POISSON MATRIX COMPLETION AND CHANGE-POINT DETECTION

A Dissertation
Presented to
The Academic Faculty

By

Yang Cao

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

Georgia Institute of Technology

August 2018

POISSON MATRIX COMPLETION AND CHANGE-POINT DETECTION

Approved by:

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

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

Dr. Xiaoming Huo
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Kamran Paynabar
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Mark Davenport
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: April 9, 2018

Truth is ever to be found in simplicity, and not in the multiplicity and confusion of things.

Isaac Newton

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Yao Xie, for introducing me to the wonderful research field of statistical signal processing and machine learning, and for her constant support and guidance during my graduate study and thesis preparation. She has not only helped me prepare my first paper from the scratch, but has also inspired me and guided me to think of many challenging and interesting problems along all my Ph.D. years. Beyond knowledge, she also helped me develop the ability to think of and solve new problems independently. She is a great advisor and a nice friend. It has been a great pleasure to work under her supervision.

I would also like to thank to Jeff Wu, Xiaoming Huo, Mark Davenport and Kamran Paynabar for serving on my thesis committee. They have provided me with much encouragement and guidance through classes and personal communications. Special thanks go to Xiaoming Huo for offering me strong recommendation during my application to Georgia Tech and helpful personal advise throughout my Ph.D. study. Special thanks go to Jeff Wu for encouraging me a lot and helping me regain the confidence when I failed to pass the comprehensive exam the first time, and for providing me considerate suggestions for my career path. I am grateful to Mark Davenport for raising helpful discussions and sharing insightful ideas about the matrix completion problem. Thanks to Kamran Paynabar for introducing me much knowledge of quality control, and for the help when I was working as a teaching assistant. I am also grateful to Zheng Wen, my mentor during the internship in Adobe, for his patient guidance and fruitful discussions. All of their expertise has been greatly beneficial to my research and career development.

I want to thank the H. Milton Stewart School of Industrial and Systems Engineering. The students, faculty and staffs have provided me with a wonderful study and work environment. My thanks go to Alan Erera and Amanda Ford for the help in my graduate semester, and go to King Harrison III and the family for their financial support through Robert Goodell Brown fellowship. I would also like to thank for the NSF CAREER CCF-1650913 Award and NSF CCF-1663709 support. I also want to thank to my undergraduate thesis advisor, Zhouwang Yang, and many professors during my Bachelor's studies at the University of Science and Technology of China (USTC). They help me lay a solid mathematical background.

I would like to say thanks to my girlfriend, Chen Feng, for her constant love and support and for spending her life with me. I would like to thank to all my friends at and outside Georgia Tech for their friendship.

Last but not least, I want to thank my parents for their love and care during my pursuit of a Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Statistical signal processing and machine learning are very important in modern science and engineering. Many theories, methods and techniques are developed to help people extract and analyze the hidden information from large amount of data, or so called Big Data. The challenge of dealing with Big Data comes from either huge amount of data or high-dimensionality of data. Considering the above two aspects, this dissertation focus on two problems: matrix completion and sequential change-point detection.

The first one is low-rank matrix completion problem. Because many survey data is capable of being represented approximately as an low-rank matrix, this model can apply to many applications such as sensor network [1], network traffic analysis [2], sensor localization [3], recommender systems [4, 5] and natural language processing [6]. Much success has been achieved, including both theories and algorithms, to learn the model. In the case when the entries of the matrix is assumed to be continuous, the authors of [7] prove that we can achieve an exact recovery of the matrix provided that the number of observed entries is at least on a logarithmic order of the total number of entries of the matrix, and they offer an efficient algorithm to recover the missing entries of the matrix. Many following work is then published to extend the results of [7] to more complicated cases when some unknown continuous random errors exist. Beyond the continuous noise, the authors in [8] first develop new techniques to handle the discrete noise, called the 1-bit matrix completion, which is crucial since it provides the theoretical basis for the quantized matrix completion. In Chapter 3, we extend this work to solve an important and fundamental case where the entries are Poisson random variables. We formulate the problems into a regularized likelihood maximization problem and solve this problem through an efficient algorithm based on the singular value decomposition. Finally, we prove the optimality of the proposed methods and demonstrate the good performance by synthetic simulations and real-data examples.

The second one is sequential change-point detection. The model is that we observe a

sequence of independent signals and would like to detect if an unexpected change happens in the system. This belongs to sequential decision making problems and we need decide sequentially whether an alarm of change detected should be raised as we obtain new observations. Due to the nature of sequential decision making, sequential change-point detection has been applied to a large number of engineering applications, such as statistical quality control [9], financial time series change detection [10], reliability [11], surveillance system [12] and system prognostic [13]. In Chapter 4, 5 and 6, we consider two related subtopics in the sequential change-point detection. One is multi-sensor gradual change-detection and another is robust sequential change-point detection. In the next two paragraphs, we explain them correspondingly.

As an enabling component for modern intelligent systems, multi-sensory monitoring has been widely deployed for large scale systems, such as manufacturing systems [14], [15], power systems [16], and biological and chemical threat detection systems [17]. The sensors acquire a stream of observations, whose distribution changes when the state of the network is shifted due to an abnormality or threat. We would like to detect the change online as soon as possible after it occurs, while controlling the false alarm rate. When the change happens, typically only a small subset of sensors are affected by the change, which is a form of sparsity. A mixture statistic which utilizes this sparsity structure of this problem is presented in [18]. The asymptotic optimality of a related mixture statistic is established in [19]. Extensions and modifications of the mixture statistic that lead to optimal detection are considered in [20]. In the above references [18, 20], the change-point is assumed to cause a shift in the means of the observations by the affected sensors, which is good for modeling an abrupt change. However, in many applications above, the change-point is an onset of system degradation, which causes a gradual change to the sensor observations. Often such a gradual change can be well approximated by a *slope* change in the means of the observations. We present a mixture procedure that detects a change-point causing a slope change to the means of the observations, which can be a model for gradual degradations. We

derive the log-likelihood ratio statistic, which becomes applying a soft-thresholding to the local statistic at each sensor and then combining the results. The mixture procedure raises an alarm whenever the statistic exceeds a prescribed threshold. Moreover, we prove the asymptotic optimality of our work and demonstrate the good performance of the proposed method using simulation and real-data examples.

In multi-sensor change-point detection problem, people consider to develop good detection procedures with a high dimension of observed signals, while the robust sequential change-point detection aims to deal with the case when the pre-change or the post-change distributions are unknown. It is necessarily to take the robustness into consideration since classic methods are sensitive to the mismatch of the distribution assumptions. We propose two techniques to handle the unknown distribution parameters. The first one is to assume a convex set for the parameters and then solve the best detection procedure via convex optimization offline before we apply the procedure sequentially. The second one is to combine the sequential detection procedure with the online convex optimization technique, namely, we learn the distribution parameters and detect the changes simultaneously. Both of the above methods improve the robustness and have better performance compared to the classic CUSUM procedure when there is a model mismatch. We prove the nearly optimality of the proposed methods, and we demonstrate the good performance of the methods by both simulation and real-data example.

CHAPTER 1
INTRODUCTION

Statistical signal processing and machine learning are very important in modern science and engineering. Many theories, methodsandtechniques are developed to help people extract and analyze the hidden information from the Big Data. Big Data has two aspects: 1) the huge number of observations;2) high-dimensionality of data. This dissertation focus on two specific topics that touch the above two aspects. The first topic is low-rank matrix completion. Because many survey data is capable of being represented approximately as a low-rank matrix, this model can apply to many applications such as sensor network, network traffic analysis,sensor localization, recommender systems and natural language processing. Even if much success has been achieved to learn the model with continuous noise, only a few works consider the quantized noise such as Poisson noise, which is applied to many real applications with count data. This motivates us to develop new algorithms and theories for the Poisson Matrix Completion. The second topic is sequential change-point detection. The model is that we observe a sequence of independent signals and would like to detect if an unexpected change happens in the system. This belongs to sequential decision-making problems and we need to decide sequentially whether an alarm of change detected should be raised as we obtain new observations. Due to the nature of sequential decision making, sequential change-point detection has been applied to a large number of engineering applications, such as statistical quality control, financial time series change detection, reliability, surveillance system and system prognostic. In the dissertation, several subtopics are considered, including multi-sensor gradual change detection and robust change-point detection via optimization techniques. For each subtopic, new theories and algorithms are developed.

In Chapter 1, I introduce the background and history for each topic and provide the insights and motivations. In Chapter 2, I review some preliminary results in mathematics, in order to make readers understand the proofs in the appendix more quickly. Specifically, I review some useful theorems in the random matrix theory since they play a crucial role in proving the performance bounds for matrix completion problem. Then, I review briefly the theory for classic one-sensor sequential change-point detection and present some recent progress about multi-sensor change-point detection. Besides, I briefly introduce the classic theory about the robust hypothesis testing and recent progress toward robust change-point detection.

In Chapter 3, I present the work about Poisson matrix completion.We extend the theory of low-rank matrix completion to the case when Poisson observations for a subset of the entries of a matrix are available, which arises in various applications with count data. We consider the usual matrix recovery formulation through maximum likelihood with proper regularization constraints on the matrix and establish theoretical upper and lower bounds on the recovery error.The bounds for matrix completion are nearly optimal up to a logarithmic factor on the order of the number of entries of the matrix. Then, we show an efficient algorithm that solves the penalized maximum likelihood approximately and demonstrates its performance on recovering solar flare images and bike sharing count data.

In Chapter 4, I present the work about multi-sensor sequential gradual change detection.We develop a mixture procedure for multi-sensor systems to monitor data streams for a change-point that causes a gradual degradation to a subset of the streams. Observations are assumed to be initially normal random variables with known constant means and variances. After the change-point, observations in the subset will have increasing or decreasing means. The subset and the rate-of-change are unknown. Our procedure uses a mixture statistics, which assumes that each sensor is affected by the change-point with probability p_0. Analytic expressions are obtained for the average run length and the expected detection delay of the mixture procedure, which are demonstrated to be quite accurate numerically. We establish the asymptotic optimality of the mixture procedure. Numerical examples demonstrate the good performance of the proposed procedure.We also discuss an adaptive mixture procedure using empirical Bayes.

In Chapter 5, I present the work about robust change detection via offline convex optimization.We address the computational challenge of finding the robust sequential change-point detection procedures when the pre- and post-change distributions are not completely specified. To tackle the difficulties of looking for least favorable distributions (LFDs) in high-dimensional settings, we present a method based on convex optimization that addresses this issue when the distributions are Gaussian with unknown parameters from pre-specified uncertainty sets. We also establish theoretical properties of our robust procedures, and numerical examples demonstrate their good performance.

In Chapter 6, I present the work about robust change detection via online convex optimization. We consider a set of detection procedures based on sequential likelihood ratios with non-anticipating estimators constructed using online convex optimization algorithms such as online mirror descent, which provides a more versatile approach to tackling complex situations where recursive maximum likelihood estimators cannot be found. When the underlying distributions belong to an exponential family and the estimators satisfy the logarithm regret property, we show that this approach is nearly second-order asymptotically optimal. This means that the upper bound for the false alarm rate of the algorithm (measured by the average-run-length) meets the lower bound asymptotically up to a log-log factor when the threshold tends to infinity. Our proof is achieved by making a connection between sequential change-point and online convex optimization and leveraging the logarithmic regret bound property of online mirror descent algorithm. Numerical and real data examples validate our theory.

1.1 Poisson matrix completion

Recovering a low-rank matrix M with Poisson observations is a key problem that arises from various real-world applications with count data, such as nuclear medicine, low-dose x-ray imaging [21], network traffic analysis [2], and call center data [22]. There the observations are Poisson counts whose intensities are determined by the matrix, either through a subset of its entries or linear combinations of its entries.

Thus far much success has been achieved in solving the matrix completion and recovery problems using nuclear norm minimization, partly inspired by the theory of compressed sensing [23, 24]. It has been shown that when M is low rank, it can be recovered from observations of a subset or a linear combination of its entries (see, e.g.[7, 25, 26, 27, 28, 29, 30, 31, 32]). Earlier work on matrix completion typically assume that the observations are noiseless, i.e., we may directly observe a subset of entries of M. In the real world, however, the observations are noisy, which is the focus of the subsequent work [33, 34, 35, 36, 37, 38], most of which consider a scenario when the observations are contaminated by Gaussian noise. The theory for low-rank matrix recovery under Poisson noise has been less developed. Moreover, the Poisson problems are quite different from their Gaussian counterpart, since under Poisson noise the variance of the noisy observations is proportional to the signal intensity. Moreover, instead of using ℓ_2 error for data fit, we need to use a highly non-linear likelihood function.

Recently there has also been work that consider the more general noise models, including noisy 1-bit observations [8], which may be viewed as a case where the observations are Bernoulli random variables whose parameters depend on a underlying low-rank matrix; [39, 40] consider the case where *all* entries of the low-rank matrix are observed and the observations are Poisson counts of the entries of the underlying matrix, and an upper bound is established (without a lower bound). In the compressed sensing literature, there is a line of research for sparse signal recovery in the presence of Poisson noise [41, 42, 43] and the corresponding performance bounds. The recently developed SCOPT [44, 45] algorithm can also be used to solve the Poisson compressed sensing of sparse signals but may not be directly applicable for Poisson matrix recovery.

We extend the theory of 1-bit low-rank matrix completion to the case with Poisson observations. The matrix recovery problem from compressive measurements is formulated as a regularized maximum likelihood estimator with Poisson likelihood. We establish performance bounds by combining techniques for recovering sparse signals under Poisson noise [41] and for establishing bounds in the case of low-rank matrices [46, 47]. Our results demonstrate that as the intensity of the signal increases, the upper bound on the normalized error decays at certain rate depending how well the matrix can be approximated by a low-rank matrix.

The matrix completion problem from partial observations is formulated as a maximum likelihood problem with proper constraints on the matrix M (nuclear norm bound $\|M\|_* \leq \alpha\sqrt{rd_1d_2}$ for some constant α and bounded entries $\beta \leq M_{ij} \leq \alpha$)[1]. We also establish upper and lower bounds on the recovery error, by adapting the arguments used for one-bit matrix completion [8]. The upper and lower bounds nearly match up to a factor on the order of $\mathcal{O}(\log(d_1d_2))$, which shows that the convex relaxation formulation for Poisson matrix completion is nearly optimal. We conjecture that such a gap is inherent to the Poisson problem in the sense that it may not be an artifact due to our proof techniques for the upper bound. Moreover, we also highlight a few important distinctions of Poisson matrix completion compared to the prior work on matrix completion in the absence of noise and with Gaussian noise: (1) Although our arguments are adapted from one-bit matrix completion (where the upper and lower bounds nearly match), in the Poisson case there will be a gap between the upper and lower bounds, possibly due to the fact that Poisson distribution is only locally sub-Gaussian. In our proof, we notice that the arguments based on bounding all moments of the observations, which usually generate tight bounds for prior results with sub-Gaussian observations, do not generate tight bounds here; (2) we will need a lower bound on each matrix entry in the maximum likelihood formulation, which can be viewed as a requirement for the lowest signal-to-noise ratio (since the signal-to-noise ratio

[1] Note that the formulation differs from the one-bit matrix completion case in that we also require a lower bound on each entry of the matrix. This is consistent with an intuition that the value of each entry can be viewed as the signal-to-noise ratio (SNR) for a Poisson observation, and hence this essentially poses a requirement for the minimum SNR.

(SNR) of a Poisson observation with intensity I is $\sqrt{I}$).

Moreover, we present a set of efficient algorithms, which can be used for both matrix recovery based on compressive measurements or based on partial observations. These include two generic (gradient decent based) algorithms: the proximal and accelerated proximal gradient descent methods, and an algorithm tailored to Poisson problems called the Penalized Maximum Likelihood Singular Value Threshold (PMLSVT) method. PMLSVT is derived by expanding the likelihood function locally in each iteration, and finding an exact solution to the local approximation problem which results in a simple singular value thresholding procedure [30]. The performance of the two generic algorithms are analyzed theoretically. PMLSVT is related to [48, 49, 50] and can be viewed as a special case where a simple closed form solution for the algorithm exists. Good performance of PMLSVT is demonstrated with synthetic and real data including solar flare images and bike sharing count data. We show that PMLSVT has much lower complexity than solving the problem directly via semidefinite program and it has fairly good accuracy.

While working on this paper we realize a parallel work [51] which also studies performance bounds for low rank matrix completion with exponential family noise and using a different approach for proof (Poisson noise is a special case of theirs). Their upper bound for the mean square error (MSE) is on the order of $\mathcal{O}\left(\log(d_1+d_2)r\max\{d_1,d_2\}/m\right)$ (our upper bound is $\mathcal{O}\left(\log(d_1d_2)[r(d_1+d_2)/m]^{1/2}\right)$), and their lower bound is on the order of $\mathcal{O}\left(r\max\{d_1,d_2\}/m\right)$ (versus our lower bound is $\mathcal{O}\left([r(d_1+d_2)/m]^{1/2}\right)$. There might be two reasons for the difference. First, our sampling model (similar to one bit matrix completion in [8]) assumes *sampling without replacement*; therefore there are at most d_1d_2 observations, and each entry may be observed at most once. In contrast, [51] assumes *sampling with replacement*; therefore there can be multiple observations for the same entry. Since our results heavily depend on the sampling model, we suspect this may be a main source of the difference. The formulations are also different. The formulation for matrix completion in our paper is a constrained optimization with an exact *upper bound on the* *matrix nuclear norm*, whereas [51] uses a regularized optimization with a regularization parameter λ (which is indirectly related to the nuclear norm of the solution), but there is no direct control of the matrix nuclear norm. Also, note that their upper and lower bounds also have a gap on the order of $\log(d_1 + d_2)$, which is consistent with our result. On the other hand, compared with the more general framework for M-estimator [52], our results are specific to the Poisson case, which may possibly be stronger but do not apply generally.

1.2 Multi-sensor slope change detection

As an enabling component for modern intelligent systems, multi-sensory monitoring has been widely deployed for large scale systems, such as manufacturing systems [14], [15], power systems [16], and biological and chemical threat detection systems [17]. The sensors acquire a stream of observations, whose distribution changes when the state of the network is shifted due to an abnormality or threat. We would like to detect the change online as soon as possible after it occurs, while controlling the false alarm rate. When the change happens, typically only a small subset of sensors are affected by the change, which is a form of sparsity. A mixture statistic which utilizes this sparsity structure of this problem is presented in [18]. The asymptotic optimality of a related mixture statistic is established in [19]. Extensions and modifications of the mixture statistic that lead to optimal detection are considered in [20].

In the above references [18, 20], the change-point is assumed to cause a shift in the means of the observations by the affected sensors, which is good for modeling an abrupt change. However, in many applications above, the change-point is an onset of system degradation, which causes a gradual change to the sensor observations. Often such a gradual change can be well approximated by a *slope* change in the means of the observations. One such example is shown in Fig. 4.1, where multiple sensors monitor an aircraft engine and each panel of figure shows the readings of one sensor. At some time a degradation initiates and causes decreasing or increasing in the means of the observations. Another example

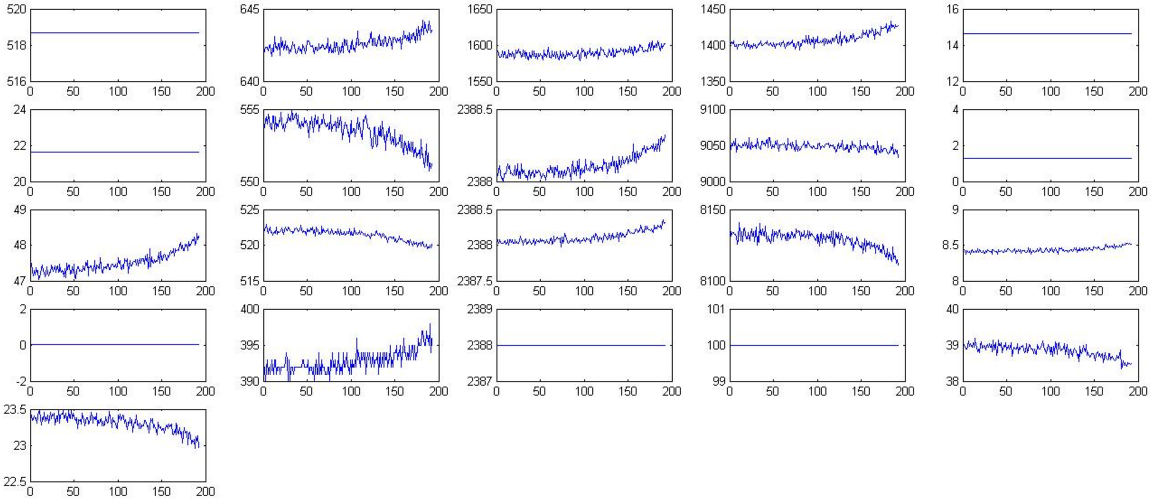


Figure 1.1: Degradation sample paths recorded by 21 sensors, generated by C-MAPSS [53]. A subset of sensors are affected by the change-point, which happens at an unknown time simultaneously and it causes a change in the slopes of the signals. The change can cause either an increase or decrease in the means.

comes from power networks, where there are thousands of sensors monitoring hundreds of transformers in the network. We would like to detect the onset of any degradation in real-time and predict the residual life time of a transformer before it breaks down and causes a major power failure.

We present a mixture procedure that detects a change-point causing a slope change to the means of the observations, which can be a model for gradual degradations. Assume the observations at each sensor are $i.i.d.$ normal random variables with constant means. After the change, observations at the sensors affected by the change-point become normal distributed with increasing or decreasing means. The subset of sensors that are affected are unknown. Moreover, their rate-of-changes are also unknown. Our mixture procedure assumes that each sensor is affected with probability p_0 independently, which is a guess for the true fraction p of sensors affected. When p_0 is small, this captures an empirical fact that typically only a small fraction of sensors are affected. With such a model, we derive the log-likelihood ratio statistic, which becomes applying a soft-thresholding to the local statistic at each sensor and then combining the results. The mixture procedure fires an alarm whenever the statistic exceeds a prescribed threshold. We consider two versions of the mixture procedure that compute the local sensor statistic differently: the *mixture CUSUM procedure* T_1, which assumes some nominal values for the unknown rate-of-change parameters, and the *mixture generalized likelihood ration (GLR) procedure* T_2, which uses the maximum likelihood estimates for these parameters. To characterize the performance of the mixture procedure, we present theoretical approximations for two commonly used performance metrics, the average run length (ARL) and the expected detection delay (EDD). Our approximations are shown to be highly accurate numerically and this is useful in choosing a threshold of the procedure. We also establish the asymptotic optimality of the mixture procedures. Good performance of the mixture procedure is demonstrated via real-data examples, including: (1) detecting a change in the trends of financial time series; (2) predicting the life of air-craft engines using the Turbofan engine degradation simulation dataset.

The mixture procedure here can be viewed as an extension of the earlier work on multi-sensor mixture procedure for detecting mean shifts [18]. The extensions of theoretical approximations to EDD and especially to ARL are highly non-trivial, because of the non-i.i.d. distributions in the slope change problem. Moreover, we also establish some new optimality results which were omitted from [18], by extending the results in [54] and [55] to handle non-$i.i.d.$ distributions in our setting. In particular, we generalize the theory to a scenario where the log likelihood ratio grows polynomially as a result of linear increase or decrease of the mean values, whereas in [18], the log-likelihood ratio grows linearly. A related recent work [19] studies optimality of the multi-sensor mixture procedure for $i.i.d.$ observations, but the results therein do not apply to the slope change case here.

1.3 Robust sequential change detection

Sequential detection of an abrupt change in a system has rich applications in practice such as statistical quality control, seismic event identification and network security monitoring (e.g, [56]). In standard settings, these change detection problems assume that we observe a sequence of signal whose distribution changes at some unknown point in time, referred to as the “change-point”. The goal is to detect the change-point with as little delay as possible, subject to the constraint that false detections occurring before the true change-point are very rare. This single sequence case is first studied by Page over 60 years ago in [57] and then many outstanding contributions are due to [58] and [59]. As the growing complexity of the systems and the enlarging number of sensors needed to monitor the systems, the detection procedures that can collect and analyze information from multiple sources are established, such as [60], [61] and [18].

Most previous work in the area of sequential change-point detection is based on an assumption that the distributions before and after the change-point are exactly specified. Under this assumption, it is well known that this optimal procedure is CUSUM procedure and many great theoretical analysis such as asymptotical and exact optimality are achieved (e.g, [59, 62, 63]). To implement CUSUM procedure in practice, people need estimate the parameters in the distributions based on historical data or domain knowledge. However, it is known in [64] that CUSUM procedure is sensitive to the model mismatch. Therefore, a more appropriate procedure should be more robust to the uncertainty of the distributions possibly caused by noise, estimation error and inevitable system error. In the following, I present two techniques to establish robust sequential detection procedure. The first one utilizes the offline convex optimization to solve the best detector before we apply the procedure. The second one utilizes the online convex optimization aiming to jointly estimating the distribution parameters and detecting the change.

1.3.1 Sequential change detection via offline convex optimization

To improve the robustness of the detection procedure, one can use robust detector to form the statistics. Robust detector dates back to Huber’s seminal work [65]. Also an asymptotic version of the robust problem was introduced in [66]. The more recent contributions [67, 68] introduce a so-called Joint Stochastic Boundedness (JSB), under which one can identify a pair of least favorable distributions (LFDs) from the uncertainty classes such that the CUSUM procedure designed for the LFDs is the optimal for the robust problem in a minimax sense. However, in the multi-dimensional setting, there remains the computational challenge to establish robust sequential detection procedures or to find the LFDs. Closed-form LFDs are found only for a few special cases (e.g,[69] and [70]) for one-dimensional case. Moreover, the JSB condition in [68] is defined on the real line, and direct extension of JSB to multi-dimensional setting becomes quite restrictive even in very simple cases. The following example illustrates the difficulty. Consider two bivariate normal distributions. Assume that Σ is a positive-definite matrix in $\mathbb{R}^{2\times 2}$, and we would like to detect a possible transition from the probability density function $\mathcal{P}_0 = \{\mathcal{N}(0, \Sigma)\}$, to a family of distributions

$$\mathcal{P}_1 = \{\mathbb{P} \mid \mathbb{P} = \mathcal{N}(\mu_1, \Sigma), \|\mu_1 - (10, 10)^T\|_2 \leq 1, \mu_1 \in \mathbb{R}^2\}.$$

In this case, it is impossible to find a distribution in $\mathcal{P}_1$ that is stochastically larger than any other distribution in $\mathcal{P}_1$ due the following Lemma 1 (see Fig. 1.2 for the illustration) which satisfies the JSB condition.

Lemma 1 (Theorem 5 in [71])**.** *Let $X \sim \mathcal{N}(\mu, \Sigma)$ and $X' \sim \mathcal{N}(\mu', \Sigma')$ be n-dimensional normally distributed random vectors. Then X' is stochastically larger than X if and only if $\mu'^{(i)} \geq \mu^{(i)}$, for all $1 \leq i \leq n$ and $\Sigma = \Sigma'$, where $\mu^{(i)}$ denotes the ith entry of μ.*

I present a method of establishing the sequential detection procedures by convex optimization. This work is inspired by the recent work for robust hypothesis testing [72]. Given the convex set to which the parameters belong, instead of identifying LFDs under restrictive assumptions, we solve the best choice of parameters by minimizing the Hellinger distance between the distributions from uncertainty classes. Then we establish the CUSUM procedure based on the parameters given by convex optimization. Even if we no longer be able to prove the optimality of our procedures as a cost for the robustness, we can provide some useful theoretical analysis and introduce the near optimality. Finally, note that we fix

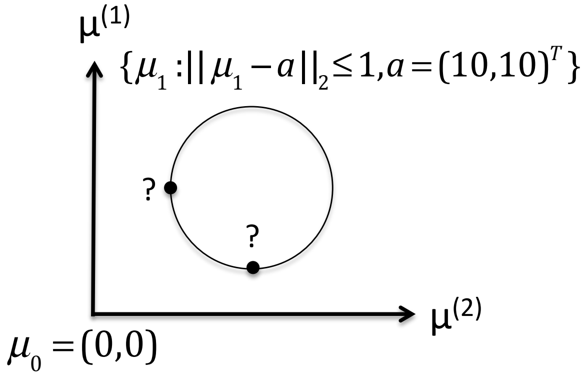


Figure 1.2: It is impossible to find a point on the circle of which all the entries are larger than those of other points on the circle.

the parametric model first (e.g, Normal and Poisson) and the uncertainty class in our work is represented by a convex set of parameters, which is different with the previous work about identifying LFDs where the uncertainty class is represented by a set of probability functions (and our approach leads to computationally more efficient methods).

Our approach is motivated by the recent work using convex optimization for hypothesis testing [72, 73]. The difference of these approaches from our work is that they treat sequential change-point detection as a multiple hypothesis test problem. Since for each time t, there are k possible change-point locations, for a fixed time horizon $t \leq T$ there are finite number of hypotheses. One may design the test such that the probability of error for each of the hypothesis is uniformly controlled and the total probability of error is less than a given level α. This approach may not be convenient to use for infinite horizon setting of the sequential change-point detection problem. In this paper, we adopt a different approach and also characterize two performance metrics that are commonly used for sequential problems: the Average Run Length (ARL) and Expected Detection Delay (EDD).

1.3.2 Sequential change detection via online convex optimization

We are interested in the sequential change-point detection problem with *known* pre-change parameters but *unknown* post-change parameters. Specifically, given a sequence of samples

X_1, X_2, $\ldots$, we assume that they are independent and identically distributed (i.i.d.) with certain distribution f_θ parameterized by θ, and the values of θ are different before and after some unknown time called the *change-point*. We further assume that the parameters before the change-point are known. This is reasonable since usually it is relatively easy to obtain the reference data for the normal state, so that the parameters in the normal state can be estimated with good accuracy. After the change-point, however, the values of the parameters switch to some *unknown* values, which represent anomalies or novelties that need to be discovered.

Motivation: Dilemma of CUSUM and generalized likelihood ratio (GLR) statistics

Consider change-point detection with unknown post-change parameters. A commonly used change-point detection method is the so-called CUSUM procedure [55] that can be derived from likelihood ratios. Assume that before the change, the samples X_i follow a distribution f_{θ_0} and after the change the samples X_i follow another distribution f_{θ_1}. CUSUM procedure has a recursive structure: initialized with $W_0 = 0$, the likelihood-ratio statistic can be computed according to $W_{t+1} = \max\{W_t + \log(f_{\theta_1}(X_{t+1})/f_{\theta_0}(X_{t+1})), 0\}$, and a change-point is detected whenever W_t exceeds a pre-specified threshold. Due to the recursive structure, CUSUM is memory and computation efficient since it does not need to store the historical data and only needs to record the value of W_t. The performance of CUSUM depends on the choice of the post-change parameter θ_1; in particular, there must be a well-defined notion of "distance" between θ_0 and θ_1. However, the choice of θ_1 is somewhat subjective. Even if in practice a reasonable choice of θ_1 is the "smallest" change-of-interest, in the multi-dimensional setting, it is hard to define what the "smallest" change would mean. Moreover, when the assumed parameter θ_1 deviates significantly from the true parameter value, CUSUM may suffer a severe performance degradation [74].

An alternative approach is the Generalized Likelihood Ratio (GLR) statistic based procedure [56]. The GLR statistic finds the maximum likelihood estimate (MLE) of the post-change parameter and plugs it back to the likelihood ratio to form the detection statistic. To be more precise, for each hypothetical change-point location k, the corresponding post-change samples are $\{X_{k+1}, \ldots, X_t\}$. Using these samples, one can form the MLE denoted as $\hat{\theta}_{k+1,t}$. Without knowing whether the change occurs and where it occurs beforehand when forming the GLR statistic, we have to maximize k over all possible change locations. The GLR statistic is given by $\max_{k<t} \sum_{i=k+1}^{t} \log(f_{\hat{\theta}_{k,t}}(X_i)/f_{\theta_0}(X_t))$, and a change is announced whenever it exceeds a pre-specified threshold. The GLR statistic is more robust than CUSUM [54], and it is particularly useful when the post-change parameter may vary from one situation to another. In simple cases, the MLE $\hat{\theta}_{k+1,t}$ may have closed-form expressions and may be evaluated recursively. For instance, when the post-change distribution is Gaussian with mean θ [75], $\hat{\theta}_{k+1,t} = (\sum_{i=k+1}^{t} X_i)/(t-k)$, and $\hat{\theta}_{k+1,t+1} = (t-k)/(t-k+1) \cdot \hat{\theta}_{k+1,t} + X_{t+1}/(t-k+1)$. However, in more complex situations, in general MLE $\hat{\theta}_{k+1,t}$ does not have recursive form and cannot be evaluated using simple summary statistics. One such instance is given in Section 1.3.2. Another instance is when there is a constraint on the MLE such as sparsity. In these cases, one has to store historical data and recompute the MLE $\hat{\theta}_{k,t}$ whenever there is new data, which is not memory efficient nor computational efficient. For these cases, as a remedy, the window-limited GLR is usually considered, where only the past w samples are stored and the maximization is restricted to be over $k \in (t-w, t]$. However, even with the window-limited GLR, one still has to recompute $\hat{\theta}_{k,t}$ using historical data whenever the new data are added.

Besides CUSUM or GLR, various online change-point detection procedures using one-sample updates have been considered, which replace with the MLE with a simple recursive estimator. The one-sample update estimate takes the form of $\hat{\theta}_{k,t} = h(X_t, \hat{\theta}_{k,t-1})$ for some function h that uses only the most recent data and the previous estimate. Then the estimates are plugged into the likelihood ratio statistic to perform detection. Online convex optimization algorithms (such as online mirror descent) are natural approach to construct these estimators (see, e.g., [76, 77]). Such a scheme provides a more versatile approach to develop detecting procedure for complex situations, where the exact MLE does not have a recursive form or even a closed-form expression. The one-sample update enjoys efficient computation, as information from the new data can be incorporated via low computational cost update. It is also memory efficient since the update only needs the most recent sample. The one sample update estimators may not correspond to the exact MLE, but they tend to result in good detection performance. However, in general there is no performance guarantees for such approach. This is the question we aim to address in this paper.

Application scenario: Social network change-point detection

The widespread use of social networks (such as Twitter) leads to a large amount of user-generated data generated continuously. One important aspect is to detect change-points in streaming social network data. These change-points may represent the collective anticipation of response to external events or system "shocks" [78]. Detecting such changes can provide a better understanding of patterns of social life. In social networks, a common form of the data is discrete events over continuous time. As a simplification, each event contains a time label and a user label in the network. In [79], the authors model discrete events using network point processes, which capture the influence between users through an *influence matrix*. We then cast the problem as detecting changes in an influence matrix, assuming that the influence matrix in the normal state (before the change) can be estimated from the reference data. After the change, the influence matrix is unknown (since it represents an anomaly) and has to be estimated online. Due to computational burden and memory constraint, since the scale of the network tends to be large, we do not want to store the entire historical data and rather compute the statistic in real-time.

Contributions

First, we present a general approach based on online convex optimization (OCO) for constructing the estimator for the one-sided sequential hypothesis test and the sequential change-point detection, in the non-anticipative approach of [75] if the MLE cannot be computed in a convenient recursive form.

Second, we provide a proof of the near second-order asymptotic optimality of this approach when a "logarithmic regret property" is satisfied and when the distributions are from an exponential family. The nearly second-order asymptotic optimality [55] means that the upper bound for performance matches the lower bound up to a log-log factor as the false-alarm rate tends to zero. Inspired by the existing connection between sequential analysis and online convex optimization in [80, 81], we prove the near optimality leveraging the logarithmic regret property of online mirror descent (OMD) and the lower bound established in statistical sequential change-point literature [82, 55]. More precisely, we provide a general upper bound for one-sided sequential hypothesis test and change-point detection procedures with the one-sample update schemes. The upper bound explicitly captures the impact of estimation on detection by an estimation algorithm dependent factor. This factor shows up as an additional term in the upper bound for the expected detection delay, and it corresponds to the regret incurred by the one-sample update estimators. This establishes an interesting linkage between sequential change-point detection and online convex optimization. Although both fields, sequential change-point detection and online convex optimization, study sequential data, the precise connection between them is not clear, partly because the performance metrics are different: the former concerns with the tradeoff between average run length and detection delay, whereas the latter focuses on bounding the cumulative loss incurred by the sequence of estimators through a regret bound [83, 81]. Synthetic examples validate the performances of one sample update schemes. Here we focus on OMD estimators, but the results can be generalized to other OCO schemes such as the online gradient descent.

CHAPTER 2
PRELIMINARIES

In this chapter, we review some fundamentals that will be used for later development. I will review some results and techniques about matrix completion. Then, some deep and useful theorems in the random matrix theory are introduced since they play a crucial role in proving the performance bounds. Next, I will review briefly the theory for classic one-sensor sequential change-point detection and present some recent progress about multi-sensor change-point detection. Finally, I will briefly introduce the classic theory about the robust hypothesis testing and recent progress about robust change-point detection.

2.1 Basic mathematics and results for matrix completion

2.1.1 Basic review of mathematical concepts and tools

In order to prove the theorems in [8] and in Chapter 3, we need have some basic understandings of mathematics including matrix norms, divergence between distributions, information theory and probability theory. Next, I will offer a quick review of these contents.

Matrix norms

As is known, the matrix norms are all equivalent in Hilbert space with certain inner product. Then, people specify them for the convenience of analysis. The Schatten norms, as one kind of specification, are widely used. Assume that M is an d_1-by-d_2 matrix, and $\sigma_1, \ldots, \sigma_{\min(d_1,d_2)}$ are the singular values of M. Then the Schatten p-norm of matrix M is defined as follows:

$$\|M\|_p = \left(\sum_{i=1}^{\min(d_1,d_2)} \sigma_i^p \right)^{1/p}.$$

Many useful norms such as Frobenius norm and Spectral norm are special cases of Schatten norms by taking various ps. Next, I explain the relationship between the norms.

When $p = 2$, The Schatten 2-norm is just the Frobenius norm, which is defined as the sum of square of all the entries of a matrix. When $p = \infty$, the Schatten ∞-norm is just the spectral norm, which is defined as the largest singular value of a matrix. When $p = 1$, the Schatten 1-norm is just the nuclear norm, which is defined as the sum of the singular values of a matrix.

Define an inner product $\langle A, B\rangle$ in matrix space $\mathbb{R}^{d_1\times d_2}$ (matrix space is a Hilbert space) as follows:

$$\langle A, B\rangle \triangleq \text{Trace}(A^\intercal B) = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} A_{ij}B_{ij}.$$

Moreover, we define the dual norms. For any (p, q) satisfying $p, q > 0$ and $1/p + 1/q = 1$, then Schatten p-norm and Schatten q-norm are called dual norms. For example, the dual norm of Spectral norm is the nuclear norm and the dual norm of Frobenius norm is itself.

Then the following Hölder's inequality holds for any pair of dual Schatten p-norms.

Lemma 2 (Theorem 2 in [84], Matrix Hölder's inequality)**.** *For any matrix $A, B \in \mathbb{R}^{d_1\times d_2}$, if $1 \le p, q \le \infty$ satisfying $1/p + 1/q = 1$, with the Schatten p-norms and inner product defined before, we have that*

$$\langle A, B\rangle \le \|A\|_p\|B\|_q.$$

This lemma is useful since we are able to bound the inner product by norms that may offer some convenience to the analysis.

Other than $p = 1, 2, \infty$, the Scattern p-norms are not very popular so in the following we denote $\|M\|_F$ as the Frobenius norm, $\|M\|$ as the spectral norm and $\|M\|_*$ as the nuclear norm. Therefore, the above lemma offers an useful result: $\langle A, B\rangle \le \|A\|\|B\|_*$, which is used in the proofs of theorems in Chapter 3.

Rademacher random variables and contraction principle

In probability theory, one popular proof technique is to take advantage of the symmetry of random variables. Rademacher random variables, as the simplest symmetric random variable, are widely used in the probability theory. To obtain a broad horizon, I suggest read the book [85]. Next, I introduce some essential concepts and results used in this thesis.

Denote B as a Banach space such that there exists a countable subset D of the unit ball or sphere of the dual space B' such that

$$\|x\| = \sup_{f \in D} |f(x)|, x \in B.$$

If B is separable, then such D exists (e.g, linear functionals of norm 1) and $\|x\|$ is a norm of x defined on B.

The Rademacher sequence (or Bernoulli sequence) is defined to be a sequence $(\epsilon_i)_{i \in \mathbb{N}}$ of independent Rademacher random variables taking the values $+1$ and -1 with equal probability. A sequence $(X_i)_{i \in \mathbb{N}}$ of random variables with values in B is called symmetric sequence if, for every $N \in \mathbb{N}$, $(\pm X_1, \dots, \pm X_N)$ has the same probability distribution with $(X_1, \dots, X_N)$ in B^N. The typical example of a symmetric sequence consists in a sequence of independent and symmetric random variables. Then, an important result about the partial sum of a symmetric sequence is given by Lévy as follows.

Lemma 3 (Proposition 2.3 in [85]). *Let (X_i) be a symmetric sequence of random variables with values in Banach space B. For every k, set $S_k = \sum_{i=1}^{k} X_i$. Then, for any integer N and $t > 0$, we have*

$$\mathbb{P}\{\max_{k \leq N} \|S_k\| > t\} \leq 2\mathbb{P}\{\|S_k\| > t\}.$$

As a consequence, note also that by integration by parts, for every $p > 0$, we have

$$\mathbb{E} \max_{k \leq N} \|S_k\|^p \leq 2\mathbb{E}\|S_k\|^p.$$

Using L(́e)vy's inequalities above, we can prove the following fundamental theorem, which is known as the contraction principle.

Theorem 1 (Theorem 4.4 in [85], Contraction principle)**.** *Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be convex. For any finite sequence (x_i) in a Banach space B and any real numbers (α_i) satisfying $|\alpha_i| \leq 1$ for all i, we have*

$$\mathbb{E}F\left(\left\|\sum_i \alpha_i \epsilon_i x_i\right\|\right) \leq \mathbb{E}F\left(\left\|\sum_i \epsilon_i x_i\right\|\right).$$

Further, for any $t > 0$,

$$\mathbb{P}\left\{\left\|\sum_i \alpha_i \epsilon_i x_i\right\| > t\right\} \leq 2\mathbb{P}\left\{\left\|\sum_i \epsilon_i x_i\right\| > t\right\}.$$

Next, we define the contraction. A map $\phi : \mathbb{R} \to \mathbb{R}$ is called a contraction if $|\phi(s) - \phi(t)| \leq |s-t|$ for all $s, t \in \mathbb{R}$. If we only look at the vector space and add some assumptions, we can replace α_i in Theorem 1 with contractions. The result is summarized in the following theorem.

Theorem 2 (Theorem 4.12 in [85])**.** *Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be convex and increasing. Let $\phi_i, i \leq N$ are contractions satisfying $\phi_i(0) = 0$. Then, for any bounded subset T in $\mathbb{R}^N$, we have*

$$\mathbb{E}F\left(\frac{1}{2}\left\|\sum_{i=1}^N \epsilon_i \phi_i(t_i)\right\|_T\right) \leq \mathbb{E}F\left(\left\|\sum_{i=1}^N \epsilon_i t_i\right\|_T\right), \tag{2.1}$$

where $\|h(t)\|_T \triangleq \sup_{t\in T} |h(t)|$.

Note that the right-hand side of (2.1) is not dependent on the contractions $\phi_i, i \leq N$. Therefore, the above theorem offers a bound without possibly complicated functions ϕ_is and then makes the analysis more easily. Finally, we offer a simple but useful result for Rademacher random variables.

Lemma 4 (Lemma 6.3 in [85])**.** *Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be convex. Then for any finite sequence (X_i) of independent mean zero random variables in B such that $\mathbb{E}F(\|X_i\|) < \infty$ for all i,* *then we have*

$$\mathbb{E}F\left(\frac{1}{2}\left\|\sum_i \epsilon_i X_i\right\|\right) \leq \mathbb{E}F\left(\left\|\sum_i X_i\right\|\right) \leq \mathbb{E}F\left(\left\|\sum_i \epsilon_i X_i\right\|\right).$$

Equipped with the above knowledge in probability theory, one should be well prepared to understand the proofs in Chapter 3.

Kullback-Leibler(KL) divergence and Hellinger distance

Similar with the Euclidean distance in vector space, KL divergence and Helliger distance are used as a measure of the difference between two probability distributions. For discrete probability distributions P and Q, KL divergence from Q to P is defined to be

$$D_{KL}(P\|Q) \triangleq \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

Then, the Hellinger distance between two discrete probability distributions P and Q is defined to be

$$H^2(P,Q) \triangleq \frac{1}{2}\sum_i (\sqrt{P(i)} - \sqrt{Q(i)})^2.$$

Note that KL divergence is not symmetric with respect to P and Q but Hellinger distance is symmetric. A simple result is that the Hellinger distance can be bounded by KL divergence (Equation (12) in Chapter 3 of [86]). I suggest reading the whole Chapter 3 in [86] to obtain more intuition and knowledge.

Fano's inequality

In information theory, Fano's inequality relates the average information lost in a noisy channel to the probability of the categorization error. In fact, Fano's inequality has its statistical interpretation. I refer to Lemma 3 in [87] for the people who are interested to the generalization of Fano's result.

2.1.2 classic matrix completion theory

When it comes to the theory of matrix completion, people need ask three questions: which matrices can be completed; which sample schemes is reasonable; which algorithm is appropriate to guarantee the completion accuracy. Matrix completion is not magic but logic. The first work that answers the above three questions clearly is presented by Dr. Candes and Dr. Recht in [7]. Next, I will introduce the basic ideas and results and their answers about the above three questions.

First, they make some assumptions about the matrices to be completed. Assume that the matrix $M \in \mathbb{R}^{d_1 \times d_2}$ is the matrix to be completed. A very important observation is that the singular vectors of M need to be sufficiently spread in order to minimize the number of observations needed to recovery M. They present a extreme example to show the intuition. Consider a rank-2 symmetric matrix M given by

$$M = \sum_{k=1}^{2} \sigma_k u_k u_k^{\intercal}, u_1 = (e_1 + e_2)/\sqrt{2}, u_2 = (e_1 - e_2)/\sqrt{2},$$

where the σ_k are arbitrary singular values and e_i is the ith canonical basis vector in Euclidean space (the vector with all entries equal to 0 but the ith equal to 1). Then M is zero everywhere except the top-left 2×2 corner. Intuitively, one can not recovery M exactly without observing the top-left corner entries.

Therefore, to explain this case mathematically, they define the coherence of a subspace $U \in \mathbb{R}^d$ of dimension r as following:

$$\mu(U) \triangleq \frac{d}{r} \cdot \max_{1 \leq i \leq d} \|\mathcal{P}_U e_i\|_2^2,$$

where $\mathcal{P}_U$ is the orthogonal projection onto subspace U. One can see that $1 \leq \mu(U) \leq d/r$ for any subspace $U \in \mathbb{R}^d$ with dimension r. They are interested in subspaces with low coherence as matrices whose column and row spaces have low coherence cannot really be in the null space of the sampling operator.

Assume the singular value decomposition of M is $M = \sum_{k=1}^{r} u_k v_k^\intercal$ and with column and row spaces denoted by U and V, respectively. In this case, $\mathcal{P}_U = \sum_{i\in[r]} u_i u_i^\intercal$ and $\mathcal{P}_V = \sum_{i\in[r]} v_i v_i^\intercal$, where $[r]$ is not an exact definition but represents the index set of spanned spaces. To answer the question about which matrices can be completed with incomplete observations, they make two assumptions:

- (A0): $\max(\mu(U), \mu(V)) \leq \mu_0$ for some positive μ_0.
- (A1): The matrix $\sum_{1\leq k\leq r} u_k v_k^\intercal$ has a maximum entry bounded by $\mu_1\sqrt{r/(d_1 d_2)}$ in absolute value for some positive μ_1.

If the matrices to be completed satisfy the above two assumptions, they prove that it is enough to set sample scheme to be randomly uniform. Finally, they consider the algorithm used to complete the low-rank matrices. Ideally, they would like to minimize the rank of matrix M. However, the minimization of rank is proved to be a NP-hard problem that can not be solved in polynomial time. Therefore, they instead try to minimize the nuclear norm of M, which is defined as the sum of singular values of M. The nuclear norm of M can be seen as a convex relaxation (similar with relax ℓ_0 norm to ℓ_1 norm in LASSO problem). They consider the following optimization problem:

$$\begin{aligned} &\text{minimize } \|X\|_* \\ &\text{subject to } X_{ij} = M_{ij}, (i,j) \in \Omega \end{aligned} \tag{2.2}$$

Then, the main result is the following theorem

Theorem 3 (Theorem 1.3 in [7])**.** *Let M be an d_1-by-d_2 matrix of rank r obeying $(A0)$ and $(A1)$ and put $d = \max(d_1, d_2)$. Suppose we observe m entries of M with locations sampled uniformly at random. Then there exist constants C, c such that if*

$$m \geq \max(\mu_1^2, \mu_0^{1/2}\mu_1, \mu_0 d^{1/4}) dr(\beta \log d),$$

for some $\beta > 2$, then the minimizer of (2.2) is unique and equal to M with probability at least $1 - cd^{-\beta}$. For $r \leq \mu_0 d^{1/5}$ and under $(A0)$ only, this estimate can be improved to

$$m \geq C\mu_0 d^{6/5} r(\beta \log d),$$

with the same probability of success.

This theorem asserts that when the coherence is low, few observations are required to recover M with high probability. Also, the theorem suggests the appropriate relationship between the rank and the dimension for successfully recovering M from randomly sampled entries. This work is very important and offers the procedures of analyzing the problem. More refinement is presented in [26].

2.1.3 1-bit matrix completion

After Dr. Candès present the theories for exact matrix completion, people continue working on this topic and extend the work to many other cases, e.g, [88, 34]. However, the famous ”Netflix Problem” brings new problems that the classic literature for matrix completion does not consider, namely, the quantized matrix completion. Assume that there is some unknown matrix whose entries each represent a rating for a particular user on a particular movie. It is only able to observe a small fraction of the total entries in the matrix since any user rates only a small subset of movies. The ”Netflix Problem” then asks scientists to predict the unseen ratings from the observed ratings. This problem is clearly a matrix completion problem but there is a difference: the ratings are integers but are not continuous values. The quantization challenges classic matrix completion theories since the classic theories all assume continuous variables with possible continuous noise (e.g, Gaussian noise). In recommender system, this phenomenon occurs frequently and even in some cases the entries are categorical. Therefore, alternative mathematical techniques are needed in order to demonstrate the performance bounds in such quantized matrix completion problem.

The first work that considers the quantized observations is written by Dr. Davenport, Dr. Plan, Dr. Berg and Dr. Wootters in [8]. They answer the three questions again for the quantized matrix completion problem. Next, I will review the new theories briefly.

First question: which matrices can be completed. Similar with classic theories, the answer in [8] is that the matrix M should be approximately low-rank. Instead of rigorously setting the rank of M, authors in [8] presents that an upper bound for the nuclear norm of $M \in \mathbb{R}^{d_1 \times d_2}$ is enough. Specifically, they assume that $\|M\|_* \leq \alpha\sqrt{rd_1d_2}$ for some positive α and r. The intuition is that if we see α as the upper bound for $\|M\|_\infty$ and see r as the upper bound for the rank of M, then the following inequalities show an upper bound for the nuclear norm of M:

$$\|M\|_* \leq \sqrt{r}\|M\|_F \leq \sqrt{rd_1d_2}\|M\|_\infty \leq \alpha\sqrt{rd_1d_2}.$$

Second question: which sample schemes is reasonable. Define a subset of indices $\Omega \in [d_1] \times [d_2]$, where $[d] = \{1, 2, \ldots, d\}$ and assume that the entries in Ω are observed. Similar with the classic settings, the authors in [8] assume that Ω follows a binomial model in which each entry $(i, j) \in [d_1] \times [d_2]$ is included in Ω with probability $n/(d_1d_2)$, independently, where n is the number of observed entries.

Third question: which algorithm is appropriate to guarantee the completion accuracy. The answer in [8] is to maximize the log-likelihood function of the optimization variables with constraints on the nuclear norm and the infinity norm of M. The log-likelihood function is dependent on the observation model. Define a differentiable function $f : \mathbb{R} \to [0, 1]$. The observation model presented in [8] is as follows: given a matrix $M \in \mathbb{R}^{d_1 \times d_2}$ and a subset of indices Ω, for any $(i, j) \in \Omega$, people observe $Y_{ij} = +1$ with probability $f(M_{ij})$ and observe $Y_{ij} = -1$ with probability $1 - f(M_{ij})$. Therefore, the log-likelihood function for this observation model is defined as follows ($X \in \mathbb{R}^{d_1 \times d_2}$ represents the optimization variables):

$$\mathcal{L}_{\Omega,Y}(X) = \sum_{(i,j)\in\Omega} \left(I(Y_{ij}=1)\log(f(X_{ij})) + I(Y_{ij}=-1)\log(1-f(X_{ij}))\right)$$

Then the authors consider solving the following convex optimization problem:

$$\begin{aligned} &\text{maximize } \mathcal{L}_{\Omega,Y}(X) \\ &\text{subject to } \|X\|_* \leq \alpha\sqrt{rd_1d_2} \quad \text{and} \quad \|X\|_\infty \leq \alpha. \end{aligned} \tag{2.3}$$

After answering the three questions, the main theorem in [8] proves the upper bound for the recovery error. Before presenting the main theorem, the authors use L_α and β_α to characterize the ”steepness” and ”flatness” of f in the observation model above. Define that

$$L_\alpha \triangleq \sup_{|x|\leq\alpha} \frac{|f'(x)|}{f(x)(1-f(x))},$$

and

$$\beta_\alpha \triangleq \sup_{|x|\leq\alpha} \frac{f(x)(1-f(x))}{(f'(x))^2}.$$

Theorem 4 (Theorem 1 in [8])**.** *Assume that $\|M\|_* \leq \alpha\sqrt{rd_1d_2}$ and $\|M\|_\infty \leq \alpha$. Suppose that Ω is chosen at random following the binomial model with $\mathbb{E}|\Omega| = n$. Suppose Y is generated by the observation model above. Consider $\widehat{M}$ as the solution to (2.3), Then with probability at least $1 - C/(d_1+d_2)$,*

$$\frac{1}{d_1d_1}\|\widehat{M} - M\|_F^2 \leq C_\alpha\sqrt{\frac{r(d_1+d_2)}{n}}\sqrt{1+\frac{(d_1+d_2)\log(d_1d_2)}{n}},$$

where $C_\alpha \triangleq C_2\alpha L_\alpha\beta_\alpha$. Above, C_1, C_2 are absolute constants.

This algorithm proves that when n is greater than $(d_1+d_2)\log(d_1d_2)$, the approximately low-rank matrix M can be recovered accurately with high probability. In fact, the algorithm offers a order of n for accurate recovery when d_1 or d_2 goes to infinity.

Moreover, the authors offer a lower bound for the recovery error and demonstrate that their method is asymptotically optimal. They design a M and proves that any algorithm can not recovery M very accurately. The result is summarized in the following theorem.

Theorem 5 (Theorem 3 in [8]). *Fix α, r, d_1, d_2 to be such that $r > 4$ and $\alpha^2 r \max(d_1, d_2) > C_0$. Suppose $f'(x)$ is decreasing for $x > 0$. Let Ω be any subset of $[d_1] \times [d_2]$ with cardinality n. Let Y be generated from the above observation model. Consider any algorithm which, for any M satisfying $M \in K \triangleq \{M \in \mathbb{R}^{d_1 \times d_2} : \|M\|_* \leq \alpha\sqrt{r d_1 d_2}$ and $\|M\|_\infty \leq \alpha\}$, takes $Y_{ij}, (i,j) \in Omega$ as inputs and returns $\widehat{M}$, there exists a $M \in K$ such that with probability at least $3/4$,*

$$\frac{1}{d_1 d_2}\|M - \widehat{M}\|_F^2 \geq \min\left\{C_1, C_2\alpha\sqrt{\beta_{3\alpha/4}}\sqrt{\frac{r\max(d_1,d_2)}{n}}\right\},$$

as long as the right-hand side of the above equation exceeds $r\alpha^2/\min(d_1, d_2)$. Above, C_1, C_2 are absolute constants.

There are several methods to solve (2.3). In [8], the authors choose Spectral projected-gradient method and alternating-direction method of multipliers (ADMM).

Poisson matrix completion is also a quantized matrix completion problem. Therefore, the work in Chapter 3 can be seen as an extension of the brilliant work in [8].

2.2 Sequential change detection

In this section, I introduce basic ideas and results about the sequential change-point detection. There are two major types of this problem: Bayesian and non-Bayesian. In Bayesian method people assumes that the location of change follows a prior distribution but in non-Bayesian method there is not such assumption. Bayesian methods are usually more easily to be analyzed mathematically but non-Bayesian methods are more generally used and more robust (even if hard to analyze). Both methods are useful. In this thesis, I only consider the non-Bayesian methods. I will introduce some classic result for one-sensor case and then introduce recent progress in multi-sensor change-point detection.

2.2.1 Basic mathematical results

At the beginning, I would like to introduce some famous results that are widely used in the analysis for sequential change-point detection problem. The most important one is Wald's equation, which is also one of the most crucial results in stochastic process.

Lemma 5 (Wald's equation)**.** *Let $x_1, x_2, \ldots$ be independent and identically distributed. For any stopping time T with $\mathbb{E}T < \infty$,*

$$\mathbb{E}\left(\sum_{i=1}^{T} x_i\right) = \mathbb{E}x_1 \cdot \mathbb{E}T.$$

This is a very good property for the stopping time. Another useful property for stopping time is about the likelihood ratio.

Lemma 6 (Walds likelihood ratio identity)**.** *Let $x_1, x_2, \ldots$ be independent and identically distributed with density f and g under $\mathbb{P}_f$ and $\mathbb{P}_g$ respectively. For any event A, we use $\mathbb{E}(X; A)$ to denote $\mathbb{E}(XI_A)$, where I_A is the indicator variable of event A which equals to 1 if A occurs and 0 otherwise. Define the likelihood ratio sequence for any n*

$$l_n = \sum_{i=1}^{n} \frac{g(x_i)}{f(x_i)}.$$

For any stopping time T and non-negative random variable Y prior to T, we have that

$$\mathbb{E}_g(Y; T < \infty) = \mathbb{E}_f(Y l_T; T < \infty).$$

In particular, if $Y = I_A$, then

$$\mathbb{P}_g(A \cap \{T < \infty\}) = \mathbb{E}_f(l_T; A \cap \{T < \infty\}).$$

The proof of Wald's likelihood ratio identity is a mathematical technique: change of measure, which establishes an equality between the two cases when the observations are drawn from f and g. When it is hard to analyze from one side, we can think of changing the measure and analyzing from the other side.

2.2.2 One-sensor change-point detection

Basic settings and concepts

Assume that a system is monitored by one sensor and the sensor returns records sequentially. We observe from the sensor a sequence of independent observations $x_1, x_2, \ldots$. At the beginning, the system is "in control" and the observations are drawn from a distribution with probability density f. At some unknown but deterministic (not random) time κ, called *change-point*, the system changes from "in control" to "out of control" and the distribution for the observations change from f to g. Usually we know ahead of time f since when the system is "in control" the sensor's records are stable and easily predicted. However, usually, it is difficult to know for sure g since many factors can make the system "out of control". Next, I will mainly introduce two type of detection procedures: cumulative sum(CUSUM) method and generalized likelihood ratio(GLR) methods, where the former one is established to solve the problem when g is known and the latter one is designed for the case when g is unknown.

Next question is about how to evaluate the performance of the detection procedures. In practice, we prefer to raise an alarm as soon as the system is out of control so that we can take appropriate action, in other words, to minimize the detection delay. In the meantime, we prefer to control the false alarm rates when the system is in control. One can imagine that if a procedure with exactly zero false alarm rate it also has infinity detection delay because of the disturb of noise. Therefore, one good detection procedure must be a procedure that achieves a good balance between the false alarm rates and detection delay. The problem of looking for such a procedure is known as a sequential change-point detection problem.

Since the observations come sequentially, the detection procedure is just a stopping time if translated to the mathematical language. In this thesis, the two names are interchangeable.

Mathematically, we can formulate the above arguments as the following hypothesis testing problem:

$$\begin{aligned} H_0: &\quad x_i \sim f, i = 1, 2, \ldots, \\ H_1: &\quad x_i \sim f, i = 1, 2, \ldots, \kappa, \\ &\quad x_i \sim g, i = \kappa + 1, \kappa + 2, \ldots. \end{aligned} \tag{2.4}$$

Then, the false alarm rate can be interpreted as the type-I error of hypothesis testing problem (2.4) and the detection delay is closely rated to the type-II error of (2.4). In fact, the change-point detection problem is highly related to the sequential hypothesis testing problem.

Under the null hypothesis probability and expectation in this case are denoted by $\mathbb{P}_\infty$ and $\mathbb{E}_\infty$, respectively. Under the alternative hypothesis probability and expectation in this case are denoted by $\mathbb{P}_\kappa$ and $\mathbb{E}_\kappa$, where κ is the change-point. Define a filtration at time t by $\mathcal{F}_t \triangleq \sigma(x_1, \ldots, x_t)$, which is a smallest sigma-algebra containing $x_1, \ldots, x_t$. Therefore, $\mathcal{F}_t$ contains all the information before time t. Next, I introduce the mathematical representations that are frequently used in the theoretical analysis of change-point detection problem. For any stopping time T,

- Average Run Length(ARL): the average number of observations between two false alarms. Larger the better and it can be represented as $\mathbb{E}_\infty\{T\}$.
- Expected Detection Delay(EDD): the average number of observations between real change-point and the alarm time. There are two popular choices of EDD. The first one is introduced by Lorden in [59], known as the "worst-case" detection delay:

  $$\text{WDD}(T) \triangleq \sup_{k \geq 0} \left(\text{esssup}\mathbb{E}_k \left\{(T - k + 1)^+ \mid \mathcal{F}_k\right\}\right),$$

  where the "esssup" is taken over $\mathcal{F}_k$, meaning that "for any possible outcome of

$(x_1, \ldots, x_k)$". The second one is introduced by Pollak in [62], known as the conditional average detection delay

$$\mathrm{CADD}(T) \triangleq \sup_{k \geq 0} \left(\mathbb{E}_k \left\{ T - k \mid T > k \right\} \right).$$

Note that the event $\{T > k\}$ is not included in $\mathcal{F}_k$ and in fact is not measurable. Therefore, the WDD and CADD are slightly different but both are meaningful. People would like to choose one of them for their analysis. Define a class $C_\gamma = \{T : \mathbb{E}_\infty\{T\} \geq \gamma\}$ with some prescribed $\gamma > 0$. The goal is to find a stopping time $T \in C_\gamma$ that minimizes WDD(T) or CADD(T). Equipped with the above notations and basic concepts, I can introduce the CUSUM and GLR detection procedures that are most widely used and are proved to be (asymptotically) optimal.

CUSUM procedure

The CUSUM procedure is first given by Page in [89]. This method requires the distributions of observations both before and after the change-point, namely, f and g. The core of CUSUM procedure is the likelihood ratio between f and g. The procedure updates the testing statistic recursively by taking new likelihood ratio into consideration. Specifically, under the settings in this section, the CUSUM procedure is a stopping time given by:

$$T_{CM} \triangleq \inf \left\{ t \geq 1 : \max_{0 \leq k < t} \sum_{i=k+1}^{t} \log \frac{g(x_i)}{f(x_i)} \geq b \right\}, \tag{2.5}$$

where b is a prescribed trigger threshold. Note that larger b means larger ARL of CUSUM procedure T_{CM}. There are two major advantages for T_{CM}: recursively computing and exact optimality. First, rewriting the procedure T_{CM}, we obtain that

$$T_{CM} \triangleq \inf \left\{ t \geq 1 : U_t \geq b \right\},$$

where

$$U_t = \max\left(0, U_{t-1} + \log\frac{g(x_t)}{f(x_t)}\right), t \geq 1, \quad U_0 = 0.$$

Therefore, the testing statistic U_t can be updated recursively. Note that in fact U_t considers all the information before time t but at time t it only needs one new computation. This property is one of the most important properties of the CUSUM procedure. Second, the CUSUM procedure enjoys good optimality that other procedure usually do not have. Dr. Moustakides proves the exact optimality in [63].

Theorem 6 (modification of Theorem 1 in [63])**.** *For any $b > 0$, the CUSUM procedure T_{CM} minimizes the worst-case detection delay WDD(T) among all stopping times T satisfying $\mathbb{E}_\infty\{T\} \geq \mathbb{E}_\infty\{T_{CM}\}$.*

Fixing b first, then $\mathbb{E}_\infty\{T_{CM}\}$ is also known (note that there is no closed form for this term). The above theorem proves that in fact the CUSUM procedure is the solution to the optimization problem $\min_{T\in C_\gamma}$ WDD(T) for any $\gamma > 0$. Even if the simple description of the exact optimality of the CUSUM procedure, the theorem is a deep result and is proved about 30 years after Page presented the CUSUM procedure. The mathematical tool behind the theorem is Markov stopping theory and I suggest read [90] if one is interested in the details. Unfortunately, the proof using Markov stopping theory is not easily extended to other detection procedures. Therefore, instead of exact optimality, people usually think of the asymptotic optimality that focus on the case in which γ goes to infinity. Note that the CUSUM procedure is also asymptotic optimal (of course since it has stronger optimality) and this result is proved by Dr. Lorden in [59]. Since the optimality in Chapter 4 of this thesis is also asymptotical optimality, I would like to introduce the basic idea of proving the asymptotical optimality, which is followed by the thinking clues from Dr. Lorden. The first step is to prove a lower bound for the WDD, for any detection procedures $T \in C_\gamma$ as $\gamma \to \infty$. The second step is to prove that the WDD of CUSUM procedure achieves the lower bound.

Theorem 7 (Theorem 3 in [59]). *Let $n(\gamma)$ be the infimum of WDD(T) as $T \in C_\gamma$. If*

$$I_1 \triangleq \int g(x) \log \frac{g(x)}{f(x)} dx < \infty,$$

then as $\gamma \to \infty$,

$$n(\gamma) = \frac{\log \gamma}{I_1} + o(\gamma).$$

Note that I_1 just defined is the KL divergence between pre and post distributions. The above theorem offers the lower bound. Applying Theorem 2 in [59] we obtain that in fact CUSUM procedure achieves this lower bound. Following this framework, one can know that the asymptotic optimal detection procedure is not unique. To compare the procedures with asymptotic optimality, one can discuss the order of γ of the $o(\gamma)$.

GLR procedure

Sometimes we do not know exactly what the post-change distribution is, in this case, we can try to use GLR procedure. GLR procedure assumes a parametric model for the post change distribution and estimate the parameters for the model based on the historic observations. The testing statistic is also the likelihood ratio but it is related to a window size w, which declares from what observations the parameters are estimated. GLR procedure is first introduced in [91].

Define that V_t as follows:

$$V_t = \max_{\theta \in \Theta} \sum_{i=k}^{t} \log \frac{f_\theta(x_i)}{f_0(x_i)},$$

where f_0 and f_θ are the density before change and the density with parameter θ, and Θ is the parameter space. For example, f_θ can be the normal distribution with mean θ and variance 1 and the density before change is just the standard normal distribution. Then the GLR procedure is given by

$$T_{GLR} \triangleq \inf \left\{ t \geq 1 : \max_{t-w \leq k < t-w_1} V_t \geq b \right\},$$

where w is the window size and we estimate the parameter θ based on $x_{t-w}, \ldots, x_{t-w_1}$. The theoretical analysis for GLR procedure is more difficult than that for CUSUM procedure because of the uncertainty of parameters. Dr. Lai offers some results in [54] (see Lemma 2).

Similar with CUSUM procedure, the GLR procedure sometimes can also be computed recursively. However, because the GLR procedure needs estimate the unknown parameters based on the observations in a window with size w, its computation complexity is w times than the CUSUM procedure. But this computation issue is not obvious since the computational ability of computers is increasing rapidly, and people care more about the performance loss caused by the mismatch between real parameters and chosen parameters in CUSUM procedure. In the next subsection, I will assume that the observations follow normal distributions and give examples of the CUSUM procedure and GLR procedure under this setting in order to help reader understand them more easily. The specific form of procedures are given by assuming N sensors since the one-sensor detection procedure is a special case when $N = 1$.

2.2.3 Multi-sensor change-point detection

Even if much progress is achieved to detect the change of one time-series from one sensor, few work has been done in the case when there are many sensors. When the signal-to-noise(SNR) ratio is low, in other words, when the signal is weak, one can not detect the change based on only one series of observations. Therefore, people use many sensors to monitor one system and detect the change by combing the information from all the sensors. One example is in the detection of DNA sequence in biometrics. Fig. 2.1 shows the signals from 10 sensors and the goal is to detect where is the useful DNA sequence (in this

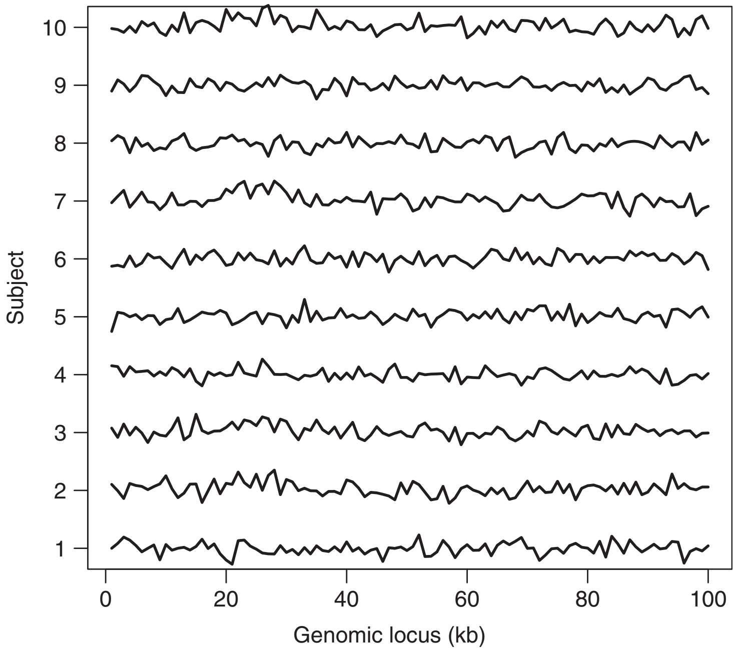


Figure 2.1: Figure 7.1 in [92]: An artificial example of a data matrix. The raw data for 10 subjects over 100 genomic markers are presented, with 1 marker set every 1 kb. Subjects 2, 3, 7, and 10 have elevated levels of the expectation for the markers in positions 20 to 30.

example we have known that the true DNA in between 20 to 30 based on expert's bio domain knowledge). It is hard to detect the true DNA even if we see the signals from all the sensors and it is impossible to distinguish the true DNA sequence from only one sensor's information. Other than this example, multi-sensor detection problem occurs frequently in quality control and security field. Therefore, effective detection procedure in multi-sensor case is necessary. Next, I will introduce three recent multi-sensor change detection procedures.

To help reader understand more easily, I assume the normal observations in the following to show an example. Specifically, we consider the following hypothesis testing problem:

$$\begin{aligned} H_0: \quad & y_{n,i} \sim \mathcal{N}(0,1), i = 1,2,\ldots, n = 1,2,\ldots,N, \\ H_1: \quad & y_{n,i} \sim \mathcal{N}(0,1), i = 1,2,\ldots,\kappa, \\ & \qquad y_{n,i} \sim \mathcal{N}(\mu_n,1), i = \kappa+1,\kappa+2,\ldots, n \in \mathcal{A}, \\ & y_{n,i} \sim \mathcal{N}(0,1), i = 1,2,\ldots, n \in \mathcal{A}^c, \end{aligned} \tag{2.6}$$

where N is the number of sensors and only the sensors in the index set $\mathcal{A}$ observes the change in the mean. Note that we do not know $\mathcal{A}$ ahead of time and $\mathcal{A}$ is also what we would like to obtain.

Multivariate CUSUM procedure

This procedure is first introduced in [93]. The idea is to run multiple CUSUM procedures simultaneously on the sensors. If one of the procedures raises an alarm then it reports that the whole system is down. Multivariate CUSUM procedure assumes that we know $\mu_n, 1 \leq n \leq N$ and assumes that **all** the sensors observe the change in the mean. The multivariate CUSUM detection procedure for problem (2.6) is given by

$$T_{M-CM} \triangleq \inf \left\{ t \geq 1 : \max_{1 \leq n \leq N} \max_{0 \leq k < t} \sum_{i=k+1}^{t} (\mu_n^2/2 - \mu_n y_{n,i}) \geq b \right\},$$

where b is a prescribed threshold.

Here comes the model mismatch: first, μ_ns may be poor guess and far away from the true values; second, the number of indexes in $\mathcal{A}$ is small so only few sensor observes the change. Although the multivariate CUSUM procedure is easily to be performed in practice, it may not have good performance if the assumptions are not similar with the real circumstances.

Mei's Sum CUSUM procedure

In order to decrease the performance loss caused by the assumption that all the sensors observe the change, Dr. Mei present a new method in [61] recently. The idea is to change the ”maximum” into ”sum” over the sensors. Specifically, for the problem (2.6), Mei's procedure is given by

$$T_{Mei} \triangleq \inf \left\{ t \geq 1 : \sum_{n=1}^{N} \left[\max_{0 \leq k < t} \sum_{i=k+1}^{t} (\mu_n^2/2 - \mu_n y_{n,i}) \right] \geq b \right\},$$

where b is a prescribed threshold. The intuition is that summation over all the sensors is one kind of combination of the information from all the sensors so it should be better than taking maximum which treats sensors independently. Moreover, it is asymptotically optimal up to first-order to detect each and every possible combination of affected data streams when the data streams are independent, no matter how many data streams are affected. This optimality is summarized in Theorem 1 in [61].

Even if Mei's procedure eases the effect caused by incorrect assumption of $\mathcal{A}$, it still assumes that we know the post-change distribution. If μ_n is badly guessed, the performance is poor. To overcome this difficulty, Dr. Xie considers the GLR procedure, but in multi-sensor case.

Xie's Multi-sensor GLR procedure

Recently, Dr. Xie and Dr. Siegmund presented the first GLR procedure [18] in multi-sensor case to solve (2.6) by introducing a sparsity factor p_0. p_0 is a prior guess for the term $|\mathcal{A}|/N$, which is the proportion of the sensors that observe the change. Instead of identifying exactly the member in $\mathcal{A}$, the guess of p_0 is more possible to be accurate. The GLR procedure is as follows:

$$T_{Xie} = \inf\left\{ t : \max_{0 \le k < t} \sum_{n=1}^{N} \log\left(1 - p_0 + p_0 \exp[\ell_n(k,t)]\right) \ge b \right\},$$

where b is a prescribed threshold and $\ell_n(k,t)$ is the likelihood ratio given by

$$\ell_n(k,t) = \frac{(\sum_{i=k+1}^{t} y_{n,i})^2}{2(t-k)}.$$

And the version with window size w is given by

$$\widetilde{T}_{Xie} = \inf\left\{ t : \max_{t-w \le k < t} \sum_{n=1}^{N} \log\left(1 - p_0 + p_0 \exp[\ell_n(k,t)]\right) \ge b \right\},$$

where b is a prescribed threshold. Note that the procedures are not dependent on $\mu_n, 1 \leq n \leq N$ but replace them with their Maximum Likelihood Estimators(MLEs).

Xie's procedure addresses the mismatches from both $\mathcal{A}$ and μ_n, however, it is slower than the above two procedures since the MLEs are updated based on w historic observations. This procedure is also first-order asymptotic optimal, which is proved in [13] recently. The proof is also one major part in Chapter 4 of this thesis. One can see that there is balance between model generalization and computational complexity. Therefore, there is no perfect detection procedure, and the ability of establishing appropriate procedures for specific problems is the most important.

2.3 Robust hypothesis testing and change detection

2.3.1 Classic theory about robust hypothesis testing

As is well known, for testing between two simple hypotheses, Neyman-Pearson test has the smallest type-II error among all the test with the same type-I error, namely, minimax. One may ask if we can find such tests when the hypotheses are not quite exactly specified. Assume that we observe finite sample $x_1, \ldots, x_n$, and we would like to test $H_0 : x_i \sim \mathbb{P}_0 \in \mathcal{P}_0, i = 1, \ldots, n$ against $H_1 : x_i \sim \mathbb{P}_1 \in \mathcal{P}_1, i = 1, \ldots, n$, where $\mathcal{P}_0$ and $\mathcal{P}_1$ are two sets of distributions that possibly contain the true distributions. The problem of finding a test to solve the above testing problem in some sense of optimality is called robust hypothesis testing problem.

Dr. Huber proved in [65] first that a robust version of probability ratio test is minimax optimal when $\mathcal{P}_0$ and $\mathcal{P}_1$ are small perturbations of true distributions. Then Dr. Huber and Dr. Strassen extended the results to more general cases by introducing the concept of 2-alternating capacity. Next I will review briefly the concepts in a language of probability theory.

2-*alternating capacity*

Let Ω be a complete separate metrizable space, $\mathcal{A}$ be its Borel-σ-algebra, $\mathcal{C}$ be the set of all probability measures on Ω and $\mathcal{P} \subset \mathcal{C}$. Then, the non-empty set $\mathcal{P}$ defines an upper probability

$$v(A) = \sup\{\mathbb{P}(A), \mathbb{P} \in \mathcal{P}\},\ A \in \mathcal{A}$$

and an lower probability

$$u(A) = \inf\{\mathbb{P}(A), \mathbb{P} \in \mathcal{P}\},\ A \in \mathcal{A},$$

then $v(A) + u(A^c) = 1$ for any $A \in \mathcal{A}$. They call *2-alternating capacity* any set function v satisfying the following properties:

- $v(\emptyset) = 0$ and $v(\Omega) = 1$.
- If $A \in B$, then $v(A) \leq v(B)$.
- If $A_n \uparrow A$, then $v(A_n) \uparrow v(A)$.
- If $\mathcal{P}$ is weakly compact and $F_n \downarrow F$, then $v(F_n) \downarrow v(F)$.
- $v(A \cup B) + v(A \cap B) \leq v(A) + v(B)$,

Least favorable distributions and probability ratio tests

Let $\mathcal{C}$ be the set of all probability measures on a complete separable metrizable space Ω. Define that

$$\mathcal{P}_j \triangleq \{\mathbb{P} \in \mathcal{C} \mid \mathcal{P}(A) \leq v_j(A) \text{ for all Borel sets } A\}, j = 1, 2,$$

where v_0 and v_1 are both 2-alternating capacities. Then there exists a pair $(\mathbb{Q}_0, \mathbb{Q}_1) \in \mathcal{P}_0 \times \mathcal{P}_1$ such that the Neyman-Pearson tests between the simple hypotheses $\mathbb{Q}_0$, $\mathbb{Q}_1$ are minimax among all the tests between $\mathcal{P}_0$ and $\mathcal{P}_1$. The pair $(\mathbb{Q}_0, \mathbb{Q}_1)$ is called least favorable distributions(LFDs) and they prove the existence of the LFDs if the distributions are ”upper bounded” by 2-alternating capacities. The main results are summarized in the following theorem.

Theorem 8 (Simplification of Corollary 4.2 in [94]). *Define that π as the probability ratio between $\mathbb{Q}_0$ and $\mathbb{Q}_1$, namely, a version of $d\mathbb{Q}_1/d\mathbb{Q}_0$. For any sample size n and any level α, the Neyman-Pearson test of level α, between $\mathbb{Q}_0$ and $\mathbb{Q}_1$, defined by*

$$\begin{aligned} \phi(x_1, \dots, x_n) =& 1 \textit{ for } \prod_{i=1}^{n} \pi(x_i) > C \\ =& \gamma \textit{ for } \prod_{i=1}^{n} \pi(x_i) = C \\ =& 0 \textit{ for } \prod_{i=1}^{n} \pi(x_i) < C, \end{aligned} \tag{2.7}$$

where C and γ are chosen such that $\mathbb{E}_{Q_0}\{\phi\} = \alpha$, is also a minimax test between $\mathcal{P}_0$ and $\mathcal{P}_1$ with the same level α and with the same minimum power.

The existence of the least favorable distributions are proved in the above theorem, how to solve for them remains a very difficult problem, especially in come complicated cases (e.g., high-dimensional sample space $\Omega = \mathbb{R}^N$ with a large N). To solve this problem, they offer a representation of $(\mathbb{Q}_0, \mathbb{Q}_1)$ in the following theorem as a solution to a minimization problem.

Theorem 9 (Theorem 6.1 in [94]). *Let Φ be a continuous and strongly convex function in $[0, 1]$, then the pair $(\mathbb{Q}_0, \mathbb{Q}_1) \in \mathcal{P}_0 \times \mathcal{P}_1$ satisfies the above theorem if and only if it minimizes*

$$\int \Phi\left(\frac{d\mathbb{P}_0}{d(\mathbb{P}_0 + \mathbb{P}_1)}\right) d(\mathbb{P}_0 + \mathbb{P}_1),$$

among all $(\mathbb{P}_0, \mathbb{P}_1) \in \mathcal{P}_0 \times \mathcal{P}_1$

Unfortunately, the above minimization problem is not easy to solve since the minimization is over a set of distributions and the ratio $\frac{d\mathbb{P}_0}{d(\mathbb{P}_0+\mathbb{P}_1)}$ can not computed easily in many circumstances.

Examples of robust testing problems with 2*-alternating capacities*

Dr. Huber presents that when the observation set is compact, several uncertainty models such as ϵ-contamination neighborhoods, total variation neighborhoods, band classes, and p-point classes are special cases of this model with different choices of capacity.

First example. Let Ω be compact. Define that $v(A) = (1-\epsilon)\mathbb{P}_0(A) + \epsilon$ for $A \neq \emptyset$, then v is a 2-alternating capacity. And the ϵ-contamination neighborhoods is just the following set of distributions:

$$\mathcal{P}_v \triangleq \{\mathbb{P} \in \mathcal{C} \mid \mathbb{P} = (1-\epsilon)\mathbb{P}_0 + \epsilon H, H \in \mathcal{C}\}.$$

Second example. Let Ω be compact. Define that $v(A) = \min(\mathbb{P}_0(A), 1)$ for $A \neq \emptyset$, then v is a 2-alternating capacity. And the total variation neighborhoods is just the following set

$$\mathcal{P}_v \triangleq \{\mathbb{P} \in \mathcal{C} \mid |\mathbb{P}(A) - \mathbb{P}_0(A)| \leq \epsilon, \forall A\}.$$

For more examples and details, one can see [94] or Chapter 10 in [69].

In summary, Dr. Huber's work offers us confidence that the probability ratio test is also minimax in a large number of robust hypothesis testing problems, even if more efficient way of finding such probability ratio tests are still needed in practice.

2.3.2 Minimax robust sequential change detection in one dimensional case

Recently in [68], Dr. Veeravalli and other authors apply the theories in robust hypothesis testing to the robust sequential change detection problem. The basic idea is to find the least favorable distributions first and then use CUSUM procedure to detect the change. Since the testing statistic in CUSUM procedure is just the probability ratio of pre and post change distributions, Dr. Veeravalli and other authors prove the optimality of their detection procedures by extending Dr. Huber's work.

The main difference between robust hypothesis testing and robust sequential change detection is the assumption on the number of sample size. In hypothesis testing problem, the sample size is fixed ahead of time, however, the sample size in unknown and the observations come sequentially in the change detection problem. Other than this difference, the change detection problem uses different metrics to characterize the optimality. Therefore, even if it seems like a straightforward extension, it remains many difficulties if one wants to offer a rigorous proof. For this reason, in [68], the authors only consider the case when the distributions are on the real line $\mathbb{R}$, an assumption that simplifies the analysis much. Next, I review their results briefly and show some difficulties not addressed in high-dimensional case.

Joint Stochastic Boundedness(JSB)

The main result in [68] is highly dependent on the assumption called Joint Stochastic Boundedness. First, I review the definition of stochastic order of random variables. If and X and X' are two real-valued random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that

$$\mathbb{P}(X \geq t) \geq \mathbb{P}(X' \geq t), \forall t \in \mathbb{R},$$

then we say that the random variable X is stochastically larger than the random variable X', denoted by $X \succ X'$.

Definition 1 (Joint Stochastic Boundedness in [67]). *Consider the pair $(\mathcal{P}_0, \mathcal{P}_1)$ of classes of distributions defined on a measurable space $(\Omega, \mathcal{F})$. Let $(\overline{\nu}_0, \underline{\nu}_1) \in \mathcal{P}_0 \times \mathcal{P}_1$ be some pair of distributions from this pair of classes such that $\underline{\nu}_1$ is absolutely continuous with respect to $\overline{\nu}_0$. The pair $(\mathcal{P}_0, \mathcal{P}_1)$ is said to be Joint Stochastic Bounded by $(\overline{\nu}_0, \underline{\nu}_1)$ if for all $(\nu_0, \nu_1) \in \mathcal{P}_0 \times \mathcal{P}_1$,*

$$\overline{\nu}_0 \succ \nu_0, \quad \nu_1 \succ \underline{\nu}_1.$$

And $(\overline{\nu}_0, \underline{\nu}_1)$ *is the pair of LFDs for the uncertainty classes* $(\mathcal{P}_0, \mathcal{P}_1)$*.*

The asymptotical optimal robust CUSUM procedure in one dimensional case

Next, I introduce the main result in [68], which claims that the robust CUSUM procedure is asymptotical optimal in a minimax sense if the JSB assumption is satisfied. Since the pre-change and post-change distributions are uncertain, we change a little bit the notation defined before in the one-sensor change detection. When the change happens at k and the pre-change and post-change distributions are ν_0 and ν_1 respectively, the probability law is denoted by $\mathbb{P}_k^{\nu_0,\nu_1}$ and the expectation is denoted by $\mathbb{E}_k^{\nu_0,\nu_1}$. When there is no change, the notations are changed to $\mathbb{P}_\infty^{\nu_0}$ and $\mathbb{E}_\infty^{\nu_0}$, respectively.

Theorem 10 (Theorem III.2 in [68])**.** *Suppose the following conditions hold: (i) The pair* $(\mathcal{P}_0, \mathcal{P}_1)$ *is Joint Stochastic Bounded by* $(\overline{\nu}_0, \underline{\nu}_1)$*. (ii) All distributions* $\nu_0 \in \mathcal{P}_0$ *are absolutely continuous with respect to* $\overline{\nu}_0$*. (iii) The function* $L^*(\cdot)$*, representing the log-likelihood ratio between* $\underline{\nu}_1$ *and* $\overline{\nu}_0$*, is continuous over the support of* $\overline{\nu}_0$*. Define that the robust CUSUM procedure* T_{R-CM} *as follows:*

$$T_{R-CM} \triangleq \inf\left\{t > 0 : \max_{0\le k\le t} \sum_{i=k+1}^{t} L^*(x_i) \ge b\right\},$$

where x_i*s are the observations and* b *is the prescribed threshold chosen such that* $\mathbb{E}_\infty^{\overline{\nu}_0}\{T_{R-CM}\} = \gamma$*. Then* T_{R-CM} *minimizes*

$$\sup_{\nu_0\in\mathcal{P}_0} \sup_{\nu_1\in\mathcal{P}_1} \textit{WDD}^{\nu_0,\nu_1}(T)$$

among all the stopping times in $\{T : \sup_{\nu_0\in\mathcal{P}_0} \mathbb{E}_\infty^{\nu_0} T \ge \gamma\}$*, where* $\textit{WDD}^{\nu_0,\nu_1}(T)$ *is defined as follows (just change the expectation operator)*

$$\textit{WDD}(T) \triangleq \sup_{k\ge 0} \left(\textit{esssup}\mathbb{E}_k^{\nu_0,\nu_1}\left\{(T-k+1)^+ \mid \mathcal{F}_k\right\}\right).$$

This theorem shows that in one-dimensional case, we need two steps to find the asymptotic optimal procedure. The first step is to find the pair LFDs $(\overline{\nu}_0, \underline{\nu}_1)$ given two uncertainty classes $\mathcal{P}_0$ and $\mathcal{P}_1$. The second step is to construct classic CUSUM procedure that is used in the case when the pre-change and post-change distributions are $\overline{\nu}_0$ and $\underline{\nu}_1$ respectively.

Difficulty in high-dimensional sample space

The above JSB assumption is defined only in the case where the distributions are defined on the real line. One may ask why not extend it to the high-dimensional case with the extended definition for ”stochastically larger”. Next, I give an counterexample to show that the ”high-dimension” JSB assumption can not be satisfied even in very simple cases.

First, I introduce the definition of multivariate stochastic order. If and X and X' are two real-valued random variables defined on a probability space $(\Omega^N, \mathcal{F}, \mathbb{P})$ such that

$$\mathbb{E}f(X) \geq \mathbb{E}f(Y), \forall \text{ increasing, bounded function } f : \mathbb{R}^N \to \mathbb{R},$$

then we say that the random variable X is multivariate stochastically larger than the random variable X', denoted by $X \succ X'$. We can see that this definition is a direct extension and when $N = 1$ the definition is equivalent with the definition of stochastic order for the distributions on the real line.

Second, we keep the description about JSB assumption except replacing the definition of stochastic order with the definition of multivariate stochastically larger, and show that this assumption is too strong to be satisfied in high dimensional case. Consider $\Omega = \mathbb{R}^2$ and $\mathcal{P}_0$ and $\mathcal{P}_1$ are two sets of bivariate normal distributions. Assume that Σ is a positive-definite matrix in $\mathbb{R}^{2\times 2}$,

$$\mathcal{P}_0 = \{\mathbb{P} \mid \mathbb{P} = \mathcal{N}(\mu_0, \Sigma), \|\mu_0\|_2 \leq 1, \mu_0 \in \mathbb{R}^2\},$$

and

$$\mathcal{P}_1 = \{\mathbb{P} \mid \mathbb{P} = \mathcal{N}(\mu_1, \Sigma), \|\mu_1 - 10\|_2 \leq 1, \mu_1 \in \mathbb{R}^2\}.$$

Next lemma shows easily that we can not find a $\overline{\nu}_0 \in \mathcal{P}_0$ such that $\overline{\nu}_0 \succ \nu_0, \forall \nu_0 \in \mathcal{P}_0$.

Similarly, we can not find $\underline{\nu}_1 \in \mathcal{P}_1$ such that $\underline{\nu}_1 \succ \nu_1, \forall \nu_1 \in \mathcal{P}_1$.

Lemma 7 (Theorem 5 in [71])**.** *Let $X \sim \mathcal{N}(\mu, \Sigma)$ and $X' \sim \mathcal{N}(\mu', \Sigma')$ be n-dimensional normally distributed random vectors. Then $X' \succ X$ if and only if $\mu'_i \geq \mu_i$, for all $1 \leq i \leq n$ and $\Sigma = \Sigma'$.*

Therefore, the JSB assumption for multivariate normal distributions is not satisfied in $\mathbb{R}^2$ even in such a simple case. This is because the definition of multivariate stochastic order is so strong that it needs all the coordinates to satisfy certain property of stochastic order.

2.3.3 Robust optimization for robust hypothesis testing

Recently, Dr. Nemirovski, Dr. Juditsky and Dr. Goldenshluger presents another clue of solving the robust hypothesis testing problem in [72], from the robust optimization point of view. Instead of minimizing over a set of distributions, the authors first specifies a parametric model and then solving a convex optimization problem with variables in very nice parameter space such as $\mathbb{R}^N$. This offers a practical and more efficient method of finding the LFDs, especially in high dimensional case. Next, I basically introduce the ideas of the authors.

Good observation scheme

The authors first makes some general assumptions and then claims their main results based on the assumptions. Assume a parametric family $\mathcal{P} = \{\mathbb{P}_\mu, \mu \in \mathcal{M}\}$ of probability distributions on a sample space Ω and an observation $\omega \sim \mathbb{P}_\mu$ with unknown parameter $\mu \in \mathcal{M}$. The collection $((\Omega, \mathbb{P}), \{p_\mu, \mu \in \mathcal{M}\}, \mathcal{F})$ is called a *good observation scheme* if the following four assumptions are satisfied:

1. $\mathcal{M} \in \mathbb{R}^m$ is a convex set which coincides with its relative interior

2. Ω is a Polish space equipped with a Borel σ-additive σ-finite measure $\mathbb{P}$ and the support of $\mathbb{P}$ is Ω. Define that $p_\mu(\omega)$ as the density of $\mathbb{P}_\mu \in \mathcal{P}$ with respect to $\mathbb{P}$. Then p_μ is continuous, positive and locally uniformly summable in $\mu \in \mathcal{M}$ and $\omega \in \Omega$.

3. Given a finite-dimensional linear space $\mathcal{F}$ of continuous functions on Ω containing constants such that $\log(p_\mu/p_\nu) \in \mathcal{F}$ whenever $\mu, \nu \in \mathcal{M}$.

4. For every $\phi \in \mathcal{F}$, the function

$$F_\phi(\mu) = \log\left(\int_\Omega \exp\{\phi(\omega)\} p_\mu(\omega)\mathbb{P}(d\omega)\right)$$

   is well defined and concave in $\mu \in \mathcal{M}$.

Even if long description about the assumptions, in most popular cases the above assumptions are satisfied. For example, the assumption 3 is equivalent to to saying that distributions $\mathbb{P}_\mu, \mu \in \mathcal{M}$ form an exponential family with continuous minimal sufficient statistics. Therefore, the exponential family is a good parametric family that is included in the good observation scheme. In the end, I will review the case when the distributions are Gaussian.

One sample simple test by convex optimization

The simplest case is when there is only one sample and we would like to test between two simple hypotheses. On the top of a good observation scheme, given two nonempty convex compact sets $X, Y \in \mathcal{M}$ and one sample ω, then the testing problem is

$$H_0 : \omega \sim \mathbb{P}_\mu, \mu \in X, \text{ against } H_1 : \omega \sim \mathbb{P}_\mu, \mu \in Y.$$

Assume a detector $\phi(\cdot) \in \mathcal{F}$ and the decision rule based on the sample ω is: it accepts H_0 if $\phi(\omega) \geq 0$ and it accepts H_1 if $\phi(\omega) < 0$. The quality of the detector ϕ is characterized by the error probabilities. The robust type-I error is then defined as

$$\epsilon_X(\phi) \triangleq \sup_{x \in X} \mathbb{P}_x\{\omega : \phi(\omega) < 0\},$$

and the robust type-II error is defined as

$$\epsilon_Y(\phi) \triangleq \sup_{y \in Y} \mathbb{P}_y\{\omega : \phi(\omega) \geq 0\}.$$

However, the above two terms do not satisfy good convex or concave properties. Therefore, the authors convexify the problem by defining another type of error (noting the assumption 4 in the good observation scheme). The risk $\epsilon(\phi)$ of detector ϕ on (H_0, H_1) is defined to be the smallest ϵ satisfying the following two conditions:

$$\int_\Omega \exp\{\phi(-\omega)\} p_x(\omega) \mathbb{P}(d\omega) \leq \epsilon(\phi), \forall x \in X,$$

and

$$\int_\Omega \exp\{\phi(\omega)\} p_y(\omega) \mathbb{P}(d\omega) \leq \epsilon(\phi), \forall y \in Y.$$

Using Markov inequality, one can prove that $\max\{\epsilon_X(\phi), \epsilon_Y(\phi)\} \leq \epsilon$. The authors then minimize $\epsilon(\phi)$ instead of the classic type-I and type-II errors, which transforms the problem into a convex setting.

Nearly optimal detector for one sample case

In the just described situation and under the above assumptions, define that $\Phi : \mathcal{F} \times (X \times Y) \to \mathbb{R}$ as follows

$$\Phi(\phi, [x; y]) \triangleq \log\left(\int_\Omega \exp\{-\phi(\omega)\} p_x(\omega) \mathbb{P}(d\omega)\right) + \log\left(\int_\Omega \exp\{\phi(\omega)\} p_y(\omega) \mathbb{P}(d\omega)\right).$$

Then the nearly optimal detector can be found by solving the saddle point of $\Phi(\phi, [x; y])$. The existence of saddle point and the simple form of the detector is summarized in the following main theorem.

Theorem 11 (Simplification of Theorem 2.1 in [72])**.** *The theorem contains three parts: existence of saddle points; nearly optimality; and specification of the detector.*

1. *$\Phi(\phi, [x; y])$ is continuous in its domain. It is convex in $\phi(\cdot) \in \mathcal{F}$ and is concave in $[x; y] \in X \times Y$. The saddle point (min in ϕ and max in $[x; y]$) exists and is denoted by $(\phi_*, [x_*; y_*])$. Then the following equation holds*

$$\int_\Omega \exp\{\phi_*(\omega)\} p_{x_*}(\omega)\mathbb{P}(d\omega) \leq \epsilon(\phi_*) = \int_\Omega \exp\{\phi_*(\omega)\} p_{y_*}(\omega)\mathbb{P}(d\omega) \leq \epsilon(\phi_*),$$

   and the common value of the two quantities is denoted by ϵ_*. Then $\epsilon(\phi_*) \leq \epsilon_*$.

2. *Let ϵ be a positive value such that there exists a test for deciding between two simple hypotheses $H_0 : \omega \sim p_{x_*}$ and $H_1 : \omega \sim p_{y_*}$ with the sum of type-I and type-II error less than 2ϵ, then $\epsilon_* \leq 2\sqrt{\epsilon(1-\epsilon)}$.*

3. *In fact, the detector ϕ_* is just the probability ratio test. Specifically*

$$\phi_* = \frac{1}{2} \log \frac{p_{x_*}}{p_{y_*}}.$$

The above theorem offers an efficient method of finding the good detector for the robust hypothesis testing problem. For example, when $\omega \sim \mathcal{N}(\mu, \Sigma)$ with unknown μ and known Σ, defining that $\mathcal{F}$ as the space of all affine functions on $\mathbb{R}^m$, then the solution of $[x; y]$ is just the minimizer of the following optimization problem

$$[x_*; y_*] = \arg \max_{x \in X, y \in Y} \left[-\frac{1}{4}(x-y)^\intercal \Sigma^{-1}(x-y) \right],$$

which is the nearest pair of two points in X and Y in the sense of Euclidean distance.

K-Repeated observations

I review the case when there is only one sample. Next, I introduce the case when K sample points are collected repeatedly. Good observation schemes admit naturally defined direct products. Assume that ϕ_* is the nearly optimal detector for one sample. Then for K-repeated sample points, the detector is

$$\phi_*^{(K)} = \sum_{k=1}^{K} \phi_*(\omega_k),$$

and the risk is $\epsilon_*^{(K)} = \epsilon_*^K$. Note that when the sample size increases the risk decreases exponentially.

CHAPTER 3
POISSON MATRIX COMPLETION

In this chapter, I present the work about poisson matrix completion. This work is mainly summarized in [95]. Section 3.1 sets up the formalism for Poisson matrix completion. Section 3.2 presents matrix recovery based on constrained maximum likelihood and establishes the upper and lower bounds for the recovery accuracy. Section 3.3 presents the PMLSVT algorithm that solves the maximum likelihood approximately and demonstrates its performance on recovering solar flare images and bike sharing count data. All proofs are delegated to the appendix.

The notation in this chapter is standard. For reader's convenience, I present the notations again here. In particular, $\mathbb{R}_+$ denotes the set of positive real numbers and $\mathbb{Z}_+^m$ denotes a m-dimensional vector with positive integer entries; $[\![d]\!] = \{1, 2, \ldots, d\}$; $(x)^+ = \max\{x, 0\}$ for any scalar x; Let $[x]_j$ denote the jth element of a vector x; $\mathbb{I}\{[\varepsilon]\}$ is the indicator function for an event ε; $|A|$ denotes the number of elements in a set A; diag$\{x\}$ denotes a diagonal matrix with entries of a vector x being its diagonal entries; $\mathbf{1}_{d_1 \times d_2}$ denotes an d_1-by-d_2 matrix of all ones. Let $\|x\|_1$, $\|x\|_2$ denote the ℓ_1 and ℓ_2 norms of a vector x. Let entries of a matrix X be denoted by X_{ij} or $[X]_{ij}$. For a matrix $X = [x_1, \ldots, x_n]$ with x_j being the jth column, vec$(X) = [x_1^\intercal, \ldots, x_n^\intercal]^\intercal$ denote vectorized matrix. Let $\|X\|$ be the spectral norm which is the largest absolute singular value, $\|X\|_F = (\sum_{i,j} X_{ij}^2)^{1/2}$ be the Frobenius norm, $\|X\|_*$ be the nuclear norm which is the sum of the singular values, $\|X\|_{1,1} = \sum_i \sum_j |X_{ij}|$ be the ℓ_1 norm, and finally $\|X\|_\infty = \max_{ij} |X_{ij}|$ be the infinity norm of the matrix. Let rank(X) denote the rank of a matrix X. We say that a random variable Z follows the Poisson distribution with a parameter λ (or $Z \sim$ Poisson(λ)) if its probability mass function $\mathbb{P}(Z = k) = e^{-\lambda}\lambda^k/(k!)$). Finally, let $\mathbb{E}[Z]$ denote the expectation of a random variable Z.

The only set of non-conventional notations that we use is the following. By a slight abuse of notation, we denote the Kullback-Leibler (KL) divergence between two Poisson distributions with parameters p and q, $p, q \in \mathbb{R}_+$ as

$$D(p\|q) \triangleq p\log(p/q) - (p-q),$$

and denote the Hellinger distance between two Poisson distributions with parameters p and q as

$$d_H^2(p,q) \triangleq 2 - 2\exp\left\{-\frac{1}{2}\left(\sqrt{p}-\sqrt{q}\right)^2\right\}.$$

It should be understood that the KL distance and the Hellinger distance are defined between two distributions and here the arguments p and q are merely parameters of the Poisson distributions since we restrict our attention to Poisson. Based on this, we also denote, by a slight abuse of notation, the average KL and Hellinger distances for two sets of Poisson distributions whose parameters are determined by entries of two matrices P, $Q \in \mathbb{R}_+^{d_1\times d_2}$:

$$D(P\|Q) \triangleq \frac{1}{d_1d_2}\sum_{i,j} D(P_{ij}\|Q_{ij}),$$

$$d_H^2(P,Q) \triangleq \frac{1}{d_1d_2}\sum_{i,j} d_H^2(P_{ij},Q_{ij}).$$

3.1 Formulation

3.1.1 Matrix completion

A related problem is matrix completion. Given a matrix $M \in \mathbb{R}_+^{d_1\times d_2}$ consisting of positive entries, we obtain noisy observations for a subset of its entries on an index set $\Omega \subset [\![d_1]\!] \times [\![d_2]\!]$. The indices are randomly selected with $\mathbb{E}[|\Omega|] = m$. In other words, $\mathbb{I}\{(i,j)\in\Omega\}$ are i.i.d. Bernoulli random variables with parameter $m/(d_1d_2)$. The observations are Poisson counts of the observed matrix entries and they are mutually independent

$$Y_{ij} \sim \text{Poisson}(M_{ij}), \quad \forall (i,j) \in \Omega. \tag{3.1}$$

Our goal is to recover M from the Poisson observations $\{Y_{ij}\}_{(i,j)\in\Omega}$.

The following assumptions are made for the matrix completion problem. First, we set an upper bound $\alpha > 0$ for each entry $M_{ij} \leq \alpha$ to entail that the recovery problem is well-posed [36]. This assumption is also reasonable in practice; for instance, M may represent an image which is usually not too spiky. The second assumption is characteristic to Poisson matrix completion: we set a lower bound $\beta > 0$ for each entry $M_{ij} \geq \beta$. This entry-wise lower bound is required by our later analysis (so that the cost function is Lipschitz), and it also has an interpretation of a minimum required signal-to-noise ratio (SNR), since SNR of a Poisson observation with intensity I is given by $\sqrt{I}$. Third, we make a similar assumption to one-bit matrix completion [8]; the nuclear norm of M is upper bounded $\|M\|_* \leq \alpha\sqrt{rd_1d_2}$. This is a relaxation of the assumption that M has a rank exactly r (some small integer). This particular choice arises from the following consideration. If $M_{ij} \leq \alpha$ and $\text{rank}(M) \leq r$, then

$$\|M\|_* \leq \sqrt{r}\|M\|_F \leq \sqrt{rd_1d_2}\|M\|_\infty \leq \alpha\sqrt{rd_1d_2}.$$

We consider a formulation by maximizing the log-likelihood function of the optimization variable X given our observations subject to a set of convex constraints. In the matrix completion problem, the log-likelihood function is given by

$$F_{\Omega,Y}(X) = \sum_{(i,j)\in\Omega} Y_{ij} \log X_{ij} - X_{ij}, \tag{3.2}$$

where the subscript Ω and Y indicate the data involved in the maximum likelihood function

F. Based on previous assumptions, we define a candidate set

$$\begin{aligned}\mathcal{S} \triangleq \Big\{ X \in \mathbb{R}_{+}^{d_1 \times d_2} : \|X\|_* \leq \alpha\sqrt{r d_1 d_2}, \\ \beta \leq X_{ij} \leq \alpha, \forall (i,j) \in [\![d_1]\!] \times [\![d_2]\!] \Big\}.\end{aligned} \tag{3.3}$$

An estimator $\widehat{M}$ can be obtained by solving the following convex optimization problem:

$$\widehat{M} = \arg\max_{X \in \mathcal{S}} F_{\Omega,Y}(X). \quad \{\text{matrix completion}\} \tag{3.4}$$

3.1.2 Relation of two formulations

Note that the matrix completion problem can also be formulated as a regularized maximum likelihood function problem. However, we consider the current formulation for the convenience of drawing connections, respectively, between Poisson matrix recovery and Poisson compressed sensing studied in[41], as well as Poisson matrix completion and one-bit matrix completion studied in [8].

Indeed, these two formulations in the forms of nuclear norm regularized maximum likelihood function and (3.4) are related by the well-known duality theory in optimization (see, e.g. [96]). Consider a penalized convex optimization problem:

$$\min_x f(x) + \lambda g(x), \quad \lambda \geq 0, \tag{3.5}$$

and the constrained convex optimization problem:

$$\min_x f(x) \quad \text{subject to} \quad g(x) \leq c. \tag{3.6}$$

Denote x^* as the solution to (4.1). Then x^* is also the solution to (3.6), if we set $c = g(x^*)$. Conversely, denote x^* as the solution to problem (3.6). We can interpret $\lambda \geq 0$ as the the Lagrange multiplier and consider the Lagrange dual problem. Under Slater's condition (i.e.

there exists at least one x such that $g(x) < c$), there is at least one λ such that x^* is also the solution to problem (4.1). Therefore, (4.1) and (3.6) are equivalent in the sense that the two problems have the same minimizer for properly chosen parameters. More details can be found in [97]. Using the suggestion by Theorem 1 in [97], we choose λ to be around $1/\alpha\sqrt{rd_1d_2}$.

3.2 Performance Bounds

In the following, we use the squared error

$$R(M, \widehat{M}) \triangleq \|M - \widehat{M}\|_F^2, \tag{3.7}$$

as a performance metric for both matrix recovery and matrix completion problems.

For matrix completion, we first establish an upper bound for estimator in (3.4), and then present an information theoretic lower bound which nearly matches the upper bound up to a logarithmic factor $\mathcal{O}(d_1d_2)$.

Theorem 12 (Matrix completion; upper bound)**.** *Assume $M \in \mathcal{S}$, Ω is chosen at random following our Bernoulli sampling model with $\mathbb{E}[|\Omega|] = m$, and $\widehat{M}$ is the solution to (3.4). Then with a probability exceeding $(1 - C/(d_1d_2))$, we have*

$$\begin{aligned} &\frac{1}{d_1d_2}R(M, \widehat{M}) \leq C'\left(\frac{8\alpha T}{1-e^{-T}}\right) \cdot (\frac{\alpha\sqrt{r}}{\beta})\cdot \\ &\big(\alpha(e^2-2) + 3\log(d_1d_2)\big) \cdot \left(\frac{d_1+d_2}{m}\right)^{1/2} \cdot \\ &\left[1 + \frac{(d_1+d_2)\log(d_1d_2)}{m}\right]^{1/2}. \end{aligned} \tag{3.8}$$

If $m \geq (d_1 + d_2)\log(d_1 d_2)$, then (3.8) simplifies to

$$\begin{aligned}\frac{1}{d_1 d_2} R(M, \widehat{M}) \leq \sqrt{2} C' \left(\frac{8\alpha T}{1-e^{-T}}\right) \cdot \left(\frac{\alpha\sqrt{r}}{\beta}\right) \cdot \\ \big(\alpha(e^2-2) + 3\log(d_1 d_2)\big) \cdot \left(\frac{d_1+d_2}{m}\right)^{1/2}.\end{aligned} \tag{3.9}$$

Above, C', C are absolute constants and T depends only on α and β. Here the expectation and probability are with respect to the random Poisson observations and Bernoulli sampling model.

The proof of Theorem 12 is an extension of the ingenious arguments for one-bit matrix completion [8]. The extension for Poisson case here is nontrivial for various aforementioned reasons (notably the non sub-Gaussian and only locally sub-Gaussian nature of the Poisson observations). An outline of our proof is as follows. First, we establish an upper bound for the Kullback-Leibler (KL) divergence $D(M\|X)$ for any $X \in \mathcal{S}$ by applying Lemma 13 given in the appendix. Second, we find an upper bound for the Hellinger distance $d_H^2(M, \widehat{M})$ using the fact that the KL divergence can be bounded from below by the Hellinger distance. Finally, we bound the mean squared error in Lemma 14 via the Hellinger distance.

Remark 1. *Fixing m, α and β, the upper bounds (3.8) and (3.9) in Theorem 12 increase as the upper bound on the nuclear norm increases, which is proportional to $\sqrt{r d_1 d_2}$. This is consistent with the intuition that our method is better at dealing with approximately low-rank matrices than with nearly full rank matrices. On the other hand, fixing d_1, d_2, α, β and r, the upper bound decreases as m increases, which is also consistent with the intuition that the recovery is more accurate with more observations.*

Remark 2. *Fixing α, β and r, the upper bounds (3.8) and (3.9) on the mean-square-error per entry can be arbitrarily small, in the sense that the they tend to zero as d_1 and d_2 go to infinity and the number of the measurements $m = \mathcal{O}((d_1 + d_2)\log^{\delta}(d_1 d_2))$ ($m \leq d_1 d_2$) for $\delta > 2$.*

We may obtain an upper bound on the KL divergence (which may reflect the true distribution error) as a consequence of Theorem 12.

Corollary 1 (Upper bound for KL divergence)**.** *Assume $M \in \mathcal{S}$, Ω is chosen at random following the Bernoulli sampling model with $\mathbb{E}[|\Omega|] = m$, and $\widehat{M}$ is the solution to (3.4). Then with a probability exceeding $(1 - C/(d_1 d_2))$,*

$$\begin{aligned} D(M \| \widehat{M}) \leq & 2C' \left(\alpha\sqrt{r}/\beta\right) \left(\alpha(e^2 - 2) + 3\log(d_1 d_2)\right) \cdot \\ & \left(\frac{d_1 + d_2}{m}\right)^{1/2} \cdot \left[1 + \frac{(d_1 + d_2)\log(d_1 d_2)}{m}\right]^{1/2}. \end{aligned} \tag{3.10}$$

Above, C and C' are absolute constants. Here the expectation and probability are with respect to the random Poisson observations and Bernoulli sampling model.

The following theorem establishes a lower bound and demonstrates that there exists an $M \in \mathcal{S}$ such that *any* recovery method cannot achieve a mean square error per entry less than the order of $\mathcal{O}(\sqrt{r \max\{d_1, d_2\}/m})$.

Theorem 13 (Matrix completion; lower bound)**.** *Fix α, β, r, d_1, and d_2 to be such that $\alpha \geq 1$, $\alpha \geq 2\beta$, $r \geq 4$, and $\alpha^2 r \max\{d_1, d_2\} \geq C_0$. Fix Ω_0 be an arbitrary subset of $[\![d_1]\!] \times [\![d_2]\!]$ with cardinality m. Consider any algorithm which, for any $M \in \mathcal{S}$, returns an estimator $\widehat{M}$. Then there exists $M \in \mathcal{S}$ such that with probability at least $3/4$,*

$$\begin{aligned} & \frac{1}{d_1 d_2} R(M, \widehat{M}) \\ & \geq \min\left\{\frac{1}{256}, C_2 \alpha^{3/2} \left[\frac{r \max\{d_1, d_2\}}{m}\right]^{1/2}\right\}, \end{aligned} \tag{3.11}$$

as long as the right-hand side of (3.11) exceeds $C_1 r \alpha^2 / \min\{d_1, d_2\}$, where C_0, C_1 and C_2 are absolute constants. Here the probability is with respect to the random Poisson observations only.

Similar to [8, 98], the proof of Theorem 13 relies on information theoretic arguments outlined as follows. First we find a set of matrices $\chi \subset \mathcal{S}$ so that the distance between any $X^{(i)}, X^{(j)} \in \chi$, identified as $\|X^{(i)} - X^{(j)}\|_F$, is sufficiently large. Suppose we obtain measurements of a selected matrix in χ and recover it using an arbitrary method. Then we could determine which element of χ was chosen, if the recovered matrix is sufficiently close to the original one. However, there will be a lower bound on how close the recovered matrix can be to the original matrix, since due to Fano's inequality the probability of correctly identifying the chosen matrix is small.

Remark 3. *Fixing α, β and r, the conditions in the statement of Theorem 3 can be satisfied if we choose sufficiently large d_1 and d_2.*

Remark 4. *When $m \geq (d_1 + d_2)\log(d_1 d_2)$, the ratio between the upper bound in (3.9) and the lower bound in (3.11) is on the order of $\mathcal{O}(\log(d_1 d_2))$. Hence, the lower bound matches the upper bound up to a logarithmic factor.*

Our formulation and results for Poisson matrix completion are inspired by one-bit matrix completion [8], yet with several important distinctions. In one-bit matrix completion, the value of each observation Y_{ij} is binary-valued and hence bounded; whereas in our problem, each observation is a Poisson random variable which is unbounded and, hence, the arguments involve bounding measurements have to be changed. In particular, we need to bound $\max_{ij} Y_{ij}$ when Y_{ij} is a Poisson random variable with intensity M_{ij}. Moreover, the Poisson likelihood function is non Lipschitz (due to a bad point when M_{ij} tends to zero), and hence we need to introduce a lower bound on each entry of the matrix M_{ij}, which can be interpreted as the lowest required SNR. Other distinctions also include analysis taking into account of the property of the Poisson likelihood function, and using the KL divergence as well as the Hellinger distance that are different from those for the Bernoulli random variable as used in [8].

3.3 Algorithms

In this section we develop efficient algorithms to solve the matrix completion problems (3.4). The problem (3.4) is semidefinite program (SDP), as they are nuclear norm minimization problems with convex feasible domains. Hence, we may solve it, for example, via the interior-point method [99]. Although the interior-point method returns an exact solution to (3.4), it does not scale well with the dimensions of the matrix d_1 and d_2 as the complexity of solving SDP is $O(d_1^3 + d_1 d_2^3 + d_1^2 d_2^2)$. Therefore, we develop two set of algorithms that can solve both problems faster than the interior point methods. These algorithms including the generic gradient descent based methods, and a Penalized Maximum Likelihood Singular Value Threshold (PMLSVT) method tailored to our problem. We analyze the performance of the generic methods. Although there is no theoretical performance guarantee, PMLSVT is computationally preferable under our assumptions. Another possible algorithm not cover here is the non-monotone spectral projected-gradient method [100, 8].

3.3.1 Generic methods

Here we only focus on solving the matrix completion problem (3.4) by proximal-gradient.

First, rewrite $\mathcal{S}$ in (3.3) as the intersection of two closed and convex sets in $\mathbb{R}^{d_1 \times d_2}$:

$$\begin{aligned}\Gamma_1 \triangleq \{X \in \mathbb{R}^{d_1 \times d_2} :& \beta \le X_{ij} \le \alpha, \\ & \forall (i,j) \in [\![d_1]\!] \times [\![d_2]\!]\},\end{aligned} \tag{3.12}$$

and

$$\Gamma_2 \triangleq \{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_* \le \alpha\sqrt{r d_1 d_2}\},$$

where the first set is a box and the second set is a nuclear norm ball. Let $f(X) \triangleq -F_{\Omega,Y}(X)$ be the negative log-likelihood function. Then optimization problem (3.4) is equivalent to

$$\widehat{M} = \arg \min_{X \in \Gamma_1 \bigcap \Gamma_2} f(X). \tag{3.13}$$

Noticing that the search space $\mathcal{S} = \Gamma_1 \bigcap \Gamma_2$ is closed and convex and $f(X)$ is a convex function, we can use proximal gradient methods to solve (3.13). Let $\mathbb{I}_\Gamma(X)$ be an extended function that takes value zero if $X \in \Gamma$ and value ∞ if $X \notin \Gamma$. Then (3.13) is equivalent to

$$\widehat{M} = \arg \min_{X \in \mathbb{R}^{d_1 \times d_2}} f(X) + \mathbb{I}_{\Gamma_1 \bigcap \Gamma_2}(X). \tag{3.14}$$

To guarantee the convergence of proximal gradient method, we need the Lipschitz constant $L > 0$ such that

$$\|\nabla f(U) - \nabla f(V)\|_F \leq L\|U - V\|_F, \quad \forall U, V \in \mathcal{S}. \tag{3.15}$$

Hence, $L = \alpha/\beta^2$ by the definition of our problem. Define the orthogonal projection of a matrix X onto a convex set $\widetilde{\Gamma}$ as

$$\Pi_{\widetilde{\Gamma}}(X) \triangleq \arg \min_{Z \in \widetilde{\Gamma}} \|Z - X\|_F^2.$$

Proximal gradient

Initialize the algorithm by $[X_0]_{ij} = Y_{ij}$ for $(i, j) \in \Omega$, and $[X_0]_{ij} = (\alpha + \beta)/2$ otherwise. Then iterate using

$$X_k = \Pi_{\mathcal{S}}(X_{k-1} - (1/L)\nabla f(X_{k-1})). \tag{3.16}$$

This proximal gradient method has a linear convergence rate:

Proposition 1 (Convergence of proximal gradient)**.** *Let $\{X_k\}$ be the sequence generated by (3.16). Then for any $k > 1$,*

$$f(X_k) - f(\widehat{M}) \leq \frac{L\|X_0 - \widehat{M}\|_F^2}{2k}.$$

Accelerated proximal gradient

Although proximal gradient can be easily implemented, it converges slowly when the Lipschitz constant L is large. In such scenarios, we may use Nesterov's accelerated method [101]. With the same initialization as above, we perform the following two projections at the kth iteration:

$$\begin{aligned} X_k &= \Pi_{\mathcal{S}}(Z_{k-1} - (1/L)\nabla f(Z_{k-1})), \\ Z_k &= X_k + ((k-1)/(k+2))\,(X_k - X_{k-1}). \end{aligned} \tag{3.17}$$

Nesterov's accelerated method converges faster:

Proposition 2 (Convergence of accelerated proximal gradient)**.** *Let $\{X_k\}$ be the sequence generated by (3.17). Then for any $k > 1$,*

$$f(X_k) - f(\widehat{M}) \leq \frac{2L\|X_0 - \widehat{M}\|_F^2}{(k+1)^2}.$$

Alternating projection

To use the above two methods, we need to specify ways to perform projection onto the space $\mathcal{S}$. Since $\mathcal{S}$ is an intersection of two convex sets, we may use alternating projection to compute a sequence that converges to the intersection of Γ_1 and Γ_2. Let U_0 be the matrix to be projected onto $\mathcal{S}$. Specifically, the following two steps are performed at the jth iteration: $V_j = \Pi_{\Gamma_2}(U_{j-1})$ and $U_j = \Pi_{\Gamma_1}(V_j)$, until $\|V_j - U_j\|_F$ is less than a user-specified error tolerance. Alternating projection is efficient if there exist some closed forms for projection onto the convex sets, which is true in our problems. Projection onto the box constraint Γ_1 is quite simple: $[\Pi_{\Gamma_1}(Y)]_{ij}$ assumes value β if $Y_{ij} < \beta$ and assumes value α if $Y_{ij} > \alpha$, and otherwise maintains the same value Y_{ij} if $\beta \leq Y_{ij} \leq \alpha$. Projection onto Γ_2, the nuclear norm ball, can be achieved by projecting the vector of singular values onto a ℓ_1 norm ball via scaling [30] [102].

3.3.2 Penalized maximum likelihood singular value threshold (PMLSVT)

We also develop an algorithm, referred to as PMLSVT, tailored to solving our Poisson problems. PMLSVT differs from the classical projected gradient in that instead of computing the exact gradient, it approximates the cost function by expanding it using a Taylor expansion up to the second order. The resulted approximate problem with a nuclear norm regularization term has a simple closed form solution using Theorem 2.1 in [30]. Therefore, PMLSVT does not perform gradient descent directly, but it has a simple form and good numerical accuracy as verified by numerical examples.

The algorithm is similar to the fast iterative shrinkage-thresholding algorithm (FISTA) [103] and its extension to matrix case with Frobenius error [48]. Similar to the construction in [37] and [49], using λ_0 and λ_1 as regularizing parameters and the convex sets Γ_0 and Γ_1 defined earlier in (**??**) and (3.12), we may rewrite (**??**) and (3.4) as

$$\widehat{M} = \arg\min_{X \in \Gamma_i} f_i(X) + \lambda_i \|X\|_*, \quad i = 0, 1, \tag{3.18}$$

respectively, where $f_0(X) = -\sum_{i=1}^{m} \{y_i \log[\mathcal{A}X]_i - [\mathcal{A}X]_i\}$ and $f_1(X) = -F_{\Omega,Y}(X)$.

The PMLSVT algorithm can be derived as follows (similar to [48]). For simplicity, we denote $f(X)$ for the $f_0(X)$ or $f_1(X)$. In the kth iteration, we may form a Taylor expansion of $f(X)$ around X_{k-1} while keeping up to the second term and then solve

$$X_k = \arg\min_{X} \left[Q_{t_k}(X, X_{k-1}) + \lambda \|X\|_*\right], \tag{3.19}$$

with

$$Q_{t_k}(X, X_{k-1}) \triangleq f(X_{k-1}) + \langle X - X_{k-1}, \nabla f(X_{k-1}) \rangle \\ + \frac{t_k}{2} \|X - X_{k-1}\|_F^2, \tag{3.20}$$

where ∇f is the gradient of f, t_k is the reciprocal of the step size at the kth iteration, which we will specify later. By dropping and introducing terms independent of M whenever needed (more details can be found in [104]), (3.19) is equivalent to

$$X_k = \\ \arg\min_X \left[\frac{1}{2} \left\| X - \left(X_{k-1} - \frac{1}{t_k} \nabla f(X_{k-1}) \right) \right\|_F^2 + \frac{\lambda}{t_k} \|X\|_* \right]. \tag{3.21}$$

Recall $d = \min\{d_1, d_2\}$. For a matrix $Z \in \mathbb{R}^{d_1 \times d_2}$, let its singular value decomposition be $Z = U \Sigma V^\intercal$, where $U \in \mathbb{R}^{d_1 \times d}$, $V \in \mathbb{R}^{d_2 \times d}$, $\Sigma = \text{diag}\{[\sigma_1, \ldots, \sigma_d]^\intercal\}$, and σ_i is a singular value of the matrix Z. For each $\tau \geq 0$, define the singular value thresholding operator as:

$$D_\tau(Z) \triangleq U \text{diag}\{[(\sigma_1 - \tau)^+, \ldots, (\sigma_d - \tau)^+]^\intercal\} V^\intercal.$$

To obtain a closed form solution to (3.21), we use the following proposition proved in [30]:

Proposition 3 (Theorem 2.1 in [30])**.** *For each* $\tau \ \geq 0$*, and* $Z \in \mathbb{R}^{d_1 \times d_2}$*:*

$$D_\tau(Z) = \arg\min_{X \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2} \|X - Z\|_F^2 + \tau \|Z\|_* \right\}. \tag{3.22}$$

Due to Proposition 3, the exact solution to (3.21) is given by

$$X_k = D_{\lambda/t_k} \left(X_{k-1} - \frac{1}{t_k} \nabla f(X_{k-1}) \right). \tag{3.23}$$

The PMLSVT algorithm is summarized in Algorithm 1. The initialization method for matrix recovery problem is suggested by [105] and that for matrix completion problem is to choose an arbitrary element in the set Γ_1. For a matrix Z, define a projection of Z onto Γ_0 as follows:

$$\mathcal{P}(Z) \triangleq \frac{I}{\|(Z)^+\|_{1,1}}(Z)^+,$$

where $(i,j)th$ entry of $(Z)^+$ is $(Z_{ij})^+$. In the algorithm description, t is the reciprocal of the step size, $\eta > 1$ is a scale parameter to change the step size, and K is the maximum number of iterations, which is user specified: a large K leads to more accurate solution, and a small K obtains the coarse solution quickly. If the cost function value does not decrease, the step size is shortened to change the singular values more conservatively. The algorithm terminates when the absolute difference in the cost function values between two consecutive iterations is less than $0.5/K$. Convergence of the PMLSVT algorithm cannot be readily established; however, Proposition 1 and Proposition 2 above may shed some light on this.

Algorithm 1 PMLSVT for Poisson matrix recovery and completion

```
Initialize: The maximum number of iterations $K$, parameters $\alpha$, $\beta$, $\eta$, and $t$.
$[X]_{ij} \leftarrow Y_{ij}$ for $(i,j) \in \Omega$ and $[X]_{ij} \leftarrow (\alpha+\beta)/2$ otherwise {matrix completion}
for $k = 1, 2, \dots K$ do
    $C \leftarrow X - (1/t)\nabla f(X)$
    $C = U\Sigma V^\intercal$ {singular value decomposition}
    $[\Sigma]_{ii} \leftarrow ([\Sigma]_{ii} - \lambda/t)^+$, $i = 1, \dots, d$
    $X' \leftarrow X$ {record previous step}
    $X \leftarrow \mathcal{P}(U\Sigma V^\intercal)$ {matrix recovery}
    $X \leftarrow \Pi_{\Gamma_1}(U\Sigma V^\intercal)$ {matrix completion}
    If $f(X) > Q_t(X, X')$ then $t \leftarrow \eta t$, go to 4.
    If $|f(X) - Q_t(X, X')| < 0.5/K$ then exit;
end for
```

Remark 5. *At each iteration, the complexity of PMLSVT (Algorithm 1) is on the order of $O(d_1^2 d_2 + d_2^3)$ (which comes from performing singular value decomposition). This is much lower than the complexity of solving an SDP, which is on the order of $O(d_1^3 + d_1 d_2^3 + d_1^2 d_2^2)$. In particular, for a d-by-d matrix, PMLSVT algorithm has a complexity $\mathcal{O}(d^3)$, which is lower than the complexity $\mathcal{O}(d^4)$ of solving an SDP. One may also use an approximate SVD* *method[106] and a better choice for step sizes [103] to accelerate PMLSVT.*

3.4 Numerical examples

In all the examples below we use our PMLSVT algorithm to solve the optimization problems. All the numerical examples are run on a laptop with 2.40Hz dual-core CPU and 8GB RAM.

3.4.1 Synthetic data based on solar flare image

We demonstrate the good performance of the PMLSVT algorithm for matrix completion on the same solar flare image as in the previous section. Set $\alpha = 200$ and $\beta = 1$ in this case. Suppose the entries are sampled via the Bernoulli model such that $\mathbb{E}[|\Omega|] = m$. Set $p \triangleq m/(d_1 d_2)$ in the sampling model. Set $t = 10^{-4}$ and $\eta = 1.1$ for PMLSVT. Fig. 3.1 shows the results when roughly 80%, 50% and 30% of the matrix entries are observed. Even when about 50% of the entries are missing, the recovery results is fairly good. When there are only about 30% of the entries are observed, PMLSVT still recovers the main features in the image. It is also quite fast: the run times for all three examples are less than 1.2 seconds.

3.4.2 Bike sharing count data

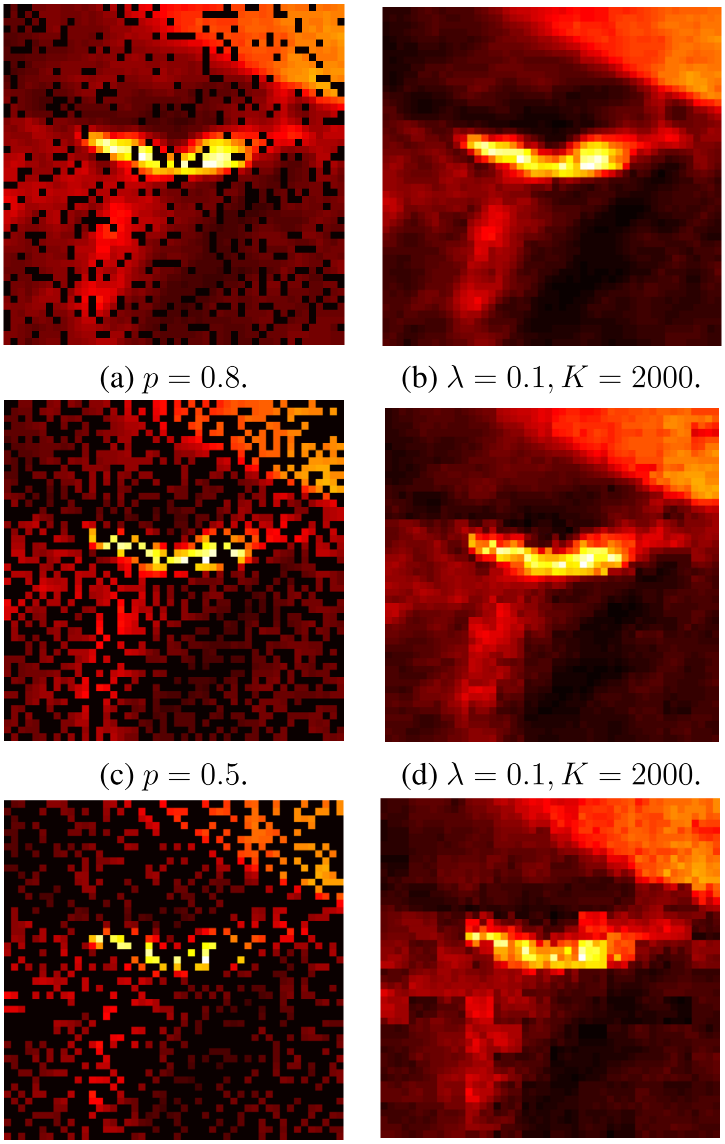
To demonstrate the performance of our algorithm on real data, we consider the bike sharing data set[1], which consists of 17379 bike sharing counts aggregated on hourly basis between the years 2011 and 2012 in Capital bike share system with the corresponding weather and seasonal information. We collect countings of 24 hours over 105 Saturdays into a 24-by-105 matrix M ($d_1 = 24$ and $d_2 = 105$). The resulted matrix is nearly low-rank. Assuming that only a fraction of the entries of this matrix are known (each entry is observed with probability 0.5 and, hence, roughly half of the entries are observed), and that the counting numbers follow Poisson distributions with unknown intensities. We aim at recover the unknown intensities, i.e., filling the missing data and performing denoising. We use PMLSVT with

[1] The data can be downloaded at
http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset[107].

the following parameters: $\alpha = 1000$, $\beta = 1$, $t = 10^{-4}$, $\eta = 1.1$, $K = 4000$ and $\lambda = 100$. In this case there is no "ground truth" for the intensities, and it is hard to measure the accuracy of recovered matrix. Instead, we are interested in identifying interesting patterns in the recovered results. As shown in Fig. 3.2(b), there are two clear increases in the counting numbers after the 17th and the 63th Saturday, which may not be easily identified from the original data in Fig. 3.2(a) with missing data and Poisson randomness.

3.5 Conclusions

In this paper, we have studied matrix recovery and completion problems when the data are Poisson random counts. We considered a maximum likelihood formulation with constrained nuclear norm of the matrix and entries of the matrix, and presented upper and lower bounds for the proposed estimators. We also developed a set of new algorithms, and in particular the efficient the Poisson noise Maximal Likelihood Singular Value Thresholding (PMLSV) algorithm. We have demonstrated its accuracy and efficiency compared with the semi-definite program (SDP) and tested on real data examples of solar flare images and bike sharing data.



(a) $p = 0.8$.

(b) $\lambda = 0.1, K = 2000$.

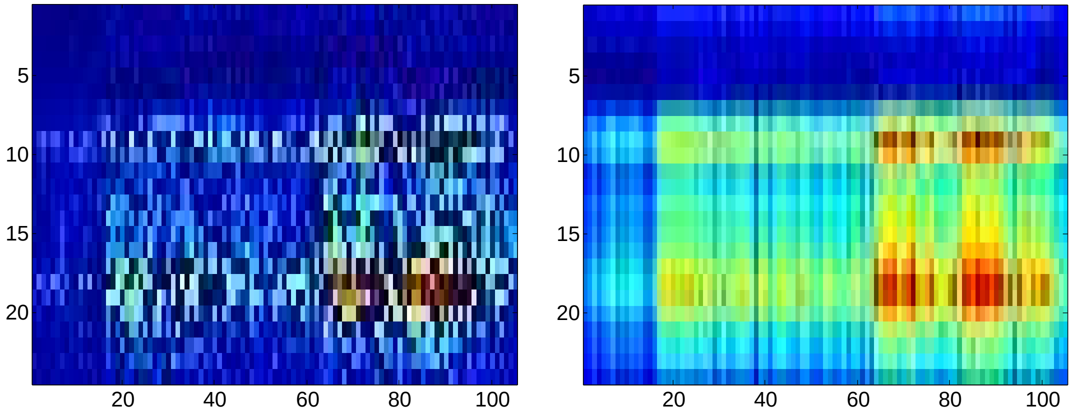
(c) $p = 0.5$.

(d) $\lambda = 0.1, K = 2000$.

(e) $p = 0.3$.

(f) $\lambda = 0.1, K = 2000$.

Figure 3.1: Matrix completion from partial observations: (a), (c), and (e): 80%, 50% and 30% of entries observed (dark spots represent missing entries); (b), (d), and (f): images formed by complete matrix with $\lambda = 0.1$ and no more than 2000 iterations, and the run times of the PMLSVT algorithm are 1.176595, 1.110226 and 1.097281 seconds, respectively.



(a) original data, $p = 0.5$.

(b) $\lambda = 100, K = 4000$.

Figure 3.2: Bike sharing count data: (a): observed matrix M with 50% missing entries; (b): recovered matrix with $\lambda = 100$ and 4000 iterations, with an elapsed time of 3.147153 seconds.

CHAPTER 4
MULTI-SENSOR SLOPE CHANGE DETECTION

In this chapter, I present the details about sequential gradual change detection in the multi-sensor case. The work is mainly summarized in [13]. Section 4.1 sets up the formalism of the problem. Section 4.2 presents our mixture procedures for slope change detection, and Section 4.3 presents theoretical approximations to its ARL and EDD, which are validated by numerical examples. Section 4.4 establishes the first order asymptotic optimality. Section 4.5 shows real-data examples. Finally, Section 4.6 presents an extension of the mixture procedure that adaptively chooses p_0. All proofs are delegated to the appendix.

4.1 Assumptions and formulation

Given N sensors. For the nth sensor $n = 1, 2, \ldots, N$, denote the sequence of observations by $y_{n,i}$, $i = 1, 2, \ldots$. Under the hypothesis of no change, the observations at the nth sensor have a *known* mean μ_n and a *known* variance σ_n^2. Probability and expectation in this case are denoted by $\mathbb{P}_\infty$ and $\mathbb{E}_\infty$, respectively. Alternatively, there exists an *unknown* change-point that occurs at time κ, $0 \leq \kappa < \infty$, and it affects an *unknown* subset $\mathcal{A} \subseteq \{1, 2, \ldots, N\}$ of sensors *simultaneously*. The fraction of affected sensors is given by $p = |\mathcal{A}|/N$. For each affected sensor $n \in \mathcal{A}$, the mean of the observations $y_{n,t}$ changes linearly from the change-point time $\kappa + 1$ and is given by $\mu_n + c_n(t - \kappa)$ for all $t > \kappa$, and the variance remains σ_n^2. For each unaffected sensor, the distribution stays the same. Here the *unknown* rate-of-change c_n can differ across sensors and it can be either positive or negative. The probability and expectation in this case are denoted by $\mathbb{P}_\kappa^{\mathcal{A}}$ and $\mathbb{E}_\kappa^{\mathcal{A}}$, respectively. In particular, $\kappa = 0$ denotes an immediate change occurring at the initial time. The above setting can formulate as the following hypothesis testing problem:

$$
\begin{aligned}
H_0: \quad & y_{n,i} \sim \mathcal{N}(\mu_n, \sigma_n^2), i = 1, 2, \ldots, n = 1, 2, \ldots, N, \\
H_1: \quad & y_{n,i} \sim \mathcal{N}(\mu_n, \sigma_n^2), i = 1, 2, \ldots, \kappa, \\
& \quad y_{n,i} \sim \mathcal{N}(\mu_n + c_n(i - \kappa), \sigma_n^2), i = \kappa + 1, \kappa + 2, \ldots, n \in \mathcal{A}, \\
& y_{n,i} \sim \mathcal{N}(\mu_n, \sigma_n^2), i = 1, 2, \ldots, n \in \mathcal{A}^c.
\end{aligned}
\tag{4.1}
$$

Our goal is to establish a stopping rule that stops as soon as possible after a change-point occurs and avoids raising false alarms when there is no change. We will make these statements more rigorous in Section 4.3 and Section 4.4. Here, for simplicity, we assume that all sensors are affected by the change simultaneously. This ignores the fact that there can be delays across sensors. For asynchronous sensors, one possible approach is to adopt the scheme in [108], which claims a change-point whenever the any sensor detects a change. We plan investigate the issue of delays in our future work.

A related problem is to detect a change in a linear regression model. One such example is a change-point in the trend of the stock price illustrated in Fig. 4.6(a). This can be casted into a slope change detection problem, if we fit a linear regression model under H_0 (e.g., using historical data) and subtract it from the sequence. The residuals after the subtraction will have zero means before the change-point, and their means will increase or decrease linearly after the change-point.

4.2 Detection procedures

Since the observations are independent, for an assumed change-point location $\kappa = k$ and an affected sensor $n \in \mathcal{A}$, the log-likelihood for observations up to time $t > k$ is given by

$$
\ell_n(k, t, c_n) = \frac{1}{2\sigma_n^2} \sum_{i=k+1}^{t} \left[2c_n(y_{n,i} - \mu_n)(i - k) - c_n^2 (i - k)^2 \right]. \tag{4.2}
$$

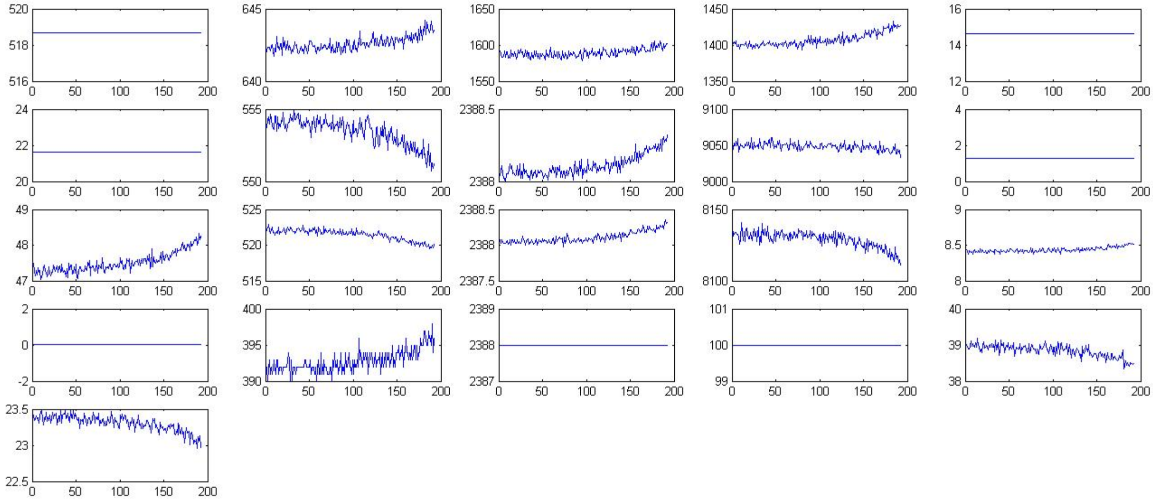


Figure 4.1: Degradation sample paths recorded by 21 sensors, generated by C-MAPSS [53]. A subset of sensors are affected by the change-point, which happens at an unknown time simultaneously and it causes a change in the slopes of the signals. The change can cause either an increase or decrease in the means.

Motived by the mixture procedure in [18] and [109] to exploit an empirical fact that typically only a subset of sensors are affected by the change-point, we assume that each sensor is affected with probability $p_0 \in (0, 1]$ independently. In this setting, the log likelihood of all N sensors is given by

$$\sum_{n=1}^{N} \log\left(1 - p_0 + p_0 \exp\left[\ell_n(k, t, c_n)\right]\right). \tag{4.3}$$

Using (4.3), we may derive several change-point detection rules.

Since the rate-of-change c_n is unknown, One possibility is to set c_n equal to some nominal post-change value δ_n and define the stopping rule, referred to as the *mixture CUSUM* procedure:

$$T_1 = \inf\left\{t : \max_{0 \le k < t} \sum_{n=1}^{N} \log\left(1 - p_0 + p_0 \exp[\ell_n(k, t, \delta_n)]\right) \ge b\right\}, \tag{4.4}$$

where b is a threshold typically prescribed to satisfy the average run length (ARL) requirement (formal definition of ARL is given in Section 4.3).

Another possibility is to replace c_n by its maximum likelihood estimator. Given the current number of observations t and a putative change-point location k, by setting the derivative of the log likelihood function (4.2) to 0, we may solve for the maximum likelihood estimator:

$$\hat{c}_n(k,t) = \frac{\sum_{i=k+1}^{t}(i-k)(y_{n,i}-\mu_n)}{\sum_{i=k+1}^{t}(i-k)^2}. \tag{4.5}$$

Define $\tau = t-k$ to be the number of samples after the change-point k. Denote the sum of squares from 1 to τ, and the weighted sum of data as, respectively,

$$A_\tau = \sum_{i=1}^{\tau} i^2, \qquad W_{n,k,t} = \sum_{i=k+1}^{t}(i-k)(y_{n,i}-\mu_n)/\sigma_n.$$

Let

$$U_{n,k,t} = (A_\tau)^{-1/2}\, W_{n,k,t}. \tag{4.6}$$

Substitution of (4.5) into (4.2) gives the log generalized likelihood ratio (GLR) statistic at each sensor:

$$\ell_n(k,t,\hat{c}_n) = U_{n,k,t}^2/2, \tag{4.7}$$

and we define the *mixture GLR* procedure as

$$T_2 = \inf\left\{t : \max_{0\le k<t}\sum_{n=1}^{N}\log\left(1-p_0+p_0\exp\left[U_{n,k,t}^2/2\right]\right)\ge b\right\}, \tag{4.8}$$

where b is a prescribed threshold.

Remark 6 (Window limited procedures.)**.** *In the following we use* window limited *versions of T_1 and T_2, where the maximum for the statistic is restricted to a window $t-w \le k \le t-w'$ for suitable choices of window size w and w'. In the following, we use $\widetilde{T}$ to denote a window-limited version of a procedure T. By searching only over a window of the past $w-w'+1$*

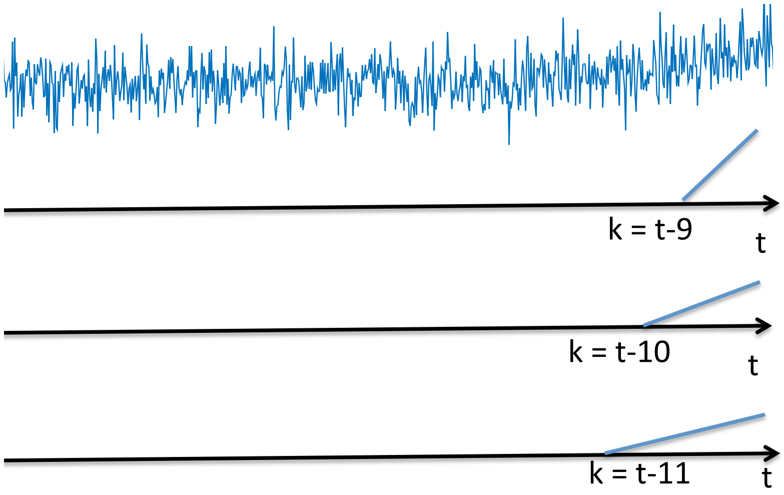


Figure 4.2: Matched filter interpretation of the generalized likelihood ratio statistic at each sensor $U_{n,k,t} = A_\tau^{-1} \sum_{i=k+1}^{t} (i-k)(y_{n,i} - \mu_n)/\sigma_n$: data at each sensor is matched with a triangle-shaped signal that starts at a hypothesized change-point time k and ends at the current time t. The slope of the triangle is A_τ^{-1}, so that the ℓ_2 norm of the triangle signal is one.

samples, this reduces the memory requirements to implement the stopping rule, and it also sets a minimum level of change that we want to detect. The choice of w may depend on b and sometimes we need make additional assumptions on w for the purpose of establishing the asymptotic results below. More discussions about the choice of w can be found in [110] and [54]. The other parameter w' is the minimum number of observations needed for computing the maximum likelihood estimator for parameters. In the following, we set $w' = 1$.

Remark 7 (Relation to mean shift.)**.** *For the mean-shift multi-sensor change-point detection [18], the detection statistic depends on a key quantify, which is the average of the samples in the time window $[k+1, t]$. Note that in the slope change case, the detection statistic has a similar structure, except that the key quantity is replaced by a weighted average of the samples in the window: $(t-k)^{-1/2} \sum_{i=k+1}^{t} (y_{n,i} - \mu_n)/\sigma_n$. This has an interpretation of "matched filtering", as illustrated in Fig. 4.2: each data stream is matched with a triangle shaped signal starting at a potential change-point time k that represents a possible slope* *change.*

Remark 8 (Recursive computation.)**.** *The quantity $W_{n,k,t}$ involved in the detection statistic for (4.8) can be calculated recursively,*

$$W_{n,k,t+1} = W_{n,k,t} + (t+1-k)\left((y_{n,t+1} - \mu_n)/\sigma_n\right),$$

where $W_{n,t,t} \triangleq 0$. This facilitates online implementation of the detection procedure. The quantity A_τ can be pre-computed since it is data-independent.

Remark 9 (Extension to correlated sensors.)**.** *The mixture procedure (4.8) can be easily extended to the case where sensors are correlated with a known covariance matrix. Define a vector of observations $\mathbf{y}_i = [y_{1,i}, \ldots, y_{N,i}]^\intercal$ for all sensors at time i. When there is no change, $\mathbf{y}_i$ follows a normal distribution with a mean vector $\mu = [\mu_1, \ldots, \mu_N]^\intercal$ and a covariance matrix Σ_0. Alternatively, there may exist a change-point at time κ such that after the change, the observation vectors are normally distributed with mean vector $\mu + (i - \kappa)\mathbf{c}$, $\mathbf{c} = [c_1, \ldots, c_n]^\intercal$ and the covariance matrix remains Σ_0 for all $i > \kappa$. We can whiten the signal vector by $\widetilde{\mathbf{y}}_i \triangleq \Sigma_0^{-1/2}(\mathbf{y}_i - \mu)$, where $\Sigma_0^{-1/2}$ is the square-root of the positive definite covariance matrix that may be computed via its eigen-decomposition. The coordinates of $\widetilde{\mathbf{y}}_i$ are independent and the problem then becomes the original hypothesis testing problem (4.1) with all sensors being affected simultaneously by the change-point, the rate-of-change vector is $\Sigma_0^{-1/2}\mathbf{c}$, the mean vector is zero before the change, and the covariance remains an identity matrix before and after the change. Hence, after the transform, we may apply the mixture procedure with $p_0 = 1$ on $\widetilde{\mathbf{y}}_i$.*

4.3 Theoretical properties of the detection procedures

In this section we develop theoretical properties of the mixture procedure. We use two standard performance metrics (1) the expected value of the stopping time when there is no change, the average run length (ARL); (2) the expected detection delay (EDD), defined to be the expected stopping time in the extreme case where a change occurs immediately at $\kappa = 0$. Since the observations are *i.i.d.* under the null, the EDD provides an upper bound on the expected delay after a change-point until detection occurs when the change occurs later in the sequence of observations (this is also a commonly used fact in change-point detection work [18]). An efficient detection procedure should have a large ARL and meanwhile a small EDD. Our approximation to the ARL is shown below to be accurate. In practice, we usually fix ARL to be a large constant, and set the threshold b in (4.8) accordingly. The accurate approximation here can be used to find the threshold analytically. Approximation for EDD shows its dependence on a quantity that plays a role of the Kullback-Leibler (KL) divergence, which links to the optimality results in Section 4.4.

4.3.1 Average run length (ARL)

We present an accurate approximation for ARL of a window limited version of the stopping rule in (4.8), which we denote as $\widetilde{T}_2$. Let

$$g(x) \triangleq \log(1 - p_0 + p_0 \exp(x^2/2)), \tag{4.9}$$

and

$$\psi(\theta) = \log \mathbb{E}\{\exp[\theta g(Z)]\},$$

where Z has a standard normal distribution. Also let

$$\gamma(\theta) = \frac{1}{2}\theta^2 \mathbb{E}\left\{[\dot{g}(Z)]^2 \exp\left[\theta g(Z) - \psi(\theta)\right]\right\},$$

and

$$H(N, \theta) = \frac{\theta[2\pi\ddot{\psi}(\theta)]^{1/2}}{\gamma^2(\theta)N^{1/2}} \exp\{N[\theta\dot{\psi}(\theta) - \psi(\theta)]\},$$

where the dot $\dot{f}$ and double-dot $\ddot{f}$ denote the first-order and second-order derivatives of a function f, respectively. Denote by $\phi(x)$ and $\Phi(x)$ the standard normal density function and its distribution function, respectively. Also define a special function $\nu(x) = 2x^{-2}\exp[-2\sum_{n=1}^{\infty} n^{-1}\Phi(-|x|n^{1/2}/2)]$. For numerical purposes an accurate approximation is given by [111]

$$\nu(x) \approx \frac{(2/x)[\Phi(x/2) - 1/2]}{(x/2)\Phi(x/2) + \phi(x/2)}.$$

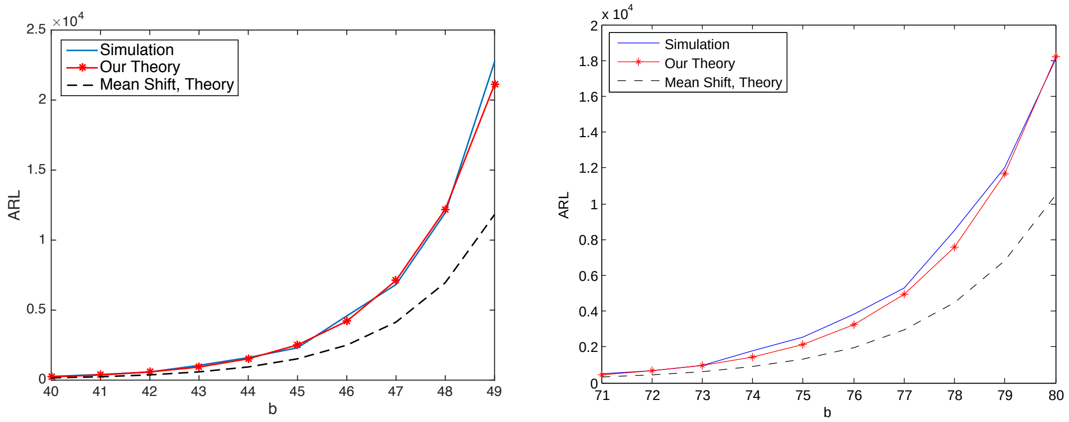
Theorem 14 (ARL of $\widetilde{T}_2$.)**.** *Assume that $N \to \infty$ and $b \to \infty$ with b/N fixed. Let θ be defined by $\dot{\psi}(\theta) = b/N$. For a window limited stopping rule of (4.8) with $w = o(b^r)$ for some positive integer r, we have*

$$\mathbb{E}_\infty\{\widetilde{T}_2\} = H(N,\theta) \cdot \left[\int_{\sqrt{2N/(4w/3)^{1/2}}}^{\sqrt{2N/(4/3)^{1/2}}} y\nu^2(y\sqrt{\gamma(\theta)})dy\right]^{-1} + o(1). \tag{4.10}$$

The proof of Theorem 14 is an extension of the proofs in [18] and [92] using the change of measure techniques. To illustrate the accuracy of approximation given in Theorem 14, we perform 500 Monte Carlo trials with $p_0 = 0.3$, and $w = 200$. Figs. 4.3(a) and (b) compare the simulated and theoretical approximation of ARL given in Theorem 14 when $N = 100$ and $N = 200$, respectively. Note that expression (4.10) takes a similar form as the ARL approximation obtained in [18] for the multi-sensor mean-shift case, and only differs in the upper and lower limits in the integration. In Figs. 4.3(a) and (b) we also plot the approximate ARL for the mean shift case in [18], which shows the importance of having the corrected integration upper and lower limits in our approximation. In practice, ARL is usually set to 5000 and 10000. Table 4.1 compares the thresholds obtained theoretically and from simulation at these two ARL levels, which demonstrates the accuracy of our approximation.

4.3.2 Expected detection delay (EDD)

After a change-point occurs, we are interested in the expected number of additional observations required for detection. In this section we establish an approximation upper bound to



(a) (b)

Figure 4.3: (a) Comparison of theoretical and simulated ARL when (a): $N = 100$, $p_0 = 0.3$, and $w = 200$; (b): $N = 200$, $p_0 = 0.3$, and $w = 200$.

Table 4.1: Theoretical versus simulated thresholds for $p_0 = 0.3$, $N = 100$ or 200, and $w = 200$.

	ARL	Theory b	Simulated ARL	Simulated b
$N = 100$	5000	46.34	5024	46.31
	10000	47.64	10037	47.60
$N = 200$	5000	77.04	5035	76.89
	10000	78.66	10058	78.59

the expected detection delay. Define a quantity

$$\Delta = \left(\sum_{n \in \mathcal{A}} c_n^2 / \sigma_n^2 \right)^{1/2}, \tag{4.11}$$

which roughly captures the total signal-to-noise ratio of all affected sensors.

Theorem 15 (EDD of $\widetilde{T}_2$.)**.** *Suppose $b \to \infty$, with other parameters held fixed. Let U be a standard normal random variable. If the window length w is sufficiently large and greater than $(6b/\Delta^2)^{1/3}$, then*

$$\mathbb{E}_0^{\mathcal{A}}\{\widetilde{T}_2\} \leq \left\{ \frac{b - |\mathcal{A}| \log p_0 - (N - |\mathcal{A}|)\mathbb{E}\{g(U)\}}{\Delta^2/6} \right\}^{1/3} + o(1), \tag{4.12}$$

where $\mathbb{E}_0^{\mathcal{A}}$ is defined at the beginning of Section 4.1. To demonstrate the accuracy of

(4.12), we perform 500 Monte Carlo trials. In each trial, we let the change-point happen at the initial time and randomly select Np sensors affected by the change and set the rate-of-change $c_n = c$ for a constant c, $n \in \mathcal{A}$. The thresholds for each procedure are set so that their ARLs are equal to 5000. Fig. 4.4 shows EDD versus c, where our upper bound turns out to be an accurate approximation to EDD.

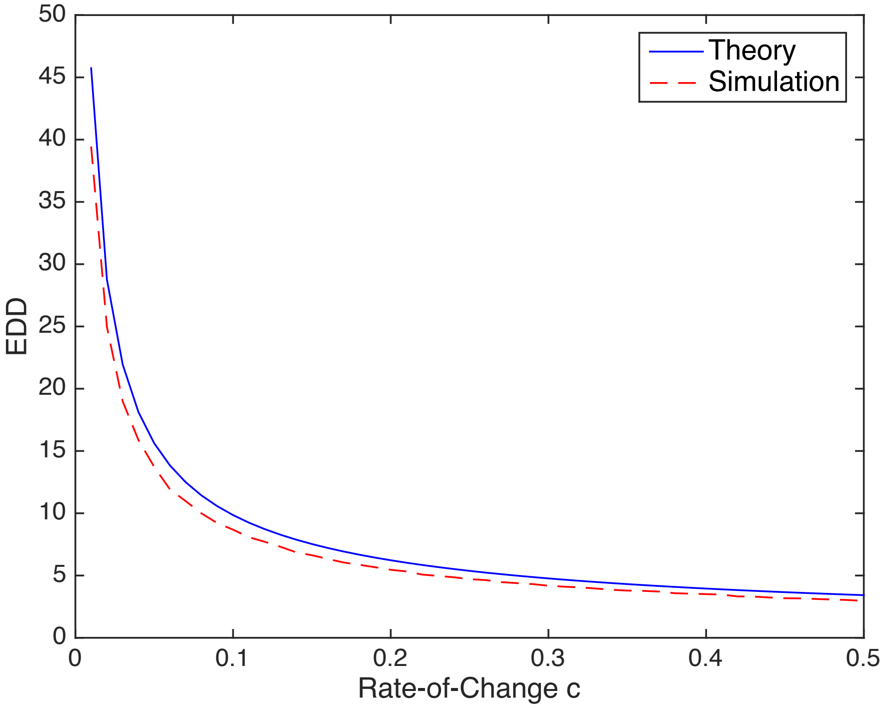


Figure 4.4: Comparison of theoretical and simulated EDD when $N = 100$, $p_0 = 0.3$, $p = 0.3$, and $w = 200$. All rate-of-change $c_n = c$ for affected sensors.

4.4 Optimality

In this section, we prove that our detection procedures: T_1 and the window limited versions $\widetilde{T}_1$ and $\widetilde{T}_2$ are asymptotically first order optimal. The optimality proofs here extends the results in [55], [54], for our multi-sensor *non-i.i.d.* data setting. The *non-i.i.d.*ness is due to the fact that under the alternative, the means of the samples change linearly as the number of post-change samples grows. Following the classic setup, we consider a class of detection procedures with their ARL greater than some constant γ, and then find an optimal procedure within such a class to minimize the detection delay. Since it is difficult to establish an uniformly optimal procedure for any given γ, we consider the asymptotic optimality when γ tends to infinity.

We first study a general setup with non-$i.i.d.$ distributions for the multi-sensor problem, and establish optimality of two general procedures related to T_1 and T_2. Then we specialize the results to the multi-sensor slope-change detection problem. In particular, we generalize the lower bound for the detection delay from the single sensor case (Theorem 8.2.2 in [55] and Theorem 1 in [54]) to our multi-sensor case. We also generalize the result therein to our setting where the log-likelihood ratio grows polynomially on the order of j^q for $q \geq 1$ as the number of post-change observations j grows (in the classic setting $q = 1$); this is used to account for the non-stationarity in our problem.

4.4.1 Setup for general non-$i.i.d.$ case

Consider a setup for the multi-sensor problem with non-$i.i.d.$ data. Assume there are N sensors that are independent (or with known covariance matrix so the observations can be whitened across sensors), and that the change-point affects all sensors simultaneously. Observations at the nth sensor are denoted by $x_{n,t}$ over time $t = 1, 2, \ldots$. If there is no change, $x_{n,t}$ are distributed according to conditional densities $f_{n,t}(x_{n,t}|x_{n,[1,t-1]})$, where $x_{n,[1,t-1]}$ = $(x_{n,1}, \ldots, x_{n,t-1})$ (this allows the distributions at time t to be dependent on the previous observations). Alternatively, if a change-point occurs at time κ and the nth sensor is affected, $x_{n,t}$ are distributed according to conditional densities $f_{n,t}(x_{n,t}|x_{n,[1,t-1]})$ for $t = 1, \ldots, \kappa$, and are according to $g_{n,t}^{(\kappa)}(x_{n,t}|x_{n,[1,t-1]})$ for $t > \kappa$. Note that the post-change densities are allowed to be dependent on the change-point κ. Define a filtration at time t by $\mathcal{F}_t = \sigma(x_{1,[1,t]}, \ldots, x_{N,[1,t]})$. Again, assume a subset $\mathcal{A} \subseteq \{1, 2, \ldots, N\}$ of sensors are affected by the change-point. Similar to Section 4.1, with a slight abuse of notation, we denote $\mathbb{P}_\infty$, $\mathbb{E}_\infty$, $\mathbb{P}_\kappa^\mathcal{A}$ and $\mathbb{E}_\kappa^\mathcal{A}$ as the probability and expectation when there is no change, or when a change occurs at time κ and a set $\mathcal{A}$ of sensors are affected by the change, with the understanding that here the probability measures are defined using the conditional densities.

4.4.2 Optimality criteria

We adopt two commonly used minimax criteria to establish the optimality of a detection procedure T. Similar to Chapter 8.2.5 of [55], we consider two criterions associated with the m-th moment of the detection delay for $m \geq 1$. The first criterion is motivated by Lorden's work [59], which minimizes the worst-case delay

$$\text{ESM}_m^{\mathcal{A}}(T) \triangleq \sup_{0 \leq k < \infty} \text{esssup}\, \mathbb{E}_k^{\mathcal{A}} \left\{ [(T-k)^+]^m | \mathcal{F}_k \right\}, \tag{4.13}$$

where "esssup" denotes the measure theoretic supremum that excluded points of measure zero. In other words, the definition (4.13) first maximizes over all possible trajectories of observations up to the change-point and then over the change-point time. The second criterion is motivated by Pollak's work [62], which minimizes the maximal conditional average detection delay

$$\text{SM}_m^{\mathcal{A}}(T) \triangleq \sup_{0 \leq k < \infty} \mathbb{E}_k^{\mathcal{A}} \left\{ (T-k)^m | T > k \right\}. \tag{4.14}$$

The extended Pollak's criterion (4.14) is not as strict as the extended Lorden's criterion in the sense that $\text{SM}_m^{\mathcal{A}}(T) \leq \text{ESM}_m^{\mathcal{A}}(T)$, and we prefer (4.14) since it is connected to the conventional decision theoretic approach and the resulted optimization problem can possibly be solved by a least favorable prior approach. The EDD defined earlier in Section 4.3 can be viewed as ESM_m and SM_m for $m = 1$, and the supremum over k happens when $k = 0$.

Define $C(\gamma)$ to be a class of detection procedures with their ARL greater than γ:

$$C(\gamma) \triangleq \{T : \mathbb{E}_\infty\{T\} \geq \gamma\}.$$

A procedure T is optimal, if it belongs to $C(\gamma)$ and minimizes $\text{ESM}_m(T)$ or $\text{SM}_m(T)$.

4.4.3 Optimality for general non-$i.i.d$ setup

Under the above assumptions, the log-likelihood ratio for each sensor is given by

$$\lambda_{n,k,t} = \sum_{i=k+1}^{t} \log \frac{g_{n,i}^{(k)}(x_{n,i}|x_{n,[1,i-1]})}{f_{n,i}(x_{n,i}|x_{n,[1,i-1]})}.$$

For any set $\mathcal{A}$ of affected sensors, the log-likelihood ratio is given by

$$\lambda_{\mathcal{A},k,t} = \sum_{n\in\mathcal{A}} \lambda_{n,k,t}. \tag{4.15}$$

We first establish an lower bound for any detection procedure. The constant $I_{\mathcal{A}}$ below can be understood intuitively as a surrogate for the Kullback-Leibler (KL) divergence in the hypothesis problem. When the observations are $i.i.d.$, $I_{\mathcal{A}}$ is precisely the KL divergence [54].

Theorem 16 (General lower bound.). *For any $\mathcal{A} \subseteq \{1,\ldots,N\}$ such that there exists some $q \geq 1$, $j^{-q}\lambda_{\mathcal{A},k,k+j}$ converges in probability to a positive constant $I_{\mathcal{A}} \in (0,\infty)$ under $\mathbb{P}_k^{\mathcal{A}}$,*

$$\frac{1}{j^q}\lambda_{\mathcal{A},k,k+j} \xrightarrow[j\to\infty]{\mathbb{P}_k^{\mathcal{A}}} I_{\mathcal{A}}, \tag{4.16}$$

and in addition, for all $\varepsilon > 0$, for an arbitrary $M \to \infty$

$$\sup_{0\leq k<\infty} \textit{esssup}\, \mathbb{P}_k^{\mathcal{A}} \left\{ M^{-q} \max_{0\leq j<M} \lambda_{\mathcal{A},k,k+j} \geq (1+\varepsilon)I_{\mathcal{A}} \middle| \mathcal{F}_k \right\} \xrightarrow[M\to\infty]{} 0. \tag{4.17}$$

Then,

(i) for all $0 < \varepsilon < 1$, there exists some $k \geq 0$ such that

$$\lim_{\gamma\to\infty} \sup_{T\in C(\gamma)} \mathbb{P}_k^{\mathcal{A}} \left\{ k < T < k + (1-\varepsilon)(I_{\mathcal{A}}^{-1}\log\gamma)^{\frac{1}{q}} \middle| T > k \right\} = 0. \tag{4.18}$$

(ii) for all $m \ge 1$,

$$\liminf_{\gamma\to\infty} \frac{\inf_{T\in C(\gamma)} \mathit{ESM}_m^{\mathcal{A}}(T)}{(\log\gamma)^{m/q}} \ge \liminf_{\gamma\to\infty} \frac{\inf_{T\in C(\gamma)} \mathit{SM}_m^{\mathcal{A}}(T)}{(\log\gamma)^{m/q}} \ge \frac{1}{I_{\mathcal{A}}^{m/q}}. \tag{4.19}$$

Consider a general mixture CUSUM procedure related to T_1, which has also been studied in [109] and [18]:

$$T_{\mathrm{CS}} = \inf\left\{ t : \max_{0\le k<t} \sum_{n=1}^{N} \log(1-p_0+p_0\exp(\lambda_{n,k,t})) \ge b \right\}, \tag{4.20}$$

where b is a prescribed threshold. The following lemma shows that for an appropriate choice of the threshold b, T_{CS} has an ARL lower bounded by γ and, hence, for such thresholds it belongs to $C(\gamma)$.

Lemma 8. *For any $p_0 \in (0,1]$, $T_{\mathrm{CS}}(b) \in C(\gamma)$, provided $b \ge \log\gamma$.*

Theorem 17 (Optimality of T_{CS}.)**.** *For any $\mathcal{A} \subseteq \{1,\dots,N\}$ such that there exists some $q \ge 1$ and a finite positive number $I_{\mathcal{A}} \in (0,\infty)$ for which (4.17) holds, and for all $\varepsilon \in (0,1)$ and $t \ge 0$,*

$$\sup_{0\le k<t} \mathit{esssup}\, \mathbb{P}_k^{\mathcal{A}}\left(j^{-q}\lambda_{\mathcal{A},k,k+j} < I_{\mathcal{A}}(1-\varepsilon) \middle| \mathcal{F}_k\right) \xrightarrow[j\to\infty]{} 0. \tag{4.21}$$

If $b \ge \log\gamma$ and $b = \mathcal{O}(\log\gamma)$, then T_{CS} is asymptotically minimax in the class $C(\gamma)$ in the sense of minimizing $\mathit{ESM}_m^{\mathcal{A}}(T)$ and $\mathit{SM}_m^{\mathcal{A}}(T)$ for all $m \ge 1$ to the first order as $\gamma \to \infty$.

We can also prove that the window-limited version $\widetilde{T}_{\mathrm{CS}}$ is asymptotically optimal. Since the window length affects ARL and the detection delay, in the following we denote this dependence more explicitly by w_γ.

Corollary 2 (Optimality of $\widetilde{T}_{\mathrm{CS}}$.)**.** *Assume the conditions in Theorem 17 hold and in addition,*

$$\liminf_{\gamma\to\infty} \frac{w_\gamma}{(\log\gamma/I_{\mathcal{A}})^{1/q}} > 1. \tag{4.22}$$

If $b \geq \log \gamma$ and $b = \mathcal{O}(\log \gamma)$, then $\widetilde{T}_{\text{CS}}(b)$ is asymptotically minimax in the class $C(\gamma)$ in the sense of minimizing ESM$_m^{\mathcal{A}}(T)$ and SM$_m^{\mathcal{A}}(T)$ for all $m \geq 1$ to the first order as $\gamma \to \infty$.

Intuitively, this means that the window length should be greater than the first order approximation to the detection delay $[(\log \gamma)/I_{\mathcal{A}}]^{1/q}$. Note that our earlier result (4.12) for the expected detection delay of the multi-sensor case is of this form for $q = 3$ and $I_{\mathcal{A}} = \Delta^2/6$.

Similarly, we may consider a general mixture GLR procedure related to T_2 as in [18]. Denote the log-likelihood (4.15) as $\lambda_{\mathcal{A},k,t}(\theta)$ to emphasize its dependence on an unknown parameter θ. The mixture GLR procedure maximizes θ over a parameter space Θ before combining them across all sensors. Unfortunately, we are unable to establish the asymptotic optimality for the general GLR procedure and its window limited version, due to a lack of martingale property.

4.4.4 Optimality for multi-sensor slope change

Note that T_1 and $\widetilde{T}_1$ correspond to special cases of T_{CS}, $\widetilde{T}_{\text{CS}}$, so we can use Theorem 17 and Corollary 2 to show their optimality by checking conditions. Although we are not able to establish optimality of the general mixture GLR procedure as mentioned above, we can prove the optimality for $\widetilde{T}_2$ by exploiting the structure of the problem.

Lemma 9 (Lower bound.). *For the multi-sensor slope change detection problem in (4.1), for a non-empty set $\mathcal{A} \subseteq \{1, \ldots, N\}$, the conditions of Theorem 16 are satisfied when $q = 3$ and $I_{\mathcal{A}} = \Delta^2/6$.*

The following lemma plays a similar role as the general version Lemma 8 in our multi-sensor case in (4.1), and it shows that for a properly chosen threshold b, ARL of $\widetilde{T}_2$ is lower bounded by γ and, hence, for such threshold it belongs to $C(\gamma)$.

Lemma 10. *For any $p_0 \in (0, 1]$, $\widetilde{T}_2(b) \in C(\gamma)$, provided*

$$b \geq N/2 - 4 \log \left[1 - (1 - 1/\gamma)^{1/w_\gamma}\right].$$

Remark 10 (Implication on window length.)**.** *Lemma 10 shows that to have $b = O(\log \gamma)$, we need* $\log w_\gamma = o(\log \gamma)$.

Theorem 18 (Asymptotical optimality of T_1, $\widetilde{T}_1$ and $\widetilde{T}_2$.)**.** *Consider the multi-sensor slope change detection problem (4.1).*

(i) If $b \geq \log \gamma$ and $b = \mathcal{O}(\log \gamma)$, then $T_1(b)$ is asymptotically minimax in class $C(\gamma)$ in the sense of minimizing expected moments ESM$^{\mathcal{A}}_m(T)$ and SM$^{\mathcal{A}}_m(T)$ for all $m \geq 1$ to the first order as $\gamma \to \infty$.

(ii) In addition to conditions in (i), if the window length satisfies

$$\liminf_{\gamma \to \infty} \frac{w_\gamma}{[6(\log \gamma)/\Delta^2]^{1/3}} > 1, \tag{4.23}$$

then $\widetilde{T}_1(b)$ is asymptotically minimax in class $C(\gamma)$ in the sense of minimizing expected moments ESM$^{\mathcal{A}}_m(T)$ and SM$^{\mathcal{A}}_m(T)$ for all $m \geq 1$ to first order as $\gamma \to \infty$.

(iii) If $b \geq N/2 - 4 \log[1 - (1 - 1/\gamma)^{1/w_\gamma}]$, $b = \mathcal{O}(\log \gamma)$, the window length satisfies $\log(w_\gamma) = o(\log \gamma)$ *and (4.23) holds, then $\widetilde{T}_2(b)$ is asymptotically minimax in class $C(\gamma)$ in the sense of minimizing ESM$^{\mathcal{A}}_m(T)$ and SM$^{\mathcal{A}}_m(T)$ for $m = 1$ to first order as $\gamma \to \infty$.*

Remark 11. *Above we prove the optimality of $T_1(b)$ and $\widetilde{T}_1(b)$ for $m \geq 1$. However, we can only prove the optimality of $\widetilde{T}_2(b)$ for a special case $m = 1$, due to a lack of martingale properties here.*

4.5 Numerical Examples

4.5.1 Comparison with mean-shift GLR procedures

We compare the mixture procedure for slope change detection, with the classic multivariate CUSUM [112] and the mixture procedure for mean shift detection [18]. The multivariate CUSUM essentially forms a CUSUM statistic at each sensor, and raises an alarm whenever a single sensor statistic hits the threshold. As commented earlier in Remark 7, the only difference between $\widetilde{T}_2$ and the mixture procedure for mean shift in [18] is how $U_{n,k,t}$ is defined. Following the steps for deriving (A.44), we can show that the mean shift mixture procedure is also asymptotically optimal for the slope change detection problem. Here, our numerical example verifies this, and show that the improvement of EDD by using $\widetilde{T}_2$ versus the multi-variate CUSUM and the mean-shift mixture procedure is not significant. However, the mean-shift mixture procedure fails to estimate the change-point time accurately due to model mismatch. Fig. 4.5 shows the mean square error for estimating the change-point time κ, using the multi-chart CUSUM, the mean-shift mixture procedure, and $\widetilde{T}_2$, respectively. Note that $\widetilde{T}_2$ has a significant improvement.

4.5.2 Financial time series

In the earlier example illustrated in Fig. 4.6(a), the goal is to detect a trend change online. Clearly a change-point occurs at time 8000 in the stock price, and such a change-point is verifiable. Fig. 4.6(b) shows that there is a peak in the bid size versus the ask size, which usually indicates a change in the trend of the price (possible with some delay). To illustrate the performance of our method in this financial dataset, we plot the detection statistics by using a “single-sensor”, i.e., using only one data stream, and by using “multi-sensor” scheme, i.e. using data from multiple streams, which in this case correspond to 8 factors (e.g, stock price, total volume, bid size and bid price, as well ask size and ask price). In fact, only 4 factors out of 8 factors contain the change-point. Fig. 4.6(c) plots the statistic if we

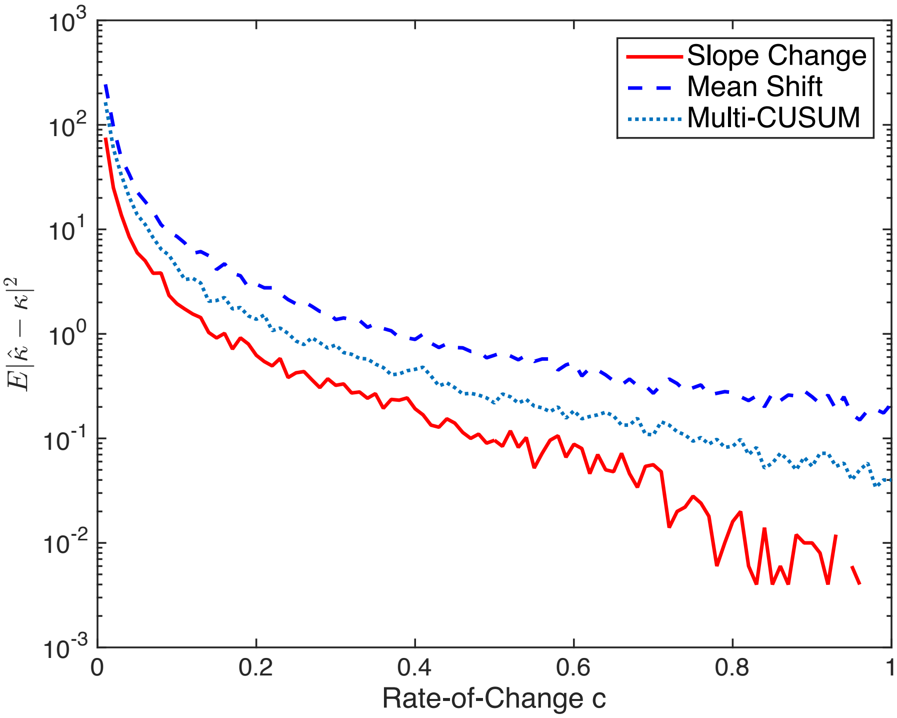
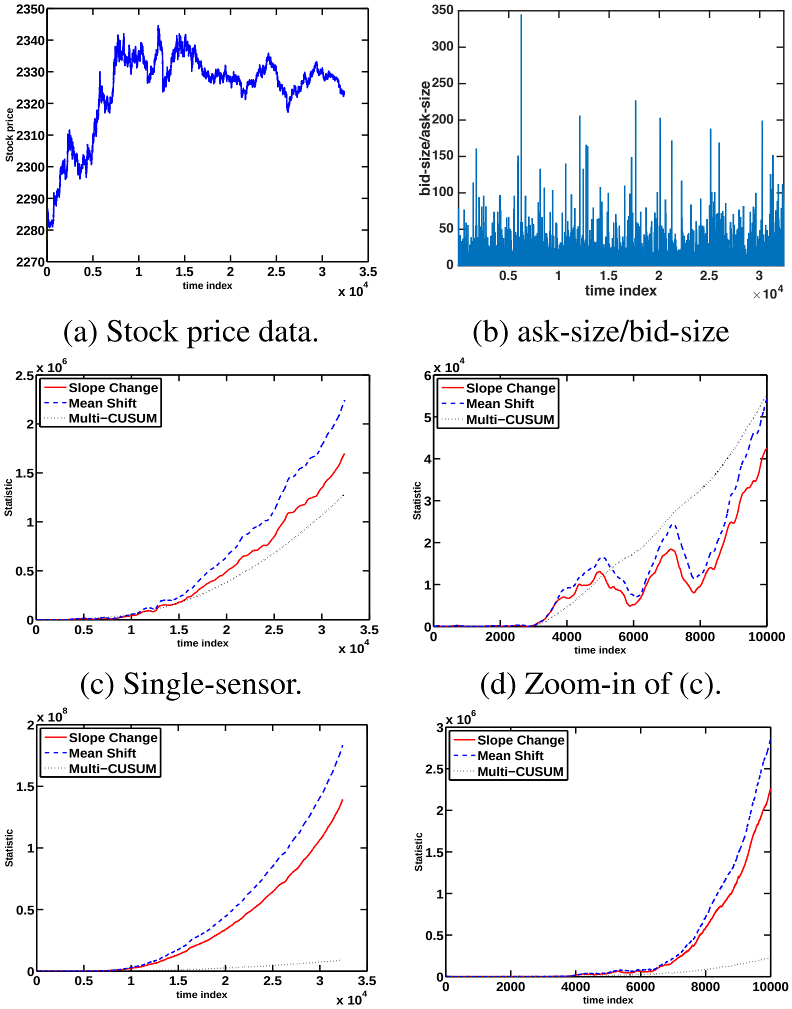


Figure 4.5: Comparison of mean square error for estimating the change-point time for the mixture procedure tailored to slope change $\widetilde{T}_2$, the mixture procedure with mean shift, and multi-chart CUSUM, when $N = 100$, $p_0 = 0.3$, $p = 0.5$ and $w = 200$.

use only a single-sensor. Fig. 4.6(e) illustrates the statistic when we use all the 8 factors and preprocess by whitening with the covariance of the factors as described in Section 9. Comparing Fig. 4.6(c) and Fig. 4.6(e), we can see that the sample path of the statistic designed for multi-sensor is smoother than that of the statistic designed for single-sensor. This means that the multi-sensor statistic is more robust to the noise than single-sensor statistic, and this is consistent with intuition since we take advantage of more information. Looking at Fig. 4.6(e), after the major trend change (around sample index 8000), the multi-chart CUSUM statistic rises the slowest. Although it appears, the slope-change mixture procedure rises a bit slower than the mean-shift mixture procedure, we demonstrate in simulation that for fixed ARL these two procedures have similar EDDs, and also in Fig. 5 that the slope-change mixture procedure has a better performance in estimating k^* than the mean-shift mixture procedure. Therefore, the slope-change mixture procedure is still preferrable.



(a) Stock price data.

(b) ask-size/bid-size

(c) Single-sensor.

(d) Zoom-in of (c).

(e) Multi-sensor, correlation.

(f) Zoom-in of (e).

Figure 4.6: Statistic for detecting trend changes in financial time series with $w = 500$ for both single sensor and multi-sensor procedures, and $p_0 = 1$ for the multi-sensor procedure.

4.5.3 Aircraft engine multi-sensor prognostic

We present an engine prognostic example using the aircraft turbofan engine dataset simulated by NASA[1]. In the dataset, multiple sensors measure different physical properties of the aircraft engine to detect a faulty condition and to predict the whole life time. The dataset

[1] Data can be downloaded from http://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository/

contains 100 training systems and 100 testing systems. Each system is monitored by $N = 21$ sensors. In the training dataset, we have a complete sample path from the initial time to failure for each of the 21 sensors of each training system. In the testing dataset, we only have partial sample paths (i.e., the system fails eventually but we have not observed that yet and it still has a remaining life). Our goal is to predict the whole life for the test systems using available observations. The dataset also provides ground truth, i.e., the actual failure times (or equivalently the whole life) of the testing systems.

We first apply our mixture procedures to each training system j, $j = 1, \ldots, 100$, to estimate a change-point location κ_j (which corresponds to the maximizer of k in the definition of $\widetilde{T}_2$ when the procedure stops), and the rate-of-change at nth sensor for the jth system $\hat{c}_{n,j}$ using (4.5). Then fit a simple survival model using $\hat{\kappa}_j$ and $\hat{c}_{n,j}$ as regressors in determining the remaining life. We build a model for the Time-To-Failure (TTF) Y_j of system j based a log location-normal model, which is commonly used in reliability theory [14]: $\mathbb{P}\{Y_j \leq y\} = \Phi\left[(\log(y) - \pi_j)/\eta\right]$, where η is a user specified scale parameter that is assumed to be the same for each system, π_j is the location parameter that is assumed to be a linear function of the rate-of-change: $\pi_j = \beta_0 + \sum_{n=1}^{N} \beta_j \hat{c}_{n,j}$, where $(\beta_0, \beta_1, \ldots, \beta_N)$ is a vector of the regression coefficients that are estimated by maximum likelihood. Next, we apply the mixture procedure on the jth testing system to estimate the change-point time $\hat{\kappa}_j$ and the rate-of-change $\hat{c}_{n,j}$, and substitute them into the fitted models to determine a TTF using the mean value. The whole life of the jth system is estimated as $\hat{\kappa}_j$ plus its mean TTF.

We use the relative prediction error as performance metric, which is the absolute difference between the estimated life and the actual whole life, divided by the actual whole life. Fig. 4.7 shows the box-plot of the relative prediction error versus threshold b. Our method based on change-point detection works well and it has a mean relative prediction error around 10%. Here the choice of the threshold b has a tradeoff: the relative prediction error decreases with a larger b; however, a larger b also causes a longer detection delay.

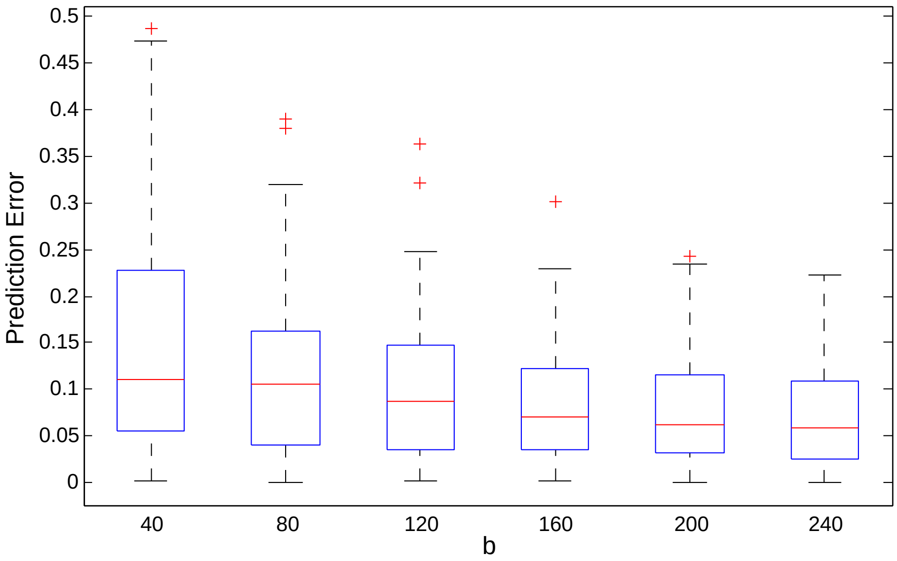


Figure 4.7: Aircraft engine prognostic example: box-plot for relative prediction error of the estimated life time of the engine versus threshold b.

4.6 Discussion: adaptive choice of p_0

The mixture procedure assumes that a fraction p_0 of the sensors are affected by the change. In practice, p_0 can be different from p which is the actual fraction of sensors affected. The performance of the procedure is fairly robust to the choice of p_0. Fig. 4.8 compares the simulated EDD of a mixture procedure with a fixed p_0 value, versus a mixture procedure when setting $p_0 = p$ if we know the true fraction of affected sensors. Again, thresholds are chosen such that ARL for all cases are 5000. Note that the detection delay is the smallest if p_0 matches p; however, EDD in these two settings are fairly close when $p_0 \neq p$.

Still, we may improve the performance of the mixture procedure by adapting the parameter p_0 using a method based on empirical Bayes. Assume each sensor is affected with probability p_0, but now p_0 itself is a random variable with Beta distribution Beta(α, β). This also allows the probability of being affected to be different at each sensor. With sequential data, we may update by computing a posterior distribution of p_0 using data in the following way. Choosing a constant a, we believe that the nth sensor is likely to be affected by the change-point if $U_{n,k,t}$ is larger than a. Let $\mathbb{I}\{\cdot\}$ denote an indicator function. For each t,

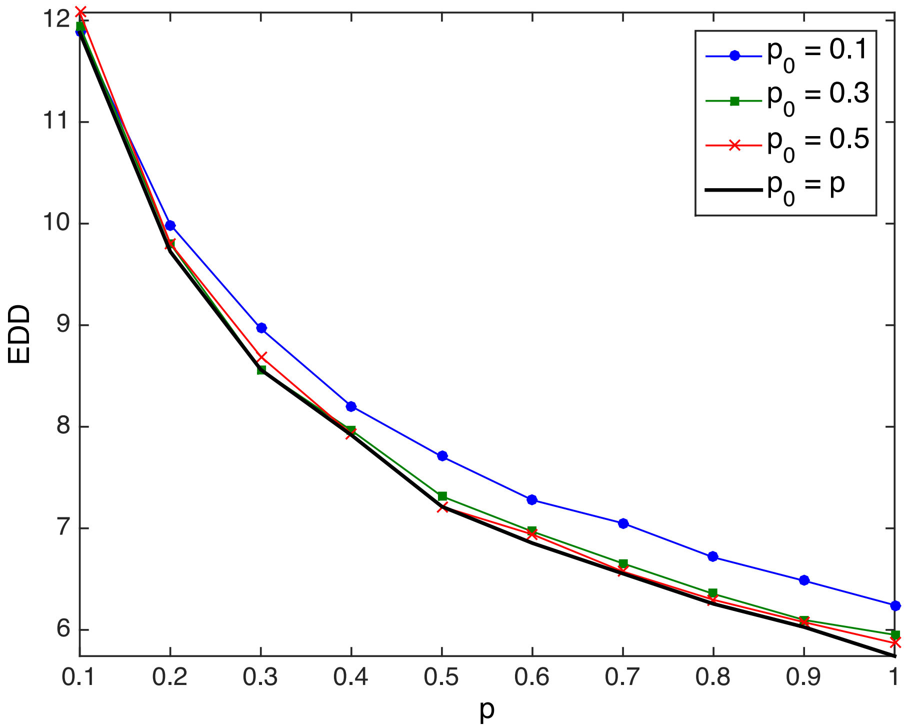


Figure 4.8: Simulated EDD for a mixture procedure with p_0 set to a fixed value, versus a mixture procedure with $p_0 = p$ equal to the true fraction of affected sensors, when $c_n = 0.1$, $N = 100$ and $w = 200$.

assume $s_{n,t} = \mathbb{I}\left\{\max_{t-w\leq k<t} U_{n,k,t} > a\right\}$ is a Bernoulli random variable with parameter p_n. Due to conjugacy, the posterior of p_0 at the nth sensor, given $s_{n,t}$ up to time t, is also a Beta distribution with parameters Beta$(s_{n,t}+\alpha, 1-s_{n,t}+\beta)$. An adaptive mixture procedure can be formed using the posterior mean of p_0, which is given by $\rho_n \triangleq (s_{n,t}+\alpha)/(\alpha+\beta+1)$:

$$T_{\text{adaptive}} = \inf\left\{t : \max_{t-w\leq k<t} \sum_{n=1}^{N} \log(1-\rho_n+\rho_n \exp(U_{n,k,t}^2/2)) \geq b\right\}, \tag{4.24}$$

where b is a prescribed threshold.

We compare the performance of $\widetilde{T}_{\text{adaptive}}$ with its non-adaptive counterpart $\widetilde{T}_2$ by numerical simulations. Assume $N = 100$ and there are 10 sensors affected from the initial time with a rate-of-change $c_n = c$. The parameters for $\widetilde{T}_{\text{adaptive}}$ are $\alpha = 1, \beta = 1$ and $a = 2$. Again, the thresholds are set so that the simulated ARL for both procedures are 5000. Table

4.6 shows that $\widetilde{T}_{\mathrm{adaptive}}$ has a much smaller EDD than $\widetilde{T}_2$ when signal is weak with a relative improvement around 20%. However, it is more difficult to analyze ARL of the adaptive method theoretically.

Table 4.2: Comparing EDD of $\widetilde{T}_2$ and $\widetilde{T}_{\mathrm{adaptive}}$.

Rate-of-change	0.01	0.03	0.05	0.07	0.09
Non-Adaptive $\widetilde{T}_2$	54.15	26.24	18.75	14.98	12.74
Adaptive $\widetilde{T}_{\mathrm{adaptive}}$	38.56	20.28	14.42	12.17	10.13

CHAPTER 5
ROBUST SEQUENTIAL CHANGE-POINT DETECTION WITH OFFLINE CONVEX OPTIMIZATION

In this chapter, I present the robust sequential change-point detection with offline convex optimization. This work is summarized in [113]. I address the computational challenge of finding the robust sequential change-point detection procedures when the pre- and post-change distributions are not completely specified. Earlier works [67, 68] establish the general conditions for robust procedures which include finding a pair of least favorable distributions (LFDs). However, in the multi-dimensional setting, it is hard to find such LFDs computationally. I present a method based on convex optimization that addresses this issue when the distributions are Gaussian with unknown parameters from pre-specified uncertainty sets. I also establish theoretical properties of our robust procedures, and numerical examples demonstrate their good performance.

5.1 Formulation

5.1.1 General setup

Assume that we observe a sequence of observations $\{\xi_i\}_{i=1}^{\infty}$ that take values in $\mathcal{X}$. Denote $\mathcal{P}(\mathcal{X})$ as the set of all the probability distributions on $\mathcal{X}$ and assume that there are two known distributions $\nu_0, \nu_1 \in \mathcal{P}(\mathcal{X})$. If there is no change, the observations are drawn i.i.d. from distribution ν_0. The probability and expectation in this case are denoted by $\mathbb{P}_{\infty}^{\nu_0}$ and $\mathbb{E}_{\infty}^{\nu_0}$, respectively. Alternatively, the i.i.d. observations $\xi_i \sim \nu_0$ for $i = 1, \ldots, \kappa - 1$, and at some *unknown* change-point κ, the distributions of the observations switch abruptly to ν_1, namely, $\xi_i \sim \nu_1$ for $i = \kappa, \kappa + 1, \ldots$. The observations are independent conditioned on the change-point κ. The probability and expectation in this case are denoted by $\mathbb{P}_{\kappa}^{\nu_0,\nu_1}$ and

$\mathbb{E}_{\kappa}^{\nu_0,\nu_1}$, respectively. In particular, $\kappa = 0$ denotes an immediate change occurring at the initial time.

A sequential change detection procedure is characterized by a stopping time T with respect to the observation sequence. To evaluate the performance of the detection procedure T, two performance measures are widely used: the average run length (ARL) and the expected detection delay (EDD). There are three commonly used mathematical formulations about ARL and EDD: Lorden's worst-case formulation in [59], Pollak's average worst-case formulation in [62] and the Bayesian formulation in [58]. In this paper, we adopt the Lorden's formulation, where the worst-case EDD of a detection procedure T is defined as follows:

$$\mathrm{WDD}(T) = \sup_{k\geq 1} \operatorname{esssup} \mathbb{E}_{k}^{\nu_0,\nu_1} \left[(T-k+1)^{+} \mid \mathcal{F}_{k-1} \right], \tag{5.1}$$

where $(x)^{+} = \max(x, 0)$. The quantity in (5.1) is called the worst-case EDD as a result of the two supreme appearing in (5.1). The first supreme means that the detection delay is taken over all possible locations of the change-point k and the second essential supreme means that the detection delay is taken over all possible realizations of the observations before the change-point k. ARL can be interpreted as the mean time between two false alarms, denoted by $\mathbb{E}_{\infty}^{\nu_0}[T]$. In practice, one usually fixes a lower bound γ for the ARL and denotes $C(\gamma)$ as the set of stopping times with ARL larger than $\gamma > 0$, in other words, $C(\gamma) = \{T : \mathbb{E}_{\infty}^{\nu_0}[T] \geq \gamma\}$. Then, our goal is to solve the following problem:

$$\min_{T \in C(\gamma)} \ \mathrm{WDD}(T). \tag{5.2}$$

In [59] and [63], it has been proven that the cumulative sum (CUSUM) procedure [57] is both the asymptotically optimal solution as $\gamma \to \infty$ and the exact optimal solution to (5.2) for any given $\gamma > 0$. Hence, in the following, we will focus on CUSUM-type procedures.

Now we consider the case when ν_0 and ν_1 are not specified exactly but belong to two classes of distributions $\mathcal{P}_0, \mathcal{P}_1 \in \mathcal{P}(\mathcal{X})$, respectively (e.g., [68]). Denote $C(\mathcal{P}_0, \gamma) = \{T :$

$\mathbb{E}_{\infty}^{\nu_0}[T] \geq \gamma, \forall \nu_0 \in \mathcal{P}_0\}$ as the set of all candidate stopping times whose ARL is lower bounded by γ. Then our goal is to solve the following robust version of (5.2):

$$\min_{T \in C(\mathcal{P}_0, \gamma)} \sup_{\nu_0 \in \mathcal{P}_0, \nu_1 \in \mathcal{P}_1} \text{WDD}(T), \tag{5.3}$$

Mean change: Assume that we observe a sequence of d-dimensional multivariate normal distribution with a known covariance matrix that does change. At some time κ, the mean vector switches from $\mu_0, \mu_0 \in \mathcal{M}_0$ to $\mu_1, \mu_1 \in \mathcal{M}_1$, where $\mathcal{M}_0$ and $\mathcal{M}_1$ are two known convex sets in $\mathbb{R}^d$ that are user-specified beforehand. The observations are independent conditioned on the change-point κ. Mathematically, we formulate the problem as the following hypothesis testing problem:

$$\begin{aligned} H_0 \quad &: \quad \xi_i \sim \mathcal{N}(\mu_0, \Sigma), \mu_0 \in \mathcal{M}_0, i = 1, 2, \ldots \\ H_1 \quad &: \quad \xi_i \sim \mathcal{N}(\mu_0, \Sigma), \mu_0 \in \mathcal{M}_0, i = 1, 2, \ldots, \kappa \\ &\quad\;\; \xi_i \sim \mathcal{N}(\mu_1, \Sigma), \mu_1 \in \mathcal{M}_1, i = \kappa+1, \kappa+2, \ldots, \end{aligned} \tag{5.4}$$

where Σ is the known positive definite covariance matrix. Here, the mean vector μ_0 and μ_1 can be any element in the convex sets $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively. For example, in the context of quality control, $\mathcal{M}_0$ can be defined as the set of all the allowable mean vectors if the system is in-control and $\mathcal{M}_1$ denotes the set of all the possible mean vectors if the system is out-of-control. Our goal is to identify the occurrence of the change as fast as possible subject to the false alarm constraints.

Covariance matrix change: Similarly, we may come up with a formulation when both the mean and the covariance matrix of the observations change. Assume a sequence of d-dimensional multivariate normal observations. At some time κ, the mean vector changes from $\mu_0, \mu_0 \in \mathcal{M}_0$ to $\mu_1, \mu_1 \in \mathcal{M}_1$ and the covariance matrix changes from $\Theta_0, \Theta_0 \in \mathcal{U}_0$ to $\Theta_1, \Theta_1 \in \mathcal{U}_1$, where $\mathcal{M}_0$ and $\mathcal{M}_1$ are two known convex sets in $\mathbb{R}^d$, $\mathcal{U}_0$ and $\mathcal{U}_1$ are two known convex sets in $\mathbb{S}_+^d$, which are user-specified beforehand. We formulate the problem as the following hypothesis testing problem:

$$\begin{aligned}
H_0: \quad & \xi_i \sim \mathcal{N}(\mu_0, \Theta_0), \mu_0 \in \mathcal{M}_0, \Theta_0 \in \mathcal{U}_0, i = 1, 2, \ldots \\
H_1: \quad & \xi_i \sim \mathcal{N}(\mu_0, \Theta_0), \mu_0 \in \mathcal{M}_0, \Theta_0 \in \mathcal{U}_0, i = 1, 2, \ldots, \kappa, \\
& \xi_i \sim \mathcal{N}(\mu_1, \Theta_1), \mu_1 \in \mathcal{M}_1, \Theta_1 \in \mathcal{U}_1, \\
& i = \kappa + 1, \kappa + 2, \ldots.
\end{aligned} \tag{5.5}$$

Even if the formulation (5.5) looks similar to the formulation (5.4), (5.5) is much more difficult than (5.4). For instance, a natural approach is to use sample mean and sample covariance matrices from the in-control and out-of-control data (there usually are these training data available in certain form) as the parameters before and after the change when designing the procedures. Then the uncertainty sets represents the estimation “precision”, which depend on the sample size and how the estimators are constructed. Mean vectors can usually be estimated up to good precision. However, it is much harder to estimate high-dimensional covariance matrix accurately (see, e.g, [114], [115], and [116]). Fortunately, most of the existing methods can guarantee that the true covariance matrix belongs to a convex set in $\mathbb{S}_+^d$, which enables us to reasonably construct uncertainty sets for covariance matrices.

5.2 Main results

5.2.1 Robust procedure for detecting mean change

For the robust version for mean shift detection (5.4), we consider a CUSUM-type procedure. CUSUM procedure needs specified likelihood ratio for two *singleton* pre-change and post-change distributions. Here, we solve a convex optimization problem to identify an appropriate pairs of parameters for the pre-change and post-change distributions, and use them to form the CUSUM procedure.

Let $\mathcal{P}_0 = \{\mathcal{N}(\mu_0, \Sigma), \mu \in \mathcal{M}_0\}$ and $\mathcal{P}_1 = \{\mathcal{N}(\mu_1, \Sigma), \mu \in \mathcal{M}_1\}$. Specifically, denote

(μ_0^*, μ_1^*) as the solution to the following convex optimization problem:

$$(\mu_0^*, \mu_1^*) = \underset{\mu_0 \in \mathcal{M}_0, \mu_1 \in \mathcal{M}_1}{\arg\min} (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1). \tag{5.6}$$

In other words, μ_0^* and μ_1^* are two points in $\mathcal{M}_0$ and $\mathcal{M}_1$ with the minimal Mahalanobis distance.

Our detection procedure is given as follows:

$$T_1 = \inf \left\{ t > 0 : \max_{1 \le k \le t} \sum_{i=k}^{t} \frac{1}{2} L^*(\xi_i) \ge b \right\}, \tag{5.7}$$

where L^* denotes the likelihood ratio between $\nu_1^* \sim \mathcal{N}(\mu_1^*, \Sigma)$ and $\nu_0^* \sim \mathcal{N}(\mu_0^*, \Sigma)$. The threshold b is chosen such that $\mathbb{E}_\infty^{\nu_0}[T_1] \ge \gamma$ for all $\nu_0 \in \mathcal{P}_0$ and a prescribed lower bound γ for ARL. We can show the following relationship between γ and b, which offers a guideline about how to determine b given any γ.

Theorem 19 (ARL)**.** *For any $\nu_0 \in \mathcal{P}_0$, for the detection procedure T_1 defined in (5.7), we have that $\mathbb{E}_\infty^{\nu_0}[T_1] \ge \gamma$ as long as*

$$b \ge \log \gamma + \log \frac{\epsilon^*}{1 - \epsilon^*}, \tag{5.8}$$

where

$$\epsilon^* = \exp(-\frac{1}{8}(\mu_0^* - \mu_1^*)^T \Sigma^{-1} (\mu_0^* - \mu_1^*)). \tag{5.9}$$

Remark 12. *When $\mathcal{P}_0 = \{\nu_0\}$ and $\mathcal{P}_1 = \{\nu_1\}$ are two singletons, T_1 is just the classic CUSUM procedure and the classic analysis tells us that if $b \ge \log \gamma$ then $\mathbb{E}_\infty^{\nu_0}[T_1] \ge \gamma$. The additional second term $\log(\epsilon^*/(1 - \epsilon^*))$ in (5.8) can be seen as a cost for the uncertainty. Specifically, ϵ^* is the upper bound for the Type-I and Type-II error for the one sample composite hypothesis testing problem: $H_0 : \xi \sim \nu_0, \nu_0 \in \mathcal{P}_0$ versus $H_1 : \xi \sim \nu_1, \nu_1 \in \mathcal{P}_1$.*

Next, we prove an upper bound for the worst-case detection delay as the threshold b goes to infinity.

Theorem 20 (EDD)**.** *For any $\nu_0 \in \mathcal{P}_0$ and $\nu_1 \in \mathcal{P}_1$, for the detection procedure T_1 defined in (5.7), as $b \to \infty$, we have that*

$$\mathit{WDD}(T_1) \leq \frac{b}{1-\epsilon^*}(1+o(1)),$$

where ϵ^* is defined in (5.9) and $o(1)$ is a vanishing term as $b \to \infty$. Therefore, as $\gamma \to \infty$, we can have both $\mathbb{E}_\infty^{\nu_0}[T_1] \geq \gamma$ and

$$\mathit{WDD}(T_1) \leq \frac{\log \gamma}{1-\epsilon^*}(1+o(1)),$$

where ϵ^* is defined in (5.9) and $o(1)$ is a vanishing term as $\gamma \to \infty$.

Remark 13. *Note that $1-\epsilon^*$ is just the Hellinger distance between the two multivariate normal distributions found by solving the convex optimization problem: $\mathcal{N}(\mu_0^*, \Sigma)$ and $\mathcal{N}(\mu_1^*, \Sigma)$. When $\mathcal{P}_0 = \{\nu_0\}$ and $\mathcal{P}_1 = \{\nu_1\}$ are two singletons, the classic analysis tells that the WDD(T_1) is asymptotically upper bounded by $2b/I$, where I is the Kullback-Leibler(KL) divergence between pre-change and post-change distributions. The Hellinger distance plays a similar role with the KL divergence as the denominator in Lorden's work [59]. Since KL divergence is known to be bounded below by Hellinger distance, our upper bound is a little bit looser. This can also be seen as the cost for uncertainty.*

Remark 14. *Define that $\bar{\nu}_0$ and $\bar{\nu}_1$ are true pre-change and post-change distributions. Since we can interpret the robust detection procedure T_1 as a repeated one-sided sequential probability ratio test (SPRT) between $\nu_0^* = \mathcal{N}(\mu_0^*, \Sigma)$ and $\nu_1^* = \mathcal{N}(\mu_1^*, \Sigma)$, we in fact can obtain that the WDD of T_1 is asymptotically upper bounded by $2b/(KL(\bar{\nu}_1 \| \nu_0^*) - KL(\bar{\nu}_1 \| \nu_1^*))$. As stated in the seminal work [68], compared with the optimal CUSUM* *procedure between $\bar{\nu}_0$ and $\bar{\nu}_1$, WDD(T_1) is asymptotically larger by a factor no more than*

$$\frac{KL(\bar{\nu}_0\|\bar{\nu}_1)}{KL(\bar{\nu}_1\|\nu_0^*) - KL(\bar{\nu}_1\|\nu_1^*)}.$$

Furthermore, as a consequence of theorem 20, for any two true pre-change and post-change distributions $\bar{\nu}_0$ and $\bar{\nu}_1$, we have that WDD(T_1) is asymptotically larger by a factor no more than $KL(\bar{\nu}_0\|\bar{\nu}_1)/[2(1-\epsilon^*)]$. When the Mahalanobis distance between $\mathcal{M}_0$ and $\mathcal{M}_1$ increases, ϵ^* in (5.9) becomes smaller and then factor above decreases, which means that our procedure moves closer to the optimal one. This is consistent with our intuition that one can detect the change more easily when the change is more obvious.

5.2.2 Robust procedure for detecting covariance change

Next, consider the case when both the mean vector and the covariance matrix of a multivariate normal distribution change and they belong to some uncertainty sets. In this case, we may consider linear and quadratic detectors, parameterized by vector h and matrix H defined below, as suggested in [73]. We include the original derivation from [73] below.

First we define the cost function, which can be viewed as exponential loss function which relates to the type-I and type-II error in the test (in the fixed sample size scenario). Let $\|\cdot\|$ denote the spectral norm and $\|\cdot\|_F$ the Frobenius norm, respectively. Let $\mathcal{U}$ be a convex compact set contained in the interior of the cone S_+^d of positive semidefinite $d\times d$ matrices in the space S^d of symmetric $d\times d$ matrices. Let $\Theta_*\in S_+^d$ be such that $\Theta_*\succeq\Theta$ for all $\Theta\in\mathcal{U}$, and let $\delta\in[0,2]$ be such that

$$\|\Theta^{1/2}\Theta_*^{-1/2} - I_d\| \leq \delta \;\; \forall\Theta\in\mathcal{U}. \tag{5.10}$$

Let $\mathcal{Z}$ be a nonempty convex compact subset of the set $\mathcal{Z}^+ = \{Z\in S_+^{d+1} : Z_{d+1,d+1}=1\}$, and let

$$\phi_{\mathcal{Z}}(Y) \triangleq \max_{Z\in\mathcal{Z}} \mathrm{Tr}(ZY) \tag{5.11}$$

be the support function of $\mathcal{Z}$. These specify the closed convex set

$$\mathcal{H} = \mathcal{H}^{\beta} := \{(h, H) \in \mathbb{R}^d \times \mathbb{S}^d : -\beta\Theta_*^{-1} \preceq H \preceq \beta\Theta_*^{-1}\}, \tag{5.12}$$

and the function $\Phi_{\mathcal{Z}} : \mathcal{H} \times \mathcal{U} \to \mathbb{R}$,

$$\begin{aligned} &\Phi_{\mathcal{Z}}(h, H; \Theta) = \\ &-\frac{1}{2}\log \mathrm{Det}(I - \Theta_*^{1/2} H \Theta_*^{1/2}) + \frac{1}{2}\mathrm{Tr}([\Theta - \Theta_*]H) \\ +&\frac{\delta(2+\delta)}{2(1 - \|\Theta_*^{1/2} H \Theta_*^{1/2}\|)}\|\Theta_*^{1/2} H \Theta_*^{1/2}\|_F^2 \\ +&\frac{1}{2}\phi_{\mathcal{Z}}\left(\left[\left[\begin{array}{c|c} H & h \\ \hline h^T & \end{array}\right] + [H, h]^T \left[\Theta_*^{-1} - H\right]^{-1} [H, h]\right]\right). \end{aligned} \tag{5.13}$$

Then, we have that $\Phi_{\mathcal{Z}}$ is continuous on its domain, convex in $(h, H) \in \mathcal{H}$ and concave in $\Theta \in \mathcal{U}$.

Next, we specify the uncertainty sets for the pre-change and post-change multivariate normal distributions. Given two collections of data as above: $(\mathcal{U}_\chi, \Theta_*^{(\chi)}, \beta_\chi, \mathcal{Z}_\chi)$, $\chi = 0, 1$, we define that

$$\begin{aligned} \mathcal{G}_\chi =& \{N(\mu, \Theta) : \Theta \in \mathcal{U}_\chi \\ & \quad \exists u : \mu = [u; 1], [u; 1][u; 1]^T \in \mathcal{Z}_\chi\},\ \chi = 0, 1. \end{aligned} \tag{5.14}$$

Now to solve for the quadratic detector (h, H), which will be applied on each individual samples and then used to construct the CUSUM recursion, we consider the convex-concave saddle point problem

$$\begin{aligned} \mathcal{SV} = & \min_{(h,H) \in \mathcal{H}_0 \cap \mathcal{H}_1} \max_{\Theta_0 \in \mathcal{U}_0, \Theta_1 \in \mathcal{U}_1} \\ & \underbrace{\tfrac{1}{2}\left[\Phi_{\mathcal{Z}_0}(-h, -H; \Theta_0) + \Phi_{\mathcal{Z}_1}(h, H; \Theta_1)\right]}_{\Phi(h, H; \Theta_0, \Theta_1)}. \end{aligned} \tag{5.15}$$

A saddle point $(H_*, h_*; \Theta_0^*, \Theta_1^*)$ in this problem does exist, which corresponds to the parameters of the quadratic detector and the picked worst-case parameters. We obtain the following quadratic detector

$$\begin{aligned}\phi^*(\xi) =& \frac{1}{2}\xi^T H_* \xi + h_*^T \xi + \\ & \underbrace{\frac{1}{2}\left[\Phi_{\mathcal{Z}_0}(-h_*, -H_*; \Theta_0^*) - \Phi_{\mathcal{Z}_1}(h_*, H_*; \Theta_1^*)\right]}_{a},\end{aligned} \tag{5.16}$$

Given above (which is pre-solved before we have seen any data), now given a sequence of data, we may evaluate ϕ^* in (5.16) for each sample and define our detection procedure as follows:

$$T_2 = \inf\left\{t > 0 : \max_{1\le k\le t} \sum_{i=k}^{t} (-\phi^*(\xi_i)) \ge b\right\}, \tag{5.17}$$

where b is a prescribed threshold.

Corollary 3 (ARL). *For any $\nu_0 \in \mathcal{G}_0$, for the detection procedure T_2 defined in (5.17), we have that $\mathbb{E}_\infty^{\nu_0}[T_2] \ge \gamma$ as long as*

$$b \ge \log\gamma + \log\frac{\epsilon^*}{1-\epsilon^*},$$

where

$$\epsilon^* = \exp(\mathcal{SV}) \tag{5.18}$$

and $\mathcal{SV}$ is defined in (5.15).

Corollary 4 (EDD). *For any $\nu_0 \in \mathcal{G}_0$ and $\nu_1 \in \mathcal{G}_1$, for the detection procedure T_2 defined in (5.17), as $b \to \infty$, we have that*

$$\mathit{WDD}(T_2) \le \frac{b}{1-\epsilon^*}(1 + o(1)),$$

where ϵ^* is defined in (5.18) and $o(1)$ is a vanishing term as $b \to \infty$. Therefore, as $\gamma \to \infty$, *we can have both* $\mathbb{E}_{\infty}^{\nu_0}[T_1] \geq \gamma$ *and*

$$WDD(T_2) \leq \frac{\log \gamma}{1-\epsilon^*}(1+o(1)),$$

where ϵ^* *is defined in (5.18) and* $o(1)$ *is a vanishing term as* $\gamma \to \infty$*.*

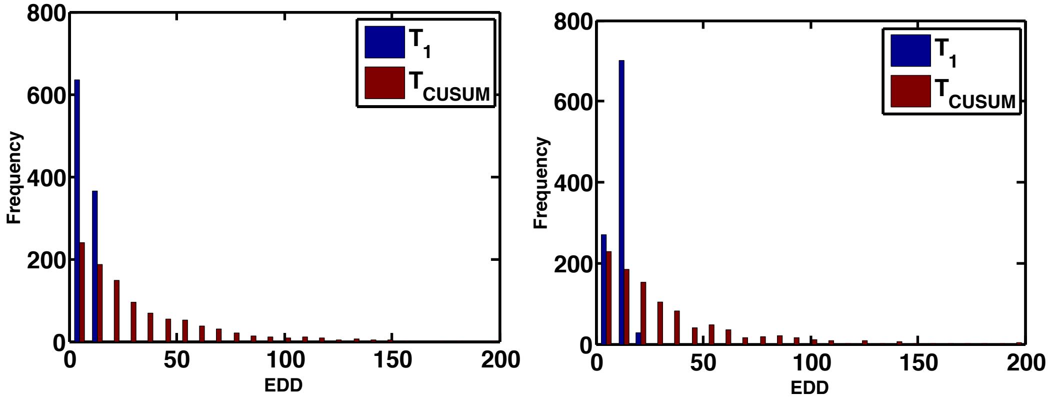


Figure 5.1: Histogram of detection delay. T_1 is established by solving (5.6) with: (Left) $\mathcal{M}_0 = \{0\}$ and $\mathcal{M}_1 = \{x \in \mathbb{R}^d : \|x - \mathbf{1}\|_1 \leq 27\}$; (Right) $\mathcal{M}_0 = \{0\}$ and $\mathcal{M}_1 = \{x \in \mathbb{R}^d : \|x - \mathbf{1}\|_2^2 \leq 27\}$. T_{CUSUM} is established by choosing pre-change distribution $\mathcal{N}(0, I)$ and post-change distribution $\mathcal{N}(\mathbf{1}, I)$. The result has 500 Monte Carlo trials.

5.3 Numerical examples

In this section, we numerically compare our procedures with the corresponding classic CUSUM procedure. In all the following experiments, we set the dimension $d = 30$ and choose b's such that the ARL of T_1 and T_{CUSUM} are both 5000. The classic CUSUM procedure are formed using randomly chosen pre-change and post-change distributions from the uncertainty sets. In the following, we denote **1** as an all-one vector.

5.3.1 Mean change detection

Assume $\mathcal{M}_0 = \{0\}$ and $\Sigma = I$ in (5.4). In the first example, set $\mathcal{M}_1 = \{x \in \mathbb{R}^d : \|x - \mathbf{1}\|_1 \leq 27\}$ in (5.4). We run 1000 experiments and for each run we choose a mean vector μ whose entries are random from $[0.1, 0.5]$, then generate the post-change observations from $\mathcal{N}(\mu, I)$. For classic CUSUM, we specify the pre-change distribution as $\mathcal{N}(0, I)$ and the post-change distribution as $\mathcal{N}(\mathbf{1}, I)$. Then we obtain 1000 simulated detection delays of T_1 and T_{CUSUM}, whose histograms are shown in Fig. 5.1. **The mean and standard deviation of detection delay of** T_1 **are** 7.6 **and** 2.3**, and those of** T_{CUSUM} **is** 32.2 **and** 30.1**, respectively**. In this case, T_1 performs much better than T_{CUSUM} since it is difficult to choose a good post-change distribution in $\mathcal{M}_1$ that is close to the true post-change distribution.

In the second example, the only difference between the second and the first example is that we replace the norm in $\mathcal{M}_1$ from ℓ_1 to ℓ_2. Set $\mathcal{M}_1 = \{x \in \mathbb{R}^d : \|x - \mathbf{1}\|_2^2 \leq 27\}$ in (5.4). We run 1000 experiments, and for each run we choose a mean vector μ whose entries are random from $[0.1, 0.5]$, then generate the post-change observations from $\mathcal{N}(\mu, I)$. For classic CUSUM, we specify the pre-change distribution to be $\mathcal{N}(0, I)$, and the post-change distribution to be $\mathcal{N}(\mathbf{1}, I)$. Then we obtain 1000 simulated detection delays of T_1 and T_{CUSUM}, whose histograms are shown in Fig. 5.1 (Right panel). **The mean and standard deviation of detection delay of** T_1 **is** 10.3 **and** 2.9**, and those of** T_{CUSUM} **is** 32.1 **and** 31.0**, respectively**. In this case, T_1 again performs much better than T_{CUSUM}.

5.3.2 Covariance matrix change detection

Consider $\mathcal{M}_0 = \mathcal{M}_1 = \{0\}$ and $\mathcal{U}_0 = \{I\}$ in (5.5). In the first example, we set $\mathcal{U}_1 = \{I + \sigma V, \sigma \in [0.5, 1]\}$ in (5.5), where V is a known matrix with diagonal entries $V_{i,i} = 0, i = 1, \ldots, d$ and off-diagonal entries $V_{i,j} = \exp(-(i-j)^2), i, j = 1, \ldots, d, i \neq j$. We run 500 experiments and for each run we randomly choose $\sigma \in [0.5, 1]$ and then generate the post-change observations from $\mathcal{N}(0, I + \sigma V)$. For classic CUSUM, we specify the pre-change distribution as $\mathcal{N}(0, I)$ and the post-change distribution as $\mathcal{N}(0, I + 0.75V)$. Then we obtain 500 experiments for T_2 and T_{CUSUM}, whose histograms are shown in Fig. 5.2. **The mean and standard deviation of detection delay of** T_2 **is** 9.10 **and** 4.21**, and those of** T_{CUSUM} **is** 8.28 **and** 5.10. In this case, there is no obvious difference between the two detection procedures, which means that T_2 performs almost as well as classical CUSUM procedure. The reason is that the set $\mathcal{U}_1$ is so small that the cost for mis-specified model is not large.

In the second example, consider the case with larger uncertainty sets: $\mathcal{U}_1 = \{\Theta \in \mathbb{S}_+^d : \|\Theta\|_2 \leq 0.5\}$ in (5.5). Again, we run 500 experiments and for each run we randomly choose a $\Sigma \in \mathcal{U}_1$ and generate the post-change observations from $\mathcal{N}(0, \Sigma)$. For classic CUSUM, we randomly choose a matrix in $\mathcal{U}_1$ as the covariance matrix of its post-change normal distribution. Then, we obtain the detection delays of T_2 and T_{CUSUM}, whose histogram are shown in Fig. 5.2. **The mean and standard deviation of detection delay of T_2 is 2.06 and 0.33, and those of T_{CUSUM} is 10.28 and 9.22.** In this case, T_2 outperforms T_{CUSUM} since $\mathcal{U}_1$ is a large convex set and cost for a misspecified model is greater. Note that for the above two choices of $\mathcal{U}_1$, (5.15) can be solved by first removing the inner maximum since the maximum is achieved at the boundary of $\mathcal{U}_1$. Then solving saddle point is equivalent to solving a convex optimization.

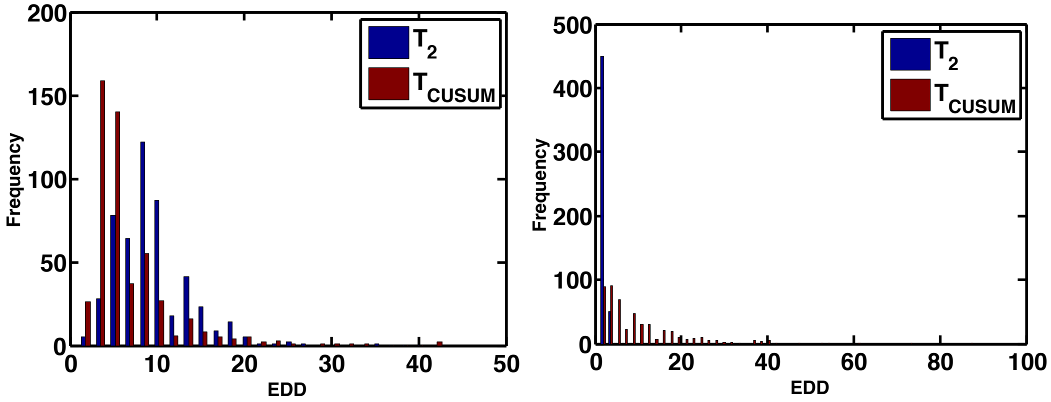


Figure 5.2: Histogram of detection delay. (Left) T_2 is established by solving (5.15) with $\Theta_*^{(0)} = I$, $\Theta_*^{(1)} = 2I$, $\Theta_* = 2I$ and $\beta_0 = \beta_1 = 0.5$. T_{CUSUM} is established by choosing pre-change distribution $\mathcal{N}(0, I)$ and post-change distribution $\mathcal{N}(0, I + 0.75V)$. (Right) T_2 is established by solving (5.15) with $\Theta_*^{(0)} = I$, $\Theta_*^{(1)} = 0.5I$ and $\beta_0 = \beta_1 = 0.5$. T_{CUSUM} is established by choosing pre-change distribution $\mathcal{N}(0, I)$ and post-change distribution $\mathcal{N}(0, \Sigma)$, where Σ is randomly chosen from $\mathcal{U}_1 = \{\Theta \in \mathbb{S}_+^d : \|\Theta\|_2 \leq 0.5\}$. The result has 500 Monte Carlo trials.

5.4 Conclusions

In this chapter, we propose robust detection procedures for detecting the change for mean vectors and covariance matrices, when they belong to some convex uncertainty sets. The proposed procedures are similar to classic CUSUM procedure, and the task is to determine appropriate pre-change and post-change distributions. Prior works look for pairs of least favorable distributions and use them to establish CUSUM procedure, but this is difficult in the multi-dimensional case. In this paper, we solve the pre-change and post-change distributions by convex optimization and this method is very efficient in both one dimensional and high dimensional cases. Moreover, from the asymptotic analysis, we obtain useful results to characterize the ARLs and EDDs and show the upper bound for the cost of robustness. Another advantage of our method is that the proposed procedures can be implemented recursively. Compared to Generalized Likelihood Ratio (GLR) methods, our method does not require solving a possibly complicated nonconvex optimization problem to estimate the parameters at each time.

CHAPTER 6
ROBUST SEQUENTIAL CHANGE-POINT DETECTION WITH ONLINE CONVEX OPTIMIZATION

In this chapter, I present the robust sequential change-point detection via online convex optimization. This work is summarized in [117].

6.1 Preliminaries

Assume a sequence of i.i.d. random variables $X_1, X_2, \ldots$ with a probability density function of a parametric form f_θ. The parameter θ may be unknown. Consider two related problems: one-sided sequential hypothesis test and sequential change-point detection. The detection statistic relies on a sequence estimators $\{\hat{\theta}_t\}$ constructed using online mirror descent. The OMD uses simple *one-sample update*: the update from $\hat{\theta}_{t-1}$ to $\hat{\theta}_t$ only uses the current sample X_t. This is the main difference from the traditional generalized likelihood ratio (GLR) statistic [54], where each $\hat{\theta}_t$ is estimated using historical samples. In the following, we present detailed descriptions for two problems. We will consider exponential family distributions and present our non-anticipating estimator based on the one-sample estimate.

6.1.1 One-sided sequential hypothesis test

First, we consider a one-sided sequential hypothesis test where the goal is only to reject the null hypothesis. This is a special case of the change-detection problem where the change-point can be either 0 or ∞ (meaning it never occurs). Studying this special case will given us an important intermediate step towards solving the sequential change-detection problem.

Consider the null hypothesis $\mathsf{H}_0 : \theta = \theta_0$ versus the alternative $\mathsf{H}_1 : \theta \neq \theta_0$. Hence the parameter under the alternative distribution is unknown. The classic approach to solve this problem is the one-sided sequential probablity-ratio test (SPRT) [118]: at each time, given samples $\{X_1, X_2, \ldots, X_t\}$, the decision is either to reject H_0 or taking more samples if the rejection decision cannot be made confidently. Here, we introduce a *modified* one-sided SPRT with a sequence of *non-anticipating* plug-in estimators:

$$\hat{\theta}_t := \hat{\theta}_t(X_1, \ldots, X_t), \quad t = 1, 2, \ldots. \tag{6.1}$$

Define the test statistic at time t as

$$\Lambda_t = \prod_{i=1}^{t} \frac{f_{\hat{\theta}_{i-1}}(X_i)}{f_{\theta_0}(X_i)}, \quad i \geq 1. \tag{6.2}$$

The test statistic has a simple recursive implementation:

$$\Lambda_t = \Lambda_{t-1} \cdot \frac{f_{\hat{\theta}_{t-1}}(X_t)}{f_{\theta_0}(X_t}.$$

Define a sequence of σ-algebras $\{\mathcal{F}_t\}_{t\geq 1}$ where $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$. The test statistic has the martingale property due to its non-anticipating nature: $\mathbb{E}[\Lambda_t \mid \mathcal{F}_{t-1}] = \Lambda_{t-1}$, where the expectation is taken when $X_1, \ldots$ are i.i.d. random variables drawn from f_{θ_0}. The decision rule is a stopping time

$$\tau(b) = \min\{t \geq 1 : \log \Lambda_t \geq b\}, \tag{6.3}$$

where $b > 0$ is a pre-specified threshold. We reject the null hypothesis whenever the statistic exceeds the threshold. The goal is to reject the null hypothesis using as few samples as possible under the false-alarm rate (or Type-I error) constraint.

6.1.2 Sequential change-point detection

Now we consider the sequential change-point detection problem. A change may occur at an unknown time ν which alters the underlying distribution of the data. One would like to detect such a change as quickly as possible. Formally, change-point detection can be cast into the following hypothesis test:

$$\begin{aligned} &\mathsf{H}_0: \; X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} f_{\theta_0}, \\ &\mathsf{H}_1: \; X_1, \ldots, X_\nu \overset{\text{i.i.d.}}{\sim} f_{\theta_0}, \quad X_{\nu+1}, X_{\nu+2}, \ldots \overset{\text{i.i.d.}}{\sim} f_{\theta}, \end{aligned} \tag{6.4}$$

Here we assume an unknown θ to represent the anomaly. The goal is to detect the change as quickly as possible after it occurs under the false-alarm rate constraint. We will consider likelihood ratio based detection procedures adapted from two types of existing ones, which we call the adaptive CUSUM (ACM), and the adaptive SRRS (ASR) procedures.

For change-point detection, the post-change parameter is estimated using post-change samples. This means that, for each putative change-point location before the current time $k < t$, the post-change samples are $\{X_k, \ldots, X_t\}$; with a slight abuse of notation, the post-change parameter is estimated as

$$\hat{\theta}_{k,i} = \hat{\theta}_{k,i}(X_k, \ldots, X_i), \quad i \geq k. \tag{6.5}$$

Therefore, for $k = 1$, $\hat{\theta}_{k,i}$ becomes $\hat{\theta}_i$ defined in (6.2) for the one-sided SPRT. Initialize with $\hat{\theta}_{k,k-1} = \theta_0$. The likelihood ratio at time t for a hypothetical change-point location k is given by

$$\Lambda_{k,t} = \prod_{i=k}^{t} \frac{f_{\hat{\theta}_{k,i-1}}(X_i)}{f_{\theta_0}(X_i)}, \tag{6.6}$$

where $\Lambda_{k,t}$ can be computed recursively similar to (6.2).

Since we do not know the change-point location ν, from the maximum likelihood principle, we take the maximum of the statistics over all possible values of k. This gives the ACM procedure:

$$T_{\text{ACM}}(b_1) = \inf\left\{t \geq 1 : \max_{1 \leq k \leq t} \log \Lambda_{k,t} > b_1\right\}, \tag{6.7}$$

where b_1 is a pre-specified threshold. Similarly, by replacing the maximization over k in (6.7) with summation, we obtain the following ASR procedure [75], which can be interpreted as a Bayesian statistic similar to the Shiryaev-Roberts procedure.

$$T_{\mathrm{ASR}}(b_2) = \inf\left\{t \geq 1 : \log\left(\sum_{k=1}^{t} \Lambda_{k,t}\right) > b_2\right\}, \tag{6.8}$$

where b_2 is a pre-specified threshold. The computations of $\Lambda_{k,t}$ and estimator $\{\hat{\theta}_t\}$, $\{\hat{\theta}_{k,t}\}$ are discussed later in section 6.1.4. For a fixed k, the comparison between our methods and GLR is illustrated in Figure 6.1.

Remark 15. *In practice, to prevent the memory and computation complexity from blowing up as time t goes to infinity, we can use window-limited version of the detection procedures in (6.7) and (6.8). The window-limited versions are obtained by replacing $\max_{1\leq k\leq t}$ with $\max_{t-w\leq k\leq t}$ in (6.7) and by replacing $\sum_{k=1}^{t}$ with $\sum_{k=t-w}^{t}$ in (6.8). Here w is a prescribed window size. Even if we do not provide theoretical analysis to the window-limited versions, we refer the readers to [54] for the choice of w the window-limited GLR procedures.*

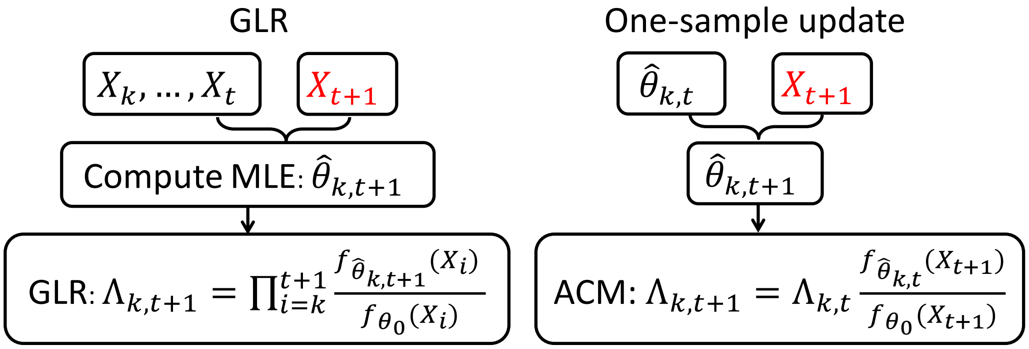

GLR
$X_k, \ldots, X_t$
X_{t+1}
Compute MLE: $\hat{\theta}_{k,t+1}$
GLR: $\Lambda_{k,t+1} = \prod_{i=k}^{t+1} \frac{f_{\hat{\theta}_{k,t+1}}(X_i)}{f_{\theta_0}(X_i)}$
One-sample update
$\hat{\theta}_{k,t}$
X_{t+1}
$\hat{\theta}_{k,t+1}$
ACM: $\Lambda_{k,t+1} = \Lambda_{k,t} \frac{f_{\hat{\theta}_{k,t}}(X_{t+1})}{f_{\theta_0}(X_{t+1})}$

Figure 6.1: Comparison of the update scheme for GLR and our methods when a new sample arrives.

6.1.3 Exponential family

In this paper, we focus on f_θ being the exponential family for the following reasons: (i) exponential family [77] represents a very rich class of parametric and even many nonparametric statistical models [119]; (ii) the negative log-likelihood function for exponential family $-\log f_\theta(x)$ is convex, and this allows us to perform online convex optimization. Some useful properties of the exponential family are briefly summarized below, and full proofs can be found in [120, 77].

Consider an observation space $\mathcal{X}$ equipped with a sigma algebra $\mathcal{B}$ and a sigma finite measure H on $(\mathcal{X}, \mathcal{B})$. Assume the number of parameters is d. Let $x^\intercal$ denote the transpose of a vector or matrix. Let $\phi : \mathcal{X} \to \mathbb{R}^d$ be an H-measurable function $\phi(x) = (\phi_1(x), \ldots, \phi_d(x))^\intercal$. Here $\phi(x)$ corresponds to the sufficient statistic for θ. Let Θ denote the parameter space in $\mathbb{R}^d$. Let $\{\mathcal{P}_\theta, \theta \in \Theta\}$ be a set of probability distributions with respect to the measure H. Then, $\{\mathcal{P}_\theta, \theta \in \Theta\}$ is said to be a multivariate exponential family with natural parameter θ, if the probability density function of each $f_\theta \in \mathcal{P}_\theta$ with respect to H can be expressed as $f_\theta(x) = \exp\{\theta^\intercal \phi(x) - \Phi(\theta)\}$. In the definition, the so-called log-partition function is given by

$$\Phi(\theta) := \log \int_{\mathcal{X}} \exp(\theta^\intercal \phi(x)) dH(x).$$

To make sure $f_\theta(x)$ a well-defined probability density, we consider the following two sets for parameters:

$$\Theta = \{\theta \in \mathbb{R}^d : \log \int_{\mathcal{X}} \exp(\theta^\intercal \phi(x)) dH(x) < +\infty\},$$

and

$$\Theta_\sigma = \{\theta \in \Theta : \nabla^2 \Phi(\theta) \succeq \sigma I_{d\times d}\}.$$

Note that $-\log f_\theta(x)$ is σ-strongly convex over Θ_σ. Its gradient corresponds to $\nabla\Phi(\theta) = \mathbb{E}_\theta[\phi(X)]$, and the Hessian $\nabla^2\Phi(\theta)$ corresponds to the covariance matrix of the vector $\phi(X)$. Therefore, $\nabla^2\Phi(\theta)$ is positive semidefinite and $\Phi(\theta)$ is convex. Moreover, Φ is a *Legendre function*, which means that it is strongly convex, continuous differentiable and essentially smooth [120]. The Legendre-Fenchel dual Φ^* is defined as

$$\Phi^*(z) = \sup_{u \in \Theta}\{u^\intercal z - \Phi(u)\}.$$

The mappings $\nabla\Phi^*$ is an inverse mapping of $\nabla\Phi$ [121]. Moreover, if Φ is a strongly convex function, then $\nabla\Phi^* = (\nabla\Phi)^{-1}$.

A general measure of proximity used in the OMD is the so-called *Bregman divergence* B_F, which is a nonnegative function induced by a Legendre function F (see, e.g., [120, 77]) defined as

$$B_F(u, v) := F(u) - F(v) - \langle \nabla F(v), u - v\rangle. \tag{6.9}$$

For exponential family, a natural choice of the Bregman divergence is the Kullback-Leibler (KL) divergence. Define $\mathbb{E}_\theta$ as the expectation when X is a random variable with density f_θ and $I(\theta_1, \theta_2)$ as the KL divergence between two distributions with densities f_{θ_1} and f_{θ_2} for any $\theta_1, \theta_2 \in \Theta$. Then

$$I(\theta_1, \theta_2) = \mathbb{E}_{\theta_1}\left[\log(f_{\theta_1}(X)/f_{\theta_2}(X))\right]. \tag{6.10}$$

It can be shown that, for exponential family, $I(\theta_1, \theta_2) = \Phi(\theta_2) - \Phi(\theta_1) - (\theta_2 - \theta_1)^\intercal \nabla\Phi(\theta_1)$. Using the definition (6.9), this means that B_Φ

$$B_\Phi(\theta_1, \theta_2) := I(\theta_2, \theta_1) \tag{6.11}$$

is a Bregman divergence. This property is useful to constructing mirror descent estimator for the exponential family [122, 121].

6.1.4 Online convex optimization (OCO) algorithms for non-anticipating estimators

Online convex optimization (OCO) algorithms [81] can be interpreted as a player who makes sequential decisions. At the time of each decision, the outcomes are unknown to the player.

After committing to a decision, the decision maker suffers a loss that can be adversarially chosen. An OCO algorithm makes decisions, which, based on the observed outcomes, minimizes the *regret* that is the difference between the total loss that has incurred relatively to that of the best fixed decision in hindsight. To design non-anticipating estimators, we consider OCO algorithms with likelihood-based regret functions. We iteratively estimate the parameters at the time when a one new observation becomes available based on *the maximum likelihood principle*, and hence the loss incurred corresponds to the negative log-likelihood of the new sample evaluated at the estimator $\ell_t(\theta) := -\log f_\theta(X_t)$, which corresponds to the log-loss in [80]. Given samples $X_1, \ldots, X_t$, the regret for a sequence of estimators $\{\hat{\theta}_i\}_{i=1}^t$ generated by a *likelihood-based OCO algorithm* a is defined as

$$\mathcal{R}_t^{\mathsf{a}} = \sum_{i=1}^{t} \{-\log f_{\hat{\theta}_{i-1}}(X_i)\} - \inf_{\tilde{\theta}\in\Theta} \sum_{i=1}^{t} \{-\log f_{\tilde{\theta}}(X_i)\}. \tag{6.12}$$

Below we omit the superscript a occasionally for notational simplicity.

In this paper, we consider a generic OCO procedure called the online mirror descent algorithms (OMD) [81, 123]. Next, we discuss how to construct the non-anticipating estimators $\{\hat{\theta}_t\}_{t\geq 1}$ in (6.1), and $\{\hat{\theta}_{k,t}\}, k = 1, 2, \ldots, t-1$ in (6.5) using OMD. The main idea of OMD is the following. At each time step, the estimator $\hat{\theta}_{t-1}$ is updated using the new sample X_t, by balancing the tendency to stay close to the previous estimate against the tendency to move in the direction of the greatest local decrease of the loss function. For the loss function defined above, a sequence of OMD estimator is constructed by

$$\hat{\theta}_t = \arg\min_{u\in\Gamma} [u^\intercal \nabla \ell_t(\hat{\theta}_{t-1}) + \frac{1}{\eta_i} B_\Phi(u, \hat{\theta}_{t-1})], \tag{6.13}$$

where B_Φ is defined in (6.11). Here $\Gamma \subset \Theta_\sigma$ is a closed convex set, which is problem-specific and encourages certain parameter structure such as sparsity.

Remark 16. *Similar to (6.13), for any fixed k, we can compute $\{\hat{\theta}_{k,t}\}_{t\geq 1}$ via OMD for sequential change-point detection. The only difference is that $\{\hat{\theta}_{k,t}\}_{t\geq 1}$ is computed if we* *use X_k as our first sample and then apply the recursive update (6.13) on $X_{k+1}, \ldots$. For $\hat{\theta}_t$, we use X_1 as our first sample.*

There is an equivalent form of OMD, presented as the original formulation in [122]. The equivalent form is sometimes easier to use for algorithm development, and it consists of four steps: (1) compute the dual variable: $\hat{\mu}_{t-1} = \nabla\Phi(\hat{\theta}_{t-1})$; (2) perform the dual update: $\hat{\mu}_t = \hat{\mu}_{t-1} - \eta_t \nabla \ell_t(\hat{\theta}_{t-1})$; (3) compute the primal variable: $\tilde{\theta}_t = (\nabla\Phi)^*(\hat{\mu}_t)$; (4) perform the projected primal update: $\hat{\theta}_t = \arg\min_{u \in \Gamma} B_\Phi(u, \tilde{\theta}_t)$. The equivalence between the above form for OMD and the nonlinear projected subgradient approach in (6.13) is proved in [121]. We adopt this approach when deriving our algorithm and follow the same strategy as [76]. Algorithm 2 summarizes the steps[1].

Algorithm 2 Online mirror-descent for non-anticipating estimators

Function Exponential family specifications $\phi(x), \Phi(x)$ and $f_\theta(x)$; initial parameter value θ_0; sequence of data $X_1, \ldots, X_t, \ldots$; a closed, convex set for parameter $\Gamma \subset \Theta_\sigma$; a decreasing sequence $\{\eta_t\}_{t \geq 1}$ of strictly positive step-sizes.

```
$\hat{\theta}_0 = \theta_0, \Lambda_0 = 1$. {Initialization}
for all $t = 1, 2, \ldots,$ do
    Acquire a new observation $X_t$
    Compute loss $\ell_t(\hat{\theta}_{t-1}) \triangleq -\log f_{\hat{\theta}_{t-1}}(X_t) = \Phi(\hat{\theta}_{t-1}) - \hat{\theta}_{t-1}^\intercal \phi(X_t)$
    Compute likelihood ratio $\Lambda_t = \Lambda_{t-1} \dot{f}_{\hat{\theta}_{t-1}}(X_t) / f_{\theta_0}(X_t)$
    $\hat{\mu}_{t-1} = \nabla\Phi(\hat{\theta}_{t-1}), \hat{\mu}_t = \hat{\mu}_{t-1} - \eta_t(\hat{\mu}_{t-1} - \phi(X_t))$ {Dual update}
    $\tilde{\theta}_t = (\nabla\Phi)^*(\hat{\mu}_t)$
    $\hat{\theta}_t = \arg\min_{u \in \Gamma} B_\Phi(u, \tilde{\theta}_t)$ {Projected primal update}
end for
return $\{\hat{\theta}_t\}_{t \geq 1}$ and $\{\Lambda_t\}_{t \geq 1}$.
```

For strongly convex loss function, the regret of many OCO algorithms, including the OMD, has the property that $\mathcal{R}_n \leq C \log n$ for some constant C (depend on f_θ and Θ_σ) and any positive integer n [124, 77]. Note that for exponential family, the loss function is the

[1] The implementation of the code can be downloaded at `http://www2.isye.gatech.edu/~yxie77/one-sample-update-code.zip`.

negative log-likelihood function, which is strongly convex over Θ_σ. Hence, we can have the logarithmic regret property.

6.2 Nearly second-order asymptotic optimality of one-sample update schemes

Below we prove the *nearly second-order asymptotic optimality* of the one-sample update schemes. More precisely, the nearly second-order asymptotic optimality means that the algorithm obtains the lower performance bound asymptotically up to a log-log factor in the false-alarm rate, as the false-alarm rate tends to zero (in many cases the log-log factor is a small number).

We first introduce some necessary notations. Denote $\mathbb{P}_{\theta,\nu}$ and $\mathbb{E}_{\theta,\nu}$ as the probability measure and the expectation when the change occurs at time ν and the post-change parameter is θ, i.e., when $X_1, \ldots, X_\nu$ are i.i.d. random variables with density f_{θ_0} and $X_{\nu+1}, X_{\nu+2}, \ldots$ are i.i.d. random variables with density f_θ. Moreover, let $\mathbb{P}_\infty$ and $\mathbb{E}_\infty$ denote the probability measure when there is no change, i.e., $X_1, X_2, \ldots$ are i.i.d. random variables with density f_{θ_0}. Finally, let $\mathcal{F}_t$ denote the σ-algebra generated by $X_1, \ldots, X_t$ for $t \geq 1$.

6.2.1 "One-sided" Sequential hypothesis test

Recall that the decision rule for sequential hypothesis test is a stopping time $\tau(b)$ defined in (6.3). The two standard performance metrics are the false-alarm rate, denoted as $\mathbb{P}_\infty(\tau(b) < \infty)$, and the expected detection delay (i.e., the expected number of samples needed to reject the null), denoted as $\mathbb{E}_{\theta,0}[\tau(b)]$. A meaningful test should have both small $\mathbb{P}_\infty(\tau(b) < \infty)$ and small $\mathbb{E}_{\theta,0}[\tau(b)]$. Usually, one adjusts the threshold b to control the false-alarm rate to be below a certain level.

Our main result is the following. As has been observed by [125], there is a loss in the statistical efficiency by using one-sample update estimators relative to the GLR approach using the entire samples $X_1, \ldots, X_t$ in the past. The theorem below shows that this loss corresponds to the expected regret given in (6.12).

Theorem 21 (Upper bound for OCO based SPRT)**.** *Let $\{\hat{\theta}_t\}_{t\geq 1}$ be a sequence of non-anticipating estimators generated by an OCO algorithm a. As $b \to \infty$,*

$$\mathbb{E}_{\theta,0}[\tau(b)] \leq \frac{b}{I(\theta,\theta_0)} + \frac{\mathbb{E}_{\theta,0}\left[\mathcal{R}^{\mathsf{a}}_{\tau(b)}\right]}{I(\theta,\theta_0)} + O(1) \tag{6.14}$$

Here $O(1)$ is a term upper-bounded by an absolute constant as $b \to \infty$.

The main idea of the proof is to decompose the statistic defining $\tau(b)$, $\log \Lambda(t)$, into a few terms that form martingales, and then invoke the Wald's Theorem for the stopped process.

Remark 17. *The inequality (6.14) is valid for a sequence of non-anticipating estimators generated by an OCO algorithm. Moreover, (6.14) gives an explicit connection between the expected detection delay for the one-sided sequential hypothesis testing (left-hand side of (6.14)) and the regret for the OCO (the second term on the right-hand side of (6.14)). This illustrates clearly the impact of estimation on detection by an estimation algorithm dependent factor.*

Note that in the statement of the Theorem 21, the stopping time $\tau(b)$ appears on the right-hand side of the inequality (6.14). For OMD, the expected sample size is usually small. By comparing with specific regret bound $\mathcal{R}_{\tau(b)}$, we can bound $\mathbb{E}_{\theta,0}[\tau(b)]$ as discussed in Section 6.3. The most important case is that when the estimation algorithm has a logarithmic expected regret. For the exponential family, as shown in section 6.2.3, Algorithm 2 can achieve $\mathbb{E}_{\theta,0}[\mathcal{R}_n] \leq C \log n$ for any positive integer n. To obtain a more specific order of the upper bound for $\mathbb{E}_{\theta,0}[\tau_b]$ when b grows, we establish an upper bound for $\mathbb{E}_{\theta,0}[\tau_b]$ as a function of b, to obtain the following Corollary 5.

Corollary 5. *Let $\{\hat{\theta}_t\}_{t\geq 1}$ be a sequence of non-anticipating estimators generated by an OCO algorithm a. Assume that $\mathbb{E}_{\theta,0}[\mathcal{R}^{\mathsf{a}}_n] \leq C \log n$ for any positive integer n and some* *constant $C > 0$, we have*

$$\mathbb{E}_{\theta,0}[\tau(b)] \leq \frac{b}{I(\theta,\theta_0)} + \frac{C\log b}{I(\theta,\theta_0)}(1+o(1)). \tag{6.15}$$

Here $o(1)$ is a vanishing term as $b \to \infty$.

Corollary 5 shows that other than the well known first-order approximation $b/I(\theta,\theta_0)$ [59, 75], the expected detection delay $\mathbb{E}_{\theta,0}[\tau(b)]$ is bounded by an additional term that is on the order of $\log(b)$ if the estimation algorithm has a logarithmic regret. This $\log b$ term plays an important role in establishing the optimality properties later. To show the optimality properties for the detection procedures, we first select a set of detection procedures with false-alarm rates lower than a prescribed value, and then prove that among all the procedures in the set, the expected detection delays of our proposed procedures are the smallest. Thus, we can choose a threshold b to uniformly control the false-alarm rate of $\tau(b)$.

Lemma 11 (false-alarm rate of $\tau(b)$)**.** *Let $\{\hat{\theta}_t\}_{t\geq 1}$ be any sequence of non-anticipating estimators. For any $b > 0$, $\mathbb{P}_\infty(\tau(b) < \infty) \leq \exp(-b)$.*

Lemma 11 shows that as b increases the false-alarm rate of $\tau(b)$ decays exponentially fast. We can set $b = \log(1/\alpha)$ to make the false-alarm rate of $\tau(b)$ less than some $\alpha > 0$. Next, leveraging an existing lower bound for general SPRT presented in Section 5.5.1.1 in [55], we establish the nearly second-order asymptotic optimality of OMD based SPRT as follows:

Corollary 6 (Nearly second-order optimality of OCO based SPRT)**.** *Let $\{\hat{\theta}_t\}_{t\geq 1}$ be a sequence of non-anticipating estimators generated by an OCO algorithm a. Assume that $\mathbb{E}_{\theta,0}[\mathcal{R}_n^{\mathsf{a}}] \leq C\log n$ for any positive integer n and some constant $C > 0$. Define a set $C(\alpha) = \{T : \mathbb{P}_\infty(T < \infty) \leq \alpha\}$. For $b = \log(1/\alpha)$, due to Lemma 11, $\tau(b) \in C(\alpha)$. For such a choice, $\tau(b)$ is nearly second-order asymptotic optimal in the sense that for any*

$\theta \in \Theta_\sigma - \{\theta_0\}$, as $\alpha \to 0$,

$$\mathbb{E}_{\theta,0}[\tau(b)] - \inf_{T \in C(\alpha)} \mathbb{E}_{\theta,0}[T] = O(\log(\log(1/\alpha))). \tag{6.16}$$

The result means that, compared with any procedure (including the optimal procedure) calibrated to have a false-alarm rate less than α, our procedure incurs an at most $\log(\log(1/\alpha))$ increase in the expected detection delay, which is usually a small number. For instance, even for a conservative case when we set $\alpha = 10^{-5}$ to control the false-alarm rate, the number is $\log(\log(1/\alpha)) = 2.44$.

6.2.2 Sequential change-point detection

Now we proceed the proof by leveraging the close connection [59] between the sequential change-point detection and the one-sided hypothesis test. For sequential change-point detection, the two commonly used performance metrics [55] are the average run length (ARL), denoted by $\mathbb{E}_\infty[T]$; and the maximal conditional average delay to detection (CADD), denoted by $\sup_{\nu \geq 0} \mathbb{E}_{\theta,\nu}[T - \nu \mid T > \nu]$. ARL is the expected number of samples between two successive false alarms, and CADD is the expected number of samples needed to detect the change after it occurs. A good procedure should have a large ARL and a small CADD. Similar to the one-sided hypothesis test, one usually choose the threshold large enough so that ARL is larger than a pre-specified level.

Similar to Theorem 21, we provide an upper bound for the CADD of our ASR and ACM procedures.

Theorem 22. *Consider the change-point detection procedure $T_{\mathrm{ACM}}(b_1)$ in (6.7) and $T_{\mathrm{ASR}}(b_2)$ in (6.8). For any fixed k, let $\{\hat{\theta}_{k,t}\}_{t \geq 1}$ be a sequence of non-anticipating estimators generated* *by an OCO algorithm* a*. Let* $b_1 = b_2 = b$*, as* $b \to \infty$ *we have that*

$$\begin{aligned}
&\sup_{\nu\geq 0} \mathbb{E}_{\theta,\nu}[T_{\mathrm{ASR}}(b) - \nu \mid T_{\mathrm{ASR}}(b) > \nu] \leq \sup_{\nu\geq 0} \mathbb{E}_{\theta,\nu}[T_{\mathrm{ACM}}(b) - \nu \mid T_{\mathrm{ACM}}(b) > \nu] \\
\leq\ & (I(\theta,\theta_0))^{-1} \left(b + \mathbb{E}_{\theta,0}\left[\mathcal{R}^{\mathsf{a}}_{\tau(b)}\right] + O(1) \right).
\end{aligned} \tag{6.17}$$

To prove Theorem 22, we relate the ASR and ACM procedures to the one-sided hypothesis test and use the fact that when the measure $\mathbb{P}_\infty$ is known, $\sup_{\nu\geq 0} \mathbb{E}_{\theta,\nu}[T - \nu \mid T > \nu]$ is attained at $\nu = 0$ for both the ASR and the ACM procedures. Above, we may apply a similar argument as in Corollary 5 to remove the dependence on $\tau(b)$ on the right-hand-side of the inequality. We establish the following lower bound for the ARL of the detection procedures, which is needed for proving Corollary 7:

Lemma 12 (ARL)**.** *Consider the change-point detection procedure* $T_{\mathrm{ACM}}(b_1)$ *in (6.7) and* $T_{\mathrm{ASR}}(b_2)$ *in (6.8). For any fixed* k*, let* $\{\hat{\theta}_{k,t}\}_{t\geq 1}$ *be any sequence of non-anticipating estimators. Let* $b_1 = b_2 = b$*, given a prescribed lower bound* $\gamma > 0$ *for the ARL, we have*

$$\mathbb{E}_\infty[T_{\mathrm{ACM}}(b)] \geq \mathbb{E}_\infty[T_{\mathrm{ASR}}(b)] \geq \gamma,$$

provided that $b \geq \log\gamma$*.*

Lemma 12 shows that given a required lower bound γ for ARL, we can choose $b = \log\gamma$ to make the ARL be greater than γ. This is consistent with earlier works [126, 75] which show that the smallest threshold b such that $\mathbb{E}_\infty[T_{ACM}(b)] \geq \gamma$ is approximate $\log\gamma$. However, the bound in Lamma 12 is not tight, since in practice we can set $b = \rho\log\gamma$ for some $\rho \in (0,1)$ to ensure that ARL is greater than γ.

Combing the upper bound in Theorem 22 with an existing lower bound for the CADD of SRRS procedure in [82], we obtain the following optimality properties.

Corollary 7 (Nearly second-order asymptotic optimality of ACM and ASR)**.** *Consider the change-point detection procedure* $T_{\mathrm{ACM}}(b_1)$ *in (6.7) and* $T_{\mathrm{ASR}}(b_2)$ *in (6.8). For any fixed* k*,* *let* $\{\hat{\theta}_{k,t}\}_{t\geq 1}$ *be a sequence of non-anticipating estimators generated by an OCO algorithm* a*. Assume that* $\mathbb{E}_{\theta,0}[\mathcal{R}_n^{\mathsf{a}}] \leq C \log n$ *for any positive integer* n *and some constant* $C > 0$*. Let* $b_1 = b_2 = b$*. Define* $S(\gamma) = \{T : \mathbb{E}_\infty[T] \geq \gamma\}$*. For* $b = \log \gamma$*, due to Lemma 12, both* $T_{\mathrm{ASR}}(b)$ *and* $T_{\mathrm{ACM}}(b)$ *belong to* $S(\gamma)$*. For such* b*, both* $T_{\mathrm{ASR}}(b)$ *and* $T_{\mathrm{ACM}}(b)$ *are nearly second-order asymptotic optimal in the sense that for any* $\theta \in \Theta - \{\theta_0\}$

$$\begin{aligned}&\sup_{\nu\geq 1} \mathbb{E}_{\theta,\nu}[T_{\mathrm{ASR}}(b) - \nu + 1 \mid T_{\mathrm{ASR}}(b) \geq \nu] \\ &\quad - \inf_{T(b)\in S(\gamma)} \sup_{\nu\geq 1} \mathbb{E}_{\theta,\nu}[T(b) - \nu + 1 \mid T(b) \geq \nu] = O(\log\log\gamma).\end{aligned} \tag{6.18}$$

A similar expression holds for $T_{\mathrm{ACM}}(b)$*.*

The result means that, compared with any procedure (including the optimal procedure) calibrated to have a fixed ARL larger than γ, our procedure incurs an at most $\log(\log\gamma)$ increase in the CADD. Comparing (6.18) with (6.16), we note that the ARL γ plays the same role as $1/\alpha$ because $1/\gamma$ is roughly the false-alarm rate for sequential change-point detection [59].

6.2.3 Example: Regret bound for specific cases

In this subsection, we show that the regret bound $\mathcal{R}_t$ can be expressed as a weighted sum of Bregman divergences between two consecutive estimators. This form of $\mathcal{R}_t$ is useful to show the logarithmic regret for OMD. The following result comes as a modification of [83].

Theorem 23. *Assume that* $X_1, X_2, \ldots$ *are i.i.d. random variables with density function* $f_\theta(x)$*. Let* $\eta_i = 1/i$ *in Algorithm 2. Assume that* $\{\hat{\theta}_i\}_{i\geq 1}, \{\hat{\mu}_i\}_{i\geq 1}$ *are obtained using Algorithm 2 and* $\hat{\theta}_i = \tilde{\theta}_i$ *(defined in step 7 and 8 of Algorithm 2) for any* $i \geq 1$*. Then for any* $\theta_0 \in \Theta$ *and* $t \geq 1$*,*

$$\mathcal{R}_t = \sum_{i=1}^t i \cdot B_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) = \frac{1}{2}\sum_{i=1}^t i \cdot (\hat{\mu}_i - \hat{\mu}_{i-1})^\intercal [\nabla^2 \Phi^*(\tilde{\mu}_i)](\hat{\mu}_i - \hat{\mu}_{i-1}),$$

where $\tilde{\mu}_i = \lambda\hat{\mu}_i + (1-\lambda)\hat{\mu}_{i-1}$*, for some* $\lambda \in (0,1)$.

Next, we use Theorem 23 on a concrete example. The multivariate normal distribution, denoted by $\mathcal{N}(\theta, I_d)$, is parametrized by an unknown mean parameter θ and a known covariance matrix I_d (I_d is a $d \times d$ identity matrix). Following the notations in subsection 6.1.3, we know that $\phi(x) = x$, $dH(x) = (1/\sqrt{|2\pi I_d|}) \cdot \exp(-x^\intercal x/2)$, $\Theta = \Theta_\sigma = \mathbb{R}^d$ for any $\sigma < 2$, $\Phi(\theta) = (1/2)\theta^\intercal\theta$, $\mu = \theta$ and $\Phi^*(\mu) = (1/2)\mu^\intercal\mu$, where $|\cdot|$ denotes the determinant of a matrix, and H is a probability measure under which the sample follows $\mathcal{N}(0, I_d)$). When the covariance matrix is known to be some $\Sigma \neq I_d$, one can "whiten" the vectors by multiplying $\Sigma^{-1/2}$ to obtain the situation here.

Corollary 8 (Upper bound for the expected regret, Gaussian)**.** *Assume* $X_1, X_2, \ldots$ *are i.i.d. following* $\mathcal{N}(\theta, I_d)$ *with some* $\theta \in \mathbb{R}^d$*. Assume that* $\{\hat{\theta}_i\}_{i\geq 1}, \{\hat{\mu}_i\}_{i\geq 1}$ *are obtained using Algorithm 2 with* $\eta_i = 1/i$ *and* $\Gamma = \mathbb{R}^d$*. For any* $t > 0$*, we have that for some constant* $C_1 > 0$ *that depends on* θ*,*

$$\mathbb{E}_{\theta,0}[\mathcal{R}_t] \leq C_1 d \log t/2.$$

The following calculations justify Corollary 8, which also serve as an example of how to use regret bound. First, the assumption $\hat{\theta}_t = \tilde{\theta}_t$ in Theorem 23 is satisfied for the following reasons. Consider $\Gamma = \mathbb{R}^d$ is the full space. According to Algorithm 2, using the non-negativity of the Bregman divergence, we have $\hat{\theta}_t = \arg\min_{u\in\Gamma} B_\Phi(u, \tilde{\theta}_t) = \tilde{\theta}_t$. Then the regret bound can be written as

$$\begin{aligned}\mathcal{R}_t =& \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^\intercal(\hat{\mu}_1 - \hat{\mu}_0) + \frac{1}{2}\sum_{i=2}^{t}[i \cdot (\hat{\mu}_i - \hat{\mu}_{i-1})^\intercal(\hat{\mu}_i - \hat{\mu}_{i-1})] \\ =& \frac{1}{2}(X_1 - \theta_0)^\intercal(X_1 - \theta_0) + \frac{1}{2}\sum_{i=2}^{t}(\hat{\mu}_i - \hat{\mu}_{i-1})^\intercal(\phi(X_i) - \hat{\mu}_{i-1}).\end{aligned}$$

Since the step-size $\eta_i = 1/i$, the second term in the above equation can be written as:

$$\begin{aligned}
&\frac{1}{2}\sum_{i=2}^{t}(\hat{\mu}_i - \hat{\mu}_{i-1})^{\intercal}(\phi(X_i) - \hat{\mu}_{i-1}) \\
=&\frac{1}{2}\sum_{i=2}^{t}(\hat{\mu}_i - \hat{\mu}_{i-1})^{\intercal}(\phi(X_i) + \hat{\mu}_i) - \sum_{i=2}^{t}\frac{1}{2}(\hat{\mu}_i - \hat{\mu}_{i-1})^{\intercal}(\hat{\mu}_{i-1} + \hat{\mu}_i) \\
=&\sum_{i=2}^{t}\frac{1}{2(i-1)}(\phi(X_i) - \hat{\mu}_i)^{\intercal}(\phi(X_i) + \hat{\mu}_i) + \sum_{i=2}^{t}\frac{1}{2}(\|\hat{\mu}_{i-1}\|^2 - \|\hat{\mu}_i\|^2) \\
=&\sum_{i=2}^{t}\frac{1}{2(i-1)}\left\|X_i\right\|^2 - \sum_{i=2}^{t}\frac{1}{2(i-1)}\left\|\hat{\mu}_i\right\|^2 + \frac{1}{2}\left\|\hat{\mu}_1\right\|^2 - \frac{1}{2}\left\|\hat{\mu}_t\right\|^2.
\end{aligned}$$

Combining above, we have

$$\mathbb{E}_{\theta,0}[\mathcal{R}_t] \leq \frac{1}{2}\mathbb{E}_{\theta,0}[(X_1 - \theta_0)^{\intercal}(X_1 - \theta_0)] + \frac{1}{2}\sum_{i=2}^{t}\frac{1}{i-1}\mathbb{E}_{\theta,0}[\|X_i\|^2] + \frac{1}{2}\mathbb{E}_{\theta,0}[\|X_1\|^2].$$

Finally, since $\mathbb{E}_{\theta,0}[\|X_i\|^2] = d(1+\theta^2)$ for any $i \geq 1$, we obtain desired result. Thus, with i.i.d. multivariate normal samples, the expected regret grows logarithmically with the number of samples.

Using the similar calculations, we can also bound the expected regret in the general case. As shown in the proof above for Corollary 8, the dominating term for $\mathcal{R}_t$ can be rewritten as

$$\sum_{i=2}^{t}\frac{1}{2(i-1)}(\phi(X_i) - \hat{\mu}_i)^{\intercal}[\nabla^2\Phi^*(\tilde{\mu}_i)](\phi(X_i) + \hat{\mu}_i),$$

where $\tilde{\mu}_i$ is a convex combination of $\hat{\mu}_{i-1}$ and $\hat{\mu}_i$. For an arbitrary distribution, the term $(\phi(X_i) - \hat{\mu}_i)^{\intercal}[\nabla^2\Phi^*(\tilde{\mu}_i)](\phi(X_i) + \hat{\mu}_i)$ can be viewed as a local normal distribution with the changing curvature $\nabla^2\Phi^*(\tilde{\mu}_i)$. Thus, it is possible to prove case-by-case the $O(\log t)$-style bounds by making more assumptions about the distributions. Recall the notation Θ_σ in subsection 6.1.3 such that $-\log f_\theta(x)$ is σ-strongly convex over Θ_σ. Let $\|\cdot\|_2$ denote the ℓ_2 norm. Moreover, we assume that the true parameter belongs to a set Γ that is a closed and convex subset of Θ_σ such that $\sup_{\theta\in\Gamma}\|\nabla\Phi(\theta)\|_2 \leq M$ for some constant M. Thus, one can show that $-\log f_\theta(x)$ is not only σ-strongly convex but also M-strongly smooth over Γ. Theorem 3 in [77] shows that for all $\theta \in \Gamma$ and $n \geq 1$, consider that $\{\hat{\theta}_i\}_{i\geq 1}$ is obtained by OMD, then

$$\mathbb{E}_{\theta,0}[\mathcal{R}_n] \leq \frac{\mathbb{E}_{\theta,0}\left[\left(\frac{1}{2}\max_{1\leq i\leq n}\|X_i\|_2 + \frac{1}{2}M\right)^2\right]}{\sigma} \cdot (\log n + 1).$$

Therefore, for any bounded distributions within the exponential family, we achieve a logarithmic regret. This logarithmic regret is valid for Bernoulli distribution, Beta distribution and some truncated versions of classic distributions (e.g., truncated Gaussian distribution, truncated Gamma distribution and truncated Geometric distribution analyzed in [127]).

6.3 Numerical examples

In this section, we present some synthetic examples to demonstrate the good performance of our methods. We will focus on ACM and ASR for sequential change-point detection. In the following, we consider the window-limited versions (see Remark 15) of ACM and ASR with window size $w = 100$. Recall that when the measure $\mathbb{P}_\infty$ is known, $\sup_{\nu\geq 0}\mathbb{E}_{\theta,\nu}[T-\nu \mid T>\nu]$ is attained at $\nu = 0$ for both ASR and ACM procedures (a proof can be found in the proof of Theorem 22). Therefore, in the following experiments we define the expected detection delay (EDD) as $\mathbb{E}_{\theta,0}[T]$ for a stopping time T. To compare the performance between different detection procedures, we determine the threshold for each detection procedure by Monte-Carlo simulations such that the ARL for each procedure is about 10000. Below, we denote $\|\cdot\|_2$, $\|\cdot\|_1$ and $\|\cdot\|_0$ as the ℓ_2 norm, ℓ_1 norm and ℓ_0 norm defined as the number of non-zero entries, respectively. The following experiments are all run on the same Macbook Air with an Intel i7 Core CPU.

6.3.1 Detecting sparse mean-shift of multivariate normal distribution

We consider detect the sparse mean shift for multivariate normal distribution. Specifically, we assume that the pre-change distribution is $\mathcal{N}(0, I_d)$ and the post-change distribution is $\mathcal{N}(\theta, I_d)$ for some unknown $\theta \in \{\theta \in \mathbb{R}^d : \|\theta\|_0 \leq s\}$, where s is called the *sparsity* of the mean shift. Sparse mean shift detection is of particular interest in sensor networks [18, 109]. For this Gaussian case, the Bregman divergence is given by $B_\Phi(\theta_1, \theta_2) = I(\theta_2, \theta_1) = \|\theta_1 - \theta_2\|_2^2/2$. Therefore, the projection onto Γ in Algorithm 2 is a Euclidean projection onto a convex set, which in many cases can be implemented efficiently. As a frequently used convex relaxation of the ℓ_0-norm ball, we set $\Gamma = \{\theta : \|\theta\|_1 \leq s\}$ (it is known that imposing an ℓ_1 constraint leads to sparse solution; see, e.g., [48]). Then, the projection onto ℓ_1 ball can be computed very efficiently via a simple soft-thresholding technique [102].

Two benchmark procedures are the CUSUM and the GLR. For the CUSUM procedure, we specify a nominal post-change mean, which is an all-one vector. If knowing the post-change mean is sparse, we can also use the shrinkage estimator presented in [128], which performs hard or soft thresholding of the estimated post-change mean parameter. Our procedures are $T_{ASR}(b)$ and $T_{ACM}(b)$ with $\Gamma = \mathbb{R}^d$ and $\Gamma = \{\theta : \|\theta\|_1 \leq 5\}$. In the following experiments, we run 10000 Monte Carlo trials to obtain each simulated EDD.

In the experiments, we set $d = 20$. The post-change distributions are $\mathcal{N}(\theta, I_d)$, where $100p\%$ entry of θ is 1 and others are 0, and the location of nonzero entries are random. Table 6.1 shows the EDDs versus the proportion p. Note that our procedures incur little performance loss compared with the GLR procedure and the CUSUM procedure. Notably, $T_{ACM}(b)$ with $\Gamma = \{\theta : \|\theta\|_1 \leq 5\}$ performs almost the same as the GLR procedure and much better than the CUSUM procedure when p is small. This shows the advantage of projection when the true parameter is sparse.

	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$	$p = 0.6$
CUSUM	188.60	146.45	64.30	18.97	7.18	3.77
Shrinkage	17.19	9.25	6.38	4.96	4.07	3.55
GLR	19.10	10.09	7.00	5.49	4.50	3.86
ASR	45.22	19.55	12.62	8.90	7.02	5.90
ACM	45.60	19.93	12.50	9.00	7.03	5.87
ASR-$\ell 1$	45.81	19.94	12.45	8.92	6.97	5.89
ACM-$\ell 1$	19.24	10.17	7.51	6.11	5.41	4.92

Table 6.1: Comparison of the EDDs in detecting the sparse mean shift of multivariate Gaussian distribution. Below, "CUSUM": CUSUM procedure with pre-specified all-one vector as post-change parameter; "Shrinkage": component-wise shrinkage estimator in [128]; "GLR": GLR procedure; "ASR": $T_{\rm ASR}(b)$ with $\Gamma = \mathbb{R}^d$; "ACM": $T_{\rm ACM}(b)$ with $\Gamma = \mathbb{R}^d$; "ASR-L1": $T_{\rm ASR}(b)$ with $\Gamma = \{\theta : \|\theta\|_1 \leq 5\}$; "ACM-L1": $T_{ACM}(b)$ with $\Gamma = \{\theta : \|\theta\|_1 \leq 5\}$. p is the proportion of non-zero entries in θ. We run 10000 Monte Carlo trials to obtain each value. For each value, the standard deviation is less than one half of the value.

6.3.2 Detecting the scale change in Gamma distribution

We consider an example that detects the scale change in Gamma distributions. Assume that we observe a sequence $X_1, X_2 \ldots$ of samples drawn from Gamma(α, β) for some $\alpha, \beta > 0$, with the probability density function given by $f_{\alpha,\beta}(x) = \exp(-x\beta)x^{\alpha-1}\beta^{\alpha}/\tilde{\Gamma}(\alpha)$ (to avoid confusion with the Γ parameter in Algorithm 2 we use $\tilde{\Gamma}(\cdot)$ to denote the Gamma function). The parameter α^{-1} is called the dispersion parameter that scales the loss and the divergences. For simplicity, we fix $\alpha = 1$ just like we fix the variance in the Gaussian case. The specifications in the Algorthm 2 are as follows: $\theta = -\beta$, $\Theta = (-\infty, 0)$, $\phi(x) = x$, $dH(x) = 1$, $\Phi(\theta) = -\log(-\theta)$, $\mu = -1/\theta$ and $\Phi^*(\mu) = -1 - \log \mu$. Assume that the pre-change distribution is Gamma$(1, 1)$ and the post-change distribution is Gamma$(1, \beta)$ for some unknown $\beta > 0$. We compare our algorithms with CUSUM, GLR and non-ancitipating estimator based on the method of moment (MOM) estimator in [75]. For the CUSUM procedure, we specify the post-change β to be 2. The results are shown in Table 6.2. CUSUM fails to detect the change when $\beta = 0.1$, which is far away from the pre-specified post-change parameter $\beta = 2$. We can see that performance loss of the proposed ACM method compared with GLR and MOM is very small.

	$\beta = 0.1$	$\beta = 0.5$	$\beta = 2$	$\beta = 5$	$\beta = 10$
CUSUM	NaN	481.2	33.75	14.37	12.04
MOM	3.41	32.87	40.86	11.42	7.21
GLR	2.40	23.79	33.29	9.07	5.67
ASR	3.95	32.34	45.18	13.45	8.55
ACM	3.70	31.80	47.20	12.42	7.87

Table 6.2: Comparison of the EDDs in detecting the scale change in Gamma distribution. Below, “CUSUM”: CUSUM procedure with pre-specified post-change parameter $\beta = 2$; “MOM”: Method of Moments estimator method; “GLR”: GLR procedure; “ASR”: $T_{\text{ASR}}(b)$ with $\Gamma = (-\infty, 0)$; “ACM”: $T_{\text{ACM}}(b)$ with $\Gamma = (-\infty, 0)$. We run 10000 Monte Carlo trials to obtain each value. For each value, the standard deviation is less than one half of the value.

6.3.3 Communication-rate change detection with Erdos-Renyi model

Next, we consider a problem to detect the communication-rate change in a network, which is a model for social network data. Suppose we observe communication between nodes in a network over time, represented as a sequence of (symmetric) adjacency matrices of the network. At time t, if node i and node j communicates, then the adjacency matrix has 1 on the ijth and jith entries (thus it forms an undirected graph). The nodes that do not communicate have 0 on the corresponding entries. We model such communication patterns using the Erdos-Renyi random graph model. Each edge has a fixed probability of being present or absent, independently of the other edges. Under the null hypothesis, each edge is a Bernoulli random variable that takes values 1 with known probability p and value 0 with probability $1 - p$. Under the alternative hypothesis, there exists an unknown time κ, after which a small subset of edges occur with an unknown and different probability $p' \neq p$.

In the experiments, we set $N = 20$ and $d = 190$. For the pre-change parameters, we set $p_i = 0.2$ for all $i = 1, \ldots, d$. For the post-change parameters, we randomly select n out of the 190 edges, denoted by $\mathcal{E}$, and set $p_i = 0.8$ for $i \in \mathcal{E}$ and $p_i = 0.2$ for $i \notin \mathcal{E}$. As said before, let the change happen at time $\nu = 0$ (since the upper bound for EDD is achieved at $\nu = 0$ as argued in the proof of Theorem 22). To implement CUSUM, we specify the post-change parameters $p_i = 0.8$ for all $i = 1, \ldots, d$.

The results are shown in Table 6.3. Our procedures are better than CUSUM procedure when n is small since the post-change parameters used in CUSUM procedure is far from the true parameter. Compared with GLR procedure, our methods have a small performance loss, and the loss is almost negligible as n approaches to $d = 190$.

	$n = 78$	$n = 100$	$n = 120$	$n = 150$	$n = 170$	$n = 190$
CUSUM	473.11	2.06	2.00	2.00	2.00	2.00
GLR	2.00	1.96	1.27	1.00	1.00	1.00
ASR	8.64	6.39	5.08	3.92	3.36	2.94
ACM	8.67	6.37	5.07	3.88	3.32	2.94

Table 6.3: Comparison of the EDDs in detecting the changes of the communication-rates in a network. Below, “CUSUM”: CUSUM procedure with pre-specified post-change parameters $p = 0.8$; “GLR”: GLR procedure; “ASR”: $T_{\text{ASR}}(b)$ with $\Gamma = \mathbb{R}$; “ACM”: $T_{\text{ACM}}(b)$ with $\Gamma = \mathbb{R}$. We run 10000 Monte Carlo trials to obtain each value. For each value, the standard deviation is less than one half of the value.

Below are the specifications of Algorithm 2 in this case. For Bernoulli distribution with unknown parameter p, the natural parameter θ is equal to $\log(p/(1-p))$. Thus, we have $\Theta = \mathbb{R}$, $\phi(x) = x$, $dH(x) = 1$, $\Phi(\theta) = \log(1 + \exp(\theta))$, $\mu = \exp(\theta)/(1 + \exp(\theta))$ and $\Phi^*(\mu) = \mu \log \mu + (1-\mu)\log(1-\mu)$.

6.3.4 Point process change-point detection: Poisson to Hawkes processes

In this example, to illustrate the situation in Section 1.3.2, we consider a case where a homogeneous Poisson process switches to a Hawkes process (see, e.g., [79]); this can be viewed as a simplest case in Section 1.3.2 with one node. We construct ACM and ASR procedures. In this case, the MLE for the unknown post-change parameter cannot be found in close-form, yet ACM and ASR can be easily constructed and give reasonably good performance, although our theory no longer holds in this case due to the lack of i.i.d. samples.

The Hawkes process can be viewed as a non-homogeneous Poisson process where the intensity is influenced by historical events. The data consist of a sequence of events occurring at times $\{t_1, t_2, \ldots, t_n\}$ before a time horizon T: $t_i \leq T$. Assume the intensity of the Poisson process is λ_s, $s \in (0, T)$ and there may exists a change-point $\kappa \in (0, T)$ such that the process changes. The null and alternative hypothesis tests are

$$\begin{cases} \mathsf{H}_0: & \lambda_s = \mu, \quad 0 < s < T; \\ \mathsf{H}_1: & \lambda_s = \mu, \quad 0 < s < \kappa, \\ & \lambda_s = \mu + \theta \sum_{\kappa < t_j < s} \varphi(s - t_j), \quad \kappa < s < T, \end{cases}$$

where μ is a known baseline intensity, $\theta > 0$ is unknown magnitude of the change, $\varphi(s) = \beta e^{-\beta s}$ is the normalized kernel function with pre-specified parameter $\beta > 0$, which captures the influence from the past events. We treat the post-change influence parameter θ as unknown since it represents an anomaly.

We first use a sliding window to convert the event times into a sequence of vectors with overlapping events. Assume of size of the sliding window is L. For a given scanning time $T_i \leq T$, we map all the events in $[T_i - L, T_i]$ to a vector $X_i = [t_{(1)}, \ldots, t_{(m_i)}]^\mathsf{T}$, $t_{(i)} \in [T_i - L, T_i]$, where m_i is the number of events falling into the window. Note that X_i can have different length for different i. Consider a set of scanning times $T_1, T_2, \ldots, T_t$. This maps the event times into a sequence of vectors $X_1, X_2, \ldots, X_t$ of lengthes m_1, m_2, ..., m_t. These scanning times can be arbitrary; here we set them to be event times so that there are at least one sample per sliding window.

For a hypothetical change-point location k, it can be shown that the log-likelihood ratio (between the Hawkes process and the Poisson process) as a function of θ, is given by

$$\ell(\theta|X_i) = \sum_{t_q \in (T_i - L, T_i)} \log \left[\mu + \theta \sum_{t_j \in (T_i - L, t_q)} \beta e^{-\beta(t_q - t_j)} \right] - \mu L - \theta \sum_{t_q \in (T_i - L, T_i)} \left[1 - e^{-\beta(T_i - t_q)} \right]. \tag{6.19}$$

Now based on this sliding window approach, we can approximate the original change-point detection problem as the following. Without change, $X_1, \ldots, X_t$ are sampled from a Poisson process. Under the alternative, the change occurs at some time such that $X_1, \ldots, X_\kappa$ are sampled from a Poisson process, and $X_{\kappa+1}, \ldots, X_t$ are sampled from a Hawkes process with parameter θ, rather than a Poisson process. We define the estimator of θ, for assumed change-point location $\kappa = k$ as follows

$$\hat{\theta}_{k,i} \triangleq \hat{\theta}_{k,i}(X_k, \ldots, X_i) = \hat{\theta}_{k,i}(t_\ell \in [T_k, T_i]) \tag{6.20}$$

Now, consider $k \in [i-w, i-1]$, and keep w estimators: $\hat{\theta}_{i-w,i}, \ldots, \hat{\theta}_{i-1,i}$. The update for each estimator is based on stochastic gradient descent. By taking derivative with respect to θ, we have

$$\frac{\partial \ell(\theta|X_i)}{\partial \alpha} = \sum_{t_q \in (T_i - L, T_i)} \frac{\sum_{t_j \in (T_i-L, t_q)} \beta e^{-\beta(t_q - t_j)}}{\mu + \theta \sum_{t_j \in (T_i - L, t_q)} \beta e^{-\beta(t_q - t_j)}} - \sum_{t_q \in (T_i - L, T_i)} \left[1 - e^{-\beta(T_i - t_q)}\right],$$

Note that there is no close form expression for the MLE, which the solution to the above equation. We perform stochastic gradient descent instead

$$\hat{\theta}_{k,i+1} = \hat{\theta}_{k,i} - \gamma \frac{\partial \ell(\theta|X_{i+1})}{\partial \theta}\Big|_{\theta = \hat{\theta}_{k,i}}, \quad k = i - w + 1, i - w, \ldots, i,$$

where $\gamma > 0$ is the step-size. Now we can apply the ACM and ASR procedures, by using the fact that $f_{\hat{\theta}_{k,t}}(X_{t+1})/f_{\theta_0}(X_{t+1}) = \ell(\hat{\theta}_{k,t}|X_{t+1})$ and calculating using (6.19).

Table. 6.4 shows the EDD for different α. Here we choose the threshold such that ARL is 5000. We see that the scheme has a reasonably good performance, the detection delay decreases as the true signal strength θ increases.

	$\theta = 0.4$	$\theta = 0.5$	$\theta = 0.5$	$\theta = 0.7$
ACM	33.03	27.75	20.39	16.16
ASR	38.59	24.96	20.17	13.91

Table 6.4: Point process change-point detection: EDD of ACM and ASR procedures for various values of true θ; ARL of the procedure is controlled to be 5000 by selecting threshold via Monte Carlo simulation.

6.4 Conclusion

In this chapter, we consider sequential hypothesis testing and change-point detection with computationally efficient one-sample update schemes obtained from online mirror descent. We show that the loss of the statistical efficiency caused by the online mirror descent estimator (replacing the exact maximum likelihood estimator using the complete historical data) is related to the regret incurred by the online convex optimization procedure. The result can be generalized to any estimation method with logarithmic regret bound. This result sheds lights on the relationship between the statistical detection procedures and the online convex optimization.

Appendices

APPENDIX A
PROOFS

A.1 Proofs for matrix completion

To begin, we first recall some definitions from introduction and explain some additional notation that we will need for the proofs. For two probability distributions $\mathcal{P}$ and $\mathcal{Q}$ on a countable set A, $D(\mathcal{P}||\mathcal{Q})$ will denote the Kullback-Leibler (KL) divergence

$$D(\mathcal{P}||\mathcal{Q}) = \sum_{x \in A} \mathcal{P}(x) \log\left(\frac{\mathcal{P}(x)}{\mathcal{Q}(x)}\right),$$

where $\mathcal{P}(x)$ denotes the probability of the outcome x under the distribution $\mathcal{P}$. In the following, we will abuse this notation slightly, to mean the KL divergence between two Poisson distributions with different parameters (the arguments in the notations denote parameters of the Poisson distributions), in the following two ways. First, for scalar inputs $p, q \in \mathbb{R}_+$, we will set $D(p\|q) \triangleq p\log(p/q) - (p-q)$, which gives the KL divergence between two Poisson probability distributions. Second, we allow the KL divergence to act on matrices via the average KL divergence over their entries: for two matrices P, $Q \in \mathbb{R}_+^{d_1 \times d_2}$, we define

$$D(P\|Q) \triangleq \frac{1}{d_1 d_2} \sum_{i,j} D(P_{ij}\|Q_{ij}).$$

For two probability distributions $\mathcal{P}$ and $\mathcal{Q}$ on a countable set A, $d_H^2(\mathcal{P}, \mathcal{Q})$ will denote the Hellinger distance

$$d_H^2(\mathcal{P}, \mathcal{Q}) = \sum_{x \in A} \left(\sqrt{\mathcal{P}(x)} - \sqrt{\mathcal{Q}(x)}\right)^2.$$

Similarly, we abuse this notation slightly to denote the Hellinger distance between two Poisson distributions with different parameters (the arguments in the notation denote parameters of the Poisson distributions). We use the Hellinger distance between two

Poisson distributions, which, for two scalars $p, q \in \mathbb{R}_+$, is given by, $d_H^2(p,q) \triangleq 2 - 2\exp\left\{-\frac{1}{2}\left(\sqrt{p}-\sqrt{q}\right)^2\right\}$. For matrices P, $Q \in \mathbb{R}_+^{d_1\times d_2}$, the average Hellinger distance is defined by

$$d_H^2(P,Q) \triangleq \frac{1}{d_1 d_2}\sum_{i,j} d_H^2(P_{ij}, Q_{ij}).$$

A.1.1 Proof of Theorem 12

To prove Theorem 12, the key is to establish the concentration inequality (Lemma 13) and the lower bound for the average Hellinger distance (Lemma 14).

Lemma 13. *Let $F_{\Omega,Y}(X)$ be the likelihood function defined in (4.2) and $\mathcal{S}$ be the set defined in (3.3), then*

$$\begin{aligned}\mathbb{P}\Bigg\{&\sup_{X\in\mathcal{S}} |F_{\Omega,Y}(X) - \mathbb{E}[F_{\Omega,Y}(X)]| \\ &\geq C'\left(\alpha\sqrt{r}/\beta\right)\left(\alpha(e^2-2)+3\log(d_1d_2)\right)\cdot \\ &\left(\sqrt{m(d_1+d_2)+d_1d_2\log(d_1d_2)}\right)\Bigg\} \leq \frac{C}{d_1d_2},\end{aligned} \tag{A.1}$$

where C' and C are absolute positive constants and the probability and the expectation are both over the choice of Ω and the draw of Y.

Lemma 14. *For any two matrices $P, Q \in \mathcal{S}$, we have*

$$d_H^2(P,Q) \geq \frac{1-e^{-T}}{4\alpha T}\frac{\|P-Q\|_F^2}{d_1d_2},$$

where $T = \frac{1}{8\beta}(\alpha-\beta)^2$.

We will prove Lemma 13 and Lemma 14 below, but first we use them in proving Theorem 12.

Proof of Theorem 12. To begin, notice that for any choice of $X \in \mathcal{S}$,

$$\begin{aligned} & \mathbb{E}\left[F_{\Omega,Y}(X) - F_{\Omega,Y}(M)\right] \\ &= \frac{m}{d_1 d_2} \sum_{i,j} \left[M_{ij} \log\left(\frac{X_{ij}}{M_{ij}}\right) - (X_{ij} - M_{ij}) \right] \\ &= -\frac{m}{d_1 d_2} \sum_{i,j} \left[M_{ij} \log\left(\frac{M_{ij}}{X_{ij}}\right) - (M_{ij} - X_{ij}) \right] \\ &= -\frac{m}{d_1 d_2} \sum_{i,j} D\left(M_{ij} \| X_{ij}\right) = -mD(M\|X), \end{aligned} \tag{A.2}$$

where the expectation is over both Ω and Y.

On the other hand, note that by assumption the true matrix $M \in \mathcal{S}$. Then for any $Z \in \mathcal{S}$, consider the difference below

$$\begin{aligned} & F_{\Omega,Y}(Z) - F_{\Omega,Y}(M) \\ &= F_{\Omega,Y}(Z) + \mathbb{E}[F_{\Omega,Y}(Z)] - \mathbb{E}[F_{\Omega,Y}(Z)] \\ &\quad + \mathbb{E}[F_{\Omega,Y}(M)] - \mathbb{E}[F_{\Omega,Y}(M)] - F_{\Omega,Y}(M) \\ &= \mathbb{E}[F_{\Omega,Y}(Z)] - \mathbb{E}[F_{\Omega,Y}(M)] + \\ &\quad F_{\Omega,Y}(Z) - \mathbb{E}[F_{\Omega,Y}(Z)] + \mathbb{E}[F_{\Omega,Y}(M)] - F_{\Omega,Y}(M) \\ &\le \mathbb{E}\left[F_{\Omega,Y}(Z) - F_{\Omega,Y}(M)\right] + \\ &\quad |F_{\Omega,Y}(Z) - \mathbb{E}[F_{\Omega,Y}(Z)]| + |F_{\Omega,Y}(M) - \mathbb{E}[F_{\Omega,Y}(M)]| \\ &\le -mD(M\|Z) + 2 \sup_{X \in \mathcal{S}} \left| F_{\Omega,Y}(X) - \mathbb{E}[F_{\Omega,Y}(X)] \right|, \end{aligned} \tag{A.3}$$

where the second equality is to rearrange terms, the first inequality is due to triangle inequality, the last inequality is due to (A.2) and the fact that

$$|F_{\Omega,Y}(Z) - \mathbb{E}[F_{\Omega,Y}(Z)]| \le \sup_{X \in \mathcal{S}} |F_{\Omega,Y}(X) - \mathbb{E}[F_{\Omega,Y}(X)]|$$

and

$$|F_{\Omega,Y}(M) - \mathbb{E}[F_{\Omega,Y}(M)]| \leq \sup_{X \in \mathcal{S}} |F_{\Omega,Y}(X) - \mathbb{E}[F_{\Omega,Y}(X)]| \,.$$

Moreover, from the definition of $\widehat{M}$, we also have that $\widehat{M} \in \mathcal{S}$ and $F_{\Omega,Y}(\widehat{M}) \geq F_{\Omega,Y}(M)$. Thus, by substituting $\widehat{M}$ for Z in (A.3), we obtain

$$0 \leq -mD(M\|\widehat{M}) + 2 \sup_{X \in \mathcal{S}} |F_{\Omega,Y}(X) - \mathbb{E}[F_{\Omega,Y}(X)]| \,.$$

To bound the second term in the above expression, we apply Lemma 13, and obtain that with probability at least $1 - C/(d_1 d_2)$, we have

$$\begin{aligned} 0 \leq -mD(M\|\widehat{M}) + 2C' \left(\alpha\sqrt{r}/\beta\right) \left(\alpha(e^2-2) + 3\log(d_1 d_2)\right) \cdot \\ \left(\sqrt{m(d_1+d_2) + d_1 d_2 \log(d_1 d_2)}\right) . \end{aligned}$$

After rearranging terms, and use the fact that $\sqrt{d_1 d_2} \leq d_1 + d_2$, we obtain

$$\begin{aligned} D(M\|\widehat{M}) \leq & 2C' \left(\alpha\sqrt{r}/\beta\right) \left(\alpha(e^2-2) + 3\log(d_1 d_2)\right) \cdot \\ & \left(\sqrt{\frac{d_1+d_2}{m}}\sqrt{1 + \frac{(d_1+d_2)\log(d_1 d_2)}{m}}\right) . \end{aligned} \tag{A.4}$$

Note that the KL divergence can be bounded below by the Hellinger distance (Chapter 3 in [86]). Using our notation to denote the parameters of the Poisson distributions in the argument of the distance, we have

$$d_H^2(p, q) \leq D(p\|q), \tag{A.5}$$

for any two scalars $p, q \in \mathbb{R}_+$ that denote the parameters of the Poisson distributions. Thus, (A.4) together with (A.5) lead to

$$\begin{aligned} d_H^2(M,\widehat{M}) \leq 2C'\left(\alpha\sqrt{r}/\beta\right)\left(\alpha(e^2-2)+3\log(d_1d_2)\right)\cdot \\ \left(\sqrt{\frac{d_1+d_2}{m}}\sqrt{1+\frac{(d_1+d_2)\log(d_1d_2)}{m}}\right). \end{aligned} \tag{A.6}$$

Finally, Theorem 12 follows immediately from Lemma 14.

□

Next, we will establish a tail bound for Poisson distribution with the method of establishing Chernoff bounds. And this result will be used for proving Lemma 13.

Lemma 15 (Tail bound for Poisson)**.** *For $Y \sim$ Poisson(λ) with $\lambda \leq \alpha$, $\mathbb{P}(Y-\lambda \geq t) \leq e^{-t}$, for all $t \geq t_0$ where $t_0 \triangleq \alpha(e^2-3)$.*

Proof of Lemma 15. The proof below is a specialized version of Chernoff bound for Poisson random variable [129] when λ is upper bounded by a constant. For any $\theta \geq 0$, we have

$$\begin{aligned} &\mathbb{P}\left(Y-\lambda \geq t\right) = \mathbb{P}\left(Y \geq t+\lambda\right) \\ =& \mathbb{P}\left(\theta Y \geq \theta\left(t+\lambda\right)\right) = \mathbb{P}\left(\exp\left(\theta Y\right) \geq \exp\left(\theta\left(t+\lambda\right)\right)\right) \\ \leq& \exp\left(-\theta\left(t+\lambda\right)\right)\mathbb{E}\left(e^{\theta Y}\right) = \exp(-\theta(\lambda+t))\cdot\exp\left(\lambda(e^\theta-1)\right), \end{aligned}$$

where we have used Markov's inequality and the moment generating function for Poisson random variable above. Now let $\theta = 2$, we have

$$\exp(t)\cdot\mathbb{P}\left(Y-\lambda \geq t\right) \leq \exp\left(-t+\lambda(e^2-3)\right).$$

Given $t_0 \triangleq \alpha(e^2-3)$, then for all $t \geq t_0$, we have $\exp(t)\cdot\mathbb{P}\left(Y-\lambda \geq t\right) \leq 1$. It follows that $\mathbb{P}\left(Y-\lambda \geq t\right) \leq e^{-t}$ when $t \geq t_0 \geq \alpha(e^2-3)$. □

Proof of Lemma 13. We begin by noting that for any $h > 0$, by using Markov's inequality we have that

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{X\in\mathcal{S}}|F_{\Omega,Y}(X)-\mathbb{E}[F_{\Omega,Y}(X)]|\right.\\
&\qquad \geq C'\left(\alpha\sqrt{r}/\beta\right)\left(\alpha(e^2-2)+3\log(d_1d_2)\right)\cdot\\
&\qquad \left.\left(\sqrt{m(d_1+d_2)+d_1d_2\log(d_1d_2)}\right)\right\}\\
&=\mathbb{P}\left\{\sup_{X\in\mathcal{S}}|F_{\Omega,Y}(X)-\mathbb{E}[F_{\Omega,Y}(X)]\,|^h\right.\\
&\qquad \geq \left(C'\left(\alpha\sqrt{r}/\beta\right)\left(\alpha(e^2-2)+3\log(d_1d_2)\right)\cdot\right.\\
&\qquad \left.\left.\left(\sqrt{m(d_1+d_2)+d_1d_2\log(d_1d_2)}\right)\right)^h\right\}\\
&\leq \mathbb{E}\left[\sup_{X\in\mathcal{S}}|F_{\Omega,Y}(X)-E[F_{\Omega,Y}(X)]|^h\right]/\\
&\quad \{\left(C'\left(\alpha\sqrt{r}/\beta\right)\left(\alpha(e^2-2)+3\log(d_1d_2)\right)\cdot\right.\\
&\quad \left.\left(\sqrt{m(d_1+d_2)+d_1d_2\log(d_1d_2)}\right)\right)^h\}.
\end{aligned} \tag{A.7}$$

The bound in (A.1) follows by combining this with an upper bound on $\mathbb{E}\Big[\sup_{X\in\mathcal{S}}|F_{\Omega,Y}(X)-E[F_{\Omega,Y}(X)]|^h$ and setting $h=\log(d_1d_2)$.

Let ϵ_{ij}s be i.i.d. Rademacher random variables. In the following derivation, the first inequality is due the Radamacher symmetrization argument (Lemma 6.3 in [85]) and the second inequality is due to the power mean inequality: $(a+b)^h\leq 2^{h-1}(a^h+b^h)$ if $a,b>0$ and $h\geq 1$. Then we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{X\in\mathcal{S}}|F_{\Omega,Y}(X)-\mathbb{E}F_{\Omega,Y}(X)|^h\right]\\
&\leq 2^h\mathbb{E}\left[\sup_{X\in\mathcal{S}}\left|\sum_{i,j}\epsilon_{ij}\mathbb{I}\{[(i,j)\in\Omega]\}(Y_{ij}\log X_{ij}-X_{ij})\right|^h\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq 2^h \mathbb{E}\left[2^{h-1}\left(\sup_{X\in\mathcal{S}}\left|\sum_{i,j}\epsilon_{ij}\mathbb{I}\{[(i,j)\in\Omega]\}(Y_{ij}(-\log X_{ij}))\right|^h\right)\right.\\
&\quad\left.+\,2^{h-1}\left(\sup_{X\in\mathcal{S}}\left|\sum_{i,j}\epsilon_{ij}\mathbb{I}\{[(i,j)\in\Omega]\}X_{ij}\right|^h\right)\right]\\
&= 2^{2h-1}\mathbb{E}\left[\sup_{X\in\mathcal{S}}\left|\sum_{i,j}\epsilon_{ij}\mathbb{I}\{[(i,j)\in\Omega]\}(Y_{ij}(-\log X_{ij}))\right|^h\right]\\
&\quad+\,2^{2h-1}\mathbb{E}\left[\sup_{X\in\mathcal{S}}\left|\sum_{i,j}\epsilon_{ij}\mathbb{I}\{[(i,j)\in\Omega]\}X_{ij}\right|^h\right],
\end{aligned}
\tag{A.8}
$$

where the expectation are over both Ω and Y.

To bound the first term of (A.8) with the assumption that $\|X\|_* \leq \alpha\sqrt{rd_1d_2}$, we use the contraction principle (Theorem 4.12 in [85]). Let $\phi(t) = -\beta\log(t+1)$. We know $\phi(0) = 0$ and $|\phi^{'}(t)| = |\beta/(t+1)|$, so $|\phi^{'}(t)| \leq 1$ if $t \geq \beta - 1$. Setting $Z = X - \mathbf{1}_{d_1\times d_2}$, then we have $Z_{ij} \geq \beta - 1, \forall(i,j) \in [\![d_1]\!] \times [\![d_2]\!]$ and $\|Z\|_* \leq \alpha\sqrt{rd_1d_2} + \sqrt{d_1d_2}$ by triangle inequality.

Therefore, $\phi(Z_{ij})$ is a contraction and it vanishes at 0. We obtain that

$$\begin{aligned}
&2^{2h-1}\mathbb{E}\left[\sup_{X\in\mathcal{S}}\left|\sum_{i,j}\epsilon_{ij}\mathbb{I}\{[(i,j)\in\Omega]\}(Y_{ij}(-\log X_{ij}))\right|^{h}\right]\\
&\leq 2^{2h-1}\mathbb{E}\left[\max_{i,j}Y_{ij}^{h}\right]\cdot\\
&\quad\mathbb{E}\left[\sup_{X\in\mathcal{S}}\left|\sum_{i,j}\epsilon_{ij}\mathbb{I}\{[(i,j)\in\Omega]\}((-\log X_{ij}))\right|^{h}\right]\\
&= 2^{2h-1}\mathbb{E}\left[\max_{i,j}Y_{ij}^{h}\right]\cdot\\
&\quad\mathbb{E}\left[\sup_{X\in\mathcal{S}}\left|\sum_{i,j}\epsilon_{ij}\mathbb{I}\{[(i,j)\in\Omega]\}\left(\frac{1}{\beta}\phi(Z_{ij})\right)\right|^{h}\right]\\
&\leq 2^{2h-1}\left(\frac{2}{\beta}\right)^{h}\mathbb{E}\left[\max_{i,j}Y_{ij}^{h}\right]\cdot\\
&\quad\mathbb{E}\left[\sup_{X\in\mathcal{S}}\left|\sum_{i,j}\epsilon_{ij}\mathbb{I}\{[(i,j)\in\Omega]\}Z_{ij})\right|^{h}\right]\\
&= 2^{2h-1}\left(\frac{2}{\beta}\right)^{h}\mathbb{E}\left[\max_{i,j}Y_{ij}^{h}\right]\mathbb{E}\left[\sup_{X\in\mathcal{S}}|\langle\Delta_{\Omega}\circ E,Z\rangle|^{h}\right],
\end{aligned}\tag{A.9}$$

where E denotes the matrix with entries given by ϵ_{ij}, Δ_{Ω} denotes the indicator matrix for Ω and $\circ$ denotes the Hadamard product.

The dual norm of spectral norm is nuclear norm. Using the Hölder's inequality for Schatten norms in [130], which is, $|\langle A,B\rangle|\leq\|A\|\|B\|_{*}$, we have

$$\begin{aligned}
&2^{2h-1}\mathbb{E}\left[\sup_{X\in\mathcal{S}}\left|\sum_{i,j}\epsilon_{ij}\mathbb{I}\{[(i,j)\in\Omega]\}(Y_{ij}(-\log X_{ij}))\right|^{h}\right]\\
&\leq 2^{2h-1}\left(\frac{2}{\beta}\right)^{h}\mathbb{E}\left[\max_{i,j}Y_{ij}^{h}\right]\mathbb{E}\left[\sup_{X\in\mathcal{S}}\|E\circ\Delta_{\Omega}\|^{h}\|Z\|_{*}^{h}\right]\\
&\leq 2^{2h-1}\left(\frac{2}{\beta}\right)^{h}\left(\alpha\sqrt{r}+1\right)^{h}\left(\sqrt{d_1d_2}\right)^{h}\mathbb{E}\left[\max_{i,j}Y_{ij}^{h}\right]\cdot\\
&\quad\mathbb{E}\left[\|E\circ\Delta_{\Omega}\|^{h}\right],
\end{aligned}\tag{A.10}$$

Similarly, the second term of (A.8) can be bounded as follows:

$$\begin{aligned}
&2^{2h-1}\mathbb{E}\left[\sup_{X\in\mathcal{S}}\left|\sum_{i,j}\epsilon_{ij}\mathbb{I}\{[(i,j)\in\Omega]\}X_{ij}\right|^h\right]\\
&\leq 2^{2h-1}\mathbb{E}\left[\sup_{X\in\mathcal{S}}\|E\circ\Delta_\Omega\|^h\|X\|_*^h\right]\\
&\leq 2^{2h-1}\left(\alpha\sqrt{r}\right)^h\left(\sqrt{d_1d_2}\right)^h\mathbb{E}\left[\|E\circ\Delta_\Omega\|^h\right].
\end{aligned}\tag{A.11}$$

Plugging (A.10) and (A.11) into (A.8), we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{X\in\mathcal{S}}|F_{\Omega,Y}(X)-\mathbb{E}F_{\Omega,Y}(X)|^h\right]\\
&\leq 2^{2h-1}\left(\alpha\sqrt{r}+1\right)^h\left(\sqrt{d_1d_2}\right)^h\mathbb{E}\left[\|E\circ\Delta_\Omega\|^h\right]\cdot\\
&\left(\left(\frac{2}{\beta}\right)^h\mathbb{E}\left[\max_{i,j}Y_{ij}^h\right]+1\right).
\end{aligned}\tag{A.12}$$

To bound $\mathbb{E}\left[\|E\circ\Delta_\Omega\|^h\right]$, we use the very first inequality on Page 215 of [8]:

$$\begin{aligned}
&\mathbb{E}\left[\|E\circ\Delta_\Omega\|^h\right]\\
&\leq C_0\left(2(1+\sqrt{6})\right)^h\left(\sqrt{\frac{m(d_1+d_2)+d_1d_2\log(d_1d_2)}{d_1d_2}}\right)^h
\end{aligned}$$

for some constant C_0. Therefore, the only term we need to bound is $\mathbb{E}\left[\max_{i,j}Y_{ij}^h\right]$.

From Lemma 15, if $t\geq t_0$, then for any $(i,j)\in[\![d_1]\!]\times[\![d_2]\!]$, the following inequality holds since $t_0>\alpha$:

$$\begin{aligned}
&\mathbb{P}\left(|Y_{ij}-M_{ij}|\geq t\right)\\
&=\mathbb{P}\left(Y_{ij}\geq M_{ij}+t\right)+\mathbb{P}\left(Y_{ij}\leq M_{ij}-t\right)\\
&\leq\ \exp(-t)+0=\mathbb{P}(W_{ij}\geq t),
\end{aligned}\tag{A.13}$$

where W_{ij}s are independent standard exponential random variables. Because $|Y_{ij}-M_{ij}|$s and W_{ij}'s are all non-negative random variables and $\max(x_1, x_2, \ldots, x_n)$ is an increasing function defined on $\mathbb{R}^n$, we have, for any $h \geq 1$,

$$\mathbb{P}\left(\max_{i,j} |Y_{ij} - M_{ij}|^h \geq t\right) \leq \mathbb{P}(\max_{i,j} W_{ij}^h \geq t), \tag{A.14}$$

for any $t \geq (t_0)^h$.

Below we use the fact that for any positive random variable q, we can write $\mathbb{E}[q] = \int_0^\infty \mathbb{P}(q \geq t)dt$, and then

$$\begin{aligned}
&\mathbb{E}\left[\max_{i,j} Y_{ij}^h\right] \\
&\leq 2^{2h-1}\left(\alpha^h + \mathbb{E}\left[\max_{i,j} |Y_{ij} - M_{ij}|^h\right]\right) \\
&= 2^{2h-1}\left(\alpha^h + \int_0^\infty \mathbb{P}\left(\max_{i,j} |Y_{ij} - M_{ij}|^h \geq t\right) dt\right) \\
&\leq 2^{2h-1}\left(\alpha^h + (t_0)^h + \int_{(t_0)^h}^\infty \mathbb{P}\left(\max_{i,j} |Y_{ij} - M_{ij}|^h \geq t\right) dt\right) \\
&\leq 2^{2h-1}\left(\alpha^h + (t_0)^h + \int_{(t_0)^h}^\infty \mathbb{P}\left(\max_{i,j} W_{ij}^h \geq t\right) dt\right) \\
&\leq 2^{2h-1}\left(\alpha^h + (t_0)^h + \mathbb{E}\left[\max_{i,j} W_{ij}^h\right]\right).
\end{aligned} \tag{A.15}$$

Above, first we use the triangle inequality and the power mean inequality, then along with independence, we use (A.14) in the third inequality. By some standard computations for exponential random variables,

$$\mathbb{E}\left[\max_{i,j} W_{ij}^h\right] \leq 2h! + \log^h(d_1 d_2). \tag{A.16}$$

Thus, we have

$$\mathbb{E}\left[\max_{i,j} Y_{ij}^h\right] \leq 2^{2h-1}\left(\alpha^h + (t_0)^h + 2h! + \log^h(d_1 d_2)\right). \tag{A.17}$$

Therefore, combining (A.17) and (A.12), we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{X\in\mathcal{S}} |F_{\Omega,Y}(X) - \mathbb{E}[F_{\Omega,Y}(X)]|^h\right] \\
&\leq 2^{4h-1}\left(\alpha\sqrt{r}+1\right)^h \left(\sqrt{d_1 d_2}\right)^h \mathbb{E}\left[\|E\circ\Delta_\Omega\|^h\right]\cdot \qquad \text{(A.18)}\\
&\quad\left(\frac{2}{\beta}\right)^h \left(\alpha^h + (t_0)^h + 2h! + \log^h(d_1d_2)\right).
\end{aligned}$$

Then,

$$\begin{aligned}
&\left(\mathbb{E}\left[\sup_{X\in\mathcal{S}} |F_{\Omega,Y}(X) - \mathbb{E}[F_{\Omega,Y}(X)]|^h\right]\right)^{\frac{1}{h}} \\
&\leq 16\left(\alpha\sqrt{r}+1\right)\left(\sqrt{d_1d_2}\right)\mathbb{E}\left[\|E\circ\Delta_\Omega\|^h\right]^{\frac{1}{h}}\cdot \\
&\quad\left(\frac{2}{\beta}\right)(\alpha + t_0 + 2h + \log(d_1d_2)) \\
&\leq 16\left(\frac{2}{\beta}\right)\left(\alpha\sqrt{r}+1\right)\left(\sqrt{d_1d_2}\right)\mathbb{E}\left[\|E\circ\Delta_\Omega\|^h\right]^{\frac{1}{h}}\cdot \qquad \text{(A.19)}\\
&\quad\left(\alpha(e^2-2) + 3\log(d_1d_2)\right) \\
&\leq 128\left(1+\sqrt{6}\right)C_0^{\frac{1}{h}}\left(\frac{\alpha\sqrt{r}}{\beta}\right)\left(\alpha(e^2-2)+3\log(d_1d_2)\right)\cdot \\
&\quad\left(\sqrt{m(d_1+d_2)+d_1d_2\log(d_1d_2)}\right).
\end{aligned}$$

where we use the fact that $(a^h+b^h+c^h+d^h)^{1/h} \leq a+b+c+d$ if $a,b,c,d>0$ in the first inequality and we take $h=\log(d_1d_2)\geq 1$ in the second and the third inequality.

Plugging this into (A.7), we obtain that the probability in (A.7) is upper bounded by

$$C_0\left(\frac{128(1+\sqrt{6})}{C'}\right)^{\log(d_1d_2)} \leq \frac{C_0}{d_1d_2},$$

provided that $C'\geq 128\left(1+\sqrt{6}\right)e$, which establishes this lemma.

□

Proof of Lemma 14. Assuming x is any entry in P and y is any entry in Q, then $\beta\leq x,y\leq\alpha$ and $0\leq|x-y|\leq\alpha-\beta$. By the mean value theorem there exists an $\xi(x,y)\in[\beta,\alpha]$

such that

$$\frac{1}{2}(\sqrt{x}-\sqrt{y})^2 = \frac{1}{2}\left(\frac{1}{2\sqrt{\xi(x,y)}}(x-y)\right)^2$$
$$= \frac{1}{8\xi(x,y)}(x-y)^2 \leq T.$$

The function $f(z) = 1 - e^{-z}$ is concave in $[0, +\infty]$, so if $z \in [0, T]$, we may bound it from below with a linear function

$$1 - e^{-z} \geq \frac{1-e^{-T}}{T}z. \tag{A.20}$$

Plugging $z = \frac{1}{2}(\sqrt{x}-\sqrt{y})^2 = \frac{1}{8\xi(x,y)}(x-y)^2$ into (A.20), we have

$$\begin{aligned} &2 - 2\exp\left(-\frac{1}{2}(\sqrt{x}-\sqrt{y})^2\right) \geq \frac{1-e^{-T}}{T}\frac{1}{4\xi(x,y)}(x-y)^2 \\ &\geq \frac{1-e^{-T}}{T}\frac{1}{4\alpha}(x-y)^2. \end{aligned} \tag{A.21}$$

Note that (A.21) holds for any x and y. This concludes the proof. □

A.1.2 Proof of Theorem 13

Before providing the proof, we first establish two useful lemmas. First, we consider the construction of the set χ.

Lemma 16 (Lemma A.3 in [8]). *Let*

$$H \triangleq \left\{X : \|X\|_* \leq \alpha\sqrt{rd_1d_2}, \|X\|_\infty \leq \alpha\right\}$$

and $\gamma \leq 1$ be such that r/γ^2 is an integer. Suppose $r/\gamma^2 \leq d_1$, then we may construct a set $\chi \in H$ of size

$$|\chi| \geq \exp\left(\frac{rd_2}{16\gamma^2}\right) \tag{A.22}$$

with the following properties:

1. For all $X \in \chi$, each entry has $|X_{ij}| = \alpha\gamma$.

2. For all $X^{(i)}$,$X^{(j)} \in \chi$, $i \neq j$,

$$\|X^{(i)} - X^{(j)}\|_F^2 > \alpha^2\gamma^2 d_1 d_2/2.$$

Second, we consider about the KL divergence.

Lemma 17. *For $x, y > 0$, $D(x\|y) \leq (y-x)^2/y$.*

Proof of Lemma 17. First assume $x \leq y$. Let $z = y - x$. Then $z \geq 0$ and $D(x\|x+z) = x\log\frac{x}{x+z} + z$. Taking the first derivative of this with respect to z, we have $\frac{\partial}{\partial z}D(x\|x+z) = \frac{z}{x+z}$. Thus, by Taylor's theorem, there is some $\xi \in [0, z]$ so that $D(x\|y) = D(x\|x) + z \cdot \frac{\xi}{x+\xi}$. Since the $z\xi/(x+\xi)$ increases in ξ, we may replace ξ with z and obtain $D(x\|y) \leq \frac{(y-x)^2}{y}$. For $x > y$, with the similar argument we may conclude that for $z = y - x < 0$ there is some $\xi \in [z, 0]$ so that $D(x\|y) = D(x\|x) + z \cdot \frac{\xi}{x+\xi}$. Since $z < 0$ and $\xi/(x+\xi)$ increases in ξ, then $z\xi/(x+\xi)$ decreases in ξ. We may also replace ξ with z and this proves the lemma. □

Next, we show how Lemma 16 and Lemma 17 imply Theorem 13. We prove the theorem by contradiction.

Proof of Theorem 13. Without loss of generality, assume $d_2 \geq d_1$. We choose $\epsilon > 0$ such that

$$\epsilon^2 = \min\left\{\frac{1}{256}, C_2\alpha^{3/2}\sqrt{\frac{rd_2}{m}}\right\}, \tag{A.23}$$

where C_2 is an absolute constant that will be be specified later. Next, use Lemma 16 to construct a set χ, choosing γ such that r/γ^2 is an integer and

$$\frac{4\sqrt{2}\epsilon}{\alpha} \leq \gamma \leq \frac{8\epsilon}{\alpha}.$$

We can make such a choice because

$$\frac{\alpha^2 r}{64\epsilon^2} \leq \frac{r}{\gamma^2} \leq \frac{\alpha^2 r}{32\epsilon^2}$$

and

$$\frac{\alpha^2 r}{32\epsilon^2} - \frac{\alpha^2 r}{64\epsilon^2} = \frac{\alpha^2 r}{64\epsilon^2} > 4\alpha^2 r > 1.$$

We verify that such a choice for γ satisfies the the requirements of Lemma 16. Indeed, since $\epsilon \leq \frac{1}{16}$ and $\alpha \geq 1$, $\gamma \leq \frac{1}{2} < 1$. Further, by assumption of the theorem that the right-hand side of (A.23) is larger than $C_1 r\alpha^2/d_1$, which implies $r/\gamma^2 \leq d_1$ for an appropriate choice of C_1.

Let $\chi'_{\alpha/2,\gamma}$ be the set whose existence is guaranteed by Lemma 16, with this choice of γ and with $\alpha/2$ instead of α. Then we can construct χ by defining

$$\chi \triangleq \left\{ X' + \alpha\left(1 - \frac{\gamma}{2}\right)\mathbf{1}_{d_1\times d_2} : X' \in \chi'_{\alpha/2,\gamma} \right\},$$

where $\mathbf{1}_{d_1\times d_2}$ denotes an d_1-by-d_2 matrix of all ones. Note that χ has the same size as $\chi'_{\alpha/2,\gamma}$, i.e.$|\chi|$ satisfies (A.22). χ also has the same bound on pairwise distances

$$\|X^{(i)} - X^{(j)}\|_F^2 \geq \frac{\alpha^2}{4}\frac{\gamma^2 d_1 d_2}{2} \geq 4 d_1 d_2 \epsilon^2, \tag{A.24}$$

for any two matrices $X^{(i)}, X^{(j)} \in \chi$. Define $\alpha' \triangleq (1-\gamma)\alpha$, then every entry of $X \in \chi$ has $X_{ij} \in \{\alpha, \alpha'\}$. Since we assume $r \geq 4$ in theorem statement, for any $X \in \chi$, we have that for some $X' \in \chi'_{\alpha/2,\gamma}$,

$$\begin{aligned} \|X\|_* &= \|X' + \alpha\left(1 - \frac{\gamma}{2}\right)\mathbf{1}_{d_1\times d_2}\|_* \\ &\leq \frac{\alpha}{2}\sqrt{r d_1 d_2} + \alpha\sqrt{d_1 d_2} \leq \alpha\sqrt{r d_1 d_2}. \end{aligned}$$

Since we choose γ less than $1/2$, we have that α' is greater than $\alpha/2$. Therefore, from the assumption that $\beta \leq \alpha/2$, we conclude that $\chi \subset \mathcal{S}$.

Now suppose for the sake of a contradiction that there exists an algorithm such that for any $X \in \mathcal{S}$, when given access to the measurements on Ω_0, returns $\widehat{X}$ such that

$$\frac{1}{d_1 d_2}\|X - \widehat{X}\|_F^2 < \epsilon^2 \tag{A.25}$$

with probability at least $1/4$. We will imagine running this algorithm on a matrix X chosen uniformly at random from χ. Let

$$X^* = \arg\min_{Z \in \chi} \|Z - \widehat{X}\|_F^2.$$

By the same argument as that in [8], we can claim that $X^* = X$ as long as (A.25) holds. Indeed, for any $X' \in \chi$ with $X' \neq X$, from (A.24) and (A.25), we have that

$$\|X' - \widehat{X}\|_F \geq \|X' - X\|_F - \|X - \widehat{X}\|_F > \sqrt{d_1 d_2}\epsilon.$$

At the same time, since $X \in \chi$ is a candidate for X^*, we have that

$$\|X^* - \widehat{X}\|_F \leq \|X - \widehat{X}\|_F \leq \sqrt{d_1 d_2}\epsilon.$$

Thus, if (A.25) holds, then $\|X^* - \widehat{X}\|_F < \|X' - \widehat{X}\|_F$ for any $X' \in \chi$ with $X' \neq X$, and hence we must have $X^* = X$.

Using the assumption that (A.25) holds with probability at least $1/4$, we have that

$$\mathbb{P}(X^* \neq X) \leq \frac{3}{4}. \tag{A.26}$$

We will show that this probability must in fact be large, generating a contradiction.

By a variant of Fano's inequality in [87], we have

$$\max_{X \in S} \mathbb{P}(X^* \neq X) \geq 1 - \frac{\max_{X^{(k)} \neq X^{(l)}} \widetilde{D}(X^{(k)} \| X^{(l)}) + 1}{\log |\chi|}, \tag{A.27}$$

where

$$\widetilde{D}(X^{(k)} \| X^{(l)}) \triangleq \sum_{(i,j) \in \Omega_0} D(X_{ij}^{(k)} \| X_{ij}^{(l)}),$$

and the maximum is taken over all pairs of different matrices $X^{(k)}$ and $X^{(l)}$ in χ. For any such pairs $X^{(k)}, X^{(l)} \in \chi$, $D(X_{ij}^{(k)} \| X_{ij}^{(l)})$ is either 0, $D(\alpha \| \alpha')$, or $D(\alpha' \| \alpha)$ for $(i,j) \in [\![d_1]\!] \times [\![d_2]\!]$. Define an upper bound on the KL divergence quantities

$$D \triangleq \max_{X^{(k)} \neq X^{(l)}} \widetilde{D}(X^{(k)} \| X^{(l)}).$$

By the assumption that $|\Omega_0| = m$, using Lemma 17 and the fact that $\alpha' < \alpha$, we have

$$D \leq \frac{m(\gamma\alpha)^2}{\alpha'} \leq \frac{64 m \epsilon^2}{\alpha'}.$$

Combining (A.26) and (A.27), we have that

$$\begin{aligned} \frac{1}{4} &\leq 1 - \mathbb{P}(X \neq X^*) \leq \frac{D+1}{\log |\chi|} \\ &\leq 16\gamma^2 \left(\frac{\frac{64 m \epsilon^2}{\alpha'} + 1}{r d_2} \right) \leq 1024 \epsilon^2 \left(\frac{\frac{64 m \epsilon^2}{\alpha'} + 1}{\alpha^2 r d_2} \right). \end{aligned} \tag{A.28}$$

We now show that for appropriate values of C_0 and C_2, this leads to a contradiction. Suppose $64 m \epsilon^2 \leq \alpha'$, then with (A.28), we have that

$$\frac{1}{4} \leq 1024 \epsilon^2 \frac{2}{\alpha^2 r d_2},$$

which together with (A.23) implies that $\alpha^2 r d_2 \leq 32$. If set $C_0 > 32$, this would lead to a contradiction. Suppose $64m\epsilon^2 > \alpha'$, (A.28) simplifies to

$$\frac{1}{4} < 1024\epsilon^2 \left(\frac{128m\epsilon^2}{(1-\gamma)\alpha^3 r d_2} \right).$$

Since $1 - \gamma > 1/2$, we have

$$\epsilon^2 > \frac{\alpha^{3/2}}{1024} \sqrt{\frac{rd_2}{m}}.$$

Setting $C_2 \leq 1/1024$ in (A.23) leads to a contradiction. Therefore, (A.25) must be incorrect with probability at least $3/4$, which proves the theorem.

□

A.1.3 Proofs of Proposition 1 and Proposition 2

Lemma 18. *If f is a closed convex function satisfying Lipschitz condition (3.15), then for any $X, Y \in \mathcal{S}$, the following inequality holds:*

$$f(Y) \leq f(X) + \langle \nabla f(X), Y - X \rangle + \frac{L}{2} \|Y - X\|_F^2.$$

Proof. Let $Z = Y - X$, then we have

$$\begin{aligned} f(Y) &= f(X) + \langle \nabla f(X), Z \rangle \\ &\qquad + \int_0^1 \langle \nabla f(X + tV) - \nabla f(X), Z \rangle \, dt \\ &\leq f(X) + \langle \nabla f(X), Z \rangle \\ &\qquad + \int_0^1 \|f(X + tV) - \nabla f(X)\|_F \|Z\|_F \, dt \\ &\leq f(X) + \langle \nabla f(X), Z \rangle + \int_0^1 Lt \|Z\|_F^2 \, dt \\ &= f(X) + \langle \nabla f(X), Y - X \rangle + \frac{L}{2} \|Y - X\|_F^2, \end{aligned}$$

where we use the Taylor expansion with integral remainder in the first line, the fact that dual norm of Frobenius norm is itself in the second line and Lipschitz condition in the third line. □

In the following, proofs for Proposition 1 and Proposition 2 use results from [101].

Proof of Proposition 1. It is well known that the proximal mapping of a $Y \in \mathcal{S}$ associated with a closed convex function h is given by

$$\text{prox}_{th}(Y) \triangleq \arg\min_{X} \left(t \cdot h(X) + \frac{1}{2}\|X - Y\|_F^2 \right),$$

where $t > 0$ is a multiplier. In our case, $h(P) = \mathbb{I}_{\mathcal{S}}(P)$. Define for each $P \in \mathcal{S}$ that

$$G_t(P) \triangleq \frac{1}{t}\left(P - \text{prox}_{th}\left(P - t\nabla f(P)\right)\right),$$

then by the characterization of subgradient,

$$G_t(P) - \nabla f(P) \in \partial h(P), \tag{A.29}$$

where $\partial h(P)$ is the subdifferential of h at P. Noticing that $P - tG_t(P) \in \mathcal{S}$, then from Lemma 18 we have

$$f(P - tG_t(P)) \leq f(P) - \langle \nabla f(P), tG_t(P)\rangle + \frac{t}{2}\|G_t(P)\|_F^2, \tag{A.30}$$

for all $0 \leq t \leq 1/L$. Define that $g(P) \triangleq f(P) + h(P)$. Combining (A.29) and (A.30) and using the fact that f and h are convex functions, we have for any $Z \in \mathcal{S}$ and $0 \leq t \leq 1/L$,

$$g(P - tG_t(P)) \leq g(Z) + \langle G_t(P), P - Z\rangle - \frac{t}{2}\|G_t(P)\|_F^2, \tag{A.31}$$

which is analogous to inequality (3.3) in [101]. Taking $Z = \widehat{M}$ and $P = X_k$ in (A.31), then for any $k \geq 0$,

$$\begin{aligned} g(X_{k+1}) - g(\widehat{M}) &\leq \langle G_t(X_k), X_k - \widehat{M} \rangle - \frac{t}{2}\|G_t(X_k)\|_F^2 \\ &= \frac{1}{2t}\left(\|X_k - \widehat{M}\|_F^2 - \|X_{k+1} - \widehat{M}\|_F^2\right), \end{aligned} \tag{A.32}$$

where we use the fact that $\langle P, P \rangle = \|P\|_F^2$. By taking $Z = X_k$ and $P = X_k$ in (A.31) we know that $g(X_{k+1}) < g(X_k)$ for any $k \geq 0$. Thus, taking $t = 1/L$, we have,

$$\begin{aligned} g(X_k) - g(\widehat{M}) &\leq \frac{1}{k}\sum_{i=0}^{k-1}\left(g(X_{i+1}) - g(\widehat{M})\right) \\ \leq \frac{L}{2k}\sum_{i=0}^{k-1}&\left(\|X_i - \widehat{M}\|_F^2 - \|X_{i+1} - \widehat{M}\|_F^2\right) \leq \frac{L\|X_0 - \widehat{M}\|_F^2}{2k}. \end{aligned} \tag{A.33}$$

Since $X_k \in \mathcal{S}$ for any $k \geq 0$ and $\widehat{M} \in \mathcal{S}$, we have that $h(X_k) = 0$ for any $k \geq 0$ and $h(\widehat{M}) = 0$, which completes the proof.

□

Proof of Proposition 2. The definitions and notations in the proof of Proposition 1 are also valid in the proof of Proposition 2.

Define $V_0 \triangleq X_0$ and for any $k \geq 1$,

$$a_k \triangleq \frac{2}{k+1},\ V_k \triangleq X_{k-1} + \frac{1}{a_k}\left(X_k - X_{k-1}\right).$$

For any $0 \leq t \leq 1/L$, noticing that

$$X_k = Z_{k-1} - tG_t(Z_{k-1}),$$

then we can rewrite V_k as

$$V_k = V_{k-1} - \frac{t}{a_k}G_t(Z_{k-1}).$$

Taking $Z = X_{k-1}$ and $Z = \widehat{M}$ in (A.31) and making convex combination we have

$$\begin{aligned} g(X_k) &\leq (1 - a_k)g(X_{k-1}) + a_k g(\widehat{M}) \\ &\qquad + a_k \langle G_t(Z_{k-1}), V_{k-1} - \widehat{M} \rangle - \frac{t}{2} \| G_t(Z_{k-1}) \|_F^2 \\ &= (1 - a_k)g(X_{k-1}) + a_k g(\widehat{M}) \\ &\qquad + \frac{a_k^2}{2t} \left(\|V_{k-1} - \widehat{M}\|_F^2 - \|V_k - \widehat{M}\|_F^2 \right). \end{aligned} \tag{A.34}$$

After rearranging terms, we have

$$\begin{aligned} &\frac{1}{a_k^2}(g(X_k) - g(\widehat{M})) + \frac{1}{2t}\|V_k - \widehat{M}\|_F^2 \leq \\ &\frac{1 - a_k}{a_k^2}(g(X_{k-1}) - g(\widehat{M})) + \frac{1}{2t}\|V_{k-1} - \widehat{M}\|_F^2. \end{aligned} \tag{A.35}$$

Notice that $(1 - a_k)/(a_k^2) \leq 1/(a_{k-1}^2)$ for any $k \geq 1$. Applying inequality (A.35) recursively,

$$\frac{1}{a_k^2}(g(X_k) - g(\widehat{M})) + \frac{1}{2t}\|V_k - \widehat{M}\|_F^2 \leq \frac{1}{2t}\|X_0 - \widehat{M}\|_F^2. \tag{A.36}$$

Taking $t = 1/L$, we have

$$g(X_k) - g(\widehat{M}) \leq \frac{2L\|X_0 - \widehat{M}\|_F^2}{(k+1)^2}.$$

Since $X_k \in \mathcal{S}$ for any $k \geq 0$ and $\widehat{M} \in \mathcal{S}$, we have that $h(X_k) = 0$ for any $k \geq 0$ and $h(\widehat{M}) = 0$, which comcludes the proof.

□

A.2 Proofs for multi-sensor slope change

A.2.1 An informal derivation of Theorem 14: ARL

We first obtain an approximation to the probability that the stopping time is greater than some big constant m. Such an approximation is obtained using a general method for computing first passing probabilities first introduced in [92] and developed in [109]. The method relies on measure transformations that shift the distribution of each sensor over a window that contains the hypothesized post-change samples. More technical details to make the proofs more rigorous are omitted. These details have been described and proved in [109].

In the following, let $\tau = t - k$. Define the log moment-generating-function $\psi_\tau(\theta) = \log \mathbb{E} \exp\{\theta g(U_{n,k,t})\}$. Recall that $U_{n,k,t}$ is a generic standardized sum over all observations within a window of size τ in one sensor, and the parameter $\theta = \theta_\tau$ is selected by solving the equation

$$\dot{\psi}_\tau(\theta) = b/N.$$

Since $U_{n,k,t}$ is a standardized weighted sum of τ independent random variables, ψ_τ converges to a limit as $\tau \to \infty$, and θ_τ converges to a limiting value. We denote this limiting value by θ.

Denote the density function under the null as $\mathbb{P}$. The transformed distribution for all sequences at a fixed current time t and at a hypothesized change-point time k (and hence there are τ hypothesized post-change samples) is denoted by $\mathbb{P}_t^k$ and is defined via

$$d\mathbb{P}_t^k = \exp\left[\theta_\tau \sum_{n=1}^{N} g(U_{n,k,t}) - N\psi_\tau(\theta_\tau)\right] d\mathbb{P}.$$

Let

$$\ell_{N,k,t} = \log(d\mathbb{P}_t^k / d\mathbb{P}) = \theta_\tau \sum_{n=1}^{N} g(U_{n,k,t}) - N\psi_\tau(\theta_\tau).$$

Let the region

$$D = \{(t,k) : 0 < t < t_0, 1 \leq t - k \leq w\}$$

be the set of all possible change-point times and time up to a horizon m. Let

$$A = \left\{ \max_{(t,k) \in D} \sum_{n=1}^{N} g(U_{n,k,t}) \geq b \right\}.$$

be the event of interest. Hence, we have

$$\begin{aligned}\mathbb{P}\{A\} &= \sum_{(t,k)\in D} \mathbb{E}\left\{\frac{e^{\ell_{N,k,t}}}{\sum_{(t',k')\in D} e^{\ell_{N,k',t'}}}; A\right\} = \sum_{(t,k)\in D} \mathbb{E}_t^k\left\{\left(\sum_{(t',k')\in D} e^{\ell_{N,k',t'}}\right)^{-1}; A\right\}\\ &= \sum_{(t,k)\in D} e^{-N(\theta_\tau b-\psi_\tau(\theta_\tau))} \times \underbrace{\mathbb{E}_t^k\left\{\frac{M_{N,k,t}}{S_{N,k,t}} e^{-\tilde{\ell}_{N,k,t}-\log M_{N,k,t}}; \tilde{\ell}_N + \log M_{N,k,t} \geq 0\right\}}_{I}\end{aligned} \tag{A.37}$$

where

$$\tilde{\ell}_{N,k,t} = \sum_{n=1}^{N} \theta_\tau[g(U_{n,k,t}) - b],$$

$$S_{N,k,t} = \sum_{(t',k')\in D} e^{\sum_{n=1}^{N} \theta_\tau[g(U_{n,k',t'})-g(U_{n,k,t})]},$$

$$M_{N,k,t} = \max_{(t',k')\in D} e^{\sum_{n=1}^{N} \theta_\tau[g(U_{n,k',t'})-g(U_{n,k,t})]}.$$

As explained in [109], under certain verifiable assumptions, a “localization lemma” allows simplifying the quantities of the form in I into a much simpler expression of the form

$$\sigma_{N,\tau}^{-1}(2\pi)^{-1/2}\mathbb{E}\{M/S\},$$

where $\sigma_{N,\tau}$ is the $\mathbb{P}_s^\tau$ standard deviation of $\tilde{\ell}_N$ and $\mathbb{E}[M/S]$ is the limit of $\mathbb{E}\{M_{N,k,t}/S_{N,k,t}\}$ as $N \to \infty$. This reduction relies on the fact that, for large N and m, the “local” processes $M_{N,k,t}$ and $S_{N,k,t}$ are approximately independent of the “global” process $\tilde{\ell}_N$. This allows the expectation to be decomposed into the expectation of M_N/S_N times the expectation involving $\tilde{\ell}_N + \log M_N$, treating $\log M_N$ as a constant.

Let $\tau' = t' - k'$, and denote by $z_{n,i} = (y_{n,i} - \mu_n)/\sigma_n$ which are i.i.d. normal random variables, $i = 1, 2, \ldots$. Note that, use Taylor expansion up to the first order, we obtain

$$\begin{aligned}
&\sum_{n=1}^{N} \theta_\tau [g(U_{n,k',t'}) - g(U_{n,k,t})] \approx \sum_{n=1}^{N} \theta_\tau \dot{g}(U_{n,k,t})[U_{n,k',t'} - U_{n,k,t}] \\
=&\sum_{n=1}^{N} \theta_\tau \dot{g}(U_{n,k,t})[A_{\tau'}^{-1/2} W_{n,k',t'} - A_\tau^{-1/2} W_{n,k',t'} + A_\tau^{-1/2} W_{n,k',t'} - A_\tau^{-1/2} W_{n,k,t}] \\
=&\sum_{n=1}^{N} \frac{\theta_\tau \dot{g}(U_{n,k,t})}{\sqrt{A_{\tau'}}} (\sum_{j=1}^{\tau'} j z_{n,t'-\tau'+j} - \sqrt{\frac{A_{\tau'}}{A_\tau}} \sum_{j=1}^{\tau'} j z_{n,t'-\tau'+j}) \\
&\qquad + \sum_{n=1}^{N} \theta_\tau \dot{g}(U_{n,k,t}) A_\tau^{-1/2} \left(\sum_{i=1}^{\tau'} i z_{n,t'-\tau'+i} - \sum_{i=1}^{\tau} i z_{n,t-\tau+i} \right)
\end{aligned} \tag{A.38}$$

Note that in the above expression, the first term has two weighted data sequences running backwards from t' and when τ and τ' both tends to infinity they tend to cancel with each other. Hence, asymptotically we need to consider the second term. Observe that one may let $t' - k' = \tau$ and $\theta = \lim_{\tau \to \infty} \theta_\tau$ for θ_τ in the definition of the increments and still maintain the required level of accuracy. When $\tau = u$ the first term in the above expression, and the second term consists of two terms that are highly correlated. The second term can be rewritten as

$$A_\tau^{-1/2} \theta_\tau \left[\sum_{n=1}^{N} \dot{g}(U_{n,k,t}) W_{n,k',t'} - \sum_{n'=1}^{N} \dot{g}(U_{n',k,t}) W_{n',k,t} \right]. \tag{A.39}$$

Since all sensors are assumed to be independent (or has been whitened by a known covariance matrix so the transformed coordinates are independent), so the covariance between the two terms is given by

$$\text{Cov}\left(\sum_{n=1}^{N} \dot{g}(U_{n,k,t}) W_{n,k,t}, \sum_{n'=1}^{N} \dot{g}(U_{n',k,t}) W_{n',k,t} \right) = \sum_{n=1}^{N} [\dot{g}(U_{n',k,t})]^2 \text{Cov}(W_{n,k,t}, W_{n',k',t'}). \tag{A.40}$$

For each n, let $k < k' < t < t'$ and $t - k = t' - k' = \tau$, and define $u \triangleq k' - k$ and

$s \triangleq t - k'$. We have

$$A_\tau^{-1}\text{Cov}(W_{n,k,t}, W_{n,k',t'}) = \mathbb{E}\left\{ \frac{\left(\sum_{i=k+1}^{t}(i-k)z_{n,i}\right)\left(\sum_{i=k'+1}^{t'}(i-k')z_{ni}\right)}{\sum_{i=1}^{\tau} i^2} \right\}$$

$$= \mathbb{E}\left\{ \frac{\sum_{i=k'+1}^{t}(i-k)(i-k')z_{n,i}^2}{\sum_{i=1}^{\tau} i^2} \right\} = \frac{\sum_{i=1}^{s} i^2 + u\sum_{i=1}^{s} i}{\sum_{i=1}^{\tau} i^2}.$$

By choosing $u = \sqrt{\tau}$, we know that the expression above is approximately on the order of

$$1 - \frac{(k'-k)+(t'-t)}{2\left(\frac{2}{3}\tau^2 + \frac{1}{3}\tau\right)} \approx 1 - \frac{(k'-k)+(t'-t)}{\frac{4}{3}\tau^2}.$$

Let $\eta \triangleq \frac{4}{3}\tau^2$. Hence, by summarizing the derivations above and applying the law of large number, we have that when $N \to \infty$ and $\tau \to \infty$, the covariance between the two terms become

$$\text{cov}\left(\sum_{n=1}^{N}\theta_\tau g(U_{n,k',t'}), \sum_{n'=1}^{N}\theta_\tau g(U_{n',k,t})\right) \approx \theta^2 N \cdot [1 - \frac{1}{\eta}(k'-k) - \frac{1}{\eta}(t'-t)].$$

This shows that the two-dimensional random walk decouples in the change-point time k' and the time index t' and the variance of the increments in these two directions are the same and are both equal to $\theta^2 N/\eta$. Hence, the random walk along these two coordiates are asymptotically independent and it becomes similar to the case studied in [109]. Compare this with (the equation following equation (A.4) in [109]), note that the only difference is that here the variance of the increment is proportional to $3/(4\tau^2)$ instead of τ, so we may follow a similar chains of calculation as in the proof in Chapter 7 of [92], [109] [18], the final result corresponds to modifying the upper and lower limit by changing the window length expression to be $\sqrt{4/3}$ and $\sqrt{4w/3}$.

A.2.2 An informal derivation of Theorem 15: EDD

Recall that $U_{n,k,t}$ is defined in (4.6), let $z_{n,i} = (y_{n,i} - \mu_n)/\sigma_n$. Then for $n \in \mathcal{A}$, $z_{n,i}$ are i.i.d. normal random variables with mean $c_n i/\sigma_n$ and unit variance, and for $n \in \mathcal{A}^c$, $z_{n,i}$ are i.i.d. standard normal random variables. Since we may write

$$U_{n,k,t} = \frac{\sum_{i=k+1}^{t}(i-k)z_{n,i}}{\sqrt{\sum_{i=k+1}^{t}(i-k)^2}}. \tag{A.41}$$

For any time t and $n \in \mathcal{A}$, we have

$$\mathbb{E}_0^{\mathcal{A}}\{U_{n,0,t}^2\} = 1 + \left(\frac{c_n}{\sigma_n}\right)^2 \sum_{i=1}^{t} i^2 = 1 + \left(\frac{c_n}{\sigma_n}\right)^2 \frac{t(t+1)(2t+1)}{6} = \left(\frac{c_n}{\sigma_n}\right)^2 \frac{t^3}{3} + o(t^3), \tag{A.42}$$

which grows cubically with respect to time. For the unaffected sensors, $n \in \mathcal{A}^c$, $\mathbb{E}_0^{\mathcal{A}}\{U_{n,0,t}^2\} = 1$. Hence, the value of the detection statistic will be dominated by those affected sensors.

On the other hand, note that when x is large,

$$g(x) = \log(1 - p_0 + p_0 e^{x^2/2}) = \log p_0 + \frac{x^2}{2} + \log\left(\frac{1-p_0}{p_0} e^{-x^2/2}\right) \approx \frac{x^2}{2} + \log p_0.$$

Then the expectation of the statistic in (4.8) can be computed if w is sufficiently large (at least larger than the expected detection delay), as follows:

$$\mathbb{E}_0^{\mathcal{A}}\left\{\max_{k<t} \sum_{n=1}^{N} g\left(U_{n,k,t}\right)\right\} \approx \left(|\mathcal{A}| \log p_0 + \frac{1}{2}\sum_{n\in\mathcal{A}} \mathbb{E}_0^{\mathcal{A}}\left\{U_{n,k,t}^2\right\} + \frac{(N - |\mathcal{A}|)}{2}\right),$$

At the stopping time, if we ignore of the overshoot of the threshold over b, the value statistic is b. Use Wald's identity [131] and if we ignore the overshoot of the statistic over the threshold b, we may obtain a first order approximation as $b \to \infty$, by solving

$$|\mathcal{A}| \log p_0 + \frac{N - |\mathcal{A}|}{2} + \frac{\mathbb{E}_0^{\mathcal{A}}\{T^3\}}{6}\left[\sum_{n\in\mathcal{A}} \left(\frac{c_n}{\sigma_n}\right)^2\right] = b. \tag{A.43}$$

From Jensen's inequality, we know that $\mathbb{E}_0^{\mathcal{A}}\{T_2^3\} \geq (\mathbb{E}_0^{\mathcal{A}}\{T_2\})^3$. Therefore, a first-order approximation for the expected detection delay is given by

$$\mathbb{E}_0^{\mathcal{A}}\{T_2\} \leq \left(\frac{b - N\log p_0 - (N - |\mathcal{A}|)\mathbb{E}\{g(U)\}}{\Delta^2/6}\right)^{1/3} + o(1). \tag{A.44}$$

A.2.3 Proof for Optimality

Proof of Theorem 16. The proof starts by a change of measure from $\mathbb{P}_\infty$ to $\mathbb{P}_k^{\mathcal{A}}$. For any stopping time $T \in C(\gamma)$, we have that for any $K_\gamma > 0$, $C > 0$ and $\varepsilon \in (0,1)$,

$$\begin{aligned}
&\mathbb{P}_\infty\left\{k < T < k + (1-\varepsilon)K_\gamma | T > k\right\} \\
&= \mathbb{E}_k^{\mathcal{A}}\left\{\mathbb{I}_{\{k<T<k+(1-\varepsilon)K_\gamma\}}\exp(-\lambda_{\mathcal{A},k,T})\big|T > k\right\} \\
&\geq \mathbb{E}_k^{\mathcal{A}}\left\{\mathbb{I}_{\{k<T<k+(1-\varepsilon)K_\gamma,\lambda_{\mathcal{A},k,T}<C\}}\exp(-\lambda_{\mathcal{A},k,T})\big|T > k\right\} \\
&\geq e^{-C}\mathbb{P}_k^{\mathcal{A}}\left\{k < T < k + (1-\varepsilon)K_\gamma, \max_{k<j<k+(1-\varepsilon)K_\gamma}\lambda_{\mathcal{A},k,j} < C\middle|T > k\right\} \\
&\geq e^{-C}\left[\mathbb{P}_k^{\mathcal{A}}\left\{T < k + (1-\varepsilon)K_\gamma | T > k\right\} - \right. \\
&\qquad \left.\mathbb{P}_k^{\mathcal{A}}\left\{\max_{1\leq j<(1-\varepsilon)K_\gamma}\lambda_{\mathcal{A},k,k+j} \geq C\middle|T > k\right\}\right],
\end{aligned} \tag{A.45}$$

where $\mathbb{I}_{\{A\}}$ is the indicator function of any event A, the first equality is Wald's likelihood ratio identity and the last inequality uses the fact that for any event A and B and probability measure $\mathbb{P}$, $\mathbb{P}(A\bigcap B) \geq \mathbb{P}(A) - \mathbb{P}(B^c)$.

From (A.45) we have for any $\varepsilon \in (0,1)$

$$\mathbb{P}_k^{\mathcal{A}}\left\{T < k + (1-\varepsilon)K_\gamma | T > k\right\} \leq p_{\gamma,\varepsilon}^{(k)}(T) + \beta_{\gamma,\varepsilon}^{(k)}(T), \tag{A.46}$$

where

$$p_{\gamma,\varepsilon}^{(k)}(T) = e^C \mathbb{P}_\infty \left\{ T < k + (1-\varepsilon)K_\gamma | T > k \right\},$$
$$\beta_{\gamma,\varepsilon}^{(k)}(T) = \mathbb{P}_k^{\mathcal{A}} \left\{ \max_{1 \le j < (1-\varepsilon)K_\gamma} \lambda_{\mathcal{A},k,k+j} \ge C \middle| T > k \right\}.$$

Next, we want to show that both $p_{\gamma,\varepsilon}^{(k)}(T)$ and $\beta_{\gamma,\varepsilon}^{(k)}(T)$ converge to zero for any $T \in C(\gamma)$ and any $k \ge 0$ as γ goes to infinity.

First, choosing $C = (1+\varepsilon)I[(1-\varepsilon)K_\gamma]^q$, then we have

$$\begin{aligned} \beta_{\gamma,\varepsilon}^{(k)}(T) &= \mathbb{P}_k \left\{ [(1-\varepsilon)K_\gamma]^{-q} \max_{1 \le j < (1-\varepsilon)K_\gamma} \lambda_{\mathcal{A},k,k+j} \ge (1+\varepsilon)I \middle| T > k \right\} \\ &\le \operatorname{esssup} \mathbb{P}_k^{\mathcal{A}} \left\{ [(1-\varepsilon)K_\gamma]^{-q} \max_{1 \le j < (1-\varepsilon)K_\gamma} \lambda_{\mathcal{A},k,k+j} \ge (1+\varepsilon)I \middle| \mathcal{F}_k \right\}. \end{aligned} \tag{A.47}$$

By the assumption (4.17), we have

$$\sup_{0 \le k < \infty} \beta_{\gamma,\varepsilon}^{(k)} \xrightarrow[\gamma \to \infty]{} 0. \tag{A.48}$$

Second, by Lemma 6.3.1 in [55], we know that for any $T \in C(\gamma)$ there exists a $k \ge 0$, possibly depending on γ, such that

$$\mathbb{P}_\infty \left\{ T < k + (1-\varepsilon)K_\gamma | T > k \right\} \le (1-\varepsilon)K_\gamma/\gamma.$$

Choosing $K_\gamma = (I^{-1}\log\gamma)^{1/q}$, then we have

$$C = (1+\varepsilon)I(1-\varepsilon)^q I^{-1} \log\gamma = (1-\varepsilon^2)(1-\varepsilon)^{q-1}\log\gamma,$$

and therefore,

$$\begin{aligned} p_{\gamma,\varepsilon}^{(k)}(T) &\le \gamma^{(1-\varepsilon^2)(1-\varepsilon)^{q-1}}(1-\varepsilon)K_\gamma/\gamma \\ &= (1-\varepsilon)(I^{-1}\log\gamma)^{1/q}\gamma^{(1-\varepsilon^2)(1-\varepsilon)^{q-1}-1} \xrightarrow[\gamma \to \infty]{} 0, \end{aligned} \tag{A.49}$$

where the last convergence holds since for any $q \geq 1$ and $\varepsilon \in (0,1)$ we have $(1-\varepsilon^2)(1-\varepsilon)^{q-1} < 1$. Therefore, for every $\varepsilon \in (0,1)$ and for any $T \in C(\gamma)$ we have that for some $k \geq 0$,

$$\mathbb{P}_k^{\mathcal{A}} \{T < k + (1-\varepsilon)K_\gamma | T > k\} \xrightarrow[\gamma\to\infty]{} 0,$$

which proves (4.18).

Next, to prove (4.19), since

$$\text{ESM}_m^{\mathcal{A}}(T) \geq \text{SM}_m^{\mathcal{A}}(T) \geq \sup_{0\leq k<\infty} \mathbb{E}_k^{\mathcal{A}} \left\{[(T-k)^+]^m \middle| T > k\right\},$$

it is suffice to show that for any $T \in C(\gamma)$,

$$\sup_{0\leq k<\infty} \mathbb{E}_k^{\mathcal{A}} \left\{[(T-k)^+]^m \middle| T > k\right\} \geq [I^{-1}\log\gamma]^{m/q}(1+o(1)) \text{ as } \gamma \to 0, \tag{A.50}$$

where the residual term $o(1)$ does not depend on T. Using the result (4.18) just proved, we can have that for any $\varepsilon \in (0,1)$, there exists some $k \geq 0$ such that

$$\inf_{T\in C(\gamma)} \mathbb{P}_k^{\mathcal{A}} \left\{T - k \geq (1-\varepsilon)(I^{-1}\log\gamma)^{\frac{1}{q}} \middle| T > k\right\} \xrightarrow[\gamma\to\infty]{} 1.$$

Therefore, by also Chebyshev inequality, for any $\varepsilon \in (0,1)$ and $T \in C(\gamma)$, there exist some $k \geq 0$ such that

$$\begin{aligned} &\mathbb{E}_k^{\mathcal{A}} \left\{[(T-k)^+]^m \middle| T > k\right\} \\ &\geq \left[(1-\varepsilon)(I^{-1}\log\gamma)^{\frac{1}{q}}\right]^m \mathbb{P}_k^{\mathcal{A}} \left\{T - k \geq (1-\varepsilon)(I^{-1}\log\gamma)^{\frac{1}{q}} \middle| T > k\right\} \\ &\geq \left[(1-\varepsilon)^m (I^{-1}\log\gamma)^{m/q}\right](1+o(1)), \text{ as } \gamma \to \infty, \end{aligned} \tag{A.51}$$

where the residual term does not depend on T. Since we can arbitrarily choose $\varepsilon \in (0,1)$ such that the (A.51) holds, so we have (A.50), which completes the proof.

□

Proof of Lemma 8. Rewrite $T_{\mathrm{CS}}(b)$ as

$$T_{\mathrm{CS}}(b) = \inf \left\{ t : \max_{0 \le k < t} \prod_{n=1}^{N} (1 - p_0 + p_0 \exp(\lambda_{n,k,t})) \ge e^b \right\} \tag{A.52}$$

and define $T_{\mathrm{SR}}(b)$ an extended Shiryaev-Roberts (SR) procedure as follows:

$$T_{\mathrm{SR}}(b) = \inf \left\{ t : R_t \ge e^b \right\}, \tag{A.53}$$

where

$$R_t = \sum_{k=1}^{t-1} \prod_{n=1}^{N} (1 - p_0 + p_0 \exp(\lambda_{n,k,t})), t = 1, 2, \ldots; R_0 = 0.$$

Clearly, $T_{\mathrm{CS}}(b) \ge T_{\mathrm{SR}}(b)$. Therefore, it is sufficient to show that $T_{\mathrm{SR}}(b) \in C(\gamma)$ if $b \ge \log \gamma$.

Noticing the martingale properties of the likelihood ratios, we have

$$\mathbb{E}_\infty \left\{ \exp(\lambda_{n,k,t}) | \mathcal{F}_{t-1} \right\} = 1 \tag{A.54}$$

for all $n = 1, 2, \ldots, N$, $t > 0$ and $0 \le k < t$. Moreover, noticing that

$$R_t = \sum_{k=1}^{t-2} \prod_{n=1}^{N} (1 - p_0 + p_0 \exp(\lambda_{n,k,t-1} + \lambda_{n,t-1,t})) + \prod_{n=1}^{N} (1 - p_0 + p_0 \exp(\lambda_{n,t-1,t})), \tag{A.55}$$

then combining (A.54) we have for all $t > 0$,

$$\begin{aligned} \mathbb{E}_\infty \left\{ R_t | \mathcal{F}_{t-1} \right\} =& \sum_{k=1}^{t-2} \prod_{n=1}^{N} (1 - p_0 + p_0 \exp(\lambda_{n,k,t-1}) \cdot 1) + 1 \\ =& R_{t-1} + 1. \end{aligned} \tag{A.56}$$

Therefore, the statistic $\{R_t - t\}_{t>0}$ is a $(P_\infty, \mathcal{F}_t)$-martingale with zero mean. If $\mathbb{E}_\infty \{T_{SR}(b)\} = \infty$ then the theorem is naturally correct, so we only suppose that $\mathbb{E}_\infty \{T_{\mathrm{SR}}(b)\} < \infty$ and thus $\mathbb{E}_\infty \left\{ R_{T_{\mathrm{SR}}(b)} - T_{\mathrm{SR}}(b) \right\}$ exists. Next, since $0 \le R_t < e^b$ on the event $\{T_{\mathrm{SR}}(b) > t\}$, we have

$$\liminf_{t\to\infty} \int_{\{T_{\mathrm{SR}}(b)>t\}} |R_t - t| \, d\mathbb{P}_\infty = 0.$$

Now we can apply the optional sampling theorem to have $\mathbb{E}_\infty \left\{ R_{T_{\mathrm{SR(b)}}} \right\} = \mathbb{E}_\infty \{T_{\mathrm{SR}}(b)\}$. By the definition of stopping time $T_{\mathrm{SR}}(b)$, we have $R_{T_{\mathrm{SR}}(b)} > e^b$. Thus, we have $\mathbb{E}_\infty \{T_{\mathrm{CS}}(b)\} \geq \mathbb{E}_\infty \{T_{\mathrm{SR}}(b)\} > e^b$, which shows that $\mathbb{E}_\infty \{T_{\mathrm{CS}}(b)\} > \gamma$ if $b \geq \log \gamma$.

□

Proof of Theorem 17. First, we notice that if $b \geq \log \gamma$

$$E_\infty \left\{ \widetilde{T}_{\mathrm{CS}}(b) \right\} \geq E_\infty \{T_{\mathrm{CS}}(b)\} \geq \gamma.$$

Therefore, by Theorem 16, it is sufficient to show that if $b \geq \log \gamma$ and $b = \mathcal{O}(\log \gamma)$, then

$$\mathrm{ESM}_m^{\mathcal{A}}(T_{\mathrm{CS}}(b)) \leq \left(\frac{\log \gamma}{I_{\mathcal{A}}} \right)^{m/q} (1 + o(1)) \text{ as } \gamma \to \infty. \tag{A.57}$$

Equivalently, it is sufficient to prove that

$$\mathrm{ESM}_m^{\mathcal{A}}(T_{\mathrm{CS}}(b)) \leq \left(\frac{b}{I_{\mathcal{A}}} \right)^{m/q} (1 + o(1)) \text{ as } b \to \infty. \tag{A.58}$$

To start with, we consider a special case when $p_0 = 1$ in T_{CS} and denote it by

$$T_{\mathrm{CS2}}(b) = \inf \left\{ t > 0 : \max_{0 \leq k < t} \sum_{n=1}^{N} \lambda_{n,k,t} \geq b \right\}.$$

Next, we will prove an asymptotical upper bound for the detection delay of $T_{CS2}(b)$.

Let

$$G_b = \left\lfloor \left(\frac{b}{I_{\mathcal{A}}(1-\varepsilon)} \right)^{1/q} \right\rfloor, \tag{A.59}$$

and then $(G_b)^q \leq b/[I_{\mathcal{A}}(1-\varepsilon)]$. Noticing that under $\mathbb{P}_k^{\mathcal{A}}$, we have $\sum_{n=1}^{N} \lambda_{n,k,t} = \lambda_{\mathcal{A},k,t}$ almost surely since the the log-likelihood ratios are 0 for the sensors that are not affected.

Therefore, by (4.21) we can have that for any $\varepsilon \in (0,1)$, $t \geq 0$ and some sufficiently large b,

$$\begin{aligned}
&\sup_{0\leq k<t} \operatorname{esssup} \mathbb{P}_k^{\mathcal{A}} \left\{ \sum_{n=1}^{N} \lambda_{n,k,k+G_b} < (G_b)^q I_{\mathcal{A}}(1-\varepsilon) \middle| \mathcal{F}_k \right\} \\
&\leq \sup_{0\leq k<t} \operatorname{esssup} \mathbb{P}_k^{\mathcal{A}} \left\{ \lambda_{\mathcal{A},k,k+G_b} < (G_b)^q I_{\mathcal{A}}(1-\varepsilon) | \mathcal{F}_k \right\} \\
&\leq \sup_{0\leq k<t} \operatorname{esssup} \mathbb{P}_k^{\mathcal{A}} \left\{ \lambda_{\mathcal{A},k,k+G_b} < b | \mathcal{F}_k \right\} \leq \varepsilon.
\end{aligned} \tag{A.60}$$

Then, for any $k \geq 0$ and integer $l \geq 1$, we can use (A.60) l times by conditioning on $\left(X_{n,1}, \ldots, X_{n,k+(l_0-1)G_b}\right)$, $n = 1, 2, \ldots, N$ for $l_0 = l, l-1, \ldots, 1$ in succession (see [54]) to have

$$\begin{aligned}
&\operatorname{esssup} \mathbb{P}_k^{\mathcal{A}} \left\{ T_{\mathrm{CS2}}(b) - k > lG_b | \mathcal{F}_k \right\} \\
&\leq \operatorname{esssup} \mathbb{P}_k^{\mathcal{A}} \left\{ \sum_{n=1}^{N} \lambda_{n,k+(l_0-1)G_b+1,k+l_0 G_b}, l_0 = 1, \ldots, l \middle| \mathcal{F}_k \right\} \leq \varepsilon^l.
\end{aligned} \tag{A.61}$$

Therefore, for sufficiently large b and any $\varepsilon \in (0,1)$, we have

$$\begin{aligned}
\mathrm{ESM}_m(T_{\mathrm{CS2}}(b)) &\leq \sum_{l=0}^{\infty} \left\{ [(l+1)G_b]^m - (lG_b)^m \right\} \cdot \\
&\qquad \sup_{0\leq k<\infty} \operatorname{esssup} \mathbb{P}_k \left\{ [(T_{\mathrm{CS2}} - k)^+]^m > (lG_b)^m \middle| \mathcal{F}_k \right\} \\
&\leq (G_b)^m \sum_{l=0}^{\infty} [(l+1)^m - l^m] \varepsilon^l \\
&= (G_b)^m (1 + o(1)) \text{ as } b \to \infty,
\end{aligned} \tag{A.62}$$

where the first inequality can be known directly from the geometric interpretation of expectation of discrete nonnegative random variables and the last equality holds since for any given $m \geq 1$, $[(l+1)^m - l^m]^{1/l} \to 1$ as $l \to \infty$ so that the radius of convergence is 1. Using the fact that $(G_b)^m \leq [b/I(1-\varepsilon)]^{m/q}$ we prove (A.58) for the case $p_0 = 1$.

Next, we will deal with the case when $p_0 \in (0, 1)$. Rewrite $T_{CS2}(b)$ as

$$T_{CS2}(b) = \inf\left\{t : \max_{0\leq k<t}\left(N\log p_0 + \sum_{n=1}^{N}\lambda_{n,t}^{k}\right) > b + N\log p_0\right\},$$

then

$$T_{CS2}(b - N\log p_0) = \inf\left\{t : \max_{0\leq k<t}\left(N\log p_0 + \sum_{n=1}^{N}\lambda_{n,t}^{k}\right) > b\right\}.$$

Clearly, $\mathrm{ESM}_m^{\mathcal{A}}(T_{\mathrm{CS}}(b)) \leq \mathrm{ESM}_m^{\mathcal{A}}(T_{CS2}(b - N\log p_0))$, and thus

$$\mathrm{ESM}_m^{\mathcal{A}}(T_{\mathrm{CS}}(b)) \leq \left(\frac{b - N\log p_0}{I_{\mathcal{A}}}\right)^{m/q}(1 + o(1)). \tag{A.63}$$

Therefore, we can claim that (A.58) holds for any fixed $p_0 \in (0, 1]$ since N and p_0 are constants. If $b \geq \log\gamma$ and $b = \mathcal{O}(\log\gamma)$, $T_{\mathrm{CS}}(b)$ belongs to $C(\gamma)$ and $\mathrm{ESM}_m^{\mathcal{A}}(T_{\mathrm{CS}})$ achieves its lower bound.

□

Proof of Corollary 2. The main steps are almost the same with that in the proof of Theorem 17. The only different thing is that we need the condition $w_\gamma \geq G_b$ (defined in (A.59)) in order to make (A.61) be correct for any $k \geq 0$ and any integer $l \geq 1$. And the additional assumption (4.22) ensures this. □

Proof of Lemma 9. Consider testing problem (4.1), then for any $k \geq 0$ and $j \geq 1$,

$$\lambda_{\mathcal{A},k,k+j} = \sum_{n\in\mathcal{A}}\frac{1}{\sigma_n^2}\sum_{i=k+1}^{k+j}\left\{c_n(i-k)(y_{n,i} - \mu_j) - \frac{c_n^2(i-k)^2}{2}\right\}.$$

We define, for each $n \in \mathcal{A}$ and for all $l = 1, \ldots, j$,

$$X_{n,l}^{(k)} = \frac{1}{\sigma_n^2}\left\{c_n l(y_{n,l+k} - \mu_j) - \frac{c_n^2 l^2}{2}\right\}.$$

Then we have

$$\lambda_{\mathcal{A},k,k+j} = \sum_{l=1}^{j} \sum_{n\in\mathcal{A}} X_{n,l}^{(k)} = \sum_{l=1}^{j} X_{\mathcal{A},l}^{(k)},$$

where we define $X_{\mathcal{A},l}^{(k)} = \sum_{n\in\mathcal{A}} X_{n,l}^{(k)}$.

Under probability measure $\mathbb{P}_k^{\mathcal{A}}$, we easily know that $(X_{\mathcal{A},l}^{(k)})_{l=1}^{j}$ are independent variables which follow normal distribution $N((l^2/2)\sum_{n\in\mathcal{A}} c_n^2, l^2 \sum_{n\in\mathcal{A}} c_n^2)$. Other simple computation tells us that

$$\mathbb{E}_k^{\mathcal{A}} \left\{ (X_{\mathcal{A},l}^{(k)})^2 \right\} < \infty, \ \forall l = 1, \ldots, j,$$

and under probability measure $\mathbb{P}_k^{\mathcal{A}}$,

$$\sum_{l=1}^{\infty} \text{Var}\left(\frac{X_{\mathcal{A},l}^{(k)}}{l^3} \right) < \infty,$$

where Var(X) denotes the variance of random variable X. Therefore, combining Kroneckers lemma with the Kolmogorov convergence criteria, we have immediately a strong law of large numbers which tells us that

$$\frac{1}{j^3}\lambda_{\mathcal{A},k,k+j} \xrightarrow[j\to\infty]{a.s.} \sum_{n\in\mathcal{A}} \frac{c_n^2}{6\sigma_n^2}.$$

Finally, we complete the proof by using the fact that all the observations are independent.

□

Proof of Lemma 10. First, define $y_{n,t}^k = \frac{\sum_{i=k+1}^{t}(y_{n,i}-\mu_j)}{\sigma_j \sum_{i=k+1}^{t}(i-k)^2}$, then

$$\begin{aligned} \mathbb{P}_\infty \left\{ \widetilde{T}_2(b) > t_0 \right\} &\geq \mathbb{P}_\infty \left\{ \max_{0<t\leq t_0} \max_{\max(0,t-m_\gamma)\leq k<t_0} \sum_{n=1}^{N} \frac{(y_{n,t_0}^k)^2}{2} < b \right\} \\ &\geq \left[\mathbb{P}_\infty\{Y < 2b\} \right]^{w_\gamma t_0}, \end{aligned} \tag{A.64}$$

where Y is a random variable with χ_N^2 distribution. Then, since $\widetilde{T}_2(b)$ is a non-negative discrete random variable, we have

$$\begin{aligned}\mathbb{E}_\infty\left\{\widetilde{T}_2(b)\right\} &= \sum_{t_0=0}^{\infty} \mathbb{P}_\infty\left\{\widetilde{T}_2(b) > t_0\right\} \\ &\geq \sum_{t_0=0}^{\infty} [\mathbb{P}_\infty\{Y < 2b\}]^{w_\gamma t_0} = \frac{1}{1 - [\mathbb{P}_\infty\{Y < 2b\}]^{w_\gamma}}.\end{aligned} \tag{A.65}$$

Then if we can choose some b so that

$$\mathbb{P}_\infty\{Y \geq 2b\} \leq 1 - \left(1 - \frac{1}{\gamma}\right)^{1/m_\gamma},$$

we can claim that $\mathbb{E}_\infty\left\{\widetilde{T}_2(b)\right\} \geq \gamma$ and thus $\widetilde{T}_2(b) \in C(\gamma)$. To choose appropriate threshold b, we need use the tail bound for the χ^2_N distribution. Since χ^2_1 is sub-exponential with parameter $(2\sqrt{N}, 4)$, it is well known that $\mathbb{P}_\infty\{Y \geq 2b\} \leq \exp(-\frac{2b-N}{8})$ if $b \geq N$. If we set

$$b \geq \frac{N}{2} - 4\log\left[1 - \left(1 - \frac{1}{\gamma}\right)^{1/m_\gamma}\right]$$

then $\widetilde{T}_1(b) \in C(\gamma)$. □

Proof of Theorem 18. By Lemma 9, we can use Theorem 16 to obtain a lower bound for the detection delays of arbitrary procedures in $C(\gamma)$. Specifically, for all $m \geq 1$,

$$\liminf_{\gamma\to\infty} \inf_{T\in C(\gamma)} \mathrm{ESM}_m^{\mathcal{A}}(T) \geq \liminf_{\gamma\to\infty} \inf_{T\in C(\gamma)} \mathrm{SM}_m^{\mathcal{A}}(T) \geq \left(\frac{\log\gamma}{I_{\mathcal{A}}}\right)^{m/q}. \tag{A.66}$$

(i) Since $T_1(b)$ is a specified mixture CUSUM procedure for testing problem (4.1) and the observations are independent, the optimality is an immediate corollary from Theorem 17.

(ii) Since $\widetilde{T}_1(b)$ is a specified window-limited mixture CUSUM procedure for testing problem (4.1) and the observations are independent, the optimality is an immediate result from Corollary 2.

(iii) The assumption that $\log w_\gamma = o(\log\gamma)$ ensures that $b \geq \frac{N}{2} - 4\log\left[1 - \left(1 - \frac{1}{\gamma}\right)^{1/m_\gamma}\right]$ and $b = \mathcal{O}(\log\gamma)$ can be satisfied simultaneously. Since the observations are independent, then $\mathrm{ESM}_1^{\mathcal{A}}(\widetilde{T}_2(b)) = \mathrm{SM}_1^{\mathcal{A}}(\widetilde{T}_2(b)) = \mathbb{E}_0^{\mathcal{A}}[\widetilde{T}_2(b)]$. The optimality of $\widetilde{T}_2(b)$ is an immediate result from Lemma 10 and the first order approximation of the detection delays in (A.44).

□

A.3 Proofs for sequential change detection with offline convex optimization

In the following, we denote $\mathbb{E}_{\xi\sim\nu}[f(\xi)]$ as the expected value of $f(\xi)$ when ξ follows some distribution ν.

Proof of Theorem 12. Define that $\phi^* \triangleq -\frac{1}{2}L^*$. From Theorem 2.1 in [72], we have that

$$\mathbb{E}_{\xi\sim\nu_0}[\exp(-\phi^*(\xi))] \leq \epsilon^*, \ \ \forall \nu_0 \in \mathcal{P}_0, \tag{A.67}$$

$$\mathbb{E}_{\xi\sim\nu_1}[\exp(\phi^*(\xi))] \leq \epsilon^*, \ \ \forall \nu_1 \in \mathcal{P}_1, \tag{A.68}$$

where ϵ^* is the solution to the equation

$$\mathbb{E}_{\xi\sim\nu_0^*}[\exp(-\phi^*(\xi))] = \mathbb{E}_{\xi\sim\nu_1^*}[\exp(\phi^*(\xi))],$$

or equivalently, it is defined in (5.9).

Define a stopping time $T = \inf\{t > 0 : \sum_{i=1}^{t} -\phi^*(\xi_t) > b\}$, then T_1 in (5.7) is the same procedure as T and the arguments about T are also true for T_1. Following the definition of

T, for any $m > 0$, we have that

$$\begin{aligned}\mathbb{P}_{\infty}^{\nu_0}(T \leq m) \leq & \mathbb{P}_{\infty}^{\nu_0}\left(\bigcup_{k=1}^{m}\left\{\sum_{i=1}^{k} -\phi^*(\xi_i) > b\right\}\right) \\ \leq & \sum_{k=1}^{m} \mathbb{P}_{\infty}^{\nu_0}\left(\sum_{i=1}^{k} -\phi^*(\xi_i) > b\right) \\ = & \sum_{k=1}^{m} \mathbb{P}_{\infty}^{\nu_0}\left(\sum_{i=1}^{k}\left(-\phi^*(\xi_i) - \frac{b}{k}\right) > 0\right).\end{aligned} \tag{A.69}$$

Fix m and k, we define that $\widetilde{\phi}^* = \phi^* + b/k$ and then we use Chernoff inequality and inequality (A.67) to obtain that

$$\begin{aligned}\mathbb{P}_{\xi\sim\nu}(-\widetilde{\phi}^*(\xi) > 0) \leq & \frac{\mathbb{E}_{\xi\sim\nu}[\exp(-\widetilde{\phi}^*(\xi))]}{1} \\ \leq & \exp(-\frac{b}{k})\epsilon^*,\ \forall \nu \in \mathcal{P}_0.\end{aligned} \tag{A.70}$$

Under $H_0, \xi_i \sim \nu_0 \in \mathcal{P}_0, i = 1, \ldots, m$ and ξ_is are independent. If we apply the shifted detector $\widetilde{\phi}^*$ on the independent variables $\xi_1, \xi_2, \ldots, \xi_k$, from the result for k-repeated observations (Section 2.4 in [72]) , we can have that

$$\mathbb{P}_{\infty}^{\nu_0}\left(\sum_{i=1}^{k}\left(-\phi^*(\xi_i) - \frac{b}{k}\right) > 0\right) \leq \left(\exp\left(-\frac{b}{k}\right)\epsilon^*\right)^k.$$

Then, we have that

$$\begin{aligned}\mathbb{P}_{\infty}^{\nu_0}(T \leq m) \leq & \sum_{k=1}^{m}\left(\exp\left(-\frac{b}{k}\right)\epsilon^*\right)^k \\ = & \sum_{k=1}^{m} \exp\left(-b\right)\left(\epsilon^*\right)^k, \\ = & \exp(-b) \cdot \frac{\epsilon^* - (\epsilon^*)^{m+1}}{1 - \epsilon^*}.\end{aligned} \tag{A.71}$$

Letting m go to infinity, we have that

$$\mathbb{P}_{\infty}^{\nu_0}(T < \infty) = \exp(-b) \cdot \frac{\epsilon^*}{(1-\epsilon^*)}.$$

Applying Theorem 2 in [59], we have that

$$\mathbb{E}_{\infty}^{\nu_0}(T) \geq \frac{1}{\mathbb{P}_{\infty}^{\nu_0}(T < \infty)} = \exp(b) \cdot \frac{1-\epsilon^*}{\epsilon^*},$$

which concludes our result. $\square$

Proof of Theorem 20. Similar with the proof for Theorem 12, we define that $\phi^* = -\frac{1}{2}L^*$, $S_t = \sum_{i=1}^{t} -\phi^*(\xi_t)$ and a stopping time $T = \inf\{t > 0 : S_t > b\}$. Then T is the same as T_1. Noticing that under $\mathbb{P}_0^{\nu_0,\nu_1}$, $\xi_1, \xi_2, \dots$ is a sequence of i.i.d random variables following some distribution $\nu_1 \in \mathcal{P}_1$, the well known Wald's equality (e.g, [131]) shows that

$$\mathbb{E}_0^{\nu_0,\nu_1}[T] = \frac{\mathbb{E}_0^{\nu_0,\nu_1}[S_T]}{\mathbb{E}_{\xi_1\sim\nu_1}[-\phi_*(\xi_1)]} = \frac{b + \mathbb{E}_0^{\nu_0,\nu_1}[S_T - b]}{\mathbb{E}_{\xi_1\sim\nu_1}[-\phi_*(\xi_1)]},$$

where $\mathbb{E}_0^{\nu_0,\nu_1}[S_T - b]$ is the expected overshoot above the decision boundary.

Combining (A.68) and the fact that for any $x \in \mathbb{R}$, $-x \geq 1 - \exp(x)$, we have that

$$\mathbb{E}_{\xi_1\sim\nu_1}[-\phi_*(\xi_1)] \geq 1 - \mathbb{E}_{\xi_1\sim\nu_1}[\exp(\phi_*(\xi_1))] \geq 1 - \epsilon^*.$$

To estimate the overshoot, we apply (8.18) and (8.50) in [131] to show that as $b \to \infty$, the following limit holds,

$$\mathbb{E}_0^{\nu_0,\nu_1}[S_T - b] \to \frac{\mathbb{E}_{\xi_1\sim\nu_1}[\phi^*(\xi_1)^2]}{2\mathbb{E}_{\xi_1\sim\nu_1}[\phi^*(\xi_1)]} - \sum_{n=1}^{\infty} \frac{\mathbb{E}_0^{\nu_0,\nu_1}[S_n^-]}{n},$$

where $x^- = -\min(x, 0)$.

By the assumption made in the statement, we have that for some $M > 0$, $\mathbb{E}_{\xi_1\sim\nu_1}[\phi_*^2(\xi_1)] \leq$

M. Therefore, as $b \to \infty$, we have that $\mathbb{E}_0^{\nu_0,\nu_1}[S_T - b] = o(b)$. Combing the Theorem 2 in [59], we conclude the result.

□

Proof of Corollary 3 and 4. When ϕ^* is obtained from (5.16), from the Proposition 4.1 in [73], we have that

$$\mathbb{E}_{\xi\sim\nu_0}[\exp(-\phi^*(\xi))] \le \epsilon^*, \ \ \forall \nu_0 \in \mathcal{G}_0, \tag{A.72}$$

$$\mathbb{E}_{\xi\sim\nu_1}[\exp(\phi^*(\xi))] \le \epsilon^*, \ \ \forall \nu_1 \in \mathcal{G}_1, \tag{A.73}$$

where ϵ^* is defined in (5.18). Then, following the same proof routine as Theorem 12 and 20, we conclude the results.

□

A.4 Proofs for Sequential change detection via online convex optimization

Proof of Theorem 21. In the proof, for the simplicity of notation we use N to denote $\tau(b)$. Recall θ is the true parameter. Define that

$$S_t^\theta = \sum_{i=1}^{t} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}.$$

Then under the measure $\mathbb{P}_{\theta,0}$, S_t is a random walk with i.i.d. increment. Then, by Wald's identity (e.g., [131]) we have that

$$\mathbb{E}_{\theta,0}[S_N^\theta] = \mathbb{E}_{\theta,0}[N] \cdot I(\theta, \theta_0). \tag{A.74}$$

On the other hand, let θ_N^* denote the MLE based on $(X_1, \ldots, X_N)$. The key to the proof is to decompose the stopped process S_N^θ as a summation of three terms as follows:

$$S_N^\theta = \sum_{i=1}^N \log \frac{f_\theta(X_i)}{f_{\theta_N^*}(X_i)} + \sum_{i=1}^N \log \frac{f_{\theta_N^*}(X_i)}{f_{\hat{\theta}_{i-1}}(X_i)} + \sum_{i=1}^N \log \frac{f_{\hat{\theta}_{i-1}}(X_i)}{f_{\theta_0}(X_i)}, \tag{A.75}$$

Note that the first term of the decomposition on the right-hand side of (A.75) is always non-positive since

$$\sum_{i=1}^N \log \frac{f_\theta(X_i)}{f_{\theta_N^*}(X_i)} = \sum_{i=1}^N \log f_\theta(X_i) - \sup_{\tilde{\theta}\in\Theta} \sum_{i=1}^N \log f_{\tilde{\theta}}(X_i) \le 0.$$

Therefore we have

$$\mathbb{E}_{\theta,0}[S_N^\theta] \le \mathbb{E}_{\theta,0}\left[\sum_{i=1}^N \log \frac{f_{\theta_N^*}(X_i)}{f_{\hat{\theta}_{i-1}}(X_i)}\right] + \mathbb{E}_{\theta,0}\left[\sum_{i=1}^N \log \frac{f_{\hat{\theta}_{i-1}}(X_i)}{f_{\theta_0}(X_i)}\right].$$

Now consider the third term in the decomposition (A.75). Similar to the proof of equation (5.109) in [55], we claim that its expectation under measure $\mathbb{P}_{\theta,0}$ is upper bounded by $b/I(\theta,\theta_0) + O(1)$ as $b \to \infty$. Next, we prove the claim. For any positive integer n, we further decompose the third term in (A.75) as

$$\sum_{i=1}^n \log \frac{f_{\hat{\theta}_{i-1}}(X_i)}{f_{\theta_0}(X_i)} = M_n(\theta) - G_n(\theta) + m_n(\theta,\theta_0) + nI(\theta,\theta_0), \tag{A.76}$$

where

$$M_n(\theta) = \sum_{i=1}^n \log \frac{f_{\hat{\theta}_{i-1}}(X_i)}{f_\theta(X_i)} + G_n(\theta),$$

$$G_n(\theta) = \sum_{i=1}^n I(\theta, \hat{\theta}_{i-1}),$$

and

$$m_n(\theta,\theta_0) = \sum_{i=1}^n \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} - nI(\theta,\theta_0).$$

The decomposition of (A.76) consists of stochastic processes $\{M_n(\theta)\}$ and $\{m_n(\theta,\theta_0)\}$, which are both $\mathbb{P}_{\theta,0}$-martingales with zero expectation, i.e., $\mathbb{E}_{\theta,0}[M_n(\theta)] = \mathbb{E}_{\theta,0}[m_n(\theta,\theta_0)] = 0$ for any positive integer n. Since for exponential family, the log-partition function $\Phi(\theta)$ is bounded, by the inequalities for martingales [132] we have that

$$\mathbb{E}_{\theta,0}|M_n(\theta)| \leq C_1\sqrt{n}, \quad \mathbb{E}_{\theta,0}|m_n(\theta,\theta_0)| \leq C_2\sqrt{n}, \tag{A.77}$$

where C_1 and C_2 are two absolute constants that do not depend on n. Moreover, we observe that for all $\theta \in \Theta$,

$$\mathbb{E}_{\theta,0}[G_n(\theta)] \leq \mathbb{E}_{\theta,0}\left[\max_{\tilde{\theta}\in\Theta} G_n(\tilde{\theta})\right] = \mathbb{E}_{\theta,0}[\mathcal{R}_n(\theta)] \leq C\log n.$$

Therefore, applying (A.77), we have that $n^{-1}G_n(\theta), n^{-1}M_n(\theta)$ and $n^{-1}m_n(\theta,\theta_0)$ converge to 0 almost surely. Moreover, the convergence is $\mathbb{P}_{\theta,0}$-r-quickly for $r = 1$. We say that $n^{-1}A_n$ converges $\mathbb{P}_{\theta,0}$-r-quickly to a constant I if $\mathbb{E}_{\theta,0}[\mathcal{G}(\epsilon)]^r < \infty$ for all $\epsilon > 0$, where $\mathcal{G}(\epsilon) = \sup\{n \geq 1 : |n^{-1}A_n - I| > \epsilon\}$ is the last time when $n^{-1}A_n$ leaves the interval $[I-\epsilon, I+\epsilon]$ (for more details, we refer the readers to Section 2.4.3 of [55]). Therefore, dividing both sides of (A.76) by n, we obtain $n^{-1}\sum_{i=1}^n \log(f_{\hat{\theta}_{i-1}}(X_i)/f_{\theta_0}(X_i))$ converges $\mathbb{P}_{\theta,0}$-1-quickly to $I(\theta,\theta_0)$.

For $\epsilon > 0$, we now define the last entry time

$$L(\epsilon) = \sup\left\{n \geq 1 : \left|\frac{1}{I(\theta,\theta_0)}\sum_{i=1}^n \log\frac{f_{\hat{\theta}_{i-1}}(X_i)}{f_{\theta_0}(X_i)} - n\right| > \epsilon n\right\}.$$

By the definition of $\mathbb{P}_{\theta,0}$-1-quickly convergence and the finiteness of $I(\theta,\theta_0)$, we have that $\mathbb{E}_{\theta,0}[L(\epsilon)] < +\infty$ for all $\epsilon > 0$. In the following, define a scaled threshold $\tilde{b} = b/I(\theta,\theta_0)$. Observe that conditioning on the event $\{L(\epsilon)+1 < N < +\infty\}$, we have that

$$(1-\epsilon)(N-1)I(\theta,\theta_0) < \sum_{i=1}^{N-1}\log\frac{f_{\hat{\theta}_{i-1}}(X_i)}{f_{\theta_0}(X_i)} < b.$$

Therefore, conditioning on the event $\{L(\epsilon)+1 < N < +\infty\}$, we have that $N < 1+b/(1-\epsilon)$. Hence, for any $0 < \epsilon < 1$, we have

$$N \leq 1+\mathbb{I}(\{N > L(\epsilon)+1\})\cdot\frac{\tilde{b}}{1-\epsilon}+\mathbb{I}(\{N \leq L(\epsilon)+1\})\cdot L(\epsilon) \leq 1+\frac{\tilde{b}}{1-\epsilon}+L(\epsilon). \quad \text{(A.78)}$$

Since $\mathbb{E}_{\theta,0}[L(\epsilon)] < \infty$ for any $\epsilon > 0$, from (A.78) above, we have that the third term in (A.75) is upper bounded by $\tilde{b} + \mathsf{O}(1)$.

Finally, the second term in (A.75) can be written as

$$\sum_{i=1}^{N} \log \frac{f_{\theta_N^*}(X_i)}{f_{\hat{\theta}_{i-1}}(X_i)} = \sum_{i=1}^{N} -\log f_{\hat{\theta}_{i-1}}(X_i) - \inf_{\tilde{\theta}\in\Theta} \sum_{i=1}^{N} -\log f_{\tilde{\theta}}(X_i),$$

which is just the regret defined in (6.12) for the online estimators: $\mathcal{R}_t$, when the loss function is defined to be the negative likelihood function. Then, the theorem is proven by combining the above analysis for the three terms in (A.75) and (A.74). □

Proof of Corollary 5. First, we can relate the expected regret at the stopping time to the expected stopping time, using the following chain of equalities and inequalities

$$\mathbb{E}_{\theta,0}[\mathcal{R}_{\tau(b)}] = \mathbb{E}_{\theta,0}[\mathbb{E}_{\theta,0}[\mathcal{R}_n \mid \tau(b) = n]] \leq \mathbb{E}_{\theta,0}[C \log \tau(b)] \leq C \log \mathbb{E}_{\theta,0}[\tau(b)], \quad \text{(A.79)}$$

where the first equality uses iterative expectation, the first inequality uses the assumption of the logarithmic regret in the statement of Corollary 5, and the second inequality is due to Jensen's inequality. Let $\alpha = (b + O(1))/I(\theta, \theta_0)$, $\beta = C/I(\theta, \theta_0)$ and $x = \mathbb{E}_{\theta,0}[\tau(b)]$. Applying (A.79), the upper bound in equation (6.14) becomes $x \leq \alpha + \beta \log(x)$. From this, we have $x \leq O(\alpha)$. Taking logarithm on both sides and using the fact that $\max\{a_1 + a_2\} \leq a_1+a_2 \leq 2\max\{a_1, a_2\}$ for $a_1, a_2 \geq 0$, $\log(x) \leq \max\{\log(2\alpha), \log(2\beta \log x)\} \leq \log(\alpha)+o(\log b)$. Therefore, we have that $x \leq \alpha + \beta(\log(\alpha) + o(\log b))$. Using this argument, we obtain

$$\mathbb{E}_{\theta,0}[\tau(b)] \leq \frac{b}{I(\theta,\theta_0)} + \frac{C \log b}{I(\theta,\theta_0)}(1 + o(1)). \quad \text{(A.80)}$$

Note that a similar argument can be found in [128]. □

Next we will establish a few Lemmas useful for proving theorem 22 for sequential detection procedures. Define a measure $\mathbb{Q}$ on $(\mathcal{X}^\infty, \mathcal{B}^\infty)$ under which the probability density of X_i conditional on $\mathcal{F}_{i-1}$ is $f_{\hat{\theta}_{i-1}}$. Then for any event $A \in \mathcal{F}_i$, we have that $\mathbb{Q}(A) = \int_A \Lambda_i d\mathbb{P}_\infty$. The following lemma shows that the restriction of $\mathbb{Q}$ to $\mathcal{F}_i$ is well defined.

Lemma 19. *Let $\mathbb{Q}_i$ be the restriction of $\mathbb{Q}$ to $\mathcal{F}_i$. Then for any $A \in \mathcal{F}_k$ and any $i \geq k$, $\mathbb{Q}_i(A) = \mathbb{Q}_k(A)$.*

Proof of Lemma 11. To bound the term $\mathbb{P}_\infty(\tau(b) < \infty)$, we need take advantage of the martingale property of Λ_t in (6.2). The major technique is the combination of change of measure and Wald's likelihood ratio identity [131]. The proofs are a combination of the results in [125] and [75] and the reader can find a complete proof in [125]. For purpose of completeness we copy those proofs here.

Define the $L_i = d\mathbb{P}_i/d\mathbb{Q}_i$ as the Radon-Nikodym derivative, where $\mathbb{P}_i$ and $\mathbb{Q}_i$ are the restriction of $\mathbb{P}_\infty$ and $\mathbb{Q}$ to $\mathcal{F}_i$, respectively. Then we have that $L_i = (\Lambda_i)^{-1}$ for any $i \geq 1$ (note that Λ_i is defined in (6.2)). Combining the Lemma 19 and the Wald's likelihood ratio identity, we have that

$$\mathbb{P}_\infty(A \cap \{\tau(b) < \infty\}) = \mathbb{E}_Q\left[\mathbb{I}(\{\tau(b) < \infty\}) \cdot L_{\tau(b)}\right], \forall A \in \mathcal{F}_{\tau(b)}, \tag{A.81}$$

where $\mathbb{I}(E)$ is an indicator function that is equal to 1 for any $\omega \in E$ and is equal to 0 otherwise. By the definition of $\tau(b)$ we have that $L_{\tau(b)} \leq \exp(-b)$. Taking $A = \mathcal{X}^\infty$ in (A.81) we prove that $\mathbb{P}_\infty(\tau(b) < \infty) \leq \exp(-b)$. □

Proof of Corollary 6. Using (5.180) and (5.188) in [55], which are about asymptotic performance of open-ended tests. Since our problem is a special case of the problem in [55], we can obtain

$$\inf_{T \in C(\alpha)} \mathbb{E}_{\theta,0}[T] = \frac{\log \alpha}{I(\theta,\theta_0)} + \frac{\log(\log(1/\alpha))}{2I(\theta,\theta_0)}(1+o(1)).$$

Combing the above result and the right-hand side of (6.15), we prove the corollary. □

Proof of Theorem 22. From (A.83), we have that for any $\nu \geq 1$,

$$\mathbb{E}_{\theta,\nu}[T_{ASR}(b) - \nu \mid T_{ASR}(b) > \nu] \leq \mathbb{E}_{\theta,\nu}[T_{ACM}(b) - \nu \mid T_{ACM}(b) > \nu].$$

Therefore, to prove the theorem using Theorem 21, it suffices to show that

$$\sup_{\nu \geq 0} \mathbb{E}_{\theta,\nu}[T_{ACM}(b) - \nu \mid T_{ACM}(b) > \nu] \leq \mathbb{E}_{\theta,0}[\tau(b)].$$

Using an argument similar to the remarks in [75], we have that the supreme of detection delay over all change locations is achieved by the case when change occurs at the first instance,

$$\sup_{\nu \geq 0} \mathbb{E}_{\theta,\nu}[T_{ACM}(b) - \nu \mid T_{ACM}(b) > \nu] = \mathbb{E}_{\theta,0}[T_{ACM}(b)]. \tag{A.82}$$

This is a slight modification (a small change on the subscripts) of the remarks in [75] but for the purpose of completeness and clearness we write the details in the following. Notice that since θ_0 is known, for any $j \geq 1$, the distribution of $\{\max_{j+1 \leq k \leq t} \Lambda_{k,t}\}_{t=j+1}^{\infty}$ under $\mathbb{P}_{\theta,j}$ conditional on $\mathcal{F}_j$ is the same as the distribution of $\{\max_{1 \leq k \leq t} \Lambda_{k,t}\}_{t=1}^{\infty}$ under $\mathbb{P}_{\theta,0}$. Below, we use a renewal property of the ACM procedure. Define

$$T_{ACM}^{(j)}(b) = \inf\{t > j : \max_{j+1 \leq k \leq t} \log \Lambda_{k,t} > b\}.$$

Then we have that $\mathbb{E}_{\theta,0}[T_{ACM}(b)] = \mathbb{E}_{\theta,j}[T_{ACM}^{(j)}(b) - j \mid T_{ACM}^{(j)}(b) > j]$. However, $\max_{1 \leq k \leq t} \log \Lambda_{k,t} \geq \max_{j+1 \leq k \leq t} \Lambda_{k,t}$ for any $t > j$. Therefore, $T_{ACM}^{(j)}(b) \geq T_{ACM}(b)$

conditioning on $\{T_{ACM}(b) > j\}$. So that for all $j \geq 1$,

$$\mathbb{E}_{\theta,0}[T_{ACM}(b)] = \mathbb{E}_{\theta,j}[T_{ACM}^{(j)}(b) - j \mid T_{ACM}(b) > j] \geq \mathbb{E}_{\theta,j}[T_{ACM}(b) - j \mid T_{ACM}(b) > j].$$

Thus, to prove (A.82), it suffices to show that $\mathbb{E}_{\theta,0}[T_{ACM}(b)] \leq \mathbb{E}_{\theta,0}[\tau(b)]$. To show this, define $\tau(b)^{(t)}$ as the new stopping time that applies the one-sided sequential hypothesis testing procedure $\tau(b)$ to data $\{X_i\}_{i=t}^{\infty}$. Then we have that in fact $T_{ACM}(b) = \min_{t \geq 1}\{\tau(b)^{(t)} + t - 1\}$, this relationship was developed in [59]. Thus, $T_{ACM}(b) \leq \tau(b)^{(1)} + 1 - 1 = \tau(b)$, and $\mathbb{E}_{\theta,0}[T_{ACM}(b)] \leq \mathbb{E}_{\theta,0}[\tau(b)]$. □

Proof of Lemma 12. This is a classic result proved by using the martingale property and the proof routine can be found in many textbooks such as [55]. First, rewrite $T_{ASR}(b)$ as

$$T_{ASR}(b) = \inf\left\{t \geq 1 : \log\left(\sum_{k=1}^{t} \Lambda_{k,t}\right) > b\right\}.$$

Next, since

$$\log\left(\sum_{k=1}^{t} \Lambda_{k,t}\right) > \log\left(\max_{1 \leq k \leq t} \Lambda_{k,t}\right) = \max_{1 \leq k \leq t} \log \Lambda_{k,t}, \tag{A.83}$$

we have $\mathbb{E}_\infty[T_{ACM}(b)] \geq \mathbb{E}_\infty[T_{ASR}(b)]$. So it suffices to show that $\mathbb{E}_\infty[T_{ASR}(b)] \geq \gamma$, if $b \geq \log \gamma$. Define $R_t = \sum_{k=1}^{t} \Lambda_{k,t}$. Direct computation shows that

$$\begin{aligned}
\mathbb{E}_\infty[R_t \mid \mathcal{F}_{t-1}] =& \mathbb{E}_\infty\left[\Lambda_{t,t} + \sum_{k=1}^{t-1} \Lambda_{k,t} \mid \mathcal{F}_{t-1}\right] \\
=& \mathbb{E}_\infty\left[1 + \sum_{k=1}^{t-1} \Lambda_{k,t-1} \cdot \log \frac{f_{\hat{\theta}_{t-1}}(X_t)}{f_{\theta_0}(X_t)} \mid \mathcal{F}_{t-1}\right] \\
=& 1 + \sum_{k=1}^{t-1} \Lambda_{k,t-1} \cdot \mathbb{E}_\infty\left[\log \frac{f_{\hat{\theta}_{t-1}}(X_t)}{f_{\theta_0}(X_t)} \mid \mathcal{F}_{t-1}\right] \\
=& 1 + R_{t-1}.
\end{aligned}$$

Therefore, $\{R_t - t\}_{t \geq 1}$ is a $(\mathbb{P}_\infty, \mathcal{F}_t)$-martingale with zero mean. Suppose that $\mathbb{E}_\infty[T_{ASR}(b)] <$

∞ (otherwise the statement of proposition is trivial), then we have that

$$\sum_{t=1}^{\infty} \mathbb{P}_{\infty}(T_{ASR}(b) \geq t) < \infty. \tag{A.84}$$

(A.84) leads to the fact that $\mathbb{P}_{\infty}(T_{ASR}(b)) \geq t = o(t^{-1})$ and the fact that $0 \leq R_t \leq \exp(b)$ conditioning on the event $\{T_{ASR}(b) > t\}$, we have that

$$\liminf_{t\to\infty} \int_{\{T_{ASR}(b)>t\}} |R_t - t| d\mathbb{P}_{\infty} \leq \liminf_{t\to\infty} \ (\exp(b) + t)\mathbb{P}_{\infty}(T_{ASR}(b) \geq t) = 0.$$

Therefore, we can apply the optional stopping theorem for martingales, to obtain that $\mathbb{E}_{\infty}[R_{T_{ASR}(b)}] = \mathbb{E}_{\infty}[T_{ASR}(b)]$. By the definition of $T_{ASR}(b)$, $R_{T_{ASR}(b)} > \exp(b)$ we have that $\mathbb{E}_{\infty}[T_{ASR}(b)] > \exp(b)$. Therefore, if $b \geq \log\gamma$, we have that $\mathbb{E}_{\infty}[T_{ACM}(b)] \geq \mathbb{E}_{\infty}[T_{ASR}(b)] \geq \gamma$. □

Proof of Corollary 7. Our Theorem 1 and the remarks in [82] show that the minimum worst-case detection delay, given a fixed ARL level γ, is given by

$$\inf_{T(b)\in S(\gamma)} \sup_{\nu\geq 1} \mathbb{E}_{\theta,\nu}[T(b) - \nu + 1 \mid T(b) \geq \nu] = \frac{\log\gamma}{I(\theta,\theta_0)} + \frac{d\log\log\gamma}{2I(\theta,\theta_0)}(1 + o(1)). \tag{A.85}$$

It can be shown that the infimum is attained by choosing $T(b)$ as a weighted Shiryayev-Roberts detection procedure, with a careful choice of the weight over the parameter space Θ. Combing (A.85) with the right-hand side of (6.15), we prove the corollary. □

The following derivation borrows ideas from [83]. First, we derive concise forms of the two terms in the definition of R_t in (6.12).

Lemma 20. *Assume that $X_1, X_2, \ldots$ are i.i.d. random variables with density function $f_\theta(x)$, and assume decreasing step-size $\eta_i = 1/i$ in Algorithm 2. Given $\{\hat{\theta}_i\}_{i\geq 1}, \{\hat{\mu}_i\}_{i\geq 1}$ generated by Algorithm 2. If $\hat{\theta}_i = \tilde{\theta}_i$ for any $i \geq 1$, then for any null distribution parameter $\theta_0 \in \Theta$*

and $t \geq 1$,

$$\sum_{i=1}^{t}\{-\log f_{\hat{\theta}_{i-1}}(X_i)\} = \sum_{i=1}^{t} iB_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) - t\Phi^*(\hat{\mu}_t). \tag{A.86}$$

Moreover, for any $t \geq 1$,

$$\inf_{\tilde{\theta}\in\Theta}\sum_{i=1}^{t}\{-\log f_{\tilde{\theta}}(X_i)\} = -t\Phi^*(\hat{\mu}), \tag{A.87}$$

where $\hat{\mu} = (1/t) \cdot \sum_{i=1}^{t} \phi(X_i)$.

By subtracting the expressions in (A.86) and (A.87), we obtain the following result which shows that the regret can be represented by a weighted sum of the Bregman divergences between two consecutive estimators.

Proof of Lemma 20. By the definition of the Legendre-Fenchel dual function we have that $\Phi^*(\mu) = \theta^\intercal\mu - \Phi(\theta)$ for any $\theta \in \Theta$. By this definition, and choosing $\eta_i = 1/i$, we have that for any $i \geq 1$

$$\begin{aligned}
&-\log f_{\hat{\theta}_{i-1}}(X_i)\\
&= \Phi(\hat{\theta}_{i-1}) - \hat{\theta}_{i-1}^\intercal\phi(X_i)\\
&= \hat{\theta}_{i-1}^\intercal(\hat{\mu}_{t-1} - \phi(X_i)) - \Phi^*(\hat{\mu}_{i-1}) = \frac{1}{\eta_i}\hat{\theta}_{i-1}^\intercal(\hat{\mu}_{i-1} - \hat{\mu}_i) - \Phi^*(\hat{\mu}_{i-1})\\
&= \frac{1}{\eta_i}(\Phi^*(\hat{\mu}_i) - \Phi^*(\hat{\mu}_{i-1})) - \hat{\theta}_{i-1}^\intercal(\hat{\mu}_i - \hat{\mu}_{i-1}) - \frac{1}{\eta_i}\Phi^*(\hat{\mu}_i) + \left(\frac{1}{\eta_i} - 1\right)\Phi^*(\hat{\mu}_{i-1})\\
&= \frac{1}{\eta_i}B_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) + \frac{1}{\eta_{i-1}}\Phi^*(\hat{\mu}_{i-1}) - \frac{1}{\eta_i}\Phi^*(\hat{\mu}_i),
\end{aligned} \tag{A.88}$$

where we use the update rule in Line 6 of Algorithm 2 and the assumption $\hat{\theta}_i = \tilde{\theta}_i$ to have the third equation. We define $1/\eta_0 = 0$ in the last equation. Now summing the terms in

(A.88), where the second term form a telescopic series, over i from 1 to t, we have that

$$\begin{aligned}\sum_{i=1}^{t}\{-\log f_{\hat{\theta}_{i-1}}(X_i)\} =& \sum_{i=1}^{t}\frac{1}{\eta_i}B_{\Phi^*}(\hat{\mu}_i,\hat{\mu}_{i-1}) + \frac{1}{\eta_0}\Phi^*(\hat{\mu}_0) - \frac{1}{\eta_t}\Phi^*(\hat{\mu}_t)\\ =& \sum_{i=1}^{t}\frac{1}{\eta_i}B_{\Phi^*}(\hat{\mu}_i,\hat{\mu}_{i-1}) - t\Phi^*(\hat{\mu}_t).\end{aligned}$$

Moreover, from the definition we have that

$$\sum_{i=1}^{t}\{-\log f_{\theta}(X_i)\} = \sum_{i=1}^{t}\left[\Phi(\theta) - \theta^{\intercal}\phi(X_i)\right].$$

Taking the first derivative of $\sum_{i=1}^{t}\{-\log f_{\theta}(X_i)\}$ with respect to θ and setting it to 0, we find $\hat{\mu}$, the stationary point, given by

$$\hat{\mu} = \nabla\Phi(\theta) = \frac{1}{t}\sum_{i=1}^{t}\phi(X_i).$$

Similarly, using the expression of the dual function, and plugging $\hat{\mu}$ back into the equation, we have that

$$\inf_{\tilde{\theta}\in\Theta}\sum_{i=1}^{t}\{-\log f_{\tilde{\theta}}(X_i)\} = t\cdot\theta^{\intercal}\hat{\mu} - t\Phi^*(\hat{\mu}) - \sum_{i=1}^{t}\theta^{\intercal}\phi(X_i) = -t\Phi^*(\hat{\mu}).$$

□

Proof of Theorem 23. By choosing the step-size $\eta_i = 1/i$ for any $i \geq 1$ in Algorithm 2, and assuming $\hat{\theta}_i = \tilde{\theta}_i$ for any $i \geq 1$, we have by induction that

$$\hat{\mu}_t = \frac{1}{t}\sum_{i=1}^{t}\phi(X_i) = \hat{\mu}.$$

Subtracting (A.86) by (A.87), we obtain

$$
\begin{aligned}
\mathcal{R}_t &= \sum_{i=1}^{t}\{-\log f_{\hat{\theta}_{i-1}}(X_i)\} - \inf_{\tilde{\theta}\in\Theta}\sum_{i=1}^{t}\{-\log f_{\tilde{\theta}}(X_i)\} \\
&= \sum_{i=1}^{t} iB_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) - t\Phi^*(\hat{\mu}_t) + t\Phi^*(\hat{\mu}) \\
&= \sum_{i=1}^{t} iB_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) \\
&= \sum_{i=1}^{t} i[\Phi^*(\hat{\mu}_i) - \Phi^*(\hat{\mu}_{i-1}) - \langle \nabla\Phi^*(\hat{\mu}_{i-1}), \hat{\mu}_i - \hat{\mu}_{i-1}\rangle] \\
&= \frac{1}{2}\sum_{i=1}^{t} i\cdot(\hat{\mu}_i - \hat{\mu}_{i-1})^\intercal[\nabla^2\Phi^*(\tilde{\mu}_i)](\hat{\mu}_i - \hat{\mu}_{i-1}).
\end{aligned}
$$

The final equality is obtained by Taylor expansion. □

REFERENCES

[1] A. Ahmed and J. Romberg, "Compressive multiplexing of correlated signals," *IEEE Transactions on Information Theory*, vol. 61, no. 1, pp. 479–498, 2015.

[2] J. A. Bazerque, G. Mateos, and G. B. Giannakis, "Inference of poisson count processes using low-rank tensor data," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 5989 –5993.

[3] P. Biswas, T.-C. Lian, T.-C. Wang, and Y. Ye, "Semidefinite programming based algorithms for sensor network localization," *ACM Transactions on Sensor Networks (TOSN)*, vol. 2, no. 2, pp. 188–220, 2006.

[4] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61–70, 1992.

[5] P. Resnick and H. R. Varian, "Recommender systems," *Communications of the ACM*, vol. 40, no. 3, pp. 56–58, 1997.

[6] S. T. Dumais, "Latent semantic analysis," *Annual review of information science and technology*, vol. 38, no. 1, pp. 188–230, 2004.

[7] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics (FOCS)*, vol. 9, no. 6, pp. 717–772, 2009.

[8] M. A. Davenport, Y. Plan, E. v. d. Berg, and M. Wootters, "1-bit matrix completion," *Information and Inference*, 2014.

[9] D. C. Montgomery, *Statistical quality control*. Wiley New York, 2009, vol. 7.

[10] I. Berkes, E. Gombay, L. Horváth, and P. Kokoszka, "Sequential change-point detection in garch (p, q) models," *Econometric Theory*, vol. 20, no. 06, pp. 1140–1167, 2004.

[11] C. Zou, Y. Zhang, and Z. Wang, "A control chart based on a change-point model for monitoring linear profiles," *IIE transactions*, vol. 38, no. 12, pp. 1093–1103, 2006.

[12] A. G. Tartakovsky and V. V. Veeravalli, "Change-point detection in multichannel and distributed systems," *Applied Sequential Methodologies: Real-World Examples with Data Analysis*, vol. 173, pp. 339–370, 2004.

[13] Y. Cao, Y. Xie, and N. Gebraeel, “Multi-sensor slope change detection,” *Annals of Operations Research*, pp. 1–27, 2016.

[14] X. Fang, R. Zhou, and N. Gebraeel, “An adaptive functional regression-based prognostic model for applications with missing data,” *Reliability Engineering & System Safety*, vol. 133, pp. 266–274, 2015.

[15] K. Liu, N. Gebraeel, and J. Shi, “A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis,” *Automation Science and Engineering, IEEE Transactions on*, vol. 10, no. 3, pp. 652–664, 2013.

[16] V. Veeravalli, “Quickest detection and isolation of line outages in power systems,” in *International Workshop on Sequential Methods (IWSM)*, 2015.

[17] E. King. (2012). Veeravalli to develop sensor networks for chemical, biological threat detection.

[18] Y. Xie and D. Siegmund, “Sequential multi-sensor change-point detection,” *The Annals of Statistics*, vol. 41, no. 2, pp. 670–692, 2013.

[19] G. Fellouris and G. Sokolov, “Multisensor quickest detection,” *arXiv:1410.3815*, 2014.

[20] H. Chan, “Optimal detection in multi-stream data,” *arXiv:1506.08504*, 2015.

[21] D. J. Brady, *Optical imaging and spectroscopy*. John Wiley & Sons, 2009.

[22] H. Shen and J. Z. Huang, “Analysis of call centre arrival data using singular value decomposition,” *Applied Stochastic Models in Business and Industry*, vol. 21, pp. 251–263, 2005.

[23] E. J. Candès and T. Tao, “Near-optimal signal recovery from random projections: universal encoding strategies?” *IEEE Trans. Info. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[24] D. L. Donoho, “Compressed sensing,” *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[25] R. H. Keshavan, A. Montanari, and S. Oh, “Matrix completion from a few entries,” *IEEE Trans. Info. Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[26] E. J. Candès and T. Tao, “The power of convex relaxation: near-optimal matrix completion,” *IEEE Trans. Info. Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[27] W. Dai and O. Milenkovic, “Set: An algorithm for consistent matrix completion,” in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 3646–3649.

[28] B. Recht, M. Fazel, and P. A. Parrilo, “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[29] B. Recht, “A simpler approach to matrix completion,” *J. Machine Learning Research*, vol. 12, pp. 3413–3430, 2011.

[30] J.-F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM J. Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[31] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, “Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix,” *Comp. Adv. Multi-Sensor Adaptive Processing (CAMSAP)*, vol. 61, 2009.

[32] R. Mazumder, T. Hastie, and R. Tibshirani, “Spectral regularization algorithms for learning large incomplete matrices,” *J. Machine Learning Research*, vol. 11, pp. 2287–2322, 2010.

[33] R. Keshavan, A. Montanari, and S. Oh, “Matrix completion from noisy entries,” in *Adv. Neural Information Processing Systems (NIPS)*, 2009, pp. 952–960.

[34] E. J. Candes and Y. Plan, “Matrix completion with noise,” *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[35] S. Negahban, M. J. Wainwright, *et al.*, “Estimation of (near) low-rank matrices with noise and high-dimensional scaling,” *Ann. Stats.*, vol. 39, no. 2, pp. 1069–1097, 2011.

[36] S. Negahban and M. J. Wainwright, “Restricted strong convexity and weighted matrix completion: optimal bounds with noise,” *J. Machine Learning Research*, vol. 13, no. 1, pp. 1665–1697, 2012.

[37] A. Rohde, A. B. Tsybakov, *et al.*, “Estimation of high-dimensional low-rank matrices,” *Ann. Stats.*, vol. 39, no. 2, pp. 887–930, 2011.

[38] A. Soni, S. Jain, J. Haupt, and S. Gonella, “Error bounds for maximum likelihood matrix completion under sparse factor models,” in *IEEE Global Conf. Sig. and Info. Proc. (GlobalSIP)*, 2014.

[39] ——, “Noisy matrix completion under sparse factor models,” *arXiv:1411.0282*, 2014.

[40] A. Soni and J. Haupt, “Estimation error guarantees for poisson denoising with sparse and structured dictionary models,” in *IEEE Int. Symp. Info. Theory (ISIT)*, IEEE, 2014, pp. 2002–2006.

[41] M. Raginsky, R. M. Willett, Z. T. Harmany, and R. F. Marcia, “Compressed sensing performance bounds under poisson noise,” *IEEE Trans. Signal Processing*, vol. 58, no. 8, pp. 3990–4002, 2010.

[42] M. Raginsky, S. Jafarpour, Z. T. Harmany, R. F. Marcia, R. M. Willett, and R. Calderbank, “Performance bounds for expander-based compressed sensing in Poisson noise,” *IEEE Trans. Sig. Proc.*, vol. 59, no. 9, pp. 4139–4153, 2011.

[43] X. Jiang, G. Raskutti, and R. Willett, “Minimax optimal rates for poisson inverse problems with physical constraints,” *arXiv:1403.6532*, 2014.

[44] Q. Tran-Dinh, A. Kyrillidis, and V. Cevher, “A proximal newton framework for composite minimization: graph learning without cholesky decomposition and matrix inversions,” *Proc. 30th Int. Conf. Machine Learning (ICML)*, 2013.

[45] ——, “Composite self-concordant minimization,” *The Journal of Machine Learning Research*, vol. 16, no. 1, 371416, 2015.

[46] Y. Plan, “Compressed sensing, sparse approximation, and low-rank matrix estimation,” PhD thesis, California Institute of Technology, 2011.

[47] E. J. Candes and Y. Plan, “Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements,” *IEEE Trans. Info. Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[48] S. Ji and J. Ye, “An accelerated gradient method for trace norm minimization,” in *Proc. 26th Ann. Int. Conf. on Machine Learning*, ACM, 2009, pp. 457–464.

[49] M. J. Wainwright, “Structured regularizers for high-dimensional problems: statistical and computational issues,” *Annual Review of Statistics and Its Application*, vol. 1, pp. 233–253, 2014.

[50] A. Agarwal, S. Negahban, and M. J. Wainwright, “Fast global convergence rates of gradient methods for high-dimensional statistical recovery,” in *Adv. Neural Information Processing Systems (NIPS)*, 2010, pp. 37–45.

[51] J. Lafond, “Low rank matrix completion with exponential family noise,” *arXiv:1502.06919*, 2015.

[52] M. J. Wainwright, “Structured regularizers for high-dimensional problems: statistical and computational issues,” *Annual Review of Statistics and its Applications*, pp. 233–253, 2014.

[53] A. Saxena, K. Goebel, D. Simon, and N. Eklund, “Damage propation modeling for aircraft engine run-to-failure simulation,” in *Int. Conf. Prognostics and Health Management (PHM)*, 2008.

[54] T. L. Lai, “Information bounds and quick detection of parameter changes in stochastic systems,” *Information Theory, IEEE Transactions on*, vol. 44, no. 7, pp. 2917–2929, 1998.

[55] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014.

[56] M. Basseville, I. V. Nikiforov, *et al.*, *Detection of abrupt changes: theory and application*. Prentice Hall Englewood Cliffs, 1993, vol. 104.

[57] E. Page, “Continuous inspection schemes,” *Biometrika*, vol. 41, pp. 100–115, 1954.

[58] A. N. Shiryaev, “On optimum methods in quickest detection problems,” *Theory of Probability & Its Applications*, vol. 8, no. 1, pp. 22–46, 1963.

[59] G. Lorden, “Procedures for reacting to a change in distribution,” *The Annals of Mathematical Statistics*, pp. 1897–1908, 1971.

[60] A. G. Tartakovsky and V. V. Veeravalli, “An efficient sequential procedure for detecting changes in multichannel and distributed systems,” in *Information Fusion, 2002. Proceedings of the Fifth International Conference on*, IEEE, vol. 1, 2002, pp. 41–48.

[61] Y. Mei, “Efficient scalable schemes for monitoring a large number of data streams,” *Biometrika*, vol. 97, no. 2, pp. 419–433, 2010.

[62] M. Pollak, “Optimal detection of a change in distribution,” *The Annals of Statistics*, pp. 206–227, 1985.

[63] G. V. Moustakides, “Optimal stopping times for detecting changes in distributions,” *The Annals of Statistics*, pp. 1379–1387, 1986.

[64] Z. G. Stoumbos, M. R. Reynolds Jr, T. P. Ryan, and W. H. Woodall, “The state of statistical process control as we proceed into the 21st century,” *Journal of the American Statistical Association*, vol. 95, no. 451, pp. 992–998, 2000.

[65] P. J. Huber, “A robust version of the probability ratio test,” *The Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.

[66] R. Crow and S. Schwartz, “On robust quickest detection procedures,” in *Information Theory, 1994. Proceedings., 1994 IEEE International Symposium on*, IEEE, 1994, p. 258.

[67] V. V. Veeravalli, T. Basar, and H. V. Poor, “Minimax robust decentralized detection,” *Information Theory, IEEE Transactions on*, vol. 40, no. 1, pp. 35–40, 1994.

[68] J. Unnikrishnan, V. V. Veeravalli, and S. P. Meyn, “Minimax robust quickest change detection,” *Information Theory, IEEE Transactions on*, vol. 57, no. 3, pp. 1604–1614, 2011.

[69] P. J. Huber, *Robust statistics. 1981.*

[70] B. C. Levy, *Principles of signal detection and parameter estimation*. Springer Science & Business Media, 2008.

[71] A. Müller, “Stochastic ordering of multivariate normal distributions,” *Annals of the Institute of Statistical Mathematics*, vol. 53, no. 3, pp. 567–575, 2001.

[72] A. Goldenshluger, A. Juditsky, A. Nemirovski, *et al.*, “Hypothesis testing by convex optimization,” *Electronic Journal of Statistics*, vol. 9, no. 2, pp. 1645–1712, 2015.

[73] Y. Cao, A. Nemirovski, Y. Xie, V. Guigues, A. Juditsky, *et al.*, “Change detection via affine and quadratic detectors,” *Electronic Journal of Statistics*, vol. 12, no. 1, pp. 1–57, 2018.

[74] P. Granjon, “The cusum algorithm-a small review,” 2013.

[75] G. Lorden and M. Pollak, “Nonanticipating estimation applied to sequential analysis and changepoint detection,” *Annals of statistics*, pp. 1422–1454, 2005.

[76] M. Raginsky, R. M. F, J. Silva, and R. Willett, “Sequential probability assignment via online convex programming using exponential families,” in *IEEE International Symposium on Information Theory*, IEEE, 2009, pp. 1338–1342.

[77] M. Raginsky, R. Willet, 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.

[78] L. Peel and A. Clauset, “Detecting change points in the large-scale structure of evolving networks,” in *29th AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

[79] S. Li, Y. Xie, M. Farajtabar, A. Verma, and L. Song, “Detecting weak changes in dynamic events over networks,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 2, pp. 346–359, 2017.

[80] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.

[81] E. Hazan, “Introduction to online convex optimization,” *Foundations and Trends in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

[82] D. Siegmund and B. Yakir, “Minimax optimality of the Shiryayev-Roberts change-point detection rule,” *Journal of Statistical Planning and Inference*, vol. 138, no. 9, pp. 2815–2825, 2008.

[83] K. Azoury and M. Warmuth, “Relative loss bounds for on-line density estimation with the exponential family of distributions,” *Machine Learning*, vol. 43, no. 3, pp. 211–246, 2001.

[84] B. Baumgartner, “An inequality for the trace of matrix products, using absolute values,” *arXiv preprint arXiv:1106.6189*, 2011.

[85] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991, vol. 23.

[86] D. Pollard, *A User’s guide to measure theoretic probability*. Cambridge University Press, 2002, vol. 8.

[87] B. Yu, “Assouad, Fano, and le cam,” in *Festschrift for Lucien Le Cam*, Springer, 1997, pp. 423–435.

[88] D. Gross, “Recovering low-rank matrices from few coefficients in any basis,” *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[89] E. S. Page, “Continuous inspection scheme,” *Biometrika*, vol. 41, no. 1/2, pp. 100 –115, 1954.

[90] H. V. Poor and O. Hadjiliadis, *Quickest detection*. Cambridge University Press, 2008.

[91] A. S. Willsky and H. L. Jones, “A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems,” *Automatic Control, IEEE Transactions on*, vol. 21, no. 1, pp. 108–112, 1976.

[92] B. Yakir, *Extremes in random fields: a theory and its applications*. John Wiley & Sons, 2013.

[93] W. H. Woodall and M. M. Ncube, “Multivariate cusum quality-control procedures,” *Technometrics*, vol. 27, no. 3, pp. 285–292, 1985.

[94] P. J. Huber and V. Strassen, “Minimax tests and the neyman-pearson lemma for capacities,” *The Annals of Statistics*, pp. 251–263, 1973.

[95] Y. Cao and Y. Xie, “Poisson matrix recovery and completion,” *IEEE Transactions on Signal Processing*, vol. 64, no. 6, pp. 1609–1620, 2016.

[96] S Boyd, *Convex optimization*. Cambridge Univ Press, 2004.

[97] M. Kloft, U. Brefeld, P. Laskov, K.-R. Müller, A. Zien, and S. Sonnenburg, “Efficient and accurate lp-norm multiple kernel learning,” in *Advances in neural information processing systems*, 2009.

[98] E. J. Candes and M. A. Davenport, “How well can we estimate a sparse vector?” *Applied and Computational Harmonic Analysis (ACHA)*, vol. 34, no. 2, pp. 317–323, 2013.

[99] Z. Liu and L. Vandenberghe, “Interior-point method for nuclear norm approximation with application to system identification,” *SIAM J. Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1235–1256, 2009.

[100] E. G. Birgin, J. M. Martínez, and M. Raydan, “Nonmonotone spectral projected gradient methods on convex sets,” *SIAM J. Optimization*, vol. 10, no. 4, pp. 1196–1211, 2000.

[101] L. El Ghaoui, *Lecture notes for ee227a: algorithms for large-scale convex optimization*, University of California, Berkeley, Berkeley, CA, 2010.

[102] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, “Efficient projections onto the l 1-ball for learning in high dimensions,” in *Proc. 25th Int. Conf. on Machine Learning (ICML)*, ACM, 2008, pp. 272–279.

[103] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM J. Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[104] Y. Cao and Y. Xie, “Low-rank matrix recovery in poisson noise,” in *IEEE Global Conf. Sig. and Info. Proc. (GlobalSIP)*, 2014, pp. 384–388.

[105] P. Jain, P. Netrapalli, and S. Sanghavi, “Low-rank matrix completion using alternating minimization,” in *Proc. 45th ACM Symp. on Theory of Computing (STOC)*, 2013, pp. 665–674.

[106] G. Lerman, M. McCoy, J. A. Tropp, and T. Zhang, "Robust computation of linear models, or how to find a needle in a haystack," *arXiv:1202.4044*, 2012.

[107] H. Fanaee-T and J. Gama, "Event labeling combining ensemble detectors and background knowledge," *Prog. Artificial Intelligence*, pp. 1–15, 2013.

[108] O. Hadjiliadis, H. Zhang, and H. V. Poor, "One shot schemes for decentralized quickest change detection," *Information Theory, IEEE Transactions on*, vol. 55, no. 7, pp. 3346–3359, 2009.

[109] D. Siegmund, B. Yakir, N. R. Zhang, *et al.*, "Detecting simultaneous variant intervals in aligned sequences," *The Annals of Applied Statistics*, vol. 5, no. 2A, pp. 645–668, 2011.

[110] 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.

[111] D. Siegmund and B. Yakir, *The Statistics of Gene Mapping*. Springer, 2007.

[112] W. H. Woodall and M. M. Ncube, "Multivariate CUSUM qualify control procedures," *Technometrics*, vol. 27, no. 3, 1985.

[113] Y. Cao and Y. Xie, "Robust sequential change-point detection by convex optimization," in *Information Theory (ISIT), 2017 IEEE International Symposium on*, IEEE, 2017, pp. 1287–1291.

[114] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, pp. 199–227, 2008.

[115] P. Ravikumar, M. J. Wainwright, G. Raskutti, B. Yu, *et al.*, "High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[116] J. Fan, Y. Liao, and M. Mincheva, "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 4, pp. 603–680, 2013.

[117] Y. Cao, L. Xie, Y. Xie, and H. Xu, "Sequential change-point detection via online convex optimization," *Entropy*, vol. 20, no. 2, p. 108, 2018.

[118] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *The Annals of Mathematical Statistics*, pp. 326–339, 1948.

[119] A. Barron and C.-H. Sheu, "Approximation of density functions by sequences of exponential families," *Annals of Statistics*, pp. 1347–1369, 1991.

[120] M. J. Wainwright, M. I. Jordan, *et al.*, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[121] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[122] A. Nemirovskii, D. Yudin, and E. Dawson, *Problem complexity and method efficiency in optimization*. Wiley, 1983.

[123] S. Shalev-Shwartz *et al.*, "Online learning and online convex optimization," *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[124] A. Agarwal and J. C. Duchi, "Stochastic optimization with non-i.i.d. noise," 2011.

[125] T.-Z. Lai, "Likelihood ratio identities and their applications to sequential analysis," *Sequential Analysis*, vol. 23, no. 4, pp. 467–497, 2004.

[126] M. Pollak, "Average run lengths of an optimal method of detecting a change in distribution," *The Annals of Statistics*, pp. 749–779, 1987.

[127] I. M. Alqanoo, "On the truncated distributions within the exponential family," *Department of Applied Statistics, Al- Azhar University - Gaza*, 2014.

[128] Y. Wang and Y. Mei, "Large-scale multi-stream quickest change detection via shrinkage post-change estimation," *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6926–6938, 2015.

[129] J. Canny, *Lecture notes for cs174: combinatorics and discrete probability*, University of California, Berkeley, Berkeley, CA, 2001.

[130] J. Watrous, *Lectures nots for cs766: theory of quantum information*, University of Waterloo, 2011.

[131] D. Siegmund, *Sequential analysis: tests and confidence intervals*. Springer, 1985.

[132] R. Lipster and A. Shiryayev, "Theory of martingales," 1989.