0.0102	0.0195	**0.0186**
2	0.0284	**0.0373**	0.0445	0.0017	**0.0203**	0.0212
3	0.0483	**0.0580**	0.0606	0.0021	0.0254	**0.0195**
4	0.0407	**0.0364**	0.0356	0.0017	**0.0178**	0.0224
5	0.0508	**0.0529**	0.0648	0.0042	0.0220	**0.0165**
6	0.1957	0.2088	**0.1834**	0.0131	0.0208	**0.0144**
7	0.0970	0.1368	**0.1224**	0.0068	0.0229	**0.0191**
8	0.0419	0.0580	**0.0563**	0.0021	**0.0127**	0.0182
9	0.0148	**0.0161**	0.0220	0.0013	**0.0165**	0.0212
10	0.0584	**0.0729**	0.0805	0.0055	0.0258	**0.0178**
11	0.1266	**0.1525**	0.1529	0.0106	0.0195	**0.0169**
12	0.1364	**0.1266**	0.1186	0.0102	0.0191	**0.0169**
13	0.0322	0.0521	**0.0445**	0.0021	0.0229	**0.0224**
14	0.0627	0.0868	**0.0834**	0.0055	0.0212	**0.0195**
15	0.0208	**0.0224**	0.0280	0.0008	0.0241	**0.0216**
16	0.0144	0.0203	**0.0178**	0.0013	**0.0203**	0.0212

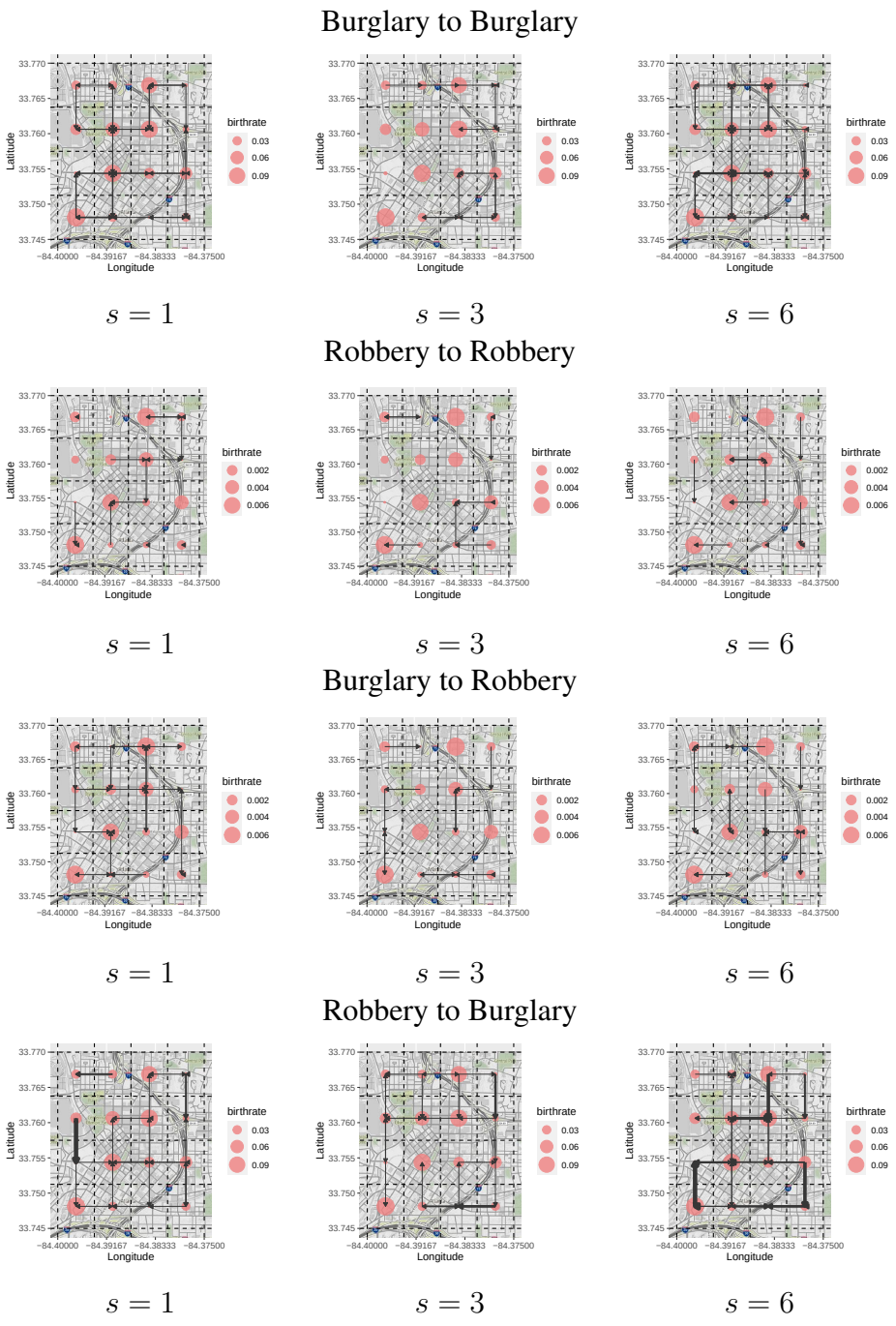


Figure 6.10: Robbery and burglary in downtown Atlanta: recovered spatio-temporal interactions, using LS estimates without additional constraint on the shapes of the interaction functions.

Appendices

APPENDIX A

PROOFS FOR CHAPTER 3

Proof of Theorem 3.2. In order to prove Theorem 3.2, we need the following lemma to characterize the local correlation between largest eigenvalue statistics.

Lemma A.1 (Approximation of local correlation)**.** *Let* $c_1 = \mathbb{E}[W_1] = -1.21$*,* $c_2 = \sqrt{\mathrm{Var}(W_1)} = 1.27$*, and*

$$\beta_{k,w} = 1 + \frac{\left(1 + c_1 k^{-\frac{1}{6}}/\sqrt{w}\right)\left(2 + c_1 k^{-\frac{1}{6}}/\sqrt{w}\right)}{c_2^2 k^{-\frac{1}{3}}/w}.$$

Then we have,

$$\mathrm{Corr}(Z_t, Z_{t+\delta}) \leq 1 - \beta_{k,w}\vartheta + o(\vartheta),$$

where $\vartheta = \delta/w$ *and* $\mathrm{corr}(X, Y)$ *stands for the Pearson's correlation*

$$\mathrm{Corr}(X, Y) = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

Proof to Lemma A.1. Under the pre-change measure, $x_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_k)$. For $\delta \in \mathbb{Z}^+$, let

$$P = \sum_{i=t-w+1}^{t-w+\delta} x_i x_i^\top, \quad Q = \sum_{i=t-w+\delta+1}^{t} x_i x_i^\top, \quad R = \sum_{i=t+1}^{t+\delta} x_i x_i^\top.$$

Then P, Q and R are independent random matrices. Now we also want to give a general

upper bound for the covariance between Z_t and $Z_{t+\delta}$. Then we have

$$\begin{aligned}\mathbb{E}[Z_t Z_{t+\delta}] &= \mathbb{E}[\lambda_{\max}(\hat{\Sigma}_{t,w})\lambda_{\max}(\hat{\Sigma}_{t+\delta,w})] = \mathbb{E}[\lambda_{\max}(P+Q)\lambda_{\max}(Q+R)] \\ &\leq \mathbb{E}\left\{[\lambda_{\max}(P)+\lambda_{\max}(Q)]\,[\lambda_{\max}(Q)+\lambda_{\max}(R)]\right\} \\ &= \mathbb{E}[\lambda^2_{\max}(Q)] + \mathbb{E}[\lambda_{\max}(Q)]\mathbb{E}[\lambda_{\max}(R)] \\ &\quad + \mathbb{E}[\lambda_{\max}(P)]\{\mathbb{E}[\lambda_{\max}(Q)] + \mathbb{E}[\lambda_{\max}(R)]\},\end{aligned}$$

where the inequality is due to the fact that the largest eigenvalue of the sum of two nonnegative definite matrices is upper bounded by the sum of the corresponding largest eigenvalues of the two matrices. The mean and second-order moments can be computed using the Tracy-Widom law depicted in (3.13).

Since k is a fixed constant, we just write μ_n and σ_n instead of $\mu_{n,k}$ and $\sigma_{n,k}$ to simplify our notation. We first consider the covariance term $\mathbb{E}[Z_t Z_{t+\delta}]$ and decompose it into four parts as following:

$$\frac{1}{w^2}\mathbb{E}[Z_t Z_{t+\delta}] \leq A + B + C + D,$$

where

$$\begin{aligned}A &= \left(\frac{\mu_{w(1-\vartheta)} + c_1\sigma_{w(1-\vartheta)}}{w}\right)^2, \\ B &= \left(\frac{c_2\sigma_{w(1-\vartheta)}}{w}\right)^2, \\ C &= 2\left[\frac{\mu_{w(1-\vartheta)} + c_1\sigma_{w(1-\vartheta)}}{w}\right]\left(\frac{\mu_{w\vartheta} + c_1\sigma_{w\vartheta}}{w}\right), \\ D &= \left(\frac{\mu_{w\vartheta} + c_1\sigma_{w\vartheta}}{w}\right)^2.\end{aligned}$$

First, the common terms $\mu_{w(1-\vartheta)}/w$ and $\sigma_{w(1-\vartheta)}/w$ can be written as

$$\begin{aligned}\frac{\mu_{w(1-\vartheta)}}{w} &= \frac{1}{w}\left[\sqrt{w(1-\vartheta)-1} + \sqrt{k}\right]^2 = \frac{w(1-\vartheta)-1}{w}\left[1 + \sqrt{\frac{k}{w(1-\vartheta)-1}}\right]^2 \\ &\doteq \frac{w(1-\vartheta)-1}{w} \doteq 1-\vartheta.\end{aligned}$$

where the second term in the bracket was ignored because $k/w = o(1)$. Moreover, we have

$$\frac{\sigma_{w(1-\vartheta)}}{w} = \frac{1}{w}\left(\sqrt{w(1-\vartheta)-1}+\sqrt{k}\right)\cdot\left(\frac{1}{\sqrt{w(1-\vartheta)-1}}+\frac{1}{\sqrt{k}}\right)^{1/3},$$

after extracting the term $\sqrt{w(1-\vartheta)-1}$ from the first bracket and $1/\sqrt{k}$ from the second bracket, we obtain

$$\frac{\sigma_{w(1-\vartheta)}}{w} = \frac{k^{-\frac{1}{6}}}{w}\sqrt{w(1-\vartheta)-1}\left(1+\sqrt{\frac{k}{w(1-\vartheta)-1}}\right)^{\frac{4}{3}} \doteq \sqrt{\frac{1-\vartheta}{w}}k^{-\frac{1}{6}},$$

where the second term in the bracket was ignored because $k/w = o(1)$. Plug these two terms into the first part we have:

$$\begin{aligned} A &= \left(\frac{\mu_{w(1-\vartheta)}+c_1\sigma_{w(1-\vartheta)}}{w}\right)^2 \doteq \left(1-\vartheta+c_1\sqrt{\frac{1-\vartheta}{w}}k^{-\frac{1}{6}}\right)^2 \\ &= (1-\vartheta)\left(1-\vartheta+2c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\sqrt{1-\vartheta}+c_1^2\frac{k^{-\frac{1}{3}}}{w}\right). \end{aligned}$$

Since ϑ is relatively small, $\sqrt{1-\vartheta} = 1-\frac{1}{2}\vartheta+o(\vartheta)$ by Taylor expansion. Then the term is approximately

$$\begin{aligned} A &\doteq (1-\vartheta)\left(1-\vartheta+2c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}(1-\frac{1}{2}\vartheta+o(\vartheta))+c_1^2\frac{k^{-\frac{1}{3}}}{w}\right) \\ &= \left(1+c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\right)^2-\left(1+c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\right)\left(2+c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\right)\vartheta+o(\vartheta), \end{aligned}$$

where the higher order terms of ϑ are included in $o(\vartheta)$. Parts C and D can be considered

negligible, since C is order $\mathcal{O}(\vartheta)$ and D is order $o(\vartheta)$. In summary, we have

$$\begin{aligned}
&\mathrm{Corr}(Z_t, Z_{t+\delta}) \\
=&\frac{\mathbb{E}[Z_t Z_{t+\delta}] - \mathbb{E}[Z_t]\mathbb{E}[Z_{t+\delta}]}{\sqrt{\mathrm{Var}(Z_t)}\sqrt{\mathrm{Var}(Z_{t+\delta})}} \\
\lesssim&\frac{1}{\left(\frac{c_2 k^{-\frac{1}{6}}}{\sqrt{w}}\right)^2}\left\{\left(1 + c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\right)^2 + c_2^2\frac{1-\vartheta}{w}k^{-\frac{1}{3}} - \left(1 + c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\right)\left(2 + c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\right)\vartheta\right. \\
&\qquad \left. - \left(1 + c_1\frac{k^{-\frac{1}{6}}}{\sqrt{w}}\right)^2 + o(\vartheta)\right\} \\
=&1 - \beta_{k,w}\vartheta + o(\vartheta).
\end{aligned}$$

This completes the proof of Lemma A.1. □

The key of proving Theorem 3.2 is to quantify the tail probability of the detection statistic. However, this probability is very small when the threshold is large [120]. Therefore we use the change-of-measure technique in [179] to recenter the process mean to the threshold, so that the tail probability becomes much higher. First, the detection statistic is standardized by:

$$Z_t' = \frac{Z_t - \mathbb{E}_\infty[Z_t]}{\mathrm{Var}_\infty(Z_t)},$$

here $\mathbb{E}_\infty[Z_t]$ and $\mathrm{Var}_\infty(Z_t)$ depends only on k and w, but does not depend on t. Then Z_t' has zero mean and unit variance under the $\mathbb{P}_\infty$ measure. We are interested in finding the probability $\mathbb{P}_\infty(T_{\mathrm{E}} \leq M) = \mathbb{P}_\infty(\max_{1\leqslant t\leqslant M} Z_t' > b)$. We now prove Theorem 3.2 in four steps.

Step 1. Exponential tilting. Denote the cumulant generating function of Z_t' by

$$\psi(a) = \log \mathbb{E}_\infty[e^{aZ_t'}].$$

Define a family of new measures

$$\frac{d\mathbb{P}_t}{d\mathbb{P}_\infty} = \exp\{aZ'_t - \psi(a)\},$$

where $\mathbb{P}_t$ denotes the new measure after the transformation. The new measure takes the form of the exponential family, and a can be viewed as the natural parameter. It can be verified that $\mathbb{P}_t$ is indeed a probability measure since

$$\int d\mathbb{P}_t = \int \exp\{aZ'_t - \psi(a)\} d\mathbb{P} = 1.$$

It can also be shown that $\dot{\psi}(a)$ is the expected value of Z'_t under $\mathbb{P}_t$, since

$$\dot{\psi}(a) = \frac{\mathbb{E}_\infty[Z'_t e^{aZ'_t}]}{\mathbb{E}_\infty[e^{aZ'_t}]} = \mathbb{E}_\infty[Z'_t e^{aZ'_t - \psi(a)}] = \mathbb{E}_t[Z'_t],$$

and similarly $\ddot{\psi}(a)$ is the variance under the tilted measure. We use the Gaussian approximation for Z'_t, then its log moment generating function is $\psi(a) = a^2/2$. We set $a = b$ such that $\dot{\psi}(a) = \mathbb{E}_t[Z'_t] = b$, therefore the tail probability after measure transformation will become much larger. Given this choice, the transformed measure is given by $d\mathbb{P}_t = \exp(bZ'_t - b^2/2) d\mathbb{P}_\infty$. We also define, for each t, the log-likelihood ratio $\log(d\mathbb{P}_t/d\mathbb{P}_\infty)$ of the form

$$\ell_t = bZ'_t - \frac{1}{2}b^2.$$

Step 2. Change-of-measure by the likelihood ratio identity. Now we convert the original problem of finding the small probability that the maximum of a random field exceeds a large threshold, to another problem: finding an alternative measure under which the event

happens with a much higher probability. By likelihood ratio identity, we have:

$$\begin{aligned}
\mathbb{P}_\infty(\max_{1\leqslant m\leqslant M} Z'_m \geq b) &= \mathbb{E}_\infty[\mathbb{1}_{\{\max_{1\leqslant m\leqslant M} Z'_m\geq b\}}] = \mathbb{E}_\infty\left[\frac{\sum_{t=1}^M e^{\ell_t}}{\sum_{n=1}^M e^{\ell_n}}\cdot \mathbb{1}_{\{\max_{1\leqslant m\leqslant M} Z'_m\geq b\}}\right] \\
&= \sum_{t=1}^M \mathbb{E}_\infty\left[\frac{e^{\ell_t}}{\sum_n e^{\ell_n}}\cdot \mathbb{1}_{\{\max_{1\leqslant m\leqslant M} Z'_m\geq b\}}\right] \\
&= \sum_{t=1}^M \mathbb{E}_t\left[\frac{1}{\sum_n e^{\ell_n}}\cdot \mathbb{1}_{\{\max_{1\leqslant m\leqslant M} Z'_m\geq b\}}\right] \\
&= e^{-b^2/2}\sum_{t=1}^M \mathbb{E}_t\left[\frac{M_t}{S_t}e^{-(\tilde{\ell}_t+\log M_t)}\cdot \mathbb{1}_{\{\tilde{\ell}_t+\log M_t\geq 0\}}\right],
\end{aligned} \tag{A.1}$$

where M_t and S_t in the last step is defined as the maximum and sum of likelihood ratio differences as:

$$M_t = \max_{m\in\{1,\ldots,M\}} e^{\ell_m-\ell_t}, \quad S_t = \sum_{m\in\{1,\ldots,M\}} e^{\ell_m-\ell_t}.$$

And $\tilde{\ell}_t$ is defined as the re-centered likelihood ratio, or the so-called global term:

$$\tilde{\ell}_t = b(Z'_t - b).$$

The last equation in (A.1) converts the tail probability to a product of two terms: a deterministic term $e^{-b^2/2}$ associated with the large deviation rate, and a sum of expectations under the transformed measures. The expectation involves a product of the ratio M_t/S_t and an exponential function that depends on $\tilde{\ell}_t$, which plays the role of a weight. Under the new measure $\mathbb{P}_t$, $\tilde{\ell}_t$ has zero mean and variance equal to b^2 and it dominates the other term $\log M_t$, hence, the probability of exceeding zero is much higher. Next, we characterize the limiting ratio and the other factors precisely, by the localization theorem.

Step 3. Establish properties of local and global terms. In (A.1), our target probability has been decomposed into terms that only depend on (i) the local field $\{\ell_m - \ell_t\}, 1 \leqslant m \leqslant M$, which are the differences between the log-likelihood ratios with parameter t and m, and (ii) the global term $\tilde{\ell}_t$, which is the centered and scaled likelihood ratio with parameter t.

We need to first establish some useful properties of the local field and global term before applying the localization theorem. We will eventually show that the local field and the global term are asymptotically independent.

For the local field $\{\ell_m - \ell_t\}$, let $r_{m,t}$ denote the correlation between Z'_m and Z'_t, then we have

$$\mathbb{E}_t(\ell_m - \ell_t) = -b^2(1 - r_{m,t}),$$

$$\mathrm{Var}_t(\ell_m - \ell_t) = 2b^2(1 - r_{m,t}),$$

$$\mathrm{Cov}_t(\ell_{m_1} - \ell_t, \ell_{m_2} - \ell_t) = b^2(1 + r_{m_1,m_2} - r_{m_1,t} - r_{m_2,t}).$$

We have Lemma A.1 to characterize the local correlation, which offers reasonably good approximation for $\mathbb{E}[Z_t Z_{t+\delta}]$ and leads to $r_{m,t} \approx 1 - |m - t|\, \beta_{k,w}/w$.

Since we assume Z'_t is approximately Gaussian, the local field $\ell_m - \ell_t$ and the global term $\tilde{\ell}_t$ are also approximately Gaussian. Therefore, when $|\delta|$ is small (i.e., in the neighborhood of zero), we can approximate the local field using a two-sided Gaussian random walk with drift $b^2\beta_{k,w}/w$ and variance of the increment equal to $2b^2\beta_{k,w}/w$:

$$\ell_{t+\delta} - \ell_t \stackrel{\Delta}{=} b\sqrt{\frac{2\beta_{k,w}}{w}} \sum_{i=1}^{|\delta|} \xi_i - b^2 \frac{\beta_{k,w}}{w}\delta, \delta = \pm 1, \pm 2, \ldots,$$

where ξ_i are i.i.d. standard normal random variables.

Step 4. Approximation using localization theorem. From the argument in [182], let $\hat{M}_t$ and $\hat{S}_t$ denote the maximization and summation restricted to the small neighborhood of t. Then they are asymptotically independent of the global term $\tilde{\ell}_t$. Moreover, under the tilted measure,

$$\mathbb{E}_t[\tilde{\ell}_t] = 0, \quad \mathrm{Var}_t[\tilde{\ell}_t] = b^2.$$

Therefore the density $\mathbb{P}_t(\tilde{\ell}_t)$ can be approximated by $1/\sqrt{2\pi b^2}$ in a neighborhood of radius $o(1/b)$ of zeros. The inner expectation in (A.1) can be approximated as

$$\mathbb{E}_t\left[\frac{M_t}{S_t} e^{-(\tilde{\ell}_t + \log M_t)} \cdot \mathbb{1}_{\{\tilde{\ell}_t + \log M_t \geq 0\}}\right] \doteq \frac{\mathbb{E}_t(\hat{M}_t/\hat{S}_t)}{b\sqrt{2\pi b^2}}.$$

By [182], the expectation $\mathbb{E}_t(\hat{M}_t/\hat{S}_t)$ equals to $b^2\beta_{k,w}\nu(b\sqrt{2\beta_{k,w}/w})/w$ in the asymptotic regime, not depending on t. Substitute into (A.1), we have

$$\begin{aligned}\mathbb{P}_\infty(T \leqslant M) &= \mathbb{P}_\infty\left(\max_{1\leqslant t\leqslant M} Z'_t > b\right)\\ &= e^{-b^2/2}\sum_{t=1}^{M}\mathbb{E}_t\left[\frac{M_t}{S_t}e^{-[\tilde{\ell}_t+\log M_t]}\cdot \mathbb{1}_{\{\tilde{\ell}_t+\log M_t\geqslant 0\}}\right]\\ &\doteq Mb\phi(b)\beta_{k,w}\nu(b\sqrt{2\beta_{k,w}/w})/w,\end{aligned}$$

where $\nu(\cdot)$ is the function defined in (3.15). From the above cumulative distribution function, we can approximate T as exponential distribution, yielding the mean value

$$1/[b\phi(b)\beta_{k,w}\nu(b\sqrt{2\beta_{k,w}/w})/w].$$

Since Z'_t is standardized, here the threshold b need to be converted to the original threshold using a simple formula

$$b' = \left[b - (\mu_{w,k} + c_1\sigma_{w,k})\right]/(c_2\sigma_{w,k}).$$

This completes the proof. □

Proof of Theorem 3.3. We first relate the largest eigenvalue procedure to a CUSUM procedure, note that

$$\lambda_{\max}(\hat{\Sigma}_{t,w}) = \max_{\|q\|=1} q^\top\hat{\Sigma}_{t,w}q. \tag{A.2}$$

For each q, we have

$$q^\top\hat{\Sigma}_{t,w}q = \sum_{i=t-w+1}^{t}(q^\top x_i)^2.$$

According to the Grothendieck's Inequality [76], the q that attains the maximum in equation (A.2) is very close to u under the alternative. Therefore, assuming the optimal q always equals to u will only cause a small error but will bring great convenience to our analysis.

Now we have under $\mathbb{P}_\infty$, $q^\top x_i \sim \mathcal{N}(0, \sigma^2)$ and under $\mathbb{P}_0$, $q^\top x_i \sim \mathcal{N}(0, \sigma^2 + \theta)$. Let f_∞ denote the pdf of $\mathcal{N}(0, \sigma^2)$ and f_0 the pdf of $\mathcal{N}(0, \sigma^2 + \theta)$. For each observation y, we can derive the one-sample log-likelihood ratio:

$$\log \frac{f_0(y)}{f_\infty(y)} = -\frac{1}{2}\log(1+\rho) + \frac{1}{2\sigma^2}\left(1 - \frac{1}{1+\rho}\right) y^2.$$

Define the CUSUM procedure

$$\widetilde{T} = \inf\left\{t : \max_{0\le k<t} \sum_{i=k+1}^{t} \left[\frac{1}{2\sigma^2}\left(1 - \frac{1}{1+\rho}\right)(q^\top x_i)^2 - \frac{\log(1+\rho)}{2}\right] \ge b'\right\},$$

where $b' = \frac{1}{2\sigma^2}(1 - \frac{1}{1+\rho})(b - \frac{\sigma^2 \log(1+\rho)}{1-1/(1+\rho)})w$, we then have

$$\mathbb{E}_0[T_{\mathrm{E}}] \ge \mathbb{E}_0[\widetilde{T}].$$

Since $\widetilde{T}$ is a CUSUM procedure with

$$\int \log\left[\frac{f_0(y)}{f_\infty(y)}\right] f_0(y)dy = -\frac{1}{2}\log(1+\rho) + \frac{\rho}{2},$$

by [180], we have:

$$\mathbb{E}_0[\widetilde{T}] = \frac{e^{-b'} + b' - 1}{-\log(1+\rho)/2 + \rho/2}.$$

This completes the proof. □

Proof of Lemma 3.1. Under the pre-change distribution we can write

$$\mathbb{E}_\infty[(\hat{u}_t^\top x_t)^2] = \mathbb{E}_\infty[\hat{u}_t^\top \mathbb{E}_\infty[x_t x_t^\top]\hat{u}_t] = \sigma^2 \mathbb{E}_\infty[\hat{u}_t^\top \hat{u}_t] = \sigma^2,$$

where the first equation is due to the independence of x_t and $\hat{u}_t$, the next one due to x_t having covariance $\sigma^2 I_k$ and the last equality due to $\hat{u}_t$ being of unit-norm.

Under the alternative regime we are going to use Central Limit Theorem (CLT) argu-

ments [7, 142] that describe the statistical behavior of the estimator $\hat{u}_t$. We will assume that the window size w is sufficiently large so that CLT approximations are possible for $\hat{u}_t$. The required result appears in the next lemma.

Lemma A.2. *Suppose vectors $x_1, \ldots, x_w$ are of dimension k and follow the distribution $\mathcal{N}(0, \sigma^2 I_k + \theta u u^\top)$. Let $\hat{\varphi}_w$ be the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix $(1/w)(x_1 x_1^\top + \cdots + x_w x_w^\top)$, then, as $w \to \infty$, we have the following CLT version for $\hat{\varphi}_w$*

$$\sqrt{w}(\hat{\varphi}_w - u) \to \mathcal{N}\left(0, \frac{1+\rho}{\rho^2}(I_k - uu^\top)\right).$$

Proof of Lemma A.2. We have the following asymptotic distribution [7]:

$$\frac{1}{\sqrt{w}}(\hat{\varphi}_w - u) \xrightarrow{d} \mathcal{N}\Big(0, \sum_{j=2}^{k} \frac{\lambda_1 \lambda_j}{(\lambda_1 - \lambda_j)^2} \nu_j \nu_j^\top\Big),$$

where λ_j are the jth largest eigenvalue of the true covariance matrix and ν_j are the corresponding eigenvector. In our case the true covariance matrix is $\sigma^2 I_k + \theta u u^\top$, therefore $\lambda_1 = \sigma^2 + \theta$ and $\lambda_j = \sigma^2$ for $j \geq 2$, and $\{\nu_j, j \geq 2\}$ is a basis of the orthogonal space of u. Thus we have

$$\sum_{j=2}^{k} \frac{\lambda_1 \lambda_j}{(\lambda_1 - \lambda_j)^2} \nu_j \nu_j^\top = \frac{\sigma^2(\sigma^2 + \theta)}{\theta^2} \sum_{j=2}^{k} \nu_j \nu_j^\top = \frac{(1+\rho)}{\rho^2}(I_k - uu^\top).$$

□

Lemma A.2 provides an asymptotic statistical description of the *un-normalized* estimate of u. More precisely it characterizes the estimation error $v_w = \hat{\varphi}_w - u$. In our case we estimate the eigenvector from the matrix $\hat{\Sigma}_{t+w,w}$ but, as mentioned before, we adopt a *normalized* (unit norm) version $\hat{u}_t$. Therefore if we fix w at a sufficiently large value and v_t denotes the estimation error of the un-normalized estimate at time t then, from Lemma A.2,

we can deduce

$$\hat{u}_t = \frac{\hat{\varphi}_t}{\|\hat{\varphi}_t\|} = \frac{u + v_t}{\|u + v_t\|}, \quad v_t \sim \mathcal{N}\left(0, \frac{1+\rho}{w\rho^2}(I_k - uu^\top)\right).$$

Consequently

$$\begin{aligned}
\mathbb{E}_0\left[(\hat{u}_t^\top x_t)^2\right] &= \sigma^2 \mathbb{E}_0\left[\hat{u}_t^\top (I_k + \rho uu^\top)\hat{u}_t\right] \\
&= \sigma^2(1 + \rho\mathbb{E}_0[(\hat{u}_t^\top u)^2]) = \sigma^2\left(1 + \rho\mathbb{E}_0\left[\frac{1}{1+\|v_t\|^2}\right]\right) \\
&\approx \sigma^2\left(1 + \rho\mathbb{E}_0\left[1 - \|v_t\|^2\right]\right) = \sigma^2(1+\rho)\left(1 - \frac{k-1}{w\rho}\right),
\end{aligned}$$

with the $o(\cdot)$ term being negligible compared to the other two when $k/w \ll 1$, where $a = o(b)$ denotes that $a/b \to 0$. For the approximate equality we used the fact that to a first order approximation we can write $1/(1+\|v_t\|^2) \approx 1 - \|v_t\|^2$ because $\|v_t\|^2$ is of the order of $1/w$ while the approximation error is of higher order. This completes the proof. □

Proof of Proposition 3.1. We first evaluate the expectation in (3.23) to demonstrate the relationship between δ_∞ and d depicted in (3.25). Using standard computations involving Gaussian random vectors we can write

$$\begin{aligned}
\mathbb{E}_\infty[e^{\delta_\infty[(\hat{u}_t^\top x_t)^2 - d]}] &= e^{-\delta_\infty d}\mathbb{E}_\infty\left[\mathbb{E}_\infty[e^{\delta_\infty(\hat{u}_t^\top x_t)^2}|\hat{u}_t]\right] \\
&= e^{-\delta_\infty d}\mathbb{E}_\infty\left[\int e^{\delta_\infty x_t^\top(\hat{u}_t\hat{u}_t^\top)x_t} \cdot \frac{e^{-x_t^\top x_t/(2\sigma^2)}}{\sqrt{(2\pi)^k\sigma^{2k}}}dx_t\right] \\
&= \frac{e^{-\delta_\infty d}}{\sqrt{1 - 2\sigma^2\delta_\infty}},
\end{aligned}$$

which is equivalent to (3.25). To compute the integral we used the standard technique of "completing the square" in the exponent and with proper normalization generate an alternative Gaussian pdf which integrates to 1. The interesting observation is that the result of the integration does not actually depend on $\hat{u}_t$.

If we use the value for d in terms of δ_∞ then as we argued in the text we obtain for

EDD the expression appearing in (3.26). We can now fix w and optimize EDD with respect to δ_∞. This is a straightforward process since it amounts in maximizing the denominator. Taking the derivative and equating to 0 yields the optimum δ_∞:

$$\delta_\infty^* = \frac{1}{2\sigma^2}\left(1 - \frac{1}{(1+\rho)\left(1 - \frac{k-1}{w\rho}\right)}\right).$$

Substituting this value in (3.26) produces (3.27).

The next step consists in minimizing (3.27) with respect to w. Again taking the derivative and equating to 0 we can show that the optimum window size is the w^* depicted in Proposition 3.1. □

APPENDIX B

PROOFS FOR CHAPTER 4

Proof of Proposition 4.1. Observe that the expectation of the empirical distribution of N-element sample drawn from a distribution $r \in \Delta_n$ on Ω is r, and the covariance matrix is

$$C_{r,N} = \frac{1}{N}[\mathrm{Diag}\{r\} - rr^\top] \preceq \frac{1}{N}\mathrm{Diag}\{r\}.$$

Representing $\omega = p + \alpha, \omega' = p + \alpha', \zeta = q + \beta, \zeta' = q + \beta'$, we have

$$\chi = (p-q)^\top \Sigma (p-q) + \underbrace{(\alpha-\beta)^\top \Sigma (p-q)}_{B} + \underbrace{(\alpha'-\beta')^\top \Sigma (p-q)}_{B'} + \underbrace{(\alpha-\beta)^\top \Sigma (\alpha'-\beta')}_{C}.$$

Since α and β are zero-mean and independent, the covariance matrix of $\alpha - \beta$ (and $\alpha' - \beta'$) is:

$$C_{p,L} + C_{q,R} = M^{-1}\left[2\gamma[\mathrm{Diag}\{p\} - pp^\top] + 2\bar{\gamma}[\mathrm{Diag}\{q\} - qq^\top]\right] \preceq 2M^{-1}\mathrm{Diag}\{\gamma p + \bar{\gamma} q\},$$

whence

$$\begin{aligned}
\mathbb{E}\{B^2\} &= (p-q)^\top \Sigma [C_{p,L} + C_{q,R}] \Sigma (p-q) \\
&\leq (p-q)^\top \Sigma [2M^{-1}\mathrm{Diag}\{\gamma p + \bar{\gamma} q\}] \Sigma (p-q) \\
&= 2M^{-1} \textstyle\sum_i \sigma_i^2 (p_i - q_i)^2 (\gamma p_i + \bar{\gamma} q_i),
\end{aligned}$$

and similarly $\mathbb{E}\{[B']^2\} \leq 2M^{-1}\sum_i \sigma_i^2(p_i - q_i)^2(\gamma p_i + \bar{\gamma} q_i)$. Moreover,

$$\begin{aligned}
\mathbb{E}\{C^2\} &= \mathbb{E}\left\{[\textstyle\sum_i \sigma_i(\alpha_i - \beta_i)(\alpha_i' - \beta_i')]^2\right\} \\
&= \textstyle\sum_{i,j} \sigma_i\sigma_j \mathbb{E}\{(\alpha_i - \beta_i)(\alpha_j - \beta_j)\}\mathbb{E}\{(\alpha_i' - \beta_i')(\alpha_j' - \beta_j')\} \\
&= \textstyle\sum_{i,j} \sigma_i\sigma_j [C_{p,L} + C_{q,R}]_{ij}^2 \\
&= \textstyle\sum_i \sigma_i^2 [C_{p,L} + C_{q,R}]_{ii}^2 + \sum_{i\neq j} \sigma_i\sigma_j [C_{p,L} + C_{q,R}]_{ij}^2 \\
&= 4M^{-2}\left[\textstyle\sum_i \sigma_i^2 \left[\gamma[p_i - p_i^2] + \bar{\gamma}[q_i - q_i^2]\right]^2 + \sum_{i\neq j} \sigma_i\sigma_j[\gamma p_i p_j + \bar{\gamma} q_i q_j]^2\right] \\
&\leq 4M^{-2}\left[\textstyle\sum_i \sigma_i^2[\gamma p_i + \bar{\gamma} q_i]^2 + \sum_{i\neq j} \sigma_i\sigma_j[\gamma p_i p_j + \bar{\gamma} q_i q_j]^2\right] \\
&\leq 4M^{-2}\left[\gamma\textstyle\sum_i \sigma_i^2 p_i^2 + \bar{\gamma}\sum_i \sigma_i^2 q_i^2 + \gamma\sum_{i\neq j} \sigma_i\sigma_j p_i^2 p_j^2 + \bar{\gamma}\sum_{i\neq j} \sigma_i\sigma_j q_i^2 q_j^2\right] \\
&\leq 4M^{-2}\left[\gamma\textstyle\sum_i \sigma_i^2 p_i^2 + \bar{\gamma}\sum_i \sigma_i^2 q_i^2 + \gamma\left[\sum_i \sigma_i p_i^2\right]^2 + \bar{\gamma}\left[\sum_i \sigma_i q_i^2\right]^2\right].
\end{aligned}$$

Note that

$$[\textstyle\sum_i \sigma_i p_i^2]^2 = [\sum_i (\sigma_i p_i) p_i]^2 \leq [\sum_i \sigma_i^2 p_i^2][\sum_i p_i^2] \leq \sum_i \sigma_i^2 p_i^2,$$

which combines with the previous computation to imply that

$$\mathbb{E}\{C^2\} \leq 8M^{-2}\left[\gamma\textstyle\sum_i \sigma_i^2 p_i^2 + \bar{\gamma}\sum_i \sigma_i^2 q_i^2\right].$$

Consequently, by applying Chebyshev's inequality to B, B', and C, respectively, for every $\theta \geq 1$, the probability $\pi(\theta)$ of the event

$$\begin{aligned}
|\chi - \textstyle\sum_i \sigma_i(p_i - q_i)^2| \leq 2\sqrt{2}\theta[&\sqrt{M^{-1}\textstyle\sum_i \sigma_i^2(p_i - q_i)^2(\gamma p_i + \bar{\gamma} q_i)} + \\
&M^{-1}\sqrt{\gamma\textstyle\sum_i \sigma_i^2 p_i^2 + \bar{\gamma}\sum_i \sigma_i^2 q_i^2}]
\end{aligned}$$

is at least $1 - 3/\theta^2$. By inspecting the derivation, it is immediately seen that when $p = q$, the lower bound $1 - 3/\theta^2$ on $\pi(\theta)$ can be improved to $1 - 1/\theta^2$, so that

$$\pi(\theta) \geq \begin{cases} 1 - 3/\theta^2, p \neq q, \\ 1 - 1/\theta^2, p = q. \end{cases}$$

Taking into account what $\mathcal{T}$ is, the conclusion of Proposition 4.1 follows. □

Remark B.1. Note we do not use the standard "Poissonization" approach which assumes that, rather than drawing N independent samples from a distribution, first select N' from Poisson distribution with mean value N, and then draw N' samples. Such Poissonization makes the number of times different elements occur in the sample independent, simplifying the analysis. Instead, we model the empirical distribution directly by considering the dependence in the covariance matrix $C_{r,N}$.

Proof of Proposition 4.2. We prove the Proposition in two steps.

Step 1. Let ξ, ξ' be empirical distributions of observations in two consecutive Q-element segments of sample X^1 that are generated from distribution p. Setting $\eta = \xi - p$, $\eta' = \xi' - p$, we have

$$\xi^\top \xi' = p^\top p + \underbrace{p^\top \eta}_{B} + \underbrace{p^\top \eta'}_{B'} + \underbrace{\eta^\top \eta'}_{C}.$$

Since η and η' are zero-mean vectors and independent of each other, with covariance matrix $C_{p,Q} = Q^{-1}[\mathrm{Diag}\{p\} - pp^\top] \preceq Q^{-1}\mathrm{Diag}\{p\}$, we have

$$\mathbb{E}\{B^2\} = \mathbb{E}\{[B']^2\} = p^\top C_{p,Q} p \preceq Q^{-1} p^\top \mathrm{Diag}\{p\} p \leq Q^{-1} \textstyle\sum_i p_i^3,$$

and

$$\begin{aligned}\mathbb{E}\{C^2\} &= \textstyle\sum_{i,j} \mathbb{E}\{\eta_i \eta_i' \eta_j \eta_j'\} \\ &= \textstyle\sum_{i,j} \left[\mathbb{E}\{\eta_i \eta_j\}\right]^2 = Q^{-2}\left[\sum_i [p_i - p_i^2]^2 + \sum_{i \neq j} p_i^2 p_j^2\right] \leq 2Q^{-2} \textstyle\sum_i p_i^2.\end{aligned}$$

Consequently, by applying Chebyshev's inequality to B, B', and C, respectively, the probability of the event

$$|\xi^\top \xi' - p^\top p| > 3\left[2Q^{-1/2}\sqrt{\textstyle\sum_i p_i^3} + \sqrt{2}Q^{-1}\|p\|_2\right] \tag{B.1}$$

is $\leq 1/3$. Taking into account that $\sum_i p_i^3 \leq \|p\|_2^2 \|p\|_\infty \leq \|p\|_2^3$, we have proved the

following Lemma.

Lemma B.1 (Concentration Inequality for $\xi^\top \xi'$)**.** *Assume that there exists $\rho \in \mathbb{R}_+$ such that the distribution p satisfies the relation $\|p\|_2 \leq \sqrt{2}\rho$, and a positive integer Q be such that*

$$3\left[2^{7/4}Q^{-1/2}\rho^{3/2} + 2Q^{-1}\rho\right] \leq \frac{1}{3}\rho^2. \tag{B.2}$$

When ξ, ξ' are empirical distributions of two consecutive segments, of cardinality Q each, generated from distribution p, we have

$$\mathbb{P}\left\{|\xi^\top \xi' - \|p\|_2^2| > \frac{1}{3}\rho^2\right\} \leq \frac{1}{3}. \tag{B.3}$$

Step 2. The parameters $Q = Q_i$ and $\rho = \rho_i$ of i-th stage of the training-step in Algorithm 1 satisfy (B.2). Recalling that by the definition of $i(p)$ we have $\|p\|_2 \leq \sqrt{2}\rho_i$, $1 \leq i \leq i(p)$. Invoking Lemma B.1 and the definition of S in (4.6), we conclude that the probability of the event

$$\mathcal{E} : |\Theta_i - \|p\|_2^2| \leq \frac{1}{3}\rho_i^2, \, \forall 1 \leq i \leq i(p),$$

is at least $1 - \delta$. Assume that this event takes place. By the definition of $i(p)$, we have $\rho_{i(p)}^2 \leq \|p\|_2^2$, and since we are in the case of $\mathcal{E}$, we have also $|\Theta_{i(p)} - \|p\|_2^2| \leq \frac{1}{3}\rho_{i(p)}^2$, whence

$$\Theta_{i(p)} \geq \frac{2}{3}\|p\|_2^2 \geq \frac{2}{3}\rho_{i(p)}^2.$$

We see that under our assumption the trial run ends up with a success at some stage $k \leq i(p)$, so that

$$\Theta_k \geq \frac{2}{3}\rho_k^2 \text{ and } |\Theta_k - \|p\|_2^2| \leq \frac{1}{3}\rho_k^2$$

(the second relation holds true since we are in the case of event $\mathcal{E}$). As a result,

$$\varrho = \Theta_k + \frac{1}{3}\rho_k^2 \geq \|p\|_2^2,$$

and $|\Theta_k - \|p\|_2^2| \le \rho_k^2/3 \le \Theta_k/2$, implying that $\Theta_k \le 2\|p\|^2$, whence

$$\varrho = \Theta_k + \frac{1}{3}\rho_k^2 \le \Theta_k + \frac{1}{2}\Theta_k = \frac{3}{2}\Theta_k \le 3\|p\|_2^2.$$

We see that, with probability at least $1-\delta$, the number of stages in the training-step is at most $i(p)$, and the output ϱ of the test satisfies (4.5). Besides this, from (B.2) it is immediately seen that $Q_i \le O(1)/\rho_i$, so that $Q_{i(p)} \le O(1)/\|p\|_2$ due to $\rho_{i(p)} \ge \|p\|_2/\sqrt{2}$. Thus, when the training-step stops before or at stage $i(p)$, the total number of observations used in training-step indeed does not exceed $4SQ_{i(p)} \le O(1)S/\|p\|_2$. Note that by the definition of S in (4.6), we have

$$S \le O(1)\ln(\ln(n)/\delta), \tag{B.4}$$

which implies (4.7). □

Proof of Theorem 4.1. By Proposition 4.2, with properly selected $O(1)$ in (4.9), the probability for the training-step to be successful is at least $1-\delta$, and there is enough observations to perform the K individual tests of the testing stage. From now on we assume that $O(1)$ in (4.9) meets this requirement.

For $k \le i(p)$, let $\mathcal{E}_k$ be the condition stating that the training-step is successful and terminates at stage k. Note that this is a condition on the first $N_k = 4SQ_k$ observations of the sample set X^1. Let us fix a realization of these N_k observations satisfying condition $\mathcal{E}_k$; from now on, speaking about probabilities of various events, we mean probabilities taken with respect to conditional, the above realization given, probability distribution of the remaining $N - N_k$ observations in sample X^1 and the entire N observations in X^2.

We first prove the type-I risk is at most α. Note that we are in the situation when the training-step was successful, hence $\varrho \ge \|p\|_2^2$. Consequently, the threshold (4.8) satisfies relation (4.3) with $\theta = 3$ and $\sigma_i \equiv 1$, implying by the first claim in Proposition 4.1 (where we set $L = R = M$) that when $p = q$, the probability to claim H_1 by a particular one of

the K individual tests is at most $1/9$. By the definition of $K(\alpha, \beta)$, we conclude that the type-I risk is indeed at most α.

We then prove the type-II risk is at most β whenever the condition (4.10) holds. Assume that $\|p - q\|_2 \geq \epsilon \|p\|_2$ with some $\epsilon > 0$, and set $L = R = M$, $\theta = 3$ and $\sigma_i \equiv 1$. With ℓ given by (4.8), the inequality (4.4) reads

$$\begin{aligned} \|p - q\|_2^2 > 6\sqrt{2}\big[M^{-1}\sqrt{\varrho} + M^{-1/2}\sqrt{\textstyle\sum_i (p_i - q_i)^2(p_i + q_i)/2} + \\ M^{-1}\sqrt{(\|p\|_2^2 + \|q\|_2^2)/2}\big]. \end{aligned} \tag{B.5}$$

Note that the condition (B.5) ensures that the power of every individual test is at least $2/3$; thus, due to the choice of $K(\alpha, \beta)$, the type-II risk is at most β. It only remains to verify that condition (4.10) implies the validity of (B.5). Since we are in the situation that the training-step is successful, the condition (4.5) holds and in particular, $\|p\|_2 \leq \sqrt{\varrho} \leq \sqrt{3}\|p\|_2$, implying that the right hand side in (B.5) is at most

$$\mathcal{R} = O(1)\left[M^{-1}[\sqrt{\varrho} + \|q\|_2] + M^{-1/2}\|p - q\|_2\sqrt{\sqrt{\varrho} + \|q\|_2}\right],$$

and therefore in order to ensure the validity of (B.5), it suffices to ensure that

$$\|p - q\|_2^2 \geq O(1)M^{-1}[\sqrt{\varrho} + \|q\|_2]. \tag{B.6}$$

First consider the case when $\|q\|_2 \leq 2\|p\|_2$, which combines with (4.5) to imply that the right hand side in (B.6) is $\leq O(1)M^{-1}\sqrt{\varrho}$. By (4.5) and $\|p - q\|_2 \geq \epsilon\|p\|_2$, the left hand side in (B.6) is at least $O(1)\epsilon^2\varrho$, so that (B.6) indeed is implied by (4.10), provided that the absolute constant factor in the latter relation is selected properly. Then consider the case when $\|q\|_2 \geq 2\|p\|_2$. In this case, by (4.5), the right hand side in (B.6) is at most $O(1)M^{-1}\|q\|_2$, and the left hand side in (B.6) is at least $O(1)\|q\|_2^2$, implying that (B.6) holds true when $M \geq O(1)/\|q\|_2$. The validity of the latter condition, in view of $\|q\|_2 \geq 2\|p\|_2$, clearly is guaranteed by the validity of (4.10), provided $O(1)$ in the latter

relation is selected properly. □

Proof of Proposition 4.3. Assuming n is even, consider the following two scenarios on distributions p, q from which the two sets of sample X^1 and X^2 are independently generated:

1. Both samples are i.i.d. drawn from the uniform distribution on Ω.

2. The nature draws, independently of each other, two $n/2$-element subsets, Ω_1 and Ω_2, of Ω, from the uniform distribution on the family of all subsets of Ω of cardinality $n/2$; the N-observation samples X^1 are i.i.d. drawn from the uniform distribution on Ω_1, and the N-observation samples X^2 are i.i.d. drawn from the uniform distribution on Ω_2.

In the first scenario, the hypothesis H_0 is true; in the second, there is a *significant* difference between p and q – with probability close to 1 when n is large enough, we have the $\|p - q\|_2 \geq \epsilon\|p\|_2 = \epsilon\sqrt{2/n}$ for any ϵ small enough, e.g., for $0 < \epsilon < 1/2$. Denote the union of two sets of samples as $x^{2N} := (x_1, ..., x_{2N})$. It follows that if there exists a test $\mathcal{T}$ obeying the premise of Proposition 4.1, then there exists a low risk test deciding on whether the entire $2N$-element sample x^{2N} shown to us is generated according to the first or the second scenario.

Specifically, given x^{2N}, let us split it into two halves and apply to the two resulting N-observation samples the test $\mathcal{T}$; if the test claim H_1, we conclude that x^{2N} is generated according to the second scenario, otherwise we claim that x^{2N} is generated according to the first scenario. When x^{2N} is generated by the first scenario, the probability for $\mathcal{T}$ to claim H_1 is at most α, that is, the probability to reject the first scenario when it is true is at most α. On the other hand, when x^{2N} is generated according to the second scenario, the conditional, p and q given, probability for $\mathcal{T}$ to accept H_0 should be at most β, provided that $\|p - q\|_2 \geq \epsilon\|p\|_2$ for a given $\epsilon \in (0, 1/2)$; when n is large, the probability for the condition $\|p - q\|_2 \geq \epsilon\|p\|_2$ to hold true approaches 1, so that for large enough values of n, the probability for the condition to hold is at least $1 - \beta$ and therefore the probability of

claiming a sample generated according to the second scenario as one generated according to the first one, is at most 2β. Thus, for properly selected n_0 and all $n \geq n_0$, given x^{2N}, we can decide with risk $\leq 2\beta$ on the scenario resulted in x^{2N}.

On the other hand, consider the distribution of x^{2N}. The corresponding observation space is the space Ω^{2N} of $2N$-element sequences with entries from Ω. Let $\widehat{\Omega}$ be the part of Ω^{2N} comprised of sequences with *all entries different from each other*, and $\widetilde{\Omega}$ be the complement of $\widehat{\Omega}$ in Ω^{2N}. Let also P_1 and P_2 be the distributions of our observations under the first and the second scenarios, and P be the distribution on Ω^{2N} which assigns equal masses to all points from $\widehat{\Omega}$ and zero masses to the points outside of $\widehat{\Omega}$. By evident symmetry reasons, we have

$$P_i = (1-\epsilon_i)P + \epsilon_i Q_i,\ i = 1, 2,$$

where Q_1 and Q_2 are probability distributions supported on $\widetilde{\Omega}$, and ϵ_i is the probability, under scenario i, to observe $2N$-element sample in $\widetilde{\Omega}$. We clearly have

$$\epsilon_1 \leq \frac{N(2N-1)}{n},\ \epsilon_2 \leq \frac{4N(2N-1)}{n}.$$

Indeed, for a fixed pair of indexes t_1, t_2, $1 \leq t_1 < t_2 \leq 2N$, the probability to get $x_{t_1} = x_{t_2}$ in x^{2N} is $1/n$ under the first scenario and is at most $4/n$ under the second scenario, while the number of pairs t_1, t_2 in question is $N(2N-1)$. We see that

$$\begin{array}{l}\sum_{\zeta^{2N}\in\Omega^{2N}} \min\{P_1(\zeta^{2N}), P_2(\zeta^{2N})\} = \min\{1-\epsilon_1, 1-\epsilon_2\} \sum_{\zeta^{2N}\in\widehat{\Omega}} P(\zeta^{2N}) \\ \qquad + \sum_{\zeta^{2N}\in\widetilde{\Omega}} \min\{\epsilon_1 Q_1(\zeta^{2N}), \epsilon_2 Q_2(\zeta^{2N})\} \geq \min\{1-\epsilon_1, 1-\epsilon_2\},\end{array}$$

implying that our scenarios cannot be decided upon with risk $\leq \min\{1-\epsilon_1, 1-\epsilon_2\}/2$. When $N \ll \sqrt{n}$, both ϵ_1 and ϵ_2 are small, so that it is impossible to decide on our scenarios with risks α (and 2β). The bottom line is that under the premise of Proposition 4.1, we either have $n \leq n_0$, or $N \geq O(1)\sqrt{n}$ with properly selected constant $O(1)$ that depends on

α, β, ϵ, and the conclusion follows. Note that here the dependence on reliability parameters is logarithmic and is not our focus here since here we aim to study the sample optimality with respect to the cardinality n of the set Ω. $\square$

Proof of Proposition 4.5. In order to approximate the correlation between statistics $\chi_{t,k}$ and $\chi_{\tau,s}$, we consider a simple case when the sample size $t-k = \tau-s = 2m$. Denote the cardinality of the non-overlapping part in two time windows as $\delta = |t-\tau| = |k-s|$. Without loss of generality, we consider two sequences, $\{x_1, x_2, \ldots, x_{4m}\}$ and $\{x_{1+\delta}, x_2, \ldots, x_{4m+\delta}\}$, with $4m - \delta$ overlapping elements, as illustrated and grouped in Figure B.1.

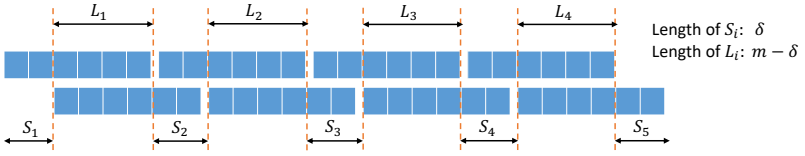


Figure B.1: Sliding window illustration.

Recall that $\mathrm{Var}(\chi_{t,k}) = 4\left[\sum_{i=1}^n \sigma_i^2 p_i^2 (1-p_i)^2 + \sum_{i\neq j} \sigma_i\sigma_j p_i^2 p_j^2\right]$ from the previous computations. Therefore, it remains to compute $\mathbb{E}[\chi_{t,k}\chi_{\tau,s}]$. For simplicity, we compute $m^2\mathbb{E}[\chi_t\chi_{t+\delta}]$ by writing it as a summation of several indicator functions as follows:

$$
\begin{aligned}
&m^2\mathbb{E}[\chi_{t,k}\chi_{\tau,s}] \\
=&\mathbb{E}\left[\sum_{i=1}^n \sigma_i \left(\mathbb{1}_i\{S_1\} + \mathbb{1}_i\{L_1\} - \mathbb{1}_i\{S_3\} - \mathbb{1}_i\{L_3\}\right)\left(\mathbb{1}_i\{S_2\} + \mathbb{1}_i\{L_2\} - \mathbb{1}_i\{S_4\} - \mathbb{1}_i\{L_4\}\right)\right] \\
&\cdot\left[\sum_{i=1}^n \sigma_i \left(\mathbb{1}_i\{L_1\} + \mathbb{1}_i\{S_2\} - \mathbb{1}_i\{L_3\} - \mathbb{1}_i\{S_4\}\right)\left(\mathbb{1}_i\{L_2\} + \mathbb{1}_i\{S_3\} - \mathbb{1}_i\{L_4\} - \mathbb{1}_i\{S_5\}\right)\right],
\end{aligned}
$$

where $\mathbb{1}_i\{\mathcal{S}\} = \sum_{k\in\mathcal{S}} \mathbb{1}\{x_k = i\}$ for a given index set $\mathcal{S}$. Since x_i are i.i.d. random variables, we have $\mathbb{E}(\mathbb{1}_i\{\mathcal{S}\}) = |\mathcal{S}|p_i$, $\mathrm{Var}(\mathbb{1}_i\{\mathcal{S}\}) = |\mathcal{S}|p_i(1-p_i)$, and $\mathbb{E}(\mathbb{1}_i\{\mathcal{S}\}\mathbb{1}_j\{\mathcal{S}\}) = |\mathcal{S}|(|\mathcal{S}|-1)p_ip_j$, with $|\mathcal{S}|$ denotes the cardinality. More specifically, for the decomposition shown in Figure B.1, we have $|S_i| = \delta$ and $|L_i| = m - \delta$. Substitute these into the above formulation, we have

$$
m^2\mathbb{E}[\chi_{t,k}\chi_{\tau,s}] = [4(m-\delta)^2 - 2\delta^2]\left[\sum_{i=1}^n \sigma_i^2 p_i^2 (1-p_i)^2 + \sum_{i\neq j} \sigma_i\sigma_j p_i^2 p_j^2\right].
$$

Therefore, the correlation of the statistic $\chi_{t,k}$ and $\chi_{\tau,s}$ is

$$\mathrm{Corr}(\chi_{t,k}, \chi_{\tau,s}) = \frac{4(m-\delta)^2 - 2\delta^2}{4m^2} = 1 - 2\frac{\delta}{m} + \frac{\delta^2}{2m^2}.$$

Substitute $\delta = |t-\tau| = |k-s|$ into the above equation then we complete the proof. □

Proof of Theorem 4.2. The proof is based on a general method for computing first passage probabilities first introduced in [218] and further developed in [182] and [179], and commonly used in similar problems [214, 120, 35]. First of all, it is worth mentioning that the probability measure in the following proof always stands for the nominal case where all samples are from the same distribution p. We define the variable

$$Z_\tau = \tau(\xi_\tau - \eta_\tau)^\top \Sigma(\xi'_\tau - \eta'_\tau),$$

where $\xi_\tau, \eta_\tau, \xi'_\tau, \eta'_\tau$ are empirical distributions of four independent (non-overlapping) sequences with equal length τ. Recall that $\mathbb{E}[Z_\tau] = 0$ and $\mathrm{Var}(Z_\tau) = \sigma_p^2$. We denote the moment generating function as

$$\psi_\tau(\theta) = \log \mathbb{E}\left[\exp\{\theta Z_\tau\}\right],$$

and select $\theta = \theta_\tau$ by solving the equation $\dot{\psi}_\tau(\theta) = b$. Since Z_τ is defined by a function of 4τ independent random samples, ϕ_τ converges to a limit as $\tau \to \infty$ and θ_τ converges to a limiting value, denoted by θ. The transformed distribution for all sequences at a fixed center position k and the window size τ is denoted by $\mathbb{P}_k^\tau$ and is defined by

$$d\mathbb{P}_k^\tau = \exp\{\theta Z_k^\tau - \psi_\tau(\theta_\tau)\} d\mathbb{P},$$

where $Z_k^\tau = \tau(\xi_{k,\tau} - \eta_{k,\tau})^\top \Sigma(\xi'_{k,\tau} - \eta'_{k,\tau})$ is the statistic for location k and window size τ, as indicated in Figure 4.2.

Let

$$\ell(k,\tau) := \log(d\mathbb{P}_k^\tau / d\mathbb{P}) = \theta Z_k^\tau - \psi_\tau(\theta_\tau).$$

Denote $D = \{(k,\tau) : 0 \le k \le m, \lceil m_0/2 \rceil \le \tau \le \lceil m_1/2 \rceil\}$ be the set of all possible windows in the scan. Let $A = \{\max_{(k,\tau)\in D} Z_k^\tau \ge b\}$ be the event of interests (the event $\{\mathcal{T}' \le m\}$), i.e., the procedures stop before time m.

By measure transformation, we have

$$\begin{aligned}
\mathbb{P}(A) &= \sum_{(k,\tau)\in D} \mathbb{E}\left[\exp[\ell(k,\tau)] \Big(\sum_{(k',\tau')\in D} \exp[\ell(k',\tau')] \Big)^{-1}; A \right] \\
&= \sum_{(k,\tau)\in D} \mathbb{E}_k^\tau \left[\Big(\sum_{(k',\tau')\in D} \exp[\ell(k',\tau')] \Big)^{-1}; A \right] \\
&= \sum_{(k,\tau)\in D} e^{\tilde{\ell}(k,\tau)-\ell(k,\tau)} \times \mathbb{E}_k^\tau \left[\frac{\max_{k',\tau'} e^{\ell(k',\tau')-\ell(k,\tau)}}{\sum_{k',\tau'} e^{\ell(k',\tau')-\ell(k,\tau)}} e^{-\tilde{\ell}(k,\tau)-[\max_{k',\tau'}\ell(k',\tau')-\ell(k,\tau)]}; A \right] \\
&= e^{-\theta_\tau \dot{\psi}_\tau(\theta_\tau)+\psi_\tau(\theta_\tau)} \times \sum_{(k,\tau)\in D} \mathbb{E}_k^\tau \left[\frac{M(k,\tau)}{S(k,\tau)} e^{-\tilde{\ell}(k,\tau)-\log M(k,\tau)}; A \right],
\end{aligned} \tag{B.7}$$

where

$$\begin{aligned}
\tilde{\ell}(k,\tau) &= \theta_\tau [Z_k^\tau - \dot{\psi}_\tau(\theta_\tau)], \\
M(k,\tau) &= \max_{k',\tau'} \exp\{\theta_\tau (Z_{k'}^{\tau'} - Z_k^\tau)\}, \\
S(k,\tau) &= \textstyle\sum_{k',\tau'} \exp\{\theta_\tau (Z_{k'}^{\tau'} - Z_k^\tau)\}.
\end{aligned}$$

Since k, τ are fixed in much of the following analysis, we suppress the dependence of the notation on k, τ and simply writet $\tilde{\ell}, S, M$. Under certain verifiable assumptions [179], a localization lemma allows us to simplify the expectation

$$\mathbb{E}_k^\tau \left[\frac{M}{S} e^{-\tilde{\ell}-\log M}; \tilde{\ell} + \log M \ge 0 \right]$$

into a simpler form

$$\frac{1}{\sqrt{2\pi\sigma_\tau^2}} \mathbb{E}\left[\frac{M}{S}\right],$$

where σ_τ^2 stands for the variance of $\tilde{\ell}$ under measure $\mathbb{P}_k^\tau$. The reduction relies on the fact that for large m, the local processes M and S are approximately independent of the global process $\tilde{\ell}$. Such independence allows the above decomposition into the expectation of M/S times the expectation involving $\tilde{\ell} + \log M$, treating $\log M$ essentially as a constant.

We first consider the process M and S and derive the expectation $\mathbb{E}[M/S]$ following [182]. The difference between $Z_{k'}^{\tau'}$ and Z_k^τ can be writted in the form

$$\begin{aligned} Z_{k'}^{\tau'} - Z_k^\tau &= \tau'(\xi_{k',\tau'} - \eta_{k',\tau'})^\top \Sigma(\xi'_{k',\tau'} - \eta'_{k',\tau'}) - \tau(\xi_{k,\tau} - \eta_{k,\tau})^\top \Sigma(\xi'_{k,\tau} - \eta'_{k,\tau}) \\ &= \tau' \left[(\xi_{k',\tau'} - \eta_{k',\tau'})^\top \Sigma(\xi'_{k',\tau'} - \eta'_{k',\tau'}) - (\xi_{k,\tau} - \eta_{k,\tau})^\top \Sigma(\xi'_{k,\tau} - \eta'_{k,\tau})\right] \\ &\quad + (\tau' - \tau)(\xi_{k,\tau} - \eta_{k,\tau})^\top \Sigma(\xi'_{k,\tau} - \eta'_{k,\tau}). \end{aligned}$$

Observe that one may let $\tau' = \tau$ and substitute $\theta = \lim_{\tau\to\infty} \theta_\tau$ for θ_τ in the definition of the increments and still maintain the required level of accuracy. When $\tau' = \tau$, the second term in the above expression vanish and the first term consists of two terms that are highly correlated. As characterized in Proposition 4.5, when $\tau' = \tau$, the covariance between the two terms is given by

$$\mathrm{Cov}(\theta_\tau Z_{k'}^{\tau'}, \theta_\tau Z_k^\tau) = \theta_\tau^2 \mathbb{E}[Z_{k'}^{\tau'}, Z_k^\tau] = \theta_\tau^2 \sigma_p^2 \left(1 - 2\frac{|k' - k|}{\tau} + \frac{|k' - k|^2}{2\tau^2}\right).$$

When τ is large, we have that the correlation depends on the difference $|k' - k|$ in a linear form, which shows that we have the random walk in the change time k, and the variance of the increment equals to $2\theta_\tau^2\sigma_p^2/\tau$. Following [182], we have

$$\mathbb{E}[M/S] = [\theta_\tau^2 \sigma_p^2/\tau \nu([2\theta_\tau \sigma_p^2/\tau]^{1/2})]^2.$$

Moreover, the process $\tilde{\ell}$ is zero-mean and has variance $\sigma_\tau^2 = \mathrm{Var}_k^\tau(\tilde{\ell}) = \theta_\tau^2 \ddot{\psi}(\theta_\tau)$ under

the measure $\mathbb{P}_k^\tau$. Substituting the result for the expectations in (B.7) yields

$$\mathbb{P}(\mathcal{T}' \leq m) = 2 \sum_{\tau=\lceil m_0/2 \rceil}^{\lceil m_1/2 \rceil} (m-2\tau) e^{-\theta_\tau \dot{\psi}_\tau(\theta_\tau)+\psi_\tau(\theta_\tau)} \frac{[\theta_\tau^2 \sigma_p^2/\tau \nu([2\theta_\tau \sigma_p^2/\tau]^{1/2})]^2}{[2\pi\theta_\tau^2 \ddot{\psi}_\tau(\theta_\tau)]^{1/2}}.$$

In the limiting case, Z_k^τ can be well approximated using Gaussian distribution $\mathcal{N}(0, \sigma_p^2)$. The moment generating function then becomes $\psi(\theta) = \theta^2\sigma_p^2/2$, and the limiting $\theta = b/\sigma_p^2$, as the solution to $\dot{\psi}(\theta) = b$. Furthermore, the summation term can be approximated by an integral, to obtain

$$\begin{aligned}\mathbb{P}(\mathcal{T}' \leq m) &= 2 \sum_{\tau=\lceil m_0/2 \rceil}^{\lceil m_1/2 \rceil} (m-2\tau) e^{-b^2/(2\sigma_p^2)} [2\pi b^2/\sigma_p^2]^{-1/2} [b^2/(\tau\sigma_p^2) \nu([2b^2/(\tau\sigma_p^2)]^{1/2})]^2 \\ &\approx 4e^{-b^2/(2\sigma_p^2)} [2\pi b^2/\sigma_p^2]^{-1/2} [b^2/\sigma_p^2]^2 \int_{m_0/m}^{m_1/m} \nu^2([4b^2/(mt\sigma_p^2)]^{1/2})(1-t)dt/t^2. \end{aligned} \tag{B.8}$$

Here it is assumed that m is large, but small enough that the right-hand side of (B.8) converges to 0 when $b \to \infty$. Changing variables in the integrand, we can rewrite this approximation as

$$\mathbb{P}\{\mathcal{T}' \leq m\} \approx m \times 2e^{-b^2/(2\sigma_p^2)} [2\pi b^2/\sigma_p^2]^{-1/2} [b^2/\sigma_p^2] \int_{[4b^2/(m_1\sigma_p^2)]^{1/2}}^{[4b^2/(m_0\sigma_p^2)]^{1/2}} y\nu^2(y)dy. \tag{B.9}$$

From the arguments in [181, 184], we know that $\mathcal{T}'$ is asymptotically exponentially distributed and is uniformly integrable. Hence if λ denotes the factor multiplying m on the right-hand side of (B.9), then for large m, in the range where $m\lambda$ is bounded away from 0 and $+\infty$, $\mathbb{P}\{\mathcal{T}' \leq m\} - [1 - \exp(-\lambda m)] \to 0$. Consequently, $\mathbb{E}[\mathcal{T}'] \approx 1/\lambda$, thereby we complete the proof. Here we omit some technical details needed to make the derivation rigorous. Those details have been described and proved in [179]. □

Proof of Theorem 4.3. Recall that $\chi_{t,0}$ is defined in (4.13). For any time t, we have

$$\mathbb{E}_0[\chi_{t,0}] = \frac{t}{2}(p-q)^\top \Sigma (p-q),$$

which grows linearly with respect to time. At the stopping time $\mathcal{T} = T$, the expectation of the window-limited statistic in (4.15) can be computed if m_1 is sufficiently large (at least larger than the expected detection delay):

$$\mathbb{E}[\max_{0\le k\le T} \chi_{t,k}] \approx \mathbb{E}[\chi_{t,0}] = \frac{T}{2}(p-q)^\top \Sigma (p-q).$$

On the other hand, we have that

$$\mathbb{E}[\max_{0\le k\le T} \chi_{t,k}] = b + \mathbb{E}[\max_{0\le k\le T} \chi_{t,k} - b].$$

If we ignore the overshoot of the threshold over b since it is a of order $o(b)$ when $b \to \infty$ (detailed analysis for overshoot has been developed in [180]), then we obtain a first-order approximation as $b \to \infty$, by solving

$$\frac{\mathbb{E}_0[\mathcal{T}]}{2}(p-q)^\top \Sigma (p-q) = b(1+o(1)).$$

Therefore, a first-order approximation for the expected detection delay is given by

$$\mathbb{E}_0[\mathcal{T}] = \frac{b(1+o(1))}{(p-q)^\top \Sigma (p-q)/2}.$$

□

Proof of Proposition 4.6. For the minimization problem that defines $f(\sigma) := \min_{S\in\mathcal{S}} \text{Tr}(\Sigma S)$, we introduce matrix variable $T \in \mathbb{R}^{n\times n}$ such that $T_{ij} \ge |S_{ij}|$ and $\sum_{i=1}^n \sum_{j=1}^n T_{ij} \le 4$. We will jointly minimize over S and T in $f(\sigma)$. The Lagrangian function of the problem $f(\sigma)$

writes

$$
\begin{aligned}
&\mathcal{L}(\sigma, S, P, T, \lambda, W, U, \xi, \{x_k\}_{k=1}^K) \\
=& \text{Tr}(\Sigma S) - \text{Tr}(PS) - r\text{Tr}(SJ) + \lambda(\rho^2 - \text{Tr}(S)) \\
&+ \textstyle\sum_k x_k(\text{Tr}(SQ_k) - 4) + \xi(\text{Tr}(TJ) - 4) - \text{Tr}(U(T-S)) - \text{Tr}(W(T+S)) \\
=& \text{Tr}\left(\left(\Sigma - P - rJ - \lambda I_n + \textstyle\sum_{k=1}^K x_k Q_k + U - W\right) S\right) \\
&+ \lambda\rho^2 - 4\textstyle\sum_{k=1}^K x_k - 4\xi + \text{Tr}\left((\xi J - U - W)T\right),
\end{aligned}
$$

where $J \in \mathbb{R}^{n\times n}$ is a matrix with all elements equal to 1, and we have

$$
f(\sigma) = \min_{S,T} \max_{\substack{P\succcurlyeq 0, U\geq 0, W\geq 0 \\ \lambda\geq 0, \xi\geq 0, x_k\geq 0, r\in\mathbb{R}}} \mathcal{L}(\sigma, S, P, T, \lambda, W, U, \xi, \{x_k\}_{k=1}^K).
$$

Then we have the dual problem of $f(\sigma)$ can be represented as

$$
\begin{aligned}
\max\ & \lambda\rho^2 - 4\textstyle\sum_k x_k - 4\xi \\
\text{s.t.}\ & \lambda \geq 0, P \succcurlyeq 0, \xi \geq 0, x_k \geq 0, U \geq 0, W \geq 0, r \in \mathbb{R}, \\
& \textstyle\sum_k x_k Q_k + U - W - P - rJ - \lambda I_n \succcurlyeq -\Sigma, \\
& U_{ij} + W_{ij} \leq \xi, 1 \leq i \leq n, 1 \leq j \leq n.
\end{aligned}
$$

Next we derive the dual form for $g(\sigma)$, similarly, we write the Lagrangian function as

$$
\begin{aligned}
&\mathcal{L}(\Sigma, P, \Lambda, V, \nu, \{\mu_k\}_{k=1}^K) \\
=& \text{Tr}((\Sigma^2 + \Lambda + V)P) + \nu(\text{Tr}(PJ) - 1) + \textstyle\sum_k \mu_k(1 - \text{Tr}(PQ_k)),
\end{aligned}
$$

and we have

$$
g(\sigma) = \max_{P} \min_{\Lambda\succcurlyeq 0, V\geq 0, \nu\in\mathbb{R}, \mu_k\geq 0} \mathcal{L}(\Sigma, P, \Lambda, V, \nu, \{\mu_k\}_{k=1}^K).
$$

Then we have the dual form of $g(\sigma)$ is

$$\begin{aligned} \min\ & \textstyle\sum_k \mu_k - \nu \\ \text{s.t. } & \Lambda \succcurlyeq 0, V \geq 0, \mu_k \geq 0, 1 \leq k \leq K, \\ & -\Lambda - V + \textstyle\sum_k \mu_k Q_k - \nu J \succcurlyeq \Sigma^2. \end{aligned}$$

Then the constraint $g(\sigma) \leq 1$ can be simplified to: $\exists \Lambda \succcurlyeq 0, V \geq 0, \mu_k \geq 0, \nu \in \mathbb{R}$, such that

$$\textstyle\sum_k \mu_k - \nu \leq 1,\ -\Lambda - V + \sum_k \mu_k Q_k - \nu J \succcurlyeq \Sigma^2.$$

Combine with the dual form of $f(\sigma)$, we have the problem (4.21) is equivalent to

$$\begin{aligned} \max\ & \lambda\rho^2 - 4\textstyle\sum_k x_k - 4\xi \\ \text{s.t. } & \lambda \geq 0, P \succcurlyeq 0, \xi \geq 0, x_k \geq 0, U \geq 0, W \geq 0, r \in \mathbb{R}, \\ & \textstyle\sum_k x_k Q_k + U - W - P - rJ - \lambda I_n \succcurlyeq -\Sigma, \\ & U_{ij} + W_{ij} \leq \xi, 1 \leq i \leq n, 1 \leq j \leq n, \\ & \textstyle\sum_k \mu_k - \nu \leq 1, \\ & \Lambda \succcurlyeq 0, V \geq 0, \mu_k \geq 0, 1 \leq k \leq K, \\ & -\Lambda - V + \textstyle\sum_k \mu_k Q_k - \nu J \succcurlyeq \Sigma^2. \end{aligned}$$

□

APPENDIX C

PROOFS FOR CHAPTER 5

Proof of Lemma 5.1. Note that the probability measures P_1, P_2 are absolutely continuous with respect to $P_1 + P_2$, hence we have

$$\begin{aligned}
&\inf_{\pi} \Phi(\pi; P_1, P_2) \\
=&\inf_{\pi} \int_{\Omega} \left[(1-\pi(\omega))\tfrac{dP_1}{d(P_1+P_2)}(\omega) + \pi(\omega)\tfrac{dP_2}{d(P_1+P_2)}(\omega)\right] d(P_1+P_2)(\omega) \\
=&\inf_{\pi} \int_{\Omega_0} \left[(1-\pi(\omega))\tfrac{dP_1}{d(P_1+P_2)}(\omega) + \pi(\omega)\tfrac{dP_2}{d(P_1+P_2)}(\omega)\right] d(P_1+P_2)(\omega) \\
=&\int_{\Omega_0} \inf_{0\leqslant x\leqslant 1} \left[(1-x)\tfrac{dP_1}{d(P_1+P_2)}(\omega) + x\tfrac{dP_2}{d(P_1+P_2)}(\omega)\right] d(P_1+P_2)(\omega),
\end{aligned} \tag{C.1}$$

where the second equality holds because the integral depends only on the subset $\Omega_0 := \left\{\omega \in \Omega : 0 < \frac{dP_k}{d(P_1+P_2)}(\omega) < 1, k = 1, 2\right\}$, on which P_1, P_2 are absolutely continuous with respect to each other; the third equality is due to the Interchangeability Principle (Theorem 7.80, [173]).

For any ω, the infimum $\pi^*(\omega)$ of the inner minimization in (C.1) is attained at 0 or 1. Therefore, for any $\omega \in \Omega$,

$$(1-\pi^*(\omega))\tfrac{dP_1}{d(P_1+P_2)}(\omega) + \pi^*(\omega)\tfrac{dP_2}{d(P_1+P_2)}(\omega) = \min\left\{\tfrac{dP_1}{d(P_1+P_2)}(\omega), \tfrac{dP_2}{d(P_1+P_2)}(\omega)\right\}.$$

This completes the proof. □

Proof of Lemma 5.2. Denote by $L^1(\mu)$ the space of all integrable functions with respect to the measure μ. Using Lagrangian and Kantorovich's duality (Theorem 5.10, [201]), we

rewrite the LFD problem as:

$$
\begin{aligned}
&\sup_{P_1\in\mathcal{P}_1,P_2\in\mathcal{P}_2}\ \psi(P_1,P_2)\\
&=\sup_{\substack{P_1\in\mathscr{P}(\Omega)\\P_2\in\mathscr{P}(\Omega)}}\ \inf_{\lambda_1,\lambda_2\geq 0}\left\{\psi(P_1,P_2)+\sum_{k=1}^{2}\lambda_k\theta_k-\sum_{k=1}^{2}\lambda_k\sup_{\substack{u_k\in\mathbb{R}^{n_k}\\v_k\in L^1(P_k)}}\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}u_k^i\right.\right.\\
&\qquad\left.\left.+\int_\Omega v_k dP_k : u_k^i+v_k(\omega)\leq c(\omega,\widehat{\omega}_k^i),\ \forall 1\leq i\leq n_k,\forall\omega\in\Omega\right\}\right\}\\
&=\sup_{\substack{P_1\in\mathscr{P}(\Omega)\\P_2\in\mathscr{P}(\Omega)}}\ \inf_{\substack{\lambda_1,\lambda_2\geq 0\\u_k\in\mathbb{R}^{n_k}\\v_k\in L^1(P_k)}}\left\{\psi(P_1,P_2)+\sum_{k=1}^{2}\lambda_k\theta_k-\sum_{k=1}^{2}\lambda_k\left(\frac{1}{n_k}\sum_{i=1}^{n_k}u_k^i+\int_\Omega v_k dP_k\right):\right.\\
&\qquad\left.u_k^i+v_k(\omega)\leq c(\omega,\widehat{\omega}_k^i),\ \forall 1\leq i\leq n_k,\forall\omega\in\Omega\right\}\\
&=\sup_{\substack{P_1\in\mathscr{P}(\Omega)\\P_2\in\mathscr{P}(\Omega)}}\ \inf_{\substack{\lambda_1,\lambda_2\geq 0\\u_k\in\mathbb{R}^{n_k}\\v_k\in L^1(P_k)}}\left\{\psi(P_1,P_2)+\sum_{k=1}^{2}\lambda_k\theta_k-\sum_{k=1}^{2}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}u_k^i+\int_\Omega v_k dP_k\right):\right.\\
&\qquad\left.u_k^i+v_k(\omega)\leq \lambda_k c(\omega,\widehat{\omega}_k^i),\ \forall 1\leq i\leq n_k,\forall\omega\in\Omega\right\},
\end{aligned}
$$

where the second equality holds by combining the innermost supreme problem with the infimum problem; and the third equality holds by replacing $\lambda_k u_k^i$ with u_k^i and $\lambda_k v_k$ with v_k (note that such change of variable is valid even when $\lambda_k = 0$). Furthermore, since the objective function is non-increasing in v_k, we can replace v_k with $\min_{1\leq i\leq n_k}\{\lambda_k c(\omega,\widehat{\omega}_k^i) - u_k^i\}$ without changing the optimal value. Interchanging $\sup$ and $\inf$ yields

$$
\begin{aligned}
\sup_{P_1\in\mathcal{P}_1,P_2\in\mathcal{P}_2}\psi(P_1,P_2)\leq \inf_{\substack{\lambda_1,\lambda_2\geq 0\\u_k\in\mathbb{R}^{n_k}}}&\left\{\sum_{k=1}^{2}\lambda_k\theta_k-\sum_{k=1}^{2}\frac{1}{n_k}\sum_{i=1}^{n_k}u_k^i+\sup_{\substack{P_1\in\mathscr{P}(\Omega)\\P_2\in\mathscr{P}(\Omega)}}\left\{\psi(P_1,P_2)\right.\right.\\
&\left.\left.-\int_\Omega\sum_{k=1}^{2}\min_{1\leq i\leq n_k}\{\lambda_k c(\omega,\widehat{\omega}_k^i)-u_k^i\}dP_k\right\}\right\}.
\end{aligned}
\tag{C.2}
$$

Now let us study the inner supremum in (C.2). For a given distribution (P_1, P_2) and any $\omega \in \text{supp } P_1 \cup \text{supp } P_2$, where $\text{supp } P$ denotes the support of the distribution P, let

$i_k(\omega) = \arg\min_i\{\lambda_k c(\omega, \widehat{\omega}_k^i) - u_k^i\}$, $k = 1, 2$, set

$$T(\omega) := \begin{cases} \widehat{\omega}_1^{i_1(\omega)}, & \text{if } \lambda_1 \frac{dP_1}{d(P_1+P_2)}(\omega) \geq \lambda_2 \frac{dP_2}{d(P_1+P_2)}(\omega), \\ \widehat{\omega}_2^{i_2(\omega)}, & \text{if } \lambda_1 \frac{dP_1}{d(P_1+P_2)}(\omega) < \lambda_2 \frac{dP_2}{d(P_1+P_2)}(\omega), \end{cases}$$

whence

$$T(\omega) \in \arg\min_{\omega' \in \Omega} \left\{ \sum_{k=1}^{2} \left[\lambda_k c(\omega', \widehat{\omega}_k^{i_k(\omega)}) - u_k^{i_k(\omega)}\right] \frac{dP_k}{d(P_1+P_2)}(\omega) \right\}.$$

By definition we have $T(\omega) \in \widehat{\Omega}$. Define another solution (P_1', P_2') such that $P_k'(B) = P_k\{\omega \in \Omega : T(\omega) \in B\}$ for any Borel set $B \subset \widehat{\Omega}$. It follows that

$$\begin{aligned} &\sum_{k=1}^{2} \int_{\widehat{\Omega}} \min_{1 \leq i \leq n_k} \{\lambda_k c(\omega, \widehat{\omega}_k^i) - u_k^i\} dP_k'(\omega) \\ =& \sum_{k=1}^{2} \int_{\Omega} \min_{1 \leq i \leq n_k} \{\lambda_k c(T(\omega), \widehat{\omega}_k^i) - u_k^i\} dP_k(\omega) \\ \leq& \sum_{k=1}^{2} \int_{\Omega} \left(\lambda_k c(T(\omega), \widehat{\omega}_k^{i_k(\omega)}) - u_k^{i_k(\omega)}\right) dP_k(\omega) \\ \leq& \sum_{k=1}^{2} \int_{\Omega} \left(\lambda_k c(\omega, \widehat{\omega}_k^{i_k(\omega)}) - u_k^{i_k(\omega)}\right) dP_k(\omega). \end{aligned}$$

In addition, by a simple fact that $\sum_i \min\{x_i, y_i\} \leq \min\{\sum_i x_i, \sum_i y_i\}$ for any series $\{x_i, y_i\}$, we have

$$\begin{aligned} \psi(P_1, P_2) &= \int_{\Omega} \min\left\{ \frac{dP_1}{d(P_1+P_2)}(\omega), \frac{dP_2}{d(P_1+P_2)}(\omega) \right\} d(P_1+P_2)(\omega) \\ &\leq \sum_{\widehat{\omega} \in \widehat{\Omega}} \min\{P_1\{\omega \in \Omega : T(\omega) = \widehat{\omega}\}, P_2\{\omega \in \Omega : T(\omega) = \widehat{\omega}\}\} \\ &= \sum_{\widehat{\omega} \in \widehat{\Omega}} \min\{P_1'(\widehat{\omega}), P_2'(\widehat{\omega})\} \\ &= \psi(P_1', P_2'). \end{aligned}$$

Hence (P_1', P_2') yields an objective value no worse than (P_1, P_2) for the inner supremum in

(C.2). This suggests that in order to solve the inner supremum of (C.2), it suffices to only consider (P_1, P_2) with $\mathrm{supp}\ P_1 \subset \widehat{\Omega}$ and $\mathrm{supp}\ P_2 \subset \widehat{\Omega}$.

For $l = 1, \ldots, n$, set $p_k^l = P_k(\widehat{\omega}^l)$, and note that $\gamma_k \in \Gamma(P_k, Q_{k,n})$ can be identified with a non-negative matrix $\gamma_k \in \mathbb{R}_+^{n \times n}$ with each column and row summing up to 1. Thus, the inner supremum in (C.2) can now be equivalently written as

$$\sup_{\substack{p_1, p_2 \in \mathbb{R}_+^n \\ \sum_l p_1^l = 1, \sum_l p_2^l = 1}} \left\{ \sum_{l=1}^{n} \min\left\{p_1^l, p_2^l\right\} - \sum_{k=1}^{2} \sum_{l=1}^{n} p_k^l \min_{1 \le i \le n_k} \{\lambda_k c(\widehat{\omega}^l, \widehat{\omega}_k^i) - u_k^i\} \right\}.$$

It follows that

$$\begin{aligned} &\sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \psi(P_1, P_2) \\ &\le \inf_{\lambda_1, \lambda_2 \ge 0} \left\{ \sum_{k=1}^{2} \lambda_k \theta_k - \sum_{k=1}^{2} \frac{1}{n_k} \sum_{i=1}^{n_k} u_k^i + \sup_{\substack{p_1, p_2 \in \mathbb{R}_+^n \\ \sum_l p_1^l = 1, \sum_l p_2^l = 1}} \left\{ \sum_{l=1}^{n} \min\left\{p_1^l, p_2^l\right\} \right.\right. \\ &\qquad \left.\left. - \sum_{k=1}^{2} \sum_{l=1}^{n} p_k^l \min_{1 \le i \le n_k} \{\lambda_k c(\widehat{\omega}^l, \widehat{\omega}_k^i) - u_k^i\} \right\}. \end{aligned}$$

Applying the Lagrangian duality for finite-dimensional convex programming on the right-hand side yields

$$\sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \psi(P_1, P_2) \le \sup_{P_1 \in \widehat{\mathcal{P}}_1, P_2 \in \widehat{\mathcal{P}}_2} \psi(P_1, P_2),$$

Observe that both sides have the same objective function, but the feasible region of the right-hand side is a subset of that of the left-hand side, and thus the right-hand side should be no greater than the left-hand side, i.e., the above inequality should hold as equality. Thereby we complete the proof. □

Proof of Theorem 5.1. Note that from Lemma 5.2, we have

$$\begin{aligned} \sup_{P_1 \in \widehat{\mathcal{P}}_1, P_2 \in \widehat{\mathcal{P}}_2} \inf_{\pi: \Omega \to [0,1]} \Phi(\pi; P_1, P_2) &= \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \inf_{\pi: \Omega \to [0,1]} \Phi(\pi; P_1, P_2) \\ &\le \inf_{\pi: \Omega \to [0,1]} \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \Phi(\pi; P_1, P_2). \end{aligned}$$

Let us prove the other direction.

To begin with, we identify $\widehat{\pi} \in [0,1]^n$ with a function on $\widehat{\Omega}$. Using Theorem 2.3, we have:

$$\begin{aligned}\sup_{P_1 \in \widehat{\mathcal{P}}_1} \mathbb{E}_{P_1}[1-\widehat{\pi}] &= \min_{\lambda_1 \geq 0} \left\{ \lambda_1 \theta_1 + \frac{1}{n_1} \sum_{l=1}^{n_1} \max_{1 \leq m \leq n} \{1 - \widehat{\pi}_m - \lambda_1 c(\widehat{\omega}^l, \widehat{\omega}^m)\} \right\}, \\ \sup_{P_2 \in \widehat{\mathcal{P}}_2} \mathbb{E}_{P_2}[\widehat{\pi}] &= \min_{\lambda_2 \geq 0} \left\{ \lambda_2 \theta_2 + \frac{1}{n_2} \sum_{l=1}^{n_2} \max_{1 \leq m \leq n} \{\widehat{\pi}_m - \lambda_2 c(\widehat{\omega}^l, \widehat{\omega}^m)\} \right\}.\end{aligned} \tag{C.3}$$

Let λ_1^* and λ_2^* be respectively the minimizers of the two problems in (C.3). Observe that the right-hand sides of (C.3) and (5.9) are identical. Hence (5.9) implies that $\widehat{\pi}^*$ defined in the statement of Theorem 5.1 satisfies

$$\mathbb{E}_{P_1^*}[1-\widehat{\pi}^*] = \sup_{P_1 \in \widehat{\mathcal{P}}_1} \mathbb{E}_{P_1}[1-\widehat{\pi}^*], \;\; \mathbb{E}_{P_2^*}[\widehat{\pi}^*] = \sup_{P_2 \in \widehat{\mathcal{P}}_2} \mathbb{E}_{P_2}[\widehat{\pi}^*],$$

and thus

$$\sup_{P_1 \in \widehat{\mathcal{P}}_1, P_2 \in \widehat{\mathcal{P}}_2} \Phi(\widehat{\pi}^*; P_1, P_2) = \sup_{P_1 \in \widehat{\mathcal{P}}_1, P_2 \in \widehat{\mathcal{P}}_2} \inf_{\widehat{\pi} \in [0,1]^n} \Phi(\widehat{\pi}; P_1, P_2). \tag{C.4}$$

Hence $(\widehat{\pi}^*; P_1^*, P_2^*)$ solves the above finite-dimensional convex-concave saddle point problem that always has an optimal solution, which verifies the well-definedness of $\widehat{\pi}^*$.

On the other hand, for the π^* defined in the statement of Theorem 5.1, the optimization problem for finding worst-case risk are decoupled and admits the following equivalent reformulations (Theorem 2.3)

$$\begin{aligned}\sup_{P_1 \in \mathcal{P}_1} \mathbb{E}_{P_1}[1-\pi^*(\omega)] &= \min_{\lambda_1 \geq 0} \left\{ \lambda_1 \theta_1 + \frac{1}{n_1} \sum_{i=1}^{n_1} \sup_{\omega \in \Omega} \left\{ 1 - \pi^*(\omega) - \lambda_1 c(\omega, \widehat{\omega}_1^i) \right\} \right\}, \\ \sup_{P_2 \in \mathcal{P}_2} \mathbb{E}_{P_2}[\pi^*(\omega)] &= \min_{\lambda_2 \geq 0} \left\{ \lambda_2 \theta_2 + \frac{1}{n_2} \sum_{i=1}^{n_2} \sup_{\omega \in \Omega} \left\{ \pi^*(\omega) - \lambda_2 c(\omega, \widehat{\omega}_2^i) \right\} \right\}.\end{aligned} \tag{C.5}$$

Comparing (C.3) and (C.5), if we can prove π^* satisfies

$$\begin{aligned}
&\sup_{\omega\in\Omega}\left\{1-\pi^*(\omega)-\lambda_1^* c(\omega,\widehat{\omega}_1^i)\right\}\le\max_{\omega\in\widehat{\Omega}}\left\{1-\widehat{\pi}^*(\omega)-\lambda_1^* c(\omega,\widehat{\omega}_1^i)\right\},\ \forall 1\le i\le n_1,\\
&\sup_{\omega\in\Omega}\left\{\pi^*(\omega)-\lambda_2^* c(\omega,\widehat{\omega}_2^i)\right\}\le\max_{\omega\in\widehat{\Omega}}\left\{\widehat{\pi}^*(\omega)-\lambda_2^* c(\omega,\widehat{\omega}_2^i)\right\},\ \forall 1\le i\le n_2,
\end{aligned}\tag{C.6}$$

then π^* would be an optimal solution to (5.2) since

$$\begin{aligned}
\inf_{\pi:\Omega\to[0,1]}\sup_{P_1\in\mathcal{P}_1,P_2\in\mathcal{P}_2}\Phi(\pi;P_1,P_2)&\le\sup_{P_1\in\mathcal{P}_1,P_2\in\mathcal{P}_2}\Phi(\pi^*;P_1,P_2)\\
&\le\sup_{P_1\in\widehat{\mathcal{P}}_1,P_2\in\widehat{\mathcal{P}}_2}\Phi(\widehat{\pi}^*;P_1,P_2)\\
&=\sup_{P_1\in\mathcal{P}_1,P_2\in\mathcal{P}_2}\inf_{\pi:\Omega\to[0,1]}\Phi(\pi;P_1,P_2).
\end{aligned}$$

To show (C.6), for π^* restricted on the empirical support $\widehat{\Omega}$, we have

$$\begin{aligned}
&\sup_{\omega\in\widehat{\Omega}}\left\{1-\pi^*(\omega)-\lambda_1^* c(\omega,\widehat{\omega}_1^i)\right\}=\max_{\omega\in\widehat{\Omega}}\left\{1-\widehat{\pi}^*(\omega)-\lambda_1^* c(\omega,\widehat{\omega}_1^i)\right\},\ \forall 1\le i\le n_1,\\
&\sup_{\omega\in\widehat{\Omega}}\left\{\pi^*(\omega)-\lambda_2^* c(\omega,\widehat{\omega}_2^i)\right\}=\max_{\omega\in\widehat{\Omega}}\left\{\widehat{\pi}^*(\omega)-\lambda_2^* c(\omega,\widehat{\omega}_2^i)\right\},\ \forall 1\le i\le n_2.
\end{aligned}$$

Indeed, this holds by construction $\pi^*(\omega)=\widehat{\pi}^*(\omega)$ for $\omega\in\widehat{\Omega}$. It remains to show (C.6) also holds outside of $\widehat{\Omega}$:

$$\begin{aligned}
&\sup_{\omega\notin\widehat{\Omega}}\left\{1-\pi^*(\omega)-\lambda_1^* c(\omega,\widehat{\omega}_1^i)\right\}\le\max_{\omega\in\widehat{\Omega}}\left\{1-\widehat{\pi}^*(\omega)-\lambda_1^* c(\omega,\widehat{\omega}_1^i)\right\},\ \forall 1\le i\le n_1,\\
&\sup_{\omega\notin\widehat{\Omega}}\left\{\pi^*(\omega)-\lambda_1^* c(\omega,\widehat{\omega}_2^i)\right\}\le\max_{\omega\in\widehat{\Omega}}\left\{\widehat{\pi}^*(\omega)-\lambda_2^* c(\omega,\widehat{\omega}_2^i)\right\},\ \forall 1\le i\le n_2.
\end{aligned}$$

To prove this, note that it is equivalent to that $\forall\omega\notin\widehat{\Omega}$:

$$\begin{aligned}
&\pi^*(\omega)\ge\min_{\widehat{\omega}\in\widehat{\Omega}}\left\{\pi^*(\widehat{\omega})+\lambda_1^* c(\widehat{\omega},\widehat{\omega}_1^i)\right\}-\lambda_1^* c(\omega,\widehat{\omega}_1^i),\quad\forall i=1,\dots,n_1,\\
&\pi^*(\omega)\le\lambda_2^* c(\omega,\widehat{\omega}_2^j)-\min_{\widehat{\omega}\in\widehat{\Omega}}\left\{\lambda_2^* c(\widehat{\omega},\widehat{\omega}_2^j)-\pi^*(\widehat{\omega})\right\},\quad\forall j=1,\dots,n_2.
\end{aligned}\tag{C.7}$$

Observe that $\forall \omega \in \Omega$, $\forall i = 1, \ldots, n_1$ and $\forall j = 1, \ldots, n_2$, we have:

$$\begin{aligned}
&\min_{\widehat{\omega} \in \widehat{\Omega}} \left\{ \pi^*(\widehat{\omega}) + \lambda_1^* c(\widehat{\omega}, \widehat{\omega}_1^i) \right\} + \min_{\widehat{\omega} \in \widehat{\Omega}} \left\{ \lambda_2^* c(\widehat{\omega}, \widehat{\omega}_2^j) - \pi^*(\widehat{\omega}) \right\} \\
&\leq \begin{cases} \pi^*(\widehat{\omega}_2^j) + \lambda_1^* c(\widehat{\omega}_2^j, \widehat{\omega}_1^i) - \pi^*(\widehat{\omega}_2^j), & \lambda_1^* \leq \lambda_2^*, \\ \pi^*(\widehat{\omega}_1^i) + \lambda_2^* c(\widehat{\omega}_1^i, \widehat{\omega}_2^j) - \pi^*(\widehat{\omega}_1^i), & \lambda_1^* > \lambda_2^*, \end{cases} \\
&= \min\{\lambda_1^*, \lambda_2^*\} c(\widehat{\omega}_1^i, \widehat{\omega}_2^j) \\
&\leq \lambda_1^* c(\omega, \widehat{\omega}_1^i) + \lambda_2^* c(\omega, \widehat{\omega}_2^j),
\end{aligned}$$

where we have used the triangle inequality of the metric $c(\cdot, \cdot)$. And we note that

$$\min_{\widehat{\omega} \in \widehat{\Omega}} \left\{ \pi^*(\widehat{\omega}) + \lambda_1^* c(\widehat{\omega}, \widehat{\omega}_1^i) \right\} - \lambda_1^* c(\omega, \widehat{\omega}_1^i) \leq \pi^*(\widehat{\omega}_1^i) \leq 1,$$

and

$$\lambda_2^* c(\omega, \widehat{\omega}_2^j) - \min_{\widehat{\omega} \in \widehat{\Omega}} \left\{ \lambda_2^* c(\widehat{\omega}, \widehat{\omega}_2^j) - \pi^*(\widehat{\omega}) \right\} \geq \pi^*(\widehat{\omega}_2^j) \geq 0,$$

since $\pi^*(\omega) = \widehat{\pi}^*(\omega) \in [0, 1]$ for $\omega \in \widehat{\Omega}$. Therefore we always have $l(\omega) \leq u(\omega)$ and (C.7) always admits a feasible solution, as defined in the Theorem statement. □

Proof of Proposition 5.1. Given batch samples $\omega_1, \ldots, \omega_m$ sampled i.i.d. from the true distribution P_1°, define Boolean random variables $\xi_i, 1 \leq i \leq m$ as:

$$\xi_i = \begin{cases} 1 & \pi_1(\omega_i) = 0; \\ 0 & \pi_1(\omega_i) = 1, \end{cases}$$

more specifically, the random variable $\xi_i = 1$ if and only if the test, as applied to observation ω_i, rejects hypothesis H_0.

Further, by construction of the Majority test, if the hypothesis H_0 is rejected, then the number of i's with ξ_i = 1 is at least $m/2$. Thus, the probability to reject H_0 is not greater than the probability of the event: in m random Bernoulli trials with probability ϵ^*

of success, the total number of successes is $\geq m/2$. The probability of this event clearly does not exceed:

$$\sum_{m/2 \leq i \leq m} \binom{m}{i} (\epsilon^*)^i (1-\epsilon^*)^{m-i}.$$

When $\epsilon^* < 1/2$, by the Chernoff bound, we have

$$\sum_{m/2 \leq i \leq m} \binom{m}{i} (\epsilon^*)^i (1-\epsilon^*)^{m-i} \leq \exp\left\{-D(1/2||\epsilon^*)m\right\},$$

where $D(1/2||\epsilon^*) = \frac{1}{2}\log\frac{1}{2\epsilon^*} + \frac{1}{2}\log\frac{1}{2(1-\epsilon^*)}$ is the relative entropy between two Bernoulli distributions with "success" probabilities being $1/2$ and ϵ^* respectively. It is easy to see that $D(1/2||\epsilon^*) > 0$. Therefore, the risk goes to 0 exponentially fast, in the order of $\exp\{-D(1/2||\epsilon^*)m\}$ as $m \to \infty$.

□

Proof of Lemma 5.3. We first establish an optimality condition (Lemma C.1) for the constraint

$$\pi^\circ \in \operatorname*{arg\,min}_{\pi:\Omega\to[0,1]} \Phi(\pi; P_1, P_2).$$

Without causing confusion, we simply write $\pi^\circ \in \arg\min_\pi \Phi(\pi; P_1, P_2)$ in subsequent proofs.

Lemma C.1. *Let π° be the oracle test. For any $P_1, P_2 \in \mathscr{P}(\Omega)$, the constraint $\pi^\circ \in \arg\min_{\pi:\Omega\to[0,1]} \Phi(\pi; P_1, P_2)$ holds if and only if*

$$\sup_{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{B}_+(\Omega)} \int_\Omega [\boldsymbol{\alpha}_2(\omega)\mathbb{1}_{\Omega_2^\circ}(\omega) - \boldsymbol{\alpha}_1(\omega)\mathbb{1}_{\Omega_1^\circ}(\omega)](dP_1 - dP_2)(\omega) = 0. \tag{C.8}$$

Proof. We first prove the necessity. Suppose $\pi^\circ \in \arg\min_\pi \Phi(\pi; P_1, P_2)$. Then by definition for all randomized test π, we have

$$\Phi(\pi; P_1, P_2) \geq \Phi(\pi^\circ; P_1, P_2).$$

For any $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{B}_+(\Omega)$, there exists a small enough $\epsilon > 0$ such that the following perturbed π° is still a randomized test:

$$\pi^\circ(\omega) + \epsilon[\boldsymbol{\alpha}_2(\omega)\mathbb{1}_{\Omega_2^\circ}(\omega) - \boldsymbol{\alpha}_1(\omega)\mathbb{1}_{\Omega_1^\circ}(\omega)] = \begin{cases} 1 - \epsilon\boldsymbol{\alpha}_1(\omega), & \omega \in \Omega_1^\circ, \\ \epsilon\boldsymbol{\alpha}_2(\omega), & \omega \in \Omega_2^\circ, \end{cases}$$

which means that the probability of accepting hypothesis H_0 is reduced on Ω_1°, and probability of accepting hypothesis H_0 is increased on Ω_2°. Recall $\boldsymbol{\alpha} = \boldsymbol{\alpha}_2\mathbb{1}_{\Omega_2^\circ} - \boldsymbol{\alpha}_1\mathbb{1}_{\Omega_1^\circ}$. The optimality of π° implies that

$$\mathbb{E}_{P_1}[1 - \pi^\circ(\omega) - \epsilon\boldsymbol{\alpha}(\omega)] + \mathbb{E}_{P_2}[\pi^\circ(\omega) + \epsilon\boldsymbol{\alpha}(\omega)] \geq \mathbb{E}_{P_1}[1 - \pi^\circ(\omega)] + \mathbb{E}_{P_2}[\pi^\circ(\omega)].$$

Dividing ϵ on both sides gives $\mathbb{E}_{P_1}[\boldsymbol{\alpha}(\omega)] - \mathbb{E}_{P_2}[\boldsymbol{\alpha}(\omega)] \leq 0$. Moreover, the equality in (C.8) holds by taking $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 \equiv 0$, which proves (C.8).

Next, we prove the sufficiency. Suppose (C.8) holds. For any randomized test π, set $\tilde{\boldsymbol{\alpha}} := \pi - \pi^\circ$. Pick $\tilde{\boldsymbol{\alpha}}_1, \tilde{\boldsymbol{\alpha}}_2 \in \mathcal{B}_+(\Omega)$ such that

$$\tilde{\boldsymbol{\alpha}}_1(\omega) = \begin{cases} 1 - \pi(\omega) & \text{if } \omega \in \Omega_1^\circ, \\ 0 & \text{otherwise;} \end{cases} \qquad \tilde{\boldsymbol{\alpha}}_2(\omega) = \begin{cases} \pi(\omega) & \text{if } \omega \in \Omega_2^\circ, \\ 0 & \text{otherwise.} \end{cases}$$

Then by the definition of π°, we have $\tilde{\boldsymbol{\alpha}}(\omega) = \tilde{\boldsymbol{\alpha}}_2(\omega)\mathbb{1}_{\Omega_2^\circ}(\omega) - \tilde{\boldsymbol{\alpha}}_1(\omega)\mathbb{1}_{\Omega_1^\circ}(\omega)$ for all $\omega \in \Omega$. It follows that $\mathbb{E}_{P_1}[\tilde{\boldsymbol{\alpha}}(\omega)] - \mathbb{E}_{P_2}[\tilde{\boldsymbol{\alpha}}(\omega)] \leq 0$, and consequently,

$$\begin{aligned} &\mathbb{E}_{P_1}[1 - \pi(\omega)] + \mathbb{E}_{P_2}[\pi(\omega)] \\ =&\mathbb{E}_{P_1}[1 - \pi^\circ(\omega) - \tilde{\boldsymbol{\alpha}}(\omega)] + \mathbb{E}_{P_2}[\pi^\circ(\omega) + \tilde{\boldsymbol{\alpha}}(\omega)] \\ =&\mathbb{E}_{P_1}[1 - \pi^\circ(\omega)] + \mathbb{E}_{P_2}[\pi^\circ(\omega)] - (\mathbb{E}_{P_1}[\tilde{\boldsymbol{\alpha}}(\omega)] - \mathbb{E}_{P_2}[\tilde{\boldsymbol{\alpha}}(\omega)]) \\ \geq&\mathbb{E}_{P_1}[1 - \pi^\circ(\omega)] + \mathbb{E}_{P_2}[\pi^\circ(\omega)]. \end{aligned}$$

This indicates that the risk of any test π is greater than or equal to the risk of π°, implying

$\pi^\circ \in \arg\min_\pi \Phi(\pi; P_1, P_2)$. Therefore we have completed the proof. $\square$

Let us proceed by defining the Lagrangian function

$$\begin{aligned} &L(P_1, P_2; \lambda_1, \lambda_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) \\ &:= \sum_{k=1}^{2} \lambda_k \mathsf{W}(P_k, Q_{k,n_k}) + \sum_{k=1}^{2} \sum_{j \neq k} \left\{ \mathbb{E}_{P_k}[\boldsymbol{\alpha}_j(\omega) \mathbb{1}_{\Omega_j^\circ}(\omega) - \boldsymbol{\alpha}_k(\omega) \mathbb{1}_{\Omega_k^\circ}(\omega)] \right\}, \end{aligned} \tag{C.9}$$

where the second term is equivalent to $\int_\Omega [\boldsymbol{\alpha}_2(\omega) \mathbb{1}_{\Omega_2^\circ}(\omega) - \boldsymbol{\alpha}_1(\omega) \mathbb{1}_{\Omega_1^\circ}(\omega)](dP_1 - dP_2)(\omega)$. Using Lemma C.1, if $\pi^\circ \notin \arg\min_\pi \Phi(\pi; P_1, P_2)$, then there exists functions $\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2' \in \mathcal{B}_+(\Omega)$ such that $\sum_{k=1}^2 \sum_{j \neq k} \mathbb{E}_{P_k}[\boldsymbol{\alpha}_j'(\omega) \mathbb{1}_{\Omega_j^\circ}(\omega) - \boldsymbol{\alpha}_k'(\omega) \mathbb{1}_{\Omega_k^\circ}(\omega)] > 0$, whence

$$\sup_{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{B}_+(\Omega)} L(P_1, P_2; \lambda_1, \lambda_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) \geq \lim_{t \to \infty} L(P_1, P_2; \lambda_1, \lambda_2, t\boldsymbol{\alpha}_1', t\boldsymbol{\alpha}_2') = +\infty.$$

Therefore, we arrive at an equivalent formulation for the profile function F_{n_1,n_2} defined in (5.12):

$$\mathsf{F}_{n_1,n_2} = \inf_{P_1, P_2 \in \mathscr{P}(\Omega)} \sup_{\substack{\lambda_1, \lambda_2 \geq 0 \\ \lambda_1 + \lambda_2 \leq 1 \\ \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{B}_+(\Omega)}} L(P_1, P_2; \lambda_1, \lambda_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2). \tag{C.10}$$

In what follows, we prove the strong duality (i.e. exchanging of $\sup$ and $\inf$) in five steps. We start by showing the weak duality and simplify the dual formulation of F_{n_1,n_2}. Next, we show that it suffices to restrict the feasible region of $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ from $\mathcal{B}_+(\Omega)$ to $\mathcal{B}_+(\Omega) \cap \mathrm{Lip}(\Omega)$, which eventually leads to the set $\mathcal{A}$ defined in (5.13), and prove the strong duality by assuming the support Ω is compact. Finally, we relax the compactness assumption.

Step 1. (Weak duality.) Exchanging $\inf$ and $\sup$ in Equation (C.10) yields

$$\mathsf{F}_{n_1,n_2} \geq \sup_{\substack{\lambda_1, \lambda_2 \geq 0 \\ \lambda_1 + \lambda_2 \leq 1 \\ \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{B}_+(\Omega)}} \inf_{P_1, P_2 \in \mathscr{P}(\Omega)} L(P_1, P_2; \lambda_1, \lambda_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2). \tag{C.11}$$

Let us simplify the right-hand side by deriving a closed-form solution to the inner $\inf$

problem. Recall that $\Gamma(P, Q)$ denotes the collection of all Borel probability measures on $\Omega \times \Omega$ with marginal distributions P and Q. By the definition of Wasserstein metric, since the empirical distribution Q_{k,n_k} is supported on a finite set $\widehat{\Omega}_k = \{\widehat{\omega}_k^1, \ldots, \widehat{\omega}_k^{n_k}\}$ for $k = 1, 2$, we have

$$\lambda_k \mathsf{W}(P_k, Q_{k,n_k}) = \inf_{\gamma_k \in \Gamma(P_k, Q_{k,n_k})} \left\{ \sum_{i=1}^{n_k} \int_\Omega \lambda_k c(\omega, \widehat{\omega}_k^i) d\gamma_k(\omega, \widehat{\omega}_k^i) \right\}.$$

Moreover, for any distribution $\gamma_k \in \Gamma(P_k, Q_{k,n_k})$, $k = 1, 2$, we have

$$\begin{aligned} &\sum_{j \neq k} \left\{ \mathbb{E}_{P_k}[\boldsymbol{\alpha}_j(\omega) \mathbb{1}_{\Omega_j^\circ}(\omega) - \boldsymbol{\alpha}_k(\omega) \mathbb{1}_{\Omega_k^\circ}(\omega)] \right\} \\ =& \sum_{i=1}^{n_k} \int_\Omega \sum_{j \neq k} [\boldsymbol{\alpha}_j(\omega) \mathbb{1}_{\Omega_j^\circ}(\omega) - \boldsymbol{\alpha}_k(\omega) \mathbb{1}_{\Omega_k^\circ}(\omega)] d\gamma_k(\omega, \widehat{\omega}_k^i). \end{aligned}$$

Substituting the above equations to (C.9), it follows that:

$$\begin{aligned} &L(P_1, P_2; \lambda_1, \lambda_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) \\ =& \sum_{k=1}^{2} \inf_{\gamma_k \in \Gamma(P_k, Q_{k,n_k})} \Big\{ \sum_{i=1}^{n_k} \int_\Omega \big[\lambda_k c(\omega, \widehat{\omega}_k^i) \\ &\qquad + \sum_{j \neq k} (\boldsymbol{\alpha}_j(\omega) \mathbb{1}_{\Omega_j^\circ}(\omega) - \boldsymbol{\alpha}_k(\omega) \mathbb{1}_{\Omega_k^\circ}(\omega)) \big] d\gamma_k(\omega, \widehat{\omega}_k^i) \Big\}. \end{aligned}$$

Thereby for fixed λ_1, λ_2, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\inf_{P_1, P_2} L(P_1, P_2; \lambda_1, \lambda_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ can be expressed equivalently as a minimization problem over γ_k, whose first marginal distribution can be arbitrary

and second marginal is the empirical distribution Q_{k,n_k}, $k = 1, 2$:

$$\begin{aligned}
&\inf_{P_1,P_2} L(P_1, P_2; \lambda_1, \lambda_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) \\
=&\sum_{k=1}^{2} \inf_{\gamma_k \in \Gamma(\cdot, Q_{k,n_k})} \Big\{ \sum_{i=1}^{n_k} \int_{\Omega} \big[\lambda_k c(\omega, \widehat{\omega}_k^i) \\
&\qquad + \sum_{j \neq k} (\boldsymbol{\alpha}_j(\omega) \mathbb{1}_{\Omega_j^\circ}(\omega) - \boldsymbol{\alpha}_k(\omega) \mathbb{1}_{\Omega_k^\circ}(\omega))\big] d\gamma_k(\omega, \widehat{\omega}_k^i) \Big\} \\
=&\sum_{k=1}^{2} \frac{1}{n_k} \sum_{i=1}^{n_k} \inf_{\omega \in \Omega} \big\{ \lambda_k c(\omega, \widehat{\omega}_k^i) + \sum_{j \neq k} (\boldsymbol{\alpha}_j(\omega) \mathbb{1}_{\Omega_j^\circ}(\omega) - \boldsymbol{\alpha}_k(\omega) \mathbb{1}_{\Omega_k^\circ}(\omega)) \big\},
\end{aligned}$$

where $\Gamma(\cdot, Q_{k,n_k})$ denotes the collection of all Borel probability measures on $\Omega \times \Omega$ with second marginal being Q_{k,n_k}, and the last equality is attained by picking

$$\gamma_k(\omega_k^i, \widehat{\omega}_k^i) = \frac{1}{n_k}, \quad i = 1, \ldots, n_k, \ k = 1, 2,$$

where

$$\omega_k^i \in \arg\min_{\omega \in \Omega} \Big\{ \lambda_k c(\omega, \widehat{\omega}_k^i) + \sum_{j \neq k} (\boldsymbol{\alpha}_j(\omega) \mathbb{1}_{\Omega_j^\circ}(\omega) - \boldsymbol{\alpha}_k(\omega) \mathbb{1}_{\Omega_k^\circ}(\omega)) \Big\}.$$

If the minimizer does not exist, we can argue similarly using a sequence of approximate minimizers. If there are multiple minimizers, we can simply choose one of them or distribute the probability mass $1/n_k$ uniformly on the optimal solution set. Therefore, we have the right-hand side of (C.11) equals to

$$\sup_{\substack{\lambda_1, \lambda_2 \geq 0 \\ \lambda_1 + \lambda_2 \leq 1 \\ \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{B}_+(\Omega)}} \sum_{k=1}^{2} \mathbb{E}_{\widehat{\omega}_k \sim Q_{k,n_k}} \Big[\inf_{\omega \in \Omega} \Big\{ \lambda_k c(\omega, \widehat{\omega}_k) + \sum_{j \neq k} [\boldsymbol{\alpha}_j(\omega) \mathbb{1}_{\Omega_j^\circ}(\omega) - \boldsymbol{\alpha}_k(\omega) \mathbb{1}_{\Omega_k^\circ}(\omega)] \Big\} \Big]. \tag{C.12}$$

In the sequel, we will refer to the right-hand side of (C.12) as the dual problem.

Step 2. (Restricting on the subset $\mathcal{A}$ as defined in (5.13).) We first prove that we can restrict $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ on the space of Lipschitz continuous functions without affecting the optimal value.

For any feasible solution $(\lambda_1, \lambda_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ of the dual problem in (C.12) such that the dual objective is finite, let us construct a modification $(\lambda_1, \lambda_2, \tilde{\boldsymbol{\alpha}}_1, \tilde{\boldsymbol{\alpha}}_2)$ which yields an objective value no worse than $(\lambda_1, \lambda_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$, but enjoys a nicer continuity property. For $i = 1, 2, \ldots, n_1$, set

$$\begin{aligned}\phi_1(\widehat{\omega}_1^i) :&= \inf_{\omega\in\Omega}\{\lambda_1 c(\omega, \widehat{\omega}_1^i) + \boldsymbol{\alpha}_2(\omega)\mathbb{1}_{\Omega_2^\circ}(\omega) - \boldsymbol{\alpha}_1(\omega)\mathbb{1}_{\Omega_1^\circ}(\omega)\} \\ &= \min\Big\{\inf_{\omega\in\Omega_1^\circ}\{\lambda_1 c(\omega, \widehat{\omega}_1^i) - \boldsymbol{\alpha}_1(\omega)\}, \inf_{\omega\in\Omega_2^\circ}\{\lambda_1 c(\omega, \widehat{\omega}_1^i) + \boldsymbol{\alpha}_2(\omega)\}\Big\}.\end{aligned}$$

It follows that

$$\boldsymbol{\alpha}_1(\omega) \le \lambda_1 c(\omega, \widehat{\omega}_1^i) - \phi_1(\widehat{\omega}_1^i), \quad \forall\omega \in \Omega_1^\circ,\ \forall i = 1, \ldots, n_1. \tag{C.13}$$

Define another function $\tilde{\boldsymbol{\alpha}}_1$ as

$$\tilde{\boldsymbol{\alpha}}_1(\omega) = \min_{i=1,\ldots,n_1}\left\{\lambda_1 c(\omega, \widehat{\omega}_1^i) - \phi_1(\widehat{\omega}_1^i)\right\}, \quad \forall\omega \in \Omega_1^\circ. \tag{C.14}$$

This yields $\boldsymbol{\alpha}_1(\omega) \le \tilde{\boldsymbol{\alpha}}_1(\omega)$ for all $\omega \in \Omega_1^\circ$, due to (C.13). Moreover, the objective value in (C.12) associated with $(\lambda_1, \lambda_2, \tilde{\boldsymbol{\alpha}}_1, \boldsymbol{\alpha}_2)$ is no less than the value associated with $(\lambda_1, \lambda_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ since

$$\begin{aligned}\phi_1(\widehat{\omega}_1^i) &\le \lambda_1 c(\omega, \widehat{\omega}_1^i) - \tilde{\boldsymbol{\alpha}}_1(\omega), \quad \forall\omega \in \Omega_1^\circ,\ \forall i = 1, \ldots, n_1, \\ \lambda_2 c(\omega, \widehat{\omega}_2^j) + \boldsymbol{\alpha}_1(\omega) &\le \lambda_2 c(\omega, \widehat{\omega}_2^j) + \tilde{\boldsymbol{\alpha}}_1(\omega), \quad \forall\omega \in \Omega_1^\circ,\ \forall j = i, \ldots, n_2.\end{aligned}$$

Furthermore, the function $\tilde{\boldsymbol{\alpha}}_1$ defined in this way is Lipschitz with constant λ_1. Indeed, for any two points $\xi, \eta \in \Omega_1^\circ$, let i_1 and i_2 be the indices at which the minimum are attained in

the definition (C.14) for ξ and η, respectively. We have

$$\begin{aligned}\tilde{\boldsymbol{\alpha}}_1(\xi) - \tilde{\boldsymbol{\alpha}}_1(\eta) &= [\lambda_1 c(\xi, \widehat{\omega}_1^{i_1}) - \phi_1(\widehat{\omega}_1^{i_1})] - [\lambda_1 c(\eta, \widehat{\omega}_1^{i_2}) - \phi_1(\widehat{\omega}_1^{i_2})] \\ &\leq [\lambda_1 c(\xi, \widehat{\omega}_1^{i_2}) - \phi_1(\widehat{\omega}_1^{i_2})] - [\lambda_1 c(\eta, \widehat{\omega}_1^{i_2}) - \phi_1(\widehat{\omega}_1^{i_2})] \\ &= \lambda_1 [c(\xi, \widehat{\omega}_1^{i_2}) - c(\eta, \widehat{\omega}_1^{i_2})] \\ &\leq \lambda_1 c(\xi, \eta),\end{aligned}$$

where the last inequality is due to the triangle inequality of the metric $c(\cdot, \cdot)$; and the same inequality holds for $\tilde{\boldsymbol{\alpha}}_1(\eta) - \tilde{\boldsymbol{\alpha}}_1(\xi)$. In a similar fashion, for $j = 1, 2, \ldots, n_2$, define

$$\begin{aligned}\phi_2(\widehat{\omega}_2^j) :&= \inf_{\omega} \{\lambda_2 c(\omega, \widehat{\omega}_2^j) + \tilde{\boldsymbol{\alpha}}_1(\omega) \mathbb{1}_{\Omega_1^\circ}(\omega) - \boldsymbol{\alpha}_2(\omega) \mathbb{1}_{\Omega_2^\circ}(\omega)\} \\ &= \min \Big\{ \inf_{\omega \in \Omega_1^\circ} \{\lambda_2 c(\omega, \widehat{\omega}_2^j) + \tilde{\boldsymbol{\alpha}}_1(\omega)\}, \inf_{\omega \in \Omega_2^\circ} \{\lambda_2 c(\omega, \widehat{\omega}_2^j) - \boldsymbol{\alpha}_2(\omega)\} \Big\},\end{aligned}$$

and set

$$\tilde{\boldsymbol{\alpha}}_2(\omega) := \min_{j=1,\ldots,n_2} \left\{ \lambda_2 c(\omega, \widehat{\omega}_2^j) - \phi_2(\widehat{\omega}_2^j) \right\}, \quad \forall \omega \in \Omega_2^\circ. \tag{C.15}$$

Then $\boldsymbol{\alpha}_2(\omega) \leq \tilde{\boldsymbol{\alpha}}_2(\omega)$ for all $\omega \in \Omega_2^\circ$ and the objective value associated with $(\lambda_1, \lambda_2, \tilde{\boldsymbol{\alpha}}_1, \tilde{\boldsymbol{\alpha}}_2)$ is no less than the objective value associated with $(\lambda_1, \lambda_2, \tilde{\boldsymbol{\alpha}}_1, \boldsymbol{\alpha}_2)$; and $\tilde{\boldsymbol{\alpha}}_2$ is Lipschitz with constant λ_2. Since we are in the region $\{\lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 \leq 1\}$, the argument above proves that without loss of generality we can restrict $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ on the set of 1-Lipschitz continuous functions.

Observe that the objective value does not change if we shift $\boldsymbol{\alpha}_k$ by any constant C_k, $k = 1, 2$. Hence, without loss of generality, we can only consider those satisfying $\boldsymbol{\alpha}_k(\omega_k^\circ) = 0$ without affecting the optimal value, where $\omega_k^\circ \in \Omega_k^\circ$, $k = 1, 2$. By the above argument, we have shown that it suffices to restrict the feasible region on $\mathcal{A}$.

Step 3. (Strong duality for compact space.) Now assume Ω is compact. We aim to prove the strong duality by applying Sion's minimax theorem. Observe that $L(P_1, P_2; \lambda_1, \lambda_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ defined in (C.9) is convex in P_k, linear in λ_k and $\boldsymbol{\alpha}_k$; by Prokhorov's theorem [152], the

convex space $\mathcal{P}(\Omega) \times \mathcal{P}(\Omega)$ is compact since Ω is relatively compact with respect to the weak topology; the space $\{\lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 \leq 1\}$ is also a convex compact space. The feasible region of $\boldsymbol{\alpha}_k, k = 1, 2$ belongs to a linear topological space under the sup-norm. This justifies the conditions for Sion's minimax theorem, thereby we can exchange sup and inf in (C.9) when Ω is compact.

Step 4. (Relaxing the compactness assumption when the cost is bounded.) We now relax the compactness assumption made in the previous step, using a technique similar to the proof of Theorem 1.3 in [200]. We temporarily assume the cost function $c(\cdot, \cdot)$ is bounded by a positive constant C and is uniformly continuous. We will relax the bounded assumption later. We already have the weak duality:

$$
\begin{aligned}
v_1 &:= \inf_{P_1, P_2 \in \mathscr{P}(\Omega)} \sup_{\substack{\lambda_1, \lambda_2 \geq 0 \\ \lambda_1 + \lambda_2 \leq 1 \\ \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{B}_+(\Omega)}} L(P_1, P_2; \lambda_1, \lambda_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) \\
&\geq \sup_{\substack{\lambda_1, \lambda_2 \geq 0 \\ \lambda_1 + \lambda_2 \leq 1 \\ \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{B}_+(\Omega)}} \sum_{k=1}^{2} \mathbb{E}_{\widehat{\omega}_k \sim Q_{k,n_k}} \Big[\inf_{\omega \in \Omega} \Big\{ \lambda_k c(\omega, \widehat{\omega}_k) + \sum_{j \neq k} [\boldsymbol{\alpha}_j(\omega) \mathbb{1}_{\Omega_j^\circ}(\omega) - \boldsymbol{\alpha}_k(\omega) \mathbb{1}_{\Omega_k^\circ}(\omega)] \Big\} \Big] \\
&=: v_2.
\end{aligned}
$$

In the following we show that $v_1 \leq v_2$.

For any $\epsilon > 0$, let $\Omega^\epsilon \subset \Omega$ be a compact subset sufficiently large, such that $P_k^\circ(\Omega \backslash \Omega^\epsilon) \leq \epsilon$ and $Q_{k,n_k}(\Omega^\epsilon) = 1$, $k = 1, 2$. This is always possible since Q_{k,n_k} is the empirical distribution and with finite support. Then the previous steps imply that the strong duality

holds on Ω^ϵ:

$$
\begin{aligned}
v_1^\epsilon &:= \inf_{P_1,P_2\in\mathscr{P}(\Omega^\epsilon)} \sup_{\substack{\lambda_1,\lambda_2\geq 0\\ \lambda_1+\lambda_2\leq 1\\ \boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2\in\mathcal{B}_+(\Omega^\epsilon)}} L(P_1,P_2;\lambda_1,\lambda_2,\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2)\\
&= \sup_{\substack{\lambda_1,\lambda_2\geq 0\\ \lambda_1+\lambda_2\leq 1\\ \boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2\in\mathcal{B}_+(\Omega^\epsilon)}} \sum_{k=1}^{2} \mathbb{E}_{\widehat{\omega}_k\sim Q_{k,n_k}}\Big[\inf_{\omega\in\Omega^\epsilon}\Big\{\lambda_k c(\omega,\widehat{\omega}_k)+\sum_{j\neq k}[\boldsymbol{\alpha}_j(\omega)\mathbb{1}_{\Omega_j^\circ}(\omega)-\boldsymbol{\alpha}_k(\omega)\mathbb{1}_{\Omega_k^\circ}(\omega)]\Big\}\Big]\\
&=: v_2^\epsilon.
\end{aligned}
$$

Consider the $\inf\sup$ problem defining v_1. For the optimal solution $(P_1^\epsilon, P_2^\epsilon)$ to the $\inf\sup$ problem that induces v_1^ϵ, we define distributions $\tilde{P}_1, \tilde{P}_2$ via

$$
\tilde{P}_k(A) = P_k^\circ(\Omega^\epsilon)\cdot P_k^\epsilon(A\cap\Omega^\epsilon) + P_k^\circ(A\cap(\Omega\setminus\Omega^\epsilon)), \quad \forall \text{ Borel set } A\subset\Omega.
$$

Recall $\boldsymbol{\alpha} = \boldsymbol{\alpha}_2\mathbb{1}_{\Omega_2^\circ} - \boldsymbol{\alpha}_1\mathbb{1}_{\Omega_1^\circ}$. We compare the Lagrangian function L defined in (C.9) associated with $(\tilde{P}_1,\tilde{P}_2)$ and $(P_1^\epsilon,P_2^\epsilon)$. For the first term in (C.9), we have that

$$
\mathsf{W}(\tilde{P}_k, Q_{k,n_k}) \leq P_k^\circ(\Omega^\epsilon)\mathsf{W}(P_k^\epsilon, Q_{k,n_k}) + CP_k^\circ(\Omega\setminus\Omega^\epsilon) \leq \mathsf{W}(P_k^\epsilon, Q_{k,n_k}) + C\epsilon.
$$

For the second term in (C.9), we have

$$
\begin{aligned}
&\int_\Omega \boldsymbol{\alpha}(\omega)\,(\tilde{P}_1-\tilde{P}_2)(d\omega)\\
=&\int_{\Omega^\epsilon}\boldsymbol{\alpha}(\omega)\,(P_1^\circ(\Omega^\epsilon)P_1^\epsilon - P_2^\circ(\Omega^\epsilon)P_2^\epsilon)(d\omega) + \int_{\Omega\setminus\Omega^\epsilon}\boldsymbol{\alpha}(\omega)\,(P_1^\circ-P_2^\circ)(d\omega).
\end{aligned}
$$

By definition of $\Omega_1^\circ, \Omega_2^\circ$, we have $\int_{\Omega\setminus\Omega^\epsilon} \boldsymbol{\alpha}(\omega)(P_1^\circ - P_2^\circ)(d\omega) \leq 0$. Moreover,

$$\begin{aligned}
&\int_{\Omega^\epsilon} \boldsymbol{\alpha}(\omega)\left(P_1^\circ(\Omega^\epsilon)P_1^\epsilon - P_2^\circ(\Omega^\epsilon)P_2^\epsilon\right)(d\omega)\\
&= \begin{cases} P_1^\circ(\Omega^\epsilon)\int_{\Omega^\epsilon} \boldsymbol{\alpha}(\omega)(P_1^\epsilon - P_2^\epsilon)(d\omega) - (P_2^\circ(\Omega^\epsilon) - P_1^\circ(\Omega^\epsilon))\int_{\Omega^\epsilon} \boldsymbol{\alpha}(\omega)P_2^\epsilon(d\omega), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } P_1^\circ(\Omega^\epsilon) \leq P_2^\circ(\Omega^\epsilon); \\ P_2^\circ(\Omega^\epsilon)\int_{\Omega^\epsilon} \boldsymbol{\alpha}(\omega)(P_1^\epsilon - P_2^\epsilon)(d\omega) + (P_1^\circ(\Omega^\epsilon) - P_2^\circ(\Omega^\epsilon))\int_{\Omega^\epsilon} \boldsymbol{\alpha}(\omega)P_1^\epsilon(d\omega), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } P_1^\circ(\Omega^\epsilon) > P_2^\circ(\Omega^\epsilon). \end{cases}
\end{aligned}$$

By definition $\int_{\Omega^\epsilon} \boldsymbol{\alpha}(\omega)(P_1^\epsilon - P_2^\epsilon)(d\omega) \leq 0$ and $P_k^\circ(\Omega^\epsilon) \geq 1 - \epsilon$, thereby

$$P_k^\circ(\Omega^\epsilon)\int_{\Omega^\epsilon} \boldsymbol{\alpha}(\omega)(P_1^\epsilon - P_2^\epsilon)(d\omega) \leq (1-\epsilon)\int_{\Omega^\epsilon} \boldsymbol{\alpha}(\omega)(dP_1^\epsilon - dP_2^\epsilon)(\omega) \leq 0.$$

Moreover, since $P_k^\circ(\Omega^\epsilon) \geq 1 - \epsilon$, we have $|P_1^\circ(\Omega^\epsilon) - P_2^\circ(\Omega^\epsilon)| \leq \epsilon$, consequently we have

$$|P_1^\circ(\Omega^\epsilon) - P_2^\circ(\Omega^\epsilon)|\int_{\Omega^\epsilon} \boldsymbol{\alpha}(\omega)dP_k^\epsilon(\omega) \leq \epsilon\int_{\Omega} c(\omega,\omega_k^0)dP_k^\epsilon(\omega) \leq C\epsilon,$$

where the last inequality is due to the 1-Lipschitz property of $\boldsymbol{\alpha}_k$ and C may be a different constant. Combining with previous inequality that $\mathsf{W}(\tilde{P}_k, Q_{k,n_k}) \leq \mathsf{W}(P_k^\epsilon, Q_{k,n_k}) + C\epsilon$, $k = 1, 2$, we have

$$v_1 \leq v_1^\epsilon + 2C\epsilon.$$

Now consider the dual problem defining v_2. Let $(\boldsymbol{\alpha}_1^\epsilon, \boldsymbol{\alpha}_2^\epsilon)$ be the optimal solution to the dual problem supported on the subset Ω^ϵ. We will construct an approximate maximizer $(\tilde{\boldsymbol{\alpha}}_1, \tilde{\boldsymbol{\alpha}}_2)$ of the original dual problem from $(\boldsymbol{\alpha}_1^\epsilon, \boldsymbol{\alpha}_2^\epsilon)$. To this end, let us define

$$\begin{aligned}
\phi_1^\epsilon(\widehat{\omega}_1^i) &:= \min\Big\{\inf_{\omega\in\Omega_1^\circ\cap\Omega^\epsilon}\{\lambda_1 c(\omega,\widehat{\omega}_1^i) - \boldsymbol{\alpha}_1^\epsilon(\omega)\}, \inf_{\omega\in\Omega_2^\circ\cap\Omega^\epsilon}\{\lambda_1 c(\omega,\widehat{\omega}_1^i) + \boldsymbol{\alpha}_2^\epsilon(\omega)\}\Big\},\\
\phi_2^\epsilon(\widehat{\omega}_2^j) &:= \min\Big\{\inf_{\omega\in\Omega_1^\circ\cap\Omega^\epsilon}\{\lambda_2 c(\omega,\widehat{\omega}_2^j) + \boldsymbol{\alpha}_1^\epsilon(\omega)\}, \inf_{\omega\in\Omega_2^\circ\cap\Omega^\epsilon}\{\lambda_2 c(\omega,\widehat{\omega}_2^j) - \boldsymbol{\alpha}_2^\epsilon(\omega)\}\Big\}.
\end{aligned}$$

From the above equations we have that $\boldsymbol{\alpha}_1^\epsilon, \boldsymbol{\alpha}_2^\epsilon$ satisfy:

$$\begin{aligned}
\boldsymbol{\alpha}_1^\epsilon(\omega) &\leq \lambda_1 c(\omega, \widehat{\omega}_1^i) - \phi_1^\epsilon(\widehat{\omega}_1^i), \quad \forall \omega \in \Omega^\epsilon,\ i = 1, \ldots, n_1, \\
\boldsymbol{\alpha}_2^\epsilon(\omega) &\leq \lambda_2 c(\omega, \widehat{\omega}_2^j) - \phi_2^\epsilon(\widehat{\omega}_2^j), \quad \forall \omega \in \Omega^\epsilon,\ j = 1, \ldots, n_2.
\end{aligned} \tag{C.16}$$

Define $\tilde{\boldsymbol{\alpha}}_1, \tilde{\boldsymbol{\alpha}}_2$ as

$$\begin{aligned}
\tilde{\boldsymbol{\alpha}}_1(\omega) &= \min_{1 \leq i \leq n_1} \{\lambda_1 c(\omega, \widehat{\omega}_1^i) - \phi_1^\epsilon(\widehat{\omega}_1^i)\}, \\
\tilde{\boldsymbol{\alpha}}_2(\omega) &= \min_{1 \leq j \leq n_2} \{\lambda_2 c(\omega, \widehat{\omega}_2^j) - \phi_2^\epsilon(\widehat{\omega}_2^j)\}.
\end{aligned} \tag{C.17}$$

This implies that $\phi_k^\epsilon(\widehat{\omega}_k^i) \leq \inf_{\omega \in \Omega_k^\circ} \{\lambda_k c(\omega, \widehat{\omega}_k^i) - \tilde{\boldsymbol{\alpha}}_k(\omega)\}$, $k = 1, 2$, $i = 1, \ldots, n_k$. Comparing (C.17) and (C.16), we have that $\tilde{\boldsymbol{\alpha}}_k(\omega) \geq \boldsymbol{\alpha}_k^\epsilon(\omega)$, $k = 1, 2$, for $\omega \in \Omega^\epsilon$. Consequently, we have

$$\begin{aligned}
\phi_1^\epsilon(\widehat{\omega}_1^i) &\leq \min \Big\{ \inf_{\omega \in \Omega_1^\circ \cap \Omega^\epsilon} \{\lambda_1 c(\omega, \widehat{\omega}_1^i) - \tilde{\boldsymbol{\alpha}}_1(\omega)\}, \inf_{\omega \in \Omega_2^\circ \cap \Omega^\epsilon} \{\lambda_1 c(\omega, \widehat{\omega}_1^i) + \tilde{\boldsymbol{\alpha}}_2(\omega)\} \Big\}, \\
\phi_2^\epsilon(\widehat{\omega}_2^j) &\leq \inf \Big\{ \min_{\omega \in \Omega_1^\circ \cap \Omega^\epsilon} \{\lambda_2 c(\omega, \widehat{\omega}_2^j) + \tilde{\boldsymbol{\alpha}}_1(\omega)\}, \inf_{\omega \in \Omega_2^\circ \cap \Omega^\epsilon} \{\lambda_2 c(\omega, \widehat{\omega}_2^j) - \tilde{\boldsymbol{\alpha}}_2(\omega)\} \Big\}.
\end{aligned}$$

Moreover, we can choose Ω^ϵ sufficiently large so that for every $\omega \in \Omega_2^\circ \cap (\Omega \setminus \Omega^\epsilon)$,

$$\begin{aligned}
\lambda_1 c(\omega, \widehat{\omega}_1^i) + \tilde{\boldsymbol{\alpha}}_2(\omega) &= \lambda_1 c(\omega, \widehat{\omega}_1^i) + \lambda_2 c(\omega, \widehat{\omega}_2^j) - \phi_2^\epsilon(\widehat{\omega}_2^j) \\
&\geq \inf_{\omega \in \Omega_2^\circ \cap \Omega^\epsilon} \{\lambda_1 c(\omega, \widehat{\omega}_1^i) + \boldsymbol{\alpha}_2^\epsilon(\omega)\} \geq \phi_1^\epsilon(\widehat{\omega}_1^i),
\end{aligned}$$

where j is the minimizer in the definition (C.17). Combining these together, we have

$$\begin{aligned}
\phi_1^\epsilon(\widehat{\omega}_1^i) &\leq \min \Big\{ \inf_{\omega \in \Omega_1^\circ} \{\lambda_1 c(\omega, \widehat{\omega}_1^i) - \tilde{\boldsymbol{\alpha}}_1(\omega)\}, \inf_{\omega \in \Omega_2^\circ} \{\lambda_1 c(\omega, \widehat{\omega}_1^i) + \tilde{\boldsymbol{\alpha}}_2(\omega)\} \Big\}, \\
\phi_2^\epsilon(\widehat{\omega}_2^j) &\leq \min \Big\{ \inf_{\omega \in \Omega_1^\circ} \{\lambda_2 c(\omega, \widehat{\omega}_2^j) + \tilde{\boldsymbol{\alpha}}_1(\omega)\}, \inf_{\omega \in \Omega_2^\circ} \{\lambda_2 c(\omega, \widehat{\omega}_2^j) - \tilde{\boldsymbol{\alpha}}_2(\omega)\} \Big\}.
\end{aligned}$$

Therefore, from $\tilde{\boldsymbol{\alpha}}_1, \tilde{\boldsymbol{\alpha}}_2$ defined in (C.17), we see $v_2 \geq v_2^\epsilon$. Combine with previous argument, we have

$$v_2^\epsilon \leq v_2 \leq v_1 \leq v_1^\epsilon + 2C\epsilon.$$

By letting $\epsilon \to 0$, we have shown the strong duality, provided that the cost function is bounded.

Step 5. (Relaxing the bounded cost assumption.) Next, we turn to the general case with cost function by writing $c := \sup_m c_m$, where $c_m(x,y) = \min\{c(x,y), m\}$ is the truncated cost function that are bounded for each $m \in \mathbb{N}$. Let v_1^m be the optimal value of the primal problem under cost c_m, and v_2^m denote the optimal value of the dual problem under cost c_m. More specifically, let

$$\begin{aligned} &L^m(P_1, P_2; \lambda_1, \lambda_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) \\ :=& \sum_{k=1}^{2} \lambda_k \mathsf{W}^m(P_k, Q_{k,n_k}) + \sum_{k=1}^{2} \sum_{j \neq k} \left\{ \mathbb{E}_{P_k}[\boldsymbol{\alpha}_j(\omega) \mathbb{1}_{\Omega_j^\circ}(\omega) - \boldsymbol{\alpha}_k(\omega) \mathbb{1}_{\Omega_k^\circ}(\omega)] \right\}, \end{aligned}$$

where $\mathsf{W}^m(P_k, Q_{k,n_k})$ is the Wasserstein distance associated with cost function $c_m(\cdot,\cdot)$. Define

$$\begin{aligned} v_1^m :=& \inf_{P_1, P_2 \in \mathscr{P}(\Omega)} \sup_{\substack{\lambda_1, \lambda_2 \geq 0 \\ \lambda_1 + \lambda_2 \leq 1 \\ \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{B}_+(\Omega)}} L^m(P_1, P_2; \lambda_1, \lambda_2, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) \\ =& \sup_{\substack{\lambda_1, \lambda_2 \geq 0 \\ \lambda_1 + \lambda_2 \leq 1 \\ \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{B}_+(\Omega)}} \sum_{k=1}^{2} \mathbb{E}_{\widehat{\omega}_k \sim Q_{k,n_k}} \left[\inf_{\omega \in \Omega} \left\{ \lambda_k c_m(\omega, \widehat{\omega}_k) + \sum_{j \neq k} [\boldsymbol{\alpha}_j(\omega) \mathbb{1}_{\Omega_j^\circ}(\omega) - \boldsymbol{\alpha}_k(\omega) \mathbb{1}_{\Omega_k^\circ}(\omega)] \right\} \right] \\ =:& \, v_2^m. \end{aligned}$$

We have proved $v_1^m = v_2^m$ in previous steps. And clearly we have $v_2^m \leq v_2$ since $c_m \leq c$, leading to $v_1^m = v_2^m \leq v_2 \leq v_1$, so we only need to show $v_1 = \sup_m v_1^m$.

Observe that $\mathsf{W}^m(P_k, Q_{k,n_k})$ is a non-decreasing sequence bounded above by $\mathsf{W}(P_k, Q_{k,n_k})$. If $\{(P_{1,l}^m, P_{2,l}^m)\}_{l \in \mathbb{N}}$ is a minimizing sequence for the problem v_1^m, then we can extract a subsequence that converges weakly to some probability measure P_1^m, P_2^m [200].

We claim that the sequence $\{P_k^m\}_{m \in \mathbb{N}}$ is relatively compact with respect to the weak topology, $k = 1, 2$. To show this, suppose $\{P_k^m\}_{m \in \mathbb{N}}$ is not relatively compact, then there exists $\epsilon > 0$ such that for any compact set A and any $m_0 \in \mathbb{N}$, there exists $m >$

m_0 such that $P_k^m(A) \geq \epsilon$. We choose $m_0 = \lceil \mathsf{W}(Q_{k,n_k}, P_k^\circ)/\epsilon \rceil$ and a set A such that $\inf_{\omega \in A, \widehat{\omega} \in \widehat{\Omega}} c(\omega, \widehat{\omega}) \geq m_0$. Then for any $m > m_0$, we have

$$\begin{aligned}\mathsf{W}^m(Q_{k,n_k}, P_k^m) &= \min_{\gamma \in \Gamma(P_k^m, Q_{k,n_k})} \left\{ \mathbb{E}_{(\omega,\omega') \sim \gamma} \left[c_m(\omega, \omega') \right] \right\} \\ &> m_0 P_k^m(A) \geq m_0 \epsilon \geq \mathsf{W}(Q_{k,n_k}, P_k^\circ),\end{aligned}$$

while at the same time we have

$$\mathsf{W}^m(Q_{k,n_k}, P_1^m) \leq \mathsf{W}^m(Q_{k,n_k}, P_1^\circ) \leq \mathsf{W}(Q_{k,n_k}, P_1^\circ),$$

which is a contradiction. Therefore $\{P_k^m\}_{m \in \mathbb{N}}$ is relatively compact and we can extract a subsequence that converges to some probability measure P_k^*.

For any $m_1 > m_2$, we have $\mathsf{W}^{m_1}(P_k^{m_1}, Q_{k,n_k}) \geq \mathsf{W}^{m_2}(P_k^{m_1}, Q_{k,n_k})$, and

$$\limsup_{m_1 \to \infty} \mathsf{W}^{m_1}(P_k^{m_1}, Q_{k,n_k}) \geq \limsup_{m_1 \to \infty} \mathsf{W}^{m_2}(P_k^{m_1}, Q_{k,n_k}) \geq \mathsf{W}^{m_2}(P_k^*, Q_{k,n_k}).$$

Moreover, $\mathsf{W}^{m_2}(P_k^*, Q_{k,n_k})$ is a non-decreasing sequence and converges to $\mathsf{W}(P_k^*, Q_{k,n_k})$ as $m_2 \to \infty$, hence:

$$\limsup_{m \to \infty} v_1^m = \limsup_{m \to \infty} \mathsf{W}^m(P_k^m, Q_{k,n_k}) \geq \mathsf{W}(P_k^*, Q_{k,n_k}) = v_1.$$

Thereby we complete the proof. □

APPENDIX D

PROOFS FOR CHAPTER 6

Proof of Lemma 6.2. We start with describing an application of the Bernstein inequality for martingales (cf., e.g., [12, 64, 60, 21]) in our situation. Let $\omega_i, i = ..., 0, 1, 2, ...$ be a sequence of random binary vectors in $\mathbb{R}^m$ such that the conditional distribution of the j-th component ω_{ij}, $j = 1, ..., m$, of ω_i given ω^{i-1} is Bernoulli distribution with parameter $\mu_{ij} = \mathbb{E}_{|\omega^{i-1}}\{\omega_{ij}\}$. Now, consider the sequence of Boolean vectors γ_i, $i = 1, 2, ...$, $\gamma_i \in \mathbb{R}^m$, such that γ_i is $|\omega^{i-1}$-measurable with $\sum_j \gamma_i^j \leq 1$ a.s.. Finally, let $\zeta_i = \gamma_i^\top \omega_i - \gamma_i^\top \mu_i$; note that, in this case,

$$\mathbb{E}_{|\omega^{i-1}}\{\zeta_i\} = 0,\ \sigma_i^2 := \mathbb{E}_{|\omega^{i-1}}\{\zeta_i^2\} = \gamma_i^\top \mu_i(1 - \gamma_i^\top \mu_i) \leq \tfrac{1}{4},\ \text{and } |\zeta_i| \leq 1 \ \text{ a.s..}$$

Denote $\bar{\mu}_N = \frac{1}{N}\sum_{i=1}^N \gamma_i^\top \mu_i$, $\bar{\nu}_N = \frac{1}{N}\sum_{i=1}^N \gamma_i^\top \omega_i$, $\bar{s}_N = \frac{1}{N}\sum_{i=1}^N \sigma_i^2$, and $\bar{\zeta}_N = \frac{1}{N}\sum_{i=1}^N \zeta_i$.

Lemma D.1. *Let $0 < \underline{s} < \overline{s} < \infty$, and let $y > 1$. One has*

$$\mathbb{P}\left\{|\bar{\zeta}_N| \geq \sqrt{\frac{2y\bar{s}_N}{N}} + \frac{y}{3N},\ \underline{s} \leq \bar{s}_N \leq \overline{s}\right\} \leq 2e\big(y \ln\left(\overline{s}/\underline{s}\right) + 1\big)e^{-y}, \tag{D.1}$$

and, as a consequence,

$$\mathbb{P}\left\{|\bar{\zeta}_N| \geq \sqrt{\frac{2y\bar{s}_N}{N}} + \frac{y}{3N}\right\} \leq 2e\big(y\big[\ln((y-1)N\big) + 2\big] + 2\big)e^{-y}. \tag{D.2}$$

Moreover, we have

$$\mathbb{P}\left\{\underline{\psi}(\bar{\nu}_N, N; y) \leq \bar{\mu}_N \leq \overline{\psi}(\bar{\nu}_N, N; y)\right\} \geq 1 - 2e\big(y\big[\ln((y-1)N\big) + 2\big] + 2\big)e^{-y}, \tag{D.3}$$

where

$$\underline{\psi}(\nu,N;y)=\begin{cases}(N+2y)^{-1}\left[N\nu+\frac{2y}{3}-\sqrt{2N\nu y+\frac{y^2}{3}-\frac{2y}{N}\left(\frac{y}{3}-\nu N\right)^2}\right],\ \nu>\frac{y}{3N},\\ 0,\quad \textit{otherwise;}\end{cases}$$
$$\overline{\psi}(\nu,N;y)=\begin{cases}(N+2y)^{-1}\left[N\nu+\frac{4y}{3}+\sqrt{2N\nu y+\frac{5y^2}{3}-\frac{2y}{N}\left(\frac{y}{3}+\nu N\right)^2}\right],\ \nu<1-\frac{y}{3N},\\ 1,\quad \textit{otherwise,}\end{cases} \tag{D.4}$$

so that

$$\begin{aligned}&\mathbb{P}\left\{\bar{\nu}_N-\overline{\psi}(\bar{\nu}_N,N;y)\le\bar{\zeta}_N\le\bar{\nu}_N-\underline{\psi}(\bar{\nu}_N,N;y)\right\}\\ &\ge 1-2e\big(y\big[\ln((y-1)N\big)+2\big]+2\big)e^{-y}.\end{aligned} \tag{D.5}$$

Proof of Lemma D.1. Utilizing Bernstein's inequality for martingales (cf., e.g., [21, Theorem 3.14]) we obtain for all $z>0$ and $s>0$,

$$\mathbb{P}\left\{\left|\sum_{i=1}^N\zeta_i\right|\ge\sqrt{2zs}+\frac{z}{3},\ \sum_{i=1}^N\sigma_i^2\le s\right\}\le 2e^{-z}. \tag{D.6}$$

We conclude that

$$\mathbb{P}\left\{|\bar{\zeta}_N|\ge\sqrt{\frac{2\bar{s}_N}{N}z(1+z^{-1})}+\frac{z}{3N},\ \bar{s}_N\in[s,(1+z^{-1})s]\right\}\le 2e^{-z},$$

implying that for $y=z+1>1$

$$\mathbb{P}\left\{|\bar{\zeta}_N|\ge\sqrt{\frac{2y\bar{s}_N}{N}}+\frac{y}{3N},\ \bar{s}_N\in\big[s,(y-1)^{-1}ys\big]\right\}\le 2e^{-y+1}. \tag{D.7}$$

Let now $s^j=\min\left\{\overline{s},\left(\frac{y}{y-1}\right)^j s^0\right\}$, $j=0,...,J$, with $s^0=\underline{s}$, $s^J=\overline{s}$, and $J=\big\lfloor\ln\left(\overline{s}/\underline{s}\right)\ln^{-1}\left((y-1)^{-1}y\right)\big\rfloor$. Note that $\ln\left(1+1/(y-1)\right)\ge 1/y$ for $y>1$, so that

$$J\le\ln\left(\overline{s}/\underline{s}\right)\ln^{-1}\left((y-1)^{-1}y\right)+1\le y\ln\left(\overline{s}/\underline{s}\right)+1.$$

On the other hand, due to (D.7),

$$\begin{aligned}
&\mathbb{P}\left\{|\bar{\zeta}_N| \geq \sqrt{\frac{2y\bar{s}_N}{N}} + \frac{y}{3N},\ \underline{s} \leq \bar{s}_N \leq \overline{s}\right\} \\
&\leq \sum_{j=1}^{J} \mathbb{P}\left\{|\bar{\zeta}_N| \geq \sqrt{\frac{2y\bar{s}_N}{N}} + \frac{y}{3N},\ \bar{s}_N \in \left[s^j, s^{j+1}\right]\right\} \leq 2Je^{-y+1} \\
&\leq 2e\big(y\ln\left(\overline{s}/\underline{s}\right) + 1\big)e^{-y}
\end{aligned}$$

what is (D.1). Let us put $s = (18z)^{-1}$ in (D.6); together with $y = z + 1 > 1$, we get

$$\mathbb{P}\left\{|\bar{\zeta}_N| \geq \frac{y}{3N},\ \bar{s}_N \leq \frac{1}{18N(y-1)}\right\} \leq 2e^{-y+1}. \tag{D.8}$$

Furthermore, we have $\bar{s}_N \leq 1/4$ a.s.. When substituting $\underline{s} = (18(y-1))^{-1}$ and $\overline{s} = N/4$ into (D.1) we obtain

$$\mathbb{P}\left\{|\bar{\zeta}_N| \geq \sqrt{\frac{2y\bar{s}_N}{N}} + \frac{y}{3N},\ \bar{s}_N \geq \frac{1}{18N(y-1)}\right\} \leq 2e\big(y\ln\left(\tfrac{9}{2}(y-1)N\right) + 1\big)e^{-y}.$$

Finally, when taking into account (D.8) we conclude with

$$\begin{aligned}
&\mathbb{P}\left\{|\bar{\zeta}_N| \geq \sqrt{\frac{2y\bar{s}_N}{N}} + \frac{y}{3N}\right\} \\
&\leq 2e\big(y\ln\left(\tfrac{9}{2}(y-1)N\right) + 2\big)e^{-y} \leq 2e\big(y\big[\ln((y-1)N\big) + 2\big] + 2\big)e^{-y}.
\end{aligned}$$

Next, we observe that $\bar{s}_N \leq \bar{\mu}_N(1-\bar{\mu}_N)$, and replacing $\bar{s}_N$ in (D.2) with this upper bound come to the inequality:

$$\mathbb{P}\left\{|\bar{\zeta}_N| \geq \sqrt{\frac{2y\bar{\mu}_N(1-\bar{\mu}_N)}{N}} + \frac{y}{3N}\right\} \leq 2e\big(y\big[\ln((y-1)N\big) + 2\big] + 2\big)e^{-y}.$$

In other words, there exist a subset $\overline{\Omega}^N$ of the space Ω^N of realizations ω^N of probability at

least $1-2e\big(y\ln\big((y-1)n\big)+4\big)e^{-y}$ and such for all $\omega^N\in\overline{\Omega}^N$ one has

$$|\bar{\zeta}_N|\leq\sqrt{\frac{2y\bar{\mu}_N(1-\bar{\mu}_N)}{N}}+\frac{y}{3N}. \tag{D.9}$$

Observe that $\bar{\mu}_n$ can be eliminated from the above inequalities: when denoting $\nu_i=\gamma_i^\top\omega_i$ with $\bar{\nu}_N=\frac{1}{N}\sum_{i=1}^N\nu_i=\bar{\zeta}_N+\bar{\mu}_N$, by simple algebra we deduce from (D.9) that

$$\underline{\psi}(\bar{\nu}_N,I;y)\leq\bar{\mu}_N\leq\overline{\psi}(\bar{\nu}_N,I;y)$$

where $\underline{\psi}(\cdot)$ and $\overline{\psi}(\cdot)$ are as in (D.4). We conclude that for $\omega^N\in\overline{\Omega}^N$

$$\bar{\nu}_N-\overline{\psi}(\bar{\nu}_N,I;y)\leq\bar{\zeta}_N\leq\bar{\nu}_N-\underline{\psi}(\bar{\nu}_N,I;y)$$

what implies (D.5). □

Now, in the premise of Lemma 6.2, let us fix $k\in\{1,...,\kappa\}$, and let us denote $\gamma_i^\top=[\eta(\omega_{i-d}^{i-1})]_k=\mathrm{Row}_k[\eta(\omega_{i-d}^{i-1})]$, the k-th row of $\eta(\omega_{i-d}^{i-1})$. We set $\nu_i=\gamma_i^\top\omega_i=[\eta(\omega_{i-d}^{i-1})]_k\omega_i$. Note that conditional distribution of the r.v. ν_i given ω^{i-1} is Bernoulli distribution with parameter $\mu_i=\mathbb{E}_{|\omega^{i-1}}\{\nu_i\}=[\eta(\omega_{i-d}^{i-1})]_k\eta^\top(\omega_{i-d}^{i-1})\beta$. Defining, as above, $\zeta_i=\nu_i-\mu_i$, $\bar{\zeta}_N=\frac{1}{N}\sum_{i=1}^N\zeta_i=(F_{\omega^N}(\beta))_k$, the k-th component of the field $F_{\omega^N}(\beta)$, $\bar{\nu}_N=\frac{1}{N}\sum_{i=1}^N\nu_i=a[\omega^N]_k$, and $\bar{\mu}_N=\frac{1}{N}\sum_{i=1}^N\mu_i=\frac{1}{N}\sum_{i=1}^N[\eta(\omega_{i-d}^{i-1})]_k\eta^\top(\omega_{i-d}^{i-1})\beta=(A[\omega^N]\beta)_k$, the k-th component of $A[\omega^N]\beta$, and utilizing bound (D.5) of Lemma D.1 we conclude that for any $y>1$, $(F_{\omega^N}(\beta))_k$, $k=1,...,\kappa$, satisfy, with probability at least $1-2e\big(y\big[\ln((y-1)N\big)+2\big]+2\big)e^{-y}$, the bound

$$\bar{\nu}_N-\overline{\psi}(\bar{\nu}_N,N;y)\leq(F_{\omega^N}(\beta))_k\leq\bar{\nu}_N-\underline{\psi}(\bar{\nu}_N,N;y)$$

where $\underline{\psi}(\cdot)$ and $\overline{\psi}(\cdot)$ are as in (D.4). □

REFERENCES

[1] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "Sublinear algorithms for outlier detection and generalized closeness testing," in *Proceedings of the International Symposium on Information Theory*, IEEE, 2014, pp. 3200–3204.

[2] J. Acharya, Z. Sun, and H. Zhang, "Differentially private testing of identity and closeness of discrete distributions," in *Proceedings of the Advances in Neural Information Processing Systems*, 2018, pp. 6878–6891.

[3] F. B. Alt, "Multivariate quality control," *Encyclopedia of statistical science*, vol. 6, pp. 110–122, 1985.

[4] F. B. Alt and N. D. Smith, "17 multivariate process control," *Handbook of statistics*, vol. 7, pp. 333–351, 1988.

[5] S.-i. Amari, *Information Geometry and Its Applications*. Springer, 2016, vol. 194.

[6] D. Amoreses, "Applying a change-point detection method on frequency-magnitude distributions," *Bulletin of the Seismological Society of America*, vol. 97, no. 5, pp. 1742–1749, 2007.

[7] T. W. Anderson, "Asymptotic theory for principal component analysis," *Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 122–148, 1963.

[8] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, *et al.*, "BACH: Grand challenge on breast cancer histology images," *Medical Image Analysis*, vol. 56, pp. 122–139, 2019.

[9] E. Arias-Castro, S. Bubeck, and G. Lugosi, "Detection of correlations," *Annals of Statistics*, vol. 40, no. 1, pp. 412–435, 2012.

[10] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, NSW, Australia, 2017, pp. 214–223.

[11] V. Avanesov and N. Buzun, "Change-point detection in high-dimensional covariance structure," *Electronic Journal of Statistics*, vol. 12, no. 2, pp. 3254–3294, 2018.

[12] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal, Second Series*, vol. 19, no. 3, pp. 357–367, 1967.

[13] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *Journal of Multivariate Analysis*, vol. 97, no. 6, pp. 1382–1408, 2006.

[14] L. Balzano, Y. Chi, and Y. M. Lu, "Streaming pca and subspace tracking: The missing data case," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1293–1310, 2018.

[15] I. Barnett and J.-P. Onnela, "Change point detection in correlation networks," *Scientific Reports*, vol. 6, 2015.

[16] M. Baron, "Early detection of epidemics as a sequential change-point problem," in *Longevity, Aging, and Degradation Models in Reliability, Public Health, Medicine and Biology*, 2004, pp. 31–43.

[17] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, 1993.

[18] A. Basu, I. R. Harris, N. L. Hjort, and M. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, 1998.

[19] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing that distributions are close," in *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, IEEE, 2000, pp. 259–269.

[20] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing closeness of discrete distributions," *Journal of the ACM (JACM)*, vol. 60, no. 1, pp. 1–25, 2013.

[21] B. Bercu, B. Delyon, and E. Rio, *Concentration inequalities for sums and martingales*. Springer, 2015.

[22] Q. Berthet and P. Rigollet, "Optimal detection of sparse principal components in high dimension," *Annals of Statistics*, vol. 41, no. 4, pp. 1780–1815, 2013.

[23] B. Bhattacharya and G. Valiant, "Testing closeness with unequal sized samples," in *Proceedings of the Advances in Neural Information Processing Systems*, 2015, pp. 2611–2619.

[24] P. J. Bickel, "A distribution free version of the Smirnov two sample test in the p-variate case," *Annals of Mathematical Statistics*, vol. 40, no. 1, pp. 1–23, 1969.

[25] P. J. Bickel and L. Breiman, "Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test," *Annals of Probability*, vol. 11, no. 1, pp. 185–214, 1983.

[26] J. Blanchet, Y. Kang, and K. Murthy, "Robust Wasserstein profile inference and applications to machine learning," *Journal of Applied Probability*, vol. 56, no. 3, pp. 830–857, 2019.

[27] J. Blanchet and K. Murthy, "Quantifying distributional model risk via optimal transport," *Mathematics of Operations Research*, vol. 44, no. 2, pp. 565–600, 2019. eprint: https://doi.org/10.1287/moor.2018.0936.

[28] O. Bodnar and W. Schmid, "Multivariate control charts based on a projection approach," *Allgemeines Statistisches Archiv*, vol. 89, no. 1, pp. 75–93, 2005.

[29] G. Boracchi, D. Carrera, C. Cervellera, and D. Maccio, "QuantTree: Histograms for change detection in multivariate data streams," in *Proceedings of the International Conference on Machine Learning*, vol. 80, 2018, pp. 638–647.

[30] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[31] G. Canas and L. Rosasco, "Learning probability measures with respect to optimal transport metrics," in *Proceedings of the Advances in Neural Information Processing Systems 25*, 2012, pp. 2492–2500.

[32] C. L. Canonne, "A survey on distribution testing: Your data is big. But is it blue?" *Theory of Computing*, pp. 1–100, 2020.

[33] Y. Cao, A. Nemirovski, Y. Xie, V. Guigues, and A. Juditsky, "Change detection via affine and quadratic detectors," *Electronic Journal of Statistics*, vol. 12, no. 1, pp. 1–57, 2018.

[34] Y. Cao and Y. Xie, "Robust sequential change-point detection by convex optimization.," in *Proceedings of the International Symposium on Information Theory*, IEEE, 2017, pp. 1287–1291.

[35] Y. Cao, Y. Xie, and N. Gebraeel, "Multi-sensor slope change detection," *Annals of Operations Research*, vol. 263, no. 1-2, pp. 163–189, 2018.

[36] L. K. Chan and J. Zhang, "Cumulative sum control charts for the covariance matrix," *Statistica Sinica*, pp. 767–790, 2001.

[37] S.-O. Chan, I. Diakonikolas, P. Valiant, and G. Valiant, "Optimal algorithms for testing closeness of discrete distributions," in *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, 2014, pp. 1193–1203.

[38] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection for discrete sequences: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 5, pp. 823–839, 2010.

[39] H. Chen, "Sequential change-point detection based on nearest neighbors," *Annals of Statistics*, vol. 47, no. 3, pp. 1381–1407, 2019.

[40] J. Chen and A. Gupta, "Statistical inference of covariance change points in Gaussian model," *Statistics*, vol. 38, no. 1, pp. 17–28, 2004.

[41] J. Chen, S.-H. Kim, and Y. Xie, "S^3T: An efficient score-statistic for spatio-temporal surveillance," *arXiv preprint arXiv:1706.05331*, 2017.

[42] S. Chen, A. Shojaie, E. Shea-Brown, and D. Witten, "The multivariate Hawkes process in high dimensions: Beyond mutual excitation," *arXiv preprint arXiv:1707.04928*, 2017.

[43] Y. C. Chen, T. Banerjee, A. D. Dominguez-Garcia, and V. V. Veeravalli, "Quickest line outage detection and identification," *IEEE Transactions on Power Systems*, vol. 31, no. 1, pp. 749–758, 2015.

[44] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493–507, 1952.

[45] E. Chiauzzi, C. Rodarte, and P. DasMahapatra, "Patient-centered activity monitoring in the self-management of chronic health conditions," *BMC medicine*, vol. 13, no. 1, p. 77, 2015.

[46] Y. S. Chow, H. Robbins, and D. Siegmund, *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin, 1971.

[47] A. Cichocki and S.-i. Amari, "Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, 2010.

[48] W. J. Conover, *Practical Nonparametric Statistics*. John Wiley & Sons, 1998, vol. 350.

[49] H. Cramér, *Mathematical Methods of Statistics*. Princeton University Press, 1946, vol. 43.

[50] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations research*, vol. 58, no. 3, pp. 595–612, 2010.

[51] C. A. Di Vittorio and A. P. Georgakakos, "Land cover classification and wetland inundation mapping using modis," *Remote Sensing of Environment*, vol. 204, pp. 1–17, 2018.

[52] I. Diakonikolas, D. M. Kane, and V. Nikishkin, "Near-optimal closeness testing of discrete histogram distributions," *arXiv preprint arXiv:1703.01913*, 2017.

[53] J. Duchi and H. Namkoong, "Learning models with uniform performance via distributionally robust optimization," *arXiv preprint arXiv:1810.08750*, 2018.

[54] A. Edelman and Y. Wang, "Random matrix theory and its innovative applications," in *Advances in Applied Mathematics, Modeling, and Computational Science*, Springer, 2013, pp. 91–116.

[55] M. Eichler, R. Dahlhaus, and J. Dueck, "Graphical modeling for multivariate Hawkes processes with nonparametric link functions," *Journal of Time Series Analysis*, vol. 38, no. 2, pp. 225–242, 2017.

[56] P. Embrechts, T. Liniger, and L. Lin, "Multivariate Hawkes processes: An application to financial data," *Journal of Applied Probability*, vol. 48, no. A, pp. 367–378, 2011.

[57] Ş. Ertekin, C. Rudin, and T. H. McCormick, "Reactive point processes: A new approach to predicting power failures in underground electrical systems," *Annals of Applied Statistics*, vol. 9, no. 1, pp. 122–144, 2015.

[58] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1, pp. 115–166, 2018.

[59] D. Evans, A. J. Jones, and W. M. Schmidt, "Asymptotic moments of near–neighbour distance distributions," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 458, no. 2028, pp. 2839–2849, 2002.

[60] X. Fan, I. Grama, and Q. Liu, "Hoeffding's inequality for supermartingales," *Stochastic Processes and their Applications*, vol. 122, no. 10, pp. 3545–3559, 2012.

[61] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 1737–1746.

[62] N. Fournier and A. Guillin, "On the rate of convergence in Wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, vol. 162, no. 3, pp. 707–738, 2015.

[63] E. W. Fox, F. P. Schoenberg, J. S. Gordon, *et al.*, "Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences," *Annals of Applied Statistics*, vol. 10, no. 3, pp. 1725–1756, 2016.

[64] D. A. Freedman, "On tail probabilities for martingales," *Annals of Probability*, pp. 100–118, 1975.

[65] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," *Annals of Statistics*, vol. 7, no. 4, pp. 697–717, 1979.

[66] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, 10. New York: Springer Series in Statistics, 2001, vol. 1.

[67] R. Gao, "Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality," *arXiv preprint arXiv:2009.04382*, 2020.

[68] R. Gao and A. J. Kleywegt, "Distributionally robust stochastic optimization with wasserstein distance," *arXiv preprint arXiv:1604.02199*, 2016.

[69] R. Gao, L. Xie, Y. Xie, and H. Xu, "Robust hypothesis testing using Wasserstein uncertainty sets," in *Proceedings of the Advances in Neural Information Processing Systems*, 2018, pp. 7902–7912.

[70] S. Geman, "A limit theorem for the norm of random matrices," *Annals of Probability*, vol. 8, no. 2, pp. 252–261, 1980.

[71] J. Goh and M. Sim, "Distributionally robust optimization and its tractable approximations," *Operations research*, vol. 58, no. 4-part-1, pp. 902–917, 2010.

[72] A. Goldenshluger, A. Juditsky, and A. Nemirovski, "Hypothesis testing by convex optimization," *Electronic Journal of Statistics*, vol. 9, no. 2, pp. 1645–1712, 2015.

[73] O. Goldreich and D. Ron, "On testing expansion in bounded-degree graphs," in *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, Springer, 2011, pp. 68–75.

[74] M. Gomez-Rodriguez, J. Leskovec, and B. Schölkopf, "Modeling information propagation with survival theory," in *International Conference on Machine Learning*, 2013, pp. 666–674.

[75] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[76] O. Guédon and R. Vershynin, "Community detection in sparse networks via Grothendieck's inequality," *Probability Theory and Related Fields*, vol. 165, no. 3-4, pp. 1025–1049, 2016.

[77] G. Gül and A. M. Zoubir, "Minimax robust hypothesis testing," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5572–5587, 2017.

[78] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proceedings of the Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[79] J. Hajek, Z. Sidak, and P. Sen, *Theory of Rank Tests*. Academic Press, 1999.

[80] E. C. Hall and R. M. Willett, "Tracking dynamic point processes on networks," *IEEE Transactions on Information Theory*, vol. 62, no. 7, pp. 4327–4346, 2016.

[81] D. Hallac, J. Leskovec, and S. Boyd, "Network lasso: Clustering and optimization in large graphs," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 387–396.

[82] N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard, "Lasso and probabilistic inequalities for multivariate point processes," *Bernoulli*, vol. 21, no. 1, pp. 83–143, 2015.

[83] A. G. Hawkes, "Point spectra of some mutually exciting point processes," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 33, no. 3, pp. 438–443, 1971.

[84] ——, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[85] A. G. Hawkes and D. Oakes, "A cluster process representation of a self-exciting process," *Journal of Applied Probability*, vol. 11, no. 3, pp. 493–503, 1974.

[86] J. D. Healy, "A note on multivariate CUSUM procedures," *Technometrics*, vol. 29, no. 4, pp. 409–412, 1987.

[87] N. Henze, "A multivariate two-sample test based on the number of nearest neighbor type coincidences," *Annals of Statistics*, vol. 16, no. 2, pp. 772–783, 1988.

[88] H. Hotelling, "Multivariate quality control," *Techniques of statistical analysis*, 1947.

[89] P. J. Huber, "A robust version of the probability ratio test," *Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.

[90] ——, "Projection pursuit," *Annals of Statistics*, pp. 435–475, 1985.

[91] P. J. Huber and V. Strassen, "Minimax tests and the Neyman-Pearson lemma for capacities," *Annals of Statistics*, vol. 1, no. 2, pp. 251–263, 1973.

[92] J. E. Jackson, "Quality control methods for several related variables," *Technometrics*, vol. 1, no. 4, pp. 359–377, 1959.

[93] J. E. Jackson and G. S. Mudholkar, "Control procedures for residuals associated with principal component analysis," *Technometrics*, vol. 21, no. 3, pp. 341–349, 1979.

[94] T. Jiang, K. Leder, and G. Xu, "Rare-event analysis for extremal eigenvalues of white Wishart matrices," *Annals of Statistics*, vol. 45, no. 4, pp. 1609–1637, 2017.

[95] Y. Jiao, Y. Chen, and Y. Gu, "Subspace Change-Point Detection: A New Model and Solution," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 6, pp. 1224–1239, 2018.

[96] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160 035, 2016.

[97] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Annals of Statistics*, vol. 29, no. 2, pp. 295–327, 2001.

[98] A. B. Juditsky and A. Nemirovski, "Signal recovery by stochastic optimization," *Automation and Remote Control*, vol. 80, no. 10, pp. 1878–1893, 2019.

[99] A. Juditsky and A. Nemirovski, *Statistical Inference via Convex Optimization*. Princeton University Press, 2020, vol. 69.

[100] A. Juditsky, A. Nemirovski, L. Xie, and Y. Xie, "Convex parameter recovery for interacting marked processes," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 799–813, 2020.

[101] V. Konev and S. Pergamenshchikov, "On asymptotic minimaxity of fixed accuracy esimators for autoregression parameters I. Stable process," *Mathematical Methods of Statistics*, vol. 5, no. 2, pp. 125–153, 1996.

[102] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[103] M. Kuperman and G. Abramson, "Small world effect in an epidemiological model," *Physical Review Letters*, vol. 86, no. 13, p. 2909, 2001.

[104] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

[105] E. L. Lai, D. Moyer, B. Yuan, E. Fox, B. Hunter, A. L. Bertozzi, and P. J. Brantingham, "Topic time series analysis of microblogs," *IMA Journal of Applied Mathematics*, vol. 81, no. 3, pp. 409–431, 2016.

[106] L. Lai, Y. Fan, and H. V. Poor, "Quickest detection in cognitive radio: A sequential change detection framework," in *Proceedings of the IEEE Global Telecommunications Conference*, 2008, pp. 1–5.

[107] T. L. Lai, "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2917–2929, 1998.

[108] T. L. Lai, "Sequential changepoint detection in quality control and dynamical systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 613–658, 1995.

[109] ——, "Sequential analysis: Some classical problems and new challenges," *Statistica Sinica*, vol. 11, no. 2, pp. 303–350, 2001.

[110] T. L. Lai and J. Z. Shan, "Efficient recursive algorithms for detection of abrupt changes in signals and control systems," *IEEE Transactions on Automatic Control*, vol. 44, no. 5, pp. 952–966, 1999.

[111] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *ACM SIGCOMM Computer Communication Review*, ACM, vol. 34, 2004, pp. 219–230.

[112] H. Lam, "Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization," *Operations Research*, vol. 67, no. 4, pp. 1090–1105, 2019.

[113] R. J. Larsen, *Statistics in the Real World: A Book of Examples*. New York: Macmillan, 1976.

[114] T. S. Lau, W. P. Tay, and V. V. Veeravalli, "A binning approach to quickest change detection with unknown post-change distribution," *IEEE Transactions on Signal Processing*, vol. 67, no. 3, pp. 609–621, 2018.

[115] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[116] K.-C. Lee and D. Kriegman, "Online learning of probabilistic appearance manifolds for video-based recognition and tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 852–859.

[117] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[118] E. Levina and P. Bickel, "The earth mover's distance is the mallows distance: Some insights from statistics," in *Proceedings of the Eighth International Conference on Computer Vision*, IEEE, vol. 2, 2001, pp. 251–256.

[119] B. C. Levy, "Robust hypothesis testing with a relative entropy tolerance," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 413–421, 2009.

[120] S. Li, Y. Xie, H. Dai, and L. Song, "M-statistic for kernel change-point detection," in *Advances in Neural Information Processing Systems*, 2015, pp. 3366–3374.

[121] S. Li, Y. Xie, M. Farajtabar, A. Verma, and L. Song, "Detecting changes in dynamic events over networks," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 2, pp. 346–359, 2017.

[122] Z. Li, Z. Peng, D. Hollis, L. Zhu, and J. McClellan, "High-resolution seismic event detection using local similarity for large-n arrays," *Scientific reports*, vol. 8, no. 1, p. 1646, 2018.

[123] R. Liptser and V. Spokoiny, "Deviation probability bound for martingales with applications to statistical estimation," *Statistics & probability letters*, vol. 46, no. 4, pp. 347–357, 2000.

[124] J. W. Lockhart, G. M. Weiss, J. C. Xue, S. T. Gallagher, A. B. Grosner, and T. T. Pulickal, "Design considerations for the WISDM smart phone-based sensor mining architecture," in *Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data*, ACM, 2011, pp. 25–33.

[125] G. Lorden, "Procedures for reacting to a change in distribution," *Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1897–1908, 1971.

[126] ——, "Open-ended tests for Koopman-Darmois families," *Annals of Statistics*, vol. 1, no. 4, pp. 633–643, 1973.

[127] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of mathematical statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[128] R. Mises and H. Pollaczek-Geiringer, "Praktische Verfahren der Gleichungsauflösung.," *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 9, no. 1, pp. 58–77, 1929.

[129] G. Mohler, "Modeling and estimation of multi-source clustering in crime and security data," *Annals of Applied Statistics*, vol. 7, no. 3, pp. 1525–1539, 2013.

[130] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, "Self-exciting point process modeling of crime," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 100–108, 2011.

[131] J. Moller and R. P. Waagepetersen, *Statistical inference and simulation for spatial point processes*. CRC Press, 2003.

[132] D. C. Montgomery, *Introduction to Statistical Quality Control*. John Wiley & Sons, 2007.

[133] P. Moulin and V. V. Veeravalli, *Statistical Inference for Engineers and Data Scientists*. Cambridge University Press, 2018.

[134] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Annals of Statistics*, vol. 14, no. 4, pp. 1379–1387, 1986.

[135] J. W. Mueller and T. Jaakkola, "Principal differences analysis: Interpretable characterization of differences between distributions," in *Proceedings of the Advances in Neural Information Processing Systems*, 2015, pp. 1702–1710.

[136] Y. Nesterov, "Semidefinite relaxation and nonconvex quadratic optimization," *Optimization Methods and Software*, vol. 9, no. 1-3, pp. 141–160, 1998.

[137] J. Neuburger, K. Walker, C. Sherlaw-Johnson, J. van der Meulen, and D. A. Cromwell, "Comparison of control charts for monitoring clinical performance using binary data," *BMJ quality & safety*, vol. 26, no. 11, pp. 919–928, 2017.

[138] J. Neyman and E. S. Pearson, "Ix. on the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694-706, pp. 289–337, 1933.

[139] Nurjahan, F. Nizam, S. Chaki, S. Al Mamun, and M. S. Kaiser, "Attack detection and prevention in the cyber physical system," in *Proceedings of the International Conference on Computer Communication and Informatics (ICCCI)*, 2016, pp. 1–6.

[140] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[141] ——, "A test for a change in a parameter occurring at an unknown point," *Biometrika*, vol. 42, no. 3/4, pp. 523–527, 1955.

[142] D. Paul, "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," *Statistica Sinica*, vol. 17, no. 4, pp. 1617–1642, 2007.

[143] L. Peel and A. Clauset, "Detecting change points in the large-scale structure of evolving networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015, pp. 2914–2920.

[144] L. Pelkowitz and S. Schwarts, "Asymptotically optimum sample size for quickest detection," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-23, no. 2, pp. 263–272, 1987.

[145] M. D. Penrose and J. Yukich, "Laws of large numbers and nearest neighbor distances," in *Advances in directional and linear statistics*, Springer, 2011, pp. 189–199.

[146] M. D. Penrose and J. E. Yukich, "Weak laws of large numbers in geometric probability," *Annals of Applied Probability*, vol. 13, no. 1, pp. 277–303, 2003.

[147] J. J. Pignatiello Jr and G. C. Runger, "Comparisons of multivariate cusum charts," *Journal of quality technology*, vol. 22, no. 3, pp. 173–186, 1990.

[148] J. Pitkin, I. Manolopoulou, and G. Ross, "Bayesian hierarchical modelling of sparse count processes in retail analytics," *arXiv preprint arXiv:1805.05657*, 2018.

[149] M. Pollak, "Optimal detection of a change in distribution," *Annals of Statistics*, vol. 13, no. 1, pp. 206–227, Mar. 1985.

[150] A. S. Polunchenko and A. G. Tartakovsky, "On optimality of the shiryaev–roberts procedure for detecting a change in distribution," *The Annals of Statistics*, vol. 38, no. 6, pp. 3445–3457, 2010.

[151] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge University Press, 2008.

[152] Y. V. Prokhorov, "Convergence of random processes and limit theorems in probability theory," *Theory of Probability & Its Applications*, vol. 1, no. 2, pp. 157–214, 1956.

[153] A. Python, J. Illian, C. Jones-Todd, and M. Blangiardo, "A Bayesian approach to modelling fine-scale spatial dynamics of non-state terrorism: World study, 2002-2013," *arXiv preprint arXiv:1610.01215*, 2016.

[154] M. Qu, F. Y. Shih, J. Jing, and H. Wang, "Automatic solar filament detection using image processing techniques," *Solar Physics*, vol. 228, no. 1-2, pp. 119–135, 2005.

[155] J. Rabin, G. Peyré, J. Delon, and M. Bernot, "Wasserstein barycenter and its application to texture mixing," in *Proceedings of the International Conference on Scale Space and Variational Methods in Computer Vision*, Springer, 2011, pp. 435–446.

[156] V. Raghavan and V. V. Veeravalli, "Quickest change detection of a Markov process across a sensor array," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1961–1981, 2010.

[157] M. Raginsky, R. Willett, C. Horn, J. Silva, and R. Marcia, "Sequential anomaly detection in the presence of noise and limited feedback," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5544–5562, 2012.

[158] H. Rahimian and S. Mehrotra, "Distributionally robust optimization: A review," *arXiv preprint arXiv:1908.05659*, 2019.

[159] R. Rajkumar, I. Lee, L. Sha, and J. Stankovic, "Cyber-physical systems: The next computing revolution," in *Proceedings of the Design Automation Conference*, 2010, pp. 731–736.

[160] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman, "On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 3571–3577.

[161] A. Ramdas, N. Trillos, and M. Cuturi, "On Wasserstein two-sample testing and related families of nonparametric tests," *Entropy*, vol. 19, no. 2, p. 47, 2017.

[162] A. Reinhart, "A review of self-exciting spatio-temporal point processes and their applications," *arXiv preprint arXiv:1708.02647*, 2017.

[163] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, The Regents of the University of California, 1961.

[164] Y. Ritov, "Decision theoretic optimality of the CUSUM procedure," *Annals of Statistics*, vol. 18, no. 3, pp. 1464–1469, 1990.

[165] R. Rubinfeld, "Taming big probability distributions," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 19, no. 1, pp. 24–28, 2012.

[166] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[167] W. Rudin, "Real and complex analysis (mcgraw-hill international editions: Mathematics series)," 1987.

[168] G. C. Runger, "Projections and the u 2 multivariate control chart," *Journal of Quality Technology*, vol. 28, no. 3, pp. 313–319, 1996.

[169] M. F. Schilling, "Multivariate two-sample tests based on nearest neighbors," *Journal of the American Statistical Association*, vol. 81, no. 395, pp. 799–806, 1986.

[170] F. Scholz and A. Zhu, *Ksamples: K-sample rank tests and their combinations*, 2019.

[171] G. Schwarz, "Estimating the dimension of a model," *Annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[172] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn, "Distributionally robust logistic regression," in *Proceedings of the Advances in Neural Information Processing Systems*, 2015, pp. 1576–1584.

[173] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[174] J. Shen and N. Zhang, "Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing," *Annals of Applied Statistics*, vol. 6, no. 2, pp. 476–496, 2012.

[175] J. Shi, *Stream of Variation Modeling and Analysis for Multistage Manufacturing Processes*. CRC press, 2006.

[176] A. N. Shiryaev, "On optimum methods in quickest detection problems," *Theory of Probability & Its Applications*, vol. 8, no. 1, pp. 22–46, 1963.

[177] M. B. Short, M. R. D'orsogna, V. B. Pasour, G. E. Tita, P. J. Brantingham, A. L. Bertozzi, and L. B. Chayes, "A statistical model of criminal behavior," *Mathematical Models and Methods in Applied Sciences*, vol. 18, no. supp01, pp. 1249–1267, 2008.

[178] N. Si, J. Blanchet, S. Ghosh, and M. Squillante, "Quantifying the empirical wasserstein distance to a set of measures: Beating the curse of dimensionality," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[179] D. Siegmund, B. Yakir, and N. Zhang, "Tail approximations for maxima of random fields by likelihood ratio transformations," *Sequential Analysis*, vol. 29, no. 3, pp. 245–262, 2010.

[180] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*. Springer Science & Business Media, 1985.

[181] D. Siegmund and E. Venkatraman, "Using the generalized likelihood ratio statistic for sequential detection of a change-point," *Annals of Statistics*, vol. 23, no. 1, pp. 255–271, 1995.

[182] D. Siegmund and B. Yakir, "Tail probabilities for the null distribution of scanning statistics," *Bernoulli*, vol. 6, no. 2, pp. 191–213, 2000.

[183] ——, *The Statistics of Gene Mapping*. Springer Science & Business Media, 2007.

[184] ——, "Detecting the emergence of a signal in a noisy image," *Statistics and Its Interface*, vol. 1, no. 1, pp. 3–12, 2008.

[185] A. Sinha, H. Namkoong, and J. Duchi, "Certifying some distributional robustness with principled adversarial training," *arXiv preprint arXiv:1710.10571*, 2017.

[186] M. Sion, "On general minimax theorems," *Pacific Journal of mathematics*, vol. 8, no. 1, pp. 171–176, 1958.

[187] N. V. Smirnov, "Estimate of deviation between empirical distribution functions in two independent samples," *Bulletin Moscow University*, vol. 2, no. 2, pp. 3–16, 1939.

[188] G. Sprint, D. J. Cook, and M. Schmitter-Edgecombe, "Unsupervised detection and analysis of changes in everyday physical activity data," *Journal of biomedical informatics*, vol. 63, pp. 54–65, 2016.

[189] A. Stomakhin, M. B. Short, and A. L. Bertozzi, "Reconstruction of missing data in social networks based on temporal patterns of interactions," *Inverse Problems*, vol. 27, no. 11, p. 115 013, 2011.

[190] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

[191] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[192] G. J. Székely and M. L. Rizzo, "Testing for equal distributions in high dimension," *InterStat*, vol. 5, no. 16.10, pp. 1249–1272, 2004.

[193] A. Tank, E. B. Fox, and A. Shojaie, "Granger causality networks for categorical time series," *arXiv preprint arXiv:1706.02781*, 2017.

[194] A. G. Tartakovsky, *Sequential Change Detection and Hypothesis Testing: General Non-iid Stochastic Models and Asymptotically Optimal Rules*. ser. Monographs on Statistics and Applied Probability 165. Boca Raton, London, New York: Chapman & Hall/CRC Press, Taylor & Francis Group, 2020.

[195] A. G. Tartakovsky, M. Pollak, and A. Polunchenko, "Third-order asymptotic optimality of the generalized Shiryaev–Roberts changepoint detection procedures," *Theory of Probability & Its Applications*, vol. 56, no. 3, pp. 457–484, 2012.

[196] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. ser. Monographs on Statistics and Applied Probability 136. Boca Raton, London, New York: Chapman & Hall/CRC Press, Taylor & Francis Group, 2015.

[197] C. Tracy and H. Widom, "On orthogonal and symplectic matrix ensembles," *Comm. Math. Phys.*, vol. 177, pp. 727–754, 1996.

[198] P. Valiant, "Testing symmetric properties of distributions," *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1927–1968, 2011.

[199] V. V. Veeravalli and T. Banerjee, "Quickest change detection," *Academic Press Library in Signal Processing: Array and Statistical Signal Processing*, vol. 3, pp. 209–256, 2013.

[200] C. Villani, *Topics in Optimal Transportation*, 58. American Mathematical Society, 2003.

[201] ——, *Optimal Transport: Old and New*. Springer Science & Business Media, 2008, vol. 338.

[202] A. R. Wade, "Explicit laws of large numbers for random nearest-neighbour-type graphs," *Advances in Applied Probability*, vol. 39, no. 2, pp. 326–342, 2007.

[203] D. Wang, Y. Yu, and A. Rinaldo, "Optimal covariance change point localization in high dimension," *arXiv preprint arXiv:1712.09912*, 2017.

[204] G. M. Weiss and J. W. Lockhart, "The impact of personalization on smartphone-based activity recognition," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[205] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Operations Research*, vol. 62, no. 6, pp. 1358–1376, 2014.

[206] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[207] A. Willsky and H. Jones, "A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems," *IEEE Transactions on Automatic Control*, vol. 21, no. 1, pp. 108–112, 1976.

[208] M. A. Woodbury, "Inverting modified matrices," *Memorandum report*, vol. 42, no. 106, p. 336, 1950.

[209] L. Xie, G. V. Moustakides, and Y. Xie, "First-order optimal sequential subspace change-point detection," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, IEEE, 2018, pp. 111–115.

[210] L. Xie and Y. Xie, "Sequential change detection by optimal weighted ℓ_2 divergence," *IEEE Journal on Selected Areas in Information Theory*, 2021.

[211] L. Xie, Y. Xie, and G. V. Moustakides, "Asynchronous multi-sensor change-point detection for seismic tremors," in *Proceedings of the International Symposium on Information Theory*, 2019, pp. 787–791.

[212] ——, "Sequential subspace change point detection," *Sequential Analysis*, vol. 39, no. 3, pp. 307–335, 2020.

[213] L. Xie, S. Zou, Y. Xie, and V. V. Veeravalli, "Sequential (quickest) change detection: Classical results and new directions," *IEEE Journal on Selected Areas in Information Theory*, 2021.

[214] Y. Xie and D. Siegmund, "Sequential multi-sensor change-point detection," *Annals of Statistics*, vol. 41, no. 2, pp. 670–692, Apr. 2013.

[215] H. Xu and H. Zha, "Thap: A Matlab toolkit for learning with hawkes processes," *arXiv preprint arXiv:1708.09252*, 2017.

[216] B. Yakir, "Multi-channel change-point detection statistic with applications in DNA copy-number variation and sequential monitoring," in *Proceedings of Second International Workshop in Sequential Methodologies*, 2009, pp. 15–17.

[217] ——, *Extremes in random fields: A theory and its applications*. John Wiley & Sons, 2013.

[218] B. Yakir and M. Pollak, "A new representation for a renewal-theoretic constant appearing in asymptotic approximations of large deviations," *Annals of Applied Probability*, vol. 8, no. 3, pp. 749–774, 1998.

[219] Y.-Q. Yin, Z.-D. Bai, and P. R. Krishnaiah, "On the limit of the largest eigenvalue of the large dimensional sample covariance matrix," *Probability theory and related fields*, vol. 78, no. 4, pp. 509–521, 1988.

[220] B. Yuan, H. Li, A. L. Bertozzi, P. J. Brantingham, and M. A. Porter, "Multivariate spatiotemporal Hawkes processes and network reconstruction," *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 2, pp. 356–382, 2019.

[221] C. Zhang, H. Yan, S. Lee, and J. Shi, "Dynamic multivariate functional data modeling via sparse subspace learning," *Technometrics*, pp. 1–14, 2020.

[222] R. Zhang, Y. Mei, and J. Shi, "Robust real-time monitoring of high-dimensional data streams," *arXiv preprint arXiv:1906.02265*, 2019.

[223] K. Zhou, H. Zha, and L. Song, "Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes," in *Artificial Intelligence and Statistics*, 2013, pp. 641–649.

[224] S. Zhu and Y. Xie, "Crime linkage detection by spatial-temporal-textual point processes," *arXiv preprint arXiv:1902.00440*, 2019.

VITA

Liyan Xie was born in Dangshan, Anhui Province, China in July 1995. She received the B.Sc. in Statistics from University of Science and Technology of China in 2016. Afterwards, she joined the Ph.D. program in the School of Industrial and Systems Engineering at Georgia Institute of Technology, under supervision of Prof. Yao Xie. With a fulfilling period and many stimulating experiences, she has completed her Ph.D. studies and now she is ready for new adventures. These will take her to The Chinese University of Hong Kong, Shenzhen, where she will be an assistant professor in the School of Data Science.