Online Heavy-tailed Change-point detection

Abishek Sankararaman*[1] **Balakrishnan (Murali) Narayanaswamy**[1]

[1]AWS AI Labs

Abstract

We study algorithms for online change-point detection (OCPD), where samples that are potentially heavy-tailed, are presented one at a time and a change in the underlying mean must be detected as early as possible. We present an algorithm based on clipped Stochastic Gradient Descent (SGD), that works even if we only assume that the second moment of the data generating process is bounded. We derive guarantees on worst-case, finite-sample false-positive rate (FPR) over the family of all distributions with bounded second moment. Thus, our method is the first OCPD algorithm that guarantees finite-sample FPR, even if the data is high dimensional and the underlying distributions are heavy-tailed. The technical contribution of our paper is to show that clipped-SGD can estimate the mean of a random vector and simultaneously provide confidence bounds at all confidence values. We combine this robust estimate with a union bound argument and construct a sequential change-point algorithm with finite-sample FPR guarantees. We show empirically that our algorithm works well in a variety of situations, whether the underlying data are heavy-tailed, light-tailed, high dimensional or discrete. No other algorithm achieves bounded FPR theoretically or empirically, over all settings we study simultaneously.

1 INTRODUCTION

Online change-point detection (OCPD) is a fundamental problem in statistics where instantiations of a random variable are presented one after another and we want to detect if some parameter or statistic corresponding to the underlying data generating distribution has changed. This problem has been widely studied in machine learning, mathematical statistics and information theory over the past century. In part, this is due to the wide-ranging applications of OCPD to computational biology [Muggeo and Adelfio, 2011], online advertising [Zhang et al., 2017], cyber-security [Osanaiye et al., 2016, Kurt et al., 2018, Polunchenko et al., 2012], cloud-computing [Maghakian et al., 2019], finance [Lavielle and Teyssiere, 2007], medical diagnostics [Yang et al., 2006, Gao et al., 2018] and robotics [Konidaris et al., 2010]. We refer interested readers to the recent surveys of [Aminikhanghahi and Cook, 2017] and [Xie et al., 2021] for details of applications of OCPD. These surveys build upon the classical texts in change-point detection obtained over the last decade [Basseville et al., 1993, Tartakovsky, 1991, Krichevsky and Trofimov, 1981].

Classical results for OCPD have focused on algorithms that assume known distributions for either one or both of the pre- and post-change data [Wald, 1992, Page, 1954, Shiryaev, 2007, Lorden, 1971, Pollak, 1985, Ritov, 1990, Moustakides, 1986, Tartakovsky, 1991]. In recent years, algorithms have been developed for cases when the pre- and post- change distributions are unknown, but belong to a parametric class such as the exponential family [Lai and Xing, 2010, Fryzlewicz, 2014, Frick et al., 2014, Cho, 2016]. Non-parametric algorithms have been developed in [Padilla et al., 2021, Madrid Padilla et al., 2021] and the references therein, but they only give asymptotic guarantees. The algorithms of [Adams and MacKay, 2007, Lai and Xing, 2010, Maillard, 2019, Alami et al., 2020] have finite-sample guarantees, but either rely on parametric assumptions such as an exponential family, or on tail assumptions such as sub-gaussian distribution families. The works of [Bhatt et al., 2022] and [Li and Yu, 2021] build upon the work in [Niu and Zhang, 2012], and give algorithms for multiple change-points with possibly heavy-tailed data in the *offline* case with all data available up-front. The works of [Wang and Ramdas, 2022, 2023, Shekhar and Ramdas, 2023] give OCPD algorithms

* Correspondence to Abishek Sankararaman : abisanka@amazon.com

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

for heavy-tailed, but uni-variate data.

In many modern applications such as cloud-computing and monitoring, data is known to often be heavy-tailed [Nair et al., 2022, Loiseau et al., 2010, Nizam et al., 2016] and too complex to model with any simple parametric family [Barnett and Onnela, 2016, Hallac et al., 2015, Dartmann et al., 2019]. Given the velocity, variety and volume of modern data streams, performance of change-point detection is measured through false-positive rates in order to combat alert fatigue [Ruff et al., 2021], and algorithms must work for streams that have multiple change points. Motivated by these requirements, we seek an OCPD algorithm that simultaneously meet the following desiderata : it *(i)* detects multiple change-points, *(ii)* makes no parametric assumptions on the distribution of data, *(iii)* works with potentially heavy-tailed data, *(iv)* works for high-dimensional data streams, and *(v)* guarantees finite sample FPR.

1.1 MAIN CONTRIBUTIONS

Our paper is the first to give an online algorithm satisfying all the 5 desiderata listed above. Specifically, our algorithm gives finite sample guarantees for FPR and detection-delay without assuming that data comes from a specific parametric family or assuming strong tail conditions - such as that the data have sub-gaussian distributions. No previous algorithm for OCPD simultaneously achieves all desiderata. Our main technical contribution is to provide a `clipped-SGD` algorithm with finite sample confidence bounds for heavy-tailed mean estimation, *that hold for all confidence values simultaneously*, a result of independent interest. We use these bounds to build a OCPD algorithm with finite sample FPR.

We further show good empirical performance across a variety of data streams with heavy-tailed, light-tailed, high dimensional or discrete distributions. However while our algorithm is designed to work across different distributions, we observe theoretically and empirically that when data has additional structure such as being one-dimensional with sub-gaussian tails or is binary, then specialized OCPD algorithms for those cases yield better results than our method. Closing these gaps is an ongoing direction of research.

2 PROBLEM SETUP

At each time t, a random vector $X_t \in \mathbb{R}^d$ is revealed to an OCPD algorithm. X_t has a probability measure and expectation denoted by $\mathbb{P}_t$ and $\mathbb{E}_t$ respectively, and mean $\mathbb{E}_t[X_t] \in \mathbb{R}^d$. Subsequently, using all the samples observed so far - $X_1, \cdots, X_t$ - the algorithm outputs a binary decision denoting whether a change in mean has occurred since time $t = 1$ or the last time a change was output by the algorithm, whichever is larger. The goal of the OCPD algorithm is identify the change points as quickly as possible after they occur, with bounded false-positive rate (FPR). The observed datum $(X_t)_{t \geq 1}$ are independent, although not identically distributed with piece-wise constant mean.

Definition 2.1 (Piece-wise constant mean process)**.** Let T be the time horizon (stream-length) and let $Q_T < T$ be the total number of change-points. A set of strictly increasing time-points $1 < \tau_1 < \tau_2 \cdots < \tau_{Q_T+1} := T + 1$ are called change-points, if for all $c \in \{1, \cdots, Q_T\}$

- $\forall t \in [1, T]$, $X_t \sim \mathbb{P}_t$ independently.
- $\forall t \in [\tau_c, \tau_{c+1})$, the mean $\mathbb{E}_t[X_t] := \theta_c$ of the observation is constant and does not depend on t.
- $\forall c \in [1, Q_T]$, $\theta_c \neq \theta_{c+1}$.

Thus, a piece-wise constant mean process is identified by the quadruple $\mathfrak{M} := (T, Q_T, (\tau_c)_{c=1}^{Q_T}, (\mathbb{P}_t)_{t=1}^{T})$. Throughout, we use probability and expectation operators $\mathbb{P}$ and $\mathbb{E}$, to denote the joint product probability distribution $(\mathbb{P}_t)_{t=1}^{T}$.

2.1 ASSUMPTIONS

Let $\mathcal{P}$ be a family of probability measures on $\mathbb{R}^d$ such that the probability distributions $\mathbb{P}_t$, for all t, are from this family, i.e., $\mathbb{P}_t \in \mathcal{P}$, $\forall t \in [1, T]$. Throughout this paper, we make the following non-parametric assumptions on the family $\mathcal{P}$.

Assumption 2.2. There exists a convex compact set $\Theta \subset \mathbb{R}^d$ known to the algorithm, such that for all $\mathbb{P} \in \mathcal{P}$, $\mathbb{E}_{X \in \mathbb{P}}[X] \in \Theta$. In words, the mean of all the distributions in the family belong to a known bounded set Θ such that $\max_{\theta_1, \theta_2 \in \Theta} \|\theta_1 - \theta_2\| := G$.

Assumption 2.3. There exists $\sigma > 0$ known to the algorithm, such that for all $\mathbb{P} \in \mathcal{P}$ and $\theta \in \Theta$, $\mathbb{E}_{X \sim \mathbb{P}}[\|X - \mathbb{E}_{X \sim \mathbb{P}}[X]\|_2^2] \leq \sigma^2$. In words, the second moment is uniformly bounded for all distributions in $\mathcal{P}$.

These assumptions are very general and encompass a wide range of families such as any bounded distribution, the set of sub-Gaussian distributions and heavy-tailed distributions that do not have finite higher moments. We seek algorithms that work without knowing the length of the data stream, the number of change-points and that do not make any assumptions on the underlying distributions generating the samples, beyond Assumptions 2.2 and 2.3.

2.2 PERFORMANCE MEASURES

Any OCPD algorithm is measured by two performance metrics – *(i)* False-positive rate and *(ii)* Detection delay. We set notation to define these measures.

Notation 2.4. For every $1 \leq r \leq s < T$, we denote by $X_{r:s} := (X_r, X_{r+1}, \cdots, X_s)$ to be the set of observed vectors from time r to time s, with both end-points r and s inclusive.

Definition 2.5 (OCPD algorithm)**.** A sequence of measurable functions $\mathcal{A} := (\mathcal{A}_t)_{t \geq 1}$ is called an OCPD algorithm if for every time $t \geq 1$, $\mathcal{A}_t \in \{0, 1\}$ and is measurable with respect to the sigma algebra generated by $X_{1:t}$. The interpretation is that if $\mathcal{A}_t = 1$ for some t, then the algorithm has detected a change at time t and if $\mathcal{A}_t = 0$, no change is detected at time t.

Notation 2.6. For an OCPD algorithm $\mathcal{A}$ and for all $t \in [T]$, denote by $R^{(\mathcal{A})}(t) \in \mathbb{N}$ to be the random variable denoting the number of detections made till time t, i.e., $R^{(\mathcal{A})}(t) = \sum_{s=1}^{t} \mathcal{A}_s$.

Notation 2.7. For an OCPD algorithm $\mathcal{A}$, and every $r \in \mathbb{N}$ and, denote by $t_r^{(\mathcal{A})}$ as the stopping time

$$t_r^{(\mathcal{A})} := \min(\inf\{t \in [0, T] \text{ s.t. } R^{(\mathcal{A})}(t) \geq r\}, T + 1),$$

where the $\inf$ of an empty set is defined to be ∞. In words, $t_r^{(\mathcal{A})}$ is the stopping time when the OCPD algorithm detects a change for the rth time, or $T + 1$, whichever is larger.

Definition 2.8 (False Positive Detection)**.** The rth detection of an OCPD algorithm $\mathcal{A}$ is said to be a False Positive, if there exists no change-point between the $r - 1$th and the rth detection. Formally, denote by the indicator (random) variable $\chi_r^{(\mathcal{A})} = \mathbf{1}(\nexists c \in [1, Q_T] \text{ s.t. } \tau_c \in (t_{r-1}^{(\mathcal{A})}, t_r^{(\mathcal{A})}])$ to denote if the rth detection of $\mathcal{A}$ is a false-positive. Note that by definition, on the event that $R^{(\mathcal{A})}(T) < r$, $\chi_r^{(\mathcal{A})} = 0$.

Definition 2.9 (False Positive Rate (FPR))**.** An OCPD algorithm $\mathcal{A}$ is said to have false-positive rate bounded by $\delta \in (0, 1)$ if

$$\sup_{\mathfrak{M}} \mathbb{E}\left[\frac{\sum_{r=1}^{T} \chi_r^{(\mathcal{A})}}{R^{(\mathcal{A})}(T)} \mathbf{1}(R^{(\mathcal{A})}(T) > 0)\right] \leq \delta. \qquad (1)$$

In words, an OCPD algorithm $\mathcal{A}$ has bounded false positive rate, if for every piece-wise constant mean process $\mathfrak{M}$, the expected fraction of false-positives made by the algorithm $\mathcal{A}$ is bounded by δ. In Equation (1), we take the sum till T because that is the maximum number of possible change points detected. If an algorithm only detects $s < T$ change points, then by definition $\chi_r^{(\mathcal{A})} = 0$ for all $r > s$.

Definition 2.10 (Worst-case Detection Delay)**.** For $n \in \mathbb{N}$ and $\Delta > 0$, let $X_1, X_2, \cdots, X_n, X_{n+1}, \cdots$ be an infinite stream with the following distribution. For every $t < n$, $X_t \overset{\text{ind}}{\sim} \mathbb{P}_t$ with $\mathbb{E}_{X \sim \mathbb{P}_t}[X] = \theta_1 \in \Theta$ and for every $t \geq n$, $X_t \overset{\text{ind}}{\sim} \mathbb{P}_t$ with $\mathbb{E}_{X \sim \mathbb{P}_t}[X] = \theta_2 \in \Theta$ with $\|\theta_1 - \theta_2\| = \Delta$. Let $\mathfrak{M}^{(n,\Delta)}$ denote all such infinite piece-wise constant mean process. An algorithm $\mathcal{A}$ is said to have worst-case detection delay $\mathcal{D}(\Delta, n, \delta')$, if

$$\sup_{\mathfrak{M}^{(n,\Delta)}} \mathbb{P}\Big[\inf\{t > n \colon \mathcal{A}_t = 1\} - n \geq \mathcal{D}(\Delta, n, \delta')\Big] \leq \delta' \qquad (2)$$

holds for all $n \in \mathbb{N}$, $\Delta > 0$ and $\delta' \in (0, 1)$.

In words, the detection delay function $\mathcal{D}(\Delta, n, \delta')$ is such that for every admissible process $\mathfrak{M}^{(n,\Delta)}$ that has a single change-point at time n with jump magnitude Δ, algorithm $\mathcal{A}$ detects the change-point before time $n + \mathcal{D}(\Delta, n, \delta')$, with probability at-least $1 - \delta'$. Note that the delay metric is measured on data streams with exactly one change-point. Defining detection delay for streams with multiple change-points is ambiguous as there could be missed detections, with only a subset of the change-points being detected [Alami et al., 2020], [Maillard, 2019]. The main question this paper studies is

For each $\delta \in (0, 1)$, does there exists an OCPD algorithm with FPR bounded by δ and having small worst-case detection-delay that only makes Assumptions 2.2 and 2.3 ? .

Observe that it is trivial to achieve a FPR of 0 for example the constant function where $\mathcal{A}(\cdot) = 0$, i.e., an algorithm that never detects change-point at all. However, this algorithm has a worst-case detection-delay of ∞, i.e., $\mathcal{D}(\Delta, n, \delta') = +\infty$ for all $\Delta > 0$, $n \in \mathbb{N}$ and $\delta' \in (0, 1)$. Thus, the challenge is to design an algorithm that satisfies the FPR constraint of δ while having small, finite worst-case detection delay, without making parametric assumptions on the underlying data generating distributions.

3 ONLINE ROBUST MEAN ESTIMATION

The central workhorse of our change-point detection algorithm is heavy-tailed online mean estimation. Suppose $X_1, X_2, \cdots$ are a sequence of independent random vectors, with the means $\mathbb{E}_t[X_t] = \theta^* \in \Theta$ being a constant independent of time t. Let $(\widehat{\theta}_t)_{t \geq 1}$ be a sequence of random variables such that $\widehat{\theta}_t$ is an estimate of θ based on the samples $X_1, \cdots, X_t$ defined through clipped-SGD algorithm described as follows. For a given non-negative sequence $(\eta_t)_{t \geq 1}$ and $\lambda > 0$, the estimate $\widehat{\theta}_0 \in \Theta$ is arbitrary, $\widehat{\theta}_t$ for each $t \geq 1$, is given by

$$\widehat{\theta}_t := \prod_{\Theta}(\widehat{\theta}_{t-1} - \eta_t \text{clip}(X_t - \widehat{\theta}_{t-1}, \lambda)), \qquad (3)$$

where, $\prod_{\Theta}$ is the projection operator onto the convex compact set Θ and for every $x \in \mathbb{R}^d$ and $\lambda > 0$,

$$\text{clip}(x, \lambda) = x \min\left(1, \frac{\lambda}{\|x\|}\right). \qquad (4)$$

Our main result on the convergence of the estimator $\widehat{\theta}_t$ to the true θ^* with increasing number of samples t is the following.

Theorem 3.1. *For all times $t \geq 1$, when clipped SGD in Equation (3) is run with $\lambda = 2G$ and $\eta_t = \frac{2}{(t+\gamma)}$ for $\gamma = \max\left(120\lambda\sigma(\sigma + 1), 320\sigma^2 + 1\right)$, then for every $t \geq 1$*

and every $\delta \in (0,1)$,

$$\mathbb{P}\left[\|\widehat{\theta}_t - \theta^*\|_2^2 \geq \mathcal{B}(t,\delta)\right] \leq \frac{\delta}{t(t+1)},$$

where

$$\mathcal{B}(t,\delta) := C_t \Bigg[\frac{\gamma^2 G^2}{(t+1)^2} + \left(\frac{16\sigma^2}{\lambda} + 4\sigma^2 \right) \frac{1}{2(t+1)} + \frac{96\lambda^2 \ln\left(\frac{2t^2(t+1)}{\delta}\right)\sigma(\sigma+1)}{(t+\gamma)\sqrt{t+1}} \Bigg], \quad (5)$$

and $C_t = \max\left(\frac{1024\sigma^4}{G^2\lambda^2}, \frac{8\lambda\sqrt{\ln\left(\frac{2t^2(t+1)}{\delta}\right)}}{\gamma^2 G}\right)$.

Corollary 3.2. *There exists an universal constant $A > 0$ such that for all $t \geq 1$, when clipped SGD in Equation (3) is run with parameters in Theorem 3.1*

$$\mathbb{P}\left[\|\widehat{\theta}_t - \theta^*\| \geq A \max\left(\frac{\sigma^3}{\sqrt{t}}, \frac{\sigma\sqrt{\ln\left(\frac{t^3}{\delta}\right)}}{\sqrt{t}}\right)\right] \leq \frac{\delta}{t(t+1)},$$

holds for every $\delta \in (0,1)$.

Proof is in Appendix in Section B and uses tools from [Bubeck, 2015], [Gorbunov et al., 2020, Tsai et al., 2022] and [Victor, 1999].

Remark 3.3. Compared to [Tsai et al., 2022], we do not need the failure probability δ in the input and we can give simultaneous confidence intervals for all failure probabilities δ. In contrast, the algorithm of [Tsai et al., 2022] requires $\delta \in (0,1)$ as an input and only guarantees that the estimate mean is close to the true mean, upto error probability of δ. However, the bound in Theorem 3.1 is off by logarithmic factors compared to [Tsai et al., 2022]. Concretely, $C_t = O(1)$ for the algorithm of [Tsai et al., 2022], while it is $O(\log(t/\delta))$ for us. This is the price to have confidence intervals hold for all failure probabilities simultaneously as opposed to just having one single failure probability.

Remark 3.4. Compared to the setting of [Tsai et al., 2022], our setting is *weaker* as we assume that the domain Θ is compact with finite diameter G. This is what enables us to use an appropriately tuned learning rate and clipping parameter to make the algorithm any-time and obtain confidence intervals at all failure probabilities simultaneously. It is an open question whether the assumption that Θ is compact can be relaxed and if we can still make guarantees confidence interval holding for all failure probabilities δ for all t for heavy tailed distributions.

Remark 3.5. The constants in Theorem 3.1 are not optimal. In Section 5, we suggest an alternative set of constants that work well empirically across variety of settings.

Remark 3.6. There have been significant recent advances in robust mean estimation [Diakonikolas et al., 2020, Lugosi and Mendelson, 2021, Depersin and Lecué, 2022, Cherapanamjeri et al., 2020, Diakonikolas et al., 2022], that are known to provide near optimal error bounds. However, unlike our method, none of these algorithms can give confidence bounds for all confidence values simultaneously.

Remark 3.7. Theorem 4.3 in [Devroye et al., 2016] proves that it is impossible to get a finite sample confidence bound to hold for all $\delta \in (0,1)$. Our result does not contradict this since the restriction on the allowable δ is *implicit* in Theorem 3.1. Equation (5) gives that, for every $t \in \mathbb{N}$, as $\delta \searrow 0$, $\mathcal{B}(t,\delta) \nearrow \infty$. However, from Assumption 2.2, if $\mathcal{B}(t,\delta) \geq G$, then the statement of Theorem 3.1 is vacuous. Thus, Theorem 3.1 gives a non-vacuous bounds only for $\delta \in (\delta_{min}^{(t)}, 1)$ where $\delta_{min}^{(t)} := \inf_{\delta > 0}\{\mathcal{B}(t,\delta) < G\}$.

3.1 UNIFORM OVER TIME BOUND

As a corollary of Theorem 3.1, we get the following bound that holds uniformly over all time.

Corollary 3.8. *There exists an universal constant $A > 0$ such that, when clipped SGD in Equation (3) is run with parameters in Theorem 3.1,*

$$\mathbb{P}\left[\exists t \in \mathbb{N} : \|\widehat{\theta}_t - \theta^*\| \geq A \max\left(\frac{\sigma^3}{\sqrt{t}}, \frac{\sigma\sqrt{\ln\left(\frac{t^3}{\delta}\right)}}{\sqrt{t}}\right)\right] \leq \delta,$$

holds for every $\delta \in (0,1)$.

The proof follows by taking an union bound over all $t \geq 1$, i.e., summing over $t \geq 1$ on both the LHS and RHS of Corollary 3.8 and noticing that $\sum_{t \geq 1} \frac{1}{t(t+1)} = 1$. The bounds in Theorem 3.1 and Corollary 3.8 are *dimension free*, i.e., the term d does not appear in the bounds. The moment bound σ plays the role of dimension. In particular, suppose that all distributions in the family $\mathcal{P}$ have covariance matrices bounded in the positive semi-definite sense by $\Sigma \in \mathbb{R}^{d \times d}$. In this case, by definition, $\sigma^2 \leq \text{Trace}(\Sigma)$ and plays the role of dimension.

In the special case when the samples $(X_t)_{t \geq 1}$ are i.i.d. with sub-gaussian distributions with mean θ^* and covariance matrix Σ, [Abbasi-yadkori et al., 2011, Maillard, 2019, Chowdhury et al., 2022] show that for all $\delta \in (0,1)$,

$$\mathbb{P}\left[\exists t \in \mathbb{N} : \left\|\frac{1}{t}\sum_{s=1}^{t} X_s - \theta^*\right\| \geq \sqrt{2\lambda_{max}(\Sigma)\left(1 + \frac{1}{t}\right)\ln\left(\frac{(t+1)^d}{\delta}\right)}\right] \leq \delta, \quad (6)$$

holds, where $\lambda_{max}(\Sigma)$ is the highest eigen-value of the covariance matrix Σ. Thus, for the special case of sub-gaussian distributions, Equation (6) has a better dependence on time t compared to our Corollary 3.8. The improved dependence on time arises as Equation (6) is based on the construction of a self normalized martingale and using the martingale stopping theorem to obtain uniform over time bounds while Corollary 3.8 is based on a simple union bound.

However, Equation (6) is not dimension free and depends on the scale of the problem through the term $d\lambda_{max}(\Sigma)$ which by definition is larger than Trace(Σ). In many high dimensional settings, $d\lambda_{max}(\Sigma)$ is much larger than Trace(Σ) and thus algorithms and bounds depending explicitly on d is undesirable [Wainwright, 2019, Lugosi and Mendelson, 2019]. For the uni-variate heavy-tailed distributions, a sequence of works [Wang and Ramdas, 2022, 2023] establish confidence bounds with sharp dependence on time by extending the martingale recipe developed in [Howard et al., 2021]. In our draft, we are able to get dimension free bounds for heavy-tailed distributions, but at the cost of a compactness Assumption 2.2 that are not needed in [Abbasi-yadkori et al., 2011]. It is an open question if we can get dimension-free bounds with the improved time-dependence of the kind in Equation (6) without the compactness assumption.

4 CHANGE-POINT DETECTION ALGORITHM

Our algorithm is described in Algorithm 1 and is based on the following idea. A change point is detected in the time-interval $[r, t]$ if there exists $r < s < t$ such that confidence interval around the estimated mean of the observations $X_{r:s}$ is separated from the confidence interval around the estimated mean of the observations $X_{s+1:t}$. Further, in order to accommodate multiple change-points, the algorithm *restarts* after every change detection, similar to [Alami et al., 2020]. It is known that standard empirical mean is a poor estimator when the underlying distributions can potentially be heavy-tailed, as its confidence interval under only assumptions in 2.3 is wide [Lugosi and Mendelson, 2019]. To attain better confidence intervals, we use the `clipped-SGD` Algorithm 1 that gives a confidence interval for the estimated mean for every failure probability $\delta \in (0, 1)$ simultaneously. Having multiple confidence intervals is crucial as we show that adaptively testing different intervals of times at different carefully chosen confidence intervals (Line 8 of Algorithm 1) leads to the bounded FPR guarantee.

4.1 CONNECTIONS TO GLR

Restating our algorithm, a change point is detected in a time-interval $[t_0, t]$ if

$$\exists s \in (t_0, t) \text{ s.t. } \|\widehat{\theta}_{t_0:s} - \widehat{\theta}_{s+1:t}\|^2 \geq \mathcal{C}(t_0, s, t, \delta),$$

where the function $\mathcal{C}(\cdot)$ is given in Line 8 of Algorithm 1. In the above re-statement, the estimates $\widehat{\theta}_{t_0:s}$ and $\widehat{\theta}_{s+1:t}$ are robust estimates of the mean based on the set of observations $\{X_{t0}, \cdots, X_s\}$ and $\{X_{s+1}, \cdots, X_t\}$ respectively. The `Improved-GLR` of [Maillard, 2019] uses a detector that is structurally similar to the above equation except that they *(i)* use the empirical mean as they are dealing with sub-gaussian random variables, and *(ii)* use a function $\mathcal{C}(\cdot)$ derived from the Laplace method that gives confidence bounds with better dependence on time, but is not dimension free. In contrast, we use the robust mean estimator given by clipped-SGD and the function $\mathcal{C}(\cdot)$ is derived from the confidence guarantees that only require the existence of the second moment and make no other tail assumptions and yields dimension free bounds. The cost however is that the confidence bound derived from clipped SGD has a weaker dependence on time compared to that obtained by the Laplace's method [Maillard, 2019].

4.2 FALSE-POSITIVE GUARANTEE

We will prove the following result on Algorithm 1. For a given process $\mathfrak{M}$, and every $r \in \mathbb{N}$, denote by the deterministic time $\tau_c^{(r)} := \inf\{\tau_c : \tau_c > r\}$ be the first change-point after time r.

Theorem 4.1 (False Positives). *When Algorithm 1 is run with parameters $\lambda = 2G$, $\eta_t = \frac{2}{(t+\gamma)}$ for $\gamma = \max\left(120\lambda\sigma(\sigma+1), 320\sigma^2+1\right)$ and $\delta \in (0, 1)$,*

$$\sup_{\mathfrak{M},r} \mathbb{P}[\exists t \in [r, \tau_c^{(r)}), \text{ s.t. } \mathcal{A}_t = 1 | \mathcal{A}_r = 1] \leq \delta,$$

holds almost-surely.

Proof is in Appendix in Section C.1. This result states that with probability at-most δ, a true change-point *does not* lie between any two consecutive detections made by the algorithm. This theorem implies the following lemma.

Lemma 4.2. *Under the conditions of Theorem 4.1, the FPR condition in Equation (1) holds.*

The proof is in the Appendix in Section C.2. We emphasize that the guarantee in 4.2 is a *worst-case guarantee*. In other words, no matter the underlying distribution, as long as Assumptions 2.2 and 2.3 are met, Algorithm 1 will not have more than a δ fraction of false-positives.

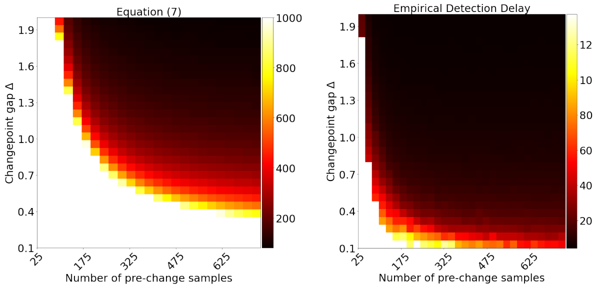
4.3 WORST-CASE DETECTION DELAY GUARANTEE

Lemma 4.3. *If Algorithm 1 is run with the parameters from Theorem 4.1, then for every $n \in \mathbb{N}$, $\Delta > 0$ and $\delta' \in (0, 1)$*

$$\mathcal{D}(n, \Delta, \delta') \leq \inf \left\{ d \in \mathbb{N} : \Delta^2 \geq \mathcal{B}\left(n-1, \frac{\delta'}{2}\right) + \right.$$

Algorithm 1 Online `Clipped-SGD` Change Point Detection

```
Input: $(\eta_t)_{t\geq 1}$, $\lambda > 0$, $\theta_0 \in \Theta$, $\delta \in (0,1)$ the FPR guarantee
$r \leftarrow 1$
$\widehat{\theta}_{t,t-1} \leftarrow \theta_0$, for all $t \geq 1$.
Set Num-change-points $\leftarrow 0$
for each time $t = 1, 2, \cdots,$ do
    Receive sample $X_t$
    $\widehat{\theta}_{s,t} \leftarrow \prod_\theta(\widehat{\theta}_{s,t-1} - \eta_{t-s}\text{clip}(X_t - \widehat{\theta}_{s,t-1}, \lambda))$, for every $r \leq s \leq t$.
    if $\exists s \in (r,t)$ such that $\|\widehat{\theta}_{r:s} - \widehat{\theta}_{s+1:t}\|_2^2 > \mathcal{B}\left(s-r, \frac{\delta}{2(t-r)(t-r+1)}\right) + \mathcal{B}\left(t-s-1, \frac{\delta}{2(t-r)(t-r+1)}\right)$ {$B(\cdot,\cdot)$ is defined in Equation (5)} then
        Set $\mathcal{A}_t \leftarrow 1$ {Change point detected}
        $r \leftarrow t+1$
        Set Num-change-points $\leftarrow$Num-change-points $+1$ {Increment number of change-points detected}
    else
        Set $\mathcal{A}_t \leftarrow 0$
    end if
end for
```



(a) (b)

Figure 1: Figure (a) plots the heat-map of $\mathcal{D}(n, \Delta, \delta')$ from Lemma 4.3 for fixed $\delta' = 0.1$. The white cells represent infinity. Figure (b) plots the 90th quantile ($\delta' = 0.1$) of the observed delay for Pareto distribution $d = 32$ over 30 runs. As can be seen, the observed detection delay in (b) is much smaller than the worst case delay in (a).

$$\mathcal{B}\left(d, \frac{\delta'}{2}\right) + \mathcal{B}\left(n-1, \frac{\delta}{2(n+d+1)(n+d)}\right) + \mathcal{B}\left(d, \frac{\delta}{2(n+d+1)(n+d)}\right)\Bigg\}, \quad (7)$$

where $\mathcal{D}(\cdot)$ and $\mathcal{B}(\cdot)$ are in Eqns (2) and (5) respectively.

Proof is in the Appendix in Section D. Lemma 4.3 is an *upper bound on the worst case delay*. In other words, for any pre- and post-change distribution with norm of the means differing by Δ, Algorithm 1 will detect this change within delay of $\mathcal{D}(n, \Delta, \delta')$ with probability at-least $1 - \delta'$.

For many specific choices of pre- and post-change distribution families however, we expect the observed detection delay to be much smaller than predicted by Lemma 4.3. This bound is conservative as it is worst-case over all distributions. In Figure 1a we plot the bound in Lemma 4.3 for a fixed $\delta' = \delta = 0.1$ as n and Δ varies. We use the constants given in Section 5.1 to plot Figure 1a. In Figure 1b, we plot the empirically observed detection delay for a sequence of 32 dimensional Pareto distributed random vectors with shape parameter 2.01. As can be seen in Figure 1, the observed detection delay is much smaller than that indicated by Lemma 4.3, which is a worst case over all distributions.

Remark 4.4. In the special case when the observations are Bernoulli random variables, the `R-BOCPD` algorithm of [Alami et al., 2020] gives a smaller detection delay compared to ours – our detection delay bound in 4.3 has additional poly-logarithmic factors of $\log(n/\delta)$ and sub-optimal constants compared to `R-BOCPD`. However, our bound holds for *any* family of distributions, including high-dimensional and heavy tailed ones, while `R-BOCPD` can only be applied for Bernoulli distributions.

Corollary 4.5 (Un-detectable Change)**.** *If* $\Delta \leq \mathcal{O}\left(\frac{\log\left(\frac{n}{\delta}\right)}{\sqrt{n}}\right)$*, then* $\mathcal{D}(n, \Delta, \delta') \leq \infty$ *for all* $\delta' \in (0,1)$*, the delay bound in Lemma 4.3 is vacuous.*

Remark 4.6. The undetecable region consists of the grey/white areas of Figure 1a. However, since Lemma 4.3 is only an upper-bound, the fact that $\mathcal{D}(n, \Delta, \delta') = \infty$ *does not imply* that our algorithm cannot detect the change (cf. Figure 1b).

Remark 4.7. In the case of sub-gaussian, exponential families, [Maillard, 2019] give a lower bound for changes that not detectable by *any* algorithm. When Algorithm 1 is applied to sub-gaussian random variables from an exponential family, the detection-delay bound in Lemma 4.3 is sub-optimal by poly-logarithmic factors in $\log(n/\delta)$ compared to the lower bound. However, Algorithm 1 and the delay bound in Lemma 4.3 holds for any class of distributions subject to Assumptions 2.3 and 2.2, while the bounds in

[Maillard, 2019] only applies to sub-gaussian observations from a known exponential family.

Remark 4.8. In parallel work, the FCS detector of [Shekhar and Ramdas, 2023], when combined with the heavy-tailed Catoni-style confidence sequences of [Wang and Ramdas, 2023] is shown to detect univariate mean changes as long as $\Delta \preceq \sqrt{\log(\log(n)/\alpha)/n}$. Whether this rate is achievable in multivariate settings is left for future work

4.4 CHANGE-POINT LOCALIZATION

In practice, it is also crucial to identify the location where the change point occurred. In this section we describe how to modify Algorithm 1 to also output the estimate of the location of change in addition to just detecting the existence of a change. Recall that for every $r \in \mathbb{N}$, $\tau_r^{(\mathcal{A})} \in \mathbb{N} \cup \{\infty\}$ is the stopping time denoting the rth time, Algorithm $\mathcal{A}$ detects a change point. We modify Algorithm 1 by additionally outputting for every $r \in \mathbb{N}$, a time interval $[s_{r;1}^{(\mathcal{A})}, s_{2;r}^{(\mathcal{A})}] \subseteq [\tau_{r-1}^{(\mathcal{A})}, \tau_r^{(\mathcal{A})}]$ such that this is an interval that contains a change-point τ_c.

In order do so, we need an additional definition. For every $r < s < t$ and $\delta \in (0,1)$, denote by $\mathfrak{B}(r,s,t,\delta) \in \{0,1\}$ as the indicator variable that

$$\mathfrak{B}(r,s,t,\delta) = \mathbf{1}\left(\|\widehat{\theta}_{r:s} - \widehat{\theta}_{s+1:t}\|_2^2 > \mathfrak{B}_1 + \mathfrak{B}_2 \right), \quad (8)$$

where $\mathfrak{B}1 = \mathcal{B}\left(s-r, \frac{\delta}{2(t-r)(t-r+1)}\right)$ and $\mathfrak{B}2 = \mathcal{B}\left(t-s-1, \frac{\delta}{2(t-r)(t-r+1)}\right)$. The estimates of the location of change in a time-interval $[r,t]$ is all those time instants $s \in [r,t]$ such that $\mathfrak{B}(r,s,t,\delta) = 1$. Line 12 in Algorithm 2 in Section A in the Appendix, precisely defines the estimator. The empirical performance of this method is shown in Figure 3. We observe that this produces an accurate and sharp estimate of the change-point location in simulations.

5 EXPERIMENTS

In this section we give numerical evidence to show that Algorithm 1 can be applied across variety of settings. Line 8 of Algorithm 1 relies on confidence bounds for high-dimensional estimation where the global constants are not optimized. This is an artifact of the proof analysis in robust estimation [Lugosi and Mendelson, 2019, Vershynin, 2018]. Thus, we modify the absolute constants used in Theorem 4.1 as follows. We use $\gamma = \max\left(4\lambda\sigma(\sigma+1), 8\sigma^2+1\right)$ with the color red highlighting the changes from the definition in Theorem 4.1. The constant C_t is modified as follows $C_t = \max(\frac{0.5\sigma^4}{G^2\lambda^2}, \frac{1\lambda\sqrt{\ln\left(\frac{2t^2(t+1)}{\delta}\right)}}{\gamma^2 G})$. In addition, we use the following definition of $\mathcal{B}(\cdot,\cdot)$

$$\mathcal{B}(t,\delta) := C_t\left[\frac{\gamma^2 G^2}{t+1} + \left(\frac{2\sigma^2}{\lambda} + 1\sigma^2\right)\frac{1}{2(t+1)} + \frac{2\lambda^2 \ln\left(\frac{2t^2(t+1)}{\delta}\right)\sigma(\sigma+1)}{(t+\gamma)\sqrt{t+1}}\right], \quad (9)$$

where C_t and γ are the modified values stated above. Further, in all simulations we assume $\Theta = \mathbb{R}^d$ to be the whole plane.

5.1 SYNTHETIC SIMULATIONS

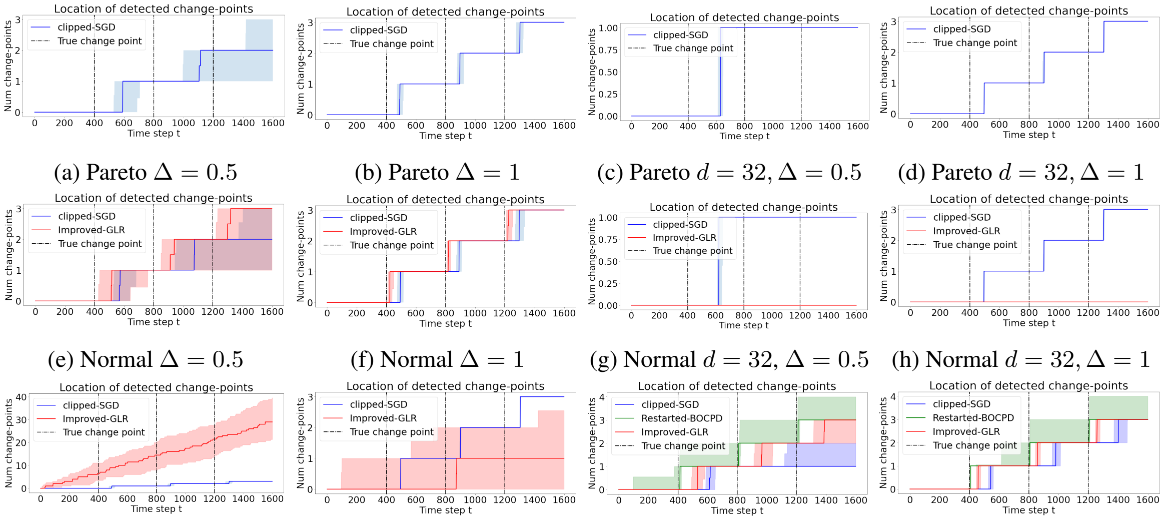
Here, demonstrate that Algorithm 1 with choice of hyper-parameters in Equation (9) is practical and can be applied across a variety of data generating distributions – either heavy-tailed, or high-dimensional or both and still obtains bounded false-positive rates and a much lower detection delay compared to what the conservative bound in Lemma 4.3 would indicate.

5.1.1 Setup

In Figure 2, we construct synthetic situations and introduce change-points with each change lasting 400 time-units. In all experiments, we choose the family of distributions $\mathfrak{M}$ such that $\sigma = 1$, $G = 12$. At each time t, a sample is drawn from the appropriate distribution that we detail below and presented to the change-point algorithm. The true-change points and the median detection times along with the 95 percentile upper and lower confidence bands are show in Figure 2. These are estimated by averaging 30 independent runs for each setting in Figure 2.

Heavy-tailed distribution: In Figures 2a, 2b and 2i, the sample at every time-point is drawn from a Pareto distribution with shape-parameter 2.01. This implies that the third central moment of the distribution is infinity. The mean of the samples in the time-durations $t \in [0,400) \cup [800,1200)$ is 0 in all figures and the mean at times $t \in [400,800) \cup [1200,1600)$ is $\Delta = 0.5, 1, 1$ respectively in Figures 2a, 2b and 2i. In Figure 2c, 2d and 2j, we consider the observation at time t to be 32 dimensional isotropic random vector with norm having Pareto distributions with shape parameter 2.01. The mean vector at times $[0,400) \cup [800,1200)$ is $0 \in \mathbb{R}^{32}$ and at times $t \in [400,800) \cup [1200,1600)$ is $\frac{\Delta}{\sqrt{32}}[1,\cdots,1] \in \mathbb{R}^{32}$, where $\Delta = 0.5, 1, 1$ respectively in Figures 2c, 2d and 2j respectively.

Gaussian distribution: In Figures 2e and 2f the sample at every time-point is drawn from a unit variance Gaussian distribution. The mean of the samples in the time-durations $t \in [0,400) \cup [800,1200)$ in all three figures is 0 and the mean at times $t \in [400,800) \cup [1200,1600)$ in the two figures 2e and 2f are $\Delta = 0.5$ and $\Delta = 1$ respectively. In Figures 2g and 2h we consider the observation at time t to be 32 dimensional isotropic gaussian random vector with co-variance on each axis being $1/\sqrt{32}$. The mean vector

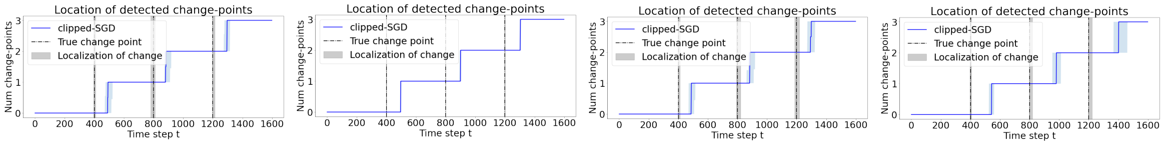


(a) Pareto $\Delta = 0.5$ (b) Pareto $\Delta = 1$ (c) Pareto $d = 32, \Delta = 0.5$ (d) Pareto $d = 32, \Delta = 1$

(e) Normal $\Delta = 0.5$ (f) Normal $\Delta = 1$ (g) Normal $d = 32, \Delta = 0.5$ (h) Normal $d = 32, \Delta = 1$

(i) Pareto $\Delta = 1$ (j) Pareto $d = 32, \Delta = 1$ (k) Bernoulli with $\Delta = 0.5$ (l) Bernoulli with $\Delta = 0.7$

Figure 2: Empirical performance of Algorithm 1 in a variety of scenarios. Exact details of each plot in Section 5.1.



(a) Pareto $\Delta = 1$ (b) Pareto $d = 32, \Delta = 1$ (c) Normal $\Delta = 1$ (d) Bernoulli $\Delta = 0.7$

Figure 3: Plots showing that by Algorithm 2 can detect and localize change-points across a variety of settings.

at times $[0, 400) \cup [800, 1200)$ is $0 \in \mathbb{R}^{32}$ and at times $t \in [400, 1600) \cup [1200, 1600)$ is $\frac{\Delta}{\sqrt{32}}[1, \cdots, 1] \in \mathbb{R}^{32}$.

Bernoulli distribution: In Figures 2k and 2l, the data was $\{0, 1\}$ valued Bernoulli random variable with means at times $[0, 400) \cup [800, 1200)$ was 0.7 and 0.85 respectively in the two figures, and the means at times $[400, 800) \cup [1200, 1600)$ are $0.3, 0.15$ respectively in the two figures.

5.1.2 Baselines

We consider the `Improved-GLR` of [Maillard, 2019] and `R-BOCPD` of [Alami et al., 2020] as baselines since they have been empirically demonstrated to be state-of-art, and are the only other algorithms to possess finite sample, non-asymptotic FPR guarantees. The `Improved-GLR` can be applied to any distribution, while its theoretical guarantees only hold for sub-gaussian distributions. The `R-BOCPD` algorithm is only applicable to binary data, and thus we only use it on the Bernoulli distributed setting.

5.1.3 Results

Figure 2 shows that our algorithm is the only one to attain bounded FPR across heavy-tailed, Gaussian, high dimensional and Bernoulli distribution.

For Pareto distribution, Figures 2h and 2j show that the `Improved-GLR` algorithm has a large number of False Positives. Intuitively this occurs because the `Improved-GLR` algorithm assumes sub-gaussian tails and thus large deviations that are typical for the heavy-tailed Pareto distributions are mistaken for a change. (See also Figure 6). In contrast, from Figures 2a, 2b, 2c, 2d and 2j, we see that our algorithm consistently attains bounded false-positive rates and finite detection delay guarantees across choices of Δ and dimension d.

On gaussian distributed data, both our algorithm 1 and the `Improved-GLR` obtains similar performance in-terms of false-positive rates. However, the the median detection time of our algorithm is larger than the 95th percentile detection time of `Improved-GLR`. In Bernoulli distributed data, all methods attain similar False-positive guarantees; however, the specialized algorithm of `R-BOCPD` is superior in terms of detection delay compared to ours and the `Improved-GLR`.

In Table 1, we summarize Figure 2 by measuring *regret*. For any OCPD algorithm $\mathcal{A}$, we can define a function $R^{(\mathcal{A})} : [T] \to \mathbb{N}$ where $R^{(\mathcal{A})}(t) = \sum_{s \leq t} \mathcal{A}_s$ is the total number

Distribution	**d**	**Δ**	**Algorithm** 1	**Improved GLR** [Maillard, 2019]	**R-BOCPD** [Alami et al., 2020]
Normal	1	1	274 ± 38	$\mathbf{64 \pm 45}$	N/A
	32	1	$\mathbf{300 \pm 6}$	2400 ± 0	
	1	0.5	694 ± 191	$\mathbf{356 \pm 150}$	
	32	0.5	$\mathbf{1427 \pm 14}$	2400 ± 1	
Pareto	1	1	$\mathbf{296 \pm 35}$	19913 ± 8143	N/A
	32	1	$\mathbf{302 \pm 7}$	1616 ± 921	
	1	0.5	$\mathbf{868 \pm 365}$	1891 ± 663	
	32	0.5	$\mathbf{1431 \pm 14}$	1667 ± 653	
Bernoulli	-	0.7	515 ± 49	181 ± 23	$\mathbf{23 \pm 479}$
	-	0.5	1509 ± 53	1466 ± 762	$\mathbf{63 \pm 380}$

Table 1: Quantitative summary of Figure 2 by comparing regret, where lower is better. Our method achieves lower regret across variety of settings of distribution, dimension and change magnitude.

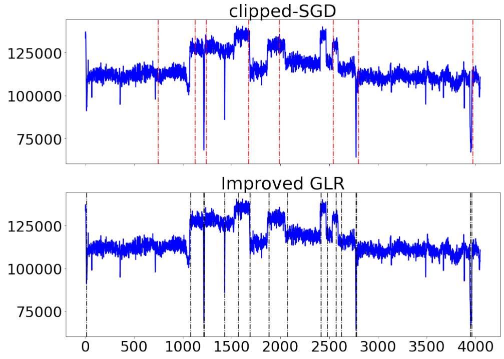


Figure 4: Performance of change-point detection of Algorithm 1 and the `Improved-GLR` on real data.

of change-points detected upto time t. Similarly, for any $t \in [T]$, the ground-truth function $R^*(t) = \max\{c : \tau_c \leq t\}$ is the number of true changes till time t. The regret of algorithm $\mathcal{A}$ is defined as $\sum_{t=1}^{T} |R^{\mathcal{A}}(t) - R^*(t)|$. This measure is non-negative and is 0 if and only if the output of the algorithm matches the ground truth. In Table 1, we give the median value of regret along with 95% confidence interval. We observe in Table 1 that our method achieves lower regret across a variety of situations - whether the data is heavy-tailed, light tailed high dimensional or discrete.

5.1.4 Change-point localization

In Figure 3, we demonstrate sharpness of change-point localization (detailed in Algorithm 2). The setting in Figure 3 is identical to that of Figure 2 with the boundary of the shaded region representing the 5th quantile for the starting point and the 95th quantile for the ending point of the change location interval output in Line 12 of Algorithm 2. The localization region is biased towards the right, which is expected since our algorithm is designed to minimize false positives even in the worst-case.

5.2 REAL-DATA

In Figure 4 we show the performance of Algorithm 1 and the `Improved-GLR` on the well-log dataset [Ó Ruanaidh and Fitzgerald, 1996]. This dataset consists of 4050 measurements in the range $[6 \times 10^4, 10^5]$ of nuclear-magnetic-response taken during drilling of a well. The data are used to interpret the geophysical structure of the rock surrounding the well. The variations in mean reflect the stratification of the earth's crust. We process the data by dividing it by $10^{4.5}$ and run Algorithm 1 with $G = 10$, $\sigma = 1$ and `Improved GLR` with $\sigma = 1$. The detected change-points are shown in Figure 4. Figure 4 shows that Algorithm 1 is comparable to `Improved-GLR` in terms of false-positives.

6 CONCLUSIONS

We introduced a new method based on clipped-SGD, to detect change-points with guaranteed finite-sample FPR, without parametric or tail assumptions. The key technical contribution is to give an anytime online mean estimation algorithm, that provides a confidence bound for the mean at all confidence levels simultaneously. We also give a finite-sample, high probability bound on the detection delay as a function of the gap between the means and number of pre-change observations. We further corroborate empirically that ours is the only algorithm to detect change-points with bounded FPR, across multi-dimensional heavy tailed, gaussian or binary-valued data streams.

Our work opens several interesting directions for future work. Obtaining sharp confidence intervals for estimating the mean of a random vector without the existence of variance was shown in [Cherapanamjeri et al., 2022, Wang and Ramdas, 2022]. Extending the tools from therein to further relax the second moment assumption we considered is a natural direction of future work. Another open question is to see if the martingale methods can be extended to the high-dimensional to get dimension free confidence bounds. Further, we observe in simulations that our method attains 'sharp' localization empirically. Understanding the threeway trade-off between sharpness of localization, FPR and detection delay is an important area of future work.

Acknowledgements AS thanks Aaditya Ramdas for several useful comments that improved the presentation.

REFERENCES

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL `https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf`.

Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Réda Alami, Odalric Maillard, and Raphael Féraud. Restarted bayesian online change-point detector achieves optimal detection delay. In *International conference on machine learning*, pages 211–221. PMLR, 2020.

Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

Ian Barnett and Jukka-Pekka Onnela. Change point detection in correlation networks. *Scientific reports*, 6(1): 18893, 2016.

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

Sujay Bhatt, Guanhua Fang, and Ping Li. Offline change detection under contamination. In *Uncertainty in Artificial Intelligence*, pages 191–201. PMLR, 2022.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Yeshwanth Cherapanamjeri, Samuel B Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 601–609, 2020.

Yeshwanth Cherapanamjeri, Nilesh Tripuraneni, Peter Bartlett, and Michael Jordan. Optimal mean estimation without a variance. In *Conference on Learning Theory*, pages 356–357. PMLR, 2022.

Haeran Cho. Change-point detection in panel data via double cusum statistic. 2016.

Sayak Ray Chowdhury, Patrick Saux, Odalric-Ambrym Maillard, and Aditya Gopalan. Bregman deviations of generic exponential families. *arXiv preprint arXiv:2201.07306*, 2022.

Guido Dartmann, Houbing Song, and Anke Schmeink. *Big data analytics for cyber-physical systems: machine learning for the internet of things*. Elsevier, 2019.

Jules Depersin and Guillaume Lecué. Robust sub-gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics*, 50(1):511–536, 2022.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I Oliveira. Sub-gaussian mean estimators. 2016.

Ilias Diakonikolas, Daniel M Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. *Advances in Neural Information Processing Systems*, 33:1830–1840, 2020.

Ilias Diakonikolas, Daniel M Kane, Ankit Pensia, and Thanasis Pittas. Streaming algorithms for high-dimensional robust statistics. In *International Conference on Machine Learning*, pages 5061–5117. PMLR, 2022.

Klaus Frick, Axel Munk, and Hannes Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 495–580, 2014.

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. 2014.

Zhen Gao, Guoliang Lu, Peng Yan, Chen Lyu, Xueyong Li, Wei Shang, Zhaohong Xie, and Wanming Zhang. Automatic change detection for real-time monitoring of eeg signals. *Frontiers in physiology*, 9:325, 2018.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

David Hallac, Jure Leskovec, and Stephen Boyd. Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 387–396, 2015.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49, 2021.

George Konidaris, Scott Kuindersma, Roderic Grupen, and Andrew Barto. Constructing skill trees for reinforcement learning agents from demonstration trajectories. *Advances in neural information processing systems*, 23, 2010.

Raphail Krichevsky and Victor Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

Barış Kurt, Çağatay Yıldız, Taha Yusuf Ceritli, Bülent Sankur, and Ali Taylan Cemgil. A bayesian change point model for detecting sip-based ddos attacks. *Digital Signal Processing*, 77:48–62, 2018.

Tze Leung Lai and Haipeng Xing. Sequential change-point detection when the pre-and post-change parameters are unknown. *Sequential analysis*, 29(2):162–175, 2010.

Marc Lavielle and Gilles Teyssiere. Adaptive detection of multiple change-points in asset price volatility. *Long memory in economics*, pages 129–156, 2007.

Mengchu Li and Yi Yu. Adversarially robust change point detection. *Advances in Neural Information Processing Systems*, 34:22955–22967, 2021.

Patrick Loiseau, Paulo Gonçalves, Guillaume Dewaele, Pierre Borgnat, Patrice Abry, and Pascale Vicat-Blanc Primet. Investigating self-similarity and heavy-tailed distributions on a large-scale experimental facility. *IEEE/ACM Transactions on Networking*, 18(4):1261–1274, 2010.

Gary Lorden. Procedures for reacting to a change in distribution. *The annals of mathematical statistics*, pages 1897–1908, 1971.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. 2021.

Oscar Hernan Madrid Padilla, Yi Yu, Daren Wang, and Alessandro Rinaldo. Optimal nonparametric change point analysis. 2021.

Jessica Maghakian, Joshua Comden, and Zhenhua Liu. Online optimization in the non-stationary cloud: Change point detection for resource provisioning. In *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2019.

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

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

Vito MR Muggeo and Giada Adelfio. Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, 27(2):161–166, 2011.

Jayakrishnan Nair, Adam Wierman, and Bert Zwart. *The fundamentals of heavy tails: Properties, emergence, and estimation*, volume 53. Cambridge University Press, 2022.

Yue S Niu and Heping Zhang. The screening and ranking algorithm to detect dna copy number variations. *The annals of applied statistics*, 6(3):1306, 2012.

Farhana Nizam, Shudarshon Chaki, Shamim Al Mamun, M Shamim Kaiser, et al. Attack detection and prevention in the cyber physical system. In *2016 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6. IEEE, 2016.

J. J. K. Ó Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer, 1996.

Opeyemi Osanaiye, Kim-Kwang Raymond Choo, and Mqhele Dlodlo. Change-point cloud ddos detection using packet inter-arrival time. In *2016 8th Computer Science and Electronic Engineering (CEEC)*, pages 204–209. IEEE, 2016.

Oscar Hernan Madrid Padilla, Yi Yu, Daren Wang, and Alessandro Rinaldo. Optimal nonparametric multivariate change point detection and localization. *IEEE Transactions on Information Theory*, 68(3):1922–1944, 2021.

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

Moshe Pollak. Optimal detection of a change in distribution. *The Annals of Statistics*, pages 206–227, 1985.

Aleksey S Polunchenko, Alexander G Tartakovsky, and Nitis Mukhopadhyay. Nearly optimal change-point detection with an application to cybersecurity. *Sequential Analysis*, 31(3):409–435, 2012.

Ya'acov Ritov. Decision theoretic optimality of the cusum procedure. *The Annals of Statistics*, pages 1464–1469, 1990.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

Shubhanshu Shekhar and Aaditya Ramdas. Sequential change detection via backward confidence sequences. *arXiv preprint arXiv:2302.02544*, 2023.

Albert N Shiryaev. *Optimal stopping rules*, volume 8. Springer Science & Business Media, 2007.

AG Tartakovsky. Sequential methods in the theory of information systems, 1991.

Che-Ping Tsai, Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. Heavy-tailed streaming statistical estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 1251–1282. PMLR, 2022.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

H Victor. A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27(1): 537–564, 1999.

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

Abraham Wald. *Sequential tests of statistical hypotheses*. Springer, 1992.

Hongjian Wang and Aaditya Ramdas. Catoni-style confidence sequences for heavy-tailed mean estimation. *arXiv preprint arXiv:2202.01250*, 2022.

Hongjian Wang and Aaditya Ramdas. Huber-robust confidence sequences. In *International Conference on Artificial Intelligence and Statistics*, pages 9662–9679. PMLR, 2023.

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

Ping Yang, Guy Dumont, and John Mark Ansermino. Adaptive change detection in heart rate trend monitoring in anesthetized children. *IEEE transactions on biomedical engineering*, 53(11):2211–2219, 2006.

Jie Zhang, Zhi Wei, Zhenyu Yan, MengChu Zhou, and Abhishek Pani. Online change-point detection in sparse time series with application to online advertising. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49 (6):1141–1151, 2017.

A CHANGE-POINT LOCALIZATION

Algorithm 2 Online `Clipped-SGD` Change Point Detection and Localization

```
Input: $(\eta_t)_{t\geq 1}$, $\lambda > 0$, $\theta_0 \in \Theta$, $\delta \in (0,1)$ FPR guarantee
$r \leftarrow 1$
$\widehat{\theta}_{t,t-1} \leftarrow \theta_0$, for all $t \geq 1$.
Set $\tau_c^{(0)} \leftarrow 0$
Set Num-change-points $\leftarrow 0$
for each time $t = 1, 2, \cdots$, do
    Receive sample $X_t$
    $\widehat{\theta}_{s,t} \leftarrow \prod_{\theta}(\widehat{\theta}_{s,t-1} - \eta_{t-s}\text{clip}(X_t - \widehat{\theta}_{s,t-1}, \lambda))$, for every $r \leq s \leq t$.
    if $\exists s \in (r,t)$ such that $\|\widehat{\theta}_{r:s} - \widehat{\theta}_{s+1:t}\|_2^2 > \mathcal{B}\left(s-r, \frac{\delta}{2(t-r)(t-r+1)}\right) + \mathcal{B}\left(t-s-1, \frac{\delta}{2(t-r)(t-r+1)}\right)$ {$B(\cdot,\cdot)$ is defined in Equation (5)} then
        Set Restart$_t \leftarrow 1$ {Change point detected}
        Set Num-change-points $\leftarrow$Num-change-points $+1$ {Increment number of change-points detected}
        Output time interval $[\inf\{s \in (r,t)$ s.t. $\mathfrak{B}(r,s,t,\delta) = 1\}, \sup\{s \in (r,t)$ s.t. $\mathfrak{B}(r,s,t,\delta) = 1\}]$ as the location of the change-point {$\mathfrak{B}()$ defined in Equation (8)}
        $r \leftarrow t+1$
    else
        Set Restart$_t \leftarrow 0$
    end if
end for
```

B PROOF FOR ROBUST ESTIMATION IN THEOREM 3.1

We follow the same proof architecture as that of Proof of [Tsai et al., 2022]. Throughout the proof, we let $m = 1$ to be the strong convexity parameter of the quadratic loss function $x \to \frac{1}{2}\|x - x_0\|^2$, for some $x_0 \in \mathbb{R}^d$.

Fix a time $t \in \mathbb{N}$. We define a sequence of random variable $(\psi_t)_{t\geq 1}$ as follows.

$$\psi_t := \text{clip}((X_t - \widehat{\theta}_{t-1}), \lambda) - (\theta^* - \widehat{\theta}_{t-1}),$$

Consider any time t. We have

$$\begin{aligned}
\|\theta_t - \theta^*\|_2^2 &= \|\prod_{\Theta}(\widehat{\theta}_{t-1} - \eta_t\text{clip}(X_t - \widehat{\theta}_{t-1}, \lambda)) - \theta^*\|_2^2, && (10)\\
&\overset{(a)}{\leq} \|\widehat{\theta}_{t-1} - \eta_t\text{clip}(X_t - \widehat{\theta}_{t-1}, \lambda) - \theta^*\|_2^2, && (11)\\
&= \|\widehat{\theta}_{t-1} - \eta_t(\psi_t + (\theta^* - \widehat{\theta}_{t-1})) - \theta^*\|_2^2, \\
&= \|\widehat{\theta}_{t-1} - \theta^*\|_2^2 + \eta_t^2\|\psi_t + (\theta^* - \widehat{\theta}_{t-1})\|_2^2 - 2\eta_t\langle\widehat{\theta}_{t-1} - \theta^*, \psi_t + (\theta^* - \widehat{\theta}_{t-1})\rangle, \\
&\overset{(b)}{\leq} \|\widehat{\theta}_{t-1} - \theta^*\|_2^2 + 2\eta_t^2\|\psi_t\|_2^2 + 2\eta_t^2\|(\theta^* - \widehat{\theta}_{t-1})\|_2^2 - 2\eta_t\langle\widehat{\theta}_{t-1} - \theta^*, \psi_t + (\theta^* - \widehat{\theta}_{t-1})\rangle, && (12)
\end{aligned}$$

Step (a) follows since Θ is a convex set, $\|\mathcal{P}_{\Theta}(\widehat{\theta}_t) - \theta^*\| \leq \|\widehat{\theta}_t - \theta^*\|$, since $\theta^* \in \Theta$. In step (b), we use the fact that $\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$, for all $a, b \in \mathbb{R}^d$. Substituting Equation (44) into (12), we get that

$$\begin{aligned}
\|\theta^* - \theta_t\|_2^2 \leq \|\widehat{\theta}_{t-1} - \theta^*\|_2^2 &+ 2\eta_t^2\|\psi_t\|_2^2 - 2\eta_t\langle\widehat{\theta}_{t-1} - \theta_t^*, \psi_t\rangle \\
&+ 2\eta_t^2\left((M+m)\langle(\theta^* - \widehat{\theta}_{t-1}), \widehat{\theta}_{t-1} - \theta_t^*\rangle - mM\|\widehat{\theta}_{t-1} - \theta^*\|_2^2\right) - 2\eta_t\langle(\theta^* - \widehat{\theta}_{t-1}), \widehat{\theta}_{t-1} - \theta_t^*\rangle.
\end{aligned}$$

Re-arranging the equation above yields

$$\|\theta^* - \theta_t\|_2^2 \leq (1 - 2\eta_t^2 mM)\|\widehat{\theta}_{t-1} - \theta^*\|_2^2 + 2\eta_t^2\|\psi_t\|_2^2 - 2\eta_t\langle\widehat{\theta}_{t-1} - \theta^*, \psi_t\rangle$$

$$- 2\eta_t(1 - \eta_t ((M + m)) \langle(\theta^* - \widehat{\theta}_{t-1}), \widehat{\theta}_{t-1} - \theta^*\rangle.$$

Further substituting Equation (43) into the display above yields that

$$\begin{aligned}
\|\theta^* - \widehat{\theta}_t\|_2^2 &\leq (1 - 2\eta_t m + 2\eta_t^2 m^2)\|\widehat{\theta}_{t-1} - \theta^*\|_2^2 + 2\eta_t^2\|\psi_t\|_2^2 - 2\eta_t\langle\widehat{\theta}_{t-1} - \theta^*, \psi_t\rangle, \\
&\leq (1 - \eta_t m)\|\widehat{\theta}_{t-1} - \theta^*\|_2^2 + 2\eta_t^2\|\psi_t\|_2^2 - 2\eta_t\langle\widehat{\theta}_{t-1} - \theta^*, \psi_t\rangle,
\end{aligned}$$

where the inequality comes from the fact that if $\eta_t m < 1 \implies 2\eta_t m - 2\eta_t^2 m^2 > \eta m.$

$$\|\theta^* - \widehat{\theta}_t\|_2^2 \leq (1 - \eta_t m)\|\widehat{\theta}_{t-1} - \theta^*\|_2^2 + 2\eta_t^2\|\psi_t\|_2^2 - 2\eta_t\langle\widehat{\theta}_{t-1} - \theta^*, \psi_t\rangle. \tag{13}$$

Unrolling the recursion yields,

$$\begin{aligned}
\|\theta^* - \widehat{\theta}_t\|_2^2 \leq \prod_{u=1}^{t}(1 - \eta_u m)\|\theta_1 - \theta^*\|_2^2 &+ 2\eta_t^2\sum_{s=1}^{t-1}\prod_{u=1}^{s}(1 - \eta_{t-u+1} m)\|\psi_{t-s+1}\|_2^2 \\
&-2\eta_t\sum_{s=1}^{t-1}\prod_{u=1}^{s}(1 - \eta_{t-u+1} m)\langle\theta_{t-s} - \theta^*, \psi_{t-s+1}\rangle.
\end{aligned}$$

Using the fact that $\prod_{u=1}^{s}(1 - \eta_{t-u+1}m) = \frac{(t-s+\gamma-3)(t-s+\gamma-2)}{(t+\gamma)(t+\gamma-1)}$, we get that

$$\begin{align}
\|\theta^* - \widehat{\theta}_t\|_2^2 \leq \frac{(\gamma-2)(\gamma-1)\|\theta_1 - \theta^*\|_2^2}{(t+\gamma)(t+\gamma-1)} \tag{14} \\
-2\eta_t\sum_{s=1}^{t-1}\frac{(t-s+\gamma-3)(t-s+\gamma-2)\langle\theta_{t-s} - \theta^*, \psi_{t-s+1}\rangle}{(t+\gamma)(t+\gamma-1)}. \tag{15}
\end{align}$$

Denote by $\psi_t := \psi_t^{(b)} + \psi_t^{(v)}$, where $\psi_t^{(b)} := \mathbb{E}_{Z_t}[\psi_t|\mathcal{F}_{t-1}]$ and $\psi_t^{(v)} := \psi_t - \psi_t^{(b)}$. Using this in the display above and using that fact that $\|a + b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$, we get

$$\begin{aligned}
\|\theta_t^* - \theta\|_2^2 \leq \frac{(\gamma-1)(\gamma-2)\|\theta_1 - \theta^*\|_2^2}{(t+\gamma)(t+\gamma-1)} &+ 4\eta_t^2\sum_{s=1}^{t-1}\frac{(t-s+\gamma-3)(t-s+\gamma-2)\|\psi_{t-s+1}\|_2^2}{(t+\gamma)(t+\gamma-1)} + \\
&- 2\eta_t\sum_{s=1}^{t-1}\frac{(t-s+\gamma-3)(t-s+\gamma-2)\langle\theta_{t-s} - \theta^*, \psi_{t-s+1}^{(b)}\rangle}{(t+\gamma)(t+\gamma-1)} \\
&- 2\eta_t\sum_{s=1}^{t-1}\frac{(t-s+\gamma-3)(t-s+\gamma-2)\langle\theta_{t-s} - \theta^*, \psi_{t-s+1}^{(v)}\rangle}{(t+\gamma)(t+\gamma-1)}.
\end{aligned}$$

Further simplifying by adding and subtracting $\mathbb{E}_{Z_t}[\|\psi_t^{(v)}\|_2^2|\mathcal{F}_{t-1}]$ to be above display, we get

$$\begin{align}
\|\theta^* - \widehat{\theta}_t\|_2^2 \leq \frac{(\gamma-1)(\gamma-2)\|\theta_1 - \theta^*\|_2^2}{(t+\gamma)(t+\gamma-1)} + 4\eta_t^2\sum_{s=1}^{t-1}\frac{(t-s+\gamma-3)(t-s+\gamma-2)\|\psi_{t-s+1}^{(b)}\|_2^2}{(t+\gamma)(t+\gamma-1)} \tag{16} \\
+ 4\eta_t^2\sum_{s=1}^{t-1}\frac{(t-s+\gamma-3)(t-s+\gamma-2)\mathbb{E}_{Z_{t-s+1}}[\|\psi_{t-s+1}^{(v)}\|_2^2|\mathcal{F}_{t-s}]}{(t+\gamma)(t+\gamma-1)} \notag \\
+ 4\eta_t^2\sum_{s=1}^{t-1}\frac{(t-s+\gamma-3)(t-s+\gamma-2)(\|\psi_{t-s+1}^{(v)}\|_2^2 - \mathbb{E}_{Z_{t-s+1}}[\|\psi_{t-s+1}^{(v)}\|_2^2|\mathcal{F}_{t-s}])}{(t+\gamma)(t+\gamma-1)} \tag{17} \\
- 2\eta_t\sum_{s=1}^{t-1}\frac{(t-s+\gamma-3)(t-s+\gamma-2)\langle\theta_{t-s} - \theta^*, \psi_{t-s+1}^{(b)}\rangle}{(t+\gamma)(t+\gamma-1)} \notag \\
- 2\eta\sum_{s=1}^{t-1}\frac{(t-s+\gamma-3)(t-s+\gamma-2)\langle\theta_{t-s} - \theta^*, \psi_{t-s+1}^{(v)}\rangle}{(t+\gamma)(t+\gamma-1)}. \tag{18}
\end{align}$$

Lemma B.1 (Lemma F.5 [Gorbunov et al., 2020]). *If $\lambda \geq 2G$, the following inequalities hold almost-surely for all times t.*

$$\|\psi_t^{(v)}\| \leq 2\lambda \mathbf{1}_{\sigma > 0} \tag{19}$$

$$\|\psi_t^{(b)}\|_2 \leq \frac{4\sigma^2}{\lambda} \tag{20}$$

$$\mathbb{E}_{Z_t}[\|\psi_t^{(v)}\|_2^2 | \mathcal{F}_{t-1}] \leq 10\sigma^2 \tag{21}$$

Simplifying Equation (18) using bounds in Lemma B.1, along with the fact that for all $1 \leq s \leq t$ and $\gamma \geq 1$, $\frac{(t-s+\gamma-3)(t-s+\gamma-2)}{(t+\gamma)(t+\gamma-1)} \leq \frac{t-s+\gamma}{t+\gamma}$ we get

$$\begin{aligned}
\|\theta^* - \widehat{\theta}_t\|_2^2 \leq{}& \frac{(\gamma-1)(\gamma-2)\|\theta_1 - \theta^*\|_2^2}{(t+\gamma)(t+\gamma-1)} + \frac{16\eta_t^2\sigma^2}{\lambda}\sum_{s=1}^{t-1}\frac{t-s+\gamma}{t+\gamma} + 4\eta_t^2\sigma^2\sum_{s=1}^{t-1}\frac{t-s+\gamma}{t+\gamma} \\
&+ 4\eta_t^2\sum_{s=1}^{t-1}\frac{(t-s+\gamma)(\|\psi_{t-s+1}^{(v)}\|_2^2 - \mathbb{E}_{Z_{t-s+1}}[\|\psi_{t-s+1}^{(v)}\|_2^2|\mathcal{F}_{t-s+1}])}{t+\gamma} \\
&+ 2\eta_t\sum_{s=1}^{t-1}\frac{(t-s+\gamma)\|\theta_{t-s} - \theta^*\|\|\psi_{t-s+1}^{(b)}\|}{t+\gamma} + -2\eta_t\sum_{s=1}^{t-1}\frac{(t-s+\gamma)\langle\theta_{t-s} - \theta^*, \psi_{t-s+1}^{(v)}\rangle}{t+\gamma}.
\end{aligned} \tag{22}$$

Further applying the bound that $\|\psi_t^{(b)}\| \leq \frac{4\sigma^2}{\lambda}$

$$\begin{aligned}
\|\theta^* - \widehat{\theta}_t\|_2^2 \leq{}& \frac{(\gamma-1)(\gamma-2)\|\theta_1 - \theta^*\|_2^2}{(t+\gamma)(t+\gamma-1)} + \underbrace{\left(\frac{16\eta_t^2\sigma^2}{\lambda} + 4\eta_t^2\sigma^2\right)\sum_{s=1}^{t-1}\frac{t-s+1}{t+\gamma}}_{\text{Term 1}} \\
&\underbrace{+ 4\eta_t^2\sum_{s=1}^{t-1}\frac{(t-s+\gamma)(\|\psi_{t-s+1}^{(v)}\|_2^2 - \mathbb{E}_{Z_{t-s+1}}[\|\psi_{t-s+1}^{(v)}\|_2^2|\mathcal{F}_{t-s+1}])}{t+\gamma}}_{\text{Term 2}} \\
&\underbrace{+ \frac{8\sigma^2\eta_t}{\lambda}\sum_{s=1}^{t-1}\frac{(t-s+\gamma)\|\theta_{t-s} - \theta^*\|}{t+\gamma}}_{\text{Term 3}}\underbrace{-2\eta_t\sum_{s=1}^{t-1}\frac{(t-s+\gamma)\langle\theta_{t-s} - \theta^*, \psi_{t-s+1}^{(v)}\rangle}{t+\gamma}}_{\text{Term 4}}.
\end{aligned} \tag{23}$$

B.1 PROBABILISTIC ANALYSIS

Definitions

For every $t \geq 1$, denote by the constant

$$C_t = \max\left(\frac{1024\sigma^4}{G^2m^2\lambda^2}, \frac{8\lambda\sqrt{\ln\left(\frac{2t^3}{\delta}\right)}}{\gamma^2 G}\right). \tag{24}$$

Denote by the deterministic constant $\xi_u^{(t)}$ for $u = 1, \cdots, t$ as

$$\left(\xi_u^{(t)}\right)^2 := C_t\left[\left(\frac{16\sigma^2}{\lambda} + 4\sigma^2\right)\frac{1}{2m^2(u+1)} + \frac{96\lambda^2\ln\left(\frac{2t^3}{\delta}\right)\sigma(\sigma+1)}{m(u+\gamma)\sqrt{u+1}}\right]. \tag{25}$$

From the definition, the following in-equalities hold.

Proposition B.2. *For all times* $u \in \{1, \cdots, t\}$*,*

$$\sum_{s=1}^{u-1} (u - s + \gamma)\xi_s^{(t)} \le 2(u + \gamma)\sqrt{u + 1}\xi_u^{(t)}, \tag{26}$$

$$\sum_{s=1}^{u-1} (\xi_s^{(t)})^2 \le 2(u + 1)\ln(u + 1)(\xi_u^{(t)})^2 \tag{27}$$

Proof. This follows from the following fact.

Proposition B.3. *For all* $u \in \mathbb{N}$ *and* $\gamma \ge 0$*, we have*

$$\sum_{s=1}^{u-1} \frac{u - s + \gamma}{\sqrt{u + 1}} \le 2(u + \gamma)\sqrt{u + 1}.$$

□

For each time $u \in \{1, \cdots, t\}$, denote by the random variable $\nu_u^{(t)}$ by

$$\nu_u^{(t)} := \begin{cases} \theta_u - \theta^* & \text{if } \|\theta_u - \theta^*\|^2 \le (\xi_u^{(t)})^2 + \frac{C_t\gamma^2G^2}{(u+1)} \\ 0 & \text{if otherwise} \end{cases}$$

For every $u \in \{1, \cdots, t\}$, denote by the event $\mathcal{E}_{u;1}^{(t)}$ to be the one in which the following inequality holds for all $u \in \{1, \cdots, t\}$.

$$\mathcal{E}_{u;1}^{(t)} := \Bigg\{ 4\eta_t^2 \sum_{s=1}^{u-1} \frac{(u - s + \gamma)(\|\psi_{u-s+1}^{(v)}\|_2^2 - \mathbb{E}_{Z_{u-s+1}}[\|\psi_{u-s+1}^{(v)}\|_2^2|\mathcal{F}_{u-s+1}])}{t + \gamma} \le \frac{96\lambda^2 \ln\left(\frac{2t^2(t+1)}{\delta}\right)\sigma(\sigma + 1)}{m(u + \gamma)\sqrt{u + 1}} \Bigg\}. \tag{28}$$

and $\mathcal{E}_{u;2}^{(t)}$ as

$$\mathcal{E}_{u;2}^{(t)} := \Bigg\{ -2\eta_u \sum_{s=1}^{u-1} \frac{(u - s + \gamma)\langle \upsilon_{u-s}, \psi_{u-s+1}^{(v)}\rangle}{t + \gamma} \le \frac{\xi_u^{(t)} \ln\left(\frac{2t^2(t+1)}{\delta}\right)}{10\sqrt{u + 1}} + \frac{C_u\gamma^2G^2}{4(u + 1)}. \Bigg\} \tag{29}$$

Denote by the event $\mathcal{E}^{(t)}$ as

$$\mathcal{E}^{(t)} := \bigcap_{u=1}^{t} \left(\mathcal{E}_{u;1}^{(t)} \cap \mathcal{E}_{u;2}^{(t)} \right). \tag{30}$$

Lemma B.4. *For all* $t \ge 1$*,*

$$\mathbb{P}[\mathcal{E}^{(t)}] \ge 1 - \frac{\delta}{t(t + 1)}.$$

We now prove by induction hypothesis that

Lemma B.5. *For every* t*, under the event* $\mathcal{E}^{(t)}$*, the following holds.*

$$\|\widehat{\theta}_u - \theta^*\|_2^2 \le \frac{C_t\gamma^2G^2}{(u + 1)^2} + (\xi_u^{(t)})^2, \tag{31}$$

for all $u \in \{1, \cdots, t\}$.

Proof. Proof of Lemma B.1. We will prove this lemma by induction on u by analyzing Equation (23). The base-case of $u = 1$ holds trivially with probability 1 since $C_t > 1, \forall t \geq 1$ and $\gamma > 2$.

Now, assume that on the event $\mathcal{E}^{(t)}$, the induction hypothesis in Equation (31) holds for all times $1, \cdots, u-1$. We prove this by expanding Equation (23) and bounding each of the terms.

Term 1

It is easy to verify that

$$\begin{aligned}\left(\frac{16\eta_u^2\sigma^2}{\lambda} + 4\eta_u^2\sigma^2\right)\sum_{s=1}^{u-1}\frac{u-s+\gamma}{u+\gamma} &\leq \left(\frac{16\sigma^2}{\lambda} + 4\sigma^2\right)\frac{u}{2m^2(u+\gamma)^2}, \\ &\leq \frac{\left(\frac{16\sigma^2}{\lambda} + 4\sigma^2\right)}{2m^2(u+1)}.\end{aligned}$$

The last inequality follows since $\gamma^2 > 1$.

Term 2

First notice that

$$4\eta_u^2\sum_{s=1}^{u-1}\frac{(u-s+\gamma)(\|\psi_{u-s+1}^{(v)}\|_2^2 - \mathbb{E}_{Z_{u-s+1}}[\|\psi_{u-s+1}^{(v)}\|_2^2|\mathcal{F}_{u-s+1}])}{t+\gamma} \leq$$
$$\frac{4\eta_u}{u+\gamma}\sum_{s=1}^{u-1}(\|\psi_{u-s+1}^{(v)}\|_2^2 - \mathbb{E}_{Z_{u-s+1}}[\|\psi_{u-s+1}^{(v)}\|_2^2|\mathcal{F}_{u-s+1}])$$

From the definition of event $\mathcal{E}^{(t)}$ in Equation (30), we get that

$$\text{Term 2} \leq \frac{96\lambda^2\ln\left(\frac{2t^2(t+1)}{\delta}\right)\sigma(\sigma+1)}{m(u+\gamma)\sqrt{u+1}}.$$

Term 3

$$\begin{aligned}\frac{8\sigma^2\eta_u}{\lambda}\sum_{s=1}^{u-1}\frac{(u-s+\gamma)\|\theta_{u-s}-\theta^*\|}{u+\gamma} &\leq \frac{8\sigma^2}{m\lambda(u+\gamma)^2}\sum_{s=1}^{u-1}\left((u-s+\gamma)\xi_{u-s}^{(t)} + \sqrt{C_t}\gamma G\frac{(u-s+\gamma)}{(u-s+1)}\right), \\ &\overset{(27)}{\leq} \frac{16\sigma^2\sqrt{(u+1)}\xi_u^{(t)}}{m(u+\gamma)} + \frac{8\sqrt{C_t}\sigma^2\gamma^2 Gu}{m\lambda(u+\gamma)^2}, \\ &\leq \frac{16\sigma^2\sqrt{(u+1)}\xi_u^{(t)}}{m(u+\gamma)} + \frac{8\sqrt{C_t}\sigma^2\gamma^2 G}{m\lambda(u+\gamma)}, \\ &\overset{(a)}{\leq} \frac{\xi_u^{(t)}}{10\sqrt{u+1}} + \frac{C_t\gamma^2G^2}{4(u+1)}.\end{aligned}$$

The last inequality follows since $\gamma \geq \frac{320\sigma^2}{m} + 1 \implies \frac{8\sigma^2(u+1)^{1/2}\log(u+1)}{m(u+\gamma)} \leq \frac{1}{10\sqrt{u+1}}$, for all $u \leq t$ and the fact that $C_t \geq \frac{1024\sigma^4}{G^2m^2\lambda^2}$.

Term 4

The definition of event $\mathcal{E}^{(t)}$ in Equation (30) gives that Term $4 \leq \frac{\xi_u^{(t)} \ln\left(\frac{2t^2(t+1)}{\delta}\right)}{10\sqrt{u+1}} + \frac{C_t\gamma^2G^2}{4(u+1)}$

Now, adding in the bounds together into Equation (23),

$$\|\widehat{\theta}_u - \theta^*\|_2^2 \leq \frac{\gamma^2G^2}{u+1} + \frac{\left(\frac{16\sigma^2}{\lambda} + 4\sigma^2\right)}{2m^2(u+1)} + \frac{\xi_u^{(t)}}{10\sqrt{u+1}} + \frac{1600\lambda^2 \ln\left(\frac{2t^2(t+1)}{\delta}\right)\sigma(\sigma+1)}{m(u+\gamma)\sqrt{u+1}} + \frac{\xi_u^{(t)} \ln\left(\frac{2t^2(t+1)}{\delta}\right)}{10\sqrt{u+1}} + \frac{C_t\gamma^2G^2}{2(u+1)}.$$

Now using the fact that $\frac{\xi_u^{(t)} \ln\left(\frac{2t^3}{\delta}\right)}{\sqrt{u+1}} \leq (\xi_u^{(t)})^2$, we get that

$$\|\widehat{\theta}_u - \theta^*\|_2^2 \leq \left(1 + \frac{C_t}{2}\right)\frac{\gamma^2G^2}{u+1} + \frac{\left(\frac{16\sigma^2}{\lambda} + 4\sigma^2\right)}{2m^2(u+1)} + \frac{(\xi_u^{(t)})^2}{5} + \frac{96\lambda^2 \ln\left(\frac{2t^2(t+1)}{\delta}\right)\sigma(\sigma+1)}{m(u+\gamma)\sqrt{u+1}}.$$

Substituting the definition of $\xi_u^{(t)}$ from Equation (25), we get that

$$\begin{aligned}\|\widehat{\theta}_u - \theta^*\|_2^2 &\leq \left(1 + \frac{C_t}{2}\right)\left[\frac{\gamma^2G^2}{u+1} + \frac{\left(\frac{16\sigma^2}{\lambda} + 4\sigma^2\right)}{2m^2(u+1)} + \frac{96\lambda^2 \ln\left(\frac{2t^2(t+1)}{\delta}\right)\sigma(\sigma+1)}{m(u+\gamma)\sqrt{u+1}}\right], \\ &\leq (\xi_u^{(t)})^2 + \frac{C_t\gamma^2G^2}{u+1}.\end{aligned}$$

The last inequality follows since $C_t = \max\left(\frac{1024\sigma^4}{G^2m^2\lambda^2}, \frac{8\lambda\sqrt{\ln\left(\frac{2t^3}{\delta}\right)}}{\gamma^2G}\right) \implies C_t \geq 2.$

□

□

B.2 PROOF OF LEMMA B.4

We first reproduce an useful result.

Lemma B.6 (Freedman's inequality[Victor, 1999]). *Suppose $Y_1, \cdots, Y_T$ is a bounded martingale with respect to a filtration $(\mathcal{F}_t)_{t=0}^T$ with $\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = 0$ and $\mathbb{P}[|Y_t| \leq B] = 1$ for all $t \in \{1, \cdots, T\}$. Denote by $V_s := \sum_{n=1}^s \mathit{Var}(Y_n|\mathcal{F}_{n-1})$ be the sum of conditional variances. Then, for every $a, v > 0$,*

$$\mathbb{P}\left(\exists n \in [1, T] \textit{ such that } \sum_{t=1}^n Y_t \geq a \textit{ and } V_n \leq v\right) \leq \exp\left(\frac{-a^2}{2(v + Ba)}\right). \tag{32}$$

Re-arranging the above inequality, we see that if

$$a \geq B\ln\left(\frac{2T}{\delta}\right) + \sqrt{\left(B\ln\left(\frac{2T}{\delta}\right)\right)^2 + 2v\ln\left(\frac{2T}{\delta}\right)}, \tag{33}$$

then the RHS of Equation (32) is bounded above by $\frac{\delta}{2}$.

Proof of Lemma B.4. **Proof of Equation (28)**

Fix a $u \in \{1, \cdots, t\}$. For $s \in \{1, \cdots, u-1\}$, denote by the random variable $Y_s^{(u)} := \frac{(u-s+\gamma)}{u+\gamma}(\|\psi_{u-s+1}^{(v)}\|_2^2 - \mathbb{E}_{Z_{u-s+1}}[\|\psi_{u-s+1}^{(v)}\|_2^2|\mathcal{F}_{u-s}])$. Thus,

$$4\eta_u^2 \sum_{s=1}^{u-1} \frac{(u-s+\gamma)(\|\psi_{u-s+1}^{(v)}\|_2^2 - \mathbb{E}_{Z_{u-s+1}}[\|\psi_{u-s+1}^{(v)}\|_2^2|\mathcal{F}_{u-s+1}])}{u+\gamma} \leq 4\eta_u^2 \sum_{s=1}^{u-1} Y_s^{(u)}.$$

Observe that the sequence $(Y_s^{(u)})_{s=1}^{u-1}$ is a martingale difference sequence with respect to the filtration $(\mathcal{G}_s)_{s=1}^{t-1}$, where $\mathcal{G}_s := \mathcal{F}_{u-s}$. Furthermore, observe that with probability 1, $|Y_s^{(u)}| \leq 4\lambda^2 \mathbf{1}_{\sigma>0} + 4\lambda^2 \mathbf{1}_{\sigma>0} \leq 8\lambda^2 \mathbf{1}_{\sigma>0}$. We can bound the conditional variance as

$$\begin{aligned}
\sum_{s=1}^{u-1} \mathrm{Var}(Y_s^{(u)}|\mathcal{G}_s) &\leq \sum_{s=1}^{u-1} \left(\frac{(u-s+\gamma)}{u+\gamma}\right)^2 \mathbb{E}_{Z_{u-s}}[(\|\psi_{u-s+1}^{(v)}\|_2^2 - \mathbb{E}_{Z_{u-s+1}}[\|\psi_{u-s+1}^{(v)}\|_2^2|\mathcal{F}_{u-s}])^2|\mathcal{F}_{u-s}], \\
&\overset{19}{\leq} 8\lambda^2 \sum_{s=1}^{u-1} \mathbb{E}_{Z_{u-s}}[|\|\psi_{u-s+1}^{(v)}\|_2^2 - \mathbb{E}_{Z_{u-s+1}}[\|\psi_{u-s+1}^{(v)}\|_2^2|\mathcal{F}_{u-s}]||\mathcal{F}_{u-s}], \\
&\leq 8\lambda^2 \sum_{s=1}^{u-1} 2\mathbb{E}_{Z_{u-s}}[|\|\psi_{u-s+1}^{(v)}\|_2^2|\mathcal{F}_{u-s}], \\
&\overset{21}{\leq} 160\lambda^2\sigma^2(u-1).
\end{aligned}$$

Now, putting $B := 8\lambda^2$ and $v = 160\lambda^2\sigma^2 u$, we get from Equation (33) that with probability at-least $1 - \delta/(2t^2(t+1))$,

$$\begin{aligned}
\sum_{s=1}^{u-1} Y_s^{(u)} &\leq 8\lambda^2 \ln\left(\frac{2t^2(t+1)}{\delta}\right)\mathbf{1}_{\sigma>0} + \sqrt{\left(8\lambda^2 \ln\left(\frac{2t^2(t+1)}{\delta}\right)\mathbf{1}_{\sigma>0}\right)^2 + 160\lambda^2\sigma^2 u \ln\left(\frac{2t^2(t+1)}{\delta}\right)}, \\
&\overset{(a)}{\leq} 32\lambda^2 \ln\left(\frac{2t^2(t+1)}{\delta}\right)\sigma(\sigma+1)\sqrt{u+1}.
\end{aligned}$$

Step (a) follows from the fact that $\lambda \geq 1$. Thus, we have with probability at-least $1 - \frac{\delta}{2t^2(t+1)}$,

$$\begin{aligned}
4\eta_u^2 \sum_{s=1}^{u-1} \frac{(u-s+\gamma)(\|\psi_{u-s+1}^{(v)}\|_2^2 - \mathbb{E}_{Z_{u-s+1}}[\|\psi_{u-s+1}^{(v)}\|_2^2|\mathcal{F}_{u-s+1}])}{u+\gamma} &\leq 96\eta_u^2\lambda^2 \ln\left(\frac{2t^2(t+1)}{\delta}\right)\sigma(\sigma+1)\sqrt{u+1}, \\
&\leq \frac{96\lambda^2 \ln\left(\frac{2t^2(t+1)}{\delta}\right)\sigma(\sigma+1)\sqrt{u+1}}{m^2(u+\gamma)^2}, \\
&\leq \frac{96\lambda^2 \ln\left(\frac{2t^2(t+1)}{\delta}\right)\sigma(\sigma+1)}{m^2(u+\gamma)\sqrt{u+1}}.
\end{aligned}$$

Now taking an union bound over all $u \in \{1, \cdots, t\}$ yields that with probability at-least $1 - \frac{\delta}{2t(t+1)}$, for all time $u \in \{1, \cdots, t\}$,

$$4\eta_u^2 \sum_{s=1}^{u-1} \frac{(t-s+\gamma)(\|\psi_{u-s+1}^{(v)}\|_2^2 - \mathbb{E}_{Z_{u-s+1}}[\|\psi_{u-s+1}^{(v)}\|_2^2|\mathcal{F}_{u-s+1}])}{t+\gamma} \leq \frac{96\lambda^2 \ln\left(\frac{2t^2(t+1)}{\delta}\right)\sigma(\sigma+1)}{m(u+\gamma)\sqrt{u+1}}$$

Proof of Equation (29)

$$-2\eta_u \sum_{s=1}^{u-1} \frac{(u-s+\gamma)\langle \upsilon_{u-s}, \psi_{u-s+1}^{(v)}\rangle}{u+\gamma} \leq \frac{2}{m(u+\gamma)^2} \sum_{s=1}^{u-1} \langle \theta_{u-s} - \theta^*, \psi_{u-s+1}^{(v)}\rangle$$

Fix a $u \in \{1, \cdots, t\}$ and denote by $Y_s^{(u)} := (u-s+\gamma)\langle \theta_{u-s} - \theta^*, \psi^{(v)}_{u-s+1}\rangle$. Since θ_{u-s} is measurable with respect to the sigma-algebra generated by $\mathcal{F}_{u-s}$, the conditional expectation $\mathbb{E}[Y_s^{(u)}|\mathcal{F}_{u-s}] = 0$. Thus, $(Y_s^{(u)})_{s=1}^{u-1}$ is a martingale difference sequence with respect to the filtration $(\mathcal{F}_{u-s})_{s=1}^{u-1}$. Furthermore, we have from Equation (19) that $|Y_s^{(u)}| \leq 2(u-s+\gamma)\left(\xi^{(t)}_{u-s} + \frac{\gamma R_1}{(u+\gamma-1)}\right)\lambda \leq 2\lambda(u+\gamma)\xi_t^{(t)} + 2\lambda\gamma G$. We can now bound the sum of conditional variances as

$$\begin{aligned}\sum_{s=1}^{u-1} \text{Var}(Y_s^{(u)}|\mathcal{F}_{u-s}) &\leq \sum_{s=1}^{u-1} 4(u-s+\gamma)^2(\xi^{(t)}_{u-s})^2\lambda^2\sigma^2 + 4\lambda^2 G^2, \\ &\overset{(27)}{\leq} 12\lambda^2\sigma^2(u+\gamma)^2(u+1)\log(u+1)(\xi_u^{(t)})^2 + 4\lambda^2\gamma^2G^2u.\end{aligned}$$

Step (a) follows since $\eta m < 1$. Now applying the bound in Equation (33) with $B := 2\lambda(u+\gamma)\xi_u^{(t)} + 2\lambda G$ and $v = 12\lambda^2\sigma^2(u+\gamma)^2(u+1)\log(u+1)(\xi_u^{(t)})^2 + 4\lambda^2\gamma^2G^2u$, we get that with probability at-least $1 - \delta/(2t^2(t+1))$,

$$\begin{aligned}\sum_{s=1}^{u-1}(u-s+\gamma)\langle \upsilon_{u-s}, \psi^{(v)}_{u-s+1}\rangle &\leq 2\lambda\left((u+\gamma)\xi_u^{(t)} + R_1\right)\ln\left(\frac{2t^2(t+1)}{\delta}\right) + \left[\left(2\lambda\left((u+\gamma)\xi_u^{(t)} + G\right)\ln\left(\frac{2t^2(t+1)}{\delta}\right)\right)^2\right. \\ &\quad \left. + \left(\lambda^2\sigma^2(u+\gamma)^2(u+1)\log(u+1)(\xi_u^{(t)})^2 + 4\lambda^2\gamma^2G^2(u+1)\right)\ln\left(\frac{2t^2(t+1)}{\delta}\right)\right]^{\frac{1}{2}}, \\ &\leq 6(u+\gamma)\sqrt{u+1}\log(u+1)(\xi_u^{(t)})\lambda\sigma(\sigma+1)\ln\left(\frac{2t^2(t+1)}{\delta}\right) + 2\lambda\gamma G\sqrt{(u+1)\ln\left(\frac{2t^2(t+1)}{\delta}\right)}.\end{aligned}$$

Thus,

$$\begin{aligned}-2\eta_u\sum_{s=1}^{u-1}\frac{(u-s+\gamma)\langle \upsilon_{u-s}, \psi^{(v)}_{u-s+1}\rangle}{u+\gamma} &\leq \frac{12\sqrt{u+1}\log(u+1)(\xi_u^{(t)})\lambda\sigma(\sigma+1)\ln\left(\frac{2t^2(t+1)}{\delta}\right)}{(u+\gamma)} + \frac{C_t\gamma G}{10(u+1)}, \\ &\leq \frac{\xi_u^{(t)}\ln\left(\frac{2t^2(t+1)}{\delta}\right)}{10\sqrt{u+1}} + \frac{C_tG}{10(u+1)}.\end{aligned}$$

The first inequality follows since $C_t \geq \frac{8\lambda\sqrt{\ln\left(\frac{2t^3}{\delta}\right)}}{\gamma^2 G}$. The last inequality follows since for all times $u \leq t$, we have

$$\frac{12\sqrt{u+1}\log(u+1)\lambda\sigma(\sigma+1)\ln\left(\frac{2t^2(t+1)}{\delta}\right)}{(u+\gamma)} \leq \frac{\ln\left(\frac{2t^2(t+1)}{\delta}\right)}{10}$$

as a consequence of $\gamma \geq 120\lambda\sigma(\sigma+1)$.

□

C PROOFS FROM SECTION 4.2

C.1 PROOF OF THEOREM 4.1

We bound this probability using the result of 3.1 and a simple union bound argument. For any process $\mathfrak{M}$, observe that

$$\begin{aligned}\mathbb{P}[\exists t \in [r+1, \tau_c^{(r)}) \text{ s.t.} \mathcal{A}_t = 1|\mathcal{A}_r = 1] &= \mathbb{P}[\cup_{t=r+1}^{\tau_c-1}\mathcal{A}_t = 1|\mathcal{A}_r = 1] \\ &\leq \sum_{t=r+1}^{\tau_c-1}\mathbb{P}[\mathcal{A}_t = 1|\mathcal{A}_r = 1]. \qquad (34)\end{aligned}$$

We now examine the above Equation to bound it. For any fixed $t \in (r, \tau_c^{(r)})$

$$
\begin{aligned}
\mathbb{P}[\mathcal{A}_t = 1|\mathcal{A}_r = 1] &= \mathbb{P}\left[\bigcup_{s=r+1}^{t-1} \|\widehat{\theta}_{r:s} - \widehat{\theta}_{s+1:t}\| \geq \mathcal{B}\left(s-r, \frac{\delta}{2t(t+1)}\right) + \mathcal{B}\left(t-s-1, \frac{\delta}{2t(t+1)}\right)\right], \\
&\leq \sum_{s=r+1}^{t-1}\left(\mathbb{P}\left[\|\widehat{\theta}_{r:s} - \theta_{c-1}\| \geq \mathcal{B}\left(s-r, \frac{\delta}{2t(t+1)}\right)\right] + \mathbb{P}\left[\|\widehat{\theta}_{s+1:t} - \theta_{c-1}\| \geq \mathcal{B}\left(t-s-1, \frac{\delta}{2t(t+1)}\right)\right]\right), \\
&\overset{(a)}{\leq} \sum_{s=r+1}^{t-1}\left(\frac{\delta}{2t(t+1)(s-r)(s-r+1)} + \frac{\delta}{2t(t+1)(t-s-1)(t-s)}\right), \\
&= \frac{\delta}{2t(t+1)}\left(\sum_{s=r+1}^{t-1}\frac{1}{(s-r)(s-r+1)} + \sum_{s=r+1}^{t-1}\frac{1}{(t-s-1)(t-s)}\right), \\
&\leq \frac{\delta}{2t(t+1)}\left(\sum_{s=1}^{t-1-r}\frac{1}{s(s+1)} + \sum_{s=1}^{t-1-r}\frac{1}{s(s+1)}\right), \\
&\overset{(b)}{\leq} \frac{\delta}{t(t+1)}.
\end{aligned} \tag{35}
$$

Since for all $t < \tau_c^{(r)}$, the mean of the random variables $X_{r+1}, \cdots, X_t$ are identical and equal to θ_{c-1} (see notation in Section 2), Theorem 3.1 gives rise to inequality (a). Step (b) follows from the fact that $\sum_{s\geq 1}\frac{1}{s(s+1)} = 1$. Now substituting the bound from Equation (35) into Equation (34), we get that

$$
\begin{aligned}
\mathbb{P}[\exists t \in [r+1, \tau_c^{(r)}) \text{ s.t. } \mathcal{A}_t = 1|\mathcal{A}_r = 1] &\leq \sum_{t=r+1}^{\tau_c - 1}\frac{\delta}{t(t+1)}, \\
&\leq \sum_{t\geq 1}\frac{\delta}{t(t+1)}, \\
&= \delta.
\end{aligned}
$$

Since the above bound holds for all r and process $\mathfrak{M}$, we have

$$
\sup_{\mathfrak{M}, r}\mathbb{P}[\exists t \in [r+1, \tau_c^{(r)}) \text{ s.t.} \mathcal{A}_t = 1|\mathcal{A}_r = 1] \leq \delta.
$$

C.2 PROOF OF LEMMA 4.2

Recall from the definition that the rth detection is false if

$$
\chi_r^{(A)} = \mathbf{1}(\nexists c \text{ s.t. } \tau_c \in (t_{r-1}^{(A)}, t_r^{(A)}]).
$$

We will show that $\mathbb{E}[\chi_r^{(A)}] \leq \delta$. This will then conclude the proof of the lemma.

$$
\begin{aligned}
\mathbb{E}[\chi_r^{(A)}] &= \mathbb{P}[\nexists c \text{ s.t. } \tau_c \in (t_{r-1}^{(A)}, t_r^{(A)}]], \\
&= \mathbb{E}\left[\mathbb{P}[\nexists c \text{ s.t. } \tau_c^{(s)} \in (s, t_r^{(A)}]]\middle| t_{r-1}^{(A)} = s\right], \\
&\leq \mathbb{E}\left[\mathbb{P}[\cup_{t=s+1}^{\infty}\tau_c^{(s)} = t, t_r^{(A)} < t]\middle| t_{r-1}^{(A)} = s\right], \\
&\leq \mathbb{E}\left[\mathbb{P}[\exists t \in [s+1, \tau_c^{(s)}), \mathcal{A}_t = 1]\middle| t_{r-1}^{(A)} = s\right], \\
&\overset{(a)}{\leq} \mathbb{E}\left[\mathbb{P}[\exists t \in [s+1, \tau_c^{(s)}), \mathcal{A}_t = 1|\mathcal{A}_s = 1]\middle| t_{r-1}^{(A)} = s\right],
\end{aligned}
$$

$$\overset{(b)}{\leq} \delta.$$

Inequality (a) follows from the fact that on the event $t_{r-1}^{(\mathcal{A})} = s$, $\mathcal{A}_s = 1$. Inequality (b) follows from Theorem 4.1.

D PROOF OF LEMMA 4.3

The proof follows from a straightforward application of Theorem 3.1 as follows. Let $n \in \mathbb{N}, \Delta > 0$ and $\delta' \in (0,1)$ be arbitrary.

$$\begin{aligned}\mathbb{P}[\mathcal{D}(n,\Delta,\delta') \geq d] &= \mathbb{P}[\cap_{s=1}^{n+d} \mathcal{A}(X_{1:s}) = 0], \\ &= \mathbb{P}\left[\bigcap_{s=1}^{n+d} \|\widehat{\theta}_{1:s} - \widehat{\theta}_{s+1:n+d}\|_2^2 \leq \mathcal{B}\left(s, \frac{\delta}{2(n+d)(n+d+1)}\right) + \mathcal{B}\left(n+d-s-1, \frac{\delta}{2(n+d)(n+d+1)}\right)\right], \\ &\leq \mathbb{P}\left[\|\widehat{\theta}_{1:n-1} - \widehat{\theta}_{n:n+d}\|_2^2 \leq \mathcal{B}\left(n-1, \frac{\delta}{2(n+d)(n+d+1)}\right) + \mathcal{B}\left(d, \frac{\delta}{2(n+d)(n+d+1)}\right)\right].\end{aligned} \tag{36}$$

From triangle-inequality, we know that

$$\begin{aligned}\|\widehat{\theta}_{1:n-1} - \widehat{\theta}_{n:n+d}\|_2^2 &\geq \|\theta_1 - \theta_2\|_2^2 - \|\widehat{\theta}_{1:n-1} - \theta_1\|_2^2 - \|\widehat{\theta}_{n:n+d} - \theta_2\|_2^2, \\ &= \Delta^2 - \|\widehat{\theta}_{1:n-1} - \theta_1\|_2^2 - \|\widehat{\theta}_{n:n+d} - \theta_2\|_2^2.\end{aligned} \tag{37}$$

Thus, substituting Equation (37 into Equation (36), we get that

$$\mathbb{P}[\mathcal{D}(n,\Delta,\delta') \geq d] \leq \mathbb{P}\Bigg[\Delta^2 - \|\widehat{\theta}_{1:n-1} - \theta_1\|_2^2 - \|\widehat{\theta}_{n:n+d} - \theta_2\|_2^2 \leq \mathcal{B}\left(n-1, \frac{\delta}{2(n+d)(n+d+1)}\right) + \mathcal{B}\left(d, \frac{\delta}{2(n+d)(n+d+1)}\right)\Bigg].$$

Denote by the events $\mathcal{E}_i$ for $i \in \{1,2\}$ as

$$\begin{aligned}\mathcal{E}_1 &:= \left\{\|\widehat{\theta}_{1:n-1} - \theta_1\|_2^2 > \mathcal{B}\left(n-1, \frac{\delta'}{2}\right)\right\}, \\ \mathcal{E}_2 &:= \left\{\|\widehat{\theta}_{n:n+d} - \theta_2\|_2^2 > \mathcal{B}\left(d, \frac{\delta'}{2}\right)\right\},\end{aligned}$$

Denote by $\mathcal{E} := \mathcal{E}_1 \cup \mathcal{E}_2$. Theorem 3.1 gives that $\mathbb{P}[\mathcal{E}_1] \leq \frac{\delta'}{2(n(n+1))} \leq \frac{\delta'}{2}$ and $\mathbb{P}[\mathcal{E}_2] \leq \frac{\delta'}{2d(d+1)} \leq \frac{\delta'}{2}$. Thus, an union bound gives that $\mathbb{P}[\mathcal{E}] \leq \delta'$. Let $d' \in \mathcal{G}$ be arbitrary, where

$$\mathcal{G} := \left\{d \in \mathbb{N} : \Delta^2 \geq \mathcal{B}\left(n-1, \frac{\delta'}{2}\right) + \mathcal{B}\left(d, \frac{\delta'}{2}\right) + \mathcal{B}\left(n, \frac{\delta}{2(n+d+1)(n+d)}\right) + \mathcal{B}\left(d, \frac{\delta}{2(n+d+1)(n+d)}\right)\right\} \tag{38}$$

Claim : If the event $\mathcal{E}^c$ holds, then $\mathcal{D}(n,\Delta,\delta) \leq d$ for all $d \in \mathcal{G}$.

Suppose $d \in \mathcal{G}$ and event $\mathcal{E}^c$ holds. Then, we know by triangle inequality in Equation (37) that

$$\begin{aligned}\|\widehat{\theta}_{1:n-1} - \widehat{\theta}_{n:n+d}\|_2^2 &\geq \|\theta_1 - \theta_2\|_2^2 - \|\widehat{\theta}_{1:n-1} - \theta_1\|_2^2 - \|\widehat{\theta}_{n:n+d} - \theta_2\|_2^2, \\ &= \Delta^2 - \|\widehat{\theta}_{1:n-1} - \theta_1\|_2^2 - \|\widehat{\theta}_{n:n+d} - \theta_2\|_2^2,\end{aligned} \tag{39}$$

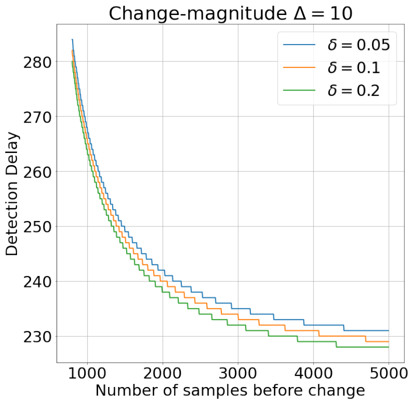


Figure 5: Plot of $\mathcal{D}(n, \Delta, \delta')$ in Lemma 4.3 for fixed $\Delta = 10, \delta = 0.1$.

$$\overset{(a)}{\geq} \Delta^2 - \mathcal{B}\left(n-1, \frac{\delta'}{2}\right) - \mathcal{B}\left(d, \frac{\delta'}{2}\right), \tag{40}$$

$$\overset{(b)}{\geq} \mathcal{B}\left(n, \frac{\delta}{2(n+d+1)(n+d)}\right) + \mathcal{B}\left(d, \frac{\delta}{2(n+d+1)(n+d)}\right). \tag{41}$$

Step (a) follows from the definition of event $\mathcal{E}$ and on the assumption of the claim that event $\mathcal{E}^c$ holds. Step (b) follows from the fact that $d \in \mathcal{G}$ is arbitrary (cf. Equation (38). The last step says from Line 8 of Algorithm 1 that if no detection has been made till time $n + d$, then under the event $\mathcal{E}^c$, time step d is a detection time. Since event $\mathcal{E}^c$ holds with probability at-least $1 - \delta'$, this concludes the proof.

D.1 USEFUL CONVEXITY BASED INEQUALITIES

Let $f : \Theta \to \mathbb{R}$ be a strongly convex function with strong convexity parameters $0 < m \leq M < \infty$. Denote by $\theta^* := \arg\min_{\theta\in\Theta} f(\theta)$. Since $f(\cdot)$ is convex and Θ is convex and compact, the existence and uniqueness of θ^* is guaranteed. Strong convexity gives that for any $\widehat{\theta}_{t-1} \in \Theta$,

$$f(\theta^*) \geq f(\widehat{\theta}_{t-1}) + \langle \nabla f(\widehat{\theta}_{t-1}), \theta^* - \widehat{\theta}_{t-1} \rangle + \frac{m}{2} \|\theta^* - \widehat{\theta}_{t-1}\|_2^2. \tag{42}$$

Further since $\theta^* = \arg\min_{\theta\in\Theta} f(\theta)$., we have that

$$f(\widehat{\theta}_{t-1}) - f(\theta_t^*) \geq \frac{m}{2} \|\widehat{\theta}_{t-1} - \theta^*\|_2^2.$$

Putting these two together, we see that

$$\langle \nabla f(\widehat{\theta}_{t-1}), \widehat{\theta}_{t-1} - \theta^* \rangle \geq m \|\widehat{\theta}_{t-1} - \theta^*\|_2^2. \tag{43}$$

Also, We further use the following lemma.

Lemma D.1 (Lemma 3.11 from [Bubeck, 2015]). *Let $g : \mathbb{R}^d \to \mathbb{R}$ be a M smooth and m strongly convex function. Then for all $x, y \in \mathbb{R}^d$,*

$$\langle \nabla g(x) - \nabla g(y), x - y \rangle \geq \frac{mM}{M+m} \|x - y\|_2^2 + \frac{1}{M+m} \|\nabla g(x) - \nabla g(y)\|_2^2.$$

By substituting $x = \widehat{\theta}_{t-1}$, $y = \theta_t^*$ and $g(\cdot) = f(\cdot)$ and by leveraging the fact that $\nabla f(\theta^*) = 0$, we get the inequality that

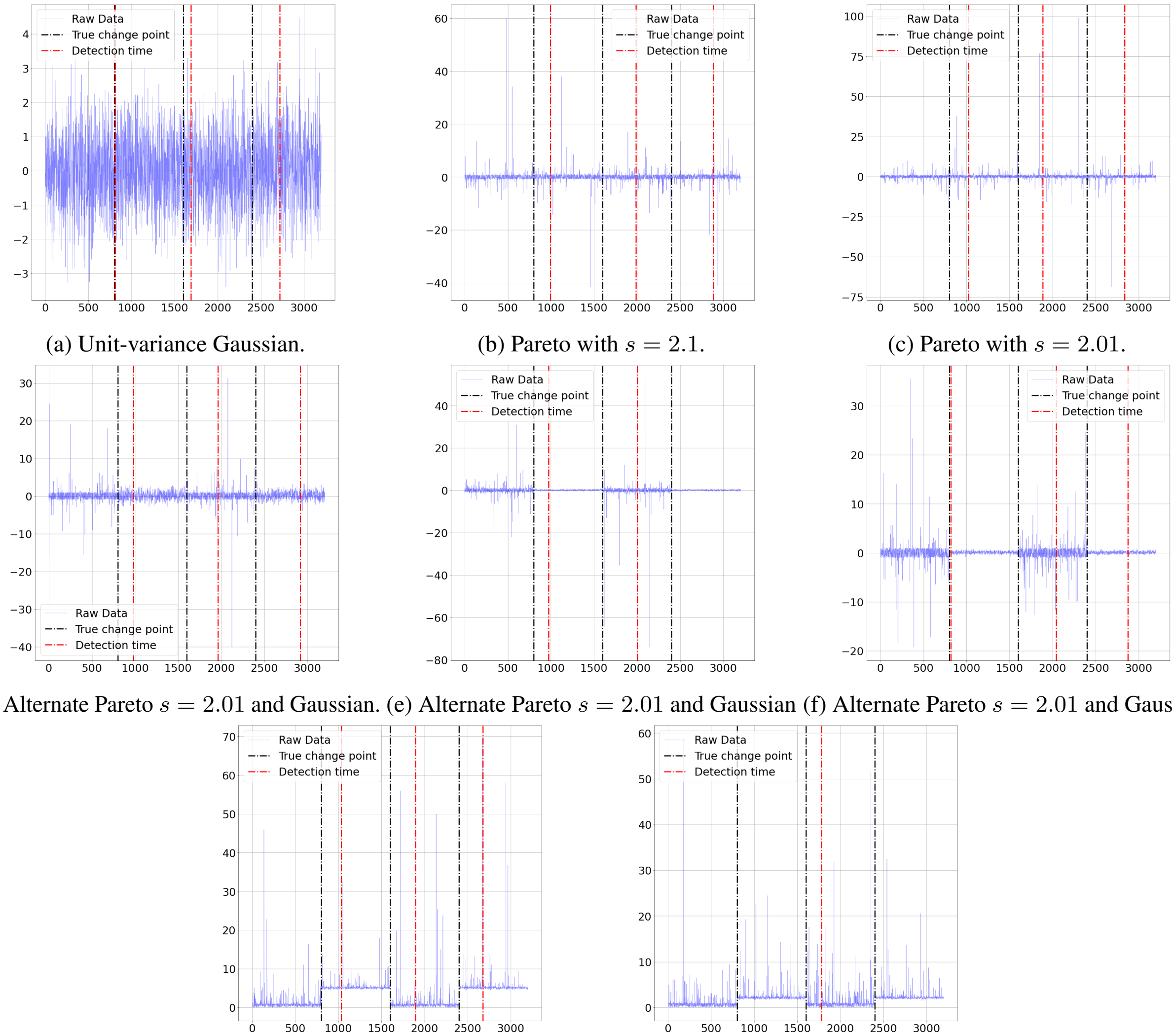
$$\langle \nabla f(\widehat{\theta}_{t-1}), \widehat{\theta}_{t-1} - \theta^* \rangle \geq \frac{mM}{m+M} \|\widehat{\theta}_{t-1} - \theta^*\|_2^2 + \frac{1}{M+m} \|\nabla f(\widehat{\theta}_{t-1})\|_2^2.$$

Re-arranging, we see that

$$\|\nabla f(\widehat{\theta}_{t-1})\|_2^2 \leq (M+m) \langle \nabla f(\widehat{\theta}_{t-1}), \widehat{\theta}_{t-1} - \theta^* \rangle - mM \|\widehat{\theta}_{t-1} - \theta^*\|_2^2. \tag{44}$$

E ADDITIONAL SIMULATIONS

In Figure 6, we plot a sample path of observed data and mark out the true change-points and the detected time-instants by Algorithm 1. The plots indicate that although visually identifying the change in the means is hard, our change-point detection algorithm is able to consistently across variety of distribution families.



(a) Unit-variance Gaussian.

(b) Pareto with $s = 2.1$.

(c) Pareto with $s = 2.01$.

(d) Alternate Pareto $s = 2.01$ and Gaussian.

(e) Alternate Pareto $s = 2.01$ and Gaussian

(f) Alternate Pareto $s = 2.01$ and Gaussian

(g) Pareto $s = 2.01, d = 15, \Delta = 5$

(h) Pareto $s = 2.01, d = 15, \Delta = 2$

Figure 6: In all plots, we choose the change-point gap to be $\Delta = 0.1$ and $\delta = 0.05$ except (g) and (h) where $\Delta = 5$ and 2 respectively. In plots (g) and (h), we plot the norm of the observed random vector and thus the Y-axis is non-negative. We see missed detection in Figures (e) and (h) with the last change-point on the right being missed. We do not observe False-positives in these plots.