Distributionally Robust Quickest Change Detection using Wasserstein Uncertainty Sets

Liyan Xie* CUHK-Shenzhen

Yuchen Liang* Ohio State University

Venugopal V. Veeravalli University of Illinois Urbana-Champaign

Abstract

The problem of quickest detection of a change in the distribution of streaming data is considered. It is assumed that the pre-change distribution is known, while the only information about the post-change is through a (small) set of labeled data. This post-change data is used in a data-driven minimax robust framework, where an uncertainty set for the post-change distribution is constructed. The robust change detection problem is studied in an asymptotic setting where the mean time to false alarm goes to infinity. It is shown that the least favorable distribution (LFD) is an exponentially tilted version of the pre-change density and can be obtained efficiently. A Cumulative Sum (CuSum) test based on the LFD, which is referred to as the distributionally robust (DR) CuSum test, is then shown to be asymptotically robust. The results are extended to the case with multiple post-change uncertainty sets and validated using synthetic and real data examples.

1 INTRODUCTION

Given sequential observations, the problem of quickest change detection (QCD) is to detect a potential change in their distribution that occurs at some change-point as quickly as possible, while not making too many false alarms (Siegmund, 1985; Basseville and Nikiforov, 1993; Poor and Hadjiliadis, 2008; Tartakovsky et al., 2015). The QCD problem is of fundamental importance in statistics, and has seen a wide range of applications (Veeravalli and Banerjee, 2013; Xie et al., 2021).

In the classical formulation of the QCD problem (Page, 1954), it is assumed that the observations are independent and identically distributed (i.i.d.) with *known* pre- and post-change distributions. In many applications of QCD, while it is reasonable to assume that the pre-change distribution is known (can be estimated accurately), the post-change distribution is rarely completely known. However, we may have access to a limited set of data corresponding to post-change.

There has been a large body of work on the QCD problem when the pre- and/or post-change distributions have *parametric* uncertainty. The most prevalent approach to dealing with parametric uncertainty is the generalized likelihood ratio (GLR) approach, introduced in Lorden (1971) for the special case where the pre-change distribution is known and the post-change distribution has an unknown parameter. The GLR approach for the QCD problem with general parametric distributions is studied in Lai (1998) and Lai and Xing (2010). An alternative approach to dealing with parametric uncertainty is the mixture-based approach, which was proposed and studied in Pollak (1978).

The QCD problem has also been studied in the *non-parametric* setting. In Li et al. (2015), a test is proposed that compares the kernel maximum mean discrepancy (MMD) within a window to a given threshold. Another approach has been to estimate the log-likelihood ratio through a pre-collected training set. This includes direct kernel estimation (Kawahara and Sugiyama, 2012), neural network estimation (Moustakides and Basioti, 2019), and density ratio estimation (Adiga and Tandon, 2022; Sugiyama et al., 2008; Kawahara and Sugiyama, 2009). More recently, a non-parametric GLR test based on density estimation has been developed for the case where the post-change distribution is completely unknown without any pre-collected post-change training samples (Liang and Veeravalli, 2023).

Another line of work for dealing with non-parametric distributional uncertainty is the one based on *minimax* robust detection, in which it is assumed that the pre- and post-change distributions come from disjoint uncertainty classes. This approach is of particular interest when distributional robustness is one of the objectives

*Equal contribution.

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

of the QCD formulation. Under certain conditions on the uncertainty classes, e.g., joint stochastic boundedness (Moulin and Veeravalli, 2018), low-complexity solutions to the minimax robust QCD problem can be found (Unnikrishnan et al., 2011). Under more general conditions, in particular, weak stochastic boundedness, a solution that is asymptotically close to the minimax robust solution can be found (Molloy and Ford, 2017).

In the literature of robust hypothesis testing, a variety of uncertainty sets have been considered in a non-parametric way. One line of work is where the uncertainty set is constructed by selecting a nominal distribution as the center and choosing a deviation measure such that the set includes all distributions whose deviation from the nominal does not exceed a positive constant. Examples include the ϵ-contamination model (Huber, 1965) and the KL-divergence sets (Levy, 2008). In the data-driven setting, the nominal distribution is often chosen as the empirical distribution of training samples. The uncertainty sets that have been used in the literature include the Wasserstein uncertainty sets (Gao et al., 2018), the kernel MMD sets (Sun and Zou, 2021), and the Sinkhorn sets (Wang and Xie, 2022). Some work also constructs the uncertainty set according to pre-specified constraints, such as moment constraints (Magesh et al., 2023).

In this paper, we consider a *data-driven minimax robust QCD* problem, where the pre-change distribution is assumed to be *known*, and the only knowledge about the post-change distribution is through a limited set of data corresponding to one or more possible post-change scenarios. For each possible post-change scenario, we define an empirical distribution using training data collected under this scenario. Then we construct the corresponding Wasserstein uncertainty set to contain all distributions such that their Wasserstein distance from the empirical distribution does not exceed some specified value (i.e., radius). Our goal is to find the asymptotically optimal robust detection procedure that minimizes the worst-case detection delay over the uncertainty set, while satisfying the false alarm constraints. We focus on the asymptotic setting where the mean time to false alarm goes to infinity.

Our contributions can be summarized as follows.

1. We characterize the least favorable distribution (LFD) within the Wasserstein uncertainty set in closed-form. We therefore establish that the Cumulative Sum (CuSum) test based on the LFD, which we refer to as the distributionally robust (DR) CuSum test, is asymptotically robust. We also characterize the size of radius through empirical concentration inequalities of Wasserstein distance.
2. We extend the DR-CuSum test to construct an asymptotically robust solution for the case where the post-change uncertainty set is a union of multiple Wasserstein uncertainty sets.
3. We show that DR-CuSum can outperform existing benchmarks using simulated Gaussian data and a real human activity dataset.

2 PROBLEM SETUP

Let $\{X_k,\ k \in \mathbb{N}\}$ be a sequence of independent random vectors whose values are observed sequentially, with $\mathscr{X}$ denoting the observation space, i.e., $X_k \in \mathscr{X}$ for all $k \in \mathbb{N}$. Let $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, $k \in \mathbb{N}$, be the filtration, with $\mathcal{F}_0$ denoting the trivial sigma algebra. Let P and Q be probability measures on $\mathscr{X}$. At some unknown (yet deterministic) time ν, the data-generating distribution changes from Q to P, i.e.,

$$\begin{aligned} &X_k \overset{\text{iid}}{\sim} Q, \quad k = 1, 2, \ldots, \nu - 1, \\ &X_k \overset{\text{iid}}{\sim} P, \quad k = \nu, \nu + 1, \ldots \end{aligned} \tag{1}$$

We assume the pre-change measure Q is *known*, while only partial knowledge of the post-change measure P is available through a set of labeled (training) data.

Let $\mathbb{P}_\nu^P$ denote the probability measure on the data sequence when the change-point is ν, and the pre- and post-change measures are Q and P, respectively, and let $\mathbb{E}_\nu^P$ denote the corresponding expectation. Denote $\mathbb{P}_\infty$ and $\mathbb{E}_\infty$ as the probability and expectation operator when there is no change (i.e., $\nu = \infty$). For brevity, we write $\mathbb{P}^P$ and $\mathbb{E}^P$ as the probability and expectation when all samples are generated from P (i.e., $\nu = 1$).

The goal in QCD is to raise an alarm after the unknown change-point ν as quickly as possible, while keeping the false alarm rate below a pre-specified level. The detection is performed through a *stopping time* τ on the observation sequence at which the change is declared.

2.1 QCD Problem and CuSum Test

False Alarm Measure. We measure the false alarm performance of a QCD test (stopping time) τ in terms of its mean time to false alarm $\mathbb{E}_\infty[\tau]$, and we denote by $\mathcal{C}(\gamma)$ the set of all tests for which the mean time to false alarm is at least γ, i.e.,

$$\mathcal{C}(\gamma) = \{\tau : \mathbb{E}_\infty[\tau] \geq \gamma\}. \tag{2}$$

Delay Measure. We use the commonly used worst-case delay measure (WADD) in Lorden (1971). Specifically, for post-change distribution P and test τ, we set[1]

$$\mathsf{WADD}^P(\tau) = \sup_{\nu \geq 1} \operatorname{ess\,sup}\ \mathbb{E}_\nu^P \left[(\tau - \nu + 1)^+ | \mathcal{F}_{\nu-1} \right]. \tag{3}$$

QCD Optimization Problem. When both Q and P are known *a priori*, the optimization problem of interest is

$$\inf_{\tau \in \mathcal{C}(\gamma)} \mathsf{WADD}^P(\tau). \tag{4}$$

The Cumulative Sum (CuSum) test (Page, 1954) is proved to solve the problem (4) exactly (Moustakides, 1986). The stopping time of the CuSum test is

$$\tau_b = \inf \{ k \in \mathbb{N} : S_k \geq b \}, \tag{5}$$

with the CuSum statistic calculated recursively as:

$$S_0 = 0,\ S_k = (S_{k-1})^+ + \log \frac{p(X_k)}{q(X_k)}, k \geq 1, \tag{6}$$

and b is chosen to meet the false alarm constraint of γ. Here p, q are the respective probability density functions (pdfs) of the measures P and Q with respect to some common dominating measure.

2.2 Asymptotically Minimax Robust QCD

As mentioned previously, we have limited knowledge about the post-change distribution P. One way to deal with this distributional uncertainty is to assume that $P \in \mathcal{P}$, where $\mathcal{P}$ is a family of probability measures representing potential post-change distributions. In the minimax robust QCD formulation, the goal is to solve the following optimization problem,

$$\inf_{\tau \in C(\gamma)} \sup_{P \in \mathcal{P}} \mathsf{WADD}^P(\tau), \tag{7}$$

where $C(\gamma)$ is as defined in (2). As is standard practice in the analysis of QCD procedures, we are primarily interested in the asymptotically optimal solution to (7) as $\gamma \to \infty$. A solution $\tau^* \in C(\gamma)$ is called *first-order asymptotically minimax robust* for (7) if

$$\sup_{P \in \mathcal{P}} \mathsf{WADD}^P(\tau^*) = \inf_{\tau \in C(\gamma)} \sup_{P \in \mathcal{P}} \mathsf{WADD}^P(\tau) \cdot (1 + o(1)),$$

where, as throughout this paper, $o(1) \to 0$ as $\gamma \to \infty$.

Solving the asymptotically minimax robust solution to robust QCD problems is facilitated by the following weak stochastic boundedness (WSB) condition.

Definition 2.1 (Weak Stochastic Boundedness (Molloy and Ford, 2017))**.** Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be sets of distributions on a common measurable space where $\mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$. The pair $(\mathcal{P}_0, \mathcal{P}_1)$ is said to be weakly stochastically bounded by the pair of distributions (P_0^*, P_1^*) if

$$\mathsf{KL}(P_1^* || P_0^*) \leq \mathsf{KL}(P_1 || P_0^*) - \mathsf{KL}(P_1 || P_1^*),\ \forall P_1 \in \mathcal{P}_1,$$

and

$$\mathbb{E}^{P_0} \left[\frac{dP_1^*}{dP_0^*}(X) \right] \leq \mathbb{E}^{P_0^*} \left[\frac{dP_1^*}{dP_0^*}(X) \right] = 1,\ \forall P_0 \in \mathcal{P}_0,$$

where

$$\mathsf{KL}(P || Q) = \int_{\mathcal{X}} \log \left(\frac{dP}{dQ}(x) \right) dP(x),$$

and $\frac{dP}{dQ}(x)$ is the Radon-Nikodym derivative of P with respect to Q with $\frac{dP}{dQ}(x) = \infty$ when the derivative does not exist. The pair of distributions (P_0^*, P_1^*) are called *least favorable distributions (LFDs).*

Intuitively, the LFDs can be viewed as a representative pair of distributions within uncertainty sets on which the stopping time reaches the *worst-case* performance. In this paper, we assume that the pre-change distribution is known; this corresponds to the special case that $\mathcal{P}_0$ is a singleton. The following lemma follows directly from (Molloy and Ford, 2017, Prop. 1 (iii)).

Lemma 2.1. *For the singleton set $\mathcal{Q} = \{Q\}$ and a* convex *set of distributions $\mathcal{P}$ where $Q \notin \mathcal{P}$, $(\mathcal{Q}, \mathcal{P})$ is weakly stochastically bounded by the pair of distributions (Q, P^*), where*

$$P^* = \arg\min_{P \in \mathcal{P}} \mathsf{KL}(P || Q). \tag{8}$$

Let p^*, q be pdfs of P^* and Q, respectively, with respect to a common dominating measure. Then applying (Molloy and Ford, 2017, Theorem 3), we conclude that the first-order asymptotically minimax robust solution to (7), as $\gamma \to \infty$, is given by the CuSum test with pre-change pdf q and post-change pdf p^*, i.e.,

$$\tau_{\mathrm{DR}} = \inf \{ k \in \mathbb{N} : S_k \geq b \}, \tag{9}$$

with S_k satisfying the recursion (for $k \geq 1$):

$$S_k = (S_{k-1})^+ + \log \frac{p^*(X_k)}{q(X_k)}, \quad S_0 = 0, \tag{10}$$

and b chosen to meet the false alarm constraint of γ.

3 MINIMAX ROBUST QCD UNDER SINGLE UNCERTAINTY SET

We are interested in a *data-driven* version of the minimax robust QCD problem, where the only knowledge about the post-change distribution is through a limited set of labeled data. We use this data to construct a

[1] Alternatively we may also consider the Pollak's measure (Pollak, 1985) $\mathsf{CADD}^P(\tau) = \sup_{\nu \geq 1}\ \mathbb{E}_\nu^P[\tau - \nu | \tau \geq \nu]$.

Wasserstein uncertainty set for the post-change distribution. We begin by considering the simplest case where there is only *one* possible post-change scenario. Our main finding is that under the Wasserstein uncertainty set, the density of the post-change LFD, i.e., the solution to (8), is an exponentially tilted version of the pre-change density.

3.1 The Wasserstein Uncertainty Model

Suppose we have n training data $\{\omega_1, \ldots, \omega_n\}$ that are independently sampled from the post-change regime, then we choose the nominal distribution of the uncertainty set to be the empirical distribution of those historical samples, i.e., $\widehat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\omega_i}$, where δ_ω corresponds to the Dirac measure at ω.

We construct the uncertainty set $\mathcal{P}_n$ as the set of all probability measures that are close to $\widehat{P}_n$ with respect to the Wasserstein distance $\mathsf{W}_s(\cdot,\cdot)$,

$$\mathcal{P}_n = \{P \in \mathscr{P}_s : \mathsf{W}_s(P, \widehat{P}_n) \le r_s\}, \tag{11}$$

where $\mathscr{P}_s$ is the set of all Borel probability measures P on the sample space $\mathscr{X}$ such that $\int_{\mathscr{X}} c^s(x, x_0) dP(x) < \infty$ holds for all $x_0 \in \mathscr{X}$, where $c(\cdot,\cdot) : \mathscr{X} \times \mathscr{X} \to \mathbb{R}_+$ is a metric and $r_s \ge 0$ is the radius parameter controlling the size of the uncertainty set. Here, for any two probability measures $P, \tilde{P} \in \mathscr{P}_s$, their Wasserstein distance (of order s) with metric $c(\cdot,\cdot)$ equals to

$$\mathsf{W}_s(P, \tilde{P}) := \left\{ \min_{\Gamma \in \Pi(P,\tilde{P})} \int_{\mathscr{X}\times\mathscr{X}} c^s(\omega, \tilde{\omega}) d\Gamma(\omega, \tilde{\omega}) \right\}^{1/s}, \tag{12}$$

where $\Pi(P, \tilde{P})$ is the set of all joint probability measures on $\mathscr{X} \times \mathscr{X}$ with marginal distributions P and $\tilde{P}$, respectively (Villani, 2003). In this work, we restrict $s \ge 1$ and mainly use $s = 2$ in numerical experiments.

It is worthwhile mentioning that we chose Wasserstein distance due to its unique advantages. It can handle divergences between discrete and continuous distributions, which is essential for our use of empirical (discrete) distributions as the center. It also incorporates data geometry via the transportation cost, aligning well with our data-driven method. In the subsection following, we omit the dependency on the sample size n and write the empirical distribution and uncertainty set as $\widehat{P}$ and $\mathcal{P}$, respectively.

3.2 Least Favorable Distribution

To obtain the LFD, the goal is to solve the following optimization problem (as in Equation (8)),

$$\min_P \mathsf{KL}(P||Q), \text{ such that } \mathsf{W}_s(P, \widehat{P}) \le r_s. \tag{13}$$

We note that (13) resembles the Optimistic Kullback Leibler defined in Dewaskar et al. (2023), albeit in a completely different context. Below, we show that the optimal solution to (13) can be found as an exponential tilting of the pre-change distribution Q. In the following, we assume that Q has pdf q with respect to the Lebesgue measure μ.

Theorem 3.1. *The pdf of the least favorable distribution to the problem* (13)*, with respect to the dominating Lebesgue measure* μ*, satisfies* $p^*(x) \propto q(x)e^{-C_{\lambda^*,u^*}(x)}$*, where* $q(x)$ *is the pre-change pdf with respect to* μ*, and* $\lambda^* \ge 0, u^* \in \mathbb{R}^n$ *are the optimizers of the following convex problem,*

$$\max_{\lambda \ge 0, u \in \mathbb{R}^n} \left\{ -\lambda r_s^s + \frac{1}{n}\sum_{i=1}^n u_i - \log \eta(\lambda, u) \right\}, \tag{14}$$

with

$$C_{\lambda,u}(x) := \min_{1\le i\le n} \{\lambda c^s(x, \omega_i) - u_i\}, \tag{15}$$

and

$$\eta(\lambda, u) := \int q(x) e^{-C_{\lambda,u}(x)} d\mu(x).$$

After we obtain the pdf p^* of the least favorable distribution, we can construct the log-likelihood ratio at each sample X_k as:

$$\log \frac{p^*(X_k)}{q(X_k)} = -C_{\lambda^*,u^*}(X_k) - \log \eta(\lambda^*, u^*), \tag{16}$$

which, after substitution into (10), gives the *Distributionally Robust CuSum* (*DR-CuSum*) statistics under Wasserstein uncertainty sets. And the corresponding stopping time in (9), the *DR-CuSum test*, denoted by τ_{DR}, is the first-order asymptotically minimax robust solution to problem (7).

By exploiting the convexity property, the optimal dual variables λ^* and u^* (thus the term $\eta(\lambda^*, u^*)$) can be pre-computed from (14). Therefore, the DR-CuSum statistics can be easily updated online, resulting in an efficient test for detecting the change. More specifically, the computational complexity of the proposed method is nearly the same as CuSum in the detection phase; however, the proposed method also needs an *offline* training phase in which we solve for the LFD efficiently via convex optimization.

Remark 3.1 (Unknown Pre-change Distribution)**.** In some applications, the pre-change distribution may also need to be estimated from historical data $\xi_1, \ldots, \xi_N$ in the pre-change regime, with N generally being much larger than n. The optimization problem (14) is easily adaptable to such case, because the integral $\eta(\lambda, u) = \int q(x)e^{-C_{\lambda,u}(x)} d\mu(x)$ can be approximated directly by sample average $\frac{1}{N}\sum_{i=1}^N e^{-C_{\lambda,u}(\xi_i)}$, without have to estimate the pre-change density. After obtaining $\eta(\lambda, u)$, we can then update the DR-CuSum statistics by accumulating the log-likelihood ratio in (16), without knowing the exact pre-change density.

4 RADIUS AND DETECTION DELAY

In this section, we discuss the choice of the radius parameter r_s in constructing the uncertainty set $\mathcal{P}_n$ in (11), which is an essential parameter to balance the robustness and effectiveness of the DR-CuSum test. We adopt the commonly used principle from distributionally robust optimization (DRO) of choosing the radius such that the true data distribution is included within the uncertainty set with high probability (Kuhn et al., 2019; Mohajerin Esfahani and Kuhn, 2018).

We consider the ideal case where the empirical samples $\{\omega_1, \ldots, \omega_n\}$ are sampled from the true post-change distribution P. The idea is to guarantee that the Wasserstein set contains the true post-change distribution but *not* the pre-change distribution. We first list a known empirical concentration result for Wasserstein distance, under the $T_s(c)$ inequality condition below (more details and examples of Gaussian and discrete distributions are given in Appendix A.2).

Definition 4.1 ($T_s(c)$ inequality (Raginsky and Sason, 2013))**.** We say that a probability measure P satisfies an L^s transportation-cost inequality with constant $c > 0$ (which is referred to as the $T_s(c)$ inequality), if for every probability measure $Q \ll P$ we have

$$\mathsf{W}_s(P, Q) \leq \sqrt{2c\mathsf{KL}(Q||P)}.$$

Theorem 4.1 (Empirical Concentration (Bolley et al., 2007))**.** *Let $s \in [1, 2]$ and let P be a probability measure on $\mathbb{R}^d$ satisfying a $T_s(c)$ inequality, then for any $d' > d$ and $c' > c$, there exists some constants N_0, depending only on c', d' and some square-exponential moment of P, such that for any $\epsilon > 0$ and $n \geq N_0 \max(\epsilon^{-(d'+2)}, 1)$,*

$$\mathbb{P}^P\{\mathsf{W}_s(P, \widehat{P}_n) > \epsilon\} \leq e^{-\gamma_s n \epsilon^2/2c'},$$

where $\gamma_s = 1$ if $s \in [1, 2)$ and $\gamma_s = 3 - 2\sqrt{2}$ if $s = 2$. Here recall $\mathbb{P}^P$ is the probability measure on the samples that are distributed i.i.d. as P.

We first give an *upper bound* requirement for the radius which guarantees that the pre-change distribution is excluded from the post-change uncertainty set ($Q \notin \mathcal{P}_n$) with high probability, thus making the detection problem valid. Since the pre-change distribution Q is known, given any empirical measure $\widehat{P}_n$, we can calculate $\mathsf{W}_s(Q, \widehat{P}_n)$ and select the radius r_s such that

$$r_s < \mathsf{W}_s(Q, \widehat{P}_n). \tag{17}$$

However, it is worthwhile emphasizing that for theoretical considerations below, the set $\mathcal{P}_n$ is essentially random due to the randomness of empirical samples. The corollary below calculates an *upper bound* for the radius considering such randomness.

Corollary 4.1 (Upper Bound for Radius)**.** *Fix $\delta \in (0, 1)$ and $s \in [1, 2]$. Suppose that the pre- and post-change distributions Q, P are probability measures on $\mathbb{R}^d$, and that P satisfies the $T_s(c)$ inequality. Suppose that $n \geq N_0 \max(r_s^{-(d+2)}, 1)$ where N_0 is the same as in Theorem 4.1. Then, if we set the radius as*

$$r_s \leq \overline{r}_{\delta,n} := \mathsf{W}_s(P, Q) - \sqrt{\frac{2|\log \delta|c}{\gamma_s n}},$$

it is guaranteed that with probability at least $1 - \delta$ we have $Q \notin \mathcal{P}_n$.

When we lack knowledge of $\mathsf{W}_s(P, Q)$, as might be the case in practice, the upper bound $\overline{r}_{\delta,n}$ is only of theoretical interest, and we can use Equation (17) to determine a proper radius.

Next, we present a *lower bound* for r_s to guarantee $P \in \mathcal{P}_n$ with high probability. We also characterize the delay performance when such a condition is satisfied.

Corollary 4.2 (Lower Bound for Radius)**.** *Under the same conditions as in Corollary 4.1, if we set the radius*

$$r_s \geq \underline{r}_{\delta,n} := \sqrt{\frac{2|\log \delta|c}{\gamma_s n}},$$

it is guaranteed that with probability at least $1 - \delta$, we have $P \in \mathcal{P}_n$.

To guarantee the existence of r_s that satisfies the upper bound in Corollary 4.1 and lower bound in Corollary 4.2 at the same time, we give the following necessary requirement on the minimum number of training samples.

Lemma 4.1. *Suppose the same conditions as in Corollary 4.1 hold. Additionally, suppose*

$$n \geq \underline{n}_\delta := \frac{8|\log \delta|c}{\gamma_s(\mathsf{W}_s(P, Q))^2}, \tag{18}$$

where $\underline{n}_\delta$ is the least number of samples to guarantee that $\overline{r}_{\delta,n} \geq \underline{r}_{\delta,n}$. Now, if r_s satisfies

$$\underline{r}_{\delta,n} \leq r_s \leq \overline{r}_{\delta,n},$$

then

$$\mathbb{P}^P\left(\{P \in \mathcal{P}_n\} \cap \{Q \notin \mathcal{P}_n\}\right) \geq 1 - 2\delta.$$

In the following, we write the LFD as P_n^* and its pdf as p_n^*. Lemma 4.2 establishes an asymptotic upper bound on the worst-case detection delay of DR-CuSum test.

Lemma 4.2. *Suppose* $\mathbb{E}^P\big[\big(\log(p_n^*(X_1)/q(X_1))\big)^2\big] < \infty$. *Fix* $\delta \in (0,1)$ *and* $s \in [1,2]$. *Suppose that the pre- and post-change distributions* Q, P *are probability measures on* $\mathbb{R}^d$, *and they both satisfy the* $T_s(c)$ *inequality. Suppose that* $n \geq (N_0(r_s^{-(d+2)} \vee 1)) \vee \underline{n}_\delta$ *where* N_0 *is the same as in Theorem 4.1 and* $\underline{n}_\delta$ *is defined in* (18). *Then, if the chosen radius* r_s *satisfies*

$$\underline{r}_{\delta,n} \leq r_s \leq \overline{r}_{\delta,n},$$

it is guaranteed that with probability at least $1-2\delta$, *the worst-case detection delay of the DR-CuSum test* τ_{DR} *with threshold* $b = \log\gamma$ *can be upper bounded as*

$$\begin{aligned}\sup_{P\in\mathcal{P}_n} \mathsf{WADD}^P(\tau_{\mathrm{DR}}) &\leq \frac{\log\gamma}{\mathsf{KL}(P_n^*||Q)}\cdot(1+o(1))\\ &\leq \frac{2c\log\gamma}{(\mathsf{W}_s(P,Q)-2r_s)^2}\cdot(1+o(1)),\end{aligned} \tag{19}$$

as $\gamma\to\infty$.

We note that the dimensionality d affects the algorithm and results in two ways: (i) The ideal training sample size n depends on d, since the empirical concentration of the Wasserstein distance depends on d as shown in Theorem 4.1; (ii) The selection of the radius and the detection delay are implicitly affected by d through the Wasserstein distance and KL divergence.

Example 4.1. For the special case, where the pre- and post-change distributions are Gaussian, with $Q = N(\mu_0,1)$ and $P = N(\mu_1,1)$, we have $\mathsf{KL}(P||Q) = \frac{1}{2}(\mu_1-\mu_0)^2 = \frac{1}{2}\mathsf{W}_2^2(Q,P)$, and the $T_2(1)$ inequality holds equality. This means that the delay of the DR-CuSum procedure is, with probability at least $1-2\delta$, bounded from above as

$$\begin{aligned}&\frac{\log\gamma}{\mathsf{KL}(P_n^*||Q)}(1+o(1)) \leq \frac{\log\gamma}{\frac{(\mathsf{W}_2(P,Q)-2r_2)^2}{2}}(1+o(1))\\ &= \frac{\log\gamma}{\mathsf{KL}(P||Q)-2r_2\sqrt{2\mathsf{KL}(P||Q)}+2r_2^2}(1+o(1)).\end{aligned}$$

From Corollary 4.2, we may choose the radius as its lower bound with $r_2 = \sqrt{2|\log\delta|c'/(\gamma_2 n)} = O(n^{-1/2})$, which means that for n sufficiently large, we have that the delay of DR-CuSum test will match the optimal delay, $[(\log\gamma)/\mathsf{KL}(P||Q)]\cdot(1+o(1))$, asymptotically.

5 MINIMAX ROBUST QCD UNDER MULTIPLE UNCERTAINTY SETS

We extend the results of Section 3 to the more general case with multiple post-change scenarios as follows. Suppose there are $M \geq 1$ potential post-change scenarios, and we have a set of training samples $\{\omega_1^{(m)},\ldots,\omega_{n_m}^{(m)}\}$ that are independently sampled from the m-th scenario, with $\widehat{P}_{n_m}^{(m)}$ being their empirical distribution. The uncertainty set $\mathcal{P}_{n_m}^{(m)}$ for the m-th post-change scenario, similar to (11), is now defined as

$$\mathcal{P}_{n_m}^{(m)} := \{P\in\mathscr{P}_s : \mathsf{W}_s(P,\widehat{P}_{n_m}^{(m)}) \leq r_{s,m}\}, \tag{20}$$

where $r_{s,m}\geq 0$ is the radius parameter controlling the size of the m-th uncertainty set. With a slight abuse of notation, we define $\mathcal{P} := \cup_{m=1}^M \mathcal{P}_{n_m}^{(m)}$ as the union of all the uncertainty sets in the remainder of this section.

5.1 Asymptotically Optimal Stopping Time

Based on Theorem 3.1, we can find M LFDs, denoted as $P_{(1)}^*,\ldots,P_{(M)}^*$, one for each Wasserstein uncertainty set. The LFD $P_{(m)}^*$ for the m-th uncertainty set is an exponential tilting of Q and has pdf $p_{(m)}^*(x) = q(x)\exp\{-C_{\lambda_m^*,u_m^*}^{(m)}(x)-\eta^{(m)}(\lambda_m^*,u_m^*)\}$, where λ_m^*, u_m^* are the solution to

$$\sup_{\lambda\geq 0, u\in\mathbb{R}^{n_m}} \bigg\{-\lambda r_{s,m}^s + \frac{1}{n_m}\sum_{j=1}^{n_m} u_j - \log\eta^{(m)}(\lambda,u)\bigg\},$$

where $C_{\lambda,u}^{(m)}(x) := \min_{1\leq j\leq n_m}\{\lambda c^s(x,\omega_j^{(m)})-u_j\}$ and $\eta^{(m)}(\lambda,u) := \int q(x)\exp\{-C_{\lambda,u}^{(m)}(x)\}d\mu(x)$. The log-likelihood ratio under scenario m equals

$$\log\frac{p_{(m)}^*(x)}{q(x)} = -C_{\lambda_m^*,u_m^*}^{(m)}(x) - \log\eta^{(m)}(\lambda_m^*,u_m^*).$$

Given online samples $\{X_k,\ k\in\mathbb{N}\}$, the detection statistic for the m-th uncertainty set can be computed recursively as

$$S_k^{(m)} = (S_{k-1}^{(m)})^+ + \log\frac{p_{(m)}^*(X_k)}{q(X_k)}, \quad \forall m = 1\ldots,M. \tag{21}$$

The *DR-CuSum* stopping time under multiple post-change scenarios is then defined as

$$\tau_{\mathrm{DR}}(b) := \inf\left\{k\in\mathbb{N} : \max_{m=1,\ldots,M} S_k^{(m)} \geq b\right\}, \tag{22}$$

where b is chosen to meet the false alarm constraint. In the following Lemma 5.1 and Theorem 5.1, we investigate the asymptotic optimality properties of this DR-CuSum test. The proofs of these results are provided in the Appendix.

Lemma 5.1. *The mean time to false alarm of the test in* (22) *satisfies* $\mathbb{E}_\infty[\tau_{\mathrm{DR}}(b)] \geq e^b/M$.

Theorem 5.1 (Asymptotic Minimax Robustness)**.** *Write*

$$I^* := \min_{m=1,2,\ldots,M} \mathsf{KL}(P_{(m)}^*||Q).$$

Then, the test in (22) *with threshold* $b_\gamma = \log(M\gamma)$ *solves the problem in* (7) *asymptotically as* $\gamma \to \infty$, *with the asymptotic worst-case delay being*

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} \mathsf{WADD}^P(\tau_{\mathrm{DR}}(b_\gamma)) \\
&= \inf_{\tau'\in C(\gamma)} \sup_{P\in\mathcal{P}} \mathsf{WADD}^P(\tau') \cdot (1+o(1)) \\
&= \frac{\log\gamma}{I^*}\cdot(1+o(1)).
\end{aligned}$$

6 NUMERICAL RESULTS

6.1 Synthetic Data Examples

We validate the performance of the DR-CuSum test (22) through a Gaussian simulation. We use the cost function $c(x,x') = \|x-x'\|_2$ and order $s = 2$ in the Wasserstein distance. The true pre- and post-change distributions are $\mathcal{N}(0,1)$ and $\mathcal{N}(0.5,1)$, respectively.

Comparison with CuSum Type Tests and Effect of Radius: We simulate the case of a single post-change scenario ($M = 1$). We first compare the performances for the following three CuSum type tests all have a recursive structure that facilitates implementation (i.e., they have similar computational complexities during the detection phase):

1. The exact CuSum test with *known* pre- and post-change distributions. This is the optimal procedure and provides us with a lower bound for the WADD.
2. The CuSum test that has knowledge of the Gaussian model, and uses the training data to produce a MLE of the post-change mean *and* variance.
3. The proposed DR-CuSum test defined in (22), with different choices of radius.

In Fig. 1, we study the effect of radius under two sizes of post-change training samples *a priori*: small sample size ($n = 25$) and large sample size ($n = 150$). When the number of training samples is small, the DR-CuSum test outperforms the Gaussian MLE CuSum test with various choices of radii. We emphasize that, unlike the latter test, the DR-CuSum test does *not* assume any knowledge of the parametric model for the post-change distribution. This highlights the effectiveness of the DR-CuSum test in dealing with distributional uncertainty, especially in data-driven and non-parametric settings.

In Fig. 2, we numerically study the effect of radius when the empirical samples are drawn from a *mismatched* Gaussian distribution: $\mathcal{N}(0.75,1)$, while the true post-change distribution for test sequences is still $\mathcal{N}(0.5,1)$. We see in Fig. 2 that the model mismatch causes a non-trivial effect on the optimal radius selection, where the

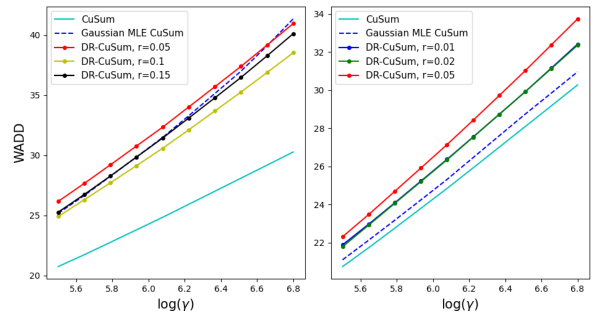

Figure 1: Comparison of detection delay, averaged over 30 sets of different training samples. The left plot corresponds to the case with $n = 25$ post-change training samples, and the right plot corresponds to that with $n = 150$. The tests shown are exact CuSum (cyan), CuSum with Gaussian MLE (blue dashes), and DR-CuSum with various radii.

DR-CuSum test with a larger radius is more robust under distributional mismatch. Also, with a proper choice of radius, we see that the DR-CuSum test outperforms the Gaussian MLE test, which, we again emphasize, knows the parametric model for the post-change distribution. This highlights the effectiveness of DR-CuSum test in dealing with training data mismatch, which is common in data-driven applications.

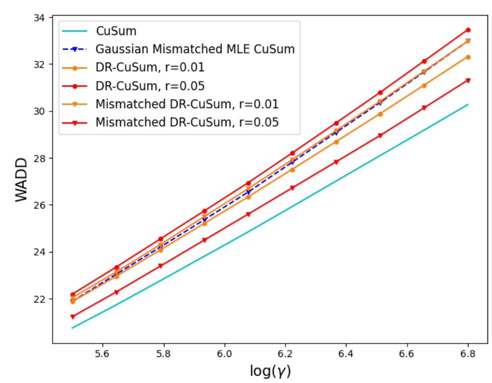

Figure 2: Comparison of detection delay, averaged over 40 sets of $n = 150$ mismatched post-change training samples. The tests shown are exact CuSum (cyan), mismatched CuSum with Gaussian MLE (blue dashes), and both matched (in circle markers) and mismatched (in triangle markers) DR-CuSum tests with two radii.

Comparison with NGLR-CuSum test: We also compare the performance of the DR-CuSum test with the NGLR-CuSum test (Liang and Veeravalli, 2023), which also assumed no knowledge about the post-change distribution. We compare their performance with d dimensional observations. The pre-change distribution is $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, where $\mathbf{I}_d$ is the identity matrix.

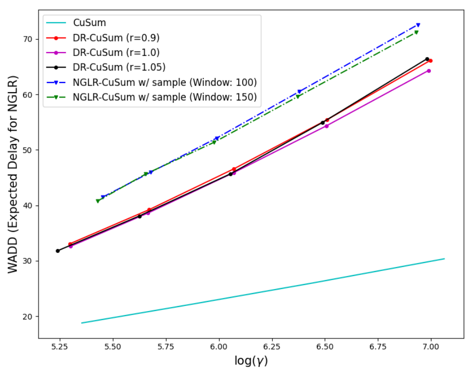

Figure 3: Comparison of DR-CuSum tests (solid lines) with the modified NGLR-CuSum tests (dashed lines) with $d = 3$ and $a = 0.3$. The number of post-change training samples $n = 25$. The leave-one-out KDE with a Gaussian product kernel (defined in (33)) is used in the NGLR-CuSum test, with the bandwidth parameter $h_m = 30^{-1/7}, \forall m = 1, \dots, 3$. The average performance over 30 different sets of training samples is reported.

The post-change distribution is $\mathcal{N}(\boldsymbol{a}, \mathbf{I}_d)$. Here $\boldsymbol{a} \in \mathbb{R}^d$ denotes a vector with all elements being $a \in \mathbb{R}$.

In Fig. 3, we see that with 3-d observations, the DR-CuSum test (with the optimal radius) performs better than the modified NGLR-CuSum test. This is because kernel density estimation becomes less accurate in higher dimensions. Also, it is observed that the DR-CuSum test is computationally much less expensive than the modified NGLR-CuSum test. The kernel density estimation is very computationally demanding in higher dimensions. In comparison, while the DR-CuSum test also suffers from a more expensive offline computation, its online computational requirements only go up modestly due to the increase in dimension. Indeed, the DR-CuSum requires only $O(nd)$ operations to compute $C_{\lambda^*,u^*}(\boldsymbol{X}_k)$ for each new sample $\boldsymbol{X}_k$. More implementation details of the NGLR-CuSum test and a one-dimensional numerical result can be found in Appendix B.

6.2 Real Data Example

We apply the DR-CuSum test to a real data example of human activity detection using the WISDM's Actitracker activity prediction dataset (Lockhart et al., 2011). The attribute at each time is a three-dimensional vector containing the acceleration in x-, y-, and z-axes. We select "Walking" as the nominal pre-change state and our goal is to detect a change to the "Jogging" state (post-change) as quickly as possible.

We mainly compare the proposed DR-CuSum with the NGLR-CuSum test, which is also non-parametric and does not impose any post-change assumptions. For the NGLR-CuSum, we first fit a Gaussian distribution as the pre-change using available historical samples. For the DR-CuSum test, following Remark 3.1, we directly solve for the LFD P^* using the pre-change samples without estimating the pre-change density. In such a real data scenario, we have a fixed set of post-change training samples. Therefore, we can use (17) to select a proper radius that guarantees that the pre-change distribution is excluded from the uncertainty set. We first visualize the trajectory of the DR-CuSum detection statistics for a particular user. Then we provide the comparison of average detection delay (over multiple users) at the end of this subsection.

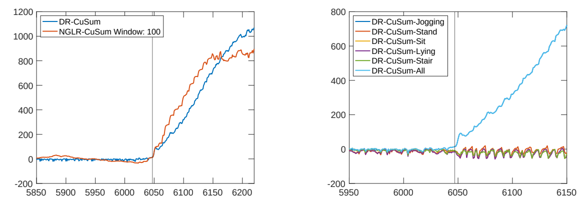

Figure 4: Detection statistics of DR-CuSum and NGLR-CuSum tests. Number of empirical samples from each scenario $n = 5$. Wasserstein order $s = 1$. Radius $r = 8$. Number of scenarios $M = 5$ (right plot only). The vertical line represents the true change-point.

In the first scenario, we consider $M = 1$ with "Jogging" being the true post-change activity. We select $n = 5$ samples from the Jogging state of a specific user to construct the empirical distribution $\widehat{P}$. This represents the case where number of training samples available is very small. The DR-CuSum test, along with the baseline NGLR-CuSum test, is applied on *another* user's data to monitor his/her activity change from Walking to Jogging. Fig. 4 (Left) shows an example where there is a clearer dichotomy for the DR-CuSum statistic before and after the change, and the DR-CuSum statistics is more stable under the pre-change regime.

In the second scenario, we consider $M = 5$ with the post-change activity being one of the Jogging, Stairs, Sitting, Standing, and LyingDown state. We construct the M uncertainty sets using $n = 5$ empirical samples from each state and solve for LFDs $P_1^*, \dots, P_M^*$. The resulting DR-CuSum detection statistics in Fig. 4 (Right) show that the maximum statistic in (22) is helpful not only for change detection, but also for change isolation.

We also compare the average detection delay of DR-CuSum test with that of the NGLR-CuSum. We focus on the detection of activity change from Walking to Jogging and select 86 user sequences that contain such activity change. For each of these 86 sequences, we use n empirical data randomly selected from the post-change state to construct the uncertainty set for the

Table 1: Average detection delay of DR-CuSum and NGLR-CuSum test on 86 user sequences. The bandwidth for NGLR-CuSum is selected as $h_i = W^{-1/(d+4)}\hat{\sigma}_i$, where $W = 100$ is the window size, $d = 3$ is the data dimension, and $\hat{\sigma}_i$ is the estimated standard deviation from pre-change data. The threshold is chosen as the upper 1% quantile of the detection statistics for the pre-change samples. The experiments are repeated ten times and the average detection delay is reported, with standard deviations in parentheses.

	DR-CuSum	NGLR-CuSum
$n = 10$, $r = 2$	25.61 (4.08)	81.26 (0.51)
$n = 20$, $r = 1.5$	16.88 (2.74)	74.77 (13.62)

DR-CuSum test, and for density estimation in NGLR-CuSum. We repeat such procedures ten times for each user to account for the randomness in post-change empirical samples. The average detection delay and standard deviation are reported in Table 1. We can see that the DR-CuSum test tends to have a smaller detection delay than NGLR-CuSum.

7 CONCLUSION AND FUTURE WORK

We developed an asymptotically minimax robust procedure for QCD, which we refer to as the distributionally robust (DR) CuSum test, in the setting where the post-change distribution belongs to a union of data-driven Wasserstein uncertainty sets. We showed that the DR-CuSum test, which makes no distributional assumptions about the post-change, outperforms the Gaussian MLE CuSum test and NGLR-CuSum test. Our theoretical findings can be extended to the non-stationary setting where the post-change observations are independent but not necessarily identically distributed; we leave this extension for future research.

Acknowledgements

The work of Liyan Xie was partially supported by UDF01002142 and 2023SC0019 through the Chinese University of Hong Kong, Shenzhen. The work of Yuchen Liang and Venugopal V. Veeravalli was supported by the U.S. National Science Foundation under grant ECCS-2033900, and by the U.S. Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196, through the University of Illinois at Urbana-Champaign.

References

Adiga, S. and Tandon, R. (2022). Unsupervised change detection using dre-cusum. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*, pages 1103–1110. IEEE.

Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Application*, volume 104. Prentice Hall Englewood Cliffs.

Bolley, F., Guillin, A., and Villani, C. (2007). Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137:541–593.

Dewaskar, M., Tosh, C., Knoblauch, J., and Dunson, D. B. (2023). Robustifying likelihoods by optimistically re-weighting data. *arXiv preprint arXiv:2303.10525*.

Gao, R., Xie, L., Xie, Y., and Xu, H. (2018). Robust hypothesis testing using Wasserstein uncertainty sets. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 7902–7912.

Huber, P. J. (1965). A robust version of the probability ratio test. *Annals of Mathematical Statistics*, 36(6):1753–1758.

Kawahara, Y. and Sugiyama, M. (2009). Change-point detection in time-series data by direct density-ratio estimation. In *Proceedings of the 2009 SIAM international conference on data mining*, pages 389–400. SIAM.

Kawahara, Y. and Sugiyama, M. (2012). Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(2):114–127.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. (2019). Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs.

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

Lai, T. L. and Xing, H. (2010). Sequential change-point detection when the pre- and post-change parameters are unknown. *Sequential Analysis*, 29(2):162–175.

Levy, B. C. (2008). *Principles of Signal Detection and Parameter Estimation*. Springer Science & Business Media.

Li, S., Xie, Y., Dai, H., and Song, L. (2015). M-statistic for kernel change-point detection. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28, pages 3366–3374. Curran Associates, Inc.

Liang, Y. and Veeravalli, V. V. (2023). Quickest change detection with leave-one-out density estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Lockhart, J. W., Weiss, G. M., Xue, J. C., Gallagher, S. T., Grosner, A. B., and Pulickal, T. T. (2011). Design considerations for the WISDM smart phone-based sensor mining architecture. In *Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data*, pages 25–33. ACM.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, 42(6):1897–1908.

Luenberger, D. G. (1969). *Optimization by Vector Space Methods*. Series in decision and control. John Wiley & Sons, Inc., 605 Third Ave, New York, NY, US.

Magesh, A., Sun, Z., Veeravalli, V. V., and Zou, S. (2023). Robust hypothesis testing with moment constrained uncertainty sets. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Mohajerin Esfahani, P. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166.

Molloy, T. L. and Ford, J. J. (2017). Misspecified and asymptotically minimax robust quickest change detection. *IEEE Transactions on Signal Processing*, 65(21):5730–5742.

Moulin, P. and Veeravalli, V. V. (2018). *Statistical Inference for Engineers and Data Scientists*. Cambridge University Press.

Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *Annals of Statistics*, 14(4):1379–1387.

Moustakides, G. V. and Basioti, K. (2019). Training neural networks for likelihood/density ratio estimation.

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

Pollak, M. (1978). Optimality and almost optimality of mixture stopping rules. *Annals of Statistics*, 6(4):910–916.

Pollak, M. (1985). Optimal detection of a change in distribution. *Annals of Statistics*, 13(1):206–227.

Poor, H. V. and Hadjiliadis, O. (2008). *Quickest Detection*. Cambridge University Press.

Raginsky, M. and Sason, I. (2013). Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends® in Communications and Information Theory*, 10(1-2):1–246.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer Series in Statistics. Springer.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., Von Bünau, P., and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746.

Sun, Z. and Zou, S. (2021). A data-driven approach to robust hypothesis testing using kernel MMD uncertainty sets. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 3056–3061. IEEE.

Tartakovsky, A., Nikiforov, I., and Basseville, M. (2015). *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.

Unnikrishnan, J., Veeravalli, V. V., and Meyn, S. P. (2011). Minimax robust quickest change detection. *IEEE Transactions on Information Theory*, 57(3):1604–1614.

Veeravalli, V. V. and Banerjee, T. (2013). Quickest change detection. *Academic Press Library in Signal Processing: Array and Statistical Signal Processing*, 3:209–256.

Villani, C. (2003). *Topics in Optimal Transportation*. Number 58. American Mathematical Society.

Wang, J. and Xie, Y. (2022). A data-driven approach to robust hypothesis testing using Sinkhorn uncertainty sets. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 3315–3320. IEEE.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer, 233 Spring Street, New York, NY 10013, USA.

Xie, L., Zou, S., Xie, Y., and Veeravalli, V. V. (2021). Sequential (quickest) change detection: Classical results and new directions. *IEEE Journal on Selected Areas in Information Theory*, 2(2):494–514.

Checklist

1. For all models and algorithms presented, check if you include:

(a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A TECHNICAL PROOFS

A.1 Proofs for Section 3

We first present the following lemma which is used in the proof of Theorem 3.1.

Lemma A.1. *Given constants $c_1, c_2, \dots, c_n \in \mathbb{R}$, consider the following optimization problem:*

$$\min_{a_1,a_2,\dots,a_n \geq 0} f(a_1, a_2, \dots, a_n) := \Big(\sum_{i=1}^n a_i\Big) \log\Big(\sum_{i=1}^n a_i\Big) + \sum_{i=1}^n c_i a_i, \tag{23}$$

where by convention, we let $0 \log 0 = 0$. Then, the minimizer $(a_1^*, \dots, a_n^*)$ satisfies

$$\sum_{i=1}^n a_i^* = e^{-\min_{i=1,\dots,n} c_i - 1},$$

and the optimal objective value is $f(a_1^*, a_2^*, \dots, a_n^*) = -e^{-\min_{i=1,\dots,n} c_i - 1}$.

Proof. We first note that it suffices to consider the case where all the c_i's are distinct, i.e., $c_i \neq c_j$, $\forall i \neq j$. This is due to the fact that if any pair of c_i and c_j are equal, we can re-define our set of constants to be the unique values in $\{c_1, \dots, c_n\}$, let's denote these as $\{c'_1, \dots, c'_k\}$ $(k < n)$. We then define new variables $\tilde{a}_i = \sum_{j: c_j = c'_i} a_j$, $i = 1, \dots, k$. The problem (23) thus becomes equivalent to:

$$\min_{\tilde{a}_1,\dots,\tilde{a}_k \geq 0} f(\tilde{a}_1, \dots, \tilde{a}_k) := (\sum_{i=1}^k \tilde{a}_i) \log(\sum_{i=1}^k \tilde{a}_i) + \sum_{i=1}^k c'_i \tilde{a}_i,$$

with distinct c'_i values, and this new problem has the same optimal value as the original problem in (23). Therefore, we can safely assume that $c_1, \dots, c_n$ are different from each other for the remaining proof, i.e., $c_i \neq c_j$, $\forall i \neq j$.

We introduce Lagrangian multipliers $\lambda_i \geq 0$ for $i = 1, \dots, n$ for the constraints. The corresponding Lagrangian function is then given by

$$L(a_1, \dots, a_n, \lambda_1, \dots, \lambda_n) = \Big(\sum_{i=1}^n a_i\Big) \log\Big(\sum_{i=1}^n a_i\Big) + \sum_{i=1}^n c_i a_i - \sum_{i=1}^n \lambda_i a_i.$$

By applying the Karush–Kuhn–Tucker condition, the optimal solution $(a_1^*, \dots, a_n^*)$ must satisfy the gradient condition

$$\frac{\partial L}{\partial a_i^*} = 1 + \log\Big(\sum_{i=1}^n a_i^*\Big) + c_i - \lambda_i = 0, \ \forall i = 1, 2, \dots, n, \tag{24}$$

and the complementary slackness conditions

$$\lambda_i a_i^* = 0, \ \forall i = 1, 2, \dots, n. \tag{25}$$

From (24), we deduce that

$$\lambda_i = 1 + \log\Big(\sum_{i=1}^n a_i^*\Big) + c_i, \ \forall i = 1, 2, \dots, n,$$

which implies that $\lambda_1, \dots, \lambda_n$ are distinct. Now, we consider two scenarios:

(i) If $\lambda_i \neq 0$, $\forall i$, then from (25) we get $a_1^* = a_2^* = \cdots = a_n^* = 0$ and the objective value is zero.

(ii) If there exists i_0 such that $\lambda_{i_0} = 0$, then from $\lambda_i \geq 0$, $\forall i$, we have that

$$i_0 = \arg\min \lambda_i = \arg\min c_i.$$

Additionally, by (25), we have $a_j^* = 0$ for $j \neq i_0$ since $\lambda_j \neq \lambda_{i_0} = 0$, and the corresponding $a_{i_0}^* = e^{-c_{i_0} - 1}$ from (24), yielding an objective value of $-e^{-c_{i_0} - 1} < 0$.

Therefore, when $c_1, \dots, c_n$ are distinct and $i_0 = \arg\min c_i$, the minimizer is given by $a_{i_0}^* = e^{-c_{i_0} - 1}$ and $a_j^* = 0$, $\forall j \neq i_0$, and the optimal value is $-e^{-c_{i_0} - 1}$. In summary, the optimal solution $(a_1^*, \dots, a_n^*)$ to (23) satisfies $\sum_{i=1}^n a_i^* = e^{-\min_{i=1,\dots,n} c_i - 1}$, and the corresponding optimal objective value is $-e^{-\min_{i=1,\dots,n} c_i - 1}$. $\square$

Proof of Theorem 3.1

Proof. We first consider $s = 1$, with the radius being denoted by r_1. Denote by $\Pi(P, \widehat{P})$ the space of all joint distributions on $\mathscr{X} \times \mathscr{X}$. Note that the empirical distribution $\widehat{P}$ is *discrete* with finite support $\{\omega_1, \ldots, \omega_n\}$. Without loss of generality, we assume that P^*, the optimal solution to (13), is *absolutely continuous* with respect to the pre-change measure Q because otherwise $\mathsf{KL}(P^*||Q) = \infty$. Since Q is dominated by μ, P^* is also dominated by μ.

Therefore, we consider all joint distributions $\Pi(P, \widehat{P})$ with a continuous marginal P (with respect to μ) and a discrete marginal $\widehat{P}$. Their joint distribution $\Pi(P, \widehat{P})$ can be characterized by the mixed joint density, denoted as

$$\pi(x, \omega_i) = \frac{1}{n} f_i(x), \text{ where } f_i(x) \geq 0, \int_{\mathscr{X}} f_i d\mu(x) = 1, \ \forall i = 1, 2, \ldots, n. \tag{26}$$

Here the term $1/n$ corresponds to the probability mass function of its second marginal, while $f_i(x)$ can be viewed as the conditional density function (with respect to the same dominating measure μ) of the first variable given that the second variable equals ω_i. Thus $\sum_{i=1}^n \pi(x, \omega_i) = \frac{1}{n} \sum_{i=1}^n f_i(x)$ is the probability density function (with respect to μ) of its first marginal P, which we denote as p. Also, define $\mathcal{P}_\mu$ to be the set of all distributions absolutely continuous with respect to measure μ. Note that $\mathcal{P}_\mu$ is trivially convex.

To solve the constrained minimization problem in (13), we note that $\mathsf{KL}(P||Q)$ is a convex functional in P, and $\mathsf{W}_1(P, \widehat{P}) - r_1$ is a convex mapping of P into $\mathbb{R}$. Since $\mathcal{P}_\mu$ is dense with respect to the Wasserstein distance for Lebesgue measure μ, meaning that there exists a distribution in $\mathcal{P}_\mu$ that is arbitrarily close to any given measure P. Thus it is easy to find some $P_1^\circ \ll \mu$ close to the empirical samples such that $\mathsf{W}_1(P_1^\circ, \widehat{P}) \leq r_1$. Also, $\inf \left\{ \mathsf{KL}(P||Q) : P \in \mathcal{P}_\mu, \mathsf{W}_1(P, \widehat{P}) \leq r_1 \right\} \geq 0 > -\infty$. Then by Lagrange duality (Luenberger, 1969, Sec 8.6 Thm 1) we have

$$\inf \left\{ \mathsf{KL}(P||Q) : P \in \mathcal{P}_\mu, \mathsf{W}_1(P, \widehat{P}) \leq r_1 \right\} = \max_{\lambda \geq 0} \inf_{P \in \mathcal{P}_\mu} \left(\mathsf{KL}(P||Q) + \lambda \mathsf{W}_1(P, \widehat{P}) - \lambda r_1 \right), \tag{27}$$

and this maximum on the right-hand side is achieved at some $\lambda^* \geq 0$.

Using the definition of Wasserstein distance, we have

$$\mathsf{W}_1(P, \widehat{P}) = \inf_{\pi \in \Pi(P, \widehat{P})} \sum_{i=1}^n \int_{\mathscr{X}} c(x, \omega_i) \pi(x, \omega_i) d\mu(x),$$

which, after substituting into (27), results in the following dual optimization problem to (13),

$$\max_{\lambda \geq 0} \left(-\lambda r_1 + \inf_\pi \int_{\mathscr{X}} \Big(\sum_{i=1}^n \pi(x, \omega_i) \Big) \log \frac{\sum_{i=1}^n \pi(x, \omega_i)}{q(x)} d\mu(x) + \sum_{i=1}^n \int_{\mathscr{X}} \lambda c(x, \omega_i) \pi(x, \omega_i) d\mu(x) \right).$$

Note that the inner problem can be written as

$$\begin{aligned} &\inf_\pi \left(\int_{\mathscr{X}} \Big(\sum_{i=1}^n \pi(x, \omega_i) \Big) \log \frac{\sum_{i=1}^n \pi(x, \omega_i)}{q(x)} d\mu(x) + \sum_{i=1}^n \int_{\mathscr{X}} \lambda c(x, \omega_i) \pi(x, \omega_i) d\mu(x) \right) \\ &s.t \quad \int_{\mathscr{X}} \pi(x, \omega_i) d\mu(x) = \frac{1}{n}, \forall i = 1, 2, \ldots, n, \ \sum_{i=1}^n \int_{\mathscr{X}} \pi(x, \omega_i) d\mu(x) = 1. \end{aligned}$$

By the definition of mixed joint density $\pi(x, \omega_i)$ in (26), the above problem is equivalent to the following optimization problem over non-negative functions $f_1, f_2, \ldots, f_n$,

$$\begin{aligned} &\inf_{f_1, \ldots, f_n} \left(\frac{1}{n} \int_{\mathscr{X}} \Big(\sum_{i=1}^n f_i(x) \Big) \log \frac{\frac{1}{n} \sum_{i=1}^n f_i(x)}{q(x)} d\mu(x) + \frac{1}{n} \sum_{i=1}^n \int_{\mathscr{X}} \lambda c(x, \omega_i) f_i(x) d\mu(x) \right) \\ &s.t \quad \int_{\mathscr{X}} f_i(x) d\mu(x) = 1, \forall i = 1, 2, \ldots, n, \ \frac{1}{n} \sum_{i=1}^n \int_{\mathscr{X}} f_i(x) d\mu(x) = 1. \end{aligned}$$

We now introduce Lagrangian multipliers $u_i \in \mathbb{R}, i = 1, 2, \ldots, n$ and $\check{\eta} \in \mathbb{R}$ for the constraints. By strong duality (Luenberger, 1969), we have that the value of the above minimization problem becomes

$$\max_{u_1,\ldots,u_n} \inf_{f_1,\ldots,f_n} \left(\frac{1}{n} \int_{\mathscr{X}} \sum_{i=1}^n f_i(x) \Big(\log \frac{\sum_{i=1}^n \frac{1}{n} f_i(x)}{q(x)} + \lambda c(x, \omega_i) - u_i + \check{\eta} \Big) d\mu(x) + \frac{1}{n} \sum_{i=1}^n u_i - \check{\eta} \right).$$

We can then solve the inner infimum problem for each value x to get the optimal $f_1^*(x), \ldots, f_n^*(x)$, since the function inside the integral only depends on the particular value of x. From Lemma A.1 in the appendix, we have that for each x, the inner minimization problem over $f_1(x), \ldots, f_n(x)$ has optimal solution satisfying

$$\frac{1}{n} \sum_{i=1}^n f_i^*(x) = q(x) e^{-\min_i(\lambda c(x,\omega_i) - u_i) - \check{\eta} - 1} = q(x) e^{-C_{\lambda,u}(x) - \check{\eta} - 1},$$

where the last equality is due to the definition in (15), and the corresponding optimum value for each x is $-q(x)e^{-C_{\lambda,u}(x)-\check{\eta}-1}$. Moreover, to satisfy the constraint, the optimal Lagrangian multiplier $\check{\eta}$ must satisfy

$$\check{\eta} + 1 = \log \left(\int_{\mathscr{X}} q(x) e^{-C_{\lambda,u}(x)} d\mu(x) \right).$$

Therefore, $\frac{1}{n}\sum_{i=1}^n f_i^*(x)$ is a probability density function and the corresponding objective value satisfies

$$\begin{aligned} &\frac{1}{n} \int_{\mathscr{X}} \sum_{i=1}^n f_i^*(x) \Big(\log \frac{\sum_{i=1}^n \frac{1}{n} f_i^*(x)}{q(x)} + \lambda c(x, \omega_i) - u_i + \check{\eta} \Big) d\mu(x) - \check{\eta} \\ &= - \int_{\mathscr{X}} q(x) e^{-C_{\lambda,u}(x) - \check{\eta} - 1} d\mu(x) - \check{\eta} = -1 - \check{\eta} \\ &= - \log \left(\int_{\mathscr{X}} q(x) e^{-C_{\lambda,u}(x)} d\mu(x) \right) =: - \log \eta(\lambda, u), \end{aligned}$$

where for notational simplicity we have defined $\eta(\lambda, u) := \int q(x) e^{-C_{\lambda,u}(x)} d\mu(x)$. The resulting outer maximization problem is as in (14). After solving the dual optimization problem (14) and obtaining the optimal dual variable λ^*, u^*, we arrive at the optimal solution to the problem in (13), which is $p^* = p^{\lambda^*,u^*}(x) = \frac{1}{n}\sum_{i=1}^n f_i^*(x) \propto q(x)e^{-C_{\lambda^*,u^*}(x)}$, or more specifically

$$p^*(x) = q(x) e^{-C_{\lambda^*,u^*}(x) - \eta(\lambda^*,u^*)}.$$

In the case of a general order $s \geq 1$, we will have $c^s(x, \omega_i)$ in the above arguments, and the proof follows similarly. The resulting LFD pdf $p^*(x)$ is still an exponentially tilting of $q(x)$. □

Example A.1. For illustrative purposes, we study the LFD under the setting where $c(x, x') = \|x - x'\|_2$ and the Wasserstein order $s = 2$. We also assume univariate data and the standard normal pre-change distribution, i.e., $Q = \mathcal{N}(0, 1)$, and the dominating measure μ is the Lebesgue measure on $\mathbb{R}$. We first derive a closed-form solution of LFD for the extreme case where the number of empirical samples $n = 1$. In this case, the function $C_{\lambda,u}(x)$ defined in Theorem 3.1 equals $C_{\lambda,u}(x) = \lambda(x - \omega_1)^2 - u,\ \forall x$. Then,

$$\eta(\lambda, u) = \int \frac{1}{\sqrt{2\pi}} e^{-x^2/2 - \lambda x^2 + 2\lambda\omega_1 x - \lambda\omega_1^2 + u} dx = \frac{1}{\sqrt{1+2\lambda}} e^{-\frac{\lambda}{1+2\lambda}\omega_1^2 + u},$$

and the optimal solution to problem (14) is given by

$$\lambda^* = \frac{\omega_1^2}{\sqrt{1 + 4r\omega_1^2 - 1}} - \frac{1}{2}, \quad \text{if } r_2 \leq 1 + \omega_1^2,$$

and $\lambda^* = 0$ otherwise. Note that a large radius yields $\lambda^* = 0$ and the LFD will thus be identical to the pre-change distribution. In practice, the radius has to be carefully chosen to avoid such scenarios so that the robust detection problem is well-defined.

For general $n > 1$, we provide an efficient LFD-solving algorithm based on the following decomposition

$$\eta(\lambda, u) = \sum_{i=1}^{n} \int_{I_i} q(x) e^{-\lambda c^s(x,\omega_i)+u_i} dx,$$

where $I_i := \{x \in \mathbb{R} : \lambda c^s(x, \omega_i) - u_i \leq \lambda c^s(x, \omega_j) - u_j,\ \forall j \neq i\}$. Under previous conditions that $s = 2$ and $c(x, x') = \|x - x'\|_2$, for $i = 1, \ldots, n$, we have that

$$\lambda(x - \omega_i)^2 - u_i \leq \lambda(x - \omega_j)^2 - u_j$$

which is equivalent to

$$\begin{array}{ll} 2(\omega_j - \omega_i)x \leq \frac{u_i - u_j}{\lambda} + \omega_j^2 - \omega_i^2,\ \forall j \neq i & \text{if } \lambda > 0 \\ u_i \geq u_j & \text{if } \lambda = 0 \end{array}$$

This implies that I_i is a connected interval, i.e. $I_i = [\underline{l}_i, \bar{l}_i]$. When $\lambda > 0$, we have

$$\underline{l}_i = \max_{j:\omega_j < \omega_i} \left\{ \frac{u_i - u_j}{2\lambda(\omega_j - \omega_i)} + \frac{\omega_j + \omega_i}{2} \right\},$$
$$\bar{l}_i = \min_{j:\omega_j > \omega_i} \left\{ \frac{u_i - u_j}{2\lambda(\omega_j - \omega_i)} + \frac{\omega_j + \omega_i}{2} \right\},$$

and the decomposition yields

$$\eta(\lambda, u) = \sum_{i=1}^{n} \mathbb{I}\left\{\underline{l}_i < \bar{l}_i\right\} \frac{\exp\left(u_i - \frac{\lambda\omega_i^2}{1+2\lambda}\right)}{2\sqrt{2\lambda+1}} \left(\operatorname{erf}\left(\frac{2\lambda(\bar{l}_i - \omega_i) + \bar{l}_i}{\sqrt{4\lambda+2}}\right) - \operatorname{erf}\left(\frac{2\lambda(\underline{l}_i - \omega_i) + \underline{l}_i}{\sqrt{4\lambda+2}}\right)\right),$$

where $\operatorname{erf}(z) := \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$ is the error function. When $\lambda = 0$, $I_i = \mathbb{R}$ if $i = \arg\max u_i$ and $I_i = \emptyset$ otherwise, and thus $\eta(0, u) = \exp(\max_i u_i)$.

For general pre-change distributions, it is not guaranteed that we can find analytical solutions for the LFD for $n > 1$. However, we note that the solution to the optimization problem in (14) is easy to compute numerically for any n, s, and r_s, regardless of the type of the pre-change distribution. This is due to the convexity of the problem in (14).

A.2 Proofs for Section 4

We first present example distributions that satisfy the Transportation-Cost Inequality in Definition 4.1, to demonstrate the wide applicability of the results in Section 4.

Examples of $T_1(c)$ inequality: For a discrete sample space with the Hamming metric $c(x, y) = 1_{\{x \neq y\}}$, the W_1 distance satisfies the following inequality

$$\mathsf{W}_1(P, Q) = \|P - Q\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2}\mathsf{KL}(Q||P)},$$

which is a consequence of Pinsker's inequality. Hence, the $T_1(1/4)$ inequality holds for every probability measure on the discrete sample space.

Examples of $T_2(c)$ inequality: For $\mathcal{X} = \mathbb{R}^n$ and $c(x, y) = \|x - y\|_2$, the standard n-dimensional Gaussian distribution satisfies the $T_2(1)$ inequality, i.e., $\mathsf{W}_2(P, Q) \leq \sqrt{2\mathsf{KL}(Q||P)}$ for P being the n-dimensional Gaussian distribution and Q being any distribution satisfying $Q \ll P$. More generally, if $P = N(\mu, \Sigma)$, where μ is the mean vector and Σ is the covariance matrix, then P satisfies the $T_2(c)$ inequality for $c = 1/2\kappa$, where $\kappa \leq \min_i \lambda_i(\Sigma^{-1})$, representing the smallest eigenvalue of the inverse covariance matrix.

Proof of Corollary 4.1

Proof. From the triangle inequality satisfied by the Wasserstein distance, we have

$$\mathsf{W}_s(Q, \widehat{P}_n) \geq \mathsf{W}_s(P, Q) - \mathsf{W}_s(P, \widehat{P}_n),$$

and thus

$$\begin{aligned}\mathbb{P}^P\{\mathsf{W}_s(Q, \widehat{P}_n) < r_s\} &\leq \mathbb{P}^P\{\mathsf{W}_s(P, \widehat{P}_n) > \mathsf{W}_s(P, Q) - r_s\} \\ &\leq \exp\left(-\gamma_s n(\mathsf{W}_s(P, Q) - r_s)^2/2c\right),\end{aligned}$$

where the last inequality follows from Theorem 4.1. Now, if

$$r_s \leq \mathsf{W}_s(P, Q) - \sqrt{\frac{2|\log \delta| c}{\gamma_s n}},$$

then $\mathbb{P}^P\{Q \in \mathcal{P}_n\} = \mathbb{P}^P\{\mathsf{W}_s(Q, \widehat{P}_n) < r_s\} \leq \delta$. □

Proof of Lemma 4.1

Proof. We first note that when (18) holds we have $\underline{r}_{\delta,n} \leq \overline{r}_{\delta,n}$. Then the result directly follows from the fact that for any two events A, B,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A \cap B^c) \geq \mathbb{P}(A) - \mathbb{P}(B^c).$$

Now, letting $A = \{P \in \mathcal{P}_n\}$ and $B = \{Q \notin \mathcal{P}_n\}$, and applying Corollaries 4.1 and 4.2, we get the desired result. □

Proof of Lemma 4.2

Proof. Throughout the proof we use the result from Lemma 4.1, i.e., under the given conditions, with probability at least $1 - 2\delta$, we have $P \in \mathcal{P}_n$ and $Q \notin \mathcal{P}_n$. To prove the first inequality in (19), note that when $Q \notin \mathcal{P}_n$ and for any $P \in \mathcal{P}_n$,

$$\mathbb{E}^P\left[\log \frac{p_n^*(X_1)}{q(X_1)}\right] = \mathsf{KL}(P||Q) - \mathsf{KL}(P||P_n^*) \overset{(i)}{\geq} \mathsf{KL}(P_n^*||Q) \overset{(ii)}{>} 0,$$

where (i) follows from the WSB condition in Lemma 2.1, and (ii) follows from the fact that $Q \notin \mathcal{P}_n$. Thus, by (Siegmund, 1985, Prop. 8.21), $\mathbb{E}_1^P[\tau_{\mathrm{DR}}] < \infty$. By Wald's identity,

$$\mathbb{E}^P[\tau_{\mathrm{DR}}] = \frac{\mathbb{E}^P[S_{\tau_{\mathrm{DR}}}]}{\mathbb{E}^P\left[\log \frac{p_n^*(X_1)}{q(X_1)}\right]} \leq \frac{\mathbb{E}^P[S_{\tau_{\mathrm{DR}}}]}{\mathsf{KL}(P_n^*||Q)}, \quad \forall P \in \mathcal{P}_n.$$

Since by assumption $\mathbb{E}^P\left[\log \frac{p_n^*(X_1)}{q(X_1)}\right]^2 < \infty$, following classical renewal analysis (e.g., see Siegmund (1985)), we have

$$\mathbb{E}^P[S_{\tau_{\mathrm{DR}}}] = \log \gamma + \mathbb{E}^P[S_{\tau_{\mathrm{DR}}} - \log \gamma] = \log \gamma \cdot (1 + o(1)), \quad \text{as } \gamma \to \infty.$$

Finally, we note that $\mathsf{WADD}^P(\tau_{\mathrm{DR}}) \leq \mathbb{E}_1^P[\tau_{\mathrm{DR}}]$, $\forall P \in \mathcal{P}_n$, since $S_k \geq 0$ for all $k \geq 1$ almost surely and S_k has a recursive update structure. This completes the proof of the first inequality.

For the second inequality in (19), due to the triangle inequality for the Wasserstein metric, we have

$$\mathsf{W}_s(\widehat{P}_n, Q) \leq \mathsf{W}_s(\widehat{P}_n, P_n^*) + \mathsf{W}_s(P_n^*, Q) \overset{(iii)}{\leq} r_s + \mathsf{W}_s(P_n^*, Q), \tag{28}$$

where (iii) is due to $P_n^* \in \mathcal{P}_n$, and thus $\mathsf{W}_s(\widehat{P}_n, P_n^*) \leq r_s$. On the other hand, under the event $P \in \mathcal{P}_n$, we have $\mathsf{W}_s(\widehat{P}_n, P) \leq r_s$, which implies

$$\mathsf{W}_s(\widehat{P}_n, Q) \geq \mathsf{W}_s(P, Q) - \mathsf{W}_s(\widehat{P}_n, P) \geq \mathsf{W}_s(P, Q) - r_s. \tag{29}$$

Combining (28) and (29) we have

$$\mathsf{W}_s(P_n^*, Q) \geq \mathsf{W}_s(\widehat{P}_n, Q) - r_s \geq \mathsf{W}_s(P, Q) - 2r_s. \tag{30}$$

Further, since Q satisfies the $T_s(c)$ inequality, we have

$$\frac{(\mathsf{W}_s(P_n^*, Q))^2}{2c} \leq \mathsf{KL}(P_n^* || Q).$$

Combining this with (30), we obtain

$$\mathsf{KL}(P_n^* || Q) \geq \frac{(\mathsf{W}_s(P, Q) - 2r_s)^2}{2c}.$$

Substituting this into the first inequality in (19) completes the proof. □

A.3 Proofs for Section 5

Proof of Lemma 5.1

Proof. We define the following quantity

$$Z_j^{(m)} := \log \frac{p_{(m)}^*(X_j)}{q(X_j)}.$$

First, we can assume that $\mathbb{E}_\infty[\tau_{\mathrm{DR}}(b)] < \infty$, as otherwise the statement would be trivial. Consider the following test based on the Shiryaev-Roberts (SR) statistic,

$$\tau_b^R := \inf \left\{ k \in \mathbb{N} : \sum_{m=1}^{M} \sum_{n=1}^{k} \prod_{j=n}^{k} e^{Z_j^{(m)}} =: \sum_{m=1}^{M} R_k^{(m)} \geq e^b \right\}.$$

Note that $\tau_b^R \leq \tau_{\mathrm{DR}}(b)$ since $\sum_{m=1}^M R_k^{(m)} \geq \max_{m=1,\dots,M} S_k^{(m)}$. Hence, $\mathbb{E}_\infty\left[\tau_b^R\right] < \infty$. Denoting $R_k := \sum_{m=1}^M R_k^{(m)}$, we have

$$\mathbb{E}_\infty\left[R_k | \mathcal{F}_{k-1}\right] = \sum_{m=1}^{M} \mathbb{E}_\infty\left[(1 + R_{k-1}^{(m)}) e^{Z_k^{(m)}} | \mathcal{F}_{k-1}\right] = \sum_{m=1}^{M} (1 + R_{k-1}^{(m)}) = M + R_{k-1},$$

which implies that the sequence $\{(R_k - Mk)\}_{k \geq 1}$ forms a martingale. Furthermore, since $R_k \in (0, e^b)$ almost surely on the event $\{\tau_b^R > k\}$, we have for any $k \geq 1$,

$$\begin{aligned} \mathbb{E}_\infty\left[\left|(R_{k+1} - M(k+1)) - (R_k - Mk)\right| \Big| \mathcal{F}_k\right] &= \mathbb{E}_\infty\left[\left|R_{k+1} - R_k - M\right| \Big| \mathcal{F}_k\right] \\ &\leq \mathbb{E}_\infty\left[R_{k+1} | \mathcal{F}_k\right] + (R_k + M) = 2(R_k + M) \leq 2(e^b + M) \end{aligned}$$

almost surely on the event $\{\tau_b^R > k\}$. Therefore, by the Optional Stopping Theorem,

$$\mathbb{E}_\infty[R_{\tau_b^R}] = M \mathbb{E}_\infty[\tau_b^R].$$

Since $R_{\tau_b^R} \geq e^b$, it follows that $\mathbb{E}_\infty[\tau_b^R] \geq M^{-1} e^b$ and consequently

$$\mathbb{E}_\infty[\tau_{\mathrm{DR}}(b)] \geq \mathbb{E}_\infty[\tau_b^R] \geq M^{-1} e^b.$$

□

Proof of Theorem 5.1

Proof. Let $\mathcal{P}_m := \mathcal{P}_{n_m}^{(m)}$. Recall that $P^*_{(m)}$ represents the LFD for the m-th class and that $p^*_{(m)}$ denotes its density with respect to the dominating measure μ. For the lower bound, we have

$$\inf_{\tau' \in C(\gamma)} \max_{i=1,\dots,M} \sup_{P_i \in \mathcal{P}_i} \mathsf{WADD}^{P_i}(\tau') \geq \max_{i=1,\dots,M} \sup_{P_i \in \mathcal{P}_i} \inf_{\tau' \in C(\gamma)} \mathsf{WADD}^{P_i}(\tau') \geq \frac{\log \gamma}{I^*}(1+o(1)),$$

where the last inequality follows from (Lai, 1998, Thm. 1) and the weak law of large numbers for independent random variables.

For the upper bound, we consider the detection rule $\tau_{\mathrm{DR}}(b)$ defined in (22). For any distribution $P_i \in \mathcal{P}_i$ we have

$$\mathbb{E}^{P_i}\left[\log \frac{p^*_{(i)}(X)}{q(X)}\right] = \mathsf{KL}(P_i||Q) - \mathsf{KL}(P_i||P^*_{(i)}) \geq \mathsf{KL}(P^*_{(i)}||Q),$$

where the last inequality is a consequence of the weak stochastic boundedness condition. Therefore, according to (Lai, 1998, Thm. 4(ii)), we have

$$\max_{i=1,\dots,M} \sup_{P_i \in \mathcal{P}_i} \mathsf{WADD}^{P_i}(\tau_{\mathrm{DR}}(b)) \leq \frac{b}{I^*}(1+o(1)).$$

By selecting $b = b_\gamma = \log \gamma + \log M$, we obtain

$$\max_{i=1,\dots,M} \sup_{P_i \in \mathcal{P}_i} \mathsf{WADD}^{P_i}[\tau_{\mathrm{DR}}(b_\gamma)] \leq \frac{\log \gamma + \log M}{I^*}(1+o(1)) = \frac{\log \gamma}{I^*}(1+o(1)).$$

Finally, from Lemma 5.1, the proof is complete since $\tau_{\mathrm{DR}}(b_\gamma) \in C(\gamma)$ when $b = b_\gamma$. □

B IMPLEMENTATION DETAILS AND ADDITIONAL EXPERIMENTS

The original NGLR-CuSum test, as introduced in Liang and Veeravalli (2023), does not use any post-change training samples. For a fair comparison with the proposed DR-CuSum test, we define and implement a modified version that uses the post-change training samples in the following.

The NGLR-CuSum test in Liang and Veeravalli (2023) is defined as

$$\tau_{\mathrm{NGLR}}(b) := \inf\left\{k \geq 1 : \max_{(k-W)^+ < \ell \leq k} \sum_{j=\ell}^{k} \log \frac{\widehat{p}^{k,\ell}_{-j}(X_j)}{q(X_j)} \geq b\right\}, \tag{31}$$

where W is the window size, and if we assume using a kernel density estimator (KDE) with some kernel function $K(\cdot)$ and bandwidth h (Wasserman, 2006), the leave-one-out density estimate is

$$\widehat{p}^{k,\ell}_{-j}(X_j) := \frac{1}{(k-\ell)h} \sum_{\substack{i=\ell \\ i \neq j}}^{k} K\left(\frac{X_i - X_j}{h}\right), \quad \forall j \in [\ell, k].$$

To utilize the post-change training samples, we similarly define the modified NGLR-CuSum test as:

$$\tau_{\mathrm{NGLRws}}(b) := \inf\left\{k \geq 1 : \max_{(k-W)^+ < \ell \leq k} \left(\sum_{j=\ell}^{k} \log \frac{\widehat{p}^{k,\ell,\omega}_{-j}(X_j)}{q(X_j)} + \sum_{i=1}^{n} \log \frac{\widehat{p}^{k,\ell,\omega}_{-j}(\omega_i)}{q(\omega_i)}\right) \geq b\right\}, \tag{32}$$

where

$$\widehat{p}^{k,\ell,\omega}_{-j}(X_j) := \frac{1}{(k-\ell+n)h}\left(\sum_{\substack{i=\ell \\ i \neq j}}^{k} K\left(\frac{X_i - X_j}{h}\right) + \sum_{i=1}^{n} K\left(\frac{\omega_i - X_j}{h}\right)\right), \ \forall j \in [\ell, k].$$

We compare the detection delay of the DR-CuSum and NGLR-CuSum tests. Note that we simulate the detection delay under the setting $\nu = 1$, i.e., all samples are from the post-change regime. We emphasize that due to the recursive structure of the DR-CuSum statistics and the independence in observations, the worst-case value of the change-point for computing the WADD in (3) is $\nu = 1$. This allows us to estimate the worst-case delays of the DR-CuSum test by simulating the post-change distribution from time 1. However, the choice of change-point $\nu = 1$ does not guarantee a worst-case delay for the NGLR-CuSum test.

In Fig. 5, we compare the detection delay (simulated under the case $\nu = 1$) of DR-CuSum and NGLR-CuSum test. We use the same setting for pre- and post-change distribution as in Fig 1, i.e., the true pre- and post-change distributions are $\mathcal{N}(0, 1)$ and $\mathcal{N}(0.5, 1)$, respectively. We see that given the training samples, the DR-CuSum test (with the optimal radius) performs slightly worse than the modified NGLR-CuSum test. However, due to the recursive CuSum update structure, the DR-CuSum test is computationally less expensive than the latter at inference time.

For the multi-dimensional data as in Fig 3, we use the following product kernel in the leave-one-out density estimate. For any $j \in [\ell, k]$,

$$\widehat{p}_{-j}^{k,\ell,\omega}(\boldsymbol{X}_j) := \frac{1}{(k-\ell+n)\prod_{m=1}^{d} h_m} \times \left(\sum_{\substack{i=\ell \\ i \neq j}}^{k} \prod_{m=1}^{d} K\left(\frac{\boldsymbol{X}_i^{(m)} - \boldsymbol{X}_j^{(m)}}{h_m} \right) + \sum_{i=1}^{n} \prod_{m=1}^{d} K\left(\frac{\boldsymbol{\omega}_i^{(m)} - \boldsymbol{X}_j^{(m)}}{h_m} \right) \right). \tag{33}$$

where $\boldsymbol{x}^{(m)}$ denotes the m-th element of vector x and h_m denotes the kernel bandwidth for the m-th element.

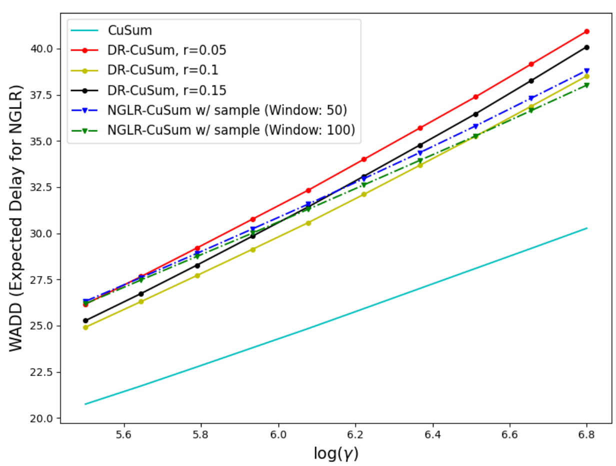

Figure 5: Comparison of DR-CuSum tests (solid lines) with the modified NGLR-CuSum tests (dashed lines) defined in (32). The number of post-change training samples $n = 25$. The KDE with a Gaussian kernel is used in the NGLR-CuSum test, with the bandwidth parameter $h = 50^{-0.2}$. All tests are first evaluated on the same set of post-change training samples, and then the average performance over 30 different sets of training samples is reported.