Conformal changepoint localization

Rohan Hore*[1] and Aaditya Ramdas[1]

[1]Department of Statistics and Data Science, Carnegie Mellon University

December 2, 2025

Abstract

We study the problem of offline changepoint localization in a distribution-free setting. One observes a vector of data with a single changepoint, assuming that the data before and after the changepoint are i.i.d. (or more generally exchangeable) from arbitrary and unknown distributions. The goal is to produce a finite-sample confidence set for the index at which the change occurs without making any other assumptions. Existing methods often rely on parametric assumptions, tail conditions, or asymptotic approximations, or only produce point estimates. In contrast, our distribution-free algorithm, CONformal CHangepoint localization (CONCH), only leverages exchangeability arguments to construct confidence sets with finite sample coverage. By proving a *conformal Neyman–Pearson lemma*, we derive principled score functions that yield informative (small) sets. Moreover, with such score functions, the normalized length of the confidence set shrinks to zero under weak assumptions. We also establish a universality result showing that any distribution-free changepoint localization method must be an instance of CONCH. Experiments suggest that CONCH delivers precise confidence sets even in challenging settings involving images or text.

1 Introduction

In this paper, we study the problem of offline changepoint localization, where we are given an ordered sequence of data and are told that the underlying data-generating distribution has changed at some unknown index, called the *changepoint*. Unless explicitly mentioned, in this work, we assume that there is a single changepoint. As a simple illustration, suppose that the data are drawn independently from some distribution P_0 before the changepoint and from a different distribution $P_1 \neq P_0$ thereafter. The objective is to localize the changepoint, i.e., give a confidence set that contains this changepoint with high probability.

Changepoint localization is substantially more challenging than the related task of changepoint detection: merely identifying whether a change has occurred. Yet in domains such as operations

[*] Corresponding author: rhore@andrew.cmu.edu

engineering, econometrics, and biostatistics, the ability to retrospectively pinpoint the time of distributional change is often critical. Consider, for instance, a manufacturing context: quality measurements of a component may remain stable until a machine begins to malfunction, after which the measurements exhibit a systematic shift. Once the production batch has concluded, it becomes essential to determine when this shift first arose in order to diagnose the source of the malfunction and implement corrective measures.

1.1 Existing approaches

Given its wide practical relevance, offline changepoint analysis has been extensively studied; see Truong et al. [2020], Duggins [2010] for surveys. Classical methods such as CUSUM [Page, 1955] and conformal martingales [Vovk et al., 2003] primarily address the online detection problem rather than retrospective localization.

Likelihood-based procedures assume specific parametric models (e.g., Gaussian mean-shift, linear regression) [Kim and Siegmund, 1989, Quandt, 1958, Gurevich and Vexler, 2006] and mostly focus on detection. More recent post-detection localization techniques [Saha and Ramdas, 2025] still rely on restrictive model assumptions, such as known and non-overlapping pre-change and post-change families.

Several nonparametric methods achieve localization only asymptotically, including SMUCE [Frick et al., 2014], regression-based approaches [Xu et al., 2024], and Gaussian mean-shift intervals [Fotopoulos et al., 2010], among others [Bhattacharyya and Johnson, 1968, Zou et al., 2007]. The construction in Verzelen et al. [2023] attains theoretical optimality but involves non-computable constants, limiting practical use.

Bootstrap-based approaches [Cho and Kirch, 2022] target mean shifts but lack finite-sample validity and are computationally intensive. Rank-based nonparametric tests [Pettitt, 1979, Ross and Adams, 2012] are distribution-free for detection but do not provide confidence sets for localization and often have low power without additional structure. Multi-changepoint algorithms [Anastasiou and Fryzlewicz, 2022, Truong et al., 2020] typically adopt "isolate-detect" strategies and return only point estimates.

Conformal martingale methods [Vovk et al., 2003, Volkhonskiy et al., 2017, Vovk, 2021, Vovk et al., 2021, Nouretdinov et al., 2021, Shin et al., 2023] provide powerful tools for online detection but do not yield confidence sets for localization. Recently, MCP localization [Dandapanthula and Ramdas, 2025] introduced the first 'truly' distribution-free approach to changepoint localization using a matrix of conformal p-values. However, it often produces wider confidence intervals than appear to be necessary, motivating the need for sharper, yet valid, distribution-free alternatives.

Overall, existing approaches are constrained by model assumptions and either focus on point estimation rather than localization or trade statistical efficiency for distribution-free validity. In this work, we close this gap by proposing a simple yet principled framework for changepoint localization that is distribution-free, finite-sample valid, and yields informative confidence sets. The formal objective of distribution-free confidence sets is introduced in Section 2.

1.2 Our contributions

The main contributions of this work are summarized below:

- We introduce CONCH (CONformal CHangepoint localization) in Algorithm 1, a novel framework that, given any changepoint plausibility score S and a confidence level $1 - \alpha$, produces a finite-sample, distribution-free confidence set for the true changepoint without making any restrictive assumptions on the pre- and post-change distributions.
- While our framework is valid for any choice of score function, offering great flexibility to the user, its statistical performance can be substantially improved by employing scores tailored to the problem at hand. We derive an expression of optimal score function based on a novel "Conformal Neyman–Pearson" lemma (Lemma 4.2), which may be of independent interest.
- While the optimal score requires oracle knowledge, we propose practical "near-optimal" alternatives learnable from data that yield narrow confidence sets.
- We further show that, under mild regularity conditions, the normalized length of the confidence set converges to zero as the sample size grows (Theorem 5.2). In addition, when the true likelihood ratio between the pre- and post-change distributions is available, the confidence set is stochastically bounded, i.e., has length $\mathcal{O}_P(1)$ (Theorem 5.1).
- Next, we show that CONCH has a universality property (Theorem 6.1): any distribution-free confidence set for the changepoint is an instance of our framework. This further enables us to give a simple calibration technique (Algorithm 2) that turns any heuristic confidence set into a distribution-free valid confidence set for changepoint.
- Algorithm 3 extends the general CONCH framework to multiple-changepoint localization, broadening the applicability of our framework.
- We illustrate the practical effectiveness of CONCH across a range of synthetic and real-world datasets, covering both image and text domains. In all cases, the resulting CONCH confidence sets concentrate tightly around the true changepoint. Notably, our framework can wrap around any black-box classifier trained to distinguish pre- and post-change samples, yielding informative confidence sets even under subtle distributional shifts.

Organization of the paper. The remainder of the paper is organized as follows. Section 2 formally defines the problem of distribution-free changepoint localization. Section 3 introduces our general framework, CONCH, and formally presents the algorithm. Section 4 develops an optimal score and provides guidance on selecting practical near-optimal choices that yield narrow confidence sets. Section 5 shows that, under mild regularity conditions, these confidence sets indeed shrink, enabling sharp localization. Section 6 establishes a universality result for CONCH. Section 6.1 builds on this foundation to introduce a calibration procedure that turns any localization method into a valid distribution-free one. Section 7 extends the CONCH framework to the multiple-changepoint setting by wrapping it around any 'nice' segmentation algorithm. Section 8 presents evaluations on synthetic and real-world data, demonstrating the practical utility of our framework.

2 Distribution-free changepoint localization

In this section, we formally describe the problem of distribution-free offline changepoint localization and introduce the necessary notation. Throughout, $\mathbb{N}$ denotes the natural numbers and, for $K \in \mathbb{N}$, we write $[K] := 1, \ldots, K$. For any set S, let $\mathcal{M}(S)$ denote the collection of probability measures on S, and let 2^S denote its power set. We use $\overset{d}{=}$ to denote equality in distribution.

With this notation in place, consider an ordered sequence of $\mathcal{X}$-valued random variables $\mathbf{X} = (X_1, \ldots, X_n)$ for some $n \in \mathbb{N}$. Likewise, we use $\mathbf{x} = (x_1, \ldots, x_n)$ to denote a generic element of $\mathcal{X}^n$. We assume that there exists an unknown changepoint $\xi \in [n-1]$ such that

$$(X_1, \ldots, X_\xi) \sim \mathcal{P}_{0,\xi}, \qquad (X_{\xi+1}, \ldots, X_n) \sim \mathcal{P}_{1,\xi},$$

where $\mathcal{P}_{0,\xi} \in \mathcal{M}(\mathcal{X}^\xi)$ and $\mathcal{P}_{1,\xi} \in \mathcal{M}(\mathcal{X}^{n-\xi})$ denote the pre-change and post-change distributions, respectively. We write the joint distribution as $\mathcal{P} = \mathcal{P}_{0,\xi} \times \mathcal{P}_{1,\xi}$. In line with the distribution-free perspective, we impose no structural assumptions on $\mathcal{P}_{0,\xi}$ or $\mathcal{P}_{1,\xi}$ beyond the following.

Assumption 1. $\mathcal{P}_{0,\xi}$ and $\mathcal{P}_{1,\xi}$ are exchangeable. Specifically, for any permutations $\pi_L : [\xi] \to [\xi]$ and $\pi_R : [n] \setminus [\xi] \to [n] \setminus [\xi]$, it holds that the pre-change and post-change segments are independent, i.e., $\mathcal{P}_{0,\xi} \perp \mathcal{P}_{1,\xi}$ and that

$$(X_1, \ldots, X_\xi) \overset{d}{=} (X_{\pi_L(1)}, \ldots, X_{\pi_L(\xi)}), \quad (X_{\xi+1}, \ldots, X_n) \overset{d}{=} (X_{\pi_R(\xi+1)}, \ldots, X_{\pi_R(n)}).$$

In words, Assumption 1 requires that the distribution of $\mathbf{X}$ is invariant under arbitrary permutations of the entries to the left of ξ and, independently, under permutations of those to its right. A canonical example, mentioned in the introduction, is the i.i.d. changepoint model: the *pre-change* observations $(X_1, \ldots, X_\xi)$ are i.i.d. from some P_0, and independently, the *post-change* observations $(X_{\xi+1}, \ldots, X_n)$ are i.i.d. from some P_1.

For any $t \in [n-1]$, let $\mathcal{H}_{0,t}$ denote the hypothesis that t is the true changepoint and that the distributions $\mathcal{P}_{0,t}$ and $\mathcal{P}_{1,t}$ satisfy Assumption 1. We write $\mathbb{P}_t$ and $\mathbb{E}_t$ to denote probability and expectation, respectively, under this model class. We can now formally define what it means to construct a distribution-free confidence set for the changepoint.

Definition 1. Fix $\alpha \in (0,1)$. A mapping $\mathcal{C}_{1-\alpha} : \mathcal{X}^n \to 2^{[n-1]}$ is called a *distribution-free confidence set for changepoint* at level $1-\alpha$ if $\mathbb{P}_\xi(\,\xi \in \mathcal{C}_{1-\alpha}(\mathbf{X})\,) \;\geq\; 1-\alpha$.

Assumption 1 is considerably weaker than the working assumptions underlying most existing changepoint localization methods reviewed in Section 1.1. Prior approaches typically rely on strong parametric models or asymptotic approximations, in contrast to the minimal nature of our assumption. While some recent methods [Dandapanthula and Ramdas, 2025] offer distribution-free guarantees under similarly mild conditions, they generally yield more diffuse confidence sets. Our approach instead enables sharper localization while retaining finite-sample validity, making it a significant contribution in this direction. The next section formally introduces our method.

Algorithm 1: CONCH: conformal changepoint localization algorithm

```
Input: $(X_t)_{t=1}^n$ (data), $1-\alpha$ (target coverage), $S : \mathcal{X}^n \to \mathbb{R}^{n-1}$ (CPP score)
Output: $\mathcal{C}_{1-\alpha}^{\text{CONCH}}$ (CONCH confidence set at level $1-\alpha$)
for $t \in [n-1]$ do
    $\Pi_t \leftarrow \{\pi \in \mathcal{S}_n : \text{for all } i \le t,\ \pi(i) \le t \ \text{ and for all } i > t,\ \pi(i) > t\}$;
    foreach $\pi \in \Pi_t$ do
        Evaluate $S_t(\pi(\mathbf{X}))$;
    end
    $p_t \leftarrow \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\{S_t(\pi(\mathbf{X})) \le S_t(\mathbf{X})\}$;
end
$\mathcal{C}_{1-\alpha}^{\text{CONCH}} \leftarrow \{\, t \in [n-1] : p_t > \alpha \,\}$;
return $\mathcal{C}_{1-\alpha}^{\text{CONCH}}$
```

3 Conformal changepoint localization

This section develops a conformal framework for localizing the true changepoint. Conformal p-values, originally introduced by [Vovk et al., 1999, Shafer and Vovk, 2008] in the context of distribution-free predictive inference, have since been extended to a wide range of problems including outlier detection [Bates et al., 2023], post-prediction screening [Jin and Candès, 2023], and conditional two-sample testing [Wu et al., 2024], among others. Motivated by these developments, we adapt conformal p-values to our problem in an efficient manner, yielding a general framework to construct confidence sets that satisfy (1). We call this the CONformal CHangepoint localization (CONCH) algorithm, formally given in Algorithm 1. Our framework relies on two key components:

- **ChangePoint Plausibility (CPP) score:** We call any mapping $S : \mathcal{X}^n \to \mathbb{R}^{n-1}$ a changepoint plausibility score. Intuitively, for each candidate index $t \in [n-1]$, $S_t = (S(\cdots))_t$ assigns a score to quantify the chance that t is indeed a changepoint; a larger S_t indicates a stronger plausibility of t being a changepoint.
- **Split-permutation group:** For any $t \in [n-1]$, we define the reduced set of permutations

$$\Pi_t := \Big\{\pi \in \mathcal{S}_n : \pi(i) \le t \ \text{ for all } i \le t, \ \ \pi(i) > t \ \text{ for all } i > t\Big\}. \tag{3.1}$$

Any $\pi \in \Pi_t$ permutes indices to the left and right of t, without mixing indices across t.

Note that, if t is indeed the true changepoint, elements of Π_t preserve the pre- and post-change exchangeability. The validity of our framework crucially relies on this observation. More precisely, starting from any user-specified CPP score S, we define a conformal p-value p_t for each index $t \in [n-1]$ by looking at the normalized rank of the true score $S_t(\mathbf{X})$ within the set of all permuted scores, $\{S_t(\pi(\mathbf{X})) : \pi \in \Pi_t\}$, i.e.,

$$p_t := \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\{S_t(\pi(\mathbf{X})) \le S_t(\mathbf{X})\}\,. \tag{3.2}$$

Intuitively, under $\mathcal{H}_{0,t}$, every permutation $\pi \in \Pi_t$ is equally likely and therefore, p_t is super-uniform under the null $\mathcal{H}_{0,t}$, a result we formally establish in Theorem 3.1. For brevity, the proof is deferred to Appendix B.1. Finally, the changepoint confidence set is then constructed by thresholding these p-values at a pre-specified level α:

$$\mathcal{C}_{1-\alpha}^{\text{CONCH}} := \{t \in [n-1] : p_t > \alpha\}.$$

Theorem 3.1. *For each $t \in [n]$, p_t in (3.2) is a valid p-value under $\mathcal{H}_{0,t}$, i.e., for any $\alpha \in (0,1)$, $\mathbb{P}_\xi\left(p_\xi \leq \alpha\right) \leq \alpha$. Consequently, $\mathcal{C}_{1-\alpha}^{\text{CONCH}}$ is a distribution-free confidence set for changepoint.*

We highlight that the CONCH algorithm does not impose any restriction on the choice of CPP score, thereby providing significant flexibility for users to design their own plausibility measure. In particular, the score function may depend non-trivially on the multiset $\{X_1, \ldots, X_n\}$. For readers familiar with the distinction between full and split conformal methods in the setting of predictive inference, this corresponds to an adaptation of the full conformal approach to our problem.

However, the choice of score function is closely tied to the power of CONCH, or equivalently to the length of the resulting confidence sets. In the following section, we discuss how to select score functions that yield narrow and informative sets. Before doing so, we conclude this section with a few remarks on the practical implementation of the framework.

Remark 3.1 (Monte-Carlo p-values)**.** Note that the p-value p_t in (3.2) is defined using the full permutation set Π_t. Therefore, computing all permuted scores becomes prohibitively expensive when the sample size is large. A practical remedy is to use a Monte-Carlo approximation of p_t, formally described in (A.1). This approximation substantially reduces the computational burden while maintaining the same validity guarantees as in Theorem 3.1; see Theorem A.1 for details.

Remark 3.2 (Exact validity)**.** Theorem 3.1 ensures that the CONCH p-value in (3.2) is super-uniform, which implies that the resulting confidence set $\mathcal{C}_{1-\alpha}^{\text{CONCH}}$ may be conservative in finite samples. Nonetheless, in our experiments (Section 8), we find that these p-values still lead to sharp localization across all empirical settings considered. Moreover, employing a randomized approximation of p_t, as in (A.2), yields an exactly uniform p-value, thereby providing exact finite-sample validity (Theorem A.2) for the resulting confidence set.

Remark 3.3 (Time-reversal symmetry)**.** The CONCH confidence set is invariant under the reversal of the timeline. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be the reversal of $\mathbf{X} = (X_1, \ldots, X_n)$, i.e., $Y_i := X_{n-i+1}$. Localizing $\xi \in [n]$ given $\mathbf{X}$ is then equivalent to localizing $n - \xi$ given $\mathbf{Y}$. Indeed, the permutation group Π_t acting on $\mathbf{X}$ corresponds to Π_{n-t} acting on $\mathbf{Y}$, and with the score functions defined accordingly, the CONCH confidence set computed from $\mathbf{Y}$ is exactly the image of the CONCH confidence set from $\mathbf{X}$ under the map $t \mapsto n - t$.

4 Guidelines for choosing the CPP score

The CONCH confidence sets introduced earlier remain valid for any choice of CPP score, offering substantial flexibility. However, a well-chosen score yields a narrower and more informative set. In this section, we first establish general properties of CPP scores and then derive an optimal score. Although this optimal score requires oracle knowledge, we propose practical alternatives that closely approximate it. Proofs of the results in this section are deferred to Appendix B.2.

Proposition 4.1. *Fix $n \in \mathbb{N}$ and $\alpha \in (0,1)$.*

(i) **(Symmetry yields trivial p-values).** *Fix $t \in [n-1]$. If S is t-symmetric, i.e., satisfies $S_t(\cdot) = S_t(\pi(\cdot))$ for all $\pi \in \Pi_t$, then p-value p_t in (3.2) equals 1, and $\mathbb{P}(t \in \mathcal{C}^{\mathrm{CONCH}}_{1-\alpha}) = 1$.*

(ii) ***(Conformal data-processing inequality).*** *Let C_1 be the CONCH set based on score S, with p-values $p_{t,1}$. For any non-decreasing $f : \mathbb{R} \to \mathbb{R}$, let C_2 be the corresponding set based on $f(S)$, with p-values $p_{t,2}$. Then $p_{t,1} \le p_{t,2}$ for all $t \in [n-1]$, and therefore $C_1 \subseteq C_2$.*

Part (i) shows that t-symmetric CPP scores yield trivial conch p-values, regardless of whether $\mathcal{H}_{0,t}$ holds, and therefore lead to overly conservative confidence sets. Such scores should be avoided in practice. Part (ii) shows a monotonicity property of CONCH: applying any non-decreasing transformation to the CPP score can only enlarge the resulting set, and any *strictly* increasing transformation leaves the set unchanged. These properties guide practical choices of CPP scores that yield informative confidence sets.

For the remainder of this section, we focus on the canonical setting, namely the i.i.d. changepoint model. Specifically, let $\mathcal{P}_{\mathrm{IID}}$ denote the class of distributions for which there exists $\xi \in [n-1]$ such that

$$\mathcal{P}_{0,\xi} = \otimes_{t=1}^{\xi} P_0, \quad \mathcal{P}_{1,\xi} = \otimes_{t=\xi+1}^{n} P_1,$$

where P_0 and P_1 admit densities f_0 and f_1 with respect to a common dominating measure ν on $\mathcal{X}$.

4.1 Optimal CPP score function

In this section, we establish the optimal CPP score function, assuming the knowledge of both densities f_0, f_1, and the true changepoint ξ. By framing the task of identifying an optimal score as a testing problem with a point null and a point alternative, we can directly apply the classical Neyman–Pearson (NP) lemma. This yields a similar optimality result tailored to the setting of distribution-free changepoint localization, which we call the *second*[1] *Conformal NP Lemma*. Before stating the lemma formally, we first set up some notation.

For any $t \in [n-1]$, let $\mathcal{X}_{L,t} := \{X_1, \ldots, X_t\}$ and $\mathcal{X}_{R,t} := \{X_{t+1}, \ldots, X_n\}$ denote the (unordered) left and right multisets. Let $\mathcal{P}^{(t)}_{\mathbf{X}} = \mathcal{P}_{0,t} \times \mathcal{P}_{1,t}$ be the law of $\mathbf{X} = (X_1, \ldots, X_n)$ corresponding to a changepoint at t under the i.i.d. model class $\mathcal{P}_{\mathrm{IID}}$. We define $\mathcal{P}^{(t)}_{\mathbf{X}|\mathcal{X}_{L,t},\mathcal{X}_{R,t}}$ for the associated conditional distribution of $\mathbf{X}$ given $(\mathcal{X}_{L,t}, \mathcal{X}_{R,t})$, and write $\mathcal{H}'_t$ to hypothesize that $\mathbf{X} \mid (\mathcal{X}_{L,t}, \mathcal{X}_{R,t}) \sim$

[1] The first instance of such a Conformal NP Lemma appears in Dandapanthula and Ramdas [2025], which establishes an analogous NP optimality result for a conformal p-value-based changepoint test.

$\mathcal{P}^{(t)}_{\mathbf{X}|\mathcal{X}_{L,t},\mathcal{X}_{R,t}}$ for any $t \in [n-1]$. Suppose we want to test $\mathcal{H}'_t$ using conformal p-values. Given a score $s : \mathcal{X}^n \to \mathbb{R}$ and the permutation set Π_t, we therefore define the randomized conformal p-value

$$p_t(s) = \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\left\{s(\pi(\mathbf{X})) < s(\mathbf{X})\right\} \; + \; U \cdot \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\left\{s(\pi(\mathbf{X})) = s(\mathbf{X})\right\}, \tag{4.1}$$

where $U \sim \text{Unif}(0,1)$ is independent of $\mathbf{X}$ and the permutations. By Theorem A.2, $p_t(s)$ is Unif$[0,1]$ under $\mathcal{H}'_t$. Consequently, the test $\phi_t(\mathbf{X}; s) = \mathbb{1}\left\{p_t(s) \le \alpha)\right\}$ controls type I error at level α for the null $\mathcal{H}'_t$ with any score function s. We note that we use the randomized p-value here to utilize the entire type I budget and maximize power with an appropriate score function.

Now, we seek an optimal score $s^\star$ such that the corresponding test $\phi_t(\mathbf{X}; s^\star)$ achieves maximum power against an alternative $\mathcal{H}'_r$ (with $r \neq t$), which posits that r rather than t is the true changepoint. The second Conformal Neyman–Pearson lemma, stated below, formally establishes that the likelihood ratio $s^\star(\cdot) = \mathcal{P}^{(t)}_X(\cdot)/\mathcal{P}^{(r)}_X(\cdot)$ defines the optimal test.

Lemma 4.2 (second Conformal NP lemma)**.** *Fix $t, r \in [n-1]$ with $t \neq r$. The power, $\mathbb{E}_{\mathcal{H}'_r}[\phi_t(\mathbf{X}; s)]$, is maximized by the score function*

$$s^\star(x_1, \ldots, x_n) := \frac{\prod_{i \le t} f_0(x_i) \prod_{i > t} f_1(x_i)}{\prod_{i \le r} f_0(x_i) \prod_{i > r} f_1(x_i)}.$$

Finally, the Conformal NP lemma can be leveraged within the CONCH framework to derive the CPP score that would yield the narrowest confidence set. We observe that the conformal p-value in (3.2) must be valid under $\mathcal{H}_{0,t}$, while be sufficiently small to sharply detect the true changepoint $\xi \neq t$ under $\mathcal{H}_{0,\xi}$. Since only the t-th component of CPP score, S_t, determines p_t, the task of optimizing S_t boils down to finding the optimal test for $\mathcal{H}'_t$ v.s $\mathcal{H}'_\xi$.

We make this connection precise in the theorem below. For notational convenience, we let $C^{\text{CONCH}}_{1-\alpha}(S)$ denote the randomized CONCH set (A.3) constructed with the CPP score S.

Theorem 4.3. *Any strictly increasing transformation of the CPP score S^{OPT} defined by*

$$S^{\text{OPT}}_t(x_1, \ldots, x_n) = \frac{\prod_{i \le t} f_0(x_i) \prod_{i > t} f_1(x_i)}{\prod_{i \le \xi} f_0(x_i) \prod_{i > \xi} f_1(x_i)} \tag{4.2}$$

achieves the minimum expected length of the CONCH confidence set. In particular, for any $\xi \in [n-1]$ and score $S : \mathcal{X}^n \to \mathbb{R}^{n-1}$, $\mathbb{E}_{\mathcal{H}_{0,\xi} \cap \mathcal{P}_{\text{IID}}}\big[\,|\bar{C}^{CONCH}_{1-\alpha}(S)|\,\big] \;\ge\; \mathbb{E}_{\mathcal{H}_{0,\xi} \cap \mathcal{P}_{\text{IID}}}\big[\,|\bar{C}^{CONCH}_{1-\alpha}(S^{\text{OPT}})|\,\big]$.

The optimal CPP score function (4.2) depends on the unknown pre-change and post-change densities f_0 and f_1 as well as the true changepoint ξ, and is therefore not directly implementable in practice. In the next subsection, we propose score functions that closely mimic the optimal score, thus providing 'near-optimal' performance in practice.

4.2 Practical choices for CPP score

(1) Weighted mean difference. If P_0 and P_1 differ only by a location shift, we may take

$$S_t(x_1, \cdots, x_n) = \left| \frac{\sum_{i=1}^t w_{t,i} x_i}{\sum_{i=1}^t w_{t,i}} - \frac{\sum_{i>t}^n w_{t,i} x_i}{\sum_{i>t}^n w_{t,i}} \right|. \tag{4.3}$$

The weights $\{w_{t,i}\}$ are introduced to break the t-wise symmetry property, and therefore to avoid trivial confidence sets. Intuitively, observations closer to the t-th index should receive more weight when defining the score at t. Reasonable choices for weights include: $w_{t,i} = 1 - (|i-t|/n)$ or $w_{t,i} = \exp(-|i-t|/n)$. If $t \in [n-1]$ is believed to be a changepoint, the weighted means on the left and right sides should differ substantially, producing a high CPP score at t as required.

(2) Oracle log likelihood-ratio (LLR). Suppose f_0 and f_1 are known. Then, the denominator of the optimal CPP score function in (4.2) can be approximated by evaluating the complete likelihood at MLE $\hat{t}$ instead of the true changepoint ξ. This yields the CPP score given by

$$S_t(x_1, \cdots, x_n) = \log \left(\frac{\prod_{i \le t} f_0(x_i) \prod_{i>t} f_1(x_i)}{\prod_{i \le \hat{\xi}(\mathbf{x})} f_0(x_i) \prod_{i > \hat{\xi}(\mathbf{x})} f_1(x_i)} \right), \tag{4.4}$$

where

$$\hat{\xi}_{\mathrm{OR}}(\mathbf{x}) \in \operatorname*{argmax}_{s \in [n-1]} \log\Big(\prod_{i \le s} f_0(x_i) \prod_{i>s} f_1(x_i)\Big) \tag{4.5}$$

is the MLE[2] of the changepoint. If $t \in [n-1]$ is indeed the changepoint, then $\hat{\xi} \approx t$ and S_t will be large, indicating strong plausibility for a change. Since this score closely approximates (4.2), it is expected to sharply localize the changepoint, as verified in experiments.

(3) Learned LLR. When f_0 and f_1 are unknown, for each $t \in [n-1]$, one can plug in estimates (parametric or non-parametric) $\hat{f}_{t,0}$ and $\hat{f}_{t,1}$, and instead consider the CPP score given by

$$S_t(x_1, \cdots, x_n) = \log \left(\frac{\prod_{i \le t} \hat{f}_{t,0}(x_i) \prod_{i>t} \hat{f}_{t,1}(x_i)}{\prod_{i \le \hat{\xi}(\mathbf{x})} \hat{f}_{\hat{\xi}(\mathbf{x}),0}(x_i) \prod_{i > \hat{\xi}(\mathbf{x})} \hat{f}_{\hat{\xi}(\mathbf{x}),1}(x_i)} \right) \tag{4.6}$$

with $\hat{\xi}(\mathbf{x}) \in \operatorname{argmax}_{s \in [n-1]} \log(\prod_{i \le s} \hat{f}_{s,0}(x_i) \prod_{i>s} \hat{f}_{s,1}(x_i))$ being the corresponding MLE.

(4) Classifier based LLR. Instead of estimating the densities f_0 and f_1 directly, one can train a binary classifier $\hat{g}$ to distinguish post-change from pre-change samples (labeled $Y = 1$ and $Y = 0$, respectively). By Bayes' rule, we have $\log(f_1(x)/f_0(x)) = \log(\mathbb{P}(Y = 1|X = x)/\mathbb{P}(Y = 0 \mid X = x)) - \log(\pi_1/\pi_0)$, where π_1 and π_0 are class priors. If $\hat{g}$ is trained on balanced data and we write

[2] Note that the MLE estimator $\hat{\xi}_{\mathrm{OR}}(\mathbf{x})$ is a function of the observed data $(x_1, \ldots, x_n)$. Thus, for each permutation π, computing the permuted score $S_t(\pi(\mathbf{X}))$ requires first computing $\hat{\xi}_{\mathrm{OR}}(\pi(\mathbf{X}))$, MLE estimate on the permuted data.

$\hat{g}(x) \in (0,1)$ to denote the predicted probability of post-change membership, then

$$\log \frac{f_1(x)}{f_0(x)} \approx \operatorname{logit} \hat{g}(x) := \log \frac{\hat{g}(x)}{1-\hat{g}(x)}.$$

The log odds components in (4.4) can then be approximated by the classifier logits to define a practically implementable CPP score. While the choice of classifiers does not affect the validity of our method, a well-trained classifier improves power.

5 Asymptotic sharpness of CONCH confidence sets

In this section, we establish that as the sample size grows, the CONCH confidence sets narrow down to a small neighborhood around the true changepoint ξ_n, enabling sharp localization. We consider the i.i.d. changepoint model: for each n, we observe $\mathcal{D}_n = (X_{1,n}, \ldots, X_{n,n}) \in \mathcal{X}^n$ with a single changepoint at $\xi_n \in [n-1]$, and

$$(X_{1,n}, \ldots, X_{\xi_n,n}) \sim \otimes_{t=1}^{\xi_n} P_0, \qquad (X_{\xi_n+1,n}, \ldots, X_{n,n}) \sim \otimes_{t=\xi_n+1}^{n} P_1,$$

where P_0 and P_1 admit densities f_0 and f_1 with respect to a common dominating measure ν on $\mathcal{X}$. Let $(p_{t,1}, \ldots, p_{t,n})$ denote the CONCH p-values based on $\mathcal{D}_n$, and $\mathcal{C}_{n,1-\alpha}^{\mathrm{CONCH}}$ be the corresponding CONCH confidence set at sample size n. In this section, we prove that under mild regularity conditions, with the practical CPP score functions laid out in Section 4.2, the normalized length of CONCH confidence sets converges to 0 as $n \to \infty$.

5.1 Sharpness of CONCH with oracle LLR score

Let $\ell(x) = \log(f_0(x)/f_1(x))$ denote the log-likelihood ratio. We start with the case when we have access to the true ℓ, and below in the theorem, we formally establish that the length of the corresponding CONCH confidence set is stochastically bounded.

Theorem 5.1 (Sharpness with oracle LLR score). *Suppose in the above setting, there exists $\tau \in (0,1)$ such that $\xi_n/n \to \tau$ as $n \to \infty$, and $\mathrm{Var}_{X\sim P_0}(\ell(X)), \mathrm{Var}_{X\sim P_1}(\ell(X)) \in (0,\infty)$. Then, for CONCH p-values with score (4.4), there exists a constant $\kappa > 0$ (independent of n) such that*

$$\max_{|t_n - \xi_n| \geq \kappa} p_{t_n,n} \xrightarrow{P} 0 \quad \textit{as} \quad n \to \infty.$$

Consequently, if $\mathcal{C}_{n,1-\alpha}^{CONCH}$ denotes the corresponding confidence set, then $|\mathcal{C}_{n,1-\alpha}^{CONCH}| = \mathrm{O}_P(1)$, meaning that it is asymptotically sharp i.e., $|\,\mathcal{C}_{n,1-\alpha}^{CONCH}|/(n-1) \xrightarrow{P} 0$ as $n \to \infty$.

The proof of this theorem appears in Appendix B.3.2. This result further establishes that the oracle LLR score is asymptotically optimal. The requirement that $\xi_n/n \to \tau \in (0,1)$ ensures that the changepoint is in the interior, making the task of changepoint localization non-trivial.

5.2 Sharpness of CONCH with learned LLR score

While the oracle LLR score yields asymptotic sharpness of CONCH, in practice we rarely have access to the true log-likelihood ratio ℓ. This naturally leads us to obtain $\hat{\ell}_n$, an estimate of ℓ. However, learning $\hat{\ell}_n$ and running CONCH on the same data $\mathcal{D}_n$ induces strong dependencies among the CONCH p-values. Instead, for simpler analysis, suppose we estimate the log-likelihood ratio $\hat{\ell}_n(x)$ using a separate independent[3] dataset $\mathcal{D}'_n$. Then, consider the CPP score

$$S_t(\mathbf{x}) = \begin{cases} \sum_{i=t+1}^{\hat{\xi}_n(\mathbf{x})} -\hat{\ell}_n(x_i) & \text{if } t \leq \hat{\xi}_n(\mathbf{x}) \\ \sum_{i=\hat{\xi}_n(\mathbf{x})+1}^{t} \hat{\ell}_n(x_i) & \text{if } t > \hat{\xi}_n(\mathbf{x}), \end{cases} \tag{5.1}$$

where $\hat{\xi}_n(\mathbf{x})$, the corresponding MLE estimate for changepoint, is given by

$$\hat{\xi}_n(\mathbf{x}) \in \operatorname*{argmax}_{s \in [n-1]} \sum_{i=1}^{s} \hat{\ell}_n(x_s). \tag{5.2}$$

Note that this score coincides with (4.6), except that here $\hat{\ell}_n$ is learned on an independent dataset. This formulation accommodates both approaches from Section 4.2: learning the densities f_0 and f_1 separately, or training a classifier to approximate $\ell(\cdot)$ directly. Under a mild consistency assumption on $\hat{\ell}_n$, we can prove that the resulting CONCH confidence set is asymptotically sharp. The precise statement is given in the theorem below, with the proof provided in Appendix B.3.3.

Theorem 5.2 (Asymptotic sharpness)**.** *In the setting of Theorem 5.1, suppose $\hat{\ell}_n$ satisfies*

$$\mathbb{E}_{\substack{\mathcal{D}'_n, X \sim P, \\ X \perp \mathcal{D}'_n}} \left[|\hat{\ell}_n(X) - \ell(X)|^2 \right] \to 0 \qquad \text{as } n \to \infty, \tag{5.3}$$

for $P \in \{P_0, P_1\}$. Then, it follows that $|\mathcal{C}^{CONCH}_{n,1-\alpha}|/(n-1) \xrightarrow{P} 0$ as $n \to \infty$.

We note that the assumption in (5.3) can be easily satisfied. In what follows, we describe a simple procedure to achieve this using a sample splitting scheme: given $3n$ observations, define three equal splits of the data as $\mathcal{D}_{(1)} = \{X_{3k-2} : k \in [n]\}$, $\mathcal{D}_{(2)} = \{X_{3k-1} : k \in [n]\}$ and $\mathcal{D}_{(3)} = \{X_{3k} : k \in [n]\}$. A natural approach is to first estimate the component densities f_0 and f_1 from $\mathcal{D}_{(1)} \cup \mathcal{D}_{(2)}$ by $\hat{f}_{0,n}$ and $\hat{f}_{1,n}$. Then, we may define $\hat{\ell}_n = \log(\hat{f}_{0,n}/\hat{f}_{1,n})$, and finally run CONCH on $\mathcal{D}_{(3)}$. Now, $\hat{\ell}_n$ will satisfy (5.3) if we can find the density estimates that achieve L_1-consistency, i.e.,

$$\|\hat{f}_{0,n} - f_0\|_1,\ \|\hat{f}_{1,n} - f_1\|_1 = o_P(1). \tag{5.4}$$

To get estimates $\hat{f}_{0,n}$ and $\hat{f}_{1,n}$, one may first obtain a preliminary changepoint estimate $\xi^\dagger_n$ from $\mathcal{D}_{(1)}$ such that $|\xi^\dagger_n - \xi_n| = \mathrm{o}_P(n)$. This is rather a mild requirement, and can be achieved

[3] In practice, we often have prior labeled data on which a classifier can be trained and then be used to construct $\hat{\ell}_n(x)$. Otherwise, we can adopt a sample-splitting scheme: given $2n$ observations, we may use one half (odd indices), $\mathcal{D}_{(1)} = \{X_{2k-1} : k \in [n]\}$, to learn $\hat{\ell}_n(\cdot)$, while the other half (even indices), $\mathcal{D}_{(2)} = \{X_{2k} : k \in [n]\}$, to run CONCH.

by the kernel changepoint detection method [Garreau and Arlot, 2018], or the MCP localization [Bhattacharyya and Ramdas, 2025]. With such a preliminary point estimator in place, the points in $\mathcal{D}_{(2)}$ that are to the left and right of $\xi_n^\dagger$ contain i.i.d. samples from P_0 and P_1, respectively, with at most an $o_P(1)$ fraction of contamination (from the opposite distribution). Estimating f_0 and f_1 from these segments is therefore a problem of robust density estimation under a vanishing contamination model, a setting that is well studied [Chen et al., 2016, Uppal et al., 2020] and for which there exist estimators satisfying (5.4).

While Theorem 5.2 assumes that $\hat{\ell}_n$ is learned on data independent of $\mathcal{D}_n$, CONCH would typically use $\hat{\ell}_n$ that is learned on $\mathcal{D}_n$ itself. Nevertheless, we expect that asymptotic sharpness should continue to hold provided the learned estimator is suitably stable, but since a formal width analysis in such a setting may be significantly more complicated, we leave it to future work.

6 Universality of the CONCH algorithm

In earlier sections, we have established CONCH as a flexible framework for constructing distribution-free confidence sets for the changepoint. One may naturally ask: is CONCH one of many such distribution-free changepoint localization approaches that one may come up with? In this section, we give a conclusive answer to this question, which is that CONCH truly captures the full class of distribution-free changepoint localization methods.

Theorem 6.1. *Fix $\alpha \in (0,1)$. Let $C : \mathcal{X}^n \to [n]$ be any procedure that maps $\mathbf{X} \in \mathcal{X}^n$ to a set $C(\mathbf{X}) \subseteq [n]$ such that $\mathbb{P}_\xi(\xi \in C(\mathbf{X})) \geq 1-\alpha$. Then, there exists a CPP score function $S : \mathcal{X}^n \to \mathbb{R}^{n-1}$ such that C coincides exactly with the set $\mathcal{C}_{1-\alpha}^{CONCH}$ constructed with the score S.*

The proof of this theorem is provided in Appendix B.4. In simple terms, this result states that a particular choice of CPP score leads to a specific instance within the universal class of valid procedures for changepoint localization.

Moreover, this universality naturally yields an algorithm for refining existing confidence sets: CONCH can wrap around any heuristic or model-based set, calibrating it to achieve exact distribution-free coverage while preserving the original method's structural or modeling advantages.

6.1 Application: calibration of heuristic confidence sets

Suppose we are given a confidence set $C : \mathcal{X}^n \to 2^{[n-1]}$ that may or may not be valid, even asymptotically; for instance, one produced by a Bayesian or Bootstrap method. We can construct a CPP score function based on the given set and thereby obtain a distribution-free, finite-sample valid confidence set by running CONCH. Two such natural constructions of CPP score are:

- **Set membership score.** Define $S_t(x_1,\ldots,x_n) = \mathbb{1}\{t \in C(x_1,\ldots,x_n)\}$, which records only whether t is included in the given confidence set.
- **Set distance score.** Define $S_t(x_1,\ldots,x_n) = \min_{\ell \in C(x_1,\ldots,x_n)} |t-\ell|$, which refines the membership score by measuring the distance of t to the nearest index included in the set.

Algorithm 2: CONCH-CAL: CONCH calibration algorithm

```
Input: $(X_t)_{t=1}^n$ (dataset), $t_0 \in [n-1]$ (point estimate), and $\mathrm{pval} : \mathcal{X}^n \to [0,1]^{n-1}$
       ($p$-value function)
Output: $\mathcal{C}_{1-\alpha}^{\text{CONCH-CAL}}$ (CONCH-CAL confidence set at level $1-\alpha$)
Define $\hat{S} : \mathcal{X}^n \to \mathbb{R}^{n-1}$ as in (6.1) ;
for $t \in [n-1]$ do
    Compute CONCH $p$-value $p_t$ as in (3.2) with $S_t$ replaced by $\hat{S}_t$;
end
$\mathcal{C}_{1-\alpha}^{\text{CONCH-CAL}} \leftarrow \{\, t \in [n-1] : p_t > \alpha \,\}$;
return $\mathcal{C}_{1-\alpha}^{\text{CONCH-CAL}}$
```

Running the CONCH algorithm with either score yields a valid confidence set. However, by Proposition 4.1 (*ii*), the set-distance score always produces a narrower set than the set-membership score. Nonetheless, both scores are fairly coarse and often lead to wide confidence sets. In particular, for indices near 0 or n, these scores can induce t-symmetry, producing artificially inflated p-values in those regions (Proposition 4.1 (*i*)). Since this behavior is undesirable in practice, we next introduce a more informative CPP score that yields sharper confidence sets.

Most existing model-based or resampling-based approaches also produce a point estimate t_0. Moreover, in many cases, they first construct a p-value function $\mathrm{pval} : \mathcal{X}^n \to [0,1]^{n-1}$, which is then appropriately thresholded to form the confidence set C. Both components (t_0, pval) can be combined to define a more informative CPP score,

$$\hat{S}_t(x_1, \ldots, x_n) = \frac{\mathrm{pval}(x_1, \ldots, x_n; t)}{\mathrm{pval}(x_1, \ldots, x_n; t_0)}. \tag{6.1}$$

Applying CONCH with this score yields what we refer to as the CONCH-CAL algorithm, formally presented in Algorithm 2. By construction, this produces a valid distribution-free confidence set while retaining the original method's assessment of the changepoint. In practice, this allows analysts to exploit the strengths of bootstrap or Bayesian methods, such as their interpretability, while simultaneously ensuring exact finite-sample coverage.

We note that the point estimate t_0 depends on the ordered sequence $(x_1, \ldots, x_n)$, and thus the denominator $\mathrm{pval}(\cdot, \ldots, \cdot; t_0)$ is not invariant under permutations. Although one could in principle use $\mathrm{pval}(\cdot, \ldots, \cdot; t)$ directly as the CPP score in CONCH, this approach typically inherits the same conservativeness observed with set-membership and set-distance scores.

7 Extension: localization of multiple changepoints

In this section, we extend CONCH to the setting of multiple changepoint localization. The key observation is that, given any sufficiently consistent segmentation algorithm, the sequence can be partitioned into disjoint segments so that each segment contains, with high probability, at most one changepoint. We can then run CONCH independently within each segment and aggregate the

Algorithm 3: CONCH-SEG: Segmentwise CONCH for Multiple Changepoints

```
Input: $(X_t)_{t=1}^n$ (data); $\hat{K}$ and $0 = \hat{\xi}_0 < \hat{\xi}_1 < \cdots < \hat{\xi}_{\hat{K}} < n = \hat{\xi}_{\hat{K}+1}$ (estimated
       changepoints); $S : \cup_{m \in \mathbb{N}} \mathcal{X}^m \to \mathbb{R}^m$ (CPP score)
Output: $\mathcal{C}_{1-\alpha}^{\text{CONCH-SEG}}$ (overall confidence set at level $1-\alpha$)
Compute $(\tilde{X}_0, \dots, \tilde{X}_{\hat{K}})$ as in (7.2);
Initialize $\mathcal{C} \leftarrow \varnothing$;
for $\ell \in [\hat{K}]$ do
    $(L_\ell, R_\ell) \leftarrow (\tilde{X}_{\ell-1},\, \tilde{X}_\ell)$;
    Let $X^{(\ell)} \leftarrow (X_{L_\ell}, \dots, X_{R_\ell})$;
    Define score $S^{(\ell)} : \mathcal{X}^{R_\ell - L_\ell + 1} \to \mathbb{R}^{R_\ell - L_\ell}$;
    Compute CONCH $p$-values $\{p_t : t \in [L_\ell,\, R_\ell - 1]\}$ as in (3.2), using $S^{(\ell)}$ on $X^{(\ell)}$;
    Set $\mathcal{C}_\ell \leftarrow \{\, t \in [L_\ell,\, R_\ell - 1] : p_t > \alpha \,\}$;
    Update $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}_\ell$;
end
return $\mathcal{C}_{1-\alpha}^{\mathit{CONCH\text{-}SEG}} \leftarrow \mathcal{C}$
```

resulting sets to obtain an overall confidence set.

Formally, suppose there exist $K \in [n]$ and changepoints $0 = \xi_0 < \xi_1 < \cdots < \xi_K < n = \xi_{K+1}$, such that for each $\ell \in \{0, 1, \dots, K\}$,

$$(X_{\xi_\ell + 1}, \dots, X_{\xi_{\ell+1}}) \sim \mathcal{P}^{(\ell)}, \qquad \mathcal{P}^{(\ell)} \in \mathcal{M}(\mathcal{X}^{\xi_{\ell+1} - \xi_\ell}). \tag{7.1}$$

To be consistent with Assumption 1, we assume each $\mathcal{P}^{(\ell)}$ is exchangeable and that the collection $\{\mathcal{P}^{(0)}, \dots, \mathcal{P}^{(K)}\}$ is pairwise independent. Further, suppose a segmentation algorithm returns (a) an estimate $\hat{K}$ of the number of changepoints, and (b) an ordered sequence of estimated changepoints $0 = \hat{\xi}_0 < \hat{\xi}_1 < \cdots < \hat{\xi}_{\hat{K}} < n = \hat{\xi}_{\hat{K}+1}$, such that $\hat{K} \approx K$ and $\hat{\xi}_\ell \approx \xi_\ell$ for all $\ell \in [K]$.

Based on these estimates, we discretize the timeline $\{1, \dots, n\}$ into $\hat{K}$ data-dependent segments centered at the $\hat{\xi}_\ell$'s. Specifically, for $\ell \in \{0, \dots, \hat{K}\}$, define

$$\tilde{X}_\ell := \begin{cases} 1, & \text{if } \ell = 0, \\ \lfloor \frac{1}{2}(\hat{\xi}_\ell + \hat{\xi}_{\ell+1}) \rfloor, & \text{if } \ell \in [\hat{K} - 1], \\ n, & \text{if } \ell = \hat{K}. \end{cases} \tag{7.2}$$

We then take the ℓ-th segment to be $[\tilde{X}_{\ell-1},\, \tilde{X}_\ell]$ for $\ell \in [\hat{K}]$. Running CONCH independently on each segment and aggregating the segmentwise sets yields our overall confidence set. The resulting procedure, denoted CONCH-SEG, is summarized in Algorithm 3.

Kernel-based changepoint detection (KCPD) methods [Harchaoui and Cappé, 2007, Arlot et al., 2019] provide consistent estimators of changepoints under mild conditions [e.g., Garreau and Arlot, 2018, Diaz-Rodriguez and Jia, 2025]. Consequently, the CONCH framework can be seamlessly wrapped around a KCPD routine to construct confidence sets in the multiple-changepoint setting.

In Appendix C.2, we consider a Gaussian mean-shift model with multiple changepoints and empirically show that CONCH-SEG, when wrapped around a KCPD algorithm, sharply localizes the changepoints, demonstrating that this extension is both practical and statistically powerful.

Moreover, Appendix B.5 establishes asymptotic validity of a cross-fitted variant of CONCH-SEG. A direct analysis is hard because segmentation and CONCH are applied to the same data, potentially violating Assumption 1. Cross-fitting uses disjoint folds of the data to estimate changepoints and then run CONCH within the estimated segments. This restores the required independence and exchangeability structure and yields an asymptotic coverage guarantee.

8 Experiments

We evaluate the performance of CONCH through synthetic simulations and real-data applications to images (CIFAR-100, MNIST, DomainNet) and text (SST-2)[4]. Throughout, the reported CONCH confidence sets are produced by the CONCH-MC procedure (Algorithm 4). Although the resulting p-values are conservative under the null, the method nevertheless produces narrow, informative confidence sets with sharp changepoint localization in all settings considered.

8.1 Numerical simulations

8.1.1 Detecting Gaussian mean-shift

We begin with the most well-studied setting for changepoint analysis, namely the Gaussian mean-shift model, to illustrate the behavior of our proposed CONCH framework. Specifically, we generate a sequence of $n = 1000$ i.i.d. observations with a changepoint at $\xi = 400$: the pre-change distribution is $\mathcal{P}_{0,\xi} = \bigotimes_{t=1}^{\xi} \mathcal{N}(-1,1)$, while the post-change distribution is $\mathcal{P}_{1,\xi} = \bigotimes_{t=\xi+1}^{n} \mathcal{N}(1,1)$. We evaluate CONCH using four choices of CPP scores, introduced earlier in Section 4.2: (a) weighted mean difference, with a specified weight function, (b) oracle log-likelihood ratio (LLR), (c)parametrically learned LLR, assuming knowledge of the Gaussian family and (d) nonparametrically learned LLR, via kernel density estimates.

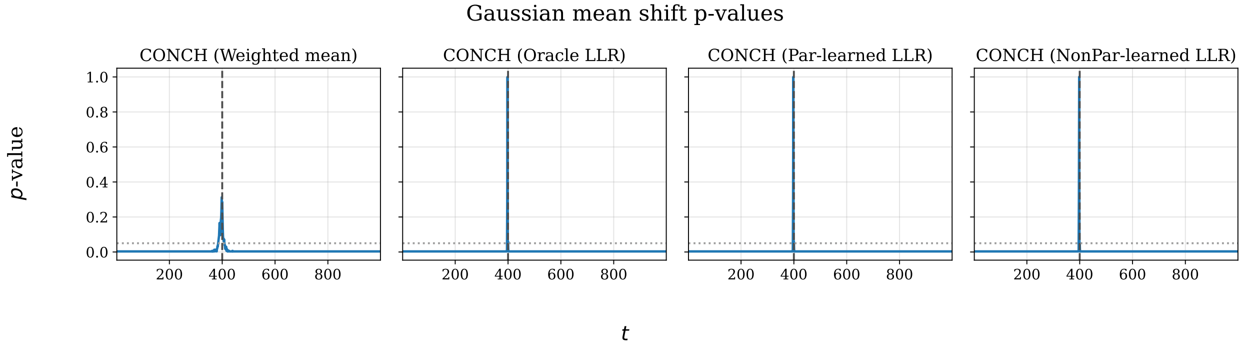

Figure 1: Distribution of conformal p-values (Gaussian mean-shift) for different methods.

[4] All code required to reproduce our experiments is available at https://github.com/rohanhore/CONCH.

Figure 1 displays the distribution of the resulting p-values produced by each method. CONCH produces sharply localized confidence sets across all score choices. The weighted-mean score results in the widest interval, $[385, 408]$, whereas all three LLR-based scores (oracle, parametrically learned and non-parametrically learned) yield a much narrower set $\{397, 398, 400\}$.

Overall, these results highlight two key features: (i) the validity of CONCH is preserved regardless of the choice of score, and (ii) more informative scores lead to sharper localization. Appendix C.1 presents an additional comparison with Dandapanthula and Ramdas [2025].

8.1.2 Refinement of resampling-based confidence sets using CONCH-CAL

We recall that the CONCH-CAL procedure (Algorithm 2) can refine confidence sets that were not originally designed with distribution-free validity guarantees. In particular, the Gaussian mean-shift model has been extensively studied, and several bootstrap-based methods provide asymptotically valid intervals that perform well in practice. However, under even mild model misspecification, these intervals can become overly wide or may miss the true changepoint ξ. In the following experiment, we consider two settings with $n = 500$ observations, and a changepoint at $\xi = 200$:

(i) Gaussian mean-shift model: $\mathcal{P}_{0,\xi} = \bigotimes_{t=1}^{\xi} \mathcal{N}(-1, 3)$ and $\mathcal{P}_{1,\xi} = \bigotimes_{t=\xi+1}^{n} \mathcal{N}(1, 3)$

(ii) Laplace mean-shift model: $\mathcal{P}_{0,\xi} = \bigotimes_{t=1}^{\xi} \text{Laplace}(-1, 3)$ and $\mathcal{P}_{1,\xi} = \bigotimes_{t=\xi+1}^{n} \text{Laplace}(1, 3)$.

In both these settings, we compute the bootstrap confidence intervals, and then evaluate how the CONCH-CAL procedure refines them to produce a distribution-free narrow confidence set. In both experiments, we employ the residual bootstrap scheme to construct the initial confidence sets: for each replicate, the changepoint is re-estimated on a resampled sequence formed from centered residuals, producing an empirical distribution of $\hat{\tau}$ from which the p-values are obtained.

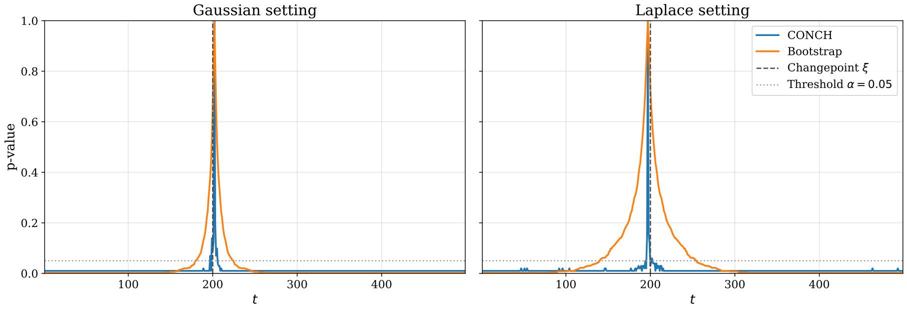

Figure 2: Refinement of bootstrap-based confidence sets using CONCH-CAL under Gaussian and Laplace mean-shift models.

Under the correct model class, the Gaussian model, the bootstrap interval $[180, 224]$ is largely unchanged by CONCH-CAL to produce a similar interval $[197, 205]$. When the model is misspecified, in the Laplace model, the bootstrap interval $[140, 258]$ is originally quite inflated by heavy-tailed noise. However, it is reduced to $[196, 202]$ after calibration, enabling more precise localization. The bootstrap p-values in general are much diffuse, while those from CONCH-CAL remain sharply concentrated near the true changepoint, highlighting robustness across distributional regimes.

8.2 Real data experiments

8.2.1 DomainNet: detecting domain shift

In this experiment, we tackle the problem of detecting a domain shift using the DomainNet dataset [Peng et al., 2019], consisting of six diverse domains (e.g., real, sketch, painting). Among these, we use the *real* and *sketch* domains to construct a changepoint detection setting. Further, we convert all images to grayscale to remove color cues and further increase the similarity between classes. Specifically, before the changepoint ($\xi = 350$), we observe samples from the real domain, and after ξ, we observe samples from the sketch domain, totaling 800 samples (Figure 3).

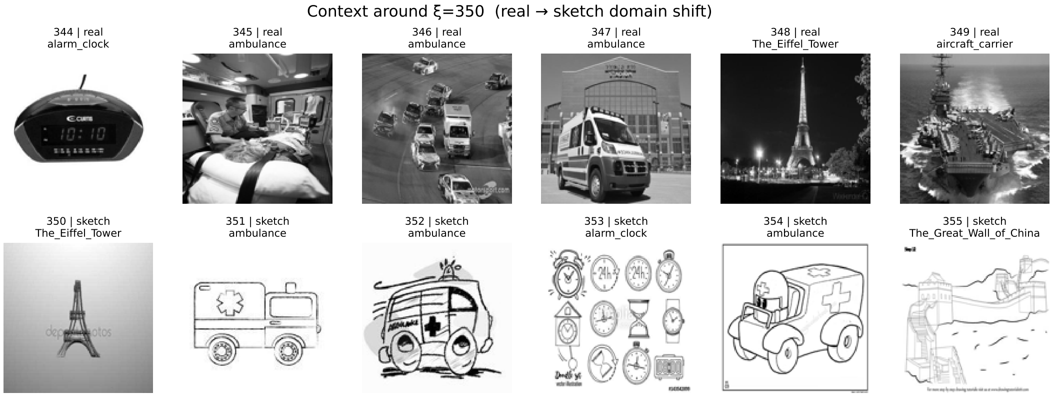

Figure 3: Illustration of the DomainNet changepoint setup: samples switch from the real to the sketch domain at $\xi = 350$ ($n = 800$). Images are drawn from the DomainNet dataset, which was collected via online search; class labels may not perfectly align with visual semantics, making the domain-shift detection problem more challenging.

We first train a CNN-based classifier to distinguish real images from hand-drawn sketches. Although the classifier provides substantial discriminative information, it does not directly translate into distribution-free guarantees for changepoint localization. The CONCH framework bridges this gap by converting classifier outputs into a principled, distribution-free procedure, yielding a narrow confidence set $[350, 351]$ that consistently contains the true changepoint (Figure 4).

8.2.2 SST-2: detecting sentiment change using language models

We next demonstrate our method on text data, showing that it can localize changepoints in language settings. Using the Stanford Sentiment Treebank (SST-2) dataset of movie reviews with binary sentiment labels [Socher et al., 2013], we simulate a shift from predominantly positive to predominantly negative sentiment, mirroring real-world tasks such as detecting changes in customer feedback or public opinion. We observe $n = 1000$ reviews with a changepoint at $\xi = 400$: before ξ, reviews are i.i.d. positive (P_0); after ξ, reviews are i.i.d. negative (P_1). For example:

- $t = 399$ (positive): “juicy writer”
- $t = 400$ (positive): “intricately structured and well-realized drama”

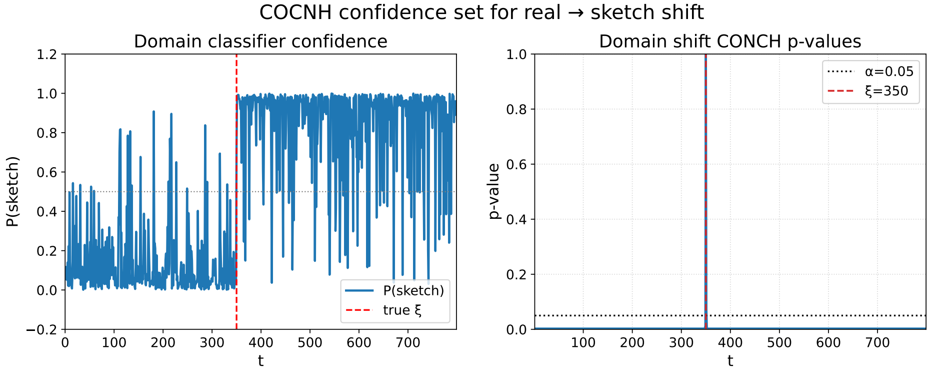

Figure 4: p-values for domain shift detection between real and sketch domains: classifier scores (left) and CONCH p-values (right)

- $t = 401$ (negative): "painfully "
- $t = 402$ (negative): "than most of jaglom's self-conscious and gratingly irritating films"

First, we find a DistilBERT model fine-tuned for sentiment classification [Sanh et al., 2019], and then the corresponding model logits are used to build a CPP score for our CONCH method, which yields a 95% confidence set $[400, 401]$ (Figure 5, left panel), effectively pinpointing the changepoint. Even under a subtler scenario, where sentiment shifts only from 60% positive to 40% positive, we obtain a nontrivial 95% confidence set $[326, 463]$ (Figure 5, right panel), demonstrating sharp localization of the changepoint in complex settings.

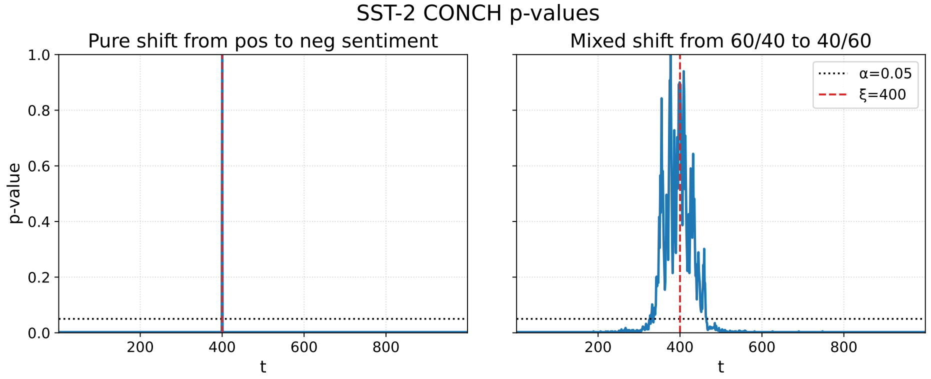

Figure 5: CONCH p-values for sentiment shift in SST-2: from positive to negative reviews at $\xi = 400$ (left), and from 60% positive to 40% positive (right).

Additional experiments Appendix C presents supplementary experiments spanning several changepoint detection settings. We first illustrate CONCH-SEG for multiple changepoint localization in Gaussian mean-shift models, then consider a two-urn example (urn-shift detection), a digit-class shift on MNIST, and a class-shift experiment on CIFAR-100, demonstrating the robustness and flexibility of our approach.

9 Conclusion

In this work, we introduced CONCH, a framework for distribution-free offline changepoint localization. Our approach leverages conformal p-values to construct confidence sets with finite-sample, distribution-free guarantees. We provided design guidelines, including principled choices of score functions and a Monte Carlo approximation to the full-permutation p-value, that enhance both the power and practicality of the framework. With an appropriate score function, the CONCH confidence sets shrink with sample size, as one would expect from any useful localization procedure.

We established a universality result positioning CONCH as a canonical method for distribution-free offline changepoint localization. This, in turn, paves the way for (i) a simple calibration procedure that can wrap around any localization algorithm to yield valid confidence sets, and (ii) an extension of the CONCH framework to multiple-changepoint settings.

While our extension of the conformal localization framework to the multiple-changepoint setting is natural, it relies crucially on first obtaining a consistent segmentation, which is a nontrivial task in its own right. A promising direction for future work is to adapt techniques such as wild binary segmentation [Fryzlewicz, 2014], and extend CONCH directly to the multiple-changepoint case, thereby broadening its scope and applicability.

References

Andreas Anastasiou and Piotr Fryzlewicz. Detecting multiple generalized change-points by isolating single ones. *Metrika*, 85(2):141–174, 2022.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of machine learning research*, 20(162):1–56, 2019.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Gouri K Bhattacharyya and Richard A Johnson. Nonparametric tests for shift at an unknown time point. *The Annals of Mathematical Statistics*, pages 1731–1743, 1968.

Swapnaneel Bhattacharyya and Aaditya Ramdas. Theoretical guarantees for change localization using conformal p-values. *arXiv preprint arXiv:2510.08749*, 2025.

A. E. Brockwell. Universal residuals: A multivariate transformation. *Statistics & Probability Letters*, 77(14):1473–1478, 2007.

Mengjie Chen, Chao Gao, and Zhao Ren. A general decision theory for huber's ε-contamination model. *Electronic Journal of Statistics*, 10:3752–3774, 2016.

Haeran Cho and Claudia Kirch. Bootstrap confidence intervals for multiple change points based on moving sum procedures. *Computational Statistics & Data Analysis*, 175:107552, 2022.

Sanjit Dandapanthula and Aaditya Ramdas. Offline changepoint localization using a matrix of conformal p-values. *arXiv preprint arXiv:2505.00292*, 2025.

Li Deng. The MNIST database of handwritten digit images for machine learning research. *Signal Processing Magazine, IEEE*, 29:141–142, 11 2012.

Jairo Diaz-Rodriguez and Mumin Jia. Consistent kernel change-point detection under m-dependence for text segmentation. *arXiv preprint arXiv:2510.03437*, 2025.

Jonathan W Duggins. *Parametric Resampling Methods for Retrospective Changepoint Analysis.* PhD thesis, Virginia Polytechnic Institute and State University, 2010.

Stergios B Fotopoulos, Venkata K Jandhyala, and Elena Khapalova. Exact asymptotic distribution of change-point mle for change in the mean of Gaussian sequences. *The Annals of Applied Statistics*, pages 1081–1104, 2010.

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

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014. ISSN 00905364.

Damien Garreau and Sylvain Arlot. Consistent change-point detection with kernels. *arXiv preprint arXiv:1612.04740*, 2018.

Gregory Gurevich and Albert Vexler. Guaranteed maximum likelihood splitting tests of a linear regression model. *Statistics*, 40(6):465–484, 2006.

Zaid Harchaoui and Olivier Cappé. Retrospective multiple change-point estimation with kernels. In *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 768–772. IEEE, 2007.

Matthew T Harrison. Conservative hypothesis tests and confidence intervals using importance sampling. *Biometrika*, 99(1):57–69, 2012.

Ying Jin and Emmanuel J Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

Hyune-Ju Kim and David Siegmund. The likelihood ratio test for a change-point in simple linear regression. *Biometrika*, 76(3):409–423, 1989.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Erich Leo Lehmann and Joseph P Romano. *Testing statistical hypotheses.* Springer, 2005.

Ilia Nouretdinov, Vladimir Vovk, and Alex Gammerman. Conformal changepoint detection in continuous model situations. In *Conformal and Probabilistic Prediction and Applications*, pages 300–302. Proceedings of Machine Learning Research, 2021.

Ewan Stafford Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527, 1955.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1406–1415, 2019.

Anthony N Pettitt. A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(2):126–135, 1979.

Richard E Quandt. The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, 53(284):873–880, 1958.

Gordon J Ross and Niall M Adams. Two nonparametric control charts for detecting arbitrary distribution changes. *Journal of Quality Technology*, 44(2):102–116, 2012.

Aytijhya Saha and Aaditya Ramdas. Post-detection inference for sequential changepoint localization. *arXiv preprint arXiv:2502.06096*, 2025.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. E-detectors: A nonparametric framework for sequential change detection. *The New England J of Stat. in Data Sci.*, 2(2):229–260, 2023. ISSN 2693-7166.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, 2013.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

Ananya Uppal, Shashank Singh, and Barnabas Poczos. Robust density estimation under besov ipm losses. *Advances in Neural Information Processing Systems*, 33:5345–5355, 2020.

Nicolas Verzelen, Magalie Fromont, Matthieu Lerasle, and Patricia Reynaud-Bouret. Optimal change-point detection and localization. *The Annals of Statistics*, 51(4):1586–1610, 2023.

Denis Volkhonskiy, Evgeny Burnaev, Ilia Nouretdinov, Alexander Gammerman, and Vladimir Vovk. Inductive conformal martingales for change-point detection. In *Conformal and Probabilistic Prediction and Applications*, pages 132–153. PMLR, 2017.

Vladimir Vovk. Testing randomness online. *Statistical Science*, 36(4):595–611, 2021.

Vladimir Vovk, Ilia Nouretdinov, and Alexander Gammerman. Testing exchangeability on-line. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 768–775, 2003.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Ernst Ahlberg, Lars Carlsson, and Alex Gammerman. Retrain or not retrain: Conformal test martingales for change-point detection. In *Conformal and Probabilistic Prediction and Applications*, pages 191–210. PMLR, 2021.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 444–453, 1999.

Xiaoyang Wu, Lin Lu, Zhaojun Wang, and Changliang Zou. Conditional testing based on localized conformal p-values. *arXiv preprint arXiv:2409.16829*, 2024.

Haotian Xu, Daren Wang, Zifeng Zhao, and Yi Yu. Change-point inference in high-dimensional regression models under temporal dependence. *The Annals of Statistics*, 52(3):999–1026, 2024.

Changliang Zou, Yukun Liu, Peng Qin, and Zhaojun Wang. Empirical likelihood ratio test for the change-point problem. *Statistics & Probability Letters*, 77(4):374–382, 2007.

Algorithm 4: CONCH-MC: CONCH with random permutations

```
Input: $(X_t)_{t=1}^n$ (dataset), $1-\alpha$ (target coverage), $M$ (number of permutations) and
       $S : \mathcal{X}^n \to \mathbb{R}^n$ (CPP score function)
Output: $\mathcal{C}_{1-\alpha}^{\text{CONCH-MC}}$ (CONCH-MC confidence set at level $1-\alpha$)
for $t \in [n-1]$ do
    $\Pi_t \leftarrow \{\pi \in \mathcal{S}_n : \text{for all } i \le t,\ \pi(i) \le t \text{ and for all } i > t,\ \pi(i) > t\}$;
    for $k \in [M]$ do
        Sample $\pi^{(k)} \sim \Pi_t$;
        Evaluate $S_t(\pi^{(k)}(\mathbf{X}))$;
    end
    $\tilde{p}_t \leftarrow \frac{1}{M+1}(1 + \sum_{k=1}^{M} \mathbb{1}\left\{S_t(\pi^{(k)}(\mathbf{X})) \le S_t(\mathbf{X})\right\}$);
end
$\mathcal{C}_{1-\alpha}^{\text{CONCH-MC}} \leftarrow \{t \in [n-1] : \tilde{p}_t > \alpha\}$
return $\mathcal{C}_{1-\alpha}^{\text{CONCH-MC}}$
```

A Variants of CONCH algorithm

In this section, we describe a few variants of the general CONCH framework that highlight the practicality and usefulness of our changepoint localization algorithm.

A.1 CONCH-MC: randomized variant for scalability

To compute the CONCH p-value p_t in (3.2), one must enumerate all permutations in Π_t and compute the corresponding score $S_t(\pi(\mathbf{X}))$ for each π. For large n, this may be computationally expensive. To reduce computational burden, we proceed as follows: sample $\pi^{(1)}, \ldots, \pi^{(M)} \overset{\text{i.i.d.}}{\sim} \text{Unif}(\Pi_t)$, and then calculate:

$$\tilde{p}_t := \frac{1 + \sum_{k=1}^{M} \mathbb{1}\left\{S_t(\pi^{(k)}(\mathbf{X})) \le S_t(\mathbf{X})\right\}}{1+M}. \tag{A.1}$$

This yields a *randomized* confidence set $\{\, t \in [n-1] : \tilde{p}_t > \alpha \,\}$. We refer to this procedure as CONCH-MC, presented formally in Algorithm 4. Similar to CONCH, any randomly sampled $\pi \in \Pi_t$ preserves pre-change and post-change exchangeability under $\mathcal{H}_{0,t}$, thereby providing us with a valid p-value $\tilde{p}_t$, as we state in Theorem A.1 and prove in Appendix B.1.

Theorem A.1. *For any $t \in [n]$, p_t defined in (A.1) is a valid p-value under $\mathcal{H}_{0,t}$. In particular, for any $\alpha \in (0,1)$, $\mathbb{P}_\xi(\tilde{p}_\xi \le \alpha) \le \alpha$. Consequently, $\mathcal{C}_{1-\alpha}^{\text{CONCH-MC}}$ is a distribution-free confidence set for changepoint.*

A.2 Achieving exact validity for CONCH confidence sets

While both p-values p_t and $\tilde{p}_t$ in (3.2) and (A.1) control the Type I error under $\mathcal{H}_{0,t}$ at level α, it is sometimes desirable to attain *exact* level-α validity. Achieving exact validity can yield more powerful or sharper procedures. To this end, we introduce a randomized refinement of the p-values that guarantees exact validity. Specifically, define

$$\bar{p}_t := \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\left\{S_t(\pi(\mathbf{X})) < S_t(\mathbf{X})\right\} + U \cdot \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\left\{S_t(\pi(\mathbf{X})) = S_t(\mathbf{X})\right\}, \tag{A.2}$$

where $U \sim \text{Unif}[0,1]$. Plugging these randomized p-values into the general CONCH framework yields the confidence set

$$\bar{\mathcal{C}}_{1-\alpha}^{\text{CONCH}} = \{\, t \in [n-1] : \bar{p}_t > \alpha \,\}, \tag{A.3}$$

which attains confidence exactly $1-\alpha$, as formalized below.

Theorem A.2. *For each $t \in [n-1]$, $\bar{p}_t$ defined in (A.2) is a valid p-value under $\mathcal{H}_{0,t}$. In particular, for any $\alpha \in (0,1)$,*

$$\mathbb{P}_\xi(\bar{p}_\xi \leq \alpha) = \alpha.$$

Consequently, $\mathbb{P}_\xi\Big(\xi \in \bar{\mathcal{C}}_{1-\alpha}^{\text{CONCH}}\Big) = 1-\alpha$ where $\bar{\mathcal{C}}_{1-\alpha}^{\text{CONCH}}$ is as defined in (A.3).

The proof of the result is deferred to Appendix B.1, where we actually prove a stronger version of this theorem. In particular, the validity of the p-value $\bar{p}_\xi$ holds even conditional on the multisets $\{X_1, \ldots, X_\xi\}$ and $\{X_{\xi+1}, \ldots, X_n\}$.

B Proofs

In this section, we present proofs of the main results stated in the paper, along with several auxiliary results that support them.

B.1 Proving coverage guarantees for CONCH

B.1.1 Proof of Theorem 3.1

First, observe that under the null $\mathcal{H}_{0t}$, $\pi(\mathbf{X}) \overset{d}{=} \mathbf{X}$ for any $\pi \in \Pi_t$. We define a function $p_t : \mathcal{X}^n \to [0,1]$ by

$$p_t(\mathbf{x}) := \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\left\{S_t(\pi(\mathbf{x})) \leq S_t(\mathbf{x})\right\}.$$

Further, note that $p_t \equiv p_t(\mathbf{X})$. Therefore,

$$\begin{aligned}
\mathbb{P}_t\left(p_t(\mathbf{X}) \leq \alpha\right) &= \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{P}_t\left(p_t(\pi(\mathbf{X})) \leq \alpha\right) \\
&= \mathbb{E}_t\left[\frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\left\{p_t(\pi(\mathbf{X})) \leq \alpha\right\}\right] \\
&= \mathbb{E}_t\left[\frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\left\{\frac{1}{|\Pi_t|} \sum_{\pi' \in \Pi_t} \mathbb{1}\left\{S_t(\pi'(\mathbf{x})) \leq S_t(\pi(\mathbf{x}))\right\} \leq \alpha\right\}\right] \leq \alpha,
\end{aligned}$$

where the penultimate step follows by noting that $\pi \circ \Pi_t = \Pi_t$, and the last inequality follows by Harrison [2012, Lemma 3]. This completes the proof. □

B.1.2 Proof of Theorem A.1

Given permutations $\pi_{1,t}, \ldots, \pi_{M,t} \in \Pi_t$, we define the fucntion

$$\tilde{p}_t(\mathbf{x}; \pi_{1,t}, \ldots, \pi_{M,t}) := \frac{1 + \sum_{k=1}^{M} \mathbb{1}\left\{s_t(\pi_{k,t}(\mathbf{x})) \le s_t(\mathbf{x})\right\}}{1+M},$$

Consider an additional uniform draw $\pi_{0,t}$ from Π_t.

Hence, note that with $\pi_{1,t}, \ldots, \pi_{M,t} \overset{iid}{\sim} \mathrm{Unif}(\Pi_t)$, we have that

$$(\pi_{1,t}, \ldots, \pi_{M,t}) \overset{d}{=} (\pi_{0,t} \circ \pi_{1,t}, \ldots, \pi_{0,t} \circ \pi_{M,t}).$$

Moreover, conditional on $\pi_{0,t}, \pi_{1,t}, \ldots, \pi_{M,t}$, $\mathbf{X} \overset{d}{=} \pi_{0,t}(\mathbf{X})$ under the null $\mathcal{H}_{0,t}$. Consequently,

$$\tilde{p}_t(\mathbf{X}; \pi_{1,t}, \ldots, \pi_{M,t}) \overset{d}{=} \tilde{p}_t(\mathbf{X}; \pi_{0,t} \circ \pi_{1,t}, \ldots, \pi_{0,t} \circ \pi_{M,t}) \overset{d}{=} \tilde{p}_t(\pi_{0,t}(\mathbf{X}); \pi_{0,t} \circ \pi_{1,t}, \ldots, \pi_{0,t} \circ \pi_{M,t}).$$

Finally, note that for $\tilde{p}_t$, defined in (A.1), $\tilde{p}_t \equiv \tilde{p}_t(\mathbf{X}; \pi_{1,t}, \ldots, \pi_{M,t})$, and therefore,

$$\begin{aligned}\tilde{p}_t(\mathbf{X}; \pi_{1,t}, \ldots, \pi_{M,t}) &\overset{d}{=} \tilde{p}_t(\pi_{0,t}(\mathbf{X}); \pi_{0,t} \circ \pi_{1,t}, \ldots, \pi_{0,t} \circ \pi_{M,t}) \\ &= \frac{1 + \sum_{k=1}^{M} \mathbb{1}\left\{s_t(\pi_{k,t}(\mathbf{X})) \le s_t(\pi_{0,t}(\mathbf{X}))\right\}}{M+1} \\ &= \frac{\sum_{k=0}^{M} \mathbb{1}\left\{s_t(\pi_{k,t}(\mathbf{X})) \le s_t(\pi_{0,t}(\mathbf{X}))\right\}}{M+1},\end{aligned}$$

i.e., the rank of $s_t(\pi_{0,t}(\mathbf{X}))$ in the exchangeable collection $\{s_t(\pi_{0,t}(\mathbf{X})), s_t(\pi_{1,t}(\mathbf{X})), \ldots, s_t(\pi_{M,t}(\mathbf{X}))\}$. Consequently,

$$\mathbb{P}_t\left(\tilde{p}_t = \tilde{p}_t(\mathbf{X}; \pi_{1,t}, \ldots, \pi_{M,t}) \le \alpha\right) \le \alpha.$$

This completes the proof. □

B.1.3 Proof of Theorem A.2

We begin by letting F denote the distribution of $S_t(\pi(\mathbf{X}))$ conditional on the multisets $M_{\mathrm{left}} := \{X_1, \ldots, X_t\}$ and $M_{\mathrm{right}} := \{X_{t+1}, \ldots, X_n\}$, where $\pi \sim \mathrm{Unif}(\Pi_t)$. Then

$$\bar{p}_t = \lim_{y \uparrow S_t(\mathbf{X})} F(y) + U\big(F(S_t(\mathbf{X})) - \lim_{y \uparrow S_t(\mathbf{X})} F(y)\big).$$

Under $\mathcal{H}_{0,t}$, we have $S_t(\mathbf{X}) \overset{d}{=} S_t(\pi(\mathbf{X}))$ conditional on M_{left} and M_{right}. Hence, by Dandapanthula and Ramdas [2025, Lemma E.1], the p-value $\bar{p}_t$, conditional on M_{left} and M_{right}, follows Unif$[0, 1]$

(see also Brockwell, 2007). Therefore,

$$\mathbb{P}_t(\bar{p}_t \leq \alpha) = \mathbb{E}_t[\mathbb{P}_t(\bar{p}_t \leq \alpha \mid \{X_1, \ldots, X_n\})] = \mathbb{E}_t[\alpha] = \alpha.$$

This completes the proof.

B.2 Proving properties of the CPP score and its optimal form

B.2.1 Proof of Proposition 4.1

The first part of the result follows immediately by noting that when S_t satisfies the t-symmetry, then by (3.2) p_t is identically equal to 1, as required.

For the second part, fix $t \in [n-1]$. By definition (see (3.2)),

$$p_{t,1} = \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\left\{S_t(\pi(\mathbf{X})) \leq S_t(\mathbf{X})\right\}, \quad p_{t,2} = \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\left\{f(S_t(\pi(\mathbf{X}))) \leq f(S_t(\mathbf{X}))\right\}.$$

Since f is non-decreasing,

$$S_t(\pi(\mathbf{X})) \leq S_t(\mathbf{X}) \implies f(S_t(\pi(\mathbf{X}))) \leq f(S_t(\mathbf{X})),$$

and therefore $p_{t,1} \leq p_{t,2}$. As this holds for all $t \in [n-1]$, it further follows that $C_1 \subseteq C_2$.

B.2.2 Proof of Lemma 4.2: second conformal NP lemma

In the setup of Section 4.1, we consider the following hypothesis testing problem:

$$\mathcal{H}_0' : \mathbf{X} \mid \mathcal{X}_{L,t}, \mathcal{X}_{R,t} \sim \mathcal{P}^{[t]}_{\mathbf{X}|\mathcal{X}_{L,t},\mathcal{X}_{R,t}} \quad \text{v.s.} \quad \mathcal{H}_1' : \mathbf{X} \mid \mathcal{X}_{L,t}, \mathcal{X}_{R,t} \sim \mathcal{P}^{[r]}_{\mathbf{X}|\mathcal{X}_{L,t},\mathcal{X}_{R,t}}.$$

Given samples $\mathbf{X} \in \mathbb{R}^n$, observe that

$$\frac{\mathsf{d}(\mathcal{P}^{[r]}_{\mathbf{X}|\mathcal{X}_{L,t},\mathcal{X}_{R,t}})}{\mathsf{d}(\mathcal{P}^{[t]}_{\mathbf{X}|\mathcal{X}_{L,t},\mathcal{X}_{R,t}})}(\mathbf{X}) \propto \frac{\mathsf{d}(\mathcal{P}^{[r]}_{\mathbf{X}})}{\mathsf{d}(\mathcal{P}^{[r]}_{\mathbf{X}})} = \frac{\prod_{i \leq r} f_0(X_i) \prod_{i > r} f_1(X_i)}{\prod_{i \leq t} f_0(X_i) \prod_{i > t} f_1(X_i)} = s^\star(\mathbf{X})^{-1}.$$

By the Neyman–Pearson lemma [Lehmann and Romano, 2005, Theorem 3.2.1 (ii)], any test $\phi(\mathbf{X})$ that attains exact validity at level α under $\mathcal{H}_0'$ and satisfies

$$\phi(\mathbf{X}) = \begin{cases} 1 & \text{if } s^\star(\mathbf{X})^{-1} > \tau_\alpha, \\ 0 & \text{if } s^\star(\mathbf{X})^{-1} < \tau_\alpha, \end{cases} \tag{B.1}$$

for an appropriate threshold $\tau_\alpha \in \mathbb{R}$, is most powerful for testing $\mathcal{H}_0'$ against $\mathcal{H}_1'$.

As discussed in Section 4.1, the test $\phi_t(\cdot; s) = \mathbb{1}\{p_t(s) \leq \alpha\}$ controls the Type I error exactly at level α under $\mathcal{H}_0'$ for any score function s. Therefore, to establish the optimality of $s^\star$, it suffices to show that $\phi_t(\cdot; s^\star)$ admits the form given in (B.1).

Define $\mathbf{X}_\pi = \pi(\mathbf{X})$ for $\pi \sim \mathrm{Unif}(\Pi_t)$, and let $F_{s^\star(\mathbf{X}_\pi)}$ denote the conditional cumulative distribution function of $s^\star(\mathbf{X}_\pi)$ given $\mathbf{X}$. Set

$$\tau_\alpha := \inf\{y \in \mathbb{R} : F_{s^\star(\mathbf{X}_\pi)}(y) \geq \alpha\}.$$

By the definition of p_t in (4.1), we have

$$\begin{aligned} s^\star(\mathbf{X})^{-1} > \tau_\alpha &\implies p_t(s^\star) \leq \alpha, \\ s^\star(\mathbf{X})^{-1} < \tau_\alpha &\implies p_t(s^\star) > \alpha, \end{aligned}$$

as required. This completes the proof. □

B.2.3 Proof of Theorem 4.3

We observe that only the t-th coordinate of CPP score S_t determines the CONCH p-value $\bar{p}_t$ defined in (A.2). Therefore, with the notation laid out in Section 4.1, we can write

$$n - \mathbb{E}_{\mathcal{H}_{0,\xi} \cap \mathcal{P}_{\mathrm{IID}}}[\bar{C}^{\mathrm{CONCH}}_{1-\alpha}(S)] = \sum_{t=1}^{n} \mathbb{E}_{\mathcal{H}_{0,\xi} \cap \mathcal{P}_{\mathrm{IID}}}[\mathbb{1}\{p_t(S_t) \leq \alpha\}].$$

Finally, note that for any $j \in [n-1]$, $\mathcal{H}_{0,j} \cap \mathcal{P}_{\mathrm{IID}} = \mathcal{P}^{[\xi]}$. Hence, by tower law, we have that

$$\mathbb{E}_{\mathcal{P}^{[\xi]}}[\mathbb{1}\{p_t(S_t) \leq \alpha\}] = \mathbb{E}\Big[\mathbb{E}\big[\,\mathbb{1}\{p_t(S_t) \leq \alpha\} \mid \{X_1, \ldots, X_\xi\}, \{X_{\xi+1}, \ldots, X_n\}\,\big]\Big]$$

Moreover, the p-value $\bar{p}_t = p_t(S_t)$ must be valid under $\mathcal{H}'_t$. Hence, applying Lemma 4.2, the optimal form of S_t^{OPT} follows readily.

B.3 Proving asymptotic sharpness of CONCH confidence set

In this section, we give the proof for the asymptotic sharpness of CONCH confidence sets for oracle LLR score (Theorem 5.1) and for learned LLR score (Theorem 5.2). First, we start with a few notation.

Notation Recall that $\ell(x) = f_0(x)/f_1(x)$, and hence observe that $\mathrm{KL}(P_0\|P_1) = \mathbb{E}_{X\sim P_0}[\ell(X)]$ and $\mathrm{KL}(P_1\|P_0) = \mathbb{E}_{X\sim P_1}[-\ell(X)]$, where $\mathrm{KL}(P\|Q)$ denotes the Kullback-Leibler divergence between distributions P and Q. Moreover, we define the Jeffreys divergence as

$$\mathrm{J}(P_0, P_1) = \mathrm{KL}(P_0\|P_1) + \mathrm{KL}(P_1\|P_0).$$

The corresponding var-entropy measures are given by $\sigma_0^2 := \mathrm{Var}_{X\sim P_0}(\ell(X))$ and $\sigma_1^2 := \mathrm{Var}_{X\sim P_1}(\ell(X))$. We let $\sigma_\star$ denote $\max\{\sigma_0, \sigma_1\}$. Next, given an estimate of ℓ, $\hat{\ell}_n$, for any $k \in [n-1]$, we define the empirical averages

$$\hat{\mu}_{k,n,L} := \frac{1}{k}\sum_{i=1}^{k} \hat{\ell}_n(X_i), \qquad \hat{\mu}_{k,n,R} := \frac{1}{n-k}\sum_{i=k+1}^{n} -\hat{\ell}_n(X_i), \qquad \hat{v}_n := \frac{1}{n}\sum_{i=1}^{n} \hat{\ell}_n^2(X_i).$$

These quantities in turn enable an empirical upper bound on the CONCH p-values. Finally, we define

$$\Gamma_0 := \mathbb{E}_{\substack{\mathcal{D}'_n, X\sim P_0, \\ X\perp\mathcal{D}'_n}}[|\hat{\ell}_n(X) - \ell(X)|^2], \quad \Gamma_1 := \mathbb{E}_{\substack{\mathcal{D}'_n X\sim P_1, \\ X\perp\mathcal{D}'_n,}}[|\hat{\ell}_n(X) - \ell(X)|^2].$$

Now, instead of working with the true CONCH p-values, we would work with a 'close' approximation of the same. We define them formally below. Throughout, we write $\hat{\xi}_n \equiv \hat{\xi}_n(\mathbf{X})$ where $\hat{\xi}_n(\mathbf{x})$ is as defined in (5.2). Consider the score function,

$$S_t^{(n)}(x_1, \dots, x_n) = \begin{cases} \sum_{i=t+1}^{\hat{\xi}_n} -\hat{\ell}_n(X_i) & \text{if } t \le \hat{\xi}_n, \\ \sum_{i=\hat{\xi}_n+1}^{t} \hat{\ell}_n(X_i) & \text{if } t > \hat{\xi}_n. \end{cases} \tag{B.2}$$

Based on this score, we define the permutation p-values $(\tilde{p}_{1,n}, \dots, \tilde{p}_{n-1,n})$ as

$$\tilde{p}_{t,n} := \frac{1}{|\Pi_t|}\sum_{\pi\in\Pi_t} \mathbb{1}\left\{S_t^{(n)}(\pi(\mathbf{X})) \le S_t^{(n)}(\mathbf{X})\right\}. \tag{B.3}$$

The only distinction between these p-values and the CONCH p-values $\{p_{t,n}\}$ is that here, when evaluating $S_t^{(n)}(\pi(\mathbf{x}))$, we *do not* recompute $\hat{\xi}_n$ under the permutation. These 'frozen-estimator' p-values $\{\tilde{p}_{t,n}\}$ will serve as intermediate quantities in our analysis.

The rest of the section is organized as follows:

- Intuitively, the CONCH confidence score (5.1) builds a confidence set around the MLE estimate $\hat{\xi}_n(\mathbf{X})$. Therefore, to obtain sharp CONCH confidence sets, the initial MLE estimates must themselves be weakly consistent. In Appendix B.3.1, we show that indeed $\hat{\xi}_n$ is at most $\mathrm{o}_P(\sqrt{n})$ away from ξ_n.
- In Appendix B.3.2, we then prove that the CONCH confidence set computed with the oracle LLR score has $\mathrm{O}_P(1)$ length. This establishes the optimality and asymptotic sharpness of the oracle LLR score.
- Next, in Appendix B.3.3, we extend the asymptotic sharpness result to the setting where ℓ is replaced by an estimate $\hat{\ell}_n$, showing that the normalized length of the corresponding CONCH confidence set converges to 0.
- Finally, in Appendix B.3.4, we state and prove all auxiliary lemmas needed for the preceding results.

B.3.1 Consistency of MLE estimates

Theorem B.1. *Suppose, at sample size n, $\hat{\xi}_{n,\mathrm{OR}} := \hat{\xi}_{n,\mathrm{OR}}(\mathbf{X})$ be as defined in (4.5), and that $0 < \sigma_\star < \infty$. Then, it holds that*

$$|\hat{\xi}_{n,\mathrm{OR}} - \xi_n| = \mathrm{O}_P(1).$$

Proof. We start with recalling that

$$\hat{\xi}_{n,\mathrm{OR}} = \hat{\xi}_{n,\mathrm{OR}}(\mathbf{X}) \in \operatorname*{argmax}_{s\in[n-1]} L(s), \qquad \text{where } L(s) = \sum_{i=1}^{s} \ell(X_i).$$

For any $t > \xi_n$, we can write $L(t) = L(\xi_n) + \sum_{s=\xi_n+1}^{t} \ell(X_s)$. By a union bound, for any $M \in \mathbb{N}$,

$$\begin{aligned}\mathbb{P}(\,\hat{\xi}_{n,\mathrm{OR}} \geq \xi_n + M) &= \mathbb{P}\big(\cup_{t\,\geq\,\xi_n+M}\{L(t) \geq L(\xi_n)\}\big)\\ &\leq \sum_{t\,\geq\,\xi_n+M}^{n} \mathbb{P}\left(L(t) \geq L(\xi_n)\right) = \sum_{t\,\geq\,\xi_n+M}^{n} \mathbb{P}\Big(\sum_{s=\xi_n+1}^{t} \ell(X_s) \geq 0\Big).\end{aligned}$$

Observe that $\{X_s\}_{s=\xi_n+1}^{t}$ are i.i.d samples from P_1. Therefore, by Lemma B.4, there exists $\gamma > 0$ such that

$$\mathbb{P}(\,\hat{\xi}_{n,\mathrm{OR}} \geq \xi_n + M) \leq \sum_{t\,\geq\,\xi_n+M}^{n} e^{-\gamma\,(t-\xi_n)} \leq C_0\, C_1^M,$$

for some $C_0 > 0$ and $C_1 < 1$. Similarly, we can show that there exists $C_0' > 0$ and $C_1' < 1$ such that $\mathbb{P}(\,\hat{\xi}_{n,\mathrm{OR}} \leq \xi_n - M) \leq C_0'\, {C_1'}^M$. Together, this yields $|\hat{\xi}_{n,\mathrm{OR}} - \xi_n| = \mathrm{O}_P(1)$, as required. □

Theorem B.2. *Suppose, $\hat{\ell}_n$ satisfies (5.3). Then, $\hat{\xi}_n := \hat{\xi}_n(\mathbf{X})$, as defined in (5.2) satisfies*

$$|\hat{\xi}_n - \xi_n| = \mathrm{o}_P(\sqrt{n}).$$

Proof. We start by defining

$$S_k^{(1)} := \sum_{s=\xi_n+1}^{\xi_n+k} \hat{\ell}_n(X_s) - \ell(X_s), \qquad S_k^{(0)} := \sum_{s=1}^{k} \hat{\ell}_n(X_s) - \ell(X_s).$$

Next, fix $\delta > 0$ and let $\mathcal{A}$ be the event that

$$\max_{k=1,\ldots,n-\xi_n} |S_k^{(1)}| \leq (k + \sqrt{n}/\delta) \cdot \Gamma_1^{1/2}, \qquad \text{and} \qquad \max_{k=1,\ldots,\xi_n} |S_k^{(0)}| \leq (k + \sqrt{n}/\delta) \cdot \Gamma_0^{1/2}.$$

By Lemma B.5, note that $P(\mathcal{A}^c) \leq 4\delta$. Now, recall that

$$\hat{\xi}_n = \hat{\xi}_n(\mathbf{X}) \in \operatorname*{argmax}_{s\in[n-1]} \hat{L}(s), \qquad \hat{L}(s) := \sum_{i=1}^{s} \hat{\ell}_n(X_s).$$

For any $t > \xi_n$, write $\hat{L}(t) = \hat{L}(\xi_n) + \sum_{s=\xi_n+1}^{t} \hat{\ell}_n(X_s)$. Take any $\varepsilon > 0$. By a union bound,

$$\begin{aligned}
\mathbb{P}\,(\,\hat{\xi}_n \geq \xi_n + \varepsilon\sqrt{n}) &\leq \mathbb{P}\,(\cup_{t\,\geq\,\xi_n+\varepsilon\sqrt{n}}\{\hat{L}(t) \geq \hat{L}(\xi_n)\},\, \mathcal{A}) + P(\mathcal{A}^c) \\
&\leq \sum_{t\,\geq\,\xi_n+\varepsilon\sqrt{n}}^{n} \mathbb{P}\,(\hat{L}(t) \geq \hat{L}(\xi_n),\, \mathcal{A}) + 4\delta \\
&= \sum_{t\,\geq\,\xi_n+\varepsilon\sqrt{n}}^{n} \mathbb{P}\Big(\sum_{s=\xi_n+1}^{t} \hat{\ell}_n(X_s) \geq 0,\, \mathcal{A}\Big) + 4\delta.
\end{aligned}$$

Recall that $\{X_s\}_{s=\xi_n+1}^{t}$ are i.i.d samples from P_1. Therefore, for any $t \geq \xi_n + \varepsilon\sqrt{n}$,

$$\begin{aligned}
\mathbb{P}\Big(\sum_{s=\xi_n+1}^{t} \hat{\ell}_n(X_s) \geq 0,\, \mathcal{A}\Big) &\leq \mathbb{P}\Big(\sum_{s=\xi_n+1}^{t} \ell(X_s) \geq -|S^{(1)}_{t-\xi_n}|,\, \mathcal{A}\Big) \\
&\leq \mathbb{P}\Big(\frac{1}{t-\xi_n}\sum_{s=\xi_n+1}^{t} \ell(X_s) \geq -(1+\frac{1}{\varepsilon\delta})\cdot \Gamma_1^{1/2}\Big).
\end{aligned}$$

By (5.3), $\Gamma_1 \to 0$ as $n \to \infty$. Hnece, for sufficiently large n, by Lemma B.4, there exists $\gamma > 0$ such that for any $t \geq \xi_n + \varepsilon\sqrt{n}$,

$$\mathbb{P}\Big(\sum_{s=\xi_n+1}^{t} \hat{\ell}_n(X_s) \geq 0,\, \mathcal{A}\Big) \leq e^{-\gamma\,(t-\xi_n)} \leq e^{-\gamma\,\varepsilon\sqrt{n}}$$

Hence, it follows that for sufficiently large n,

$$\mathbb{P}(\,\hat{\xi}_n \geq \xi_n + \varepsilon\sqrt{n}) \leq n\, e^{-\gamma\,\varepsilon\sqrt{n}} + 4\delta.$$

Similarly, we can show that

$$\mathbb{P}(\,\hat{\xi}_n \geq \xi_n - \varepsilon\sqrt{n}) \leq n\, e^{-\gamma\,\varepsilon\sqrt{n}} + 4\delta.$$

Together, since ε and δ are arbitrary, this yields $|\hat{\xi}_{n,\mathrm{OR}} - \xi_n| = \mathrm{o}_P(\sqrt{n})$, as required. □

B.3.2 Asymptotic sharpness for oracle LLR score

We start by observing that, by symmetry, it suffices to show that for some $\kappa > 0$,

$$\max_{t_n \geq \xi_n+\kappa} \; p_{t_n,n} \xrightarrow{P} 0,$$

as $n \to \infty$. The proof is now split into three key steps. First, in Step 1, we relate the CONCH p-values to the 'frozen-estimator' $\tilde{p}$-values, and restrict to a high-probability, non-growing set for the changepoint $\hat{\xi}_n$. Next, in Step 2, we use a data-dependent bound on the $\tilde{p}$-values to reduce the problem to deriving the concentration of a few empirical terms. Finally, in Step 3, we show that one of these key empirical terms has a negative drift uniformly over all indices $t_n \geq \xi_n + \kappa$ with high probability and complete the proof.

Step 1: Relate to $\tilde{p}$-values, and restrict to a high-probability set for $\hat{\xi}_n$. First note that, by Lemma B.8, for each $i \in [n-1]$, $p_{i,n} \leq \tilde{p}_{i,n}$. Therefore, fix any $\eta, \delta \in (0,1)$, and note that it is enough to show that there exists a $\kappa > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{t_n \geq \xi_n + \kappa} \tilde{p}_{t_n,n} > \delta \right) \leq 5\eta.$$

For the oracle LLR score 4.4, we have $\hat{\ell}_n \equiv \ell$ for all $n \in \mathbb{N}$. Consequently,

$$\hat{\xi}_n = \hat{\xi}_{n,\mathrm{OR}}, \qquad \Gamma_0 = 0, \quad \Gamma_1 = 0,$$

where $\hat{\xi}_{n,\mathrm{OR}}$ is as in (4.5). Further, by Theorem B.1, $\hat{\xi}_{n,\mathrm{OR}}$ satisfies $|\hat{\xi}_{n,\mathrm{OR}} - \xi_n| = \mathrm{O}_P(1)$. Since $\xi_n/n \to \tau$, there exist $C_0 > 0$ and $N_0 \in \mathbb{N}$ such that for all $n > N_0$,

$$\mathbb{P}\left(|\hat{\xi}_n - \xi_n| > C_0 \right) \leq \eta, \qquad \frac{\xi_n}{n} \geq \tau^2.$$

Now suppose $n > N_0$ and $\kappa > C_0$. By the law of total probability,

$$\begin{aligned} \mathbb{P}\left(\max_{t_n \geq \xi_n + \kappa} \tilde{p}_{t_n,n} > \delta \right) &\leq \mathbb{P}\left(\max_{t_n \geq \xi_n + \kappa} \tilde{p}_{t_n,n} > \delta,\ |\hat{\xi}_{n,\mathrm{OR}} - \xi_n| \leq C_0 \right) + \eta \\ &\leq \sum_{k_n = \xi_n C_0}^{\xi_n + C_0} \mathbb{P}\left(\max_{t_n \geq \xi_n + \kappa} \tilde{p}_{t_n,n} > \delta,\ \hat{\xi}_{n,\mathrm{OR}} = k_n \right) + \eta. \end{aligned} \tag{B.4}$$

Hence, it suffices to upper bound each of the summands in (B.4).

Step 2: Data-dependent upper bound on $\tilde{p}$-values, and its implication. Fix any k_n such that $\xi_n - C_0 \leq k_n \leq \xi_n + C_0$. By Lemma B.6, for any $t_n \in [n-1]$,

$$\tilde{p}_{t_n,n} \leq \frac{n}{t_n} \frac{\hat{v}_n}{\hat{m}_n \, \Delta^2_{t_n,\hat{\xi}_n}} \mathbb{1}_{\Delta_{t_n,\hat{\xi}_n} < 0} + \mathbb{1}_{\Delta_{t_n,\hat{\xi}_n} \geq 0},$$

where

$$\Delta_{t_n,j} := \frac{1}{\hat{m}_n} \sum_{i=j+1}^{t_n} \hat{\ell}_n(X_i) - \hat{\mu}_{t_n,n,L}, \qquad \hat{m}_n := |t_n - \hat{\xi}_n|.$$

Since $\kappa > C_0$, on the event $\{\hat{\xi}_{n,\mathrm{OR}} = k_n\}$ we have

$$\hat{m}_n = t_n - k_n \geq (\kappa - C_0), \qquad \frac{n}{t_n} \leq \frac{n}{\xi_n} \leq \frac{1}{\tau^2}.$$

Therefore, on $\{\hat{\xi}_{n,\mathrm{OR}} = k_n\}$, the empirical bound on $\tilde{p}_{t_n,n}$ (uniform over $t_n \geq \xi_n + \kappa$) becomes

$$\max_{t_n \geq \xi_n + \kappa} \tilde{p}_{t_n,n} \leq \max_{t_n \geq \xi_n + \kappa} \left\{ \frac{1}{\tau^2} \frac{\hat{v}_n}{(\kappa - C_0)\, \Delta_{t_n,k_n}^2} \mathbb{1}_{\Delta_{t_n,k_n} < 0} \;+\; \mathbb{1}_{\Delta_{t_n,k_n} \geq 0} \right\}.$$

Furthermore, since $\epsilon_{2,n} = 0$ for all $n \in \mathbb{N}$, Lemma B.7 implies that with probability at least $1 - \eta$,

$$\hat{v}_n \leq \frac{2(\sigma_\star^2 + \mathrm{J}(P_0, P_1)^2)}{\eta}.$$

Therefore, using (B.4) and the fact that $\delta \in (0,1)$, a union bound yields

$$\begin{aligned}
&\mathbb{P}\left(\max_{t_n \geq \xi_n + \kappa} \; \tilde{p}_{t_n,n} > \delta \right) \\
&\quad \leq \sum_{k_n = \xi_n - C_0}^{\xi_n + C_0} \mathbb{P}\left(\max_{t_n \geq \xi_n + \kappa} \left\{ \frac{1}{\tau^2} \frac{2(\sigma_\star^2 + \mathrm{J}(P_0, P_1)^2)}{\eta(\kappa - C_0)\, \Delta_{t_n,k_n}^2} \mathbb{1}_{\Delta_{t_n,k_n} < 0} \;+\; \mathbb{1}_{\Delta_{t_n,k_n} \geq 0} \right\} > \delta,\; \hat{\xi}_{n,\mathrm{OR}} = k_n \right) + 2\eta \\
&\quad \leq \sum_{k_n = \xi_n - C_0}^{\xi_n + C_0} \mathbb{P}\left(\max_{t_n \geq \xi_n + \kappa} \Delta_{t_n,k_n} > -\left(\frac{2(\sigma_\star^2 + \mathrm{J}(P_0, P_1)^2)}{\tau^2 \delta \eta (\kappa - C_0)} \right)^{1/2} \right) + 2\eta. \qquad \text{(B.5)}
\end{aligned}$$

Hence, we only need to argue that with high probability, Δ_{t_n,k_n} has a negative drift uniformly over $t_n \geq \xi_n + \kappa$ for every $\xi_n - C_0 \leq k_n \leq \xi_n + C_0$.

Step 3: Proving uniform negative drift of Δ_{t_n,k_n}, and completing the proof. Fix any k_n such that $\xi_n - C_0 \leq k_n \leq \xi_n + C_0$. We can write

$$\Delta_{t_n,k_n} = \frac{1}{t_n - k_n} \sum_{i=k_n+1}^{t_n} \ell(X_i) - \hat{\mu}_{t_n,n,L} = \sum_{i=1}^{t_n} a_i\, \ell(X_i),$$

where

$$a_i := -\frac{1}{t_n} \mathbb{1}_{i \leq k_n} + \left(\frac{1}{t_n - k_n} - \frac{1}{t_n} \right) \mathbb{1}_{i > k_n}.$$

For such k_n, we may decompose Δ_{t_n,k_n} as

$$\Delta_{t_n,k_n} = -\frac{1}{t_n} \sum_{i=1}^{\xi_n - C_0} \ell(X_i) \;+\; \sum_{i=\xi_n - C_0}^{\xi_n + C_0} a_i \ell(X_i) \;+\; \frac{k_n}{(t_n - k_n) t_n} \sum_{i=\xi_n + C_0}^{t_n} \ell(X_i).$$

Now, we define

$$S_1 := -\min_{t_n \geq \xi_n + \kappa} \frac{1}{t_n} \sum_{i=1}^{\xi_n - C_0} \ell(X_i), \qquad S_2 := \sum_{i=\xi_n - C_0}^{\xi_n + C_0} \frac{1}{\kappa - C_0} |\ell(X_i)|,$$

$$\text{and} \quad S_3 := \max_{t_n \geq \xi_n + \kappa} \frac{k_n}{(t_n - k_n) t_n} \sum_{i=\xi_n + C_0}^{t_n} \ell(X_i).$$

Noting that $|a_i| \leq \frac{1}{t_n - k_n} \leq \frac{1}{\kappa - C_0}$, we obtain

$$\max_{t_n \geq \xi_n + \kappa} \Delta_{t_n, k_n} \leq \; S_1 \; + \; S_2 \; + \; S_3. \tag{B.6}$$

Let $N_1 \in N$ such that $N_1 > N_0$ and $N_1 > 2C_0/\tau^2$. Since $X_1, \dots, X_{\xi_n - C_0} \overset{iid}{\sim} P_0$ and $X_{\xi_n + C_0}, \dots, X_{t_n} \overset{iid}{\sim} P_1$, Lemma B.4 yields constants $c, \gamma > 0$ such that

$$\mathbb{P}\left(\frac{1}{\xi_n - C_0} \sum_{i=1}^{\xi_n - C_0} \ell(X_i) \geq c\, J(P_0, P_1)\right) \geq 1 - e^{-\gamma(\xi_n - C_0)},$$
$$\mathbb{P}\left(\frac{1}{t_n - \xi_n - C_0} \sum_{i=\xi_n + C_0}^{t_n} \ell(X_i) \leq -c\, J(P_0, P_1)\right) \geq 1 - e^{-\gamma(t_n - \xi_n - C_0)}.$$

These imply high-probability bounds on S_1 and S_3. For S_1, since $t_n \leq n$, with probability $1 - e^{-\gamma(\xi_n - C_0)}$, for $n > N_1$,

$$S_1 \leq -\frac{\xi_n - C_0}{n} c\, J(P_0, P_1) \leq -\frac{\tau^2}{2} c\, J(P_0, P_1)$$

For S_3, with probability at least $1 - e^{-\gamma(t_n - \xi_n - C_0)}$,

$$\frac{k_n}{(t_n - k_n)t_n} \sum_{i=\xi_n + C_0}^{t_n} \ell(X_i) \leq -\frac{k_n(t_n - \xi_n - C_0)}{(t_n - k_n)t_n} cJ(P_0, P_1) \leq 0.$$

Hence,

$$\mathbb{P}(S_3 \leq 0) \geq 1 - \sum_{t_n > \xi_n + \kappa} e^{-\gamma(t_n - \xi_n - C_0)} \geq 1 - \frac{e^{-\gamma(\kappa - C_0)}}{1 - e^{-\gamma}}.$$

Finally, by Markov's inequality and Cauchy–Schwarz, with probability at least $1 - \eta/C_0$,

$$\begin{aligned} S_2 \leq \frac{C_0}{\eta(\kappa - C_0)} \cdot \sum_{i=\xi_n - C_0}^{\xi_n + C_0} \mathbb{E}[|\ell(X_i)|] &\leq \frac{C_0}{\eta(\kappa - C_0)} \cdot \sum_{i=\xi_n - C_0}^{\xi_n + C_0} \mathbb{E}[|\ell(X_i)|^2])^{1/2} \\ &\leq \frac{C_0}{\eta(\kappa - C_0)} \sum_{i=\xi_n - C_0}^{\xi_n + C_0} (\sigma_\star^2 + \mathrm{J}^2(P_0, P_1))^{1/2} \\ &\leq \frac{2C_0^2(\sigma_\star^2 + \mathrm{J}^2(P_0, P_1))^{1/2}}{(\kappa - C_0)\eta}, \end{aligned}$$

Combining the bounds on S_1, S_2, and S_3, if

$$-\frac{\tau^2}{2} cJ(P_0, P_1) + \frac{2C_0^2(\sigma_\star^2 + \mathrm{J}^2(P_0, P_1))^{1/2}}{(\kappa - C_0)\eta} \leq -\left(\frac{4(\sigma_\star^2 + \mathrm{J}(P_0, P_1)^2)}{\tau^2 \delta \eta(\kappa - C_0)}\right)^{1/2}, \tag{B.7}$$

then it follows that

$$\max_{t_n \geq \xi_n + \kappa} \Delta_{t_n, k_n} \leq S_1 + S_2 + S_3 \leq -\left(\frac{4(\sigma_\star^2 + \mathrm{J}(P_0, P_1)^2)}{\tau^2 \delta \eta(\kappa - C_0)}\right)^{1/2}.$$

Hence, using (B.5) and a union bound,

$$\begin{aligned}\mathbb{P}\left(\max_{t_n \geq \xi_n + \kappa} \tilde{p}_{t_n, n} > \delta\right) &\leq \sum_{k_n = \xi_n - C_0}^{\xi_n + C_0} \mathbb{P}\left(\max_{t_n \geq \xi_n + \kappa} \Delta_{t_n, k_n} > -\left(\frac{4(\sigma_\star^2 + \mathrm{J}(P_0, P_1)^2)}{\tau^2 \delta \eta(\kappa - C_0)}\right)^{1/2}\right) + 2\eta \\ &\leq \sum_{k_n = \xi_n - C_0}^{\xi_n + C_0} \left(e^{-\gamma(\xi_n - C_0)} + \frac{e^{-\gamma(\kappa - C_0)}}{1 - e^{-\gamma}} + \frac{\eta}{C_0}\right) + 2\eta \\ &\leq 2C_0\, e^{-\gamma(n\tau^2 - C_0)} + \frac{2C_0}{1 - e^{-\gamma}}\, e^{-\gamma(\kappa - C_0)} + 4\eta,\end{aligned}$$

where the last step uses $\xi_n \geq \tau^2 n$. Now choose κ large enough (independet on n) so that

$$\text{(B.7) holds,} \qquad \text{and} \qquad \frac{2C_0}{1 - e^{-\gamma}}\, e^{-\gamma(\kappa - C_0)} \leq \eta.$$

For such a choice of κ,

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{t_n \geq \xi_n + \kappa} \tilde{p}_{t_n, n} > \delta\right) \leq 5\eta,$$

as required. □

B.3.3 Asymptotic sharpness of CONCH with learned LLR score

Theorem B.3. *In the setting of Theorem 5.2, the p-values $\{\tilde{p}_{t,n}\}$ defined in (B.3) satisfy*

$$\mathbb{E}\left[\frac{1}{n-1} \sum_{i=1}^{n-1} \tilde{p}_{i,n}\right] \longrightarrow 0 \qquad \text{as } n \to \infty.$$

Proof. We start with noting that it suffices to prove that

$$\lim_{n \to \infty} \mathbb{E}\left[\frac{1}{n-1} \sum_{i=1}^{n-1} \tilde{p}_{i,n}\right] \leq 7\eta,$$

for any $\eta \in (0, 1)$. Towards that, fix $\eta \in (0, 1)$. The proof is split into a few key steps. We first separate the indices that are far away from the changepoint from the ones that are in a $\sqrt{n}$ neighborhood. The latter contribute negligibly to the sum. Therefore, in Step 2, we first get a data-dependent upper bound on the p-values corresponding to 'far-indices'. In Step 3, we prove that one of the key terms in that upper bound has a negative drift, which then, in Step 4, allows us to put all the pieces together and argue that all these 'far-indices' p-values contribute at most $O(n^{1/2})$ to the sum.

Step 1: Split all indices into a $\sqrt{n}$-neighborhood around ξ_n, and its complement. By Theorem B.2, $|\hat{\xi}_n - \xi_n| = \mathrm{o}_P(n^{1/2})$. Moreover, by (5.3) and since $\xi_n/n \to \tau$, there exists a $N_0 \in \mathbb{N}$ such that for all $n \geq N_0$,

$$\tau^2 \leq \xi_n/n \leq \sqrt{\tau}, \qquad \mathbb{P}\left(|\hat{\xi}_n - \xi_n| > n^{1/2}\right) \leq \eta, \quad \text{and} \quad \frac{\Gamma_0}{\eta\,\mathbb{E}_{X\sim P_0}[\ell^2(X)]}, \frac{\Gamma_1}{\eta\,\mathbb{E}_{X\sim P_1}[\ell^2(X)]} \leq \frac{1}{4}.$$

Now, suppose $n > N_0$. Fix κ such that $\kappa - 1 > 2/\tau^2$. We partition all indices in $[n-1]$ into the 'far-indices'

$$\mathcal{I} := \{i : \ |i - \xi_n| \geq \kappa\, n^{1/2}\}$$

and its complement 'near-indices' $[n-1] \setminus \mathcal{I}$. Since $\tilde{p}_{i,n} \leq 1$ for all $i \in [n-1]$, by the tower law

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n-1}\tilde{p}_{i,n}\right] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i\in\mathcal{I}}\tilde{p}_{i,n}\right] + \mathbb{E}\left[\frac{1}{n-1}\sum_{i\in[n-1]\setminus\mathcal{I}}\tilde{p}_{i,n}\right] \\
&\leq \mathbb{E}\left[\frac{1}{n-1}\sum_{i\in\mathcal{I}}\tilde{p}_{i,n}\right] + \frac{2\kappa\, n^{1/2}+1}{n-1} \\
&\leq \mathbb{E}\Bigg[\frac{1}{n-1}\sum_{i\in\mathcal{I}}\tilde{p}_{i,n}\cdot \mathbb{1}\left\{|\hat{\xi}_n - \xi_n| \leq n^{1/2}\right\}\Bigg] + \eta + \frac{2\kappa\, n^{1/2}+1}{n-1}. \qquad \text{(B.8)}
\end{aligned}$$

Therefore, now it suffices to bound the first term in (B.8).

Step 2: Data-dependent upper bound on $\tilde{p}_{\cdot,n}$ for all far indices. Firstly, by Lemma B.6, we have that

$$\tilde{p}_{t_n,n} \leq \frac{n}{t_n}\,\frac{\hat{v}_n}{\hat{m}_n\,\Delta^2_{t_n,\tilde{\xi}_n}}\mathbb{1}_{\Delta_{t_n,\tilde{\xi}_n}<0} + \mathbb{1}_{\Delta_{t_n,\tilde{\xi}_n}\geq 0}, \qquad \text{(B.9)}$$

where we let

$$\Delta_{t_n,\hat{\xi}_n} := \frac{1}{\hat{m}_n}\sum_{i=\hat{\xi}_n+1}^{t_n}\hat{\ell}_n(X_i) - \hat{\mu}_{t_n,n,L}, \qquad \hat{m}_n := |t_n - \hat{\xi}_n|.$$

Now, we fix any 'far-index' $t_n \in \mathcal{I}$. Without loss of generality, we may assume $t_n > \xi_n$; the case $t_n < \xi_n$ is symmetric. First, observe that

$$\mathbb{E}\left[\tilde{p}_{t_n,n}\cdot \mathbb{1}\left\{|\hat{\xi}_n - \xi_n| \leq n^{1/2}\right\}\right] = \sum_{|k_n-\xi_n|\leq n^{1/2}}\mathbb{E}\Big[\tilde{p}_{t_n,n}\cdot\mathbb{1}\left\{\hat{\xi}_n = k_n\right\}\Big]. \qquad \text{(B.10)}$$

Therefore, it suffices to bound each of the summands from above. Fix a k_n such that $|k_n - \xi_n| \leq n^{1/2}$. Note that on the event $\{\hat{\xi}_n = k_n\}$, $\hat{m}_n = t_n - k_n$. Consequently, we have for all $n \geq N_0$,

$$\hat{m}_n \geq (\kappa-1)\, n^{1/2} \geq (2/\tau^2)n^{1/2}, \qquad \frac{n}{t_n} \leq \frac{n}{\xi_n} \leq \frac{1}{\tau^2}.$$

Therefore, it follows that

$$\mathbb{E}\Big[\tilde{p}_{t_n,n} \cdot \mathbb{1}\left\{\hat{\xi}_n = k_n,\, \mathcal{A}\right\}\Big] \;\le\; \mathbb{E}\left[\left(\frac{1}{2n^{1/2}}\, \frac{\hat{v}_n}{\Delta^2_{t_n,k_n}}\, \mathbb{1}_{\Delta_{t_n,k_n}<0,\, \mathcal{A}} + \mathbb{1}_{\Delta_{t_n,k_n}\ge 0}\right) \cdot \mathbb{1}\left\{\hat{\xi}_n = k_n,\, \mathcal{A}\right\}\right], \tag{B.11}$$

for any measurable $\mathcal{A}$. Now, if we can establish high-probability upper and lower bounds on $\hat{v}_n$ and $\Delta^2_{t_n,k_n}$, the upper bound on the expression on the left-hand side would follow. The following steps indeed follow this strategy.

Step 3: Proving negative drift of Δ_{t_n,k_n}. We write

$$\Delta_{t_n,k_n} = \frac{1}{t_n - k_n} \sum_{i=k_n+1}^{t_n} \hat{\ell}_n(X_i) - \hat{\mu}_{t_n,n,L} = \sum_{i=1}^{t_n} a_i\, \hat{\ell}_n(X_i),$$

where we let $a_i = -\frac{1}{t_n}\mathbb{1}_{i\le k_n} + \left(\frac{1}{t_n-k_n} - \frac{1}{t_n}\right)\mathbb{1}_{i>k_n}$ for all $i \in [t_n]$. Next, note that

$$\begin{aligned}\mathbb{P}&\left(\sum_{i=1}^{t_n} a_i\, \hat{\ell}_n(X_i) \;\ge\; -\frac{\tau^2}{8}\, \mathrm{J}(P_0,P_1)\right)\\ &\qquad\qquad \le \mathbb{P}\left(\sum_{i=1}^{t_n} a_i\, \ell(X_i) \;\ge\; -\frac{\tau^2}{8}\, \mathrm{J}(P_0,P_1) - \sum_{i=1}^{t_n} |a_i| \cdot |\ell(X_i) - \hat{\ell}_n(X_i)|\right).\end{aligned}$$

Next, we observe that

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{t_n} |a_i| \cdot |\ell(X_i) - \hat{\ell}_n(X_i)|\right] &\le \frac{1}{t_n}\sum_{i=1}^{t_n}\big(\mathbb{E}_{\substack{\mathcal{D}'_n, X\sim P_0,\\ X\perp \mathcal{D}'_n}}[|\hat{\ell}_n(X) - \ell(X)|] + \mathbb{E}_{\substack{\mathcal{D}'_n, X\sim P_1,\\ X\perp \mathcal{D}'_n}}[|\hat{\ell}_n(X) - \ell(X)|]\big)\\ &\quad + \frac{1}{t_n - k_n}\sum_{i=k_n+1}^{t_n}\big(\mathbb{E}_{\substack{\mathcal{D}'_n, X\sim P_0,\\ X\perp \mathcal{D}'_n}}[|\hat{\ell}_n(X) - \ell(X)|] + \mathbb{E}_{\substack{\mathcal{D}'_n X\sim P_1,\\ X\perp \mathcal{D}'_n,}}[|\hat{\ell}_n(X) - \ell(X)|]\big)\\ &\le 2 \max_{P=P_0,P_1} \mathbb{E}_{\substack{\mathcal{D}'_n, X\sim P,\\ X\perp \mathcal{D}'_n}}\Big[|\hat{\ell}_n(X) - \ell(X)|\Big].\end{aligned}$$

Therefore, by Markov's inequality, with probability at least $1-\eta$,

$$\sum_{i=1}^{t_n} |a_i| \cdot |\ell(X_i) - \hat{\ell}_n(X_i)| \le (2/\eta) \max_{P=P_0,P_1} \mathbb{E}_{\substack{\mathcal{D}'_n, X\sim P,\\ X\perp \mathcal{D}'_n}}\Big[|\hat{\ell}_n(X) - \ell(X)|\Big].$$

Further, by (5.3) and the Cauchy-Schwarz inequality, for $P \in \{P_0, P_1\}$,

$$\mathbb{E}_{\substack{\mathcal{D}'_n, X\sim P,\\ X\perp \mathcal{D}'_n}}\Big[|\hat{\ell}_n(X) - \ell(X)|\Big] \le \mathbb{E}_{\substack{\mathcal{D}'_n, X\sim P,\\ X\perp \mathcal{D}'_n}}\Big[|\hat{\ell}_n(X) - \ell(X)|^2\Big]^{1/2} = \mathrm{o}_P(1) \quad \text{as } n\to\infty.$$

Consequently, it follows that

$$\mathbb{P}\left(\sum_{i=1}^{t_n} |a_i| \cdot |\ell(X_i) - \hat{\ell}_n(X_i)| \le \frac{\tau^2}{8}\,\mathrm{J}(P_0, P_1)\right) \ge 1 - \eta,$$

and therefore, we have

$$\mathbb{P}\left(\sum_{i=1}^{t_n} a_i\,\hat{\ell}_n(X_i) \;\ge\; -\frac{\tau^2}{8}\,\mathrm{J}(P_0, P_1)\right) \le \mathbb{P}\left(\sum_{i=1}^{t_n} a_i\,\ell(X_i) \;\ge\; -\frac{\tau^2}{4}\,\mathrm{J}(P_0, P_1)\right) + \eta.$$

Now, we need to bound the probability term in the right. Note that $\mathbb{E}_{P_1}\ell(X) = -\mathrm{KL}(P_1\|P_0)$ and $\mathbb{E}_{P_0}\ell(X) = \mathrm{KL}(P_0\|P_1)$, and obtain

$$\mathbb{E}\left[\sum_{i=1}^{t_n} a_i\,\ell(X_i)\right] = \begin{cases} -\dfrac{\xi_n}{t_n}\,\mathrm{J}(P_0, P_1), & k_n > \xi_n, \\ -\dfrac{\xi_n}{t_n}\,\mathrm{J}(P_0, P_1) + \dfrac{(\hat{m}_n - m_n)_+}{\hat{m}_n}\,\mathrm{J}(P_0, P_1), & k_n \le \xi_n, \end{cases}$$

where $a_+ = \max\{a, 0\}$. Moreover,

$$\frac{(\hat{m}_n - m_n)_+}{\hat{m}_n} \le \frac{1}{\kappa - 1}, \qquad \mathbb{E}\left[\sum_{i=1}^{t_n} a_i\,\ell(X_i)\right] \le -\frac{\tau^2}{2}\,\mathrm{J}(P_0, P_1) < 0.$$

The variance of $\sum_{i=1}^{t_n} a_i\,\ell(X_i)$ is bounded by $\sigma_\star^2 \sum_{i=1}^{t_n} a_i^2$. Since $k_n/t_n \le 1$ and $\hat{m}_n < t_n$,

$$\sum_{i=1}^{t_n} a_i^2 = \frac{k_n}{t_n^2} + \frac{k_n^2}{\hat{m}_n t_n^2} \le \frac{1}{t_n} + \frac{1}{\hat{m}_n} \le \frac{2}{\hat{m}_n} \le \frac{\tau^2}{n^{1/2}}.$$

By Chebyshev's inequality, therefore with probability at least $1 - \eta$,

$$\sum_{i=1}^{t_n} a_i\,\ell(X_i) \;\le\; -\frac{\tau^2}{2}\,\mathrm{J}(P_0, P_1) + \tau\sigma_\star\sqrt{\frac{1}{\eta\, n^{1/2}}}.$$

Hence there exists $N_2 > N_1$ such that for all $n \ge N_2$,

$$\mathbb{P}\left(\Delta_{t_n,k_n} \;\le\; -\frac{\tau^2}{8}\,\mathrm{J}(P_0, P_1)\right) \ge 1 - 2\eta. \tag{B.12}$$

Step 4: Completing the proof. Now, we may put all the pieces together to get a high probability upper bound on $\tilde{p}_{t_n,n}$ for all 'far-indices' t_n. Suppose, $n > N_2$. By Lemma B.7 and noting that $\varepsilon_{2,n}/\eta \le 1/4$, we have

$$\mathbb{P}\left(\hat{v}_n \le \frac{8(\sigma_\star^2 + \mathrm{J}(P_0, P_1)^2)}{\eta}\right) \ge 1 - 4\eta.$$

Now, we take $\mathcal{A}$ to be the event that

$$\hat{v}_n \leq \frac{8(\sigma_\star^2 + \mathrm{J}(P_0, P_1)^2)}{\eta}, \qquad \Delta^2_{t_n,k_n} \geq \frac{\tau^4}{64}\,\mathrm{J}^2(P_0,P_1), \qquad \Delta_{t_n,k_n} < 0.$$

Hence, by (B.11), we have that

$$\mathbb{E}\Big[\tilde{p}_{t_n,n} \cdot \mathbb{1}\left\{\hat{\xi}_n = k_n,\, \mathcal{A}\right\}\Big] \;\leq\; \mathbb{E}\left[\left(\frac{1}{2n^{1/2}} \cdot \frac{8(\sigma_\star^2 + \mathrm{J}(P_0,P_1)^2)}{\eta} \cdot \frac{64}{\tau^4 \mathrm{J}^2(P_0,P_1)}\right) \cdot \mathbb{1}\left\{\hat{\xi}_n = k_n,\, \mathcal{A}\right\}\right].$$

Since $\tilde{p}_{t_n,n} \leq 1$, further

$$\mathbb{E}\Big[\tilde{p}_{t_n,n} \cdot \mathbb{1}\left\{\hat{\xi}_n = k_n\right\}\Big] \;\leq\; \frac{256\,(\sigma_\star^2 + \mathrm{J}^2(P_0,P_1))}{\tau^4\,\eta\,\mathrm{J}^2(P_0,P_1)\,n^{1/2}} \cdot \mathbb{P}(\hat{\xi}_n = k_n) + 6\eta,$$

where we note that, by union bound, $\mathbb{P}(\mathcal{A}) \geq 1 - 6\eta$. Consequently, by (B.10),

$$\mathbb{E}\left[\tilde{p}_{t_n,n} \cdot \mathbb{1}\left\{|\hat{\xi}_n - \xi_n| \leq n^{1/2}\right\}\right] \leq \frac{256\,(\sigma_\star^2 + \mathrm{J}^2(P_0,P_1))}{\tau^4\,\eta\,\mathrm{J}^2(P_0,P_1)\,n^{1/2}} + 6\eta.$$

Finally, the above inequality holds for any $t_n \in \mathcal{I}$. Therefore by (B.8),

$$\mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n-1} \tilde{p}_{i,n}\right] \leq \frac{256\,(\sigma_\star^2 + \mathrm{J}^2(P_0,P_1))}{\tau^4\,\eta\,\mathrm{J}^2(P_0,P_1)\,n^{1/2}} + 7\eta + \frac{2\kappa\, n^{1/2} + 1}{n-1}.$$

Consequently,

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n-1} \tilde{p}_{i,n}\right] \leq 7\eta,$$

as required.

□

Proof of Theorem 5.2 Fix $\eta > 0$. By Markov's inequality, we have that

$$\mathbb{P}\left(\frac{|\,\mathcal{C}^{\mathrm{CONCH}}_{n,1-\alpha}\,|}{n-1} \geq \eta\right) \leq \frac{\mathbb{E}\left[|\,\mathcal{C}^{\mathrm{CONCH}}_{n,1-\alpha}\,|\right]}{(n-1)\eta}.$$

Next, note that

$$\frac{1}{n-1}\,\mathbb{E}\left[|\,\mathcal{C}^{\mathrm{CONCH}}_{n,1-\alpha}\,|\right] \;=\; \frac{1}{n-1}\sum_{i=1}^{n-1} \mathbb{P}\left(p_{i,n} > \alpha\right).$$

By applying Markov's inequality again, we have

$$\frac{1}{n-1}\,\mathbb{E}\left[|\,\mathcal{C}^{\mathrm{CONCH}}_{n,1-\alpha}\,|\right] \leq \frac{1}{\alpha}\cdot\frac{1}{(n-1)}\sum_{i=1}^{n-1}\mathbb{E}[p_{i,n}] \leq \frac{1}{\alpha}\cdot\frac{1}{(n-1)}\sum_{i=1}^{n-1}\mathbb{E}[\tilde{p}_{i,n}],$$

where the last step follows by Lemma B.8. Hence, by Theorem B.3, it follows that

$$\mathbb{P}\left(\frac{|\,\mathcal{C}^{\text{CONCH}}_{n,1-\alpha}\,|}{n-1} \geq \eta\right) \longrightarrow 0, \qquad \text{as } n \to \infty.$$

Since η is arbitrary, the result follows. □

B.3.4 Auxiliary lemmas

Lemma B.4. *Fix $k \in \mathbb{N}$. Suppose, $X_1, \ldots, X_k \overset{iid}{\sim} P_1$ and $Y_1, \ldots, Y_k \overset{iid}{\sim} P_0$ and that $0 <$ $\mathrm{KL}(P_0\|P_1)$, $\mathrm{KL}(P_1\|P_0) < \infty$. Then, there exists $c > 0$ and $\gamma > 0$ such that*

$$\mathbb{P}\bigg(\frac{1}{k}\sum_{i=1}^{k}\ell(X_i) \geq -cJ(P_0, P_1)\bigg) \leq e^{-\gamma\, k}, \qquad \mathbb{P}\bigg(\frac{1}{k}\sum_{i=1}^{k}\ell(Y_i) \leq cJ(P_0, P_1)\bigg) \leq e^{-\gamma\, k}.$$

Proof. To prove the first part, we start with observing that for any $\theta \in (0, 1)$, we have that by Hölder's inequality,

$$M(\theta) := \mathbb{E}_{X\sim P_1}[e^{\theta\ell(X)}] = \int f_0^{\theta}(x) f_1^{1-\theta}(x)\, \mathsf{d}x \leq \left(\int f_0(x)\, \mathsf{d}x\right)^{\theta} \left(\int f_1(x)\, \mathsf{d}x\right)^{1-\theta} \leq 1.$$

Moreover, we have $M(0) = 1$ and

$$\lim_{\theta\to 0^+} M'(\theta) = \mathbb{E}_{X\sim P_1}[\ell(X)] = -\mathrm{KL}(P_1\|P_0) < 0.$$

Now, by Chernoff's bound, we have that for any $\theta \in (0, 1)$,

$$\begin{aligned}\mathbb{P}\bigg(\sum_{i=1}^{k}\ell(X_i) \geq -(k/2)\mathrm{KL}(P_1\|P_0)\bigg) &\leq \exp\left(\theta(k/2)\mathrm{KL}(P_1\|P_0)\right)\, \mathbb{E}_{X_1,\cdots,X_k\sim P_1}\left[\exp(\sum_{i=1}^{k}\theta\ell(X_i))\right] \\ &\leq \exp\left(k\log(M(\theta)) + \theta(k/2)\mathrm{KL}(P_1\|P_0)\right)\end{aligned}$$

We call $f(\theta) = \log(M(\theta)) + (\theta/2)\,\mathrm{KL}(P_1\|P_0)$ and note that $f(0) = 0$ and $\lim_{\theta\to 0^+} f'(\theta) < 0$, implying that there exists $\theta_\star \in (0, 1)$ such that $f(\theta_\star) = -\gamma$ with $\gamma > 0$. Consequently, it follows that

$$\mathbb{P}\bigg(\frac{1}{k}\sum_{i=1}^{k}\ell(X_i) \geq -(1/2)\,\mathrm{KL}(P_1\|P_0)\bigg) \leq e^{-\gamma\, k}$$

Finally, since $0 < \mathrm{KL}(P_1\|P_0)$, $\mathrm{KL}(P_1\|P_0) < \infty$, there exists $c < 1$ such that

$$\frac{\mathrm{KL}(P_1\|P_0)}{2} \geq c\, J(P_0, P_1).$$

Therefore, the first part follows. The second part follows by noting that $E_{Y\sim P_0}[e^{-\theta\ell(Y)}] \leq 1$ for all $\theta \in (0, 1)$, and the rest of the argument is similar to that of the first part. □

Lemma B.5. *Fix $\delta \in (0,1)$. Suppose $0 < \Gamma_0, \Gamma_1 < \infty$. Then, we have that*

$$\mathbb{P}\left(\max_{k \leq n-\xi_n} \left|\sum_{s=\xi_n+1}^{\xi_n+k} \hat{\ell}_n(X_s) - \ell(X_s)\right| \leq (k + \sqrt{n}/\delta) \cdot \Gamma_1^{1/2}\right) \geq 1 - 2\delta, \text{ and}$$

$$\mathbb{P}\left(\max_{k=1,\dots,\xi_n} \left|\sum_{s=1}^{k} \hat{\ell}_n(X_s) - \ell(X_s)\right| \leq (k + \sqrt{n}/\delta) \cdot \Gamma_0^{1/2}\right) \geq 1 - 2\delta.$$

Proof. We define the partial sum

$$S_k^{(1)} := \sum_{s=\xi_n+1}^{\xi_n+k} \hat{\ell}_n(X_s) - \ell(X_s),$$

for $k = 1, \dots, n - \xi_n$. Note that conditional on $\mathcal{D}_n'$, each of the summands has variance bounded by $\mathbb{E}_{X\sim P_1}[|\hat{\ell}_n(X) - \ell(X)|^2 \mid \mathcal{D}_n']$. Fix any $\delta \in (0,1)$, and note that by Kolmogorov's inequality,

$$\mathbb{P}\left(\max_{k \leq n-\xi_n} |S_k^{(1)} - E[S_k^{(1)} \mid \mathcal{D}_n']| \geq \frac{\left(n\,\mathbb{E}_{X\sim P_1}[|\hat{\ell}_n(X) - \ell(X)|^2 \mid \mathcal{D}_n']\right)^{1/2}}{\sqrt{\delta}} \,\middle|\, \mathcal{D}_n'\right) \leq \delta.$$

Moreover, by the triangle inequality and the Cauchy-Schwarz inequality.

$$\begin{aligned}|S_k^{(1)} - E[S_k^{(1)} \mid \mathcal{D}_n']| &\leq |S_k^{(1)}| + k\,|\mathbb{E}_{X\sim P_1}[\hat{\ell}_n(X) - \ell(X) \mid \mathcal{D}_n']| \\ &\leq |S_k^{(1)}| + k\left(\mathbb{E}_{X\sim P_1}[|\hat{\ell}_n(X) - \ell(X)|^2 \mid \mathcal{D}_n']\right)^{1/2}.\end{aligned}$$

By Markov's inequality, further with probability at least $1 - \delta$,

$$\mathbb{E}_{X\sim P_1}[|\hat{\ell}_n(X) - \ell(X)|^2 \mid \mathcal{D}_n'] \leq \frac{\mathbb{E}_{X\sim P_1}[|\hat{\ell}_n(X) - \ell(X)|^2]}{\delta} = \frac{\Gamma_1^{1/2}}{\delta}.$$

Therefore, in aggregate,

$$\mathbb{P}\left(\max_{k \leq n-\xi_n} |S_k^{(1)}| \leq (k + \sqrt{n}/\delta) \cdot \Gamma_1^{1/2}\right) \geq 1 - 2\delta.$$

This concludes the proof of the first part. The second part follows likewise. □

Lemma B.6. *For $k \in [n-1]$, we have deterministically*

$$\tilde{p}_{k,n} \leq \left(\frac{n}{\max\{k, n-k\}} \cdot \frac{\hat{v}_n}{|k - \hat{\xi}_n|\,\Delta_{k,\hat{\xi}_n}^2} \cdot \mathbb{1}_{\Delta_{k,\hat{\xi}_n} < 0}\right) + \mathbb{1}_{\Delta_{k,\hat{\xi}_n} \geq 0}, \tag{B.13}$$

where

$$\Delta_{k,\hat{\xi}_n} := \frac{S_k^{(n)}(\mathbf{X})}{|k - \hat{\xi}_n|} - \hat{\mu}_{k,n,L}\mathbb{1}_{k \geq \hat{\xi}_n} - \hat{\mu}_{k,n,R}\mathbb{1}_{k < \hat{\xi}_n}.$$

Proof. Fix $k \in [n-1]$ and set $\hat{m}_n = |k - \hat{\xi}_n|$. Without loss of generality, assume $k > \hat{\xi}_n$. Then

$$S_k^{(n)}(\mathbf{x}) = \sum_{i=\hat{\xi}_n(\mathbf{x})+1}^{k} \hat{\ell}_n(x_i),$$

a sum of $\hat{\ell}_n$ over a block of $\hat{m}_n$ observations. Let $\mathcal{X}_L := \{X_1, \dots, X_k\}$ be the multiset of the first k observations. The permutation p-value (B.3) is given by

$$\tilde{p}_{k,n} := \mathbb{P}_{\pi \sim \mathrm{Unif}(\Pi_k)} \Big(S_k^{(n)}(\pi(\mathbf{X})) \le S_k^{(n)}(\mathbf{X}) \;\Big|\; \mathbf{X} \Big).$$

Since $k \ge \hat{\xi}_n$, sampling π uniformly from Π_k is equivalent to sampling without replacement (WOR) from $\mathcal{X}_L$. Thus,

$$\tilde{p}_{k,n} = \mathbb{P}\left(\frac{1}{\hat{m}_n} \sum_{j=1}^{\hat{m}_n} \hat{\ell}_n(\tilde{X}_j) \le \frac{S_k^{(n)}(\mathbf{X})}{\hat{m}_n} \;\middle|\; \mathbf{X} \right),$$

where $\tilde{X}_1, \dots, \tilde{X}_{\hat{m}_n}$ are drawn WOR from $\mathcal{X}_L$. Hence,

$$\mathbb{E}\left[\frac{1}{\hat{m}_n} \sum_{j=1}^{\hat{m}_n} \hat{\ell}_n(\tilde{X}_j) \;\middle|\; \mathbf{X} \right] = \hat{\mu}_{k,n,L}, \quad \mathrm{Var}\left(\frac{1}{\hat{m}_n} \sum_{j=1}^{\hat{m}_n} \hat{\ell}_n(\tilde{X}_j) \;\middle|\; \mathbf{X} \right) = \frac{v_{k,n}}{\hat{m}_n} \cdot \frac{n - \hat{m}_n}{n-1},$$

where $v_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \hat{\ell}_n^2(X_i) - (\hat{\mu}_{k,n,L})^2$. Note that

$$v_{k,n} \le \frac{1}{k} \sum_{i=1}^{k} \hat{\ell}_n^2(X_i) \le \frac{n}{k}\, \hat{v}_n.$$

Now, we can write $\tilde{p}_{k,n}$ as

$$\mathbb{P}\left(\frac{1}{\hat{m}_n} \sum_{j=1}^{\hat{m}_n} \hat{\ell}_n(\tilde{X}_j) \le \frac{S_k^{(n)}(\mathbf{X})}{\hat{m}_n} \;\middle|\; \mathbf{X} \right) \;\le\; \mathbb{P}\left(\frac{1}{\hat{m}_n} \sum_{j=1}^{\hat{m}_n} \hat{\ell}_n(\tilde{X}_j) - \hat{\mu}_{k,n,L} \le \frac{S_k^{(n)}(\mathbf{X})}{\hat{m}_n} - \hat{\mu}_{k,n,L} \;\middle|\; \mathbf{X} \right).$$

Observe that $\Delta_{k,\hat{\xi}_n} = \frac{S_k^{(n)}(\mathbf{X})}{\hat{m}_n} - \hat{\mu}_{k,n,L}$, and if $\Delta_{k,\hat{\xi}_n} < 0$, then by Chebyshev's inequality,

$$\tilde{p}_{k,n} \le \frac{v_{k,n}}{\hat{m}_n\, \Delta_{k,\hat{\xi}_n}^2} \cdot \frac{n - \hat{m}_n}{n-1} \;\le\; \frac{n}{k}\, \frac{\hat{v}_n}{\hat{m}_n\, \Delta_{k,\hat{\xi}_n}^2}.$$

On the other hand, if $\Delta_{k,\hat{\xi}_n} \ge 0$, then $\tilde{p}_{t_n,n} \le 1$. This proves the lemma. □

Lemma B.7. *In the setting of Section 5, suppose* $0 < \sigma_\star < \infty$. *Then, for any* $\eta \in (0,1)$,

$$\mathbb{P}\left(\hat{v}_n \le \frac{2(\sigma_\star^2 + \mathrm{J}^2(P_0, P_1))(1 + \varepsilon_{2,n}/\eta + 2\sqrt{\varepsilon_{2,n}/\eta}\,)}{\eta} \right) \ge 1 - 4\eta,$$

where we let

$$\varepsilon_{2,n} = \max\left\{\frac{\Gamma_0^2}{\mathbb{E}_{X\sim P_0}[\ell^2(X)]}, \frac{\Gamma_1^2}{\mathbb{E}_{X\sim P_1}[\ell^2(X)]}\right\}.$$

Proof. Fix $\eta \in (0,1)$. We can write

$$\begin{aligned}\hat{v}_n &= \frac{\xi_n}{n}\cdot\frac{1}{\xi_n}\sum_{i=1}^{\xi_n}\hat{\ell}_n^2(X_i) \;+\; \frac{n-\xi_n}{n}\cdot\frac{1}{n-\xi_n}\sum_{i=\xi_n+1}^{n}\hat{\ell}_n^2(X_i)\\ &\le \frac{1}{\xi_n}\sum_{i=1}^{\xi_n}\hat{\ell}_n^2(X_i) \;+\; \frac{1}{n-\xi_n}\sum_{i=\xi_n+1}^{n}\hat{\ell}_n^2(X_i).\end{aligned}$$

We first derive an upper bound on the first term. Recall that conditional on $\mathcal{D}_n'$, $\hat{\ell}_n^2(X_1),\dots,\hat{\ell}_n^2(X_{\xi_n})$ are i.i.d. with mean $\mathbb{E}_{X\sim P_0}[\hat{\ell}_n^2(X)\mid\mathcal{D}_n']$. Moreover, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}_{X\sim P_0}[\hat{\ell}_n^2(X)\mid\mathcal{D}_n'] &= \mathbb{E}_{X\sim P_0}\left[(\ell(X)+|\hat{\ell}_n(X)-\ell(X)|)^2\mid\mathcal{D}_n'\right]\\ &\le \mathbb{E}_{X\sim P_0}[\ell^2(X)] + \mathbb{E}_{X\sim P_0}[|\hat{\ell}_n(X)-\ell(X)|^2\mid\mathcal{D}_n']\\ &\qquad + 2\mathbb{E}_{X\sim P_0}[|\ell(X)|\,|\hat{\ell}_n(X)-\ell(X)|\mid\mathcal{D}_n']\\ &\le \mathbb{E}_{X\sim P_0}[\ell^2(X)] + \mathbb{E}_{X\sim P_0}[|\hat{\ell}_n(X)-\ell(X)|^2\mid\mathcal{D}_n']\\ &\qquad + 2(\mathbb{E}_{X\sim P_0}[|\ell(X)|^2])^{1/2}\,(\mathbb{E}_{X\sim P_0}[|\hat{\ell}_n(X)-\ell_n(X)|^2\mid\mathcal{D}_n'])^{1/2}.\end{aligned}$$

By Markov's inequality, with probability at least $1-\eta$,

$$\mathbb{E}_{X\sim P_0}[\hat{\ell}_n^2(X)\mid\mathcal{D}_n'] \le \mathbb{E}_{X\sim P_0}[\ell^2(X)](1+\varepsilon_{2,n}/\eta+2\sqrt{\varepsilon_{2,n}/\eta}\,).$$

Further, we observe that

$$\mathbb{E}_{X\sim P_0}[\ell^2(X)] = \mathrm{Var}_{X\sim P_0}(\ell(X)) + (\mathbb{E}_{X\sim P_0}[\ell(X)])^2 = \sigma_0^2 + \mathrm{KL}^2(P_0\|P_1).$$

By Markov's inequality, therefore

$$\mathbb{P}\left(\frac{1}{\xi_n}\sum_{i=1}^{\xi_n}\ell^2(X_i) \ge \frac{(\sigma_0^2+\mathrm{KL}^2(P_0\|P_1))(1+\varepsilon_{2,n}/\eta+2\sqrt{\varepsilon_{2,n}/\eta}\,)}{\eta}\right) \le 2\eta.$$

Similarly, we can show that

$$\mathbb{P}\left(\frac{1}{n-\xi_n}\sum_{i=\xi_n+1}^{n}\delta^2(X_i) \ge \frac{(\sigma_1^2+\mathrm{KL}^2(P_1\|P_0))(1+\varepsilon_{2,n}/\eta+2\sqrt{\varepsilon_{2,n}/\eta}\,)}{\eta}\right) \le 2\eta.$$

Since $\sigma_0, \sigma_1 \le \sigma_\star$ and $\mathrm{KL}(P_0\|P_1)$, $\mathrm{KL}(P_1\|P_0) \le \mathrm{J}(P_0,P_1)$, a union bound therefore gives us

$$\mathbb{P}\left(\hat{v}_n \le \frac{2(\sigma_\star^2+\mathrm{J}^2(P_0,P_1))(1+\varepsilon_{2,n}/\eta+2\sqrt{\varepsilon_{2,n}/\eta}\,)}{\eta}\right) \ge 1-4\eta,$$

as required. □

Lemma B.8. *Let $\{p_{t,n}\}$ be the CONCH p-values computed based on CPP score (5.1). Then, it holds that deterministically,*

$$p_{t,n} \ \leq \ \tilde{p}_{t,n}, \qquad \textit{for} \ \ t \in [n-1].$$

Proof. We start with observing that the original score $S_t(\mathbf{X})$ and frozen score $S_t^{(n)}(\mathbf{X})$ can be equivalently written as

$$S_t(\mathbf{x}) \ = \ \sum_{i=1}^{t} \hat{\ell}_n(x_i) - \sum_{i=1}^{\hat{\xi}_n(\mathbf{x})} \hat{\ell}_n(x_i), \qquad S_t^{(n)}(\mathbf{x}) \ = \ \sum_{i=1}^{t} \hat{\ell}_n(x_i) - \sum_{i=1}^{\hat{\xi}_n} \hat{\ell}_n(x_i),$$

where $\hat{\xi}_n = \hat{\xi}_n(\mathbf{X})$. Recall that by (5.2), for any $\mathbf{x} \in \mathcal{X}^n$,

$$\hat{\xi}_n(\mathbf{x}) \in \operatorname*{argmax}_{s \in [n-1]} \sum_{i=1}^{s} \hat{\ell}_n(x_s).$$

Hence, for any permutation π, we have that

$$\sum_{i=1}^{\hat{\xi}_n(\pi(\mathbf{X}))} \hat{\ell}_n(\pi(X_s)) \ \geq \sum_{i=1}^{\hat{\xi}_n(\mathbf{X})} \hat{\ell}_n(\pi(X_s)).$$

Consequently, for any $t \in [n-1]$ and any permutation π of $[n]$,

$$S_t(\pi(\mathbf{X})) \leq S_t^{(n)}(\pi(\mathbf{X})), \qquad S_t(\mathbf{X}) = S_t^{(n)}(\mathbf{X}).$$

Hence, for all $t \in [n-1]$, deterministically, $p_{t,n} \leq \tilde{p}_{t,n}$ as required. □

B.4 Proof of Theorem 6.1: Universality Theorem

The proof of this universality theorem is inspired by the classical universality result for full-conformal procedures in the predictive inference framework (see Vovk et al., 2005, Chapter 2.4; Angelopoulos et al., 2024, Theorem 9.6).

Fix $n \in \mathbb{N}$. First, given the confidence set C, we consider the CPP score

$$S_t(\mathbf{x}) = \mathbb{1}\{t \in C(\mathbf{x})\} \in \{0, 1\},$$

for any $t \in [n-1]$. We will show that the CONCH confidence set constructed from this score, denoted $\mathcal{C}_{1-\alpha}^{\text{CONCH}}$, coincides exactly with the given confidence set C.

We start with showing that $\mathcal{C}_{1-\alpha}^{\text{CONCH}}(\mathbf{X}) \supseteq C(\mathbf{X})$; that is, if $t \in C(\mathbf{X})$, then it holds that $p_t > \alpha$, where p_t is as defined in (3.2). This is immediate by observing that if $t \in C(\mathbf{X})$, then $S_t(\mathbf{X}) = 1$, and consequently,

$$p_t = \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\{S_t(\pi(\mathbf{X})) \le S_t(\mathbf{X})\} = \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\{S_t(\pi(\mathbf{X})) \le 1\} = 1.$$

Next, we show that $\mathcal{C}_{1-\alpha}^{\mathrm{CONCH}}(\mathbf{X}) \subseteq C(\mathbf{X})$, i.e., if $t \notin C(\mathbf{X})$, then $p_t \le \alpha$. To that end, we first claim that for any $t \in [n-1]$ and any vector $\mathbf{x} \in \mathcal{X}^n$,

$$\frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\{t \in C(\pi(\mathbf{x}))\} \ge 1 - \alpha. \tag{B.14}$$

We now prove this claim. Fix $t \in [n-1]$, and sample π uniformly from the set of permutations Π_t. Define $\tilde{\mathbf{X}} := (\tilde{X}_1, \ldots, \tilde{X}_n) := \pi(\mathbf{x})$. Conditional on the multisets $\{x_1, \ldots, x_t\}$ and $\{x_{t+1}, \ldots, x_n\}$, we have

$$\tilde{X}_1, \ldots, \tilde{X}_t \text{ are exchangeable, and } \tilde{X}_{t+1}, \ldots, \tilde{X}_n \text{ are exchangeable.}$$

Moreover, conditional on the multisets, $(\tilde{X}_1, \ldots, \tilde{X}_t)$ and $(\tilde{X}_{t+1}, \ldots, \tilde{X}_n)$ are independent, implying that the sampling process of $\tilde{\mathbf{X}}$ satisfies Assumption 1. Consequently,

$$\mathbb{P}_{\pi \sim \mathrm{Unif}(\Pi_t)}\big(t \in C(\tilde{\mathbf{X}}) \,\big|\, \{x_1, \ldots, x_t\}, \{x_{t+1}, \ldots, x_n\}\big) \ge 1 - \alpha,$$

or equivalently, (B.14) holds.

Returning to the main proof, observe that if $t \notin C(\mathbf{X})$, then $S_t(\mathbf{X}) = 0$. Consequently,

$$\begin{aligned} p_t &= \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\{S_t(\pi(\mathbf{X})) \le S_t(\mathbf{X})\} \\ &= \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\{S_t(\pi(\mathbf{X})) \le 0\} = \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\{t \notin C(\pi(\mathbf{X}))\} \le \alpha, \end{aligned}$$

where the last step follows from (B.14). This completes the proof. □

B.5 Proving asymptotic validity of CONCH-SEG

In this section, we establish the asymptotic validity for the extension of CONCH to multiple changepoint localization. A direct analysis of CONCH-SEG from Section 7 is hard because segmentation and CONCH are applied to the same observations, potentially violating Assumption 1. To decouple these steps, we introduce a close variant of the same, CONCH-SEG-`crossfit`, formally given in Algorithm 5.

In particular, we partition the index set into two disjoint folds

$$\mathcal{I}_1 := \{t \in [n] : t \text{ odd}\}, \quad \text{and} \quad \mathcal{I}_2 := \{t \in [n] : t \text{ even}\}.$$

Algorithm 5: CONCH-SEG-crossfit

```
Input: $(X_t)_{t=1}^n$ (data); $S : \bigcup_{m\in\mathbb{N}} \mathcal{X}^m \to \mathbb{R}^m$ (CPP score function); segmentation
       algorithm SEG
Output: $\mathcal{C}_{1-\alpha}^{\text{CONCH-SEG-crossfit}}$
$\mathcal{I}_1 \leftarrow \{t \le n : \ t \text{ odd}\}, \ \mathcal{I}_2 \leftarrow \{t \le n : \ t \text{ even}\}$;
$\mathcal{C} \leftarrow \varnothing$;
for $r \in \{1, 2\}$ do
    $(\hat{K}^{(r)}, \hat{\xi}_1^{(r)}, \ldots, \hat{\xi}_{\hat{K}^{(r)}}^{(r)}) \leftarrow \texttt{SEG}((X_t)_{t\in\mathcal{I}_r})$;
    Compute $(\tilde{X}_0^{(r)}, \ldots, \tilde{X}_{\hat{K}^{(r)}}^{(r)})$ by (7.2) based on $(\hat{\xi}_1^{(r)}, \ldots, \hat{\xi}_{\hat{K}^{(r)}}^{(r)})$;
    $J_\ell^{(r)} \leftarrow [\,\tilde{X}_{\ell-1}^{(r)}, \, \tilde{X}_\ell^{(r)}\,] \cap \mathcal{I}_{3-r}$ for $\ell \in [\hat{K}^{(r)}]$;
    for $\ell \in [\hat{K}^{(r)}]$ do
        Let $X^{(r,\ell)}$ be the subsequence $(X_t)_{t\in J_\ell^{(r)}}$ ordered by increasing index $t$;
        Define $S^{(r,\ell)} : \mathcal{X}^{|J_\ell^{(r)}|} \to \mathbb{R}^{|J_\ell^{(r)}|-1}$;
        Compute CONCH $p$-values $\{p_t : t \in J_\ell^{(r)} \setminus \{\max J_\ell^{(r)}\}\}$ as in (3.2) using score
         $S^{(r,\ell)}$ on $X^{(r,\ell)}$;
        $\mathcal{C}_\ell^{(r)} \leftarrow \{\, t \in J_\ell^{(r)} : p_t > \alpha \,\}$;
        $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}_\ell^{(r)}$;
    end
end
return $\mathcal{C}_{1-\alpha}^{\textit{CONCH-SEG-crossfit}} \leftarrow \mathcal{C}$
```

For $r \in \{1, 2\}$, we run the segmentation algorithm on $\mathcal{I}_r$ to obtain $\hat{K}^{(r)}$ and

$$0 = \hat{\xi}_0^{(r)} < \hat{\xi}_1^{(r)} < \cdots < \hat{\xi}_{\hat{K}^{(r)}}^{(r)} < n = \hat{\xi}_{\hat{K}^{(r)}+1}^{(r)}.$$

Next, we form the segment boundaries as in (7.2), based on the sequence $(\hat{\xi}_0^{(r)}, \ldots, \hat{\xi}_{\hat{K}^{(r)}}^{(r)})$, to obtain

$$1 = \tilde{X}_0^{(r)} < \tilde{X}_1^{(r)} < \cdots < \tilde{X}_{\hat{K}^{(r)}-1}^{(r)} < \tilde{X}_{\hat{K}^{(r)}}^{(r)} = n,$$

and define disjoint segments $J_\ell^{(r)} := [\,\tilde{X}_{\ell-1}^{(r)}, \, \tilde{X}_\ell^{(r)}\,] \cap \mathcal{I}_{3-r}$ for $\ell \in [\hat{K}^{(r)}]$. We then run CONCH on the restricted segment $J_\ell^{(r)}$ to produce a segmentwise confidence set, and aggregate across segments and folds. In particular, when segmentation is performed on $\mathcal{I}_1$ (respectively, $\mathcal{I}_2$), CONCH is run on $\mathcal{I}_2$ (respectively, $\mathcal{I}_1$). This construction restores the exchangeability structure required for validity, as established in the following theorem.

Theorem B.9 (Asymptotic coverage of CONCH-SEG-crossfit)**.** *Fix $\alpha \in (0, 1)$ and consider the multiple-changepoint model in (7.1). Suppose the segmentation algorithm, for each $r \in \{1, 2\}$, satisfies:*

(a) ***Consistent changepoint count.*** $\mathbb{P}(\hat{K}^{(r)} = K) \to 1$ *as* $n \to \infty$.

(b) ***Vanishing normalized localization error.***

$$\max_{k\in[\hat{K}^{(r)}]} \min_{j\in[K]} \frac{|\hat{\xi}_k^{(r)} - \xi_j|}{\min_{j\in[K-1]}(\xi_{j+1}-\xi_j)} \xrightarrow{P} 0 \quad \textit{as } n \to \infty.$$

(c) ***No clustering of estimated changepoints.*** *There exists $\eta > 0$, such that*

$$\mathbb{P}\left(\frac{\min_{k\in[\hat{K}^{(r)}-1]}(\hat{\xi}_{k+1}^{(r)} - \hat{\xi}_k^{(r)})}{\min_{j\in[K-1]}(\xi_{j+1}-\xi_j)} > \eta \right) \to 1 \quad \textit{as } n \to \infty.$$

Then, for any $k \in [K]$,

$$\mathbb{P}(\xi_k \in \mathcal{C}_{1-\alpha}^{\textit{CONCH-SEG-crossfit}}) \geq 1 - \alpha - \mathrm{o}(1) \qquad \textit{as } n \to \infty,$$

where the probability is taken under the model class (7.1).

Proof. Fix $k \in [K]$ and, without loss of generality, assume ξ_k is odd (the even case is symmetric). Recall that in the cross-fitted construction, fold $r = 2$ applies the segmentation algorithm to the even indices and runs CONCH on the odd indices. Define the event

$$\mathcal{G} := \left\{ \hat{K}^{(2)} = K, \;\; \xi_{k-1} < \tilde{X}_{k-1}^{(2)} < \xi_k < \tilde{X}_k^{(2)} < \xi_{k+1} \right\},$$

namely, the (even-fold) midpoint boundaries $\tilde{X}_{k-1}^{(2)}$ and $\tilde{X}_k^{(2)}$ lie on opposite sides of the true changepoint ξ_k. On $\mathcal{G}$, ξ_k falls inside $[\tilde{X}_{k-1}^{(2)}, \tilde{X}_k^{(2)}]$, so by construction,

$$\{\xi_k \in \mathcal{C}_{1-\alpha}^{\text{CONCH-SEG-}\texttt{crossfit}}\} = \{\xi_k \in \mathcal{C}_k^{(2)}\},$$

where $\mathcal{C}_k^{(2)}$ is the CONCH confidence set computed on $J_k^{(2)}$, the odd-index subsequence of $[\tilde{X}_{k-1}^{(2)}, \tilde{X}_k^{(2)}]$.

Hence, by the law of total probability and tower law, we have

$$\begin{aligned} \mathbb{P}(\xi_k \in \mathcal{C}_{1-\alpha}^{\text{CONCH-SEG-}\texttt{crossfit}}) &\geq \mathbb{P}(\xi_k \in \mathcal{C}_k^{(2)} \mid \mathcal{G})\,\mathbb{P}(\mathcal{G}) \\ &= \mathbb{E}\Big[\mathbb{P}(\xi_k \in \mathcal{C}_k^{(2)} \mid \mathcal{G}, (X_t)_{t\in\mathcal{I}_2}) \;\Big|\; \mathcal{G}\Big]\,\mathbb{P}(\mathcal{G}) \end{aligned} \tag{B.15}$$

Conditional on $\mathcal{G}$ and $(X_t)_{t\in\mathcal{I}_2}$, the CONCH algorithm is run on an independent set of observations with indices in $J_k^{(2)}$. Therefore, $(X_t)_{t\in J_k^{(2)}}$ satisfies Assumption 1 with a single changepoint at ξ_k. Consequently, by Theorem 3.1, we have

$$\mathbb{P}(\xi_k \in \mathcal{C}_k^{(2)} \mid \mathcal{G}, (X_t)_{t\in\mathcal{I}_2}) \geq 1 - \alpha \quad \text{almost surely.}$$

Plugging this into (B.15) yields

$$\mathbb{P}(\xi_k \in \mathcal{C}_{1-\alpha}^{\text{CONCH-SEG-}\texttt{crossfit}}) \geq (1-\alpha)\mathbb{P}(\mathcal{G}) \geq (1-\alpha) - \mathbb{P}(\mathcal{G}^c).$$

Now, it remains to show $\mathbb{P}(\mathcal{G}) \to 1$. To that end, let $\Delta := \min_{j \in [K-1]}(\xi_{j+1} - \xi_j)$ and define the event

$$\mathcal{A} := \Big\{\hat{K}^{(2)} = K,\ \min_{k \in [\hat{K}^{(2)}-1]} (\hat{\xi}_{k+1}^{(2)} - \hat{\xi}_k^{(r)}) > \eta\Delta \\ \text{and } \max_{j \in [K]} \min_{i \in [K]} |\hat{\xi}_j^{(2)} - \xi_i| \leq \min\{\Delta/8, \eta\Delta/2\}\Big\}.$$

By the theorem hypothesis, $\mathbb{P}(\mathcal{A}) \to 1$. Now on $\mathcal{A}$, for each j choose $i_j \in [K]$ attaining the inner minimum, i.e.,

$$\min_{i \in [K]} |\hat{\xi}_j^{(2)} - \xi_i| \ = \ |\hat{\xi}_j^{(2)} - \xi_{i_j}|.$$

Firstly observe that $i_{j+1} \neq i_j$ for all $j \in [K]$. If that's not the case, then

$$\eta\Delta < \hat{\xi}_{j+1}^{(2)} - \hat{\xi}_j^{(2)} \leq |\hat{\xi}_{j+1}^{(2)} - \xi_{i_{j+1}}| + |\hat{\xi}_j^{(2)} - \xi_{i_j}| \leq \eta\Delta,$$

leading to a contradiction.

We further claim that $i_1 < i_2 < \ldots < i_K$. If instead, $i_{j+1} < i_j$ for some $j \in [K-1]$, then

$$\begin{aligned} \hat{\xi}_{j+1}^{(2)} - \hat{\xi}_j^{(2)} &= (\hat{\xi}_{j+1}^{(2)} - \xi_{i_{j+1}}) + (\xi_{i_{j+1}} - \xi_{i_j}) + (\xi_{i_j} - \hat{\xi}_j^{(2)}) \\ &\leq \frac{\Delta}{8} - \Delta + \frac{\Delta}{8} = -\frac{3\Delta}{4}, \end{aligned}$$

contradicting $\hat{\xi}_{j+1}^{(2)} > \hat{\xi}_j^{(2)}$. Hence, we must have $i_j = j$ for all j, and

$$|\hat{\xi}_j^{(2)} - \xi_j| \leq \Delta/8, \qquad \text{for } j \in [K]. \tag{B.16}$$

Consequently, it follows that for $j \in [K]$,

$$\begin{aligned} \hat{\xi}_{j+1}^{(2)} - \hat{\xi}_j^{(2)} &= (\hat{\xi}_{j+1}^{(2)} - \xi_{j+1}) + (\xi_{j+1} - \xi_j) + (\xi_j - \hat{\xi}_j^{(2)}) \\ &\geq -\frac{\Delta}{8} + \Delta - \frac{\Delta}{8} = \frac{3\Delta}{4}. \end{aligned}$$

Let $\tilde{X}_j^{(2)}$ denote the midpoint between $\hat{\xi}_j^{(2)}$ and $\hat{\xi}_{j+1}^{(2)}$ (and set $\tilde{X}_0^{(2)} = 1$, $\tilde{X}_K^{(2)} = n$ as in (7.2)). Then

$$\tilde{X}_j^{(2)} - \hat{\xi}_j^{(2)} = \frac{1}{2}(\hat{\xi}_{j+1}^{(2)} - \hat{\xi}_j^{(2)}) \geq \frac{3\Delta}{8}, \qquad \hat{\xi}_j^{(2)} - \tilde{X}_{j-1}^{(2)} = \frac{1}{2}(\hat{\xi}_j^{(2)} - \hat{\xi}_{j-1}^{(2)}) \geq \frac{3\Delta}{8}.$$

Together with (B.16), this implies for each $j \in [K]$ that

$$\tilde{X}_{j-1}^{(2)} < \xi_j < \tilde{X}_j^{(2)}.$$

Hence $\mathcal{A} \subseteq \mathcal{G}$, and thus $\mathbb{P}(\mathcal{G}) \to 1$. Therefore,

$$\mathbb{P}(\xi_k \in \mathcal{C}_{1-\alpha}^{\text{CONCH-SEG-}\texttt{crossfit}}) \geq (1-\alpha) - \mathbb{P}(\mathcal{G}^c) = 1 - \alpha - \mathrm{o}(1),$$

as claimed. □

In particular, kernel-based changepoint detection (KCPD) yields estimators that satisfy the consistency conditions (a) and (b), stated in the above theorem, under mild regularity assumptions [see, e.g., Garreau and Arlot, 2018, Diaz-Rodriguez and Jia, 2025]. This supports the large-sample validity of the CONCH-SEG-`crossfit` procedure when wrapped around a KCPD algorithm. However, in our experiments, whether or not we employ cross-fitting has little effect on power; see Appendix C.2 for empirical evidence.

C Additional Experiments

C.1 Gaussian mean-shift: comparison with Dandapanthula and Ramdas [2025]

In the Gaussian mean-shift setting described in Section 8.1.1, we compare our framework against the changepoint localization method of Dandapanthula and Ramdas [2025], which also constructs distribution-free confidence sets for changepoints using a matrix of conformal p-values. Figure 6 displays the p-value distributions from both methods. Their approach yields a confidence set over $[362, 432]$, which is broader than the widest interval obtained by CONCH (using the weighted-mean score).

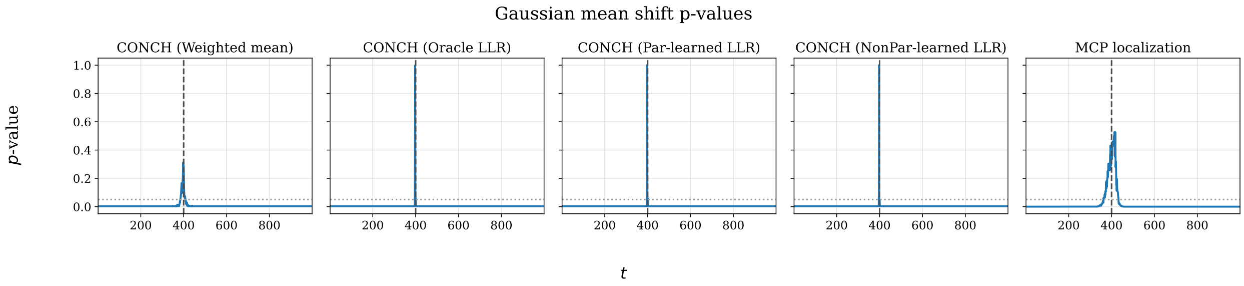

Figure 6: p-value distributions from Dandapanthula and Ramdas [2025] and CONCH under the Gaussian mean-shift model.

C.2 Gaussian mean-shift: localization of multiple changepoints

We consider a multiple-changepoint Gaussian mean-shift model to illustrate the performance of CONCH-SEG (Algorithm 3). In particular, we generate $n = 1500$ observations with true changepoints at $\xi_1 = 150$, $\xi_2 = 500$, $\xi_3 = 820$, and $\xi_4 = 1100$, segment means $\mu_1 = -1$, $\mu_2 = 0.5$, $\mu_3 = 1.5$, $\mu_4 = -2$, and $\mu_5 = -1$, and common variance $\sigma^2 = 1$, i.e.,

$$X_t \sim \mathcal{N}(\mu_j, \sigma^2) \quad \text{for } t \in (\xi_{j-1}, \xi_j], \qquad \text{and } \ \xi_0 = 0, \ \xi_5 = n.$$

Initial changepoint estimates are obtained via KCPD [Garreau and Arlot, 2018] with a Gaussian kernel, yielding the sequence $(150, 497, 820, 1091)$. As expected, when the pre- and post-change distributions around a changepoint are more distinct, KCPD detects the changepoint more accurately, whereas the estimate is offset by a small margin when the adjacent distributions are more similar.

We then apply CONCH-SEG, wrapping the CONCH framework around KCPD and using the parametric CPP score specialized to the Gaussian family with known variance (see (4.6)). The resulting confidence set is

$$[145, 152] \,\cup\, [483, 515] \,\cup\, [820, 821] \,\cup\, [1084, 1103].$$

Figure 7 displays the observed sequence (left) and the corresponding conformal p-values (right). The procedure sharply localizes all the changepoints. The sets around ξ_2 and ξ_4 are wider than the sets around ξ_1 and ξ_3 because the distributions on either side are more similar.

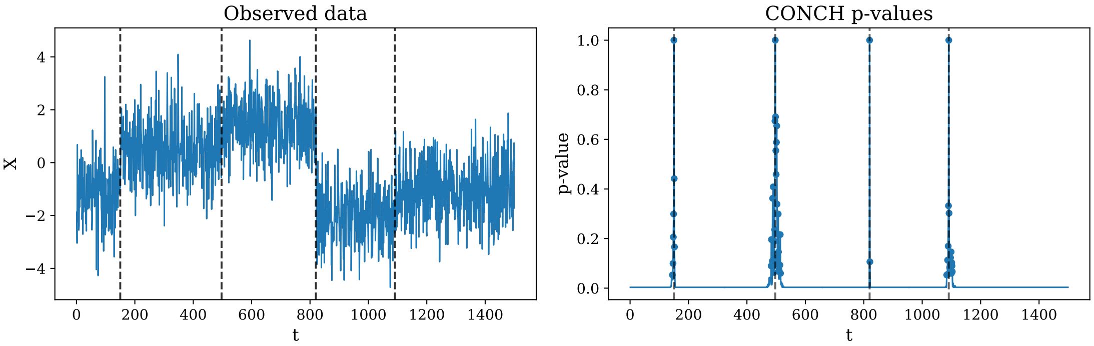

Figure 7: Left: observed sequence with multiple changepoints. Right: conformal p-values from CONCH-SEG using the parametric LLR score.

C.3 Two urns model: effect of dissimilarity between $\mathcal{P}_{0,\xi}$ and $\mathcal{P}_{1,\xi}$ on confidence set length

While we have demonstrated the performance of CONCH on a variety of changepoint detection tasks, our experiments so far have focused on i.i.d. settings, that is, changepoint models within $\mathcal{P}_{\text{IID}}$. In what follows, we go beyond the i.i.d. assumption and show that the CONCH framework requires only exchangeability to produce valid confidence sets.

To illustrate this, we evaluate the performance of the CONCH confidence sets on a two-urn model with finite populations. Specifically, we consider two urns, each containing 2500 balls colored either red or blue. The proportions of red balls in the first and second urns are $0.5 - \delta$ and $0.5 + \delta$, respectively, for some $\delta \in (0, 0.5)$. We draw balls without replacement: the first $\xi = 350$ draws come from urn 1, and the remaining from urn 2, yielding a total of $n = 800$ observations. Our goal is to detect the changepoint ξ. We use the weighted mean difference as the CPP score, and for each $\delta \in \{0.05, 0.10, \ldots, 0.50\}$, we run the CONCH-MC algorithm (Algorithm 4) with $M = 300$ permutations to obtain confidence sets.

When δ is small, the pre-change and post-change distributions are nearly indistinguishable.

Consequently, no method can sharply localize the changepoint, including CONCH confidence sets. As δ increases, the two distributions become more distinct. In the extreme case $\delta = 1$, the first urn contains only blue balls and the second only red balls, allowing perfect localization with absolute confidence. Accordingly, the average length of the CONCH confidence sets decreases with δ, as shown in the right panel of Figure 8, where the shaded region denotes one standard error around the mean. Across the whole collection of δ values, the true change-point $\xi = 350$ lies within the reported confidence set, demonstrating the validity of our procedure (left panel of Figure 8).

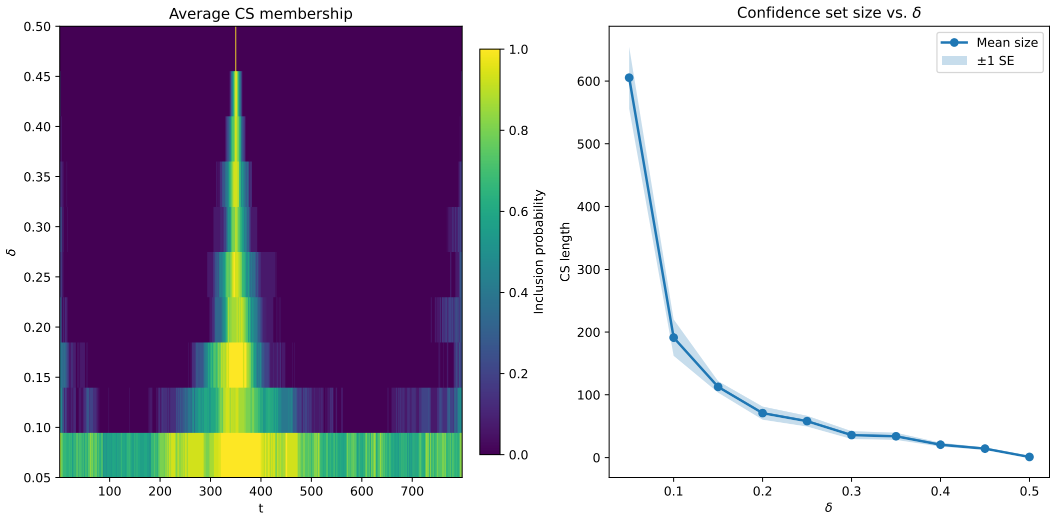

Figure 8: Two-urn changepoint experiment: CONCH confidence sets across δ values. Left: confidence sets always contain $\xi = 350$; right: average confidence set length decreases as dissimilarity δ increases.

C.4 MNIST: detect change in digits

We conduct a simulation based on the MNIST handwritten digits dataset [Deng, 2012] to evaluate the performance of CONCH for a digit shift localization. In particular, suppose we observe a sequence of 1,000 images: the first $\xi = 400$ observations consist of i.i.d. samples of the digit “1”, and the latter observations are i.i.d. samples of the digit “7” (Figure 9).

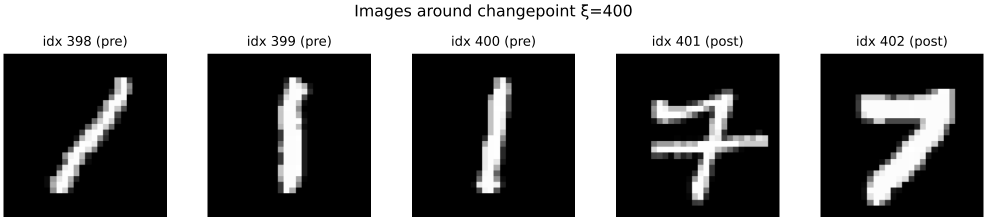

Figure 9: Illustration of MNIST changepoint setup: digit class shifts from ‘1’ to ‘7’ at $\xi = 400$ ($n = 1000$).

As in our main experiments, we use a classifier based log-likelihood ratio as CPP score in our

CONCH algorithm. Specifically, we employ a pretrained convolutional neural network classifier to distinguish between the two digits; its logits define the CPP score, which is then passed to CONCH to produce a confidence interval for the changepoint.

Although the handwritten digits "1" and "7" often exhibit substantial visual similarity, our approach accurately detects the changepoint, yielding an exceptionally narrow, in fact singleton confidence set $\{400\}$ (Figure 10). We remark that the sharp localization here is partially a consequence of the strong classifier, which can confidently distinguish between the two digits. In the next section, we investigate how classifier strength influences the width of CONCH confidence sets on the CIFAR-100 dataset.

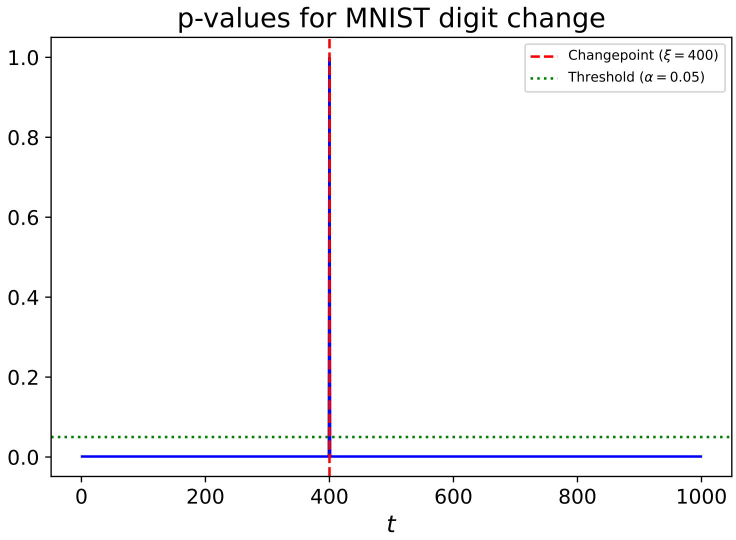

Figure 10: p-values for digit shift detection in MNIST: from digit '1' to digit '7' at $\xi = 400$

C.5 CIFAR100: classifier strength affects power of CONCH

We simulate a class-shift scenario using the CIFAR-100 image dataset [Krizhevsky et al., 2009] to evaluate CONCH under a challenging setting. Specifically, we construct a sequence of $n = 1,000$ observations with a changepoint at $\xi = 400$: the pre-change distribution $\mathbb{P}_{0,\xi}$ consists of i.i.d. images of bears, while the post-change distribution $\mathbb{P}_{1,\xi}$ consists of i.i.d. images of beavers (Figure 11). Because bears and beavers share many visual attributes, accurately localizing the changepoint is a non-trivial task.

We pre-train a small three-block convolutional network with a lightweight classification head. We first train this network for 5 epochs to obtain a weak classifier and then train it further for an additional 20 epochs to obtain a stronger classifier. The resulting logits from each model define a CPP score, which we pass to CONCH to produce a changepoint confidence interval.

Figure 12 reports the p-value distributions and confidence sets produced by CONCH. As anticipated, the stronger classifier yields sharper separation between the two classes, leading to a much

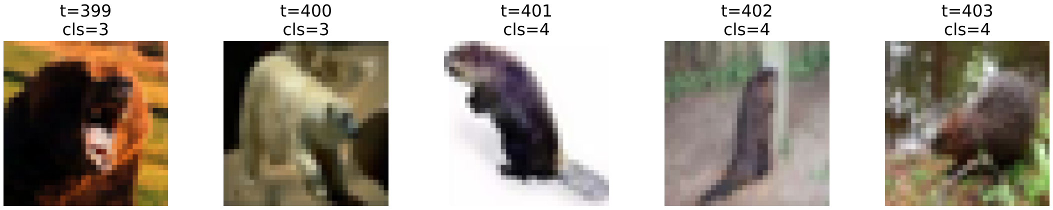

Figure 11: Illustration of CIFAR-100 changepoint setup: sequence shifts from bear images to beaver images at $\xi = 400$ ($n = 1000$).

narrower confidence set $[398, 405] \cup \{408, 415, 419\}$ compared to the weaker model's wider interval $[387, 434]$. This experiment highlights both the sensitivity of CONCH to classifier quality and its ability to localize changepoints even under subtle visual differences between classes.

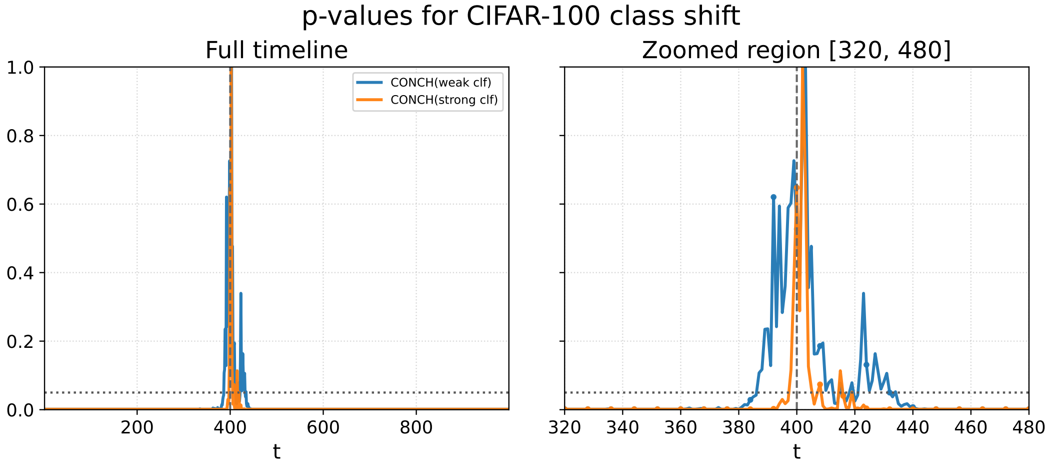

Figure 12: CONCH p-values for CIFAR-100 class shift (bear → beaver): weak vs. strong classifiers over the full timeline (left) and a zoomed window around $\xi = 400$ (right).