Nearly second-order optimality of online joint detection and estimation via one-sample update schemes

Yang Cao Georgia Tech
Liyan Xie Georgia Tech
Yao Xie Georgia Tech
Huan Xu Georgia Tech

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 change-point detection and online convex optimization and leveraging the logarithmic regret bound property of online mirror descent algorithm. Numerical examples validate our theory.

1 Introduction

Sequential analysis is a classic topic in statistics concerning *online* inference from a sequence of observations. The goal is to make a 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 [Siegmund, 1985]. They arise from various applications including online anomaly detection, statistical quality control, biosurveillance, financial arbitrage detection and network security monitoring (see, e.g., Siegmund [2013], Tartakovsky et al. [2014]).

We are interested in sequential change-point detection when there are unknown parameters for data distribution. For instance, in change-point detection, given a sequence of samples X_1, X_2, ..., 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

First, we explore the 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 [Page, 1954]. 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 prespecified 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

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

θ_1 deviates significantly from the true parameter value, CUSUM may suffer a severe performance degradation [Granjon, 2013].

An alternative approach is the Generalized Likelihood Ratio (GLR) statistic [Basseville et al., 1993]. 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 [Lai, 1998], 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 to re-estimates $\hat{\theta}_{k,t}$ and re-compute the summation $\sum_{i=k+1}^{t} \log(f_{\hat{\theta}_{k,t}}(X_i)/f_{\theta_0}(X_t))$ whenever there is new data. As a remedy, the frequently used window-limited GLR restricts the maximization over $k \in (1, t]$ to be over $k \in (t-w, t]$. However, this does not help eliminate the time of the re-computation of the summation.

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 Raginsky et al. [2009, 2012]). 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 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 [Tartakovsky et al., 2014] 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 [Lorden and Pollak, 2005] from detecting Gaussian mean shift and Gamma shape shift to the exponential family with constrained parameters. Moreover, we provide a more general framework to prove the second-order optimality beyond Gaussian assumption [Lorden and Pollak, 2008], through linking the statistical efficiency with the regret bound for the online optimization algorithm. 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 change-point detection literature [Siegmund and Yakir, 2008, Tartakovsky et al., 2014]. 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 change-point detection* 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

[1] Although both fields, sequential change-point detection and online convex optimization, study sequential data, the precise connection between them is not clear, partly because the performance metrics are different: the former concerns with the tradeoff between false-alarm-rate and detection delay, whereas the latter focuses on bounding the cumulative loss incurred by the sequence of estimators through regret bound [Azoury and Warmuth, 2001, Hazan, 2016].

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.

	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	✓	✓	✓

Table 1: Comparison of three approaches.

1.3 Literature and related work

Besides GLR procedure [Lai, 1995, 1998], another approach aiming to address the unknown post-change parameter issue is called the Shiryayev-Roberts-Robbins-Siegmund (SRRS) procedure [Lorden and Pollak, 2005]. The main idea of SRRS dates back to the power one sequential test [Robbins and Siegmund, 1974]: 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. The SRRS procedure is then extended to general exponential family [Lorden and Pollak, 2008] and the first order optimality of their procedures are well established. Even if our work also extends the original SRRS procedure [Robbins and Siegmund, 1974] to the general exponential family, two big differences should be noticed. First, our non-anticipating estimator is different from the original SRRS [Robbins and Siegmund, 1974] and the extended version [Lorden and Pollak, 2008] 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. This approximation further allows us to take the parameter structure into consideration such as the sparsity and the smoothness of parameters. Second, different with Lorden and Pollak [2005] and Lorden and Pollak [2008] of which the second-order optimality is only proved for Gaussian and Gamma distributions, our work establishes nearly second-order optimality for our procedure for all the distributions in the exponential family.

With unknown parameters, Pollak [1987] 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 [Yilmaz et al., 2015, Yılmaz et al., 2016] studies 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 [Yilmaz et al., 2014], image observations [Vo et al., 2010], MIMO radar [Tajer et al., 2010], etc. Yilmaz et al. [2015] demonstrate 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, Yılmaz et al. [2016] generalizes the results. There are also other approaches solving the joint detection-estimation problem using multiple hypotheses testing Baygun and Hero [1995], Vo et al. [2010] and Bayesian formulation Moustakides et al. [2012]. 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 Raginsky et al. [2009], which develops an efficient detector for the exponential family using online mirror descent and proves a logarithmic regret bound, and Raginsky et al. [2012], which dynamically adjusts the detection threshold to allow feedbacks about decision outcomes. 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). In Li et al. [2017], the authors develop a one-sample update scheme to estimate the influence matrix and then form the likelihood ratio detection statistic to detect changes in the social network. However, theoretical performance of such one-sample update schemes has not been well-understood.

2 Problem formulation

Assume a sequence of i.i.d. random variables $X_1, X_2, \dots$ 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 [Lai, 1998], where each $\hat{\theta}_t$ is estimated using historical samples. In the following, we present detailed descriptions of 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 probability-ratio test (SPRT) [Wald and Wolfowitz, 1948]: at each time, given samples $\{X_1, X_2, \dots, X_t\}$, the decision is either to accept H_0, accept H_1, or taking more samples if neither hypotheses can be resolved confidently. Here, we introduce *modified SPRT* with a sequence of *non-anticipating* plug-in estimators:

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

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: $\mathbb{E}_{f_{\theta_0}}[\Lambda_t|\Lambda_{t-1}] = \Lambda_{t-1}$. 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:

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

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 $\{X_k, \dots, X_t\}$; with a slight abuse of notation, the post-change parameter is estimated as

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

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 [Lorden and Pollak, 2005], 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 estimators $\{\hat{\theta}_t\}$, $\{\hat{\theta}_{k,t}\}$ are discussed later in section 2.4.

2.3 Exponential family

In this paper, we focus on f_θ being the exponential family. 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 of 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.

Two more terms used in this paper are the dual function of Φ and the Bregman divergence between two distributions. Based on Wainwright and Jordan [2008], for the exponential family the Legendre-Fenchel dual Φ^* is defined as $\Phi^*(z) := \sup_{u\in\Theta}\{u^\intercal z - \Phi(u)\}$, and the Bregman divergence induced by Φ is defined as $B_\Phi(\theta_1, \theta_2) := I(\theta_2, \theta_1)$, where $I(u, v)$ is the Kullback-Leibler (KL) divergence between $f_u(x)$ and $f_v(x)$.

2.4 Online mirror descent (OMD) for non-anticipating estimators

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 (OCO) optimization problem. Our problem of finding maximum likelihood estimator can be cast into an OCO 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

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

Here $\Gamma \subset \Theta_\sigma$ is a closed convex set, which is problem-specific and encourages certain parameter structure such as sparsity and smoothness. Similarly, $\hat{\theta}_{k,t}$ can be constructed via OMD for sequential change-point detection. The only difference is that $\hat{\theta}_{k,t}$ is computed if we use X_k as our first observation and then apply the recursive update (9) on $X_{k+1}, \ldots$ (for $\hat{\theta}_t$ we use X_1 as our first observation).

There is an equivalent form of OMD, presented as the original formulation [Nemirovskii et al., 1983]. 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 by Beck and Teboulle [2003]. We adopt this approach when deriving our algorithm and follow the same strategy as Raginsky et al. [2009]. Our algorithm is presented in Algorithm 1.

A standard performance metric for OCO 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

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

For strongly convex loss function, the regret of many OCO 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 [Agarwal and Duchi, 2011, Raginsky et al., 2012]. 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.

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

Below we prove the *nearly second-order optimality* of the one-sample update schemes. 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,

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 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: $\ell_t(\hat{\theta}_{t-1}) = \Phi(\hat{\theta}_{t-1}) - \hat{\theta}_{t-1}^\intercal \phi(X_t)$
  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 $\{\hat{\theta}_t\}_{t\geq 1}$ and $\{\Lambda_t\}_{t\geq 1}$.

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 converge to the true parameter θ as we collect more data. This is reflected by the following regularity condition (similar assumption has been made in equation (5.84) from Tartakovsky et al. [2014])

$$\sum_{t=1}^{\infty} (\mathbb{E}_{\theta,0}[I(\theta, \hat{\theta}_t)])^r < \infty, \tag{11}$$

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 Lai [2004], 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 OCO based SPRT)**.** *Given a sequence of estimator $\{\hat{\theta}_t\}_{t\geq 1}$ generated by any OCO algorithm of which the regret is $\mathcal{R}_n$ for each n, with $\hat{\theta}_0 = \theta_0$. When (11) holds, as $b \to \infty$,*

$$\mathbb{E}_{\theta,0}[\tau(b)] \leq \frac{b}{I(\theta, \theta_0)} + \frac{\mathbb{E}_{\theta,0}[\mathcal{R}_{\tau(b)}]}{I(\theta, \theta_0)} + O(1) \tag{12}$$

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.

The result of Theorem 1 is very significant for the following two reasons. First, it is very general because we obtain an upper bound for the expected number of observations needed to make correct decisions for any reasonable OCO algorithm. Second, equation (12) shows a clear connection between the classic metric for sequential hypothesis testing (left-hand side of (12)) and the classic metric for OCO algorithm (the second term on the right-hand side of (12)). This bridges the gap between the sequential hypothesis testing and the online optimization, two different but important fields.

Though the stopping time $\tau(b)$ appears on both sides of the inequality. This is not an issue since we can see later in Corollary 1 that the $\tau(b)$ in the right-hand side of (12) can be replaced by $\log b$ if the estimation algorithm has a logarithmic expected regret. This logarithmic expected regret, as shown in the supplementary materials, can be achieved by OMD for the exponential family. Specifically, Algorithm 1 can guarantee that $\mathbb{E}_{\theta,0}[R_n] \leq C \log n$ for any positive integer n.

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 (11) holds, we have

$$\mathbb{E}_{\theta,0}[\tau(b)] \leq \frac{b}{I(\theta,\theta_0)} + \frac{C\log b}{I(\theta,\theta_0)}(1+o(1)). \quad (13)$$

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. Moreover, Lemma 1 is valid generally for any non-anticipating estimators.

Leveraging an existing lower bound for general SPRT presented in Section 5.5.1.1 of Tartakovsky et al. [2014], we establish the nearly second-order optimality of OCO based SPRT as follows:

Corollary 2 (Nearly second-order optimality of OCO based SPRT)**.** *Consider 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 $C > 0$, and (11) holds. 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 $\alpha \to 0$,*

$$\mathbb{E}_{\theta,0}[\tau(b)] - \inf_{T\in C(\alpha)} \mathbb{E}_{\theta,0}[T] = O(\log(\log(1/\alpha))). \quad (14)$$

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 at most an increase in the expected sample size to make correct decisions on the order of $\log(\log(1/\alpha))$, which is usually a small number. For instance, even for a conservative choice $\alpha = 10^{-5}$ when controlling the false alarm, we have $\log(\log(1/\alpha)) = 2.44$.

3.2 Sequential change-point detection

Following the similar routine as before, we consider the sequential change-point detection problem. Here the two commonly used performance metrics [Tartakovsky et al., 2014] 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 sequential hypothesis test, 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 [Lorden, 1971] 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.

Theorem 2. *Consider the change-point detection procedure $T_{\mathrm{ASR}}(b)$ in (8) and $T_{\mathrm{ACM}}(b)$ in (7). For any sequence of estimator $\{\hat{\theta}_t\}_{t\geq 1}$ generated by any OCO algorithm of which the regret is $\mathcal{R}_n$ for each n, with $\hat{\theta}_0 = \theta_0$. As $b \to \infty$, if (11) 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.

Then, 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 3.

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

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

Lemma 2 guides us how to choose b appropriately. For example, given a required lower bound γ for ARL, one can choose $b = \log\gamma$ to satisfy the ARL constraint. Lemma 2 is valid for any non-anticipating estimators.

Combing the upper bound in Theorem 2 with an existing lower bound for the EDD of SRRS procedure from Siegmund and Yakir [2008], we obtain the following nearly second-order optimality.

Corollary 3 (Nearly second-order optimality of OCO based ACM and ASR)**.** *Consider 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 $C > 0$, and (11) holds. 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 $\theta \in \Theta - \{\theta_0\}$

$$\begin{aligned} &\sup_{\nu\geq 1} \mathbb{E}_{\theta,\nu}[T_{\mathrm{ASR}}(b) - \nu + 1 \mid T_{\mathrm{ASR}}(b) \geq \nu] \\ - &\inf_{T(b)\in S(\gamma)} \sup_{\nu\geq 1} \mathbb{E}_{\theta,\nu}[T(b) - \nu + 1 \mid T(b) \geq \nu] \quad (15) \\ = \ & O(\log\log\gamma). \end{aligned}$$

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

Similar with Corollary 2, this result means that compared with the optimal ones among all the detection procedures of which the ARLs are larger than γ, our procedure incurs at most an increase in the expected detection delay on the order of $O(\log\log\gamma)$. This number $\log\log\gamma$ is also usually small even for a very large γ. Furthermore, comparing (15) with (14), we note that the lower bound γ for the ARL plays the same role as $1/\alpha$ because $1/\gamma$ is roughly the false-alarm rate for sequential change-point detection [Lorden, 1971].

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. 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, i.e., equation (8) in the supplementary material). Therefore, in the following experiments, we define the expected detection delay (EDD) as $\mathbb{E}_{\theta,0}[T]$ for a stopping time T. In other words, we assume that the change happens at the very beginning of the sequence.

We consider detecting the sparse mean shift in multivariate normal distribution. Sparse mean shift means that only a small part of entries of the post-change mean vector are non-zero. This setting is of particular interest in sensor network or DNA sequence detection [Xie and Siegmund, 2013, Siegmund et al., 2011]. Below, $\|\cdot\|_2$ means the ℓ_2 norm, $\|\cdot\|_1$ means the ℓ_1 norm, $\|\cdot\|_0$ means the ℓ_0 norm defined as the number of non-zero entries.

In this setting, we have that $B_\Phi(\theta_1,\theta_2) = I(\theta_2,\theta_1) = \|\theta_1-\theta_2\|_2^2/2$. 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 $\Gamma = \{\theta : \|\theta\|_1 \leq s\}$ where s is a prescribed radius of the ℓ_1 ball. The projection onto ℓ_1 ball can be obtained via simple soft-thresholding [Duchi et al., 2008]. This encourages the detection of sparse mean vectors since Γ can be viewed as the convex relaxation of $\{\theta : \|\theta\|_0 \leq s\}$.

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 $\Gamma = \{\theta : \|\theta\|_1 \leq s\}$. 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 and the location of nonzero entries are deterministic. 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 $\Gamma = \{\theta : \|\theta\|_1 \leq 5\}$ 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.

	$p=0.1$	$p=0.3$	$p=0.4$	$p=0.6$
CUSUM	188.60	64.30	18.97	3.77
GLR	19.10	7.00	5.49	3.86
ASR	45.22	12.62	8.90	5.90
ACM	45.60	12.50	9.00	5.87
ASR-ℓ1	20.81	9.45	7.42	5.09
ACM-ℓ1	19.24	7.51	6.11	4.92

Table 2: Comparison of OMD based methods 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": $T_{\mathrm{ASR}}(b)$ with $\Gamma = \mathbb{R}^d$; "ACM": $T_{\mathrm{ACM}}(b)$ with $\Gamma = \mathbb{R}^d$; "ASR-L1": $T_{\mathrm{ASR}}(b)$ with $\Gamma = \{\theta : \|\theta\|_1 \leq 5\}$; "ACM-L1": $T_{ACM}(b)$ with $\Gamma = \{\theta : \|\theta\|_1 \leq 5\}$. p is the proportion of nonzero entries in θ. The value for each point is averaged over 10000 Monte Carlo trials. For each value in the table, the standard deviation is less than one half of the value.

Acknowledgements

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

References

D. Siegmund. *Sequential analysis: tests and confidence intervals*. Springer-Verlag, 1985.

D. Siegmund. Change-points: From sequential detection to biologu and back. *Sequential analysis*, (1), 2013.

A. Tartakovsky, I. Nikiforov, and M. Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014.

ES Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Pierre Granjon. The cusum algorithm-a small review. 2013.

Michèle Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs, 1993.

T.-Z. Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, 44(7):2917–2929, 1998.

M. Raginsky, R. Marcia F, J. Silva, and R. Willett. Sequential probability assignment via online convex programming using exponential families. In *IEEE International Symposium on Information Theory*, pages 1338–1342. IEEE, 2009.

M. Raginsky, R. Willet, C. Horn, J. Silva, and R. Marcia. Sequential anomaly detection in the presence of noise and limited feedback. *IEEE Transactions on Information Theory*, 58(8):5544–5562, 2012.

G. Lorden and M. Pollak. Nonanticipating estimation applied to sequential analysis and changepoint detection. *Annals of statistics*, pages 1422–1454, 2005.

Gary Lorden and Moshe Pollak. Sequential change-point detection procedures that are nearly optimal and computationally simple. *Sequential Analysis*, 27 (4):476–512, 2008.

D. Siegmund and B. Yakir. Minimax optimality of the Shiryayev-Roberts change-point detection rule. *Journal of Statistical Planning and Inference*, 138(9): 2815–2825, 2008.

K. Azoury and M. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3): 211–246, 2001.

E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

Tze Leung Lai. Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 613–658, 1995.

H. Robbins and D. Siegmund. The expected sample size of some tests of power one. *The Annals of Statistics*, pages 415–436, 1974.

M. Pollak. Average run lengths of an optimal method of detecting a change in distribution. *The Annals of Statistics*, pages 749–779, 1987.

Yasin Yilmaz, George V Moustakides, and Xiaodong Wang. Sequential joint detection and estimation. *Theory of Probability & Its Applications*, 59(3):452–465, 2015.

Yasin Yılmaz, Shang Li, and Xiaodong Wang. Sequential joint detection and estimation: Optimum tests and applications. *IEEE Transactions on Signal Processing*, 64(20):5311–5326, 2016.

Yasin Yilmaz, Ziyu Guo, and Xiaodong Wang. Sequential joint spectrum sensing and channel estimation for dynamic spectrum access. *IEEE Journal on Selected Areas in Communications*, 32(11):2000–2012, 2014.

Ba-Ngu Vo, Ba-Tuong Vo, Nam-Trung Pham, and David Suter. Joint detection and estimation of multiple objects from image observations. *IEEE Transactions on Signal Processing*, 58(10):5129–5141, 2010.

Ali Tajer, Guido H Jajamovich, Xiaodong Wang, and George V Moustakides. Optimal joint target detection and parameter estimation by mimo radar. *IEEE Journal of Selected Topics in Signal Processing*, 4(1): 127–145, 2010.

Bulent Baygun and Alfred O Hero. Optimal simultaneous detection and estimation under a false alarm constraint. *IEEE Transactions on Information Theory*, 41(3):688–703, 1995.

George V Moustakides, Guido H Jajamovich, Ali Tajer, and Xiaodong Wang. Joint detection and estimation: Optimum tests and applications. *IEEE Transactions on Information Theory*, 58(7):4215–4229, 2012.

S. Li, Y. Xie, M. Farajtabar, A. Verma, and L. Song. Detecting weak changes in dynamic events over networks. *IEEE Transactions on Signal and Information Processing over Networks*, 3(2):346–359, 2017.

Abraham Wald and Jacob Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, pages 326–339, 1948.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

A. Nemirovskii, D. Yudin, and E. Dawson. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

A. Agarwal and J. C. Duchi. Stochastic optimization with non-i.i.d. noise. 2011.

T.-Z. Lai. Likelihood ratio identities and their applications to sequential analysis. *Sequential Analysis*, 23 (4):467–497, 2004.

G. Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, pages 1897–1908, 1971.

Y. Xie and D. Siegmund. Sequential multi-sensor change-point detection. *The Annals of Statistics*, 41(2):670–692, 2013.

David Siegmund, Benjamin Yakir, and Nancy R Zhang. Detecting simultaneous variant intervals in aligned sequences. *The Annals of Applied Statistics*, pages 645–668, 2011.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *International Conference on Machine learning (ICML)*, pages 272–279. ACM, 2008.