

Minimax Robust Quickest Change Detection using Wasserstein Ambiguity Sets

Liyan Xie

School of Data Science

The Chinese University of Hong Kong, Shenzhen

Shenzhen, China

xieliyan@cuhk.edu.cn

Abstract—We study the robust quickest change detection under unknown pre- and post-change distributions. To deal with uncertainties in the data-generating distributions, we formulate two data-driven ambiguity sets based on the Wasserstein distance, without any parametric assumptions. The minimax robust test is constructed as the CUSUM test under least favorable distributions, a representative pair of distributions in the ambiguity sets. We show that the minimax robust test can be obtained in a tractable way and is asymptotically optimal. We investigate the effectiveness of the proposed robust test over existing methods, including the generalized likelihood ratio test and the robust test under KL divergence based ambiguity sets.

Index Terms—CUSUM test, Least favorable distributions, Robust change detection, Wasserstein metric

I. INTRODUCTION

Quickest change detection aims to detect a potential change-point from sequential data and is widely applicable in signal processing and statistical problems [1]–[3]. Classical approaches, such as the well-known cumulative sum (CUSUM) test [4], are usually designed for cases where the pre- and post-change distributions are exactly known. When the post-change distribution is unknown, the generalized likelihood ratio (GLR) test [5] is commonly used, in which the post-change distributions are sequentially estimated based on maximum likelihood.

However, the maximum likelihood estimate may deviate significantly from the true parameter if we only have limited data samples or the observations are contaminated [6], [7]. We aim to overcome this limitation by considering a robust quickest change detection problem by constructing *ambiguity sets* for the distribution estimates. The goal is to find the minimax robust test that minimizes the worst-case detection delay over the ambiguity sets [8]. In [6], it is proved that an exact minimax robust optimal test does not hold for the robust sequential detection problem in general. Therefore, most work focus on finding the asymptotically optimal test [9], [10].

The minimax robust change detection has been studied in [10] and [11] with two ambiguity sets that are given in a priori, for the pre- and post-change distributions, respectively. In [11], it is proved that under the joint stochastic boundedness condition on the pre- and post-change distributional ambiguity sets, the detection rule based on least favorable distributions (LFDs) are minimax robust under several performance metrics. Although the joint stochastic boundedness condition can be satisfied and verified for several classical types of ambiguity

sets, it is difficult to verify for modern types of ambiguity sets, e.g., the KL ambiguity sets. Later in [10], the problem is solved by proving a weaker condition on the ambiguity sets, and asymptotic optimal solutions are proposed. A recent work [12] studies the change detection with uncertain distributions from the Bayesian perspective by applying the uncertain likelihood ratio [13] test. However, the posterior prediction distribution cannot be calculated when the parametric model is unknown or insufficient to model the data distribution.

The main contribution of this work is a non-parametric method for minimax robust quickest change detection based on Wasserstein ambiguity sets [14]. The key advantage is that the proposed method does not require complete knowledge about pre- and post-change distributions and parametric assumptions. Moreover, the resulting LFDs from the Wasserstein ambiguity sets are proved to be efficiently solvable, and thus the proposed test can be applied to a wide range of applications.

The remainder of this paper is organized as follows. Section II details the problem set-up, including the performance criteria and the construction of the ambiguity sets. Section III derives a tractable formulation to find the LFDs and the minimax optimal test. Section IV demonstrates the proposed detection procedure using synthetic data. Section V concludes the paper with possible future directions.

II. PROBLEM SETUP

The quickest change detection problem can be formulated as follows. Given observations $\{x_t, t = 1, 2, \dots\}$ in the sample space \mathcal{X} , we aim to detect the change-point τ at which the data-generating distribution changes from μ to ν :

$$\begin{aligned} x_t &\stackrel{\text{iid}}{\sim} \mu, & t = 1, 2, \dots, \tau - 1, \\ x_t &\stackrel{\text{iid}}{\sim} \nu, & t = \tau, \tau + 1, \dots \end{aligned} \quad (1)$$

We consider the case where τ is *unknown* but is a deterministic value. An important quantity for the detection problem (1) is the Kullback-Leibler (KL) divergence defined as follows.

Definition 1 (KL divergence [15]). *The KL divergence between two probability distributions ν and μ is:*

$$\text{KL}(\nu||\mu) = \int \{\log(d\nu(x)/d\mu(x))\}d\nu(x).$$

Let $\mathcal{P}(\mathcal{X})$ denote the family of all probability distributions supported on the sample space \mathcal{X} . Assume there exists a

probability space $(\mathcal{X}, \mathcal{F}, \mathbb{P}_\tau^{\mu, \nu})$ where $\mathbb{P}_\tau^{\mu, \nu}$ denotes the probability measure when the change-point equals to τ and the pre- and post-change probability measures being μ and ν , respectively. In particular, \mathbb{P}_∞^μ and \mathbb{E}_∞^μ denote the probability and expectation when there is no change-point (i.e., $\tau = \infty$) and the pre-change distribution being μ . Similarly, \mathbb{P}_0^ν and \mathbb{E}_0^ν denote the probability and expectation when all samples are generated from the post-change distribution ν .

Our goal is to detect the unknown change-point τ as quickly as possible while at the same time keeping the false alarm rate below a pre-specified level. Usually, the detection is performed by designing a *stopping time* on the data sequence [16]. A stopping time with respect to the random data sequence $\{x_t\}_t$ is a random variable T such that for any n , the event $\{T = n\}$ belongs to the sigma-algebra generated by $\{x_1, \dots, x_n\}$.

A. Performance Criteria

We typically focus on two criteria to measure the performance of a stopping time T . One is the average run length (ARL) used to measure the average time between consecutive false alarms, defined as $\mathbb{E}_\infty^\mu[T]$. Usually we impose certain lower bound γ on the ARL and only consider the stopping times satisfying $\mathbb{E}_\infty^\mu[T] \geq \gamma$. The other criteria is the detection delay. There are two main measures for the detection delay, the Lorden's measure [17] and the Pollak's measure [18].

The Lorden's measure for detection delay is defined as the worst-case average detection delay (WADD), which is the supremum of the average delay conditioned on the worst-case historical data and change-point:

$$\text{WADD}^{\mu, \nu}(T) = \sup_{n \geq 1} \text{ess sup} \mathbb{E}_n^{\mu, \nu} [(T - n)^+ | X_1, \dots, X_{n-1}]. \quad (2)$$

A less conservative characterization of detection delay is proposed by Pollak [18] as the conditional average detection delay (CADD) conditioned on the event that $\{T \geq n\}$:

$$\text{CADD}^{\mu, \nu}(T) = \sup_{n \geq 1} \mathbb{E}_n^{\mu, \nu} [T - n | T \geq n]. \quad (3)$$

B. Uncertainty Model

Consider the case when the pre- and post-change probability measure μ and ν in (1) are *unknown*. This typically happens in real data applications, especially for data with complex structures or of high-dimensionality. To deal with the uncertainties in distributions, we construct two ambiguity sets $\mathcal{P}_{\mu_0}, \mathcal{P}_{\nu_0}$ for pre- and post-change distributions, respectively.

Assume we have a nominal distribution μ_0 and ν_0 for pre- and post-change, and the ambiguity sets $\mathcal{P}_{\mu_0}, \mathcal{P}_{\nu_0}$ are the collection of probabilities measures that are close to μ_0, ν_0 with respect to certain divergence measures $D(\cdot, \cdot)$:

$$\begin{aligned} \mathcal{P}_{\mu_0} &= \{\mu \in \mathcal{P}(\mathcal{X}) : D(\mu, \mu_0) \leq r_1\}, \\ \mathcal{P}_{\nu_0} &= \{\nu \in \mathcal{P}(\mathcal{X}) : D(\nu, \nu_0) \leq r_2\}, \end{aligned} \quad (4)$$

where $r_1, r_2 \geq 0$ are the radius parameter controlling the size of ambiguity sets. Some commonly used divergence measures $D(\cdot, \cdot)$ include the KL divergence [19], [20], Total-Variation distance [6], [21], [22], Wasserstein metric [7], [23], [24], etc.

In this paper, we consider a fully *data-driven* and *non-parametric* setting where (i) the nominal distribution is set as the *empirical distribution* from historical data, and (ii) the ambiguity sets are constructed using the *Wasserstein distance*.

In the data-driven case, suppose we have a set of training samples $\{x_1, \dots, x_{n_1}\}$ that are i.i.d. sampled from the pre-change regime, and $\{y_1, \dots, y_{n_2}\}$ that are i.i.d. sampled from the post-change regime, the nominal distribution is set as the empirical distribution of those historical samples, i.e., $\mu_0 = (\sum_{i=1}^{n_1} \delta_{x_i})/n_1$, $\nu_0 = (\sum_{i=1}^{n_2} \delta_{y_i})/n_2$, where δ_x denotes the Dirac point mass concentrated on x for each $x \in \mathcal{X}$, i.e., $\delta_x(A) = \mathbb{I}\{x \in A\}$ for any Borel measurable set and $\mathbb{I}\{\cdot\}$ is the indicator function.

Remark 1. *The historical data used here is additional available data before we start the detection procedure for problem (1). If we have no access to historical data in post-change regime beforehand, we may consider construing the post-change ambiguity sets adaptively with sequential observations, and the detailed discussion will be left for future work.*

Moreover, the Wasserstein metric we use in this paper is defined as follows. For two given distributions $P, Q \in \mathcal{P}(\mathcal{X})$, their Wasserstein distance (of order 1) equals to [14]:

$$W(P, Q) := \min_{\Gamma \in \Pi(P, Q)} \mathbb{E}_{(\omega, \omega') \sim \Gamma} [c(\omega, \omega')], \quad (5)$$

where $c(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ is a metric, and $\Pi(P, Q)$ is the set of all joint probability distributions on $\mathcal{X} \times \mathcal{X}$ with marginal distributions P and Q .

Substitute the nominal distribution as the empirical distributions and the divergence measure as the Wasserstein metric, we construct the ambiguity sets as in (4):

$$\begin{aligned} \mathcal{P}_{\mu_0} &= \{\mu \in \mathcal{P}(\mathcal{X}) : W(\mu, \mu_0) \leq r_1\}, \\ \mathcal{P}_{\nu_0} &= \{\nu \in \mathcal{P}(\mathcal{X}) : W(\nu, \nu_0) \leq r_2\}. \end{aligned} \quad (6)$$

C. Minimax Robust Change Detection

Under the ambiguity sets (6), we aim to find the robust optimal stopping time that solves the following problem:

$$\inf_{T \in C(\gamma, \mathcal{P}_{\mu_0})} \sup_{\mu \in \mathcal{P}_{\mu_0}, \nu \in \mathcal{P}_{\nu_0}} \text{WADD}^{\mu, \nu}(T), \quad (7)$$

where $C(\gamma, \mathcal{P}_{\mu_0})$ is the set containing all stopping times T that satisfies $\mathbb{E}_\infty^\mu[T] \geq \gamma, \forall \mu \in \mathcal{P}_{\mu_0}$. Similarly, the corresponding problem defined using CADD is:

$$\inf_{T \in C(\gamma, \mathcal{P}_{\mu_0})} \sup_{\mu \in \mathcal{P}_{\mu_0}, \nu \in \mathcal{P}_{\nu_0}} \text{CADD}^{\mu, \nu}(T). \quad (8)$$

In general, it may be challenging to exactly solve the problems (7) and (8). Therefore, asymptotically optimal solutions for the above problems are often investigated in practice. A solution $T_0 \in C(\gamma, \mathcal{P}_{\mu_0})$ is called first-order asymptotic optimal [16] for (7) (and similarly defined for (8)) if:

$$\lim_{\gamma \rightarrow \infty} \frac{\sup_{\mu \in \mathcal{P}_{\mu_0}, \nu \in \mathcal{P}_{\nu_0}} \text{WADD}^{\mu, \nu}(T_0)}{\inf_{T \in C(\gamma, \mathcal{P}_{\mu_0})} \sup_{\mu \in \mathcal{P}_{\mu_0}, \nu \in \mathcal{P}_{\nu_0}} \text{WADD}^{\mu, \nu}(T)} = 1.$$

Remark 2. *The choice of the radius r_1, r_2 is crucial for the minimax detection problem. There is a tradeoff between*

model robustness and detection performance. A large radius will lead to a more robust detection but also a larger detection delay. Empirically, we can use cross-validation to set the radius. Theoretically, we may analyze the concentration of the Wasserstein distance to determine the appropriate radius [25].

III. OPTIMAL STOPPING TIME AND THEORETICAL GUARANTEE

In this section, we derive the asymptotic optimal stopping time that solves the problem (7) and (8). Based on previous results established in [10], the optimal stopping time can be constructed based on a pair of distributions in the ambiguity sets $(\mathcal{P}_{\mu_0}, \mathcal{P}_{\nu_0})$, which are called the *least favorable distributions* (LFD). We first list the conditions to find such a pair of LFDs and show that they can be efficiently solved under the Wasserstein ambiguity sets (6). Then we construct the optimal stopping time, which is a CUSUM test [4] based on LFDs.

A. Least Favorable Distributions

There are two types of conditions for finding the LFDs, which can be viewed as a representative pair of distributions within $(\mathcal{P}_{\mu_0}, \mathcal{P}_{\nu_0})$ on which the stopping time reaches the worst-case performance. The first condition, joint stochastic boundedness, was proposed in [11] as follows.

Definition 2 (Joint stochastic boundedness [11]). *A pair of ambiguity sets $(\mathcal{P}_{\mu_0}, \mathcal{P}_{\nu_0})$ is jointly stochastically bounded by the pair of distributions $(\tilde{\mu}, \tilde{\nu})$ if $\forall \nu \in \mathcal{P}_{\nu_0}$,*

$$\mathbb{P}^{\tilde{\nu}} \left(\log \frac{d\tilde{\nu}}{d\mu} (X) \geq x \right) \leq \mathbb{P}^{\nu} \left(\log \frac{d\tilde{\nu}}{d\mu} (X) \geq x \right), \forall x \in \mathbb{R},$$

and $\forall \mu \in \mathcal{P}_{\mu_0}$,

$$\mathbb{P}^{\mu} \left(\log \frac{d\tilde{\nu}}{d\mu} (X) \geq x \right) \leq \mathbb{P}^{\tilde{\mu}} \left(\log \frac{d\tilde{\nu}}{d\mu} (X) \geq x \right), \forall x \in \mathbb{R}.$$

This condition was later relaxed by [10] as follows.

Definition 3 (Weak stochastic boundedness [10]). *A pair of ambiguity sets $(\mathcal{P}_{\mu_0}, \mathcal{P}_{\nu_0})$ is weakly stochastically bounded by the pair of distributions $(\tilde{\mu}, \tilde{\nu})$ if*

$$\text{KL}(\tilde{\nu} || \tilde{\mu}) \leq \text{KL}(\nu || \tilde{\mu}) - \text{KL}(\nu || \tilde{\nu}), \forall \nu \in \mathcal{P}_{\nu_0}, \quad (9)$$

and

$$\mathbb{E}^{\mu} \left[\frac{d\tilde{\nu}}{d\mu} (X) \right] \leq \mathbb{E}^{\tilde{\mu}} \left[\frac{d\tilde{\nu}}{d\mu} (X) \right] = 1, \forall \mu \in \mathcal{P}_{\mu_0}. \quad (10)$$

In [10], it was shown that finding the pair of distributions that satisfies the weak stochastic boundedness condition is equivalent to finding the pair of distributions that minimizes the pairwise KL divergence between ambiguity sets. More specifically, the LFDs $(\tilde{\mu}, \tilde{\nu})$ satisfying (9) is a solution to:

$$\min_{\mu \in \mathcal{P}_{\mu_0}, \nu \in \mathcal{P}_{\nu_0}} \text{KL}(\nu || \mu). \quad (11)$$

Our main finding is that under the Wasserstein ambiguity sets (6), the pair of distributions such that the sets are weakly stochastic bounded can be found through the following convex optimization problem efficiently. Denote $n = n_1 + n_2$,

$\{z_1, \dots, z_n\}$ as the union of pre- and post-change historical data in the order of $\{x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}\}$, and $\mathbf{1}_n$ as a n -dimensional column vector with all entries equal to one.

Theorem 1 (LFD). *The pair of LFD solving (11) can be found by the following finite-dimensional convex program*

$$\begin{aligned} & \min_{\substack{p_1, p_2 \in \mathbb{R}_+^n \\ \Gamma_1, \Gamma_2 \in \mathbb{R}^{n \times n}}} \sum_{l=1}^n p_2^l \log(p_2^l / p_1^l) \\ & \text{subject to} \quad \sum_{l=1}^n \sum_{m=1}^n \Gamma_{k,l,m} c(z_l, z_m) \leq r_k, \quad k = 1, 2; \\ & \quad \Gamma_{1,l,m} \mathbf{1}_n = \mu_0, \quad \Gamma_{2,l,m} \mathbf{1}_n = \nu_0; \\ & \quad \Gamma_{1,l,m}^\top \mathbf{1}_n = p_1, \quad \Gamma_{2,l,m}^\top \mathbf{1}_n = p_2. \end{aligned} \quad (12)$$

Proof. See the Appendix. \square

Note that the optimization problem in (12) is the problem (11) for discrete distributions supported on the joint empirical samples $\{z_1, \dots, z_n\}$. The variables Γ_1, Γ_2 are two matrices representing how the probability mass is transported between the empirical distribution μ_0, ν_0 and the desired LFD p_1, p_2 . Instead of solving the infinite-dimensional problem (11), we can now solve the finite-dimensional optimization problems (12) which can be solved efficiently using off-the-shelf solvers. It is a linear program when c is ℓ_1 or ℓ_∞ norms, and a conic program when c is ℓ_2 norm, and the complexity quadratically depends on n . It is worth mentioning that the equivalence between (12) and (11) is not obvious and depends on the properties of the objective function and ambiguity sets.

B. Optimal Stopping Time

Once we find $(\tilde{\mu}, \tilde{\nu})$ by which the ambiguity sets are weakly stochastically bounded, the optimal stopping time that solves the problem (7) asymptotically can be constructed as the CUSUM procedure [4] based on $(\tilde{\mu}, \tilde{\nu})$. The detection statistic can be computed recursively as

$$S_t = (S_{t-1})^+ + \log \frac{d\tilde{\nu}}{d\tilde{\mu}} (x_t), \quad S_0 = 0, \quad (13)$$

and stopping time is therefore defined as

$$\mathcal{T} := \inf\{t : S_t \geq b\}, \quad (14)$$

where b is a pre-specified threshold such that the average run length meets the desired lower bound γ .

Theorem 2 (Asymptotical Optimality). *Consider the ambiguity sets $\mathcal{P}_{\mu_0}, \mathcal{P}_{\nu_0}$ in (6), and suppose the pair of distributions $(\tilde{\mu}, \tilde{\nu})$ found through (12) satisfies the condition (10). Then the CUSUM test (13)-(14) under $(\tilde{\mu}, \tilde{\nu})$ with threshold $b = |\log \gamma|$ solves (7) and (8) asymptotically as $\gamma \rightarrow \infty$.*

Proof. From Theorem 1, the pair $(\tilde{\mu}, \tilde{\nu})$ is a solution to (11), which is equivalent to (9). If $(\tilde{\mu}, \tilde{\nu})$ also satisfies (10), then the ambiguity sets $\mathcal{P}_{\mu_0}, \mathcal{P}_{\nu_0}$ are weakly stochastically bounded by $(\tilde{\mu}, \tilde{\nu})$. From Theorem 3 in [10], we have the desired results. \square

Note that the CUSUM procedure as in (13) with threshold $b = \lceil \log \gamma \rceil$ is asymptotically optimal for both Lorden's and Pollak's formulations when the true distribution is $\tilde{\mu}$ and $\tilde{\nu}$. The results in Theorem 2 means that when we have two ambiguity sets, the CUSUM test based on the LFDs are minimax robust asymptotically optimal for the robust Lorden's (7) and Pollak's formulations (8). When the true distributions differ from the LFDs, the price we pay in performance loss is due to the robustness that we would like to guarantee.

C. Extensions and Modifications

We note that the LFDs found through (12) is only supported on the historical data used to construct the empirical distributions. When applied to new observations that are outside the support of those empirical distributions, we need to modify the algorithm to make it applicable in real scenarios. Here we mention two possible methods.

1) *Kernel convolution*: Firstly, we may interpolate the discrete LFDs within the entire sample space \mathcal{X} , through, for example, kernel convolution. And then apply the modified LFDs to calculate the detection statistics for new observations. A simple example is to convolve with the Gaussian kernel $K_h(x, y) = \exp\{-(x - y)^2 / (2h^2)\} / \sqrt{2\pi h^2}$, with a carefully chosen kernel bandwidth. More specifically, the smoothed LFDs after convolution are

$$\tilde{\nu}'(x) = \int \tilde{\nu}(y) K_h(x, y) dy, \quad \tilde{\mu}'(x) = \int \tilde{\mu}(y) K_h(x, y) dy.$$

Thus the detection statistic in (13) becomes $S_t = (S_{t-1})^+ + \log\{d\tilde{\nu}'(x_t)/d\tilde{\mu}'(x_t)\}$.

2) *Binning approach*: Second is a binning approach, which has been used previously in change-point detection problems but for different purposes [26]. In detail, we could partition the sample space \mathcal{X} into L exclusive and exhaustive regions, $\mathcal{X}_1, \dots, \mathcal{X}_L$, satisfying $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$ and $\cup_{i=1}^L \mathcal{X}_i = \mathcal{X}$. In this way, we convert any continuous distribution into discrete ones and the LFDs can then be used naturally for new observations.

IV. NUMERICAL RESULTS

In this section, we investigate the performance of the proposed robust test based on Wasserstein ambiguity sets (which we call Robust-Was CUSUM in this section). For illustrative purposes, we consider a simple Gaussian mean shift example where the data distribution changes from $\mu = \mathcal{N}(0, 1)$ to $\nu_m = \mathcal{N}(m, 1)$ with the post-change mean m takes two possible values 0.5 and 1, representing different signal-to-noise ratios. We randomly generate 50 samples from the pre-change distribution $\mathcal{N}(0, 1)$ and 50 samples from the post-change distribution $\mathcal{N}(m, 1)$. Then we construct the ambiguity sets based on the Wasserstein metric as shown in (6). Then we solve the convex programming problem (12) to find the LFDs $\tilde{\mu}$ and $\tilde{\nu}$. The Robust-Was CUSUM test is constructed based on the LFDs, according to the definition (13) and (14).

In the first result, we use the convolution approach to extend the LFDs to the whole sample space and then calculate the resulting CUSUM statistic. We compare the performance of

the Robust-Was CUSUM test with the exact CUSUM and the GLR test. In detail, the exact CUSUM test is constructed assuming full knowledge of the true distributions, i.e., the exact CUSUM statistic is defined as in (13) using true distributions ν and μ . Moreover, the exact CUSUM test is the optimal test in the sense that it has the smallest detection delay and thus serves as the information-theoretic lower bound to the detection delay [17], [27]. The GLR test is designed for the case when the post-change parameter m is unknown. The parameter is estimated using maximum likelihood estimate and plugged into the log-likelihood ratio to calculate the GLR statistics. Moreover, to increase the efficiency, we adopt the window-limited GLR approach with the test statistic [5]:

$$S_t^G := \max_{t-W \leq k \leq t} \max_{m \in \mathbb{R}} \sum_{i=k}^t \log \frac{d\nu_m}{d\mu}(x_i),$$

where W is the window size and is chosen at 50, the same as the number of empirical observations used in Robust-Was CUSUM. The radii parameters r_1 and r_2 are set to be equal. We select smaller radii for smaller post-change mean, since the empirical samples tend to be closer as m decreases and we need two ambiguity sets to have an empty intersection. The kernel bandwidth parameter as in Section III-C1 is chosen as $h = 0.25$. Moreover, each time after solving the LFDs, we verify that the condition (10) indeed holds.

We plot the expected detection delay versus average run length for different methods, averaged over 10000 times, as shown in Fig. 1. We see that the robust CUSUM derived from the Wasserstein ambiguity sets has a smaller delay than the GLR test.

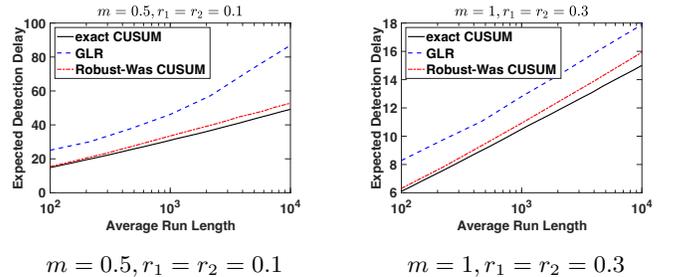


Fig. 1. The detection delay comparison with exact CUSUM and GLR, under different signal-to-noise ratios.

We then compare the performance of the proposed method under the binning approach detailed in Section III-C2, with bin size $L = 20$. We select the breakpoints such that the resulting discretized pre-change distributions is a uniform distribution. In such case, we can also compare with the robust CUSUM test based on KL ambiguity sets [10], where the two ambiguity sets are constructed using the KL divergence and LFDs are again found through (9) and (10). The detection delay shown in Fig. 2 shows that the KL robust CUSUM test tends to have a larger detection delay and the proposed Robust-Was CUSUM test still has a better performance.

We also compare the performance when the observations are contaminated. We add a uniform noise (contamination)

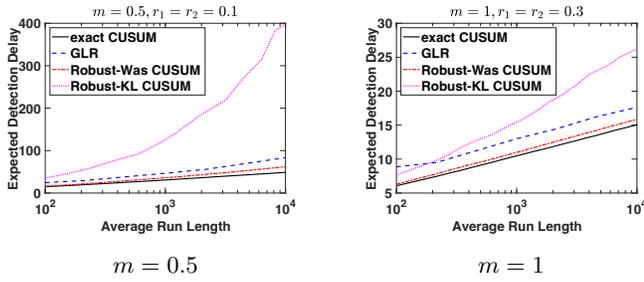


Fig. 2. The detection delay comparison with exact CUSUM, GLR, and the robust test under KL ambiguity sets, after data binning.

into the observations. The contamination follows the uniform distributions on the interval $[-\epsilon, 0]$. We test five values for ϵ from 0.1 to 0.5, representing different strength levels of the contamination. The average detection delay is plotted in Fig. 3. The exact CUSUM algorithm is no longer optimal when the observations are contaminated, since there is a mismatch between the distribution used to construct CUSUM statistics and the true data distribution after contamination. From Fig. 3, we see that the proposed method may even have a smaller detection delay than the exact CUSUM method.

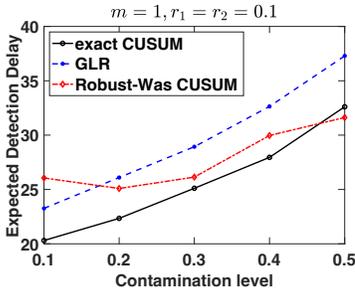


Fig. 3. The detection delay comparison with exact CUSUM and GLR, after data binning and for contaminated data. ARL fixed at 50000.

V. CONCLUSIONS AND DISCUSSIONS

We applied Wasserstein ambiguity sets to robust quickest change detection. This also brought new questions worth investigating. First, it would be of great importance to study a data-driven and precise characterization of the radii for future work. Second, the real data are usually under contamination; thus it would be interesting to study the theoretical performance of the robust test under contaminated data or outliers. Third, the LFDs solved in this work are discrete distributions, it would be worthwhile to study the theoretical loss or explore different ways to better fix such problem.

VI. APPENDIX

Proof of Theorem 1. Denote by $L^1(\mu)$ the space of all integrable functions with respect to the measure μ . Using the Kantorovich duality [28], the Wasserstein distance equals:

$$W(\mu, \nu) = \sup_{\substack{(\phi, \psi) \in L^1(\mu) \times L^1(\nu) \\ \phi(x) + \psi(y) \leq c(x, y) \\ \forall x, y}} \left(\int_{\mathcal{X}} \phi(x) d\mu + \int_{\mathcal{X}} \psi(y) d\nu \right).$$

Following [7], we rewrite the problem using the Lagrangian of the optimization problem (11) and the above equation:

$$\begin{aligned} & \inf_{\mu, \nu \in \mathcal{P}(\Omega)} \sup_{\substack{\lambda_1, \lambda_2 \geq 0 \\ u_1 \in \mathbb{R}^{n_1}, u_2 \in \mathbb{R}^{n_2} \\ v_1 \in L^1(\mu), v_2 \in L^1(\nu)}} \left\{ \text{KL}(\nu || \mu) - \lambda_1 r_1 - \lambda_2 r_2 + \right. \\ & \frac{1}{n_1} \sum_{i=1}^{n_1} u_1^i + \frac{1}{n_2} \sum_{i=1}^{n_2} u_2^i + \int_{\mathcal{X}} v_1(x) d\mu + \int_{\mathcal{X}} v_2(x) d\nu : \\ & u_1^i + v_1(\xi) \leq \lambda_1 c(\xi, x_i), \quad \forall 1 \leq i \leq n_1, \forall \xi \in \mathcal{X}, \\ & u_2^i + v_2(\xi) \leq \lambda_2 c(\xi, y_i), \quad \forall 1 \leq i \leq n_2, \forall \xi \in \mathcal{X} \}. \end{aligned}$$

Furthermore, since the objective function is increasing in v_1, v_2 , we can replace v_1 with $\min_{1 \leq i \leq n_1} \{\lambda_1 c(\xi, x_i) - u_1^i\}$ and replace v_2 with $\min_{1 \leq i \leq n_2} \{\lambda_2 c(\xi, y_i) - u_2^i\}$. Interchanging sup and inf, we have

$$\begin{aligned} & \inf_{\mu \in \mathcal{P}_{\mu_0}, \nu \in \mathcal{P}_{\nu_0}} \text{KL}(\nu || \mu) \\ & \geq \sup_{\substack{\lambda_1, \lambda_2 \geq 0 \\ u_1 \in \mathbb{R}^{n_1}, u_2 \in \mathbb{R}^{n_2}}} \left\{ -\lambda_1 r_1 - \lambda_2 r_2 + \frac{1}{n_1} \sum_{i=1}^{n_1} u_1^i + \frac{1}{n_2} \sum_{i=1}^{n_2} u_2^i + \right. \\ & \inf_{\mu, \nu \in \mathcal{P}(\mathcal{X})} \left\{ \text{KL}(\nu || \mu) + \int_{\mathcal{X}} \min_{1 \leq i \leq n_1} \{\lambda_1 c(\xi, x_i) - u_1^i\} d\mu(\xi) \right. \\ & \left. \left. + \int_{\mathcal{X}} \min_{1 \leq i \leq n_2} \{\lambda_2 c(\xi, y_i) - u_2^i\} d\nu(\xi) \right\} \right\}. \end{aligned}$$

For the inner infimum problem, note that $\forall (\mu, \nu)$ and $\forall \xi \in \text{supp}(\mu) \cup \text{supp}(\nu)$, let $i_1(\xi) = \arg \min_i \{\lambda_1 c(\xi, x_i) - u_1^i\}$, $i_2(\xi) = \arg \min_i \{\lambda_2 c(\xi, y_i) - u_2^i\}$, set

$$T(\xi) := \begin{cases} x_{i_1(\xi)}, & \text{if } \lambda_1 d\mu(\xi) \geq \lambda_2 d\nu(\xi), \\ y_{i_2(\xi)}, & \text{if } \lambda_1 d\mu(\xi) < \lambda_2 d\nu(\xi), \end{cases}$$

then $T(\xi)$ belongs to the minimum of $\min_{1 \leq i \leq n_1} \{\lambda_1 c(\xi, x_i) - u_1^i\} d\mu(\xi) + \min_{1 \leq i \leq n_2} \{\lambda_2 c(\xi, y_i) - u_2^i\} d\nu(\xi)$. Moreover, construct another distributions (μ', ν') such that $\mu'(B) = \mu\{\xi \in \mathcal{X} : T(\xi) \in B\}$ and $\nu'(B) = \nu\{\xi \in \mathcal{X} : T(\xi) \in B\}$ for any Borel set $B \subset \hat{\mathcal{X}} := \{z_1, \dots, z_n\}$. Then it is easy to see that

$$\begin{aligned} & \int_{\hat{\mathcal{X}}} \min_{1 \leq i \leq n_1} \{\lambda_1 c(\xi, x_i) - u_1^i\} d\mu'(\xi) + \\ & \int_{\hat{\mathcal{X}}} \min_{1 \leq i \leq n_2} \{\lambda_2 c(\xi, y_i) - u_2^i\} d\nu'(\xi) \\ & \leq \int_{\mathcal{X}} \min_{1 \leq i \leq n_1} \{\lambda_1 c(\xi, x_i) - u_1^i\} d\mu(\xi) + \\ & \int_{\mathcal{X}} \min_{1 \leq i \leq n_2} \{\lambda_2 c(\xi, y_i) - u_2^i\} d\nu(\xi) \end{aligned}$$

In addition, for ν that is absolutely continuous with respect to μ , we have $\text{KL}(\nu' || \mu') \leq \text{KL}(\nu || \mu)$ since the KL divergence is a convex function.

Hence (μ', ν') yields an objective value no worse than (μ, ν) for the inner infimum problem. This means that it suffices to only consider (μ, ν) supported on the empirical set $\hat{\mathcal{X}}$. Following a similar argument as in [7], the optimization problem can be reduced to a finite-dimensional convex optimization problem as shown in Theorem 1. \square

REFERENCES

- [1] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. ser. Monographs on Statistics and Applied Probability 136. Boca Raton, London, New York: Chapman & Hall/CRC Press, Taylor & Francis Group, 2015.
- [2] D. O. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*, ser. Springer Series in Statistics. Springer, 1985.
- [3] V. V. Veeravalli and T. Banerjee, “Quickest change detection,” *Academic Press Library in Signal Processing: Array and Statistical Signal Processing*, vol. 3, pp. 209–256, 2013.
- [4] E. S. Page, “Continuous inspection schemes,” *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.
- [5] T. L. Lai, “Information bounds and quick detection of parameter changes in stochastic systems,” *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2917–2929, 1998.
- [6] P. J. Huber, “A robust version of the probability ratio test,” *Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.
- [7] R. Gao, L. Xie, Y. Xie, and H. Xu, “Robust hypothesis testing using Wasserstein uncertainty sets,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 7902–7912.
- [8] M. Fauß, A. M. Zoubir, and H. V. Poor, “Minimax robust detection: Classic results and recent advances,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 2252–2283, 2021.
- [9] Z. Sun and S. Zou, “A data-driven approach to robust hypothesis testing using kernel MMD uncertainty sets,” in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 3056–3061.
- [10] T. L. Molloy and J. J. Ford, “Misspecified and asymptotically minimax robust quickest change detection,” *IEEE Transactions on Signal Processing*, vol. 65, no. 21, pp. 5730–5742, 2017.
- [11] J. Unnikrishnan, V. V. Veeravalli, and S. P. Meyn, “Minimax robust quickest change detection,” *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1604–1614, 2011.
- [12] J. Z. Hare, L. Kaplan, and V. V. Veeravalli, “Toward uncertainty aware quickest change detection,” in *2021 IEEE 24th International Conference on Information Fusion (FUSION)*. IEEE, 2021, pp. 1–8.
- [13] J. Z. Hare, C. A. Uribe, L. Kaplan, and A. Jadbabaie, “Non-Bayesian social learning with uncertain models,” *IEEE Transactions on Signal Processing*, vol. 68, pp. 4178–4193, 2020.
- [14] C. Villani, *Topics in Optimal Transportation*. American Mathematical Society, 2003, no. 58.
- [15] S. Kullback and R. A. Leibler, “On information and sufficiency,” *Annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [16] L. Xie, S. Zou, Y. Xie, and V. V. Veeravalli, “Sequential (quickest) change detection: Classical results and new directions,” *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 2, pp. 494–514, 2021.
- [17] G. Lorden, “Procedures for reacting to a change in distribution,” *Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1897–1908, 1971.
- [18] M. Pollak, “Optimal detection of a change in distribution,” *Annals of Statistics*, vol. 13, no. 1, pp. 206–227, 1985.
- [19] G. Gül and A. M. Zoubir, “Minimax robust hypothesis testing,” *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5572–5587, 2017.
- [20] B. C. Levy, “Robust hypothesis testing with a relative entropy tolerance,” *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 413–421, 2009.
- [21] P. J. Huber and V. Strassen, “Minimax tests and the Neyman-Pearson lemma for capacities,” *Annals of Statistics*, vol. 1, no. 2, pp. 251–263, 1973.
- [22] M. Fauß, A. M. Zoubir, and H. V. Poor, “Minimax optimal sequential hypothesis tests for Markov processes,” *Annals of Statistics*, vol. 48, no. 5, pp. 2599–2621, 2020.
- [23] R. Gao and A. J. Kleywegt, “Distributionally robust stochastic optimization with Wasserstein distance,” *arXiv preprint arXiv:1604.02199*, 2016.
- [24] P. M. Esfahani and D. Kuhn, “Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations,” *Mathematical Programming*, vol. 171, no. 1, pp. 115–166, 2018.
- [25] N. Fournier and A. Guillin, “On the rate of convergence in Wasserstein distance of the empirical measure,” *Probability Theory and Related Fields*, vol. 162, no. 3, pp. 707–738, 2015.
- [26] T. S. Lau, W. P. Tay, and V. V. Veeravalli, “A binning approach to quickest change detection with unknown post-change distribution,” *IEEE Transactions on Signal Processing*, vol. 67, no. 3, pp. 609–621, 2018.
- [27] G. V. Moustakides, “Optimal stopping times for detecting changes in distributions,” *Annals of Statistics*, pp. 1379–1387, 1986.
- [28] C. Villani, *Optimal Transport: Old and New*. Springer Science & Business Media, 2008, vol. 338.