Nonparametric Multi Change Point Detection for Markov Chains via Adaptive Clustering

Imon Banerjee
Northwestern University

Jiaqi Lei
Northwestern University

Sanjay Mehrotra
Northwestern University

Abstract

Offline change point detection tries to detect *time points* of distribution change in a given data sequence; and is now routinely used in signal processing, speech processing, climatology etc. Despite this broad applicability across economics, computer science, and planetary sciences, rigorous, nonparametric techniques for change point detection with non-independent and identically distributed (i.i.d.) datasets has remained elusive. This paper establishes such guarantees by proposing a nonparametric clustering algorithm which can accurately obtain the change points from a given Markovian dataset of length n. It does so by bridging together two different components of mathematical statistics; Rademacher complexities of Markov chains, and adaptive clustering via penalisation. Our first result uses recent advances in Rademacher complexities of regenerating Markov chains to derive a Dvoretzky Kiefer Wolfowitz (DKW) type inequality for the empirical distribution of the Markov chain. We then use this to show that an adaptive clustering algorithm recovers the correct change points for a Markovian sequence. We establish the tightness of our rates by showing that they essentially coincide with the best known rates for i.i.d. data. We end the paper by discussing the computational considerations of the problem.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Change point analysis—due first to the seminal work of Page (Page, 1954, 1955)—is a well-established area of study that focuses on identifying points within a data sequence where significant structural changes occur. Yet traditional methods suffer from in two key aspects: (i) the theoretical properties for such methods are most studied when the number of changes are known (and typically a single change), and (ii) the relevant method enjoy the strongest guarantees in the i.i.d. setting, with some oracle based online techniques (Bansal and Papantoni-Kazakos 1986, Lai 1998, Veeravalli 2012) used in the dependent-offline context through restarts; whereas most industrial applications (like process engineering, climate dynamics etc.) have dependent (and particularly, Markovian) data streams. This paper addresses both of these shortcomings; we show that without prior knowledge about the number of change points, minimising the *penalised* inter-cluster variance leads to detecting the correct number of change points for Markovian data-sets.

To set the stage, we introduce some notation. Let $X_1, \ldots, X_n$ be samples from a real valued Markov chain on $[0, 1]$. Let $\tau_1 < \tau_2 < \cdots < \tau_{K_n} \in \{1, \ldots, n\}$ be the true set of change points, with $F_1, \ldots, F_{K_n+1}$ being the corresponding stationary distributions. For any given L and any sequence of estimated change points $\tau'_1 < \tau'_2 < \cdots < \tau'_L$, the empirical counterparts of F_i is given by $\hat{F}^{\tau_i}_{\tau_{i-1}}(u) := \sum_{j=\tau_{i-1}}^{\tau_i} \mathbb{1}[X_j \leq u]/(\tau_i - \tau_{i-1})$. Note that, K_n is allowed to increase with n. The risk function we consider is total clustering variance (defined formally in §3)

$$\int \sum_{i=1}^{L} (\tau_i - \tau_{i-1}) \hat{F}^{\tau_i}_{\tau_{i-1}}(u) \left(1 - \hat{F}^{\tau_i}_{\tau_{i-1}}(u)\right) d\hat{F}^{n}_{0}(u). \quad (1.1)$$

To motivate this risk function, consider its population analogue. As $n \to \infty$, let $\tau_i/n \approx \theta_i$ denote the true (scaled) change points, and let $\hat{\theta}_i$ be their estimates. Suppose each observation X_j is independently assigned to cluster i, that is, drawn from $F_i(u)$ over the interval $(\hat{\theta}_{i-1}, \hat{\theta}_i]$. Let $\mathcal{X}_i = \#\{X_j : j \in (\hat{\theta}_{i-1}, \hat{\theta}_i]\}$. Observe that $\mathcal{X}_i$ counts the number of samples in cluster i and is

distributed as

$$\mathcal{X}_i \sim \text{Binomial}(n(\hat{\theta}_i - \hat{\theta}_{i-1}), F_i(u)),$$

Then,

$$\text{Var}(\mathcal{X}_i) = n(\hat{\theta}_i - \hat{\theta}_{i-1})F_i(u)(1 - F_i(u)).$$

Summing across clusters yields

$$\sum_i \text{Var}(\mathcal{X}_i) = \sum_i n(\hat{\theta}_i - \hat{\theta}_{i-1})F_i(u)(1 - F_i(u)),$$

which is precisely the population counterpart of eq. (1.1).

We can understand why the cluster variance is a good metric by looking at the population counterpart. Assume that there is only a single change point τ_1 such that $\tau_1/n \approx \theta$. Now suppose the candidate for the estimated change point is τ_1' such that $\tau_1'/n \approx \hat{\theta}$, with the population counterpart of the estimate of F_1 being $F_{1,\hat{\theta}}$ and F_2 being $F_{2,\hat{\theta}}$ as below:

$$F_{1,\hat{\theta}} = \frac{\min(\hat{\theta}, \theta)F_1(u) + \max(\hat{\theta} - \theta, 0)F_2(u)}{\min(\theta_0, \theta) + \max(\hat{\theta} - \theta, 0)} \quad \text{and}$$

$$F_{2,\hat{\theta}} = \frac{\max(\theta - \hat{\theta}, 0)F_1(u) + \min(1 - \hat{\theta}, 1 - \theta)F_2(u)}{\max(\theta - \hat{\theta}, 0) + \min(1 - \hat{\theta}, 1 - \theta)}.$$

It can then be verified that for each u, the clustering variance

$$\hat{\theta}F_{1,\hat{\theta}}(u)\left(1 - F_{1,\hat{\theta}}(u)\right) + (1 - \hat{\theta})F_{2,\hat{\theta}}(u)\left(1 - F_{2,\hat{\theta}}(u)\right)$$

decreases as $\hat{\theta} \to \theta$ from both sides.

The pathway towards proving the consistency of the previous mechanism hinges on two main results: (i) A Dvoretzky Kiefer Wolfowitz inequality for regenerating Markov chains (Theorem 1), and (ii) the monotonicity of the risk function with respect to spurious detection (Lemma 4). The proof of the former was by using recent advances in computing the Rademacher complexities of regenerating Markov chains, whereas the second is a consequence of a careful analysis of the risk function (as defined below in eq. (3.2)). We discuss ways to further improve our results, which leads to an open question in the Poissonian concentration of Markov chains (further details in §4.1).

As mentioned above, change point detection has typically been seen through the lenses of hypothesis testing, based on Kolmogorov-Smirnoff (KS), Cramér-von Mises, or Anderson-Darling tests. But in the multi-change point detection, this reveals additional pathologies. A crucial assumption in existing literature attempting to use KS (Padilla et al., 2021) as a baseline for change point detection is access to multiple data streams. In fact, a close inspection of Theorem 3.1 (and in particular equation 3.2) Padilla et al. (2021) reveals a condition on a tuning parameter, which, in absence of multiple data streams, becomes vacuous, by having a lower bound that is increasing and an upper bound that is an universal constant.

Finally, we extensively discuss the computational considerations of our method; providing two distinct mixed-integer binary formulations which all recover the correct number of solutions, as opposed to the standing baseline in the field—PELT(Killick et al., 2012)—which consistently overestimates the correct number of change points.

We now move on to briefly mention our key contributions in this paper.

- **DKW inequality for regerating Markov chains.** We establish a Hoeffding-style tail inequality for the suprema of the empirical distribution (Dvoretzky Kiefer Wolfowitz inequality) in Theorem 1. Corollary 1 establishes that this inequality is essentially sharp; and devolves to the regular DKW inequality for empirical processes for i.i.d. data (up to log terms). Along the way we also discuss the question of a Bennett-style inequality, which to the best of our knowledge was first identified in 2000 (Samson, 2000) but has remained open since, and its implications in this paper (see §4.1 for more details).

- **Consistent multi-change point detection for Markovian data.** Exploiting Theorem 1, we establish the first offline change point detection mechanism for regenerating Markov chains. Our rates are tight, in the sense that when the data is i.i.d., they match the known rates in literature (more details in §4.2).

- **Computational considerations.** We illustrate the effectiveness of the proposed method by comparing it with a common baseline: the Pruned Exact Linear Time (PELT) method (Killick et al., 2012). Using simulated nonstationary Markov chain data, we evaluate runtime and detection accuracy of different methods.(see §5.1).

The rest of the paper is organised as follows: §2 outlines a comprehensive discussion of relevant research work. We formally introduce the model and relevant notations in §3. §4 hold our key theorems and the sketches of their proofs, while the full proofs have been deferred to appendix due to lack of space. The computational considerations are housed in §5, and finally, in §6, we discuss the broader impacts and the future outlooks of our work.

2 BACKGROUND AND RELATED RESEARCH

Broadly speaking, change point detection methods can be categorised as *online* (detecting changes in realtime) or *offline* (segmenting retrospectively once all data are observed), the latter of which is the focus of this paper. Online methods prioritize early detection, while offline methods typically aim to identify multiple changes simultaneously.

The earliest contribution on change point detection in the Gaussian setting are due to Page (1954, 1955). Since then, change point methodology has been applied in speech processing (Desobry et al., 2005; Harchaoui et al., 2009), finance (Bai and Perron, 1998; Frick et al., 2014), bioinformatics (Hocking et al., 2013; Maidstone et al., 2017), climatology (Maidstone, 2016; Verbesselt et al., 2010), and network traffic analysis (Lévy-Leduc and Roueff, 2009; Lung-Yut-Fong et al., 2012), among others. We point the reader to standard references like the monographs Basseville and Nikiforov (1993); Brodsky and Darkhovsky (1993); Csörgö and Horváth (1997); Chen and Gupta (2011) and the surveys Truong et al. (2020); Lavielle and Teyssière (2007); Jandhyala et al. (2013); Haynes et al. (2017) for more details on the theory and methods of change point analysis.

Reliable detection of multiple changes is a comparatively recent development. A key contribution is Zou et al. (2014) (see also Fryzlewicz (2014)), who proposed a nonparametric maximum likelihood method with BIC-based model selection, establishing consistency and optimal rates. More recently, Padilla et al. (2021); Madrid Padilla et al. (2022) analyzed univariate and multivariate data using the Kolmogorov-Smirnov (KS) distance, showing that wild binary segmentation achieves nearly minimax rate-optimality in multi-stream settings.

However, possibly due to the absence of suitable mathematical tools (Bertail and Portier, 2019), the problem of offline nonparametric multiple change point detection for Markov chains remains open, with limited progress under parametric settings such as time-series models (Fryzlewicz, 2017). We get the following open question.

Open Question. Is it possible to design a *nonparametric change point detection method for Markov chains* on general state spaces with mild assumptions, and achieving optimal convergence?

Remark 1. *The choice of a nonparametric approach in this paper is motivated primarily due to the fact that apart from a few developments in time-series and queueing contexts, parametric models for Markov chains remain largely undeveloped.*

3 Problem Statement

In order to formalize our results, we introduce some notations. A σ-algebra $\mathcal{E} := \sigma(E)$ of a set E is said to be countably generated if there exists a countable set $\mathcal{C}$ such that $\mathcal{E} \subseteq \sigma(\mathcal{C})$. Then, we call $(E, \mathcal{E})$ to be a countably generated state space. Our choice of $(E, \mathcal{E}) = ([0,1], \mathcal{B}_{[0,1]})$, with $\mathcal{B}_{[0,1]}$ denoting the Borel σ-algebra. In particular, $([0,1], \mathcal{B}_{[0,1]})$ is countably generated. We define the Orlicz norm of a random variable X as $\|X\|_{\psi_1} = \text{argmin}\{\lambda > 0 : \mathbb{E}\left[e^{(X/\lambda)}\right] \leq 1\}$.

Data generating process. Let us assume that we have data that is a sample from Markov chains, defined on a continuous state space $[0,1]$, that change at K_n (unknown) time points. The transition densities and invariant distributions of these Markov chains are denoted as $\jmath_1, \jmath_2, \ldots, \jmath_{K_n+1}$; $F_1, F_2, \ldots, F_{K_n+1}$, respectively. We have $X_1^1, \ldots, X_{n_1}^1$ samples from the first transition density, $X_1^2, \ldots, X_{n_2}^2$ samples from the second transition density, and so on. In the offline change point detection problem, we do not have access to n_1, n_2, etc. and observe our data as a single sample $X_1, \ldots, X_n$ with $n = n_1 + n_2 + \ldots$, and

$$\begin{aligned} X_1 = X_1^1, X_2 = X_2^1, \ldots, X_{n_1} = X_{n_1}^1, \\ X_{n_1+1} = X_1^2, \ldots, X_{n_1+n_2} = X_{n_2}^2, \ldots. \end{aligned} \tag{3.1}$$

With $\tau_1 = n_1, \tau_2 = n_1 + n_2, \ldots$ as the reparametrizations of the true (but unknown) change points, our problem then is to estimate the τ_i's from a single data-stream $X_1, \ldots, X_n$. We recall from the introduction that $\tau_1 < \tau_2 < \cdots < \tau_{K_n}$ is the ordered set of true change points, and remind the readers that K_n is allowed to increase with n. The empirical counterparts of F_i is given by $\hat{F}_{\tau_{i-1}}^{\tau_i}(u) := \sum_{j=\tau_{i-1}}^{\tau_i} \mathbb{1}[X_j \leq u]/(\tau_i - \tau_{i-1})$ and let $\hat{F}_n(u) := \sum_{i=1}^{n} \mathbb{1}[X_i \leq u]/n$ the empirical distribution of the whole sample.

Let L be an integer estimating the number of change points with the candidate change points being $\tau_1', \ldots, \tau_L'$. For any $u \in [0,1]$ and $\tau_i' < \tau_j'$, we define $\hat{F}_{\tau_i'}^{\tau_j'} := \sum_{p=\tau_i'}^{\tau_j'} \mathbb{1}[X_p \leq u]/(\tau_j' - \tau_i')$ as the empirical distribution between τ_i' and τ_j'. By $X_{(1)}, \ldots, X_{(n)}$, we denote the order statistics of $X_1, \ldots, X_n$.

Our choice of loss metric is the nonparametric clustering variance defined as:

$$\begin{aligned} &R_n(\tau_1', \ldots, \tau_L') \\ &= \sum_{i=0}^{L-1} (\tau_{i+1}' - \tau_i') \int_{X_{(1)}}^{X_{(n)}} \hat{F}_{\tau_i'}^{\tau_{i+1}'}(u)(1 - \hat{F}_{\tau_i'}^{\tau_{i+1}'}(u)) d\hat{F}_n(u) \end{aligned} \tag{3.2}$$

When $\tau'_1, \ldots, \tau'_L$ is not ordered, then

$$R_n(\tau'_1, \ldots, \tau'_L) := R_n\left(\tau'_{(1)}, \ldots, \tau'_{(L)}\right) \tag{3.3}$$

where $\tau'_{(1)} < \tau'_{(2)} < \cdots < \tau'_{(L)}$ is the order statistics for $\tau'_1, \ldots, \tau'_L$.

Our objective is the risk penalised by the Bayesian information criterion

$$\mathrm{BIC_L} := \min_{\tau'_1 < \cdots < \tau'_L} R_n(\tau'_1, \ldots, \tau'_L) + L\zeta_n, \qquad \text{(BIC)}$$

where ζ_n is a suitably increasing sequence given explicitly in Theorem 2. Our choice of estimators will be

$$\hat{K}_N := \mathrm{argmin}_L \, \mathrm{BIC_L} \qquad \text{and}$$
$$\mathcal{G}_n(L) := (\hat{\tau}_1, \ldots \hat{\tau}_L) = \mathrm{argmin}_{\tau'_1 < \cdots < \tau'_L} R_n(\tau'_1, \ldots, \tau'_L),$$

with $\mathcal{G}_n(\hat{K}_n)$ being the final estimated change points.

Remark 2. *Observe that solving eq. (BIC) by enumeration is NP-hard. In §5, we provide a mixed integer program to minimise this objective.*

Theorem 2, proves that optimising the previous objective function achieves the optimal rate of detecting change points.

Before proceeding we briefly justify our choice of integrating measure $d\hat{F}_n(u)$ in eq. (3.2). Prior work (Zhang, 2002, 2006) highlights the empirical benefit of using weighted measures. In nonparametric change-point analysis, $d\hat{F}_n(u)$ is effective when adjacent distributions differ near their medians. However, as noted by Zou et al. (2014) in the i.i.d. setting, it can underperform when differences lie in the tails, since little information is captured in the integral. In such cases, using

$$\frac{d\hat{F}_n(u)}{\hat{F}_n(u)(1 - \hat{F}_n(u))}$$

offers greater detection power. This improvement, however, relies on a Poissonian concentration inequality for i.i.d. processes (Wellner, 1978), with no clear analogue in the Markovian case. Deriving a Markovian counterpart to Lemma 1 of Wellner (1978) appears infeasible without new tools, so we instead adopt the unweighted integrating measure which is amenable to the DKW inequality. We emphasize that such weighting only improves empirical power: Theorem 2 already guarantees the asymptotically optimal rate. We now turn to the formal statements of our main results.

4 THEORETICAL RESULTS

As stated in the introduction, our first objective will be to provide a DKW inequality for the empirical suprema of a regenerating Markov chains. To that end, we provide a brief introduction to regenerating Markov chains.

4.1 DKW Inequality for Regenerating Markov chains

We give the following definition of atomic ψ-irreducible Markov chain (we refer the readers to (Meyn and Tweedie, 2012, page 89) for the formal definition of ψ-irreducibility).

Definition 1 (Regenerative/Atomic chain)**.** *A ψ-irreducible, aperiodic Markov chain X with a transition probability distribution $P(\cdot, \cdot)$ is* regenerative *(or* atomic*) if there exists a measurable set A (the* atom*) with $\Psi(A) > 0$ (for some measure Ψ) such that*

$$P(x, \cdot) = P(y, \cdot), \qquad \forall\, x, y \in A.$$

Remark 3. *The set A is called the Ψ-atom. In chains with finitely many states any single state may serve as an atom.*

Intuitively, ψ-irredicibility extends the classical notion of irreducibility for finite state Markov chains to infinite state spaces, whereas Ψ-atoms are sets from which the transitions behave homogeneously. That is, the probability of transition to any set is equal for any two starting points inside an atom. For finite state Markov chains, the atoms are the individual states (i.e. singletons). Informally, a regenerating Markov chain is a ψ-irreducible Markov chain which has at least one Ψ-atom that is repeatedly visited (with the inter-arrival times termed as the regeneration time). Conditions on the moment generating function of the regeneration time characterises the ergodic properties of the Markov chain as can be seen in Assumption 1 below.

We now formalise the previous intuition. Extend the sample space by introducing a sequence $(Y_m)_{m \in \mathbb{N}}$ of independent Bernoulli random variables with success probability δ. Our construction relies on a mixture representation of the transition kernel on a set S:

$$P(x, S) = \delta\, \Psi(S) + (1 - \delta) \frac{P(x, S) - \delta\, \Psi(S)}{1 - \delta},$$

where the first term is independent of the starting point. The validity of this construction is ensured by the existence of the atom and we refer to (Meyn and Tweedie, 2012, Chapter 4) for more details. This independence is key since it guarantees regeneration when that component is selected. In other words, each time the chain visits S, we randomly reassign the transition probability P as follows:

- If $X_m \in S$ and $Y_m = 1$ (which occurs with probability $\delta \in (0, 1)$), then the next state X_{m+1} is generated according to the measure Ψ.

- If $X_m \in S$ and $Y_m = 0$ (with probability $1-\delta$), then X_{m+1} is drawn from the probability measure

$$(1-\delta)^{-1}\Big(P(X_m,\cdot) - \delta\,\Psi(\cdot)\Big).$$

The resulting bivariate process

$$Z_m = (X_m, Y_m),$$

known as the *split chain*, takes values in $E \times \{0,1\}$ and is itself atomic, with the atom defined as

$$A = S \times \{1\}.$$

We then define the *regeneration times* recursively by setting

$$\rho_A(1) = \inf\{m \geq 1 : Z_m \in A\},$$

and for $j \geq 2$,

$$\rho_A(j) = \inf\{m > \rho_A(j-1) : Z_m \in A\}.$$

It is well known that the split chain Z inherits aperiodicity and ψ-irreducibility from the original chain X. Furthermore, by the recurrence property, the regeneration times have finite expectation; that is, one has (Azaïs et al., 2016, Lemma A1)

$$\sup_{z\in A} \mathbb{E}_z[\rho_A(j)] < \infty \quad \text{and} \quad \mathbb{E}_\nu[\rho_A(j)] < \infty$$

for any initial measure ν on A and any integer j.

Regeneration theory (Meyn and Tweedie, 2012) shows that, given the sequence $(\rho_A(j))_{j\geq 1}$, the sample path can be divided into blocks (or cycles) defined by

$$B_j = (X_{\rho_A(j)+1}, \ldots, X_{\rho_A(j+1)}), \quad j \geq 2,$$

corresponding to successive visits to the regeneration set A. The strong Markov property then ensures that both the regeneration times and the blocks $\{B_j\}_{j\geq 1}$ form independent and identically distributed (i.i.d.) sequences (Meyn and Tweedie, 2012, Chapter 13). Moreover, with $\mathbb{E}_A$ to be the expectation when the Markov chain is initialised from the atom A, the lengths of these blocks are i.i.d. with mean $\mathbb{E}_A[\rho_A(2)]$. Furthermore, B_1 is independent of $B_2, B_3, \ldots$.

Assumption 1. *There exists a constant $\lambda > 0$ for which*

$$\mathbb{E}_A[\exp(\lambda \rho_A(2))] < \infty. \tag{4.1}$$

Moreover, for this choice of λ, we assume that

$$\kappa_\lambda := \frac{2\,\mathbb{E}_A[\exp(\lambda\rho_A(2))]}{\lambda} > \frac{1}{2}. \tag{4.2}$$

Assumption 1 ensures that the regeneration time has an exponential moment, a property equivalent to geometric ergodicity, the uniform Doeblin condition, or the Foster-Lyapunov drift criterion (see Theorem 16.0.2 in Meyn–Tweedie). Under this assumption, classical convergence results such as the LIL and CLT remain valid (Chapter 17). Next, we define

$$\rho_o := \min\{\|\rho_A(1)\|_{\psi_1}, \|\rho_A(2)\|_{\psi_1}\}. \tag{4.3}$$

Observe that under Assumption 1, the sets

$$\{\lambda > 0 : \mathbb{E}\left[e^{(\rho_A(1)/\lambda)}\right] \leq 1\} \text{ and } \{\lambda > 0 : \mathbb{E}\left[e^{(\rho_A(2)/\lambda)}\right] \leq 1\}$$

are non-empty. It follows that $\rho_o < \infty$.

We now proceed to formally state the main result of this section. Let $\mathcal{S} \subset \mathbb{R}$ and define $\mathcal{F}_\mathcal{S} := \{\mathbb{1}[\cdot < t], t \in \mathcal{S}\}$ be the set of all half-interval functions on $\mathcal{S}$. The following theorem is proved in §A.7.

Theorem 1. *Let $Y_1, \ldots, Y_n$ be a sequence of random variables from a Markov chain satisfying Assumption 1 with stationary distribution π. Let Y be a random variable with distribution π. Define*

$$Z := \sup_{f\in\mathcal{F}_{[0,1]}} \left| \frac{1}{n}\sum_{i=1}^{n} f(Y_i) - \mathbb{E}_\pi\left[f(Y)\right] \right|.$$

Then, for all $t > 0$

$$\mathbb{P}(Z > t) \leq \kappa(\rho,\lambda)\exp\left(-\frac{\kappa(\rho,\lambda)nt^2}{\log n}\right) \tag{4.4}$$

where, constant $\kappa(\rho,\lambda)$ depend only on $\mathbb{E}_A[\rho_A^2(2)]$, $\mathbb{E}_v[\rho_A(2)]$, ρ_o, and λ (as defined in Eq. (4.1), (4.2), and (4.3)).

Remark 4. *The extra $\log n$ term appears commonly when transitioning from i.i.d random variables to Markov chains (Samson, 2000). Observing that the regeneration time ρ of any i.i.d. process is (deterministically) 1, the following corollary shows that our previous result is tight up to $\log$ terms.*

Corollary 1. *Let $Y_1, \ldots, Y_n$ be a sequence of i.i.d. random variables from distribution π. Let Y be a random variable with distribution π. Then, with Z defined as in Theorem 1, and an universal constant $\mathbb{C}$;*

$$\mathbb{P}(Z > t) \leq \mathbb{C}\exp\left(-\frac{\mathbb{C}nt^2}{\log n}\right) \tag{4.5}$$

On Poissonian Tail Concentration: Observe that Theorem 1 provides a Höeffding (sub-Gaussian) concentration inequality (ignoring constants)

$$\mathbb{P}(Z_n > t) \lesssim e^{-nt^2/\log n}$$

for the suprema of the empirical distribution, which is sharp when t is large. However, for small values of t, a sharper bound can be obtained from so called Bennett (Poissonian) concentration inequalities

$$\mathbb{P}(Z_n > t) \lesssim e^{-nt\log(1+t)}.$$

Owing to the fact that such inequalities for the empirical suprema of i.i.d. processes were known due to Michael Talagrand, (Talagrand, 1996), our initial objective in this paper was to derive a Bennett inequality for regenerating Markov chains and then use it to derive the consistency of the *weighted* risk function described in §3. However, upon further investigation, multiple papers (Samson, 2000; Adamczak, 2008) point out that such a result requires novel approaches towards empirical process theory, which makes it beyond the scope of the current work. The particular pathology is the unavailability of the Hoffmann-Jørgensen inequality (more specifically, a counterpart to Theorem 6.21 in Ledoux and Talagrand) for the Bennett-Orlicz norm (Wellner, 2017). However, we point out that availability of any such result will be immediately applicable; and proving the consistency of the weighted risk function would be a straightforward extension of the proof in §A.1.

Remark 5. *The previous challenge is not purely theoretical, since it clearly leads to a pathway of improving the risk function for change point detection of Markov chains.*

4.2 Change Point Detection

Now we proceed to state the main result of this section. To that end, we make some assumptions, beginning with the regeneration times of the underlying Markov chains. Recall from eq. (3.1) that X_j^i was used to denote the j-th sample of the i-th Markov chain, with $K_n + 1$ many Markov chains being available in total.

Assumption 2. *All of the Markov chains X_j^i satisfy Assumption 1 with constants κ_i for $i \in \{1, \ldots, K_n + 1\}$*

Remark 6. *Observe it is not sufficient to simply impose Assumption 1 the Markov chain $X_1, \ldots, X_n$, since one of X_j^i's may be transient while the whole chain stays regenerating.*

We next make the following standard assumption (see Zou et al. (2014, A1-A3)) of sufficient gap between any two consecutive changes:

Assumption 3. *Let $\beta_n := \min_{1\le k\le K_n+1}(\tau_k - \tau_{k-1})$. We assume that*

$$\beta_n \xrightarrow{n\to\infty} \infty \quad \text{and} \quad \hat{F}_n(u) \xrightarrow{n\to\infty} F(u)$$

almost everywhere uniformly in u, for some distribution F in the convex hull generated by $\{F_1, \ldots, F_{K_n+1}\}$.

Finally, for all $r \in \{1, \ldots, K_n + 1\}$ define $\eta_r(u) := (F_{r-1}(u) - F_r(u))^2$. Our next assumption is on the minimum amount of change between any two F_i's. It can be thought of as the minimum signal strength indicating change.

Assumption 4. *F_k are continuous and distinct for all k and define*

$$\eta_{\min} := \min_{1\le r\le K_n+1} \int_0^1 \eta_r(u)dF(u).$$

Note that, $\eta_{\min} > 0$.

Remark 7. *We assume that we know the true value of $\eta_{\min}$. In practice, we can replace it by a large enough constant. By a careful inspection of the proof of Theorem 2 (and in particular eq. (A.14)) one can see that any constant larger than $\eta_{\min}$ is actually sufficient for detecting the correct number of change points.*

To formalise our main result we introduce some notation. Let $L > 0$ be an integer and define $\tau'_1, \ldots, \tau'_{K_n}$ be a sequence of time points. We define the set $\mathcal{C}_{K_n}(\delta_n)$ as the set of all $(\tau'_1, \ldots, \tau'_{K_n})$ which are at most δ_n away from the true change point $\tau_1, \ldots, \tau_{K_n}$. Formally,

$$\mathcal{C}_{K_n}(\delta_n) := \{(\tau'_1, \ldots, \tau'_{K_n}) : 1 < \tau'_1 < \cdots < \tau'_{K_n} \le n, \\ |\tau'_s - \tau_s| \le \delta_n \; \forall\, 1 \le s \le K_n\}$$

for some positive sequence δ_n. Next, for a fixed L, define $\mathcal{G}_n(L)$ as the set of time points which minimizes $\mathrm{BIC}_{\mathrm{L}}$. Formally,

$$\begin{aligned}\mathcal{G}_n(L) =& (\hat{\tau}_1, \ldots, \hat{\tau}_L) := \operatorname{argmin}_{\tau'_1<\cdots<\tau'_L} \mathrm{BIC}_{\mathrm{L}} \\ =& -\operatorname{argmax}_{\tau'_1<\cdots<\tau'_L} R_n(\tau'_1, \ldots, \tau'_L) + L\zeta_n.\end{aligned}$$

Then, we have the following theorem, proved in §A.1.

Theorem 2. *Under Assumptions 1, 3, and 4, and if $K_n^3(\log K_n)^2(\log \delta_n)^2/\delta_n = \mathcal{O}(1)$ and $\delta_n/\beta_n \to 0$, then with any sequence $\zeta_N \to \infty$, we have*

$$\mathbb{P}\left(\mathcal{G}_n(K_n) \in \mathcal{C}_{K_n}(\delta_n)\right) \xrightarrow{n\to\infty} 1.$$

Furthermore, let $K_n \le \bar{K}$ for some known $\bar{K}$. Then, with $\zeta_n = \kappa_n \bar{K}^3(\log \bar{K})^2(\log \beta_n)^2/\beta_n$,

$$\hat{K}_n := \arg\min_L \mathcal{G}_n(L)$$

satisfies $\mathbb{P}\left(\hat{K}_n = K_n\right) \to 1$, where κ_n is a constant that depends only on $\kappa_i, i \in \{1, \ldots, K_n + 1\}$.

Theorem 2 provides the asymptotic consistency of $\hat{K}_n$, which is the estimated number of change points. It implies that the localization accuracy δ_n may be taken as any sequence satisfying

$$K_n^3(\log K_n)^2(\log \delta_n)^2/\delta_n = O(1), \qquad \delta_n/\beta_n \to 0.$$

In particular, if $K_n = O(1)$, then $\delta_n = O(1)$ is admissible, yielding constant-error localizationthe best achievable rate in nonparametric change-point detection. When K_n grows, the attainable δ_n remains of order β_n up to polylogarithmic factors, matching known optimal rates in the i.i.d. literature (Zou et al., 2014).

Now consider the independent case. Here the regeneration time is 1, with the atom equal to the entire state space. Hence any i.i.d. sequence satisfies Assumption 2 with deterministic regeneration time 1 ($\rho_A(j) = 1$ for all j). In particular, κ_n in the Theorem 2 becomes an universal constant, and we recover the following corollary.

Corollary 2. *Let X_i be independent. Then, under Assumptions 1, 3, and 4 if $K_n^3(\log K_n)^2(\log \delta_n)^2/\delta_n = \mathcal{O}(1)$ and $\delta_n/\beta_n \to 0$, then with any sequence $\zeta_N \to \infty$, we have*

$$\mathbb{P}\left(\mathcal{G}_n(K_n) \in \mathcal{C}_{K_n}(\delta_n)\right) \xrightarrow{n\to\infty} 1.$$

Furthermore, let $K_n \le \bar{K}$ for some known $\bar{K}$. Then, with $\zeta_n = \bar{K}^3(\log \bar{K})^2(\log \beta_n)^2/\beta_n$,

$$\hat{K}_n := \operatorname{argmin}_L \mathcal{G}_n(L)$$

satisfies $\mathbb{P}\left(\hat{K}_n = K_n\right) \to 1$.

Before presenting a sketch of the proof of the theorem, we discuss some aspects of the previous result.

Tightness: A direct comparison of Corollary 2 with Zou et al. (2014, Theorem 1) reveals that the rate of β_n do not suffer when we move from i.i.d. to Markovian data, but rather adjusts for the Markovianity in the obvious way, through a constant which depends upon the regeneration time, and is an universal constant for i.i.d. data.

Choice of penalty: We further remark that β_n is typically unavailable to the practitioner, and setting the penalty $\zeta_n = (\log n)^{2+c}$ for a small constant $c \ge 0$ seems to work well in practice (Zou et al., 2014). We use $c = 0$ in §5. We now provide a sketch of proof of Theorem 2.

Sketch of Proof of Theorem 2 To guide the readers through the broad steps of the proof, we provide a sketch of the proof. For the full details, see §A.1.

1. Use Theorem 1 to develop a deviation bound for the clustering variance risk (Lemma 3).
2. Use Lemma 3 to bound the rate of growth of the risk R_n of detecting superfluous change points (Lemmas 4 and 5).
3. Use Lemma 3 to bound the rate of growth of the clustering variance of detecting superfluous change points (Lemma 6)
4. Use Lemmas 4, 5, and 6 to bound the probability of detecting less than the correct number of change points (Lemma 7).
5. Using contradiction, bound the probability of detecting more than the correct number of change points (Lemma 8). Theorem 2 follows by combining Lemmas 7 and 8.

5 COMPUTATION

We recast the risk in (3.2) as a nonlinear binary optimization problem. Let $z_{i,l} \in \{0,1\}$ indicate whether time point i is assigned to segment $l \in \{1,\dots,L+1\}$. The segmentation constraints enforce single assignment, contiguity, and a minimum segment length. The resulting formulation both encodes (3.2) exactly (for fixed L) and readily accommodates additional structural constraints that are awkward for classical dynamic programming. We therefore write the following proposition, whose proof can be found in §A.10.

Proposition 1. *Let $a_{i,u} := \mathbb{1}[X_i \le X_{(u)}]$ and*

$$\mathcal{Z}_{i,l} := \Big\{ z_{i,l} \in \{0,1\} \;\Big|\; \sum_{l=1}^{L+1} z_{i,l} = 1,\ \sum_{i=1}^{n} z_{i,l} \ge 3,\ z_{i,l} \le \sum_{l'\ge l} z_{i+1,l'} \Big\}.$$

Consider

$$\begin{aligned} \min \quad & \sum_{l=1}^{L+1}\sum_{u=1}^{n}\left(\sum_{i=1}^{n} a_{i,u} z_{i,l}\right)\left(1 - \frac{\sum_{i=1}^{n} a_{i,u} z_{i,l}}{\sum_{i=1}^{n} z_{i,l}}\right) \\ \text{s.t.} \quad & z_{i,l} \in \mathcal{Z}_{i,l} \qquad \forall i, l. \end{aligned} \tag{5.1}$$

Then $z^\star$ is a solution of (5.1) if and only if the induced change points $\tau'_l - \tau'_{l-1} = \sum_{i=1}^{n} z^\star_{i,l}$ solve (3.2); moreover, the objective values of (5.1) and (3.2) coincide.

To reduce runtime, we use the following bilinear reformulation, formalized in Proposition 2, which preserves optimality (see §A.11 for the proof) while avoiding division inside the objective, and can be solved with off-the-shelf solvers like Gurobi. A detailed empirical comparison of runtimes is provided in §5.1.

Proposition 2. *Consider the bilinear reformulation:*

$$\begin{aligned}
\min \quad & \sum_{l=1}^{L+1} \sum_{u=1}^{n} s_{u,l} \\
s.t. \quad & \sum_{i=1}^{n} k_l z_{i,l} = 1 \quad \forall l, \\
& f_{u,l} = \sum_{i=1}^{n} a_{i,u} k_l z_{i,l}, \text{ and } d_{u,l} + f_{u,l} = 1 \; \forall u, l, \\
& s_{u,l} = \sum_{i=1}^{n} a_{i,u} d_{u,l} z_{i,l} \; \forall u, \forall l; \quad z_{i,l} \in \mathcal{Z}_{i,l} \quad \forall i, \forall l, \\
& k_l, \; f_{u,l}, \; d_{u,l}, \; s_{u,l} \geq 0 \quad \forall u, \forall l,
\end{aligned} \tag{5.2}$$

then z^* is a solution of (5.1) if and only if $(z^*, k^*, f^*, d^*, s^*)$ is a solution of (5.2) and the objectives are the same.

5.1 Simulation

We compare (5.1), its bilinear reformulation (5.2), and PELT on a nonstationary Markov chain. We generate $n = 250$ time points partitioned into 4 segments of lengths $0.1n, 0.2n, 0.3n, 0.4n$ (with the corresponding time points τ_i being 25,75,150 respectively). On segment l, arrivals and departures follow the arrival rate λ_l and departure rate μ_l are drawn from uniform distributions:

$$\lambda_l \sim \text{Uniform}(0, \alpha_\lambda), \quad \mu_l \sim \text{Uniform}(0, \alpha_\mu),$$

where $\alpha_\lambda, \alpha_\mu > 0$ are scaling factors controlling variability. At each time point i, let $l(i)$ denote the segment containing point i. The system evolves according to

$$A_i \sim \text{Poisson}(\lambda_{l(i)}), \quad D_i \sim \text{Binomial}(N_{i-1}, \mu_{l(i)}),$$

The population is updated recursively by adding arrivals and subtracting departures. Be specific, $N_i = \max\{0, N_{i-1} + A_i - D_i\}$. All PELT results are obtained using the changepoint.np package, which implements the nonparametric PELT method of Zou et al. (2014). Experiments were run using the default MBIC penalty, which provided the strongest empirical performance. Note that PELT guarantees pruning efficiency (Killick et al. (2012), Thm. 3.1) but not correctness; in contrast, our MIP formulations exactly minimize the clustering-variance risk.

Figure 1 shows that our method recovers the true change points exactly, $\hat{\tau} = [25, 75, 150]$, while PELT over-segments ($\hat{\tau} = [25, 37, 46, 72, 151, 161, 176, 204]$). Across n, (5.2) matches the accuracy of (5.1) at substantially lower cost, whereas PELT is fastest but consistently overestimates L (Table 1).

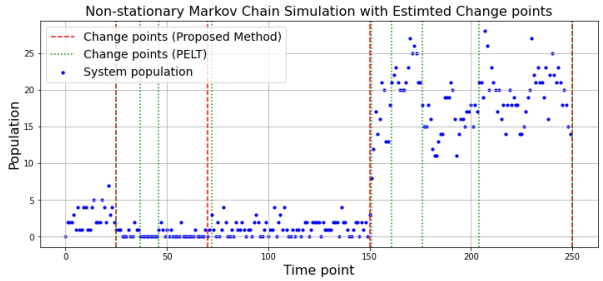

Figure 1: Illustration of the simulated nonstationary Markov chain data. Blue dots represent the population size at each time point. Red lines (our proposed method) mark $\hat{\tau} = [25, 75, 150]$, which match the true change points. Green lines (the PELT method) mark $\hat{\tau} = [25, 37, 46, 72, 151, 161, 176, 204]$, which overestimate the total number of change points.

We now compare runtime and accuracy. Across all n, both (5.1) and (5.2) exactly recover the true change points, with the bilinear reformulation achieving a substantial speedup over (5.1). PELT is the fastest but consistently over-segments. Thus, (5.2) offers the best trade-off, retaining accuracy while reducing runtime, whereas PELT sacrifices reliability for speed (Table 1).

Table 1: Runtime and change points of the optimization frameworks and the PELT method with different numbers of time points n. (5.1)-Time (seconds) is the runtime for (5.1), which is slowest, while (5.2)-Time (seconds) is the runtime for (5.2), which is much faster; both recover the true change points. PELT-Time (seconds) is the runtime and PELT-L is the number of change points estimated by the PELT method. The PELT method is the fastest but consistently estimates more than the correct number of change points.

n	(5.1)-Time	(5.2)-Time	PELT-Time	PELT-L
50	5.49	0.69	0.03	7
100	9.98	1.42	0.07	8
250	30.42	9.43	0.35	9
400	180.38	28.71	1.14	22
500	302.76	92.59	1.53	26

6 CONCLUSION

Clustering for Markov chains has some recent developments (Lee et al., 2025), and we present a nonparametric framework for multiple change-point detection in Markov chains that couples a DKW-type inequality for regenerating chains with an adaptive clustering criterion. The theory yields consistent estimation of both the number and locations of changes with rates that match the i.i.d. benchmark up to logarithmic factors, indicating that dependence need not fundamentally hinder detectability.

On the computational side, we provide exact optimization formulations that achieve high accuracy; the bilin-

ear model (5.2) offers a practical speedaccuracy tradeoff and outperforms a PELT baseline that tends to oversegment in our experiments.

Two directions appear especially promising. First, Bennetttype (Poissonian) concentration for Markov chains would directly enable weighted risks with greater power in tail-difference regimes (see §4.1). Second, enriching the optimization with robustness and domain constraints (e.g., minimum dwell times, forbidden transitions, or outlier resistance) would broaden applicability without sacrificing exactness.

Limitations and future outlook. Online methods such as CUSUM require knowledge of the pre- and post-change distributions and use their KL-divergence to set thresholds. Offline method addresses the complementary regime where these distributions are unknown, trading immediate detection for identifiability. As we note earlier, online methods with unknown pre- and post-change distributions being an evolving field.

We also remark that the multivariate extension remains open, since meaningful generalizations likely require kernel-based empirical processes or alternative notions of dependence, as the univariate DKW framework does not transfer directly. Developing such tools is an interesting direction, and we plan to investigate it in a future work.

Acknowledgements

The first author acknowledges the IEMS Alumni Fellowship at Northwestern University for financial support during the conduct of this research. The second and last authors acknowledge support from NIAID grant R01AI168144. The authors also thank the four anonymous reviewers for their useful comments and suggestions, which significantly improved the readability of the paper.

Bibliography

Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13(none):1000–1034. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

Azaïs, R., Delyon, B., and Portier, F. (2016). Integral estimation based on markovian design. arXiv:1609.01165.

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

Basseville, M. and Nikiforov, I. (1993). *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs.

Bertail, P. and Portier, F. (2019). Rademacher complexity for Markov chains: Applications to kernel smoothing and MetropolisHastings. *Bernoulli*, 25(4B):3912–3938.

Borrero, J. S., Gillen, C., and Prokopyev, O. A. (2016). A simple technique to improve linearized reformulations of fractional (hyperbolic) 0–1 programming problems. *Operations Research Letters*, 44(4):479–486.

Brodsky, B. E. and Darkhovsky, B. S. (1993). *Nonparametric methods in change point problems*. Springer Netherlands.

Chen, J. and Gupta, A. K. (2011). *Parametric statistical change point analysis: With applications to genetics, medicine, and finance*. Springer Science & Business Media.

Csörgö, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*. Chichester, New York.

Desobry, F., Davy, M., and Doncarli, C. (2005). An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 53(8):2961–2974.

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 76(3):495–580.

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

Fryzlewicz, P. (2017). breakfast: multiple change-point detection and segmentation.

Harchaoui, Z., Vallet, F., Lung-Yut-Fong, A., and Cappé, O. (2009). A regularized kernel-based approach to unsupervised audio segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1665–1668, Taipei, Taiwan.

Haynes, K., Eckley, I. A., and Fearnhead, P. (2017). Computationally efficient changepoint detection for a range of penalties. *Journal of Computational and Graphical Statistics*, 26(1):134–143.

Hocking, T., Schleiermacher, G., Janoueix-Lerosey, I., Boeva, V., Cappo, J., Delattre, O., Bach, F., and Vert, J.-P. (2013). Learning smoothing models of copy number profiles using breakpoint annotations. *BMC Bioinformatics*, 14(1):164.

Jandhyala, V., Fotopoulos, S., Macneill, I., and Liu, P. (2013). Inference for single and multiple changepoints in time series. *Journal of Time Series Analysis*, 34(4):423–446.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Lavielle, M. and Teyssière, G. (2007). Adaptive detection of multiple change-points in asset price volatility. In *Long-Memory in Economics*, pages 129–156. Springer Verlag, Berlin, Germany.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*. Springer, Berlin, Heidelberg.

Lee, J., Jedra, Y., Proutière, A., and Yun, S.-Y. (2025). Near-optimal clustering in mixture of markov chains. *arXiv preprint arXiv:2506.01324*.

Lévy-Leduc, C. and Roueff, F. (2009). Detection and localization of change-points in high-dimensional network traffic data. *The Annals of Applied Statistics*, 3(2):637–662.

Lung-Yut-Fong, A., Lévy-Leduc, C., and Cappé, O. (2012). Distributed detection/localization of change-points in high-dimensional network traffic data. *Statistics and Computing*, 22(2):485–496.

Madrid Padilla, O. H., Yu, Y., Wang, D., and Rinaldo, A. (2022). Optimal Nonparametric Multivariate Change Point Detection and Localization. *IEEE Transactions on Information Theory*, 68(3):1922–1944.

Maidstone, R. (2016). *Efficient analysis of complex changepoint problems*. Lancaster University (United Kingdom).

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

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Padilla, O. H. M., Yu, Y., Wang, D., and Rinaldo, A. (2021). Optimal nonparametric change point analysis. *Electronic Journal of Statistics*, 15(1):1154–1201. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41:100–105.

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42:523–527.

Samson, P.-M. (2000). Concentration of measure inequalities for Markov chains and Φ-mixing processes. *The Annals of Probability*, 28(1):416–461. Publisher: Institute of Mathematical Statistics.

Sen, B. (2018). A Gentle Introduction to Empirical Process Theory and Applications. *Lecture Notes, Columbia University*, 11.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126(3):505–563.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299. arXiv:1801.00718 [cs].

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Verbesselt, J., Hyndman, R., Newnham, G., and Culvenor, D. (2010). Detecting trend and seasonal changes in satellite images time series. *Remote Sensing of Environment*, (114):106–115.

Wellner, J. A. (1978). Limit theorems for the ratio of the empirical distribution function to the true distribution function. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 45(1):73–88.

Wellner, J. A. (2017). The Bennett-Orlicz Norm. *Sankhya A*, 79(2):355–383.

Zhang, J. (2002). Powerful Goodness-of-Fit Tests Based on the Likelihood Ratio. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(2):281–294. Publisher: [Royal Statistical Society, Wiley].

Zhang, J. (2006). Powerful Two-Sample Tests Based on the Likelihood Ratio. *Technometrics*, 48(1):95–103. Publisher: ASA Website _eprint: https://doi.org/10.1198/004017005000000328.

Zou, C., Yin, G., Feng, L., and Wang, Z. (2014). Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics*, 42(3):970–1002.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   [Yes. See §5.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   [Not applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   [Not applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   [Yes. See §4.]
   (b) Complete proofs of all theoretical results.
   [Yes. See §A.1.]
   (c) Clear explanations of any assumptions.
   [Yes. See §4.]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
   [Yes. See §5.]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
   [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
   [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets.
   [Yes. See §A.1.]
   (b) The license information of the assets, if applicable.
   [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable.
   [Not Applicable]
   (d) Information about consent from data providers/curators.
   [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
   [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots.
   [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
   [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
   [Not Applicable]

Supplementary Materials for "Nonparametric Multi Change Point Detection for Markov Chains via Adaptive Clustering"

A Proofs

A.1 Proof of Theorem 2

This section is dedicated to the proof of Theorem 2. As detailed in the sketch of the proof, Theorem 2 will follow as a consequence of a series of 6 lemmas (Lemma 3-8). In this section, we only prove Lemmas 7 and 8, while the proofs of Lemmas 3-5 are not directly related and thus deferred till later in this section. Before presenting our results, we introduce the $\mathcal{O}(\cdot,\cdot,\cdot)$ notation for convenience.

Definition 2. *A sequence of random variables Z_n is said to be $\mathcal{O}_p(a_n, b_n, \kappa_n^\dagger)$ if*

$$\lim_{n\to\infty} b_n \mathbb{P}(|Z_n| > \kappa_n^\dagger a_n) \leq \varepsilon \tag{A.1}$$

where a_n, b_n are sequence of positive real numbers, and $\kappa_n^\dagger > 0$ is assumed to be independent of n.

We now state Lemmas 3-6 starting with Lemma 3 provides a deviation bound for the empirical distribution in a neighborhood of the change point. It is proved in §A.4.

Lemma 3. *Let $l < k$ be two time points and let $n_{kl} := l - k$, and*

$$\xi_m(k,l) := n_{kl} \int_{X_{(1)}}^{X_{(n)}} \left(\hat{F}_k^l(u) - F_m(u)\right)^2 d\hat{F}_n(u), \tag{A.2}$$

and assume that the conditions in Assumptions 3, and 1 hold. Then, with δ_n as specified in the hypothesis of Theorem 2, $\kappa^\star$, κ are as in the statement of Theorem 1, $\kappa_n^\dagger$ be a large enough positive constant such that $\kappa_i(\rho,\lambda)\exp(-\kappa_i(\rho,\lambda)\kappa_n^\dagger \log x) < \varepsilon/x$, for all $i \in 1,\ldots,K_n+1$, and ε is as in eq. (A.1), we have

$$\sup_{\tau_{m-1}\leq k<l\leq\tau_m} \xi_m(k,l) = \mathcal{O}_p(u_n, K_n, \kappa_n^\dagger), \text{ where } u_n := 8\log(K_n\delta_n^2)\log\delta_n. \tag{A.3}$$

Before stating the next result, we introduce $S_n(\tau_1',\ldots,\tau_L')$ as the sum of the between time points $(\tau_1,\ldots,\tau_L)$. Formally,

$$S_n(\tau_1',\tau_2',\ldots,\tau_L') := (\tau_{i+1}' - \tau_i') \sum_{i=1}^{L-1} \int_{X_{(1)}}^{X_{(n)}} \hat{F}_{\tau_i'}^{\tau_{i+1}'}(u)\left(1 - \hat{F}_{\tau_i'}^{\tau_{i+1}'}(u)\right) d\hat{F}_n(u) \tag{A.4}$$

Next, we require another Lemma which establishes the monotonicity property of the risk and is proved in §A.5.

Lemma 4. *Under Assumptions 3 and 1, and for any integer $s \in \{1,\ldots,K_n\}$ and for all $L \geq 1$, such that $\tau_s < \tau_1' < \tau_2' < \cdots < \tau_L' < \tau_{s+1}$, we have*

$$0 \geq R_n(\tau_s, \tau_1', \ldots, \tau_L', \tau_{s+1}) - R_n(\tau_s, \tau_{s+1}) = \mathcal{O}_p(u_n^{(L)}, K_n, \kappa_n^\dagger)$$

where $u_n^{(L)} = 8\kappa_n^\dagger L\log(LK_n\delta_n^2)\log\delta_n$, and $\kappa_n^\dagger$ is as in Lemma 3.

The following lemma (proved in §A.6) is a generalisation of the previous lemma.

Lemma 5. *Under Assumptions 3 and 1, let $\tilde{\tau}_1 < \cdots < \tilde{\tau}_s$ be a collection of s time points and let $\tau_1' < \cdots < \tau_p'$ be a collection of p time points. Then,*

$$R_n(\tilde{\tau}_1,\ldots,\tilde{\tau}_s) \geq R_n(\tilde{\tau}_1,\ldots,\tilde{\tau}_s,\tau_1',\ldots,\tau_p')$$

where the risk for an unordered sequence of time points $\tau_1,\ldots,\tau_s,\tau_1',\ldots,\tau_p'$ is defined as in eq. (3.3).

Since

$$S_n(\tau_s, \tau'_1, \ldots, \tau'_L, \tau_{s+1}) - S_n(\tau_s, \tau_{s+1}) = R_n(\tau_s, \tau'_1, \ldots, \tau'_L, \tau_{s+1}) - R_n(\tau_s, \tau_{s+1})$$

the following Lemma can be proved as a corollary to Lemma 4.

Lemma 6. *Under Assumptions 3 and 1, and for any integer $s \in \{1, \ldots, K_n\}$ and for all $L \geq 1$, such that $\tau_s < \tau'_1 < \tau'_2 < \cdots < \tau'_L < \tau_{s+1}$, we have*

$$0 \geq S_n(\tau_s, \tau'_1, \ldots, \tau'_L, \tau_{s+1}) - S_n(\tau_s, \tau_{s+1}) = \mathcal{O}_p(u_n^{(L)}, K_n, \kappa_n^\dagger)$$

where $u_n^{(L)} = 8\kappa_n^\dagger L \log(LK_n\delta_n^2) \log \delta_n$, and $\kappa_n^\dagger$ is as in Lemma 3.

We move on to stating our next result which proves that under Assumptions 3, 4, and 1, we detect at least the correct number of change points.

Lemma 7. *Under Assumptions 3, 4, and 1, and the hypothesis of Theorem 2, we have $\mathbb{P}(\hat{K}_n \geq K_n) \to 1$.*

A.2 Proof of Lemma 7

Proof. The proof of this lemma is divided into two steps. We will first prove a deviation bound of $\mathrm{BIC}_{\mathrm{K_n}}$ from $\mathrm{BIC}_{\mathrm{L}}$. Then the rest of the proof follows using an union bound on the events of incurring an error for every $1 < L < K_n$. We begin with the first step.

Step I: We will first prove the following statement via induction: For any $L < K_n$

$$\mathrm{BIC}_{\mathrm{L}} - \mathrm{BIC}_{\mathrm{K_n}} \geq 3(K_n - L)\eta_{\min} - (K_n - L)(K_n + 5)\mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger) - (K_n - L)\zeta_n. \tag{A.5}$$

For $r = 1, \ldots, K_n$ let $\mathcal{B}_r(L, \delta_n)$ be the set of all estimated change points for which at least one τ_r is δ_n away from every estimate. Formally, we define

$$\mathcal{B}_r(L, \delta_n) := \{(\tau'_1, \ldots, \tau'_L) : \tau'_1 < \cdots < \tau'_L \text{ and } |\tau'_s - \tau_r| > \delta_n \forall\, 1 \leq s \leq L\}. \tag{A.6}$$

For $L = K_n - 1$, by pigeon-hole principle, there exists at least one τ_r such that $|\tau'_i - \tau_r|$ is large for all i. Therefore, the estimated change points $(\hat{\tau}_1, \ldots, \hat{\tau}_L) \in \mathcal{B}_r(L, \delta_n)$ for some r.

Let $(\tau'_1, \ldots, \tau'_L)$ be any element of $\mathcal{B}_r(L, \delta_n)$. Let $\tau_{K_n+1} = n$ and for $i = 0, \ldots, r-1, r+2, \ldots, K_n$ let $\{\tau'_{i,1}, \ldots, \tau'_{i,\max}\}$ be the largest ordered subset of $\{\tau'_1, \ldots, \tau'_L\}$ with all values between τ_i and τ_{i+1}. If the set $\{\tau'_{i,1}, \ldots, \tau'_{i,\max}\}$ is empty, we trivially define $\tau'_{i,1}$ as τ_{i+1} and $\tau'_{i,\max}$ as τ_{i-1}. We define

$$\begin{aligned} T_i &:= S_n(\tau_i, \tau'_{i,1}, \ldots, \tau'_{i,\max}, \tau_{i+1}) \\ T_r &:= S_n(\tau_{r-1}, \tau_{r-\delta_n}) \\ T_{r+1} &:= S_n(\tau_{r+\delta_n}, \tau_{r+1}) \end{aligned}$$

and T_{K_n+2} as defined below in eq. (A.8). Now, using Lemma 5, we have

$$\begin{aligned} R_n(\tau'_1, \ldots, \tau'_L) &\geq R_n(\tau'_1, \ldots, \tau'_L, \tau_1, \ldots, \tau_{r-1}, \tau_{r-\delta_n}, \tau_{r+\delta_n}, \tau_{r+1}, \ldots, \tau_{K_n}) \\ &= T_0 + T_1 + \cdots + T_{K_n+2} \end{aligned} \tag{A.7}$$

with the last equality following from the definition of T_i's. It follows using Lemma 6 that for all $i = 0, \ldots, r-1, r+2, \ldots, K_n$

$$S_n(\tau_i, \tau_{i+1}) \geq T_i \geq S_n(\tau_i, \tau_{i+1}) + \mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger).$$

It follows by trivially subtracting $\mathcal{O}_p$ terms that,

$$\begin{aligned} T_r &\geq S_n(\tau_{r-1}, \tau_{r-\delta_n}) + \mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger) \\ T_{r+1} &\geq S_n(\tau_{r+\delta_n}, \tau_{r+1}) + \mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger). \end{aligned}$$

Finally, we are left with T_{K_n+2} which is

$$\begin{aligned} T_{K_n+2} &= S_n(\tau_r - \delta_n, \tau_r + \delta_n) + S_n(\tau_r + \delta_n, \tau_{r+1}) \\ &= S_n(\tau_r - \delta_n, \tau_r) + S_n(\tau_r, \tau_r + \delta_n) + \Delta_S, \end{aligned} \tag{A.8}$$

where (with $F_{r-1,1/2}$ as defined in eq. (A.9))

$$F_{r-1,1/2}(u) = \frac{F_{r-1}(u) + F_r(u)}{2}. \tag{A.9}$$

$$\begin{aligned} \Delta_S &:= S_n(\tau_r - \delta_n, \tau_r + \delta_n) - S_n(\tau_r - \delta_n, \tau_r) - S_n(\tau_r, \tau_r + \delta_n) \\ &= 2\delta_n \int_{X_{(1)}}^{X_{(n)}} \hat{F}^{\tau_r+\delta_n}_{\tau_r-\delta_n}(u)\left(1 - \hat{F}^{\tau_r+\delta_n}_{\tau_r-\delta_n}(u)\right) d\hat{F}_n(u) \\ &\quad - \delta_n \int_{X_{(1)}}^{X_{(n)}} \hat{F}^{\tau_r}_{\tau_r-\delta_n}(u)\left(1 - \hat{F}^{\tau_r}_{\tau_r-\delta_n}(u)\right) d\hat{F}_n(u) \\ &\quad - \delta_n \int_{X_{(1)}}^{X_{(n)}} \hat{F}^{\tau_r+\delta_n}_{\tau_r}(u)\left(1 - \hat{F}^{\tau_r+\delta_n}_{\tau_r}(u)\right) d\hat{F}_n(u) \end{aligned} \tag{A.10}$$

$$= -\int_{X_{(1)}}^{X_{(n)}} \left[2\delta_n(\hat{F}^{\tau_r+\delta_n}_{\tau_r-\delta_n}(u))^2 - \delta_n(\hat{F}^{\tau_r}_{\tau_r-\delta_n}(u))^2 - \delta_n(\hat{F}^{\tau_r+\delta_n}_{\tau_r}(u))^2\right] d\hat{F}_n(u)$$

$$= -\int_{X_{(1)}}^{X_{(n)}} \left[2\delta_n(\hat{F}^{\tau_r+\delta_n}_{\tau_r-\delta_n}(u))^2 - \delta_n(\hat{F}^{\tau_r}_{\tau_r-\delta_n}(u))^2 - \delta_n(\hat{F}^{\tau_r+\delta_n}_{\tau_r}(u))^2 + \mathcal{H} - \mathcal{H}\right] d\hat{F}_n(u) \tag{A.11}$$

where (with the dependence on u implicit for convenience)

$$\begin{aligned} \mathcal{H} &= -4\delta_n \hat{F}^{\tau_r+\delta_n}_{\tau_r-\delta_n} F_{r-1,1/2} + 2\delta_n (F_{r-1,1/2})^2 \\ &\quad + 2\delta_n \hat{F}^{\tau_r}_{\tau_r-\delta_n} F_{r-1} - \delta_n (F_{r-1})^2 \\ &\quad + 2\delta_n \hat{F}^{\tau_r+\delta_n}_{\tau_r} F_r - \delta_n (F_r)^2 \\ &= \delta_n \left(\hat{F}^{\tau_r}_{\tau_r-\delta_n} - \hat{F}^{\tau_r+\delta_n}_{\tau_r}\right)(F_{r-1} - F_r) \\ &\xrightarrow{n\to\infty} \delta_n (F_{r-1} - F_r)^2 + o(\delta_n) \end{aligned} \tag{A.12}$$

uniformly in u.

Now using Lemma A.4, the right hand side of eq. (A.11) becomes

$$\begin{aligned} &-\int_{X_{(1)}}^{X_{(n)}} \left[2\delta_n(\hat{F}^{\tau_r+\delta_n}_{\tau_r-\delta_n}(u) - F_{r-1,1/2}(u))^2 - \delta_n(\hat{F}^{\tau_r}_{\tau_r-\delta_n}(u) - F_{r-1}(u))^2 - \delta_n(\hat{F}^{\tau_r+\delta_n}_{\tau_r}(u) - F_r(u))^2 - \mathcal{H}\right] d\hat{F}_n(u) \\ &= 3\mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger) - \int_{X_{(1)}}^{X_{(n)}} \mathcal{H} d\hat{F}_n(u). \end{aligned}$$

Observe that $d\hat{F}_n(u) = 0$ for all $u \notin [X_{(1)}, X_{(n)}]$ which, using implies Assumption 3, and eq. (A.12)

$$\begin{aligned} \int_{X_{(1)}}^{X_{(n)}} \mathcal{H} d\hat{F}_n &= \int_0^1 \mathcal{H} d\hat{F}_n \\ &\xrightarrow{n\to\infty} \delta_n \int_0^1 (F_{r-1}(u) - F_r(u))^2 dF(u) + o_p(\delta_n) \\ &= \delta_n \int \eta_r(u) dF(u) + o_p(\delta_n). \end{aligned}$$

Now it follows by substituting the previous bounds in eq. (A.7) that,

$$\begin{aligned}
&\min_{(\tau_1',\ldots,\tau_L')\in B_r(L,\delta_n)} R_n(\tau_1',\ldots,\tau_L') \\
&\geq \min_{(\tau_1',\ldots,\tau_L')\in B_r(L,\delta_n)} R_n(\tau_1',\ldots,\tau_L',\tau_1,\ldots,\tau_{r-1},\tau_{r-\delta_n},\tau_{r+\delta_n},\tau_{r+1},\ldots,\tau_{K_n}) \\
&\geq R_n(\tau_1,\ldots,\tau_{K_n}) + (K_n+5)\mathcal{O}_p(u_n^{(K_n)},K_n,\kappa_n^\dagger) - \delta_n\int \eta_r(u)dF(u) + o_p(\delta_n).
\end{aligned}$$

Let $\mathrm{BIC_{K_n}} = -R_n(\tau_1,\ldots,\tau_{K_n}) + K_n\zeta_n$ and recall from eq. (BIC) the definition of $\mathrm{BIC_L}$. Following the calculations above, we get

$$\begin{aligned}
\mathrm{BIC_L} - \mathrm{BIC_{K_n}} &\leq -\delta_n\int \eta_r(u)dF(u) - (K_n+5)\mathcal{O}_p(u_n^{(K_n)},K_n,\kappa_n^\dagger) - \zeta_n \\
&\leq -\delta_n\eta_{\min} - (K_n+5)\mathcal{O}_p(u_n^{(K_n)},K_n,\kappa_n^\dagger) - \zeta_n \qquad \text{(A.13)}
\end{aligned}$$

where the second inequality follows from the definition of $\eta_{\min}$ in Assumption 4. This proves our induction hypothesis for $L = K_n - 1$.

Now let the hypothesis of the induction (eq. (A.5)) hold true for $L = K_n - r$. Reparametrizing $K_n - r$ as K_n', we now show for $L = K_n' - 1$. It is easy to see invoking eq. (A.13) that,

$$\begin{aligned}
\mathrm{BIC_L} - \mathrm{BIC_{K_n'}} &\leq -\delta_n\eta_{\min} - (K_n+5)\mathcal{O}_p(u_n^{(K_n')},K_n',\kappa_n^\dagger) - \zeta_n \\
&\leq -\delta_n\eta_{\min} - (K_n+5)\mathcal{O}_p(u_n^{(K_n)},K_n,\kappa_n^\dagger) - \zeta_n
\end{aligned}$$

where inequality follows since for $K_n' < K_n$, $|\mathcal{O}_p(u_n^{(K_n')},K_n',\kappa_n^\dagger)| \leq |\mathcal{O}_p(u_n^{(K_n)},K_n,\kappa_n^\dagger)|$. Therefore using induction,

$$\begin{aligned}
\mathrm{BIC_{L+1}} - \mathrm{BIC_{K_n}} &= \mathrm{BIC_{L+1}} - \mathrm{BIC_L} + \mathrm{BIC_L} - \mathrm{BIC_{K_n}} \\
&\leq -\delta_n(K_n - r + 1)\eta_{\min} - K_n(K_n+5)\mathcal{O}_p(u_n^{(K_n)},K_n,\kappa_n^\dagger) - (K_n - r + 1)\zeta_n
\end{aligned}$$

which is what we required.

Step II: It now follows that, for any n,

$$\begin{aligned}
\mathbb{P}(\hat{K}_n < K_n) &= \mathbb{P}\left(\bigcup_{L=1}^{K_n-1}\{\mathrm{BIC_L} > \mathrm{BIC_{K_n}}\}\right) \\
&\leq \sum_{L=1}^{K_n-1}\mathbb{P}\left(\mathrm{BIC_L} > \mathrm{BIC_{K_n}}\right) \\
&\leq \sum_{L=1}^{K_n-1}\mathbb{P}\left(K_n(K_n+5)\mathcal{O}_p(u_n^{(K_n)},K_n,\kappa_n^\dagger) < -\delta_n\eta_{\min} - (K_n-L)\zeta_n\right) \\
&= \sum_{L=1}^{K_n-1}\mathbb{P}\left(\mathcal{O}_p(u_n^{(K_n)},K_n,\kappa_n^\dagger) < -\frac{\delta_n}{5K_n^2}\eta_{\min} - (K_n-L)\frac{\zeta_n}{5K_n^2}\right) \\
&= \sum_{L=1}^{K_n-1}\mathbb{P}\left(\mathcal{O}_p(u_n^{(K_n)},K_n,\kappa_n^\dagger) < -\frac{u_n^{(K_n)}\delta_n\log K_n}{40\kappa_n^\dagger K_n^3(\log\delta_n)^2(\log K_n)^2}\eta_{\min} - (K_n-L)\frac{\zeta_n}{5K_n^2}\right)
\end{aligned}$$

where the last inequality follows by substituting the terms from eq. (A.13).

Recall that under the hypothesis of the theorem,

$$K_n^3\log(\delta_n)^2(\log K_n)^2/\delta_n = \mathcal{O}(1), \quad \text{so that} \quad \delta_n/K_n^3(\log\delta_n)^2(\log K_n)^2 = \Omega(1).$$

Therefore, for some large n, we have

$$\frac{\delta_n \log K_n}{40\kappa_n^\dagger K_n^3 \log(\delta_n)^2 (\log K_n)^2}\eta_{\min} > 1$$

This implies

$$\begin{aligned}
&\sum_{L=1}^{K_n-1} \mathbb{P}\left(\mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger) < -\frac{u_n^{(K_n)}\delta_n \log K_n}{5K_n^3 \log(\delta_n)^2(\log K_n)^2}\eta_{\min} - (K_n - L)\frac{\zeta_n}{5K_n^2}\right) \\
&\leq \sum_{L=1}^{K_n-1} \mathbb{P}\left(\mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger) < -u_n^{(K_n)} - (K_n - L)\frac{\zeta_n}{5K_n^2}\right) \\
&\leq \sum_{L=1}^{K_n-1} \mathbb{P}\left(\mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger) < -u_n^{(K_n)}\right) \\
&= K_n \mathbb{P}\left(\mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger) < -u_n^{(K_n)}\right) \\
&< \varepsilon.
\end{aligned} \tag{A.14}$$

This completes the proof. □

The next lemma establishes that no more than the correct number of change points is detected.

Lemma 8. *Under Assumptions 3, 4, and 1, and the hypothesis of Theorem 2,* $\mathbb{P}\left(\hat{K}_n > K_n\right) \to 0$.

A.3 Proof of Lemma 8

Proof. To avoid trivialities, assume that $K_n > 0$ and consider two cases:

Case I ($(\hat{\tau}_1, \ldots, \hat{\tau}_L) \in \bigcup_{r=1}^{K_n} \mathcal{B}_r(L, \delta_n)$)*:* When $(\hat{\tau}_1, \ldots, \hat{\tau}_L) \in \mathcal{B}_r(L, \delta_n)$, for some $r \in \{1, \ldots, K_n\}$ (with $\mathcal{B}_r(L, \delta_n)$ defined as in eq. (A.6)), it follows similarly to the proof of Lemma 7 that

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{L=1}^{K_n-1} \{\mathrm{BIC_L} > \mathrm{BIC_{K_n}}\}\right) &\leq K_n \mathbb{P}\left(\mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger) > u_n^{(L)}\right) \\
&< \varepsilon.
\end{aligned}$$

Case II ($(\hat{\tau}_1, \ldots, \hat{\tau}_L) \in \mathcal{C}(L, \delta_n)$)*:* To formalize the second case, we introduce the following notation

$$\mathcal{C}(L, \delta_n) := \{(\tau_1', \ldots, \tau_L') : 1 < \tau_1' < \cdots < \tau_L' \leq n, \text{and } \exists \text{ i such that } |\tau_i' - \tau_s| \leq \delta_n \ \forall\, 1 \leq s \leq L\}$$

We will show that

$$R_n(\tau_1', \ldots, \tau_L') \geq R_n(\tau_1, \ldots, \tau_{K_n}) + 4L\mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger). \tag{A.15}$$

Following eq. (A.7),

$$\begin{aligned}
R_n(\tau_1', \ldots, \tau_L') &\geq R_n(\tau_1', \ldots, \tau_L', \tau_1, \ldots, \tau_{r-1}, \tau_r, \tau_{r+1}, \ldots, \tau_{K_n}) \\
&= T_0 + T_1 + \cdots + T_{K_n} \\
&\geq \sum_{i=0}^{K_n} S_n(\tau_i, \tau_{i+1}) + K_n \mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger) \\
&= R_n(\tau_1, \ldots, \tau_n) + K_n \mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger).
\end{aligned}$$

Therefore,

$$R_n(\tau'_1, \ldots, \tau'_L) \geq R_n(\tau_1, \ldots, \tau_{K_n}) + K_n \mathcal{O}_p(u_n, K_n, \kappa_n^\dagger).$$

As before, let $\mathrm{BIC}_{\mathrm{K_n}} = \mathrm{BIC}_\star = -R_n(\tau_1, \ldots, \tau_{K_n}) + K_n\zeta_n$. Then, for any given L,

$$\mathrm{BIC_L} - \mathrm{BIC_{K_n}} \leq -K_n \mathcal{O}_p(u_n, K_n, \kappa_n^\dagger) - (K_n - L)\zeta_n$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left(\hat{K}_n > K_n\right) &= \mathbb{P}\left(\bigcup_{L=K_n+1}^{\bar{K}} \{\mathrm{BIC_L} > \mathrm{BIC_{K_n}}\}\right) \\
&= \sum_{L=K_n+1}^{\bar{K}} \mathbb{P}\left(\mathrm{BIC_L} > \mathrm{BIC_{K_n}}\right) \\
&\leq \sum_{L=K_n+1}^{\bar{K}} \mathbb{P}\left(4K_n \mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger) > (L - K_n)\zeta_n\right) \\
&\leq \sum_{L=K_n+1}^{\bar{K}} \mathbb{P}\left(\mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger) > \left(\frac{L}{K_n} - 1\right)\frac{\zeta_n}{4}\right) \\
&\leq \bar{K}\mathbb{P}\left(\mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger) > \left(\frac{1}{K_n}\right)\frac{\zeta_n}{4}\right) \\
&\overset{(i)}{\leq} \bar{K}\mathbb{P}\left(\mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger) > u_n^{(K_n)}\right) \\
&= \frac{\bar{K}}{K_n} K_n \mathbb{P}\left(\mathcal{O}_p(u_n^{(K_n)}, K_n, \kappa_n^\dagger) > u_n^{(K_n)}\right) \\
&\overset{(i)}{<} \frac{\bar{K}}{K_n}\varepsilon
\end{aligned}$$

where (i) follows from eq. (A.1). Recall that $\bar{K}$ is bounded and we have assumed K_n to be positive. Since ε is arbitrary, the proof follows. □

Proof of Theorem 2. Now we can finally prove our main theorem. It follows by combining Lemma 7 and 8 that $\mathbb{P}\left(\hat{K}_n = K_n\right) \to 1$. □

A.4 Proof of Lemma 3

Proof. We first show that under the hypothesis of the Lemma 3,

$$\lim_{n\to\infty} K_n \mathbb{P}\left(\sup_{\tau_{m-1} \leq k < l \leq \tau_{m-1} + \delta_n} \xi_m(k, l) \geq u_n\right) < \varepsilon. \tag{A.16}$$

As before, let $n_{kl} = l - k$, $\|\hat{F}_k^l - F\|_0^1 := \sup_{u\in[0,1]} |\hat{F}_k^l(u) - F(u)|$ and E_n to be the following event

$$E_n := \bigcup_{k,l} \left\{\sqrt{n_{kl}}\|\hat{F}_k^l - F\|_0^1 > \sqrt{\kappa_n^\dagger \log(K_n\delta_n^2)\max\{\log 2, \log n_{kl}\}}\right\}.$$

Then using union bound,

$$\begin{aligned}
\mathbb{P}(E_n) &\leq \sum_{k,l} \mathbb{P}\left(\sqrt{n_{kl}}\|\hat{F}_k^l - F\|_0^1 > \sqrt{\kappa_n^\dagger \log(K_n\delta_n^2)\max\{\log 2, \log n_{kl}\}}\right) \\
&\overset{(i)}{\leq} \sum_{k,l} \kappa^\star(\rho,\lambda)\exp\left(-\kappa(\rho,\lambda)\kappa_n^\dagger \frac{n_{kl}}{\log n_{kl}} \frac{\log(K_n\delta_n^2)\max\{\log 2, \log n_{kl}\}}{n_{kl}}\right) \\
&\overset{(ii)}{\leq} \delta_n^2 \kappa^\star(\rho,\lambda)\exp\left(-\kappa(\rho,\lambda)\kappa_n^\dagger \frac{\log(K_n\delta_n^2)\log n_{kl}}{\log n_{kl}}\right) \\
&\overset{(iii)}{\leq} \delta_n^2 \kappa^\star(\rho,\lambda)\exp\left(-\kappa(\rho,\lambda)\kappa_n^\dagger \log(K_n\delta_n^2)\right) \\
&\overset{(iv)}{\leq} \frac{\varepsilon\delta_n^2}{K_n\delta_n^2} \\
&= \frac{\varepsilon}{K_n}
\end{aligned} \tag{A.17}$$

(i) follows from Theorem 1; (ii) follows because to avoid trivialities we can assume $n_{kl} \geq 2$ and because $|\{\tau_{m-1} \leq k, l \leq \tau_{m-1} + \delta_n\}| \leq \delta_n^2$; (iii) follows simplifying the fraction; (iv) follows from the definition of $\kappa_n^\dagger$ in the statement of Lemma 3.

Now, observe that for any m, k and l,

$$\xi_m(k,l) = n_{kl}\int_{X_{(1)}}^{X_{(n)}} \left|\hat{F}_k^l(u) - F(u)\right|^2 d\hat{F}_n(u) \leq n_{kl}(\|\hat{F}_k^l - F\|_0^1)^2.$$

Since $\log\delta_n \geq \log n_{kl}$, $E_n \subseteq \{\sup_{\tau_{m-1}\leq k<l\leq\tau_{m-1}+\delta_n} \xi_m(k,l) \geq u_n\}$ and it follows from eq. (A.17), that

$$K_n\mathbb{P}\left(\sup_{\tau_{m-1}\leq k<l\leq\tau_{m-1}+\delta_n} \xi_m(k,l) \geq u_n\right) \leq \varepsilon.$$

The conclusion of the lemma now follows. This completes the proof. □

A.5 Proof of Lemma 4

Proof. For convenience of notation, let $n_1^\star = \tau_1' - \tau_s$, and so on until we have $n_L^\star = \tau_L' - \tau_{L-1}'$, and $n_{L+1}^\star = \tau_{s+1} - \tau_L'$. Let $n^\star = \sum_i n_i^\star$. Then, overloading the notation of τ_s as τ_0' and τ_{s+1} as τ_{L+1}', we observe that

$$\begin{aligned}
&R_n(\tau_s, \tau_1', \ldots, \tau_L', \tau_{s+1}) - R_n(\tau_s, \tau_{s+1}) \\
&= \int_{X_{(1)}}^{X_{(n)}} \left[\sum_{i=1}^{L+1} n_i^\star \hat{F}_{\tau_i'}^{\tau_{i+1}'}(u)\left(1 - \hat{F}_{\tau_i'}^{\tau_{i+1}'}(u)\right) - n^\star \hat{F}_{\tau_s}^{\tau_{s+1}}(u)\left(1 - \hat{F}_{\tau_s}^{\tau_{s+1}}(u)\right)\right] d\hat{F}_n(u)
\end{aligned}$$

Looking at the integrand (and dropping u for convenience) we get

$$\begin{aligned}
\sum_{i=1}^{L+1} n_i^\star \hat{F}_{\tau_i'}^{\tau_{i+1}'}\left(1 - \hat{F}_{\tau_i'}^{\tau_{i+1}'}\right) - n^\star \hat{F}_{\tau_s}^{\tau_{s+1}}\left(1 - \hat{F}_{\tau_s}^{\tau_{s+1}}\right) &\overset{(i)}{=} -\sum_{i=1}^{L+1} n_i^\star (\hat{F}_{\tau_i'}^{\tau_{i+1}'})^2 + n^\star (\hat{F}_{\tau_s}^{\tau_{s+1}})^2 \\
&\overset{(ii)}{\leq} 0,
\end{aligned} \tag{A.18}$$

where (i) follows since $\sum_{i=1}^{L+1} n_i^\star \hat{F}_{\tau_i'}^{\tau_{i+1}'} = n^\star F_{\tau_s}^{\tau_{s+1}}$ and (ii) follows since $-x^2$ is a concave function. We turn to proving the order term.

We have already pointed out the fact that $\sum_{i=1}^{L+1} n_i^\star \hat{F}_{\tau_i'}^{\tau_{i+1}'} = n^\star F_{\tau_s}^{\tau_{s+1}}$ and recall that $\sum_{i=1}^{L+1} n_i^\star = n^\star$ by definition. Now consider the right hand side of eq. (A.18) and add and subtract $n^\star(F_s)^2 - 2n^\star \hat{F}_{\tau_s}^{\tau_{s+1}} F_s$ to get,

$$\begin{aligned}
-\sum_{i=1}^{L+1} n_i^\star (\hat{F}_{\tau_i'}^{\tau_{i+1}'})^2 + n^\star (\hat{F}_{\tau_s}^{\tau_{s+1}})^2 &= -\sum_{i=1}^{L+1} n_i^\star \left[(\hat{F}_{\tau_i'}^{\tau_{i+1}'})^2 + (F_s)^2 - 2n^\star \hat{F}_{\tau_i'}^{\tau_{i+1}'} F_s \right] \\
&\quad + n^\star \left[(\hat{F}_{\tau_s}^{\tau_{s+1}})^2 + (F_s)^2 - 2\hat{F}_{\tau_s}^{\tau_{s+1}} F_s \right] \\
&= -\sum_{i=1}^{L+1} n_i^\star (\hat{F}_{\tau_i'}^{\tau_{i+1}'} - F_s)^2 + \underbrace{n^\star (\hat{F}_{\tau_s}^{\tau_{s+1}} - F_s)^2}_{\geq 0} \\
&\geq -\sum_{i=1}^{L+1} n_i^\star (\hat{F}_{\tau_i'}^{\tau_{i+1}'} - F_s)^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
R_n(\tau_s, \tau_1', \ldots, \tau_L', \tau_{s+1}) - R_n(\tau_s, \tau_{s+1}) &\geq \int_{X_{(i)}}^{X_{(n)}} -\sum_{i=1}^{L+1} n_i^\star (\hat{F}_{\tau_i'}^{\tau_{i+1}'}(u) - F_s(u))^2 d\hat{F}_n(u) \\
&\geq -\sum_{i=1}^{L+1} \xi_s(\tau_i', \tau_{i+1}').
\end{aligned}$$

For any collection of positive random variables $Z_1, \ldots, Z_{L+3}$, recall the probabilistic bound

$$\mathbb{P}\left(\sum_{i=1}^{L+1} Z_i > a \right) \leq \sum_{i=1}^{L+1} \mathbb{P}\left(Z_i > \frac{a}{L+1} \right). \tag{A.19}$$

Therefore, using Lemma 3

$$\begin{aligned}
&\lim_{n\to\infty} K_n \mathbb{P}\left(|R_n(\tau_s, \tau_1', \ldots, \tau_L', \tau_{s+1}) - R_n(\tau_s, \tau_{s+1})| > 8\kappa_n^\dagger L \log(LK_n\delta_n^2) \log \delta_n \right) \\
&\leq \lim_{n\to\infty} K_n \mathbb{P}\left(\sum_{i=1}^{L+1} \xi_s(\tau_i', \tau_{i+1}') > 8\kappa_n^\dagger L \log(K_n L \delta_n^2) \log \delta_n \right) \\
&\leq \lim_{n\to\infty} K_n \mathbb{P}\left(\sum_{i=1}^{L+1} \xi_s(\tau_i', \tau_{i+1}') > 4\kappa_n^\dagger (L+1) \log(LK_n\delta_n^2) \log \delta_n \right) \\
&\leq L^{-1} \sum_{i=1}^{L+1} \lim_{n\to\infty} LK_n \mathbb{P}\left(\xi_s(\tau_i', \tau_{i+1}') > 4\kappa_n^\dagger \log(LK_n\delta_n^2) \log \delta_n \right) \\
&< L^{-1}(L+1)\,\varepsilon\,.
\end{aligned}$$

Since $n > \delta_n$, the statement of Lemma 4 is now established. □

A.6 Proof of Lemma 5

Proof. Fix one coarse interval $J = (\tilde{\tau}_j, \tilde{\tau}_{j+1})$ induced by $\{\tilde{\tau}_1, \ldots, \tilde{\tau}_s\}$. The extra points $\{\tau_1', \ldots, \tau_p'\}$ partition J into subintervals $J_1, \ldots, J_{m_j}$ with counts $n_{j,k}^\star$ (so that $\sum_{k=1}^{m_j} n_{j,k}^\star = n_J^\star$) and $p_{j,k}(u) := \hat{F}_{J_k}(u) \in [0,1]$. Set weights $w_{j,k} := n_{j,k}^\star / n_J^\star$ and the weighted average $\bar{p}_j(u) := \sum_{k=1}^{m_j} w_{j,k} p_{j,k}(u) = \hat{F}_J(u)$.

Let $f(x) = x(1-x) = x - x^2$, a concave function on $[0,1]$. Then, pointwise in u,

$$\begin{aligned}
\sum_{k=1}^{m_j} n_{j,k}^\star f\big(p_{j,k}(u)\big) - n_J^\star f\big(\bar{p}_j(u)\big) &= -\,n_J^\star \left(\sum_{k=1}^{m_j} w_{j,k}\, p_{j,k}(u)^2 - \Big(\sum_{k=1}^{m_j} w_{j,k}\, p_{j,k}(u) \Big)^2 \right) \\
&= -\,n_J^\star \operatorname{Var}_{w_j}\big(p_{j,k}(u)\big) \leq 0.
\end{aligned}$$

Equivalently,

$$\sum_{k=1}^{m_j} n^\star_{j,k}\hat{F}_{J_k}(u)\big(1-\hat{F}_{J_k}(u)\big) \le n^\star_J \hat{F}_J(u)\big(1-\hat{F}_J(u)\big).$$

Summing this inequality over all coarse intervals $J \in \mathcal{J}(\{\tilde{\tau}_1,\dots,\tilde{\tau}_s\})$ and integrating with respect to $d\hat{F}_n(u)$ yields

$$R_n(\tilde{\tau}_1,\dots,\tilde{\tau}_s,\tau'_1,\dots,\tau'_p) \le R_n(\tilde{\tau}_1,\dots,\tilde{\tau}_s).$$

Equality holds iff $\mathrm{Var}_{w_j}(p_{j,k}(u)) = 0$ for all j and a.e. u, i.e., when each coarse interval has identical within-subinterval cdfs. □

A.7 Proof of Theorem 1

Now we proceed to proving Theorem 1. Since there is always a rational number between any two real numbers, it holds almost everywhere that

$$\sup_{f\in\mathcal{F}_{[0,1]}}\left|\sum_{i=1}^n (f(Y_i)-\mathbb{E}_\pi[f(Y)])\right| \le \sup_{f\in\mathcal{F}_{[0,1]}\bigcap\mathbb{Q}}\left|\sum_{i=1}^n (f(Y_i)-\mathbb{E}_\pi[f(Y)])\right| + 2$$

Therefore,

$$\mathbb{P}(nZ > nt+\kappa\mathcal{R}(\mathcal{F}_{[0,1]}\bigcap\mathbb{Q}))) \le \mathbb{P}\left(\underbrace{\sup_{f\in\mathcal{F}_{[0,1]}\bigcap\mathbb{Q}}\left|\sum_{i=1}^n (f(Y_i)-\mathbb{E}_\pi[f(Y)])\right| > nt+\kappa\mathcal{R}(\mathcal{F}_{[0,1]}\bigcap\mathbb{Q})) - 2}_{=:\mathcal{T}}\right)$$

$t \ge 3/n$ by hypothesis. Therefore $nt-2 \ge 1$. We now state the following Lemma which is proved in §A.8.

Lemma 9. *Let $Y_1,\dots,Y_n$ be a sequence of random variables from a Markov chain with stationary distribution π, and let Y be a random variable with distribution π. Define*

$$Z' := \sup_{f\in\mathcal{F}_{[0,1]}\bigcap\mathbb{Q}}\left|\sum_{i=1}^n (f(Y_i)-\mathbb{E}_\pi[f(Y)])\right|.$$

Then, for some universal constant $\kappa > 4e$, any $\kappa_\rho > \sqrt{\mathbb{E}_A[\rho_A^2(2)]}$, $\kappa_\lambda = 2\mathbb{E}_A[\exp(\rho_A(2)\lambda)]/\lambda$, and ρ_o as defined in §4.1,

$$\mathbb{P}\left(Z' > t+\kappa\mathcal{R}(\mathcal{F}_{[0,1]}\bigcap\mathbb{Q})\right) \le \exp\left(-\frac{\mathbb{E}_A[\rho_A(2)]}{\kappa}\min\left\{\frac{t^2}{n\mathbb{E}_A[\rho_A^2(2)]}, \frac{t}{\rho_o^3\log n}\right\}\right) \text{ and}$$

$$\mathcal{R}(\mathcal{F}_{[0,1]}\bigcap\mathbb{Q})) = 2(\mathbb{E}_A[\rho_A(2)]+\mathbb{E}_\nu[\rho_A(2)]) + \kappa\left[\kappa_\rho\log\left(\frac{\kappa_\rho}{\sqrt{\mathbb{E}_A[\rho_A^2(2)]}}\right) + \sqrt{n\mathbb{E}_A[\rho_A^2(2)]\log\left(\frac{\kappa_\rho}{\sqrt{\mathbb{E}_A[\rho_A^2(2)]}}\right)}\right]$$
$$+ n\exp\left(-\kappa_\rho\lambda/2\right)\kappa_\lambda.$$

□

It now follows using Lemma 9 that the right hand side of the previous equation is bounded above by

$$\exp\left(-\frac{\mathbb{E}_A[\rho_A(2)]}{\kappa}\min\left\{\frac{(nt-2)^2}{n\mathbb{E}_A[\rho_A^2(2)]}, \frac{(nt-2)}{\rho_o^3\log n}\right\}\right)$$

By setting $\kappa_\rho = (2/\lambda)\log(n/2\kappa_\lambda)$ and observing that under Assumption 1 (EM), $1/(2\kappa_\lambda) < 1$ we get

$$\mathcal{R}(\mathcal{F}_{[0,1]}\bigcap\mathbb{Q})) \le 2(\mathbb{E}_A[\rho_A(2)]+\mathbb{E}_\nu[\rho_A(2)]) + \kappa\left[2\frac{\log(n)}{\lambda}\log\left(\frac{2\log(n)/\lambda}{\sqrt{\mathbb{E}_A[\rho_A^2(2)]}}\right) + \sqrt{n\mathbb{E}_A[\rho_A^2(2)]\log\left(\frac{2\log(n)/\lambda}{\sqrt{\mathbb{E}_A[\rho_A^2(2)]}}\right)}\right]$$

Observe that $\sqrt{\mathbb{E}_A[\rho_A^2(2)]} \geq 1$. Now, with a constant $\kappa(\tau, \lambda)$ depending on λ and $\mathbb{E}_A[\rho_A^2(2)]$, and $\mathbb{E}_v[\rho_A(2)]$ we have with some standard manipulations

$$\mathcal{R}(\mathcal{F}_{[0,1]\bigcap\mathbb{Q}})) \leq \kappa(\rho,\lambda)\sqrt{n\log n}$$

Then we have

$$\mathbb{P}(nZ > nt + \kappa(\rho,\lambda)\sqrt{n\log n})) \leq \exp\left(-\frac{\mathbb{E}_A[\rho_A(2)]}{\kappa}\min\left\{\frac{(nt-2)^2}{n\mathbb{E}_A[\rho_A^2(2)]}, \frac{(nt-2)}{\rho_o^3\log n}\right\}\right).$$

Now dividing both sides of $\mathcal{T}$ by n and trivially upper bounding 2 by 2κ, we have for some universal constant $\kappa > 0$, and for all $t > 3/n$

$$\mathbb{P}\left(Z > t + \kappa(\rho,\lambda)\sqrt{n\log n}/n\right) \leq \exp\left(-\frac{\mathbb{E}_A[\rho_A(2)]}{\kappa}\min\left\{\frac{(nt-2)^2}{n\mathbb{E}_A[\rho_A^2(2)]}, \frac{(nt-2)}{\rho_o^3\log n}\right\}\right) \tag{A.20}$$

where, for some constant $\kappa(\tau,\lambda)$ depending only on $\mathbb{E}_A[\rho_A^2(2)]$, $\mathbb{E}_v[\rho_A(2)]$, λ. Next, observe that

$$\mathbb{P}\left(Z > t\right) = \mathbb{P}\left(Z - \mathbb{E}Z > t - \mathbb{E}Z\right)$$

Since $\mathbb{E}Z < \kappa(\rho,\lambda)\sqrt{n\log n}/n = \mathcal{O}(\sqrt{\log n/n})$, there exists a constant $\kappa'(\rho,\lambda) > 3/n$ such that

$$t - \mathbb{E}Z > t - \kappa'(\rho,\lambda).$$

Then,

$$\mathbb{P}(Z > t) \leq \exp\left(-\frac{\mathbb{E}_A[\rho_A(2)]}{\kappa}\min\left\{\frac{(n(t-\kappa'(\rho,\lambda))-2)^2}{n\mathbb{E}_A[\rho_A^2(2)]}, \frac{(n(t-\kappa'(\rho,\lambda))-2)}{\rho_o^3\log n}\right\}\right).$$

We now make 2 cases.

Case I: When $t > 2\kappa'(\rho,\lambda)$, we also have $t - \kappa'(\rho,\lambda) - 2/n > t/2$, and hence

$$\mathbb{P}(Z > t) \leq \exp\left(-\frac{\mathbb{E}_A[\rho_A(2)]}{\kappa}\min\left\{\frac{(nt/2)^2}{n\mathbb{E}_A[\rho_A^2(2)]}, \frac{nt/2}{\rho_o^3\log n}\right\}\right).$$

Case II: When $0 < t \leq 2\kappa'(\rho,\lambda)$, there exists a large enough constant $\kappa^\star(\rho,\lambda)$ such that

$$\mathbb{P}(Z > t) \leq \kappa^\star(\rho,\lambda)\exp\left(-\frac{\mathbb{E}_A[\rho_A(2)]}{\kappa}\min\left\{\frac{(nt/2)^2}{n\mathbb{E}_A[\rho_A^2(2)]}, \frac{nt/2}{\rho_o^3\log n}\right\}\right).$$

It therefore follows that, for some large enough constant $\kappa(\rho,\lambda)$ and for all $t > 0$

$$\mathbb{P}(Z > t) \leq \kappa^\star(\rho,\lambda)\exp\left(-\frac{\mathbb{E}_A[\rho_A(2)]}{\kappa}\min\left\{\frac{(nt/2)^2}{n\mathbb{E}_A[\rho_A^2(2)]}, \frac{nt/2}{\rho_o^3\log n}\right\}\right).$$

Consequently,

$$\begin{aligned}\mathbb{P}(Z > t) &\leq \kappa^\star(\rho,\lambda)\exp\left(-\frac{\mathbb{E}_A[\rho_A(2)]}{\kappa}\min\left\{\frac{(nt/2)^2}{n\mathbb{E}_A[\rho_A^2(2)]}, \frac{nt/2}{\rho_o^3\log n}\right\}\right)\\ &\leq \kappa^\star(\rho,\lambda)\exp\left(-\frac{\mathbb{E}_A[\rho_A(2)]}{\kappa}\frac{n\min\{t,t^2\}}{4\log n}\min\left\{\frac{1}{\mathbb{E}_A[\rho_A^2(2)]}, \frac{1}{\rho_o^3}\right\}\right).\end{aligned}$$

Let

$$\kappa(\rho,\lambda) := \frac{\mathbb{E}_A[\rho_A(2)]}{4\kappa}\min\left\{\frac{1}{\mathbb{E}_A[\rho_A^2(2)]}, \frac{1}{\rho_o^3}\right\}.$$

It now follows that, for all $t > 0$

$$\mathbb{P}(Z > t) \leq \kappa^\star(\rho,\lambda)\exp\left(-\frac{\kappa(\rho,\lambda)n\min\{t,t^2\}}{\log n}\right).$$

A.8 Proof of Lemma 9

Observe from part (ii) of Theorem 4 Bertail and Portier (2019) that under Assumption 1, the Rademacher complexity $\mathcal{R}(\mathcal{F}_{[0,1]\bigcap\mathbb{Q}}))$ (as defined in definition 7 Bertail and Portier (2019)) for any class of VC functions with constant envelope U and characteristic (κ_1, v) can be upper bounded as

$$\mathcal{R}(\mathcal{F}_{[0,1]\bigcap\mathbb{Q}})) \leq \kappa \left[v\kappa_\rho U \log \frac{\kappa_1\kappa_\rho U}{\sigma'} + \sqrt{vn\sigma' \log \frac{\kappa_1\kappa_\rho U}{\sigma'}} \right] + nU \exp(-\kappa_\rho\lambda/2)\kappa_\lambda, \tag{A.21}$$

where $(\sigma')^2$ is any number such that

$$\sup_{f\in\mathcal{F}_{[0,1]\bigcap\mathbb{Q}}} \mathbb{E}_A \left[\left(\sum_{i=1}^{\rho_A(2)} f(Y_i) \right)^2 \right] \leq (\sigma')^2$$

and κ_ρ is any number such that $0 < \sigma' < \kappa_\rho U$. To continue with the proof, we first write the following lemma. Its proof is provided in §A.9 for completeness.

Lemma 10. *$\mathcal{F}_{[0,1]\bigcap\mathbb{Q}}$ is VC with constant envelope 1 and admissible charactaristic $(\kappa, 2)$ for some universal constant $\kappa > 4e$.*

Recall from Lemma 10 that the class of all half intervals on rationals $\mathcal{F}_{[0,1]\bigcap\mathbb{Q}}$ are VC with a constant envelope U and characteristic $(\kappa, 2)$ for some universal constant κ. Substituting this in eq. (A.21), we get

$$\mathcal{R}(\mathcal{F}_{[0,1]\bigcap\mathbb{Q}})) \leq \kappa \left[2\kappa_\rho \log \frac{\kappa_\rho}{\sigma'} + \sqrt{2n\sigma' \log \frac{\kappa_\rho}{\sigma'}} \right] + n \exp(-\kappa_\rho\lambda/2)\kappa_\lambda.$$

Next, we observe that $f(\cdot)$ are indicators of half-intervals. Hence $f(\cdot) \leq 1$ and

$$\left(\sum_{i=1}^{\rho_A(2)} f(Y_i) \right)^2 \leq \rho_A^2(2).$$

Therefore, choosing $(\sigma')^2 = \mathbb{E}_A[\rho_A^2(2)]$ suffices and we get

$$\mathcal{R}(\mathcal{F}_{[0,1]\bigcap\mathbb{Q}})) \leq \kappa \left[2\kappa_\rho \log \frac{\kappa_\rho}{\sqrt{\mathbb{E}_A[\rho_A^2(2)]}} + \sqrt{2n\sqrt{\mathbb{E}_A[\rho_A^2(2)]} \log \frac{\kappa_\rho}{\sqrt{\mathbb{E}_A[\rho_A^2(2)]}}} \right] + n \exp(-\kappa_\rho\lambda/2)\kappa_\lambda.$$

Finally, substituting this into Theorem 5 Bertail and Portier (2019) and trivially substituting $\log(x\kappa) \leq \kappa \log(x)$ for all large enough constant κ, we arrive at the required bound

$$\mathcal{R}(\mathcal{F}_{[0,1]\bigcap\mathbb{Q}})) = 2(\mathbb{E}_A[\rho_A(2)] + \mathbb{E}_\nu[\rho_A(2)]) + \kappa \left[\kappa_\rho \log \left(\frac{\kappa_\rho}{\sqrt{\mathbb{E}_A[\rho_A^2(2)]}} \right) + \sqrt{n\mathbb{E}_A[\rho_A^2(2)] \log \left(\frac{\kappa_\rho}{\sqrt{\mathbb{E}_A[\rho_A^2(2)]}} \right)} \right]$$
$$+ n \exp\left(-\kappa_\rho\lambda/2\right) \kappa_\lambda.$$

Now, using the exponential tail bound for the suprema of additive functions of regenerative Markov chains (Theorem 6 in Bertail and Portier (2019), or Theorem 7 in Adamczak (2008)), we arrive at the conclusion.

A.9 Proof of Lemma 10

To prove this lemma, we introduce the notion of VC classes which are commonly used in nonparametric statistics (Sen, 2018). The function H is an envelope for the function class $\mathcal{F}$ with metric d if $|f(x)| \leq H(x)$ for all $x \in E$ and $f \in \mathcal{F}$. For a metric space $(\mathcal{F}, d)$, the covering number $\mathcal{N}(\varepsilon, \mathcal{F}, d)$ is the minimal number of balls of size ε needed to cover $\mathcal{F}$. The metric that we use here is the $L_2(Q)$-norm denoted by $\|.\|_{L_2(Q)}$ and given by $\|f\|_{L_2(Q)} = \left(\int f^2 dQ\right)^{1/2}$.

Definition 3. *A countable class $\mathcal{F}$ of measurable functions on $E \to \mathbb{R}$ is said to be of VC-type (or Vapnik-Chervonenkis type) for an envelope H and admissible characteristic (κ, v) (positive constants) such that $\kappa \geq (3\sqrt{e})^v$ and $v \geq 1$, if for all probability measure Q on $(E, \mathcal{E})$ with $0 < \|H\|_{L_2(Q)} < \infty$ and every $0 < \varepsilon < 1$,*

$$\mathcal{N}\left(\varepsilon\|H\|_{L_2(Q)}, \mathcal{F}, \|.\|_{L_2(Q)}\right) \leq \kappa\varepsilon^{-v}.$$

Proof. We provide a proof for completeness. We begin this proof with some requisite definitions. Given a class of indicator functions $\mathcal{I}$ defined on χ, and a set $\{x_1, \ldots, x_n\} \in \chi^n$, we first define

$$\mathcal{I}(\{x_1, \ldots, x_n\}) := \{(f(x_1), \ldots, f(x_n)) \in \{0,1\}^n : f \in \mathcal{I}\}$$

The *growth function* of the $\mathcal{F}$ is then defined as

$$\Delta_n(\mathcal{I}) = \max_{\{x_1,\ldots,x_n\}\in\chi^n} |\mathcal{I}(\{x_1, \ldots, x_n\})|$$

The *VC-dimension* of $\mathcal{I}$ is then defined as

$$VC(\mathcal{I}) := \text{argmax}_n \{n : \Delta_n(\mathcal{I}) = 2^n\}$$

We will now show that $VC(\mathcal{F}_{[0,1]\bigcap\mathbb{Q}}) = 1$. Let $\{x_1, \ldots, x_n\}$ be any ordered sample. That is, $x_1 < x_2 <, \ldots, < x_n$. For any $t \in [0,1]\bigcap\mathbb{Q}$, observe that $(\mathbb{1}[x_1 < t], \mathbb{1}[x_2 < t], \ldots, \mathbb{1}[x_n < t])$ has the form $(1, 1, 1, \ldots, 1, 0, 0 \ldots, 0)$. In particular, the values of $\mathcal{F}_{[0,1]\bigcap\mathbb{Q}}(\{x_1, \ldots, x_n\})$ has to be within the following set

$$\begin{aligned}
&(0, 0, 0, \ldots, 0),\\
&(1, 0, 0, \ldots, 0),\\
&(1, 1, 0, \ldots, 0),\\
&\vdots\\
&(1, 1, 1, \ldots, 1)
\end{aligned}$$

Therefore $\Delta_n(\mathcal{F}_{[0,1]\bigcap\mathbb{Q}}) = n + 1$. This implies that

$$\begin{aligned}
VC(\mathcal{F}_{[0,1]\bigcap\mathbb{Q}}) &= \text{argmax}_n\{\Delta_n(\mathcal{F}_{[0,1]\bigcap\mathbb{Q}}) = 2^n\}\\
&= \text{argmax}_n\{n + 1 = 2^n\}\\
&= 1.
\end{aligned}$$

Now, using standard results of covering number bounds, (Theorem 7.8 of Sen (2018), see also Theorem 2.6.4 Van der Vaart (2000)) we have the following result. For some universal constant $\kappa > 0$

$$\begin{aligned}
\mathcal{N}\left(\varepsilon\|H\|_{L_2(Q)}, \mathcal{F}_{[0,1]\bigcap\mathbb{Q}}, \|.\|_{L_2(Q)}\right) &\leq \kappa \times VC(\mathcal{F}_{[0,1]\bigcap\mathbb{Q}})(4e)^{VC(\mathcal{F}_{[0,1]\bigcap\mathbb{Q}})}\left(\frac{1}{\varepsilon}\right)^{2VC(F_{[0,1]\bigcap\mathbb{Q}})}\\
&\overset{(i)}{\leq} \frac{\kappa'}{\varepsilon^2}.
\end{aligned}$$

where (i) follows by substituting $VC(\mathcal{F}_{[0,1]\bigcap\mathbb{Q}})$. This completes the proof. □

A.10 Proof of Proposition 1

We first explain the logical construction leading to the optimization model (5.1). We use the index $i = 1, \ldots, n$ to represent the time points and $l = 1, \ldots, L+1$ to represent the segment of the time points (i.e. l represents the segment of time points between τ_{l-1} and τ_l).

Instead of determining the positions of change points directly, we introduce binary decision variables to represent whether the time point i belongs to the segment l:

$$z_{i,l} \in \{0,1\} \quad l = 1, \ldots, L+1, i = 1, \ldots, n.$$

Based on the property that a time point can be assigned to only one segment, we have the constraints:

$$\sum_{l=1}^{L+1} z_{i,l} = 1 \quad i = 1, \ldots, n. \tag{A.22}$$

Since the length of a segment is assumed to be at least 3, we have the constraints:

$$\sum_{i=1}^{n} z_{i,l} \geq 3 \quad l = 1, \ldots, L+1. \tag{A.23}$$

Also note that if the time point i is assigned to the segment l, then the time point $i+1$ must either remain in the segment l or a later segment $l' > l$. This is represented by the constraints:

$$z_{i,l} \leq \sum_{l' \geq l} z_{i+1,l'} \quad l = 1, \ldots, L+1, i = 1, \ldots, n-1. \tag{A.24}$$

Proof. Assume $z^*_{i,l}$ is an optimal solution of (5.1). We now use $z^*_{i,l}$ to represent each part of (3.2) as follows:

$$\sum_{p=\tau'_{l-1}}^{\tau'_l} \mathbb{1}[X_p \leq u] := \sum_{i=1}^{n} a_{u,i} z^*_{i,l}$$

$$\tau'_l - \tau'_{l-1} := \sum_{i=1}^{n} z^*_{i,l}$$

$$\hat{F}^{\tau'_l}_{\tau'_{l-1}} := \frac{\sum_{i=1}^{n} a_{u,i} z^*_{i,l}}{\sum_{i=1}^{n} z^*_{i,l}}.$$

Then we have:

$$R_n(\tau'_1, \ldots, \tau'_L) := \sum_{l=1}^{L+1} \sum_{u=1}^{n} \left(\sum_{i=1}^{n} a_{i,u} z^*_{i,l} \right) \left(1 - \frac{\sum_{i=1}^{n} a_{i,u} z^*_{i,l}}{\sum_{i=1}^{n} z^*_{i,l}} \right),$$

where $(\tau'_1, \ldots, \tau'_L)$ are the estimated change points. Therefore, we conclude that τ' is a solution of (3.2) if z^* is a solution of (5.1) and the objective value of (5.1) is the same as (3.2).

Similarly, if $(\tau'_1, \ldots, \tau'_L)$ are the estimated change points, we can use $\tau'_l - \tau'_{l-1} = \sum_{i=1}^{n} z^*_{i,l}$ to represent the length of the segments for each l. The segmentation implies three properties: 1) each time point can be assigned to one segment; 2) the length of a segment $(\tau'_l - \tau'_{l-1})$ is at least 3 ; and 3) if a time point belongs to a given segment, the next time point cannot be assigned to an early segment. Then if $(\tau'_1, \ldots, \tau'_L)$ are the estimated change points, z^* should satisfy the constraints (A.22), (A.23) and (A.24), which means that z^* is a solution of (5.1), and (3.2) has the same value as the objective function of (5.1). □

A.11 Proof of Proposition 2

Proof. Suppose z^* solves (5.1). Following Equation R1 in Borrero et al. (2016), we linearize the fractional terms by defining

$$k^*_l = \frac{1}{\sum_{i=1}^{n} z^*_{i,l}}, \qquad \text{so that} \quad \sum_{i=1}^{n} k^*_l z^*_{i,l} = 1.$$

Then

$$\begin{aligned} f^*_{u,l} &= \frac{\sum_{i=1}^{n} a_{i,u} z^*_{i,l}}{\sum_{i=1}^{n} z^*_{i,l}} = \sum_{i=1}^{n} a_{i,u} k^*_l z^*_{i,l}, \\ d^*_{u,l} &= 1 - f^*_{u,l}, \end{aligned} \tag{A.25}$$

and hence

$$s^*_{u,l} = \left(\sum_{i=1}^{n} a_{i,u} z^*_{i,l}\right)\left(1 - \frac{\sum_{i=1}^{n} a_{i,u} z^*_{i,l}}{\sum_{i=1}^{n} z^*_{i,l}}\right) = \sum_{i=1}^{n} a_{i,u} z^*_{i,l} d^*_{u,l}.$$

Thus $(z^*, k^*, f^*, d^*, s^*)$ is feasible for (5.2) with the same objective value.

Conversely, if $(z^*, k^*, f^*, d^*, s^*)$ is feasible for (5.2), then $z^* \in \mathcal{Z}$ and satisfies all constraints of (5.1). Moreover,

$$\sum_{l=1}^{L+1}\sum_{u=1}^{n} s^*_{u,l} = \sum_{l=1}^{L+1}\sum_{u=1}^{n}\left(\sum_{i=1}^{n} a_{i,u} z^*_{i,l}\right)\left(1 - \frac{\sum_{i=1}^{n} a_{i,u} z^*_{i,l}}{\sum_{i=1}^{n} z^*_{i,l}}\right),$$

which coincides with the objective of (5.1). Therefore, z^* solves (5.1) if and only if $(z^*, k^*, f^*, d^*, s^*)$ solves (5.2), with equal objective values. This completes the proof. □