High-dimensional, multiscale online changepoint detection

Yudong Chen, Tengyao Wang and Richard J. Samworth

(March 7, 2020)

Abstract

We introduce a new method for high-dimensional, online changepoint detection in settings where a $p$-variate Gaussian data stream may undergo a change in mean. The procedure works by performing likelihood ratio tests against simple alternatives of different scales in each coordinate, and then aggregating test statistics across scales and coordinates. The algorithm is online in the sense that its worst-case computational complexity per new observation, namely $O(p^2 \log(ep))$, is independent of the number of previous observations; in practice, it may even be significantly faster than this. We prove that the patience, or average run length under the null, of our procedure is at least at the desired nominal level, and provide guarantees on its response delay under the alternative that depend on the sparsity of the vector of mean change. Simulations confirm the practical effectiveness of our proposal.

1 Introduction

Modern technology has not only allowed the collection of data sets of unprecedented size, but has also facilitated the real-time monitoring of many types of evolving processes of interest. Wearable health devices, astronomical survey telescopes, self-driving cars and transport network load-tracking systems are just a few examples of new technologies that collect large quantities of streaming data, and that provide new challenges and opportunities for statisticians.

Very often, a key feature of interest in the monitoring of a data stream is a changepoint; that is, a moment in time at which the data generating mechanism undergoes a change. Such times often represent events of interest, e.g. a change in heart function, and moreover, the accurate identification of changepoints often facilitates the decomposition of a data stream into stationary segments.

Historically, it has tended to be univariate time series that have been monitored and studied, within the well-established field of statistical process control (e.g. Duncan, 1952; Page, 1954; Barnard, 1959; Fearnhead and Liu, 2007; Oakland, 2007; Tartakovsky, Nikiforov and Basseville, 2014). These days, however, it is frequently the case that many data processes are measured simultaneously. In the context of changepoint detection, this introduces the new
challenge of borrowing strength across the different component series in an attempt to detect much smaller changes than would be possible through the observation of any individual series alone.

The field of changepoint detection and estimation also has a long history (e.g. Page, 1955), but has been undergoing a marked renaissance in recent years; entry points to the field include Csörgő and Horváth (1997) and Horváth and Rice (2014). However, the vast majority of this ever-growing literature has focused on the offline changepoint problem, where, after the entire data stream is observed, the statistician is asked to identify any changepoints retrospectively. For univariate, offline changepoint estimation, state-of-the-art methods include the Pruned Exact Linear Time method (PELT) (Killick, Fearnhead and Eckley, 2012), Narrowest-Over-Threshold (NOT) (Baranowski, Chen and Fryzlewicz, 2019), Simultaneous Multiscale Changepoint Estimator (SMUCE) (Frick, Munk and Sieling, 2014) and $\ell_0$-penalisation (Wang, Yu and Rinaldo, 2018), while work on multivariate and high-dimensional offline changepoints includes the double CUSUM method of Cho (2016), the inspect algorithm of Wang and Samworth (2018), as well as Enikeeva and Harchaoui (2019), Liu, Gao and Samworth (2019) and Padilla et al. (2019).

Despite this rich literature on offline changepoint problems, it is the online version of the problem that is arguably the more important for many applications: one would like to be able to detect a change as soon as possible after it has occurred. Of course, one option here is to apply an offline method after seeing every new observation (or batch of observations). However, this is unlikely to be a successful strategy: not only is there a difficult and highly dependent multiple testing issue to handle when using the method repeatedly on data sets of increasing size, but moreover, the storage and running time costs may frequently be prohibitive.

In this work, we are interested in algorithms for detecting changepoints in high-dimensional data that are observed sequentially. In order to avoid the trap mentioned in the previous paragraph and ensure that any methods we consider can be applied to large data streams, we will focus our attention on online algorithms. By this, we mean that the computational complexity for processing each new observation should depend only on the number of bits needed to represent the new data point observed, and not on the storage requirements of any previously observed data. This turns out to be a very stringent requirement, in the sense that finding online algorithms with good statistical performance is typically extremely challenging. Online algorithms must necessarily store only compact summaries of the historical observations, so the class of all possible procedures is severely restricted.

To set the scene for our contributions, let $X_1, X_2, \ldots$ be a sequence of independent random vectors in $\mathbb{R}^p$. Assume that for some unknown, deterministic time $z \in \mathbb{N} \cup \{0\}$, the sequence is generated according to

$$X_1, \ldots, X_z \sim N_p(\mu_-, I_p) \quad \text{and} \quad X_{z+1}, X_{z+2}, \ldots \sim N_p(\mu_+, I_p),$$

for some $\mu_-, \mu_+ \in \mathbb{R}^p$. When $\mu_+ \neq \mu_-$, we say that there is a changepoint at time $z$. In many applications, such as in industrial quality control where the distribution of relevant properties of goods in a manufacturing process under regular conditions may be well understood, we may assume that the mean before the change is known (or at least can be estimated to high accuracy using historical data). However, the vector of change, $\theta := \mu_+ - \mu_-$, is typically
unknown. Thus, for simplicity, we will work in the setting where \( \mu_- = 0 \) and \( \mu_+ = \theta \). Let \( \mathbb{P}_{z,\theta} \) denote the joint distribution of \( (X_n)_{n=1}^{\infty} \) under (1) and \( \mathbb{E}_{z,\theta} \) the expectation under this distribution. Note that when \( \theta = 0 \), the joint distribution of the data does not depend on \( z \), and we therefore let \( \mathbb{P}_{0} = \mathbb{P}_{z,0} \) denote this joint distribution (with corresponding expectation \( \mathbb{E}_{0} \)). We will then say that the data is generated under the null. By contrast, if \( \theta \neq 0 \), we will say that the data is generated under the alternative, though we emphasise that in fact the alternative is composite, being indexed by \( z \in \mathbb{N} \) and \( \theta \in \mathbb{R}^p \setminus \{0\} \). In practice, in order for our procedure to have uniformly non-trivial power, it will be necessary to work with a subset of the alternative hypothesis parameter space that is well-separated from the null, in the sense that the \( \ell_2 \)-norm of the vector of mean change, \( \vartheta := ||\theta||_2 \), is at least a known lower bound \( \beta > 0 \).

A sequential changepoint procedure is an extended stopping time \( N \) (with respect to the natural filtration) taking values in \( \mathbb{N} \cup \{\infty\} \). Equivalently, we can think of it as a family of \( \{0,1\} \)-valued estimators \( \{\hat{H}_n\}_{n=1}^{\infty} \), where \( \hat{H}_n = \hat{H}_n(X_1,\ldots,X_n) \), and where the sequence is increasing in the sense that \( \hat{H}_m(X_1,\ldots,X_m) \leq \hat{H}_n(X_1,\ldots,X_n) \) for \( m \leq n \). Here, the correspondence arises from \( \hat{H}_n = 1_{\{N \leq n\}} \) and \( N = \inf\{n \in \mathbb{N} : \hat{H}_n = 1\} \), with the usual convention that \( \inf\emptyset := \infty \).

We measure the performance of a sequential changepoint procedure via its responsiveness subject to a given upper bound on the false alarm rate, or equivalently, a lower bound on the average run length in the absence of change. Specifically, following the concepts introduced by Lorden (1971), we define the patience\(^1\) of a sequential changepoint procedure \( N \) to be \( \mathbb{E}_{0}(N) \), and its worst-case response delay\(^2\) to be

\[
\bar{E}_{wc}(N) := \sup_{z \in \mathbb{N}} \text{ess sup} \mathbb{E}_{z,\theta}\{(N-z) \lor 0 \mid X_1,\ldots,X_z\}.
\]

While controlling the worst-case response delay provides a very strong theoretical guarantee of the average detection delay of the procedure, even under the worst possible pre-change data sequence, obtaining a good bound for this quantity is often difficult. We therefore also consider the average-case response delay, or simply the response delay of a procedure \( N \), defined as

\[
\bar{E}_{\theta}(N) := \sup_{z \in \mathbb{N}} \mathbb{E}_{z,\theta}\{(N-z) \lor 0\}.
\]

We note that \( \bar{E}_{\theta}(N) \leq \bar{E}_{wc}(N) \). A good sequential changepoint procedure should have small worst- and average-case response delays, uniformly over the relevant class of alternatives \( \{\mathbb{P}_{z,\theta} : (z,\theta) \in \mathbb{N} \times \mathbb{R}^p, ||\theta||_2 \geq \beta\} \), subject to its patience being at least some suitably large, pre-determined \( \gamma > 0 \). Finally, as mentioned above, we are interested in sequential changepoint procedures that are online, so that the computational complexity per additional observation should be a function of \( p \) only.

Our main contribution in this work is to propose, in Section 2, a new algorithm called ocd (short for online changepoint detection), for high-dimensional, online changepoint detection in the above setting. The procedure works by performing likelihood ratio tests against

---

\(^1\)This is sometimes referred to as the average run length under the null or average run length to false alarm in the literature.

\(^2\)Likewise, this is sometimes referred to as the worst-worst-case average detection delay.
simple alternatives of different scales in each coordinate, and then aggregating test statistics across scales and coordinates for changepoint detection. The \texttt{ocd} algorithm has worst-case computational complexity \(O(p^2 \log(ep))\) per new observation, so satisfies our requirement for being an online algorithm. In fact, as we explain in Section 2.1, the algorithmic complexity is often even better than this. Moreover, as we illustrate in Section 4, it has extremely effective empirical performance. In terms of theoretical guarantees, it turns out to be more convenient to analyse a slight variant of our initial algorithm, which we refer to as \texttt{ocd}'. This has the same order of computational complexity per new observation as \texttt{ocd}, but enables us to ensure that whenever we are yet to declare that a change has occurred at the changepoint, only post-change observations contribute to the running test statistics. In practice, the original \texttt{ocd} algorithm also appears to have this property for typical pre-change sequences, and we argue heuristically that there is a sense in which it is more efficient than \texttt{ocd}' by a factor of at most 2.

Our theoretical analysis in Section 3 initially considers separately versions of the \texttt{ocd}' algorithm best tuned towards settings where the vector \(\theta\) of change is dense, and where it is sparse in an appropriate sense. We then present results for a combined, adaptive procedure that seeks the best of both worlds. In all cases, the appropriate version of \texttt{ocd}' has guaranteed patience, at least at the desired nominal level. In the (small-change) regime of primary interest, and when \(\vartheta\) is of the same order as \(\beta\), the response delay of \texttt{ocd}' is of order at most \(\sqrt{p}/\vartheta^2\) in the dense case, up to a poly-logarithmic factor; this can be improved to order \(s/\vartheta^2\), again up to a poly-logarithmic factor, when the effective sparsity of \(\theta\) is \(s < \sqrt{p}\).

As alluded to above, there is a paucity of prior literature on multivariate, online change-point problems, though exceptions include Tartakovsky et al. (2006), Mei (2010) and Zou et al. (2015). These works focus either on the case where both the pre- and post-change distributions are exactly known, or where, for each coordinate, both the sign and a lower bound on the magnitude of change, are known in advance. A number of methods have also been proposed that involve scanning a moving window of fixed size for changes (Xie and Siegmund, 2013; Soh and Chandrasekaran, 2017; Chan, 2017). Such methods can be effective when the signal-to-noise ratio is large enough that the change can be detected within the prescribed window, but may experience excessive response delay in other cases. Of course, the window size may be increased to compensate, but this correspondingly increases the computational complexity and storage requirements, so allowing the window size to vary with the signal strength would fail to satisfy our definition of an online algorithm.

Numerical results illustrate the performance of our \texttt{ocd} algorithm in Section 4. Proofs of our main results are given in Section 5, with various auxiliary results deferred to Section 6.

1.1 Notation

We write \(\mathbb{N}_0\) for the set of all non-negative integers. For \(d \in \mathbb{N}\), we write \([d] := \{1, \ldots, d\}\). Given \(a, b \in \mathbb{R}\), we denote \(a \vee b := \max(a, b)\) and \(a \wedge b := \min(a, b)\). For a set \(S\), we use \(1_S\) and \(|S|\) to denote its indicator function and cardinality respectively. For a real-valued function \(f\) on a totally ordered set \(S\), we write \(\text{sargmax}_{x \in S} f(x) := \min \text{argmax}_{x \in S} f(x)\), the smallest maximiser of \(f\) in set \(S\). For a vector \(v = (v^1, \ldots, v^M)^T \in \mathbb{R}^M\), we define \(\|v\|_0 := \sum_{i=1}^M 1_{\{v^i \neq 0\}}\), \(\|v\|_2 := \left\{ \sum_{i=1}^M (v^i)^2 \right\}^{1/2}\) and \(\|v\|_\infty := \max_{i \in [M]} |v^i|\). In addition, we
define $\|v^{-j}\|_2 := \{\sum_{i:i\neq j}(v^i)^2\}^{1/2}$. For a matrix $A = (A^{i,j}) \in \mathbb{R}^{d_1 \times d_2}$ and $j \in [d_2]$, we write $A^{:-j} := (A^{1,j}, \ldots, A^{d_1,j}) \in \mathbb{R}^{d_1}$ and $A^{-j:j} := (A^{1,j}, \ldots, A^{j-1,j}, A^{j+1,j}, \ldots, A^{d_1,j})^\top \in \mathbb{R}^{d_1-1}$. We use $\Phi(\cdot)$ and $\phi(\cdot)$ to denote the distribution function and density function of the standard normal distribution respectively. For two real-valued random variables $U$ and $V$, we write $U \geq_{st} V$ if $\mathbb{P}(U \leq x) \leq \mathbb{P}(V \leq x)$ for all $x \in \mathbb{R}$. We adopt the convention that an empty sum is 0.

2 An online changepoint procedure

2.1 The ocd algorithm

In this section, we describe our online changepoint procedure, ocd, in more detail. As mentioned in the introduction, the procedure aggregates likelihood ratio test statistics against simple alternatives of different scales in different coordinates. If we want to test a null of $N(0,1)$ against a simple post-change alternative distribution of $N(b,1)$ for some $b \neq 0$ in coordinate $j \in [p]$, by Page (1954), the optimal online changepoint procedure is to declare a change has occurred by time $n$ when the test statistic

$$R^j_{n,b} := \max_{0 \leq h \leq n} \sum_{i=n-h+1}^{n} b(X^j_i - b/2)$$

(2)

exceeds a certain threshold. Note that $\sum_{i=n-h+1}^{n} b(X^j_i - b/2)$ can be viewed as the likelihood ratio test statistic between the null and this simple alternative using the tail sequence $X_{n-h+1}, \ldots, X_n$. Thus $R^j_{n,b}$ can be regarded as the most extreme of these likelihood ratio statistics, over all possible starting points for the tail sequence. Write

$$t^j_{n,b} := \text{sargmax}_{0 \leq h \leq n} \sum_{i=n-h+1}^{n} b(X^j_i - b/2)$$

(3)

for the length of the tail sequence in which the associated likelihood ratio statistic (in the $j$th coordinate) is maximised. One way to aggregate across the $p$ coordinates would be to use $\sum_{j=1}^{p} R^j_{n,b}$ as a test statistic. However, this approach is not ideal for two reasons. Firstly, the exact distribution of the tail likelihood ratio statistic $R^j_{n,b}$ is hard to obtain, making it difficult to analyse the aggregated statistic under the null. More importantly, this aggregated statistic uses the same simple alternative $N(b,1)$ in all coordinates, and so even after varying the magnitude of $b$, it is only effective against a very limited set of alternative distributions in $\{\mathbb{P}_{z,\theta} : z \in \mathbb{N}, \|\theta\|_2 \geq \beta\}$, namely those for which the change is of very similar magnitude in all coordinates. To overcome these problems, our procedure uses the coordinate-wise statistics $(R^j_{n,b} : j \in [p])$, which we call ‘diagonal statistics’, to detect changes that have a large proportion of their signal concentrated in one coordinate. Then, for each $j \in [p]$, we also compute tail partial sums of length $t^j_{n,b}$ in all other coordinates $j' \neq j$, given by

$$A^j_{n,h} := \sum_{i=n-t^j_{n,h}+1}^{n} X^j_i,$$
and aggregate them to form an ‘off-diagonal statistic’ anchored at coordinate \( j \). Note that the number of summands in \( A_{n,b}^j \) depends only on the observed data in the \( j \)th coordinate, and not on the data being aggregated in the \( j \)th coordinate. These off-diagonal statistics are used to detect changes whose signal is not concentrated in a single coordinate. Intuitively, if a change has occurred and \( \theta^j / b \geq 1 \), then we can expect the tail length in coordinate \( j \) to be roughly of order \( n - z \) for sufficiently large \( n \), and this will ensure that the off-diagonal statistic anchored at coordinate \( j \) is close to the generalised likelihood ratio test statistic between the null and the composite alternative \( \{ \mathbb{P}_{z, \theta} : \| \theta \|_2 \neq 0 \} \). If, in addition, a non-trivial proportion of the signal is contained in coordinates \( [p] \setminus \{ j \} \), then this statistic will be powerful for detecting the change.

The full description of the \texttt{ocd} procedure is given in Algorithm 1. Note that for notational simplicity, we have suppressed the time dependence of many variables as they are updated recursively in the algorithm. In the following, when necessary, we will make this dependence explicit by writing \( A_{n,b}, t_{n,b}, Q_{n,b}, S_n^{\text{diag}} \) and \( S_n^{\text{off}} \) for the relevant quantities at the end of the \( n \)th iteration of the repeat loop.

By Lemma 10, \( bA_{n,b}^j - b^2 t_{n,b}^j / 2 \), as defined in the algorithm, is equal to the quantity \( R_{n,b}^j \) defined in (2) (we will also suppress its \( n \) dependence when it is clear from the context). Moreover, by Lemma 11, the two definitions of \( t_{n,b}^j \) from Algorithm 1 and (3) coincide. In the algorithm, we allow \( b \) to range over the (signed) dyadic grid \( \mathcal{B} \cup \mathcal{B}_0 \), since the maximal signal strength in individual coordinates, \( \| \theta \|_\infty \), can range from \( \vartheta / \sqrt{p} \) to \( \vartheta \). In this way, the algorithm automatically adapts to different signal strengths in each coordinate. Here, the inclusion of \( \mathcal{B}_0 \) and the extra logarithmic factors in the denominators of elements of \( \mathcal{B} \cup \mathcal{B}_0 \) appear due to technical reasons in the theoretical analysis of the algorithm.

Algorithm 1 uses \( S^{\text{diag}} \) and \( S^{\text{off}} \) to aggregate diagonal and off-diagonal statistics respectively as mentioned above, and declares that a change has occurred as soon as either of these quantities exceeds its own pre-determined threshold. As mentioned previously, \( S^{\text{diag}} \) tracks the maximum of \( R_{n,b}^j \) over all scales \( b \) and coordinates \( j \). Before introducing \( S^{\text{off}} \), we first discuss the off-diagonal statistics \( Q_{b}^j \) in Algorithm 1, which are \( \ell_2 \) aggregations of normalised tail sums \( A_{b}^j \sqrt{t_{b}^j} \vee 1 \), each hard-thresholded at level \( a \). The hard thresholding level can be chosen to detect dense or sparse signals \( \theta \); in the sparse case a non-zero \( a \) facilitates an aggregation that aims to exclude coordinates with negligible change (thereby reducing the variance of the normalised tail sums). Finally, \( S^{\text{off}} \) is computed as the maximum of the \( Q_{b}^j \) over all anchoring coordinates \( j \in [p] \) and scales \( b \in \mathcal{B} \).

Although the off-diagonal statistics described in the previous paragraph are effective for detecting changes when the signal sparsity is known, it is desirable to the practitioner to have a combined procedure that adapts to the sparsity level. This may be computed straightforwardly by tracking \( S^{\text{off}} \) for \( a = a^{\text{dense}} \) and \( a = a^{\text{sparse}} \), as well as \( S^{\text{diag}} \), and declaring a change when any of these three statistics exceeds a suitable threshold. Figure 1 illustrates the performance of this adaptive procedure, together with the time evolution of normalised versions of all three statistics tracked, in synthetic datasets both with and without a change. This adaptive procedure is analysed theoretically in Section 3.3 and empirically in Section 4.

The \texttt{ocd} procedure satisfies our definition of an online algorithm. Indeed, for each new observation \( X_n \), \texttt{ocd} updates \( t_{n,b} \in \mathbb{R}^p \) and \( A_{n,b} \in \mathbb{R}^{p \times p} \) for \( O((\log(ep)) \) different values of \( b \). It
Figure 1: Behaviour of the three normalised statistics in ocd under the null and under the alternative with different signal strength, sparsity level and assumed lower bound. A change is declared as soon as one of these three normalised statistics exceeds 1. The data were generated in the top-left panel according to $P_0$, and, in the other panels, according to $P_{z, \theta}$, with $p = 100$, $z = 300$ and $\theta = \vartheta U$, where $U$ is uniformly distributed on the union of all $s$-sparse unit spheres in $\mathbb{R}^p$ (see Section 4.2 for a more detailed description).
then computes $S^\text{diag}_n$ and $S^\text{off}_n$ via $A_{n,b}$. These steps require $O(p^2 \log(ep))$ operations. Moreover, the total storage used is $O(p^2 \log(ep))$ throughout the algorithm.

In fact, the computational complexity of ocd can often be reduced, because typically $\mathcal{T} := \{t_b^j : j \in [p], b \in \mathcal{B}\}$ has cardinality much less than $p|\mathcal{B}|$ (which is the worst case, when all elements are distinct). Correspondingly, at each time step, we need only store the $p \times |\mathcal{T}|$ matrix $(B^{k,t}_j)_{k \in [p], t \in \mathcal{T}}$ given by $B^{k,t}_j := A^{k,j}_{b}$, resulting in an improved per-iteration computational complexity and storage for ocd of $O(p|\mathcal{T}|)$. For simplicity of exposition, we have not presented this computational speed-up in Algorithm 1, and it appears to be difficult to provide theoretical guarantees on $|\mathcal{T}|$. Nevertheless we have implemented the algorithm in this form in the R package ocd (Chen, Wang and Samworth, 2020), and have found it to provide substantial computational savings in practice.

### Algorithm 1: Pseudo-code of the ocd algorithm

**Input:** $X_1, X_2 \ldots \in \mathbb{R}^p$ observed sequentially, $\beta > 0$, $a \geq 0$, $T^\text{diag} > 0$ and $T^\text{off} > 0$

**Set:** $\mathcal{B} = \left\{ \frac{\beta}{\sqrt{2^\ell \log_2(p)}} : \ell = 0, \ldots, \lfloor \log_2(p) \rfloor \right\}$, $\mathcal{B}_0 = \left\{ \frac{\beta}{\sqrt{2^\ell \log_2(p) + 1 \log_2(2p)}} : \right\}, n = 0,$

$A_b = 0 \in \mathbb{R}^{p \times p}$ and $t_b = 0 \in \mathbb{R}^p$ for all $b \in \mathcal{B} \cup \mathcal{B}_0$

**repeat**

$n \leftarrow n + 1$

observe new data vector $X_n$

**for** $(j, b) \in [p] \times (\mathcal{B} \cup \mathcal{B}_0)$ **do**

$t_b^j \leftarrow t_b^j + 1$

$A_b^{j,j} \leftarrow A_b^{j,j} + X_n$

**if** $bA_b^{j,j} - b^2t_b^j/2 \leq 0$ **then**

$t_b^j \leftarrow 0$ and $A_b^{j,j} \leftarrow 0$

**compute** $Q_b^j \leftarrow \sum_{j' \in [p] : j' \neq j} \frac{|A_b^{j',j}|^2}{t_b^{j'}} \mathbb{1}_{|A_b^{j,j}||\geq a\sqrt{t_b^j}}$

$S^\text{diag} \leftarrow \max_{(j,b) \in [p] \times (\mathcal{B} \cup \mathcal{B}_0)} (bA_b^{j,j} - b^2t_b^j/2)$

$S^\text{off} \leftarrow \max_{(j,b) \in [p] \times \mathcal{B}} Q_b^j$

**until** $S^\text{diag} \geq T^\text{diag}$ or $S^\text{off} \geq T^\text{off}$;

**Output:** $N = n$

### 2.2 A slight variant of ocd

While the ocd algorithm performs very well numerically, it turns out to be easier theoretically to analyse a slight variant, which we call ocd$, and describe in Algorithm 2. Again, we have suppressed the time dependence $n$ of many variables including $\tau_{n,b}, \bar{\tau}_{n,b}, \Lambda_{n,b}$ and $\bar{\Lambda}_{n,b}$ in the algorithm. The main difference between these two algorithms is that in ocd$, the off-diagonal statistics $Q_b^j$ are computed using tail partial sums of length $\tau_b^j$ instead of $t_b^j$. These new tail partial sums are recorded in $\Lambda_b \in \mathbb{R}^{p \times p}$.

By Lemma 16, we always have

$$t_b^j/2 \leq \tau_b^j < 3t_b^j/4$$  \hfill (4)
whenever \( t^j_b \geq 2 \). In this sense, the tail sample size used by ocd' is smaller than that of ocd by a factor of at most 2. The benefit of using a shorter tail in ocd' is that when \( n \) exceeds a known, deterministic threshold, we can be sure that whenever we have not declared that a change has occurred by time \( z \), the tail partial sum consists exclusively of post-change observations. In practice, we observe that even in Algorithm 1, the tail lengths \( t^j_{z,b} \) at the changepoint are generally very short for many coordinates, so the inclusion of a few pre-change observations in the tail partial sum calculation does not significantly affect the efficacy of the changepoint detection procedure. The practical performance of Algorithm 1 is statistically more efficient than Algorithm 2 in many settings by a factor of between 4 and 9, as suggested by (4).

Indeed, Algorithm 2 is also an online algorithm, with overall computational complexity per observation and storage remaining at \( O(p^2 \log(ep)) \) in the worst case; similar computational improvements to those mentioned for ocd at the end of Section 2.1 are also possible here.

**Algorithm 2**: Pseudo-code of the ocd' algorithm, a slight variant of ocd

\[
\text{Input: } X_1, X_2 \ldots \in \mathbb{R}^p \text{ observed sequentially, } \beta > 0, a \geq 0, T^{\text{diag}} > 0 \text{ and } T^{\text{off}} > 0.
\]
\[
\text{Set: } \mathcal{B} = \left\{ \pm \frac{\beta}{\sqrt{2^l \log_2(2p)}} : l = 0, \ldots, \lfloor \log_2 p \rfloor \right\}, \quad \mathcal{B}_0 = \left\{ \pm \frac{\beta}{\sqrt{2^l \log_2(2p)}} : n = 0 \right\},
\]
\[
A_b = \Lambda_b = \hat{\Lambda}_b = \mathbf{0} \in \mathbb{R}^{p \times p} \text{ and } t_b = \tau_b = \hat{\tau}_b = 0 \in \mathbb{R}^p \text{ for all } b \in \mathcal{B} \cup \mathcal{B}_0
\]

repeat
\[
\begin{align*}
&n \leftarrow n + 1 \\
&\text{observe new data vector } X_n \\
&\text{for } (j, b) \in [p] \times (\mathcal{B} \cup \mathcal{B}_0) \text{ do}
\end{align*}
\]
\[
\begin{align*}
&t^j_b \leftarrow t^j_{b} + 1 \text{ and } A^{j,j}_b \leftarrow A^{j,j}_b + X_n \\
&\text{set } \delta = 0 \text{ if } t^j_b \text{ is a power of } 2 \text{ and } \delta = 1 \text{ otherwise.}
\end{align*}
\]
\[
\begin{align*}
&\tau^j_b \leftarrow \tau^j_{b} + (1 - \delta) \text{ and } \Lambda^j,j_b \leftarrow \Lambda^j,j_b + X_n \\
&\hat{\tau}_b \leftarrow \sum_{j \neq j'} \left| A^{j,j'}_b \right|^2 \text{ and } \hat{\Lambda}^j,j_b \leftarrow \sum_{j \neq j'} \left| A^{j,j'}_b \right|^2 \\
&\text{if } bA^{j,j}_b - b^2 t^j_{b} / 2 \leq 0 \text{ then}
\end{align*}
\]
\[
\begin{align*}
&t^j_b \leftarrow \tau^j_b \text{ and } \hat{\tau}_b \leftarrow 0 \\
&A^{j,j}_b \leftarrow \Lambda^{j,j}_b \text{ and } \hat{\Lambda}^{j,j}_b \leftarrow 0 \\
&\text{compute } Q^j_b \leftarrow \max_{(j,b) \in [p] \times (\mathcal{B} \cup \mathcal{B}_0)} \left( bA^{j,j}_b - b^2 t^j_{b} / 2 \right)
\end{align*}
\]
\[
\begin{align*}
&S^{\text{diag}} \leftarrow \max_{(j,b) \in [p] \times \mathcal{B}} \left\{ |A^{j,j}_b|^2 \right\} \\
&S^{\text{off}} \leftarrow \max_{(j,b) \in [p] \times \mathcal{B}_0} Q^j_b \\
&\text{until } S^{\text{diag}} \geq T^{\text{diag}} \text{ or } S^{\text{off}} \geq T^{\text{off}};
\end{align*}
\]

Output: \( N = n \)

3 Theoretical analysis

As mentioned in Section 2, the input \( a \) in Algorithms 1 and 2 allows users to detect changepoints of different sparsity levels. More precisely, for any \( \theta \in \mathbb{R}^p \), we have by Lemma 15 that
there exists a smallest \( s(\theta) \in \{2^0, 2^1, \ldots, 2^{\log_2 p}\} \) such that the set

\[
S(\theta) := \left\{ j \in [p] : |\theta^j| \geq \frac{\|\theta\|_2}{\sqrt{s(\theta) \log_2 (2p)}} \right\}
\]

has cardinality at least \( s(\theta) \). On the other hand, we also have \(|S(\theta)| \leq s(\theta) \log_2 (2p)\). We call \( s(\theta) \) the effective sparsity of the vector \( \theta \) and \( S(\theta) \) its effective support. Intuitively, the sum of squares of coordinates in the effective support of \( \theta \) has the same order of magnitude as \( \|\theta\|_2^2 \), up to logarithmic factors. Moreover, if \( \theta \) is an \( s \)-sparse vector in the sense that \( \|\theta\|_0 \leq s \), then \( s(\theta) \leq s \), and the equality is attained when, for example, all non-zero coordinates have the same magnitude.

In this section, we initially analyse the theoretical performance of Algorithm 2 for two different choices of \( a \) in \( S^\text{off} = S^\text{off}(a) \), namely \( a = 0 \) and \( a = \sqrt{8 \log (p - 1)} \). We then present our combined, adaptive procedure and its performance guarantees.

Define \( N^\text{diag} := \inf\{n : S^\text{diag}_n \geq \tau^\text{diag}\} \) and \( N^\text{off} = N^\text{off}(a) := \inf\{n : S^\text{off}_n (a) \geq \tau^\text{off}\} \). Then the stopping time for our changepoint detection procedure is simply \( N = N(a) = N^\text{diag} \land N^\text{off}(a) \).

### 3.1 Dense case

Here, we analyse the changepoint detection procedure \( N = N(0) \), which, as we will see, is most suitable for detecting dense mean changes in the sense that \( s(\theta) \geq \sqrt{p} \) (though we do not assume this in our theory). In this case, when \( p \geq 2 \) and conditionally on \( \tau^\text{off}_0 \), the quantity \( Q^j_b \) follows a chi-squared distribution with \( p - 1 \) degrees of freedom under the null, provided that \( \tau^j_b \) is positive\(^3\). Motivated by Laurent and Massart (2000, Lemma 1), we choose a threshold of the form

\[
T^\text{off} := p - 1 + \tilde{T}^\text{off} + \sqrt{2(p - 1)\tilde{T}^\text{off}} =: \psi(\tilde{T}^\text{off}),
\]

say, for some \( \tilde{T}^\text{off} > 0 \).

The following theorem provides control of the patience of \( \text{ocd}' \).

**Theorem 1.** Let \( X_1, X_2, \ldots \) be generated according to \( \mathbb{P}_0 \). For any \( \gamma \geq 1 \), let \( (X_t)_{t \in \mathbb{N}} \), \( \beta > 0 \), \( a = 0 \), \( T^\text{diag} = \log\{16p\gamma \log_2 (4p)\} \) and \( T^\text{off} = \psi(\tilde{T}^\text{off}) \) with \( \tilde{T}^\text{off} = 2 \log\{16p\gamma \log_2 (2p)\} \) be the inputs of Algorithm 2, with corresponding output \( N \). Then \( \mathbb{E}_0(\gamma) \geq \gamma \).

We note that either of the two statistics \( S^\text{diag} \) and \( S^\text{off} \) may trigger a false alarm under the null. The two threshold levels \( T^\text{diag} \) and \( T^\text{off} \) are chosen so that \( \mathbb{E}_0(\gamma) \) and \( \mathbb{E}_0(\theta) \) have comparable upper bounds.

Our next result controls the response delay of \( \text{ocd}' \) in both worst-case and average senses.

**Theorem 2.** Assume that \( X_1, X_2, \ldots \) are generated according to \( \mathbb{P}_{z, \theta} \) for some \( z \) and \( \theta \) such that \( \|\theta\|_2 = \vartheta \geq \beta > 0 \) and that \( \theta \) has an effective sparsity of \( s := s(\theta) \). Then there exists a universal constant \( C > 0 \), such that the output \( N \) from Algorithm 2, with inputs \( (X_t)_{t \in \mathbb{N}} \), \( \beta \in \mathbb{N} \),

\(^3\)When \( p = 1 \), we have that \( Q^j_b = 0 \) for all \( j \in [p] \) and \( b \in \mathcal{B} \), so \( S^\text{off} = 0 \) and the off-diagonal statistic never triggers the declaration of a change. Similarly, if \( p \geq 2 \) but \( \tau^j_n, b = 0 \), then we also have \( Q^j_n, b = 0 \).
\[ \mathbb{R}^p, \ a = 0, \ T_{\text{diag}} = \log\{16p\gamma\log_2(4p)\} \text{ and } T_{\text{off}} = \psi(\tilde{T}_{\text{off}}) \text{ with } \tilde{T}_{\text{off}} = 2\log\{16p\gamma\log_2(2p)\}, \]
satisfies
\[
\mathbb{E}_\theta^{\text{wc}}(N) \leq C\left\{ \frac{\sqrt{p}\log(ep\gamma)}{\beta^2} \vee \frac{s\log(ep\gamma)\log(ep)}{\beta^2} \vee 1 \right\}. \tag{6}
\]
Furthermore, there exists \( \beta_0(s) > 0 \), depending only on \( s \), such that for all \( \beta \leq \beta_0(s) \), the output \( N \) satisfies
\[
\mathbb{E}_\theta(N) \leq C\left\{ \frac{\sqrt{p}\log(ep\gamma)}{\beta^2} \vee \frac{s\log(ep/\beta)\log(ep)}{\beta^2} \vee 1 \right\}, \tag{7}
\]
for \( s \geq 2 \), and
\[
\mathbb{E}_\theta(N) \leq C\left\{ \frac{\log(ep\gamma)\log(ep)}{\beta^2} \vee 1 \right\}, \tag{8}
\]
for \( s = 1 \).

### 3.2 Sparse case

We now assume that \( p \geq 2 \), and analyse the performance of \( N = N(\sqrt{8\log(p-1)}) \); in other words, we choose \( a = \sqrt{8\log(p-1)} \). This choice turns out to work particularly well when the vector of mean change is sparse in the sense that \( s(\theta) \leq \sqrt{p} \), though again we do not assume this in our theory. The motivation for this choice of \( a \) comes from the fact that, for fixed \( b \) and \( j \), we have \( \Lambda_b^{j,j} \mid \tau_b^j \sim N(0, \tau_b^j) \) for \( j' \in [p] \setminus \{j\} \) under the null. It is therefore natural to choose \( a \) to be of the same order as the maximum of \( p-1 \) independent and identically distributed \( N(0,1) \) random variables. The declaration threshold \( T_{\text{off}} \) is determined based on Lemma 17. Theorem 3 below shows that, in the sparse case, the patience of our procedure is also guaranteed to be at least at the nominal level \( \gamma > 0 \). In addition, as in the dense case, we can also control the response delay of \( \text{oCD' \ } \) according to Theorem 4.

**Theorem 3.** Let \( X_1, X_2, \ldots \) be generated according to \( \mathbb{P}_0 \). For any \( \gamma \geq 1 \), let \((X_t)_{t \in \mathbb{N}}, \ \beta > 0, \ a = \sqrt{8\log(p-1)}, \ T_{\text{diag}} = \log\{16p\gamma\log_2(4p)\} \) and \( T_{\text{off}} = 8\log\{16p\gamma\log_2(2p)\} \) be the inputs of Algorithm 2, with corresponding output \( N \). Then \( \mathbb{E}_\theta(N) \geq \gamma \).

**Theorem 4.** Assume that \( X_1, X_2, \ldots \) are generated according to \( \mathbb{P}_{z,\theta} \) for some \( z \) and \( \theta \) such that \( \|\theta\|_2 = \vartheta \geq \beta > 0 \) and that \( \theta \) has an effective sparsity of \( s := s(\theta) \). Then there exists a universal constant \( C > 0 \), such that the output \( N \) from Algorithm 2, with inputs \((X_t)_{t \in \mathbb{N}}, \ \beta \in \mathbb{R}^p, \ a = \sqrt{8\log(p-1)}, \ T_{\text{diag}} = \log\{16p\gamma\log_2(4p)\} \) and \( T_{\text{off}} = 8\log\{16p\gamma\log_2(2p)\} \), satisfies
\[
\mathbb{E}_\theta(N) \leq \mathbb{E}_\theta^{\text{wc}}(N) \leq C\left\{ \frac{s\log(ep\gamma)\log(ep)}{\beta^2} \vee 1 \right\}. \tag{9}
\]
Comparing Theorems 2 and 4, we see that the thresholding induced by the non-zero choice of \( a = \sqrt{8\log(p-1)} \) in Theorem 4 facilitates an improved dependence on the effective sparsity \( s \) in the bound on the response delay, whenever \( s \) is of smaller order than \( \sqrt{p} \).
3.3 Adaptive procedure

To adapt to different sparsity levels $s$, we can run ocd (or ocd') with two values of $a$ simultaneously: we choose $a = a_{\text{dense}} = 0$ to form the off-diagonal dense statistic $S_{\text{off},d}^\theta = S_{\text{off}}(a_{\text{dense}})$, and $a = a_{\text{sparse}} = \sqrt{8 \log(p - 1)}$ to form the off-diagonal sparse statistic $S_{\text{off},s}^\theta = S_{\text{off}}(a_{\text{sparse}})$. We recall that the diagonal statistic $S_{\text{diag}}$ does not depend on the choice of $\theta$. For clarity, we redefine the three stopping times here: $N_{\text{diag}} := \inf\{n : S_n^\theta \geq T_{\text{diag}}\}$, $N_{\text{off},d} := \inf\{n : S_{n,d}^\theta \geq T_{\text{off},d}\}$ and $N_{\text{off},s} := \inf\{n : S_{n,s}^\theta \geq T_{\text{off},s}\}$, for appropriately-chosen thresholds $T_{\text{diag}}$, $T_{\text{off},d}$ and $T_{\text{off},s}$. The output of this adaptive procedure is thus $N = N_{\text{diag}} \land N_{\text{off},d} \land N_{\text{off},s}$.

The following results provide patience and response delay guarantees for this adaptive procedure.

**Theorem 5.** Let $X_1, X_2, \ldots$ be generated according to $\mathbb{P}_\theta$. For any $\gamma \geq 1$, let $(X_t)_{t \in \mathbb{N}}, \beta > 0$, $T_{\text{diag}} = \log\{24p\gamma \log_2(4p)\}$, $T_{\text{off},d} = \psi(T_{\text{off},d})$ with $T_{\text{off},d} = 2\log\{24p\gamma \log_2(2p)\}$ and $T_{\text{off},s} = 8\log\{24p\gamma \log_2(2p)\}$ be the inputs of the adaptive version of Algorithm 2, with corresponding output $N$. Then $\mathbb{E}_\theta(N) \geq \gamma$.

**Theorem 6.** Assume that $X_1, X_2, \ldots$ are generated according to $\mathbb{P}_{z,\theta}$ for some $z$ and $\theta$ such that $\|\theta\|_2 = \vartheta \geq \beta > 0$ and that $\theta$ has an effective sparsity of $s := s(\theta)$. Then there exists a universal constant $C > 0$, such that the output $N$ from the adaptive version of Algorithm 2, with inputs $(X_t)_{t \in \mathbb{N}}, \beta \in \mathbb{R}_+, T_{\text{diag}} = \log\{24p\gamma \log_2(4p)\}$, $T_{\text{off},d} = \psi(T_{\text{off},d})$ with $T_{\text{off},d} = 2\log\{24p\gamma \log_2(2p)\}$ and $T_{\text{off},s} = 8\log\{24p\gamma \log_2(2p)\}$, satisfies

$$\mathbb{E}_{\theta}^{\text{wc}}(N) \leq C\left\{ \frac{s \log(ep\gamma) \log(ep)}{\beta^2} \lor 1 \right\}.$$ (10)

Furthermore, there exists $\beta_0(s) \in (0, 1/2]$, depending only on $s$, such that for all $\beta \leq \beta_0(s)$, the output $N$ satisfies

$$\mathbb{E}_{\theta}(N) \leq C\left\{ \left( \frac{\sqrt{p} \log(ep\gamma)}{\vartheta^2} \lor \frac{\sqrt{s} \log(ep\beta^{-1}) \log(ep)}{\beta^2} \right) \land \frac{s \log(ep\gamma) \log(ep)}{\beta^2} \right\},$$ (11)

for $s \geq 2$, and

$$\mathbb{E}_{\theta}(N) \leq C\frac{s \log(ep\gamma) \log(ep)}{\beta^2},$$ (12)

for $s = 1$.

Comparing these two results with the corresponding theorems in Sections 3.1 and 3.2, we see that by choosing slightly more conservative thresholds, the adaptive procedure retains the nominal patience control while (up to constant factors) achieving the best of both worlds in terms of its response delay guarantees under different sparsity regimes.

To better understand the worst-case and average-case response delay bounds in Theorem 6, it is helpful to assume that $\vartheta/C_1 \leq \beta \leq \vartheta \leq C_1$ and $\log(\gamma/\beta) \leq C_2 \log p$ for some $C_1, C_2 > 0$. Under these additional assumptions, the result of Theorem 6 takes the simpler form that for some $C > 0$, depending only on $C_1$ and $C_2$, we have

$$\mathbb{E}_{\theta}^{\text{wc}}(N) \leq C\frac{s \log^2(ep)}{\vartheta^2} \quad \text{and} \quad \mathbb{E}_{\theta}(N) \leq C\frac{s \land p^{1/2} \log^2(ep)}{\vartheta^2}.$$
In particular, the average-case response delay upper bound exhibits a phase transition when the effective sparsity level $s$ is of order $\sqrt{p}$, which is the boundary between the sparse and dense cases. Similar sparsity-related elbow effects have been observed in the minimax rate of estimating high-dimensional Gaussian means (Collier, Comminges and Tsybakov, 2017). On the other hand, we note that quadratic dependence on $\vartheta$ in the denominator is known to be optimal in the case when $p = 1$ (Lorden, 1971, Theorem 3). The different dependencies on sparsity of the worst-case and average-case response delays for the dense, sparse and adaptive versions of $\text{ocd}'$ are illustrated in Figure 2.

4 Numerical studies

In this section, we study the empirical performance of the $\text{ocd}$ algorithm and compare it with other online changepoint detection methods. Recall that the (adaptive) $\text{ocd}$ algorithm declares a change when any of the three statistics $S_{\text{diag}}^d$, $S_{\text{off},d}$ and $S_{\text{off},s}$ exceeds their respective thresholds $T_{\text{diag}}$, $T_{\text{off},d}$ and $T_{\text{off},s}$. If a priori knowledge about the signal sparsity is available, it may be slightly preferable to use $N_{\text{diag}} \land N_{\text{off},d}$ in the dense case, and $N_{\text{diag}} \land N_{\text{off},s}$ in the sparse case, but for simplicity of exposition, we will focus on the adaptive version of our $\text{ocd}$ procedure throughout the remainder of this section. While the threshold choices given in Theorem 5 guarantee that the patience of (adaptive) $\text{ocd}$ will be at least at the nominal level, in practice, they may be conservative. We therefore describe a scheme for practical choice of thresholds in Section 4.1. Recall that, in order to form $S_{\text{off},d}$ and $S_{\text{off},s}$, two different entrywise hard thresholds for $A_{t,b}^{i,j}/\sqrt{t_b}$ need to be specified. For $S_{\text{off},d}$, we choose $a = 0$ for both theoretical analysis and practical usage. For $S_{\text{off},s}$, the theoretical choice is $a = \sqrt{8 \log(p - 1)}$, 13
but since this is also slightly conservative, the choice of \( a = \sqrt{2 \log p} \) is used in our practical implementation of the algorithm, and our numerical simulations below.

### 4.1 Practical choice of declaration thresholds

The purpose of this section is to introduce an alternative to using the theoretical thresholds \( T_{\text{diag}}, T_{\text{off,d}} \) and \( T_{\text{off,s}} \) provided by Theorem 5, namely to determine the thresholds through Monte Carlo simulation. The basic idea is that since the null distribution is known, we can simulate from it to determine the patience for any given choice of thresholds. A complicating issue is the fact that the choices of the three thresholds \( T_{\text{diag}}, T_{\text{off,d}} \) and \( T_{\text{off,s}} \) are related, so that we may be able to achieve the same patience by increasing \( T_{\text{diag}} \) and decreasing \( T_{\text{off,d}} \), for example. To handle this, we first argue that the renewal nature of the processes involved means that, at least for moderately large thresholds, the times to exceedence for each of the three statistics \( S_{\text{diag}}, S_{\text{off,d}} \) and \( S_{\text{off,s}} \) are approximately exponentially distributed. Evidence to support this is provided by Figure 3, where we present QQ-plots of \( N_{\text{diag}}/m(N_{\text{diag}}) \), \( N_{\text{off,d}}/m(N_{\text{off,d}}) \) and \( N_{\text{off,s}}/m(N_{\text{off,s}}) \), where the \( m(N) \) statistics are empirical medians of the corresponding \( N \) statistics (divided by \( \log 2 \)) over 200 repetitions.

We can therefore set an individual Monte Carlo threshold for \( S_{\text{diag}} \) as follows (the other two statistics can be handled in identical fashion): for \( r \in [B] \), simulate \( X_1^{(r)}, \ldots, X_\gamma^{(r)} \) iid \( N_p(0, I_p) \) and for each \( n \in [\gamma] \), compute the diagonal statistic \( S_{n,\text{diag}}^{(r)} \) on the \( r \)th sample. Now compute \( V^{(r)} := \max_{1 \leq n \leq \gamma} S_{n,\text{diag}}^{(r)} \), and take \( \tilde{T}_{\text{diag}} \) to be the \((1/e)\)th quantile of \( \{ V^{(r)} : r \in [B] \} \). The rationale for the final step here is that if \( P_0(V^{(1)} < \tilde{T}_{\text{diag}}) = 1/e \), then \( P_0(\tilde{N}_{\text{diag}} > \gamma) = 1/e \), where \( \tilde{N}_{\text{diag}} := \min\{ n : S_{n,\text{diag}} \geq \tilde{T}_{\text{diag}} \} \). Thus, under an exponential distribution for \( \tilde{N}_{\text{diag}} \), we have that \( \tilde{N}_{\text{diag}} \) has individual patience \( \gamma \).

Having determined appropriate thresholds \( \tilde{T}_{\text{diag}}, \tilde{T}_{\text{off,d}} \) and \( \tilde{T}_{\text{off,s}} \), we can then use similar ideas to set a suitable combined threshold \( T_{\text{comb}} \). In particular, we also argue that \( N_{\text{diag}} \wedge N_{\text{off,d}} \wedge N_{\text{off,s}} \) has an approximate exponential distribution; see Figure 3 for supporting evidence. We therefore proceed as follows: for \( r \in [B] \), simulate \( \tilde{X}_1^{(r)}, \ldots, \tilde{X}_\gamma^{(r)} \) iid \( N_p(0, I_p) \) and use this new data to compute \( S_{n,\text{diag}}^{(r)} := S_{n,\text{diag}}^{(r)}/\tilde{T}_{\text{diag}}, S_{n,\text{off,d}}^{(r)} := S_{n,\text{off,d}}^{(r)}/\tilde{T}_{\text{off,d}} \) and \( S_{n,\text{off,s}}^{(r)} := S_{n,\text{off,s}}^{(r)}/\tilde{T}_{\text{off,s}} \) for each \( n \in [\gamma] \), and set \( W^{(r)} := \max\{ S_{n,\text{diag}}^{(r)} \vee S_{n,\text{off,d}}^{(r)} \vee S_{n,\text{off,s}}^{(r)} : n \in [\gamma] \} \) on the \( r \)th sample. Now take \( T_{\text{comb}} \) to be the \((1/e)\)th quantile of \( \{ W^{(r)} : r \in [B] \} \). Similar to before, our reasoning here is that if \( P_0(W^{(1)} < T_{\text{comb}}) = 1/e \), then \( N_{\text{diag}} := \min\{ n : S_{n,\text{diag}} \geq \tilde{T}_{\text{diag}} T_{\text{comb}} \} \), \( N_{\text{off,d}} := \min\{ n : S_{n,\text{off,d}} \geq \tilde{T}_{\text{off,d}} T_{\text{comb}} \} \) and \( N_{\text{off,s}} := \min\{ n : S_{n,\text{off,s}} \geq \tilde{T}_{\text{off,s}} T_{\text{comb}} \} \) satisfy

\[
P_0\left(N_{\text{diag}} \wedge N_{\text{off,d}} \wedge N_{\text{off,s}} > \gamma\right) = 1/e.
\]

Thus, under an exponential distribution for \( N_{\text{diag}} \wedge N_{\text{off,d}} \wedge N_{\text{off,s}} \), it again has the desired nominal patience. Our practical thresholds, therefore, are \( T_{\text{diag}} = \tilde{T}_{\text{diag}} T_{\text{comb}}, T_{\text{off,d}} = \tilde{T}_{\text{off,d}} T_{\text{comb}} \) and \( T_{\text{off,s}} = \tilde{T}_{\text{off,s}} T_{\text{comb}} \) for \( S_{\text{diag}}, S_{\text{off,d}} \) and \( S_{\text{off,s}} \) respectively. Table 1 confirms that, with these choices of Monte Carlo thresholds, the patience of the adaptive OCD algorithm remains at approximately the desired nominal level.
Figure 3: QQ-plots of standardised versions of $N^{\text{diag}}$, $N^{\text{off},d}$ and $N^{\text{off},s}$, as well as $N = N^{\text{diag}} \wedge N^{\text{off},d} \wedge N^{\text{off},s}$, against theoretical Exp(1) quantiles.

Table 1: Estimated run lengths under the null using the Monte Carlo thresholds described in Section 4.1 over 500 repetitions, with desired patience level $\gamma = 5000$. Algorithm is terminated after 20000 data points for each repetition. Each reported value is the average run length taken over the repetitions which have already declared prior to time 20000. For reference, $\mathbb{E}(X \mid X < 20000) \approx 4626.9$ when $X \sim \text{Exp}(1/5000)$. 

<table>
<thead>
<tr>
<th></th>
<th>$p = 100$</th>
<th>$p = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta = 2$</td>
<td>4606.2</td>
<td>4480.8</td>
</tr>
<tr>
<td>$\beta = 1/2$</td>
<td>5291.5</td>
<td>4383.6</td>
</tr>
</tbody>
</table>
A further observation is that the three individual response delays, as well as the combined response delay, are all approximately proportional to $\vartheta^{-2}$, a phenomenon which is supported by Theorem 6.

Table 3 presents corresponding results when $\beta$ is both over- and under-specified. We note
Table 3: Estimated response delays over 200 repetitions for $N^\text{diag}$, $N^\text{off,d}$ and $N^\text{off,s}$ and the response delay of the combined declaration time $N$ for ocd. Settings where $\beta$ is both over- and under-specified are given. The quickest response in each setting is given in bold. Other parameters: $p = 100$, $\gamma = 5000$, $z = 0$ and $\theta = \vartheta$, where the distribution of $U$ is described in Section 4.2.

That both $N^\text{off,d}$ and $N^\text{off,s}$ are almost unaffected by either type of misspecification. For $N^\text{diag}$, a mild over-misspecification of $\beta$ helps it to react faster, while an under-misspecification causes it to have increased response delay. However, since we can also observe that $N^\text{diag}$ rarely declares first by a large margin, the performance of ocd is highly robust to misspecification of $\beta$, especially when $s$ is not too small.

4.3 Comparison with other methods

We now compare our adaptive ocd algorithm with other online changepoint detection algorithms proposed in the literature, namely those of Mei (2010), Xie and Siegmund (2013) and Chan (2017). Since we were unable to find publicly-available implementations of any of these algorithms, we briefly describe below their methodology and the small adaptations that we made in order to allow them to be used in our settings.

Mei (2010) assumes knowledge of $\theta$, and, on observing each new data point, aggregates likelihood ratio tests in each coordinate of the null $N(0, 1)$ against an alternative of $N(\theta^j, 1)$ in the $j$th coordinate. More precisely, in the notation of (2), the algorithm declares a change when either $\sum_{j \in [p]} R^j_{n,\theta^j}$ or $\max_{j \in [p]} R^j_{n,\theta^j}$ exceeds given thresholds. In our setting where we do not know $\theta$ and only assume that $\|\theta\|_2 \geq \beta$, we replace $\sum_{j \in [p]} R^j_{n,\theta^j}$ and $\max_{j \in [p]} R^j_{n,\theta^j}$ with

$$\max \left\{ \sum_{j=1}^{p} R^j_{n,\beta/\sqrt{p}}, \sum_{j=1}^{p} R^j_{n,-\beta/\sqrt{p}} \right\}$$

and

$$\max \left\{ \max_{j \in [p]} R^j_{n,\beta/\sqrt{p}}, \max_{j \in [p]} R^j_{n,-\beta/\sqrt{p}} \right\}$$

respectively.

The algorithms of Xie and Siegmund (2013) and Chan (2017) have a similar flavour. The idea is to test the null $N_p(0, I_p)$ distribution against an alternative where the $j$th coordinate
has a \((1 - p_0)N(0, 1) + p_0N(\mu^j, 1)\) mixture distribution, for some known \(p_0 \in [0, 1]\) and unknown \(\mu^j \in \mathbb{R}\). After specifying a window size \(w\), both algorithms search for the strongest evidence against the null from the past \(r \in [w]\) observations. Specifically, writing \(Z^j_{n,r} := r^{-1/2} \sum_{i=n-r+1}^n X^j_i\) for \(n \in \mathbb{N}\), \(r \in [n]\) and \(j \in [p]\), the test statistics are of the form

\[
S^+_{XSC}(p_0, \lambda, \kappa, w) := \max_{r \in [w \wedge n]} \sum_{j=1}^p \log \left( 1 - p_0 + \lambda p_0 e^{(Z^j_{n,r} \lor 0)^2/\kappa} \right),
\]

where Xie and Siegmund (2013) take \((\lambda, \kappa, w) = (1, 2, 200)\) and Chan (2017) takes \((\lambda, \kappa, w) = (2\sqrt{2} - 2, 4, 200)\). Since such a test statistic is only effective when \(\sum_{j \in [p]} (\mu^j \lor 0)^2\) is large, we considered statistics of the form \(S^+_{XSC}(p_0, \lambda, \kappa, w) \lor S^-_{XSC}(p_0, \lambda, \kappa, w)\), where \(S^-_{XSC}(p_0, \lambda, \kappa, w)\) replaces the exponent \(Z^j_{n,r} \lor 0\) with \(Z^j_{n,r} \land 0\). An adaptive choice of \(p_0\) is not provided by the authors, but the choices \(p_0 \in \{1/\sqrt{p}, 0.1\}\) have been considered; we found the choice \(p_0 = 1/\sqrt{p}\) to be the most competitive overall, so for simplicity of exposition, present only that choice in our results.

For each of the Mei (2010), Xie and Siegmund (2013) and Chan (2017) algorithms, we determined appropriate thresholds using Monte Carlo simulation, as suggested by the authors, and in a similar fashion to the way in which we set the ocd thresholds as described in Section 4.1. This guarantees that the algorithms have approximately the nominal patience, and so allows us to compare the methods by means of the response delay.

Table 4 displays the response delays for the ocd algorithm, as well as the alternative methods described above, for \(p \in \{100, 2000\}\), \(s \in \{5, \lfloor \sqrt{p} \rfloor, p\}\) and \(\vartheta \in \{1, 0.5, 0.25\}\). In fact, we ran simulations for \(p = 1000\), \(s \in \{1, p/2\}\) and \(\vartheta \in \{1, 0.125\}\), but the results are qualitatively similar and are therefore omitted. Overall, the results reveal that ocd performs very well in comparison with existing methods, across a wide range of scenarios; in several cases it is by far the most responsive procedure, and in none of the settings considered is it outperformed by much. The Xie and Siegmund (2013) and Chan (2017) algorithms perform similarly to each other, and in most settings are both more competitive than the Mei (2010) method described above. We note that the performance of the Xie and Siegmund (2013) and Chan (2017) algorithms is relatively better when the signal-to-noise ratio is high; in these scenarios, the default window size \(w = 200\) is large enough that sufficient evidence against the null can typically be accumulated within the moving window. For lower signal-to-noise ratios, this ceases to be the case, and from time \(z + w\) onwards, the test statistic has the same marginal distribution (with no positive drift). This explains the relative deterioration in performance for those algorithms in the harder settings considered. As mentioned in the introduction, if the change in mean were known to be small, then the window size could be increased to compensate, but at additional computational expense; a further advantage of ocd, then, is that the computational time only depends on \(p\) (and not on \(\beta\) or other problem parameters).
Table 4: Estimated response delay for ocd, as well as the algorithms of Mei (2010) (Mei), Xie and Siegmund (2013) (XS) and Chan (2017) (Chan) over 200 repetitions, with $z = 0$, $\gamma = 5000$ and $\theta$ generated as described in Section 4.2. The smallest response delay is given in bold.
5 Proofs of main results

5.1 Proofs from Section 3.1

Proof of Theorem 1. Define $m := [2\gamma]$. It suffices to prove (a) $\mathbb{P}_0(N_{\text{off}} \leq m) \leq 1/4$ and (b) $\mathbb{P}_0(N_{\text{diag}} \leq m) \leq 1/4$, since then we have

$$\mathbb{E}_0(N) = \mathbb{E}_0(N_{\text{off}} \land N_{\text{diag}}) \geq 2\gamma \mathbb{P}_0(N_{\text{off}} \land N_{\text{diag}} > 2\gamma) \geq 2\gamma \{1 - \mathbb{P}_0(N_{\text{off}} \leq m) - \mathbb{P}_0(N_{\text{diag}} \leq m)\} \geq \gamma.$$

We prove the two claims below.

(a) By (5) and a union bound, we have

$$\mathbb{P}_0(N_{\text{off}} \leq m) \leq \sum_{n \in [m], j \in [p]} \sum_{b \in \mathcal{B}} \mathbb{P}_0\left(Q_{n,b}^j \geq T_{\text{off}} \mid \tau_{n,b}^j \right) = \sum_{n \in [m], j \in [p]} \mathbb{E}_0\left[ \mathbb{P}_0\left(Q_{n,b}^j \geq T_{\text{off}} \mid \tau_{n,b}^j \right) \right]. \quad (13)$$

Recall that under the null, $\Lambda_b^{k,j} \mid \tau_b^j \overset{iid}{\sim} N(0, \tau_b^j)$ for all $b \in \mathcal{B}, j \in [p]$ and $k \in [p] \setminus \{j\}$, which implies that $Q_b^j \mid \tau_b^j \sim \chi_{p-1}^2 \mathbb{1}_{\{\tau_b^j > 0\}}$. Thus, we have by Laurent and Massart (2000, Lemma 1) that for all $n \in [m], j \in [p]$ and $b \in \mathcal{B},$

$$\mathbb{P}_0\left(Q_{n,b}^j \geq T_{\text{off}} \mid \tau_{n,b}^j \right) \leq e^{-\tau_{n,b}^j / 2}. \quad (14)$$

Combining (13) and (14), we have

$$\mathbb{P}_0(N_{\text{off}} \leq m) \leq |\mathcal{B}|mpe^{-\tau_{\text{off}} / 2} \leq 1/4. \quad (15)$$

(b) For $j \in [p]$ and $b \in \mathcal{B} \cup \mathcal{B}_0$, denote $N_b^j := \inf\{n : R_{n,b}^j \geq T_{\text{diag}}\}$, where $R_{n,b}^j$ is defined by (2). By Lemma 10, we have that $R_{n,b}^j = \{R_{n-1,b}^j + b(X_n^j - b/2)\} \lor 0$, and that this process is always non-negative. Then $N_{\text{diag}} = \min\{N_b^j : j \in [p], b \in \mathcal{B} \cup \mathcal{B}_0\}$.

Now, fix some $j \in [p]$ and $b \in \mathcal{B} \cup \mathcal{B}_0$. Define $U_0 := 0$ and $U_h := \inf\{n > U_{h-1} : R_{n,b}^j \notin (0, T_{\text{diag}})\}$ for $h \in \mathbb{N}$, and let $H := \inf\{h : R_{U_h,b}^j \geq T_{\text{diag}}\}$. Then

$$N_b^j = U_H \geq H.$$

To study the distribution of $H$, consider the one-sided sequential probability ratio test of $H_{0,2} : Z_1, Z_2, \ldots \overset{iid}{\sim} N(0, 1)$ against $H_{1,2} : Z_1, Z_2, \ldots \overset{iid}{\sim} N(b, 1)$ with log-boundaries $T_{\text{diag}}$ and $-\infty$. The associated stopping time for this test is

$$N_{\text{os}} := \inf\left\{ n \in \mathbb{N} : b \sum_{t=1}^n (Z_t - b/2) \geq T_{\text{diag}} \right\}.$$

Since $(R_{U_h,b}^j)_n$ is a Markov process that renews itself every time it hits $0$, $H$ follows a geometric distribution with success probability

$$\mathbb{P}_0(R_{U_1,b}^j \geq T_{\text{diag}}) \leq \mathbb{P}_{H_{0,2}}(N_{\text{os}} < \infty) \leq e^{-T_{\text{diag}}},$$

20
where the last inequality follows from Lemma 12. Consequently,
\[ \mathbb{P}_0(N_b^j \leq m) \leq \mathbb{P}_0(H \leq m) \leq 1 - \left(1 - e^{-T_{\text{diag}}}\right)^m. \]
As the above inequality holds for all \( j \in [p] \) and \( b \in \mathcal{B} \cup \mathcal{B}_0 \), we have that
\[ \mathbb{P}_0(N^\text{diag} > m) = \mathbb{P}_0\left( \bigcap_{j \in [p], b \in \mathcal{B} \cup \mathcal{B}_0} \{N_b^j > m\}\right) = \prod_{j \in [p]} \left\{1 - \mathbb{P}_0\left( \bigcup_{b \in \mathcal{B} \cup \mathcal{B}_0} \{N_b^j \leq m\}\right)\right\} \geq \left[1 - |\mathcal{B} \cup \mathcal{B}_0|\left(1 - e^{-T_{\text{diag}}}\right)^m\right]^p \geq 1 - mp|\mathcal{B} \cup \mathcal{B}_0|e^{-T_{\text{diag}}} \geq 3/4, \quad (16) \]
as desired, where in the penultimate inequality, we twice used the fact that \((1 - x)^{\alpha} \geq 1 - \alpha x\) for all \( \alpha \geq 1 \) and \( x \in [0, 1] \).

The proof of Theorem 2 is quite involved. We first define some relevant quantities, and then state and prove some preliminary results. For \( \theta \in \mathbb{R}^p \) with effective sparsity \( s(\theta) \), there can be at most one coordinate in \( \theta \) of magnitude larger than \( \vartheta/\sqrt{2} \), so there exists \( b_* \in \{\beta/\sqrt{s(\theta) \log_2(2p)}, -\beta/\sqrt{s(\theta) \log_2(2p)}\} \subseteq \mathcal{B} \) such that
\[ J := \left\{ j \in [p] : \theta^j/b_* \geq 1 \text{ and } |\theta^j| \leq \vartheta/\sqrt{2} \right\} \quad (17) \]
has cardinality at least \( s(\theta)/2 \) (note that the condition \( \theta^j/b_* \geq 1 \) above ensures that \( \{\theta^j : j \in J\} \) all have the same sign as \( b_* \)). Both \( b_* \) and \( J \) can be chosen as functions of \( \theta \). Now, given any sequence \( X_1, X_2, \ldots \in \mathbb{R}^p \) and \( \theta \in \mathbb{R}^p \), define for any \( \alpha \in (0, 1] \) the function
\[ q(\alpha) = q(\alpha; X_1, \ldots, X_2, \theta) := \inf\{y \in \mathbb{R} : |\{j \in J : t_{j,b_*}^j \leq y\}| \geq \alpha|J|\}, \quad (18) \]
where \( t_{j,b_*}^j \) is obtained by running Algorithm 2 up to time \( z \) with \( a = 0 \) and \( T_{\text{diag}} = T_{\text{off}} = \infty \). In other words, \( q(\alpha) \) is the empirical \( \alpha \)-quantile of the tail lengths \( \{t_{j,b_*}^j : j \in J\} \) when we run the algorithm without declaring any change up to time \( z \). Recall the definition of the function \( \psi \) in (5).

**Proposition 7.** Assume that \( X_1, X_2, \ldots \) are generated according to \( \mathbb{P}_{z, \theta} \) for some \( z \) and \( \theta \) such that \( \|\theta\|_2 = \vartheta \geq \beta > 0 \) and that \( \theta \) has an effective sparsity of \( s := s(\theta) \geq 2 \). Then the output \( N \) from Algorithm 2, with input \( (X_t)_{t \in \mathbb{N}}, \beta \in \mathbb{R}^p, a = 0, T_{\text{diag}} \geq 1 \) and \( T_{\text{off}} = \psi(T_{\text{off}}) \) for \( T_{\text{off}} \geq \log(ep) \), satisfies
\[ \mathbb{E}_{z, \theta}\{(N - z) \wedge 0 \mid X_1, \ldots, X_z\} \leq \frac{396T_{\text{off}} + 65\sqrt{pT_{\text{off}}}}{\vartheta^2} + \frac{24\log_2(2p)}{\alpha \beta^2} + 3q(\alpha) + 2, \quad (19) \]
for any \( \alpha \in (0, 1] \).

**Proof.** Since the bound in (19) is positive, we may, throughout the proof and for arbitrary \( z \in \mathbb{N} \), restrict attention to realisations \( X_1 = x_1, \ldots, X_z = x_z \) for which we have not declared a change by time \( z \). In other words, we have \( N > z \). This restriction also ensures that \( q(\alpha) \)
defined in (18) by setting the thresholds to infinity is now indeed the empirical $\alpha$-quantile of the tail lengths $(t_{z, b_*}^j : j \in J)$ at the changepoint. Denote $J_\alpha := \{ j \in J : t_{z, b_*}^j \leq q(\alpha) \}$. Then we have $|J_\alpha| \geq \alpha |J| \geq \alpha s/2$.

We now fix some

$$r \geq \left\{ \frac{12(T^{\text{off}} + \sqrt{2(p - 1)T^{\text{off}}})}{\gamma^2} \vee 3q(\alpha) \right\} + 2 := r_0. \tag{20}$$

Note that $r_0 > 3q(\alpha) \geq 3t_{z, b_*}^j$ for all $j \in J_\alpha$. For $j \in J_\alpha$, we define the event

$$\Omega_j^i := \{ t_{z+\lfloor r \rfloor, b_*}^j > 2\lfloor r \rfloor / 3 \}.$$ 

Under $\mathbb{P}_{z, \theta}$, conditional on $X_1 = x_1, \ldots, X_z = x_z$, we know that $X_{z+1}, X_{z+2}, \ldots \overset{\text{ind}}{\sim} N_p(\theta, I_p)$.

Hence, by using Lemma 11 and applying Lemma 13(b) to $t_{z+\lfloor r \rfloor, b_*}^j \wedge \lfloor r \rfloor$ for $j \in J_\alpha$, we obtain

$$\mathbb{P}_{z, \theta} \left( \bigcap_{j \in J_\alpha} (\Omega_j^i)^c \middle| X_1 = x_1, \ldots, X_z = x_z \right) \leq \exp\{-|J_\alpha|b_*^2[\lfloor r \rfloor] / 12\} \leq \exp\{-\alpha s b_*^2[\lfloor r \rfloor] / 24\}. \tag{21}$$

We now work on the event $\Omega_j^i$, for some $j \in J_\alpha$. We note that (20) guarantees that $r \geq 2$, and thus $t_{z+\lfloor r \rfloor, b_*}^j \geq 2\lfloor r \rfloor / 3 \geq 2$. Then, by Lemma 16 and the fact that $r_0 > 3t_{z, b_*}^j$, we have that

$$\frac{\lfloor r \rfloor}{3} < \frac{t_{z+\lfloor r \rfloor, b_*}^j}{2} \leq \tau_{z+\lfloor r \rfloor, b_*}^j \leq \frac{3t_{z+\lfloor r \rfloor, b_*}^j}{4} \leq \frac{3(t_{z, b_*}^j + r)}{4} < r.$$ 

Hence we conclude that on the event $\Omega_j^i$,

$$2/3 \leq \lfloor r \rfloor / 3 < \tau_{z+\lfloor r \rfloor, b_*}^j \leq \lfloor r \rfloor. \tag{22}$$

Recall that $\Lambda_{z+\lfloor r \rfloor, b_*}^j \in \mathbb{R}^p$ records the tail CUSUM statistics with tail length $t_{z+\lfloor r \rfloor, b_*}^j$. We observe by (22) that on $\Omega_j^i$, only post-change observations are included in $\Lambda_{z+\lfloor r \rfloor, b_*}^j$. Hence we have that on the event $\Omega_j^i$,

$$\Lambda_{z+\lfloor r \rfloor, b_*}^{k, j} \mid \{ \tau_{z+\lfloor r \rfloor, b_*}^j, X_1 = x_1, \ldots, X_z = x_z \} \overset{\text{ind}}{\sim} N(\theta^{k, j}_{z+\lfloor r \rfloor, b_*}, \tau_{z+\lfloor r \rfloor, b_*}^j) \tag{23}$$

for $k \in [p] \setminus \{ j \}$. Therefore, on the event $\Omega_j^i$ and conditional on $\tau_{z+\lfloor r \rfloor, b_*}^j, X_1 = x_1, \ldots, X_z = x_z$, the random variable $\frac{\|\Lambda_{z+\lfloor r \rfloor, b_*}^{-j, j}\|^2}{\tau_{z+\lfloor r \rfloor, b_*}^j} \overset{\text{iid}}{\sim} \chi^2_{p - 1}$ follows a non-central chi-squared distribution with $p - 1$ degrees of freedom and noncentrality parameter $\|\theta^{-j}\|^2_{2\tau_{z+\lfloor r \rfloor, b_*}^j}$. Since $j \in J$ and $s \geq 2$, we observe, by (17) and (22) that $\|\theta^{-j}\|_{2\tau_{z+\lfloor r \rfloor, b_*}^j} \geq \sqrt{2} [\lfloor r \rfloor] / 6$ on $\Omega_j^i$. Write

$$E_{\tau}^j := \left\{ \frac{\|\Lambda_{z+\lfloor r \rfloor, b_*}^{-j, j}\|^2}{\tau_{z+\lfloor r \rfloor, b_*}^j} \vee 1 < T^{\text{off}} \right\}.$$
Then by Birgé (2001, Lemma 8.1), we have
\[
\mathbb{P}_{z, \theta}\left(E_r^j \cap \Omega_r^j \mid \tau_{z+|r|, b_*}^j, X_1 = x_1, \ldots, X_z = x_z\right)
\leq \exp\left\{-\frac{\vartheta^2[r] / 6 - \bar{T}^{\text{off}} - \sqrt{2(p - 1) \bar{T}^{\text{off}}}}{4(p - 1 + \vartheta^2[r] / 3)}\right\}.
\] (24)

Combining (21) and (24), we deduce that
\[
\mathbb{P}_{z, \theta}\left(N > z + r \mid X_1 = x_1, \ldots, X_z = x_z\right) \leq \mathbb{P}_{z, \theta}\left(N > z + |r| \mid X_1 = x_1, \ldots, X_z = x_z\right)
\leq \mathbb{P}_{z, \theta}\left(\bigcap_{j \in J_\alpha} (\Omega_r^j)^c \mid X_1 = x_1, \ldots, X_z = x_z\right) + \sum_{j \in J_\alpha} \mathbb{P}_{z, \theta}\left(E_r^j \cap \Omega_r^j \mid X_1 = x_1, \ldots, X_z = x_z\right)
\leq \exp\left\{-\frac{\alpha s b_*^2(r - 1)}{24}\right\} + p \exp\left\{-\frac{\vartheta^2(r - 1) / 6 - \bar{T}^{\text{off}} - \sqrt{2(p - 1) \bar{T}^{\text{off}}}}{4(p - 1 + \vartheta^2(r - 1) / 3)}\right\}
\leq \exp\left\{-\frac{\alpha s b_*^2(r - 1)}{24}\right\} + p \exp\left\{-\frac{\vartheta^4(r - 1)^2}{576(p - 1 + \vartheta^2(r - 1) / 3)}\right\},
\]
where the last inequality uses (20). Therefore, we have
\[
\mathbb{E}_{z, \theta}\{(N - z) \vee 0 \mid X_1 = x_1, \ldots, X_z = x_z\} = \int_0^\infty \mathbb{P}_{z, \theta}(N > z + u \mid X_1 = x_1, \ldots, X_z = x_z) \, du
\leq r_0 + \int_{r_0 - 1}^\infty \exp\left\{-\frac{\alpha s b_*^2 u}{24}\right\} + p \exp\left\{-\frac{\vartheta^4 u^2}{576(p - 1 + \vartheta^2 u / 3)}\right\} \wedge 1 \, du
\leq r_0 + \frac{24}{\alpha s b_*^2} + \int_0^\infty \left(\exp^{-\vartheta^2 u / 384}\right) \wedge 1 \, du + \int_0^\infty \left(\exp^{-\vartheta^4 u^2 / 1152}\right) \wedge 1 \, du
\leq r_0 + \frac{24}{\alpha s b_*^2} + \frac{384 \log(\exp)}{\vartheta^2} + \frac{24 \sqrt{2(p - 1) \log p}}{\vartheta^2} + \frac{12 \sqrt{2\pi(p - 1)}}{\vartheta^2}
\leq r_0 + \frac{24}{\alpha s b_*^2} + \frac{384 \log(\exp)}{\vartheta^2} + 48 \frac{(p - 1) \log(\exp)}{\vartheta^2},
\]
where the penultimate inequality follows from the fact that \(1 - \Phi(x) \leq \frac{1}{2} e^{-x^2 / 2}\) for \(x \geq 0\). The desired bound (19) follows by substituting in the expressions for \(r_0\) and \(b_*\).

The following two propositions control the residual tail length quantile term \(q(\alpha)\) in (19) in the worst-case and average-case scenarios respectively.

**Proposition 8.** Let \(X_1, X_2, \ldots, z, \theta, s, a, p\) and \(N\) be defined as in Proposition 7. On the event \(\{N > z\}\), we have
\[
q(1; X_1, \ldots, X_z, \theta) \leq \frac{8 T^{\text{diag}} s \log_2(2p)}{\beta^2}.
\]
Proof. We will show the stronger result that on the event \( \{ N > z \} \), we have

\[
t^j_{z,b} < \frac{8T^{\text{diag}}}{b^2}
\]

for all \( b \in \mathcal{B} \) and \( j \in [p] \). The desired result then follows immediately by taking \( b = b_\ast \) and restricting to the subset \( \mathcal{J} \subseteq [p] \).

Fix \( b \in \mathcal{B} \) and \( j \in [p] \). Recall from (2) and Lemma 10 the definition of \( R^j_{n,b} \) and the recursive relation \( R^j_{n,b} = \{ R^j_{n-1,b} + b(X^j_n - b/2) \} \lor 0 \). By the update procedure for \( t^j_{n,b} \) in Algorithm 2 and Lemma 11, we have

\[
R^j_{n,b} \begin{cases} 
= 0 & \text{when } n = z - t^j_{z,b}, \\
> 0 & \text{when } z - t^j_{z,b} < n \leq z.
\end{cases}
\tag{25}
\]

We claim that

\[
R^j_{n,b/2} \geq \frac{R^j_{n,b}}{2} + \frac{b^2(n - z + t^j_{z,b})}{8},
\tag{26}
\]

for all \( n \in \{ z - t^j_{z,b}, \ldots, z \} \). To see this, the claim is true when \( n = z - t^j_{z,b} \) since the right hand side of (26) is 0 by (25). Now, assume (26) is true for some \( n = m - 1 \). Then,

\[
R^j_{m,b/2} \geq R^j_{m-1,b/2} + \frac{b}{2}(X^j_m - \frac{b}{4}) \geq \frac{R^j_{m-1,b}}{2} + \frac{b^2(m - 1 - z + t^j_{z,b})}{8} + \frac{b}{2}(X^j_m - \frac{b}{4})

= \frac{R^j_{m,b}}{2} + \frac{b^2(m - z + t^j_{z,b})}{8}.
\]

This proves the claim by induction. In particular, on the event \( \{ N > z \} \), we have \( T^{\text{diag}} > R^j_{z,b/2} > b^2t^j_{z,b}/8 \) as desired. \( \square \)

**Proposition 9.** Let \( X_1, X_2, \ldots, z, \theta, s, a, p \) and \( N \) be defined as in Proposition 7. There exists \( \beta_0(s) > 0 \), depending only on \( s \), such that for all \( \beta < \beta_0(s) \), we have

\[
\mathbb{E}_{z,\theta}\{ q(s^{-1/2}; X_1, \ldots, X_z, \theta) \} \leq \frac{200s^{1/2} \log(16s^2\beta^{-2}\log_2(2p))\log_2(2p)}{\beta^2}.
\]

**Proof.** Recall the definition of \( b_\ast \) in (17). We may assume, without loss of generality that \( b_\ast = \beta/\sqrt{s \log_2(2p)} \) (the case \( b_\ast = -\beta/\sqrt{s \log_2(2p)} \) can be proved in essentially the same way). We first prove the result for \( s > 256 \). Recall that \( t^j_{z,b_\ast} = \arg\max_{0 \leq r \leq z} \sum_{i=z-r+1}^z (X^j_i - b_\ast/2) \).

Define \( Z_i := X_{z-i+1} \) for \( i \in [z] \) and let \( Z_{z+1}, Z_{z+2}, \ldots \text{iid } N_p(0, I_p) \) be independent from \( Z_1, \ldots, Z_z \). For each \( j \in [p] \), let

\[
S^j_r := \sum_{i=1}^r (Z^j_i - b_\ast/2) \quad \text{and} \quad \tilde{S}^j_r := \sum_{i=1}^r Z^j_i
\]

for \( r \in \mathbb{N} \) and define \( S^j_0 := \tilde{S}^j_0 := 0 \). Writing \( \xi^j_0 := \arg\max_{0 \leq r \leq \Delta b_\ast^{-2}} S^j_r \), \( \xi^j := \arg\max_{r \in \mathbb{N}_0} S^j_r \), and \( \tilde{\xi}^j := \arg\max_{0 \leq r \leq \Delta b_\ast^{-2}} \tilde{S}^j_r \), where \( \Delta := 8\log(16sb_\ast^{-2}) \), we note that like \( t^j_{z,b_\ast} \), these three
maxima are also uniquely attained almost surely (see the proof of Lemma 13). By construction, we have for each \( j \in [p] \)

\[
t^j_{z,b_*} = \argmax_{0 \leq r \leq z} \sum_{i=z-r+1}^{z} (X_i^j - b_/2) = \argmax_{0 \leq r \leq z} S_r^j = \argmax_{r \in \mathbb{N}_0} S_r^j = \xi^j.
\]

Writing \( q_\xi(\alpha) := \inf\{y : |\{j \in \mathcal{J} : \xi^j \leq y\}| \geq \alpha|\mathcal{J}|\} \) as the empirical \( \alpha \)-quantile of \((\xi^j : j \in \mathcal{J})\), it follows that \( q_\xi(\alpha) \leq q_\xi(\alpha) \) and so it suffices to control \( \mathbb{E}\{q_\xi(s^{-1/2})\} \) instead of \( \mathbb{E}\{q(s^{-1/2})\} \). To this end, we observe that \( \{16\Delta s^{-1/2}b_* < \xi^j \leq \Delta b_*^{-2}\} \subseteq \{16\Delta s^{-1/2}b_*^{-2} < \xi^j_0 \leq \Delta b_*^{-2}\} \) and \( \xi^j_0 \geq \xi^j_1 \), and thus

\[
\mathbb{P}(\xi^j \leq 16\Delta s^{-1/2}b_*^{-2}) \geq \mathbb{P}(\xi^j_0 \leq 16\Delta s^{-1/2}b_*^{-2}) - \mathbb{P}(\xi^j > \Delta b_*^{-2}) \\
\geq \mathbb{P}(\xi^j_0 \leq 16\Delta s^{-1/2}b_*^{-2}) - \mathbb{P}(\xi^j > \Delta b_*^{-2}).
\]

For the first term on the right hand side of \((27)\), by Donsker’s invariance principle and the continuity of the argmax map (see, e.g. van der Vaart and Wellner, 1996, Lemma 3.2.1 and Theorem 3.2.2), we have in the limit \( \beta \to 0 \) that \( \Delta b_*^{-2} \to \infty \) and so

\[
\frac{\xi^j_0}{\Delta b_*^{-2}} \xrightarrow{d} \argmax_{t \in [0,1]} B_t,
\]

where \((B_t)_{t \geq 0}\) denotes a standard Brownian motion. In particular, we can find \( \beta_0(s) > 0 \) depending only on \( s \) such that for \( \beta \leq \beta_0(s) \) and \( s > 256 \), we have

\[
\mathbb{P}(\xi^j_0 \leq 16\Delta s^{-1/2}b_*^{-2}) \geq \frac{1}{2} \mathbb{P}\left( \argmax_{t \in [0,1]} B_t \leq 16s^{-1/2} \right) = \frac{1}{\pi} \arcsin(4s^{-1/4}) \geq \frac{4s^{-1/4}}{\pi}.
\]

where in the second step we used the arcsine law for Brownian motion (see, e.g. Mörters and Peres, 2010, Theorem 5.26), and in the final step we used the fact that \( 4s^{-1/4} < 1 \).

For the second term on the right-hand side of \((27)\), we have by a union bound that

\[
\mathbb{P}(\xi^j > \Delta b_*^{-2}) \leq \sum_{r=\lceil \Delta b_*^{-2} \rceil}^{\infty} \mathbb{P}(S_r^j \geq 0) \leq \sum_{r=\lceil \Delta b_*^{-2} \rceil}^{\infty} e^{-b_*^2 r / 8} = \frac{e^{-\Delta/8}}{1 - e^{-b_*^2 / 8}}.
\]

By reducing \( \beta_0(s) > 0 \) if necessary, we can have for all \( \beta \leq \beta_0(s) \) that

\[
e^{-b_*^2 / 8} = e^{-\beta^2/(8s \log_2(2p))} \leq 1 - \beta^2/(16s \log_2(2p)) = 1 - b_*^2 / 16.
\]

Since \( \Delta = 8 \log(16sb_*^{-2}) \), we have

\[
\mathbb{P}(\xi^j > \Delta b_*^{-2}) \leq \frac{1}{s}.
\]

Substituting \((28)\) and \((29)\) into \((27)\), we have, for all \( j \in \mathcal{J} \), that

\[
\mathbb{P}(\xi^j \leq 16\Delta s^{-1/2}b_*^{-2}) \geq s^{-1/4}.
\]
As a result, \(|\{j \in \mathcal{J} : \xi^j \leq 16\Delta s^{-1/2}b_s^{-2}\}|\) is stochastically larger than \(\text{Bin}(|\mathcal{J}|, s^{-1/4})\). Thus, for \(s > 256\), we have,
\[
\mathbb{P}_{z, \theta}\{q_\xi(s^{-1/2}) > 16\Delta s^{-1/2}b_s^{-2}\} \leq \mathbb{P}\left\{\text{Bin}(|\mathcal{J}|, s^{-1/4}) \leq s^{-1/2}|\mathcal{J}|\right\} \leq e^{-s^{1/2}/2},
\]
where we have used Hoeffding’s inequality and the fact that \(|\mathcal{J}| \geq s/2\) in the last step. On the other hand, we have,
\[
\mathbb{E}_{z, \theta}\left\{q_\xi(s^{-1/2}) \bigg| q_\xi(s^{-1/2}) > 16\Delta s^{-1/2}b_s^{-2}\right\} \leq \mathbb{E}_{z, \theta}\left\{q_\xi(s^{-1/2}) \bigg| q_\xi(s^{-1/2}) \geq \Delta b_s^{-2}\right\}
\leq \mathbb{E}_{z, \theta}\left\{q_\theta(1) \bigg| q_\theta(|\mathcal{J}|^{-1}) \geq \Delta b_s^{-2}\right\} = \mathbb{E}_{z, \theta}\left\{\max_{j \in \mathcal{J}} \xi^j \bigg| \min_{j \in \mathcal{J}} \xi^j \geq \Delta b_s^{-2}\right\} \leq 32\Delta b_s^{-2},
\]
where we have used Lemma 14(b) in the second inequality and Lemma 13(d) (with \(\Delta b_s^{-2}\) taking the role of \(c\) there) in the final inequality. As a result,
\[
\mathbb{E}_{z, \theta}\{q(s^{-1/2})\} \leq \mathbb{E}_{z, \theta}\{q_\xi(s^{-1/2})\} \leq 16\Delta s^{-1/2}b_s^{-2} + 32e^{-s^{1/2}/2}\Delta b_s^{-2}
\leq \frac{200s^{1/2} \log(16s^2\beta^{-2} \log_2(2p)) \log_2(2p)}{\beta^2},
\]
where we have used in the final step the fact that \(e^{-s^{1/2}/2} \leq s^{-1/2}/100\) for \(s > 256\). This proves the desired result for \(s > 256\).

Finally, for \(s \leq 256\), we have by Lemma 13(c) that, for \(\beta < \sqrt{s}/2\),
\[
\mathbb{E}_{z, \theta}\{q(s^{-1/2})\} \leq \mathbb{E}_{z, \theta}\left\{\max_{j \in \mathcal{J}} \xi^j\right\} \leq \frac{8s \log(s^{3/2}\beta^{-1} \log_2^{1/2}(2p)) \log_2(2p)}{\beta^2}
\leq \frac{128s^{1/2} \log(16s^2\beta^{-2} \log_2(2p)) \log_2(2p)}{\beta^2},
\]
and the desired bound then follows.

We are now in a position to prove Theorem 2.

**Proof of Theorem 2.** The proof proceeds with different arguments for the case \(s \geq 2\) and the case \(s = 1\).

**Case 1: \(s \geq 2\).** Combining Propositions 7 (applied with \(\alpha = 1\)) and 8, we have
\[
\bar{\mathbb{E}}_{w}^\alpha(N) \leq \frac{396\tilde{T}^{\text{off}} + 65\sqrt{p\tilde{T}^{\text{off}}}}{\vartheta^2} + \frac{24 \log_2(2p)}{\beta^2} + \frac{24T^{\text{diag}} s \log_2(2p)}{\beta^2} + 2.
\]
The desired bound (6) then follows by substituting in the expression for \(\tilde{T}^{\text{off}}\). On the other hand, combining Propositions 7 (applied with \(\alpha = s^{-1/2}\)) and 9, we have
\[
\bar{\mathbb{E}}_{w}(N) \leq \frac{396\tilde{T}^{\text{off}} + 65\sqrt{p\tilde{T}^{\text{off}}}}{\vartheta^2} + \frac{24\sqrt{s} \log_2(2p)}{\beta^2} + \frac{600s^{1/2} \log(16s^2\beta^{-2} \log_2(2p)) \log_2(2p) + 2}{\beta^2},
\]

26
which proves (7).

Case 2: \( s = 1 \). There exists \( j_* \in [p] \) such that \( |\theta^{j_*}| \geq \sqrt[4]{\log_2(2p)} \), and recall from (17) that \( b_* := \text{sgn}(\theta^{j_*}) \beta / \sqrt{\log_2(2p)} \in \mathcal{B} \). Note that \( S^\text{diag}_{n,1} = \max_{(j,b) \in [p] \times \mathcal{B} \cup \mathcal{B}_0} R_{n,b}^j \geq R_{n,b_*}^{j_*} \). We define \( \tilde{R}_n := \sum_{i=z+1}^{z+n} b_*(X_i^j - b_*/2) \) for \( n \in \mathbb{N}_0 \). Since \( R_{n,b_*}^{j_*} \geq 0 = \tilde{R}_0 \) and \( R_n - R_{n-1} = b_*(X_{z+n}^j - b_*/2) \leq R_{n-1,b_*}^{j_*} - R_{z+n,b_*}^{j_*} \), it follows by induction that \( R_{z+n,b_*}^{j_*} \geq \tilde{R}_n \) for all \( n \in \mathbb{N}_0 \). Then, for \( n \geq \lceil 4T^\text{diag}/(b_*\theta^{j_*}) \rceil = n_0 \), we have

\[
\mathbb{P}_{z,\theta}(N > z + n \mid X_1 = x_1, \ldots, X_z = x_z) \leq \mathbb{P}_{z,\theta}(\tilde{R}_{z+n,b_*}^{j_*} \leq T^\text{diag} \mid X_1 = x_1, \ldots, X_z = x_z) \leq \mathbb{P}_{z,\theta}(\tilde{R}_n \leq T^\text{diag}) = \Phi\left(-\frac{b_*(\theta^{j_*} - b_*/2) - T^\text{diag}}{n^{1/2}b_*}\right) \leq \frac{1}{2} \exp\left\{ -\frac{2nb_*^2}{2n(b_*\theta^{j_*} - b_*/2)^2} \right\} \leq \frac{1}{2} e^{-n(\theta^{j_*})^2/32}.
\]

Therefore,

\[
\mathbb{E}_{z,\theta}\{(N - z) \vee 0 \mid X_1 = x_1, \ldots, X_z = x_z\} = \sum_{n=0}^{\infty} \mathbb{P}_{z,\theta}(N > z + n \mid X_1 = x_1, \ldots, X_z = x_z) \leq n_0 + \frac{1}{2} \sum_{n=n_0}^{\infty} e^{-n(\theta^{j_*})^2/32} \leq n_0 + \frac{1}{2} \int_0^{\infty} e^{-u(\theta^{j_*})^2/32} du \leq 1 + \frac{4T^\text{diag}}{b_*\theta^{j_*}} + \frac{16}{(\theta^{j_*})^2}.
\]

After substituting in the expressions for \( b_* \), \( \theta^{j_*} \) and \( T^\text{diag} \), we see that

\[
\mathbb{P}_\theta(N) \leq \mathbb{E}_{\theta}^\text{wc}(N) \leq 1 + \frac{4 \log(16p\gamma \log_2(4p)) \log_2(2p)}{\beta^2} + \frac{16 \log_2(2p)}{\beta^2},
\]

which proves both (6) and (8).

\[
\Box
\]

5.2 Proofs from Sections 3.2 and 3.3

**Proof of Theorem 3.** It suffices to only prove \( \mathbb{P}_0(N^\text{off} \leq m) \leq 1/4 \), since the remaining proof is identical to that of Theorem 1.

Since \( \Lambda_{b,k}^{j,j} \mid \tau_{b,j}^{i,j} \overset{\text{iid}}{\sim} N(0, \tau_{b}^{j,j}) \) for all \( b \in \mathcal{B}, j \in [p] \) and \( k \in [p] \setminus \{j\} \) under the null, by the fact that \( T^\text{off} \geq 12 \) and Lemma 17, we have

\[
\mathbb{P}_0(Q_{n,b}^j \geq T_{n,b} \mid \tau_{n,b}^j) \leq \mathbb{P}_0(Q_{n,b}^j \geq 6 + T_{n,b}^\text{off}/2 \mid \tau_{n,b}^j) \leq \exp(-T_{n,b}^\text{off}/8).
\]

Hence, it follows that

\[
\mathbb{P}_0(N^\text{off} \leq m) \leq |\mathcal{B}| m p e^{-T_{n,b}^\text{off}/8} \leq 1/4,
\]

as desired.

\[
\Box
\]

**Proof of Theorem 4.** We note that the case \( s = 1 \) in the proof of Theorem 2 does not rely on the off-diagonal statistics. Hence (30) is still valid here with \( a = \sqrt{8 \log(p - 1)} \) and the last expression in (30) again proves the desired bound (9). For the case \( s \geq 2 \), we follow exactly
the proof of Proposition 7 until (23), with the only exception that we now fix, instead of (20),

\[
  r \geq \left\{ \frac{24T^\text{off} \log_2(2p)}{\theta^2} \vee \frac{96s \log_2(2p) \log p}{\theta^2} \vee 3q(\alpha) \right\} + 2 := \tilde{r}_0. \tag{32}
\]

By the definition of the effective sparsity of \( \theta \), for a fixed \( j \in J_\alpha \),

\[
  \mathcal{L}^j := \left\{ j' \in [p] : |\theta^{j'}| \geq \frac{\theta}{s \log_2(2p)} \text{ and } j' \neq j \right\}
\]

has cardinality at least \( s - 1 \). On the event \( \Omega_j^j \), we have, by (22), that for all \( k \in \mathcal{L}^j \)

\[
  |\theta^k| \sqrt{\frac{r^j z + |r|, b_s}{z}} \geq \sqrt{\frac{\vartheta^2 |r|}{3s \log_2(2p)}} =: \tilde{a}_r.
\]

We then observe, by (32), that

\[
  \tilde{a}_r \geq \sqrt{32 \log p} > 2a. \tag{33}
\]

Now, from (23) we have on the event \( \Omega_j^j \) that, for all \( k \in \mathcal{L}^j \),

\[
  \mathbb{P}_{z, \theta} \left( \Omega_j^j \cap \left\{ |\Lambda^{k,j}_{z + |r|, b_s}| < \frac{1}{2} \tilde{a}_r \sqrt{\frac{r^j z + |r|, b_s}} \right\} \mid \tau^j z + |r|, b_s, X_1 = x_1, \ldots, X_z = x_z \right) \leq \frac{1}{2} e^{-\tilde{a}_r^2/8} =: q_r.
\]

We denote

\[
  U^j := \left\{ k \in \mathcal{L}^j : \left| |\Lambda^{k,j}_{z + |r|, b_s}| < \frac{1}{2} \tilde{a}_r \sqrt{\frac{r^j z + |r|, b_s}} \right\} \right\}.
\]

Then, by the Chernoff–Hoeffding binomial tail bound (Hoeffding, 1963, Equation (2.1)), we have

\[
  \mathbb{P}_{z, \theta} \left( \Omega_j^j \cap \left\{ U^j \geq |\mathcal{L}^j|/2 \right\} \mid \tau^j z + |r|, b_s, X_1 = x_1, \ldots, X_z = x_z \right) \leq \exp \left\{ -\frac{|\mathcal{L}^j|}{2} \log \left( \frac{1}{4q_r(1 - q_r)} \right) \right\} \leq \exp \left\{ -\frac{3|\mathcal{L}^j|\tilde{a}_r^2}{64} \right\} \leq \exp \left\{ -\frac{\vartheta^2 |r|}{128 \log_2(2p)} \right\}, \tag{34}
\]

where the penultimate inequality follows from (33). Now, on the event \( \Omega_j^j \cap \{ U^j < |\mathcal{L}^j|/2 \} \), we have

\[
  \sum_{j' \in [p] ; j' \neq j} \frac{(\Lambda^{j',j}_{z + |r|, b_s})^2}{\tau^j z + |r|, b_s} \vee 1 \left\{ |\Lambda^{j',j}_{z + |r|, b_s}| \geq \bar{a} \sqrt{\frac{r^j z + |r|, b_s}} \right\} \geq \sum_{j' \in [p] ; j' \neq j} \frac{(\Lambda^{j',j}_{z + |r|, b_s})^2}{\tau^j z + |r|, b_s} \vee 1 \left\{ |\Lambda^{j',j}_{z + |r|, b_s}| \geq \bar{a} \sqrt{\frac{r^j z + |r|, b_s}} \right\} \geq \frac{\tilde{a}_r^2}{4} \left\{ |\mathcal{L}^j| - \left( \frac{|\mathcal{L}^j|}{2} - 1 \right) \right\} \geq \frac{\tilde{a}_r^2}{4} \frac{|\mathcal{L}^j|}{2} \geq \frac{\vartheta^2 |r|}{24 \log_2(2p)} \geq T^\text{off}, \tag{35}
\]

28
where the penultimate inequality uses the fact that $|\mathcal{L}| \geq s - 1$ and the last inequality follows from (32). We now denote

$$\tilde{E}_r^j := \left\{ \sum_{j' \in [p] \setminus j} \left( \frac{A_{z+r}^{j'}}{\gamma_{z+r}} \right)^2 \right\}^{1/2} \{ |A_{z+r}^{j'}| \geq \alpha \sqrt{\gamma_{z+r}} \} < T^{\text{off}} \right\}. $$

Combining (21), (34) and (35), we deduce that

$$\mathbb{P}_{z,\theta}(N > z + r \mid X_1 = x_1, \ldots, X_z = x_z) \leq \mathbb{P}_{z,\theta}(N > z + |r| \mid X_1 = x_1, \ldots, X_z = x_z)
\leq \mathbb{P}_{z,\theta}\left( \bigcap_{j \in J_a} (\Omega_{r}^j)^c \mid X_1 = x_1, \ldots, X_z = x_z \right) + \sum_{j \in J_a} \mathbb{P}_{z,\theta}\left( \tilde{E}_r^j \cap \Omega_{r}^j \mid X_1 = x_1, \ldots, X_z = x_z \right)
\leq \exp\left\{- \frac{\alpha s b^2_r (r - 1)}{24} \right\} + p \exp\left\{- \frac{\vartheta^2 (r - 1)}{128 \log_2 (2p)} \right\}.

Therefore we have

$$\mathbb{E}_{z,\theta}\{(N - z) \vee 0 \mid X_1 = x_1, \ldots, X_z = x_z\} = \int_0^{\infty} \mathbb{P}_{z,\theta}(N > z + u \mid X_1 = x_1, \ldots, X_z = x_z) \, du
\leq \tilde{r}_0 + \int_{\tilde{r}_0}^{\infty} \left[ \exp\left\{- \frac{\alpha s b^2_r u}{24} \right\} + p \exp\left\{- \frac{\vartheta^2 u}{128 \log_2 (2p)} \right\} \right] \wedge 1 \, du
\leq \tilde{r}_0 + \frac{24}{\alpha s b^2_r} + \int_{\tilde{r}_0}^{\infty} \left( p e^{-\frac{\vartheta^2 u}{128 \log_2 (2p)}} \right) \wedge 1 \, du \leq \tilde{r}_0 + \frac{24}{\alpha s b^2_r} + \frac{128 \log_2 (2p) \log(ep)}{\vartheta^2}
\leq 24 T^{\text{off}} \log_2 (2p) + 96 s \log_2 (2p) \log p + 3q(\alpha) + \frac{24 \log_2 (2p)}{\alpha \beta^2} + \frac{128 \log_2 (2p) \log(ep)}{\vartheta^2} + 2.

Combining this with Proposition 8 (applied with $\alpha = 1$), we have, by substituting in the expression for $T^{\text{off}}$, that

$$\mathbb{E}_{\theta}(N) \leq \mathbb{E}^{\text{wc}}_{\theta}(N) \leq C \left\{ \frac{s \log(ep \gamma) \log(ep)}{\beta^2} \right\} \vee 1 \right\},
\text{for some universal constant } C > 0, \text{ as desired.}
\qed

\textbf{Proof of Theorem 5.} Let $T^{\text{off},d} = \psi(T^{\text{off},d})$. Then, similar to (15), (16) and (31), we have

$$\mathbb{P}_0(N^{\text{diag}} \leq m) \leq mp|\mathcal{B} \cup \mathcal{B}_0|e^{-T^{\text{diag}}} \leq 1/6,
\mathbb{P}_0(N^{\text{off},d} \leq m) \leq mp|\mathcal{B}|e^{-T^{\text{off},d}/2} \leq 1/6,
\mathbb{P}_0(N^{\text{off},s} \leq m) \leq mp|\mathcal{B}|e^{-T^{\text{off},s}/8} \leq 1/6.
and hence,
\[ \mathbb{E}_0(N) = \mathbb{E}_0(N_{\text{diag}} \land N_{\text{off,d}} \land N_{\text{off,s}}) \geq 2\gamma \mathbb{P}_0(N_{\text{diag}} \land N_{\text{off,d}} \land N_{\text{off,s}} > 2\gamma) \]
\[ \geq 2\gamma \{1 - \mathbb{P}_0(N_{\text{diag}} \leq m) - \mathbb{P}_0(N_{\text{off,d}} \leq m) - \mathbb{P}_0(N_{\text{off,s}} \leq m)\} \geq \gamma, \]
as desired. \hfill \Box

**Proof of Theorem 6.** We observe that
\[ \bar{\mathbb{E}}_\theta (N) = \mathbb{E}_\theta [\mathbb{E}[\mathbb{P} \mid N_{\text{diag}} \land N_{\text{off,d}} \land N_{\text{off,s}}]] \]
\[ \leq \mathbb{E}_\theta \mathbb{E}[N_{\text{diag}} \land N_{\text{off,d}}] \land \mathbb{E}_\theta \mathbb{E}[N_{\text{off,s}}], \]
and similarly for \( \mathbb{E}_\theta (N) \). The desired bounds (10), (11) and (12) are therefore direct consequences of Theorems 2 and 4 (note that the different constants in the thresholds only affect the value of the universal constant). \hfill \Box

## 6 Auxiliary results

**Lemma 10.** For \( n \in \mathbb{N}_0, b \in \mathcal{B} \cup \mathcal{B}_0 \) and \( j \in [p] \), we define \( R^j_{n,b} := bA^i_{n,b} - b^2 t^j_{n,b} / 2 \), where \( A_{n,b} \) and \( t_{n,b} \) are taken from Algorithm 2. Then
\[ R^j_{n,b} = \max_{0 \leq h \leq n} \sum_{i=n-h+1}^{n} b(X_i^j - b/2). \] (36)

**Proof.** We prove the claim by induction on \( n \). The base case \( n = 0 \) is true since, by definition, \( R^j_{0,b} = 0 \) and the sum on the right-hand side of (36) is empty. Assume (36) is true for \( n = m - 1 \). Then, by the update procedure in Algorithm 2, we have
\[ R^j_{m,b} = \{R^j_{m-1,b} + b(X_m^j - b/2)\} \lor 0 = \max_{0 \leq h \leq m-1} \sum_{i=m-h}^{m-1} b(X_i^j - b/2) + b(X_m^j - b/2) \lor 0 \]
\[ = \max_{0 \leq h \leq m-1} \sum_{i=m-h}^{m} b(X_i^j - b/2), \]
and the desired result follows. \hfill \Box

**Lemma 11.** For \( n \in \mathbb{N}_0, b \in \mathcal{B} \cup \mathcal{B}_0 \) and \( j \in [p] \), let \( t^j_{n,b} \) be defined as in Algorithm 2 and \( R^j_{n,b} \) as in Lemma 10. Then
\[ t^j_{n,b} = \min\{0 \leq i \leq n : R^j_{n-i,b} = 0\} = \text{sargmax}_{0 \leq h \leq n} \sum_{i=n-h+1}^{n} b(X_i^j - b/2). \] (37)

**Proof.** We observe from the procedure in Algorithm 2 that \( R^j_{n,b} = 0 \) if and only if \( t^j_{n,b} = 0 \) and that \( R^j_{n,b} > 0 \) if and only if \( t^j_{n,b} = t^j_{n-1,b} + 1 \). Hence,
\[ t^j_{n,b} = n - \max\{0 \leq i \leq n : R^j_{i,b} = 0\} = \min\{0 \leq i \leq n : R^j_{n-i,b} = 0\}. \]
We now prove that $t_{n,b}^j = \text{sargmax}_{0 \leq h \leq n} \sum_{i=m-h+1}^{n} b(X_i^j - b/2)$ by induction on $n$. The base case $n = 0$ is true because $t_{n,b}^j = 0$, and the sum on the right-hand side of $(37)$ is empty. Assume the claim is true for $n = m - 1$. Then, by the inductive hypothesis and Lemma 10,

$$t_{m,b}^j = (t_{m-1,b}^j + 1)1_{\{R_{m,b}^j > 0\}} = \left(\text{sargmax}_{0 \leq h \leq m-1} \sum_{i=m-h}^{m-1} b(X_i^j - b/2) + 1\right)1_{\{R_{m,b}^j > 0\}}$$

$$= \left(\text{sargmax}_{1 \leq h \leq m} \sum_{i=m-h+1}^{m} b(X_i^j - b/2)\right)1_{\{\text{max}_{0 \leq h \leq m} \sum_{i=m-h+1}^{m} b(X_i^j - b/2) > 0\}}$$

$$= \text{sargmax}_{0 \leq h \leq m} \sum_{i=m-h+1}^{m} b(X_i^j - b/2),$$

and the desired result follows.

For two distributions $P_0$ and $P_1$ on the same measurable space, the sequential probability ratio test of $H_0 : X_1, X_2, \ldots \sim P_0$ against $H_1 : X_1, X_2, \ldots \sim P_1$ with log-boundaries $a > 0$ and $b < 0$ is defined as the (extended) stopping time

$$N := \inf\left\{ n : \sum_{i=1}^{n} \log \frac{dP_1}{dP_0}(X_i) \notin (b, a) \right\},$$

together with the decision rule after stopping that accepts $H_0$ if $\sum_{i=1}^{N} \log \{dP_1/dP_0\}(X) \leq b$ and accepts $H_1$ if $\sum_{i=1}^{N} \log \{dP_1/dP_0\}(X) \geq a$.

**Lemma 12.** Suppose $N$ is the stopping time associated with the (one-sided) sequential probability ratio test of $H_0 : X_1, X_2, \ldots \sim P_0$ against $H_1 : X_1, X_2, \ldots \sim P_1$ with log-boundaries $a > 0$ and $b = -\infty$. Then

$$\mathbb{P}_0(\ N < \infty \ ) \leq e^{-a}.$$

**Proof.** Let $L_n := \prod_{i=1}^{n} (dP_1/dP_0)(X_i)$. On the event $\{N < \infty\}$, we have $L_N \geq e^a$. Therefore,

$$\mathbb{P}_0(N < \infty) = \sum_{n=1}^{\infty} \mathbb{P}_0(N = n) \leq e^{-a} \sum_{n=1}^{\infty} \mathbb{E}_0(L_n 1_{\{N=n\}}) = e^{-a} \sum_{n=0}^{\infty} \mathbb{P}_1(N = n) \leq e^{-a},$$

which proves the desired result.

We collect in the following lemma some useful bounds on both the maximum and the argmax of a Gaussian random walk with a negative drift.

**Lemma 13.** Suppose that $X_1, X_2, \ldots \sim N(-b, 1)$ for some $b > 0$. Let $S_0 := 0$, and, for $r \in \mathbb{N}$, define $S_r := \sum_{i=1}^{r} X_i$, as well as $M := \sup_{r \in \mathbb{N}_0} S_r$ and $\xi := \text{argmax}_{r \in \mathbb{N}_0} S_r$. Then $\xi$ is almost surely unique. Moreover, letting $\xi^1, \ldots, \xi^s$ denote independent copies of $\xi$, we have the following results:
(a) For any $a \geq 0$, we have
\[
\frac{3\sqrt{ab/2}}{\sqrt{2\pi(9ab/2 + 1)}} e^{-9ab/4} \mathbb{1}_{\{a \geq b\}} + \frac{2b}{\sqrt{2\pi(4b^2 + 1)}} e^{-2b^2} \mathbb{1}_{\{a < b\}} \leq \mathbb{P}(M \geq a) \leq e^{-2ab}.
\]

(b) If $c \in \mathbb{N}$ satisfies $bc \geq a \geq 0$, then
\[
\mathbb{P}\left(\max_{r \geq c} S_r \geq a\right) \leq \exp\left\{-\frac{(bc + a)^2}{2c}\right\}.
\]

(c) If $b \leq 1/2$, then
\[
\mathbb{E}\left(\max_{j \in [s]} \xi^j \mid \min_{j \in [s]} \xi^j \geq c\right) \leq 32c.
\]

Proof. To prove that $\xi$ is almost surely unique, it suffices to note that
\[
\mathbb{P}(\xi \text{ not unique}) \leq \mathbb{P}\left(\bigcup_{r_1 < r_2} \{S_{r_1} = S_{r_2}\}\right) \leq \sum_{r_1 < r_2} \mathbb{P}\left(\sum_{i=r_1+1}^{r_2} X_i = 0\right) = 0,
\]
since $\sum_{i=r_1+1}^{r_2} X_i \sim N(-b(r_2 - r_1), r_2 - r_1)$.

(a) Let $(W_t)_{t \geq 0}$ be a standard Brownian motion, so that $(S_r + rb)_{r \in \mathbb{N}} \overset{d}{=} (W_r)_{r \in \mathbb{N}}$. Thus
\[
\mathbb{P}(M \geq a) \leq \mathbb{P}\left(\sup_{t \geq 0} (W_t - bt) \geq a\right) = e^{-2ab},
\]
where the calculation for the final equality can be found in, e.g. Siegmund (1986, Proposition 2.4 and Equation (2.5)). For the lower bound, we note that
\[
\mathbb{P}(M \geq a) \geq \sup_{r \in \mathbb{N}} \mathbb{P}(S_r \geq a) = \sup_{r \in \mathbb{N}} \Phi\left(-\frac{a + br}{\sqrt{r}}\right) \geq \Phi\left(-\frac{a + br_0}{\sqrt{r_0}}\right),
\]
where $r_0 = \lceil a/b \rceil \lor 1$. If $a \geq b$, then using the fact that the function $x \mapsto (a + bx)/\sqrt{x}$ is increasing on $[\sqrt{a/b}, \infty)$, we have
\[
\frac{a + br_0}{\sqrt{r_0}} \leq \frac{a + b(a/b + 1)}{\sqrt{a/b + 1}} = 2\sqrt{b} \cdot \frac{a + b/2}{\sqrt{a+b}} \leq 2\sqrt{b} \left(a + \frac{b^2/4}{a+b}\right)^{1/2} \leq 3\sqrt{ab/2}.
\]

On the other hand, if $a < b$, then
\[
\frac{a + br_0}{\sqrt{r_0}} = a + b < 2b.
\]

32
The desired result follows from the bound \( \Phi(-x) \geq \frac{x}{\sqrt{2\pi}(x^2+1)}e^{-x^2/2} \) for all \( x > 0 \).

(b) By part (a), we have

\[
\mathbb{P}\left(\max_{r \geq c} S_r \geq a\right) = \int_{-\infty}^{\infty} \mathbb{P}\left(\max_{r \geq c} S_r \geq a \mid S_c = x\right) \frac{1}{\sqrt{2\pi c}}e^{-\frac{(x+bc)^2}{2c}} \, dx
\]

\[
\leq \int_{-\infty}^{a} e^{-2(a-x)b} \frac{1}{\sqrt{2\pi c}}e^{-\frac{(x+bc)^2}{2c}} \, dx + \int_{a}^{\infty} \frac{1}{\sqrt{2\pi c}}e^{-\frac{(x+bc)^2}{2c}} \, dx
\]

\[
= e^{-2ab} \Phi\left(-\frac{bc-a}{\sqrt{c}}\right) + \Phi\left(-\frac{bc+a}{\sqrt{c}}\right) \leq \exp\left\{ -\frac{(bc+a)^2}{2c}\right\},
\]

where in the final step we have used the fact that \( bc \geq a \) and \( \Phi(-x) \leq e^{-x^2/2} \) for \( x \geq 0 \).

(c) For any \( x \in \mathbb{N} \), we have by two union bounds that for \( b \in (0, 1/2] \),

\[
\mathbb{P}\left(\max_{j \in [s]} \xi_j^i \geq x\right) \leq \sum_{r=x}^{\infty} \mathbb{P}(\xi = r) \leq \sum_{r=x}^{\infty} \mathbb{P}(S_r \geq 0)
\]

\[
= \sum_{r=x}^{\infty} \Phi(-b\sqrt{r}) \leq \frac{e^{-rb^2/2}}{2(1-e^{-b^2/2})},
\]

Now define \( x_0 := \lceil 4b^{-2} \log(s/b) \rceil \). Then for \( b \in (0, 1/2] \),

\[
\mathbb{E}\left(\max_{j \in [s]} \xi_j^i \right) = \sum_{x=1}^{\infty} \mathbb{P}\left(\max_{j \in [s]} \xi_j^i \geq x\right) \leq x_0 - 1 + \sum_{x=x_0}^{\infty} \frac{se^{-rb^2/2}}{2(1-e^{-b^2/2})}
\]

\[
\leq \frac{4 \log(s/b)}{b^2} + \frac{se^{-x_0b^2/2}}{2(1-e^{-b^2/2})} \leq \frac{4 \log(s/b)}{b^2} + \frac{2}{b^2(1-1/16)^2}
\]

\[
\leq \frac{8 \log(s/b)}{b^2},
\]

where we have used the fact that \( 1 - e^{-x} \geq 15x/16 \) for \( x \in [0, 1/8] \).

(d) Define \( M_c := \max_{1 \leq r \leq c} S_r \) and let \( P_{S_c,M_c} \) be the joint distribution of \((S_c, M_c)\). Then

\[
\mathbb{E}\left(\max_{j \in [s]} \xi_j^i \mid \min_{j \in [s]} \xi_j^i \geq c\right) - 31c \leq \sum_{x=31c}^{\infty} \mathbb{P}\left(\max_{j \in [s]} \xi_j^i \geq x \mid \min_{j \in [s]} \xi_j^i \geq c\right)
\]

\[
\leq s \sum_{x=31c}^{\infty} \mathbb{P}(\xi \geq x \mid \xi \geq c)
\]

\[
= s \sum_{x=31c}^{\infty} \int_{\mathbb{R}^2} \mathbb{P}(\xi \geq x \mid \xi \geq c, S_c = y, M_c = a) \, dP_{S_c,M_c}(y, a)
\]

\[
\leq s \sum_{x=31c}^{\infty} \int_{\mathbb{R}^2} \mathbb{P}\left(\max_{r \geq x} S_r \geq a \mid \max_{r \geq c} S_r \geq a, S_c = y, M_c = a\right) \, dP_{S_c,M_c}(y, a)
\]

\[
= s \sum_{x=31c}^{\infty} \int_{\mathbb{R}^2} \mathbb{P}\left(\max_{r \geq x-c} S_r \geq a-y \mid M \geq a-y\right) \, dP_{S_c,M_c}(y, a).
\]
If $b(x-c)/4 \geq a - y \geq b$, then by parts (a) and (b) we have that

$$
\mathbb{P}(\max_{r \geq x-c} S_r \geq a - y \mid M \geq a - y)
\leq \exp\left\{ -\frac{(b(x-c) + (a-y))^2}{2(x-c)} \right\} \cdot \frac{9(a-y)b/2 + 1}{3\sqrt{(a-y)b/(4\pi)}} e^{9(a-y)b/4}
\leq e^{-b^2(x-c)/2+5b(a-y)/4}\left(3\sqrt{\pi(a-y)b} + \frac{2\sqrt{\pi}/3}{\sqrt{(a-y)b}(a-y)b}\right)
\leq e^{-\sqrt{2\pi}(a-y)b/2} \left(3\sqrt{(x-c)b^2} + \frac{2\sqrt{\pi}}{3b}\right).
$$

Since the function $h \mapsto he^{-h^2/2}$ is decreasing for $h \geq 1$, we have that $3\sqrt{\pi(x-c)b^2}/4 + 2\sqrt{\pi}/(3b) \leq 3e^{b(x-c)/16}/2$ for $x-c \geq 30c \geq 60b^{-2}\log(1/b)$. Thus,

$$
\mathbb{P}(\max_{r \geq x-c} S_r \geq a - y \mid M \geq a - y) \leq \frac{3}{2}e^{-b^2(x-c)/8}. \tag{39}
$$

On the other hand, if $b > a - y$ (note that this implies $b(x-c) \geq a - y$), then by parts (a) and (b) we have that

$$
\mathbb{P}(\max_{r \geq x-c} S_r \geq a - y \mid M \geq a - y) \leq \exp\left\{ -\frac{(b(x-c) + (a-y))^2}{2(x-c)} \right\} \cdot \frac{\sqrt{2\pi}(1 + 4b^2)}{2b} e^{2b^2}
\leq e^{-b^2(x-c)/2+2b^2}\left(\frac{\sqrt{2\pi}}{2b} + 2\sqrt{2\pi b}\right)
\leq \frac{\sqrt{2\pi}}{b} e^{-b^2(x-c)/4} \leq \frac{\sqrt{2\pi}}{2^{13/2}} e^{-b^2(x-c)/8},
$$

where we have used the fact that $x-c \geq 60b^{-2}\log(1/b)$ in the final bound. Combining the above display with (39), we see that for $b(x-c)/4 \geq a - y$, we always have

$$
\mathbb{P}(\max_{r \geq x-c} S_r \geq a - y \mid M \geq a - y) \leq \frac{3}{2}e^{-b^2(x-c)/8}.
$$

Thus, by part (a),

$$
\int_{\mathbb{R}^2} \mathbb{P}(\max_{r \geq x} S_r \geq a \mid M \geq a - y) \cdot \mathbb{P}(S_{c, M} = a) \cdot dP_{S_{c, M}}(y, a)
\leq \mathbb{P}(M_c \geq \frac{b(x-c)}{8}) + \mathbb{P}(S_c \leq -\frac{b(x-c)}{8})
+ \int_{\mathbb{R}^2} \mathbb{P}(\max_{r \geq x-c} S_r \geq a - y \mid M \geq a - y) \mathbb{1}_{\{a-y \leq b(x-c)/4\}} \cdot dP_{S_{c, M}}(y, a)
\leq e^{-b^2(x-c)/8} + \Phi\left(-\frac{b(x-9c)}{8\sqrt{c}}\right) + \frac{3}{2}e^{-b^2(x-c)/8} \leq 3e^{-b^2(x-c)/8},
$$

34
for \( x \geq 31c \), where in the final step we have used the Gaussian tail bound \( \Phi(-x) \leq \frac{1}{2}e^{-x^2/2} \) for \( x \geq 0 \) and the fact that
\[
\frac{(x - 9c)^2}{c} \geq \frac{(1 - 8/30)^2(x - c)^2}{c} \geq 30(1 - 8/30)^2(x - c) \geq 16(x - c).
\]
Combining with (38), we conclude that
\[
\mathbb{E}\left(\max_{j \in [s]} \xi^j \mid \min_{j \in [s]} \xi^j \geq c\right) - 31c \leq 3s \sum_{x=31c}^{\infty} e^{-b^2(x-c)/8} = \frac{3se^{-15b^2c/4}}{1 - e^{-b^2/8}} \leq \frac{3b^{15/2}}{15b^2/128} \leq c,
\]
as desired, where we have used again the fact that \( 1 - e^{-x} \geq 15x/16 \) for \( x \in [0, 1/8] \) in the penultimate inequality.

\textbf{Lemma 14.} (a) For any \( n \in \mathbb{N} \), \( 0 < p \leq q < 1 \) and \( x \in \{0, 1, \ldots, n\} \), we have
\[
\frac{\mathbb{P}(\text{Bin}(n, p) \leq x)}{\mathbb{P}(\text{Bin}(n, p/q) \leq x)} \leq \mathbb{P}(\text{Bin}(n, q) \geq x). \tag{40}
\]
(b) Let \( W_1, \ldots, W_n \) be independent and identically distributed, real-valued random variables, with corresponding order statistics \( W_{(1)} \leq \ldots \leq W_{(n)} \). Then for every \( s \geq t \) and every \( m \in [n] \), we have that
\[
\mathbb{P}(W_{(m)} \geq s \mid W_{(m)} \geq t) \leq \mathbb{P}(W_{(m)} \geq s \mid W_{(1)} \geq t).
\]
In particular, \( \mathbb{E}(W_{(m)} \mid W_{(m)} \geq t) \leq \mathbb{E}(W_{(m)} \mid W_{(1)} \geq t) \).

\textbf{Proof.} (a) Let \( g(p) \) denote the left-hand side of (40). It suffices to prove that \( g \) is an increasing function on \( (0, q] \). We may also assume that \( x \geq 1 \), because otherwise the result is clear. Now, let
\[
h(p) := \mathbb{P}(\text{Bin}(n, p) \geq x) = \sum_{r=x}^{n} \binom{n}{r} p^r (1 - p)^{n-r}.
\]
Then
\[
h'(p) = \sum_{r=x}^{n} \binom{n}{r} rp^{r-1}(1 - p)^{n-r} - \sum_{r=x}^{n-1} \binom{n}{r} (n - r)p^r (1 - p)^{n-r-1}
\]
\[
= \binom{n}{x} x! (n - x) ! p^x (1 - p)^{n-x-1} - \sum_{r=x}^{n-1} \binom{n}{r} r! (n - r - 1)! p^r (1 - p)^{n-r-1}
\]
\[
= \frac{n!}{(x - 1)! (n - x)!} p^{x-1} (1 - p)^{n-x}.
\]
We can therefore compute
\[
g'(p) = \frac{h(p/q)h'(p) - h(p)h'(p/q)/q}{h(p/q)^2},
\]
35
and we note that

\[
    h(p/q)h'(p) - h(p)h'(p/q)/q \\
    = \frac{n!}{(x-1)!(n-x)!} B^{-1} (1-p)^{n-x} \frac{n}{r} \left( \frac{p}{q} \right) (1 - \frac{p}{q})^{n-r} \\
    - \frac{n!}{(x-1)!(n-x)!} \frac{1}{q} B^{-1} (1-p)^{n-x} \frac{n}{r} \left( \frac{p}{q} \right) (1 - \frac{p}{q})^{n-r} \\
    = \frac{n! B^{-1} (1-p)^{n-x} (1-p/q)^{n-x}}{q^2 (x-1)! (n-x)!} \frac{n}{r} \left( \frac{p}{q} \right) (1 - \frac{p}{q})^{n-r} \\
    \frac{1}{(q-p)^{r-x}} - \frac{1}{p^{r-x}} \geq 0,
\]

as required.

(b) Write \( F \) for the distribution function of \( W_1 \), and let \( \bar{F} := 1 - F \). We also write \( \bar{F}(x-) := \lim_{y \nearrow x} \bar{F}(x) \). For a Borel measurable set \( A \subseteq \mathbb{R} \), let \( N(A) := \sum_{i=1}^{n} 1_{\{ W_i \in A \}} \). Then, for \( s \geq t \),

\[
    \mathbb{P}(W_{(m)} \geq s | W_{(m)} \geq t) = \frac{\mathbb{P}(W_{(m)} \geq s)}{\mathbb{P}(W_{(m)} \geq t)} = \frac{\mathbb{P}(N([s, \infty)) \geq n - m + 1)}{\mathbb{P}(N([t, \infty)) \geq n - m + 1)} \\
    = \frac{\mathbb{P}\{ \text{Bin}(n, \bar{F}(s-)) \geq n - m + 1 \}}{\mathbb{P}\{ \text{Bin}(n, \bar{F}(t-)) \geq n - m + 1 \}}.
\]

On the other hand,

\[
    \mathbb{P}(W_{(m)} \geq s | W_{(1)} \geq t) = \frac{\mathbb{P}(W_{(m)} \geq s, W_{(1)} \geq t)}{\mathbb{P}(W_{(1)} \geq t)} \\
    = \frac{\mathbb{P}\{ N((-\infty, t)) = 0, N([s, \infty)) \geq n - m + 1 \}}{\mathbb{P}\{ N((-\infty, t)) = 0 \}} \\
    = \sum_{r=n-m+1}^{n} \binom{n}{r} \bar{F}(s-)^{r} \frac{\bar{F}(t-) - \bar{F}(s-)}{\bar{F}(t-)^n} \\
    = \mathbb{P}\{ \text{Bin}(n, \bar{F}(s-)/\bar{F}(t-)) \geq n - m + 1 \}.
\]

The first conclusion therefore follows immediately from (a), and the second conclusion is an immediate consequence of the first.

\[\square\]

**Lemma 15.** Let \( v = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p \) be a unit vector. There exists \( \ell \in \{0, \ldots, \lfloor \log_2 p \rfloor \} \) such that

\[
    \left\lfloor \sum_{j \in [p]} \left\lfloor \frac{1}{2^\ell \log_2(2p)} \right\rfloor \right\rfloor \geq 2^\ell.
\]

**Proof.** The case \( p = 1 \) is trivially true, so we may assume without loss of generality that \( p \geq 2 \). Let \( L := \lfloor \log_2 p \rfloor \), \( b_\ell := 2^{-\ell} \log_2^{-1}(2p) \) and \( n_\ell := \left\lfloor \left\{ j : v_j^2 \geq b_\ell \right\} \right\rfloor \) for \( \ell \in \{0, \ldots, L\} \).
Assume for a contradiction that $n_{\ell} < 2^\ell$ for all $\ell$. Then by Fubini’s theorem we have

$$
\|v\|_2^2 = \sum_{j=1}^{p} \int_{t=0}^{1} \mathbb{1}_{\{v_j^2 \geq t\}} dt \leq n_0(1 - b_0) + \sum_{\ell=1}^{L} n_{\ell}(b_{\ell-1} - b_{\ell}) + pb_L
$$

$$
\leq \sum_{\ell=1}^{L} (2^\ell - 1)(b_{\ell-1} - b_{\ell}) + pb_L = \sum_{\ell=0}^{L-1} 2^\ell b_{\ell} + (p - 2^L + 1)b_L \leq \frac{L + 1}{\log_2(2p)} \leq 1.
$$

Note that the penultimate inequality is strict if $p + 1$ is not an integer power of 2 and the final inequality is strict if $p$ is not an integer power of 2. Since $p \geq 2$, it cannot be the case that we have equality in both equalities, so $\|v\|_2^2 < 1$, which contradicts the fact that $v$ is a unit vector. 

\[\Box\]

\textbf{Lemma 16.} Define sequences $(a_n)_{n \in \mathbb{N}_0}$ and $(b_n)_{n \in \mathbb{N}_0}$ as follows: $a_0 := b_0 := 0$, $b_n := (b_{n-1} + 1)\mathbb{1}_{\{n \not\in \{2^\ell : \ell \in \mathbb{N}_0\}\}}$ and $a_n := (a_{n-1} + 1)\mathbb{1}_{\{n \not\in \{2^\ell : \ell \in \mathbb{N}_0\}\}} + (b_{n-1} + 1)\mathbb{1}_{\{n \in \{2^\ell : \ell \in \mathbb{N}_0\}\}}$ for $n \in \mathbb{N}$. Then, we have

$$
n/2 \leq a_n < 3n/4,
$$

for all $n \geq 2$.

\textbf{Proof.} The two sequences $(a_n)_{n \in \mathbb{N}_0}$ and $(b_n)_{n \in \mathbb{N}_0}$ are tabulated below.

<table>
<thead>
<tr>
<th>$n$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>$\ldots$</th>
<th>$2^\ell$</th>
<th>$2^\ell + 1$</th>
<th>$\ldots$</th>
<th>$2^\ell + 1 - 1$</th>
<th>$\ldots$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_n$</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>$\ldots$</td>
<td>$2^\ell$</td>
<td>$2^\ell - 1$</td>
<td>$\ldots$</td>
<td>$3 \cdot 2^\ell - 1$</td>
<td>$\ldots$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b_n$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>$\ldots$</td>
<td>0</td>
<td>1</td>
<td>$\ldots$</td>
<td>$2^\ell - 1$</td>
<td>$\ldots$</td>
</tr>
</tbody>
</table>

It is clear from the definition of $(b_n)_n$ that $b_{2^\ell + i} = i$ for $\xi \in \mathbb{N}_0$ and $0 \leq i \leq 2^\ell - 1$. Consequently, we have $a_{2^\ell} = b_{2^\ell - 1} + 1 = 2^\ell - 1$ and $a_{2^\ell + i} = 2^\ell - 1 + i$ for $\xi \in \mathbb{N}$ and $1 \leq i \leq 2^\ell - 1$. Hence, we have

$$
\frac{1}{2} \leq \frac{2^\ell - 1}{2^\ell} \leq \frac{a_{2^\ell + i}}{2^\ell + i} = \frac{2^\ell - 1 + i}{2^\ell + i} \leq \frac{2^\ell - 1 + 2^\ell - 1}{2^\ell + 2^\ell - 1} < \frac{3}{4},
$$

for all $\xi \in \mathbb{N}$ and $0 \leq i \leq 2^\ell - 1$ and the desired result follows. \[\Box\]

\textbf{Lemma 17.} Let $Z_1, \ldots, Z_p \overset{iid}{\sim} N(0, 1)$. Then for any $a > 0$ and $x > 0$, we have

$$
\mathbb{P}
\left(\sum_{j=1}^{p} Z_j^2 \mathbb{1}_{\{|Z_j| \geq a\}} \geq 6pe^{-a^2/8} + 4x\right) \leq e^{-x}.
$$

\textbf{Proof.} This proof has some similarities with that of Lemma 17 of Liu, Gao and Samworth (2019). By a Chernoff bound, we have for any $u, \lambda > 0$ that,

$$
\mathbb{P}
\left(\sum_{j=1}^{p} Z_j^2 \mathbb{1}_{\{|Z_j| \geq a\}} \geq u\right) \leq e^{-\lambda u \left\{ \mathbb{E}e^{\lambda Z_1^2 \mathbb{1}_{\{|Z_j| \geq a\}}} \right\}^p}.
$$

(41)
We write $p(x) := (2\pi)^{-1/2}x^{-1/2}e^{-x/2}$ for the density of a $\chi_1^2$ distribution. For $\lambda \in (0, 1/4]$, we bound the moment generating function above as follows:

$$
\mathbb{E}e^{\lambda Z_1^2 \mathbb{1}_{\{|Z_1| \geq a\}}} = \int_{a^2}^{\infty} e^{\lambda x} p(x) \, dx \leq 1 + \int_{a^2}^{\infty} (e^{\lambda x} - 1) p(x) \, dx = 1 + \int_{a^2}^{\infty} \sum_{k=1}^{\infty} \frac{\lambda^k x^k}{k!} p(x) \, dx
$$

$$
= 1 + \int_{a^2}^{\infty} \frac{\lambda x e^{\lambda x} p(x) \, dx}{\sqrt{2\pi}} = 1 + \frac{4\lambda}{\sqrt{\pi}} \int_{a/\sqrt{2}}^{\infty} t^2 e^{-t^2/2} \, dt
$$

$$
= 1 + \sqrt{\frac{8}{\pi}} \lambda e^{-a^2/4} + 4\sqrt{2}\lambda \left( 1 - \Phi\left( \frac{a}{\sqrt{2}} \right) \right) \leq 1 + \sqrt{\frac{8}{\pi}} \lambda e^{-a^2/4} + 2\sqrt{2}\lambda e^{-a^2/4}
$$

$$
\leq 1 + \left( 2\frac{8}{\pi} e^{-1/2} + 2\sqrt{2} \right) \lambda e^{-a^2/8} \leq 1 + 5\lambda e^{-a^2/8},
$$

where we use the fact that $xe^{-x^2/4} \leq 2e^{-1/2}e^{-x^2/8}$ for $x \in \mathbb{R}$ in the penultimate inequality. Hence, by substituting this bound into (41), we have for every $u > 0$, that

$$
\mathbb{P}\left( \sum_{j=1}^{p} Z_j^2 \mathbb{1}_{\{|Z_j| \geq a\}} \geq u \right) \leq \exp\left\{ -\lambda u + p \log(1 + 5\lambda e^{-a^2/8}) \right\} \leq \exp\left\{ -\lambda u + 5p\lambda e^{-a^2/8} \right\}.
$$

We set $u = 6pe^{-a^2/8} + 4x$. If $x \leq pe^{-a^2/8}/4$, choose $\lambda = p^{-1}xe^{-a^2/8} \leq 1/4$; if $x > pe^{-a^2/8}/4$, choose $\lambda = 1/4$. In both cases, we have

$$
\mathbb{P}\left( \sum_{j=1}^{p} Z_j^2 \mathbb{1}_{\{|Z_j| \geq a\}} \geq u \right) \leq e^{-x},
$$

as required. \(\square\)

**References**


