Quickest Change Detection With Post-Change Density Estimation

Yuchen Liang, *Member, IEEE*, and Venugopal V. Veeravalli, *Fellow, IEEE*

***Abstract*— The problem of quickest change detection in a sequence of independent observations is considered. The pre-change distribution is assumed to be known, while the post-change distribution is unknown. Two tests based on post-change density estimation are developed for this problem, the window-limited non-parametric generalized likelihood ratio (NGLR) CuSum test and the non-parametric window-limited adaptive (NWLA) CuSum test. Both tests do not assume any knowledge of the post-change distribution, except that the post-change density satisfies certain smoothness conditions that allows for efficient non-parametric estimation; also, they do not require any pre-collected post-change training samples. Under certain convergence conditions on the density estimator, it is shown that both tests are first-order asymptotically optimal, as the false alarm rate goes to zero. The analysis is validated through numerical results, where both tests are compared with baseline tests that have distributional knowledge.**

***Index Terms*— Quickest change detection (QCD), non-parametric statistics, (kernel) density estimation, sequential methods.**

I. Introduction

THE problem of quickest change detection (QCD) is of fundamental importance in mathematical statistics (see, e.g., [2], [3] for an overview). Given a sequence of observations whose distribution changes at some unknown change-point, the goal is to detect the change in distribution as quickly as possible after it occurs, while controlling the false alarm rate. In classical formulations of the QCD problem, it is assumed that the pre- and post-change distributions are known, and that the observations are independent and identically distributed (i.i.d.) in the pre- and post-change regimes. However, in many practical situations, while it is reasonable to assume that we can accurately estimate the pre-change distribution, the post-change distribution is rarely completely known.

Manuscript received 21 November 2023; revised 30 March 2024; accepted 17 June 2024. Date of publication 24 June 2024; date of current version 22 October 2024. This work was supported in part by the U.S. National Science Foundation under Grant ECCS-2033900. This paper was presented in part at the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing [1]. *(Corresponding author: Venugopal V. Veeravalli.)*

Yuchen Liang was with the ECE Department and the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA. He is now with the ECE Department, The Ohio State University, Columbus, OH 43210 USA (e-mail: liang.1439@osu.edu).

Venugopal V. Veeravalli is with the ECE Department and the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: vvv@illinois.edu).

Communicated by L. Lai, Associate Editor for Signal Processing and Source Coding.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2024.3418379.

Digital Object Identifier 10.1109/TIT.2024.3418379

There have been extensive efforts to address pre- and/or post-change distributional uncertainty in QCD problems. In the case where both distributions are not fully known, one approach is to assume that the distributions are parametrized by a (low-dimensional) parameter that comes from a pre-defined parameter set, and to employ a generalized likelihood ratio (GLR) approach for detection. This approach was first introduced in [4] and later analyzed in more detail in [5]. In particular, in [5], it is assumed that the pre-change distribution is known and that the post-change distribution comes from a parametric family, with the parameter being finite-dimensional. A window-limited GLR test is proposed, which is shown to be asymptotically optimal under certain smoothness conditions. This work has recently been extended to non-stationary post-change settings [6]. For the setting considered in [5], a window-limited adaptive approach to constructing a QCD test was developed in recent work [7]. This adaptive test is also shown to achieve first-order asymptotic optimality [7]. In this paper, one of the test constructions for the case where the post-change is completely unknown is based on extending techniques introduced in [7].

We assume complete knowledge of the pre-change distribution, while not making any parametric assumptions about the post-change distribution. There has been prior work along these lines. One approach is to replace the log-likelihood ratio by some other useful statistic for distinguishing between distributions in constructing tests. Examples of this approach include the use of kernel M-statistics [8], [9], one-class SVMs [10], nearest neighbors [11], [12], and Geometric Entropy Minimization [13], [14]. In [8], a test is proposed that compares the kernel maximum mean discrepancy (MMD) within a window to a given threshold. A way to set the threshold is also proposed that meets the false alarm rate asymptotically [8]. Another approach is to estimate the log-likelihood ratio and thus the CuSum test statistic through a pre-collected training dataset. This include direct kernel estimation [15] and, more recently, neural network estimation [16]. However, the tests proposed in [8], [9], [10], [11], [12], [13], [14], [15], and [16] lack explicit performance guarantees on the detection delay. In [17], a binning approach is proposed for the QCD problem for the case where the pre-change distribution is known, and without any pre-collected training set for the post-change distribution. Asymptotic optimality of a binning based test is established for the case where the post-change distribution is distinguishable from the pre-change with binning, and where both distribu-

tions have discrete support. Our approach is based on using density estimators rather than binning. We show that the two detectors that we propose are asymptotically optimal for a far wider class of distributions, including those having continuous support.

Our contributions are as follows:

1) We propose a window-limited non-parametric generalized likelihood ratio (NGLR) CuSum test and a non-parametric window-limited adaptive (NWLA) CuSum test, both of which do not assume any knowledge of the post-change distribution (except that the post-change density satisfies certain smoothness conditions that allows for efficient non-parametric estimation), and do not require any post-change training data.
2) We characterize a generic class of density estimators that enable detection.
3) For both tests, we provide a way to set the test threshold to meet false alarm constraints (asymptotically).
4) We show that both proposed tests are first-order asymptotically optimal with the selected thresholds, as the false alarm rate goes to zero.
5) We validate our analysis through numerical results, in which we compare both tests with baseline tests that have distributional knowledge.

The rest of the paper is structured as follows. In Section II, we describe some properties required of the density estimators for asymptotically optimal QCD. In Section III, we propose the NGLR-CuSum test and analyze its theoretical performance. In Section IV, we study the performance of the NWLA-CuSum test. Both tests are analyzed under the assumption that the post-change distribution is completely unknown. In Section V, we present numerical results that validate the theoretical analysis. In Section VI, we provide some concluding remarks.

A preliminary version of the results in this paper for the NGLR-CuSum test appeared in [1].

II. Density Estimators for Quickest Change Detection

Let $X_1, X_2, \cdots \in \mathbb{R}^d$ be i.i.d. observations drawn from an unknown distribution, with probability density function (or *density*) p with respect to some dominating measure μ, and let $\mathrm{supp}(p)$ be the support of p. Let E_p and V_p denote, respectively, the expectation and variance operator on the sequence of observations, when the density corresponding to each observation is p. For two densities p and q on $\mathbb{R}^d$ with respect to μ, the Kullback-Leibler (KL) divergence is defined as:

$$D(p||q) := \int_{\mathrm{supp}(p)} \log(p(x)/q(x))p(x)d\mu(x).$$

Define $X^{[k,n]} := X_k, \ldots, X_n$. Let $\widehat{p}_{-i}^{n,k}$ be a density on $\mathbb{R}^d$ with respect to μ that is estimated using $X_{-i}^{[k,n]} := X_k, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$, where the subscript $-i$ represents that X_i, with $k \le i \le n$, is the observation that is left out from $X^{[k,n]}$. We refer to $\widehat{p}_{-i}^{n,k}$ as a leave-one-out (LOO) estimator. Note that $\widehat{p}_{-i}^{n,k}$ and X_i are independent for each $1 \le k \le i \le n$.

With some possible abuse of notation, we also define

$$\widehat{p}_n^w := \widehat{p}_{-n}^{n,n-w}$$

to be the LOO estimate of p obtained from the past w i.i.d. samples from p.

Assumption 1 (KL-Loss of Estimator): Suppose that, for large enough w, there exist constants $0 < \beta_1, C_1, C_2 < \infty$ and $0 < \beta_2 < 2$ (that depend only on the density p and the estimation procedure) such that the KL loss [18] of the density estimator satisfies

$$\text{KL-loss}(\widehat{p}_n^w) := \mathrm{E}_p\left[D(p||\widehat{p}_n^w)\right] \le \frac{C_1}{w^{\beta_1}}$$

where the KL divergence and the expectation operator E_p are taken over the randomness of X_n and $\widehat{p}_n^w$, respectively.

Assumption 2 (Vanishing Second Moment): The second moment of the log-likelihood ratio satisfies

$$\mathrm{E}_p\left[\left(\log\frac{p(X_n)}{\widehat{p}_n^w(X_n)}\right)^2\right] \le \frac{C_2}{w^{\beta_2}}.$$

Here the expectation operator E_p is taken over the randomness of both X_n and $\widehat{p}_n^w$. Recall that X_n is independent of $\widehat{p}_n^w$.

Similar assumptions to Assumptions 1 and 2 are imposed for general $\widehat{p}_{-i}^{n,k}$ as follows.

Assumption 3 (KL-loss of estimator, LOO): When $n-k$ is large enough, for each $k \le i \le n$,

$$\text{KL-loss}(\widehat{p}_{-i}^{n,k}) = \mathrm{E}_p\left[D(p||\widehat{p}_{-i}^{n,k})\right] \le \frac{C_1}{(n-k)^{\beta_1}}.$$

Assumption 4 (Vanishing Second Moment, LOO): When $n-k$ is large enough, for each $k \le i \le n$,

$$\mathrm{E}_p\left[\left(\frac{1}{(n-k+1)}\sum_{i=k}^{n}\log\frac{p(X_i)}{\widehat{p}_{-i}^{n,k}(X_i)}\right)^2\right] \le \frac{C_2}{(n-k+1)^{\beta_2}}.$$

A typical loss measure for a density estimator is the mean-integrated squared error (MISE), defined as (see, e.g., [19, Chap. 2])

$$\begin{aligned}\mathrm{MISE}(p,\widehat{p}_n^w) &= \mathrm{E}_p\left[\int(\widehat{p}_n^w(x_n) - p(x_n))^2 d\mu(x_n)\right] \\ &= \mathrm{E}_p\left[\|\widehat{p}_n^w - p\|_2^2\right]. \end{aligned} \tag{1}$$

The following lemma connects the MISE measure with the bounds in Assumptions 1–4. The proof is given in the Appendix.

Lemma 1: Suppose that there exist $\overline{\zeta}, \underline{\zeta}$ such that

$$0 < \underline{\zeta} \le p(x), \widehat{p}_n^w(x) \le \overline{\zeta} < \infty, \ \forall x \in \mathrm{supp}(p). \tag{2}$$

If the estimator achieves

$$\mathrm{MISE}(p,\widehat{p}_n^w) \le \frac{C_3}{w^{\beta_3}}, \tag{3}$$

for all w large enough and for some constants $0 < \beta_3, C_3 < \infty$, then Assumptions 1–4 are satisfied with

$$C_1 = \frac{C_3}{\underline{\zeta}}, \quad C_2 = \frac{\overline{\zeta} r C_3}{\underline{\zeta}^2}, \quad \beta_1 = \beta_2 = \beta_3$$

where

$$r := \left(\frac{\log(\underline{\zeta}/\overline{\zeta})}{(\underline{\zeta}/\overline{\zeta}) - 1} \right)^2. \tag{4}$$

In the following, for any positive functions $g(w), h(w)$, the notation $h(w) = O(g(w))$ means that $\frac{h(w)}{g(w)} \xrightarrow{w\to\infty} L < \infty$, and $h(w) = \Omega(g(w))$ means that $\frac{h(w)}{g(w)} \xrightarrow{w\to\infty} L > 0$.

Corollary 1: Suppose that (2) is satisfied with

$$\overline{\zeta} = \overline{\zeta}_w = O(w^{\overline{\beta}}), \quad \underline{\zeta} = \underline{\zeta}_w = \Omega(w^{-\underline{\beta}})$$

such that

$$\underline{\beta} < \beta_3/2, \quad \overline{\beta} < \beta_3 - 2\underline{\beta}.$$

Suppose that the estimator still achieves (3). Then, Assumptions 1–4 are still satisfied, with

$$\beta_1 = \beta_3 - \underline{\beta}, \quad \beta_2 = \beta_3 - 2\underline{\beta} - \overline{\beta} - \varrho$$

where $\varrho > 0$ is a small constant such that β_2 is still positive. The proof of this corollary is given in the Appendix.

An example of a density estimator that satisfies Assumptions 1–4 (under condition (2) and when the density satisfies some smoothness condition) is the *kernel* density estimator (KDE).

Example 1 (Kernel Density Estimator (KDE)): For the case where μ is the Lebesgue measure on $\mathbb{R}^d$, given observations $X_1, \ldots, X_w$, a product kernel can be used to estimate the density. The KDE is defined as

$$\widehat{p}_n^w(\boldsymbol{x}_n) = \frac{1}{w \prod_{i=1}^d h^{(i)}} \sum_{j=n-w}^{n-1} \prod_{i=1}^d K\left(\frac{x_n^{(i)} - X_j^{(i)}}{h^{(i)}} \right). \tag{5}$$

Here $K(\cdot) \geq 0$ is a kernel function, $x^{(i)}$, $i = 1, \ldots, d$ is the i-th element of a vector $\boldsymbol{x} \in \mathbb{R}^d$, and $\boldsymbol{h}$ is a vector for smoothing parameter. Define the γ-Hölder density class as

$$\mathcal{H}_\gamma := \Big\{ p : \int p(x)dx = 1, \exists L > 0, \\ \left| p^{(\ell)}(x_1) - p^{(\ell)}(x_2) \right| \leq L \left\| x_1 - x_2 \right\|^{\gamma - \ell}, \\ \forall x_1, x_2 \in \mathrm{supp}(p) \Big\}.$$

Here $\gamma > 0$ and $\ell = \lfloor \gamma \rfloor$. Further, let the kernel function $K(\cdot)$ satisfy

$$\int K(u)du = 1, \quad \int u^j K(u)du = 0, \ j = 1, \ldots, \ell. \tag{6}$$

Then, with a properly chosen $\boldsymbol{h}$, it can be shown that [20], [21]:

$$\sup_{p \in \mathcal{H}_\gamma} \mathrm{MISE}(p, \widehat{p}_n^w) = O(w^{-\frac{2\gamma}{2\gamma + d}}).$$

Therefore, if the condition (2) is further satisfied, from Lemma 1, we have

$$\beta_1 = \beta_2 = \frac{2\gamma}{2\gamma + d}. \tag{7}$$

We note that the actual choices of β_1 and β_2 do not affect the first-order asymptotic optimality results given in Thm 1 and Thm 2.

III. QCD With NGLR-CuSum Test

Let $X_1, X_2, \ldots, X_n, \cdots \in \mathbb{R}^d$ be a sequence of independent random variables (or vectors), and let ν be a change-point. Assume that $X_1, \ldots, X_{\nu-1}$ all have density p_0 with respect to some dominating measure μ. Furthermore, assume that $X_\nu, X_{\nu+1}, \ldots$ have densities p_1 also with respect to μ. Here p_0 is assumed to be completely known. Regarding p_1, we only assume that Assumptions 3 and 4 are satisfied. Let $(\mathcal{F}_n)_{n\geq 0}$ be the filtration, with $\mathcal{F}_0 = \{\Omega, \varnothing\}$ and $\mathcal{F}_n = \sigma\{X_\ell, 1 \leq \ell \leq n\}$ being the sigma-algebra generated by the set of n observations $X_1, \ldots, X_n$. Furthermore, let $\mathcal{F}_\infty = \sigma(X_1, X_2, \ldots)$.

Let $\mathbb{P}_\nu$ denote the probability measure on the entire sequence of observations when the change-point is ν, and let $\mathbb{E}_\nu$ denote the corresponding expectation. The change-time ν is assumed to be unknown but deterministic. The problem is to detect the change quickly, while controlling the false alarm rate. Let τ be a stopping time [22] defined on the observation sequence associated with the detection rule, i.e. τ is the time at which we stop taking observations and declare that the change has occurred.

A. QCD Problem Formulation and Classical Results

When p_1 is known, Lorden [4] proposed solving the following optimization problem to find the best stopping time τ:

$$\inf_{\tau \in \mathcal{C}_\alpha} \mathrm{WADD}(\tau) \tag{8}$$

where

$$\mathrm{WADD}(\tau) := \sup_{\nu \geq 1} \operatorname{ess\,sup} \mathbb{E}_\nu \left[(\tau - \nu + 1)^+ \,|\mathcal{F}_{\nu-1} \right] \tag{9}$$

characterizes the worst-case delay, and the constraint set is

$$\mathcal{C}_\alpha := \{\tau : \mathrm{FAR}(\tau) \leq \alpha\} \tag{10}$$

with $\mathrm{FAR}(\tau) := \frac{1}{\mathbb{E}_\infty[\tau]}$, which guarantees that the false alarm rate of the algorithm does not exceed α. Here, $\mathbb{E}_\infty[\cdot]$ is the expectation operator when the change never happens, and $(\cdot)^+ := \max\{0, \cdot\}$.

Lorden also showed that Page's Cumulative Sum (CuSum) algorithm [23] whose test statistic is given by:

$$W(n) = \max_{1 \leq k \leq n} \sum_{i=k}^n \log \frac{p_1(X_i)}{p_0(X_i)} = (W(n-1))^+ + \log \frac{p_1(X_n)}{p_0(X_n)}$$

solves the problem in (8) asymptotically as $\alpha \to 0$. The CuSum stopping rule is given by:

$$\tau_{\text{Page}}(b) := \inf\{n : W(n) \geq b\}. \tag{11}$$

It was shown by Moustakides [24] that the CuSum test is exactly optimal for the problem in (8) with some threshold b_α that meets the false alarm constraint exactly, where $b_\alpha \sim |\log \alpha|$. Here and throughout the rest of this paper, we employ standard notations as follows. Let $o(x)$ stand for a function $h(x) \geq 0$ such that $\limsup_{x \to x_0} \left| \frac{h(x)}{x} \right| = 0$. Let $\omega(x)$ stand for a function $h(x) \geq 0$ such that $\limsup_{x \to x_0} \left| \frac{h(x)}{x} \right| = \infty$. Let $O(x)$ stand for a function $h(x) \geq 0$ such that

$\limsup_{x\to x_0}\left|\frac{h(x)}{x}\right| < \infty$. Let $\Theta(x)$ stand for a function $h(x) \geq 0$ such that $\limsup_{x\to x_0}\left|\frac{h(x)}{x}\right| = L \in (0,\infty)$. Let $A_\alpha \sim B_\alpha$ be equivalent to $A_\alpha = B_\alpha(1+o(1))$. If not explicitly specified, $x \to x_0$ refers to $\alpha \to 0$ or $b \to \infty$.

Thus, we have the first-order asymptotic approximation as:

$$\inf_{\tau\in\mathcal{C}_\alpha} \mathrm{WADD}\left(\tau\right) = \mathrm{WADD}\left(\tau_{\mathrm{Page}}\left(b_\alpha\right)\right) \sim \frac{|\log\alpha|}{I} \tag{12}$$

as $\alpha \to 0$. Here we define

$$I := D(p_1||p_0).$$

When the post-change distribution has parametric uncertainties, Lai [5] generalized this performance guarantee with the following assumptions. Let $\theta \in \Theta$ be the post-change parameter, and denote the post-change density as p_1^θ. Define $\mathbb{P}_\nu^\theta$ and $\mathbb{E}_\nu^\theta$ to be the probability and expectation operator on the sequence, respectively, when the true post-change density is p_1^θ. For fixed $\theta \in \Theta$, define the worst-case average detection delay as:

$$\mathrm{WADD}^\theta\left(\tau\right) := \sup_{\nu\geq 1} \operatorname{ess\,sup} \mathbb{E}_\nu^\theta\left[\left(\tau-\nu+1\right)^+ |\mathcal{F}_{\nu-1}\right]. \tag{13}$$

Under parametric uncertainty, the goal is to find a test that belongs to $\mathcal{C}_\alpha$ (see (10) and achieves

$$\inf_{\tau\in\mathcal{C}_\alpha} \mathrm{WADD}^\theta\left(\tau\right) \tag{14}$$

for every $\theta \in \Theta$.

Define $I^\theta := D(p_1^\theta||p_0)$. Suppose that p_0 and p_1^θ satisfy the following assumptions.

Assumption 5 (Right-tail Condition for True LLR): For any $\delta > 0$,

$$\sup_{\nu\geq 1}\mathbb{P}_\nu^\theta\left\{\max_{t\leq n}\sum_{i=\nu}^{\nu+t}\log\frac{p_1^\theta(X_i)}{p_0(X_i)} \geq (1+\delta)nI^\theta\right\} \xrightarrow{n\to\infty} 0.$$

Assumption 6 (Left-tail Condition for True LLR): For any $\delta \in (0,1)$,

$$\sup_{t\geq \nu}\mathbb{P}_\nu^\theta\left\{\sum_{i=t}^{t+n}\log\frac{p_1^\theta(X_i)}{p_0(X_i)} \leq (1-\delta)nI^\theta\right\} \xrightarrow{n\to\infty} 0.$$

Then, if the window size m_α satisfies

$$\liminf m_\alpha/\left|\log\alpha\right| > \frac{1}{I^\theta} \quad \text{and} \quad \log m_\alpha = o(|\log\alpha|),$$

under some smoothness conditions [5], the window-limited GLR-CuSum test:

$$\tilde{\tau}_{\mathrm{GLR}}\left(b\right) := \inf\left\{n\geq 1 : \max_{(n-m_\alpha)^+<k\leq n}\sup_{\theta\in\Theta}\sum_{i=k}^{n}\log\frac{p_1^\theta(X_i)}{p_0(X_i)} \geq b\right\} \tag{15}$$

with test threshold $b_\alpha = |\log\alpha|\left(1+o(1)\right)$ solves the problem in (14) asymptotically as $\alpha \to 0$, for every $\theta \in \Theta$. The asymptotic performance is

$$\inf_{\tau\in\mathcal{C}_\alpha} \mathrm{WADD}^\theta\left(\tau\right) \sim \mathrm{WADD}^\theta\left(\tilde{\tau}_{\mathrm{GLR}}\left(b_\alpha\right)\right) \sim \frac{|\log\alpha|}{I^\theta}. \tag{16}$$

B. Non-Parametric GLR CuSum Test

For the case when p_1 is unknown, we define the non-parametric GLR statistic as

$$\widehat{Z}_i^{n,k} = \log\frac{\widehat{p}_{-i}^{n,k}(X_i)}{p_0(X_i)}, \ \forall k \leq i \leq n. \tag{17}$$

We remind readers of the definition of $\widehat{p}_{-i}^{n,k}$ from Section II. The non-parametric generalized likelihood ratio (NGLR) CuSum stopping rule is defined as

$$\widehat{\tau}(b) := \inf\left\{n > 1 : \max_{(n-m_b)^+<k\leq n-1}\sum_{i=k}^{n}\widehat{Z}_i^{n,k} \geq b\right\}. \tag{18}$$

Here the window size m_b is designed to satisfy the following assumption.

Assumption 7 (Minimum Window Size): For some arbitrary constant $\eta > 1$, m_b satisfies

$$\liminf m_b/b \geq \frac{\eta}{I}$$

for all large b.

In Lemma 2, we show that $\widehat{\tau}$ with a properly chosen density estimator and threshold $b = b_\alpha$ satisfies the false alarm constraint asymptotically in (10). In Lemma 3, we establish an asymptotic upper bound on $\mathrm{WADD}\left(\widehat{\tau}(b)\right)$. The proofs of the lemmas are given in the Appendix. Finally, in Theorem 1, we combine the lemmas and establish the first-order asymptotic optimality of the NGLR-CuSum test.

In order to satisfy the false alarm constraint, the following assumption is imposed on the density estimator.

Assumption 8: Suppose that $\exists\ \varsigma > 0$, such that

$$\mathbb{E}_\infty\left[\max_{n:k\leq n\leq k+m_b}\prod_{i=k}^{n}\frac{\widehat{p}_{-i}^{n,k}(X_i)}{p_0(X_i)}\right] \leq b^\varsigma$$

for each fixed $k \geq 1$ and for any large enough b.

We will elaborate on Assumption 8 in Section V-A. Intuitively, this is satisfied when the density estimator converges to the true density fast enough ($\mathbb{P}_\infty$-almost surely). Since the pre-change distribution is known, we can numerically verify Assumption 8 under the chosen window size and density estimator.

Lemma 2: Suppose Assumption 8 holds. Let b_α satisfy

$$b_\alpha - \varsigma\log b_\alpha = |\log\alpha| + \log 8. \tag{19}$$

Then,

$$\mathbb{E}_\infty\left[\widehat{\tau}(b_\alpha)\right] \geq \alpha^{-1}(1+o(1)).$$

Lemma 3: Suppose b is large enough such that Assumption 7 holds. Suppose that Assumptions 3 and 4 hold. Further, suppose Assumption 6 holds for the true log-likelihood ratio. Then,

$$\mathrm{WADD}\left(\widehat{\tau}(b)\right) \leq \frac{b}{I}(1+o(1)), \ \text{as } b\to\infty.$$

Theorem 1: Suppose that Assumptions 5 and 6 hold for the true log-likelihood ratio, and suppose that the window size satisfies Assumption 7. Suppose that Assumption 8 is satisfied for the chosen estimator. Let b_α be so selected according to

equation (19) such that $\mathrm{FAR}\left(\widehat{\tau}(b_\alpha)\right) \le \alpha(1+o(1))$, where also $b_\alpha = |\log\alpha|\,(1+o(1))$. Then $\widehat{\tau}(b_\alpha)$ solves the problem in (8) asymptotically as $\alpha \to 0$, and

$$\inf_{\tau\in\mathcal{C}_\alpha} \mathrm{WADD}\left(\tau\right) \sim \mathrm{WADD}\left(\widehat{\tau}\left(b_\alpha\right)\right) \sim \frac{|\log\alpha|}{I}.$$

Proof of Theorem 1: The asymptotic lower bound on the delay follows from Assumption 6 using [5, Thm. 1]. The asymptotic optimality of $\widehat{\tau}(b_\alpha)$ follows from Lemma 2 and Lemma 3. □

IV. QCD With NWLA CuSum Test

Although the NGLR CuSum test is shown to achieve asymptotic optimality, its computational complexity is very high. In order to obtain its statistic at each time n, a max operation needs to be performed over $k = (n-m)^+ + 1, \ldots, n$, and the corresponding LOO estimator needs to be constructed for each candidate k. In this section, we propose another test, the NWLA CuSum test, which still achieves asymptotic optimality and whose statistic has a lower computational complexity.

Define the non-parametric window-limited adaptively-estimated log-likelihood ratio as

$$\widehat{Z}_n^w = \log\frac{\widehat{p}_n^w(X_n)}{p_0(X_n)},\ \forall n > w \tag{20}$$

where $\widehat{p}_n^w$ is the output of the density estimator given input $X^{[n-w,n-1]}$. Note that $\widehat{Z}_n^w$ is independent of $\mathcal{F}_{n-w-1}$. Define the non-parametric window-limited adaptive (NWLA) CuSum statistic as:

$$\overline{W}^w(n) = \left(\overline{W}^w(n-1)\right)^+ + \widehat{Z}_n^w,\quad n > w \tag{21}$$

and $\overline{W}^w(1) = \cdots = \overline{W}^w(w) = 0$. The corresponding stopping rule is

$$\overline{\tau}(b) := \inf\left\{n > w : \overline{W}^w(n) \ge b\right\}. \tag{22}$$

Here $b = b_\alpha > 0$ is a threshold depending on the false alarm rate α. We omit the dependency of $\overline{W}$ on w for brevity.

The following observations regarding Assumption 1 are useful for the analysis in this section. If the estimated density $p = p_1$, the KL-loss bound in Assumption 1 is equivalent to

$$\widehat{I} := \mathbb{E}_1\left[\widehat{Z}_n^w\right] \ge I - \frac{C_1}{w^{\beta_1}},\quad n > w \tag{23}$$

when w is large. This guarantees that $\widehat{I} > 0$ for all sufficiently large w's.

In Lemma 4, we show that $\overline{\tau}$ with a properly chosen threshold b_α satisfies the false alarm constraint in (10). In Lemma 8, we establish an asymptotic upper bound on $\mathrm{WADD}\left(\overline{\tau}(b_\alpha)\right)$. Finally, in Theorem 2, we combine the lemmas and establish the first-order asymptotic optimality of the NWLA-CuSum test. It should be mentioned that the results in this section are similar to those in [7], in which a window-limited adaptive CuSum test is studied for the case where there is parametric uncertainty in the post-change regime. However, the results in [7] are clearly not applicable to the non-parametric setting studied here.

The proofs of Lemmas 4, 5, 7 and 8 are given in the Appendix.

Lemma 4: For any $w > 0$,

$$\mathbb{E}_\infty\left[\overline{\tau}(b)\right] \ge e^b.$$

Thus, $\overline{\tau}(\bar{b}_\alpha) \in \mathcal{C}_\alpha$ if $\bar{b}_\alpha = |\log\alpha|$.

Before introducing the main lemma on the delay, we first introduce three helping lemmas below.

Lemma 5: For any change-point $\nu \ge 1$ and $b > 0$,

$$\operatorname{ess\,sup}\mathbb{E}_\nu\left[(\overline{\tau}(b)-\nu+1)^+|\mathcal{F}_{\nu-1}\right] \le \mathbb{E}_1\left[\overline{\tau}(b)\right].$$

Lemma 6: Define

$$U_n = U_{n-1} + \widehat{Z}_n^w,\quad \forall n > w$$

with $U_1 = \cdots = U_w = 0$. Also define the stopping time

$$\tau_u(b) := \inf\{n > w : U_n \ge b\}. \tag{24}$$

Then, $\tau_u(b) \ge \overline{\tau}(b)$ on $\{\tau_u(b) < \infty\}$ for any $b > 0$.

Proof of Lemma 6: The proof is similar to [7, Lemma 5]. Note that $U_w = 0 = \overline{W}(w)$. For any $k \ge w$, if $U_k \le \overline{W}(k)$, then

$$\begin{aligned} U_{k+1} &= U_k + \widehat{Z}_k^w \le \overline{W}(k) + \widehat{Z}_k^w \\ &\le \left(\overline{W}(k)\right)^+ + \widehat{Z}_k^w = \overline{W}(k+1)\quad a.s. \end{aligned}$$

Thus by induction, $\tau_u(b) \ge \overline{\tau}(b)$ on $\{\tau_u(b) < \infty\}$. □

Lemma 7: If $\widehat{I} > 0$, then the τ_u defined in (24) satisfies $\tau_u < \infty$ almost surely under $\mathbb{P}_1$.

Now, using the lemmas above, we can upper bound the delay of the NWLA-CuSum test.

Lemma 8: Suppose that w is sufficiently large such that $\widehat{I} > 0$. Suppose further that Assumption 2 holds for the density estimator. Then,

$$\begin{aligned} &\mathbb{E}_1\left[\overline{\tau}(b)\right] \le \mathbb{E}_1\left[\tau_u(b)\right] \\ &\le \widehat{I}^{-1}\left(b + w\widehat{I} + I + \sqrt{2}\frac{C_2}{\widehat{I}w^{\beta_2}} + \left(\frac{4C_2}{\widehat{I}w^{\beta_2}}(b+I)\right)^{\frac{1}{2}}\right) \end{aligned} \tag{25}$$

where $\tau_u(b)$ is defined in (24).

Theorem 2: Suppose that $\bar{b}_\alpha = |\log\alpha|$ and the window size $w_\alpha = |\log\alpha|^\kappa$ for some $0 < \kappa < 1$. Then, under Assumptions 1 and 2, $\overline{\tau}(\bar{b}_\alpha)$ solves the problem in (8) asymptotically as $\alpha \to 0$, and the delay is upper-bounded as

$$\mathrm{WADD}\left(\overline{\tau}\left(\bar{b}_\alpha\right)\right) \le \frac{|\log\alpha|}{I}\left(1 + \Theta(|\log\alpha|^{-\rho_\kappa})\right),$$

where

$$\rho_\kappa = \min\left\{\kappa\beta_1, 1-\kappa\right\}. \tag{26}$$

Proof of Theorem 2: From Assumption 1 it follows that,

$$\begin{aligned} \frac{I}{\widehat{I}} = 1 + \frac{I-\widehat{I}}{\widehat{I}} &\le 1 + \frac{C_1}{\widehat{I}w^{\beta_1}} \\ &\le 1 + \frac{C_1}{\left(I - \frac{C_1}{w^{\beta_1}}\right)w^{\beta_1}} \\ &= 1 + \frac{C_1}{Iw^{\beta_1} - C_1}. \end{aligned}$$

Given the selected $w = w_\alpha$, $\widehat{I} > 0$ for a sufficiently small α. Define $r_\alpha := \frac{|\log\alpha|}{I}$. From Lemma 8 (in particular, (25)),

if we select $\bar{b}_\alpha = |\log \alpha|$ and $w = |\log \alpha|^\kappa$, the scaled average delay (when $\nu = 1$) can be upper bounded as:

$$\begin{aligned} & r_\alpha^{-1}\mathbb{E}_1\left[\overline{\tau}(\bar{b}_\alpha)\right] \le r_\alpha^{-1}\mathbb{E}_1\left[\tau_u(\bar{b}_\alpha)\right] \\ & \le \frac{I}{\widehat{I}}\Bigg(1 + \frac{\widehat{I}}{|\log\alpha|^{1-\kappa}} + \frac{I}{|\log\alpha|} + \sqrt{2}\frac{C_2}{\widehat{I}\,|\log\alpha|^{1+\kappa\beta_2}} \\ & \quad + \frac{1}{|\log\alpha|}\left(\frac{4C_2}{\widehat{I}\,|\log\alpha|^{\kappa\beta_2}}(|\log\alpha| + I)\right)^{\frac{1}{2}}\Bigg) \\ & \le 1 + \frac{C_1}{I\,|\log\alpha|^{\kappa\beta_1} - C_1} + \frac{I}{|\log\alpha|^{1-\kappa}} + o\left(\frac{1}{|\log\alpha|^{1-\kappa}}\right) \\ & = 1 + \Theta\left(\frac{1}{|\log\alpha|^{\rho_\kappa}}\right). \end{aligned}$$

Together with Lemmas 4–7, the result on the asymptotic delay at $\nu = 1$ establishes the asymptotic optimality. □

Remark 1: Since $\beta_1 < 1$, we have $\rho_\kappa < \frac{1}{2}$ if $\kappa \in (0,1)$. Also, from (26), to maximize ρ_κ, one can choose

$$\rho^* = \max_{\kappa\in(0,1)} \rho_\kappa, \quad \kappa^* = \arg\max_{\kappa\in(0,1)} \rho_\kappa.$$

Example 2: Let the dominating measure μ be the Lebesgue measure on $\mathbb{R}^d$. Recall the definition of $\mathcal{H}_\gamma$ in Example 1. Consider $p_0, p_1 \in \mathcal{H}_\gamma$ with bounded support and non-zero density. Consider using KDE as the estimator with some kernel that satisfies (6). Previously we showed in (7) and Lemma 1 that with a properly chosen $\boldsymbol{h}$, the optimal β_1 and β_2 are $\beta_1 = \beta_2 = \frac{2\gamma}{2\gamma+d}$. Therefore, in this case κ^* and ρ^* are

$$\kappa^* = \arg\max_{\kappa\in(0,1)} \min\left\{\frac{2\gamma}{2\gamma+d}\kappa, 1-\kappa\right\} = \frac{2\gamma+d}{4\gamma+d},$$

and

$$\rho^* = \frac{2\gamma}{4\gamma+d}.$$

V. Numerical Results

In this section, we present some numerical results for the proposed tests. Before presenting the main results, we first investigate from a numerical perspective the feasibility of Assumption 8, which is key to Lemma 2.

A. Discussion of Assumption 8

The quantity of interest is

$$Q(m) := \mathbb{E}_\infty\left[\max_{n=2,\dots,m}\prod_{i=1}^{n}\frac{\widehat{p}_{-i}^{n,1}(X_i)}{p_0(X_i)}\right]. \tag{27}$$

In Fig. 1, we study the numerical properties of $Q(m)$ for KDE with Gaussian kernels. The difference $\log(Q(m)) - 3\log m$ is plotted against m for each $m = 5, 10, \dots, 100$. In the plot, it is observed that

$$\log(Q(m)) - 3\log m < 0 \implies Q(m) \le m^3.$$

Therefore, with the chosen density estimator, if further the window size is chosen such that $m_b \le b^\vartheta(1 + o(1))$ with some $\vartheta > 1$, then Assumption 8 is satisfied with $\varsigma = 3\vartheta$.

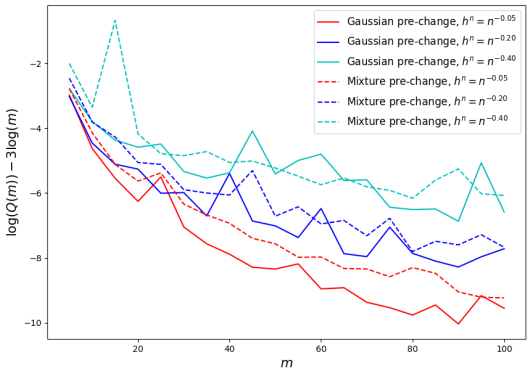

Fig. 1. $\log(Q(m)) - 3\log m$ versus m with $Q(m)$ defined in (27). For each value of m, 100000 Monte Carlo runs are performed for each $n = 2, \dots, m$, and the maximum product of likelihood ratios is averaged. The selected pre-change distributions are $\mathcal{N}(0,1)$ (solid lines) and $\frac{1}{3}(\mathcal{N}(-2,\frac{1}{4}) + \mathcal{N}(0,\frac{1}{4}) + \mathcal{N}(2,1))$ (dashed lines), with $\mathcal{N}(\mu,\sigma^2)$ denote a Gaussian with mean μ and σ^2. The KDE with Gaussian kernel is used for density estimation, with bandwidth $h = n^{-r}$ where r chosen as 0.05 (red), 0.2 (blue), and 0.4 (cyan). In both plots, it is observed that the difference trends lower as m increases, and that all simulated values are below zero.

B. Performance of NGLR-CuSum Test

In Fig. 2, we study the performance of the proposed NGLR-CuSum test (defined in (18)) through Monte Carlo (MC) simulations when the pre-change distribution is $\mathcal{N}(0,1)$. The KDE (defined in (5) with $d = 1$) with a Gaussian kernel is used to estimate the density. The actual post-change distribution is $\mathcal{N}(0.5,1)$, but this knowledge is not used in the NGLR-CuSum test. The performance of the NGLR-CuSum test is compared against that of the following tests:

1) the CuSum test (in (11)), which has full knowledge of the post-change distribution;
2) the parametric window-limited GLR-CuSum test (in (15)), in which it is assumed that the post-change distribution belongs to $\{\mathcal{N}(\theta,1)\}_{\theta\neq 0}$.

The change-point is taken to be $\nu = 1$.[1] Different window sizes are considered, among which the window size of 100 is sufficiently large to cover the full range of delay. It is seen that the expected delay of the NGLR-CuSum test is close to that of the GLR-CuSum test for all window sizes considered.

C. Performance of NWLA-CuSum Test

In Fig. 3, we study the performance of the proposed NWLA-CuSum test (defined in (22)) through Monte Carlo (MC) simulations when the pre-change distribution is $\mathcal{N}(0,1)$. The KDE (defined in (5) with $d = 1$) is used to estimate the density. The actual post-change distribution is $\mathcal{N}(0.5,1)$. This knowledge is not used in the NWLA-CuSum test. The performance of the NWLA-CuSum test is compared against that of the following tests:

1) the CuSum test (in (11)), which assumes full knowledge of the post-change distribution;

[1]Note that $\nu = 1$ may not necessarily be the worst-case value for the change-point for the NGLR-CuSum test in general. However, extensive experimentation on this Gaussian mean-change problem with different values of ν ranging from 1 to 100, with a window-size of 100, shows that $\nu = 1$ results in the largest expected delay among all ν's considered.

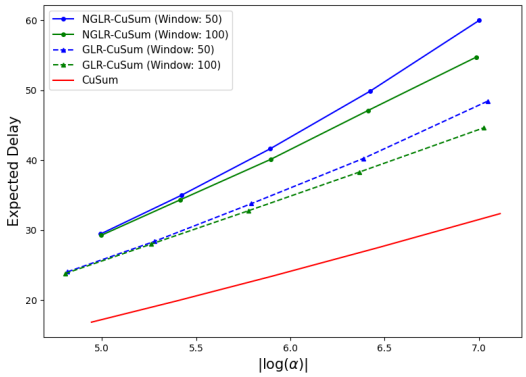

Fig. 2. Comparison of operating characteristics of the NGLR-CuSum test (solid lines) with the CuSum test (in red) and the parametric window-limited GLR-CuSum test (dashed lines) in detecting a shift in the mean of a Gaussian. The pre- and post-change distributions are $\mathcal{N}(0,1)$ and $\mathcal{N}(0.5,1)$. The change-point $\nu = 1$. The kernel width parameter $h = 10^{-1/5}$.

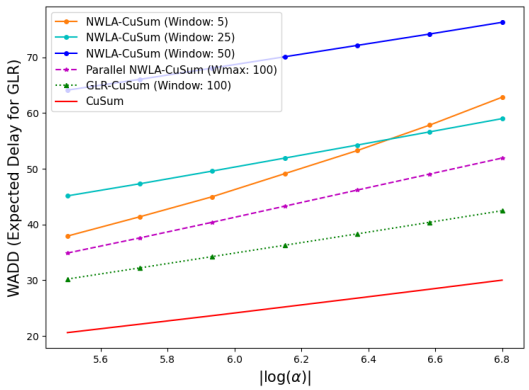

Fig. 3. Comparison of operating characteristics of the NWLA-CuSum test (solid lines) and the parallel-NWLA-CuSum test (dashed line) with the CuSum test (in red) and the parametric window-limited GLR-CuSum test (dotted line) in detecting a shift in the mean of a Gaussian. The pre- and post-change distributions are $\mathcal{N}(0,1)$ and $\mathcal{N}(0.5,1)$. The change-point $\nu = 1$. The kernel width parameter $h = w^{-1/5}$, where w is the window size.

2) the parametric window-limited GLR-CuSum test (in (15)), in which it is assumed that the post-change distribution belongs to $\{\mathcal{N}(\theta,1)\}_{\theta\neq 0}$.;
3) the parallel-NWLA-CuSum test, defined as

$$\overline{\tau}_{\text{parallel}}(b, W_{\max}) := \inf\left\{n > 1 : \max_{1\le w\le W_{\max}} \overline{W}^{w}(n) \ge b\right\}.$$

Using a similar analysis as in Section IV, it can be shown that the parallel-NWLA-CuSum test is also asymptotically optimal with the threshold chosen as $b_\alpha = |\log\alpha| + \log W_{\max}$. The change-point is taken to be $\nu = 1$, which corresponds to the worst-case expected delay for the NWLA-CuSum test (shown in Lemma 5), the parallel NWLA-CuSum test, and the CuSum test, but not necessarily for the parametric window-limited GLR-CuSum test. Different window sizes are also considered. We note that there is a trade-off to consider in the design of the window size for the NWLA-CuSum test. If the window size is too small, the post-change density might not be accurately estimated. On the other hand, if the window size is too large, the test might wait too long before its statistic starts to grow in the post-change regime. To address this trade-off, the parallel-NWLA-CuSum test could be employed without specifying a pre-defined window size, albeit at the expense of having to run more tests in parallel.

D. Comparison Between NGLR-CuSum Test and NWLA-CuSum Test

We now compare the performance between the NGLR-CuSum test and the parallel-NWLA-CuSum test. First, we compare the computational complexity of both tests if the KDE is used for density estimation. For the NGLR-CuSum test, at each input observation X_n, for all hypothesized change-points $k \in ((n-m)^+, n-1]$ (where m is the window size), the pair-wise kernel value $K(X_i, X_j)$ for each pair of $i, j \in [k, n]$ with $i \neq j$ is calculated and a LOO kernel estimate is evaluated at each point $X_k, \ldots, X_n$. Then, the LOO-estimated log-likelihood ratios at these points are summed up and the maximum sum (over k) is compared to the given threshold. Thus, the computation complexity of the NGLR-CuSum test at each time is $\Theta((n \wedge m)^3)$. For the parallel-NWLA-CuSum test, at each given X_n, $K(X_i, X_n)$ is first calculated for each $i \in [(n - W_{\max}) \vee 1, n-1]$. Then, with each possible window size $w = 1, \ldots, W_{\max} \wedge (n-1)$, a WLA kernel estimate is evaluated to update the corresponding NWLA-CuSum statistic in an efficient manner. Finally, the maximum NWLA-CuSum statistic (over w) is compared to the given threshold. Therefore, the computation complexity of the parallel-NWLA-CuSum test is $\Theta((n \wedge W_{\max})^2)$. Also note that in practice, $W_{\max}$ is usually chosen to be smaller than m, by comparing the requirement of m_b in Assumption 7 with the window condition in Thm 2.

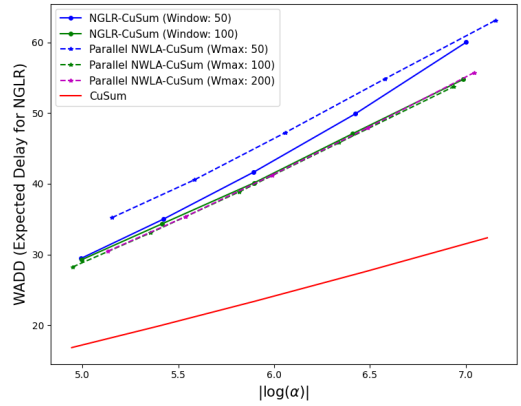

Fig. 4. Comparison of operating characteristics of the NGLR-CuSum test (solid lines) and the parallel-NWLA-CuSum test (dashed lines) with the CuSum test (in red) in detecting a shift in the mean of a Gaussian. The pre- and post-change distributions are $\mathcal{N}(0,1)$ and $\mathcal{N}(0.5,1)$. The change-point $\nu = 1$. The kernel width parameters are $h = 10^{-1/5}$ for the NGLR-CuSum test, and $h = w^{-1/5}$ for the parallel-NWLA-CuSum test, where w is the window size.

In Fig. 4, we compare the numerical performance of the NGLR-CuSum test and the parallel-NWLA-CuSum test in detecting a shift in the mean of a Gaussian. When the window size is large enough, both tests achieve similar performance at $\nu = 1$, which corresponds to the worst-case ν for the parallel-NWLA-CuSum test, but not necessarily for the NGLR-CuSum test.

VI. Conclusion

We studied a window-limited non-parametric generalized likelihood ratio (NGLR) CuSum test and a non-parametric window-limited adaptive (NWLA) CuSum test for QCD. Both tests do not assume any explicit knowledge of the post-change distribution, and do not require post-change training samples ahead of time. We characterized a generic class of density estimators that enable detection. For both tests, we provided a way to set the test thresholds to meet false alarm constraints, and we showed that the tests are first-order asymptotically optimal with the selected thresholds, as the false alarm rate goes to zero. We validated our analysis through Monte-Carlo simulations, in which we compared both tests with baseline tests that have distributional knowledge.

Appendix

Proof of Lemma 1: For brevity we write $\widehat{p}(X) = \widehat{p}_n^w(X_n)$, and note that X is independent of $\widehat{p}$. We use the fact that $\log s \le (s-1)$ to establish an upper bound on the first moment. In particular,

$$\begin{aligned}
&\mathrm{E}_p\left[\log\frac{p(X)}{\widehat{p}(X)}\right]\\
&\le \mathrm{E}_p\left[\frac{p(X)}{\widehat{p}(X)}-1\right]\\
&= \mathrm{E}_p\left[\int\frac{p^2(x)-p(x)\widehat{p}(x)}{\widehat{p}(x)}d\mu(x)\right]\\
&\overset{(*)}{=} \mathrm{E}_p\left[\int\frac{p^2(x)-2\,p(x)\widehat{p}(x)+\widehat{p}^2(x)}{\widehat{p}(x)}d\mu(x)\right]\\
&\le \frac{1}{\underline{\zeta}}\mathrm{MISE}(p,\widehat{p}) \qquad (28)
\end{aligned}$$

where $(*)$ follows by the independence between $\widehat{p}$ and X and because both p and $\widehat{p}$ are densities. This establishes Assumption 1 with $\beta_1 = \beta_3$. The proof for Assumption 3 is similar, noting the independence between $\widehat{p}_{-i}^{n,k}$ and X_i. For the second moment, note that $(\log s)^2 \le r(s-1)^2$ on $s \ge \underline{\zeta}/\overline{\zeta}$ with r as defined in (4). Thus,

$$\begin{aligned}
\mathrm{E}_p\left[\left(\log\frac{p(X)}{\widehat{p}(X)}\right)^2\right] &\le \mathrm{E}_p\left[r\left(\frac{p(X)}{\widehat{p}(X)}-1\right)^2\right]\\
&= r\mathrm{E}_p\left[\int\frac{(p(x)-\widehat{p}(x))^2}{\widehat{p}^2(x)}p(x)d\mu(x)\right]\\
&\le \frac{\overline{\zeta}r}{\underline{\zeta}^2}\mathrm{MISE}(p,\widehat{p}) \qquad (29)
\end{aligned}$$

which shows Assumption 2 with $\beta_2 = \beta_3$. Furthermore, for Assumption 4,

$$\begin{aligned}
&\mathrm{E}_p\left[\left(\frac{1}{n-k+1}\sum_{i=k}^n\log\frac{p(X_i)}{\widehat{p}_{-i}^{n,k}(X_i)}\right)^2\right]\\
&\overset{(a)}{\le} \mathrm{E}_p\left[\frac{1}{n-k+1}\sum_{i=k}^n\left(\log\frac{p(X_i)}{\widehat{p}_{-i}^{n,k}(X_i)}\right)^2\right]\\
&\overset{(b)}{=} \mathrm{E}_p\left[\left(\log\frac{p(X_n)}{\widehat{p}_{-n}^{n,k}(X_n)}\right)^2\right]\\
&\le \frac{\overline{\zeta}rC_3}{\underline{\zeta}^2(n-k+1)^{\beta_3}}. \qquad (30)
\end{aligned}$$

Here (a) follows by Jensen's inequality, and (b) follows because $\log\frac{p(X_i)}{\widehat{p}_{-i}^{n,k}(X_i)}$ has the same distribution for all $i \in [k,n]$. The proof is now complete. □

Proof of Corollary 1: Following the argument in (28), we have

$$\mathrm{E}_p\left[\log\frac{p(X)}{\widehat{p}(X)}\right] = O\left(\frac{1}{\underline{\zeta}}\mathrm{MISE}(p,\widehat{p})\right) = O(w^{-(\beta_3-\underline{\beta})}),$$

and the first moment results (i.e., that of β_1) follow immediately for Assumptions 1 and 3.

Now we turn to the second moment. From the definition of r,

$$r = \left(\frac{\log(\overline{\zeta}/\underline{\zeta})}{1-(\underline{\zeta}/\overline{\zeta})}\right)^2 \le (\log(\overline{\zeta}/\underline{\zeta}))^2 = (\overline{\beta}-\underline{\beta})^2(\log w)^2$$

Therefore, following the argument in (29), we get

$$\begin{aligned}
\mathrm{E}_p\left[\left(\log\frac{p(X)}{\widehat{p}(X)}\right)^2\right] &= O\left(\frac{\overline{\zeta}r}{\underline{\zeta}^2}\mathrm{MISE}(p,\widehat{p})\right)\\
&= O\left(\frac{w^{\overline{\beta}}}{w^{-2\underline{\beta}}w^{\beta_3}}(\log w)^2\right)\\
&= O\left(w^{-(\beta_3-2\underline{\beta}-\overline{\beta}-\varrho)}\right)
\end{aligned}$$

where $\varrho > 0$ is an arbitrarily small constant. This shows the second moment result for Assumption 2. The result for Assumption 4 is similar following the argument in (30). □

Proof of Lemma 2: Fix $\ell > 1$. For all thresholds $b > 0$,

$$\begin{aligned}
&\mathbb{P}_\infty\{\ell \le \widehat{\tau}(b) < \ell + m_b\}\\
&\overset{(i)}{\le} \mathbb{P}_\infty\Big\{\exists(k,n) \text{ with } \ell \le n < \ell+m_b,\\
&\qquad (n-m_b)^+ < k \le n-1 : \sum_{i=k}^n \widehat{Z}_i^{n,k} \ge b\Big\}\\
&\overset{(ii)}{\le} \mathbb{P}_\infty\Big\{\exists(k,n) \text{ with } (\ell-m_b)^+ < k < \ell+m_b,\\
&\qquad k+1 \le n \le k+m_b : \sum_{i=k}^n \widehat{Z}_i^{n,k} \ge b\Big\}\\
&= \mathbb{P}_\infty\Bigg\{\bigcup_{k=(\ell-m_b)^++1}^{\ell+m_b-1}\\
&\qquad \Big\{\exists n \text{ with } k+1 \le n \le k+m_b : \sum_{i=k}^n \widehat{Z}_i^{n,k} \ge b\Big\}\Bigg\}\\
&\le \sum_{k=(\ell-m_b)^++1}^{\ell+m_b-1}\\
&\qquad \mathbb{P}_\infty\Big\{\exists n \text{ with } k+1 \le n \le k+m_b : \sum_{i=k}^n \widehat{Z}_i^{n,k} \ge b\Big\}
\end{aligned}$$

$$\leq \sum_{k=(\ell-m_b)^+ +1}^{\ell+m_b-1} \mathbb{P}_\infty\{\tau_k(b) \leq k+m_b\} \tag{31}$$

where (i) follows from the definition of $\widehat{\tau}(b)$, and (ii) follows because

$$\ell \leq n < \ell + m_b, \quad (n-m_b)^+ < k \leq n-1$$

implies that

$$(\ell - m_b)^+ < k < \ell + m_b, \quad k+1 \leq n \leq k+m_b.$$

Here we define the auxiliary stopping time $\tau_k(b)$ for $k \geq 1$ as

$$\tau_k(b) := \inf\left\{ n \in [k+1, k+m_b] : \sum_{i=k}^{n} \widehat{Z}_i^{n,k} \geq b \right\} \tag{32}$$

and we define $\inf \emptyset := \infty$. Now, for each $k \in [(\ell-m_b)^+, \ell+m_b)$, we have

$$\begin{aligned}
&\mathbb{P}_\infty\{k+1 \leq \tau_k(b) \leq k+m_b\} \\
&= \int \mathbb{1}\{k+1 \leq \tau_k(b) \leq k+m_b\} d\mathbb{P}_\infty \\
&= \int \mathbb{1}\{k+1 \leq \tau_k(b) \leq k+m_b\} \times \\
&\quad \prod_{i=k}^{\tau_k(b)} \frac{\widehat{p}_{-i}^{\tau_k(b),k}(x_i)}{p_0(x_i)} \prod_{i=k}^{\tau_k(b)} \frac{p_0(x_i)}{\widehat{p}_{-i}^{\tau_k(b),k}(x_i)} d\mathbb{P}_\infty \\
&\overset{(iii)}{\leq} e^{-b} \int \mathbb{1}\{k+1 \leq \tau_k(b) \leq k+m_b\} \times \\
&\quad \prod_{i=k}^{\tau_k(b)} \frac{\widehat{p}_{-i}^{\tau_k(b),k}(x_i)}{p_0(x_i)} d\mathbb{P}_\infty
\end{aligned} \tag{33}$$

where (iii) follows from the definition of $\tau_k(b)$.

Now,

$$\begin{aligned}
&\int \mathbb{1}\{k+1 \leq \tau_k(b) \leq k+m_b\} \prod_{i=k}^{\tau_k(b)} \frac{\widehat{p}_{-i}^{\tau_k(b),k}(x_i)}{p_0(x_i)} d\mathbb{P}_\infty \\
&\leq \int \max_{n \in [k,k+m_b]} \prod_{i=k}^{n} \frac{\widehat{p}_{-i}^{n,k}(x_i)}{p_0(x_i)} d\mathbb{P}_\infty \\
&\overset{(iv)}{\leq} b^\varsigma (1+o(1))
\end{aligned}$$

where (iv) follows from Assumption 8. Combining with (31) and (33), we have

$$\sup_{\ell > 1} \mathbb{P}_\infty\{\ell \leq \widehat{\tau}(b) < \ell + m_b\} \leq 2\, m_b e^{-b} b^\varsigma (1+o(1)),$$

and by [25, Lemma 2.2(ii)],

$$\mathbb{E}_\infty[\widehat{\tau}(b)] \geq \frac{1}{8} e^b b^{-\varsigma}(1+o(1)).$$

Choosing $b = b_\alpha$ then satisfies the false alarm constraint asymptotically. □

Proof of Lemma 3: Recall that $I = D(p_1||p_0)$. Under Assumption 7, define a function δ_b such that $\delta_0 := 1 - \eta^{-1} < 1$, that $\delta_b \in (0, \delta_0)$ is decreasing in b, and that $\delta_b \searrow 0$ as $b \to \infty$. Define

$$n_b := \left\lfloor \frac{b}{I(1-\delta_b)} \right\rfloor \tag{34}$$

and thus

$$n_b < \frac{b}{I(1-\delta_0)} = \frac{\eta b}{I} \leq m_b$$

when b is large enough. If for now that we can get a large enough b to satisfy

$$\mathbb{P}_\nu\left\{ \sum_{i=n}^{n+n_b-1} \widehat{Z}_i^{n+n_b-1,n} < b \right\} < 2\delta_b^2, \ \forall (\nu, n) : n \geq \nu \geq 1. \tag{35}$$

Then in the following, we will show by induction that

$$\begin{aligned}
&\operatorname{ess\,sup} \mathbb{P}_\nu\Big\{ \widehat{\tau}(b) - \nu + 1 > kn_b \\
&\qquad \Big| \widehat{\tau}(b) - \nu + 1 > (k-\ell)n_b, \mathcal{F}_{\nu-1} \Big\} \leq (2\delta_b^2)^\ell
\end{aligned} \tag{36}$$

for all $\nu \geq 1$ and $k \geq \ell$ when b is large enough.

We will induct on the variable ℓ. The base case is where $\ell = 1$, and we get, $\forall k \geq 1$,

$$\begin{aligned}
&\operatorname{ess\,sup} \mathbb{P}_\nu\{\widehat{\tau}(b) - \nu + 1 > kn_b | \widehat{\tau}(b) - \nu + 1 > (k-1)n_b, \mathcal{F}_{\nu-1}\} \\
&\overset{(i)}{\leq} \operatorname{ess\,sup} \mathbb{P}_\nu\left\{ \widehat{\tau}(b) - \nu + 1 > kn_b \middle| \mathcal{F}_{\nu+(k-1)n_b-1} \right\} \\
&\overset{(ii)}{\leq} \operatorname{ess\,sup} \mathbb{P}_\nu\Bigg\{ \sum_{i=\nu+(k-1)n_b}^{\nu+kn_b-1} \widehat{Z}_i^{\nu+kn_b-1,\nu+(k-1)n_b} < b \\
&\qquad \Bigg| \mathcal{F}_{\nu+(k-1)n_b-1} \Bigg\} \\
&\overset{(iii)}{=} \mathbb{P}_\nu\left\{ \sum_{i=\nu+(k-1)n_b}^{\nu+kn_b-1} \widehat{Z}_i^{\nu+kn_b-1,\nu+(k-1)n_b} < b \right\} \\
&\overset{(iv)}{\leq} 2\delta_b^2.
\end{aligned} \tag{37}$$

In the series of inequalities above, (i) is by definition of essential supremum and $\widehat{\tau}(b)$, (iii) follows from independence between the event $\left\{ \sum_{i=\nu+(k-1)n_b}^{\nu+kn_b-1} \widehat{Z}_i^{\nu+kn_b-1,\nu+(k-1)n_b} < b \right\}$ and $\mathcal{F}_{\nu+(k-1)n_b-1}$, and (iv) follows from (35). The reason for (ii) is as follows. The event $\{\widehat{\tau}(b) - \nu + 1 > kn_b\}$ implies that no change has been detected until time $n = kn_b + \nu - 1$. In particular, this means that at time $n = kn_b + \nu - 1$,

$$\max_{(\nu+kn_b-1-m_b)^+ < \kappa \leq \nu+kn_b-2} \sum_{i=\kappa}^{\nu+kn_b-1} \widehat{Z}_i^{\nu+kn_b-1,\kappa} < b.$$

Now, since $n_b \leq m_b$,

$$\begin{aligned}
&\sum_{i=\nu+(k-1)n_b}^{\nu+kn_b-1} \widehat{Z}_i^{\nu+kn_b-1,\nu+(k-1)n_b} \\
&\leq \max_{(\nu+kn_b-1-m_b)^+ < \kappa \leq \nu+kn_b-2} \sum_{i=\kappa}^{\nu+kn_b-1} \widehat{Z}_i^{\nu+kn_b-1,\kappa} \\
&< b.
\end{aligned}$$

The induction base is thus established.

We now turn to the induction step. Suppose we have proved that

$$\operatorname{ess\,sup}\mathbb{P}_\nu\Big\{\widehat{\tau}(b)-\nu+1>kn_b \;\Big|\; \widehat{\tau}(b)-\nu+1>(k-\ell)n_b,\mathcal{F}_{\nu-1}\Big\}\le(2\delta_b^2)^\ell,\ \forall k\ge l.$$

Then, for $k\ge\ell+1$,

$$\begin{aligned}
&\mathbb{P}_\nu\Big\{\widehat{\tau}(b)-\nu+1>kn_b \;\Big|\; \widehat{\tau}(b)-\nu+1>(k-\ell-1)n_b,\mathcal{F}_{\nu-1}\Big\}\\
&\overset{(v)}{=}\mathbb{P}_\nu\Big\{\widehat{\tau}(b)-\nu+1>kn_b,\widehat{\tau}(b)-\nu+1>(k-1)n_b \;\Big|\; \widehat{\tau}(b)-\nu+1>(k-\ell-1)n_b,\mathcal{F}_{\nu-1}\Big\}\\
&=\mathbb{P}_\nu\Big\{\widehat{\tau}(b)-\nu+1>(k-1)n_b \;\Big|\; \widehat{\tau}(b)-\nu+1>(k-\ell-1)n_b,\mathcal{F}_{\nu-1}\Big\}\times\\
&\quad\mathbb{P}_\nu\Big\{\widehat{\tau}(b)-\nu+1>kn_b \;\Big|\; \widehat{\tau}(b)-\nu+1>(k-1)n_b,\\
&\qquad\widehat{\tau}(b)-\nu+1>(k-\ell-1)n_b,\mathcal{F}_{\nu-1}\Big\}
\end{aligned}$$

where (v) holds because $\{\widehat{\tau}(b)-\nu+1>kn_b\}\subseteq\{\widehat{\tau}(b)-\nu+1>(k-1)n_b\}$. Thus,

$$\begin{aligned}
&\operatorname{ess\,sup}\mathbb{P}_\nu\Big\{\widehat{\tau}(b)-\nu+1>kn_b \;\Big|\; \widehat{\tau}(b)-\nu+1>(k-\ell-1)n_b,\mathcal{F}_{\nu-1}\Big\}\\
&\le\operatorname{ess\,sup}\mathbb{P}_\nu\Big\{\widehat{\tau}(b)-\nu+1>(k-1)n_b \;\Big|\; \widehat{\tau}(b)-\nu+1>(k-\ell-1)n_b,\mathcal{F}_{\nu-1}\Big\}\times\\
&\quad\operatorname{ess\,sup}\mathbb{P}_\nu\Big\{\widehat{\tau}(b)-\nu+1>kn_b \;\Big|\; \widehat{\tau}(b)-\nu+1>(k-1)n_b,\widehat{\tau}(b)-\nu+1>(k-\ell-1)n_b,\mathcal{F}_{\nu-1}\Big\}\\
&\overset{(vi)}{\le}\operatorname{ess\,sup}\mathbb{P}_\nu\Big\{\widehat{\tau}(b)-\nu+1>(k-1)n_b \;\Big|\; \widehat{\tau}(b)-\nu+1>(k-\ell-1)n_b,\mathcal{F}_{\nu-1}\Big\}\times\\
&\quad\operatorname{ess\,sup}\mathbb{P}_\nu\Big\{\widehat{\tau}(b)-\nu+1>kn_b|\mathcal{F}_{\nu+(k-1)n_b-1}\Big\}\\
&\overset{(vii)}{\le}\operatorname{ess\,sup}\mathbb{P}_\nu\Big\{\widehat{\tau}(b)-\nu+1>(k-1)n_b \;\Big|\; \widehat{\tau}(b)-\nu+1>(k-\ell-1)n_b,\mathcal{F}_{\nu-1}\Big\}\times(2\delta_b^2)\\
&\le(2\delta_b^2)^{\ell+1}
\end{aligned}$$

where (vi) follows by definition of essential supremum and the fact that

$$\Big\{\widehat{\tau}(b)-\nu+1>(k-1)n_b,\ \widehat{\tau}(b)-\nu+1>(k-\ell-1)n_b,\mathcal{F}_{\nu-1}\Big\}\subset\{\mathcal{F}_{\nu+(k-1)n_b-1}\}$$

and (vii) follows by (37). Therefore, by induction, we get (36). In particular, letting $\ell=k$, we get

$$\operatorname{ess\,sup}\mathbb{P}_\nu\{\widehat{\tau}(b)-\nu+1>kn_b|\mathcal{F}_{\nu-1}\}\le(2\delta_b^2)^k,\ \forall\nu\ge1$$

for all sufficiently large b's.

Therefore, for all sufficiently large b's,

$$\begin{aligned}
&\sup_{\nu\ge1}\operatorname{ess\,sup}\mathbb{E}_\nu\left[n_b^{-1}(\widehat{\tau}(b)-\nu+1)^+|\mathcal{F}_{\nu-1}\right]\\
&\le\sum_{k=1}^{\infty}\operatorname{ess\,sup}\mathbb{P}_\nu\{\widehat{\tau}(b)-\nu+1>kn_b|\mathcal{F}_{\nu-1}\}\\
&\le\sum_{k=0}^{\infty}(2\delta_b^2)^k=\frac{1}{1-2\delta_b^2}.
\end{aligned}$$

Recall the definition of WADD in (9). As $b\to\infty$, this implies that

$$\begin{aligned}
\operatorname{WADD}(\widehat{\tau}(b))&\le\frac{n_b}{1-2\delta_b^2}\le\frac{b}{I(1-\delta_b)(1-2\delta_b^2)}\\
&=\frac{b}{I}(1+o(1)).
\end{aligned}$$

It remains to show (35). Write

$$Z_i:=\log\frac{p_1(X_i)}{p_0(X_i)}.$$

For any $n\ge\nu\ge1$ and $\epsilon>0$,

$$\begin{aligned}
&\mathbb{P}_\nu\left\{\sum_{i=n}^{n+n_b-1}\widehat{Z}_i^{n+n_b-1,n}<b\right\}\\
&=\mathbb{P}_\nu\left\{\sum_{i=n}^{n+n_b-1}\widehat{Z}_i^{n+n_b-1,n}<b,\ \sum_{i=n}^{n+n_b-1}Z_i-\widehat{Z}_i^{n+n_b-1,n}\le\epsilon\right\}+\\
&\quad\mathbb{P}_\nu\left\{\sum_{i=n}^{n+n_b-1}\widehat{Z}_i^{n+n_b-1,t}<b,\ \sum_{i=n}^{n+n_b-1}Z_i-\widehat{Z}_i^{n+n_b-1,t}\ge\epsilon\right\}\\
&\le\mathbb{P}_\nu\left\{\sum_{i=n}^{n+n_b-1}Z_i\le b+\epsilon\right\}+\\
&\quad\mathbb{P}_\nu\left\{\frac{1}{n_b}\sum_{i=n}^{n+n_b-1}\left(Z_i-\widehat{Z}_i^{n+n_b-1,t}\right)\ge\frac{\epsilon}{n_b}\right\}\\
&=\mathbb{P}_1\left\{\sum_{i=1}^{n_b}Z_i\le b+\epsilon\right\}+\\
&\quad\mathbb{P}_1\left\{\frac{1}{n_b}\sum_{i=1}^{n_b}\left(Z_i-\widehat{Z}_i^{n_b,1}\right)\ge\frac{\epsilon}{n_b}\right\}
\end{aligned}\tag{38}$$

where Z_i is the true log-likelihood ratio at time i. Observe that the first term increases with ϵ, while the second term decreases. It is important to choose a proper $\epsilon = \epsilon_b$ in order to keep both terms small. The idea in the following is that we first choose a proper $\epsilon = \epsilon_b$ by controlling the second term, and then verify that it is small enough for the first term when b becomes large.

Below, the goal is to choose $\epsilon = \epsilon_b$ and δ_b such that

$$\mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < b + \epsilon_b\right\} \le \delta_b^2$$

and

$$\mathbb{P}_1\left\{\frac{1}{n_b}\sum_{i=1}^{n_b}\left(Z_i - \widehat{Z}_i^{n_b,1}\right) \ge \frac{\epsilon_b}{n_b}\right\} \le \delta_b^2$$

hold simultaneously. In the following, write $\widehat{p}_i$ and $\widehat{Z}_i$ as short-hand notations for $\widehat{p}_{-i}^{n_b,1}$ and $\widehat{Z}_{-i}^{n_b,1}$, respectively. Note that $\mathbb{E}_1\left[Z_i - \widehat{Z}_i\right] = \mathbb{E}_1\left[D(p_1||\widehat{p}_i)\right]$. Under the conditions for the estimator in Assumptions 3 and 4, the mean and variance of $n_b^{-1}\sum_{i=1}^{n_b}(Z_i - \widehat{Z}_i)$ can be bounded as

$$\begin{aligned}\mathbb{E}_1\left[\frac{1}{n_b}\sum_{i=1}^{n_b}(Z_i - \widehat{Z}_i)\right] &= \mathbb{E}_1\left[\frac{1}{n_b}\sum_{i=1}^{n_b}\log\frac{p_1(X_i)}{\widehat{p}_i(X_i)}\right] \le \frac{C_1}{n_b^{\beta_1}}\\ \mathrm{Var}_1\left(\frac{1}{n_b}\sum_{i=1}^{n_b}(Z_i - \widehat{Z}_i)\right) &= \mathrm{Var}_1\left(\frac{1}{n_b}\sum_{i=1}^{n_b}\log\frac{p_1(X_i)}{\widehat{p}_i(X_i)}\right)\\ &\le \frac{C_2}{n_b^{\beta_2}}. \end{aligned} \tag{39}$$

Now, for any $\epsilon_b > n_b \times \mathbb{E}_1\left[D(p_1||\widehat{p}_i)\right]$,

$$\begin{aligned}&\mathbb{P}_1\left\{\frac{1}{n_b}\sum_{i=1}^{n_b}\left(Z_i - \widehat{Z}_i\right) \ge \frac{\epsilon_b}{n_b}\right\}\\ &\le \mathbb{P}_1\left\{\left|\frac{1}{n_b}\sum_{i=1}^{n_b}\left(Z_i - \widehat{Z}_i\right) - \mathbb{E}_1\left[D(p_1||\widehat{p}_i)\right]\right| \right. \qquad (40)\\ &\qquad \left. \ge \frac{\epsilon_b}{n_b} - \mathbb{E}_1\left[D(p_1||\widehat{p}_i)\right]\right\}\\ &\overset{(*)}{\le} \mathrm{Var}_1\left(\frac{1}{n_b}\sum_{i=1}^{n_b}\log\frac{p_1(X_i)}{\widehat{p}_i(X_i)}\right)\left(\frac{\epsilon_b}{n_b} - \mathbb{E}_1\left[D(p_1||\widehat{p}_i)\right]\right)^{-2}\\ &\le \frac{C_2}{n_b^{\beta_2}}\left(\frac{\epsilon_b}{n_b} - \mathbb{E}_1\left[D(p_1||\widehat{p}_i)\right]\right)^{-2}. \qquad (41)\end{aligned}$$

Here $(*)$ follows from Chebyshev's inequality. Now, (40) is less than or equal to δ_b^2 if

$$\frac{\epsilon_b}{n_b} - \mathbb{E}_1\left[D(p_1||\widehat{p}_i)\right] \ge \frac{\sqrt{C_2}}{\delta_b n_b^{\frac{\beta_2}{2}}},$$

which is equivalent to

$$\epsilon_b \ge \frac{\sqrt{C_2}n_b^{1-\beta_2/2}}{\delta_b} + n_b\mathbb{E}_1\left[D(p_1||\widehat{p}_i)\right]. \tag{42}$$

Consider the two terms on the right-hand-side of (42). Since $\mathbb{E}_1\left[D(p_1||\widehat{p}_i)\right] \le C_1 n_b^{-\beta_1}$ (from (39)), the second term in (42) is no larger than $C_1 n_b^{1-\beta_1}$. In order to choose a proper ϵ_b, there are three cases depending on the rate of the first term in (42).

- Case 1: $4\beta_1 > \beta_2$. Let

$$\epsilon_b = \frac{2\sqrt{C_2}n_b^{1-\beta_2/2}}{\delta_b}, \tag{43}$$

with δ_b as chosen below. With this ϵ_b, the first term in (38) becomes

$$\begin{aligned}&\mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < b + \epsilon_b\right\}\\ &= \mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < (1-\delta_b)n_b I + \frac{2\sqrt{C_2}n_b^{1-\beta_2/2}}{\delta_b}\right\}\\ &= \mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < (1-\delta_b)n_b I\left(1 + \frac{2\sqrt{C_2}}{(1-\delta_b)\delta_b n_b^{\beta_2/2} I}\right)\right\}\\ &\le \mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < (1-\delta_b)n_b I\left(1 + \frac{2\eta\sqrt{C_2}}{\delta_b n_b^{\beta_2/2} I}\right)\right\}\end{aligned}$$

where in the last inequality we have used the fact that $1 - \delta_b > \eta^{-1}$. Let

$$\delta_b = \frac{(4\eta^2 C_2)^{\frac{1}{4}}}{n_b^{\beta_2/4}\sqrt{I}} \iff \frac{2\eta\sqrt{C_2}}{\delta_b n_b^{\beta_2/2} I} = \delta_b. \tag{44}$$

With this chosen δ_b,

$$\mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < b + \epsilon_b\right\} \le \mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < (1-\delta_b^2)n_b I\right\}.$$

Assuming that Assumption 6 is true for the true Z_i's, we have [5, Appendix B]

$$\mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < (1-\delta_b^2)n_b I\right\} \le \delta_b^2,$$

and thus

$$\mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < b + \epsilon_b\right\} \le \delta_b^2. \tag{45}$$

Now, we verify that (42) holds for all large enough b's. With the chosen δ_b (in (44)), the first term in (42) satisfies

$$\sqrt{C_2}\delta_b^{-1}n_b^{1-\beta_2/2} = \Theta(n_b^{1-\beta_2/4}) = \omega(n_b^{1-\beta_1}).$$

Therefore, the chosen ϵ_b (in (43)) satisfies (42) for all b's large enough. As a result, from (40), we get

$$\mathbb{P}_1\left\{\frac{1}{n_b}\sum_{i=1}^{n_b}\left(Z_i - \widehat{Z}_i\right) \ge \frac{\epsilon_b}{n_b}\right\} \le \delta_b^2.$$

- Case 2: $4\beta_1 < \beta_2$. Let

$$\begin{aligned}\epsilon_b &= 2C_1 n_b^{1-\beta_1},\\ \delta_b &= \frac{2\eta C_1}{I}n_b^{-\beta_1} \iff \frac{\eta\epsilon_b}{n_b I} = \delta_b.\end{aligned} \tag{46}$$

With this choice,

$$\begin{aligned}&\mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < b + \epsilon_b\right\}\\ &= \mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < (1-\delta_b)n_b I + 2C_1 n_b^{1-\beta_1}\right\}\end{aligned}$$

$$= \mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < (1-\delta_b)n_b I\left(1+\frac{2C_1}{(1-\delta_b)n_b^{\beta_1} I}\right)\right\}$$
$$\leq \mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < (1-\delta_b)n_b I\left(1+\frac{2\eta C_1}{n_b^{\beta_1} I}\right)\right\}$$
$$= \mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < (1-\delta_b^2)n_b I\right\} \tag{47}$$

and thus, assuming Assumption 6 holds, we have

$$\mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < b+\epsilon_b\right\} \leq \delta_b^2.$$

Also, since

$$\sqrt{C_2}\delta_b^{-1} n_b^{1-\beta_2/2} = \Theta\left(n_b^{1-\beta_2/2+\beta_1}\right) = o(n_b^{1-\beta_1}),$$

the chosen ϵ_b (in (46)) satisfies (42) for all b's large enough. As a result, from (40), we get

$$\mathbb{P}_1\left\{\frac{1}{n_b}\sum_{i=1}^{n_b}\left(Z_i - \widehat{Z}_i\right) \geq \frac{\epsilon_b}{n_b}\right\} \leq \delta_b^2.$$

- Case 3: $4\beta_1 = \beta_2$. Let C_3 be a large enough constant such that

$$C_3 \geq \frac{I\sqrt{C_2}}{\eta C_3} + C_1. \tag{48}$$

Choose

$$\epsilon_b = C_3 n_b^{1-\beta_1},$$
$$\delta_b = \frac{\eta C_3}{I} n_b^{-\beta_1} \iff \frac{\eta\epsilon_b}{n_b I} = \delta_b. \tag{49}$$

Following the same line of argument as in (47), we get

$$\mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < b+\epsilon_b\right\}$$
$$= \mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < (1-\delta_b)n_b I + 2C_1 n_b^{1-\beta_1}\right\} \leq \delta_b^2.$$

Also, from (48),

$$\epsilon_b = C_3 n_b^{1-\beta_1}$$
$$\geq \frac{I\sqrt{C_2}}{\eta C_3} n_b^{1-\beta_2/2+\beta_1} + C_1 n_b^{1-\beta_1}$$
$$= \frac{\sqrt{C_2} n_b^{1-\beta_2/2}}{\delta_b} + C_1 n_b^{1-\beta_1}.$$

Therefore, (42) is satisfied for the chosen ϵ_b (in (49)), and from (40) we get

$$\mathbb{P}_1\left\{\frac{1}{n_b}\sum_{i=1}^{n_b}\left(Z_i - \widehat{Z}_i\right) \geq \frac{\epsilon_b}{n_b}\right\} \leq \delta_b^2.$$

To sum up, in all cases, we have shown the existence of ϵ_b and δ_b (that depend on β_1 and β_2) such that

$$\mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i < b+\epsilon_b\right\} \leq \delta_b^2$$

and

$$\mathbb{P}_1\left\{\frac{1}{n_b}\sum_{i=1}^{n_b}\left(Z_i - \widehat{Z}_i^{n_b,1}\right) \geq \frac{\epsilon_b}{n_b}\right\} \leq \delta_b^2$$

hold simultaneously. Continuing (38), we can write, for any (n,ν) such that $n \geq \nu \geq 1$,

$$\mathbb{P}_\nu\left\{\sum_{i=n}^{n+n_b-1} \widehat{Z}_i^{n+n_b-1,n} < b\right\}$$
$$\leq \mathbb{P}_1\left\{\sum_{i=1}^{n_b} Z_i \leq b+\epsilon_b\right\}$$
$$+ \mathbb{P}_1\left\{\frac{1}{n_b}\sum_{i=1}^{n_b}\left(Z_i - \widehat{Z}_i^{n_b,1}\right) \geq \frac{\epsilon_b}{n_b}\right\}$$
$$\leq 2\delta_b^2.$$

This is exactly what was required to be shown in (35). The proof is now complete. □

Proof of Lemma 4: Define the SR-like statistic

$$R_n = (1+R_{n-1})e^{\widehat{Z}_n^w},\ \forall n > w$$

with $R_1 = \cdots = R_w = 0$. Also define the corresponding test:

$$\overline{\tau}_R(b) := \inf\left\{n > w : R_n \geq e^b\right\}.$$

Note that the NWLA-CuSum statistic in (21) can be written equivalently as

$$e^{\overline{W}(n)} = \max\left\{1, e^{\overline{W}(n-1)}\right\} e^{\widehat{Z}_n^w},\quad n > w.$$

Therefore, for $n > w$, $R_n > e^{\overline{W}(n)}$ and thus $\overline{\tau}(b) \geq \overline{\tau}_R(b)$ on $\{\overline{\tau}(b) < \infty\}$.

Now, without loss of generality assume $\mathbb{E}_\infty[\overline{\tau}(b)] < \infty$; otherwise the statement of the lemma holds trivially. This implies that $\mathbb{E}_\infty[\overline{\tau}_R(b)] < \infty$. Observe that $R_n \in \mathcal{F}_n$ and

$$\mathbb{E}_\infty[R_n - n|\mathcal{F}_{n-1}] = (1+R_{n-1})\mathbb{E}_\infty\left[e^{\widehat{Z}_n^w}|\mathcal{F}_{n-1}\right] - n$$
$$= R_{n-1} - (n-1),\ \forall n > w.$$

The last equality follows because $\widehat{p}_n^w$ is a density given $\mathcal{F}_{n-1}$. Hence $\{R_n - n\}_{n>w}$ is a $(\mathbb{P}_\infty, \mathcal{F}_n)$-martingale. Also, for any $n > w$, since $R_n \in (0, e^b)$ almost surely on the event $\{\overline{\tau}_R(b) > n\}$, we have, for any $n > w$,

$$\mathbb{E}_\infty\left[|(R_{n+1} - (n+1)) - (R_n - n)|\,\middle|\,\mathcal{F}_n\right]$$
$$= \mathbb{E}_\infty\left[|R_{n+1} - R_n - 1|\,\middle|\,\mathcal{F}_n\right]$$
$$\leq \mathbb{E}_\infty[R_{n+1}|\mathcal{F}_n] + (R_n+1)$$
$$= 2(R_n+1)$$
$$\leq 2(e^b+1)$$

almost surely on the event $\{\overline{\tau}_R(b) > n\}$. Therefore, we can apply the optional sampling theorem and obtain

$$\mathbb{E}_\infty\left[R_{\overline{\tau}_R(b)} - \overline{\tau}_R(b)\right] = \mathbb{E}_\infty[R_{w+1} - (w+1)]$$
$$= \mathbb{E}_\infty\left[e^{\widehat{Z}_{w+1}^w}\right] - (w+1)$$
$$= -w,$$

where $\mathbb{E}_\infty\left[e^{\widehat{Z}^w_{w+1}}\right]=1$ because $\widehat{p}^w_{w+1}$ is a density given $\mathcal{F}_w$. Finally, we arrive at

$$\mathbb{E}_\infty\left[\overline{\tau}(b)\right]\geq\mathbb{E}_\infty\left[\overline{\tau}_R(b)\right]=w+\mathbb{E}_\infty\left[R_{\overline{\tau}_R(b)}\right]\geq e^b.\square$$

□

Proof of Lemma 5: The proof is similar to [7, Lemma 4]. Define a helping stopping time

$$\overline{\tau}_\nu(b):=\inf\{n\geq\nu+w:\overline{W}_\nu(n)\geq b\}$$

where

$$\overline{W}_\nu(n)=\left(\overline{W}_\nu(n-1)\right)^+\widehat{Z}^w_n,\quad n\geq\nu+w$$

with $\overline{W}_\nu(n)=0,\forall n<\nu+w$. Note that $\overline{W}(\nu+w)\geq\overline{W}_\nu(\nu+w)$. Now, if $\overline{W}(k)\geq\overline{W}_\nu(k)$, we have

$$\begin{aligned}\overline{W}(k+1)&=\left(\overline{W}(k)\right)^+\widehat{Z}^w_{k+1}\\&\geq\left(\overline{W}_\nu(k)\right)^+\widehat{Z}^w_{k+1}\\&=\overline{W}_\nu(k+1)\end{aligned}$$

as long as $\overline{W}(k)<b$. Thus, by induction,

$$\overline{W}(n)\geq\overline{W}_\nu(n),\ \forall n\geq\nu+w$$

on the event $\left\{\overline{W}(n)<b\right\}$, which implies that $\overline{\tau}(b)\leq\overline{\tau}_\nu(b)$ almost surely under $\mathbb{P}_\nu$. In the remainder of the proof, we omit "(b)" in the descriptions of the stopping times for notational brevity.

Since $\overline{\tau}-\nu+1=w+(\overline{\tau}-\nu-w+1)\leq w+(\overline{\tau}-\nu-w+1)^+$, we have

$$(\overline{\tau}-\nu+1)^+\leq w+(\overline{\tau}-\nu-w+1)^+\quad\mathbb{P}_\nu\text{–}a.s.$$

Thus,

$$\begin{aligned}&\mathbb{E}_\nu\left[(\overline{\tau}-\nu+1)^+|\mathcal{F}_{\nu-1}\right]\\&\leq w+\mathbb{E}_\nu\left[(\overline{\tau}-\nu-w+1)^+|\mathcal{F}_{\nu-1}\right]\\&=w+\mathbb{E}_\nu\Bigg[\\&\quad\mathbb{E}_\nu\left[(\overline{\tau}-\nu-w+1)^+|X_\nu,\ldots,X_{\nu+w-1},\mathcal{F}_{\nu-1}\right]\Bigg|\mathcal{F}_{\nu-1}\Bigg]\\&\overset{(*)}{\leq}w+\mathbb{E}_\nu\Bigg[\\&\quad\mathbb{E}_\nu\left[(\overline{\tau}_\nu-\nu-w+1)^+\Big|X_\nu,\ldots,X_{\nu+w-1},\mathcal{F}_{\nu-1}\right]\Bigg|\mathcal{F}_{\nu-1}\Bigg]\\&\overset{(**)}{=}w+\mathbb{E}_\nu\Bigg[\\&\quad\mathbb{E}_\nu\left[\overline{\tau}_\nu-\nu-w+1\Big|X_\nu,\ldots,X_{\nu+w-1},\mathcal{F}_{\nu-1}\right]\Bigg|\mathcal{F}_{\nu-1}\Bigg]\end{aligned}$$

where $(*)$ holds because $\overline{\tau}\leq\overline{\tau}_\nu$ almost surely (under $\mathbb{P}_\nu$), and $(**)$ holds because $\overline{\tau}_\nu\geq\nu+w-1\geq0$ almost surely (under $\mathbb{P}_\nu$).

Now, $\forall\nu\geq1$, given the information of $X_\nu,\ldots,X_{\nu+w-1}$, the event $\{\overline{\tau}_\nu\geq\nu+w\}$ is independent of $\mathcal{F}_{\nu-1}$. Thus,

$$\begin{aligned}&\operatorname{ess\,sup}\mathbb{E}_\nu\Bigg[\\&\quad\mathbb{E}_\nu\left[\overline{\tau}_\nu-\nu-w+1|X_\nu,\ldots,X_{\nu+w-1},\mathcal{F}_{\nu-1}\right]\Bigg|\mathcal{F}_{\nu-1}\Bigg]\\&=\mathbb{E}_\nu\left[\mathbb{E}_\nu\left[\overline{\tau}_\nu-\nu-w+1|X_\nu,\ldots,X_{\nu+w-1}\right]\right]\\&=\mathbb{E}_1\left[\mathbb{E}_1\left[\overline{\tau}_1-w|X_1,\ldots,X_w\right]\right]\\&=\mathbb{E}_1\left[\overline{\tau}-w\right].\end{aligned}$$

The last line holds because $\overline{\tau}_1=\overline{\tau}$ almost surely (under $\mathbb{P}_1$). The proof is now complete. □

Proof of Lemma 7: First, for any $k>0$,

$$\sum_{i=w+1}^{w+wk}\widehat{Z}^w_i=\sum_{j=1}^{w}\sum_{\ell=1}^{k}\widehat{Z}^w_{w\ell+j}$$

and given j, $\sum_{\ell=1}^k\widehat{Z}^w_{w\ell+j}$ is a sum of i.i.d. random variables under $\mathbb{P}_1$. In the following, we extend the idea of [26, Prop. 8.21] to w-dependent sequence of random variables. For any $n<\infty$,

$$\begin{aligned}\mathbb{E}_1\left[U_{\min\{\tau_u,n\}}\right]&=\mathbb{E}_1\left[\sum_{i=w+1}^{\min\{\tau_u,n\}}\widehat{Z}^w_i\right]\\&=\mathbb{E}_1\left[\sum_{j=1}^{w}\sum_{\ell=1}^{\lfloor(\min\{\tau_u,n\}-j)/w\rfloor}\widehat{Z}^w_{w\ell+j}\right]\\&=\sum_{j=1}^{w}\mathbb{E}_1\left[\sum_{\ell=1}^{\lfloor(\min\{\tau_u,n\}-j)/w\rfloor}\widehat{Z}^w_{w\ell+j}\right].\end{aligned}$$

By Wald's identity (which is applicable since $(\widehat{Z}^w_{w\ell+j})_{\ell\geq1}$ is i.i.d.), for each $j=1,\ldots,w$,

$$\begin{aligned}&\mathbb{E}_1\left[\sum_{\ell=1}^{\lfloor(\min\{\tau_u,n\}-j)/w\rfloor}\widehat{Z}^w_{w\ell+j}\right]\\&=\widehat{I}\cdot\mathbb{E}_1\left[\lfloor(\min\{\tau_u,n\}-j)/w\rfloor\right]\\&\geq\widehat{I}\left(\frac{1}{w}\mathbb{E}_1\left[\min\{\tau_u,n\}\right]-\frac{j}{w}-1\right).\end{aligned}$$

Thus,

$$\mathbb{E}_1\left[U_{\min\{\tau_u,n\}}\right]\geq\widehat{I}\left(\mathbb{E}_1\left[\min\{\tau_u,n\}\right]-\frac{3w+1}{2}\right).$$

Next we consider two cases. Suppose initially that for some $c<\infty$, $\widehat{Z}^w_i<c$, $\mathbb{P}_1$ almost surely. If this is true, since $\widehat{I}=\mathbb{E}_1\left[\widehat{Z}^w_i\right]>0$, $\forall i>w$, we have

$$\begin{aligned}&\widehat{I}\left(\frac{1}{w}\mathbb{E}_1\left[\min\{\tau_u,n\}\right]-\frac{3w+1}{2}\right)\\&\quad\leq\mathbb{E}_1\left[U_{\min\{\tau_u,n\}}\right]\leq b+c\\&\implies\mathbb{E}_1\left[\min\{\tau_u,n\}\right]\leq\frac{b+c}{\widehat{I}}+\frac{3w+1}{2}<\infty.\end{aligned}$$

Letting $n\to\infty$ we get, by Monotone Convergence Theorem,

$$\mathbb{E}_1\left[\tau_u\right]=\mathbb{E}_1\left[\lim_n\min\{\tau_u,n\}\right]=\lim_n\mathbb{E}_1\left[\min\{\tau_u,n\}\right]<\infty.$$

In general, define $\tilde{Z}^{w,c}_i:=\widehat{Z}^w_i\mathbb{1}\{\widehat{Z}^w_i<c\}+c\mathbb{1}\{\widehat{Z}^w_i>c\}$, where c is large enough such that $\mathbb{E}_1\left[\tilde{Z}^{w,c}_i\right]>0$. Similarly define

$$U^c_n:=\sum_{i=w+1}^{n}\tilde{Z}^{w,c}_i,\quad\tau^c_u:=\inf\{n>w:U^c_n\geq b\}.$$

Note that $\tilde{Z}_i^{w,c} < c$ almost surely and thus $\mathbb{E}_1[\tau_u^c] < \infty$. Since $U_n^c \le U_n$, we have $\tau_u \le \tau_u^c$ almost surely under $\mathbb{P}_1$. Therefore, $\mathbb{E}_1[\tau_u] \le \mathbb{E}_1[\tau_u^c] < \infty$. The proof is now complete. □

Proof of Lemma 8: The proof consists of two parts. In the first part, we use a similar technique as in [27, Thm 1.1] to obtain an extension of Wald's identity to the case where the samples are w-dependent. In the second part, we upper bound the overshoot using results from renewal theory. For notational brevity, we omit the dependence on w and write $\widehat{Z}_i^w = \widehat{Z}_i$.

Define

$$Y_i := \mathbb{E}_1\left[U_i - (i-w)\widehat{I}|\mathcal{F}_{i-w}\right],\ \forall i \ge w$$

and note that $Y_w = 0$. Now,

$$\begin{aligned}
&\mathbb{E}_1[Y_{i+1}|\mathcal{F}_{i-w}]\\
&= \mathbb{E}_1\left[\mathbb{E}_1\left[U_{i+1} - (i+1-w)\widehat{I}|\mathcal{F}_{i+1-w}\right]|\mathcal{F}_{i-w}\right]\\
&= \mathbb{E}_1\left[U_i + \widehat{Z}_{i+1} - (i-w)\widehat{I} - \widehat{I}|\mathcal{F}_{i-w}\right]\\
&= Y_i + \mathbb{E}_1\left[\widehat{Z}_{i+1} - \widehat{I}|\mathcal{F}_{i-w}\right]\\
&\overset{(*)}{=} Y_i + \mathbb{E}_1\left[\widehat{Z}_{i+1} - \widehat{I}\right]\\
&= Y_i,\ \forall i > w,\ \mathbb{P}_1\text{–}a.s.
\end{aligned}$$

where $(*)$ follows from independence between $\widehat{Z}_{i+1}^w$ and $\mathcal{F}_{i-w}$. This implies that $\{(Y_i, \mathcal{F}_{i-w})\}_{i\ge w}$ is a martingale. Therefore, for any finite $k > w$, $\min\{\tau_u, k\} \le k < \infty$, and thus

$$\begin{aligned}
&\mathbb{E}_1\left[U_{\min\{\tau_u,k\}} - \widehat{I}(\min\{\tau_u,k\} - w)\right]\\
&= \sum_{m=1}^{\infty} \mathbb{P}_1\{\tau_u = m\} \times \mathbb{E}_1\Big[\mathbb{E}_1\big[U_{\min\{m,k\}}\\
&\qquad - \widehat{I}(\min\{m,k\} - w)\big|\mathcal{F}_{\min\{m,k\}-w}\big]\Big|\tau_u = m\Big]\\
&= \sum_{m=1}^{\infty} \mathbb{P}_1\{\tau_u = m\}\,\mathbb{E}_1\left[Y_{\min\{m,k\}}|\tau_u = m\right]\\
&= \mathbb{E}_1\left[Y_{\min\{\tau_u,k\}}\right] \overset{(*)}{=} \mathbb{E}_1[Y_w] = 0
\end{aligned}$$

where $(*)$ follows from optional sampling theorem. This implies that

$$\mathbb{E}_1\left[U_{\min\{\tau_u,k\}}\right] = \widehat{I}(\mathbb{E}_1[\min\{\tau_u,k\}] - w). \tag{50}$$

Note that $\tau_u < \infty$ with probability 1 under $\mathbb{P}_1$ by Lemma 7. For $i > w$, let $\widehat{Z}_i^+ := \max\{0, \widehat{Z}_i\}$ and $\widehat{Z}_i^- := -\min\{0, \widehat{Z}_i\}$. Note that $\widehat{Z}_i^+, \widehat{Z}_i^- \ge 0$, $\widehat{Z}_i = \widehat{Z}_i^+ - \widehat{Z}_i^-$, and $U_n = \sum_{i=w+1}^{n}\left(\widehat{Z}_i^+ - \widehat{Z}_i^-\right), \forall n > w$. Thus, we have

$$\begin{aligned}
&\lim_{k\to\infty} \mathbb{E}_1\left[U_{\min\{\tau_u,k\}}\right]\\
&= \lim_{k\to\infty} \mathbb{E}_1\left[\sum_{i=w+1}^{\min\{\tau_u,k\}} \widehat{Z}_i^+\right] - \lim_{k\to\infty} \mathbb{E}_1\left[\sum_{i=w+1}^{\min\{\tau_u,k\}} \widehat{Z}_i^-\right]\\
&\overset{(i)}{=} \mathbb{E}_1\left[\lim_{k\to\infty}\sum_{i=w+1}^{\min\{\tau_u,k\}} \widehat{Z}_i^+\right] - \mathbb{E}_1\left[\lim_{k\to\infty}\sum_{i=w+1}^{\min\{\tau_u,k\}} \widehat{Z}_i^-\right]\\
&\overset{(ii)}{=} \mathbb{E}_1\left[\sum_{i=w+1}^{\tau_u} \widehat{Z}_i^+\right] - \mathbb{E}_1\left[\sum_{i=w+1}^{\tau_u} \widehat{Z}_i^-\right]\\
&= \mathbb{E}_1[U_{\tau_u}]
\end{aligned}$$

where (i) follows from the monotone convergence theorem, and (ii) is due to the fact that $\tau_u < \infty$ with probability 1. Also by the monotone convergence theorem,

$$\lim_{k\to\infty} \mathbb{E}_1[\min\{\tau_u,k\}] = \mathbb{E}_1\left[\lim_{k\to\infty}\min\{\tau_u,k\}\right] = \mathbb{E}_1[\tau_u].$$

Thus, taking the limit of k on both sides of (50),

$$\begin{aligned}
\mathbb{E}_1[U_{\tau_u}] &= \lim_{k\to\infty} \mathbb{E}_1\left[U_{\min\{\tau_u,k\}}\right]\\
&= \lim_{k\to\infty} \widehat{I}(\mathbb{E}_1[\min\{\tau_u,k\}] - w)\\
&= \widehat{I}(\mathbb{E}_1[\tau_u] - w).
\end{aligned} \tag{51}$$

Now, denote

$$L_i := \log\frac{\widehat{p}_i^w(X_i)}{p_1(X_i)}, \quad \forall i > w.$$

By definition we have $\mathbb{E}_1\left[\widehat{Z}_i - L_i\right] = I, \forall i > w$. The proof of (51) is also applicable to $L_{\tau_u}^2$, which gives us

$$\mathbb{E}_1\left[\sum_{i=w+1}^{\tau_u} L_i^2\right] = \mathbb{E}_1\left[L_{w+1}^2\right](\mathbb{E}_1[\tau_u] - w). \tag{52}$$

Thus,

$$\begin{aligned}
&\mathbb{E}_1[U_{\tau_u} - b]\\
&= \mathbb{E}_1\left[U_{\tau_u-1} - b + \widehat{Z}_{\tau_u}\right] < \mathbb{E}_1\left[\widehat{Z}_{\tau_u}\right] = I + \mathbb{E}_1[L_{\tau_u}]\\
&\overset{(i)}{\le} I + \sqrt{\mathbb{E}_1\left[L_{\tau_u}^2\right]} \le I + \sqrt{\mathbb{E}_1\left[\sum_{i=w+1}^{\tau_u} L_i^2\right]}\\
&\overset{(ii)}{=} I + \sqrt{(\mathbb{E}_1[\tau_u] - w)\,\mathbb{E}_1\left[L_{w+1}^2\right]}\\
&\overset{(iii)}{\le} I + \sqrt{\frac{C_2}{w^{\beta_2}}(\mathbb{E}_1[\tau_u] - w)}
\end{aligned} \tag{53}$$

for sufficiently large w. Here (i) follows from Jensen's inequality, (ii) follows from (52), and (iii) follows from Assumption 2. Denote $c_w := C_2 w^{-\beta_2}$ and $x := \mathbb{E}_1[\tau_u] - w$. The goal below is to get an upper bound for x. Combining (53) with (51), we obtain

$$\sqrt{c_w x} + I \ge \widehat{I}x - b$$

which implies that

$$\begin{aligned}
x &\le \frac{(2(b+I)\widehat{I} + c_w) + \sqrt{(2(b+I)\widehat{I} + c_w)^2 - 4\widehat{I}^2(b+I)^2}}{2\widehat{I}^2}\\
&\le \frac{2(b+I)\widehat{I} + c_w}{\widehat{I}^2}.
\end{aligned}$$

Plugging this bound into (53) gives us

$$\mathbb{E}_1[U_{\tau_u} - b] \le I + \sqrt{\frac{2(b+I)c_w\widehat{I} + c_w^2}{\widehat{I}^2}}$$

$$\leq I + \frac{2\sqrt{(b+I)c_w\widehat{I}} + \sqrt{2}c_w}{\widehat{I}}$$

where in the last inequality we use the fact that $\sqrt{u+v} \leq \sqrt{2u}+\sqrt{2v}$ for any $u, v > 0$. Therefore, combining with (51), we obtain

$$\begin{aligned}\mathbb{E}_1[\tau_u] &= w + \widehat{I}^{-1}\left(b + \mathbb{E}_1\left[U_{\tau_u} - b\right]\right) \\ &\leq w + \widehat{I}^{-1}(b+I) + \left(4C_2(b+I)w^{-\beta_2}\widehat{I}^{-3}\right)^{\frac{1}{2}} \\ &\quad + \sqrt{2}\widehat{I}^{-2}C_2 w^{-\beta_2}.\end{aligned}$$

The proof is now complete since $\mathbb{E}_1[\overline{\tau}(b)] \leq \mathbb{E}_1[\tau_u(b)]$ by Lemma 6. □

ACKNOWLEDGMENT

The authors would like to thank George Moustakides for helpful discussions regarding the NWLA-CuSum test.

REFERENCES

[1] Y. Liang and V. V. Veeravalli, "Quickest change detection with leave-one-out density estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.

[2] V. V. Veeravalli and T. Banerjee, "Quickest change detection," in *Academic Press Library in Signal Processing: Array and Statistical Signal Processing*. Cambridge, MA, USA: Academic, 2013.

[3] L. Xie, S. Zou, Y. Xie, and V. V. Veeravalli, "Sequential (quickest) change detection: Classical results and new directions," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 2, pp. 494–514, Jun. 2021.

[4] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 1897–1908, Dec. 1971.

[5] T. Leung Lai, "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2917–2929, Nov. 1998.

[6] Y. Liang, A. G. Tartakovsky, and V. V. Veeravalli, "Quickest change detection with non-stationary post-change observations," *IEEE Trans. Inf. Theory*, vol. 69, no. 5, pp. 3400–3414, May 2023.

[7] L. Xie, G. V. Moustakides, and Y. Xie, "Window-limited CUSUM for sequential change detection," *IEEE Trans. Inf. Theory*, vol. 69, no. 9, pp. 5990–6005, Sep. 2023.

[8] S. Li, Y. Xie, H. Dai, and L. Song, "M-statistic for kernel change-point detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2015, pp. 3366–3374.

[9] T. Flynn and S. Yoo, "Change detection with the kernel cumulative sum algorithm," in *Proc. IEEE 58th Conf. Decis. Control (CDC)*, Nice, France, Dec. 2019, pp. 6092–6099.

[10] F. Desobry, M. Davy, and C. Doncarli, "An online kernel change detection algorithm," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2961–2974, Aug. 2005.

[11] L. Chu and H. Chen, "Sequential change-point detection for high-dimensional and non-Euclidean data," *IEEE Trans. Signal Process.*, vol. 70, pp. 4498–4511, 2022.

[12] H. Chen, "Sequential change-point detection based on nearest neighbors," *Ann. Statist.*, vol. 47, no. 3, pp. 1381–1407, 2019.

[13] Y. Yilmaz, "Online nonparametric anomaly detection based on geometric entropy minimization," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Sep. 2017, pp. 3010–3014.

[14] M. N. Kurt, Y. Yilmaz, and X. Wang, "Real-time nonparametric anomaly detection in high-dimensional settings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2463–2479, Jul. 2021.

[15] Y. Kawahara and M. Sugiyama, "Sequential change-point detection based on direct density-ratio estimation," *Stat. Anal. Data Mining*, vol. 5, no. 2, pp. 114–127, Apr. 2012, doi: 10.1002/sam.10124.

[16] G. V. Moustakides and K. Basioti, "Training neural networks for Likelihood/Density ratio estimation," 2019, *arXiv:1911.00405*.

[17] T. S. Lau, W. P. Tay, and V. V. Veeravalli, "A binning approach to quickest change detection with unknown post-change distribution," *IEEE Trans. Signal Process.*, vol. 67, no. 3, pp. 609–621, Feb. 2019.

[18] P. Hall, "On Kullback–Leibler loss and density estimation," *Ann. Statist.*, vol. 15, no. 4, pp. 1491–1519, Dec. 1987.

[19] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, 2nd ed. Hoboken, NJ, USA: Wiley, 2015.

[20] L. Wasserman, *All of Nonparametric Statistics*. New York, NY, USA: Springer, 2006.

[21] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. New York, NY, USA: Springer, 2009.

[22] P. Moulin and V. V. Veeravalli, *Statistical Inference for Engineers and Data Scientists*. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[23] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, nos. 1–2, pp. 100–115, 1954.

[24] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Ann. Statist.*, vol. 14, no. 4, pp. 1379–1387, Dec. 1986.

[25] A. G. Tartakovsky, *Sequential Change Detection and Hypothesis Testing: General Non-I.I.D. Stochastic Models and Asymptotically Optimal Rules* (Monographs on Statistics and Applied Probability), vol. 165. Boca Raton, FL, USA: Chapman & Hall, 2020.

[26] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*. New York, NY, USA: Springer, 1985.

[27] S. Janson, "Renewal theory for M-dependent variables," *Ann. Probab.*, vol. 11, no. 3, pp. 558–568, Aug. 1983.

Yuchen Liang (Member, IEEE) received the B.Sc. degree (Hons.) in computer engineering and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana–Champaign in 2019 and 2023, respectively. He is currently a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, The Ohio State University. His research interests include machine learning, quickest change detection, non-parametric statistics, decision and estimation theory, and data science.

Venugopal V. Veeravalli (Fellow, IEEE) received the B.Tech. degree (Hons.) in electrical engineering from the Indian Institute of Technology, Bombay, in 1985, the M.S. degree in electrical engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 1987, and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana–Champaign in 1992. He joined the University of Illinois at Urbana–Champaign in 2000, where he is currently the Henry Magnuski Professor with the Department of Electrical and Computer Engineering, and where he is also affiliated with the Department of Statistics and the Coordinated Science Laboratory. Prior to joining the University of Illinois at Urbana–Champaign, he was a Faculty Member with the ECE Department, Cornell University. He was the Program Director for communications research with the U.S. National Science Foundation from 2003 to 2005. His research interests include statistical inference, machine learning, and information theory, with applications to data science, wireless communications, and sensor networks. He was elected as a fellow of the Institute of Mathematical Statistics in 2024. Among the awards he has received for research and teaching are the IEEE Browder J. Thompson Best Paper Award in 1996, the Presidential Early Career Award for Scientists and Engineers (PECASE) in 1999, the Wald Prize in Sequential Analysis in 2015 and 2019, and the Fulbright-Nokia Distinguished Chair in Information and Communication Technologies in 2023. From 2004 to 2007, he served on the Board of Governors of the IEEE Information Theory Society, where he is serving a second term. From 2011 to 2016, he served on the SPTM Technical Committee. From 2017 to 2019, he served on the Big Data SIG. He is the Editor-in-Chief of IEEE TRANSACTIONS ON INFORMATION THEORY. He has been an Associate Editor for Detection and Estimation of IEEE TRANSACTIONS ON INFORMATION THEORY and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He has been a Senior Area Editor of IEEE OPEN JOURNAL ON SIGNAL PROCESSING and an Area Editor for Statistics and Machine Learning of IEEE TRANSACTIONS ON INFORMATION THEORY. From 2010 to 2011, he was a Distinguished Lecturer of the IEEE Signal Processing Society.