Optimistic Search: Change Point Estimation for Large-scale Data via Adaptive Logarithmic Queries

Solt Kovács[1,*]   KOVACSSOLT@GMAIL.COM

Housen Li[2,3,*]   HOUSEN.LI@MATHEMATIK.UNI-GOETTINGEN.DE

Lorenz Haubner[1]   LORENZ.HAUBNER@GMAIL.COM

Axel Munk[2,3]   MUNK@MATH.UNI-GOETTINGEN.DE

Peter Bühlmann[1]   PETER.BUEHLMANN@STAT.MATH.ETHZ.CH

[1] *Seminar für Statistik, ETH Zürich, 8092 Zürich, Switzerland*

[2] *Institute for Mathematical Stochastics, University of Göttingen, 37077 Göttingen, Germany*

[3] *Cluster of Excellence Multiscale Bioimaging (MBExC), University of Göttingen, Germany*

[*] *The first two authors contributed equally to this work*

Editor: Nicolas Vayatis

Abstract

Change point estimation is often formulated as a search for the maximum of a gain function describing improved fits when segmenting the data. Searching one change point through all candidates requires $O(n)$ evaluations of the gain function for an interval with n observations. If each evaluation is computationally demanding (e.g. in high-dimensional models), this can become infeasible. Instead, we propose *optimistic search*, a methodology that only requires $O(\log n)$ evaluations of the gain function, leading to huge computational gains for massive (large-scale, high-dimensional) data for single and multiple change point estimation.

Towards solid understanding of our strategy, we investigate in detail the p-dimensional Gaussian changing means setup, including high-dimensional scenarios. For some of our proposals, we prove asymptotic minimax optimality for detecting change points and derive sharp asymptotic rates for localizing change points.

Our search strategy generalizes far beyond the theoretically analyzed setup. We illustrate, as an example, massive computational speedup in change point detection for high-dimensional Gaussian graphical models.

Keywords: fast computation, high-dimension, minimax optimality, multiple break points, sublinear runtime

1. Introduction

Change point (break point) estimation tackles the problem of estimating the locations of abrupt structural changes for ordered and noisy data, e.g., by time or space. One can distinguish between online (sequential) and offline (retrospective) detection problems. We will primarily focus on the latter setup where all ordered observations are available at once and only point to online detection in connection with our methods and results for the detection of a single change point. Applications include detecting changes in copy number variation (Olshen et al., 2004; Zhang and Siegmund, 2007; Hocking et al., 2013),

ion channels (Hotz et al., 2013), financial time series (Bai and Perron, 1998; Kim et al., 2005; Davies et al., 2012), climate data (Reeves et al., 2007), environmental monitoring systems (Londschien et al., 2021), among many others. We refer also to the recent reviews (Niu et al., 2016; Truong et al., 2020).

We focus on computational improvements of change point inference while maintaining statistical optimality. Two common algorithmic approaches are optimal partitions via dynamic programming (Jackson et al., 2005; Friedrich et al., 2008), and greedy procedures, e.g. binary segmentation (BS, Vostrikova, 1981) and its variants. The former approaches are mainly studied for univariate data. Typical examples include ℓ_0 penalization methods (cf. Boysen et al., 2009; PELT, Killick et al., 2012 and FPOP, Maidstone et al., 2017) and multiscale methods (SMUCE, Frick et al., 2014, FDRSeg, Li et al., 2016, MQS, Vanegas et al., 2022, and MUSCLE, Liu and Li, 2024), which are known to be statistically minimax optimal in a Gaussian setup. However, finding the optimal partition requires in the worst case at least quadratic run time. In contrast, BS is typically faster and easier to adapt to more general scenarios, but worse in terms of estimation performance than methods finding the optimal partitioning. The wild binary segmentation (WBS, Fryzlewicz, 2014) and its variants (e.g. narrowest over threshold, Baranowski et al., 2019) improve on estimation performance of BS, but lose some of its computational efficiency.

Already for simple univariate cases, increasingly larger data sets with long time series led to the development of computationally more efficient (univariate) approaches (Maidstone et al., 2017; Lu et al., 2017; Fryzlewicz, 2020; Kovács et al., 2023). Computational issues are even more pronounced for multivariate problems. This is in particular true for emerging high-dimensional parametric change point detection approaches (e.g. Sharpnack et al., 2013; Leonardi and Bühlmann, 2016; Roy et al., 2017; Gibberd and Nelson, 2017; Gibberd and Roy, 2017; Bybee and Atchadé, 2018; Avanesov and Buzun, 2018; Wang and Samworth, 2018; Wang et al., 2021c; Londschien et al., 2021; Yu and Chen, 2021; Wang et al., 2021b,a; Dette et al., 2022; Xu et al., 2024; Cho and Owens, 2024; Gao and Wang, 2022; Li et al., 2023; Cho et al., 2024), or multivariate non-parametric change point detection methods (Madrid Padilla et al., 2022). Many of these approaches rely on computationally costly fits of algorithms such as the lasso (Tibshirani, 1996) or the graphical lasso (Friedman et al., 2008). Performing a full grid search in order to find a single split point requires as many fits as there are observations. Even with warm-starts, neighboring fits with one additional observation are not straightforward to update (unlike e.g. means in univariate cases) and the number of fits is the main driver of computational cost. Thus, for a few hundred or thousand observations, full grid search based methods (incl. BS, its variants and even more costly dynamic programming approaches) can be very slow, beyond what is acceptable. This is a main motivation for our work, namely to avoid too many fits with piecewise stationary model structure.

1.1 Our Contribution and Related Work

The key idea is to replace the exhaustive search for a change point by an *adaptive search* that dynamically determines the next search location given the previous ones in the spirit of divide-and-conquer, requiring only a logarithmic number of operations. We call this novel methodology for searching for single change points *optimistic search* (OS). The basic idea

behind OS is based on a rather general observation: In numerous change point problems, population gain functions (describing the expected gain when splitting into two parts at a given split point) have a specific piecewise quasiconvex structure where local maxima correspond to change points, and thus the presence of a single change point implies only one global maximum.

Besides computational advantages, methods built on OS achieve a high statistical efficiency. Such a method for finding a single change point is advanced optimistic search (aOS). We will show (Theorems 2 and 6) that in a Gaussian setup it achieves asymptotically minimax optimal detection and nearly minimax optimal localization of the change point with $O(\log n)$ evaluations as opposed to full grid search methods requiring at least $O(n)$ evaluations. These results are surprising and fundamental, as asymptotically one loses nothing in terms of detectable signals and in terms of localization error compared to full grid search, while at the same hand computationally only $O(\log n)$ operations are required. In case of multiple change points, we extend this idea to develop a method (Optimistic Seeded Binary Segmentation, OSeedBS), which achieves similar computational speedup and statistical optimality as in the single change point scenario. In p-dimensional Gaussian changing mean problems, our methodology is significantly faster, e.g., finding a change point in $O(p \log n)$ run time. This requires that cumulative sums have been pre-computed, which can easily be done in online change point detection. Note that the existing (full grid search based) methods also benefit computationally from the availability of cumulative sums. Thus, we suggest to store data in cumulative sums format also for offline analysis. Even when the data are not properly stored, the worst case computation complexity of our method for finding all possible change points is $O(np)$, which is limited by the cost of calculating the cumulative sums. In particular, for the univariate setup with multiple change points, it improves the worst case computation complexity of $O(n \log n)$ that is the fastest in the literature, by a log factor. Further, the improvement becomes way more significant in more complex models where the number of model fits involved in each evaluation of the gain function are the main driver of computational cost, see Table 1 for an overview.

| | Method | Evaluations of gain function | Computation complexity for p-dimensional data | | |
			mean changes with cum. sums	mean changes w/o cum. sums	covariance changes
Single change	aOS	$O(\log n)$	$O(\log n) \cdot p$	$O(n) \cdot p$	$O(\log n) \cdot \mathrm{glasso}(p)$
	BS	$O(n)$	$O(n) \cdot p$	$O(n) \cdot p$	$O(n) \cdot \mathrm{glasso}(p)$
Multiple changes	OSeedBS	$O\big((n^\omega \log n) \wedge n\big)$	$O\big((n^\omega \log n) \wedge n\big) \cdot p$	$O(n) \cdot p$	$O\big((n^\omega \log n) \wedge n\big) \cdot \mathrm{glasso}(p)$
	SeedBS	$O(n \log n)$	$O(n \log n) \cdot p$	$O(n \log n) \cdot p$	$O(n \log n) \cdot \mathrm{glasso}(p)$

Table 1: Computation comparison of OS (aOS in Algorithm 2, and OSeedBS in Algorithm 3) and full grid search methods (BS, and SeedBS in Kovács et al., 2023). Here glasso(p) denotes the computation complexity of graphical lasso estimator of the precision matrix in $\mathbb{R}^{p \times p}$, which is typically $O(p^3)$ or worse, see Appendix C. Sample size is n, and $n^{-\omega}$ with $\omega \in [0,1]$ encodes an a-priori lower bound on the minimal length of stationary segments, which is tailored to particular applications. Note that $\omega = 1$ corresponds to the case of no such a-priori knowledge.

Some problem specific solutions for gaining computational speedup in certain high-dimensional setups have been suggested, e.g. for changing graphical models (Hallac et al., 2017; Bybee and Atchadé, 2018) or changing linear regression coefficients (Kaul et al., 2019a; Kaul et al., 2019b), but these may not be easy to adapt to other scenarios. The following two proposals may be seen as most related to our approach. Our idea of avoiding a high number of model fits is vaguely related to the procedure of Kaul et al. (2019a) for change point detection in high-dimensional linear regression. For an initial split point they fit appropriate models which are then kept for evaluating (an approximation of) the gain function for candidate split points on the full grid, but without updating the costly high-dimensional fits. However, in scenarios where the evaluation of the gain on the full grid has a comparable cost as the model fit itself (e.g. calculating and updating means in the Gaussian change in mean setup), their procedure would not lead to any speedup. In Lu et al. (2017), the authors proposed a procedure specific for the univariate case. They thin out the number of evaluations by searching on a small subsample to obtain preliminary estimates, and then use a dense sample in neighbourhoods for the final change point estimates. A rough subsample is not well suited for high-dimensional problems, unlike our approach of keeping all samples but avoiding evaluations on the full grid. This will be illustrated in the following example.

1.2 A Motivating Example

Rather than a problem specific solution, the proposed OS is a computationally attractive and general methodology for searching for a single change point. Figure 1 illustrates its computational efficiency in an example of single change point detection for a Gaussian graphical model (based on an estimator discussed in Section 6.3 and Appendix C). The aim is to find the maximum of the gain function (black curve). Evaluating the gain function at a single split point $t \in \{1, \ldots, n\}$ with $n = 2,000$ requires two graphical lasso fits: one for the segment $(1, t]$ and one for $(t, n]$. For finding the maximum, the full grid search took roughly 100 times longer than OS. The latter evaluated the gain only at two initial points (marked by two zeros) and subsequently at further 14 split points (marked by colored numbers) which are determined dynamically. Then OS returns the maximum over all considered split points. This leads to a massive reduction of computation time.

When searching for multiple change points, OS can be combined flexibly with existing algorithms as discussed in Section 4, resulting again in massive computational speedup with essentially no loss in statistical performance.

1.3 Outline and Announcing Our Results

The crucial part of our methodology is the OS introduced in Section 3. It is capable of finding a local maximum of the population gain function with $O(\log n)$ evaluations for n observations. For the sake of readability, we first introduce our methodology as well as the statistical guarantees in the classical *univariate* Gaussian change in mean setup, detailed in Section 2. In the scenario when only one change point is present, the advanced version of OS, i.e. aOS, (Theorem 2) is able to detect the change point even in the most difficult detection scenario

$$\text{minimal jump size} \times \sqrt{\text{minimal segment length}} \gtrsim \sqrt{\log\log(n)/n} \qquad (1)$$
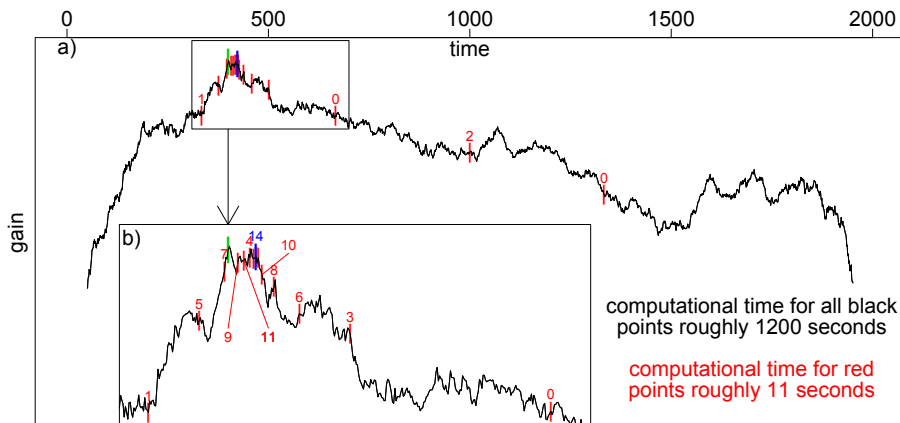
Figure 1: Finding a single change point with full grid search (black) and OS (red) in a 200×200-dimensional covariance change example with graphical lasso fits. OS starts with two evaluations marked by the two zeros and then evaluates further 14 split points adaptively, in the order marked by the numbers shown in the zoomed in b). The true underlying change point at observation 400 is marked in green and the final candidate returned by OS at observation 423 in blue. The overall maximum of the black gain curve found by full grid search is at observation 402.

thus being optimal in minimax sense. In Section 4 we extend the methodology for multiple change points and derive a minimax optimal performance result. Namely, under nearly the same condition as in (1), the number of change points is identified correctly, and the location of each change point is estimated at the best rate (up to possible log factors) that is available in literature, see Theorem 5.

We further examine general *multivariate* and *high-dimensional* Gaussian mean changes in Section 5. We again obtain via aOS the detection and the localization rates of change points, which are (nearly) the best in the literature. Interestingly, unlike the univariate and multivarite cases, the localization rate may be much faster than the rate induced by the detection problem in several high-dimensional setups, including both sparse and dense scenarios, e.g., when the signal-to-noise ratio is larger than the minimax detection rate. Our theoretical analysis is built on a careful exploitation of algorithmic structures, and a subtle treatment of the difference of certain (noncentral) chi-squared random variables resulting from the comparison of gain functions, which takes advantage of local dependence behavior (Lemmas 15 and 18 in the Appendix). Although the detailed proofs are rather technical, we stress that the main arguments can serve as a roadmap for more complicated models, which require their own treatments.

Sections 6 and 7 contain empirical simulation results and conclusions. Additional material on simulation study and high-dimensional scenarios as well as proofs are given in the Appendix. There we present several deviation inequalities for randomly weighted chi-squares which are of independent interest.

Notation. For a real number r, we define downward rounding as $\lfloor r \rfloor = \max\{n \in \mathbb{Z} : n \leq r\}$ and upward rounding as $\lceil r \rceil = \min\{n \in \mathbb{Z} : n \geq r\}$, and also define $(r)_+ := \max\{r, 0\}$.

For two sequences of positive real numbers $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n \ll b_n$, or $a_n = o(b_n)$, if $\lim_{n \to \infty} a_n/b_n = 0$, write $a_n \lesssim b_n$, or $a_n = O(b_n)$, if $\limsup_{n \to \infty} a_n/b_n < \infty$, and write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We use bold symbols for vectors and matrices to differentiate them from scalar values.

2. Gaussian Mean Shifts with Constant Variance

We consider first the simple model of univariate Gaussian changing means (Model 1). The results are later generalized to multivariate and high-dimensional scenarios in Section 5.

Model 1 (Univariate Gaussian changing means) *Assume that observations* $X_1, \ldots, X_n$ *are independent and*

$$X_{\tau_0 n+1}(= X_1), \ldots, X_{\tau_1 n} \sim \mathcal{N}(\mu_0, \sigma^2),$$
$$\vdots$$
$$X_{\tau_\kappa n+1}, \ldots, X_{\tau_{\kappa+1} n}(= X_n) \sim \mathcal{N}(\mu_\kappa, \sigma^2),$$

where $\{\tau_i : i = 1, \ldots, \kappa\}$ *is the locations of change points satisfying*

$$0 = \tau_0 < \tau_1 < \cdots < \tau_{\kappa+1} = 1 \quad and \quad \tau_i n \in \mathbb{N},$$

means $\mu_i \neq \mu_{i-1}$, $i = 1, \ldots, \kappa$ *are the levels on segments, and the standard deviation* $\sigma > 0$ *is known. Assume w.l.o.g.* $\sigma = 1$. *Further, define the minimal segment length*

$$\lambda \equiv \lambda_n = \min_{i=0,\ldots,\kappa} (\tau_{i+1} - \tau_i),$$

and the minimal jump size

$$\delta \equiv \delta_n = \min_{i=1,\ldots,\kappa} \delta_i \quad with \quad \delta_i = |\mu_i - \mu_{i-1}|.$$

The goal of change point inference is to estimate the number κ and the locations τ_i's of the true underlying change points from realizations of $X_1, \ldots, X_n$. A common criterion for determining the best split point is the CUSUM statistics (Page, 1954), defined for an interval $(l, r]$ and a split point t as

$$\text{CS}_{(l,r]}(t) = \sqrt{\frac{r-t}{(r-l)(t-l)}} \sum_{i=l+1}^{t} X_i - \sqrt{\frac{t-l}{(r-l)(r-t)}} \sum_{i=t+1}^{r} X_i, \tag{2}$$

with integers $0 \leq l < t < r \leq n$. The CUSUM statistic is the likelihood ratio test for a single change point at location t in the interval $(l, r]$ against a constant signal. The population counterpart of $|\text{CS}_{(l,r]}(\cdot)|$, i.e. replacing X_i by $\mathbb{E}[X_i]$, has its maximum at one of the underlying change points. In noisy cases, the best split point candidate when dividing the segment $(l, r]$ into two parts is the location of the maximal absolute CUSUM statistics

$$\hat{t}_{(l,r]} = \arg\max_{t \in \{l+1,\ldots,r-1\}} |\text{CS}_{(l,r]}(t)|.$$

Algorithm 1 Naive Optimistic Search (OS)

Require: $r - l > 2$, $l, r \in \mathbb{N}$; and step size $\nu \in (0, 1)$ with $1/2$ by default

1: **initialize:** $\tilde{l} \leftarrow l$, $\tilde{r} \leftarrow r$ and $t \leftarrow \lfloor (l + \nu r)/(1 + \nu) \rfloor$
2: **function** OS($\tilde{l}, t, \tilde{r} \mid \nu, l, r$)
3:     **if** $\tilde{r} - \tilde{l} \leq 5$ **then**   ▷ Stopping condition for recursion
4:         $\hat{t}_{(l,r]} \leftarrow \underset{t \in \{\tilde{l}+1,\ldots,\tilde{r}-1\}}{\arg\max} \; G_{(l,r]}(t)$ ▷ Search over all points if less than 5 remain
5:         **return** $\hat{t}_{(l,r]}$
6:     **if** $\tilde{r} - t > t - \tilde{l}$ **then**   ▷ Pick a new probe point in larger segment
7:         $w \leftarrow \lceil \tilde{r} - (\tilde{r} - t)\nu \rceil$
8:         **if** $G_{(l,r]}(w) \geq G_{(l,r]}(t)$ **then**
9:             OS($t, w, \tilde{r} \mid \nu, l, r$)
10:         **else**
11:             OS($\tilde{l}, t, w \mid \nu, l, r$)
12:     **else**
13:         $w \leftarrow \lfloor \tilde{l} + (t - \tilde{l})\nu \rfloor$
14:         **if** $G_{(l,r]}(w) \geq G_{(l,r]}(t)$ **then**
15:             OS($\tilde{l}, w, t \mid \nu, l, r$)
16:         **else**
17:             OS($w, t, \tilde{r} \mid \nu, l, r$)

We refer to the function $|\mathrm{CS}_{(l,r]}(\cdot)|$ as a gain function, denoted by $G_{(l,r]}(\cdot)$, because the square of it describes gains, namely the reductions in squared errors when fitting separate means on the left and right segments for split points in the segment $(\ell, r]$. The gain functions $G_{(l,r]}(\cdot)$ are initially defined on a discrete grid of split points, but it is convinient to extend them continuously to $t \in (l, r]$ (via e.g. linear interpolation).

3. Optimistic Search for a Single Change Point

In this section, we focus on Model 1 with a single change point ($\kappa = 1$), and introduce two versions of OS that are at the core of our methodology for multiple change points, as well.

3.1 Naive Optimistic Search

We introduce first the *naive* version of optimistic search (OS) within a segment $(l, r]$ in Algorithm 1, which is a key building block for later introduced statistically optimal methods. The procedure is similar to the golden section search (Kiefer, 1953; Avriel and Wilde, 1966, 1968), typically used to find the global extremum of unimodal functions. OS splits an interval into three segments recursively and discards one of the outer segments in each iteration. For unimodal functions with one peak this search converges to the global maximum. If there is only a single change point contained in $(l, r]$, then the (continuously embedded) population gain function is unimodal with the single peak at the true underlying change point. Optimism is required in noisy scenarios, and hence the naming of the method, as
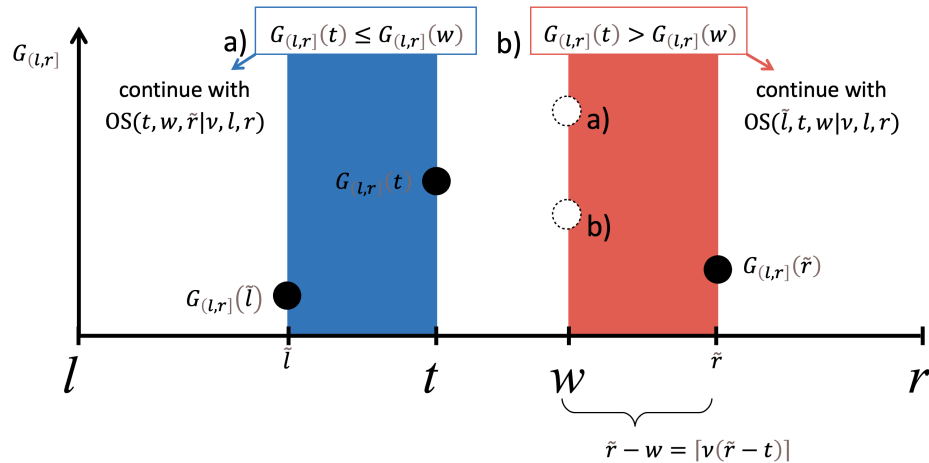
Figure 2: Naive optimistic search step $\text{OS}(\tilde{l}, t, \tilde{r} \mid \nu, l, r)$ within the current segment $(\tilde{l}, \tilde{r}] \subseteq (l, r]$ given the previous evaluation at t and step size ν. As $\tilde{r} - t > t - \tilde{l}$, the new probe point w is taken within $(t, \tilde{r}]$ as $w = \lceil \tilde{r} - (\tilde{r} - t)\nu \rceil$. Depending on the gain $G_{(l,r]}(w)$ vs. $G_{(l,r]}(t)$, one either continues with $\text{OS}(t, w, \tilde{r} \mid \nu, l, r)$ (discarding the blue part) or $\text{OS}(\tilde{l}, t, w \mid \nu, l, r)$ (discarding the red part).

noisy counterparts are rather "wiggly" functions following the shape of the population gain function only approximately.

When initializing by calling $\text{OS}(l, \lfloor (l + \nu r)/(1 + \nu) \rfloor, r \mid \nu, l, r)$ for an interval $(l, r]$, the search first probes the points $t = l + (r - l)\nu/(1 + \nu)$ and $w = r - (r - l)\nu/(1 + \nu)$ (up to rounding), i.e. the two first probe points are equally distant from l and r respectively. Depending on the gains at the probe points t and w either $(l, t]$ or $(w, r]$ is then discarded. The possible decisions for the general case when the search is already narrowed down to the sub-interval $(\tilde{l}, \tilde{r}]$ is depicted in Figure 2. Note that in general, the lengths of the two candidate intervals for discarding are not necessarily equal. In case (a) we have $G_{(l,r]}(t) < G_{(l,r]}(w)$ and hence the less promising blue area is discarded. In case (b) we have $G_{(l,r]}(t) > G_{(l,r]}(w)$ and the red area is discarded. Also note that one of the previous probe points will be one of the new boundary points while the other probe point is going to be one of the new probe points in the middle with gain that is thus at least as high as for the new boundary. This leads to a "triangular structure" that one probe point in the middle has a higher gain than both boundary points, throughout the search.

We set $\nu = 1/2$ by default, but in general ν can be interpreted as (relative) step size, expected to reflect some kind of trade off between computational performance and estimation accuracy. The choice of evaluating the last 5 points remaining (in line 3) is somewhat arbitrary and can be also set to e.g. 10 or 20. In rare cases, when $r - t = t - l$ or $G_{(l,r]}(w) = G_{(l,r]}(t)$ one can also take a (pseudo) random choice or incorporate additional information (e.g. variance in the segments for Model 1) to decide on the new probe point and the segment to discard.

Theorem 1 (Naive optimistic search) *Under Model 1 with a single change point, i.e. $\kappa = 1$, we assume that the minimal segment length λ and the minimal jump size δ satisfy*

$$\delta\lambda\sqrt{n} \geq C_0\sqrt{\log\log n} \qquad (3)$$

for some large enough constant C_0. Let $\hat{\tau} = \hat{t}_{(0,n]}/n$ be the estimated change point by OS (Algorithm 1) on $(0, n]$. Then:

$$\lim_{n\to\infty} \mathbb{P}\left\{|\hat{\tau} - \tau| \leq C_1 \frac{\log\log n}{\delta^2 n}\right\} = 1 \qquad \text{with some constant } C_1.$$

Theorem 1 states that OS detects the only change point with a localization error of order $\log\log n/(\delta^2 n)$, which is minimax optimal up to a possible log factor (see e.g. Lemma 1 in Verzelen et al., 2023). In comparison with the weakest condition to ensure consistency of change point estimation, which is $\delta\sqrt{\lambda n} \gtrsim \sqrt{\log\log n}$ (Liu et al., 2021), OS is sub-optimal. However, in the particular case that the minimal segment length does not vanish, i.e. $\lambda \asymp 1$, the condition (3) becomes $\delta\sqrt{n} \gtrsim \sqrt{\log\log n}$ and OS is then optimal. This is the situation where the length of the left segment is comparable to that of the right segment, which we thus refer to as a "balanced" scenario. In contrast, "unbalanced" scenarios are ones where the lengths of shorter and longer segments are very different. It is possible to show that the suboptimal condition (3) cannot be improved and is intrinsic to OS, rather than an artifact in our theoretical analysis (see Example 4 in Appendix B).

3.2 Advanced Optimistic Search

In Algorithm 2 we propose the *advanced optimistic search* (aOS) that improves on the naive version to tackle unbalanced cases with the change point being close to the boundary. The main idea is to check a preliminary set of dyadic locations to localize the change point approximately and then apply OS in a suitable (balanced) neighborhood around the preliminary estimate in order to achieve a better localization. The preliminary estimate and the two locations marking its neighborhood are chosen from the dyadic locations, namely, as the location of the biggest gain as well as the closest dyadic neighbors thereof (to the left and to the right). Intuitively, as the dyadic points are denser on the boundaries, the advanced search is suitable even in very unbalanced scenarios where the naive version fails. From a theoretical perspective, this modification leads to minimax optimality.

Theorem 2 (Advanced optimistic search) *Under Model 1 with a single change point, i.e. $\kappa = 1$, assume that the minimal segment length λ and the minimal jump size δ satisfy*

$$\delta\sqrt{\lambda n} \geq C_0\sqrt{\log\log n} \qquad (4)$$

for some large enough constant C_0. Let $\hat{\tau} = \hat{t}_{(0,n]}/n$ be the estimated change point by aOS (Algorithm 2) on $(0, n]$. Then:

$$\lim_{n\to\infty} \mathbb{P}\left\{|\hat{\tau} - \tau| \leq C_1 \frac{\log\log n}{\delta^2 n}\right\} = 1 \qquad \text{with some constant } C_1.$$

Algorithm 2 Advanced Optimistic Search (aOS)

Require: $r - l > 2$; $l, r \in \mathbb{N}$ and step size $\nu \in (0, 1)$ with $1/2$ by default

1: **function** aOS(ν, l, r)
2:      $k \leftarrow \lfloor \log_2((r-l)/2) \rfloor$
3:      $\mathcal{D} \leftarrow \left\{ \lfloor l + 2^{-i}(r-l) \rfloor, \lceil r - 2^{-i}(r-l) \rceil : i = 1, \ldots, k \right\}$ ▷ Dyadic locations
4:      $t_* \leftarrow \underset{t \in \mathcal{D}}{\arg \max}\, G_{(l,r]}(t)$ ▷ Find the best split point on the "dyadic" grid
5:      **if** $t_* \leq (r+l)/2$ **then**
6:          $\tilde{l} \leftarrow \lfloor t_* - (t_* - l)/2 \rfloor$ and $\tilde{r} \leftarrow \lceil t_* + (t_* - l) \rceil$ ▷ "Dyadic neighbours" of t_*
7:      **else**
8:          $\tilde{l} \leftarrow \lfloor t_* - (r - t_*) \rfloor$ and $\tilde{r} \leftarrow \lceil t_* + (r - t_*)/2 \rceil$ ▷ "Dyadic neighbours" of t_*
9:      $\hat{t}_{(l,r]} \leftarrow \text{OS}(\tilde{l}, t_*, \tilde{r} \mid \nu, l, r)$ ▷ Naive optimistic search on $(\tilde{l}, \tilde{r}]$ containing t_*
10:     **return** $\hat{t}_{(l,r]}$

Similar to OS (Theorem 1), it is shown in Theorem 2 that aOS is able to localize the only change point at the best possible rate up to a log factor, but now under a much weaker condition (4) instead. In comparison with the weakest condition $\delta\sqrt{\lambda n} \gtrsim \sqrt{\log \log n}$ in Liu et al. (2021), we lose nothing except for a possibly larger multiplying constant. Therefore, aOS possesses the (nearly) statistical minimax optimality like the full grid search, which checks every possible split point in $\{1, \ldots, n\}$. Note that aOS (and OS) only requires $O(\log n)$ evaluations of the gain function (Lemma 3 later), in sharp contrast to $O(n)$ required by the full grid search. It is a surprising fact that computational speedups come at almost no cost of statistical performance at all. In this sense, "free lunch" is possible!

The idea of preliminary check of dyadic locations dated back to Rufibach and Walther (2010) (or even earlier to wavelets) and was recently explored in Liu et al. (2021) and Kovács et al. (2023). In practice, variants of aOS/OS might be equally viable e.g. the combination of OS and aOS, referred to as *combined OS,* see Appendix A for details.

Lemma 3 *OS and aOS (Algorithms 1 and 2) terminate in $O\big(\log(r-l)\big)$ and thus at most $O(\log n)$ steps (i.e. the number of gain function evaluations).*

For the univariate Gaussian setting, the overall computational cost is only $O(\log n)$ if cumulative sums have been pre-computed, as in that case each evaluation is possible in $O(1)$ time. Otherwise the $O(n)$ cost of calculating the cumulative sums becomes dominant. We remark that availability of cumulative sums (or similar "sufficient statistics" for the evaluation of the gain function) is a practical recommendation to store data for off- and online change point problems.

4. Methodology and Theory for Multiple Change Points

We consider now the setup of multiple change points, and investigate how our methodology can be extended to such a more ambitious setup in order to still have a sublinear number of evaluations of the gain function and yet with theoretical optimality guarantees for the estimation performance. Obviously, the extension of OS to multiple change points is not straight-forward. Here we adopt the idea of Seeded Binary Segmentation (SeedBS, Kovács

et al., 2023), which searches for a single change point in various intervals with the hope that some of these intervals contain only a single change point, where the detection is "easy". While the best split point in each interval is a candidate, the decision which candidates to declare finally as change points depends on a subsequent selection step. The intervals are called seeded intervals and they are constructed deterministically (Definition 4 below). SeedBS is thus very similar to wild binary segmentation (WBS, Fryzlewicz, 2014) and the narrowest over threshold method (Baranowski et al., 2019). The latter two procedures use random intervals instead of the deterministic ones, where the former in general leads to total length and number of considered intervals to be larger and thus computationally more expensive.

Definition 4 (Seeded intervals; Kovács et al., 2023) *Let $a \in [1/2, 1)$ be a given decay parameter. Let $I_1 = (0, n]$. For $k = 2, \ldots, \lceil \log_{1/a}(n) \rceil$ (logarithm with base $1/a$) define the k-th layer $\mathcal{I}_k$ as the set of n_k intervals of length l_k that are evenly shifted by s_k:*

$$\mathcal{I}_k = \bigcup_{i=1}^{n_k} \{ (\lfloor (i-1)s_k \rfloor, \lceil (i-1)s_k + l_k \rceil] \},$$

with $n_k = 2 \lceil (1/a)^{k-1} \rceil - 1$, $l_k = na^{k-1}$, $s_k = (n - l_k)/(n_k - 1)$. The set of seeded intervals is

$$\mathcal{I} = \bigcup_{k=1}^{\lceil \log_{1/a}(n) \rceil} \mathcal{I}_k.$$

Note that $\mathcal{I}$ covers the whole range of scales and locations in an efficient way such that there are $O(n)$ intervals, which is constructed to guarantee appropriate background for different types of change points. When there is only one change point, all intervals that do not have a starting point at 1 or do not have an end point at n can be discarded, reducing the number of intervals to $O(\log n)$. In the case of multiple change points, assuming a certain minimal spacing between change points allows to discard intervals that are too short.

Algorithm 3 Optimistic Seeded Binary Segmentation (OSeedBS)

Require: a decay parameter $a \in [1/2, 1)$, a minimal segment length $m \geq 2$, and tuning parameters for the selected optimistic search (OS or aOS)

1: **function** OSEEDBS
2:     $\mathcal{I} \leftarrow$ seeded intervals with decay a and at least m observations.
3:     **for** $(l, r] \in \mathcal{I}$ **do**
4:         $\hat{t}_{(l,r]} \leftarrow$ the split point returned by the optimistic search on $(l, r]$.
5:     Apply some selection rule to $\left(\hat{t}_{(l,r]}, G_{(l,r]}(\hat{t}_{(l,r]}) \right), (l, r] \in \mathcal{I}$, to output the change point estimates $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{\kappa}}$.
6:     Post-process: refine the estimated change points by applying the optimistic search to intervals $\left((\hat{\tau}_{i-1} + \hat{\tau}_i)/2, \ (\hat{\tau}_i + \hat{\tau}_{i+1})/2 \right]$, $i = 1, \ldots, \kappa$.

The difference between SeedBS and its optimistic counterpart OSeedBS is essentially in line 4 of Algorithm 3, where we perform either OS or aOS rather than full grid search. The selection method in line 5 can be for example greedy or narrowest-over-threshold (NOT)

selection, see Kovács et al. (2023). The computational times of OSeedBS depend critically on the minimal segment length m. If $m = O(n)$, only a handful of intervals are considered, with $O(\log n)$ evaluations each, and thus $O(\log n)$ evaluations overall. For the other extreme, when m is very small, many intervals need to be generated and thus the main driver of the number of evaluations (and hence, the computational cost) is the number of considered intervals. For $m = 2$, the number of intervals and also the total number of evaluations is $O(n)$. The estimation performance of course also depends on the choice of m. If chosen too large, estimation performance will be bad as change points within short segments may not be detected. Thus, m offers some kind of trade-off between estimation performance and computational efforts. Such trade-offs are inherent also in other methods, see e.g. the number of random intervals chosen in WBS.

Theorem 5 *Under Model 1, we assume that the minimal segment length λ and the minimal jump size δ satisfy*

$$\delta\sqrt{\lambda n} \geq C_0\sqrt{\log n} \tag{5}$$

for some large enough constant C_0. Assume further that there is an a-priori known lower bound λ_* of all segment lengths, i.e.

$$\lambda \geq \lambda_* \asymp n^{-\omega} \qquad \text{for some constant } \omega \in [0, 1]. \tag{6}$$

By $\hat{\kappa}$ and $\hat{\tau}_1 < \cdots < \hat{\tau}_{\hat{\kappa}}$ denote respectively the number and the locations of estimated change points by OSeedBS (Algorithm 3), with the NOT selection method, and the seeded intervals of lengths larger than $m = \lfloor \lambda_* n/3 \rfloor$. Then:

i. *There exist constants C_1, C_2, independent of n, ω, a and m, such that, given the threshold for the selection method $\gamma = C_1\sqrt{\log n}$,*

$$\lim_{n\to\infty} \mathbb{P}\left\{\hat{\kappa} = \kappa, \ \max_{i=1,\dots,\kappa} \delta_i^2 |\hat{\tau}_i - \tau_i| \leq C_2 \frac{\log n}{n}\right\} = 1.$$

ii. *The number of evaluations is $O(\min\{n^\omega \log n, n\})$.*

We emphasize that the assumption (6) is only needed for computational efficiency ensuring a sublinear number of evaluations as specified in part ii of Theorem 5. If the data are stored in the format of cumulative sums, then the overall computational cost itself is also $O(\min\{n^\omega \log n, n\})$, i.e. it equals the number of evaluations. However, if cumulative sums are not available, then the $O(n)$ cost of calculating cumulative sums becomes dominant and the overall computational cost is $O(n)$, see Appendix F. For the statistical guarantee in part i, an assumption on the minimal spacing, i.e. (6), becomes obvious if we choose $\omega = 1$ and $m = 2$, since it is pointless to work on a higher resolution than the sampling rate $1/n$ without further model assumption, and thus in this sense it imposes no restriction at all. In case of multiple change points, the signal strength condition (5) is the weakest one that still allows for detection. It coincides with the best known results (e.g. Frick et al., 2014; Baranowski et al., 2019; Hu et al., 2023; Verzelen et al., 2023) with the only difference in multiplying constants. Following the proof in Appendix E.3, we can easily replace it by

$$\min_{i=1,\dots,\kappa} \left(\min\{\tau_{i+1} - \tau_i, \tau_i - \tau_{i-1}\}\delta_i^2 n\right) \gtrsim \log n \qquad \text{as} \quad n \to \infty.$$

This is slightly more general, as it allows for frequent large jumps and rare small jumps over long segments (cf. Cho and Kirch, 2022). However, we prefer the current version as in Theorem 5, for notational simplicity. Note, moreover, that the localization rate reported in part i of Theorem 5 is minimax optimal up to a possible log factor. In the particular case of $\kappa \asymp n^\theta$ with some constant $\theta > 0$, the derived localization rate is indeed optimal (namely, the log factor being necessary, see Verzelen et al., 2023).

WBS (Fryzlewicz, 2014), and the similar narrowest over threshold method (Baranowski et al., 2019), can also be sped up using OS (or aOS). However, in the worst case with very short segments, e.g. in frequent change point scenarios with up to $O(n)$ change points, these two methods need to draw up to $O(n^2)$ random intervals, which prohibits sublinear number of evaluations overall. Nonetheless, we expect substantial computational gains using OS (or aOS) in connection with many other multiple change point detection techniques compared to the respective full grid search based counterparts.

5. Extension to Multivariate and High-dimensional Scenarios

In the previous sections, the theoretical findings on the univariate Gaussian changing means (Model 1) reveal the *statistical insight* that the computational efficiency can be improved by almost one order (more precisely, from $O(n)$ to $O(\log n)$ evaluations) with nearly no loss of statistical efficiency, using the optimistic search strategy. We show that this is also true for Gaussian changing means problems of general and potentially high dimension.

5.1 The Multivariate Model and Some Technical Simplification

The following model and variants thereof find applications e.g. in genomics, particularly in analyzing copy number variations (cf. Zhang et al., 2010; Picard et al., 2011; Behr et al., 2018; Mallory et al., 2020).

Model 2 (Gaussian changing means) *Assume that the observation vectors* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^p$ *are independent and*

$$\boldsymbol{X}_{\tau_0 n+1}(= \boldsymbol{X}_1), \ldots, \boldsymbol{X}_{\tau_1 n} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{I}_p),$$

$$\vdots$$

$$\boldsymbol{X}_{\tau_\kappa n+1}, \ldots, \boldsymbol{X}_{\tau_{\kappa+1} n}(= \boldsymbol{X}_n) \sim \mathcal{N}(\boldsymbol{\mu}_\kappa, \boldsymbol{I}_p),$$

where $\{\tau_i : i = 1, \ldots, \kappa\}$ *is the points satisfying*

$$0 = \tau_0 < \tau_1 < \cdots < \tau_{\kappa+1} = 1 \quad and \quad \tau_i n \in \mathbb{N},$$

mean vectors $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_{i-1} \in \mathbb{R}^p$, $i = 1, \ldots, \kappa$ *are the levels on segments, and the covariance matrix is* $\boldsymbol{I}_p \in \mathbb{R}^{p \times p}$ *the identity matrix. Define the minimal segment length*

$$\lambda \equiv \lambda_n = \min_{i=0,\ldots,\kappa} (\tau_{i+1} - \tau_i),$$

and the minimal jump size

$$\delta \equiv \delta_n = \min_{i=1,\ldots,\kappa} \delta_i \quad with \quad \delta_i = \left\| \boldsymbol{\mu}_i - \boldsymbol{\mu}_{i-1} \right\|,$$

13

with $\|\cdot\|$ the Euclidean norm. In addition, assume that there is a known *integer* $s \in \{1, \ldots, p\}$ *such that, for* $i = 0, \ldots, \kappa,$

$$\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i-1}\|_0 := \#\{j = 1, \ldots, p \mid \mu_{i,j} \neq \mu_{i-1,j}\} \leq s,$$

with $\mu_{i,j}$ *the* j-*th entry of* $\boldsymbol{\mu}_i$.

In Model 2 the locations of change points are shared over p coordinates, and thus it potentially allows aggregation of detection power among different coordinates. In case that the change of means happens in only a sparse fraction of coordinates (e.g. $s \ll p$), one should focus only on the coordinates where the mean changes. The selection of changing coordinates can be achieved by a simple thresholding rule, see Liu et al. (2021). This motivates us to consider the following gain function:

$$G^o_{(l,r]}(t) := \sum_{j=1}^{p} \big(\mathrm{CS}_{(l,r],j}(t;\boldsymbol{X})^2 - \alpha^2\big)_+, \tag{7}$$

where $0 \leq l < t < r \leq n$ are integers, $\alpha \geq 0$ is a user-specified threshold, and $\mathrm{CS}_{(l,r],j}(t;\boldsymbol{X})$ is the CUSUM statistics in the j-th coordinate of $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, namely,

$$\mathrm{CS}_{(l,r],j}(t;\boldsymbol{X}) = \sqrt{\frac{r-t}{(r-l)(t-l)}} \sum_{i=l+1}^{t} X_{i,j} - \sqrt{\frac{t-l}{(r-l)(r-t)}} \sum_{i=t+1}^{r} X_{i,j},$$

with $X_{i,j}$ the j-th entry of $\boldsymbol{X}_i$.

In the gain function (7) the CUSUM statistics is utilized for change point estimation as well as for coordinate selection. This entanglement of change point estimation and coordinate selection complicates the theoretical analysis. We employ *two technical modifications* to ease the theoretical analysis.

The first is a sample splitting trick that removes the aforementioned entanglement. We split the samples from Model 2 into two independent groups, with one group at odd times, and the other at even times. One group of samples is then used for the estimation of change points, and the other for the selection of coordinates. For simplicity, we assume that there are two independent copies of samples, denoted as $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ and $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$, from Model 2. Then the modified gain function is defined as

$$G_{(l,r]}(t) := \sum_{j=1}^{p} \big(\mathrm{CS}_{(l,r],j}(t;\boldsymbol{X})^2 - 1\big) \mathbb{1}\left\{\big|\mathrm{CS}_{(l,r],j}(t;\boldsymbol{Y})\big| \geq \alpha\right\}, \tag{8}$$

with a threshold $\alpha \geq 0$ and integers $0 \leq l < t < r \leq n$.

Recall that the basic operation in optimistic searches is the comparison between a pair of locations to determine which one is more likely to be a change point. Such a comparison is done via the absolute scores determined by the gain function, but it is also feasible whenever relative scores are available. Thus, as the second modification, we introduce a relative score between two locations t and w in the form of a *comparison function*

$$\mathcal{C}_{(l,r]}(t,w) := \sum_{j=1}^{p} \Bigg(\Big(\mathrm{CS}_{(l,r],j}(t;\boldsymbol{X})^2 - \mathrm{CS}_{(l,r],j}(w;\boldsymbol{X})^2\Big) \times$$

$$\mathbb{1}\Big\{\max\big\{\big|\mathrm{CS}_{(l,r],j}(t;\boldsymbol{Y})\big|, \big|\mathrm{CS}_{(l,r],j}(w;\boldsymbol{Y})\big|\big\} \geq \alpha\Big\} \Bigg). \tag{9}$$

The location t is preferred as a change point candidate rather than w, if and only if $\mathcal{C}_{(l,r]}(t, w) \geq 0$. This second modification means that in OS and aOS (Algorithms 1 and 2) we use the comparison function in (9) instead of the gain function in (8) to decide which location is preferred. Besides, in the dyadic search in aOS (Algorithm 2, line 4), the maximum of the gain function should be replaced by the location at which it is preferable in terms of comparison function over all other dyadic locations. However, for a found change point candidate by OS or aOS, the gain function in (8) is still used to decide whether it should be selected as an estimated change point, in OSeedBS (Algorithm 3, line 5).

These two modifications are needed only when $\alpha > 0$. In case of $\alpha = 0$, the modified gain function (8) is the same as the original gain (7), and the comparison function (9) is simply the difference of the original gain function (7) at two locations.

5.2 Statistical Guarantees

Given the above two technical modifications, we can establish the following statistical properties of the optimistic searches.

Theorem 6 (Single change point) *Under Model 2 with $\kappa = 1$ (a single change point), assume that the minimal segment length λ and the minimal jump size δ satisfy*

$$n\lambda\delta^2 \geq C_0\rho_\circ(n, p, s), \tag{10a}$$

where $C_0 > 0$ is a sufficiently large constant, and

$$\rho_\circ(n, p, s) := \begin{cases} \sqrt{p \log \log n} & \text{if } s \geq \sqrt{p \log \log n}, \\ \max\left(s \log \frac{e\sqrt{p \log \log n}}{s}, \log \log n\right) & \text{if } s \leq \sqrt{p \log \log n}. \end{cases} \tag{10b}$$

Let $\hat{\tau} = \hat{t}_{(0,n]}/n$ be the estimated change point by aOS (Algorithm 2) on $(0, n]$, with the gain (8), the comparison function (9) and

$$\alpha \equiv \alpha(n, p, s) = \begin{cases} 0 & \text{if } s \geq \sqrt{p \log \log n} \text{ or } s = p, \\ \sqrt{2 \log \frac{e^2 p \log \log n}{s^2}} & \text{otherwise.} \end{cases}$$

Then, for some constant $C_1 > 0$, it holds that

$$\lim_{n \to \infty} \mathbb{P}\left\{|\hat{\tau} - \tau| \leq C_1 \max\left\{\frac{\log \log n}{\delta^2 n}, \frac{\min\{s^2, p \log \log n\}}{n^2 \lambda \delta^4}\right\}\right\} = 1.$$

Theorem 6 includes Theorem 2 as a special case when $p = 1$, with the threshold $\alpha = 0$. In general, the condition (10) is the weakest possible for the detection of a single change point (Liu et al., 2021). For the localization of the change point, since $\rho_\circ(n, p, s) \geq \sqrt{\min\{s^2, p \log \log n\}}$, Theorem 6 implies

$$\mathbb{P}\left\{|\hat{\tau} - \tau| \leq C_1 \frac{\rho_\circ(n, p, s)}{n\delta^2}\right\} \to 1, \quad \text{as } n \to \infty. \tag{11}$$

This is the induced localization rate from the detection of change points, as reported in Pilliat et al. (2023). Intuitively, the induced rate can be obtained since one may treat the

localization of a single change point up to an accuracy ε_n "equivalently" as the detection of a single change point at $t = n\varepsilon_n$ with the same jump size. This connection between detection and localization of change points leads to almost sharp localization rates in univariate and multivariate cases, but may yield suboptimal rates in case of high-dimension. In fact, if $p \gg \log\log n$, the localization rate in Theorem 6 is strictly faster than the rate in (11), in the dense scenario ($s \geq \sqrt{p\log\log n}$) when $n\lambda\delta^2 \gg \sqrt{p\log\log n}$, and in the sparse scenario ($s \leq \sqrt{p\log\log n}$) when $s \gtrsim \log\log n$.

Moreover, Theorem 6 together with the condition

$$n\lambda\delta^2 \gtrsim \min\Big\{\frac{s^2}{\log\log n}, p\Big\}, \tag{12}$$

leads to

$$\mathbb{P}\Big\{|\hat{\tau} - \tau| \lesssim \frac{\log\log n}{n\delta^2}\Big\} \to 1, \quad \text{as } n \to \infty,$$

which is not improvable except for the factor of $\log\log n$ (Wang and Samworth, 2018, Proposition 3). In the literature, stricter conditions than (12) are required for the same localization rate (ignoring the log factor), see e.g. Bhattacharjee et al. (2017), Kaul et al. (2021) and Li et al. (2023). We stress that in the low dimensional case of $p \lesssim \log\log n$, or in the highly sparse case of $s \lesssim \log\log n$, the condition (12) is simply a consequence of (10), and thus the localization rate in Theorem 6 is minimax optimal (up to a $\log\log n$ factor). However, in general, it remains unclear, whether the localization rate in Theorem 6 is optimal or not, under the weakest detection condition (10), see Section 7.

The optimistic searches can be applied to the inference of multiple change points, if one incorporates the idea of SeedBS, which results in OSeedBS (Algorithm 3), as in the univariate setup of Model 1. In particular, because of multiscale nature of seeded intervals, OSeedBS extends the statistical optimality of optimistic searches for a single change point to the general case of multiple change points.

Theorem 7 (Multiple change points) *Under Model 2, we assume that the minimal segment length λ and the minimal jump size δ satisfy*

$$n\lambda\delta^2 \geq C_0\rho(n, p, s), \tag{13a}$$

where $C_0 > 0$ is a sufficiently large constant, and

$$\rho(n, p, s) := \begin{cases} \sqrt{p\log n} & \text{if } s \geq \sqrt{p\log n}, \\ \max\Big(s\log\frac{e\sqrt{p\log n}}{s}, \log n\Big) & \text{if } s \leq \sqrt{p\log n}. \end{cases} \tag{13b}$$

By $\hat{\kappa}$ and $\hat{\tau}_1 < \cdots < \hat{\tau}_{\hat{\kappa}}$ denote respectively the number and the locations of estimated change points by OSeedBS (Algorithm 3) with the NOT selection. Set the threshold α in the gain (8) and the comparison function (9) as

$$\alpha \equiv \alpha(n, p, s) = \begin{cases} 0 & \text{if } s \geq \sqrt{p\log n} \text{ or } s = p, \\ \sqrt{2\log\frac{e^2 p\log n}{s^2}} & \text{otherwise,} \end{cases}$$

and the selection threshold γ in NOT as $\gamma = C_1\rho(n,p,s)$ for some constant $C_1 > 0$. Then, there exists a constant C_2, such that, as $n \to \infty$,

$$\mathbb{P}\left\{\hat{\kappa} = \kappa, \ |\hat{\tau}_i - \tau_i| \leq C_2 \max\left\{\frac{\log n}{\delta_i^2 n}, \frac{\min\{s^2, p\log n\}}{n^2\lambda\delta_i^4}\right\}, \ i = 1,\ldots,\kappa\right\} \to 1.$$

It is clear from the proof (Section E.3) that if all segment lengths are larger than λ_*, then Theorem 7 remains valid even if the minimal length m of seeded intervals is chosen as $m = \lfloor \lambda_* n/3 \rfloor$. As a consequence, it covers part i of Theorem 5 as a special case of $p = 1$. Such a-priori knowledge of λ_* will lead to computational speedups, see Section 5.3 later.

Since $\rho(n,p,s) \geq \sqrt{\min\{s^2, p\log n\}}$, the localization rate of Theorem 7 implies the induced rate from the detection of change points, namely,

$$\mathbb{P}\left\{\hat{\kappa} = \kappa, \ |\hat{\tau}_i - \tau_i| \leq C_2 \frac{\rho(n,p,s)}{n\delta_i^2}, \ i = 1,\ldots,\kappa\right\} \to 1, \quad \text{as } n \to \infty, \tag{14}$$

which was given in Pilliat et al. (2023). Similar to the single change point case, the localization rate in Theorem 7 can be strictly faster than the induced rate in (14). For instance, in the high-dimensional setup of $p \gg \log n$, this occurs in the dense scenario ($s \geq \sqrt{p\log n}$) when $n\lambda_i\delta_i^2 \gg \sqrt{p\log n}$, and in the sparse scenario ($s \leq \sqrt{p\log n}$) when $s \gtrsim \log n$.

The condition (13) is minimax optimal in detection of two or more change points, while the minimax optimality of localization rates remains unclear. An exception is the case of $n\lambda_i\delta_i^2 \gtrsim \min\{s^2/\log n, p\}$, the localization rate of Theorem 7 is of order $(\log n)/(\delta_i^2 n)$ and thus not improvable except a possible log factor. In high-dimension, the optimal localization rates remain yet unknown, similar to the case of a single change point, see Section 7.

Inspecting the proof of Theorem 7 (Appendix E.3), we can show that with no post-processing (line 6, Algorithm 3), the localization error rate is

$$\varepsilon_i \asymp \max\left\{\frac{\log n}{n\delta_i^2}, \ \frac{\min\{s^2, p\log n\}}{n\delta_i^2\gamma}\right\},$$

provided that the selection threshold γ satisfies

$$\rho(n,p,s) \lesssim \gamma \lesssim n\lambda\delta^2.$$

Then, if $\gamma \asymp n\lambda\delta_i^2$, the same localization rate as in Theorem 7 can be achieved, but this is not practical or feasible, as $n\lambda\delta_i^2$ is often unknown, and may be larger than $n\lambda\delta^2$. Thus, if

$$\frac{\log n}{n\delta_i^2} \ll \frac{\min\{s^2, p\log n\}}{n\lambda\delta_i^4},$$

the post-processing is necessary for the faster rate in Theorem 7. Otherwise, e.g. in the univariate case, we can drop the post-processing step for the sake of saving computation.

The tuning parameters of OSeedBS are the selection threshold γ and the threshold α, which can be chosen according to Theorem 7. One restriction in the sparse scenario is that the level of sparsity s is required (which appears in the threshold α). One may adjust the gain function by considering a proper selection of guesses on s, e.g.

$$\{1, 2, \ldots, 2^{\lceil \log_2 \sqrt{p\log n} \rceil}\} \cup \{p\}$$

17

and aggregating over such a selection, in a similar way as in Liu et al. (2021). The careful investigation is left as future research. In addition, data-driven rules for selecting γ and α can be beneficial, especially for improving finite sample performances. We expect that similar approaches (e.g. various information criteria, minimum description lengths, or heuristic methods) developed for full grid search methods would also apply to the optimistic searches.

We emphasize that the proofs of the statistical guarantees (Appendix E) are organized in a modular fashion, which handle the random perturbations in the data (Appendix E.1, particularly, Lemmas 15 and 18) separately from the specific structure of optimistic searches (Appendices E.2 and E.3). This modular approach allows the application of our proof strategies to general scenarios beyond the specific Gaussian model (Model 2). For instance, we might extend our results to general non-Gaussian and possibly temporally dependent situations, by leveraging the framework of functional dependence (introduced in Wu, 2005). To achieve this, we only need to extend Lemmas 15 and 18 to such situations, which seems feasible based on the results in Zhang and Wu (2017, Lemma C.4) and Zhang and Wu (2021, Theorem 6.6). Moreover, we might introduce spatial dependence into the Gaussian model by considering a general covariance $\boldsymbol{\Sigma}$ instead of the identity matrix $\boldsymbol{I}_p$. In the dense scenario, this modification would request to adjust the selection threshold γ according to the trace, operator norm and Frobenius norm of $\boldsymbol{\Sigma}$, which can be estimated using the sample covariance matrix (see Section 3 in Liu et al., 2021). The sparse scenario is more intricate, as various sparsity structures, such as the sparsity of $\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i-1}$, that of $\boldsymbol{\Sigma}$ and/or that of $\boldsymbol{\Sigma}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i-1})$, may come into play. Further investigation of these extensions, along the aforementioned lines of discussion, presents interesting avenues for future research.

5.3 Computation Complexity

The number of gain function evaluations remains the same as in the univariate case (Lemma 3 and Theorem 5ii). If the data is stored in cumulative sums, each evaluation of gain function involves $O(p)$ computations, which leads to an additional factor of p in run time; Otherwise, each evaluation of gain function will cost $O(np)$ computations, in which case the calculation of cumulative sums is recommended as a preprocessing of the data, which requires $O(np)$ computations only once.

Proposition 8 *Assume that the data from Model 2 is stored in the format of cumulative sums, that is, $S_{t,j} := \sum_{i=1}^{t} X_{i,j}$ with $t = 1, \ldots, n$ and $j = 1, \ldots, p$. Then:*

i. *In case of a single change point, the optimistic searches (Algorithms 1 and 2) require $O(p \log n)$ computations in the worst case.*

ii. *In case of multiple change points, assume further that*

$$\lambda \geq \lambda_* \asymp n^{-\omega} \qquad \text{for some constant } \omega \in [0, 1].$$

OSeedBS (Algorithm 3) with $m = \lfloor \lambda_* n / 3 \rfloor$ requires $O(p \min\{n^\omega \log n, n\})$ computations in the worst case.

Clearly, Proposition 8 includes Lemma 3 and part ii of Theorem 5 as a special case of $p = 1$. In comparison, the full grid search has $O(pn)$ run time for a single change point,

and SeedBS has $O(pn \log n)$ run time for multiple change points. Thus, the computational speedup based on optimistic searches can be polynomial in sample size.

Moreover, we stress that OSeedBS can be sped up by a slight modification of the procedure as follows. On every seeded interval, we first select s coordinates such that their corresponding squared CUSUM statistics evaluated at the middle point of this interval are the largest among all coordinates. Afterwards, we only consider the gain function restricted to the selected s coordinates. With this modification, OSeedBS will have an improved worst case run time $O(s \min\{n^\omega \log n, \, n\} + pn^\omega)$, and will still enjoy the same statistical guarantee of OSeedBS established in Theorem 7 (Appendix E.3).

6. Simulations

We provide a simulation study of our optimistic searches (incl. combined OS in Appendix A) on univariate changing means and high-dimensional changing covariance problems.

6.1 Single Change Point in Univariate Gaussian Means

Example 1 *Let $X_1, \ldots, X_{100} \sim \mathcal{N}(0, \sigma^2)$ and $X_{101}, \ldots, X_{100+n} \sim \mathcal{N}(0.5, \sigma^2)$ be independent observations with a single change in the mean value at observation* 100.

Simulation results are reported in Table 2 and Figure 5 in Appendix B. While OS clearly struggles when the lengths of the two segments are very unbalanced (n large, in particular with high noise level), aOS has a much better performance. However, for the more balanced scenarios (up to $n = 400$ for $\sigma = 1.5$ for example), OS performs well. The combined OS (Appendix A) has a slightly improved performance compared to aOS (in particular for the rather balanced scenarios). The full grid search has the best performance for the rather challenging scenarios that are very unbalanced and/or have a high noise level, but aOS and combined OS come remarkably close. We note that increasing absolute errors in change point location for higher values of n despite having more available observations is actually reasonable, as there are meanwhile more potential candidates for change points on the grid. This is also compatible with the theoretical bound of order $\log \log(n)/\delta^2$.

6.2 Multiple Change Points in Univariate Gaussian Means

Example 2 below describes the blocks signal (Donoho and Johnstone, 1994) with the noise level as used in Fryzlewicz (2014).

Example 2 *Consider a total of* 2048 *observations with* 11 *change points at locations* 205, 267, 308, 472, 512, 820, 902, 1332, 1557, 1598 *and* 1659 *as well as mean values* 0, 14.64, -3.66, 7.32, -7.32, 10.98, -4.39, 3.29, 19.03, 7.68, 15.37 *and* 0 *between the change points to which independent Gaussian noise with a standard deviation of $\sigma = 10$ is added.*

The results are given in Figure 3. The average performances of OS and combined OS are very close to that of the full search. Overall, it turns out that the optimistic variants of OSeedBS have a competitive performance compared to the full grid search based SeedBS. Further, as long as the minimal segment length constraints are short enough to guarantee coverage of each single change point, both SeedBS and variants of OSeedBS perform well.
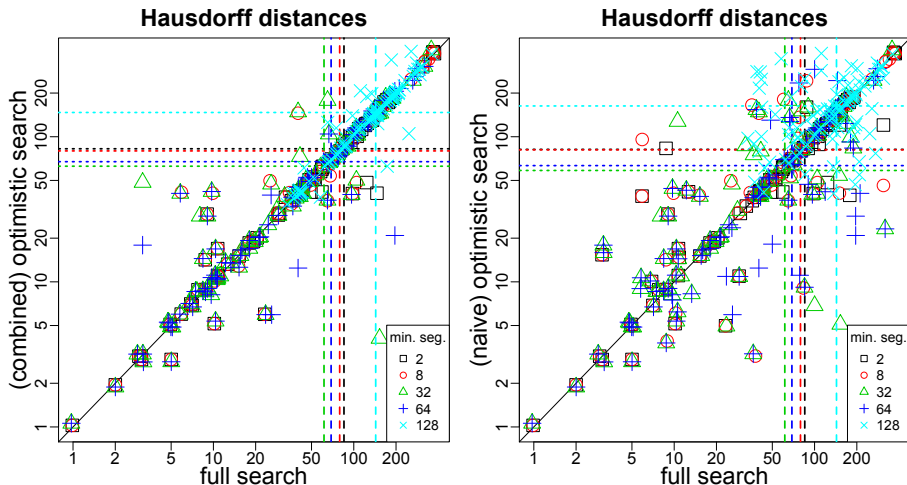
Figure 3: Results on Example 2. Pairwise plots of Hausdorff distances of the locations of the best 11 change point candidates (with greedy selection) compared to the true ones in 100 simulations for SeedBS (decay $a = 1/\sqrt{2}$) with various minimal segment length constraints and full grid search in each seeded interval (horizontal-axis) versus combined OS (vertical-axis, left) and OS (vertical-axis, right). The vertical and horizontal dashed lines indicate the average Hausdorff distances for the respective minimal segment length constraint and search method within the seeded intervals. Note the logarithmic scales on both axes.

6.3 High-dimensional Gaussian Covariance Changes

As an exploration on the potential of optimistic searches, we introduce a changing covariance setup in Model 3 with specific instances for simulations in Example 3.

Model 3 *Assume that observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent and*

$$\boldsymbol{X}_{\tau_0 n+1}(= \boldsymbol{X}_1), \ldots, \boldsymbol{X}_{\tau_1 n} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_0),$$

$$\vdots$$

$$\boldsymbol{X}_{\tau_\kappa n+1}, \ldots, \boldsymbol{X}_{\tau_{\kappa+1} n}(= \boldsymbol{X}_n) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\kappa),$$

where $\{\tau_i : i = 1, \ldots, \kappa\}$ gives the locations of change points satisfying

$$0 = \tau_0 < \tau_1 < \cdots < \tau_{\kappa+1} = 1 \quad and \quad \tau_i n \in \mathbb{N},$$

and $\boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_{i-1} \in \mathbb{R}^{p \times p}$, $i = 1, \ldots, \kappa$ are the covariances on the segments.

A simulation setup of Model 3 was considered in Kovács et al. (2023) as an example to demonstrate the computational efficiency of seeded intervals over random intervals utilized in WBS. We will show that further speedups in such computationally challenging setup for many available algorithms can be easily obtained utilizing our optimistic search strategies.

Example 3 *Let* $\Sigma_{ij} = \exp\left(-\frac{1}{2}|t_i - t_j|\right)$ *with* $t_i - t_{i-1} = 0.75, i = 2, \ldots, 20,$ *be a chain network model (Fan et al., 2009, Example 4.1) with* $p = 20$ *variables. A modified version* $\tilde{\Sigma}$ *is obtained by replacing the top left* 5×5 *block of* Σ *by a 5-dimensional identity matrix.*

(a) *As in Kovács et al. (2023), we set in Model 3* $\Sigma_0 = \Sigma$, $\Sigma_1 = \tilde{\Sigma}$, $\Sigma_2 = \Sigma$, $\Sigma_3 = \tilde{\Sigma}$, *etc. and draw* 100 *observations for each segment until obtaining a total of* $n = 2000$ *observations. Hence, there are* 20 *segments of length* 100 *each, i.e.* $\kappa = 19$.

(b) *In Model 3, we set* $\Sigma_0 = \Sigma_2 = \Sigma_4 = \Sigma$, $\Sigma_1 = \Sigma_3 = \Sigma_5 = \tilde{\Sigma}$, *and draw* 550, 300, 700, 250, 100 *and* 100 *observations for the respective segments, obtaining again a total of* $n = 2000$ *observations, but this time with* 6 *segments, i.e.* $\kappa = 5$.

We consider a gain function (defined in Appendix C) based on the multivariate Gaussian log-likelihood where the underlying precision matrices are obtained by the graphical lasso (Friedman et al., 2008; Giraud, 2022). The graphical lasso is rather costly especially when repeatedly fitting at each possible split point t on a grid. The essential problem is that the estimator $\hat{\Omega}_{(u,t]}^{\text{glasso}}$ of precision matrix for a segment $(u, t]$ cannot be efficiently updated (not even using warm starts) to obtain $\hat{\Omega}_{(u,t+1]}^{\text{glasso}}$ for the segment $(u, t + 1]$. Hence, the overall number of graphical lasso fits is the main driver of computational time.

This chosen gain function is motivated by the fact that its population version attains local maxima only at change points. More precisely, the population gain has the form of

$$G_{(l,r]}^{*}(t) = \frac{r - l}{n} \log(|\Sigma_{(l,r]}|) - \frac{t - l}{n} \log(|\Sigma_{(l,t]}|) - \frac{r - t}{n} \log(|\Sigma_{(t,r]}|), \tag{15}$$

where $\Sigma_{(l,r]}$ is the average covariance matrix on the segment $(l, r] \subseteq (0, 1]$,

$$\Sigma_{(l,r]} = \sum_{i=0}^{\kappa} \frac{|(\tau_i, \tau_{i+1}] \cap (l, r]|}{|(l, r]|} \Sigma_i \quad \text{with} \quad |(a, b]| = b - a,$$

and $|A|$ is the determinant of a matrix A.

Lemma 9 *The function* $G_{(l,r]}^{*}(\cdot)$ *defined in* (15) *is piecewise (between change points) convex, and up to the special cases as detailed in the proof (Appendix C.1) even strictly convex.*

Lemma 9 shows that in case of a single change point in $(l, r]$, $G_{(l,r]}^{*}$ is unimodal, and in case $(l, r]$ contains multiple change points, each strict local maximum is a change point.

We compare the estimation performance and computational times of various methods. In order to eliminate the effect of model selection, for all algorithms we selected greedily as many change points as the true underlying number (with some exceptions for WBS with small M). The results in Figure 4 show roughly speedups of factor 30 for OBS compared to BS, factor 35 for OWBS versus WBS and factor 10–14 for OSeedBS versus SeedBS in both of the considered setups. We provide a further discussion on the speedups of different approaches and potential benefits of combining them in Appendix D.
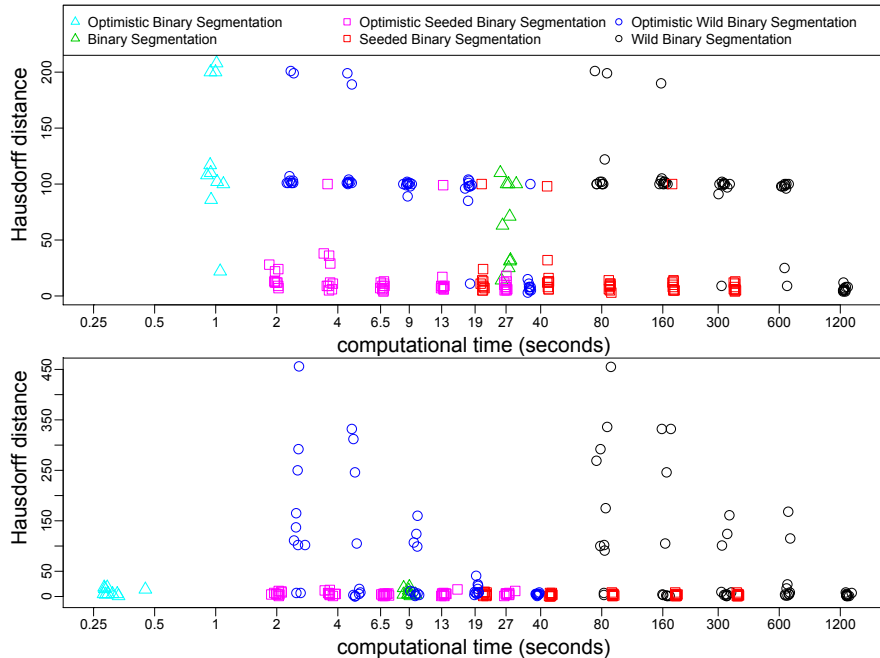
Figure 4: Estimation performances (in Hausdorff distance) and computational times on Example 3 (based on 10 simulations) with setups (a) on the top and (b) on the bottom. The symbols differentiate among the basic algorithms and the colors indicate whether full grid search or OS was used. The five point clouds for SeedBS and OSeedBS correspond to decay $a = 2^{-1}, 2^{-1/2}, 2^{-1/4}, 2^{-1/8}, 2^{-1/16}$ for the seeded intervals, while the five point clouds for WBS and OWBS correspond to $M = 100, 200, 400, 800, 1600$ random intervals. For all algorithms, the true number of change points (or the maximally many if not achieved) was used.

7. Discussion

We introduced optimistic search strategies that avoid the full grid search and thus lead to computationally fast change point detection methods in great generality. For univariate, multivariate and high-dimensional Gaussian changing means setups we proved that aOS is asymptotically minimax optimal for detecting a single change point with only a logarithmic number of evaluations of the gain function. For multiple change point problems we combined optimistic searches (OS and aOS) with seeded binary segmentation, leading to asymptotically minimax optimal detection while having superior runtime compared to existing approaches. In addition, the localization rate of change points is by far the sharpest, given the weakest possible condition on the signal-to-noise ratios. It is unclear though whether our localization rate is optimal or not in certain high-dimensional scenarios. In the literature, a faster localization rate is shown to be possible in certain regimes with much larger signal-to-noise ratio (Wang and Samworth, 2018, Theorems 1 and 2). In particular, it indicates that our localization rate is *not adaptively* minimax optimal over all possible ranges of signal-to-noise ratios. The complete understanding of localization rates is, to the

best of our knowledge, still open for high-dimensional Gaussian mean changes, which offers an interesting avenue for future research in this direction. Overall, our theoretical results reveal a surprising fact that the computational acceleration up to one order in sample size can be achieved (by optimistic searches) with nearly no loss of statistical efficiency.

Our methodology is also most relevant for complex change point detection problems with computationally expensive model fits, as demonstrated by the massive computational gains in examples involving high-dimensional graphical models.

Acknowledgments

Appendix A. Combined Optimistic Search

Combining the results of naive and advanced optimistic search (i.e. OS and aOS), thus referred to as the *combined optimistic search*, leads to slightly better empirical performance than the individual searches, but at a slightly higher computational cost, see Algorithm 4. Also from a theory point of view, the combined optimistic search enjoys the same statistical minimax optimality as the advanced version, see Remark 22 later in Appendix E.

Algorithm 4 Combined Optimistic Search

Require: $r - l > 2$; $l, r \in \mathbb{N}$ and step size $\nu \in (0, 1)$ with $1/2$ by default
1: **function** cOS(ν, l, r)
2:      $\hat{t}_0 \leftarrow$ aOS(ν, l, r) ▷ Advanced optimistic search
3:      $\hat{t}_1 \leftarrow$ OS$(l, \lfloor (l + \nu r)/(1 + \nu) \rfloor, r \mid \nu, l, r)$ ▷ Naive optimistic search
4:      **if** $G_{(l,r]}(\hat{t}_0) \geq G_{(l,r]}(\hat{t}_1)$ **then**
5:          $\hat{t}_{(l,r]} \leftarrow \hat{t}_0$
6:      **else**
7:          $\hat{t}_{(l,r]} \leftarrow \hat{t}_1$
8:      **return** $\hat{t}_{(l,r]}$

Appendix B. Additional Material on the Univariate Gaussian Simulations

Example 4 *Consider a specific example of Model 1 with $\kappa = 1$ and $\tau_1 = \lambda \leq 1/3$. For simplicity, let $\nu = 1/2$ and $n/3 \in \mathbb{N}$. In the first step of OS, we check the gain function at*

$n/3$ and $2n/3$. In order to avoid wrongly discarding $(0, n/3]$, we have to ensure

$$\left| \mathrm{CS}_{(0,n]}\left(\frac{n}{3}\right)\right| \geq \left| \mathrm{CS}_{(0,n]}\left(\frac{2n}{3}\right)\right|.$$

In fact, we have

$$\mathbb{P}\left\{\left| \mathrm{CS}_{(0,n]}\left(\frac{n}{3}\right)\right| < \mathrm{CS}_{(0,n]}\left(\frac{2n}{3}\right)\right\} \leq \mathbb{P}\left\{\left| \mathrm{CS}_{(0,n]}\left(\frac{n}{3}\right)\right| < \left| \mathrm{CS}_{(0,n]}\left(\frac{2n}{3}\right)\right|\right\}$$

$$\leq 2\,\mathbb{P}\left\{\left| \mathrm{CS}_{(0,n]}\left(\frac{n}{3}\right)\right| < \mathrm{CS}_{(0,n]}\left(\frac{2n}{3}\right)\right\}.$$

Elementary calculation using properties of the Gaussian distribution reveals

$$\mathbb{P}\left\{\left| \mathrm{CS}_{(0,n]}\left(\frac{n}{3}\right)\right| < \mathrm{CS}_{(0,n]}\left(\frac{2n}{3}\right)\right\} = \Phi\left(-\delta\lambda\sqrt{\frac{n}{2}}\right)\Phi\left(\delta\lambda\sqrt{\frac{3n}{2}}\right),$$

with Φ the standard Gaussian distribution function. Thus, if and only if $\delta\lambda\sqrt{n} \to \infty$,

$$\mathbb{P}\left\{\left| \mathrm{CS}_{(0,n]}\left(\frac{n}{3}\right)\right| \geq \left| \mathrm{CS}_{(0,n]}\left(\frac{2n}{3}\right)\right|\right\} \to 1 \qquad as \quad n \to \infty.$$

Note that $\delta\lambda\sqrt{n} \to \infty$ is, up to a log factor, equivalent to the condition (3), which guarantees that the probability of making a mistake in the first step of OS vanishes eventually.

Table 2 displays various results for Example 1 in the main text. The top part of Table 2 shows the localization error of the change point estimates found by the naive, advanced and combined optimistic search, as well as the full grid search for various choices of n (from 100 to 5000) for three different noise levels ($\sigma = 0.5, 1, 1.5$). The bottom part of Table 2 shows the number of evaluations as a measure of computation times.

Figure 5 shows found change points using various search methods in each 1000 simulations for a balanced ($n = 200$) and an unbalanced ($n = 5000$) scenario. The failure of the naive optimistic search in most cases for the unbalanced scenario is again clearly visible, while for the advanced, and in particular the combined optimistic search, the found change points very often lie exactly on the diagonal when compared to the full grid search and hence exactly the candidate proposed by the full grid search were found.

The simulation results in Table 2 and Figure 5 confirm our theoretical results that the naive optimistic search is not consistent for very unbalanced signals, while the advanced and combined versions are. Computationally, the number of evaluations for optimistic searches can be orders of magnitude smaller compared to full grid search, in particular if n is large.

Appendix C. Changes in High-dimensional Gaussian Graphical Models

We briefly describe an estimator introduced by Londschien et al. (2021) for change point detection in high-dimensional Gaussian graphical models, as this is the basis of all change point detection algorithms (BS, SeedBS, WBS and their optimistic variants) that we investigate in Section 6.3. For a segment $(u, v]$ with $0 \leq u < v \leq n$ let $\boldsymbol{S}_{(u,v]}$ denote the

Noise level	n	Average absolute estimation error for search methods			
		Naive	*Advanced*	*Combined*	*Full search*
$\sigma = 0.5$	100	3.38 (7)	2.77 (4)	2.88 (5)	3.24 (5)
	200	2.72 (4)	4.22 (7)	2.95 (5)	3.17 (5)
	300	3.43 (7)	4.45 (8)	3.21 (5)	3.16 (5)
	400	4.68 (10)	3.95 (6)	3.37 (5)	3.16 (5)
	500	6.55 (27)	4.24 (8)	3.09 (5)	3.08 (5)
	1000	13.75 (74)	3.84 (6)	3.35 (5)	3.08 (5)
	2000	171.74 (387)	3.92 (7)	3.26 (6)	3.01 (4)
	5000	1021.12 (1338)	3.92 (7)	3.52 (6)	3.05 (5)
$\sigma = 1$	100	15.86 (20)	15.26 (23)	15.07 (21)	16.79 (22)
	200	12.37 (18)	28.93 (43)	15.78 (26)	17.44 (28)
	300	19.50 (34)	26.91 (45)	19.30 (35)	17.73 (33)
	400	30.58 (56)	26.02 (54)	20.14 (42)	17.85 (37)
	500	50.09 (87)	26.97 (59)	21.06 (49)	18.80 (44)
	1000	136.75 (240)	29.70 (94)	24.59 (81)	21.24 (72)
	2000	544.70 (547)	35.73 (160)	34.16 (156)	24.21 (116)
	5000	1948.79 (1328)	48.08 (341)	51.94 (354)	38.34 (298)
$\sigma = 1.5$	100	25.24 (25)	33.95 (35)	31.70 (32)	34.19 (33)
	200	23.77 (29)	60.82 (62)	39.03 (50)	42.05 (52)
	300	41.23 (54)	65.17 (82)	50.79 (72)	48.55 (72)
	400	62.98 (85)	70.69 (107)	58.85 (95)	56.11 (93)
	500	96.54 (114)	82.27 (134)	70.03 (121)	62.41 (115)
	1000	253.11 (291)	121.14 (256)	114.73 (243)	98.52 (226)
	2000	739.92 (534)	202.01 (504)	203.74 (493)	156.51 (434)
	5000	2171.28 (1211)	436.96 (1269)	455.99 (1260)	355.35 (1123)
Noise level	n	Average number of evaluations for search methods			
		Naive	*Advanced*	*Combined*	*Full search*
$\sigma = 1$	100	16.18 (1)	25.10 (1)	41.28 (2)	199 (0)
	200	17.31 (1)	25.92 (2)	43.24 (2)	299 (0)
	500	19.08 (1)	29.34 (2)	48.43 (2)	599 (0)
	1000	19.36 (1)	30.95 (1)	50.31 (2)	1099 (0)
	2000	21.37 (1)	33.00 (1)	54.36 (2)	2099 (0)
	5000	23.69 (1)	35.02 (1)	58.71 (2)	5099 (0)

Table 2: Simulation results for Example 1. Reported are average absolute differences between the true change point at location 100 and the best single split point found by the respective search method (top) and the average number of evaluations (bottom). Values are averaged over 10,000 simulations and in parentheses the standard deviations (rounded to integers). The number of evaluations for noise levels $\sigma = 0.5, 1.5$ are omitted as they are very similar to the case $\sigma = 1$.
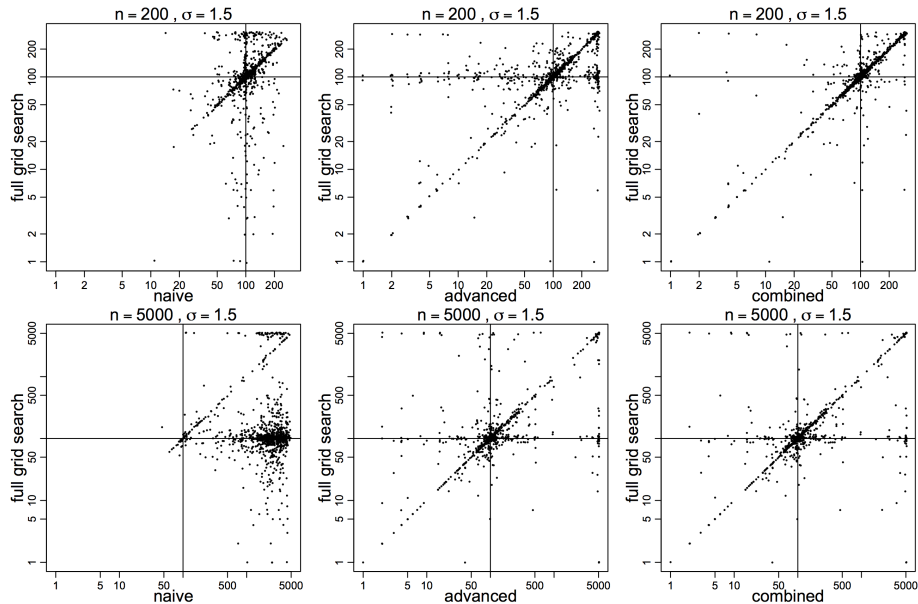
Figure 5: Pairwise plots of found change points using different optimistic search methods (horizontal-axis) versus the ones returned by the full grid search (vertical-axis) for a noise level $\sigma = 1.5$ and $n = 200$ (top) as well as $n = 5000$ (bottom) in 1000 simulations from Example 1. The vertical and horizontal lines indicate the location of the true change point at observation 100.

empirical covariance matrix within that segment. Let $0 < \epsilon < 1/2$ be some required minimal relative segment length and $\gamma > 0$ be a regularization parameter. For a segment $(u, v]$ with $v - u > 2\epsilon n$ we define the split point candidate as

$$\hat{\eta}_{(u,v]} = \underset{t \in \{u + \epsilon n, \ldots, v - \epsilon n\}}{\arg \max} L_n(\hat{\mathbf{\Omega}}^{\text{glasso}}_{(u,v]}; \mathbf{S}_{(u,v]}) - \left(L_n(\hat{\mathbf{\Omega}}^{\text{glasso}}_{(u,t]}; \mathbf{S}_{(u,t]}) + L_n(\hat{\mathbf{\Omega}}^{\text{glasso}}_{(t,v]}; \mathbf{S}_{(t,v]}) \right), \quad (16)$$

where

$$L_n(\mathbf{\Omega}; \mathbf{S}_{(u,v]}) = \frac{v - u}{n} \left(\text{Tr}(\mathbf{\Omega} \mathbf{S}_{(u,v]}) - \log(|\mathbf{\Omega}|) \right)$$

is a multivariate Gaussian log-likelihood based loss in the considered segment $(u, v]$ (scaled according to its length), and $\hat{\mathbf{\Omega}}^{\text{glasso}}_{(u,v]}$ is the graphical lasso precision matrix estimator (Friedman et al., 2008) with a scaled regularization parameter $\sqrt{n/(v - u)}\gamma$, i.e.,

$$\hat{\mathbf{\Omega}}^{\text{glasso}}_{(u,v]} = \underset{\mathbb{R}^{p \times p} \ni \mathbf{\Omega} \succ 0}{\arg \min} \text{Tr}(\mathbf{\Omega} \mathbf{S}_{(u,v]}) - \log(|\mathbf{\Omega}|) + \sqrt{(n/(v - u))}\gamma \|\mathbf{\Omega}\|_1.$$

Each split point t in (16) requires fitting two graphical lasso estimators. While various algorithms for computing exact or approximate solutions for the graphical lasso estimator exist, scalings of $O(p^3)$ (or worse) are common (assuming that the input covariance matrices have been pre-computed and that no special structures such as the block diagonal screening

26

proposed by Witten et al. (2011) as well as Mazumder and Hastie (2012a) can be exploited). Hence, the graphical lasso is rather costly especially when repeatedly fitting at each possible split point t on the grid $u + \epsilon n, \ldots, v - \epsilon n$. The essential problem is that it is not easy to re-use the estimator $\hat{\boldsymbol{\Omega}}_{(u,t]}^{\text{glasso}}$ for the segment $(u, t]$ to obtain $\hat{\boldsymbol{\Omega}}_{(u,t+1]}^{\text{glasso}}$ for the segment $(u, t+1]$. One could use $\hat{\boldsymbol{\Omega}}_{(u,t]}^{\text{glasso}}$ as a warm start, but not all algorithms that have been developed to compute graphical lasso fits are guaranteed to converge with warm starts (Mazumder and Hastie, 2012b) and even the ones that do converge would not save orders of magnitude in terms of runtime. This lack of efficient updates is common for more complex (e.g. high-dimensional) scenarios and it is in sharp contrast with e.g. change point detection in means for p-dimensional Gaussian variables. There, one needs to calculate means, but the mean for the segment $(u, t+1]$ can be updated in $O(p)$ cost if the mean for the segment $(u, t]$ is already available, and hence the computational cost is typically proportional to the total length of considered segments. In contrast, in the estimator in (16), the number of graphical lasso fits, as given by the number of considered split points, is the main driver of computational cost. Our optimistic search techniques rely on evaluating far fewer split points t than the full grid search and thus provide an option for massive computational speedups. Of course, the price to pay is having no guarantee to obtain exactly the optimal split point, but the "optimistic" approximation to $\hat{\eta}_{(u,v]}$ is still fairly good, see the simulations in Section 6.3.

In the simulations, we used the **glasso** R package, available on CRAN, for the graphical lasso fits. For all six methods, we set $\epsilon = 0.01$, i.e. skipping 20 observations on the boundaries of each considered interval and overall, no change points were searched in intervals containing less than 60 observations. We set $\gamma = 0.007$. Regularization in these examples is not essential in the sense that we do not have truly high-dimensional scenarios, but for split points close to the boundaries of the search interval and in short intervals, where the number of observations is close to p, regularization can be still helpful. We could have increased p in order to cover truly high-dimensional setups in our simulations, but given the scaling $O(p^3)$ of the graphical lasso, this very quickly goes beyond reasonable computational times for the full grid search based approaches that we want to include as references in terms of achievable estimation error.

C.1 Proof of Lemma 9

First note that in the following we only consider the case $l < \tau_i < t < \tau_{i+1} < r$, but similar arguments can be used also in the presence of a single change point in $(l, r]$ or when considering a split point t in the segment from l to the first change point within $(l, r]$. Recall that $\boldsymbol{\Sigma}_{(l,r]}$ denotes the convex combination of the covariance matrices within the segment $(l, r] \subseteq (0, 1]$ with the weights given by the relative segment lengths within $(l, r]$. In particular, for $l < \tau_i < t < \tau_{i+1} < r$,

$$\boldsymbol{\Sigma}_{(l,t]} = \frac{1}{t - l} \left((\tau_i - l) \boldsymbol{\Sigma}_{(l,\tau_i]} + (t - \tau_i) \boldsymbol{\Sigma}_{i+1} \right)$$

and

$$\boldsymbol{\Sigma}_{(t,r]} = \frac{1}{r - t} \left((\tau_{i+1} - t) \boldsymbol{\Sigma}_{i+1} + (r - \tau_{i+1}) \boldsymbol{\Sigma}_{(\tau_{i+1},r]} \right),$$

where $\boldsymbol{\Sigma}_{i+1} = \boldsymbol{\Sigma}_{(\tau_i, \tau_{i+1}]}$ is the covariance matrix in the $i+1$-st segment $(\tau_i, \tau_{i+1}]$. We seek to find the first and second derivatives of $G^*_{(l,r]}(t)$. First note that

$$\frac{\partial}{\partial t}\boldsymbol{\Sigma}_{(l,t]} = \frac{\tau_i - l}{(t-l)^2}\left(\boldsymbol{\Sigma}_{i+1} - \boldsymbol{\Sigma}_{(l,\tau_i]}\right)$$

and

$$\frac{\partial^2}{\partial t^2}\boldsymbol{\Sigma}_{(l,t]} = -\frac{2(\tau_i - l)}{(t-l)^3}\left(\boldsymbol{\Sigma}_{i+1} - \boldsymbol{\Sigma}_{(l,\tau_i]}\right) = -\frac{2}{t-l}\frac{\partial}{\partial t}\boldsymbol{\Sigma}_{(l,t]}.$$

We use the following expressions for derivatives of an invertible matrix $A(s)$ depending on s,

$$\frac{\partial}{\partial t}\log(|\boldsymbol{A}(t)|) = \mathrm{Tr}\left(\boldsymbol{A}(t)^{-1}\cdot\frac{\partial}{\partial t}\boldsymbol{A}(t)\right);$$

$$\frac{\partial}{\partial t}\mathrm{Tr}(\boldsymbol{A}(t)) = \mathrm{Tr}\left(\frac{\partial}{\partial t}\boldsymbol{A}(t)\right);$$

$$\frac{\partial}{\partial t}\boldsymbol{A}(t)^{-1} = -\boldsymbol{A}(t)^{-1}\cdot\frac{\partial}{\partial t}\boldsymbol{A}(t)\cdot\boldsymbol{A}(t)^{-1}.$$

Compute next the first and second derivatives of $G^*_{(l,r]}(t)$. Recall from equation (15) that

$$G^*_{(l,r]}(t) = \frac{r-l}{n}\log(|\boldsymbol{\Sigma}_{(l,r]}|) - \frac{t-l}{n}\log(|\boldsymbol{\Sigma}_{(l,t]}|) - \frac{r-t}{n}\log(|\boldsymbol{\Sigma}_{(t,r]}|).$$

Consider first the middle part of $G^*_{(l,r]}(t)$, i.e.

$$L^*(t) := -\frac{t-l}{n}\log(|\boldsymbol{\Sigma}_{(l,t]}|).$$

Then for the first derivative

$$\frac{\mathrm{d}}{\mathrm{d}t}L^*(t) = -\frac{1}{n}\log(|\boldsymbol{\Sigma}_{(l,t]}|) - \frac{t-l}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{(l,t]}^{-1}\cdot\frac{\partial}{\partial t}\boldsymbol{\Sigma}_{(l,t]}\right),$$

and for the second derivative

$$\begin{aligned}
&\frac{\mathrm{d}^2}{\mathrm{d}t^2}L^*(t)\\
&= -\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{(l,t]}^{-1}\cdot\frac{\partial}{\partial t}\boldsymbol{\Sigma}_{(l,t]}\right) - \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{(l,t]}^{-1}\cdot\frac{\partial}{\partial t}\boldsymbol{\Sigma}_{(l,t]}\right) - \frac{t-l}{n}\frac{\partial}{\partial t}\left(\mathrm{Tr}\left(\boldsymbol{\Sigma}_{(l,t]}^{-1}\cdot\frac{\partial}{\partial t}\boldsymbol{\Sigma}_{(l,t]}\right)\right)\\
&= -\frac{2}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{(l,t]}^{-1}\cdot\frac{\partial}{\partial t}\boldsymbol{\Sigma}_{(l,t]}\right) - \frac{t-l}{n}\mathrm{Tr}\left(-\boldsymbol{\Sigma}_{(l,t]}^{-1}\cdot\frac{\partial}{\partial t}\boldsymbol{\Sigma}_{(l,t]}\cdot\boldsymbol{\Sigma}_{(l,t]}^{-1}\cdot\frac{\partial}{\partial t}\boldsymbol{\Sigma}_{(l,t]}\right)\\
&\qquad -\frac{t-l}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{(l,t]}^{-1}\cdot\frac{\partial^2}{\partial t^2}\boldsymbol{\Sigma}_{(l,t]}\right)\\
&= \frac{t-l}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{(l,t]}^{-1}\cdot\frac{\partial}{\partial t}\boldsymbol{\Sigma}_{(l,t]}\cdot\boldsymbol{\Sigma}_{(l,t]}^{-1}\cdot\frac{\partial}{\partial t}\boldsymbol{\Sigma}_{(l,t]}\right) = \frac{t-l}{n}\left\|\boldsymbol{\Sigma}_{(l,t]}^{-1/2}\cdot\frac{\partial}{\partial t}\boldsymbol{\Sigma}_{(l,t]}\cdot\boldsymbol{\Sigma}_{(l,t]}^{-1/2}\right\|_F^2 \geq 0.
\end{aligned}$$

By symmetry, we can obtain similarly for the right part of $G^*_{(l,r]}(t)$,

$$\frac{\partial^2}{\partial t^2}\left(-\frac{r-t}{n}\log(|\mathbf{\Sigma}_{(t,r]}|)\right) = \frac{r-t}{n}\left\|\mathbf{\Sigma}_{(t,r]}^{-1/2}\cdot\frac{\partial}{\partial t}\mathbf{\Sigma}_{(t,r]}\cdot\mathbf{\Sigma}_{(t,r]}^{-1/2}\right\|_F^2 \geq 0.$$

As the left part of $G^*_{(l,r]}(t)$ is constant, we have $G^{*''}_{(l,r]}(t) \geq 0$ for $\tau_i < t < \tau_{i+1}$ and thus $G^*_{(l,r]}$ is convex in between τ_i and τ_{i+1}. Further, with the exception of the special cases of $\mathbf{\Sigma}_{(l,\tau_i]} = \mathbf{\Sigma}_{i+1} = \mathbf{\Sigma}_{(\tau_{i+1},t]}$, $G^*_{(l,r]}$ is even strictly convex in the interval $(\tau_i, \tau_{i+1}]$. For such special cases $G^*_{(l,r]}(t) = 0$ for arbitrary $t \in (\tau_i, \tau_{i+1}]$, i.e. the population gain function is flat between change points τ_i and τ_{i+1}. The special cases can only occur in the presence of two or more change points in the considered segment. In particular, in case τ_i is the single change point in $(l, r]$, $G^*_{(l,r]}$ is strictly convex in $(l, \tau_i]$ and strictly convex in $(\tau_i, r]$.

Appendix D. Computational Gains for High-dimensional Simulations

The achievable speedups using optimistic search in general are dependent on the cost of the model fit in each segment (how they depend on the number of observations n and the dimensionality p), whether there are possibilities to update neighboring fits efficiently, but also on the length of the series, the number of change points, which basic algorithm (BS, SeedBS, WBS or yet another one) is used with which specific tuning parameters, etc. Nonetheless, we would like to further comment on some of the observed computational gains in the high-dimensional simulations presented in Section 6.3 in the main text.

The biggest computational gains for optimistic search occur when the underlying search intervals are long. Random intervals have expected length $O(n)$ and thus many of them are comparably long. For these long intervals we gain a lot by optimistic searches. However, the lengths of the intervals in lower layers of seeded intervals are quite short (decaying exponentially) and what becomes dominant in that case is the number of very short intervals. For example, while there is only a single interval containing 2000 observations (first layer), there were more than sixty intervals on the lowest layer we considered with the minimally required segment length of $m = 60$ observations. This explains why the speedup for OSeedBS versus SeedBS is a factor 10–14, while for BS and WBS we could achieve factor 30 or more. Skipping the last few layers of seeded intervals would have saved considerable computational time for OSeedBS, which is important to keep in mind when interpreting the results from Figure 4 in the main text. From a practical perspective, when utilizing OSeedBS, one should thus limit the number of covered layers in order to consider fewer of the very short intervals that are a driver of computational cost. However, the minimal segment length in seeded intervals cannot be too large either, as in that case one is risking not covering each single change points sufficiently (similar to what happened in the shown examples for WBS and OWBS with a small number of random intervals M). The choice for the minimal segment length for seeded intervals might come fairly naturally in some applications, where segments below a certain size are uninteresting or when considering high-dimensional problems requiring a minimal number of observations for fitting reasonable models.

A pragmatic approach could be to combine the best of both worlds from OBS and OSeedBS. For example, find a first set of change points with fewer number of seeded intervals and then, to protect against the possibility that there could be even further change points

that were not discovered due to having chosen a too large minimal segment length, in between the first found change points from the seeded intervals, one could perform a further OBS-like search that adapts better to the number of change points within these shorter search intervals. This way adaptively one could invest more computational effort if there is evidence for further change points beyond the ones found by the rough first set of seeded intervals, but without the need to go over each and every very short interval as would be the case with further layers of seeded intervals containing very short intervals. Thus, one could keep computational advantages from OBS and at the same time exploit the better expected statistical performance of OSeedBS.

Appendix E. Proofs of Statistical Guarantees

Here we give proofs of statistical guarantees for optimistic searches in terms of consistency and localization rates. For ease of reading, we rewrite Model 2 as

$$\boldsymbol{X}_t = \boldsymbol{f}_t + \boldsymbol{\xi}_t \qquad \text{for} \quad t = 1, \ldots, n, \tag{17}$$

where $\boldsymbol{f}_t \equiv \boldsymbol{f}(t/n)$ with $\boldsymbol{f} : (0,1] \to \mathbb{R}^p$ defined as $\boldsymbol{f}(x) := \sum_{i=0}^{\kappa} \boldsymbol{\mu}_i \mathbb{1}_{(\tau_i, \tau_{i+1}]}(x)$, and $\boldsymbol{\xi}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$, i.e. independently, standard p-dimensional Gaussian distributed. Let $\boldsymbol{X} := (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top \in \mathbb{R}^{n \times p}$ be the data matrix, $\boldsymbol{F} := (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)^\top \in \mathbb{R}^{n \times p}$ the signal matrix and $\boldsymbol{\Xi} := (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)^\top \in \mathbb{R}^{n \times p}$ the noise matrix. Then, in an equivalent matrix form of (17), it holds that $\boldsymbol{X} = \boldsymbol{F} + \boldsymbol{\Xi}$. Similarly, we denote another independent sample as $\boldsymbol{Y} = \boldsymbol{F} + \tilde{\boldsymbol{\Xi}}$ (which is needed for sample splitting, see Section 5.1).

Towards a matrix-vector formulation of CUSUM statistics, we introduce

$$\boldsymbol{e}_t := \left(\underbrace{\sqrt{\frac{n-t}{nt}}, \cdots, \sqrt{\frac{n-t}{nt}}}_{t}, \underbrace{-\sqrt{\frac{t}{n(n-t)}}, \cdots, -\sqrt{\frac{t}{n(n-t)}}}_{n-t} \right)^\top \in \mathbb{R}^n,$$

for $t \in \{1, \ldots, n-1\}$, and $\boldsymbol{e}_n := (1/\sqrt{n}, \ldots, 1/\sqrt{n})^\top \in \mathbb{R}^n$. Then

$$\langle \boldsymbol{e}_t, \boldsymbol{e}_n \rangle = 0, \qquad \|\boldsymbol{e}_t\| = \|\boldsymbol{e}_n\| = 1, \qquad t = 1, \ldots, n-1,$$

where $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ are the inner product and the norm, respectively, in Euclidean spaces. Further notation is as follows. Let $\|\cdot\|_\infty$ denote the supremum norm of vectors. For real numbers a, b, let $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. Let also $\Phi(\cdot)$ be the distribution function of a standard Gaussian random variable.

In all proofs, we try to give constants as explicitly as possible, but those constants may not be optimal. The limiting behavior is considered as the sample size $n \to \infty$, and the involved quantities, including the sparsity level s, the dimension p, the signal $\boldsymbol{f}$, and thus the minimal segment length λ and the minimal jump size δ, are allowed to depend on n.

E.1 Technical Tools

We need several deviation bounds on chi-squares related quantities.

Lemma 10 (Tail of chi-squares) *Let $Y := \sum_{i=1}^{k} w_i(X_i + \mu_i)^2$, with $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and constants $w_i \geq 0$ and $\mu_i \in \mathbb{R}$. Then*

$$\mathbb{E}\left[Y\right] = \sum_{i=1}^{k} w_i(1 + \mu_i^2) \quad and \quad \mathrm{Var}\left(Y\right) = 2\sum_{i=1}^{k} w_i^2(1 + 2\mu_i^2).$$

Further, for every $x \geq 0$, it holds

$$\mathbb{P}\left\{Y \leq \mathbb{E}\left[Y\right] - x\right\} \leq \exp\left(-\frac{x^2}{2\mathrm{Var}\left(Y\right)}\right),$$

$$and \qquad \mathbb{P}\left\{Y \geq \mathbb{E}\left[Y\right] + x\right\} \leq \exp\left(-\frac{\mathrm{Var}\left(Y\right)}{4\left\|\boldsymbol{w}\right\|_{\infty}^2}\psi\left(\frac{2\left\|x\boldsymbol{w}\right\|_{\infty}}{\mathrm{Var}\left(Y\right)}\right)\right)$$

$$\leq \exp\left(-\frac{x^2}{2\mathrm{Var}\left(Y\right) + 4x\left\|\boldsymbol{w}\right\|_{\infty}}\right),$$

where $\psi(x) = 1 + x - \sqrt{1 + 2x}$.

Proof The expectation and variance are easy to compute. Note that

$$\psi(x) = \frac{x^2}{1 + x + \sqrt{1 + 2x}} \geq \frac{x^2}{2 + 2x}.$$

Then the second assertion is a reformulation of Lemma 2 in Laurent et al. (2012) or the Hanson–Wright inequality. $\blacksquare$

Lemma 11 (Tail of Bernoulli weighted chi-squares) *Let $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $B_i \overset{\text{ind.}}{\sim} \mathrm{Ber}(\beta_i)$, with $0 \leq \beta_i \leq 1$, $i = 1, \dots, k$, and $(X_i)_{i=1}^{k}$ and $(B_i)_{i=1}^{k}$ be independent. Let also*

$$Y := \sum_{i=1}^{k} \left(w_i(X_i + \mu_i)^2 - w_i(1 + \mu_i^2)\right) B_i$$

with constants $w_i \geq 0$, $\mu_i \in \mathbb{R}$. Then

$$\mathbb{E}\left[Y\right] = 0 \quad and \quad \mathrm{Var}\left(Y\right) = 2\sum_{i=1}^{k} \beta_i w_i^2(1 + 2\mu_i^2),$$

and, for every $x \geq 0$, it holds

$$\mathbb{P}\left\{Y \geq x\right\} \leq \exp\left(-\min\left\{\frac{x}{8\max_i\left(w_i(1 + 2\mu_i^2)^{1/2}\right)}, \frac{x^2}{6\mathrm{Var}\left(Y\right)}\right\}\right), \qquad (18\text{a})$$

$$\mathbb{P}\left\{Y \leq -x\right\} \leq \exp\left(-\min\left\{\frac{x}{2\max_i\left(w_i(1 + 2\mu_i^2)^{1/2}\right)}, \frac{x^2}{4\mathrm{Var}\left(Y\right)}\right\}\right). \qquad (18\text{b})$$

Further, if $\min_{1 \leq i \leq k} \beta_i \geq 1/2$, then for every $x \geq 0$,

$$\mathbb{P}\{Y \geq x\} \leq \exp\left(-\min\left\{\frac{x}{8\|\boldsymbol{w}\|_\infty}, \frac{x^2}{8\mathrm{Var}(Y)}\right\}\right), \tag{19a}$$

$$\mathbb{P}\{Y \leq -x\} \leq \exp\left(-\frac{x^2}{4\mathrm{Var}(Y)}\right). \tag{19b}$$

Proof We introduce the shorthand notation

$$Z_i := w_i(X_i + \mu_i)^2 - w_i(1 + \mu_i^2).$$

Then $Y = \sum_{i=1}^k Z_i B_i$, and $\mathbb{E}[Y] = \sum_{i=1}^k \mathbb{E}[Z_i]\mathbb{P}\{B_i = 1\} = 0$ and

$$\mathrm{Var}(Y) = \sum_{i=1}^k \mathbb{E}[Z_i^2]\mathbb{P}\{B_i = 1\} = 2\sum_{i=1}^k \beta_i w_i^2(1 + 2\mu_i^2).$$

Note that, for $-\infty < a < 1/(2w_i)$,

$$\mathbb{E}[\exp(aZ_i)] = \exp\left(\frac{2a^2 w_i^2 \mu_i^2}{1 - 2aw_i} - aw_i - \frac{1}{2}\log(1 - 2aw_i)\right),$$

and also that

$$-x - \frac{1}{2}\log(1 - 2x) \leq \begin{cases} \frac{x^2}{1-2x} & \text{if } 0 \leq x < 1/2, \\ x^2 & \text{if } x \leq 0. \end{cases} \tag{20}$$

We consider two separate cases:

- The case of general $\beta_i \in [0, 1]$. For $0 \leq x \leq (1 + 2\mu^2)^{-1/2}/4$, it holds

$$\exp\left(\frac{2x^2\mu^2}{1 - 2x} - x - \frac{1}{2}\log(1 - 2x)\right) - 1 \leq \exp\left(\frac{(1 + 2\mu^2)x^2}{1 - 2x}\right) - 1$$

$$\leq \exp\left(2(1 + 2\mu^2)x^2\right) - 1 \leq 3(1 + 2\mu^2)x^2.$$

Thus, for a such that $0 \leq 4a \max_i \left(w_i(1 + 2\mu_i^2)^{1/2}\right) \leq 1$, we obtain

$$\log(\mathbb{E}[\exp(aY)]) = \sum_{i=1}^k \log\left(\beta_i \mathbb{E}[\exp(aZ_i)] + 1 - \beta_i\right)$$

$$\leq \sum_{i=1}^k \log\left(1 + 3\beta_i(1 + 2\mu_i^2)a^2 w_i^2\right) \leq \sum_{i=1}^k 3\beta_i(1 + 2\mu_i^2)a^2 w_i^2.$$

By the Chernoff bound, we obtain

$$\log\left(\mathbb{P}\{Y \geq x\}\right) \leq \inf_{0 \leq 4a \max_i\left(w_i(1+2\mu_i^2)^{1/2}\right) \leq 1}\left(-ax + \sum_{i=1}^k 3\beta_i(1 + 2\mu_i^2)a^2 w_i^2\right)$$

$$\leq -\min\left\{\frac{x}{8\max_i\left(w_i(1 + 2\mu_i^2)^{1/2}\right)}, \frac{x^2}{12\sum_{i=1}^k \beta_i(1 + 2\mu_i^2)w_i^2}\right\}.$$

Similarly, we have, for $-1 \leq a \max_i \left(w_i(1+2\mu_i^2)^{1/2} \right) \leq 0$,

$$\log(\mathbb{E}\left[\exp(aY)\right]) \leq \sum_{i=1}^{k} \log\left(1 + 2\beta_i(1+2\mu_i^2)a^2w_i^2\right) \leq \sum_{i=1}^{k} 2\beta_i(1+2\mu_i^2)a^2w_i^2,$$

and

$$\log\left(\mathbb{P}\left\{Y \leq -x\right\}\right) \leq \inf_{-1 \leq a \max_i \left(w_i(1+2\mu_i^2)^{1/2}\right) \leq 0} \left(ax + \sum_{i=1}^{k} 2\beta_i(1+2\mu_i^2)a^2w_i^2\right)$$

$$\leq -\min\left\{\frac{x}{2\max_i\left(w_i(1+2\mu_i^2)^{1/2}\right)}, \frac{x^2}{8\sum_{i=1}^{k}\beta_i(1+2\mu_i^2)w_i^2}\right\}.$$

- The case of $\beta_i \geq 1/2$ for all $i \in \{1,\ldots,k\}$. For $0 \leq x < 1/2 \leq \beta \leq 1$,

$$1 + \beta \exp\left(\frac{2x^2\mu^2}{1-2x} - x - \frac{1}{2}\log(1-2x)\right) - \beta \leq \exp\left(\frac{2\beta x^2(1+2\mu^2)}{1-2x}\right).$$

This implies, for $0 \leq 2a\left\|w\right\|_\infty < 1$,

$$\log(\mathbb{E}\left[\exp(aY)\right]) = \sum_{i=1}^{k} \log\left(\beta_i\mathbb{E}\left[\exp(aZ_i)\right] + 1 - \beta_i\right) \leq \sum_{i=1}^{k} \frac{2\beta_i a^2 w_i^2(1+2\mu_i^2)}{1-2aw_i}.$$

Then, again by the Chernoff bound, we have

$$\log\left(\mathbb{P}\left\{Y \geq x\right\}\right) \leq \inf_{0 \leq 2a\left\|w\right\|_\infty < 1} \left(-ax + \sum_{i=1}^{k} \frac{2\beta_i a^2 w_i^2(1+2\mu_i^2)}{1-2aw_i}\right)$$

$$\leq \inf_{0 \leq 4a\left\|w\right\|_\infty < 1} \exp\left(-ax + \sum_{i=1}^{k} 4\beta_i a^2 w_i^2(1+2\mu_i^2)\right)$$

$$\leq -\min\left\{\frac{x}{8\left\|w\right\|_\infty}, \frac{x^2}{16\sum_{i=1}^{k}\beta_i(1+2\mu_i^2)w_i^2}\right\}.$$

Similarly, for $a \leq 0$, we have

$$\log(\mathbb{E}\left[\exp(aY)\right]) = \sum_{i=1}^{k} \log\left(\beta_i\mathbb{E}\left[\exp(aZ_i)\right] + 1 - \beta_i\right)$$

$$\leq \sum_{i=1}^{k} \log\left(\beta_i\exp\left(2\beta_i a^2 w_i^2(1+2\mu_i^2)\right) + 1 - \beta_i\right) \leq \sum_{i=1}^{k} 2\beta_i a^2 w_i^2(1+2\mu_i^2),$$

and

$$\log\left(\mathbb{P}\left\{Y \leq -x\right\}\right) \leq \inf_{a \leq 0} \left(ax + \sum_{i=1}^{k} 2\beta_i(1+2\mu_i^2)a^2w_i^2\right) = -\frac{x^2}{8\sum_{i=1}^{k}\beta_i(1+2\mu_i^2)w_i^2}.$$

Therefore, the assertions follow, as $\mathrm{Var}\,(Y) = 2\sum_{i=1}^{k}\beta_i(1+2\mu_i^2)w_i^2$. ∎

Remark 12 *In comparison with Lemma 10, there is an additional term $\max_i\left(w_i(1+2\mu_i^2)^{1/2}\right)$ in the bound of lower tail probability, see (18), when there are Bernoulli weights. We stress that such a term is necessary, especially when $\beta_i \searrow 0$ for every $i \in \{1,\ldots,k\}$. However, in case of $\min_i\beta_i \geq 1/2$, Y behaves the same as if there are no Bernoulli weights (i.e. $\beta_1 = \cdots = \beta_p = 1$), see (19). In particular, up to difference in constants, Lemma 11 includes Lemma 10 as a special case.*

Lemma 13 (Lower tail of Bernoulli weighted non-central chi-squares) *Let $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $B_i \overset{ind.}{\sim} \mathrm{Ber}(\beta_i)$, with $0 \leq \beta_i \leq 1$, $i = 1,\ldots,k$, and $(X_i)_{i=1}^k$ and $(B_i)_{i=1}^k$ be independent. Let also*

$$Y := \sum_{i=1}^{k} B_i w_i (X_i + \mu_i)^2$$

with constants $w_i \geq 0$, $\mu_i \in \mathbb{R}$. Assume that, for some constant $C > 0$,

$$1 - \beta_i \leq \exp\left(-Cw_i(1+\mu_i^2)\right), \quad i = 1,\ldots,k. \tag{21}$$

Then, for every $x \geq 0$, it holds that

$$\mathbb{P}\left\{Y - \sum_{i=1}^{k} w_i(1+\mu_i^2) \leq -x\right\}$$
$$\leq \exp\left(-\min\left\{\frac{Cx}{4}, \frac{x^2}{4\sum_{i=1}^{k} w_i^2(1+2\mu_i^2)}\right\}\right)\cdot \prod_{i=1}^{k}\left(1+\exp\left(-\frac{Cw_i(1+\mu_i^2)}{2}\right)\right)$$
$$\leq 2^k \exp\left(-\min\left\{\frac{Cx}{4}, \frac{x^2}{4\sum_{i=1}^{k} w_i^2(1+2\mu_i^2)}\right\}\right).$$

Proof Note first that, for $a \leq 0$,

$$\mathbb{E}\left[\exp\left(aY - a\sum_{i=1}^{k} w_i(1+\mu_i^2)\right)\right]$$
$$= \prod_{i=1}^{k}\left(\beta_i\mathbb{E}\left[\exp\left(aw_i(X_i+\mu_i)^2 - aw_i(1+\mu_i^2)\right)\right] + (1-\beta_i)\exp\left(-aw_i(1+\mu_i^2)\right)\right)$$
$$= \prod_{i=1}^{k}\left(\beta_i\exp\left(\frac{2a^2w_i^2\mu_i^2}{1-2aw_i} - aw_i - \frac{1}{2}\log(1-2aw_i)\right) + (1-\beta_i)\exp\left(-aw_i(1+\mu_i^2)\right)\right)$$
$$\leq \prod_{i=1}^{k}\left(\exp\left(a^2w_i^2(1+2\mu_i^2)\right) + \exp\left(-(a+C)w_i(1+\mu_i^2)\right)\right),$$

where the last inequality is due to (20) and the assumption in (21). Then, we apply the Chernoff bound and obtain

$$\mathbb{P}\left\{Y - \sum_{i=1}^{k} w_i(1+\mu_i^2) \leq -x\right\} \leq \inf_{a \leq 0} \mathbb{E}\left[\exp\left(aY - a\sum_{i=1}^{k} w_i(1+\mu_i^2) + ax\right)\right]$$

$$\leq \inf_{a \leq 0} e^{ax} \prod_{i=1}^{k}\Big(\exp\big(a^2 w_i^2(1+2\mu_i^2)\big) + \exp\big(-(a+C)w_i(1+\mu_i^2)\big)\Big)$$

$$\leq \inf_{-C/2 \leq a \leq 0} e^{ax} \prod_{i=1}^{k}\left(\exp\big(a^2 w_i^2(1+2\mu_i^2)\big) + \exp\Big(-\frac{Cw_i(1+\mu_i^2)}{2}\Big)\right)$$

$$\leq \inf_{-C/2 \leq a \leq 0} e^{ax} \prod_{i=1}^{k}\left(1 + \exp\Big(-\frac{Cw_i(1+\mu_i^2)}{2}\Big)\right)\exp\big(a^2 w_i^2(1+2\mu_i^2)\big)$$

$$\leq \inf_{-C/2 \leq a \leq 0} \exp\left(ax + a^2\sum_{i=1}^{k} w_i^2(1+2\mu_i^2)\right) \cdot \prod_{i=1}^{k}\left(1 + \exp\Big(-\frac{Cw_i(1+\mu_i^2)}{2}\Big)\right)$$

$$\leq \exp\left(-\min\left\{\frac{Cx}{4}, \frac{x^2}{4\sum_{i=1}^{k} w_i^2(1+2\mu_i^2)}\right\}\right) \cdot \prod_{i=1}^{k}\left(1 + \exp\Big(-\frac{Cw_i(1+\mu_i^2)}{2}\Big)\right),$$

which concludes the proof, since $\exp\big(-Cw_i(1+\mu_i^2)/2\big) \leq 1$. ∎

Remark 14 (Upper tail) *We stress that the bound of the upper tail probability of Y follows readily from Lemma 10, since*

$$\mathbb{P}\left\{Y - \sum_{i=1}^{k} w_i(1+\mu_i^2) \geq x\right\} \leq \mathbb{P}\left\{\sum_{i=1}^{k} w_i(X_i+\mu_i)^2 - \sum_{i=1}^{k} w_i(1+\mu_i^2) \geq x\right\}.$$

It is a bit unusual that the concentration inequalities here are centered at $\sum_{i=1}^{k} w_i(1+\mu_i^2)$ rather than $\mathbb{E}[Y] = \sum_{i=1}^{k} \beta_i w_i(1+\mu_i^2)$, but this makes little difference as β_i's are fairly close to one, which is assumed in (21). The current version is chosen to ease our later proofs.

Consider next concentration inequalities on the difference of (Bernoulli weighted) chi-squares. The key is to decouple the possible correlation between the involved chi-squares.

Lemma 15 (Tail of difference of chi-squares) *Let $\boldsymbol{X}_j \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{I}_n)$ be independent, with $\boldsymbol{\mu}_j \in \mathbb{R}^n$, $j = 1, \ldots, k$, and $t, r \in \{1, \ldots, n\}$ be arbitrary. Define the relative difference of t, r within the background $(0, n]$ as*

$$d(t, r) := \frac{|t - r|}{\min\{r \vee t,\ n - (r \wedge t)\}}, \tag{22}$$

which is always in $[0, 1]$. Then, for every $x \geq 0$, it holds that

$$\mathbb{P}\left\{\sum_{j=1}^{k}\langle e_t, X_j\rangle^2 - \sum_{j=1}^{k}\langle e_r, X_j\rangle^2 \leq \sum_{j=1}^{k}\langle e_t, \mu_j\rangle^2 - \sum_{j=1}^{k}\langle e_r, \mu_j\rangle^2 - x\right\}$$

$$\leq 2\exp\left(-\min\left\{\frac{x^2}{32\left(2d(t,r)\wedge 1\right)\left(k + 2\sum_{j=1}^{k}\|\mu_j\|^2\right)}, \frac{x}{16\sqrt{2d(t,r)\wedge 1}}\right\}\right).$$

Proof We first compute the eigendecomposition

$$e_t e_t^\top - e_r e_r^\top = \begin{pmatrix} u_+ & u_- \end{pmatrix}\begin{pmatrix} \theta & 0 \\ 0 & -\theta \end{pmatrix}\begin{pmatrix} u_+ & u_- \end{pmatrix}^\top = \theta u_+ u_+^\top - \theta u_- u_-^\top,$$

with $\theta = \sqrt{1 - \langle e_t, e_r\rangle^2} \geq 0$, $\|u_+\| = \|u_-\| = 1$ and $\langle u_+, u_-\rangle = 0$. More precisely,

$$u_+ = \frac{1}{2\sqrt{1 + \langle e_t, e_r\rangle}}(e_t + e_r) + \frac{1}{2\sqrt{1 - \langle e_t, e_r\rangle}}(e_t - e_r),$$

$$\text{and}\quad u_- = \frac{1}{2\sqrt{1 + \langle e_t, e_r\rangle}}(e_t + e_r) - \frac{1}{2\sqrt{1 - \langle e_t, e_r\rangle}}(e_t - e_r).$$

Elementary calculation shows that

$$d(t,r) \leq \theta^2 = \frac{n\,|t - r|}{(r \vee t)\big(n - (r \wedge t)\big)} \leq \min\left\{2d(t,r),\ 1\right\}.$$

Define $\xi_+ := \sum_{j=1}^{k}\langle u_+, X_j\rangle^2$ and $\xi_- := \sum_{j=1}^{k}\langle u_-, X_j\rangle^2$. Then ξ_+ and ξ_- are independent, because $u_+^\top(X_1, \ldots, X_k)$ and $u_-^\top(X_1, \ldots, X_k)$ are jointly Gaussian and uncorrelated, thus independent.

Note that

$$\left(\sum_{j=1}^{k}\langle e_t, X_j\rangle^2 - \sum_{j=1}^{k}\langle e_r, X_j\rangle^2\right) - \left(\sum_{j=1}^{k}\langle e_t, \mu_j\rangle^2 - \sum_{j=1}^{k}\langle e_r, \mu_j\rangle^2\right)$$

$$= \sum_{j=1}^{k}X_j^\top(e_t e_t^\top - e_r e_r^\top)X_j - \sum_{j=1}^{k}\mu_j^\top(e_t e_t^\top - e_r e_r^\top)\mu_j$$

$$= \theta(\xi_+ - \xi_-) - \theta(\mathbb{E}[\xi_+] - \mathbb{E}[\xi_-]).$$

Thus, we have

$$\mathbb{P}\left\{\theta(\xi_+ - \xi_-) \leq \theta(\mathbb{E}[\xi_+] - \mathbb{E}[\xi_-]) - x\right\}$$

$$\leq \mathbb{P}\left\{\xi_+ \leq \mathbb{E}[\xi_+] - \frac{x}{2\theta}\right\} + \mathbb{P}\left\{\xi_- \geq \mathbb{E}[\xi_-] + \frac{x}{2\theta}\right\}. \tag{23}$$

By Lemma 10, we obtain, for the first term in (23),

$$\mathbb{P}\left\{\xi_+ \leq \mathbb{E}[\xi_+] - \frac{x}{2\theta}\right\} \leq \exp\left(-\frac{x^2}{8\theta^2 \text{Var}(\xi_+)}\right) = \exp\left(-\frac{x^2}{16\theta^2\left(k + 2\sum_{j=1}^{k}\langle u_+, \mu_j\rangle^2\right)}\right)$$

$$\leq \exp\left(-\frac{x^2}{16\theta^2\left(k + 2\sum_{j=1}^k \|\boldsymbol{\mu}_j\|^2\right)}\right)$$

$$\leq \exp\left(-\frac{x^2}{16\big(2d(t,r)\wedge 1\big)\left(k + 2\sum_{j=1}^k \|\boldsymbol{\mu}_j\|^2\right)}\right),$$

and, similarly, for the second term in (23),

$$\mathbb{P}\left\{\xi_- \geq \mathbb{E}\left[\xi_-\right] + \frac{x}{2\theta}\right\}$$

$$\leq \exp\left(-\frac{x^2}{8\theta^2 \mathrm{Var}\left(\xi_-\right) + 8\theta x}\right) \leq \exp\left(-\frac{x^2}{16\theta^2\left(k + 2\sum_{j=1}^k \|\boldsymbol{\mu}_j\|^2\right) + 8\theta x}\right)$$

$$\leq \exp\left(-\min\left\{\frac{x^2}{32\big(2d(t,r)\wedge 1\big)\left(k + 2\sum_{j=1}^k \|\boldsymbol{\mu}_j\|^2\right)}, \frac{x}{16\sqrt{2d(t,r)\wedge 1}}\right\}\right).$$

These two upper bounds, together with (23), conclude the proof. ∎

Remark 16 (Centering) *Note that $\boldsymbol{u}_+, \boldsymbol{u}_- \in \mathrm{span}\,\{\boldsymbol{e}_t, \boldsymbol{e}_r\} \subseteq \big(\mathrm{span}\,\{\boldsymbol{e}_n\}\big)^\perp$. Then*

$$\left\langle \boldsymbol{u}_+, \boldsymbol{\mu}_j \right\rangle^2 = \left\langle \boldsymbol{u}_+, (\boldsymbol{I}_n - \boldsymbol{e}_n\boldsymbol{e}_n^\top)\boldsymbol{\mu}_j \right\rangle^2 \leq \left\| (\boldsymbol{I}_n - \boldsymbol{e}_n\boldsymbol{e}_n^\top)\boldsymbol{\mu}_j \right\|^2$$

and a similar bound holds for $\left\langle \boldsymbol{u}_-, \boldsymbol{\mu}_j \right\rangle^2$. Thus, Lemma 15 can be slightly improved if every $\|\boldsymbol{\mu}_j\|^2$ is replaced by $\|(\boldsymbol{I}_n - \boldsymbol{e}_n\boldsymbol{e}_n^\top)\boldsymbol{\mu}_j\|^2$.

Remark 17 (Simple bound) *A simple bound on the tail of difference of chi-squares can be derived directly from Lemma 10 as follows:*

$$\mathbb{P}\left\{\sum_{j=1}^k \left\langle \boldsymbol{e}_t, \boldsymbol{X}_j \right\rangle^2 - \sum_{j=1}^k \left\langle \boldsymbol{e}_r, \boldsymbol{X}_j \right\rangle^2 \leq \sum_{j=1}^k \left\langle \boldsymbol{e}_t, \boldsymbol{\mu}_j \right\rangle^2 - \sum_{j=1}^k \left\langle \boldsymbol{e}_r, \boldsymbol{\mu}_j \right\rangle^2 - x\right\}$$

$$\leq \mathbb{P}\left\{\sum_{j=1}^k \left\langle \boldsymbol{e}_t, \boldsymbol{X}_j \right\rangle^2 - \sum_{j=1}^k \left(\left\langle \boldsymbol{e}_t, \boldsymbol{\mu}_j \right\rangle^2 + 1\right) \leq -\frac{x}{2}\right\}$$

$$+ \mathbb{P}\left\{\sum_{j=1}^k \left\langle \boldsymbol{e}_r, \boldsymbol{X}_j \right\rangle^2 - \sum_{j=1}^k \left(\left\langle \boldsymbol{e}_r, \boldsymbol{\mu}_j \right\rangle^2 + 1\right) \geq \frac{x}{2}\right\}$$

$$\leq 2\exp\left(-\min\left\{\frac{x^2}{32\left(k + 2\sum_{j=1}^k (\left\langle \boldsymbol{e}_r, \boldsymbol{\mu}_j \right\rangle^2 \vee \left\langle \boldsymbol{e}_t, \boldsymbol{\mu}_j \right\rangle^2)\right)}, \frac{x}{16}\right\}\right) \tag{24}$$

$$\leq 2\exp\left(-\min\left\{\frac{x^2}{32\left(k + 2\sum_{j=1}^k \|\boldsymbol{\mu}_j\|^2\right)}, \frac{x}{16}\right\}\right). \tag{25}$$

This simple bound above is no better than that in Lemma 15, and the one in Lemma 15 is strictly sharper if $d(t,r) < 1/2$, and particularly, if $d(t,r) = o(1)$.

In exactly the same way as Lemma 15, we can derive the tail bound on the differences of Bernoulli weighted chi-squares.

Lemma 18 (Tail of difference of Bernoulli weighted chi-squares) *Assume the same setup as in Lemma 15, and let $B_i \overset{ind.}{\sim} \mathrm{Ber}(\beta_i)$, with $0 \le \beta_i \le 1$, $i = 1, \ldots, k$, be independent from $X_1, \ldots, X_k$. Then, for every $x \ge 0$:*

i. It holds that

$$
\mathbb{P}\left\{ \sum_{j=1}^{k} B_j \left(\langle e_t, X_j \rangle^2 - \langle e_t, \mu_j \rangle^2 \right) - \sum_{j=1}^{k} B_j \left(\langle e_r, X_j \rangle^2 - \langle e_r, \mu_j \rangle^2 \right) \le -x \right\}
$$
$$
\le 2 \exp\left(- \min\left\{ \frac{x}{16\sqrt{(2d(t,r) \wedge 1) \max_i (1 + 2\|\mu_i\|^2)}}, \right.\right.
$$
$$
\left.\left. \frac{x^2}{48(2d(t,r) \wedge 1) \sum_{j=1}^{k} \beta_j \left(1 + 2\|\mu_j\|^2\right)} \right\} \right).
$$

ii. If $\beta_i \ge 1/2$ for every $i = 1, \ldots, k$, then

$$
\mathbb{P}\left\{ \sum_{j=1}^{k} B_j \left(\langle e_t, X_j \rangle^2 - \langle e_t, \mu_j \rangle^2 \right) - \sum_{j=1}^{k} B_j \left(\langle e_r, X_j \rangle^2 - \langle e_r, \mu_j \rangle^2 \right) \le -x \right\}
$$
$$
\le 2 \exp\left(- \min\left\{ \frac{x}{16\sqrt{2d(t,r) \wedge 1}}, \frac{x^2}{64(2d(t,r) \wedge 1)\left(k + 2\sum_{j=1}^{k}\|\mu_j\|^2\right)} \right\} \right).
$$

iii. If the condition (21) holds, then

$$
\mathbb{P}\left\{ \sum_{j=1}^{k} B_j \langle e_t, X_j \rangle^2 - \sum_{j=1}^{k} B_j \langle e_r, X_j \rangle^2 \le \sum_{j=1}^{k} \langle e_t, \mu_j \rangle^2 - \sum_{j=1}^{k} \langle e_r, \mu_j \rangle^2 - x \right\}
$$
$$
\le 2 \exp\left(- \min\left\{ \frac{x^2}{32(2d(t,r) \wedge 1)\left(k + 2\sum_{j=1}^{k}\|\mu_j\|^2\right)}, \frac{(C \wedge 1/2)x}{8\sqrt{2d(t,r) \wedge 1}} \right\} \right)
$$
$$
\times \prod_{i=1}^{k} \left(1 + \exp\left(-\frac{C w_i (1 + \mu_i^2)}{2} \right) \right).
$$

The proof of Lemma 18 is omitted for brevity, since the only difference to that of Lemma 15 is to employ Lemmas 11 and 13 instead of Lemma 10. Remarks 16 and 17 apply here as well.

E.2 Single Change Point

We consider here the particular case of $\kappa = 1$, i.e. a single change point, in Models 1 and 2, and provide proofs for Theorems 1, 2 and 6. Since the involved CUSUM statistic is invariant to constant shifts, we assume, without loss of generality, in (17),

$$\boldsymbol{f}(x) = \boldsymbol{\delta}\mathbb{1}_{(0,\lambda]}(x), \qquad x \in (0,1],$$

where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)^\top \in \mathbb{R}^p$, the change point $\tau = \lambda \in (0, 1/2]$, and the jump size $\delta = \|\boldsymbol{\delta}\|$.

The following proofs rely on the observation that the localization error is no larger than the minimal length of search intervals that contain the only true change point. Moreover, assume that the search interval in a step still contains the change point, i.e. no mistake has been made yet. Then excluding the segment containing the true change point, and thus making a mistake, can only happen when both probe points lie to the left of the change point or when both probe points lie to the right of the change point. In such cases, in order to avoid wrongly excluding the segment containing the true change point, we have to ensure that the difference of population gain function at two investigated probe points is larger than the random oscillation caused by the noise with high probability.

For ease of notation, we assume the default step size $\nu = 1/2$ in optimistic searches, which implies that the three parts within each search window have relative lengths $1:1:1$, $1:1:2$ or $2:1:1$. We further assume that there is no rounding in determining the dyadic search locations and the probe points in all search intervals.

E.2.1 Naive Optimistic Search

Proof [of Theorem 1] We will prove that the assertion of theorem holds with constants

$$C_0 = 20 \quad \text{and} \quad C_1 = 3^2 2^{17}.$$

The proof consists of two steps.

Firstly, we show that, with probability tending to one, the naive optimistic searches makes no mistake whenever a search window is of length no shorter than $O(n\lambda)$. To this end, we introduce $I_1, I_2, \ldots$, such that $I_1 \supset I_2 \supset \cdots$, as the search windows of the naive optimistic search in the population case, i.e. $I_1 = (0, n]$, $I_2 = (0, 2n/3]$ and so on. It is easy to see that the change point $n\lambda$ lies in every I_k, thus no mistakes. Let

$$k_* := \min\big\{k \,:\, (t \wedge w) \le n\lambda \text{ with } t, w \text{ the probe points of } I_k\big\},$$

i.e., when one probe point drops in $[0, n\lambda]$ for the first time. The left end point of I_k, $k \le k_*$, is always 0, and $|I_k| \ge |I_{k_*}| \asymp n\lambda$. Then it is sufficient to show that the first k_* steps of the naive optimistic search coincide with the population case, with probability tending to one.

We thus define

$$\mathcal{P}_1 := \big\{(t, w) \,:\, t, w \text{ are probe points of } I_k \text{ with } t \le w \text{ for } k < k_*\big\}.$$

Note that $\mathcal{P}_1$ is deterministic, and only depends on the signal f. Recall that

$$\boldsymbol{F} = \big(f(1/n), \ldots, f(n/n)\big)^\top \in \mathbb{R}^n.$$

Fix an arbitrary pair of probe points $(t, w) \in \mathcal{P}_1$. Then it holds that

$$\langle e_w, F \rangle^2 \leq \langle e_t, F \rangle^2 = n \lambda^2 \delta^2 \frac{n - t}{t} \leq n^2 \lambda^2 \delta^2 \frac{1}{t},$$

$$\text{and} \quad \langle e_t, F \rangle^2 - \langle e_w, F \rangle^2 = n^2 \lambda^2 \delta^2 \frac{|t - w|}{tw} \geq n^2 \lambda^2 \delta^2 \frac{1}{3t}.$$

We apply inequality (25) in Remark 17 and obtain

$$\mathbb{P}\left\{\left|\mathrm{CS}_{(0,n]}(t)\right| \leq \left|\mathrm{CS}_{(0,n]}(w)\right|\right\} = \mathbb{P}\left\{\langle e_t, X \rangle^2 \leq \langle e_w, X \rangle^2\right\}$$

$$\leq \mathbb{P}\left\{\langle e_t, X \rangle^2 - \langle e_w, X \rangle^2 \leq \langle e_t, F \rangle^2 - \langle e_w, F \rangle^2 - n^2 \lambda^2 \delta^2 \frac{1}{3t}\right\}$$

$$\leq 2 \exp\left(-\min\left\{\frac{n^4 \lambda^4 \delta^4}{288(t^2 + 2tn^2 \lambda^2 \delta^2)}, \frac{n^2 \lambda^2 \delta^2}{48t}\right\}\right)$$

$$\leq 2 \exp\left(-\min\left\{\frac{n^2 \lambda^4 \delta^4}{32(1 + 6n\lambda^2 \delta^2)}, \frac{n \lambda^2 \delta^2}{16}\right\}\right),$$

where the last inequality is due to $t \leq n/3$. Since $|\mathcal{P}_1| \leq \log(2\lambda)/\log(3/4) \leq 4 \log n$, the bound above in combination with the union bound implies

$$\mathbb{P}\left\{\left|\mathrm{CS}_{(0,n]}(t)\right| \leq \left|\mathrm{CS}_{(0,n]}(w)\right| \text{ for some } (t, w) \in \mathcal{P}_1\right\}$$

$$\leq \sum_{(t,w)\in\mathcal{P}_1} \mathbb{P}\left\{\left|\mathrm{CS}_{(0,n]}(t)\right| \leq \left|\mathrm{CS}_{(0,n]}(w)\right|\right\}$$

$$\leq 8 \exp\left(-\min\left\{\frac{n^2 \lambda^4 \delta^4}{32(1 + 6n\lambda^2 \delta^2)}, \frac{n \lambda^2 \delta^2}{16}\right\} + \log \log n\right).$$

Thus, if

$$n \lambda^2 \delta^2 \geq 400 \log \log n = C_0^2 \log \log n,$$

the probability that the first k_* steps of naive optimistic search differ from the population version satisfies

$$\mathbb{P}\left\{\left|\mathrm{CS}_{(0,n]}(t)\right| \leq \left|\mathrm{CS}_{(0,n]}(w)\right| \text{ for some } (t, w) \in \mathcal{P}_1\right\} \leq \frac{8}{\log n} \to 0.$$

Secondly, we consider the search windows in the naive optimistic search that are shorter than $O(n\lambda)$, namely, later steps $k \geq k_*$. Recall that the true change point $n\lambda$ can be wrongly excluded from consecutive search intervals only when both pairs of probe points t, w lie on the same side of $n\lambda$. It is thus sufficient to consider

$$\mathcal{P}_2 := \Big\{(t, w) \; : \; t, w \text{ are pairs of probe points from steps } k \geq k_* \text{ of}$$

$$\text{the naive optimistic search, such that } t, w \text{ lie on the same side of}$$

$$\text{the change point, and } t \text{ is closer to the change point}\Big\}.$$

We fix arbitrarily a pair of probe points $(t, w) \in \mathcal{P}_2$, and assume that the first k_* steps of the naive optimistic search coincide with the population case, which happens with probability towards one as shown earlier. It follows that $t, w \leq 4n\lambda$ and further that

$$\langle \boldsymbol{e}_t, \boldsymbol{F} \rangle^2 - \langle \boldsymbol{e}_w, \boldsymbol{F} \rangle^2 = \begin{cases} n^2 \lambda^2 \delta^2 \frac{|t-w|}{tw} \geq \frac{1}{16} |t-w| \delta^2 & \text{if } t, w \geq n\lambda, \\ n^2 \delta^2 \frac{|t-w|(1-\lambda)^2}{(n-t)(n-w)} \geq \frac{1}{4} |t-w| \delta^2 & \text{if } t, w \leq n\lambda. \end{cases}$$

The relative distance $d(\cdot, \cdot)$ in (22) satisfies

$$d(t, w) = \frac{|t-w|}{\min\{t \wedge w, (n-t) \wedge (n-w)\}} \leq \frac{4|t-w|}{n\lambda}.$$

Then, by Lemma 15 we have, for $(t, w) \in \mathcal{P}_2$,

$$\mathbb{P}\left\{\left|\text{CS}_{(0,n]}(t)\right| \leq \left|\text{CS}_{(0,n]}(w)\right|\right\} = \mathbb{P}\left\{\langle \boldsymbol{e}_t, \boldsymbol{X} \rangle^2 \leq \langle \boldsymbol{e}_w, \boldsymbol{X} \rangle^2\right\}$$

$$\leq \mathbb{P}\left\{\langle \boldsymbol{e}_t, \boldsymbol{X} \rangle^2 - \langle \boldsymbol{e}_w, \boldsymbol{X} \rangle^2 \leq \langle \boldsymbol{e}_t, \boldsymbol{F} \rangle^2 - \langle \boldsymbol{e}_w, \boldsymbol{F} \rangle^2 - \frac{1}{16} |t-w| \delta^2\right\}$$

$$\leq 2\exp\left(-\min\left\{\frac{x^2}{64 d(t, r)(1 + 2n\lambda\delta^2)}, \frac{x}{16\sqrt{2d(t, r)}}\right\}\right)$$

$$\leq 2\exp\left(-\min\left\{\frac{n\lambda |t-w| \delta^4}{2^{16}(1 + 2n\lambda\delta^2)}, \frac{\sqrt{n\lambda |t-w|}\delta^2}{2^{19/2}}\right\}\right).$$

Note that $\mathcal{P}_2$ is contained in a mother set $\mathcal{P}_2^*$ of size $\leq (4\log n)^2$, and that $\mathcal{P}_2^*$ is determined only by the signal f, see later Part 2 in the proof of Theorem 6 for a formal proof. Let

$$\varepsilon_* := \frac{C_1}{2} \frac{\log\log n}{\delta^2} = 3^2 2^{16} \frac{\log\log n}{\delta^2}.$$

Then, by the union bound again, we obtain

$$\mathbb{P}\left\{\left|\text{CS}_{(0,n]}(t)\right| \leq \left|\text{CS}_{(0,n]}(w)\right| \text{ for some } (t, w) \in \mathcal{P}_2 \text{ with } |t-w| \geq \varepsilon_*\right\}$$

$$\leq \mathbb{P}\left\{\left|\text{CS}_{(0,n]}(t)\right| \leq \left|\text{CS}_{(0,n]}(w)\right| \text{ for some } (t, w) \in \mathcal{P}_2^* \text{ with } |t-w| \geq \varepsilon_*\right\}$$

$$\leq \sum_{(t,w)\in\mathcal{P}_2^*, |t-w|\geq\varepsilon_*} \mathbb{P}\left\{\left|\text{CS}_{(0,n]}(t)\right| \leq \left|\text{CS}_{(0,n]}(w)\right|\right\}$$

$$\leq 32\exp\left(-\min\left\{\frac{n\lambda\varepsilon_*\delta^4}{2^{16}(1 + 2n\lambda\delta^2)}, \frac{\sqrt{n\lambda\varepsilon}\delta^2}{2^{19/2}}\right\} + 2\log\log n\right) \leq \frac{32}{\log n} \to 0.$$

This implies an upper bound of $2\varepsilon_*$ on the localization error of the change point, which concludes the proof. ∎

E.2.2 ADVANCED AND COMBINED OPTIMISTIC SEARCHES

Since Theorem 2 is a special of Theorem 6 when $p = 1$, we only need to prove Theorem 6. **Proof** [of Theorem 6] Divide the p coordinates of observations in Model 2 into three groups:

i. The set of coordinates with large jump sizes

$$L := \left\{ j \ : \ |\delta_j|^2 \geq \frac{1}{32} \frac{\|\boldsymbol{\delta}\|^2}{s} \right\};$$

ii. The set of coordinates with small jump sizes

$$S := \left\{ j \ : \ 0 < |\delta_j|^2 < \frac{1}{32} \frac{\|\boldsymbol{\delta}\|^2}{s} \right\};$$

iii. The set of coordinates with no jumps

$$N := \{ j \ : \ \delta_j = 0 \}.$$

The constant $1/32$ above can be replaced by any constant that is sufficiently small. Clearly, L, S, N are disjoint, $|L| + |S| \leq s$ and $L \cup S \cup N = \{1, \ldots, p\}$. It holds that $L \neq \emptyset$, since

$$\sum_{j \in S} |\delta_j|^2 \leq \frac{1}{32} \|\boldsymbol{\delta}\|^2 < \|\boldsymbol{\delta}\|^2 = \sum_{j \in S \cup L} |\delta_j|^2.$$

The remaining proof is split into two parts.

Part 1. Global search over dyadic locations. In this part, we will show

$$\mathbb{P} \{ n\lambda \in (t_*/2, \ 2t_*] \} \to 1, \tag{26}$$

with t_* the output of the dyadic search (i.e. line 4 in Algorithm 2), provided that the constant C_0 in (10a) is sufficiently large. One possible choice is

$$C_0 = 2^{18} \cdot 3^2. \tag{27}$$

Since $0 < \lambda \leq 1/2$, there exists an integer $k_0 \geq 1$ such that $\lambda \in (2^{-k_0-1}, 2^{-k_0}]$. Then (26) is equivalent to

$$\mathbb{P} \left\{ t_* = 2^{-k_0-1}n \ \text{ or } \ 2^{-k_0}n \right\} \to 1.$$

Thus, we only need to show that

$$\mathbb{P} \{ \mathcal{C}(t_{k_0+1}, t_k) \leq 0 \quad \text{for some} \ \ k = k_0 + 2, \ldots, \log_2 n \} \to 0, \tag{28a}$$
$$\text{and} \quad \mathbb{P} \{ \mathcal{C}(t_{k_0}, t_k) \leq 0 \quad \text{for some} \ \ k = 1, \ldots, k_0 - 1 \} \to 0, \tag{28b}$$

where $\mathcal{C}(\cdot, \cdot) \equiv \mathcal{C}_{(0,n]}(\cdot, \cdot)$ is the comparison function defined in (9), and $t_k := 2^{-k}n$ for $k \in \{1, \ldots, \log_2 n\}$ the dyadic locations.

We first prove (28a). Note that for $k \in \{k_0 + 2, \ldots, \log_2 n\}$, i.e. on the left side of the change point, and for $j \in \{1, \ldots, p\}$,

$$\mathrm{CS}_j(t_{k_0+1}; \boldsymbol{F})^2 - \mathrm{CS}_j(t_k; \boldsymbol{F})^2 \geq \mathrm{CS}_j(t_{k_0+1}; \boldsymbol{F})^2 - \mathrm{CS}_j(t_{k_0+2}; \boldsymbol{F})^2$$
$$= n^2(1 - \lambda)^2 \delta_j^2 \frac{t_{k_0+1} - t_{k_0+2}}{(n - t_{k_0+1})(n - t_{k_0+2})}$$

$$\geq (1-\lambda)^2 \delta_j^2 \frac{n\lambda}{4(1-\lambda/2)(1-\lambda/4)} \;\geq\; \frac{2}{21} n\lambda\delta_j^2,$$

where $\mathrm{CS}_j(\cdot;\boldsymbol{F}) \equiv \mathrm{CS}_{(0,n],j}(\cdot;\boldsymbol{F})$ is the CUSUM statistics in the j-th coordinate applied to the signal matrix $\boldsymbol{F}$. Thus, it holds that

$$\sum_{j\in L}\Big(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{F})^2 - \mathrm{CS}_j(t_k;\boldsymbol{F})^2\Big)$$

$$= \sum_{j\in L\cup S}\Big(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{F})^2 - \mathrm{CS}_j(t_k;\boldsymbol{F})^2\Big) - \sum_{j\in S}\Big(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{F})^2 - \mathrm{CS}_j(t_k;\boldsymbol{F})^2\Big)$$

$$\geq \frac{2}{21}n\lambda \sum_{j\in L\cup S}\delta_j^2 - \sum_{j\in S}\mathrm{CS}_j(t_{k_0+1};\boldsymbol{F})^2$$

$$\geq \frac{2}{21}n\lambda\,\|\boldsymbol{\delta}\|^2 - |S|\,\frac{1}{32}\frac{n\lambda\,\|\boldsymbol{\delta}^2\|}{s} \;\geq\; \frac{1}{16}n\lambda\,\|\boldsymbol{\delta}\|^2\,.$$

Fix an arbitrary $k \in \{k_0+2,\ldots,\log_2 n\}$, and let

$$B_j := \mathbb{1}\Big\{\max\big(|\mathrm{CS}_j(t_{k_0+1};\boldsymbol{Y})|\,,|\mathrm{CS}_j(t_k;\boldsymbol{Y})|\big) \geq \alpha\Big\},$$

for $j = 1,\ldots,p$. Introduce

$$x := \frac{1}{16}n\lambda\,\|\boldsymbol{\delta}\|^2\,, \tag{29}$$

and we have

$$\mathbb{P}\left\{\mathcal{C}(t_{k_0+1},t_k) \leq 0\right\}$$

$$\leq \mathbb{P}\left\{\sum_{j\in L\cup S\cup N} B_j\Big(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{X})^2 - \mathrm{CS}_j(t_k;\boldsymbol{X})^2\Big) \leq \sum_{j\in L}\Big(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{F})^2 - \mathrm{CS}_j(t_k;\boldsymbol{F})^2\Big) - x\right\}$$

$$\leq \mathbb{P}\left\{\sum_{j\in L} B_j\Big(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{X})^2 - \mathrm{CS}_j(t_k;\boldsymbol{X})^2\Big) \leq \sum_{j\in L}\Big(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{F})^2 - \mathrm{CS}_j(t_k;\boldsymbol{F})^2\Big) - \frac{x}{3}\right\} \tag{30a}$$

$$+ \mathbb{P}\left\{\sum_{j\in S} B_j\Big(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{X})^2 - \mathrm{CS}_j(t_k;\boldsymbol{X})^2\Big) \leq -\frac{x}{3}\right\} \tag{30b}$$

$$+ \mathbb{P}\left\{\sum_{j\in N} B_j\Big(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{X})^2 - \mathrm{CS}_j(t_k;\boldsymbol{X})^2\Big) \leq -\frac{x}{3}\right\}. \tag{30c}$$

Next we bound the probabilities in (30a), (30b) and (30c), separately.

i. The probability in (30a) can be bounded from above by

$$\mathbb{P}\left\{\sum_{j\in L} B_j\mathrm{CS}_j(t_{k_0+1};\boldsymbol{X})^2 - \sum_{j\in L}\Big(1 + \mathrm{CS}_j(t_{k_0+1};\boldsymbol{F})^2\Big) \leq -\frac{x}{6}\right\}$$

$$+ \mathbb{P}\left\{\sum_{j \in L} B_j \mathrm{CS}_j(t_k; \boldsymbol{X})^2 - \sum_{j \in L}\left(1 + \mathrm{CS}_j(t_k; \boldsymbol{F})^2\right) \geq \frac{x}{6}\right\}. \quad (31)$$

For the first term in (31), we consider two cases separately.

- Dense case, i.e. when $s \geq \sqrt{p \log \log n}$ or $s = p$. It follows that $\alpha = 1$ and thus $B_j = 1$ for $j \in L$. By Lemma 10, we have

$$\mathbb{P}\left\{\sum_{j \in L} B_j \mathrm{CS}_j(t_{k_0+1}; \boldsymbol{X})^2 - \sum_{j \in L}\left(1 + \mathrm{CS}_j(t_{k_0+1}; \boldsymbol{F})^2\right) \leq -\frac{x}{6}\right\}$$

$$= \mathbb{P}\left\{\sum_{j \in L} \mathrm{CS}_j(t_{k_0+1}; \boldsymbol{X})^2 - \sum_{j \in L}\left(1 + \mathrm{CS}_j(t_{k_0+1}; \boldsymbol{F})^2\right) \leq -\frac{x}{6}\right\}$$

$$\leq \exp\left(-\frac{x^2}{144\left(s + 2n\lambda\|\boldsymbol{\delta}\|^2\right)}\right).$$

- Sparse case, i.e. when $s < \sqrt{p \log \log n}$ and $s \neq p$. It implies $\alpha \geq 2$. Note that, for every $j \in \{1, \ldots, p\}$,

$$|\mathrm{CS}_j(t_{k_0+1}; \boldsymbol{F})|^2 \geq n\lambda\delta_j^2 \frac{(1-\lambda)^2}{1-\lambda/2} \geq \frac{1}{3}n\lambda\delta_j^2,$$

and that $n\lambda\|\boldsymbol{\delta}\|^2 \geq C_0 s\alpha^2/4 \geq C_0 s$. For $j \in L$, by C_0 in (27), we have

$$(|\mathrm{CS}_j(t_{k_0+1}; \boldsymbol{F})| - \alpha)^2 \geq \frac{2}{9}n\lambda\delta_j^2 \geq \frac{1}{6}\left(1 + n\lambda\delta_j^2\right),$$

and then, by Mill's ratio,

$$\mathbb{P}\left\{B_j = 0\right\} \leq \mathbb{P}\left\{|\mathrm{CS}_j(t_{k_0+1}; \boldsymbol{Y})| \leq \alpha\right\}$$

$$\leq \mathbb{P}\left\{\left|\mathrm{CS}_j(t_{k_0+1}; \tilde{\boldsymbol{\Xi}})\right| \geq |\mathrm{CS}_j(t_{k_0+1}; \boldsymbol{F})| - \alpha\right\}$$

$$\leq 2\left(1 - \Phi(|\mathrm{CS}_j(t_{k_0+1}; \boldsymbol{F})| - \alpha)\right) \leq \exp\left(-\frac{(|\mathrm{CS}_j(t_{k_0+1}; \boldsymbol{F})| - \alpha)^2}{2}\right)$$

$$\leq \exp\left(-\frac{1}{12}\left(1 + n\lambda\delta_j^2\right)\right).$$

Recall that $\Phi(\cdot)$ denotes the distribution function of standard Gaussian random variable. Thus, by Lemma 13, we obtain

$$\mathbb{P}\left\{\sum_{j \in L} B_j \mathrm{CS}_j(t_{k_0+1}; \boldsymbol{X})^2 - \sum_{j \in L}\left(1 + \mathrm{CS}_j(t_{k_0+1}; \boldsymbol{F})^2\right) \leq -\frac{x}{6}\right\}$$

$$\leq 2^s \exp\left(-\min\left\{\frac{x}{288}, \frac{x^2}{144\left(s + 2n\lambda\|\boldsymbol{\delta}\|^2\right)}\right\}\right).$$

For the second term in (31), because $\mathbb{P}\{B_j = 1\} \geq 1/2$ for every $j \in L$, we apply (19) in Lemma 11 and obtain

$$\mathbb{P}\left\{\sum_{j\in L}B_j\mathrm{CS}_j(t_k;\boldsymbol{X})^2 - \sum_{j\in L}\left(1 + \mathrm{CS}_j(t_k;\boldsymbol{F})^2\right) \geq \frac{x}{6}\right\}$$

$$\leq \mathbb{P}\left\{\sum_{j\in L}B_j\mathrm{CS}_j(t_k;\boldsymbol{X})^2 - \sum_{j\in L}B_j\left(1 + \mathrm{CS}_j(t_k;\boldsymbol{F})^2\right) \geq \frac{x}{6}\right\}$$

$$\leq \exp\left(-\min\left\{\frac{x}{48}, \frac{x^2}{576\left(s + 2n\lambda\,\|\boldsymbol{\delta}\|^2\right)}\right\}\right).$$

ii. We split the probability in (30b) according to all possible values of $(B_j)_{j\in S}$, namely,

$$\mathbb{P}\left\{\sum_{j\in S}B_j\left(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{X})^2 - \mathrm{CS}_j(t_k;\boldsymbol{X})^2\right) \leq -\frac{x}{3}\right\}$$

$$= \sum_{\emptyset\neq J\subset S}\left(\mathbb{P}\left\{\sum_{j\in J}B_j\left(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{X})^2 - \mathrm{CS}_j(t_k;\boldsymbol{X})^2\right) \leq -\frac{x}{3}\,\bigg|\,B_j = 1 \text{ iff } j \in J\right\}\right.$$

$$\left.\times\,\mathbb{P}\{B_j = 1 \text{ iff } j \in J\}\right)$$

$$= \sum_{\emptyset\neq J\subset S}\mathbb{P}\left\{\sum_{j\in J}\left(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{X})^2 - \mathrm{CS}_j(t_k;\boldsymbol{X})^2\right) \leq -\frac{x}{3}\right\}\mathbb{P}\{B_j = 1 \text{ iff } j \in J\}.$$

For every $\emptyset\neq J\subset S$, we apply Lemma 10 (cf. Remark 17) and obtain

$$\mathbb{P}\left\{\sum_{j\in J}\left(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{X})^2 - \mathrm{CS}_j(t_k;\boldsymbol{X})^2\right) \leq -\frac{x}{3}\right\}$$

$$\leq \mathbb{P}\left\{\sum_{j\in J}\left(\left(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{X})^2 - \mathrm{CS}_j(t_k;\boldsymbol{X})^2\right) - \left(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{F})^2 - \mathrm{CS}_j(t_k;\boldsymbol{F})^2\right)\right) \leq -\frac{x}{3}\right\}$$

$$\leq \mathbb{P}\left\{\sum_{j\in J}\left(\mathrm{CS}_j(t_{k_0+1};\boldsymbol{X})^2 - \mathrm{CS}_j(t_{k_0+1};\boldsymbol{F})^2\right) \leq -\frac{x}{6}\right\}$$

$$+ \mathbb{P}\left\{\sum_{j\in J}\left(\mathrm{CS}_j(t_k;\boldsymbol{X})^2 - \mathrm{CS}_j(t_k;\boldsymbol{F})^2\right) \geq \frac{x}{6}\right\}$$

$$\leq 2\exp\left(-\min\left\{\frac{x}{48}, \frac{x^2}{288(s + 2n\lambda\,\|\boldsymbol{\delta}\|^2\,/21)}\right\}\right).$$

45

Thus, we have

$$
\mathbb{P}\left\{\sum_{j \in J}\left(\mathrm{CS}_j(t_{k_0+1}; \boldsymbol{X})^2 - \mathrm{CS}_j(t_k; \boldsymbol{X})^2\right) \leq -\frac{x}{3}\right\}
$$

$$
\leq \sum_{\emptyset \neq J \subset S} 2\exp\left(-\min\left\{\frac{x}{48}, \frac{x^2}{288(s + 2n\lambda \|\boldsymbol{\delta}\|^2 / 21)}\right\}\right) \mathbb{P}\left\{B_j = 1 \text{ iff } j \in J\right\}
$$

$$
= 2\exp\left(-\min\left\{\frac{x}{48}, \frac{x^2}{288(s + 2n\lambda \|\boldsymbol{\delta}\|^2 / 21)}\right\}\right).
$$

iii. Consider the probability in (30c), and note that for $j \in N$,

$$
\mathbb{P}\left\{B_j = 1\right\} \leq \mathbb{P}\left\{\left|\mathrm{CS}_j(t_{k_0+1}; \tilde{\boldsymbol{\Xi}})\right| \geq \alpha\right\} + \mathbb{P}\left\{\left|\mathrm{CS}_j(t_k; \tilde{\boldsymbol{\Xi}})\right| \geq \alpha\right\}
$$

$$
= 2\big(1 - \Phi(\alpha)\big) \leq \exp\left(-\frac{\alpha^2}{2}\right).
$$

We apply Lemma 11, more precisely, (18), and obtain

$$
\mathbb{P}\left\{\sum_{j \in N} B_j\left(\mathrm{CS}_j(t_{k_0+1}; \boldsymbol{X})^2 - \mathrm{CS}_j(t_k; \boldsymbol{X})^2\right) \leq -\frac{x}{3}\right\}
$$

$$
= \mathbb{P}\left\{\sum_{j \in N} B_j\left(\mathrm{CS}_j(t_{k_0+1}; \boldsymbol{\Xi})^2 - \mathrm{CS}_j(t_k; \boldsymbol{\Xi})^2\right) \leq -\frac{x}{3}\right\}
$$

$$
\leq \mathbb{P}\left\{\sum_{j \in N} B_j\left(\mathrm{CS}_j(t_{k_0+1}; \boldsymbol{\Xi})^2 - 1\right) \leq -\frac{x}{3}\right\} + \mathbb{P}\left\{\sum_{j \in N} B_j\left(1 - \mathrm{CS}_j(t_k; \boldsymbol{\Xi})^2\right) \leq -\frac{x}{3}\right\}
$$

$$
\leq 2\exp\left(-\min\left\{\frac{x}{48}, \frac{x^2}{432 \sum_{j \in N} \mathbb{P}\{B_j = 1\}}\right\}\right)
$$

$$
\leq 2\exp\left(-\min\left\{\frac{x}{48}, \frac{x^2}{432\, p \exp(-\alpha^2/2)}\right\}\right).
$$

Therefore, combining all the bounds above and (29), we obtain

$$
\mathbb{P}\left\{\mathcal{C}(t_{k_0+1}, t_k) \leq 0\right\}
$$

$$
\leq 2\exp\left(-\frac{x^2}{432\, p \exp(-\alpha^2/2)}\right)
$$

$$
\qquad + \big(6 + 2^s \mathbb{1}\{\alpha > 0\}\big)\exp\left(-\min\left\{\frac{x}{288}, \frac{x^2}{576(s + 2n\lambda \|\boldsymbol{\delta}\|^2)}\right\}\right)
$$

$$
= \exp\left(-\frac{n^2\lambda^2 \|\boldsymbol{\delta}\|^4}{2^{12}3^3 p \exp(-\alpha^2/2)}\right)
$$

$$+ \left(6 + 2^s \mathbb{1}\left\{\alpha > 0\right\}\right) \exp\left(-\min\left\{\frac{n\lambda \|\boldsymbol{\delta}\|^2}{2^9 3^2}, \frac{n^2\lambda^2 \|\boldsymbol{\delta}\|^4}{2^{14} 3^2 \left(s + 2n\lambda \|\boldsymbol{\delta}\|^2\right)}\right\}\right),$$

for any fixed $k \in \{k_0 + 2, \ldots, \log_2 n\}$. This together with the union bound (i.e. Boole's inequality) implies

$$\mathbb{P}\left\{\mathcal{C}(t_{k_0+1}, t_k) \leq 0 \text{ for some } k = k_0 + 2, \ldots, \log_2 n\right\}$$

$$\leq \sum_{k=k_0+2}^{\log_2 n} \mathbb{P}\left\{\mathcal{C}(t_{k_0+1}, t_k) \leq 0\right\} \leq \frac{12}{\log n} \to 0,$$

where we use

$$p \exp\left(-\frac{\alpha^2}{2}\right) = \begin{cases} p & \text{if } \alpha = 0, \\ \frac{s^2}{e^2 \log\log n} & \text{if } \alpha > 0, \end{cases} \tag{32}$$

and the assumption (10) with C_0 in (27).

Next, we consider (28b). For $k \in \{1, \ldots, k_0 - 1\}$, i.e. on the right side of the change point, and for $j \in \{1, \ldots, p\}$, it holds that

$$\mathrm{CS}_j(t_{k_0}; \boldsymbol{F})^2 - \mathrm{CS}_j(t_k; \boldsymbol{F})^2 \geq \mathrm{CS}_j(t_{k_0}; \boldsymbol{F})^2 - \mathrm{CS}_j(t_{k_0-1}; \boldsymbol{F})^2 = \frac{1}{2t_{k_0}} n^2 \lambda^2 \delta_j^2$$

$$\geq \frac{1}{4} n\lambda \delta_j^2 \geq \frac{2}{21} n\lambda \delta_j^2,$$

and also that

$$|\mathrm{CS}_j(t_{k_0}; \boldsymbol{F})|^2 = n\lambda^2 \delta_j^2 \frac{n - t_{k_0}}{t_{k_0}} \geq \frac{1}{4} n\lambda \delta_j^2.$$

Thus, for $j \in L$, we have

$$\left(|\mathrm{CS}_j(t_{k_0+1}; \boldsymbol{F})| - \alpha\right)^2 \geq \frac{2}{9} n\lambda \delta_j^2 \geq \frac{1}{6}\left(1 + n\lambda \delta_j^2\right).$$

Therefore, the same calculation as for (28a) remain valid, and thus

$$\mathbb{P}\left\{\mathcal{C}(t_{k_0}, t_k) \leq 0 \text{ for some } k = 1, \ldots, k_0 - 1\right\} \leq \frac{12}{\log n} \to 0.$$

Part 2. Advanced optimistic search. We will show that the assertion of Theorem 6 holds with C_0 given in (27) and

$$C_1 = 2^{29}\, 3^4. \tag{33}$$

Based on Part 1, we can start from the search interval $(t_*/2, 2t_*]$, which contains $n\lambda$ with probability towards one as $n \to \infty$.

We only need to consider the pairs of probe points r, t at the same side of $n\lambda$, since otherwise no matter which side is dropped off, no mistake will occur. Thus, we introduce

$$\mathcal{P} = \Big\{(t, r) \ : \ t, r \text{ are the pair of probe points in the same step of}$$

$$\text{optimistic search such that } t, r \text{ lie on the same side of}$$

47

the change point, and t is closer to the change point$\Big\}$.

That is, for $(t, r) \in \mathcal{P}$, it holds $(t - n\lambda)(r - n\lambda) \geq 0$ and $|t - n\lambda| \leq |r - n\lambda|$.

Consider arbitrarily $(t, r) \in \mathcal{P}$. Recall that the step size of optimistic search is $\nu = 1/2$, and that the relative distance $d(\cdot, \cdot)$ is defined in (22). Since $2t_* \leq n$, it always holds

$$(r \vee t) \leq \frac{2}{3}\left(2t_* - \frac{1}{2}t_*\right) + \frac{1}{2}t_* \leq \frac{3}{2}t_* \leq \frac{3}{4}n.$$

Then, using $n\lambda/2 \leq t_* \leq 2n\lambda$, we have

$$\frac{|t - r|}{4n\lambda} \leq \frac{|t - r|}{2t_*} \leq \frac{|t - r|}{r \vee t} \leq d(t, r) = \frac{|t - r|}{(t \vee r) \wedge (n - (r \wedge t))}$$

$$\leq \frac{|t - r|}{(t_*/2) \wedge (n - 3n/4)} \leq \frac{4|t - r|}{n\lambda}, \quad (34)$$

that is, $d(t, r) \asymp |t - r|/(n\lambda)$.

Given $(t, r) \in \mathcal{P}$, there are only two possible cases:

- Case: $n\lambda \leq t \leq r \leq 2t_*$. For $j \in \{1, \ldots, p\}$, we have

$$|t - r|\delta_j^2 \geq \mathrm{CS}_j(t; \boldsymbol{F})^2 - \mathrm{CS}_j(r; \boldsymbol{F})^2 = n^2\lambda^2\delta_j^2\frac{r - t}{tr} \geq \frac{1}{16}|t - r|\delta_j^2,$$

  where the second inequality is due to $t \leq r \leq 2t_* \leq 4n\lambda$.

- Case: $t_*/2 < r \leq t \leq n\lambda$. For $j \in \{1, \ldots, p\}$, we have

$$|t - r|\delta_j^2 \geq \mathrm{CS}_j(t; \boldsymbol{F})^2 - \mathrm{CS}_j(r; \boldsymbol{F})^2 = n^2(1 - \lambda)^2\delta_j^2\frac{t - r}{(n - t)(n - r)}$$

$$\geq \frac{(1 - \lambda)^2}{(1 - \lambda/4)^2}|t - r|\delta_j^2 \geq \frac{16}{49}|t - r|\delta_j^2 \geq \frac{1}{16}|t - r|\delta_j^2,$$

  where the second inequality is due to $n\lambda/4 \leq t_*/2 < r \leq t$, and the third inequality is due to $\lambda \leq 1/2$.

Thus, it follows

$$\sum_{j \in L}\left(\mathrm{CS}_j(t; \boldsymbol{F})^2 - \mathrm{CS}_j(r; \boldsymbol{F})^2\right)$$

$$= \sum_{j \in L \cup S}\left(\mathrm{CS}_j(t; \boldsymbol{F})^2 - \mathrm{CS}_j(r; \boldsymbol{F})^2\right) - \sum_{j \in S}\left(\mathrm{CS}_j(t; \boldsymbol{F})^2 - \mathrm{CS}_j(r; \boldsymbol{F})^2\right)$$

$$\geq \frac{1}{16}|t - r|\sum_{j \in L \cup S}\delta_j^2 - |t - r|\sum_{j \in S}\delta_j^2 \geq \frac{1}{16}|t - r|\|\boldsymbol{\delta}\|^2 - |S||t - r|\frac{\|\boldsymbol{\delta}\|^2}{32s} \geq \frac{1}{32}|t - r|\|\boldsymbol{\delta}\|^2.$$

Fix a pair of probe points $(t, r) \in \mathcal{P}$, and introduce

$$x := \frac{1}{32}|t - r|\|\boldsymbol{\delta}\|^2, \quad (35)$$

and, for $j \in \{1, \ldots, p\}$,

$$B_j := \mathbb{1}\Big\{\max\big(|\mathrm{CS}_j(t; \boldsymbol{Y})|, |\mathrm{CS}_j(r; \boldsymbol{Y})|\big) \geq \alpha\Big\}.$$

Then we can bound the probability of making a mistake as below:

$$\mathbb{P}\{\mathcal{C}(t, r) \leq 0\}$$

$$\leq \mathbb{P}\Big\{\sum_{j \in L \cup S \cup N} B_j\Big(\mathrm{CS}_j(t; \boldsymbol{X})^2 - \mathrm{CS}_j(r; \boldsymbol{X})^2\Big) \leq \sum_{j \in L}\Big(\mathrm{CS}_j(t; \boldsymbol{F})^2 - \mathrm{CS}_j(r; \boldsymbol{F})^2\Big) - x\Big\}$$

$$\leq \mathbb{P}\Big\{\sum_{j \in L} B_j\Big(\mathrm{CS}_j(t; \boldsymbol{X})^2 - \mathrm{CS}_j(r; \boldsymbol{X})^2\Big) \leq \sum_{j \in L}\Big(\mathrm{CS}_j(t; \boldsymbol{F})^2 - \mathrm{CS}_j(r; \boldsymbol{F})^2\Big) - \frac{x}{3}\Big\} \quad \text{(36a)}$$

$$+ \mathbb{P}\Big\{\sum_{j \in S} B_j\Big(\mathrm{CS}_j(t; \boldsymbol{X})^2 - \mathrm{CS}_j(r; \boldsymbol{X})^2\Big) \leq -\frac{x}{3}\Big\} \quad \text{(36b)}$$

$$+ \mathbb{P}\Big\{\sum_{j \in N} B_j\Big(\mathrm{CS}_j(t; \boldsymbol{X})^2 - \mathrm{CS}_j(r; \boldsymbol{X})^2\Big) \leq -\frac{x}{3}\Big\}. \quad \text{(36c)}$$

The three terms above can be bounded in a similar way as (30a), (30b) and (30c) in Part 1, respectively, but we need to decouple the correlation between $\mathrm{CS}_j(t; \boldsymbol{X})$ and $\mathrm{CS}_j(r; \boldsymbol{X})$ by means of Lemmas 15 and 18. The details are given below.

i. For (36a), we consider two cases separately.

- Dense case, i.e. when $s \geq \sqrt{p \log \log n}$ or $s = p$. Note that $\alpha = 0$ and thus $B_j = 1$ for $j \in L$. Then Lemma 15 implies

$$\mathbb{P}\Big\{\sum_{j \in L} B_j\Big(\mathrm{CS}_j(t; \boldsymbol{X})^2 - \mathrm{CS}_j(r; \boldsymbol{X})^2\Big) \leq \sum_{j \in L}\Big(\mathrm{CS}_j(t; \boldsymbol{F})^2 - \mathrm{CS}_j(r; \boldsymbol{F})^2\Big) - \frac{x}{3}\Big\}$$

$$= \mathbb{P}\Big\{\sum_{j \in L}\Big(\mathrm{CS}_j(t; \boldsymbol{X})^2 - \mathrm{CS}_j(r; \boldsymbol{X})^2\Big) \leq \sum_{j \in L}\Big(\mathrm{CS}_j(t; \boldsymbol{F})^2 - \mathrm{CS}_j(r; \boldsymbol{F})^2\Big) - \frac{x}{3}\Big\}$$

$$\leq 2 \exp\left(-\min\left\{\frac{x^2}{576\, d(t, r)\big(s + 2n\lambda \|\boldsymbol{\delta}\|^2\big)}, \frac{x}{48\sqrt{2d(t, r)}}\right\}\right).$$

- Sparse case, i.e. when $s < \sqrt{p \log \log n}$ and $s \neq p$. In this case, it holds that $\alpha \geq 2$, and thus that $n\lambda \|\boldsymbol{\delta}\|^2 \geq C_0 s \alpha^2 / 4 \geq C_0 s$. Because t is closer the change point and $t \in [t_*, 3t_*/2]$ (which is due to $\nu = 1/2$), we obtain, for $j \in \{1, \ldots, p\}$,

$$|\mathrm{CS}_j(t; \boldsymbol{F})|^2 \geq n\lambda^2 \delta_j^2 \frac{n-t}{t} \geq \frac{1}{4} n\lambda \delta_j^2.$$

Then, the choice of C_0 in (27) implies, for $j \in L$,

$$(|\mathrm{CS}_j(t; \boldsymbol{F})| - \alpha)^2 \geq \frac{2}{9} n\lambda \delta_j^2 \geq \frac{1}{6}\big(1 + n\lambda \delta_j^2\big),$$

49

and further, by Mill's ratio,

$$\mathbb{P}\left\{B_j = 0\right\} \leq \mathbb{P}\left\{|\mathrm{CS}_j(t; \boldsymbol{Y})| \leq \alpha\right\} \leq \mathbb{P}\left\{\left|\mathrm{CS}_j(t; \tilde{\boldsymbol{\Xi}})\right| \geq |\mathrm{CS}_j(t; \boldsymbol{F})| - \alpha\right\}$$

$$\leq 2\big(1 - \Phi(|\mathrm{CS}_j(t; \boldsymbol{F})| - \alpha)\big)$$

$$\leq \exp\left(-\frac{(|\mathrm{CS}_j(t; \boldsymbol{F})| - \alpha)^2}{2}\right)$$

$$\leq \exp\left(-\frac{1}{12}\big(1 + n\lambda\delta_j^2\big)\right).$$

Thus, by Lemma 18iii, we obtain

$$\mathbb{P}\left\{\sum_{j \in L} B_j\Big(\mathrm{CS}_j(t; \boldsymbol{X})^2 - \mathrm{CS}_j(r; \boldsymbol{X})^2\Big) \leq \sum_{j \in L}\Big(\mathrm{CS}_j(t; \boldsymbol{F})^2 - \mathrm{CS}_j(r; \boldsymbol{F})^2\Big) - \frac{x}{3}\right\}$$

$$\leq 2\exp\left(-\min\left\{\frac{x^2}{576\, d(t,r)\,(s + 2n\lambda\|\boldsymbol{\delta}\|^2)}, \frac{x}{288\sqrt{2d(t,r)}}\right\}\right)$$

$$\times \prod_{j \in L}\left(1 + \exp\Big(-\frac{1}{24}\big(1 + n\lambda\delta_j^2\big)\Big)\right)$$

$$\leq 2\exp\left(-\min\left\{\frac{x^2}{576\, d(t,r)\,(s + 2n\lambda\|\boldsymbol{\delta}\|^2)}, \frac{x}{288\sqrt{2d(t,r)}}\right\}\right.$$

$$\left. + s\exp\Big(-\frac{1}{768}\frac{n\lambda\|\boldsymbol{\delta}\|^2}{s}\Big)\right),$$

where the last inequality is due to the definition of L and the basic inequality $1 + x \leq \exp(x)$.

ii. We decompose the probability in (36b) into events conditioned on all possible values of $(B_j)_{j \in S}$, namely,

$$\mathbb{P}\left\{\sum_{j \in S} B_j\Big(\mathrm{CS}_j(t; \boldsymbol{X})^2 - \mathrm{CS}_j(r; \boldsymbol{X})^2\Big) \leq -\frac{x}{3}\right\}$$

$$= \sum_{\emptyset \neq J \subset S}\left(\mathbb{P}\left\{\sum_{j \in J} B_j\Big(\mathrm{CS}_j(t; \boldsymbol{X})^2 - \mathrm{CS}_j(r; \boldsymbol{X})^2\Big) \leq -\frac{x}{3} \,\middle|\, B_j = 1 \text{ iff } j \in J\right\}\right.$$

$$\left. \times \mathbb{P}\left\{B_j = 1 \text{ iff } j \in J\right\}\right)$$

$$= \sum_{\emptyset \neq J \subset S} \mathbb{P}\left\{\sum_{j \in J}\Big(\mathrm{CS}_j(t; \boldsymbol{X})^2 - \mathrm{CS}_j(r; \boldsymbol{X})^2\Big) \leq -\frac{x}{3}\right\}\mathbb{P}\left\{B_j = 1 \text{ iff } j \in J\right\}.$$

For every $\emptyset \neq J \subset S$, we obtain, by Lemma 15,

$$\mathbb{P}\left\{\sum_{j \in J}\Big(\mathrm{CS}_j(t; \boldsymbol{X})^2 - \mathrm{CS}_j(r; \boldsymbol{X})^2\Big) \leq -\frac{x}{3}\right\}$$

$$\leq \mathbb{P}\left\{\sum_{j\in J}\left(\left(\mathrm{CS}_j(t;\boldsymbol{X})^2 - \mathrm{CS}_j(r\boldsymbol{X})^2\right) - \left(\mathrm{CS}_j(t;\boldsymbol{F})^2 - \mathrm{CS}_j(r;\boldsymbol{F})^2\right)\right) \leq -\frac{x}{3}\right\}$$

$$\leq 2\exp\left(-\min\left\{\frac{x}{48\sqrt{2d(t,r)}},\ \frac{x^2}{576\,d(t,r)(s + n\lambda\,\|\boldsymbol{\delta}\|^2/16)}\right\}\right).$$

Thus, we have

$$\mathbb{P}\left\{\sum_{j\in J}\left(\mathrm{CS}_j(t;\boldsymbol{X})^2 - \mathrm{CS}_j(r;\boldsymbol{X})^2\right) \leq -\frac{x}{3}\right\}$$

$$\leq \sum_{\emptyset\neq J\subset S}\left(2\exp\left(-\min\left\{\frac{x}{48\sqrt{2d(t,r)}},\frac{x^2}{576\,d(t,r)(s+n\lambda\|\boldsymbol{\delta}\|^2)}\right\}\right)\mathbb{P}\left\{B_j=1 \text{ iff } j\in J\right\}\right)$$

$$= 2\exp\left(-\min\left\{\frac{x}{48\sqrt{2d(t,r)}},\frac{x^2}{576\,d(t,r)\left(s+n\lambda\|\boldsymbol{\delta}\|^2\right)}\right\}\right).$$

iii. For (36c), we first note that for $j \in N$,

$$\mathbb{P}\left\{B_j=1\right\} \leq \mathbb{P}\left\{\left|\mathrm{CS}_j(t;\tilde{\boldsymbol{\Xi}})\right| \geq \alpha\right\} + \mathbb{P}\left\{\left|\mathrm{CS}_j(r;\tilde{\boldsymbol{\Xi}})\right| \geq \alpha\right\}$$

$$= 2\big(1-\Phi(\alpha)\big) \leq \exp\left(-\frac{\alpha^2}{2}\right).$$

Then, Lemma 18i implies

$$\mathbb{P}\left\{\sum_{j\in N}B_j\left(\mathrm{CS}_j(t;\boldsymbol{X})^2 - \mathrm{CS}_j(r;\boldsymbol{X})^2\right) \leq -\frac{x}{3}\right\}$$

$$= \mathbb{P}\left\{\sum_{j\in N}B_j\left(\mathrm{CS}_j(t;\boldsymbol{\Xi})^2 - \mathrm{CS}_j(r;\boldsymbol{\Xi})^2\right) \leq -\frac{x}{3}\right\}$$

$$\leq 2\exp\left(-\min\left\{\frac{x}{48\sqrt{2d(t,r)}},\frac{x^2}{864\,d(t,r)\sum_{j\in N}\mathbb{P}\left\{B_j=1\right\}}\right\}\right)$$

$$\leq 2\exp\left(-\min\left\{\frac{x}{48\sqrt{2d(t,r)}},\frac{x^2}{864\,d(t,r)p\exp(-\alpha^2/2)}\right\}\right).$$

Thus, we combine all bounds above and obtain, for $(t,r)\in\mathcal{P}$,

$$\mathbb{P}\left\{\mathcal{C}(t,r)\leq 0\right\}$$

$$\leq 2\exp\left(-\frac{x^2}{864\,d(t,r)p\exp(-\alpha^2/2)}\right) + \left(4 + 2\exp\big(se^{-n\lambda\|\boldsymbol{\delta}\|^2/(768\,s)}\mathbb{1}\left\{\alpha>0\right\}\big)\right)$$

$$\times\exp\left(-\min\left\{\frac{x}{288\sqrt{2d(t,r)}},\frac{x^2}{576\,d(t,r)\left(s+2n\lambda\|\boldsymbol{\delta}\|^2\right)}\right\}\right)$$

$$\leq 2 \exp\left(-\frac{n\lambda \left| t - r \right| \|\boldsymbol{\delta}\|^4}{2^{17} 3^3 p \exp(-\alpha^2/2)} \right) + \left(4 + 2 \exp\left(s e^{-n\lambda \|\boldsymbol{\delta}\|^2/(768\,s)} \mathbb{1}\left\{ \alpha > 0 \right\} \right) \right)$$

$$\times \exp\left(-\min\left\{ \frac{\sqrt{n\lambda \left| t - r \right|} \|\boldsymbol{\delta}\|^2}{2^{11} 3^2 \sqrt{2}}, \; \frac{n\lambda \left| t - r \right| \|\boldsymbol{\delta}\|^4}{2^{18} 3^2 \left(s + 2n\lambda \|\boldsymbol{\delta}\|^2 \right)} \right\} \right),$$

where the last inequality is due to (34) and (35).

We claim that there are at most $O\left((\log n)^2\right)$ possible choices for such pairs of probe points r, t. More precisely, we will show that there exists another set $\mathcal{P}_*$ such that $\mathcal{P} \subseteq \mathcal{P}_*$ with $|\mathcal{P}_*| = O\left((\log n)^2\right)$, and $\mathcal{P}_*$ is completely determined by the signal $\boldsymbol{F}$, in particular, independent of the noise $\boldsymbol{\Xi}$. We will prove the claim in two steps.

First, we note that in each step of (naive) optimistic search, at least $1/4$ of the interval is dropped off. Thus, the optimistic search stops with at most

$$\frac{\log n}{\log(4/3)} \leq 4 \log n \quad \text{steps.}$$

However, in some steps (e.g. two probe points lying on both side of $n\lambda$ and with the same or similar distance to $n\lambda$), which part to drop off may depend not only on the signal $\boldsymbol{F}$ but also on the noise $\boldsymbol{\Xi}$. In this case, we will include both pairs of probe points that are followed by deleting the leftmost part and the rightmost part, respectively, as possible choices of probe points. This results in a bigger set of pairs of probe points than $\mathcal{P}$, which is denoted by $\mathcal{P}_*$.
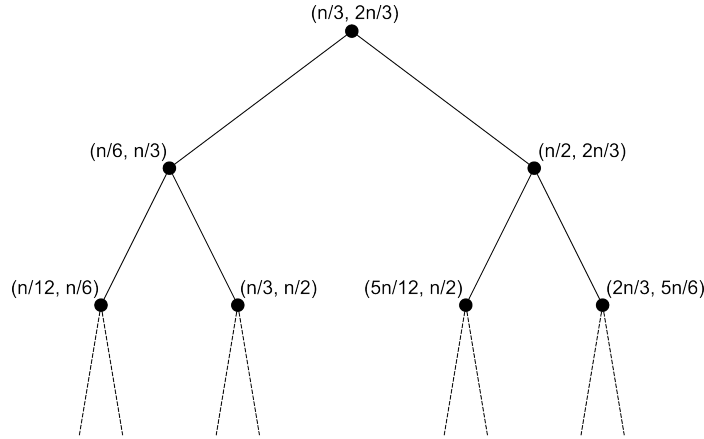

Figure 6: Binary tree structure of (naive) optimistic searches. The starting search interval is $(0, n]$. Each node represents a pair of probe points (t, r). The level of depth corresponds to the number of steps in optimistic searches.

Second, we record all possible choices of pairs of probe points by a binary tree, see Figure 6 for an illustration. Let us look into the binary tree: at each level (which consists of all nodes that have the same depth to the root), there is at most one node for which the corresponding pair of probe points have similar distance to $n\lambda$. Namely, given the signal $\boldsymbol{F}$, at each level, there is at most one node that yields two child nodes, while for the rest it is

clear that which part needs dropped off. In this way, the possible choices of pairs of probe points can be restricted to a subtree that has at most $4 \log n$ width, and thus has at most

$$(4 \log n)^2 \quad \text{nodes.}$$

That is, $|\mathcal{P}_*| \leq (4 \log n)^2$, which shows the above claim.

Thus, by the union bound, we can bound the probability of wrongly excluding any segment for all pairs of probe points r, t with $|t - r| \geq \varepsilon$, for some $\varepsilon \equiv \varepsilon(n, p, s) > 0$.

$$\mathbb{P}\left\{ \mathcal{C}(t, r) \leq 0 \text{ for some } (t, r) \in \mathcal{P} \text{ such that } |t - r| \geq \varepsilon \right\}$$

$$\leq \mathbb{P}\left\{ \mathcal{C}(t, r) \leq 0 \text{ for some } (t, r) \in \mathcal{P}_* \text{ such that } |t - r| \geq \varepsilon \right\}$$

$$\leq \sum_{(t,r) \in \mathcal{P}_*, \, |t-r| \geq \varepsilon} \mathbb{P}\left\{ \mathcal{C}(t, r) \leq 0 \right\}$$

$$\leq \sum_{(t,r) \in \mathcal{P}_*, \, |t-r| \geq \varepsilon} \left(2 \exp\left(-\frac{n\lambda \, |t - r| \, \|\boldsymbol{\delta}\|^4}{2^{17} 3^3 p \exp(-\alpha^2/2)} \right) + \left(4 + 2 \exp\left(s e^{-n\lambda \|\boldsymbol{\delta}\|^2/(768\,s)} \mathbb{1}\{\alpha > 0\} \right) \right) \right.$$

$$\left. \times \exp\left(-\min\left\{ \frac{\sqrt{n\lambda \, |t - r|} \, \|\boldsymbol{\delta}\|^2}{2^{11} 3^2 \sqrt{2}}, \, \frac{n\lambda \, |t - r| \, \|\boldsymbol{\delta}\|^4}{2^{18} 3^2 \left(s + 2n\lambda \, \|\boldsymbol{\delta}\|^2 \right)} \right\} \right) \right)$$

$$\leq (4 \log n)^2 \left(2 \exp\left(-\frac{n\lambda\varepsilon \, \|\boldsymbol{\delta}\|^4}{2^{17} 3^3 p \exp(-\alpha^2/2)} \right) + \left(4 + 2 \exp\left(s e^{-n\lambda \|\boldsymbol{\delta}\|^2/(768\,s)} \mathbb{1}\{\alpha > 0\} \right) \right) \right.$$

$$\left. \times \exp\left(-\min\left\{ \frac{\sqrt{n\lambda\varepsilon} \, \|\boldsymbol{\delta}\|^2}{2^{11} 3^2 \sqrt{2}}, \, \frac{n\lambda\varepsilon \, \|\boldsymbol{\delta}\|^4}{2^{18} 3^2 \left(s + 2n\lambda \, \|\boldsymbol{\delta}\|^2 \right)} \right\} \right) \right).$$

Thus, under assumption (10) with C_0 in (27), if

$$\varepsilon \geq \varepsilon_* := \frac{C_1}{2} \max\left\{ \frac{\log \log n}{\|\boldsymbol{\delta}\|^2}, \, \frac{\min\{s^2, p \log \log n\}}{n\lambda \, \|\boldsymbol{\delta}\|^4} \right\}, \tag{37}$$

with C_1 in (33), then by $\max_{x \geq 0} x e^{-x} = e^{-1}$ and (32) we obtain

$$\mathbb{P}\left\{ \mathcal{C}(t, r) \leq 0 \text{ for some } (t, r) \in \mathcal{P} \text{ such that } |t - r| \geq \varepsilon \right\} \leq \frac{128}{\log \log n} \to 0.$$

That is, when the probe points t, r satisfy $|t - r| \geq \varepsilon_*$, no wrong dropping-off will occur, except on an event with asymptotically vanishing probability. As a consequence, it holds with probability tending to one that $|\hat{\tau} - \lambda| \leq 2\varepsilon_*/n$, which concludes the proof. ∎

Remark 19 (Rates of probability towards one) *In fact, we have shown*

$$1 - \mathbb{P}\left\{ |\hat{\tau} - \tau| \leq C_1 \max\left\{ \frac{\log \log n}{n \, \|\boldsymbol{\delta}\|^2}, \, \frac{\min\{s^2, p \log \log n\}}{n^2 \lambda \, \|\boldsymbol{\delta}\|^4} \right\} \right\} \leq \frac{140}{\log n},$$

in the proof above. Such a rate can be improved to any polynomials of $1/\log n$ if constant C_0 is chosen larger than the one in (27).

Remark 20 (Slightly weaker assumption) *If the dyadic grid search in advanced optimistic search is modified to start from $k = 1$ (or equivalently, $t_1 = n/2$) on, and to stop at the first k_0 such that $\mathcal{C}(t_{k_0}, t_{k_0+1}) > 0$ and $\mathcal{C}(t_{k_0}, t_{k_0-1}) > 0$, then we only need to control random perturbations at pairs of dyadic points of maximal number $O\big(\log(2/\lambda)\big)$. This, together with slight modification of the proof, will allow a weaker condition*

$$
n\lambda \left\|\boldsymbol{\delta}\right\|^2 \gtrsim \begin{cases} \sqrt{p \log \log(2/\lambda)} & \text{if } s \geq \sqrt{p \log \log(2/\lambda)}, \\ \max\left(s \log \frac{e\sqrt{p \log \log(2/\lambda)}}{s}, \log \log n \right) & \text{if } s \leq \sqrt{p \log \log(2/\lambda)}. \end{cases}
$$

A similar condition appeared also in Pilliat et al. (2023).

Remark 21 (Slightly higher accuracy) *Note that in the proof we only need to control the random perturbations at the pairs of probe points until the length of search interval is ε, which leads to an upper bound $O\big(\log(n/\varepsilon)^2\big)$ on the number of pairs of probe points. This will lead to a slightly better accuracy of order*

$$
\min\left\{ \gamma > 0 \ : \ \gamma \geq \frac{\log \log(n/\gamma)}{\left\|\boldsymbol{\delta}\right\|^2} \ \text{and} \ \gamma \geq \frac{\min\{s^2, p \log \log(n/\gamma)\}}{n\lambda \left\|\boldsymbol{\delta}\right\|^4} \right\}.
$$

Further, the iterated log factor might be removable if one employs the sub-martingale property of the gain function and partitions the sampling locations into geometrically equally spaced segments, as detailed in Howard et al. (2021). This might lead to the essential multiplicity of $\mathrm{polylog}(\varepsilon)$ instead of order $(\log n)^2$, and then it would allow the removability of the iterated log factor. The careful examination will be left as part of future research.

Remark 22 (Combined optimistic search) *We note that Theorem 6 also holds for the combined optimistic search, introduced in Appendix A. This can be proven as follows: Let $\hat{t}_a$ and $\hat{t}_c$ be the outputs of advanced and combined optimistic search, respectively. Similar to the proof above, one can apply Lemmas 15 and 18 and show that*

$$
\mathbb{P}\Big\{ \mathcal{C}(\hat{t}_a, r) \leq 0 \text{ for some probe point } r \text{ in naive optimistic search s.t. } |r - n\lambda| \geq 6\varepsilon_* \Big\} \to 0,
$$

with ε_* defined in (37). Since $\mathcal{C}(\hat{t}_a, \hat{t}_c) \leq 0$, it holds that

$$
\mathbb{P}\left\{ \left| \hat{t}_c - n\lambda \right| \leq 6\varepsilon_* \right\} \to 1.
$$

E.3 Multiple Change Points

Theorem 5i is a special case of Theorem 7 for $p = 1$, so the proof of Theorem 5i is omitted.

The intuition why the OSeedBS in combination with the NOT selection performs optimally is that the selected seeded interval often contains only a single change point close to the center, due to the multiscale nature of seeded intervals. On such a seeded interval, naive and advanced (as well as combined) optimistic searches perform in a minimax optimal way.
Proof [of Theorem 7] We set the constant C_0 in assumption (13) as

$$
C_0 = 2^{34} 3^4,
$$

and the selection threshold $\gamma = C_1 \rho(n, p, s)$ with

$$C_1 = 2^{33} 3^4.$$

We will show that the assertion of the theorem holds with

$$C_2 = 2^{31} 3^4.$$

For notation simplicity, we assume there is no rounding, the decay $a = 1/2$ for seeded intervals, and the step size $\nu = 1/2$ for optimistic searches (otherwise, only the multiplying constants may be different). We split the rest of the proof into five steps.

Step 1. Consider intervals $\big((\tau_i - \lambda/2)n, \ (\tau_i + \lambda/2)n\big]$ as potential backgrounds for change points $\tau_i n$ for $i = 1, \ldots, \kappa$. By the construction of seeded intervals, we can find seeded intervals $(c_i - r_i, c_i + r_i]$ such that

$$(c_i - r_i, \ c_i + r_i] \subseteq \Big((\tau_i - \frac{1}{2}\lambda)n, \ (\tau_i + \frac{1}{2}\lambda)n\Big],$$

$$r_i \geq \frac{1}{6}\lambda n, \quad \text{and} \quad |c_i - \tau_i n| \leq \frac{1}{2}r_i.$$

Note that $|(c_i - r_i, \ c_i + r_i]| \geq \lambda n/3$ and that $(c_i - r_i, \ c_i + r_i]$ contains only a single change point $\tau_i n$ for every $i = 1, \ldots, \kappa$.

Step 2. By $\hat{\tau}_i^0$ we denote the estimated change point (scaled by $1/n$) by naive or advanced (or combined) optimistic search on $(c_i - r_i, \ c_i + r_i]$. Note that there is a single change point in $(c_i - r_i, \ c_i + r_i]$, which is closer to the center of the interval than boundaries, for every $i = 1, \ldots, \kappa$. Following similar lines as in the proof of Theorem 6, under the assumption in (13), we can show that

$$\mathbb{P}\left\{ \left|\hat{\tau}_i^0 - \tau_i\right| \leq \varepsilon_i^0 \equiv C_2^0 \left(\frac{\log n}{n\delta_i^2} \bigvee \frac{s^2 \wedge p \log n}{n^2 \lambda \delta_i^4} \right), \ i = 1, \ldots, \kappa \right\} \geq 1 - \frac{9}{n},$$

with $C_2^0 := 2^{22} 3^3$. One important difference in calculation to Appendix E.2 is the overall number of events that we need to control is now in polynomial of n rather than in polynomial of $\log n$. If $\left|\hat{\tau}_i^0 - \tau_i\right| \leq \varepsilon_i^0$, then it holds, for every $j \in \{1, \ldots, p\}$,

$$\mathrm{CS}_{(c_i - r_i, \ c_i + r_i], j}(\hat{\tau}_i^0 n; \boldsymbol{F})^2 \geq \frac{1}{16} n\lambda \delta_{i,j}^2 - \frac{3}{3 - 4\varepsilon_i^0} n\varepsilon_i^0 \delta_{i,j}^2 \geq \frac{1}{18} n\lambda \delta_{i,j}^2,$$

where the last inequality is due to $\varepsilon_i^0 \leq \lambda/432$. Thus, we obtain:

i. Dense case, i.e. when $s \geq \sqrt{p \log n}$ or $s = p$. The union bound and Lemma 10 imply

$$\mathbb{P}\left\{ G_{(c_i - r_i, \ c_i + r_i]}(\hat{\tau}_i^0 n) \geq \gamma, \ i = 1, \ldots, \kappa \right\} \geq 1 - \frac{10}{n}.$$

ii. Sparse case, i.e. when $s < \sqrt{p \log n}$ and $s \neq p$. Similar to Part 1 in the proof of Theorem 6, we split the set of coordinates into the set of large jumps, the set of small jumps and the set of no jumps, apply Lemmas 11 and 13, and then obtain

$$\mathbb{P}\left\{ G_{(c_i - r_i, \ c_i + r_i]}(\hat{\tau}_i^0 n) \geq \gamma, \ i = 1, \ldots, \kappa \right\} \geq 1 - \frac{12}{n}.$$

That is, for every $i \in \{1 \ldots, \kappa\}$, $G_{(c_i-r_i, c_i+r_i]}(\hat{\tau}_i^0 n)$ is above the selection threshold γ, with probability tending to one.

Step 3. Recall that the NOT selection rule selects the shortest seeded interval I_* which has a value of gain function above the threshold γ. Then it follows from Step 2, with probability tending to one, that the length of selected seeded interval $|I_*| \leq \lambda$, and thus that I_* contains at most one change point. By Lemmas 10 and 11 and the union bound, we consider dense and sparse cases separately, and then obtain

$$\mathbb{P}\left\{\max_{I \subseteq (0,n], I \text{ contains no change point}} \max_{t \in I} G_{(l,r]}(t) \leq 24\,\rho(n,p,s)\right\} \geq 1 - \frac{1}{6n}.$$

Since $\gamma \geq 24\,\rho(n,p,s)$, the selected seeded interval I_* contains exactly one change point, with probability tending to one.

Step 4. We rewrite the selected seeded interval I_* as $\big((\tau_i - u)n, (\tau_i + v)n\big]$ for some $i \in \{1, \ldots, \kappa\}$. We claim that, with probability tending to one,

$$\frac{uv}{u+v}n\delta_i^2 \geq \frac{\gamma}{2}. \tag{38}$$

Otherwise, we have $\max_{t \in I_*} \sum_{j=1}^p \text{CS}_{I_*,j}(t : \boldsymbol{F})^2 \leq \gamma/2$, and then, by the union bound, Lemma 10 and Remark 14, obtain (noting that $C_1 \geq 36$)

$$\mathbb{P}\left\{\max_{t \in I_*} G_{I_*}(t) \leq \gamma\right\} \geq 1 - \sum_{t \in I_*} \mathbb{P}\left\{G_{I_*}(t) \geq \gamma\right\}$$

$$\geq 1 - \sum_{t \in I_*} \mathbb{P}\left\{G_{I_*}(t) - \sum_{j=1}^p \text{CS}_{I_*,j}(t : \boldsymbol{F})^2 \geq \frac{\gamma}{2}\right\} \geq 1 - \frac{1}{n^2},$$

which contradicts with the fact that I_* is selected by NOT.

Let $\hat{\tau}_i^1$ be the estimated change point (scaled by $1/n$) by naive or advanced (or combined) optimistic search applied to I_*. Then, following similar calculations as in the proof of Theorem 6, we can see that (38) implies

$$\mathbb{P}\left\{\left|\hat{\tau}_i^1 - \tau_i\right| \leq \varepsilon_i^1 \equiv C_2^1\left(\frac{\log n}{n\delta_i^2} \bigvee \frac{s^2 \wedge p\log n}{n\gamma\delta_i^2}\right), \; i = 1, \ldots, \kappa\right\} \geq 1 - \frac{11}{n},$$

with $C_2^1 := 2^{32}3^4$. In particular, it holds that $\varepsilon_i^1 \leq \lambda/4$.

Step 5. Steps 1–4 imply that all change points $\{\tau_1, \ldots, \tau_\kappa\}$ can be estimated by $\{\hat{\tau}_1^1, \ldots, \hat{\tau}_\kappa^1\}$ with errors ε_i^1 and in particular $\hat{\kappa} \geq \kappa$, with probability tending to one. It remains to show that $\hat{\kappa} = \kappa$. To this end, we note that after the detection of κ change points, the remaining seeded intervals either contain no change point or contain one or two change points that are very close to the boundary. In the latter case, if a seeded interval contains say τ_i then we must have the distance of τ_i to one of the boundaries is no more than ε_i^1. This leads to

$$\max_{t \in I} \sum_{j=1}^p \text{CS}(t; \boldsymbol{F})^2 \leq \max_{1 \leq i \leq \kappa} n\varepsilon_i^1\delta_i^2 \leq \frac{1}{2}\gamma$$

56

for any remaining interval I. Thus, as in Step 4, we can show that the maximum value of gain function $\max_{t \in I} G_I(t)$ on every remaining seeded interval I is upper bounded by γ, more precisely,

$$\mathbb{P}\left\{\max_{I \in \mathcal{I}_r} \max_{t \in I} G_{(l,r]}(t) \leq \gamma\right\} \geq 1 - \frac{12}{n}.$$

where $\mathcal{I}_r$ denotes the collection of remaining seeded intervals. This further implies that OSeedBS will stop after κ steps, i.e. $\hat{\kappa} = \kappa$.

Step 6. Recall from Step 4 that $\varepsilon_i^1 \leq \lambda/4$. Then, with probability tending to one, it holds that each interval $I_i := \left(n(\hat{\tau}_{i-1}^1 + \hat{\tau}_i^1)/2, \; n(\hat{\tau}_i^1 + \hat{\tau}_{i+1}^1)/2\right]$ contains only one change point $n\tau_i$ such that $n\tau_i$ is at least $n\lambda/4$ apart from the boundaries of I_i. Thus, as in the proof of Theorem 6, under the assumption in (13), we can establish for naive or advanced (or combined) optimistic search the following

$$\mathbb{P}\left\{\hat{\kappa} = \kappa, \; |\hat{\tau}_i - \tau_i| \leq C_2 \left(\frac{\log n}{\delta_i^2 n} \bigvee \frac{s^2 \wedge p \log n}{n^2 \lambda \delta_i^4}\right), \; i = 1, \ldots, \kappa\right\} \geq 1 - \frac{24}{n},$$

which concludes the proof. ∎

Appendix F. Computation Complexity Analysis

In this section, we derive computation complexity analysis for optimistic searches, as stated in Lemma 3, Theorem 5ii and Proposition 8. We omit the proof of Theorem 5ii as it follows from Proposition 8 with $p = 1$.

Proof [of Lemma 3] We start with the naive optimistic search. Let I_k denote the search window at the k-th step of the naive optimistic search. The choice of probe points ensures that a segment of length at least $\nu |I_k| / 2 \wedge \nu |I_{k-1}| / 2$ will be dropped out, with ν the step size, at the k-th step. Since $I_k \subseteq I_{k-1}$, we have a proportion of at least $\nu/2$ will be removed at each step, and thus the procedure will stop in

$$\left\lceil \frac{1}{\log\left(2/(2 - \nu)\right)} \log(R - L) \right\rceil \quad \text{steps.}$$

Consider next the search over dyadic locations (in the advanced optimistic search), which requires $O\left(\log(R - L)\right)$ evaluations of gain functions, as there are in total $2\lceil \log_2(R - L) \rceil$ dyadic locations.

Thus, overall, in the worst case one has $O(\log(R - L)) \leq O(\log n)$ evaluations for the naive and the advanced optimistic search. ∎

Proof [of Proposition 8] *Part i.* It follows from Lemma 3 and the fact that given the cumulative sums each evaluation of the gain or the comparison function takes $O(p)$ computations.

Part ii. We consider first the optimistic searches on seeded intervals. By $\mathcal{I}_\omega$ we denote the set of all seeded intervals of length no less than $m \asymp n^{1-\omega}$. Then the total number of evaluations in naive and advanced (and also combined) optimistic search on all seeded intervals in $\mathcal{I}_\omega$ is of order

$$p \sum_{I \in \mathcal{I}_w} \log |I| \lesssim p \sum_{k=1}^{\lceil \log_{1/a}(n^\omega) \rceil} a^{-(k-1)} \log(a^{k-1}n) \lesssim p \int_0^{\log_{1/a}(n^\omega)} a^{-x} \log(a^x n) dx$$

$$\lesssim p \int_0^{n^\omega} \log \left(\frac{n}{x} \right) dx = n^\omega \big(1 + (1-\omega)\log n\big) \lesssim p \min\{n^\omega \log n, \, n\} .$$

Consider the NOT selection, and employ the deterministic and nested structure of seeded intervals. Each seeded interval I is equipped with a boolean variable b_I, denoting whether it is needed to be considered or not (its default value being true), and with an integer variable t_I, recording the estimated change point location (its default value being zero).

Note that the seeded intervals at k-th layer are of the same length $l_k = na^{k-1}$, and that the NOT selection favours shorter intervals. Thus, we iterate from the highest layer to the lowest layer. More precisely, at k-th layer, we check each seeded interval I:

- If its boolean variable b_I is false, we mark the seeded intervals that lie at $(k-1)$-th layer and contain t_I, with the boolean variable being false and the integer variable being t_I. Due to the construction of seeded intervals, this can be done in $O(1)$ computations.

- If its boolean variable b_I is true, we check whether the candidate change point $\hat{t}_I$ found in this seeded interval has a value of gain function above the selection threshold. If no, we do nothing. If yes, we identify $\hat{t}_I$ as an estimated change point, and mark the seeded intervals that lie at $(k-1)$-th layer and contain $\hat{t}_I$, with the boolean variable being false and the integer variable being $\hat{t}_I$. Similarly, this can be done in $O(1)$ computations.

It is clear to see that this procedure executes precisely the NOT selection. As each seeded interval is accessed at most $O(1)$ times and each access costs $O(1)$ computations, the NOT selection requires $O(|\mathcal{I}_\omega|) = O(n^\omega)$ computations. Note that this computation complexity analysis improves the one in Theorem 1 of Kovács et al. (2023) by a log factor.

The post-processing step involves optimistic searches on $\hat{\kappa}$ non-overlapping intervals, and thus requires another $O\big(p \min\{\hat{\kappa} \log n, \, n\}\big) = O\big(p \min\{n^\omega \log n, \, n\}\big)$ computations.

Thus, if cumulative sums are pre-computed, the final computation complexity is

$$O\big(p \min\{n^\omega \log n, \, n\}\big).$$

Otherwise, the computation of cumulative sums dominates, and leads to the final computation complexity of $O(pn)$. Besides, the overall memory complexity is $O(pn)$. ∎

Remark 23 (Implementation) *It is logically clean and easy to consider the optimistic search part and the selection part, separately, as in Algorithm 3. But in regard to implementation, we should merge both parts, and run optimistic searches only for the seeded intervals with boolean variables being true, though this only improves the multiplying constant in the computation complexity analysis, instead of the order.*

Remark 24 (Speedup) *We can speed up OSeedBS by considering first the middle points of seeded intervals as follows:*

i. *For a given seeded interval, we run optimistic searches only when the value of gain function at the middle of this seeded interval is above the selection threshold.*

ii. *Further, we use the middle point of seeded intervals to decide which coordinates to look at for optimistic searches, more precisely, all the coordinates that has a cumulative statistic at the middle point with absolute value greater than α.*

Note that for each change point there is a seeded interval containing it roughly at the center, which will eventually be picked by the NOT selection, as shown in the proof of Theorem 7. Then the gain function evaluated at the middle points of such seeded intervals has similar values as if it were evaluated at true change points. Thus, the same statistical guarantee remains still valid for OSeedBS with the modification above.

Let $\hat{s}$ denote the maximal number of selected coordinates. Then as $n \to \infty$ it holds that $\mathbb{P}\{\hat{s} = O(s)\} \to 1$. It is clear to see that the overall computation complexity can be improved to $O(\hat{s}\min\{n^{\omega}\log n, n\} + pn^{\omega})$, which is $O(s\min\{n^{\omega}\log n, n\} + pn^{\omega})$ with probability tending to one.

In particular, in case of a single change point, one only needs to consider seeded intervals that starting at 0 or ending at n, the number of which is $O(\log n)$. Then the naive optimistic search with the NOT selection can achieve the same statistical optimality as in Theorem 6, and together with the speedup idea above it results in $O(p\log n)$ computations, provided that cumulative sums of data are available. Thus, the naive optimistic search with the NOT selection, and the advanced optimistic search, are equivalent in terms of statistical and computational efficiency.

References

V. Avanesov and N. Buzun. Change-point detection in high-dimensional covariance structure. *Electron. J. Stat.*, 12(2):3254–3294, 2018.

M. Avriel and D. J. Wilde. Optimality proof for the symmetric Fibonacci search technique. *Fibonacci Quart.*, 4:265–269, 1966.

M. Avriel and D. J. Wilde. Golden block search for the maximum of unimodal functions. *Management Sci.*, 14:307–319, 1968.

J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998.

R. Baranowski, Y. Chen, and P. Fryzlewicz. Narrowest-over-threshold detection of multiple change points and change-point-like features. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 81 (3):649–672, 2019.

M. Behr, C. Holmes, and A. Munk. Multiscale blind source separation. *Ann. Statist.*, 46 (2):711–744, 2018.

M. Bhattacharjee, M. Banerjee, and G. Michailidis. Common change point estimation in panel data from the least squares and maximum likelihood viewpoints. *arXiv preprint arXiv:1708.05836v1*, 2017.

L. Boysen, A. Kempe, V. Liebscher, A. Munk, and O. Wittich. Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.*, 37(1):157–183, 2009.

L. Bybee and Y. Atchadé. Change-point computation for large graphical models: a scalable algorithm for Gaussian graphical models with change-points. *J. Mach. Learn. Res.*, 19 (11):1–38, 2018.

H. Cho and C. Kirch. Two-stage data segmentation permitting multiscale change points, heavy tails and dependence. *Ann. Inst. Statist. Math.*, 74(4):653–684, 2022.

H. Cho and D. Owens. High-dimensional data segmentation in regression settings permitting temporal dependence and non-Gaussianity. *Electron. J. Stat.*, 18(1):2620–2664, 2024.

H. Cho, T. Kley, and H. Li. Detection and inference of changes in high-dimensional linear regression with non-sparse structures. *arXiv preprint arXiv:2402.06915v2*, 2024.

L. Davies, C. Höhenrieder, and W. Krämer. Recursive computation of piecewise constant volatilities. *Comput. Statist. Data Anal.*, 56(11):3623–3631, 2012.

H. Dette, G. Pan, and Q. Yang. Estimating a Change Point in a Sequence of Very High-Dimensional Covariance Matrices. *J. Amer. Statist. Assoc.*, 117(537):444–454, 2022.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Stat.*, 3(2):521–541, 2009.

K. Frick, A. Munk, and H. Sieling. Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(3):495–580, 2014.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441, 2008.

F. Friedrich, A. Kempe, V. Liebscher, and G. Winkler. Complexity penalized M-estimation: fast computation. *J. Comput. Graph. Statist.*, 17(1):201–224, 2008.

P. Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, 42(6):2243–2281, 2014.

P. Fryzlewicz. Detecting possibly frequent change-points: Wild Binary Segmentation 2 and steepest-drop model selection. *J. Korean Statist. Soc.*, 49(4):1027–1070, 2020.

F. Gao and T. Wang. Sparse change detection in high-dimensional linear regression. *arXiv preprint arXiv:2208.06326v2*, 2022.

A. J. Gibberd and J. D. B. Nelson. Regularized estimation of piecewise constant Gaussian graphical models: the group-fused graphical Lasso. *J. Comput. Graph. Statist.*, 26(3): 623–634, 2017.

A. J. Gibberd and S. Roy. Multiple changepoint estimation in high-dimensional Gaussian graphical models. *arXiv preprint arXiv:1712.05786v1*, 2017.

C. Giraud. *Introduction to high-dimensional statistics*, volume 168 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2022.

D. Hallac, Y. Park, S. Boyd, and J. Leskovec. Network inference via the time-varying graphical Lasso. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 205–213, 2017.

T. Hocking, G. Rigaill, J.-P. Vert, and F. Bach. Learning sparse penalties for change-point detection using max margin interval regression. In *International Conference on Machine Learning*, pages 172–180. PMLR, 2013.

T. Hotz, O. M. Schütte, H. Sieling, T. Polupanow, U. Diederichsen, C. Steinem, and A. Munk. Idealizing ion channel recordings by a jump segmentation multiresolution filter. *IEEE Trans. Nanobiosci.*, 12(4):376–386, 2013.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *Ann. Statist.*, 49(2):1055–1080, 2021.

S. Hu, J. Huang, H. Chen, and H. P. Chan. Likelihood scores for sparse signal and change-point detection. *IEEE Trans. Inf. Theory*, 2023.

B. Jackson, J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. T. Tsai. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Process. Lett.*, 12(2):105–108, 2005.

A. Kaul, V. K. Jandhyala, and S. B. Fotopoulos. An efficient two step algorithm for high dimensional change point regression models without grid search. *J. Mach. Learn. Res.*, 20(111):1–40, 2019a.

A. Kaul, V. K. Jandhyala, and S. B. Fotopoulos. Detection and estimation of parameters in high dimensional multiple change point regression models via ℓ_1/ℓ_0 regularization and discrete optimization. *arXiv preprint arXiv:1906.04396v1*, 2019b.

A. Kaul, S. B. Fotopoulos, V. K. Jandhyala, and A. Safikhani. Inference on the change point under a high dimensional sparse mean shift. *Electron. J. Stat.*, 15(1):71–134, 2021.

J. Kiefer. Sequential minimax search for a maximum. *Proc. Amer. Math. Soc.*, 4:502–506, 1953.

R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.*, 107(500):1590–1598, 2012.

C.-J. Kim, J. C. Morley, and C. R. Nelson. The structural break in the equity premium. *J. Bus. Econom. Statist.*, 23(2):181–191, 2005.

S. Kovács, P. Bühlmann, H. Li, and A. Munk. Seeded binary segmentation: a general methodology for fast and optimal changepoint detection. *Biometrika*, 110(1):249–256, 2023.

B. Laurent, J.-M. Loubes, and C. Marteau. Non asymptotic minimax rates of testing in signal detection with heterogeneous variances. *Electron. J. Stat.*, 6:91–122, 2012.

F. Leonardi and P. Bühlmann. Computationally efficient change point detection for high-dimensional regression. *arXiv preprint arXiv:1601.03704v1*, 2016.

H. Li, A. Munk, and H. Sieling. FDR-control in multiscale change-point segmentation. *Electron. J. Stat.*, 10(1):918–959, 2016.

W. Li, D. Wang, and A. Rinaldo. Divide and conquer dynamic programming: An almost linear time change point detection methodology in high dimensions. In *International Conference on Machine Learning*, pages 20065–20148. PMLR, 2023.

H. Liu, C. Gao, and R. J. Samworth. Minimax rates in sparse, high-dimensional change point detection. *Ann. Statist.*, 49(2):1081–1112, 2021.

Z. Liu and H. Li. Multiscale quantile regression with local error control. *arXiv preprint arXiv:2403.11356v1*, 2024.

M. Londschien, S. Kovács, and P. Bühlmann. Change-point detection for graphical models in the presence of missing values. *J. Comput. Graph. Statist.*, 30(3):768–779, 2021.

Z. Lu, M. Banerjee, and G. Michailidis. Intelligent sampling for multiple change-points in exceedingly long time series with rate guarantees. *arXiv preprint arXiv:1710.07420v4*, 2017.

O. H. Madrid Padilla, Y. Yu, D. Wang, and A. Rinaldo. Optimal nonparametric multivariate change point detection and localization. *IEEE Trans. Inform. Theory*, 68(3):1922–1944, 2022.

R. Maidstone, T. Hocking, G. Rigaill, and P. Fearnhead. On optimal multiple changepoint algorithms for large data. *Stat. Comput.*, 27(2):519–533, 2017.

X. F. Mallory, M. Edrisi, N. Navin, and L. Nakhleh. Methods for copy number aberration detection from single-cell dna-sequencing data. *Genome Biol.*, 21:1–22, 2020.

R. Mazumder and T. Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *J. Mach. Learn. Res.*, 13:781–794, 2012a.

R. Mazumder and T. Hastie. The graphical lasso: new insights and alternatives. *Electron. J. Stat.*, 6:2125–2149, 2012b.

Y. S. Niu, N. Hao, and H. Zhang. Multiple change-point detection: a selective overview. *Statist. Sci.*, 31(4):611–623, 2016.

A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41:100–115, 1954.

F. Picard, E. Lebarbier, M. Hoebeke, G. Rigaill, B. Thiam, and S. Robin. Joint segmentation, calling, and normalization of multiple cgh profiles. *Biostat.*, 12(3):413–428, 2011.

E. Pilliat, A. Carpentier, and N. Verzelen. Optimal multiple change-point detection for high-dimensional data. *Electron. J. Stat.*, 17(1):1240–1315, 2023.

J. Reeves, J. Chen, X. L. Wang, R. Lund, and Q. Q. Lu. A review and comparison of changepoint detection techniques for climate data. *J. Appl. Meteorol. Climatol.*, 46:900–915, 2007.

S. Roy, Y. Atchadé, and G. Michailidis. Change point estimation in high dimensional Markov random-field models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(4):1187–1206, 2017.

K. Rufibach and G. Walther. The block criterion for multiscale inference about a density, with applications to other multiscale problems. *J. Comput. Graph. Statist.*, 19(1):175–190, 2010.

J. Sharpnack, A. Singh, and A. Rinaldo. Changepoint detection over graphs with the spectral scan statistic. In *Artificial Intelligence and Statistics*, pages 545–553. PMLR, 2013.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

C. Truong, L. Oudre, and N. Vayatis. Selective review of offline change point detection methods. *Signal Process.*, 167:107299, 2020.

L. J. Vanegas, M. Behr, and A. Munk. Multiscale quantile segmentation. *J. Amer. Statist. Assoc.*, 117(539):1384–1397, 2022.

N. Verzelen, M. Fromont, M. Lerasle, and P. Reynaud-Bouret. Optimal change-point detection and localization. *Ann. Statist.*, 51(4):1586–1610, 2023.

L. Y. Vostrikova. Detecting 'disorder' in multidimensional random processes. *Soviet Mathematics Doklady*, 24:55–59, 1981.

D. Wang, Y. Yu, and A. Rinaldo. Optimal change point detection and localization in sparse dynamic networks. *Ann. Statist.*, 49(1):203–232, 2021a.

D. Wang, Y. Yu, and A. Rinaldo. Optimal covariance change point localization in high dimensions. *Bernoulli*, 27(1):554–575, 2021b.

D. Wang, Z. Zhao, K. Z. Lin, and R. Willett. Statistically and computationally efficient change point localization in regression settings. *J. Mach. Learn. Res.*, 22:Paper No. [248], 46, 2021c.

T. Wang and R. J. Samworth. High dimensional change point estimation via sparse projection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 80(1):57–83, 2018.

D. M. Witten, J. H. Friedman, and N. Simon. New insights and faster computations for the graphical lasso. *J. Comput. Graph. Statist.*, 20(4):892–900, 2011.

W. B. Wu. Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA*, 102(40):14150–14154, 2005.

H. Xu, D. Wang, Z. Zhao, and Y. Yu. Change-point inference in high-dimensional regression models under temporal dependence. *Ann. Statist.*, 52(3):999–1026, 2024.

M. Yu and X. Chen. Finite sample change point inference and identification for high-dimensional mean vectors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 83(2):247–270, 2021.

D. Zhang and W. B. Wu. Gaussian approximation for high dimensional time series. *Ann. Statist.*, 45(5):1895–1919, 2017.

D. Zhang and W. B. Wu. Convergence of covariance and spectral density estimates for high-dimensional locally stationary processes. *Ann. Statist.*, 49(1):233–254, 2021.

N. R. Zhang and D. O. Siegmund. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007.

N. R. Zhang, D. O. Siegmund, H. Ji, and J. Z. Li. Detecting simultaneous changepoints in multiple sequences. *Biometrika*, 97(3):631–645, 2010.