Theoretical Computer Science 620 (2016) 119–133

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Nonparametric multiple change point estimation in highly dependent time series

Azadeh Khaleghi [a,*], Daniil Ryabko [b,*]

[a] Department of Mathematics & Statistics, Lancaster University, UK
[b] INRIA Lille, France

ARTICLE INFO

Article history:
Available online 11 November 2015

Keywords:
Change point analysis
Stationary ergodic time series
Unsupervised learning
Consistency

ABSTRACT

Given a heterogeneous time-series sample, the objective is to find points in time, called change points, where the probability distribution generating the data has changed. The data are assumed to have been generated by arbitrary unknown stationary ergodic distributions. No modelling, independence or mixing assumptions are made. A novel, computationally efficient, nonparametric method is proposed, and is shown to be asymptotically consistent in this general framework. The theoretical results are complemented with experimental evaluations.

1. Introduction

Change point estimation is a classical problem in mathematical statistics [1,2] which, with its broad range of applications in learning problems, has started to gain attention in the machine learning community. The problem can be introduced as follows. A given sequence

$$\boldsymbol{x} := X_1, \ldots, X_{\lfloor n\theta_1 \rfloor}, X_{\lfloor n\theta_1 \rfloor + 1}, \ldots, X_{\lfloor n\theta_2 \rfloor}, \ldots, X_{\lfloor n\theta_\kappa \rfloor + 1}, \ldots, X_n$$

is formed as the concatenation of $\kappa + 1$ non-overlapping segments, where $\kappa \in \mathbb{N}$ and $0 < \theta_1 < \cdots < \theta_\kappa < 1$. Each segment is generated by some unknown time-series (or process) distribution. The distributions that generate every pair of consecutive segments are different. The index $\lfloor n\theta_k \rfloor$ where one segment ends and another starts is called a *change point*. The parameters θ_k, $k = 1..\kappa$ specifying the change points $\lfloor n\theta_k \rfloor$ are unknown and have to be estimated.

In a typical formulation of the problem, the samples within each segment $X_{\lfloor n\theta_1 \rfloor + 1}..X_{\lfloor n\theta_2 \rfloor}$ are assumed to be i.i.d. and the change is in the mean (see, e.g., [3] for a review). In the literature on nonparametric change point methods for dependent data the form of the change and/or the nature of dependence are usually restricted; for example, settings involving time series that satisfy strong mixing conditions are often considered [1]. Moreover, the finite-dimensional marginals are almost exclusively assumed different [4,5]. Such assumptions often do not hold in real-world applications. From a machine-learning perspective, change point estimation appears to be a difficult unsupervised learning problem: an algorithm is required to locate the changes in a given sequence without any examples of correct solutions.

In this paper, we consider highly dependent time series, making as few assumptions as possible on how the data are generated. The only assumption that we make is that each segment is generated by an unknown stationary ergodic process distribution. The joint distribution over the samples can be otherwise arbitrary. We make no such assumptions as

* Corresponding authors.
E-mail addresses: a.khaleghi@lancaster.ac.uk (A. Khaleghi), Daniil@Ryabko.net (D. Ryabko).

http://dx.doi.org/10.1016/j.tcs.2015.10.041
0304-3975/

independence, finite memory or mixing. The marginal distributions of any given size before and after the change may be the same: the change refers to that in the time-series distribution.

The main result of this paper is an asymptotically consistent algorithm for estimating all κ parameters θ_k, $k = 1..\kappa$ simultaneously. We assume that κ is given, but the process distributions as well as the nature of the change are unknown. An estimate $\hat{\theta}_k$ of a change point parameter θ_k is *asymptotically consistent* if it becomes arbitrarily close to θ_k in the limit, as the length n of the sequence approaches infinity. However, the problem is *offline* and $\boldsymbol{x}$ does not grow with time. Thus, the asymptotic regime only means that the error is arbitrarily small if the sequence is sufficiently long. Real-world scenarios that correspond to this formulation include, for example, genomic data, sequences of stock-market values, high-resolution audio/video data, and all such long sequential observations with distributional changes, where the distributions are completely unknown, changes are arbitrary, but the segments are long.

While the assumption that each segment is generated by a stationary ergodic process is already very general, it can be relaxed even further. In particular, one relatively simple but meaningful generalisation that we consider is that each process is asymptotically mean stationary ergodic. This generalisation allows us to address the problem of gradual (as opposed to abrupt) change in the distribution.

In general, for stationary ergodic processes, rates of convergence are provably impossible to obtain; this already concerns the convergence of frequencies to probabilities [6]. Thus, non-asymptotic results cannot be obtained in this setting. On the other hand, this means that, unlike in more restricted settings, in our setting the algorithms are forced not to rely on any rate of convergence guarantees. We see this as an advantage of the framework, as it means that the algorithms are applicable to a much wider range of situations. Furthermore, in this setting it is provably impossible to estimate κ. This follows from the impossibility result of [7], which states that it is not possible to determine, even in the weakest asymptotic sense, whether two sequences have been generated by the same or by different stationary ergodic distributions. Thus, in this paper we assume that κ is known.

The case of $\kappa = 1$ was addressed in [8], where a simple consistent algorithm for estimating one change point was provided. The general case of $\kappa > 1$ turns out to be much more complex. With the sequence containing multiple change points, the algorithm is required to simultaneously analyse multiple segments of the input sequence, with no a-priori lower bound on their lengths. In this case the main challenge is to ensure that the algorithm is robust with respect to segments of arbitrarily small length. The problem is considerably simplified if additionally a lower bound on the minimum separation of the change points is provided. Indeed, the method of [8] for $\kappa = 1$ also relies on the knowledge of such parameter, namely, a lower bound on the minimum distance of the change point from the two end-points. With this additional information, some inference can be made even in the case where $\kappa > 1$ is unknown. Specifically, an algorithm is proposed in [9] which, without the knowledge of κ, gives an exhaustive list of candidate estimates whose first κ elements constitute asymptotically consistent estimates of the change points. Assuming additionally that the total number of different distributions generating the sample is known, it is shown in [10] that the number of change points κ can be found eventually almost surely for large enough n. In this work we do not make any such assumptions; specifically, we do not assume that a lower bound on the minimum separation of the change points is known.

Our algorithm is based on empirical estimates of the so-called distributional distance [11], which have proven useful in various statistical and learning problems involving stationary ergodic time series [8–10,12–14]. The computational complexity of our algorithm is at most quadratic in each argument. We evaluate the proposed method on synthetic data generated by processes that, while being stationary ergodic, do not belong to any of the "simpler" classes studied in the literature on such problems. In particular, the processes used in the experiments cannot be modelled as hidden Markov processes with a countable set of states. Moreover, in the considered examples the single-dimensional marginals before and after each change point are the same.

Organisation. In Section 2 we introduce preliminary notations and definitions. In Section 3 we formalise the problem and describe the general framework considered. In Section 4 we present our method, state the main consistency result, and informally describe how the algorithm works; the proof of the main result is deferred to Section 8. In Section 6 we provide some experimental evaluations. In Section 5 we discuss some theoretical extensions of the considered framework and finally we conclude in Section 7.

2. Preliminaries

Let $\mathcal{X}$ be a measurable space (the domain); in this work we let $\mathcal{X} = \mathbb{R}$, but extensions to more general spaces are straightforward. For a sequence $X_1, \dots, X_n$ we use the abbreviation $X_{1..n}$. Consider the Borel σ-algebra $\mathcal{B}$ on $\mathcal{X}^\infty$ generated by the cylinders $\{B \times \mathcal{X}^\infty : B \in B^{m,l}, m, l \in \mathbb{N}\}$, where the sets $B^{m,l}$, $m, l \in \mathbb{N}$ are obtained via the partitioning of $\mathcal{X}^m$ into cubes of dimension m and volume 2^{-ml} (starting at the origin). Let also $B^m := \cup_{l\in\mathbb{N}} B^{m,l}$. Process distributions are probability measures on the space $(\mathcal{X}^\infty, \mathcal{B})$. For $\boldsymbol{x} = X_{1..n} \in \mathcal{X}^n$ and $B \in B^m$ let $\nu(\boldsymbol{x}, B)$ denote the frequency with which $\boldsymbol{x}$ falls in B, i.e.

$$\nu(\boldsymbol{x}, B) := \frac{\mathbb{I}\{n \geq m\}}{n-m+1} \sum_{i=1}^{n-m+1} \mathbb{I}\{X_{i..i+m-1} \in B\}. \tag{1}$$

A process ρ is *stationary* if for any $i, j \in 1..n$ and $B \in B^m$, $m \in \mathbb{N}$, we have $\rho(X_{1..j} \in B) = \rho(X_{i..i+j-1} \in B)$. A stationary process ρ is called *stationary ergodic* if for all $B \in \mathcal{B}$ with probability 1 we have $\lim_{n\to\infty} \nu(X_{1..n}, B) = \rho(B)$. By virtue of the

ergodic theorem this definition can be shown to be equivalent to the usual definition given in terms of shift-invariant sets; see e.g., [11,15].

Definition 1 *(Distributional distance).* (See [11].) The distributional distance between a pair of process distributions ρ_1, ρ_2 is defined as follows

$$d(\rho_1, \rho_2) := \sum_{m,l=1}^{\infty} w_m w_l \sum_{B \in B^{m,l}} |\rho_1(B) - \rho_2(B)| .$$

We let $w_j := \frac{1}{j(j+1)}$, but any summable sequence of positive weights may be used.

In words, we partition the sets $\mathcal{X}^m$, $m \in \mathbb{N}$ into cubes of decreasing volume (indexed by l) and take a weighted sum over the differences in probabilities of all the cubes in these partitions. Different generating sets (other than cubes) can be used to define the distributional distance; here we chose cubes in order to facilitate the experimental setup. Smaller weights are given to larger m and finer partitions. We use empirical estimates of this distance, where probabilities are replaced with frequencies:

Definition 2 *(Empirical estimates of $d(\cdot,\cdot)$).* For $\boldsymbol{x}_i \in \mathcal{X}^{n_i}$ $n_i \in \mathbb{N}$, $i = 1, 2$, and a distribution ρ the empirical estimate of d are defined as

$$\hat{d}(\boldsymbol{x}, \rho) := \sum_{m=1}^{m_n} \sum_{l=1}^{l_n} w_m w_l \sum_{B \in B^{m,l}} |\nu(\boldsymbol{x}, B) - \rho(B)| , \tag{2}$$

$$\hat{d}(\boldsymbol{x}_1, \boldsymbol{x}_2) := \sum_{m=1}^{m_n} \sum_{l=1}^{l_n} w_m w_l \sum_{B \in B^{m,l}} |\nu(\boldsymbol{x}_1, B) - \nu(\boldsymbol{x}_2, B)| , \tag{3}$$

where m_n and l_n are any sequences of integers that go to infinity with n.

Remark 1. Despite the infinite summations, $\hat{d}$ can be calculated efficiently [12]. Its computational complexity is upper-bounded by $\mathcal{O}(n \text{ polylog } n)$ for $m_n := \log n$, the choice of which is justified in [13] (see also [9]).

Proposition 1 *($\hat{d}(\cdot,\cdot)$ is consistent; see [8]). Let a pair of sequences $\boldsymbol{x}_1 \in \mathcal{X}^{n_1}$ and $\boldsymbol{x}_2 \in \mathcal{X}^{n_2}$ be generated by a distribution ρ whose marginals ρ_i, $i = 1, 2$ are stationary and ergodic. Then*

$$\lim_{n_i \to \infty} \hat{d}(\boldsymbol{x}_i, \rho_j) = d(\rho_i, \rho_j), \ i, j \in 1, 2, \ \rho - a.s., \tag{4}$$

$$\lim_{n_1, n_2 \to \infty} \hat{d}(\boldsymbol{x}_1, \boldsymbol{x}_2) = d(\rho_1, \rho_2), \ \rho - a.s. \tag{5}$$

3. Problem formulation

We formalise the problem of multiple change point estimation as follows. The sequence $\boldsymbol{x} \in \mathcal{X}^n$, $n \in \mathbb{N}$ is formed as the concatenation of $\kappa + 1$ of sequences

$$X_{1..\lfloor n\theta_1 \rfloor}, X_{\lfloor n\theta_1 \rfloor + 1..\lfloor n\theta_2 \rfloor}, \ldots, X_{\lfloor n\theta_\kappa \rfloor + 1..n},$$

where $\theta_k \in (0, 1)$, $k = 1..\kappa$, and where the number of change points κ is assumed known. Denote $\theta_0 := 0$, $\theta_{\kappa+1} := 1$. Each of the sequences $\boldsymbol{x}_k := X_{\lfloor n\theta_{k-1} \rfloor + 1..\lfloor n\theta_k \rfloor}$, $k = 1..\kappa + 1$, is generated by an *unknown stationary ergodic* process distribution. Formally, consider a matrix $\mathbf{X} \in (\mathcal{X}^{\kappa+1})^\infty$ of random variables generated by some (unknown) stochastic process distribution ρ such that **1.** the marginal distribution over every one of its rows is an unknown stationary ergodic process distribution; **2.** the marginal distributions over the consecutive rows are different, so that every two consecutive rows are generated by different process distributions. The sequence $\boldsymbol{x} \in \mathcal{X}^n$ is formed as follows. First, the length $n \in \mathbb{N}$ is fixed, next for each $k = 1..\kappa + 1$ a segment $\boldsymbol{x}_k \in \mathcal{X}^{\lfloor n(\theta_k - \theta_{k-1}) \rfloor}$ is obtained as the first $\lfloor n(\theta_k - \theta_{k-1}) \rfloor$ elements of the kth row of $\mathbf{X}$.

Note that the requirements are only on the marginal distributions over the rows; the distribution ρ is otherwise completely arbitrary. The process distributions are unknown and may be dependent. Moreover, the means, variances, or, more generally, the finite-dimensional marginal distributions of any fixed size before and after the change points are not required to be different. We consider the most general scenario where the *process distributions are different*.

The parameters θ_k, $k = 1..\kappa$ specify the change points $\lfloor n\theta_k \rfloor$, which separate consecutive segments $\boldsymbol{x}_k, \boldsymbol{x}_{k+1}$ generated by different process distributions. Define the minimum separation of the change point parameters as

$$\lambda_{\min} := \min_{k=1..\kappa+1} \theta_k - \theta_{k-1}. \tag{6}$$

Algorithm 1 A multiple change point estimator.

input: $\boldsymbol{x} = X_{1..n}$, Number κ of Change points
initialize: $\eta \leftarrow 0$
for $j = 1..\log n$ **do**
  $\lambda_j \leftarrow 2^{-j}$, $\alpha_j \leftarrow \lambda_j/3$, $w_j \leftarrow 2^{-j}$ ▷ Set the step size and iteration weight
  for $t = 1..\kappa + 1$ **do**
    $b_i^{t,j} \leftarrow n\alpha_j(i + \frac{1}{t+1})$, $i = 0..\lfloor \frac{1}{\alpha_j} - \frac{1}{t+1} \rfloor$ ▷ Generate boundaries
    for $l = 0..2$ **do**
      $d_{i'} \leftarrow \Delta_{\boldsymbol{x}}(b^{t,j}_{l+3(i'-1)}, b^{t,j}_{l+3i'})$, $i' = 1..\frac{1}{3}(\lfloor \frac{1}{\alpha_j} - \frac{1}{t+1} \rfloor - l)$
      $\gamma_l \leftarrow d_{\lfloor \kappa \rfloor}$ ▷ Store the κth highest value
    end for
    $\gamma(t, j) \leftarrow \min_{l=0..2} \gamma_l$ ▷ Obtain the grid's performance score
    $\{\mu_1, \ldots, \mu_\kappa\} \leftarrow \underset{i \in 1..\lfloor \frac{1}{\alpha_j} - \frac{1}{t+1} \rfloor - 1}{\overset{k=1..\kappa}{\operatorname{argmax}}} \Delta_x(b_i^{t,j}, b_{i+1}^{t,j})$ ▷ Find κ segments of highest $\Delta_{\boldsymbol{x}}$; ($X_{b^{t,j}_{\mu_k}..b^{t,j}_{\mu_k+1}}$ is the segment with kth highest score).
    $(b^{t,j}_{[1]}, \ldots, b^{t,j}_{[\kappa]}) \leftarrow \mathbf{sort}(b^{t,j}_{\mu_1}, \ldots, b^{t,j}_{\mu_\kappa})$ ▷ Sort the selected boundaries in increasing order
    $\hat{\pi}_k^{t,j} := \Phi_{\boldsymbol{x}}(b^{t,j}_{[k]}, b^{t,j}_{[k]+1}, \alpha_j)$, $k = 1..\kappa$ ▷ Seek a change point in κ segments of highest $\Delta_{\boldsymbol{x}}$
    $\eta \leftarrow \eta + w_j\gamma(t, j)$ ▷ Update the sum of weights
  end for
end for
$\hat{\theta}_k \leftarrow \frac{1}{n\eta} \sum_{j=1}^{\log n} \sum_{t=1}^{\kappa+1} w_j \gamma(t, j) \hat{\pi}_k^{t,j}$, $k = 1..\kappa$ ▷ Calculate the final estimates
return: $\hat{\theta}_1, \ldots, \hat{\theta}_\kappa$

Since the consistency properties that we seek are asymptotic in n, we require that $\lambda_{\min} > 0$. This means that the (unknown) minimum separation of the change points is linear in n. Note that this condition is standard in the change point literature, although it may be unnecessary when simpler formulations of the problem are considered, for example when the samples within each segment are i.i.d. However, conditions of this kind are inevitable in the general setting that we consider, where the segments and the samples within each segment are allowed to be arbitrarily dependent: if the length of one of the sequences is constant or sub-linear in n then asymptotic consistency is not possible in this setting. Finally, note that we make no assumptions on the distance between the process distributions: they can be arbitrarily close.

Our goal is to devise an algorithm that provides estimates $\hat{\theta}_k$ for the parameters θ_k, $k = 1..\kappa$. The algorithm must be *asymptotically consistent* so that

$$\lim_{n\to\infty} \sup_{k=1..\kappa} |\hat{\theta}_k(n) - \theta_k| = 0 \text{ a.s.} \tag{7}$$

4. Main result

In this section we propose Algorithm 1, which, as shown in Theorem 1, is asymptotically consistent under the general assumptions stated in Section 3. The proof of the consistency result is deferred to Section 8. Here we give an intuitive description of how the algorithm works and why the consistency result holds.

Theorem 1. *Algorithm 1 is asymptotically consistent, provided that each segment $\boldsymbol{x}_k$, $k = 1..\kappa$, is generated by a stationary ergodic distribution, and that the correct number κ of change points is given:*

$$\lim_{n\to\infty} \sup_{k=1..\kappa} \left|\hat{\theta}_k(n) - \theta_k\right| = 0 \text{ a.s.}$$

The following two operators, namely, the score function $\Delta_{\boldsymbol{x}}$ and the single-change point-estimator $\Phi_{\boldsymbol{x}}$ are used in the algorithm.

Definition 3. Let $\boldsymbol{x} = X_{1..n}$ be a sequence and consider a subsequence $X_{a..b}$ of $\boldsymbol{x}$ with $a < b \in 1..n$.

i. Define the score function or the intra-subsequence distance of $X_{a..b}$ as

$$\Delta_{\boldsymbol{x}}(a, b) := \hat{d}\left(X_{a..\lfloor \frac{a+b}{2} \rfloor}, X_{\lceil \frac{a+b}{2} \rceil..b}\right) \tag{8}$$

ii. Define the single-change point estimator of $X_{a..b}$ as

$$\Phi_{\boldsymbol{x}}(a, b, \alpha) := \underset{t \in a..b}{\operatorname{argmax}} \hat{d}\left(X_{a-n\alpha..t}, X_{t..b+n\alpha}\right), \text{ where } \alpha \in (0, 1) \tag{9}$$

Let us start by giving an overview of what Algorithm 1 aims to do. The algorithm attempts to simultaneously estimate all κ change points using the single-change point-estimator $\Phi_{\boldsymbol{x}}$ given by (8) applied to appropriate segments of the sequence.

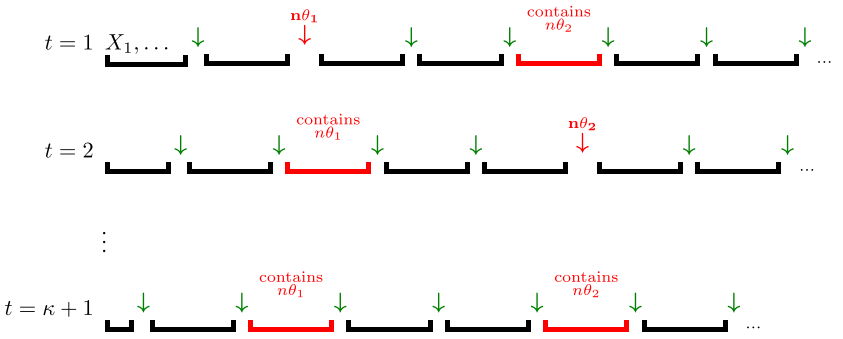

Fig. 1. For a fixed j, Algorithm 1 generates $\kappa + 1$ grids composed of segments of length $n\alpha_j$ but with distinct starting points: $n\alpha_j/(t+1)$, $t = 1..\kappa+1$, where α_j is the algorithm's guess of $\lambda_{\min}/3$. At the iteration shown in this figure, $\alpha_j \le \lambda_{\min}/3$ so that every three consecutive segments contain at most one change point. Since there are κ change points, there exists at least one grid (in this example the one corresponding to $t = \kappa + 1$) with the property that none of the change points are located at the boundaries.

In order for $\Phi_{\boldsymbol{x}}$ to produce asymptotically consistent estimates in this setting, each change point must be isolated within a segment of $\boldsymbol{x}$ whose length is a linear function of n. Moreover, each segment containing a change point must be "sufficiently far" from the rest of the change points, where "sufficiently far" means within a distance linear in n. This may be achieved by dividing $\boldsymbol{x}$ into consecutive non-overlapping segments, each of length $n\alpha$ with $\alpha := \lambda/3$ for some $\lambda \in (0, \lambda_{\min}]$, where $\lambda_{\min}$ is given by (6). Since, by definition, $\lambda_{\min}$ specifies the minimum separation of the change point parameters, the resulting partition has the property that every three consecutive segments of the partition contain *at most one* change point. However, $\lambda_{\min}$ is not known to the algorithm. Moreover, even if $\lambda \le \lambda_{\min}$, not all segments in the partition contain a change point. The algorithm uses the score function $\Delta_{\boldsymbol{x}}$ given by (8) to identify the segments that contain change points. As for $\lambda_{\min}$, instead of trying to find it, the algorithm produces many partitions of $\boldsymbol{x}$ (using different guesses of $\lambda_{\min}$), and gives a set of candidate change point estimates using each guess. Finally, a weighted combination of the candidate estimates is produced. The weights are designed to converge to zero on iterations where the algorithm's guess of a lower bound on $\lambda_{\min}$ is incorrect.

More precisely, Algorithm 1 works as follows. Given $\boldsymbol{x} \in \mathcal{X}^n$, it iterates over $j = 1..\log n$, and at each iteration it produces a guess λ_j as a lower-bound on $\lambda_{\min}$. For every fixed j, a total of $\kappa+1$ grids are generated, each composed of evenly-spaced boundaries $b_i^{t,j}$, $i = 0..\lfloor \frac{1}{\alpha_j} - \frac{1}{t+1} \rfloor$, that are $n\alpha_j$ apart for $\alpha_j := \lambda_j/3$, $\lambda_j := 2^{-j}$. This is specified in Line 6 of Algorithm 1. The grids have distinct starting positions $\frac{n\alpha_j}{t+1}$ for $t = 1..\kappa+1$. As shown in the proof of Theorem 1, this ensures that for a fixed j at least one of the grids for some $t \in 1..\kappa+1$ has the property that the change points do not lie at the boundaries. This idea is depicted in Fig. 1. Among the segments of the grid, κ segments, $X_{b^{t,j}_{[k]}..b^{t,j}_{[k]+1}}$, $k = 1..\kappa$, of highest score $\Delta_{\boldsymbol{x}}$ are selected; this is outlined in Lines 12 and 13 of the algorithm. The single-change point estimator $\Phi_{\boldsymbol{x}}$ is used to seek a candidate change point parameter in each of the selected segments. The weighted combination is given as the final estimate for every change point parameter θ_k, $k = 1..\kappa$. Two sets of weights are used, namely, an iteration weight $w_j := 2^{-j}$ and a score $\gamma(t, j)$. The former gives lower precedence to finer grids. To calculate the latter, at each iteration on j and t, for every fixed $l \in 0..2$, a partition of the grid is considered, composed of non-overlapping consecutive segments $X_{b^{t,j}_{l+3(i'-1)}..b^{t,j}_{l+3i'}}$, $i' = 1..\frac{1}{3}(\lfloor \frac{1}{\alpha_j} - \frac{1}{t+1} \rfloor - l)$ of length $n\lambda_j$. For each partition, a parameter γ_l is calculated as the κth highest intra-subsequence distance value Δ_x of its segments; the performance weight $\gamma(t, j)$ is obtained as $\min_{l=0..2} \gamma_l$; this procedure is outlined in Lines 7–11 of the algorithm. (As shown in the proof, $\gamma(t, j)$ converges to zero on iterations where either $\lambda_j > \lambda_{\min}$ or there exists some change point on the boundary of one of the segments.)

Computational complexity. The proposed method can be easily and efficiently implemented. For a fixed j, a total of $1/\alpha_j$ distance calculations are done on segments of length $3\alpha_j$, and a total of $\kappa\alpha_j n$ distance calculations are done to estimate each change point; the procedure is repeated $\kappa+1$ times. By Remark 1, and summing over $j \in 1..\log n$ iterations, the overall complexity of these calculations is bounded by $\mathcal{O}(\kappa^2 n^2 \text{ polylog } n)$. The rest of the computations are of negligible order.

5. Generalisation: AMS processes and gradual changes

In this section we argue that our results can be strengthened to a more general case, where the process distributions that generate the data are Asymptotically Mean Stationary (AMS) ergodic. We use this observation in turn to address the problem of estimating gradual as opposed to abrupt changes in the distribution of the data.

Recall that a process ρ is *stationary* if for any $i, j \in 1..n$ and $B \in B^m$, $m \in \mathbb{N}$, we have $\rho(X_{1..j} \in B) = \rho(X_{i..i+j-1} \in B)$. A process ρ is called *AMS* if for any $j \in 1..n$ and $B \in B^m, m \in \mathbb{N}$ the series $\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{n} \rho(X_{i..i+j-1} \in B)$ converges. In this case the limit, which we denote $\bar{\rho}(B)$, forms a measure $\bar{\rho}(X_{1..j} \in B) := \bar{\rho}(B)$, $B \in B^m$, $m \in \mathbb{N}$, called the *asymptotic mean* of ρ. Furthermore, for AMS processes for every $B \in B^m$, $m \in \mathbb{N}$, the frequency $\nu(X_{1..n}, B)$ converges ρ-a.s. to a random

variable with mean $\bar{\rho}(B)$. Finally, as in the case of stationary processes, if the latter random variable is a.s. constant, then ρ is called AMS ergodic. The reader is referred to [11] for more information on AMS processes.

It is easy to check that our results readily hold for the case where the unknown process distributions that generate the data are AMS ergodic, and their asymptotic means before and after the change are different. Indeed, the only property that is used in the proofs is the convergence of all frequencies. The class of all processes with AMS properties is precisely the class of all processes for which this convergence holds.

This generalisation allows us to take into consideration gradual rather than abrupt changes in distribution. So far we have considered a formulation in which the distribution is the same throughout a segment and is different between the segments. This kind of change is referred to as *abrupt*. Another formulation of the problem also considered in the literature (e.g., [1]) is when the process distributions change gradually. More formally, we are given a sequence $\boldsymbol{x} \in \mathcal{X}^n$, $n \in \mathbb{N}$ such that

$$\boldsymbol{x} := X_{1..\lfloor n\theta_1^{(1)} \rfloor}, X_{\lfloor n\theta_1^{(1)} \rfloor+1..\lfloor n\theta_1^{(2)} \rfloor}, X_{\lfloor n\theta_1^{(2)} \rfloor+1..\lfloor n\theta_2^{(1)} \rfloor}, \ldots, X_{\lfloor n\theta_k^{(2)} \rfloor+1..n}$$

has κ change points at $\lfloor n\theta_k^{(1)} \rfloor$, $k \in \kappa$. The segments $X_{\lfloor n\theta_{k-1}^{(2)} \rfloor+1..\lfloor n\theta_k^{(2)} \rfloor}$, $k = 1..\kappa$, where $\theta_0^{(2)} := 0$ and $\theta_{\kappa+1}^{(1)} := 1$, are generated by unknown, stationary ergodic process distributions. Moreover, their lengths are linear in n so that $\lambda_{\min} := \theta_k^{(2)} - \theta_{k-1}^{(2)} > 0$. The notion of gradual change is formalised by considering between every pair of consecutive segments generated by different process distributions some arbitrary sequence of $o(n)$ length, i.e. for all $k \in 1..\kappa$ we have $\theta_k^{(2)} - \theta_k^{(1)} = o(n)$, and $X_{\lfloor n\theta_1^{(1)} \rfloor+1..\lfloor n\theta_1^{(2)} \rfloor}$ is arbitrary (for example, deterministic). Observe that under this formulation the process distributions generating the segments $X_{\lfloor n\theta_{k-1}^{(1)} \rfloor+1..\lfloor n\theta_k^{(1)} \rfloor}$, $k = 1..\kappa$, where $\theta_0^{(1)} := 0$, are AMS ergodic. Thus, by the above argument the results of Theorem 1 carry over to this scenario as well, ensuring the asymptotic consistency of Algorithm 1 in this formulation. Even more generally, under the AMS ergodic assumption, such $o(n)$ segments of arbitrary data could be located anywhere within the segments, without affecting the asymptotic consistency results.

6. Experimental evaluations

In this section we evaluate our method using synthetically generated data. In order to generate the data we use stationary ergodic process distributions that do not belong to any "simpler" general class of time-series, and cannot be approximated by finite-state models. Moreover, the single-dimensional marginals of all distributions are the same throughout the generated sequence.

We generate a segment $\boldsymbol{y} := Y_1, \ldots, Y_m \in \mathbb{R}^m$, $m \in \mathbb{N}$ as follows. **1.** Fix a parameter $\alpha \in (0, 1)$ and two Gaussian distributions $\mathcal{N}_1$ and $\mathcal{N}_2$. **2.** Let r_0 be drawn randomly from $[0, 1]$. **3.** For each $i = 1..m$ obtain $r_i := r_{i-1} + \alpha \mod 1$; draw $y_i^{(j)}$ from $\mathcal{N}_j$, $j = 1, 2$. **4.** Set $Y_i := \mathbb{I}\{r_i \le 0.5\} y_i^{(1)} + \mathbb{I}\{r_i > 0.5\} y_i^{(2)}$. If α is irrational this produces a real-valued stationary ergodic time-series. We simulate α by a long double with a long mantissa. Note that deterministically setting $y_i^{(1)} = 0$ and $y_i^{(2)} = 1$, $i \in 1..m$ results in a binary sequence $\boldsymbol{x} \in \{0, 1\}^m$. Similar families are commonly used as examples in this framework; see, for example, [6].

6.1. Convergence of error-rate as a function of sequence-length n

We considered three values of κ: 4, 5 and 6. In each case, we fixed $\kappa + 1$ parameters $\alpha_1 := 0.2..$, $\alpha_2 := 0.4..$, $\alpha_3 := 0.6..$, ... (with long mantissae)[1] to correspond to different process distributions and used two Gaussian distributions $\mathcal{N}_1$ and $\mathcal{N}_2$ with means 0 and 1 respectively, and standard deviation 1. To produce $\boldsymbol{x} \in \mathbb{R}^n$ in each case we used the first κ change point parameters from the following sequence of 6 values $\theta_1 = 0.18$, $\theta_2 = 0.29$, $\theta_3 = 0.51$, $\theta_4 = 0.62$, $\theta_5 = 0.80$ and $\theta_6 = 0.91$, and respectively set $\theta_0 = 0$ and $\theta_{\kappa+1} = 1$. Notice that for $\kappa = 6$ the minimum separation $\lambda_{\min}$ between the change points is 0.09 and for $\kappa = 4, 5$ it is 0.1. Every segment of length $n_k := \lfloor n(\theta_k - \theta_{k-1}) \rfloor$, $k = 1..\kappa + 1$ with $\theta_0 := 0$, $\theta_{\kappa+1} := 1$ was generated with α_k, $k = 0..\kappa + 1$, and using $\mathcal{N}_1$ and $\mathcal{N}_2$. Fig. 2 demonstrates the average estimation error of Algorithm 1 as a function of sequence length n. We calculate the estimation error as $\sum_{k=1}^{\kappa} |\hat{\theta}_k - \theta_k|$. As can be seen in the graph, while the error converges to zero in all three cases, this convergence is on average slightly slower for larger κ.

6.2. Dependence on the minimum separation $\lambda_{\min}$ between the change points

We fixed the sequence length n, and varied $\lambda_{\min}$ to observe the average error change as a function $\lambda_{\min}$. More specifically, we generated sequences $\boldsymbol{x} \in \{0, 1\}^n$, $n = 3.0 \times 10^4$ with $\kappa = 4$ change-points as the concatenation of 5 segments of lengths $5000, n_0, 12000 - n_0, 7000, 6000$ respectively, where $n_0 = 1000, 1250, 1500, \ldots, 5000$; note that $\lambda_{\min} = n_0/n$ in this case. To generate the segments we proceeded as in the previous experiment but with binary-valued processes (letting $y_i^{(1)} = 0$

[1] $\alpha_1 = 0.22573625315372165312763512$ $\alpha_2 = 0.46545635654654376453$ $\alpha_3 = 0.67863827632786327836273628362 8$
$\alpha_4 = 0.887438463874637846343$ $\alpha_5 = 0.07283729372372987323232323$ $\alpha_6 = 0.42726387263827362832791217312893$ $\alpha_7 = \alpha_1$.

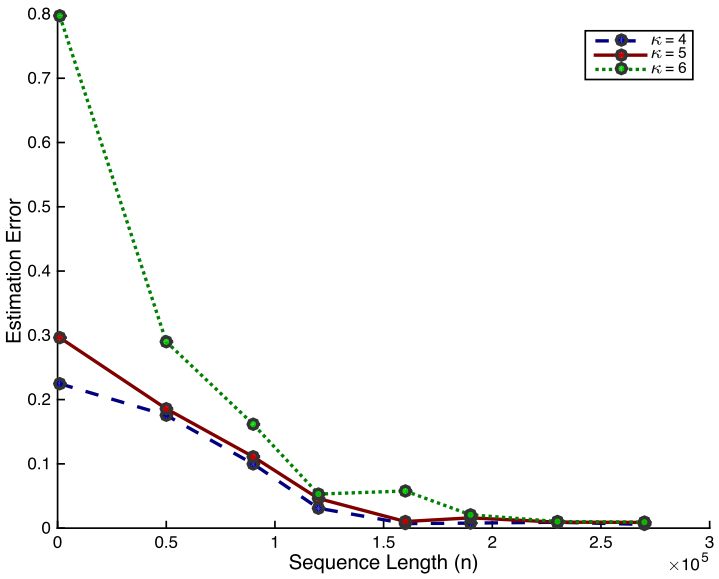

Fig. 2. Average (over 20 iterations) error of Alg1$(\boldsymbol{x}, \kappa)$, $\boldsymbol{x} \in \mathbb{R}^n$, as a function of n for $\kappa = 4, 5, 6$.

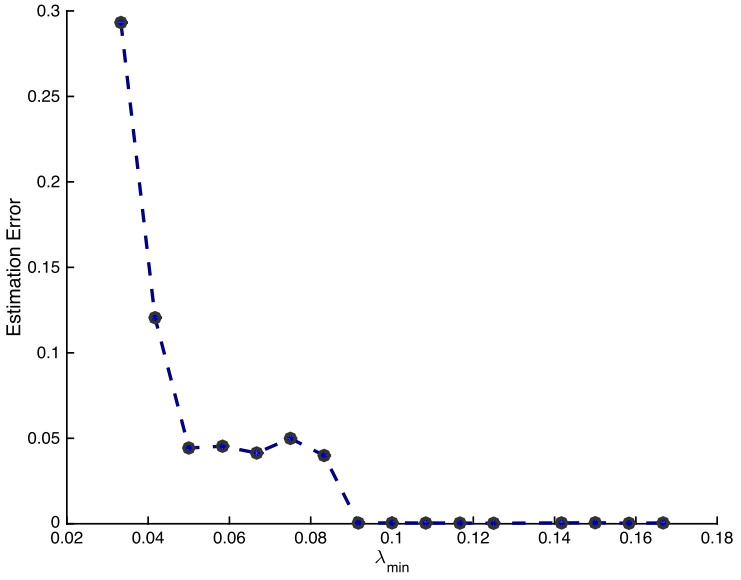

Fig. 3. Average (over 10 iterations) error of Alg1$(\boldsymbol{x}, \kappa)$, $\boldsymbol{x} \in \{0, 1\}^n$ as a function of $\lambda_{\min}$ for $n = 3.0 \times 10^4$ and $\kappa = 4$.

and $y_i^{(2)} = 1$, $i \in 1..n$), and used $\alpha_1 := 0.12..$, $\alpha_2 := 0.14..$, $\alpha_3 := 0.16..$, ... (with long mantissae)[2] as parameters for the consecutive distributions. As can be seen in Fig. 3, for a fixed n, the estimation error decreases as a function of $\lambda_{\min}$.

7. Concluding remarks

We have presented an asymptotically consistent method to locate the changes in highly dependent time-series data. As explained in the introduction, in the considered setting rates of convergence (even of frequencies to respective probabilities) are provably impossible to obtain, which is why the proposed algorithm comes only with asymptotic guarantees. At the same time, it may be interesting to analyse how fast its error converges to zero under stronger assumptions, such as i.i.d. or mixing conditions. More generally, it would be interesting to discover whether asymptotic guarantees in the considered settings can be combined with optimality (up to constant factors) under stronger assumptions. The same questions can be

2 $\alpha_1 = 0.12257362531537216531276351 2$ $\alpha_2 = 0.1465456356354654376453$ $\alpha_3 = 0.16786382763278632783627362836287 36$
$\alpha_4 = 0.1887438463874637846343$ $\alpha_5 = 0.10728372937237298732323232323.$

posed for the problems of [9,10] described in the introduction, where the number of change points is unknown. This is left for future work. Another interesting question concerns the time series distance used in the algorithms. The consistency result is established using some properties specific to the (empirical estimates of the) distributional distance. It could be interesting to see whether analogous results can be established for other distances which can be estimated consistently for stationary ergodic process distributions, such as the telescope distance of [16], or whether such results can be obtained with compression-based methods in the spirit of [17]. Finally, an interesting generalisation of the considered problems could be a one to processes stationary over more than one dimension, such as space-time stationary processes. In this case, the change-point estimation problem turns into the problem of approximating the boundaries of regions in which the data are generated by the same distribution.

8. Proof of Theorem 1

Proof Sketch. To see why Algorithm 1 works, first observe that the empirical estimate $\hat{d}(\cdot,\cdot)$ of the distributional distance is consistent. Thus, the empirical distributional distance between a given pair of sequences converges to the distributional distance between their generating processes. From this we can show that the intra-subsequence distance $\Delta_{\boldsymbol{x}}$ corresponding to the segments in the grid that do not contain a change point converges to zero. This is established in Lemma 1.(iii) below. On the other hand, since the generated grid becomes finer as a function of j, from some j on, we have $\alpha_j < \lambda_{\min}/3$ so that every three consecutive segments of the grid contain *at most* one change point. In this case, for every segment that contains a change point, the single-change-point estimator $\Phi_{\boldsymbol{x}}$ produces an estimate that, for long enough segments, becomes arbitrarily close to the true change point. This is shown in Lemma 3.(ii) below. Moreover, as follows from Lemma 3.(i), for large enough n the performance scores associated with these segments are bounded below by some non-zero constant. Thus, the κ segments of highest $\Delta_{\boldsymbol{x}}$ each contain a change point which can be estimated consistently using $\Phi_{\boldsymbol{x}}$. However, the estimates produced at a given iteration for which $\alpha_j > \lambda_{\min}/3$ may be arbitrarily bad. Moreover, recall that even for $\alpha_j \leq \lambda_{\min}/3$, an appropriate grid to provide consistent estimates must have the property that no change point lies exactly on a grid boundary. However, it is not possible to directly identify such appropriate grids. The following observation is key to their indirect identification.

Consider the partitioning of $\boldsymbol{x}$ into κ consecutive segments where there exists at least one segment with more than one change point. Since there are exactly κ change points, there must exist at least one segment in this partitioning that does not contain any change points. As follows from Lemma 1.(iii), the segment that contains no change points has an intra-subsequence distance $\Delta_{\boldsymbol{x}}$ that converges to 0. On the iterations for which $\alpha_j > \lambda_{\min}/3$, at least one of the three partitions has the property that among every set of κ segments in the partition, there is *at least* one segment that contains no change points. In this case, $\Delta_{\boldsymbol{x}}$ corresponding to the segment without a change point converges to 0. The same argument holds for the case where $\alpha_j \leq \lambda_{\min}$, while at the same time a change point happens to be located exactly at the boundary of a segment in the grid. Observe that for a fixed j, the algorithm forms a total of $\kappa+1$ different grids, with the same segment size, but distinct starting points $\frac{n\alpha_j}{t+1}$ $t = 1..\kappa+1$. Since there are κ change points, for all j such that $\alpha_j \leq \lambda_{\min}/3$ there exists at least one appropriate grid (for some $\tau \in 1..\kappa+1$), that simultaneously contains all the change points within its segments. In this case, $\gamma(\tau, j)$ converges to a non-zero constant. The final estimate $\hat{\theta}_k$ for each change point parameter θ_k is obtained as a weighted sum of the candidate estimates produced at each iteration. Two sets of weights are used in this step, namely $\gamma(t, j)$ and w_j, whose roles can be described as follows.

1. $\gamma(t, j)$ is used to penalise for the (arbitrary) results produced on iterations on $j \in 1..\log n$ and $t \in 1..\kappa+1$, where either $\alpha_j > \lambda_{\min}/3$, or, while we have $\alpha_j \leq \lambda_{\min}/3$, there exists some θ_k for some $k \in 1..\kappa$ such that $\lfloor n\theta_k \rfloor \in \{b_i^{t,j} : i = 0..\lfloor \frac{1}{\alpha_j} - \frac{1}{t+1} \rfloor\}$. As follows from the argument above, $\gamma(t, j)$ converges to zero only on these iterations, while it is bounded below by a non-zero constant on the rest.
2. w_j is used to give precedence to estimates sought in longer segments. Since the grids are finer for larger j, at some higher iterations the segments may not be long enough to produce correct estimates.

Therefore, if n is large enough the final estimates $\hat{\theta}_k$, $k = 1..\kappa$ produced by Algorithm 1 converge to the true change point parameters, θ_k, $k = 1..\kappa$. We now present a proof for Theorem 1 which in turn depends upon some technical lemmas below.

Lemma 1. *Let $\boldsymbol{x} = X_{1..n}$ be generated by a stationary ergodic process ρ. For all $\alpha \in (0,1)$ the following statements hold with ρ-probability 1:*

(i) $\lim_{n\to\infty} \sup_{\substack{b_1,b_2\in 1..n \\ b_2-b_1\geq \alpha n}} \sum_{\substack{B\in B^{m,l} \\ m,l\in 1..T}} \left|\nu(X_{b_1..b_2}, B) - \rho(B)\right| = 0$ *for every* $T \in \mathbb{N}$.

(ii) $\lim_{n\to\infty} \sup_{\substack{b_1,b_2\in 1..n \\ b_2-b_1\geq \alpha n}} \hat{d}\left(X_{b_1..b_2}, \rho\right) = 0.$

(iii) $\lim_{n\to\infty} \sup_{b_2-b_1\geq \alpha n} \Delta_{\boldsymbol{x}}(b_1, b_2) = 0.$

Proof. (i). Assume the contrary: There exists and some $\lambda > 0$, $T \in \mathbb{N}$ and sequences $b_1^{(i)} \in 1..n_i$ and $b_2^{(i)} \in 1..n_i$, $n_i, i \in \mathbb{N}$ with $b_2^{(i)} - b_1^{(i)} \geq \alpha n_i$, such that with probability $\Delta > 0$ we have

$$\limsup_{i\in\mathbb{N}} \sum_{\substack{B\in B^{m,l}\\ m,l\in 1..T}} \left|\nu(X_{b_1^{(i)}..b_2^{(i)}}, B) - \rho(B)\right| > \lambda. \tag{10}$$

From the definition of $\nu(\cdot,\cdot)$ given by (1), it is easy to see that for all $B \in B^{m,l}$, $m, l \in \mathbb{N}$ and $b_1 < b_2 \in 1..n$ we have

$$\begin{aligned}\left|\nu(X_{b_1..b_2}, B) - \rho(B)\right| &\leq \left|\left(1 - \frac{m-1}{b_2-b_1}\right)\nu(X_{b_1..b_2}, B) - \rho(B)\right| + \frac{m-1}{b_2-b_1}\\ &\leq \frac{4(m-1)}{b_2-b_1} + \sum_{i=1}^{2}\frac{b_i}{b_2-b_1}\left|\nu(X_{1..b_i}, B) - \rho(B)\right|. \end{aligned}\tag{11}$$

Fix $\varepsilon > 0$. For each $m, l \in 1..T$ we can find a finite subset $S^{m,l} \subset B^{m,l}$ such that

$$\rho(S^{m,l}) \geq 1 - \frac{\varepsilon}{T^2 w_m w_l}. \tag{12}$$

Since ρ is stationary ergodic, for every $B \in S^{m,l}$, there exists some $N(B)$ such that with probability 1 for all $t \geq N(B)$

$$\sup_{b\geq t}|\nu(X_{1..b}, B) - \rho(B)| \leq \frac{\varepsilon\rho(B)}{T^2 w_m w_l}. \tag{13}$$

Define $\zeta_0 := \min_{m,l\in 1..T}\frac{\varepsilon}{T^2 w_m w_l}$. (Note that in the particular case where $w_i = 1/i(i+1)$, $i = m, l$, we simply get $\zeta_0 = 4\varepsilon/T^2$, but we keep this parameter in its general form, independently of the specific choice of w_m and w_l.) Let $\zeta := \min\{\alpha, \zeta_0\}$ and observe that $\zeta > 0$. For every $m, l \in 1..T$ and all $t \in \mathbb{N}$ we have,

$$\sup_{\substack{b_1\leq\zeta t\\ b_2-b_1\geq\alpha t}} \frac{b_1}{b_2-b_1} \leq \frac{\zeta}{\alpha} \leq \frac{\varepsilon}{\alpha T^2 w_m w_l}. \tag{14}$$

Define $N := \max_{m,l=1..T} N(B)/\zeta$. On the other hand, by (13) for all $n \geq N$ we have

$$\sup_{b_1>\zeta n}\left|\nu(X_{1..b_1}, B) - \rho(B)\right| \leq \frac{\varepsilon\rho(B)}{T^2 w_m w_l}. \tag{15}$$

Increase N if necessary to have

$$\sum_{m,l=1}^{T} w_m w_l \frac{m}{\alpha N} \leq \varepsilon. \tag{16}$$

For all $n \geq N$ we obtain

$$\begin{aligned}&\sup_{\substack{b_1,b_2\in 1..n\\ b_2-b_1\geq\alpha n}} \sum_{m,l=1}^{T} w_m w_l \sum_{B\in B^{m,l}}\left|\nu(X_{b_1..b_2}, B) - \rho(B)\right|\\ &\leq \sup_{\substack{b_1,b_2\in 1..n\\ b_2-b_1\geq\alpha n}} \sum_{m,l=1}^{T} w_m w_l \sum_{B\in S^{m,l}}\left|\nu(X_{b_1..b_2}, B) - \rho(B)\right| + \varepsilon \qquad (17)\\ &\leq \sup_{\substack{b_1,b_2\in 1..n\\ b_2-b_1\geq\alpha n}} \sum_{m,l=1}^{T} w_m w_l \sum_{B\in S^{m,l}}\frac{b_2}{b_2-b_1}\left|\nu(X_{1..b_2}, B) - \rho(B)\right|\\ &\quad + \sup_{\substack{b_1>\zeta n\\ b_2-b_1\geq\alpha n}} \sum_{m,l=1}^{T} w_m w_l \sum_{B\in S^{m,l}}\frac{b_1}{b_2-b_1}\left|\nu(X_{1..b_1}, B) - \rho(B)\right|\\ &\quad + \sup_{\substack{b_1\leq\zeta n\\ b_2-b_1\geq\alpha n}} \sum_{m,l=1}^{T} w_m w_l \sum_{B\in S^{m,l}}\frac{b_1}{b_2-b_1}\left|\nu(X_{1..b_1}, B) - \rho(B)\right| + 5\varepsilon \qquad (18)\\ &\leq \varepsilon(3/\alpha + 5) \qquad (19)\end{aligned}$$

where, (17) follows from (12), (18) follows from (11) and (16), and (19) follows from (13), (15), (14), summing over the probabilities, and noting that $\frac{b_2}{b_2-b_1} \leq \frac{1}{\alpha}$ for all $b_2 - b_1 \geq \alpha n$. Observe that (19) holds for any $\varepsilon > 0$, and in particular for $\varepsilon \in (0, \frac{\lambda}{3/\alpha+5})$. As a result, in the latter case for all $n \geq N$ we have

$$\sup_{\substack{i\in\mathbb{N}\\ n_i\geq n}} \sum_{\substack{B\in B^{m,l}\\ m,l\in 1..T}} \left|\nu(X_{b_1^{(i)}..b_2^{(i)}}, B) - \rho(B)\right| < \lambda,$$

contradicting (10). This contradiction implies (i).

(ii). Fix $\varepsilon > 0$, $\alpha \in (0,1)$. We can find some $T \in \mathbb{N}$ such that

$$\sum_{m,l=T}^{\infty} w_m w_l \leq \varepsilon. \tag{20}$$

By (i), there exists some N such that for all $n \geq N$ we have

$$\sup_{\substack{b_1,b_2\in 1..n\\ |b_2-b_1|\geq\alpha n}} \sum_{m,l=1}^{T} \sum_{B\in B^{m,l}} \left|\nu(X_{b_1..b_2}, B) - \rho(B)\right| \leq \varepsilon. \tag{21}$$

From (20) and (21), for all $n \geq N$ we have

$$\sup_{\substack{b_1,b_2\in 1..n\\ |b_2-b_1|\geq\alpha n}} \hat{d}(X_{b_1..b_2}, \rho) \leq \sup_{\substack{b_1,b_2\in 1..n\\ |b_2-b_1|\geq\alpha n}} \sum_{m,l=1}^{T} w_m w_l \sum_{B\in B^{m,l}} \left|\nu(X_{b_1..b_2}, B) - \rho(B)\right| + \varepsilon \leq 2\varepsilon$$

and Part (ii) follows.

(iii). Fix $\varepsilon > 0$, $\alpha \in (0,1)$. Without loss of generality assume that $b_2 > b_1$. Observe that, for every $b_1 + \alpha n \leq b_2 \leq n$, we have $\frac{b_1+b_2}{2} - b_1 = b_2 - \frac{b_1+b_2}{2} \geq \alpha n/2$. Therefore, by (ii), there exists some N, such that for all $n \geq N_1$ we have

$$\sup_{b_2-b_1\geq\alpha n} \hat{d}\left(X_{b_1..\frac{b_1+b_2}{2}}, \rho\right) \leq \varepsilon,$$

$$\sup_{b_2-b_1\geq\alpha n} \hat{d}\left(X_{\frac{b_1+b_2}{2}..b_2}, \rho\right) \leq \varepsilon.$$

It remains to use the definition of $\Delta_{\boldsymbol{x}}$ given by (8) and the triangle inequality to observe that

$$\begin{aligned}\sup_{b_2-b_1\geq\alpha n} \Delta_{\boldsymbol{x}}(b_1,b_2) &= \sup_{b_2-b_1\geq\alpha n} \hat{d}\left(X_{b_1..\frac{b_1+b_2}{2}}, X_{\frac{b_1+b_2}{2}..b_2}\right)\\ &\leq \sup_{b_2-b_1\geq\alpha n} \hat{d}\left(X_{b_1..\frac{b_1+b_2}{2}}, \rho\right) + \hat{d}\left(X_{\frac{b_1+b_2}{2}..b_2}, \rho\right) \leq 2\varepsilon\end{aligned}$$

for all $n \geq N$, and (iii) follows. □

Lemma 2. *Let $\boldsymbol{x} \in \mathcal{X}^n$ have a change point at $\pi = \theta n$ for some $\theta \in (0,1)$ so that the segments $X_{1..\pi}$, $X_{\pi..n}$ are generated by ρ, ρ' respectively. If ρ, ρ' are stationary ergodic, for every $\zeta \in (0, \min\{\theta, 1-\theta\})$ with probability 1 we have*

(i) $$\lim_{n\to\infty} \sup_{\substack{b\in 1..(\theta-\zeta)n\\ t\in\pi..(1-\zeta)n}} \hat{d}\left(X_{b..t}, \frac{\pi-b}{t-b}\rho + \frac{t-\pi}{t-b}\rho'\right) = 0.$$

(ii) $$\lim_{n\to\infty} \sup_{\substack{b\in\zeta n..\pi\\ t\in(\theta+\zeta)n..n}} \hat{d}\left(X_{b..t}, \frac{\pi-b}{t-b}\rho + \frac{t-\pi}{t-b}\rho'\right) = 0.$$

Proof. (i). Fix $\varepsilon > 0$, $\theta \in (0,1)$, $\zeta \in (0, \min\{\theta, 1-\theta\})$. There exists some $T \in \mathbb{N}$ such that $\sum_{m,l=T}^{\infty} w_m w_l \leq \varepsilon$. By the definition of $\nu(\cdot,\cdot)$, for all $b \in 1..(\theta-\zeta)n$, $t \in \pi..(1-\zeta)n$ and all $B \in B^{m,l}$ $m,l \in 1..T$ we have

$$\begin{aligned}\left|\nu(X_{\pi..t}, B) - \rho'(B)\right| \leq{}& \frac{n-\pi}{t-\pi-m+1}\left|\nu(X_{\pi..n}, B) - \rho'(B)\right|\\ &+ \frac{n-t}{t-\pi-m+1}\left|\nu(X_{t..n}, B) - \rho'(B)\right| + \frac{3(m-1)}{t-\pi-m+1}.\end{aligned} \tag{22}$$

Furthermore, using the fact that $\nu(\cdot,\cdot) \leq 1$, for all $b \in 1..(\theta-\zeta)n$, $t \in \pi..(1-\zeta)n$ and $B \in B^{m,l}$ $m,l \in 1..T$ we obtain

$$\left|\nu(X_{b..t}, B) - \frac{\pi - b}{t-b}\rho(B) - \frac{t-\pi}{t-b}\rho'(B)\right|$$
$$\leq \frac{\pi - b}{t-b}\left|\nu(X_{b..\pi}, B) - \rho(B)\right| + \frac{t-\pi-m+1}{t-b}\left|\nu(X_{\pi..t}, B) - \rho'(B)\right| + \frac{3(m-1)}{t-b}. \tag{23}$$

By Part (i) of Lemma 1, there exists some N' such that for all $n \geq N'$ we have

$$\sup_{b\in 1..(\theta-\zeta)n} \sum_{m,l=1}^{T} w_m w_l \sum_{B\in B^{m,l}} |\nu(X_{b..\pi}, B) - \rho(B)| \leq \varepsilon. \tag{24}$$

Similarly, $n - t \geq \zeta n$ for all $t \in \pi..(1-\zeta)n$. Therefore, by Part (i) of Lemma 1, there exists some N'' such that for all $n \geq N''$ we have

$$\sup_{t\in \pi..(1-\zeta)n} \sum_{m,l=1}^{T} w_m w_l \sum_{B\in B^{m,l}} \left|\nu(X_{t..n}, B) - \rho'(B)\right| \leq \varepsilon. \tag{25}$$

Since $t - b \geq \zeta n$ for all $b \in 1..(\theta-\zeta)n$, $t \in \pi..(1-\zeta)n$, we have $\frac{n}{t-b} \leq \frac{1}{\zeta}$. For all $n \geq \frac{T}{\varepsilon\zeta}$, $m \in 1..T$, $b \in 1..(\theta-\zeta)n$ and $t \in \pi..(1-\zeta)n$ we have $\frac{m-1}{t-b} \leq \frac{m}{\zeta n} \leq \varepsilon$. Let $N := \max\{N', N'', \frac{T}{\varepsilon\zeta}\}$. By (22), (23), (24), (25), for all $n \geq N$ we have

$$\sup_{\substack{b\in 1..(\theta-\zeta)n \\ t\in \pi..(1-\zeta)n}} \sum_{m,l=1}^{T} w_{m,l} \sum_{B\in B^{m,l}} \left|\nu(X_{b..t}, B) - \frac{\pi-b}{t-b}\rho(B) - \frac{t-\pi}{t-b}\rho'(B)\right| \leq 3\varepsilon(2 + \frac{1}{\zeta}).$$

By this, and the definition of T, for all $n \geq N$ we have $\sup_{\substack{b\in 1..(\theta-\zeta)n \\ t\in \pi..(1-\zeta)n}} \hat{d}(X_{b..t}, \frac{\pi-b}{t-b}\rho + \frac{t-\pi}{t-b}\rho') \leq \varepsilon(7 + \frac{3}{\zeta})$ and Part (i) follows. The proof of the second part is analogous. □

Lemma 3. *Consider a sequence $\boldsymbol{x} \in \mathcal{X}^n$, $n \in \mathbb{N}$ with κ change points. Let $\boldsymbol{b} := b_1, \ldots, b_{|\boldsymbol{b}|} \in \cup_{i=1}^{n}\{1..n\}^i$, be a sequence of indices with $\min_{i\in 1..|\boldsymbol{b}|-1} b_{i+1} - b_i \geq \alpha n$ for some $\alpha \in (0,1)$, such that for some $\zeta \in (0,1)$ we have $\inf_{k=1..\kappa, b\in\boldsymbol{b}} |\frac{1}{n}b - \theta_k| \geq \zeta$.*

(i) *With probability 1 we have $\lim_{n\to\infty} \inf_{k\in 1..\kappa} \Delta_{\boldsymbol{x}}(L(k), R(k)) \geq \delta\zeta$ where $L(k) := \max_{b\leq n\theta_k}\{b \in \boldsymbol{b}\}$ and $R(k) := \max_{b>n\theta_k}\{b \in \boldsymbol{b}\}$ denote the elements of $\boldsymbol{b}$ that appear immediately to the left and to the right of $\lfloor n\theta_k \rfloor$ respectively, and δ is the minimum distance between the distinct distributions that generate $\boldsymbol{x}$.*

(ii) *Assume that we additionally have $[\frac{1}{n}L(k) - \alpha, \frac{1}{n}R(k) + \alpha] \subseteq [\theta_{k-1}, \theta_{k+1}]$. With probability 1 we obtain $\lim_{n\to\infty} \sup_{k\in 1..\kappa} |\frac{1}{n}\Phi_{\boldsymbol{x}}(L(k), R(k), \alpha) - \theta_k| = 0$.*

Proof. (i). Fix some $k \in 1..\kappa$. Define $c_k := \frac{L(k)+R(k)}{2}$. To prove Part (i), we show that with probability 1 for large enough n, we have

$$\hat{d}\left(X_{L(k)..c_k}, X_{c_k..R(k)}\right) \geq \delta\zeta. \tag{26}$$

Fix $\varepsilon > 0$. Let $\pi_k := \lfloor n\theta_k \rfloor$, $k = 1..\kappa$. To prove (26) for the case where $\pi_k \leq c_k$ we proceed as follows. As follows from the assumption of the lemma and the definition of $L(\cdot)$ and $R(\cdot)$, we have $R(k) - L(k) \geq n\alpha$, so that $R(k) - c_k \geq \frac{\alpha}{2}n$. Since by assumption of the lemma we have $\inf_{k=1..\kappa, b\in\boldsymbol{b}} |\frac{1}{n}b - \theta_k| \geq \zeta$, it follows that $\pi_{k+1} - c_k \geq (\zeta + \frac{\alpha}{2})n$. Moreover, from the same assumption we have $\frac{\pi_k - L(k)}{c_k - L(k)} \geq \frac{\pi_k - L(k)}{n} \geq \zeta$. Therefore, we obtain

$$d\left(\rho_{k+1}, \frac{\pi_k - L(k)}{c_k - L(k)}\rho_k + \frac{c_k - \pi_k}{c_k - L(k)}\rho_{k+1}\right) = \frac{\pi_k - L(k)}{c_k - L(k)} d\left(\rho_{k+1}, \rho_k\right) \geq \delta\zeta. \tag{27}$$

From the definition of $L(k)$ and $R(k)$, and our assumption that $\pi_k \leq c_k$, the segment $X_{c_k..R(k)}$ is fully generated by ρ_{k+1}. By Part (ii) of Lemma 1, there exists some N_1 such that for all $n \geq N_1$ we have

$$\hat{d}\left(X_{c_k..R(k)}, \rho_{k+1}\right) \leq \varepsilon. \tag{28}$$

By Part (i) of Lemma 2 there exists some N_2 such that for all $n \geq N_2$ we have

$$\hat{d}\left(X_{L(k)..c_k}, \frac{\pi_k - L(k)}{c_k - L(k)}\rho_k + \frac{c_k - \pi_k}{c_k - L(k)}\rho_{k+1}\right) \leq \varepsilon. \tag{29}$$

By (29) and (27) for all $n \geq \max_{i=1,2} N_i$ we obtain

$$\begin{aligned}\Delta_{\boldsymbol{x}}(L(k),R(k)) &\geq \hat{d}\left(X_{L(k)..c_k},\rho_{k+1}\right)-\hat{d}\left(X_{c_k..R(k)},\rho_{k+1}\right)\\ &\geq d\left(\rho_{k+1},\frac{\pi_k-L(k)}{c_k-L(k)}\rho_k+\frac{c_k-\pi_k}{c_k-L(k)}\rho_{k+1}\right)\\ &\quad-\hat{d}\left(X_{L(k)..c_k},\frac{\pi_k-L(k)}{c_k-L(k)}\rho_k+\frac{c_k-\pi_k}{c_k-L(k)}\rho_{k+1}\right)-\hat{d}\left(X_{c_k..R(k)},\rho_{k+1}\right)\geq\delta\zeta-2\varepsilon. \end{aligned}\tag{30}$$

Since (30) holds for every $\varepsilon>0$, this proves (26) in the case where $\pi_k\leq c_k$. The proof for $\pi_k>c_k$ is analogous. Since (26) holds for all $k\in 1..\kappa$, part (i) follows.

(ii). Fix some $k\in 1..\kappa$. Following the definition of $\Phi_{\boldsymbol{x}}$ given by (9) we have

$$\Phi(L(k)-n\alpha,R(k)+n\alpha,\alpha):=\operatorname*{argmax}_{l'\in L(k)..R(k)}\hat{d}\left(X_{L(k)-n\alpha..l'},X_{l'..R(k)+n\alpha}\right).$$

We show that for any $\beta\in(0,1)$, with probability 1 for large enough n we have

$$\hat{d}\left(X_{L(k)-n\alpha..l'},X_{l'..R(k)+n\alpha}\right)<\hat{d}\left(X_{L(k)-n\alpha..\pi_k},X_{\pi_k..R(k)+n\alpha}\right),\tag{31}$$

for all $l'\in L(k)..(1-\beta)\pi_k\cup\pi_k(1+\beta)..R(k)$. To prove (31) for $l'\in L(k)..(1-\beta)\pi_k$ we proceed as follows. Fix some $\beta\in(0,1)$ and $\varepsilon>0$. For all $l'\in L(k)..(1-\beta)\pi_k$ we have $\frac{\pi_k-l'}{R(k)+n\alpha-l'}\geq\beta$. Hence, by the definitions of $\hat{d}$ and δ we obtain

$$d\left(\rho_k,\rho_{k+1}\right)-d\left(\rho_k,\frac{\pi_k-l'}{R(k)+n\alpha-l'}\rho_k+\frac{R(k)+n\alpha-\pi_k}{R(k)+n\alpha-l'}\rho_{k+1}\right)\geq\beta\delta.\tag{32}$$

By Part (ii) of Lemma 1, there exists some N_1 such that for all $n\geq N_1$ we have

$$\sup_{l'\in L(k)..\pi_k}\hat{d}\left(X_{L(k)-n\alpha..l'},\rho_k\right)\leq\varepsilon,\tag{33}$$

$$\hat{d}\left(X_{\pi_k..R(k)+n\alpha},\rho_{k+1}\right)\leq\varepsilon.\tag{34}$$

For all $l'\in L(k)..\pi_k$ we have $l'-\pi_{k-1}\geq\alpha n$. Also, $R(k)+n\alpha\in\pi_k+n\alpha..\pi_{k+1}$. Therefore by Part (ii) of Lemma 2 there exists some N_2 such that

$$\sup_{l'\in L(k)..\pi_k}\hat{d}\left(X_{l'..R(k)+n\alpha},\frac{\pi_k-l'}{R(k)+n\alpha-l'}\rho_k+\frac{R(k)+n\alpha-\pi_k}{R(k)+n\alpha-l'}\rho_{k+1}\right)\leq\varepsilon.\tag{35}$$

By (33), (34) and the triangle inequality, for all $n\geq\max_{i=1,2}N_i$ we obtain

$$\hat{d}\left(X_{L(k)-n\alpha..\pi_k},X_{\pi_k..R(k)+n\alpha}\right)\geq\hat{d}\left(\rho_k,\rho_{k+1}\right)-2\varepsilon.\tag{36}$$

By (33), (35), and using the triangle inequality, for all $n\geq\max_{i=1,2}N_i$ we obtain

$$\begin{aligned}&\sup_{l'\in L(k)..(1-\beta)\pi_k}\hat{d}\left(X_{L(k)-n\alpha..l'},X_{l'..R(k)+n\alpha}\right)\\ &\leq\sup_{l'\in L(k)..(1-\beta)\pi_k}d\left(\rho_k,\frac{\pi_k-l'}{R(k)+n\alpha-l'}\rho_k+\frac{R(k)+n\alpha-\pi_k}{R(k)+n\alpha-l'}\rho_{k+1}\right)+2\varepsilon.\end{aligned}\tag{37}$$

Finally, from (36), (37) and (32) for all $n\geq\max_{i=1,2}N_i$ we obtain

$$\inf_{l'\in L(k)..(1-\beta)\pi_k}\hat{d}\left(X_{L(k)-n\alpha..\pi_k},X_{\pi_k..R(k)+n\alpha}\right)-\hat{d}\left(X_{L(k)-n\alpha..l'},X_{l'..R(k)+n\alpha}\right)\geq\beta\delta-4\varepsilon.\tag{38}$$

Since (38) holds for every $\varepsilon>0$, this proves (31) for $l'\in L(k)..(1-\beta)\pi_k$, $k\in 1..\kappa$. The case where $l'\in(1+\beta)\pi_k..R(k)$ is analogous; Part (ii) follows. □

Proof of Theorem 1. On each iteration on $j\in 1..\log n$ the algorithm produces a set of estimated change points. We show that on some iterations these estimates are consistent, and that estimates produced on the rest of the iterations are negligible. To this end, we will partition the set of iterations into three sets as described in Steps 1–3 below.

Define $\zeta(t,j):=\min_{k\in 1..\kappa,i\in 0..\lfloor\frac{1}{\alpha_j}-\frac{1}{t+1}\rfloor}|\alpha_j(i+\frac{1}{t+1})-\theta_k|$, $j=1..\log n$, $t\in 1..\kappa+1$; for all $i=0..\lfloor\frac{1}{\alpha_j}-\frac{1}{t+1}\rfloor$ we have $|b_i^{t,j}-\pi_k|\geq n\zeta(t,j)$.

Step 1. Fix $\varepsilon>0$. There exist some J_ε such that $\sum_{j=J_\varepsilon}^{\infty}w_j\leq\varepsilon$. J_ε is used to cut off the iterations over $j\in[1..\log n]$ where λ_j is too small for the estimates of the distributional distance between the segments to be consistent (the grids are too fine). These iterations are penalised by small weights w_j, so that the corresponding candidate estimates become negligible (their combined weight is less than ε).

Step 2. Let $J(\lambda_{\min}):=-\log(\lambda_{\min}/3)$, where $\lambda_{\min}$ is given by (6). The iterations on j for $j\in[J(\lambda_{\min}),J_\varepsilon]$ correspond to iterations where $\lambda_j\in(0,\lambda_{\min}]$ and, moreover, the segments are long enough for the estimates to be consistent as we

show below. For all $j \geq J(\lambda_{\min})$ and $t \in 1..\kappa+1$, and every θ_k, $k \in 1..\kappa$ we have $[\frac{1}{n}L(k) - \alpha_j, \frac{1}{n}R(k) + \alpha_j] \subseteq [\theta_{k-1}, \theta_{k+1}]$ where $L(\cdot)$ and $R(\cdot)$ are defined in Lemma 3. For every fixed $j \in J(\lambda_{\min})..J_\varepsilon$ we identify a subset $\mathcal{T}(j)$ of the iterations on $t = 1..\kappa+1$ at which the change point parameters θ_k, $k = 1..\kappa$ are estimated consistently and the performance scores $\gamma(t, j)$, $j \in J(\lambda_{\min})..J_\varepsilon$, $t \in \mathcal{T}(j)$ are bounded below by a nonzero constant. Moreover, we show that if the set $\mathcal{T}'(j) := \{1..\kappa+1\} \setminus \mathcal{T}(j)$ is nonempty, the performance scores $\gamma(t, j)$ for all $j \in J(\lambda_{\min})..J_\varepsilon$ and $t \in \mathcal{T}'(j)$ are arbitrarily small.

i. To define $\mathcal{T}(j)$ we proceed as follows. For every fixed $j \in J(\lambda_{\min})..J_\varepsilon$, for every θ_k, $k = 1..\kappa$ we can uniquely define $q_k \in \mathbb{N}$ and $p_k \in [0, \alpha_j)$ so that $\theta_k = q_k\alpha_j + p_k$. Therefore, for any $p \in [0, \alpha_j)$ with $p \neq p_k$, $k = 1..\kappa$, we have $\inf_{k=1..\kappa, i\in\mathbb{N}\cup\{0\}} |i\alpha_j + p - \theta_k| > 0$. Since we can only have κ distinct residues p_k, $k = 1..\kappa$, any set of $\kappa+1$ different elements of $[0, \alpha_j)$ contains at least one element p' such that $p' \neq p_k$, $k = 1..\kappa$. So, for every $j \in J(\lambda_{\min})..J_\varepsilon$ there exists at least one $t \in 1..\kappa+1$ such that $\zeta(t, j) > 0$. For every $j \in J(\lambda_{\min})..J_\varepsilon$, define

$$\mathcal{T}(j) := \{t \in 1..\kappa+1 : \zeta(t, j) > 0\}. \tag{39}$$

Let $\bar{\zeta}(j) := \min_{t\in\mathcal{T}(j)} \zeta(t, j)$ and define $\zeta_{\min} := \inf_{j\in J(\lambda_{\min})..J_\varepsilon} \bar{\zeta}(j)$. Note that $\zeta_{\min} > 0$. By Part (i) of Lemma 3, for all $j \in J(\lambda_{\min})..J_\varepsilon$ there exists some $N_1(j)$ such that for all $n \geq N_1(j)$ we have

$$\inf_{t\in\mathcal{T}(j)} \gamma(t, j) \geq \delta\bar{\zeta}(j), \tag{40}$$

where δ is the minimum distance between the distinct distributions. As specified by Algorithm 1 we have $\eta := \sum_{j=1}^{\log n\,\kappa+1} \sum_{t=1} w_j\gamma(t, j)$. By (40) for all $n \geq N_1(J_{\lambda_{\min}})$ we have

$$\eta \geq w_{J(\lambda_{\min})}\delta\bar{\zeta}(J_{\lambda_{\min}}), \tag{41}$$

which does not depend on ε. By Part (ii) of Lemma 3, there exists some $N_2(j)$ such that for all $n \geq N_2(j)$ we have

$$\sup_{k\in1..\kappa, t\in1..\mathcal{T}(j)} \frac{1}{n}\left|\hat{\pi}_k^{t,j} - \pi_k\right| \leq \varepsilon. \tag{42}$$

ii. Define $\mathcal{T}'(j) := \{1..\kappa+1\} \setminus \mathcal{T}(j)$ for $j \in J(\lambda_{\min})..J_\varepsilon$, where $\mathcal{T}(j)$ is given by (39). It may be possible for the set $\mathcal{T}'(j)$ to be nonempty on some iterations on $j \in J(\lambda_{\min})..J_\varepsilon$. Observe that by definition, for all $j \in J(\lambda_{\min})..J_\varepsilon$ such that $\mathcal{T}'(j) \neq \varnothing$, we have $\max_{t\in\mathcal{T}'(j)} \zeta(t, j) = 0$. This means that on each of these iterations, there exists some π_k for some $k \in 1..\kappa$ such that $\pi_k = b$ for some grid boundary

$$b \in \left\{ b_i^{t,j} := n\alpha_j(i + \frac{1}{t+1}),\ i = 0..\lfloor\frac{1}{\alpha_j} - \frac{1}{t+1}\rfloor,\ \alpha_j = \lambda_j/3,\ t \in \mathcal{T}'(j) \right\}$$

where the boundaries are specified by Line 6 of Algorithm 1. Since $\lambda_j \leq \lambda_{\min}$ for all $j \in J(\lambda_{\min})..J_\varepsilon$, and that $b = \pi_k$ we have $b..b + n\lambda_j \subseteq \pi_k..\pi_{k+1}$ and $b - n\lambda_j \subseteq \pi_{k-1}..\pi_k$. That is, the segments $X_{b..b+n\lambda_j}$ and $X_{b-n\lambda_j..b}$ are between two consecutive change points and are thus each generated by a single process distribution. Following Lines 6 to 9 of Algorithm 1, it is easy to see that in this case $\gamma(t, j)$ corresponds to $\max\{\Delta_{\boldsymbol{x}}(b, b + n\lambda_j), \Delta_{\boldsymbol{x}}(b - n\lambda_j, b)\}$. Since $b = \pi_k$, by Part (iii) of Lemma 1 there exists some $N_3(j)$ such that for all $n \geq N_3(j)$ we have $\max\{\Delta_{\boldsymbol{x}}(b - n\lambda_j, b), \Delta_{\boldsymbol{x}}(b, b + n\lambda_j)\} \leq \varepsilon$. Thus, for every $j \in J(\lambda_{\min})..J_\varepsilon$ and all $n \geq N_3(j)$ we have

$$\sup_{t\in\mathcal{T}'(j)\neq\varnothing} \gamma(t, j) \leq \varepsilon. \tag{43}$$

This scenario is depicted in Fig. 4.

Step 3. Consider, $j = 1..J(\lambda_{\min}) - 1$. It is desired for a grid to be such that every three consecutive segments contain at most one change point. This property is not satisfied for $j = 1..J(\lambda_{\min}) - 1$ since, by definition, on these iterations we have $\alpha_j > \lambda_j/3$. We show that for all these iterations, the performance score $\gamma(t, j)$, $1..\kappa+1$ becomes arbitrarily small; see Fig. 5. For all $j = 1..J(\lambda_{\min}) - 1$ and $t = 1..\kappa+1$, define the set of intervals $\mathcal{S}^{t,j} := \{(b_i^{t,j}, b_{i+3}^{t,j}) : i = 0..\lfloor\frac{1}{\alpha_j} - \frac{1}{t+1}\rfloor - 3\}$ and consider its partitioning into

$$\mathcal{S}_l^{t,j} := \left\{ \left(b_{l+3i'}^{t,j}, b_{l+3(i'+1)}^{t,j}\right) :\ i' = 0..\frac{1}{3}\left(\left\lfloor\frac{1}{\alpha_j} - \frac{1}{t+1}\right\rfloor - l\right) \right\},\ l = 0..2.$$

Observe that, by construction, for every fixed $l = 0..2$, every pair of indices $(b, b') \in \mathcal{S}_l^{t,j}$ specifies a segment $X_{b..b'}$ of length $3n\alpha_j$ and the elements of $\mathcal{S}_l^{t,j}$ index non-overlapping segments of $\boldsymbol{x}$. Since for all $j = 1..J(\lambda_{\min}) - 1$ we have $\alpha_j > \lambda_j/3$, $j \in 1..J(\lambda_{\min}) - 1$ and $t \in 1..\kappa+1$, there exists some $(b, b') \in \mathcal{S}^{t,j}$ such that $X_{b..b'}$ contains more than one change point. Since there are exactly κ change points, in at least one of the partitions $\mathcal{S}_l^{t,j}$ for some $l \in 0..2$ we have that within any set of κ segments there exists at least one segment that contains no change points. Note that, as specified by Lines 6–11 of Algorithm 1, we have

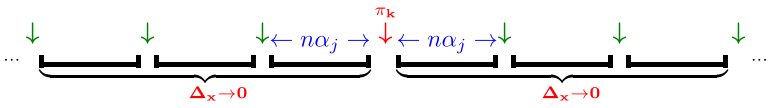

Fig. 4. The case considered in Step 2.ii of the proof: $\lambda_j < \lambda_{\min}$ and since the algorithm sets the spacing α_j between consecutive boundaries to $\lambda_j/3$, every three consecutive segments contain *at most one change point*. In the particular case depicted, one of the grid boundaries lies exactly on some change point π_k. As follows from Step 2.ii, the grid score γ assigned to such iterations converges to zero.

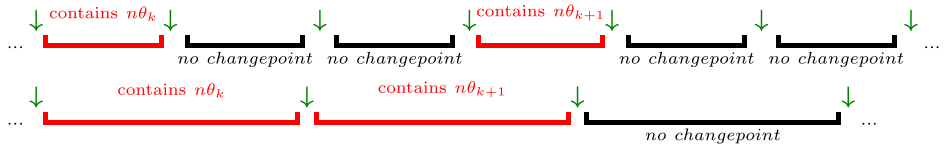

Fig. 5. Top. Desired iteration where every three consecutive grid segments contain *at most one* change point. **Bottom.** Undesired iteration where some groups of three consecutive grid segments may contain more than one change points. As follows from Step 2.i and Step 3, the algorithm indirectly distinguishes between the two scenarios. Specifically, in the former case the grid performance score γ converges to a non-zero constant, while in the latter, it converges to zero.

$$\gamma(t,j) := \min_{l=0..2} \left\{ \Delta_{\boldsymbol{x}}(b,b') : (b,b') \in \mathcal{S}_l^{t,j} \text{ s.t. } \left| \left\{ (a,a') \in \mathcal{S}_l^{t,j} : \Delta_{\boldsymbol{x}}(a,a') > \Delta_{\boldsymbol{x}}(b,b') \right\} \right| = \kappa - 1 \right\}.$$

Therefore, by Part (iii) of Lemma 1, for every $j \in 1..J(\lambda_{\min}) - 1$ there exists some $N(j)$ such that for all $n \geq N(j)$ we have

$$\sup_{t \in 1..\kappa+1} \gamma(t,j) \leq \varepsilon. \tag{44}$$

Combining the steps. Let $N := \max\{\max_{j=1..J(\lambda_{\min})-1} N(j), \max_{\substack{i=1..3 \\ j=J(\lambda_{\min})..J_\varepsilon}} N_i(j)\}$ (note that the ranges of the max operators are finite, so N is well defined). By (41), the definition of J_ε, and that $\gamma(\cdot,\cdot) \leq 1$, for all $n \geq N$ we have

$$\frac{1}{n\eta} \sum_{j=J_\varepsilon}^{\log n} \sum_{t=1}^{\kappa+1} w_j \gamma(t,j) \left| \pi_k - \hat{\pi}_k^{t,j} \right| \leq \frac{\varepsilon(\kappa+1)}{w_{J(\lambda_{\min})} \delta \bar{\zeta}(J(\lambda_{\min}))}. \tag{45}$$

Note that $\eta := \sum_{j=1}^{\log n} \sum_{t=1}^{\kappa+1} w_j \gamma(t,j)$; by (41), (42) for all $n \geq N$ we have

$$\frac{1}{n\eta} \sum_{j=J(\lambda_{\min})}^{J_\varepsilon} \sum_{t \in \mathcal{T}(j)} w_j \gamma(t,j) \left| \pi_k - \hat{\pi}_k^{t,j} \right| \leq \varepsilon. \tag{46}$$

By (41), (43) and (44) for all $n \geq N$ we obtain

$$\frac{1}{n\eta} \sum_{j=J_\varepsilon}^{\log n} \sum_{t \in \mathcal{T}'(j)} w_j \gamma(t,j) |\pi_k - \hat{\pi}_k^{t,j}| \leq \frac{\varepsilon(\kappa+1)}{w_{J(\lambda_{\min})} \delta \bar{\zeta}(J(\lambda_{\min}))}, \tag{47}$$

$$\frac{1}{n\eta} \sum_{j=1}^{J(\lambda_{\min})-1} \sum_{t=1}^{\kappa+1} w_j \gamma(t,j) |\pi_k - \hat{\pi}_k^{t,j}| \leq \frac{\varepsilon(\kappa+1)}{w_{J(\lambda_{\min})} \delta \bar{\zeta}(J(\lambda_{\min}))}. \tag{48}$$

Let $\hat{\theta}_k(n) := \frac{\hat{\pi}_k}{n}$, $k = 1..\kappa$. By (45), (46), (47) and (48) we have

$$\begin{aligned}
|\hat{\theta}_k(n) - \theta_k| \leq{}& \frac{1}{n\eta} \sum_{j=1}^{J(\lambda_{\min})-1} \sum_{t=1}^{\kappa+1} w_j \gamma(t,j) |\pi_k - \hat{\pi}_k^{t,j}| \\
&+ \frac{1}{n\eta} \sum_{j=J(\lambda_{\min})}^{J_\varepsilon} \sum_{t \in \mathcal{T}(j)} w_j \gamma(t,j) |\pi_k - \hat{\pi}_k^{t,j}| \\
&+ \frac{1}{n\eta} \sum_{j=J(\lambda_{\min})}^{J_\varepsilon} \sum_{t \in \mathcal{T}'(j)} w_j \gamma(t,j) |\pi_k - \hat{\pi}_k^{t,j}| \\
&+ \frac{1}{n\eta} \sum_{j=J_\varepsilon}^{\log n} \sum_{t=1}^{\kappa+1} w_j \gamma(t,j) |\pi_k - \hat{\pi}_k^{t,j}|
\end{aligned}$$

$$\leq \varepsilon \left(1 + \frac{3(\kappa+1)}{w_{J(\lambda_{\min})} \delta \bar{\xi}(J(\lambda_{\min}))}\right).$$

Since the choice of ε is arbitrary, the statement of the theorem follows. □

Acknowledgements

This work was supported by the French Ministry of Higher Education and Research, by FP7/2007-2013 under grant agreements 270327 (CompLACS) and 216886 (PASCAL-2), by the Nord-Pas-de-Calais Regional Council and FEDER through CPER 2007–2013, and by an INRIA Ph.D. grant to Azadeh Khaleghi.

References

[1] B. Brodsky, B. Darkhovsky, Non-parametric Methods in Change-Point Problems, Mathematics and Its Applications, Kluwer Academic Publishers, 1993.
[2] M. Basseville, I. Nikiforov, Detection of Abrupt Changes: Theory and Application, Prentice Hall Information and System Sciences Series, Prentice Hall, 1993.
[3] M. Csörgö, L. Horváth, Limit Theorems in Change-Point Analysis, Wiley Series in Probability & Statistics, 1998.
[4] E. Carlstein, S. Lele, Nonparametric change-point estimation for data from an ergodic sequence, Theory Probab. Appl. 38 (4) (1994) 726–733.
[5] L. Giraitis, R. Leipus, D. Surgailis, The change-point problem for dependent observations, J. Statist. Plann. Inference 53 (3) (1996) 297–310.
[6] P. Shields, The Ergodic Theory of Discrete Sample Paths, AMS Bookstore, 1996.
[7] D. Ryabko, Discrimination between B-processes is impossible, J. Theoret. Probab. 23 (2) (2010) 565–575.
[8] D. Ryabko, B. Ryabko, Non-parametric statistical inference for ergodic processes, IEEE Trans. Inform. Theory 56 (3) (2010) 1430–1435.
[9] A. Khaleghi, D. Ryabko, Locating changes in highly-dependent data with unknown number of change points, in: Neural Information Processing Systems, NIPS, Nevada, United States, 2012, pp. 3095–3103.
[10] Azadeh Khaleghi, Daniil Ryabko, Asymptotically consistent estimation of the number of change points in highly dependent time series, in: ICML, JMLR W&CP, Beijing, China, 2014, pp. 539–547.
[11] R. Gray, Probability, Random Processes, and Ergodic Properties, Springer Verlag, 1988.
[12] D. Ryabko, Clustering processes, in: Proceedings of the 27th International Conference on Machine Learning, ICML, Haifa, Israel, 2010, pp. 919–926.
[13] A. Khaleghi, D. Ryabko, J. Mary, P. Preux, Online clustering of processes, in: The International Conference on Artificial Intelligence and Statistics, AI&STATS, in: JMLR W&CP, vol. 22, 2012, pp. 601–609.
[14] D. Ryabko, Testing composite hypotheses about discrete ergodic processes, TEST 21 (2) (2012) 317–329.
[15] I. Csiszar, P. Shields, Notes on information theory and statistics, Foundations and Trends in Communications and Information Theory, 2004.
[16] Daniil Ryabko, Jeremie Mary, A binary-classification-based metric between time-series distributions and its use in statistical and learning problems, J. Mach. Learn. Res. 14 (2013) 2837–2856.
[17] Boris Ryabko, Applications of universal source coding to statistical analysis of time series, in: Selected Topics in Information and Coding Theory, World Scientific Publishing, 2010, pp. 289–338.