Online Adaptive Anomaly Thresholding with Confidence Sequences

Sophia Sun [1] **Abishek Sankararaman** [2] **Balakrishnan (Murali) Narayanaswamy** [2]

Abstract

Selecting appropriate thresholds for anomaly detection in online, unsupervised settings is a challenging task, especially in the presence of data distribution shifts. Addressing these challenges is critical in many practical large scale systems, such as infrastructure monitoring and network intrusion detection. This paper proposes an algorithm that connects online thresholding with constructing confidence sequences achieving (1) adaptive online threshold selection robust to distribution shifts, (2) statistical guarantees on false positive and false negative rates without any distributional assumptions, and (3) improved performance when given relevant offline data to warm-start the online algorithm, while having bounded degradation if the offline data is irrelevant. We complement our theoretical results with empirical evidence that our method outperforms commonly used baselines across synthetic and real world datasets.

1. Introduction and Motivation

Online anomaly detection (OAD) is the task of identifying deviations from the normal behavior in a streaming fashion, where samples arrive sequentially, and decisions must be made before the next sample is received. This task plays a fundamental role in various practical applications within cyber and cyber-physical systems, including maintenance (Khan et al., 2020), monitoring (Hill and Minsker, 2010), and security (Mirsky et al., 2018; Lazarevic et al., 2003). In modern environments, the sheer volume and velocity of data make it often impractical to have comprehensive labels for all anomalous samples (Siffer et al., 2017; Schmidl et al., 2022; Ren et al., 2019; Audibert et al., 2020).

In deployment, typical AD algorithms assign a real-valued anomaly score to each sample, where a higher score indicates a greater degree of anomaly (Chandola et al., 2009). The OAD algorithm first scores each sample and then calibrates a threshold to make the binary decision of whether the given sample is anomalous or not. The choice of this threshold is critical due to its significant downstream impact. In security systems, for example, failing to detect a genuine anomaly or security event can have catastrophic consequences (Ho et al., 2017a). Conversely, false positives incur costs, as each detection necessitates investigation by a security operator, potentially leading to 'alert fatigue' (Chen, 2017; Lin et al., 2018; He et al., 2023).

Given the risk of incorrect decisions, a common methodology to set the threshold is by *abstaining* from decisions for an initial 'cold-start' number of samples on the stream (Huang and Kasiviswanathan, 2015; Katz and Raz, 2023). At the end of the cold-start period, a threshold is chosen for future decision making, either by algorithm or by an expert based on the online scores during the cold-start, as well as any offline data that is available (Ren et al., 2019; Bhatia et al., 2021; Huang et al., 2022; Li et al., 2021). This approach has two drawbacks. Firstly, using a fixed cold-start duration across all streams independent of stream statistics is sub-optimal and can lead to much more abstains than necessary. Secondly, a static threshold is prone to performance loss when the data stream undergoes distribution shift - a common occurrence in many AD systems (Herley, 2022; Gama et al., 2014). Although adapting anomaly scoring to distribution drifts have been studied (Ma et al., 2018; Bhatia et al., 2022; Sankararaman et al., 2022), limited advances have occurred in adapting the threshold dynamically.

We formulate adaptive threshold selection as online quantile estimation by defining anomalies as tail quantile of anomaly scores (Steinwart et al., 2005; Cadre et al., 2013; Gan and Bailis, 2017; Siffer et al., 2017). Formally, we model that each time $t \in \mathbb{N}$, the anomaly score $S_t \in \mathbb{R}$ is independently sampled from distribution f_t. This score S_t is an anomaly if it exceeds the pth quantile of the distribution f_t. We assume the quantile level $p \in (0, 1)$ is known, while the distribution f_t as unknown to the algorithm. Knowledge of p arises either from a constraint on the rate of anomalies that can be investigated without succumbing to alert fatigue (Hassan et al., 2019; Ho et al., 2021), or from domain knowledge (Perini and Davis, 2023).

Performance of threshold selection is measured by false positives (FP), benign samples marked as anomalous, false

[1]University of California, San Diego. Work done when interning with Amazon Web Services. [2]Amazon Web Services, Santa Clara CA. Correspondence to: Sophia Sun <shs066@ucsd.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

negatives (FN), anomalous sample mark as benign, and number of samples abstained from making a decision on. Since AD outputs are typically used to make consequential decisions, e.g. engaging human operators to triage alerts, OAD systems must achieve low number of FPs and FNs, by possibly trading-off with abstains (Ho et al., 2017b). Further, these systems need to be adaptive to distribution shifts and be designed without requiring distributional assumptions (cf. *Research challenges* 5-7 of (Sadik and Gruenwald, 2014)).

1.1. Desiderata of OAD threshold selection systems

In summary, the following set of statistical requirements is desired for online threshold selection.

(i) High accuracy: Constant number of mistakes[1] independent of the stream length on an i.i.d. data-stream.

(ii) Low abstention: Abstain on a vanishing fraction of samples on an i.i.d. stream, i.e. small cold-start period.

(iii) Adaptive to distribution changes: On a non-stationary stream, number of mistakes and abstains only increase compared to the stationary setting by a constant factor that is independent of the stream length, but only proportional to the number and magnitude of distribution changes.

(iv) Learn from offline dataset: When unlabeled offline datasets are available, abstain rate is a constant independent of stream length, as long as the online distribution matches one of the offline distribution. In the worst-case of arbitrary offline dataset distribution, performance degradation compared to no offline dataset must be bounded independent of stream length and size of the dataset.

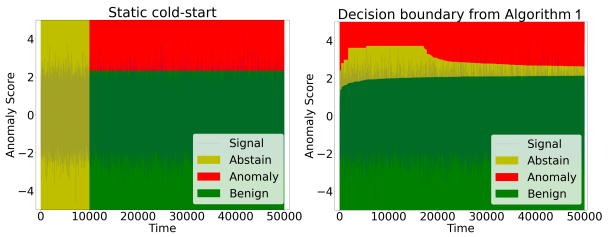
(v) Distribution and parameter free: The algorithm does not require any knowledge of the distributions or assumptions such as sub-gaussian tails or parameterized distribution family, and nevertheless must guarantee all of above.

1.2. Main result and technical contributions

Algorithm 5 is distribution and parameter free and is the *first algorithm to satisfy all the aforementioned desiderata.* Algorithm 5 does not require knowledge on the locations or number of distribution shifts nor on the online stream or the offline datasets' distributions. Nevertheless, under mild technical assumptions, the guarantees in Theorem 6 show that the performance adapts to the problem complexity. Concretely, our contributions are as follows.

1. High accuracy requires abstention. We prove that any scheme that does not abstain, cannot achieve constant FPs and FNs, even for a stationary stream. The proof follows by constructing a hard instance and a reduction from decision making to hypothesis testing on that instance.

[1]Throughout, we denote the mistakes as sum of FP and FN.



(a) (b)

Figure 1: Fig (a) is the heuristic of a fixed cold-start where all samples are abstained from and a static threshold for all future samples. Fig (b) is the decision boundary from Algorithm 6 using an adaptive cold-start.

2. Algorithm using confidence sequences (CS) to adapt threshold to distribution changes. A sample S_t is deemed as an anomaly by our algorithm (Algo. 5) if it exceeds the upper confidence of the pth quantile, is benign if it is smaller than the lower confidence, and is abstained from otherwise (Fig. 1). The crux of our algorithm is in choosing the *relevant online and offline samples* to construct the CS. Online distribution shifts are detected by constructing CS on different sub-sequences and testing if they have a non-empty intersection. Theorem 2 shows that on an i.i.d. stream of length T, no mistakes and at-most $\mathcal{O}(\sqrt{T})$ abstains occur with high probability. Theorem 4 shows that when there are distribution shifts, our algorithm incur total additional mistakes *independent* of T, without requiring any knowledge of the data distributions or change-points.

3. Relevant offline dataset improves online performance, while irrelevant offline dataset has bounded degradation. Theorem 6 proves that when offline data matching the distributions of the online stream are available, they improve abstain rate without loss of accuracy. In the *worst-case* when the offline dataset's distributions are arbitrary, the performance degradation compared to the case when the algorithm ignores offline dataset is *bounded independent of stream length or offline dataset size*. Algorithm 5 does not a priori know if the offline dataset is useful or not, but yields improved guarantees whenever the offline dataset is useful. The technical novelty is to rely on an offline dataset *only if a CS constructed on offline dataset* intersects with the online CS. This is the first algorithm for online thresholding showing improved performance in the presence of useful offline data, while guaranteeing worst-case bounds if the offline data is arbitrary.

2. Notations and Problem Formulation

We formally state the problem and performance measures.

2.1. Online Data Stream

A data-stream of length $T \in \mathbb{N}$ is a collection of T, independent, scalar valued random variables $(S_t)_{t=1}^{T}$, indexed by

time $t \in [T]$ [2]. In practice, $S_t \in \mathbb{R}$ at time t is the *anomaly score* of an input sample produced by an AD algorithm such that high scores corresponds to anomalies. For each time $t \in [T]$, f_t is the probability distribution of S_t, i.e., $S_t \sim f_t$. We assume f_t is a continuous measure without discrete atoms.[3] Since the distribution f_t is indexed by time t, $(S_t)_{t=1}^T$ is not necessarily identically distributed and thus non-stationary. We assume that the data-stream although is non-stationary, is piece-wise stationary.

Definition 2.1. (Piece-wise stationary) Let T be the time horizon and let $H_T \leq T-1$ be the total number of change points. Define a set of strictly increasing time indices $1 < \tau_1 < \tau_2 \cdots < \tau_{H_T}$ as change points; let $\tau_0 = 0$ and $\tau_{H_T+1} = T$. A piece-wise stationary data-stream $(S_t)_{t=1}^T \sim \left(T, H_T, (\tau_c)_{c=0}^{H_T}, (f^{(c)})_{c=0}^{H_T}\right)$ is such that $\forall k \in [1, H_T]$:

- $\forall t \in [\tau_k, \tau_{k+1})$, $S_t \sim f^{(k)}$ are i.i.d. with the p-quantile of $f^{(k)}$ denoted as $Q^{(k)}(p) \in \mathbb{R}$.
- $f^{(k)} \neq f^{(k+1)}$

Figure 2: The piece-wise stationary process

The special case of $H_T = 0$ corresponds to a stationary i.i.d. data stream. We remark that our algorithm does not have any information on the change-points H_T, $(\tau_t)_{t=1}^{H_T}$, nor on distributions $(f^{(c)})_{c=0}^{H_T}$.

2.2. Anomaly Definition

Definition 2.2 (Anomalies)**.** The sample $S_t \sim f_t$ at time t is an anomaly if $S_t > \tau_t$, where $\tau_t = \inf\{\tau \geq 0 : \mathbb{P}_{S\sim f_t}[S \leq \tau] \geq p\}$ and $p \in (0,1)$. We denote by $y_t = \mathbf{1}(S_t > \tau_t)$ the indicator variable that the sample S_t is an anomaly.

In words, a sample S_t is anomalous if it exceeds the threshold τ_t corresponding to the pth quantile. Thus, the thresholding algorithm needs to estimate τ_t and then use that estimate to make the binary decision of whether S_t is an anomaly. We assume the quantile value $p \in (0,1)$ to be known.

2.3. Offline Datasets

We formalize the notion of *offline datasets* that can be used in addition to using historical samples on the data-stream. Offline datasets are typically available to train the anomaly scoring systems (Bhatia et al., 2022; Audibert et al., 2020) and have been used to address the cold-start problem in previous works (Gutflaish et al., 2019; Hofmann, 1999).

[2]For any positive integer L, we denote $[L] = \{1, \cdots, L\}$

[3]Formally, $\forall p \in (0,1)$, the set $\{Q \in \mathbb{R} : \mathbb{P}_{X\sim f}[X \leq Q] = p\}$ has exactly one element.

Definition 2.3 (Offline Dataset)**.** We define an offline dataset as K sets of independent samples $\mathcal{D} := \{\{X_i^{(j)}\}_{i=1}^{N_j}\}_{j=1}^K$. Each $j \in \{1, \cdots, K\}$ set $\{X_i^{(j)}\}_{i=1}^{N_j}$ are independent samples such that all of the N_j samples are drawn from distributions having the same pth quantile denoted by $Q_{(j)}(p)$. A special case is $\{X_i^{(j)}\}_{i=1}^{N_j}$ are i.i.d., with all N_j samples having identical distributions and thus identical pth quantile.

Observe that $\mathcal{D} = \emptyset$ corresponds to the special case of no offline data availabe. We will assume that K, the number of partitions are known to the algorithm. Clustered offline datasets is a valid assumption, since typically, the offline data is pre-processed before deploying online algorithms. In practice, the clusters could correspond to different groups of signals, e.g. different sensor types, different environmental conditions or day of the week (Khan et al., 2020; Audibert et al., 2020; Sadik and Gruenwald, 2014).

2.4. Online Anomaly Thresholding Algorithm

Definition 2.4. An online anomaly thresholding algorithm $\mathcal{A}$ outputs a $\hat{y}_t = \mathcal{A}(S_t; (S_1, \ldots, S_{t-1}), \mathcal{D}) \in \{0, 1, *(\text{abstain})\}$, a decision for the score S_t depending on the history $S_1, \cdots, S_{t-1}$ and offline datasets $\mathcal{D}$ if any.

We emphasize the dependence on the offline dataset $\mathcal{D}$ in the algorithm since that can be used to make a decision at all times. However, as mentioned before, the algorithm has no information about the data-stream quadruple, nor any parametric information on the distribution from which either the scores $(S_t)_{t=1}^T$ or the offline dataset $\mathcal{D}$ are sampled from.

2.5. Performance Measures

The three performance measures of an anomaly thresholding algorithm are *(i)* False Positives FP $= \sum_{t=1}^T \mathbf{1}(\hat{y}_1 = 1, y_t = 0)$, *(ii)* False Negatives FN $= \sum_{t=1}^T \mathbf{1}(\hat{y}_0 = 1, y_t = 1)$, and *(iii)* Abstains $= \sum_{t=1}^T \mathbf{1}(\hat{y}_t = *)$. We denote by Mistakes := FP + FN as the sum of false-positives and negatives. The desiderata of high-accuracy and low abstains translate to having a constant independent of T mistakes and a sub-linear in T abstains.

3. Lower Bound: Achieving high accuracy requires abstaining

Choosing thresholds for online anomaly detection is non-trivial. In the absence of abstaining, the expected sum of FPs and FNs is at-least order $\sqrt{T}$ on an i.i.d. stream. We provide a sketch and defer details to the Appendix. Consider the standard reduction of estimation to hypothesis testing (Wainwright, 2019) where the stream is either coming from the standard gaussian or a unit variance gaussian with mean

located at $\frac{1}{\sqrt{T}}$. Denote by $Q^{(0)}(p) < Q^{(1)}(p)$ to be the pth quantile of the two distributions respectively. Lower bounds from hypothesis testing gives that at any test that looks at all T samples and identifies which of the two distributions the stream came from makes a mistake with probability at-least $1/8$. Furthermore, Chernoff bound applied to the binary random variables $\mathbf{1}_{Q^{(0)}(p)<S_t<Q^{(1)}(p)}$ gives that with probability at-least $1 - e^{\sqrt{T}/8}$, at-least $C_p\sqrt{T}$ samples among $S_1, \cdots, S_T$ lies in the range $[Q^{(0)}(p), Q^{(1)}(p)]$, where C_p is a constant depending on p. Thus, an union bound gives that in the absence of abstains, with probability at-least $1/8 - e^{\sqrt{T}/8}$, the sum of FP and FN is at-least $C_p\sqrt{T}$. A formal statement and proof in Section H in the Appendix.

4. Special Case I: Stationary stream only

In order to build up the intuition and concepts, we consider three special cases in increasing order of complexity *(i)* stationary stream without offline data *(ii)* piece-wise stationary stream without offline dataset, and *(iii)* stationary stream with offline dataset. The ideas from these special cases are developed in Section 7 to give the general algorithm.

In this section, we consider the special case when the online algorithm knows that data-stream $(S_t)_{t\geq 1}$ is an i.i.d. sequence and there is no offline dataset for warm-starting, i.e., $\mathcal{D} = \emptyset$. For this special case, we show in Theorem 2 that results from (Howard and Ramdas, 2022) yields an algorithm that guarantees 0 mistakes and $\mathcal{O}(\sqrt{T})$ abstains.

To formally state the results, we set notations. For any sequence of numbers $(y_t)_{t\geq 1}$ and $t_1 < t_2$, denote by $y_{t_1:t_2} := (y_{t_1}, \cdots, y_{t_2})$. For a positive integer t, sequence $y_{1:t}$, and $p \in (0,1)$, the pth empirical quantile of $y_{1:t}$ is given by $\widehat{Q}(p; y_{1:t}) := \frac{y^{(\lfloor pt \rfloor)} + y^{(\lceil pt \rceil)}}{2}$, where $y^{(1)} \leq \cdots \leq y^{(t)}$ is the sorted order of $y_{1:t}$.

Definition 4.1 (Confidence Sequences (CS) of a quantile)**.** For $\alpha \in (0,1)$, a level $1-\alpha$ CS of the $p \in (0,1)$ quantile of an i.i.d. sequence $(S_t)_{t\geq 1}$ with true pth quantile $Q(p) \in \mathbb{R}$, is a collection of subsets $\{C_t : t \geq 1\}$ such that for all $t \geq 1$, *(i)* $C_t \subseteq \mathbb{R}$, *(ii)* C_t is $\sigma(S_1, \ldots, S_t)$-measurable w.r.t. the first t samples and *(iii)* $\mathbb{P}(Q(p) \in \bigcap_{t\geq 1} C_t) \geq 1 - \alpha$.

Theorem 1 (Theorem 2 from (Howard and Ramdas, 2022))**.** Let $(S_t)_{t\geq 1}$ be an i.i.d. sequence of $\mathbb{R}$ valued random variables. Given $\alpha \in (0,1)$, for all $p \in (0,1)$, the sequence of sets $(C(p, \alpha, S_{1:t}))_{t\geq 1}$ is a $1-\alpha$ level CS for the pth quantile, where for any sequence of numbers $(y_{1:n})$,

$$C(p, \alpha, y_{1:n}) = \Big[\widehat{Q}(\max(p - 2u_n(\alpha), 0), y_{1:n}), \widehat{Q}(\min(p + 2u_n(\alpha), 1), y_{1:n})\Big], \quad (1)$$

where $u_{(\cdot)}(\cdot) : \mathbb{N} \times [0,1] \to [0,1]$ given by the function

$$u_t(\alpha) = 0.85\sqrt{t^{-1}[\log\log(et) + 0.8\log(1612/\alpha)]}, \quad (2)$$

and for all $\alpha \in [0,1]$, $u_0(\alpha) = 1$.

A thresholding algorithm for this special case is obtained from the definition of CS as shown in Algorithm 1.

Algorithm 1 Decision making with confidence set

Input : Sample $S \in \mathbb{R}$, Confidence set $C \subseteq \mathbb{R}$
Output : Anomaly label $\hat{y} \in \{0, 1, *\}$
if $S > \max C$ **then**
  $\hat{y} \leftarrow 1$ `// Declare an anomaly`
else if $S \in C$ **then**
  $\hat{y} \leftarrow *$ `// Abstain from decision`
else
  $\hat{y} \leftarrow 0$ `// Declare benign`
end
return $\hat{y}$

Theorem 2. Suppose for an i.i.d., data-stream $(S_t)_{t\geq 1}$, the decision at time t is given by the output of Algorithm 1 with inputs S_t and confidence $C(p, \alpha, S_{1:t-1})$. Then, with probability at-least $1 - 2\alpha$, 0 mistakes and at-most $7\sqrt{T \ln\left(\frac{1612 \ln(eT)}{\alpha^2}\right)}$ abstains are incurred.

Remark 2.1. Theorem 2 implies that points 1 and 2 from Section 1.1 are satisfied by Algorithm 6.

Remark 2.2. To bound the number of abstains, we need a different analysis compared to (Howard and Ramdas, 2022) to control the probability that a sample S_t lies within $C(p, \alpha, S_{1:t-1})$. We do this by constructing a martingale from the sequence of binary random variables $\mathbf{1}_{S_t \in C(p,\alpha,S_{1:t-1})}$ and applying Azuma's inequality.

5. Special Case II: Piece-wise stationary stream without offline data

Going beyond the i.i.d. case, we consider a piece-wise stationary stream without offline dataset. Our method in Algorithm 3 uses a change-point detection Algorithm 2 from (Shekhar and Ramdas, 2023) as a sub-routine. We need a definition to state the main result of this section.

5.1. Measure of distribution shift

Definition 5.1 (quantile shift)**.** For two continuous distributions f, g on $\mathbb{R}$ with quantile functions $Q_f(\cdot), Q_g(\cdot) : [0,1] \to \mathbb{R}$ respectively, the distance at quantile $p \in (0,1)$ denoted as $\text{Shift}(f, g, p)$ is defined as

$$\text{Shift}(f, g, p) := \sup\{\Delta \geq 0 : \max[Q_f(p - \Delta) - Q_g(p + \Delta), Q_g(p - \Delta) - Q_f(p + \Delta)] \geq 0\}.$$

Observe that for any two continuous distributions f and g, $\text{Shift}(f, g, p) \in [0, 1]$ and $\text{Shift}(f, g, p) = 0$ if and only if f and g have identical pth quantile. The following proposition shows that the quantile shift is the non-stationary measure that governs the delay in detecting a change-point.

Algorithm 2 Change detection (Shekhar and Ramdas, 2023)

Input : $p, \alpha \in (0, 1)$, sequence $S_{1:n}$
Output : Binary variable if input has a change-point
$\widetilde{C} \leftarrow \bigcap_{s \le n} C(p, \alpha, S_{1:s})\ \forall t \in [n]$
$\widehat{C} \leftarrow \bigcap_{s \le n} C(p, \alpha, S_{s:n})\ \forall t \in [n]$
return $\mathbf{1}(\widetilde{C} \bigcap \widehat{C} = \emptyset)$ `// 1 is a change-point`

Algorithm 3 Thresholding without offline data

Input : Quantile $p \in (0, 1)$, Confidence $\alpha \in (0, 1)$
Output : Anomaly labels $\hat{y}_1, \hat{y}_2, \ldots$
$\tau \leftarrow 0$; `// previous change point`
for *each time* $t \ge 1$ **do**
  Receive t^{th} input S_t
  if `Algorithm-2` $(p, \alpha, S_{\tau+1:t-1}) == 1$ **then**
    $\hat{y}_t \leftarrow *$ `// Change-point detected`
    $\tau \leftarrow t - 1$
  else
    $\hat{y}_t \leftarrow$ `Algorithm-1` $(S_t, C(p, \alpha, S_{\tau+1:t-1}))$
  end
end

Proposition 3 (Stream with a single change-point)**.** Let $(S_t)_{t=1}^T$ be a piece-wise stationary with one change-point ($H_T = 1$) and quntile shift $\Delta := \text{Shift}(f^{(0)}, f^{(1)}, p) > 0$, at time $\tau_1 = \tau \ge \frac{80}{\Delta^2} \ln\left(\frac{1612}{\alpha\Delta^2}\right)$. Denote by $\tau' \in \mathbb{N}$ as the first time when Algorithm 3 detects a change, i.e., the `if` statement evaluates to `True`. Then, with probability at-least $1 - \alpha T$, $\tau \le \tau' \le \tau + \frac{80}{\Delta^2} \ln\left(\frac{1612}{\alpha\Delta^2}\right)$.

Thus, if there are sufficient pre-change samples, then the change is detected with delay scaling with Δ^{-2}, which is minimax optimal (Maillard, 2019; Besson et al., 2022; Shekhar and Ramdas, 2023). As notation, we define detection delay as $D(\cdot, \cdot) : [0, 1] \times [0, 1] \to \mathbb{R}_+$ as

$$D(\Delta, \alpha) := \frac{80}{\Delta^2} \ln\left(\frac{1612}{\alpha\Delta^2}\right). \tag{3}$$

5.2. Mistake and abstain bounds

To give bounds on performance, we make a simplifying assumption that change-points are sufficiently far apart. We refer the readers to Figure 7 in the appendix for an illustration of the definitions of Assumption 5.1 and 6.1.

Assumption 5.1 (α-detectable)**.** For a given $\alpha \in (0, 1)$, the piece-wise stationary data-stream $(S_t)_{t=1}^T \sim \left(T, H_T, (\tau_c)_{c=0}^{H_T}, (f^{(c)})_{c=0}^{H_T}\right)$ is said to be α-detectable if

$$\tau_k - \tau_{k-1} \ge \begin{cases} D(\Delta_1, \alpha), & k = 1, \\ D(\Delta_{k-1}, \alpha) + D(\Delta_k, \alpha), & 2 \le k \le H_T \\ D(\Delta_{H_T}, \alpha), & k = H_T + 1 \end{cases}$$

holds, where $\Delta_k = \text{Shift}(f^{(k)}, f^{(k-1)}, p)$.

This assumption is standard to give guarantees for online algorithms with multiple change points on the stream (Besson et al., 2022; Sankararaman and Narayanaswamy, 2023; Cao et al., 2019; Liu et al., 2018). Assumption 5.1 is made for mathematical tractability and is not assumed in experiments. Analysis without Assumption 5.1 is an open problem (Section 7 (Besson et al., 2022)). The main result in this section is Theorem 4.

Theorem 4 (**main result for Algorithm 3**)**.** Let $(S_t)_{t=1}^T \sim \left(T, H_T, (\tau_c)_{c=0}^{H_T}, (f_c)_{c=0}^{H_T}\right)$ be a piece-wise stationary data-stream satisfying Assumption 5.1 for $\alpha \in (0, 1)$. Then, with probability at-least $1 - 2\alpha T$, Algorithm 3 satisfies *both*

- $\text{FP} + \text{FN} \le \sum_{k=1}^{H_T} D(\Delta_k, \alpha)$
- abstains $\le$

$$\sum_{k=1}^{H_T} \left[4\sqrt{(\tau_k - \tau_{k-1}) \ln\left(\tfrac{1612}{\alpha^2} \ln(3(\tau_k - \tau_{k-1}))\right)} + D(\Delta_k, \alpha)\right]$$

where for all $k \in [H_T]$, $\Delta_k = \text{Shift}(f^{(k)}, f^{(k-1)}, p)$ is the quantile shift defined in Definition (5.1) and $D(\cdot, \cdot)$ is defined in Equation (3).

Remark 4.1. This result shows that $\text{FP} + \text{FN} = H_T \cdot \mathcal{O}(\frac{1}{(\min_k \Delta_k)^2})$[4], scales only with the number of changes H_T, and abstains at-most $O(\sqrt{H_T T})$ [5]. Thus, desiderata 3 from Section 1.1 is achieved.

Remark 4.2. The bound in Theorem 4 holds with probability $1 - 2\alpha T$. If $(S_t)_{t=1}^T$ is α/T-detectable according to Assumption 5.1, and with knowledge of T, then Algorithm 3 run with input α/T instead of α yields the same order-wise guarantees as Theorem 4 holding with probability $1-2\alpha$. Corollary 11.1 in the Appendix shows that assuming $(S_t)_{t=1}^T$ being α/T detectable is only weaker by a logarithmic factor compared to the assumption of α-detectable.

Proof sketch of Theorem 4 The crux is an induction argument, that under Assumption 5.1, there exists exactly one true change-point between any two successive detected change-points in Algorithm 2. This is formalized in Lemma 3 in the Appendix. Lemma 3 along with Proposition 3 applied recursively to all change-points allows us to decompose the mistakes and abstains over the stationary segments.

[4] $\mathcal{O}(\cdot)$ ignores constants and poly-log terms in $\frac{T}{\alpha}$.

[5] Bound follows from Cauchy-Schwartz inequality that $\sqrt{x} + \sqrt{y} \le \sqrt{2(x + y)}, \forall x, y \ge 0$.

6. Special Case III: Stationary stream with offline data

In this section, we return to $(S_t)_{t\geq 1}$ being i.i.d., but give guarantees when the algorithm has access to offline datasets. Our results are based on an *offline* version of Assumption 5.1 that we state below.

Assumption 6.1 (Well separated offline dataset). For $p, \alpha \in (0,1)$, the offline dataset $\mathcal{D} = \{\{X_l^{(j)}\}_{l=1}^{N_j}\}_{j=1}^{K}$ is (p,α)-separated with resepct to a piece-wise stationary stream $\left(T, H_T, (\tau_c)_{c=0}^{H_T}, (f^{(c)})_{c=0}^{H_T}\right)$ if both: *(i)* for all $i \neq j \in [K]$, $3(u_{N_i}(\alpha) + u_{N_j}(\alpha)) < \text{Shift}(f_{(i)}, f_{(j)}, p)$, and *(ii)* for all $j \in [K]$ and stationary-segments $k \in [H_T]$, either $Q^{(k)}(p) = Q_{(j)}(p)$, or $3u_{N_j}(\alpha) < \text{Shift}(f^{(k)}, f_{(j)}, p)$.

Roughly, a well-separated offline dataset are those that have a sufficient number of samples so as to be distinguishable. Schematic illustrations in Section D.1 in the Appendix.

The main result of this section is how to make decision at time t which is summarized in Algorithm 4. In a nutshell, Algorithm 4 uses the offline dataset if and only if there exists *exactly one* of the K offline dataset whose CS intersects the CS constructed from $S_{1:t}$. The decision at time t is made by constructing a with the unique offline dataset (if it exists) and the online stream using Algorithm 1.

Algorithm 4 Decision making with offline data

Input : $p, \alpha \in (0,1)$, Online data $S_1, \ldots, S_n$, offline datasets $\mathcal{D}_1, \ldots, \mathcal{D}_K$.
Output : Anomaly label $\hat{y} \in \{0, 1, *\}$
$J_{\text{match}} \leftarrow \{j : C(p,\alpha,\mathcal{D}_j) \cap C(p,\alpha,S_{1:n-1}) \neq \emptyset\}$
if $|J_{match}| = 1$ **then**
  $\hat{y} \leftarrow$ `Algorithm-1` $(S_n, C(p,\alpha,\mathcal{D}_{J_{\text{match}}} \cup \{S_{1:n-1}\}))$
  `// use online and offline dataset`
else
  $\hat{y} \leftarrow$ `Algorithm-1` $(S_n, C(p,\alpha,\{S_{1:n-1}\}))$ `// use online data only`
end
return $\hat{y}$

We set some more notation to state the result. For all $j \in [K]$, $\widehat{\tau}_j = \inf\{\widehat{\tau} : 3(u_{\widehat{\tau}}(\alpha) + u_{N_j}(\alpha)) \leq \text{Shift}(f, f_{(j)}, p)\}$, where $\widehat{\tau}_j = +\infty$ if $\text{Shift}(f, f_{(j)}, p) = 0$. Assumption 6.1 guarantees that if $\text{Shift}(f, f_{(j)}, p) > 0$, then $\widehat{\tau}_j < \infty$ is finite independent of T.

$$\widehat{\tau} := \max\{\lceil \widehat{\tau}_j \rceil : \text{Shift}(f, f_{(j)}, p) > 0\}. \tag{4}$$

Theorem 5 (Main result for Algorithm 4). Let the online stream $(S_t)_{t\geq 1}$ be i.i.d. from distribution f and the $j \in [K]$ offline dataset with N_j i.i.d. samples have distribution $f_{(j)}$. Further, let the offline dataset be (p,α) well-separated for some $\alpha \in (0,1)$ (Def. 6.1). Let $\Delta = \min_{1\leq j\leq K} \text{Shift}(f, f_{(j)}, p)$. Then,for all times $T \in \mathbb{N}$, with probability at-least $1 - (K+T)\alpha$, if the decision $\hat{y}_t$ is made by Algorithm 4 with inputs $p, \alpha, S_{1:t}$ and $\mathcal{D}_{1:K}$, we have:

- If $\Delta = 0$, then FP + FN $= 0$ and Abstains $\leq B(N+T,\alpha) - \frac{B(N,\alpha)}{28} + \widehat{\tau}$,
- If $\Delta > 0$, then FP + FN $\leq \widehat{\tau}$ and Abstains $\leq \widehat{\tau} + B(T,\alpha)$,

where $N = \min\{N_j : \text{Shift}(f, f_{(j)}, p) = 0\}$, $B(\cdot,\cdot) : \mathbb{N} \times [0,1] \to \mathbb{R}_+$ is given by $B(Y,\alpha) = 7\sqrt{Y \ln\left(\frac{1612 \ln(3Y)}{\alpha^2}\right)}$ and $\widehat{\tau}$ is defined in Equation (4).

We now read off several remarks from this result.

Improved performance if offline dataset is useful. The case of $\Delta = 0$ implies that the online stream's pth quantile equals at-least one of the K offline dataset's pth quantile, i.e., the offline dataset is useful. In this case, Theorem 5 guarantees 0 FP and FN, and abstains at-most $\mathcal{O}(\sqrt{N+T} - \sqrt{N})$, which is much smaller compared to the $\mathcal{O}(\sqrt{T})$ bound from Algorithm 6 without offline data.

Bounded degradation if offline dataset is arbitrary. The case of $\Delta > 0$ implies that the online stream's pth quantile does not match any of the K offline dataset's pth quantile, i.e., the offline dataset is useless. In this case, Theorem 5, guarantees FP + FN $\leq \widehat{\tau}$ and abstains $\leq \mathcal{O}(\sqrt{T}) + \widehat{\tau}$. Thus performance compared to Theorem 2 is worse *only by an additive term independent of the stream length* T, thereby achieving desiderata 4 from Section 1.1. Equation (3) implies that $\widehat{\tau}$ is a sample complexity upper bound to confidently reject the hypothesis that one of the K offline dataset's pth quantile equals that of f.

Thus, Algorithm 4 adapts to the complexity yielding much lower abstains in the case when the online distribution matches one of the offline dataset, while simultaneously having bounded degradation *in the worst-case*.

Proof Sketch. The crux of the proof is that, if the online stream's quantile matches that of one of the offline dataset, then *(i)* for all times, $J_{\text{match}} \geq 1$ in Algorithm 4, and *(ii) for all times* $t \geq \widehat{\tau}$, $J_{\text{match}} = 1$. Similarly, if no offline dataset's pth quantile matches that of the online pth quantile, for all times $t \geq \widehat{\tau}$, $J_{\text{match}} = 0$ and thus Algorithm 4 will only use the online stream's samples for decision making.

7. The General Case

This setting generalizes all the special cases. For each offline dataset $j \in [K]$ and stationary segment of the online stream $k \in [H_T]$, denote by $\Delta^{(j;k)} = \text{Shift}(f_{(j)}, f^{(k)}, p)$, where $f_{(j)}$ is the distribution of offline dataset j and $f^{(k)}$ the distribution of the kth online segment (Def. 2.1). Identical to Theorem 4, for each $k \in [H_T]$, $\Delta_k = \text{Shift}(f^{(k)}, f^{(k-1)}, p)$ is

the shift between the kth and $k-1$th online segment. Similar to Theorem 5, for each $k \in [H_T]$ and offline dataset $j \in [K]$, $\widehat{\tau}_j^{(k)} = \inf\{\widehat{\tau} : 3(u_{\widehat{\tau}}(\alpha) + u_{N_j}(\alpha)) \leq \text{Shift}(f^{(k)}, f_{(j)}, p)\}$, where $\widehat{\tau}_j^{(k)} = \infty$ if $\text{Shift}(f^{(k)}, f_{(j)}, p) = 0$. Assumption 6.1 guarantees $\text{Shift}(f^{(k)}, f_{(j)}, p) > 0 \implies \widehat{\tau}_j^{(k)} < \infty$. Similar to Eq. (4), $\widehat{\tau}^{(k)} = \max\{\lceil \widehat{\tau}_j^{(k)} \rceil : \widehat{\tau}_j^{(k)} < \infty\}$.

$$\widetilde{D}_k(\alpha) = D(\Delta_k, \alpha) + \widehat{\tau}^{(k)}. \tag{5}$$

In words, $\widetilde{D}_k(\alpha)$ is the *worst-case* delay after the kth change point has occurred on the data-stream for Algorithm 5 to both *(i)* detect that the online stream has undergone a shift, and *(ii)* identify if any of the offline datasets' pth quantile matches the new pth quantile.

Algorithm 5 General thresholding algorithm

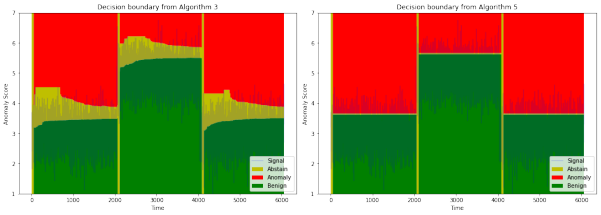
Input : $p, \alpha \in (0,1)$, offline datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$.
Output : Anomaly labels $\hat{y}_1, \hat{y}_2, \ldots$
$\tau \leftarrow 0$; `// previous change point`
for *each time* $t \geq 1$ **do**
  Receive t^{th} input S_t
  if `Algorithm-2` $(p, \alpha, S_{\tau+1:t-1}) == 1$ **then**
    $\hat{y}_t \leftarrow *$ `// Change-point detected`
    $\tau \leftarrow t - 1$
  else
    $\hat{y}_t \leftarrow$ `Algorithm-4`$(p, \alpha, S_{\tau+1:t-1}, (\mathcal{D}_{1:K}))$
  end
end

Theorem 6 (Main result of Algorithm 5). For a desired $\alpha \in (0,1)$, suppose the piece-wise stationary online stream $(S_t)_{t=1}^T \sim \left(T, H_T, (\tau_c)_{c=0}^{H_T}, (f_c)_{c=0}^{H_T}\right)$ is α-detectable, i.e., satisfies Assumption 5.1, and the offline datasets $\mathcal{D}_1, \ldots, \mathcal{D}_K$ are (p, α)-separated according to Definition 6.1. Then, for all times $T \in \mathbb{N}$, with probability at-least $1 - (K + 2T)\alpha$, Algorithm 5 satisfies *both*,

- $\text{FP} + \text{FN} \leq \sum_{k=1}^{H_T} D(\Delta_k, \alpha) + (1 - \mathbf{1}_{\text{Match}_k}) \max_{j \in [K]} \left(\lceil \widehat{\tau}_j^{(k)} \rceil \mathbf{1}_{\text{Shift}(f^{(k)}, f_{(j)}, p) > 0} \right)$
- $\text{Abstains} \leq \sum_{k=1}^{H_T} \Big[(1 - \mathbf{1}_{\text{Match}_k}) \mathcal{O}(\sqrt{\tau_k - \tau_{k-1}})^6 + \mathbf{1}_{\text{Match}_k} \left(\mathcal{O}(\sqrt{N^{(k)} + (\tau_k - \tau_{k-1})} - \sqrt{N^{(k)}}) \right) + \widetilde{D}_k(\alpha) \Big]$

where $\mathbf{1}_{\text{Match}_k} = \mathbf{1}_{Q^{(k)}(p) \in \{Q_{(1)}(p), \cdots, Q_{(K)}(p)\}}$, and $Q^{(k)}(p)$ is the pth quantile of the kth segment of the online stream (Def. 2.1), $Q_{(1)}(p), \cdots, Q_{(K)}(p)$ are the pth quantiles of the offline datasets (Definition 2.3), $N^{(k)} = \min\{N_j : Q_{(j)}(p) = Q^{(k)}(p)\}$ and $\widetilde{D}_k(\alpha)$ is defined in Equation (5).

Observe that Theorem 6 recovers Theorem 5 if $(S_t)_{t \geq 1}$ is i.i.d. Theorem 6 shows that even in the general case, FP + FN only scale with the number and complexity of change-points, and not with the stream length T.



(a) (b)

Figure 3: Decision boundaries for the same synthetic data stream, without offline dataset (a) and with (b). Incorporating datasets allows us to significantly reduce abstention.

Abstains reduced if offline data is useful. Suppose the online stream $(S_t)_{t=1}^T$ and the offline datasets $\mathcal{D}_1, \ldots, \mathcal{D}_K$ are such that for all segments $k \in [H_T]$, there exists an offline dataset $j \in [K]$ s.t. $Q^{(k)}(p) = Q_{(j)}(p)$. Theorem 6 guarantees that abstains $\leq \mathcal{O}\left(\sqrt{H_T N}\left(\sqrt{1 + \frac{T}{N}} - 1\right)\right) + \sum_{k=1}^{H_T} \widetilde{D}_k(\alpha)$, where $N = \min\{N_j : \exists k \in [H_T] : Q_{(j)}(p) = Q^{(k)}(p)\}$ is size of the smallest useful offline dataset. Thus, if the useful offline data is large ($N >> T$), the fraction of abstains is small even in the presence of non-stationarity (Fig. 3).

Bounded degradation if offline data is arbitrary. Theorem 6 proves that in the worst-case when the offline datasets are arbitrary, abstains are at-most $\mathcal{O}(\sqrt{H_T T}) + \sum_{k=1}^{H_T} \widetilde{D}_k(\alpha)$, which is comparable to Theorem 4, with an additional additive term of $\sum_{k=1}^{H_T} \widetilde{D}_k(\alpha)$. This extra term arises due to incurring delay in both identifying a change-point on the online stream, and deciding that either the new segment has the same pth quantile as one of the offline datasets or not. This penalty is incurred since Algorithm 5 does not know both the time of change-points and if the pth quantile of any offline datasets equals that of the online. Thus, Theorem 6 is adaptive, yielding low mistakes and abstains in the event that the offline dataset is useful while simultaneously bounding worst-case performance independent of T and N when the offline dataset is arbitrary.

8. Experiments

8.1. Synthetic data experiments

We evaluated our algorithm using two synthetic datasets, reflecting all scenarios discussed in this paper. The metrics are abstention percentages (Abs. %) and mistake counts (FP + FN), averaged over 1000 streams each with 2000 samples drawn from Normal distributions with random parameters, with $p = 1 - 10^{-2}$ and $\alpha = 10^{-3}$. Our comparisons against common thresholding baselines: $\tau^{30\%}$ for static thresholding, `DSpot` for dynamic thresholding, and `EQ` for empirical quantiles using online gradient descent, are presented in Table 1. Implementation details and additional experiments on Pareto distributions are in Appendix G. These results

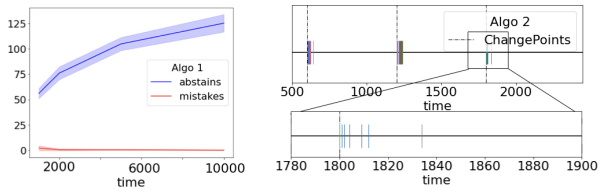


Figure 4: *Left:* algorithm 1, Abstains grows at $\mathcal{O}(\sqrt{T})$. *Right:* algorithm 2, timestamp of mistakes in 10 random streams with aligned change points. Mistakes cluster right after change points, only as detection delay.

confirm our algorithm's high accuracy, adaptability to distribution shifts, and effective learning from offline data, as outlined in Section 1's desiderata. Figure 4 illustrates how our algorithm's abstention rates and errors grow over time, corroborating our theoretical results.

shift	data		Ours	$\tau^{30\%}$	DSpot	EQ
x	x	Abs. %	12.1 ± 1.4	30	15	0
		FP+FN	**0** ± 0	5.3 ± 3.7	9.1 ± 3.6	3.9 ± 2.1
✓	x	Abs. %	32.7 ± 9.1	30	15	0
		FP+FN	**5.2** ± 1.3	195 ± 71	225 ± 95	219 ± 71
x	✓	Abs. %	9.3 ± 1.2	30	15	0
		FP+FN	**0** ± 0	5.3 ± 3.7	9.1 ± 3.6	3.9 ± 2.1
✓	✓	Abs. %	18.9 ± 4.1	30	15	0
		FP+FN	**2.0** ± 1.5	155 ± 69	173 ± 64	160 ± 64

Table 1: Synthetic dataset results. We used Algorithm 1, 3, 4, 5 as "ours" for the four settings respectively. Compared to baselines, we achieve significant less mistakes (FP+FN) with low abstain rate, especially in settings with shift.

8.2. MNIST anomaly detection

We tested our algorithms on the MNIST dataset for one-class anomaly detection to demonstrate their real-world efficacy. We designated even digits as normal and odd digits as anomalous, and created data streams by sampling (e.g. Fig 5). To show synergy of our thresholding algorithm with standard anomaly detection algorithms, we obtain anomaly scores from two models: convolutional autoencoder neural networks (NN) and isolation forests (IF), both trained on the normal class. With parameters set at $p = 0.99$ and $\alpha = 0.01$ for streams of 1000 samples, Table 2 illustrates our algorithm's improved performance over baselines.

8.3. Case study of real AD application

We also perform a case study on two real world datasets (`DS1` and `DS2`) obtained from large scale cloud computing services. Each dataset is a stream of anomaly scores obtained by applying the same black box anomaly detection algorithm (Details in Table 4). The practically motivated target in these datasets is to report 0.001% anomalies, i.e., $p = 1 - 10^{-6}$. Table 5 in the appendix reports the full result

Static	abstains	FP+FN	Shift	abstains	FP+FN
$\tau^{30\%}$	327 ± 0	12.7 ± 0.9	$\tau^{30\%}$	670 ± 0	91 ± 8.0
DSpot	150 ± 0	78.1 ± 2.5	DSpot	150 ± 0	129.5 ± 10.2
IF+A1	172 ± 16	17.6 ± 5.3	IF+A3	190 ± 39	75.7 ± 18.9
NN+A1	133 ± 4	**3.9** ± 0.7	NN+A3	339 ± 12	**9.3** ± 2.1
NN+A5	89 ± 6	**2.1** ± 0.2	NN+A5	210 ± 8	**8.8** ± 2.3

Table 2: One class MNIST result. We applied our algorithms (A as shorthand) to anomaly scores generated by Isolation Forest (IF) and neural networks (NN) and achieve much lower mistakes with moderate number of abstains.

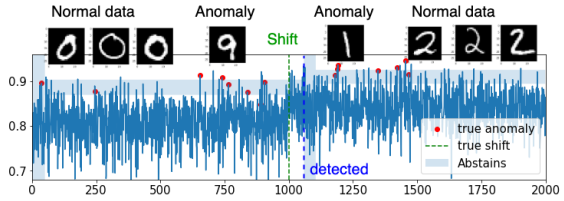


Figure 5: MNIST experiment setup and a sample stream of (NN+A5) result. We choose a different set of digits for normal and anomaly to create the distribution shift, which our algorithm detects with a small delay.

and comparison to baselines. We found that *(i)* our method achieves desired anomaly volume, *(ii)* incorporating offline data reduces the number of abstains without changing the volume of anomalies raised, and *(iii)* the volume of anomalies detected by all other methods are an order of magnitude higher leading to alert-fatigue.

9. Related Work

Dynamic thresholding for AD. AD is an extensively researched topic, where a common paradigm is to first model data likelihood and then use a threshold to decide if a data point is normal or anomalous. This threshold is typically determined by an expert or offline through cross-validation (Luo et al., 2021; Schmidl et al., 2022). Dynamic and automatic methods makes AD algorithms more practical in real-world deployment (Ali et al., 2013; Hundman et al., 2018), and are necessary where the data distribution can shift (Siffer et al., 2017). Our work is the first to provide abstain and mistake bounds on dynamic thresholding without assumptions on the underlying data distribution. Appendix F contains more related work.

Confidence Sequences (CS). We incorporate theoretical results of confidence sequences (Ramdas et al., 2022) in our analysis. (Maharaj et al., 2023; Howard and Ramdas, 2022) derived the CS algorithm for estimating quantiles online with any-time guarantees, and (Shekhar and Ramdas, 2023) leverages CS for detecting distribution shifts.

10. Discussion

Our paper is the first to systematically theoretically and empirically show that thresholding algorithm can improve the overall performance of any anomaly scoring model for online anomaly detection. Through a simple argument (Theorem 15), we demonstrate that abstaining is necessary for high accuracy.

Our work introduces a new approach to online adaptive anomaly thresholding, leveraging CS to utilize offline data and dynamically adjust thresholds in real-time data streams. We give the first statistical guarantees on abstains and mistakes in various settings with non-stationary data streams and offline datasets (Theorem 2 - 6), motivated by practical AD applications. Our results are corroborated with synthetic and real experiments.

A limitation of our work is that the theoretical results rely on (1) samples are temporally independent and (2) change points are far apart and changes are detectable. These assumptions, although standard in theoretical literature, need not necessarily hold in all applications. Our algorithms also do not apply to cases when scores are temporal dependent, or if there are gradual drifts in the score distribution. Designing algorithms that provably work for these cases likely requires new techniques and is thus left for future work.

Impact Statement

This paper presents work whose goal is to advance analysis and performance of online anomaly detection, a popular use case of machine learning. There will be application-specific potential societal consequences of our work when applied, none which we feel must be specifically highlighted here.

References

Ali, M. Q., Al-Shaer, E., Khan, H., and Khayam, S. A. (2013). Automated anomaly detector adaptation using adaptive threshold tuning. *ACM Transactions on Information and System Security (TISSEC)*, 15(4):1–30.

Audibert, J., Michiardi, P., Guyard, F., Marti, S., and Zuluaga, M. A. (2020). Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3395–3404.

Balcan, M.-F., Broder, A., and Zhang, T. (2007). Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer.

Besson, L., Kaufmann, E., Maillard, O.-A., and Seznec, J. (2022). Efficient change-point detection for tackling piecewise-stationary bandits. *Journal of Machine Learning Research*, 23(77):1–40.

Bhatia, S., Jain, A., Li, P., Kumar, R., and Hooi, B. (2021). Mstream: Fast anomaly detection in multi-aspect streams. In *Proceedings of the Web Conference 2021*, pages 3371–3382.

Bhatia, S., Jain, A., Srivastava, S., Kawaguchi, K., and Hooi, B. (2022). Memstream: Memory-based streaming anomaly detection. In *Proceedings of the ACM Web Conference 2022*, pages 610–621.

Cadre, B., Pelletier, B., and Pudlo, P. (2013). Estimation of density level sets with a given probability content. *Journal of nonparametric statistics*, 25(1):261–272.

Cao, Y., Wen, Z., Kveton, B., and Xie, Y. (2019). Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 418–427. PMLR.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58.

Chen, Y. (2017). Draining the Flood—A combat against alert fatigue. USENIX Association.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37.

Gan, E. and Bailis, P. (2017). Scalable kernel density classification via threshold-based pruning. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 945–959.

Goyal, S., Raghunathan, A., Jain, M., Simhadri, H. V., and Jain, P. (2020). Drocc: Deep robust one-class classification. In *International conference on machine learning*, pages 3711–3721. PMLR.

Gutflaish, E., Kontorovich, A., Sabato, S., Biller, O., and Sofer, O. (2019). Temporal anomaly detection: calibrating the surprise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3755–3762.

Hassan, W. U., Guo, S., Li, D., Chen, Z., Jee, K., Li, Z., and Bates, A. (2019). Nodoze: Combatting threat alert fatigue with automated provenance triage. In *network and distributed systems security symposium*.

He, Z., Hu, G., and Lee, R. B. (2023). Cloudshield: Real-time anomaly detection in the cloud. In *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy*, pages 91–102.

Herley, C. (2022). Automated detection of automated traffic. In *31st USENIX Security Symposium (USENIX Security*

22), pages 1615–1632, Boston, MA. USENIX Association.

Hill, D. J. and Minsker, B. S. (2010). Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling & Software*, 25(9):1014–1022.

Ho, G., Dhiman, M., Akhawe, D., Paxson, V., Savage, S., Voelker, G. M., and Wagner, D. (2021). Hopper: Modeling and detecting lateral movement. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 3093–3110.

Ho, G., Sharma, A., Javed, M., Paxson, V., and Wagner, D. (2017a). Detecting credential spearphishing in enterprise settings. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 469–485, Vancouver, BC. USENIX Association.

Ho, G., Sharma, A., Javed, M., Paxson, V., and Wagner, D. (2017b). Detecting credential spearphishing in enterprise settings. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 469–485.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Howard, S. R. and Ramdas, A. (2022). Sequential estimation of quantiles with applications to a/b testing and best-arm identification. *Bernoulli*, 28(3):1704–1728.

Huang, H. and Kasiviswanathan, S. P. (2015). Streaming anomaly detection using randomized matrix sketching. *Proc. VLDB Endow.*, 9(3):192–203.

Huang, T., Chen, P., and Li, R. (2022). A semi-supervised vae based active anomaly detection framework in multivariate time series for online systems. In *Proceedings of the ACM Web Conference 2022*, pages 1797–1806.

Hundman, K., Constantinou, V., Laporte, C., Colwell, I., and Soderstrom, T. (2018). Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 387–395.

Katz, Y. and Raz, D. (2023). Cold start for cloud anomaly detection. In *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, pages 1–8. IEEE.

Khan, W. Z., Rehman, M., Zangoti, H. M., Afzal, M. K., Armi, N., and Salah, K. (2020). Industrial internet of things: Recent advances, enabling technologies and open challenges. *Computers & electrical engineering*, 81:106522.

Lazarevic, A., Ertoz, L., Kumar, V., Ozgur, A., and Srivastava, J. (2003). A comparative study of anomaly detection schemes in network intrusion detection. In *Proceedings of the 2003 SIAM international conference on data mining*, pages 25–36. SIAM.

Li, J., Di, S., Shen, Y., and Chen, L. (2021). Fluxev: a fast and effective unsupervised framework for time-series anomaly detection. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 824–832.

Lin, Y., Chen, Z., Cao, C., Tang, L.-A., Zhang, K., Cheng, W., and Li, Z. (2018). Collaborative alert ranking for anomaly detection. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1987–1995.

Liu, F., Lee, J., and Shroff, N. (2018). A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Luo, Y., Xiao, Y., Cheng, L., Peng, G., and Yao, D. (2021). Deep learning-based anomaly detection in cyber-physical systems: Progress and opportunities. *ACM Computing Surveys (CSUR)*, 54(5):1–36.

Ma, M., Zhang, S., Pei, D., Huang, X., and Dai, H. (2018). Robust and rapid adaption for concept drift in software system anomaly detection. In *2018 IEEE 29th International Symposium on Software Reliability Engineering (ISSRE)*, pages 13–24. IEEE.

Maharaj, A., Sinha, R., Arbour, D., Waudby-Smith, I., Liu, S. Z., Sinha, M., Addanki, R., Ramdas, A., Garg, M., and Swaminathan, V. (2023). Anytime-valid confidence sequences in an enterprise a/b testing platform. In *Companion Proceedings of the ACM Web Conference 2023*, pages 396–400.

Maillard, O.-A. (2019). Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds. In *Algorithmic Learning Theory*, pages 610–632. PMLR.

Mirsky, Y., Doitshman, T., Elovici, Y., and Shabtai, A. (2018). Kitsune: an ensemble of autoencoders for online network intrusion detection. *arXiv preprint arXiv:1802.09089*.

Nardi, M., Valerio, L., and Passarella, A. (2022). Anomaly detection through unsupervised federated learning. In *2022 18th International Conference on Mobility, Sensing and Networking (MSN)*, pages 495–501. IEEE.

Pang, G., Shen, C., Cao, L., and Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38.

Perini, L. and Davis, J. (2023). Unsupervised anomaly detection with rejection. *arXiv preprint arXiv:2305.13189*.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2022). Game-theoretic statistics and safe anytime-valid inference. *arXiv preprint arXiv:2210.01948*.

Ren, H., Xu, B., Wang, Y., Yi, C., Huang, C., Kou, X., Xing, T., Yang, M., Tong, J., and Zhang, Q. (2019). Time-series anomaly detection service at microsoft. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3009–3017.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., and Müller, K.-R. (2021). A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795.

Sadik, S. and Gruenwald, L. (2014). Research issues in outlier detection for data streams. *Acm Sigkdd Explorations Newsletter*, 15(1):33–40.

Sankararaman, A. and Narayanaswamy, B. (2023). Online heavy-tailed change-point detection. In *Uncertainty in Artificial Intelligence*, pages 1815–1826. PMLR.

Sankararaman, A., Narayanaswamy, B., Singh, V. Y., and Song, Z. (2022). Fitness:(fine tune on new and similar samples) to detect anomalies in streams with drift and outliers. In *International Conference on Machine Learning*, pages 19153–19177. PMLR.

Schmidl, S., Wenig, P., and Papenbrock, T. (2022). Anomaly detection in time series: a comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15(9):1779–1797.

Shekhar, S. and Ramdas, A. (2023). Sequential change detection via backward confidence sequences. *arXiv preprint arXiv:2302.02544*.

Siffer, A., Fouque, P.-A., Termier, A., and Largouet, C. (2017). Anomaly detection in streams with extreme value theory. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1067–1075.

Steinwart, I., Hush, D., and Scovel, C. (2005). A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6(2).

Vishwakarma, H., Lin, H., Sala, F., and Vinayak, R. K. (2023). Promises and pitfalls of threshold-based auto-labeling. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

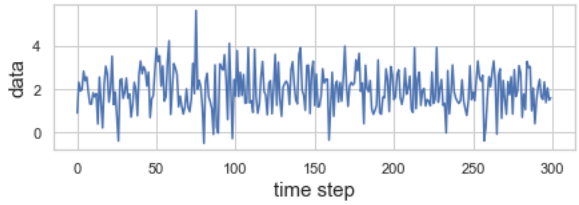
Zhang, X., Yang, T., and Srinivasan, P. (2016). Online asymmetric active learning with imbalanced data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2055–2064.

Zhu, J., Cai, S., Deng, F., Ooi, B. C., and Zhang, W. (2023). Meter: A dynamic concept adaptation framework for online anomaly detection. *arXiv preprint arXiv:2312.16831*.

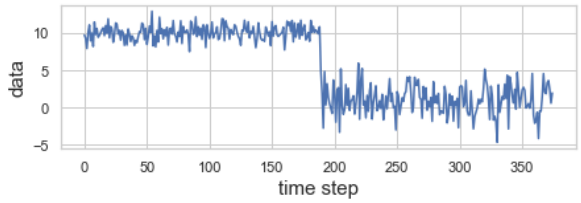
Supplementary Materials

A. Illustrations

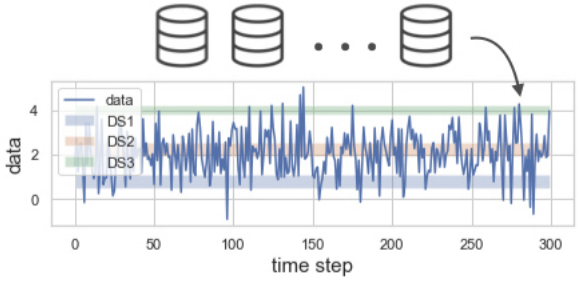
We give schematic of the various settings and special cases studied in this paper.



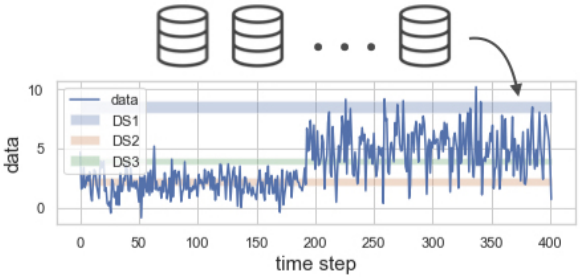
(a) Special Case I: stationary stream



(b) Special Case II: distribution shift



(c) Special Case III: offline dataset



(d) The General Case

Figure 6: Illustration of the cases we study in this paper.

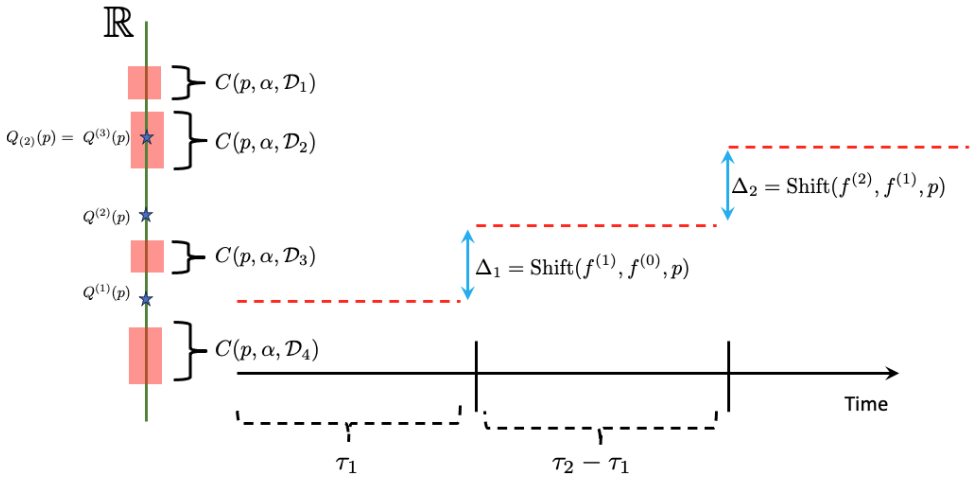


Figure 7: An illustration of Assumption 5.1 and 6.1 on a stream with 2 change-points ($H_T = 2$) and $K = 4$ offline datasets. Assumption 5.1 implies that the change-points on the stream are sufficiently far apart. Assumption 6.1 implies that the 4 CS do not intersect. Moreover, either the true quantile on the stream are outside the offline CS, i.e., $Q^{(1)}(p)$ and $Q^{(2)}(p)$ are not contained in any offline dataset's CS. Or, it is the case that the online quantile matches one of the offline dataset, i.e., $Q_{(2)}(p) = Q^{(3)}(p)$ in this example.

B. Proof of Theorem 2, stationary setting

Algorithm 6 OAD for stationary i.i.d. stream

Input : Quantile p, Confidence $1-\alpha$
Output : Anomaly labels $\hat{y}_1, \hat{y}_2, \ldots$
for *each time* $t \geq 1$ **do**
  Receive t^{th} input S_t
  Compute $C(p, \alpha, S_{1:t})$ (Eq. 1)
  $\hat{y}_t \leftarrow$ `Algorithm 1`$(S_t, C(p, \alpha, S_{1:t}))$
end

Theorem 7 (Formal version of Theorem 2)**.** If the data-source $S_1, S_2, \cdots$ are i.i.d., then with probability at-least $1-2\alpha$, for all times $T \in \mathbb{N}$, Algorithm 6 satisfies

1. 0 False positives, 0 False negatives
2. Abstains is less than or equal to $7\sqrt{T \ln\left(\frac{1612 \ln(eT)}{\alpha^2}\right)}$,
3. Abstains is at-least $\frac{1}{4}\sqrt{T \ln\left(\frac{1612 \ln(eT)}{\alpha^2}\right)}$

Proof. In order to give the proof, we set some notations.

Definition B.1 (Good event $\mathcal{E}_1$)**.**

$$\mathcal{E}_1 := \bigcap_{t \in \mathbb{N}} \bigcap_{p \in (0,1)} \{\widehat{Q}(p - u_t(\alpha), S_{1:t}) \leq Q(p) \leq \widehat{Q}(p + u_t(\alpha), S_{1:t})\}. \tag{6}$$

In words, the good event states that for all time t, the confidence sequence contains the true pth quantile. From Corollary 2 of (Howard and Ramdas, 2022), we know that $\mathbb{P}[\mathcal{E}] \geq 1-\alpha$. We remark that (Howard and Ramdas, 2022) show that even when $(S_t)_{t \geq 1}$ are independent, but not necessarily identical, $\mathbb{P}[\mathcal{E}] \geq 1-\alpha$ holds as long as the pth quantile $Q_t(p) := Q(p)$ for all $t \geq 1$. This extension relaxing the identical distribution will be crucial in the subsequent sections in the sequel.

Further, observe from the definition in Algorithm 6 and Equation (1) that for all $t \in \mathbb{N}$ and $p \in (0,1)$,

$$\{\widehat{Q}(p - u_t(\alpha), S_{1:t}) \leq Q(p) \leq \widehat{Q}(p + u_t(\alpha), S_{1:t})\} \subseteq \{Q(p) \in C(p, \alpha, S_{1:t}\}.$$

The reason for this inclusion is that the definition of $C(\cdot)$ in Equation (1) has an extra factor of 2, i.e.,

$$C(p, \alpha, S_{1:t}) := \left[\widehat{Q}(p - 2u_n(\alpha), y_{1:n}), \widehat{Q}(p + 2u_n(\alpha), y_{1:n})\right].$$

In the rest of this proof, we will assume that the good event $\mathcal{E}$ holds. It is immediately clear the following two claims hold, proving the first two statements of the theorem.

Claim 1. If event $\mathcal{E}$ holds, then Algorithm 1 will not make any False Positive or False negative detections.

It now only remains to prove the third condition of the theorem. Define the sequence of indicator random variables indicating if time t is an abstain

$$Z_t := \mathbf{1}(\widehat{y}_t = *). \tag{7}$$

From Algorithm 6, this is equivalent to

$$Z_t = \mathbf{1}(S_t \in (\widehat{Q}_t(p - 2u_{t-1}(\alpha), S_{1:t-1}), \widehat{Q}_t(p + 2u_{t-1}(\alpha), S_{1:t-1})).$$

We denote by the filtration $(\mathcal{F}_t)_{t=1}^T$ to be generated by the observed sequence $(S_t)_{t=1}^T$, i.e., for all $t \in \{1, \cdots, T\}$, $\mathcal{F}_t := \sigma(S_1, \cdots, S_t)$.

The proof of claim 1 follows from the following three lemmas. For clarity, proof for these lemmas can be found after this proof finishes.

Lemma 1 (Conditional expectation of abstains)**.** For every $t \geq 2$,

$$4u_{t-1}(\alpha) \leq \mathbb{E}[Z_t \mathbf{1}_{\mathcal{E}} | \mathcal{F}_{t-1}] \leq 6u_{t-1}(\alpha),$$

holds almost-surely.

Lemma 2 (Azuma-Hoefding bound)**.** Let $(Z_t)_{t=1}^T$ are bounded binomial variables and the sequence $(\mathcal{F}_t)_{t=1}^T$ is a probability filtration. Then for all $\varepsilon > 0$,

$$\mathbb{P}\left[\left|\sum_{s=1}^T Z_s - \mathbb{E}[Z_s|\mathcal{F}_{s-1}]\right| \geq \varepsilon\right] \leq 2\exp\left(-\frac{\varepsilon^2}{2T}\right)$$

We now apply Lemma 2 to the sequence of binary random variables $(\widetilde{Z}_t)_{t\geq 1} := (Z_t \mathbf{1}_{\mathcal{E}})_{t\geq 1}$ by setting $\varepsilon = \sqrt{T \ln\left(\frac{4}{\alpha^2}\right)}$. Thus, with probability at least $1-\alpha$, we have

$$\mathbb{P}\left[\underbrace{\left|\mathbf{1}_{\mathcal{E}}\sum_{s=1}^T Z_s - \sum_{s=1}^T \mathbb{E}[Z_s \mathbf{1}_{\mathcal{E}}|\mathcal{F}_{s-1}]\right| \leq \sqrt{T\ln\left(\frac{4}{\alpha^2}\right)}}_{\text{Event } \mathcal{G}_1}\right] \geq 1-\alpha$$

From a simple union-bound argument, we see that $\mathbb{P}[\mathcal{E} \cap \mathcal{G}_1] \geq 1 - 2\alpha$. Furthermore by definition of event $\mathcal{G}_1$, the following inequality holds almost-surely.

$$\begin{aligned}
\mathbf{1}_{\mathcal{G}}\mathbf{1}_{\mathcal{E}_1}\sum_{s=1}^T Z_s &\leq \mathbf{1}_{\mathcal{G}_1}\left(\sum_{s=1}^T \mathbb{E}[Z_s\mathbf{1}_{\mathcal{E}}|\mathcal{F}_{s-1}] + \sqrt{T\ln\left(\frac{4}{\alpha^2}\right)}\right), \\
&\overset{\text{Lem 1}}{\leq} \mathbf{1}_{\mathcal{G}_1}\left(\sum_{t=1}^T 6u_{t-1}(\alpha) + \sqrt{T\ln\left(\frac{4}{\alpha^2}\right)}\right), \\
&\leq \mathbf{1}_{\mathcal{G}_1}\left(6\sqrt{T\ln\left(\frac{1612\ln(eT)}{\alpha}\right)} + \sqrt{T\ln\left(\frac{4}{\alpha^2}\right)}\right), \\
&\leq \mathbf{1}_{\mathcal{G}_1}\left(7\sqrt{T\ln\left(\frac{1612\ln(eT)}{\alpha^2}\right)}\right).
\end{aligned}$$

In other words, the preceding display reads that on the event $\mathcal{E} \cap \mathcal{G}_1$ holds, we have

$$\sum_{s=1}^T Z_s \leq 7\sqrt{T\ln\left(\frac{1612\ln(eT)}{\alpha^2}\right)}.$$

Similarly, from the definition of $\mathcal{G}_1$, we also have

$$\begin{aligned}
\mathbf{1}_{\mathcal{G}}\mathbf{1}_{\mathcal{E}_1}\sum_{s=1}^T Z_s &\geq \mathbf{1}_{\mathcal{G}_1}\left(\sum_{s=1}^T \mathbb{E}[Z_s\mathbf{1}_{\mathcal{E}}|\mathcal{F}_{s-1}] - \sqrt{T\ln\left(\frac{4}{\alpha^2}\right)}\right), \\
&\overset{\text{Lem 1}}{\geq} \mathbf{1}_{\mathcal{G}_1}\left(\sum_{t=1}^T 2u_{t-1}(\alpha) - \sqrt{T\ln\left(\frac{4}{\alpha^2}\right)}\right),
\end{aligned}$$

$$\geq \mathbf{1}_{\mathcal{G}_1}\left(2\sqrt{T\ln\left(\frac{1612\ln(eT)}{\alpha}\right)} - \sqrt{T\ln\left(\frac{4}{\alpha^2}\right)}\right),$$
$$\geq \mathbf{1}_{\mathcal{G}_1}\left(\frac{1}{4}\sqrt{T\ln\left(\frac{1612\ln(eT)}{\alpha^2}\right)}\right).$$

The proof is complete since $\mathbb{P}[\mathcal{E}\cap\mathcal{G}_1]\geq 1-2\alpha$. □

Below we provide the proof of each lemma used in proof of Theorem 2.

Proof of Lemma 1. We can compute the conditional expectation as

$$\begin{aligned}\mathbb{E}[Z_t\mathbf{1}_{\mathcal{E}}|\mathcal{F}_{t-1}] &= \mathbb{P}[S_t\in(\widehat{Q}(p-2u_{t-1}(\alpha),S_{1:t-1}),\widehat{Q}(p+2u_{t-1}(\alpha),S_{1:t-1}))\cap\mathcal{E}|\mathcal{F}_{t-1}],\\ &\overset{(a)}{\leq}\mathbb{P}[S_t\in(Q(p-3u_{t-1}(\alpha)),Q(p+3u_{t-1}(\alpha))|\mathcal{F}_{t-1}]\\ &\overset{(b)}{\leq}\mathbb{P}[S_t\in(Q(p-3u_{t-1}(\alpha)),Q(p+3u_{t-1}(\alpha))]\\ &\overset{(c)}{\leq} 6u_{t-1}(\alpha).\end{aligned}$$

Step (a) follows from the fact that on event $\mathcal{E}$, for all times $t\in\mathbb{N}$ and quantiles $p\in(0,1)$, we have $Q(p)\leq\widehat{Q}(p+u_t(\alpha),S_{1:t})$ and $Q(p)\geq\widehat{Q}(p-u_t(\alpha),S_{1:t-1})$ holds. Step (b) follows since $(S_t)_{t\geq1}$ is an i.i.d. sequence and thus S_t is independent of $\mathcal{F}_{t-1}$. Step (c) follows from the standard fact the for any $\mathbb{R}$ valued continuous random variable, and for any quantiles $0\leq p_1\leq p_2\leq 1$, $\mathbb{P}[X\in(Q(p_1),Q(p_2))]=p_2-p_1$. Similarly, we have

$$\begin{aligned}\mathbb{E}[Z_t\mathbf{1}_{\mathcal{E}}|\mathcal{F}_{t-1}] &= \mathbb{P}[S_t\in(\widehat{Q}(p-2u_{t-1}(\alpha),S_{1:t-1}),\widehat{Q}(p+2u_{t-1}(\alpha),S_{1:t-1}))\cap\mathcal{E}|\mathcal{F}_{t-1}],\\ &\overset{(a)}{\geq}\mathbb{P}[S_t\in(Q(p-u_{t-1}(\alpha)),Q(p+u_{t-1}(\alpha))|\mathcal{F}_{t-1}]\\ &\overset{(b)}{\leq}\mathbb{P}[S_t\in(Q(p-u_{t-1}(\alpha)),Q(p+u_{t-1}(\alpha))]\\ &\overset{(c)}{\leq} 2u_{t-1}(\alpha).\end{aligned}$$

follows from the fact that on event $\mathcal{E}$, for all times $t\in\mathbb{N}$ and quantiles $p\in(0,1)$, we have $Q(p)\leq\widehat{Q}(p+u_t(\alpha),S_{1:t})$ and $Q(p)\geq\widehat{Q}(p-u_t(\alpha),S_{1:t-1})$ holds. Step (b) follows since $(S_t)_{t\geq1}$ is an i.i.d. sequence and thus S_t is independent of $\mathcal{F}_{t-1}$. Step (c) follows from the standard fact the for any $\mathbb{R}$ valued continuous random variable, and for any quantiles $0\leq p_1\leq p_2\leq 1$, $\mathbb{P}[X\in(Q(p_1),Q(p_2))]=p_2-p_1$. □

Proof of Lemma 2. Denote by the sequence of 0 mean random variables $\widetilde{Z}_t := Z_t-\mathbb{E}[Z_t|\mathcal{F}_{t-1}]$. Denote by the running sum $Y_t:=\sum_{s=1}^t\widetilde{Z}_s$. It is easy to verify that $(Y_t)_{t=1}^T$ is a martingale sequence. Further from definition, we have have that almost-surely, for all $t\in\{1,\cdots,T\}$, $|Y_t-Y_{t+1}|\leq 2$. Thus, from Azuma Hoeffding inequality for bounded martingales, we have that

$$\mathbb{P}\left[\sum_{s=1}^T\widetilde{Z}_s\geq\sqrt{T\ln\left(\frac{4}{\alpha^2}\right)}\right]\leq 2\exp\left(-\frac{2T\ln\left(\frac{4}{\alpha^2}\right)}{4T}\right), \tag{8}$$
$$\leq\frac{\alpha}{2}. \tag{9}$$

□

C. Proofs from Section 5

The proofs in this section are dependent on the following good-event, under which we will perform the analysis.

Definition C.1 (Good event 2)**.** Let $(S_t)_{t=1}^T \sim \left(T, H_T, (\tau_c)_{c=0}^{H_T}, (f^{(c)})_{c=0}^{H_T}\right)$, with quantile functions of the H_T+1 segments given by $(Q^{(c)}(\cdot))_{c=0}^{H_T}$. Define $\mathcal{E}_2$ as

$$\widehat{\mathcal{E}_2} := \bigcap_{k=1}^{H_T} \mathcal{E}_2^{(k)}, \tag{10}$$

where for each $k \in [H_T]$,

$$\mathcal{E}_2^{(k)} = \bigcap_{p\in[0,1]} \bigcap_{t_1=\tau_k}^{\tau_{k+1}} \bigcap_{t_2=t_1+1}^{\tau_{k+1}} \left\{ \widehat{Q}(p - u_{t_2-t_1}(\alpha), S_{t_1:t_2}) \le Q^{(k)}(p) \le \widehat{Q}(p - u_{t_2-t_1}(\alpha), S_{t_1:t_2}) \right\},$$

where $\widehat{Q}$ is the empirical quantile and $u_{(\cdot)}(\cdot)$ is defined in Equation (2).

In words, $\mathcal{E}_2$ is the intersection over all the $H_T + 1$ stationary segments, where for each segment the quantile estimate is close to the true quantile for all quantiles. Following Corollary 2 of (Howard and Ramdas, 2022) and an union bound, we get the following.

Proposition 8.

$$\mathbb{P}[\widehat{\mathcal{E}_2}] \ge 1 - T\alpha.$$

Proof. We analyze the complement as follows.

$$\begin{aligned}
&\mathbb{P}[\widehat{\mathcal{E}_2}^{\complement}] \\
&= \mathbb{P}\left[\cup_{k=0}^{H_T} \cup_{p\in[0,1]} \cup_{t_1=\tau_k}^{\tau_{k+1}} \cup_{t_2\ge t_1+1} \{\widehat{Q}(p - u_{t_2-t_1}(\alpha), S_{t_1:t_2}) > Q^{(k)}(p)\} \cup \{Q^{(k)}(p) > \widehat{Q}(p - u_{t_2-t_1}(\alpha), S_{t_1:t_2})\}\right], \\
&\le \sum_{k=0}^{H_T} \sum_{t_1=\tau_k}^{\tau_{k+1}} \mathbb{P}\left[\cup_{p\in[0,1]} \cup_{t_2\ge t_1+1} \{\widehat{Q}(p - u_{t_2-t_1}(\alpha), S_{t_1:t_2}) > Q^{(k)}(p)\} \cup \{Q^{(k)}(p) > \widehat{Q}(p - u_{t_2-t_1}(\alpha), S_{t_1:t_2})\}\right], \\
&\overset{(a)}{\le} \sum_{k=0}^{H_T} \sum_{t_1=\tau_k}^{\tau_{k+1}} \alpha, \\
&= T\alpha.
\end{aligned}$$

Step (a) follows from Corollary 2 of (Howard and Ramdas, 2022). □

In addition, we also need a generalization of event $\mathcal{G}_1$ to the piece-wise stationary stream.

$$\mathcal{G}_2 := \bigcap_{k=0}^{H_T} \bigcap_{t_1=\tau_k+1}^{\tau_{k+1}} \left\{ \left| \sum_{t=t_1}^{\tau_{k+1}} Z_{t:t_1} - \mathbb{E}[Z_{t:t_1}|\mathcal{F}_{t-1}] \right| \le \sqrt{(\tau_{k+1} - \tau_k) \ln\left(\frac{4}{\alpha^2}\right)} \right\}, \tag{11}$$

where

$$Z_{t:t_1} := \mathbf{1}(S_t \in (\widehat{Q}_t(p - 2u_{t-t_1}(\alpha), S_{t_1:t-1}), \widehat{Q}_t(p + 2u_{t-t_1}(\alpha), S_{t_1:t-1})), \tag{12}$$

and $\mathcal{F}_{t-1} := \sigma(S_{1:t-1})$ is the sigma-algebra generated by the first $t-1$ samples.

Proposition 9.

$$\mathbb{P}[\mathcal{G}_2] \ge 1 - T\alpha.$$

For a given k and t_1, the proof follows from an application of Azuma-Hoeffding similar to the proof of Theorem 2. Applying an union bound over k and t_1 yields the result.

In the rest of this section, denote the good event $\mathcal{E}_2$ as

$$\mathcal{E}_2 = \widehat{\mathcal{E}}_2 \bigcap \mathcal{G}_2, \tag{13}$$

where $\widehat{\mathcal{E}}_2$ is defined in Equation (10) and $\mathcal{G}_2$ in Equation (11).

Proposition 10.

$$\mathbb{P}[\widehat{\mathcal{E}}_2] \geq 1 - T\alpha.$$

The proof follows from an union bound using estimates in Proposition 8 and 9.

C.1. Proof of Proposition 3

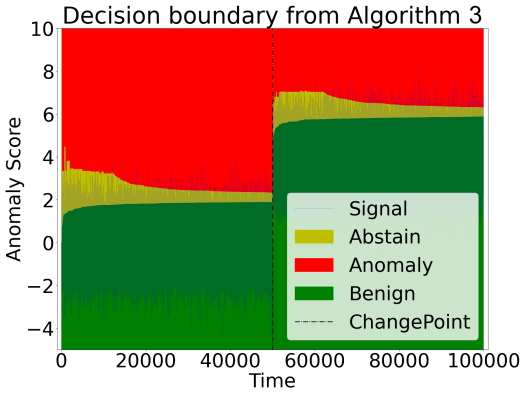


Figure 8: Algorithm 3 output on a piece-wise stationary stream.

In order to give the proof of Proposition 3, we need the following general lemma on change-detection.

Proposition 11 (Stream with a single change-point)**.** Let $(S_t)_{t=1}^T$ be a piecewise stationary stream with a single change-point, i.e., $H_T = 1$ with the change-point instant $\tau_1 = \tau \in [T]$. Denote by $\Delta := \text{Shift}(f^{(0)}, f^{(1)}, p)$ as the distribution shift magnitude. Denote by $\tau' \in \mathbb{N}$ as the first time when the `if` statement in Algorithm 3 evaluates to `True`, i.e., a change is detected for the first time by Algorithm 3. Then, under the good event $\mathcal{E}_2$ (Definition C.1),

$$\tau \leq \tau' \leq \left\lceil \inf_{\widetilde{\tau}} \left\{ u_\tau(\alpha) + u_{\widetilde{\tau}-\tau}(\alpha) \leq \frac{\Delta}{6} \right\} \right\rceil,$$

where $u_t(\alpha)$ is given in Equation (2).

We recall some notation to give the proof. The two distributions are denoted by $f^{(0)}$ and $f^{(1)}$ and their corresponding quantile functions are given by $Q^{(0)}(\cdot)$ and $Q^{(1)}(\cdot)$, where for $i \in \{0, 1\}$, the function $Q^{(i)}(\cdot) : [0, 1] \to \mathbb{R}$ is such that for all $p \in [0, 1]$, $Q^{(i)}(p)$ is the pth quantile of the distribution $f^{(i)}$. Since we assume all distributions are continuous, the existence and uniqueness of quantile functions are granted.

Proof that $\tau \leq \tau'$

Observe that under the event $\mathcal{E}_2$, we have for all $t_1 < t_2 \in [\tau]$, $Q^{(0)}(p) \in C(p, \alpha, S_{t_1:t_2})$. Thus, for all $t \in [\tau]$, $Q^{(0)}(p) \in \bigcap_{s=1}^t C(p, \alpha, S_{s:t}) \bigcap_{s=1}^t C(p, \alpha, S_{1:s})$, and in particular, $\bigcap_{s=1}^t C(p, \alpha, S_{s:t}) \bigcap_{s=1}^t C(p, \alpha, S_{1:s})$ is non-empty. Therefore, under event $\mathcal{E}_2$, for all $t \in [\tau]$, Algorithm 2 will never return `True` for the input sequence $S_{1:t}$.

Proof on the upper bound of τ'

Since $\Delta > 0$, we will assume here that $Q^{(0)}(p) < Q^{(1)}(p)$ and this is without loss of generality since the proof for the other case follows identically. To prove the upper bound, it suffices to prove that if event $\mathcal{E}_2$ holds and no change is detected till time $\tau' = \inf\{\widetilde{\tau} : u_\tau(\alpha) + u_{\widetilde{\tau}-\tau}(\alpha)\}$, then at time $\lceil \tau' \rceil$, a change point will be detected.

In order to prove this, we need to establish that $\bigcap_{s=1}^{\lceil \tau' \rceil} C(p, \alpha, S_{s:t}) \bigcap_{s=1}^{\lceil \tau' \rceil} C(p, \alpha, S_{1:s}) = \emptyset$. A sufficient condition for this is to prove that $C(p, \alpha, S_{\tau+1:\lceil \tau' \rceil}) \bigcap C(p, \alpha, S_{1:\tau}) = \emptyset$, i.e., the following implication holds

$$C(p, \alpha, S_{\tau+1:\lceil \tau' \rceil}) \bigcap C(p, \alpha, S_{1:\tau}) = \emptyset \implies \bigcap_{s=1}^{\lceil \tau' \rceil} C(p, \alpha, S_{s:t}) \bigcap_{s=1}^{\lceil \tau' \rceil} C(p, \alpha, S_{1:s}) = \emptyset. \tag{14}$$

From Equation (1), we know that

$$C(p, \alpha, S_{1:\tau}) = \left[\widehat{Q}(p - 2u_\tau(\alpha), S_{1:\tau}), \widehat{Q}(p + 2u_\tau(\alpha), S_{1:\tau}) \right].$$

Since event $\mathcal{E}_2$ holds, it follows that

$$\widehat{Q}(p + 2u_\tau(\alpha), S_{1:\tau}) \leq Q^{(0)}(p + 3u_\tau(\alpha)) \tag{15}$$

$$\widehat{Q}(p - 2u_{\lceil \tau' \rceil - \tau}(\alpha)) \geq Q^{(1)}(p - 3u_{\lceil \tau' \rceil - \tau}(\alpha)). \tag{16}$$

Further, from the definition of Δ in Definition 5.1 and the assumption that $Q^{(0)}(p) < Q^{(1)}(p)$ we know that

$$Q^{(0)}(p + \Delta) \leq Q^{(1)}(p - \Delta) \tag{17}$$

Now, from the definition of τ', we know that

$$3(u_\tau(\alpha) + u_{\lceil \tau' \rceil - \tau}(\alpha)) \leq \Delta \tag{18}$$

Now, combining the inequalities in Equations (16, 17, 18), we get that

$$\widehat{Q}(p + 2u_\tau(\alpha), S_{1:\tau}) \overset{15}{\leq} Q^{(0)}(p + 3u_\tau(\alpha)) \overset{(a)}{<} Q^{(0)}(p + 3u_\tau(\alpha) + 3u_{\lceil \tau' \rceil - \tau}(\alpha)) \overset{18}{\leq} Q^{(0)}(p + \Delta) \overset{17}{\leq}$$
$$Q^{(1)}(p - \Delta) \overset{18}{\leq} Q^{(1)}(p - 3u_{\lceil \tau' \rceil - \tau}(\alpha)) \overset{16}{\leq} \widehat{Q}(p - 2u_\tau(\alpha), S_{\tau:\lceil \tau' \rceil}). \tag{19}$$

Step (a) follows from the fact that since $f^{(0)}$ and $f^{(1)}$ are a continuous distribution, we have for all $i \in \{0, 1\}$ and a;; $0 < p_1 < p_2 < 1$, $Q^{(i)}(p_1) < Q^{(i)}(p_2)$. Thus, Equation (19) gives that under the event $\mathcal{E}_2$, $C(p, \alpha, S_{\tau+1:\lceil \tau' \rceil}) \bigcap C(p, \alpha, S_{1:\tau}) = \emptyset$, which from Equation (14) implies that if Algorithm 2 is queried with input $S_{1:\lceil \tau' \rceil}$, will return a 1, i.e., detect a change-point.

Concluding the proof of Proposition 3 as an application of Proposition 11

Proof. The proof rests on the following numerical claim.

Claim 2. If $\tau \geq \frac{80}{\Delta^2} \ln\left(\frac{1612}{\alpha\Delta^2}\right)$, then $u_\tau(\alpha) \leq \Delta/6$ holds.

Proof of Claim 2. Suppose $\tau \geq \frac{80}{\Delta^2} \ln\left(\frac{1612}{\alpha\Delta^2}\right)$. Then

$$u_\tau(\alpha) \leq \sqrt{\frac{\ln\left(\frac{1612}{\alpha}(1 + \ln(\tau))\right)}{\tau}},$$
$$\leq \frac{\Delta}{6} \frac{\sqrt{\ln\left(\frac{1612}{\alpha}(1 + \ln(\tau))\right)}}{\sqrt{5 \ln\left(\frac{1612}{\alpha\Delta^2}\right)}},$$

$$\leq \frac{\Delta}{6}\frac{\sqrt{\ln\left(\frac{1612}{\alpha}\left(1+\ln\left(\frac{20}{\Delta^2}\ln\left(\frac{1612}{\alpha\Delta^2}\right)\right)\right)\right)}}{\sqrt{5\ln\left(\frac{1612}{\alpha\Delta^2}\right)}},$$
$$\leq \frac{\Delta}{6}.$$

□

Applying Claim 2 to both τ and $\tau' - \tau$ allows use to satisfy the condition in Proposition 3.

$$\begin{aligned}
&\tau \geq \frac{80}{\Delta^2}\ln\left(\frac{1612}{\alpha\Delta^2}\right),\ \lceil\tau'\rceil - \tau \geq \frac{80}{\Delta^2}\ln\left(\frac{1612}{\alpha\Delta^2}\right)\\
\implies &u_\tau(\alpha) \leq \Delta/6,\ u_{\lceil\tau'\rceil-\tau}(\alpha) \leq \Delta/6,\\
\implies &3(u_\tau(\alpha) + u_{\lceil\tau'\rceil-\tau}(\alpha)) \leq \Delta
\end{aligned}$$

□

C.2. Proof of Theorem 4

Theorem (4)**.** Let $(S_t)_{t=1}^T \sim \left(T, H_T, (\tau_c)_{c=0}^{H_T}, (f_c)_{c=0}^{H_T}\right)$ be a piece-wise stationary data-stream satisfying Assumption 5.1 for $\alpha \in (0,1)$. Then, with probability at-least $1 - 2T\alpha$, Algorithm 3 satisfies *both*

- FP + FN $\leq \sum_{k=1}^{H_T} D(\Delta_k, \alpha) = H_T \cdot \mathcal{O}(\frac{1}{(\min_k \Delta_k)^2})$,
- Number of abstains $\leq \sum_{k=1}^{H_T}\left[7\sqrt{(\tau_k - \tau_{k-1})\ln\left(\frac{1612}{\alpha^2}\ln(3(\tau_k - \tau_{k-1}))\right)} + D(\Delta_k, \alpha)\right]$,

where for all $k \in [H_T]$, $\Delta_k = \text{Shift}(f^{(k)}, f^{(k-1)}, p)$ is the quantile shift defined in Definition (5.1) and $D(\cdot,\cdot)$ is defined in Equation (3).

At a high-level, the proof idea is to show in Lemma 3 that if Assumption 5.1 holds, then under the good event in Equation (13) *(i)* there are no false positive detections, and *(ii)* all the true changes are detected with bounded delay. This will conclude the proof of Theorem 4 since a simple union bound over Corollary 2 of (Howard and Ramdas, 2022) implies that the good event in Equation (13) holds with probability at-least $1 - T\alpha$.

In order to formalize the proof, we recall notations. As before, denote by $\tau_1 < \tau_2 \cdots < \tau_k < \tau_{k+1}$ be the set of times at which a true-change occurs. For each $k \in \{1, \cdots, H_T\}$, denote by τ'_k as the time at which Algorithm 3 detects a change for the kth time. The first observation is the lemma below which states that there are no false positives and the delay of all change detections are bounded.

Lemma 3 (No false detections and bounded delay)**.** Under the event $\mathcal{E}_2$, if Assumption 5.1 holds, then for all $k \in \{1, \cdots, H_T\}$,

- $\tau'_k \geq \tau_k$, and
- $\tau'_k - \tau_k \leq D(\Delta_k, \alpha)$.

Before giving the proof of the lemma, we show how it concludes the Proof of Theorem 4.

Proof of Theorem 4. We can break down the analysis by change points, because we know that all changes are detected before the next change occurs. Moreover, Algorithm 3 restarts a fresh copy of Algorithm 6 after each change point is detected, and thus we can use the guarantees given in Theorem 2.

Consider a subset of time points $[S_{\tau_k}, S_{\tau_{k+1}}]$ for some $k \in [1, H_T]$. We will denote the time step of detecting change point τ_k, as τ_k'. The total number of mistakes (false positives and false negatives) is the sum of mistakes before and after τ_k'. Therefore under the good event $\mathcal{E}_2$ we have:

$$\begin{aligned}
\text{FP} + \text{FN} &= \sum_{k=0}^{H_T} \left(\sum_{t=\tau_k}^{\tau_k'} \mathbf{1}(\text{Time } t \text{ is a mistake}) + \sum_{t=\tau_k'+1}^{\tau_{k+1}} \mathbf{1}(\text{Time } t \text{ is a mistake}) \right) \\
&\leq \sum_{k=0}^{H_T} (\tau_k' - \tau_k) + 0 \qquad \text{(Event } \mathcal{E}_2 \text{ definition in Equation (13))} \\
&\leq \sum_{k=0}^{H_T} D(\Delta_k, \alpha) + 0 \qquad \text{(Lemma 3)}
\end{aligned}$$

The analysis is similar for number of abstains.

$$\begin{aligned}
\text{Abstains} &= \sum_{k=0}^{H_T} \left(\sum_{t=\tau_k}^{\tau_k'} \mathbf{1}(\text{Time } t \text{ is abstained}) + \sum_{t=\tau_k'+1}^{\tau_{k+1}} \mathbf{1}(\text{Time } t \text{ is abstained}) \right) \\
&\leq \sum_{k=0}^{H_T} \left[(\tau_k' - \tau_k) + \sum_{t=\tau_k'+1}^{\tau_{k+1}} \mathbf{1}(\text{Time } t \text{ is abstained}) \right] \\
&\overset{\text{Lemma 3}}{\leq} \sum_{k=0}^{H_T} \Big[D(\Delta_k, \alpha) + \sum_{t=\tau_k'+1}^{\tau_{k+1}} \mathbf{1}(\text{Time } t \text{ is abstained}) \Big], \\
&\overset{(a)}{\leq} \sum_{k=0}^{H_T} \Big[D(\Delta_k, \alpha) + \sum_{t=\tau_k'+1}^{\tau_{k+1}} \mathbf{1}(Z_{t:\tau_k'} = 1) \Big], \\
&\overset{(b)}{\leq} \sum_{k=0}^{H_T} \Big[D(\Delta_k, \alpha) + 7\sqrt{(\tau_k - \tau_{k-1}) \ln \left(\frac{1612}{\alpha^2} \ln(3(\tau_k - \tau_{k-1})) \right)} \Big].
\end{aligned}$$

In step (a), we use the definition of $Z_{t:t_1}$ as used in Equation (11) and in step (b), we use the fact that event $\mathcal{G}_2$ holds and thus the calculations from Theorem 2 can be used. □

C.3. Proofs of Lemma 3

We prove the result by induction on $k \in \{1, \cdots, H_T\}$.

Base case of $k = 1$: Observe that under event $\mathcal{E}_2$, we have $\tau_1^{'} \geq \tau_1$. This is so since for all times $t \in \{1, \cdots, \tau_1\}$, there will be no distribution-shift detected, since by definition under event $\mathcal{E}_2$, the forward and backward CS will contain at-least the true quantile and will thus be non-empty.

Now, under Assumption 5.1, we know that $\tau_1 \geq \frac{20}{\Delta_1^2} \ln\left(\frac{1612}{\alpha\Delta_1^2}\right)$. Thus, by Proposition 3, we know that $\tau_1^{'} - \tau_1 \leq \frac{20}{\Delta_1^2} \ln\left(\frac{1612}{\alpha\Delta_1^2}\right) := D(\Delta_1, \alpha)$, where the last equality is from definition in Equation (3). This proves the induction base case.

Induction hypothesis: Now, assume that for some $k \in \{1, \cdots, H_t\}$, we have that $\tau_k^{'} - \tau_k \leq D(\Delta_k, \alpha)$. We will now show that the $(k+1)$th detection time $\tau_{k+1}^{'}$ satisfies $\tau_{k+1}^{'} - \tau_{k+1} \leq D(\Delta_{k+1}, \alpha)$.

Assumption 5.1 ensures that $\tau_{k+1} - \tau_k \geq D(\Delta_{k+1}, \alpha) + D(\Delta_k, \alpha)$. The induction hypothesis gives that $\tau_k^{'} - \tau_k \leq D(\Delta_k, \alpha)$. Thus, under Assumption 5.1 and the induction hypothesis, we have that $\tau_{k+1} - \tau_k^{'} \geq D(\Delta_{k+1}, \alpha)$. From the working in Algorithm 3, we know that at time $\tau_k^{'}$, a new instantiation of Algorithm 6 is started. Following the same arguments as for the

base-case, we know that under the good-event $\mathcal{E}_2$, $\tau^{'}_{k+1} \geq \tau_{k+1}$, i.e., there are no false-positive detection. Moreover, since this re-started version of Algorithm 6 has seen at-least $\tau_{k+1} - \tau^{'}_{k} \geq D(\Delta_{k+1}, \alpha)$ pre-change samples, Proposition 3 gives that $\tau^{'}_{k+1} - \tau_{k+1} \leq D(\Delta_{k+1}, \alpha)$. Thus, the induction hypothesis is proved.

C.4. Details on Assumption 5.1

The following corollary states that if a process $(S_t)_{t=1}^T$ is α-detectable, then it is also α/T detectable with an additional log-factor, made precise in the following corollary.

Corollary 11.1. For a given $\alpha \in (0,1)$, $T \in \mathbb{N}$, $H_T \leq T-1$ and distributions $(f^{(c)})_{c=0}^{H_T}$, if $\left(T, H_T, (\tau_c)_{c=0}^{H_T}, (f^{(c)})_{c=0}^{H_T}\right)$ is α-detectable according to Assumption 5.1 with $T \geq 9$ and $\frac{1612}{\alpha\Delta_k^2} \geq 9$ for all $k \in [H_T]$, then the process $\left(T\lceil\log(T)\rceil, H_T, (\tau_c\lceil\log(T)\rceil)_{c=0}^{H_T}, (f^{(c)})_{c=0}^{H_T}\right)$ is α/T-detectable.

Proof. We need to verify that for all $k \in [H_T]$, the bound $\lceil\log(T)\rceil(\tau_k - \tau_{k-1})$ satisfies the conditions in Definition 5.1 by with α/T in place of α. Consider $k=1$. Since $\left(T, H_T, (\tau_c)_{c=0}^{H_T}, (f^{(c)})_{c=0}^{H_T}\right)$ is α-detectable, we have

$$\tau_1 \geq D(\Delta_1, \alpha).$$

Thus, multiplying both sides by $\lceil\log T\rceil$, we get that

$$\begin{aligned}
\lceil\log T\rceil\tau_1 &\geq \lceil\log T\rceil D(\Delta_1, \alpha), \\
&= \lceil\log T\rceil \frac{80}{\Delta_1^2}\ln\left(\frac{1612}{\alpha\Delta_1^2}\right), \\
&\overset{(a)}{\geq} C_1 \ln(T)\ln\left(\frac{C_2}{\alpha}\right), \\
&\overset{(b)}{\geq} C_1 \ln\left(\frac{C_2 T}{\alpha}\right), \\
&= D\left(\Delta_1, \frac{\alpha}{T}\right)
\end{aligned}$$

In step (a), we let $C_1 = \frac{80}{\Delta_1^2}$ and $C_2 = \frac{1612}{\Delta_1^2}$. Step (b) follows since $T \geq 9$ and $\frac{1612}{\alpha\Delta_k^2} \geq 9$ for all $k \in [H_T]$. □

D. Algorithm and proof for Theorem 5, stationary stream with offline data

We outline the full algorithm for the offline data setting in algorithm 7, followed by the proof of Theorem 5. To set the ideas, we will need some definitions.

Definition D.1. For every $j \in [K]$, we denote by $Q_{(j)}(p) \in \mathbb{R}$ as the pth quantile of the jth dataset $\mathcal{D}_j := \{X_l^{(j)}\}_{l=1}^{N_j}$.

Definition D.2 (Good event 3)**.** Define the event $\mathcal{E}_3$ as

$$\mathcal{E}_3 := \mathcal{G}_2 \bigcap \Bigg\{ \bigcap_{p\in[0,1]} \bigcap_{j=1}^{K} \{Q_{(k)}(p - u_{N_j}(\alpha)) \leq \widehat{Q}(p, \{X_l^{(j)}\}_{l=1}^{N_j}) \leq Q_{(k)}(p + u_{N_j}(\alpha))\} \bigcap_{t\in\mathbb{N}} \{\widehat{Q}(p - u_t(\alpha), S_{1:t}) \leq Q(p) \leq \widehat{Q}(p + u_t(\alpha), S_{1:t})\} \Bigg\}, \quad (20)$$

where $Q(p)$ is the pth quantile of the i.i.d. sequence $(S_t)_{t\geq 1}$, and $\mathcal{G}_2$ is defined in Equation (11).

Proposition 12.

$$\mathbb{P}[\mathcal{E}_3] \geq 1 - (K + 2T)\alpha.$$

The proof of this follows using Corollary 2 of (Howard and Ramdas, 2022) and an union bound similar to the proof of Proposition 10.

Algorithm 7 Stationary stream with offline datasets

Input : Quantile q, Confidence $1-\alpha$, offline datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$.
Output : Anomaly labels $\hat{y}_1, \hat{y}_2, \ldots$
for *each time* $t \geq 1$ **do**
  Receive t^{th} input S_t
  $\hat{y}_t \leftarrow$ `Algorithm-4`$(p, \alpha, S_{1:t-1}, (\mathcal{D}_1, \cdots, \mathcal{D}_K))$
end

Theorem (5)**.** Let the online stream $(S_t)_{t\geq 1}$ be i.i.d. from distribution f and the $j \in [K]$ offline dataset with N_j i.i.d. samples have distribution $f_{(j)}$. Further, let the offline dataset be (p, α) well-separated, i.e., satisfy the condition in Definition 6.1. Let $\Delta = \min_{1\leq j\leq K} \text{Shift}(f, f_{(j)}, p)$ and $\widehat{\tau}$ is defined in Section 6. Then with probability at-least $1-(K+2T)\alpha$, for all times $T \in \mathbb{N}$, *all* of the following holds for Algorithm 7.

- If $\Delta = 0$, then,
  - FP + FN $= 0$ and
  - Abstains $\leq B(N+T, \alpha) - \frac{1}{28}B(N, \alpha) + \widehat{\tau}$,

  where $N = \min\{N_j : \text{Shift}(f, f_{(j)}, p) = 0\}$ and $B(\cdot,\cdot) : \mathbb{N} \times [0,1] \to \mathbb{R}_+$ is given by $B(Y, \alpha) = 7\sqrt{Y \ln\left(\frac{1612 \ln(3Y)}{\alpha^2}\right)}$.
- On the other hand, if $\Delta > 0$, then
  - FP + FN $\leq \widehat{\tau}$ and
  - Abstains $\leq \widehat{\tau} + B(T, \alpha)$.

Proof. The proofs are based on two structural lemmas in Lemma 4 and 5.

Lemma 4. Suppose there exists $j \in [K]$ such that the pth quantile of the offline dataset $\mathcal{D}_j$ and the online stream (S_t) are identical, i.e., Shift$(f, f_{(j)}, p) = 0$. If the offline datasets $\mathcal{D}_1, \cdots, \mathcal{D}_K$ satisfy the conditions in Definition 6.1 and the good event $\mathcal{E}_3$ holds, then for all time t, $j \in \{j' \in [K] : C(p, \alpha, \mathcal{D}_{j'}) \cap C(p, \alpha, S_{1:t-1}) \neq \emptyset\}$.

Proof. Under the hypothesis of the Lemma, $Q_{(j)}(p) = Q(p)$. Further, since event $\mathcal{E}_3$ holds, we know that for all $t \in \mathbb{N}$, $Q(p) \in C(p, \alpha, S_{1:t-1})$ and $Q(p) \in C(p, \alpha, \mathcal{D}_j)$. Thus, under event $\mathcal{E}_3$, for all $t \in \mathbb{N}$, $C(p, \alpha, \mathcal{D}_j) \cap C(p, \alpha, S_{1:t-1}) \neq \emptyset$. This concludes the proof of the lemma. □

Lemma 5. Suppose the online stream $(S_t)_{t=1}^T$ and the offline datasets $\mathcal{D}_1, \cdots, \mathcal{D}_K$ satisfy the conditions of Theorem 5. Then, for all $j' \in [K]$ such that $Q_{(j')}(p) \neq Q(p)$, and all $t \geq \widehat{\tau}_j$, $C(p, \alpha, \mathcal{D}_{j'}) \cap C(p, \alpha, S_{1:t}) = \emptyset$.

Proof. From the condition in Definition 6.1, we know that if $Q_{(j')}(p) \neq Q(p)$, then $Q(p) \notin C(p, \alpha, \mathcal{D}_{j'})$. Suppose without loss of generality, assume $Q(p) < Q_{(j')}(p)$. Consider a time $t \geq \widehat{\tau}_j$. Since the good event $\mathcal{E}_3$ holds, we know that

$$\begin{aligned}\max C(p, \alpha, S_{1:t}) = \widehat{Q}(p + 2u_t(\alpha), S_{1:t}) &\overset{(a)}{\leq} Q(p + 3u_t(\alpha)) \overset{(b)}{<} Q(p + \Delta_{j'}) \overset{(c)}{\leq} Q_{(j')}(p - \Delta_{j'}) \\ &\overset{(b)}{<} Q_{(j')}(p - 3u_{N_{j'}}(\alpha)) \overset{(a)}{\leq} \widehat{Q}(p - 2u_{N_{j'}}(\alpha), \mathcal{D}_{j'}) = \min C(p, \alpha, \mathcal{D}_{j'}).\end{aligned}$$

Step (a) follows from the good event $\mathcal{E}_3$ definition in Equation (20). Step (b) follows from the definition of $t \geq \widehat{\tau}_{j'}$ where $u_{\widehat{\tau}_{j'}} \leq \Delta_{j'}$ and step (c) follows from the definition of quantile shift in Definition 5.1. □

Proof of Theorem 5 in the case $\Delta = 0$:

Proof that FP + FN = 0. Observe from the definition of Algorithm 4, one of two possibilities occur at each time t. Either only the online stream's samples $S_{1:t-1}$ are used to make a decision on S_t, or both the online and one offline dataset is used to make a decision. If only the online stream is used for decision making, then Theorem 2 gives that time t is neither a FP

nor a FN. On the other hand, Lemma 4 gives that if an offline dataset is used to make a decision, then the *correct* offline dataset is used, i.e., the one whose pth quantile matches that of the online stream. Thus, Theorem 2 can be applied since the union of the offline and online stream is a combination of independent samples with identical pth quantile.

Proof on the abstain bound. From Lemma 5, we know that for all times $t \geq \max\{\lceil\widehat{\tau_{j'}}\rceil : Q_{(j')} \neq Q(p)\}$, the decision $\hat{y}_t$ is made by the union of the *correct* offline dataset $\mathcal{D}_j$ and the online stream $S_{1:t-1}$. Thus, the maximum number of abstains is the sum of two terms - *(i)* is the max number of abstains till time $\max\{\lceil\widehat{\tau_{j'}}\rceil : Q_{(j')} \neq Q(p)\}$ which is at-most one per time-step, and the second is the number of abstains in the last T samples of a $N+T$ length stationary stream. Theorem 2 gives that under the event $\mathcal{E}_3$, we have that at-most $B(N+T,\alpha)$ abstains occur over the entire $T+N$ time horizon, and *(ii)* at-least $\frac{1}{28}B(N,\alpha)$ abstains occurs in the first N samples. Putting these together yields the result.

Proof of Theorem 5 in the case $\Delta > 0$:

The case of $\Delta > 0$ implies that for all $j \in [K]$, $Q_{(j)}(p) \neq Q(p)$. Lemma 5 gives that for all times $t \geq \lceil\widehat{\tau_j}\rceil$, $C(p,\alpha,\mathcal{D}_j) \cap C(p,\alpha,S_{1:t}) = \emptyset$. Thus, for all times $t \geq \max_{j\in[K]}\lceil\widehat{\tau_j}\rceil$, the decision $\hat{y}_t \leftarrow \texttt{Algorithm-1}(C(p,\alpha,S_{1:t}))$. Theorem 2 gives that for all times $t \geq \max_{j\in[K]}\lceil\widehat{\tau_j}\rceil$, the number of mistakes is 0 under the good event $\mathcal{E}_3$ and the number of abstains is at-most $B(T,\alpha)$.

□

D.1. Illustration of well separated offline datasets

In Figure 7, we give a schematic representation of the well-separated assumption described in Definition 6.1. This shows that the 4 offline datasets' CS must be non-intersecting and must be such that *(i)* either the true quantile of an online segment must match one of the offline datasets, *(ii)* or the online quantile must not lie inside any of the offline dataset's CS.

E. Proof for Theorem 6, general case

To state performance guarantee of Algorithm 5, we need a definition of a good event, which occurs with high probability.

Definition E.1 (Good event 4)**.** Given a piece-wise stationary sequence $(S_t)_{t=1}^T := \left(T, H_T, (\tau_c)_{c=0}^{H_T}, (f^{(c)})_{c=0}^{H_T}\right)$, with the quantile functions of the H_T+1 different segments given by $(Q^{(c)}(\cdot))_{c=1}^{H_T}$ and K offline datasets that are (p,α) separated according to Definition 6.1, let

$$\mathcal{E}_4 = \mathcal{G}_2 \underbrace{\bigcap_{p\in[0,1]}\bigcap_{j=1}^{K}\{Q_{(k)}(p-u_{N_j}(\alpha)) \leq \widehat{Q}(p,\{X_l^{(j)}\}_{l=1}^{N_j}) \leq Q_{(k)}(p+u_{N_j}(\alpha)))\}}_{\text{Offline dataset's CS contain true quantile}}$$
$$\underbrace{\bigcap_{k=0}^{H_T}\bigcap_{t_1=\tau_k}^{\tau_{k+1}}\bigcap_{t_2=t_1+1}^{\tau_{k+1}}\left\{\widehat{Q}(p-u_{t_2-t_1}(\alpha),S_{t_1:t_2}) \leq Q^{(k)}(p) \leq \widehat{Q}(p-u_{t_2-t_1}(\alpha),S_{t_1:t_2})\right\}}_{\text{Online stream's CS contain true quantile}}, \tag{21}$$

where the pth quantile of the offline datasets $Q_{(l)}(p)$ are defined in Definition D.1 and $\mathcal{G}_2$ is defined in Equation (11).

Event $\mathcal{E}_4$ is the union of $\mathcal{E}_2$ in Equation (13) for the online stream along with the event $\mathcal{E}_3$ in Equation (20) for the offline datasets.

Proposition 13.

$$\mathbb{P}[\mathcal{E}_4] \geq 1 - \alpha(K+2T)$$

The proof follows similar to that of Proposition 10.

As before, denote by H_T as the number of change points, and the time-instants of change occuring at $1 := \tau_0 < \tau_1 < \tau_2 \cdots < \tau_{H_T} < \tau_{H_T+1} := T+1$. We denote by the kth *online stream* as the stationary samples $\{S_u : \tau_{k-1} \leq u < \tau_k\}$.

For an offline dataset $j \in \{1, \cdots, K\}$ and the online stream k, $\Delta^{(j;k)} \geq 0$ denotes the shift between the jth offline dataset and the kth online stream. As before, similar to Theorem 4, for all $k \in [H_T]$, denote by $\Delta_k = \text{Shift}(f_k, f_{k+1})$ to be the shift between the kth and $k-1$th online segment, and for $k \in [H_T]$ and $j \in [K]$, let $\Delta^{(j,k)} = \text{Shift}(f^{(k)}, f_{(j)}, p)$ be the shift between the jth offline dataset and the kth online segment.

Theorem (6)**.** Let $(S_t)_{t=1}^T \sim \left(T, H_T, (\tau_c)_{c=0}^{H_T}, (f_c)_{c=0}^{H_T}\right)$ be a piece-wise stationary stream satisfying Assumption 5.1 and offline datasets $\mathcal{D}_1, \ldots, \mathcal{D}_L$ satisfying Assumption 6.1. Then, with probability at-least $1-(K+2T)\alpha$, for all times $T \in \mathbb{N}$, Algorithm 5 satisfies

- $\text{FP} + \text{FN} \leq \sum_{k=1}^{H_T} \left(D(\Delta_k, \alpha) + (1 - \mathbf{1}_{\text{Match}_k}) \max_{j \in [K]} \left(\lceil \widehat{\tau}_j^{(k)} \rceil \mathbf{1}_{\text{Shift}(f^{(k)}, f_{(j)}, p) > 0} \right) \right)$
- Number of abstains less than

$$\sum_{k=1}^{H_T} \Big[(1 - \mathbf{1}_{\text{Match}_k}) 7 \sqrt{(\tau_k - \tau_{k-1}) \ln \left(\frac{1612}{\alpha^2} \ln(e(\tau_k - \tau_{k-1})) \right)} + \mathbf{1}_{\text{Match}_k} \left(7 \sqrt{(N+T) \ln \left(\frac{1612 \ln(e(N+T))}{\alpha^2} \right)} - 4 \sqrt{N \ln \left(\frac{1612 \ln(eN)}{\alpha^2} \right)} \right) + D(\Delta_k, \alpha) + \max_{j \in [K]} \left(\lceil \widehat{\tau}_j^{(k)} \rceil \mathbf{1}_{\text{Shift}(f^{(k)}, f_{(j)}, p) > 0} \right) \Big],$$

where $\mathbf{1}_{\text{Match}_k} = \mathbf{1}_{Q^{(k)}(p) \in \{Q_{(1)}(p), \cdots, Q_{(K)}(p)\}}$, $Q^{(k)}(p)$ is the pth quantile of the kth segment of the online stream (Defn. 2.1) and $Q_{(1)}(p), \cdots, Q_{(K)}(p)$ are the pth quantiles of the K different offline datasets (Defn. 2.3).

The proof structure follows a similar path as that of Theorem 4. Since Assumption 5.1 holds, we know that Lemma 3 holds. Thus, we have that *(i)* there are no false positive change detections, and *(ii)* all changes are detected with a short detection delay. From the description of Algorithm 5, we know that whenever a change is detected, a new instantiation of Algorithm 7 is started. Thus, the final result is obtained by summing the guarantees in Theorem 5 over the k online segments in addition to the delay taken to detect changes. The delay to detect changes are bounded by using Proposition 3 and Assumption 5.1, similar to that done in Theorem 4.

Proof of Theorem 6. Theorem 6 is the joint result of theorems 4 and 5. As in Theorem 4, Lemma 3 implies that at time $\tau_k' < \tau_{k+1}$, a new instantiation of Algorithm 7 is started. Theorem 5 states that in the time interval $[\tau_k', \tau_{k+1}]$, the kth online segment incurs a total number of mistakes bounded by the guarantee in Theorem 5 for a time horizon $\tau_{k+1} - \tau_k'$. Concretely, in the time-interval $[\tau_k', \tau_{k+1}]$, the sum of FP and FN is bounded above by 0 in the event $\mathbf{1}_{\text{Match}_k}$ holds, and by $\max_{j \in [K]} \left(\lceil \widehat{\tau}_j^{(k)} \rceil \mathbf{1}_{\text{Shift}(f^{(k)}, f_{(j)}, p) > 0} \right)$ in the case when $\mathbf{1}_{\text{Match}_k}$ does not hold. Similarly, the number of abstains in the time-interval $\tau_{k+1} - \tau_k^{'}$ is given by the guarantee of Theorem 5. Now, summing over the H_T different online segments yields the result.

For any segment $k \in [1, \ldots, H_T]$, we have

- $\text{FP}_k + \text{FN}_k \leq D(\Delta_k, \alpha) + 1 - \mathbf{1}_{\text{Match}_k}) \max_{j \in [K]} \left(\lceil \widehat{\tau}_j^{(k)} \rceil \mathbf{1}_{\text{Shift}(f^{(k)}, f_{(j)}, p) > 0} \right)$ where the first term is from detection delay in detecting the change (Proposition 3) and the second term from Theorem 5.
- If $Q^{(k)}(p)$ matches one of the dataset $\{Q_{(1)}(p), \cdots, Q_{(K)}(p)\}$, then under event $\mathcal{G}_2$, number of abstains is at-most

$$D(\Delta_k, \alpha) + \max_{j \in [K]} \left(\lceil \widehat{\tau}_j^{(k)} \rceil \mathbf{1}_{\text{Shift}(f^{(k)}, f_{(j)}, p) > 0} \right) + 7 \sqrt{(N + \tau_k - \tau_{k-1}) \ln \left(\frac{1612 \ln(e(N + \tau_k - \tau_{k-1}))}{\alpha^2} \right)} - 7 \sqrt{N \ln \left(\frac{1612 \ln(eN)}{\alpha^2} \right)},$$

by Theorem 6, case $\Delta = 0$.

- If $Q^{(k)}(p)$ matches none of the datasets:

$$\text{Abstains} \leq D(\Delta_k, \alpha) + \max_{j\in[K]} \left(\lceil \widehat{\tau}_j^{(k)} \rceil \mathbf{1}_{\text{Shift}(f^{(k)}, f_{(j)}, p)>0} \right) + 7\sqrt{(\tau_k - \tau_{k-1}) \ln \left(\frac{1612 \ln(e(\tau_k - \tau_{k-1}))}{\alpha^2} \right)}$$

  by theorem 6, case $\Delta > 0$.

The term of $D(\Delta_k, \alpha)$ always shows in all three equations above as that is the amount of time taken for the kth change-point to be detected and thus all decisions made during the times of change can potentially lead to mistakes and abstains. Combining the piece-wise results together, we have

$$\text{FP} + \text{FN} \ \leq \sum_{k=1}^{H_T} D(\Delta_k, \alpha) + (1 - \mathbf{1}_{\text{Match}_k}) \max_{j\in[K]} \left(\lceil \widehat{\tau}_j^{(k)} \rceil \mathbf{1}_{\text{Shift}(f^{(k)}, f_{(j)}, p)>0} \right)$$

Similarly, summing over the H_T online segments gives the bound on the number of abstains.

□

F. Additional Related Works

Online anomaly detection works There exist works that focuses on other aspects of online anomaly detection. (Zhu et al., 2023) models online concept drift via an ensemble of models. (Nardi et al., 2022) tackles modeling for online AD in a Federated setting. (Bhatia et al., 2022) studies the low-memory constraint of online AD, and (Goyal et al., 2020) focuses on efficiently modeling multi-dimensional data. A survey of methodologies of data modelling for AD can be found in the surveys of (Chandola et al., 2009; Ruff et al., 2021; Pang et al., 2021). Our work differs from these works in setting: we focus on dynamic thresholding of anomaly scores without having access or ability to retrain the underlying AD algorithm.

Threshold learning Using score thresholds to make binary decision of seeking labels from human experts are studied in active learning (Balcan et al., 2007; Zhang et al., 2016; Vishwakarma et al., 2023). However, active learning does not have a concept of mistakes and is thus have different desiderata compared to our study.

G. Additional experiments and experiment details

G.1. Synthetic experiments

We conduct our synthetic experiments for 1000 trails on streams each of length 1000. The normal distribution online streams are synthesized with parameters sampled uniformly form $[0, 10]$ for mean and $[0.2, 2]$ for variance. The Pareto distributions' parameters are sample uniformly from $[1, 3]$ for b, $[0, 10]$ for mean, and $[0.2, 2]$ for scale. To synthesize distribution shifts, we sample a duration for each stationary piece from a Poisson distribution with $\lambda = 300$. To simulate datasets, we prepare $L = 5$ offline datasets each of size $N = 5000$; the online stream can be generated from either one of the dataset distributions or a random one with uniform probability. The range of these parameters are chosen to simulate real world anomaly score data. The code to generate the synthetic dataset as well as implementations of our algorithms will be open sourced.

		Pareto	
shift	data	Abs. (%)	Mis. (%)
x	x	17.8 ± 7.4	.03 ± .01
✓	x	51.9 ± 35.0	.51 ± .18
x	✓	16.8 ± 4.7	.06 ± .04
✓	✓	24.3 ± 5.2	.33 ± .41

Table 3: Results of our algorithms on synthetic stream dataset generated from Pareto distribution.

G.2. Case study of real AD application

In this section we provide details on our real data experiments. We run Algorithms 3 and 5 on two large scale datasets `DS1` and `DS2` obtained from monitoring two large cloud computing services. Each dataset is a stream of anomaly scores

obtained by applying a single anomaly detection algorithm. On these datasets, we run our two general algorithms, Algorithm 3 and 5 with $p = 1 - 10^{-6}$ and $\alpha = 10^{-6}$. Algorithm 5 can use offline datasets. In order to create them, we sampled 50 streams randomly that had length at-least $20,000$ in each of the two datasets. From this collection of 100 streams, we sub-selected 37 and 23 streams from datasets `DS1` and `DS2` respectively such that no two datasets in this collection of 60 streams overlapped. This is so that Assumption 6.1 is satisfied by the offline dataset.

The description of the rest of the dataset on which we run the online algorithms is in Table 4. In addition, we also run `DSpot` (Siffer et al., 2017) with $p = 1 - 10^{-6}$ and set `num-init` to 500 and `depth` to 100 on the online algorithm dataset. All other parameters of `DSpot` was set to the suggested values in (Siffer et al., 2017). In addition to these, we also compare against two static baselines of using the first 500 and 100 samples respectively and set the threshold as the max of the observed samples. This threshold is not changed since.

The results of this experiment are reported in Table 5. From this table, we can see that incorporating offline data reduces the number of abstains for both datasets without changing the volume of anomalies raised. This shows that when there are offline datasets available to the algorithm, the online performance is improved.

	Num. of Streams	Average stream length	Total num samples
DS1	8377	10925.4	91521785
DS2	9657	10855.6	104832532

Table 4: Summary of data properties for real world AD case study.

		Algo 3	Algo 5	DSpot	$\tau^{(1)}$	$\tau^{(2)}$
DS1	Abstains ↓	13.6%	11.2%	1.2%	1.16%	0.05%
	Anomaly ↓	**.008**%	**.008**%	2.9%	0.02%	0.95%
DS2	Abstains ↓	13.3%	9.9%	1.3%	1.25%	1.08%
	Anomaly ↓	.006%	**.005**%	3.9%	0.07%	0.06%

Table 5: Case study performance. All baseline algorithms report anomaly rates much higher than the targeted 10^{-6}, overwhelming the alarm system. By incorporating offline datasets, we are able to further decrease abstains and achieve a lower rate of reported anomaly.

G.3.

We include additional experiments as follows. With the same underlying AD scoring algorithm (isolation forest for ForestCover, and a 3-layer MLP for Mammography), our confidence sequence algorithm (Algorithm 1 in paper) can achieve a higher F1 score and fewer mistakes, with similar or less number of abstains, compared to a static threshold. This further demonstrates that our thresholding algorithm improves performance across datasets and base anomaly scoring algorithms.

	ForestCover (first 1500 as holdout)		Mammography (first 300 as holdout)	
	Ours	Static Threshold	Ours	Static Threshold
Abstain %	1.5%	1.5%	2.1%	5%
FP + FN	645	1603	56	83
F1	0.39	0.33	0.51	0.43

Table 6: Performance comparison of algorithms on ForestCover and Mammography datasets

H. Proof of Lower Bound

The proof is based on the classical hypothesis testing lower bounds due to Neyman and Pearson. Throughout this section, we will fix a $T \in \mathbb{N}$ sufficiently large and let $\mathbb{P}_0 := N(0, 1)$ be the standard normal distribution and $\mathbb{P}_1 := N(\frac{1}{\sqrt{T}}, 1)$ as an

unit variance normal distribution with mean $1/\sqrt{T}$.

Theorem 14 (Neyman-Pearson)**.** For any $T \in \mathbb{N}$, given T i.i.d.. samples $S_{1:T}$, we want to distinguish between the two hypothesis $H_0 : S_{1:T} \sim \mathbb{N}(0,1)$ or $H_1 : S_{1:T} \sim \mathcal{N}(1/\sqrt{T}, 1)$. Then, for any measurable function $\mathcal{A} : \mathbb{R}^T \to \{0,1\}$,

$$\max(\mathbb{P}_0[\mathcal{A}(S_{1:T} = 1], \mathbb{P}_0[\mathcal{A}(S_{1:T} = 1]) \geq \frac{1}{4}\exp(-\frac{1}{2}),$$

where $\mathbb{P}_i$ for $i \in \{0,1\}$ is the proabability distribution under hypothesis H_i.

Lemma 6. For every $p \in (0,1)$, there exists a constant $C_p \in (0,\infty)$ depending only on p, such that for $S \sim N(0,1)$

$$\mathbb{P}[S \in (Q^{(0)}(p), Q^{(1)}(p))] \geq \frac{C_p}{\sqrt{T}},$$

where $Q^{(0)}(p)$ is the pth quantile of a $N(0,1)$ distribution and $Q^{(1)}(p)$ is the quantile of a $N(\frac{1}{\sqrt{T}}, 1)$ distribution.

Now, a standard Chernoff bound gives the following result.

Lemma 7 (Chernoff bound)**.** Let $T \in \mathbb{N}$ and suppose $S_{1:T}$ are i.i.d., $N(0,1)$ random variables. Let $p \in (0,1)$. Then,

$$\mathbb{P}\left[\sum_{t=1}^{T} \mathbf{1}_{S_t \in (Q^{(0}(p), Q^{(1)}(p))} \leq \frac{C_p\sqrt{T}}{2}\right] \leq \exp\left(\frac{\sqrt{T}}{8}\right), \tag{22}$$

where C_p is from Lemma 6.

Theorem 15 (Main lower bound)**.** Denote by two distributions $\mathbb{P}_0 := N(0,1)$ and $\mathbb{P}_1 := N(\frac{1}{\sqrt{T}}, 1)$ as unit variance guassians with means 0 and $\frac{1}{\sqrt{T}}$ respectively. For $i \in \{0,1\}$, denote by $Q^{(i)}(p)$ as the pth quantile of distribution $\mathbb{P}_i$. For a threshold $\theta \in \mathbb{R}$ and sample $x \in \mathbb{R}$, denote by the indicator function $\widehat{Y}_t^{(\theta)}(x) := \mathbf{1}(x \geq \theta)$. Similarly, denote by the indicator function $Y_t^{(i)}(x) = \mathbf{1}(x \geq Q^{(i)})$ for $i \in \{0,1\}$. Let $\mathcal{A} : \mathbb{R}^T \to \{0,1\}$ be any measurable hypothesis testing function. Then, there exists $i \in \{0,1\}$ such that when $S_{1:T}$ are i.i.d. with distribution $\mathbb{P}_i$, then with probability at-least $\frac{1}{8} - \exp\left(\frac{\sqrt{T}}{8}\right)$,

$$\sum_{t=1}^{T} \mathbf{1}(\widehat{Y}_t^{(Q^{(\mathcal{A}(S_{1:T}))}(p))}(S_t) \neq Y_t^{(i)}(S_t)) \geq C_p\sqrt{T},$$

where C_p is from Lemma 6.

In words, the above theorem states that even when the true distribution is known to be one of two possibilities and the optimal threshold is chosen using any measurable function, $O(\sqrt{T})$ mistakes is un-avoidable.

Proof. Pick $\mathcal{A} : \mathbb{R}^T \to \{0,1\}$ by a measurable function that tests the hypothesis $H_0 : S_{1:T}$ are i.i.d. with distribution $\mathbb{P}_0$ and $H_1 : S_{1:T}$ are i.i.d. with distribution $\mathbb{P}_1$. From Theorem 14, we know that for this given test $\mathcal{A}$, there exists $i \in \{0,1\}$ such that

$$\mathbb{P}_i[\mathcal{A}(S_{1:T}) \neq i] \geq \frac{1}{8}. \tag{23}$$

Without loss of generality, lets suppose $i = 0$, i.e., given a hypothesis test $\mathcal{A}$, $\mathbb{P}_0[\mathcal{A}(S_{1:T}) \neq 0] \geq \frac{1}{8}$. We will show in the rest of the proof that for this $\mathcal{A}$, if $S_{1:T}$ are i.i.d. with distribution $\mathbb{P}_0$, $\sum_{t=1}^{T} \mathbf{1}(\widehat{Y}_t^{(Q^{(\mathcal{A}(S_{1:T}))}(p))} \neq Y_t^{(0)}) \geq C_p\sqrt{T}$ holds with probability at-least $\frac{1}{8} - \exp\left(\frac{\sqrt{T}}{8}\right)$.

To do so, let $S_{1:T}$ be i.i.d. $\mathbb{P}_0$. From Lemma 7, we know that

$$\mathbb{P}_0\left[\sum_{t=1}^{T} \mathbf{1}_{S_t \in (Q^{(0}(p), Q^{(1)}(p))} \leq \frac{C_p\sqrt{T}}{2}\right] \leq \exp\left(\frac{\sqrt{T}}{8}\right).$$

Thus, from an union bound, we know that for the given $\mathcal{A}$,

$$\mathbb{P}_0\left[\underbrace{\mathcal{A}(S_{1:T}) \neq 0, \sum_{t=1}^{T} \mathbf{1}_{S_t \in (Q^{(0)}(p), Q^{(1)}(p))} \geq \frac{C_p\sqrt{T}}{2}}_{\text{Event } \mathcal{H}}\right] \geq \frac{1}{8} - \exp\left(\frac{\sqrt{T}}{8}\right).$$

Observe that under event $\mathcal{H}$, for all $x \in \mathbb{R}$, we have the equality

$$\mathbf{1}(\widehat{Y}_t^{(\mathcal{A}(S_{1:T}))}(x) \neq Y_t^{(0)}(x)) = \mathbf{1}(x \in (Q^{(0)}(p), Q^{(1)}(p)).$$

Thus, under event $\mathcal{H}$, we have

$$\sum_{t=1}^{T} \mathbf{1}(\widehat{Y}_t^{(\mathcal{A}(S_{1:T}))}(S_t) \neq Y_t^{(0)}(S_t)) = \sum_{t=1}^{T} \mathbf{1}_{S_t \in (Q^{(0)}(p), Q^{(1)}(p))} \geq \frac{C_p\sqrt{T}}{2},$$

holding with probability at-least $\frac{1}{8} - \exp\left(\frac{\sqrt{T}}{8}\right)$. □

The theorem asserts that if an algorithm does not abstain, then even in a stationary stream, $O(\sqrt{T})$ mistakes will occur with high probability.