Sequential Change-Point Detection via Online Convex Optimization

Yang Cao, Liyan Xie, Yao Xie * and **Huan Xu**

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology,
Atlanta, GA 30332, USA; caoyang@gatech.edu (Y.C.); lxie49@gatech.edu (L.X.); huan.xu@isye.gatech.edu (H.X.)
* Correspondence: yao.xie@isye.gatech.edu

Abstract: Sequential change-point detection when the distribution parameters are unknown is a fundamental problem in statistics and machine learning. When the post-change parameters are unknown, 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 a 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.

Keywords: sequential methods; change-point detection; online algorithms

1. Introduction

Sequential analysis is a classic topic in statistics concerning *online* inference from a sequence of observations. The goal is to make statistical inference *as quickly as possible*, while controlling the false-alarm rate. An important sequential analysis problem commonly studied is sequential change-point detection [1]. It arises from various applications including online anomaly detection, statistical quality control, biosurveillance, financial arbitrage detection and network security monitoring (see, e.g., [2–4]).

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.

1.1. 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 [4] 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 [5].

An alternative approach is the Generalized Likelihood Ratio (GLR) statistic based procedure [6]. The GLR statistic finds the maximum likelihood estimate (MLE) of the post-change parameter and plugs it back into 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 [7], 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 [8], $\hat{\theta}_{k+1,t} = (\sum_{i=k+1}^{t} X_i)/(t-k)$, 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 1.2. 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., [9,10]). Such a scheme provides a more versatile approach to developing a 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 an approach. This is the question we aim to address in this paper.

1.2. 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" [11]. 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 our prior work [12], we 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. A simulated example to illustrate this case is shown later in Section 4.4.

1.3. Contributions

This paper has two main 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 [8] 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 [4] 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 [13,14], 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 [4,15]. More precisely, we provide a general upper bound for the 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 is concerned 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 [14,16]. 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.

1.4. Literature and Related Work

Sequential change-point detection is a classic subject with an extensive literature. Much success has been achieved when the pre-change and post-change distributions are exactly specified. For example, the CUSUM procedure [17] with first-order asymptotic optimality [18] and exact optimality [19] in the minimax sense, and the Shiryayev-Roberts (SR) procedure [20] derived based on Bayesian principle that also enjoys various optimality properties. Both CUSUM and SR procedures rely on likelihood ratios between the specified pre-change and post-change distributions.

There are two main approaches in dealing with the unknown post-change parameters. The first one is a GLR approach [7,21–24], and the second is a mixture approach [15,25]. The GLR statistic enjoys certain optimality properties, but it can not be computed recursively in many cases [23]. To address the infinite memory issue, [7,21] studied the window-limited GLR procedure. The main advantage of the mixture approach is that it allows an easy evaluation of a threshold that guarantees the desired

false alarm constraint. A disadvantage of this approach is that sometimes there may not be a natural way of selecting the weight function, in particular when there is no conjugate prior. This motivated a third approach to this problem, which was proposed first by Robbins and Siegmund in the context of hypothesis testing, and then Lorden and Pollak [8] in the sequential change detection problem. This approach replaces the unknown parameter with some non-anticipating estimator, which can be easier to find even if there is no conjugate prior, as in the Gamma example considered in [8,25]. This work developed a modified SR procedure by introducing a prior distribution to the unknown parameters. While the non-anticipating estimator approach [8,24] enjoys recursive and thus efficient computation for the likelihood ratio based detection statistics, their approaches to constructing recursive estimators (based on MLE or method-of-moments) cannot be easily extended to more complex cases (for instance, multi-dimensional parameters with constraints). Here, we consider a general and convenient approach for constructing non-anticipating estimators based on online convex optimization which is particularly useful for these complex cases. Our work provides an alternative proof for the nearly second-order asymptotic optimality by building a connection to online convex optimization and leveraging the regret bound type of results [14]. For one-dimensional Gaussian mean shift without any constraint, we replicate the second-order asymptotic optimality, namely, Theorem 3.3 in [24]. Recent work [26] also treats the problem when the pre-change distribution has unknown parameters.

Another related problem is sequential joint estimation and detection, but the goal is different in that one aims to achieve both good detection and good estimation performance, whereas in our setting, estimation is only needed for computing the detection statistics. These works include [27] and [28], which study the joint detection and estimation problem of a specific form that arises from many applications such as spectrum sensing [29], image observations [30], and MIMO radar [31]: a linear scalar observation model with Gaussian noise, and under the alternative hypothesis there is *an unknown multiplicative parameter*. The paper of [27] demonstrates that solving the joint problem by treating detection and estimation separately with the corresponding optimal procedure does not yield an overall optimum performance, and provides an elegant closed-form optimal detector. Later on [28] generalizes the results. There are also other approaches solving the joint detection-estimation problem using multiple hypotheses testing [30,32] and Bayesian formulations [33].

Related work using online convex optimization for anomaly detection includes [9], which develops an efficient detector for the exponential family using online mirror descent and proves a logarithmic regret bound, and [10], which dynamically adjusts the detection threshold to allow feedbacks about whether decision outcome. However, these works consider a different setting that the change is a transient outlier instead of a persistent change, as assumed by the classic statistical change-point detection literature. When there is persistent change, it is important to accumulate "evidence" by pooling the post-change samples (our work considers the persistent change).

Extensive work has been done for parameter estimation in the online-setting. This includes online density estimation over the exponential family by regret minimization [9,10,16], sequential prediction of individual sequence with the logarithm loss [13,34], online prediction for time series [35], and sequential NML (SNML) prediction [34] which achieves the optimal regret bound. Our problem is different from the above, in that estimation is not the end goal; one only performs parameter estimation to plug them back into the likelihood function for detection. Moreover, a subtle but important difference of our work is that the loss function for online detecting estimation is $-f_{\hat{\theta}_i}(X_i)$, whereas our loss function is $-f_{\hat{\theta}_{i-1}}(X_i)$ in order to retain the *martingale property*, which is essential to establish the nearly second-order asymptotic optimality.

2. 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 estimator $\{\hat{\theta}_t\}$ 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 [7], 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.

2.1. 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 $H_0 : \theta = \theta_0$ versus the alternative $H_1 : \theta \neq \theta_0$. 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) [36]: 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 can not be made confidently. Here, we introduce a *modified* one-sided SPRT with a sequence of *non-anticipating* plug-in estimators:

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

Define the test statistic at time t as

$$\Lambda_t = \prod_{i=1}^{t} \frac{f_{\hat{\theta}_{i-1}}(X_i)}{f_{\theta_0}(X_i)}, \quad i \geq 1. \tag{2}$$

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

$$\tau(b) = \min\{t \geq 1 : \log \Lambda_t \geq b\}, \tag{3}$$

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.

2.2. 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:

$$
\begin{aligned}
H_0 : \ & X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} f_{\theta_0}, \\
H_1 : \ & X_1, \ldots, X_\nu \overset{\text{i.i.d.}}{\sim} f_{\theta_0}, \quad X_{\nu+1}, X_{\nu+2}, \ldots \overset{\text{i.i.d.}}{\sim} f_\theta,
\end{aligned}
\tag{4}
$$

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

$$\hat{\theta}_{k,i} = \hat{\theta}_{k,i}(X_k, \ldots, X_i), \quad i \geq k. \tag{5}$$

Therefore, for $k = 1$, $\hat{\theta}_{k,i}$ becomes $\hat{\theta}_i$ defined in (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

$$\Lambda_{k,t} = \prod_{i=k}^{t} \frac{f_{\hat{\theta}_{k,i-1}}(X_i)}{f_{\theta_0}(X_i)}, \tag{6}$$

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

$$T_{\text{ACM}}(b_1) = \inf\left\{ t \geq 1 : \max_{1 \leq k \leq t} \log \Lambda_{k,t} > b_1 \right\}, \tag{7}$$

where b_1 is a pre-specified threshold. Similarly, by replacing the maximization over k in (7) with summation, we obtain the following ASR procedure [8], which can be interpreted as a Bayesian statistic similar to the Shiryaev-Roberts procedure.

$$T_{\text{ASR}}(b_2) = \inf\left\{ t \geq 1 : \log\left(\sum_{k=1}^{t} \Lambda_{k,t} \right) > b_2 \right\}, \tag{8}$$

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 2.4. For a fixed k, the comparison between our methods and GLR is illustrated in Figure 1.

Remark 1. *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 (7) and (8). The window-limited versions are obtained by replacing $\max_{1 \leq k \leq t}$ with $\max_{t-w \leq k \leq t}$ in (7) and by replacing $\sum_{k=1}^{t}$ with $\sum_{k=t-w}^{t}$ in (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 [7] for the choice of w the window-limited GLR procedures.*
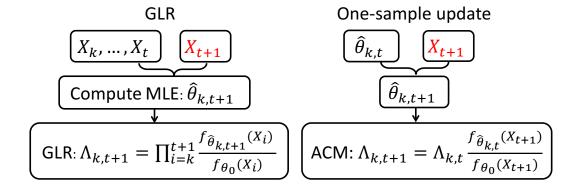

Figure 1. Comparison of the update scheme for GLR and our methods when a new sample arrives.

2.3. Exponential Family

In this paper, we focus on f_θ being the exponential family for the following reasons: (i) exponential family [10] represents a very rich class of parametric and even many nonparametric statistical models [37]; (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 [10,38].

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^T denote the transpose of a vector or matrix. Let $\phi : \mathcal{X} \to \mathbb{R}^d$ be an H-measurable function $\phi(x) = (\phi_1(x), \ldots, \phi_d(x))^\mathsf{T}$. 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 $f_\theta(x) = \exp\{\theta^\mathsf{T}\phi(x) - \Phi(\theta)\}$. In the definition, the so-called log-partition function is given by

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

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

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

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 [38]. The Legendre-Fenchel dual Φ^* is defined as

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

The mappings $\nabla\Phi^*$ is an inverse mapping of $\nabla\Phi$ [39]. 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., [10,38]) defined as

$$B_F(u, v) := F(u) - F(v) - \langle\nabla F(v), u - v\rangle. \tag{9}$$

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

$$I(\theta_1, \theta_2) = \mathbb{E}_{\theta_1}\left[\log(f_{\theta_1}(X)/f_{\theta_2}(X))\right]. \tag{10}$$

It can be shown that, for exponential family, $I(\theta_1, \theta_2) = \Phi(\theta_2) - \Phi(\theta_1) - (\theta_2 - \theta_1)^\mathsf{T}\nabla\Phi(\theta_1)$. Using the definition (9), this means that B_Φ

$$B_\Phi(\theta_1, \theta_2) := I(\theta_2, \theta_1) \tag{11}$$

is a Bregman divergence. This property is useful to constructing mirror descent estimator for the exponential family [39,40].

2.4. Online Convex Optimization (OCO) Algorithms for Non-Anticipating Estimators

Online convex optimization (OCO) algorithms [14] 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 relative 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 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 [13]. 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

$$\mathcal{R}_t^{\mathsf{a}} = \sum_{i=1}^t \{-\log f_{\hat{\theta}_{i-1}}(X_i)\} - \inf_{\tilde{\theta} \in \Theta} \sum_{i=1}^t \{-\log f_{\tilde{\theta}}(X_i)\}. \tag{12}$$

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) [14,41]. Next, we discuss how to construct the non-anticipating estimators $\{\hat{\theta}_t\}_{t\geq 1}$ in (1), and $\{\hat{\theta}_{k,t}\}, k = 1, 2, \ldots, t-1$ in (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

$$\hat{\theta}_t = \arg\min_{u \in \Gamma} [u^{\mathsf{T}} \nabla \ell_t(\hat{\theta}_{t-1}) + \frac{1}{\eta_i} B_\Phi(u, \hat{\theta}_{t-1})], \tag{13}$$

where B_Φ is defined in (11). Here $\Gamma \subset \Theta_\sigma$ is a closed convex set, which is problem-specific and encourages certain parameter structure such as sparsity.

Remark 2. *Similar to (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 (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 [40]. 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 (13) is proved in [39]. We adopt this approach when deriving our algorithm and follow the same strategy as [9]. Algorithm 1 summarizes the steps [42].

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 [10,43]. Note that for exponential family, the loss function is the negative log-likelihood function, which is strongly convex over Θ_σ. Hence, we can have the logarithmic regret property.

Algorithm 1 Online mirror-descent for non-anticipating estimators.

Require: 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.

1: $\hat{\theta}_0 = \theta_0, \Lambda_0 = 1.$ {Initialization}

2: **for all** $t = 1, 2, \ldots,$ **do**

3:     Acquire a new observation X_t

4:     Compute loss $\ell_t(\hat{\theta}_{t-1}) \triangleq -\log f_{\hat{\theta}_{t-1}}(X_t) = \Phi(\hat{\theta}_{t-1}) - \hat{\theta}_{t-1}^\mathsf{T} \phi(X_t)$

5:     Compute likelihood ratio $\Lambda_t = \Lambda_{t-1} \dot{f}_{\hat{\theta}_{t-1}}(X_t) / f_{\theta_0}(X_t)$

6:     $\hat{\mu}_{t-1} = \nabla \Phi(\hat{\theta}_{t-1}), \hat{\mu}_t = \hat{\mu}_{t-1} - \eta_t(\hat{\mu}_{t-1} - \phi(X_t))$ {Dual update}

7:     $\tilde{\theta}_t = (\nabla \Phi)^*(\hat{\mu}_t)$

8:     $\hat{\theta}_t = \arg\min_{u \in \Gamma} B_\Phi(u, \tilde{\theta}_t)$ {Projected primal update}

9: **end for**

10: **return** $\{\hat{\theta}_t\}_{t \geq 1}$ and $\{\Lambda_t\}_{t \geq 1}$.

---

3. 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 $t \geq 1$.

3.1. "One-Sided" Sequential Hypothesis Test

Recall that the decision rule for sequential hypothesis test is a stopping time $\tau(b)$ defined in (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 [23], 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 (12).

Theorem 1 (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 $b \to \infty$,*

$$\mathbb{E}_{\theta,0}[\tau(b)] \leq \frac{b}{I(\theta, \theta_0)} + \frac{\mathbb{E}_{\theta,0}\left[\mathcal{R}^a_{\tau(b)}\right]}{I(\theta, \theta_0)} + O(1) \tag{14}$$

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 3. *The inequality (14) is valid for a sequence of non-anticipating estimators generated by an OCO algorithm. Moreover, (14) gives an explicit connection between the expected detection delay for the one-sided sequential hypothesis testing (left-hand side of (14)) and the regret for the OCO (the second term on the right-hand side of (14)). This illustrates clearly the impact of estimation on detection by an estimation algorithm dependent factor.*

Note that in the statement of the Theorem 1, the stopping time $\tau(b)$ appears on the right-hand side of the inequality (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 4. The most important case is that when the estimation algorithm has a logarithmic expected regret. For the exponential family, as shown in Section 3.3, Algorithm 1 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 1.

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

$$\mathbb{E}_{\theta,0}[\tau(b)] \leq \frac{b}{I(\theta,\theta_0)} + \frac{C \log b}{I(\theta,\theta_0)}(1 + o(1)). \tag{15}$$

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

Corollary 1 shows that other than the well known first-order approximation $b/I(\theta,\theta_0)$ [8,18], 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 1 (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 1 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 [4], we establish the nearly second-order asymptotic optimality of OMD based SPRT as follows:

Corollary 2 (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* $\mathbb{E}_{\theta,0}[\mathcal{R}_n^a] \leq C \log n$ *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 1,* $\tau(b) \in C(\alpha)$*. 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* $\alpha \to 0$*,*

$$\mathbb{E}_{\theta,0}[\tau(b)] - \inf_{T \in C(\alpha)} \mathbb{E}_{\theta,0}[T] = O(\log(\log(1/\alpha))). \tag{16}$$

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 $\log(\log(1/\alpha)) = 2.44$.

3.2. Sequential Change-Point Detection

Now we proceed the proof by leveraging the close connection [18] between the sequential change-point detection and the one-sided hypothesis test. For sequential change-point detection, the two commonly used performance metrics [4] 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 1, we provide an upper bound for the CADD of our ASR and ACM procedures.

Theorem 2. *Consider the change-point detection procedure $T_{\mathrm{ACM}}(b_1)$ in (7) and $T_{\mathrm{ASR}}(b_2)$ in (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*

$$
\sup_{\nu \geq 0} \mathbb{E}_{\theta,\nu}[T_{\mathrm{ASR}}(b) - \nu \mid T_{\mathrm{ASR}}(b) > \nu] \leq \sup_{\nu \geq 0} \mathbb{E}_{\theta,\nu}[T_{\mathrm{ACM}}(b) - \nu \mid T_{\mathrm{ACM}}(b) > \nu]
$$
$$
\leq (I(\theta,\theta_0))^{-1} \left(b + \mathbb{E}_{\theta,0}\left[\mathcal{R}^a_{\tau(b)}\right] + O(1) \right). \tag{17}
$$

To prove Theorem 2, 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 1 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 3:

Lemma 2 (ARL). *Consider the change-point detection procedure $T_{\mathrm{ACM}}(b_1)$ in (7) and $T_{\mathrm{ASR}}(b_2)$ in (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 2 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 [8,25] 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 2 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 2 with an existing lower bound for the CADD of SRRS procedure in [15], we obtain the following optimality properties.

Corollary 3 (Nearly second-order asymptotic optimality of ACM and ASR). *Consider the change-point detection procedure $T_{\mathrm{ACM}}(b_1)$ in (7) and $T_{\mathrm{ASR}}(b_2)$ in (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 $\mathbb{E}_{\theta,0}[\mathcal{R}^a_n] \leq C \log n$ 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 2, 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 $\theta \in \Theta - \{\theta_0\}$*

$$
\sup_{\nu \geq 1} \mathbb{E}_{\theta,\nu}[T_{\mathrm{ASR}}(b) - \nu + 1 \mid T_{\mathrm{ASR}}(b) \geq \nu]
$$
$$
- \inf_{T(b) \in S(\gamma)} \sup_{\nu \geq 1} \mathbb{E}_{\theta,\nu}[T(b) - \nu + 1 \mid T(b) \geq \nu] = O(\log \log \gamma). \tag{18}
$$

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 (18) with (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 [18].

3.3. 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 [16].

Theorem 3. *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 1. Assume that $\{\hat{\theta}_i\}_{i \geq 1}, \{\hat{\mu}_i\}_{i \geq 1}$ are obtained using Algorithm 1 and $\hat{\theta}_i = \tilde{\theta}_i$ (defined in step 7 and 8 of Algorithm 1) for any $i \geq 1$. Then for any $\theta_0 \in \Theta$ and $t \geq 1$,*

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

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

Next, we use Theorem 3 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 Section 2.3, we know that $\phi(x) = x$, $dH(x) = (1/\sqrt{|2\pi I_d|}) \cdot \exp(-x^\mathsf{T} x/2)$, $\Theta = \Theta_\sigma = \mathbb{R}^d$ for any $\sigma < 2$, $\Phi(\theta) = (1/2)\theta^\mathsf{T}\theta$, $\mu = \theta$ and $\Phi^*(\mu) = (1/2)\mu^\mathsf{T}\mu$, 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 4 (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 1 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 4, which also serve as an example of how to use regret bound. First, the assumption $\hat{\theta}_t = \tilde{\theta}_t$ in Theorem 3 is satisfied for the following reasons. Consider $\Gamma = \mathbb{R}^d$ is the full space. According to Algorithm 1, 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

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

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

$$\frac{1}{2} \sum_{i=2}^{t} (\hat{\mu}_i - \hat{\mu}_{i-1})^\intercal (\phi(X_i) - \hat{\mu}_{i-1})$$

$$= \frac{1}{2} \sum_{i=2}^{t} (\hat{\mu}_i - \hat{\mu}_{i-1})^\intercal (\phi(X_i) + \hat{\mu}_i) - \sum_{i=2}^{t} \frac{1}{2} (\hat{\mu}_i - \hat{\mu}_{i-1})^\intercal (\hat{\mu}_{i-1} + \hat{\mu}_i)$$

$$= \sum_{i=2}^{t} \frac{1}{2(i-1)} (\phi(X_i) - \hat{\mu}_i)^\intercal (\phi(X_i) + \hat{\mu}_i) + \sum_{i=2}^{t} \frac{1}{2} (\|\hat{\mu}_{i-1}\|^2 - \|\hat{\mu}_i\|^2)$$

$$= \sum_{i=2}^{t} \frac{1}{2(i-1)} \|X_i\|^2 - \sum_{i=2}^{t} \frac{1}{2(i-1)} \|\hat{\mu}_i\|^2 + \frac{1}{2} \|\hat{\mu}_1\|^2 - \frac{1}{2} \|\hat{\mu}_t\|^2 .$$

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 the 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 4, 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 Section 2.3 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 [10] 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 [44]).

4. 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 1) 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 2). 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 $10,000$. 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.

4.1. 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 [45,46]. 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 1 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., [47]). Then, the projection onto ℓ_1 ball can be computed very efficiently via a simple soft-thresholding technique [48].

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 [49], 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 $10,000$ 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 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.

Table 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 [49]; "GLR": GLR procedure; "ASR": $T_{ASR}(b)$ with $\Gamma = \mathbb{R}^d$; "ACM": $T_{ACM}(b)$ with $\Gamma = \mathbb{R}^d$; "ASR-L1": $T_{ASR}(b)$ 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 $10,000$ Monte Carlo trials to obtain each value. For each value, the standard deviation is less than one half of the value.

	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$	$p = 0.6$
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

4.2. 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 1 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 1 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 [8]. For the CUSUM procedure, we specify the post-change β to be 2. The results are shown in Table 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.

Table 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 $10,000$ Monte Carlo trials to obtain each value. For each value, the standard deviation is less than one half of the value.

	$\beta = 0.1$	$\beta = 0.5$	$\beta = 2$	$\beta = 5$	$\beta = 10$
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

4.3. Communication-Rate Change Detection with Erdős-Rényi 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 2). 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 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 the GLR procedure, our methods have a small performance loss, and the loss is almost negligible as n approaches to $d = 190$.

Table 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 $10,000$ Monte Carlo trials to obtain each value. For each value, the standard deviation is less than one half of the value.

	$n = 78$	$n = 100$	$n = 120$	$n = 150$	$n = 170$	$n = 190$
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

Below are the specifications of Algorithm 1 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 $\Phi^*(\mu) = \mu \log \mu + (1 - \mu)\log(1 - \mu)$.

4.4. Point Process Change-Point Detection: Poisson to Hawkes Processes

In this example, to illustrate the situation in Section 1.2, we consider a case where a homogeneous Poisson process switches to a Hawkes process (see, e.g., [12]); this can be viewed as a simplest case in Section 1.2 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 $\lambda_s, s \in (0, T)$ 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 $m_1, m_2, \ldots, m_t$. 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

$$
\ell(\theta | X_i) = \sum_{t_q \in (T_i - L, T_i)} \log \left[\mu + \theta \sum_{t_j \in (T_i - L, t_q)} \beta e^{-\beta(t_q - t_j)} \right] - \mu L - \theta \sum_{t_q \in (T_i - L, T_i)} \left[1 - e^{-\beta(T_i - t_q)} \right]. \quad (19)
$$

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

$$
\hat{\theta}_{k,i} \triangleq \hat{\theta}_{k,i}(X_k, \ldots, X_i) = \hat{\theta}_{k,i}(t_\ell \in [T_k, T_i]) \quad (20)
$$

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

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

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 (19).

Table 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.

Table 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.

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

5. Conclusions

In this paper, 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.

Author Contributions: Yang Cao, Yao Xie, and Huan Xu conceived the idea and performed the theoretical part of the paper; Liyan Xie helped with numerical examples of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs

Proof of Theorem 1. 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., [1]) we have that

$$\mathbb{E}_{\theta,0}[S_N^\theta] = \mathbb{E}_{\theta,0}[N] \cdot I(\theta, \theta_0). \tag{A1}$$

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:

$$S_N^\theta = \sum_{i=1}^N \log \frac{f_\theta(X_i)}{f_{\theta_N^*}(X_i)} + \sum_{i=1}^N \log \frac{f_{\theta_N^*}(X_i)}{f_{\hat{\theta}_{i-1}}(X_i)} + \sum_{i=1}^N \log \frac{f_{\hat{\theta}_{i-1}}(X_i)}{f_{\theta_0}(X_i)}, \tag{A2}$$

Note that the first term of the decomposition on the right-hand side of (A2) 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 (A2). Similar to the proof of equation (5.109) in [4], 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 (A2) as

$$\sum_{i=1}^n \log \frac{f_{\hat{\theta}_{i-1}}(X_i)}{f_{\theta_0}(X_i)} = M_n(\theta) - G_n(\theta) + m_n(\theta,\theta_0) + nI(\theta,\theta_0), \tag{A3}$$

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 (A3) 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 [47] we have that

$$\mathbb{E}_{\theta,0}|M_n(\theta)| \le C_1\sqrt{n}, \quad \mathbb{E}_{\theta,0}|m_n(\theta,\theta_0)| \le C_2\sqrt{n}, \tag{A4}$$

where C_1 and C_2 are two absolute constants that do not depend on n. Moreover, we observe that for all $\theta \in \Theta$,

$$\mathbb{E}_{\theta,0}[G_n(\theta)] \le \mathbb{E}_{\theta,0}\left[\max_{\tilde\theta \in \Theta} G_n(\tilde\theta)\right] = \mathbb{E}_{\theta,0}[\mathcal{R}_n(\theta)] \le C\log n.$$

Therefore, applying (A4), 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 $\mathcal{G}(\epsilon) = \sup\{n \ge 1 : |n^{-1}A_n - I| > \epsilon\}$ 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 [4]). Therefore, dividing both sides of (A3) 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

$$L(\epsilon) = \sup\left\{n \ge 1 : \left|\frac{1}{I(\theta,\theta_0)}\sum_{i=1}^n \log \frac{f_{\hat{\theta}_{i-1}}(X_i)}{f_{\theta_0}(X_i)} - n\right| > \epsilon n\right\}.$$

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

$$N \leq 1 + \mathbb{I}(\{N > L(\epsilon) + 1\}) \cdot \frac{\tilde{b}}{1 - \epsilon} + \mathbb{I}(\{N \leq L(\epsilon) + 1\}) \cdot L(\epsilon) \leq 1 + \frac{\tilde{b}}{1 - \epsilon} + L(\epsilon). \tag{A5}$$

Since $\mathbb{E}_{\theta,0}[L(\epsilon)] < \infty$ for any $\epsilon > 0$, from (A5) above, we have that the third term in (A2) is upper bounded by $\tilde{b} + \mathsf{O}(1)$.

Finally, the second term in (A2) 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 (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 (A2) and (A1). $\square$

Proof of Corollary 1. First, we can relate the expected regret at the stopping time to the expected stopping time, using the following chain of equalities and inequalities

$$\mathbb{E}_{\theta,0}[\mathcal{R}_{\tau(b)}] = \mathbb{E}_{\theta,0}[\mathbb{E}_{\theta,0}[\mathcal{R}_n \mid \tau(b) = n]] \leq \mathbb{E}_{\theta,0}[C \log \tau(b)] \leq C \log \mathbb{E}_{\theta,0}[\tau(b)], \tag{A6}$$

where the first equality uses iterative expectation, the first inequality uses the assumption of the logarithmic regret in the statement of Corollary 1, 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 (A6), the upper bound in Equation (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

$$\mathbb{E}_{\theta,0}[\tau(b)] \leq \frac{b}{I(\theta,\theta_0)} + \frac{C \log b}{I(\theta,\theta_0)}(1 + o(1)). \tag{A7}$$

Note that a similar argument can be found in [49]. $\square$

Next we will establish a few Lemmas useful for proving Theorem 2 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 A1. *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 1. To bound the term $\mathbb{P}_\infty(\tau(b) < \infty)$, we need take advantage of the martingale property of Λ_t in (2). The major technique is the combination of change of measure and Wald's likelihood ratio identity [1]. The proofs are a combination of the results in [23] and [8] and the reader can find a complete proof in [23]. 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 (2)). Combining the Lemma A1 and the Wald's likelihood ratio identity, we have that

$$\mathbb{P}_\infty(A \cap \{\tau(b) < \infty\}) = \mathbb{E}_\mathbb{Q}\left[\mathbb{I}(\{\tau(b) < \infty\}) \cdot L_{\tau(b)}\right], \forall A \in \mathcal{F}_{\tau(b)}, \tag{A8}$$

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 (A8) we prove that $\mathbb{P}_\infty(\tau(b) < \infty) \leq \exp(-b)$. $\square$

Proof of Corollary 2. Using (5.180) and (5.188) in [4], which are about asymptotic performance of open-ended tests. Since our problem is a special case of the problem in [4], 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 (15), we prove the corollary. $\square$

Proof of Theorem 2. From (A10), 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 1, 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 [8], we have that the supreme of detection delay over all change locations is achieved by the case when change occurs at the first instance,

$$\sup_{\nu \geq 0} \mathbb{E}_{\theta,\nu}[T_{ACM}(b) - \nu \mid T_{ACM}(b) > \nu] = \mathbb{E}_{\theta,0}[T_{ACM}(b)]. \tag{A9}$$

This is a slight modification (a small change on the subscripts) of the remarks in [8] 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 (A9), 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 [18]. Thus, $T_{ACM}(b) \leq \tau(b)^{(1)} + 1 - 1 = \tau(b)$, and $\mathbb{E}_{\theta,0}[T_{ACM}(b)] \leq \mathbb{E}_{\theta,0}[\tau(b)]$. $\square$

Proof of Lemma 2. This is a classic result proved by using the martingale property and the proof routine can be found in many textbooks such as [4]. 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

$$\log\left(\sum_{k=1}^{t} \Lambda_{k,t}\right) > \log\left(\max_{1\leq k\leq t} \Lambda_{k,t}\right) = \max_{1\leq k\leq t} \log \Lambda_{k,t}, \tag{A10}$$

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 $\mathbb{E}_\infty[T_{ASR}(b)] < \infty$ (otherwise the statement of proposition is trivial), then we have that

$$\sum_{t=1}^{\infty} \mathbb{P}_\infty(T_{ASR}(b) \geq t) < \infty. \tag{A11}$$

Equation (A11) 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 3. Our Theorem 1 and the remarks in [15] show that the minimum worst-case detection delay, given a fixed ARL level γ, is given by

$$\inf_{T(b)\in S(\gamma)} \sup_{\nu\geq 1} \mathbb{E}_{\theta,\nu}[T(b) - \nu + 1 \mid T(b) \geq \nu] = \frac{\log\gamma}{I(\theta,\theta_0)} + \frac{d\log\log\gamma}{2I(\theta,\theta_0)}(1 + o(1)). \tag{A12}$$

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 (A12) with the right-hand side of (15), we prove the corollary. □

The following derivation borrows ideas from [16]. First, we derive concise forms of the two terms in the definition of R_t in (12).

Lemma A2. *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 1. Given $\{\hat{\theta}_i\}_{i \geq 1}, \{\hat{\mu}_i\}_{i \geq 1}$ generated by Algorithm 1. If $\hat{\theta}_i = \tilde{\theta}_i$ for any $i \geq 1$, then for any null distribution parameter $\theta_0 \in \Theta$ and $t \geq 1$,*

$$\sum_{i=1}^{t} \{ -\log f_{\hat{\theta}_{i-1}}(X_i) \} = \sum_{i=1}^{t} i B_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) - t\Phi^*(\hat{\mu}_t). \tag{A13}$$

Moreover, for any $t \geq 1$,

$$\inf_{\tilde{\theta} \in \Theta} \sum_{i=1}^{t} \{ -\log f_{\tilde{\theta}}(X_i) \} = -t\Phi^*(\hat{\mu}), \tag{A14}$$

where $\hat{\mu} = (1/t) \cdot \sum_{i=1}^{t} \phi(X_i)$.

By subtracting the expressions in (A13) and (A14), 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 A2. By the definition of the Legendre-Fenchel dual function we have that $\Phi^*(\mu) = \theta^\mathsf{T}\mu - \Phi(\theta)$ for any $\theta \in \Theta$. By this definition, and choosing $\eta_i = 1/i$, we have that for any $i \geq 1$

$$-\log f_{\hat{\theta}_{i-1}}(X_i) = \Phi(\hat{\theta}_{i-1}) - \hat{\theta}_{i-1}^\mathsf{T}\phi(X_i) = \hat{\theta}_{i-1}^\mathsf{T}(\hat{\mu}_{t-1} - \phi(X_i)) - \Phi^*(\hat{\mu}_{i-1}) = \frac{1}{\eta_i}\hat{\theta}_{i-1}^\mathsf{T}(\hat{\mu}_{i-1} - \hat{\mu}_i) - \Phi^*(\hat{\mu}_{i-1})$$

$$= \frac{1}{\eta_i}(\Phi^*(\hat{\mu}_i) - \Phi^*(\hat{\mu}_{i-1})) - \hat{\theta}_{i-1}^\mathsf{T}(\hat{\mu}_i - \hat{\mu}_{i-1}) - \frac{1}{\eta_i}\Phi^*(\hat{\mu}_i) + \left(\frac{1}{\eta_i} - 1\right)\Phi^*(\hat{\mu}_{i-1}) \tag{A15}$$

$$= \frac{1}{\eta_i}B_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) + \frac{1}{\eta_{i-1}}\Phi^*(\hat{\mu}_{i-1}) - \frac{1}{\eta_i}\Phi^*(\hat{\mu}_i),$$

where we use the update rule in Line 6 of Algorithm 1 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 (A15), where the second term form a telescopic series, over i from 1 to t, we have that

$$\sum_{i=1}^{t} \{ -\log f_{\hat{\theta}_{i-1}}(X_i) \} = \sum_{i=1}^{t} \frac{1}{\eta_i}B_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) + \frac{1}{\eta_0}\Phi^*(\hat{\mu}_0) - \frac{1}{\eta_t}\Phi^*(\hat{\mu}_t)$$

$$= \sum_{i=1}^{t} \frac{1}{\eta_i}B_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) - t\Phi^*(\hat{\mu}_t).$$

Moreover, from the definition we have that

$$\sum_{i=1}^{t} \{ -\log f_\theta(X_i) \} = \sum_{i=1}^{t} [\Phi(\theta) - \theta^\mathsf{T}\phi(X_i)].$$

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

$$\inf_{\tilde{\theta} \in \Theta} \sum_{i=1}^{t} \{ -\log f_{\tilde{\theta}}(X_i) \} = t \cdot \theta^\mathsf{T}\hat{\mu} - t\Phi^*(\hat{\mu}) - \sum_{i=1}^{t} \theta^\mathsf{T}\phi(X_i) = -t\Phi^*(\hat{\mu}).$$

□

Proof of Theorem 3. By choosing the step-size $\eta_i = 1/i$ for any $i \geq 1$ in Algorithm 1, 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 (A13) by (A14), we obtain

$$
\begin{aligned}
\mathcal{R}_t &= \sum_{i=1}^{t} \{-\log f_{\hat{\theta}_{i-1}}(X_i)\} - \inf_{\tilde{\theta} \in \Theta} \sum_{i=1}^{t} \{-\log f_{\tilde{\theta}}(X_i)\} \\
&= \sum_{i=1}^{t} i B_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) - t\Phi^*(\hat{\mu}_t) + t\Phi^*(\hat{\mu}) \\
&= \sum_{i=1}^{t} i B_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) \\
&= \sum_{i=1}^{t} i [\Phi^*(\hat{\mu}_i) - \Phi^*(\hat{\mu}_{i-1}) - \langle \nabla \Phi^*(\hat{\mu}_{i-1}), \hat{\mu}_i - \hat{\mu}_{i-1} \rangle] \\
&= \frac{1}{2} \sum_{i=1}^{t} i \cdot (\hat{\mu}_i - \hat{\mu}_{i-1})^\mathsf{T} [\nabla^2 \Phi^*(\tilde{\mu}_i)] (\hat{\mu}_i - \hat{\mu}_{i-1}).
\end{aligned}
$$

The final equality is obtained by Taylor expansion. □

References

1. Siegmund, D. *Sequential Analysis: Tests and Confidence Intervals*; Springer: Berlin, Germany, 1985.
2. Chen, J.; Gupta, A.K. *Parametric Statistical Change Point Analysis*; Birkhauser: Basel, Switzerland, 2000.
3. Siegmund, D. Change-points: From sequential detection to biologu and back. *Seq. Anal.* **2013**, *23*, 2–14.
4. Tartakovsky, A.; Nikiforov, I.; Basseville, M. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*; CRC Press: Boca Raton, FL, USA, 2014.
5. Granjon, P. The CuSum Algorithm—A Small Review. 2013. Available online: https://hal.archives-ouvertes.fr/hal-00914697 (accessed on 6 February 2018).
6. Basseville, M.; Nikiforov, I.V. *Detection of Abrupt Changes: Theory and Application*; Prentice Hall Englewood Cliffs: Upper Saddle River, NJ, USA, 1993; Volume 104.
7. Lai, T.Z. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Trans. Inf. Theory* **1998**, *44*, 2917–2929.
8. Lorden, G.; Pollak, M. Nonanticipating estimation applied to sequential analysis and changepoint detection. *Ann. Stat.* **2005**, *33*, 1422–1454.
9. Raginsky, M.; Marcia, R.F.; Silva, J.; Willett, R. Sequential probability assignment via online convex programming using exponential families. In Proceedings of the IEEE International Symposium on Information Theory, Seoul, Korea, 28 June–3 July 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 1338–1342.
10. Raginsky, M.; Willet, R.; Horn, C.; Silva, J.; Marcia, R. Sequential anomaly detection in the presence of noise and limited feedback. *IEEE Trans. Inf. Theory* **2012**, *58*, 5544–5562.
11. Peel, L.; Clauset, A. Detecting change points in the large-scale structure of evolving networks. In Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI), Austin, TX, USA, 25–30 January 2015.
12. Li, S.; Xie, Y.; Farajtabar, M.; Verma, A.; Song, L. Detecting weak changes in dynamic events over networks. *IEEE Trans. Signal Inf. Process. Over Netw.* **2017**, *3*, 346–359.
13. Cesa-Bianchi, N.; Lugosi, G. *Prediction, Learning, and Games*; Cambridge University Press: Cambridge, UK, 2006.
14. Hazan, E. Introduction to online convex optimization. *Found. Trends Optim.* **2016**, *2*, 157–325.
15. Siegmund, D.; Yakir, B. Minimax optimality of the Shiryayev-Roberts change-point detection rule. *J. Stat. Plan. Inference* **2008**, *138*, 2815–2825.
16. Azoury, K.; Warmuth, M. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.* **2001**, *43*, 211–246.

17. Page, E. Continuous inspection schemes. *Biometrika* **1954**, *41*, 100–115.

18. Lorden, G. Procedures for reacting to a change in distribution. *Ann. Math. Stat.* **1971**, *42*, 1897–1908.

19. Moustakides, G.V. Optimal stopping times for detecting changes in distributions. *Ann. Stat.* **1986**, *14*, 1379–1387.

20. Shiryaev, A.N. On optimum methods in quickest detection problems. *Theory Probab. Appl.* **1963**, *8*, 22–46.

21. Willsky, A.; Jones, H. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Trans. Autom. Control* **1976**, *21*, 108–112.

22. Lai, T.L. Sequential changepoint detection in quality control and dynamical systems. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 613–658.

23. Lai, T.Z. Likelihood ratio identities and their applications to sequential analysis. *Seq. Anal.* **2004**, *23*, 467–497.

24. Lorden, G.; Pollak, M. Sequential change-point detection procedures that are nearly optimal and computationally simple. *Seq. Anal.* **2008**, *27*, 476–512.

25. Pollak, M. Average run lengths of an optimal method of detecting a change in distribution. *Ann. Stat.* **1987**, *15*, 749–779.

26. Mei, Y. Sequential change-point detection when unknown parameter are present in the pre-change distribution. *Ann. Stat.* **2006**, *34*, 92–122.

27. Yilmaz, Y.; Moustakides, G.V.; Wang, X. Sequential joint detection and estimation. *Theory Probab. Appl.* **2015**, *59*, 452–465.

28. Yılmaz, Y.; Li, S.; Wang, X. Sequential joint detection and estimation: Optimum tests and applications. *IEEE Trans. Signal Process.* **2016**, *64*, 5311–5326.

29. Yilmaz, Y.; Guo, Z.; Wang, X. Sequential joint spectrum sensing and channel estimation for dynamic spectrum access. *IEEE J. Sel. Areas Commun.* **2014**, *32*, 2000–2012.

30. Vo, B.N.; Vo, B.T.; Pham, N.T.; Suter, D. Joint detection and estimation of multiple objects from image observations. *IEEE Trans. Signal Process.* **2010**, *58*, 5129–5141.

31. Tajer, A.; Jajamovich, G.H.; Wang, X.; Moustakides, G.V. Optimal joint target detection and parameter estimation by MIMO radar. *IEEE J. Sel. Top. Signal Process.* **2010**, *4*, 127–145.

32. Baygun, B.; Hero, A.O. Optimal simultaneous detection and estimation under a false alarm constraint. *IEEE Trans. Inf. Theory* **1995**, *41*, 688–703.

33. Moustakides, G.V.; Jajamovich, G.H.; Tajer, A.; Wang, X. Joint detection and estimation: Optimum tests and applications. *IEEE Trans. Inf. Theory* **2012**, *58*, 4215–4229.

34. Kotlowski, W.; Grünwald, P. Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In Proceedings of the Conference on Learning Theory (COLT), Budapest, Hungary, 9–11 July 2011; pp. 457–476.

35. Anava, O.; Hazan, E.; Mannor, S.; Shamir, O. Online learning for time series prediction. In Proceedings of the Conference on Learning Theory (COLT), Princeton, NJ, USA, 12–14 June 2013; pp. 1–13.

36. Wald, A.; Wolfowitz, J. Optimum character of the sequential probability ratio test. *Ann. Math. Stat.* **1948**, *19*, 326–339.

37. Barron, A.; Sheu, C.H. Approximation of density functions by sequences of exponential families. *Ann. Stat.* **1991**, *19*, 1347–1369.

38. Wainwright, M.J.; Jordan, M.I. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **2008**, *1*, 1–305.

39. Beck, A.; Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.* **2003**, *31*, 167–175.

40. Nemirovskii, A.; Yudin, D.; Dawson, E. *Problem Complexity and Method Efficiency in Optimization*; Wiley: Hoboken, NJ, USA, 1983.

41. Shalev-Shwartz, S. Online learning and online convex optimization. *Found. Trends® Mach. Learn.* **2012**, *4*, 107–194.

42. The Implementation of the Code. Available online: http://www2.isye.gatech.edu/~yxie77/one-sample-update-code.zip (accessed on 6 February 2018).

43. Agarwal, A.; Duchi, J.C. Stochastic Optimization with Non-i.i.d. Noise. 2011. Available online: http://opt.kyb.tuebingen.mpg.de/papers/opt2011_agarwal.pdf (accessed on 6 February 2018).

44. Alqanoo, I.M. *On the Truncated Distributions within the Exponential Family*; Department of Applied Statistics, Al-Azhar University—Gaza: Gaza, Gaza Strip, 2014.

45. Xie, Y.; Siegmund, D. Sequential multi-sensor change-point detection. *Ann. Stat.* **2013**, *41*, 670–692.

46. Siegmund, D.; Yakir, B.; Zhang, N.R. Detecting simultaneous variant intervals in aligned sequences. *Ann. Appl. Stat.* **2011**, *5*, 645–668.

47. Lipster, R.; Shiryayev, A. *Theory of Martingales*; Springer: Dordrecht, The Netherlands, 1989.

48. Duchi, J.; Shalev-Shwartz, S.; Singer, Y.; Chandra, T. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In Proceedings of the International Conference on Machine learning (ICML), Helsinki, Finland, 5–9 June 2008; ACM: New York, NY, USA, 2008; pp. 272–279.

49. Wang, Y.; Mei, Y. Large-scale multi-stream quickest change detection via shrinkage post-change estimation. *IEEE Trans. Inf. Theory* **2015**, *61*, 6926–6938.