Peer-reviewed version available at Entropy 2018, 20, 108; doi:10.3390/e20020108

Article

On near optimality of one-sample update for joint detection and estimation

Yang Cao[1], Liyan Xie[1], Yao Xie[1]*, Huan Xu[1]

[1] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology; {caoyang, lxie49}@gatech.edu, {yao.xie, huan.xu}@isye.gatech.edu

* Correspondence: yao.xie@isye.gatech.edu

Abstract: Sequential hypothesis test and change-point detection when the distribution parameters are unknown is a fundamental problem in statistics and machine learning. We show that for such problems, detection procedures based on sequential likelihood ratios with simple one-sample update estimates such as online mirror descent are nearly second-order optimal. This means that the upper bound for the algorithm performance meets the lower bound asymptotically up to a log-log factor in the false-alarm rate when it tends to zero. This is a blessing, since although the generalized likelihood ratio(GLR) statistics are optimal theoretically, but they cannot be computed recursively, and their exact computation usually requires infinite memory of historical data. We prove the nearly second-order optimality by making a connection between sequential analysis 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. Two related sequential analysis problems commonly studied are sequential hypothesis testing and sequential change-point detection [1]. They arise from various applications including online anomaly detection, statistical quality control, biosurveillance, financial arbitrage detection and network security monitoring (see, e.g., [2,3]).

We are interested in joint estimation and detection in sequential analysis, which occurs when there are unknown parameters for data distribution. For instance, in change-point detection, given a sequence of samples $X_1, X_2, \ldots$, a common assumption is that they are i.i.d. with certain distribution f_θ parameterized by θ, and the values of θ are different before and after the change-point. One can assume that before the change, the parameter value is θ_0. This is reasonable since, in various settings, there is a relatively large amount of background data. Thus, the parameter θ in the normal state can be estimated with good accuracy. After the change, the value of the parameter switches to an *unknown* value, and it represents an anomaly or novelty that needs to be discovered.

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

Consider change-point detection with unknown parameters. A commonly used change-point detection method is the so-called CUSUM procedure [3]. It 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. Initiate

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 efficient*, since it does not need to store the historical data and only needs to record the value of W_t. However, one possible issue with CUSUM is the choice of the post-change parameter θ_1. In practice, it is usually chosen to represent the "smallest" change-of-interest. However, this choice is somewhat subjective. In the multi-dimensional setting, it is hard to define what the "smallest" change would mean. Moreover, when the assume parameter θ_1 deviates significantly from the true parameter value, CUSUM may suffer a severe performance degradation [4].

An alternative approach is the Generalized Likelihood Ratio (GLR) statistic [5]. The GLR statistic finds the maximum likelihood estimate (MLE) of the post-change parameter and plugs it back to the likelihood ratio to form the detection statistic. To be more precise, for each hypothetical change-point location k, the corresponding post-change samples are $\{X_{k+1}, \ldots, X_t\}$. Using these samples, one can form the MLE denoted as $\hat{\theta}_{k,t}$. 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 [6], and it is particularly useful when the post-change parameter may vary from one situation to another. However, a drawback of GLR statistic is that it is *not memory efficient* and it cannot be computed recursively. Moreover, when there is a constraint on the maximum likelihood estimator (such as sparsity), MLE cannot have closed-form solution; one has to store the historical data, and re-estimates $\hat{\theta}_{k,t}$ whenever there is new data. As a remedy, the window-limited GLR is usually considered, where one only keeps the past w samples, and restrict the maximization over k to be over $(t-w, t]$. However, even with window-limited GLR, one still has to re-estimate $\hat{\theta}_{k,t}$ using historical data whenever the new data are added.

In practice, rather than CUSUM or GLR, various one-sample update schemes are used especially in machine learning literature. The one-sample update schemes perform *online estimates* of the unknown parameter, and plug the estimates into the likelihood ratio statistic to perform detection. The one-sample update takes the form of $\hat{\theta}_t = h(X_t, \hat{\theta}_{t-1})$ for some function h that uses only the most recent data and the previous estimate. Some examples of one-sample estimate schemes include online gradient descent and online mirror descent (similar scheme has been used in [7,8]). The one-sample update enjoys efficient computation, as the information from the new data can be incorporated via low computational cost update such as mirror descent, which even has closed-form solution in some cases. It is also memory efficient since the update only needs the most recent sample. Such estimator may not correspond to the exact MLE, but they tend to have good performance. An important question remains to be answered: *how much performance do we lose by using one-sample update schemes rather than the exact GLR?*

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 or response to external events or system "shocks" [9]. 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 [10], we model discrete events data using network point processes, which capture the influence between users through an *influence matrix*. We then cast the problem as detecting changes in influence matrix. We assume 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's due to an anomaly, and it has to be estimated online. Due to computational burden and memory constraint, since the scale

of the network can be large, we do not want to store the entire historical data and rather compute the statistic in real-time. In [10], we develop a one-sample update scheme to estimate the influence matrix and then form the likelihood ratio detection statistic. However, theoretical performance of such one-sample update schemes has not been well-understood.

1.3. Contributions

This paper aims to address the above question by proving the nearly second-order optimality of simple one-sample update schemes for sequential hypothesis test and change-point detection. The nearly second-order optimality [3] means that the upper bound for performance matches the lower bound up to a log-log factor. In particular, we consider likelihood ratios with plug-in online mirror descent estimator. Our approach generalizes the non-anticipating estimator framework [11] from detecting Gaussian mean shift to the exponential family with constrained parameters. Here we focus on online mirror-descent, but the result can be generalized to other schemes such as the online gradient descent. The proof leverages the logarithmic regret property of online mirror descent and the lower bound established in statistical sequential analysis literature [3,12]. Synthetic examples validate the performances of one sample update schemes.

The contributions of the paper are summarized as follows

- We provide a general upper bound for 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 bound of the estimator. This establishes an interesting linkage between *sequential analysis* and *online convex optimization*[1].
- Using our upper bound and existing lower bound, we show that the one-sample update schemes are nearly second-order optimal for the exponential family. Moreover, numerical examples verify the good performance of one-sample update schemes. They can perform better and are more robust than the likelihood ratio methods with pre-specified parameters (e.g., CUSUM for change-point detection). Moreover, they are computationally efficient alternatives of GLR statistic (which requires storing infinite samples) and cause little performance loss relative to GLR.

The comparison of three approaches is summarized in Table 1.

Table 1. Comparison of three approaches.

	Memory Efficiency	Computation Efficiency	Robust Performance
Likelihood ratio with pre-specified parameters: SPRT/CUSUM	✓	✓	
Generalized likelihood ratio (GLR) with exact MLE			✓
One-sample update schemes	✓	✓	✓

[1] Although both fields, sequential analysis and online convex optimization, study sequential data, the precise connection between them is not clear, partly because the performance metrics are different: the former concerns with the tradeoff between false-alarm-rate and detection delay, whereas the latter focuses on bounding the cumulative loss incurred by the sequence of estimators through regret bound [13,14].

1.4. Literature and related work

Sequential analysis 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 [15] with first-order asymptotic optimality [16] and exact optimality [17] in the minimax sense, the Shiryayev-Roberts (SR) procedure [18], which can be derived based on a Bayesian principle and it enjoys various optimality. Both CUSUM and SR procedures rely on likelihood ratios between the specified pre-change and post-change distributions.

The GLR [6,19] statistic enjoys certain optimality properties, but it can not be computed recursively in most cases [43]. To address the infinite memory issue, [6,20] studied the window-limited GLR procedure. Another approach aiming to address the issue is called the Shiryayev-Roberts-Robbins-Siegmund (SRRS) procedure [11]. The main idea of SRRS dates back to the power one sequential test [21]: instead of plugging in the MLE obtained using all samples up to the current moment as done in the GLR procedure, the SRRS procedure uses a sequence of non-anticipating estimators. The non-anticipating estimators are formed by dropping the most recent sample (thus the name "non-anticipating"). The advantage is that the test statistic can be computed recursively. However, there is a small loss of performance which can be bounded. The original SRRS procedure [21] was developed for Gaussian when the post-change mean is unknown. Our work extends it to the general exponential family via the adaptive SRRS (ASR) procedure. Our non-anticipating estimator is also different from the original SRRS [21] in that SRRS still uses exact MLE estimated from all but the most recent sample, whereas our estimator only approximates the MLE using one-sample update schemes.

With unknown parameters, [22] developed a modified SR procedure by introducing a prior distribution to the unknown parameters; however, the resulted detection statistic is hard to compute recursively since the prior is not a conjugate. The more recent work [23] and [24] study joint detection and estimation problem of a specific form: a linear scalar observation model with Gaussian noise, and under the alternative hypothesis there is *an unknown multiplicative parameter*. This problem arises from many applications such as spectrum sensing [25], image observations [26], MIMO radar [27], etc. [23] 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 [24] generalizes the results. There are also other approaches solving the joint detection-estimation problem using multiple hypotheses testing [26,28] and Bayesian formulation [29]. Our work differs from the above in that we consider the general form of joint detection and estimation problem, where the unknown parameter θ shows up generally as the parameter of the exponential family. Moreover, we do not aim to find the exact optimal solution. Instead, we find whether using the computationally efficient one-sample estimator for detection loses much performance.

Related work using online convex optimization for anomaly detection include [7], which develops an efficient detector for the exponential family using online mirror descent and proves a logarithmic regret bound, and [8], 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 [7,8,13], sequential prediction of individual sequence with the logarithm loss [30,31], online prediction for time series [32], and sequential NML (SNML) prediction [31] 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 optimality.

On a high level, the problem of joint detection and estimation is also related to universal source coding [33,34] or Minimum Description Length (MDL) [35,36]. In universal source coding, the goal is to minimize the cumulative Kullback-Leibler (KL) loss.

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: sequential hypothesis test and sequential change-point detection. The detection statistic relies on a sequence estimators $\{\hat{\theta}_t\}$ constructed using online mirror descent. The online mirror descent 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 [6], 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 and present our non-anticipating estimator based on the one-sample estimate.

2.1. Sequential hypothesis test

Consider null hypothesis $\mathsf{H}_0 : \theta = \theta_0$ versus the alternative $\mathsf{H}_1 : \theta \neq \theta_0$. Hence the parameter under the alternative distribution is unknown. The classic approach to solve this problem is the sequential probablity-ratio test (SPRT) [37]: at each time, given samples $\{X_1, X_2, \ldots, X_t\}$, the decision is either to accept H_0, accept H_1, or taking more samples if neithter hypotheses can be resolved confidently. Here, we introduce *modified 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 likelihood ratio 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

$$\Lambda_t = \Lambda_{t-1} \cdot f_{\hat{\theta}_{i-1}}(X_i) / f_{\theta_0}(X_i).$$

Moreover, it has a martingale property due to the non-anticipating nature of the estimator: $\mathbb{E}_{f_{\theta_0}}[\Lambda_t] = \mathbb{E}_{f_{\theta_0}}[\Lambda_{t-1}]$. 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 resolve the two hypotheses using as few samples as possible under the type-I error constraint.

2.2. Sequential change-point detection

A problem related to sequential hypothesis test is sequential change-point detection. Due to its importance in applications and different performance metrics, sequential change-point detection is usually studied separately. A change may occur at an unknown time ν which changes the underlying

distribution of the data. One would like to detect such a change as quickly as possible. Formally, change-point detection can be cast into the following hypothesis test:

$$\begin{aligned} \mathsf{H}_0 &: X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} f_{\theta_0}, \\ \mathsf{H}_1 &: X_1, \ldots, X_\nu \overset{\text{i.i.d.}}{\sim} f_{\theta_0}, \quad X_{\nu+1}, X_{\nu+2}, \ldots \overset{\text{i.i.d.}}{\sim} f_{\theta}, \end{aligned} \tag{4}$$

Here we assume θ is unknown, and it represents the anomaly. The goal is to detect the change as quickly as possible after it occurs under the false alarm constraint.

We will consider likelihood ratio based detection procedures adapted from two types of existing ones, which we call 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 SPRT. Base on this, the likelihood ratio at time t for a hypothetical change-point location k is given by

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

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:

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

where b is a pre-specified threshold.

Similarly, by replacing the maximization in (6) with summation, we obtain the following ASR procedure [11], which can be interpreted as a Bayesian statistic similar to the Shiryaev-Roberts procedure.

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

where b 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 2.3.

2.3. Online mirror descent (OMD) for non-anticipating estimators

Next, we discuss how to construct the non-anticipating estimators $\{\hat{\theta}_t\}_{t \geq 1}$ in (1), and $\{\hat{\theta}_{k,t}\}, 1 \leq k < t$ in (5) using online mirror descent (OMD). OMD is a generic procedure for solving the online convex optimization problem (OCP). Our problem of finding maximum likelihood estimator can be cast into an OCP with the loss function being the negative log-likelihood $\ell_t(\theta) := -\log f_\theta(X_t)$.

The main idea of OMD is the following. At each time step, the estimator $\hat{\theta}_{t-1}$ is updated using the new sample X_t, by balancing the tendency to stay close to the previous estimate, against the tendency to move in the direction of the greatest local decrease of the loss function. For the loss function defined above, a sequence of OMD estimator is constructed by

$$\hat{\theta}_t = \arg\min_{u \in \Gamma} [u^\intercal \nabla \ell_t(\hat{\theta}_{t-1}) + \frac{1}{\eta_i} B_\Phi(u, \hat{\theta}_{t-1})]. \tag{9}$$

Here $\Gamma \subset \Theta_\sigma$ is a closed convex set, which is problem-specific and encourages certain parameter structure such as sparsity. Similarly, $\hat{\theta}_{k,t}$ can be constructed via OMD for sequential change-point detection.

There is an equivalent form of OMD, presented as the original formulation in [42]. 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 (9) is proved in [41]. We adopt this approach when deriving our algorithm and follow the same strategy as [7]. Our algorithm is presented in Algorithm 1.

A standard performance metric for OCP is *regret*. The regret is the difference between the total cost that an online algorithm has incurred relatively to that of the best fixed decision in hindsight. Given samples $X_1, \ldots, X_t$, the regret for a sequence of estimators $\{\hat{\theta}_i\}_{i=1}^t$ is defined as

$$\mathcal{R}_t = \sum_{i=1}^{t} \{-\log f_{\hat{\theta}_{i-1}}(X_i)\} - \inf_{\tilde{\theta}\in\Theta} \sum_{i=1}^{t} \{-\log f_{\tilde{\theta}}(X_i)\}. \tag{10}$$

For strongly convex loss function, the regret of many OCP algorithms, including the online mirror descent, has the property that $R_n \leq C \log n$ for some constant C (depend on f_θ and Θ_σ) and any positive integer n [8,38]. Note that for exponential family, the loss function is the negative log-likelihood function, which is strongly convex over Θ_σ. Hence, we have the logarithmic regret property.

2.4. Exponential family

In this paper, we focus on f_θ being the exponential family for the following reasons: (i) exponential family [8] represents a very rich class of parametric and even many nonparametric statistical models [39]; (ii) the negative log-likelihood function for exponential family $-\log f_\theta(x)$ is convex, and this allows us to perform online convex optimization with nice theoretical properties. Some useful properties of the exponential family are briefly summarized below, and full proofs can be found in [8].

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

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

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

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

and

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

Note that $-\log f_\theta(x)$ is σ-strongly convex over Θ_σ. Its gradient corresponds to $\nabla\Phi(\theta) = \mathbb{E}_\theta[\phi(X)]$, and the Hessian $\nabla^2\Phi(\theta)$ corresponds to the covariance matrix of the vector $\phi(X)$. Due to this property, since the covariance matrix is positive semidefinite, $\Phi(\theta)$ 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 [40]. The Legendre-Fenchel dual Φ^* is defined as

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

The mappings $\nabla\Phi^*$ is an inverse mapping of $\nabla\Phi$ [41]. Moreover, if Φ is a strongly convex function, then $\nabla\Phi^* = (\nabla\Phi)^{-1}$.

A general measure of proximity used in online mirror descent is the so-called *Bregman divergence* B_F, which is a nonnegative function induced by a Legendre function F (see, e.g., [8,40]) defined as

$$B_F(u,v) := F(u) - F(v) - \langle \nabla F(v), u - v\rangle. \tag{11}$$

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, IntΘ as be the interior of 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

$$I(\theta_1,\theta_2) = \mathbb{E}_{\theta_1}\left[\log(f_{\theta_1}(X)/f_{\theta_2}(X))\right]. \tag{12}$$

It can be shown that, for exponential family, $I(\theta_1,\theta_2) = \Phi(\theta_2) - \Phi(\theta_1) - (\theta_2-\theta_1)^\intercal\nabla\Phi(\theta_1)$. Using the definition (11), this means that B_Φ

$$B_\Phi(\theta_1,\theta_2) := I(\theta_2,\theta_1)$$

is a Bregman divergence. This property is quite useful to constructing mirror descent estimator for the exponential family [41,42].

Algorithm 1 Online mirror-descent for maximum likelihood 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 of strictly positive step-sizes $\{\eta_t\}$.

$\hat\theta_0 = \theta_0, \Lambda_0 = 1$. {Initialization}
for all $t = 1,2,\ldots,$ **do**
  Acquire a new observation X_t
  Compute loss $\ell_t(\hat\theta_{t-1}) := -\log f_{\hat\theta_{t-1}}(X_t) = \Phi(\hat\theta_{t-1}) - \hat\theta_{t-1}^\intercal\phi(X_t)$
  Compute likelihood ratio $\Lambda_t = \Lambda_{t-1}\times f_{\hat\theta_{t-1}}(X_t)/f_{\theta_0}(X_t)$
  $\hat\mu_{t-1} = \nabla\Phi(\hat\theta_{t-1})$, $\hat\mu_t = \hat\mu_{t-1} - \eta_t(\hat\mu_{t-1} - \phi(X_t))$ {Dual update}
  $\tilde\theta_t = (\nabla\Phi)^*(\hat\mu_t)$
  $\hat\theta_t = \arg\min_{u\in\Gamma} B_\Phi(u,\tilde\theta_t)$ {Projected primal update}
end for
return $\{\hat\theta_t\}_{t\ge1}$ and $\{\Lambda_t\}_{t\ge1}$.

3. Nearly second-order optimality of one-sample update procedures

Below we prove the *nearly second-order optimality* of the one-sample update scheme based on OMD. More precisely, the nearly second-order 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). In particular, we show that the performance of $\tau(b)$ for sequential hypothesis testing, $T_{\mathrm{ACM}}(b)$ and $T_{\mathrm{ASR}}(b)$ for sequential change-point detection setting, obtain the known lower bounds established in the statistical sequential analysis literature up to a log-log factor.

We first introduce some necessary notations. Denote $\mathbb{P}_{\theta,\nu}$ and $\mathbb{E}_{\theta,\nu}$ the probability measure and 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-field generated by $X_1, \ldots, X_t$ for $t \geq 1$.

3.1. Sequential hypothesis test

The two standard performance metrics are the the type-I error (false detection probability), which is defined for sequential hypothesis testing as $\mathbb{P}_\infty(\tau(b) < \infty)$, and the expected number of samples needed to reject the null $\mathbb{E}_{\theta,0}[\tau(b)]$. Since it is possible to take infinite samples, the power of the test in (3) is one, and the type-II error is zero. 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 type-I error to be below a certain level.

Intuitively, a reasonable sequence of estimator $\{\hat{\theta}_t\}$ should move closer to the true parameter θ as we collect more data. This is reflected by the following regularity condition (similar assumption has been made in (5.84) in [3])

$$\sum_{t=1}^{\infty} (\mathbb{E}_{\theta,0}[I(\theta, \hat{\theta}_t)])^r < \infty, \tag{13}$$

for some constant $r \geq 1$ that characterizes the convergence rate of $\{\hat{\theta}_t\}$. A larger r means a slower convergence rate. This is a mild assumption that can be obtained by many estimators such as OMD.

Our main result is the following. As has been observed by [43], there is a loss in the statistical efficiency by using one-sample update estimator $\{\hat{\theta}_t\}$, relative to the GLR approach using the entire sample in the past $(X_1, \ldots, X_t)$. The theorem below shows that this loss due to one-sample update corresponds to the expected regret of the estimators $\{\hat{\theta}_t\}$.

Theorem 1 (Upper bound for OCD based SPRT)**.** *Given a sequence of estimator $\{\hat{\theta}_t\}_{t\geq 1}$ generated by OCD, with $\hat{\theta}_0 = \theta_0$. When (13) holds, as $b \to \infty$,*

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

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 invoking the Wald's Theorem for the stopped process.

Note that in the statement of the Theorem, $\tau(b)$, the stopping time, appears on both sides of the inequality. This is not an issue since the expected sample size $\mathbb{E}_{\theta,0}[\tau(b)]$ can be bounded, and it 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 $\mathbb{E}_{\theta,0}[R_n] \leq C \log n$ for any positive integer n. Equipped with this regret bound, we obtain the following Corollary 1.

Corollary 1. *For a sequence of estimators with a logarithmic expect regret bound such that $\mathbb{E}_{\theta,0}[\mathcal{R}_n] \leq C \log n$ for any positive integer n and some constant $C > 0$, when (13) holds, 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 $b \to \infty$.

Moreover, we can obtain an upper bound on the type-I error of test $\tau(b)$.

Lemma 1 (Type-I error). *For a sequence of estimators* $\{\hat{\theta}_t\}_{t\geq 0}$, $\hat{\theta}_t \in \Theta$, *given threshold* b, $\mathbb{P}_\infty(\tau(b) < \infty) \leq \exp(-b)$.

Lemma 1 sheds some lights on how to choose an appropriate b. One can choose $b = \log(1/\alpha)$ to control the type-I error to be less than α.

Leveraging an existing lower bound for general SPRT presented in Section 5.5.1.1 in [3], we establish the nearly second-order optimality of OMD based SPRT as follows:

Corollary 2 (Nearly second-order optimality of OMD based SPRT). *Given a sequence of estimators* $\{\hat{\theta}_t\}$ *generated by Algorithm 1 with* $\Gamma \subset \Theta_\sigma$. *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 optimal in the sense that for any* $\theta \in \Theta_\sigma - \{\theta_0\}$, *as* $\alpha \to 0$,

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

Here, $\lesssim$ *means the inequality ignoring constants.*

The result means that, compared with any procedure (including the optimal procedure) calibrated to have a fixed type-I error less than α, our procedure incurs an at most $\log(\log(1/\alpha))$ increase in the expected sample size, which is usually a small number. For instance, for example, a usual choice in statistics is to set $\alpha = 10^{-5}$ when controling the false alarm; then $\log(\log(1/\alpha)) = 2.44$.

3.2. Sequential change-point detection

For sequential change-point detection, the two commonly used performance metrics [3] 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. Similarly, one usually choose b large enough so that ARL is larger than a pre-specified level.

We have the following theorem bounding the detection delay, by relating the CUSUM to SPRT [16] and using 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 ASR and ACM procedures. First, using martingale property of the detection statistic, we establish the lower bound for the ARL of the detection procedures, which is needed for proving Theorem 2.

Lemma 2 (ARL). *Consider the change-point detection procedure* $T_{\text{ASR}}(b)$ *in (8) and* $T_{\text{ACM}}(b)$ *in (7). For a sequence of estimators* $\{\hat{\theta}_t\}_{t\geq 0}$, $\hat{\theta}_t \in \Theta$ *generated by OMD. Given* $\gamma > 0$, *provided that* $b \geq \log\gamma$, *we have*

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

Lemma 2 shows that given a required lower bound γ for ARL, we can choose $b = \log\gamma$ to satisfy the ARL constraint. This is consistent with earlier works[11,22] which show that the smallest threshold b such that $\mathbb{E}_\infty[T_{ACM}(b)] \geq \gamma$ is approximately $\log\gamma$. Specifically, by setting $b = \rho\log\gamma$ for some $\rho \in (0,1)$, it is sufficient to ensure that the ARL to be greater than γ.

Theorem 2. *Consider the change-point detection procedure $T_{\mathrm{ASR}}(b)$ in (8) and $T_{\mathrm{ACM}}(b)$ in (7). Using a sequence of estimators $\{\hat{\theta}_t\}_{t\geq 1}$ with $\hat{\theta}_0 = \theta_0$ generated by OMD. When $b \to \infty$, if (13) holds, we have that*

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

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.

Combing the upper bound in Theorem 2 with an existing lower bound for the EDD of SRRS procedure in [12], we obtain the following corollary.

Corollary 3 (Nearly second-order optimality of ACM and ASR)**.** *Assume that the estimators used in the stopping times are generated with respect to Algorithm 1. 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 optimal in the sense that for any $\theta \in \Theta - \{\theta_0\}$*

$$\begin{aligned}&\sup_{\nu\geq 1} \mathbb{E}_{\theta,\nu}[T_{\mathrm{ASR}}(b) - \nu + 1 \mid T_{\mathrm{ASR}}(b) \geq \nu] \\ &- \inf_{T(b)\in S(\gamma)} \sup_{\nu\geq 1} \mathbb{E}_{\theta,\nu}[T(b) - \nu + 1 \mid T(b) \geq \nu] = O(\log\log\gamma).\end{aligned} \tag{17}$$

Similar expression holds for $T_{\mathrm{ACM}}(b)$.

Comparing (17) 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 [16].

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 in the showing of the logarithmic expected regret property. This is also useful in showing how the assumptions required by Corollary 1 are satisfied. The following result comes as a modification of [13].

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 $\hat{\theta}_i = \tilde{\theta}_i$ for any $i \geq 1$. Then for any $\theta_0 \in \Theta$ and $t \geq 1$,*

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

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

Next, we demonstrate how to use Theorem 3 by a concrete example with multivariate normal distribution, $\{\mathcal{P}_\theta, \theta \in \Theta\}$ with unknown mean parameter θ, and known covariance matrix I_d (I_d is a $d \times d$ identity matrix), denoted by $\mathcal{N}(\theta, I_d)$. Here $\phi(x) = x$, $dH(x) = (1/\sqrt{|2\pi I_d|}) \cdot \exp(-x^\intercal x/2)$, $\Theta = \Theta_\sigma = \mathbb{R}^d$ for any $\sigma < 2$, $\Phi(\theta) = (1/2)\theta^\intercal\theta$, $\mu = \theta$ and $\Phi^*(\mu) = (1/2)\mu^\intercal\mu$, where $|\cdot|$ denotes the determinant of a matrix, and H is a probability measure under which the sample follows $\mathcal{N}(0, I_d)$). When the covariance matrix is known to be some $\Sigma \neq I_d$, one can "whiten" the vectors by multiplying $\Sigma^{-1/2}$ to obtain the situation here.

Corollary 4 (Upper bound for expected regret bound, 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 $\hat{\theta}_t = \arg\min_{u\in\Gamma} B_\Phi(u, \tilde{\theta}_t) = \tilde{\theta}_t$. The the regret bound can be written as

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

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

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

Combining above, we have

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

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

Using similar calculation, 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. Proofs for Bernoulli distribution and Gamma distribution can be found in [13]. Proof of OCM for covariance matrix in multivariate normal can be found in [44]. A more general solution can be found in the Theorem 3 in [8], which however requires stronger conditions.

4. Synthetic 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.

4.1. Detecting sparse mean-shift of multivariate normal distribution

We consider detecting the emergence of a sparse mean vector in multivariate normal distribution. Sparse mean shift detection is of particular interest in sensor network or DNA sequence detection. In these settings usually only a small part of entries of the post-change mean parameter are non-zero [46,47]. Below, $\|\cdot\|_2$ means the ℓ_2 norm in $\mathbb{R}^d$, $\|\cdot\|_1$ means the ℓ_1 norm, $\|\cdot\|_0$ means the ℓ_0 norm defined as the number of non-zero entries.

In this case, the Bregman divergence is equivalent to the KL divergence and is given by $B_\Phi(\theta_1, \theta_2) = I(\theta_2, \theta_1) = \|\theta_1 - \theta_2\|_2^2/2$. Equipped with this Bregman divergence, the projection onto Γ in Algorithm 1 is just a Euclidean projection onto a convex set. In many cases, the projection can be implemented efficiently. An important and useful case is $\Gamma = \{\theta : \|\theta\|_1 \le s\}$, and s is a prescribed radius of the ℓ_1 ball. The projection onto ℓ_1 ball can be obtained via simple soft-thresholding [45]. This encourages sparse post-change mean, and Γ can be viewed as the convex relaxation of $\{\theta : \|\theta\|_0 \le s\}$.

Assume that the initial samples have been normalized by subtracting mean and dividing the standard deviation, therefore, the pre-change distribution is $\mathcal{N}(0, I_d)$. To compare the performance of different procedures, we first use simulations to choose the threshold b's such that the ARLs of the procedures are all 10000. Note that ARL is an increasing function of b so this can be done by a simple bisection. Two benchmark procedures are CUSUM and GLR. For CUSUM procedure, we specify a nominal post-change mean, which is an all-one vector. Our procedures are $T_{ASR}(b)$ and $T_{ACM}(b)$ with $\Gamma = \mathbb{R}^d$ and $\Gamma = \{\theta : \|\theta\|_1 \le s\}$. In the following experiments, we run 10000 Monte Carlo trials to obtain each simulated EDD.

In the experiments, we set $d = 20$. The post-change distributions are $\mathcal{N}(\theta, I_d)$, where $100p\%$ entry of θ is 1 and others are 0, the location of nonzero entries are random. Table 2 shows the EDDs versus the proportion p of nonzero entries of post-change parameter θ. Note that our procedures incur little performance loss compared with GLR procedure and CUSUM procedure. Notably, $T_{ACM}(b)$ with $\Gamma = \{\theta : \|\theta\|_1 \le 5\}$ performs almost the same as the GLR procedure and much better than the CUSUM procedure when p is small. This also shows the advantage of projection when the true parameter is sparse.

Table 2. Comparison of one-sample update schemes versus the traditional CUSUM and GLR methods for detecting sparse mean-shift. Below, "CUSUM": CUSUM procedure with pre-specified all-one vector as post-change parameter; "GLR": GLR procedure; "ASR": $T_{\text{ASR}}(b)$ with $\Gamma = \mathbb{R}^d$; "ACM": $T_{\text{ACM}}(b)$ with $\Gamma = \mathbb{R}^d$; "ASR-L1": $T_{\text{ASR}}(b)$ with $\Gamma = \{\theta : \|\theta\|_1 \le 5\}$; "ACM-L1": $T_{ACM}(b)$ with $\Gamma = \{\theta : \|\theta\|_1 \le 5\}$. p is the proportion of non-zero entries in θ. The value for each point is averaged over 10000 Monte Carlo trials. For each point, 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
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-ℓ1	45.81	19.94	12.45	8.92	6.97	5.89
ACM-ℓ1	19.24	10.17	7.51	6.11	5.41	4.92

4.2. 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 $i \notin \mathcal{E}$. Moreover, 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 $i = 1, \ldots, d$. We select threshold b's such that the ARLs are all equal to 10000.

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 GLR procedure, our methods have a small performance loss, and the loss is almost negligible as n approaches to $d = 190$. Moreover, in implementation, our methods are much faster than GLR procedure since the computational complexity of updating the statistic is $O(t)$, compared with the $O(t^2)$ in GLR procedure.

Table 3. Comparison of EDDs in detecting change of communication-rate in a network. The results are obtained from 10000 Monte Carlo trials. For each number, the standard deviation is less than one half of the number.

	$n = 78$	$n = 100$	$n = 120$	$n = 150$	$n = 170$	$n = 190$
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$\phi(x) = x$, $dH(x) = 1$, $\Phi(\theta) = \log(1+\exp(\theta))$, $\mu = \exp(\theta)/(1+\exp(\theta))$ and $\Phi^*(\mu) = \mu\log\mu + (1-\mu)\log(1-\mu)$.

5. Conclusion

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.

Acknowledgments: This research was supported in part by National Science Foundation (NSF) NSF CCF-1442635, CMMI-1538746, NSF CAREER CCF-1650913 to Yao Xie.

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

1. Siegmund, D. *Sequential analysis: tests and confidence intervals*; Springer-Verlag, 1985.
2. Siegmund, D. Change-points: From sequential detection to biologu and back. *Sequential analysis* **2013**.
3. Tartakovsky, A.; Nikiforov, I.; Basseville, M. *Sequential analysis: Hypothesis testing and changepoint detection*; CRC Press, 2014.

4. Granjon, P. The CuSum algorithm-a small review **2013**.
5. Basseville, M.; Nikiforov, I.V.; others. *Detection of abrupt changes: theory and application*; Vol. 104, Prentice Hall Englewood Cliffs, 1993.
6. Lai, T.Z. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory* **1998**, *44*, 2917–2929.
7. Raginsky, M.; F, R.M.; Silva, J.; Willett, R. Sequential probability assignment via online convex programming using exponential families. IEEE International Symposium on Information Theory. IEEE, 2009, pp. 1338–1342.
8. Raginsky, M.; Willet, R.; Horn, C.; Silva, J.; Marcia, R. Sequential anomaly detection in the presence of noise and limited feedback. *IEEE Transactions on Information Theory* **2012**, *58*, 5544–5562.
9. Peel, L.; Clauset, A. Detecting change points in the large-scale structure of evolving networks. 29th AAAI Conference on Artificial Intelligence (AAAI), 2015.
10. Li, S.; Xie, Y.; Farajtabar, M.; Verma, A.; Song, L. Detecting weak changes in dynamic events over networks. *IEEE Transactions on Signal and Information Processing over Networks* **2017**, *3*, 346–359.
11. Lorden, G.; Pollak, M. Nonanticipating estimation applied to sequential analysis and changepoint detection. *Annals of statistics* **2005**, pp. 1422–1454.
12. Siegmund, D.; Yakir, B. Minimax optimality of the Shiryayev-Roberts change-point detection rule. *Journal of Statistical Planning and Inference* **2008**, *138*, 2815–2825.
13. Azoury, K.; Warmuth, M. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning* **2001**, *43*, 211–246.
14. Hazan, E. Introduction to online convex optimization. *Foundations and Trends in Optimization* **2016**, *2*, 157–325.
15. Page, E. Continuous inspection schemes. *Biometrika* **1954**, *41*, 100–115.
16. Lorden, G. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics* **1971**, pp. 1897–1908.
17. Moustakides, G.V. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics* **1986**, pp. 1379–1387.
18. Shiryaev, A.N. On optimum methods in quickest detection problems. *Theory of Probability & Its Applications* **1963**, *8*, 22–46.
19. Lai, T.L. Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society. Series B (Methodological)* **1995**, pp. 613–658.
20. Willsky, A.; Jones, H. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic control* **1976**, *21*, 108–112.
21. Robbins, H.; Siegmund, D. The expected sample size of some tests of power one. *The Annals of Statistics* **1974**, pp. 415–436.
22. Pollak, M. Average run lengths of an optimal method of detecting a change in distribution. *The Annals of Statistics* **1987**, pp. 749–779.
23. Yilmaz, Y.; Moustakides, G.V.; Wang, X. Sequential joint detection and estimation. *Theory of Probability & Its Applications* **2015**, *59*, 452–465.
24. Yılmaz, Y.; Li, S.; Wang, X. Sequential joint detection and estimation: Optimum tests and applications. *IEEE Transactions on Signal Processing* **2016**, *64*, 5311–5326.
25. Yilmaz, Y.; Guo, Z.; Wang, X. Sequential joint spectrum sensing and channel estimation for dynamic spectrum access. *IEEE Journal on Selected Areas in Communications* **2014**, *32*, 2000–2012.
26. Vo, B.N.; Vo, B.T.; Pham, N.T.; Suter, D. Joint detection and estimation of multiple objects from image observations. *IEEE Transactions on Signal Processing* **2010**, *58*, 5129–5141.
27. Tajer, A.; Jajamovich, G.H.; Wang, X.; Moustakides, G.V. Optimal joint target detection and parameter estimation by MIMO radar. *IEEE Journal of Selected Topics in Signal Processing* **2010**, *4*, 127–145.
28. Baygun, B.; Hero, A.O. Optimal simultaneous detection and estimation under a false alarm constraint. *IEEE Transactions on Information Theory* **1995**, *41*, 688–703.
29. Moustakides, G.V.; Jajamovich, G.H.; Tajer, A.; Wang, X. Joint detection and estimation: Optimum tests and applications. *IEEE Transactions on Information Theory* **2012**, *58*, 4215–4229.
30. Cesa-Bianchi, N.; Lugosi, G. *Prediction, learning, and games*; Cambridge university press, 2006.

31. Kotlowski, W.; Grünwald, P. Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. Proc. Conference on Learning Theory (COLT), 2011, pp. 457–476.
32. O. Anava, E. Hazan, S.M.; Shamir, O. Online learning for time series prediction. Conference on Learning Theory (COLT), 2013, pp. 1–13.
33. Cover, T.M.; Thomas, J.A. *Elements of information theory*; John Wiley & Sons, 2012.
34. Cesa-Bianchi, N.; Lugosi, G. Worst-case bounds for the logarithmic loss of predictors. *Machine Learning* **2001**, *43*, 247–264.
35. Rissanen, J. *Minimum description length principle*; Wiley Online Library, 1985.
36. Barron, A.; Rissanen, J.; Yu, B. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory* **1998**, *44*, 2743–2760.
37. Wald, A.; Wolfowitz, J. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics* **1948**, pp. 326–339.
38. Agarwal, A.; Duchi, J.C. Stochastic optimization with non-i.i.d. noise **2011**.
39. Barron, A.; Sheu, C.H. Approximation of density functions by sequences of exponential families. *Annals of Statistics* **1991**, pp. 1347–1369.
40. Wainwright, M.J.; Jordan, M.I.; others. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **2008**, *1*, 1–305.
41. Beck, A.; Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters* **2003**, *31*, 167–175.
42. Nemirovskii, A.; Yudin, D.; Dawson, E. *Problem complexity and method efficiency in optimization*; Wiley, 1983.
43. Lai, T.Z. Likelihood ratio identities and their applications to sequential analysis. *Sequential Analysis* **2004**, *23*, 467–497.
44. Dasgupta, S.; Hsu, D. On-line estimation with the multivariate Gaussian distribution. *Learning Theory* **2007**, pp. 278–292.
45. Duchi, J.; Shalev-Shwartz, S.; Singer, Y.; Chandra, T. Efficient projections onto the ℓ_1-ball for learning in high dimensions. International Conference on Machine learning (ICML). ACM, 2008, pp. 272–279.
46. Xie, Y.; Siegmund, D. Sequential multi-sensor change-point detection. *The Annals of Statistics* **2013**, *41*, 670–692.
47. Siegmund, D.; Yakir, B.; Zhang, N.R. Detecting simultaneous variant intervals in aligned sequences. *The Annals of Applied Statistics* **2011**, pp. 645–668.
48. Lipster, R.; Shiryayev, A. Theory of martingales **1989**.

Appendix 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

$$\sum_{i=1}^{N} \log \frac{f_\theta(X_i)}{f_{\theta_N^*}(X_i)} = \sum_{i=1}^{N} \log f_\theta(X_i) - \sup_{\tilde{\theta} \in \Theta} \sum_{i=1}^{N} \log f_{\tilde{\theta}}(X_i) \leq 0.$$

Therefore we have

$$\mathbb{E}_{\theta,0}[S_N^\theta] \le \mathbb{E}_{\theta,0}[\sum_{i=1}^{N} \log \frac{f_{\theta_N^*}(X_i)}{f_{\hat{\theta}_{i-1}}(X_i)}] + \mathbb{E}_{\theta,0}[\sum_{i=1}^{N} \log \frac{f_{\hat{\theta}_{i-1}}(X_i)}{f_{\theta_0}(X_i)}].$$

Now consider the third term in the decomposition (A2). Similar to the proof of equation (5.109) in [3], we obtain that under the condition (13), its expectation under measure $\mathbb{P}_{\theta,0}$ is upper bounded by $b/I(\theta,\theta_0) + O(1)$ as $b \to \infty$. Then, for any positive integer n, we may further decompose the third term in (A2) as

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

where

$$M_n(\theta) = \sum_{i=1}^{n} \log \frac{f_{\hat{\theta}_{i-1}}(X_i)}{f_{\theta}(X_i)} + R_n(\theta),$$

$$R_n(\theta) = \sum_{i=1}^{n} I(\theta, \hat{\theta}_{i-1}),$$

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 [48] 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. Applying (A4), together with condition (13), we have that $n^{-1}R_n(\theta), n^{-1}M_n(\theta)$ and $n^{-1}m_n(\theta,\theta_0)$ converge to 0 almost surely. Moreover, the convergence is $\mathbb{P}_{\theta,0}$-r-quickly for $r = 1$ (For the definition of r-quick convergence, refer to Section 2.4.3 in [3]). 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 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 1-quickly convergence, 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

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

Since $\mathbb{E}_{\theta,0}[L(\epsilon)] < \infty$ for any $\epsilon > 0$, from (A5) above, we have that the third term in (A4) is upper bounded by $\tilde{b} + O(1)$.

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 (10) 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 (A4) and (A1). □

Proof of Corollary 1. Let $\alpha = (b + O(1))/I(\theta, \theta_0)$, $\beta = C/I(\theta, \theta_0)$ and $x = \mathbb{E}_{\theta,0}[\tau(b)]$. Applying Jensen's inequality, 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

$$\mathbb{E}_{\theta,0}[\tau(b)] \leq \frac{b}{I(\theta, \theta_0)} + \frac{C \log b}{I(\theta, \theta_0)}(1 + o(1)). \tag{A6}$$

□

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 proof is based on the method presented in [43] and [11].

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

$$\mathbb{P}_\infty(A \cap \{\tau(b) < \infty\}) = \mathbb{E}_Q\left[\mathbb{I}(\{\tau(b) < \infty\}) \cdot L_{\tau(b)}\right], \forall A \in \mathcal{F}_{\tau(b)}, \tag{A7}$$

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 (A7) we prove that $\mathbb{P}_\infty(\tau(b) < \infty) \leq \exp(-b)$. □

Proof of Corollary 2. Using (5.180) and (5.188) in [3], which are about asymptotic performance of open-ended tests. Since our problem is a special case of the problem in [3], 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. □

Proof of Theorem 2. From (A9), 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 [11], we have that the supreme of detection delay over all change locations is achieved by the case when change occurs at the first instance.

$$\sup_{\nu\geq 0}\mathbb{E}_{\theta,\nu}[T_{ACM}(b)-\nu \mid T_{ACM}(b)>\nu]=\mathbb{E}_{\theta,0}[T_{ACM}(b)]. \tag{A8}$$

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 (A8), 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 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 [16]. Thus, $T_{ACM}(b)\leq \tau(b)^{(1)}+1-1=\tau(b)$, and $\mathbb{E}_{\theta,0}[T_{ACM}(b)]\leq \mathbb{E}_{\theta,0}[\tau(b)]$. □

Proof of Lemma 2. 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

$$\log\left(\sum_{k=1}^{t}\Lambda_{k,t}\right)>\log\left(\max_{1\leq k\leq t}\Lambda_{k,t}\right)=\max_{1\leq k\leq t}\log\Lambda_{k,t}, \tag{A9}$$

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

$$\sum_{t=1}^{\infty} \mathbb{P}_\infty(T_{ASR}(b) \geq t) < \infty. \tag{A10}$$

(A10) 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 martingale, 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 [12] show that the minimum worst-case detection delay, given a fixed ARL level γ, is given by

$$\inf_{T(b)\in S(\gamma)} \sup_{\nu\geq 1} \mathbb{E}_{\theta,\nu}[T(b) - \nu + 1 \mid T(b) \geq \nu] = \frac{\log\gamma}{I(\theta,\theta_0)} + \frac{d\log\log\gamma}{2I(\theta,\theta_0)}(1 + o(1)). \tag{A11}$$

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 (A11) with the right-hand side of (15), we prove the corollary. □

The following derivation borrows ideas from [13]. First, we derive concise forms of the two terms in the definition of R_t in (10).

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 $t \geq 1$,*

$$\sum_{i=1}^{t} \{-\log f_{\hat{\theta}_{i-1}}(X_i)\} = \sum_{i=1}^{t} iB_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) - t\Phi^*(\hat{\mu}_t). \tag{A12}$$

Moreover, for any $t \geq 1$,

$$\inf_{\tilde{\theta}\in\Theta} \sum_{i=1}^{t} \{-\log f_{\tilde{\theta}}(X_i)\} = -t\Phi^*(\hat{\mu}), \tag{A13}$$

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

By subtracting the expressions in (A12) and (A13), 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 $\Phi^*(\mu) = \theta^\intercal \mu - \Phi(\theta)$ for any $\theta \in \Theta$. By this definition, and choosing $\eta_i = 1/i$, we have that for any $i \geq 1$

$$\begin{aligned}
&-\log f_{\hat{\theta}_{i-1}}(X_i) = \Phi(\hat{\theta}_{i-1}) - \hat{\theta}_{t-1}^\intercal \phi(X_i) = \hat{\theta}_{i-1}^\intercal(\hat{\mu}_{t-1} - \phi(X_i)) - \Phi^*(\hat{\mu}_{i-1}) = \frac{1}{\eta_i}\hat{\theta}_{i-1}^\intercal(\hat{\mu}_{i-1} - \hat{\mu}_i) - \Phi^*(\hat{\mu}_{i-1}) \\
&= \frac{1}{\eta_i}(\Phi^*(\hat{\mu}_i) - \Phi^*(\hat{\mu}_{i-1})) - \hat{\theta}_{i-1}^\intercal(\hat{\mu}_i - \hat{\mu}_{i-1}) - \frac{1}{\eta_i}\Phi^*(\hat{\mu}_i) + \left(\frac{1}{\eta_i} - 1\right)\Phi^*(\hat{\mu}_{i-1}) \\
&= \frac{1}{\eta_i} B_{\Phi^*}(\hat{\mu}_i, \hat{\mu}_{i-1}) + \frac{1}{\eta_{i-1}}\Phi^*(\hat{\mu}_{i-1}) - \frac{1}{\eta_i}\Phi^*(\hat{\mu}_i),
\end{aligned} \tag{A14}$$

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 (A14), where the second term form a telescopic series, over i from 1 to t, we have that

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

Moreover, from the definition we have that

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

Taking the first derivative of $\sum_{i=1}^{t}\{-\log f_\theta(X_i)\}$ with respect to θ and setting it to 0, we find $\hat{\mu}$, the stationary point, given by

$$\hat{\mu} = \nabla\Phi(\theta) = \frac{1}{t}\sum_{i=1}^{t}\phi(X_i).$$

Similarly, using the expression of the dual function, and plugging $\hat{\mu}$ back into the equation, we have that

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

513 □

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 (A12) by (A13), we obtain

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

The final equality is obtained by Taylor expansion. □