Conformal changepoint localization

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

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

October 7, 2025

Abstract

We study the problem of *offline changepoint localization*, where the goal is to identify the index at which the data-generating distribution changes. Existing methods often rely on restrictive parametric assumptions or asymptotic approximations, limiting their practical applicability. To address this, we propose a distribution-free framework, CONformal CHangepoint localization (CONCH), which leverages conformal p-values to efficiently construct valid confidence sets for the changepoint. Under mild assumptions of exchangeability within each segment and independence across segments, CONCH guarantees finite-sample coverage. By proving a conformal Neyman–Pearson Lemma, we derive principled score functions that yield narrow and informative confidence sets. We further establish a universality result showing that any distribution-free changepoint localization method can be viewed as an instance of CONCH. Experiments on synthetic and real data confirm that CONCH delivers precise and reliable confidence sets even in challenging settings.

Contents

*Corresponding author: rhore@andrew.cmu.edu

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*. 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 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

Offline changepoint analysis has been extensively studied due to its wide practical relevance; 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. In practice, 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, focus mainly on detection 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 fully 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), a framework that, given any $\mathbb{R}^{n-1}$-valued changepoint plausibility measure S and a confidence level $1-\alpha$, produces a finite-sample valid confidence set for the 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 in hand. We propose practically applicable 'near-optimal' score functions that yield the narrowest confidence sets, based on a novel "Conformal Neyman–Pearson" lemma, which may be of independent interest.

- We show that CONCH has a universality property: any distribution-free confidence set is an instance of our framework. Moreover, we provide a simple calibration procedure that can turn any heuristic changepoint localization methods into a truly distribution-

free, valid confidence set.

- We demonstrate the practical utility of CONCH on diverse synthetic and real-world datasets. In particular, our method can wrap around any black-box classifier trained to distinguish pre- and post-change samples, producing informative confidence sets even when the change is subtle.

Organization of the paper The rest of the paper is organized as follows. In Section 2 we formally define the problem of distribution-free changepoint localization. Section 3 introduces our general framework, CONCH, and give algorithms for its practical implementation. In Section 4, we provide guidance on selecting score functions that would yield narrow confidence sets. Section 5 establishes a universality result for CONCH, and Section 6 builds on this foundation to introduce a calibration procedure that turns any localization method into a valid distribution-free one. Section 7 then presents empirical evaluations on synthetic and real datasets, demonstrating the applicability of our framework.

2 Distribution-free changepoint localization

In this section, we formally describe the problem of distribution-free offline changepoint localization. We begin by introducing some notation. Throughout the paper, $\mathbb{N}$ denotes the set of 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 the power set of S. Finally, 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 $n \in \mathbb{N}$. 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 π_L :

$[\xi] \to [\xi]$ and $\pi_R : [n] \setminus [\xi] \to [n] \setminus [\xi]$, it holds that

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

Moreover, the pre-change and post-change segments are independent: $\mathcal{P}_{0,\xi} \perp \mathcal{P}_{1,\xi}$.

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 corresponding distributions $\mathcal{P}_{0,t}$ and $\mathcal{P}_{1,t}$ satisfy Assumption 1. We can now formally define what it means to construct a distribution-free confidence set for 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

$$\mathbb{P}_{\mathcal{H}_{0,\xi}}(\, \xi \in \mathcal{C}_{1-\alpha}(\mathbf{X}) \,) \;\geq\; 1 - \alpha. \tag{2.1}$$

Assumption 1 is considerably weaker than the working assumptions typically required by methods, reviewed in Section 1.1. Existing approaches to changepoint localization often rely on strong parametric models or asymptotic approximations, which starkly contrast the minimal nature of our assumption. Yet, to the best of our knowledge, no existing method achieves distribution-free finite-sample validity under such mild conditions, making our work a significant contribution in this direction. The next section formally introduces our approach and describes its main components.

3 Conformal changepoint localization

This section develops a conformal framework for localizing a 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. Building on these developments, we adapt conformal p-values to the changepoint localization problem in an efficient manner, yielding confidence sets for the changepoint with guarantees as in (2.1).

3.1 General framework of CONCH algorithm

Algorithm 1: CONCH: conformal changepoint localization algorithm

Input: $(X_t)_{t=1}^n$ (dataset) and $S : \mathcal{X}^n \to \mathbb{R}^{n-1}$ (CPP score function)
Output: $\mathcal{C}_{1-\alpha}^{\mathrm{CONCH}}$ (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
$\mathcal{C}_{1-\alpha}^{\mathrm{CONCH}} \leftarrow \{\, t \in [n-1] : p_t > \alpha \,\}$;
return $\mathcal{C}_{1-\alpha}^{\mathrm{CONCH}}$

Building upon the machinery of conformal p-values, we first present the general framework for distribution-free changepoint localization, namely the Conformal changepoint localization (CONCH) algorithm. 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 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 $t \in [n-1]$, 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 $\pi \in \Pi_t$ freely permutes indices to the left and right of t independently while never mixing across the split.

Note that, if t is indeed the true changepoint, elements of Π_t preserve the pre-change and post-change exchangeability. Our framework crucially depends 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 $S_t(\mathbf{X})$ 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, or equivalently, p_t is super-

uniform under the null $\mathcal{H}_{0,t}$, a result we formally establish in Theorem 3.1. The changepoint confidence set is then given by thresholding these p-values at level α:

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

which attains the distribution-free validity in (2.1).

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

3.2 CONCH-MC: randomized approximation 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. That being said, here is a quick remedy: to improve efficiency, we may sample $\pi^{(1)}, \ldots, \pi^{(M)} \overset{\text{i.i.d.}}{\sim} \text{Unif}(\Pi_t)$, and then use a Monte Carlo approximation to p_t, in particular,

$$\tilde{p}_t := \frac{1 + \sum_{k=1}^{M} \mathbb{1}\left\{S_t(\pi^{(k)}(\mathbf{X})) \leq S_t(\mathbf{X})\right\}}{1+M}. \tag{3.3}$$

This yields the *randomized* confidence set $\{\, t \in [n-1] : \tilde{p}_t > \alpha \,\}$. We refer to this procedure as CONCH-MC, presented formally in Algorithm 2. Similar to the underlying principle of 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 establish formally in Theorem 3.2.

Theorem 3.2. *For any $t \in [n]$, p_t defined in (3.3) is a valid p-value under $\mathcal{H}_{0,t}$. In particular, for any $\alpha \in (0,1)$, $\mathbb{P}_{\mathcal{H}_{0,\xi}}(\tilde{p}_\xi \leq \alpha) \leq \alpha$. Consequently, $\mathcal{C}_{1-\alpha}^{\text{CONCH-MC}}$ is a distribution-free confidence set for changepoint.*

Remark 1. 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 entire sequence $(X_1, \ldots, X_n)$. 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 the changepoint localization setting.

Algorithm 2: CONCH-MC: CONCH with random permutations

Input: $(X_t)_{t=1}^n$ (dataset), 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
  $\tilde{p}_t \leftarrow \frac{1}{M+1}\Big(1 + \sum_{k=1}^M \mathbb{1}\left\{S_t(\pi^{(k)}(\mathbf{X})) \le S_t(\mathbf{X})\right\}\Big)$;
end
$\mathcal{C}_{1-\alpha}^{\text{CONCH-MC}} \leftarrow \{t \in [n-1] : \tilde{p}_t > \alpha\}$
return $\mathcal{C}_{1-\alpha}^{\text{CONCH-MC}}$

4 Guidelines for choosing the CPP score

Both CONCH and CONCH-MC retain validity as in (2.1) for any choice of CPP score, offering substantial flexibility in constructing valid confidence sets. This, however, naturally raises the question: how should one choose a score that yields narrow and informative sets? In what follows, we establish a few general properties of CPP scores, derive an optimal score, which in turn depends on oracle knowledge. Finally, we give concrete proposals of practical score functions that closely mimic this ideal.

We begin with two general properties that explain the influence of CPP score on the resulting CONCH set.

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 the t-th component S_t of the CPP score satisfies*

$$S_t(\cdot) = S_t(\pi(\cdot)) \qquad \text{for every } \pi \in \Pi_t, \tag{4.1}$$

then the p-values p_t and $\tilde{p}_t$ defined in (3.2) and (3.3) are identically 1. Consequently,

$$\mathbb{P}\Big(t \in \mathcal{C}_{1-\alpha}^{\text{CONCH}}\Big) \;=\; \mathbb{P}\Big(t \in \mathcal{C}_{1-\alpha}^{\text{CONCH-MC}}\Big) \;=\; 1.$$

(ii) ***(Conformal data-processing inequality).*** *Let C_1 denote the CONCH (or CONCH-MC) confidence set at level $1-\alpha$ based on a CPP score S, and let $\{p_{1,1}, \ldots, p_{n-1,1}\}$*

be the corresponding conformal p-values. For any non-decreasing function $f : \mathbb{R} \to \mathbb{R}$, let C_2 be the CONCH confidence set at the same level based on the transformed score $f(S)$, with corresponding conformal p-values $\{p_{1,2}, \ldots, p_{n-1,2}\}$. Then,

$$p_{t,1} \leq p_{t,2} \quad \textit{for all } t \in [n-1],$$

and consequently $C_1 \subseteq C_2$.

Part (i) of the proposition shows that t-wise symmetry, i.e., (4.1), yields trivial conformal p-values regardless of whether $\mathcal{H}_{0,t}$ holds, and therefore leads to overly conservative sets; therefore, such scores should be avoided in practice. Part (ii) of the result establishes a monotonicity property of CONCH: applying any non-decreasing transformation can only enlarge the set. In particular, any *strictly* increasing transformation on the CPP score leaves the confidence set unchanged. These properties help us make practical choices of the CPP score that yield meaningful confidence sets in practice.

For the remainder of this section, we focus on the canonical setting, namely the i.i.d. changepoint model. Specifically, let $\mathcal{P}_{\text{IID}}$ 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 involving 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 'Conformal NP Lemma'.

For any $t \in [n-1]$, we write $\mathcal{H}'_t : \mathbf{X} \sim \otimes_{j=1}^{t} P_0 \times \otimes_{j=t+1}^{n} P_1$ to hypothesize that t is the changepoint under the model class $\mathcal{P}_{\text{IID}}$. Suppose we want to test the null $\mathcal{H}'_t$ using conformal p-values. In particular, we take a score function $s : \mathcal{X}^n \to \mathbb{R}$, and define the conformal p-value

$$p_t(s) = \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\left\{s(\pi(\mathbf{X})) \leq s(\mathbf{X})\right\}. \tag{4.2}$$

By Theorem 3.1, $p_t(s)$ is a valid p-value under $\mathcal{H}'_t$. Consequently, $\phi_s(\mathbf{X}) = \mathbb{1}\{p_t(s) \leq \alpha)\}$ is a valid test at level α for the null $\mathcal{H}'_t$ with any score function s.

Now, we seek to find the optimal score $s^\star$ such that the test $\phi_{s^\star}$ has maximum power against an alternative $\mathcal{H}'_r$ (with $r \neq t$), which hypothesizes that instead r is the changepoint. The Conformal NP lemma, stated below, formally establishes that a log likelihood-ratio (LLR) based score function $s^\star$ gives the optimal test.

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

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

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 write $C^{\mathrm{CONCH}}_{1-\alpha}(S)$ to denote the CONCH confidence set constructed with CPP score S.

Theorem 4.3. *The CPP score S^{OPT} defined by*

$$S^{\mathrm{OPT}}_t(x_1, \ldots, x_n) = \log\left(\frac{\prod_{i\leq t} f_0(x_i) \prod_{i>t} f_1(x_i)}{\prod_{i\leq \xi} f_0(x_i) \prod_{i>\xi} f_1(x_i)}\right) \tag{4.3}$$

achieves the minimum expected length of the CONCH confidence set. In particular, for any score function $S : \mathcal{X}^n \to \mathbb{R}^{n-1}$,

$$\mathbb{E}_{\mathcal{H}_{0,\xi} \cap \mathcal{P}_{\mathrm{IID}}}[C^{CONCH}_{1-\alpha}(S)] \geq \mathbb{E}_{\mathcal{H}_{0,\xi} \cap \mathcal{P}_{\mathrm{IID}}}[C^{CONCH}_{1-\alpha}(S^{\mathrm{OPT}})].$$

4.2 Practical choices for CPP score

The optimal CPP score function (4.3) 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 this section, we propose score functions that closely mimic the optimal score, thus providing 'near-optimal' performance in practice.

We now describe a few principled choices for CPP scores in this setting.

(1) Weighted mean difference. If densities f_0 and f_1 differ merely by a location shift, a natural CPP score is given by

$$S_t(x_1, \cdots, x_n) = \left| \frac{\sum_{i=1}^{t} w_{t,i} x_i}{\sum_{i=1}^{t} w_{t,i}} - \frac{\sum_{i>t}^{n} w_{t,i} x_i}{\sum_{i>t}^{n} w_{t,i}} \right|. \tag{4.4}$$

The weights $\{w_{t,i}\}$ are introduced to break the t-wise symmetry property, (4.1), 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. Common choices for weights include:

$$w_{t,i} = 1 - \frac{|i-t|}{n} \quad \text{or} \quad w_{t,i} = \exp\big(-|i-t|/n\big).$$

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 LLR. Suppose f_0 and f_1 are known. Then, the optimal CPP score function in (4.3) can be approximated by evaluating the complete likelihood at MLE $\hat{t}$ instead of the true changepoint ξ. Therefore, we may take the CPP score given by

$$S_t(x_1, \cdots, x_n) = \log\left(\frac{\prod_{i \le t} f_0(x_i) \prod_{i>t} f_1(x_i)}{\prod_{i \le \hat{t}} f_0(x_i) \prod_{i>\hat{t}} f_1(x_i)} \right), \tag{4.5}$$

where

$$\hat{t} := \operatorname*{argmax}_{s \in [n-1]} \log\Big(\prod_{i \le s} f_0(x_i) \prod_{i>s} f_1(x_i) \Big)$$

is the MLE estimate of the changepoint. If $t \in [n-1]$ is indeed the changepoint, then $\hat{t} \approx t$ and S_t will be large, indicating strong plausibility for a change. Since this score closely approximates (4.3), it is expected to sharply localize the true changepoint, as verified in our experiments too.

(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

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

with $\tilde{t} = \operatorname{argmax}_{s \in [n-1]} \log\Big(\prod_{i \le s} \hat{f}_{s,0}(x_i) \prod_{i>s} \hat{f}_{s,1}(x_i) \Big)$ 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

$$\log \frac{f_1(x)}{f_0(x)} = \log \frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0 \mid X = x)} - \log \frac{\pi_1}{\pi_0},$$

where π_1 and π_0 are class priors. If $\hat{g}$ is trained on balanced data and we write $\hat{g}(x) \in (0, 1)$ to denote the predicted probability of post-change membership, then we obtain the approximation

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

The log odds components in (4.5) 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 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 approaches, or does it truly capture the full class of distribution-free changepoint localization methods? In this section, we give a conclusive answer to this question. In fact, we establish a universality property of CONCH, which states that any procedure satisfying the coverage guarantee in (2.1) can be realized as an instance of our CONCH framework.

Theorem 5.1. *Fix $\alpha \in (0, 1)$. Let C be any procedure that maps a dataset $\mathbf{X}$ to a confidence set $C(\mathbf{X})$ such that*

$$\mathbb{P}_{\mathcal{H}_{0,\xi}}\Big(\xi \in C(\mathbf{X})\Big) \geq 1 - \alpha.$$

Then there exists a CPP score function $S : \mathcal{X}^n \to \mathbb{R}^{n-1}$ such that the distribution-free confidence set C coincides exactly the set $\mathcal{C}_{1-\alpha}^{\texttt{CONCH}}$ constructed with the score S.

This result establishes CONCHas a canonical framework for distribution-free changepoint inference: a particular choice of CPP score leads to a specific instance within the universal class of valid procedures for changepoint localization.

Moreover, it provides a practical recipe for calibrating *any* heuristic confidence set. In particular, confidence sets constructed from model-based or resampling-based methods, whether

or not they are asymptotically valid, can be "wrapped" to obtain rigorous distribution-free guarantees. We formalize this calibration procedure in the next section.

6 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, it could be one obtained from a Bayesian or bootstrap-based method. Guided by the general CONCH framework, we can construct a CPP score function from such a set and thereby obtain a distribution-free, finite-sample valid confidence set. Two natural constructions of CPP score are as follows:

- **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 in the set.

Running the CONCH algorithm with either score yields a valid confidence set by Theorem 3.1. Moreover, by Proposition 4.1 (ii), the set distance score always produces a narrower confidence set than the set membership score.

However, both score functions are relatively coarse and often lead to wide confidence sets. In particular, for indices close to n or 0, these scores frequently induce t-wise symmetry (cf. (4.1)), resulting in artificially inflated p-values in that region (by 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 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 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 3. 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

Algorithm 3: 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 $\mathcal{C}_{1-\alpha}^{\text{CONCH-CAL}}$

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 shortcomings observed with set-membership and set-distance scores, and yields a conservative confidence set.

7 Experiments

In this section, we evaluate the performance of CONCH through a series of experiments, including synthetic simulations and applications to real datasets involving images (CIFAR-100, MNIST, DomainNet), text (SST-2). Throughout this section, when we refer to CONCH confidence sets, we specifically mean those obtained using the CONCH-MC algorithm (Algorithm 2), as one would in practice. Across all these settings, the CONCH framework consistently produces informative and narrow confidence sets, sharply localizing the changepoint.

7.1 Numerical simulations

7.1.1 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 $\mathcal{P}_{1,\xi} = \bigotimes_{t=\xi+1}^{n} \mathcal{N}(1, 1)$. In this setup, changepoint localization reduces to detecting a mean

shift in the Gaussian family with scale parameter 1.

We evaluate CONCH using four choices of CPP scores, introduced earlier in Section 4.2:

(i) weighted mean difference, with a specified weight function,

(ii) oracle log-likelihood ratio (LLR),

(iii) parametrically learned LLR, assuming knowledge of the Gaussian family,

(iv) nonparametrically learned LLR, using kernel density estimates.

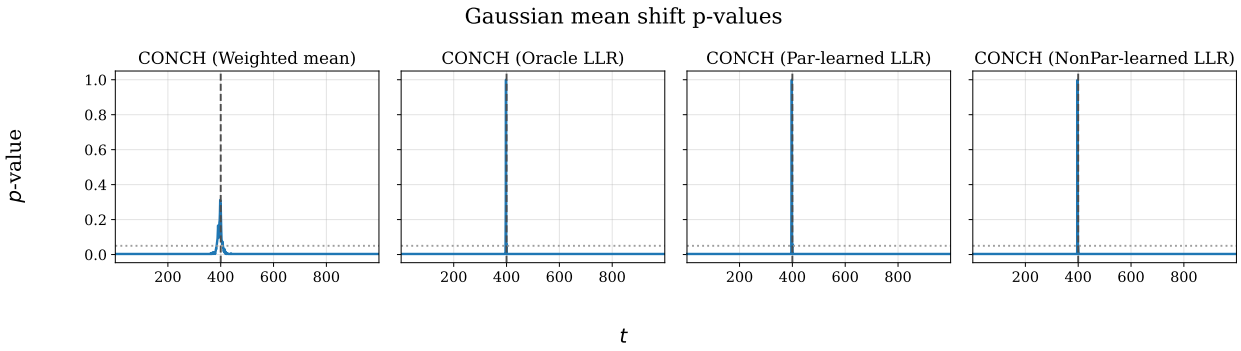

Figure 1: Distribution of conformal p-values under the Gaussian mean-shift model for different methods.

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 substantially sharper localization.

Appendix B.1 presents an additional comparison between the CONCH intervals and those from Dandapanthula and Ramdas [2025].

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

We demonstrate that the CONCH-CAL procedure (Algorithm 3) can refine confidence sets that were not originally designed with distribution-free validity guarantees. 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 mild model misspecification, these intervals can become overly wide or may miss the true changepoint ξ.

If a confidence set is valid, then the set-membership score from Section 6 should reproduce the same set. In contrast, CONCH-CAL leverages p-values, and a finer notion of CPP score building on them. This provides a principled mechanism to recalibrate and refine existing confidence sets, thereby yielding sharper, distribution-free intervals.

In both experiments, we use the same 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 percentile-based intervals and p-values are obtained.

We consider two settings: (i) the Gaussian mean-shift model with $n = 500$ and $\xi = 200$, where $\mathcal{P}_{0,\xi} = \bigotimes_{t=1}^{\xi} \mathcal{N}(-1, 3)$ and $\mathcal{P}_{1,\xi} = \bigotimes_{t=\xi+1}^{n} \mathcal{N}(1, 3)$, and (ii) a Laplace mean-shift model with $n = 500$ i.i.d. observations and the same changepoint, where $\mathcal{P}_{0,\xi} = \bigotimes_{t=1}^{\xi} \text{Laplace}(-1, 3)$ and $\mathcal{P}_{1,\xi} = \bigotimes_{t=\xi+1}^{n} \text{Laplace}(1, 3)$.

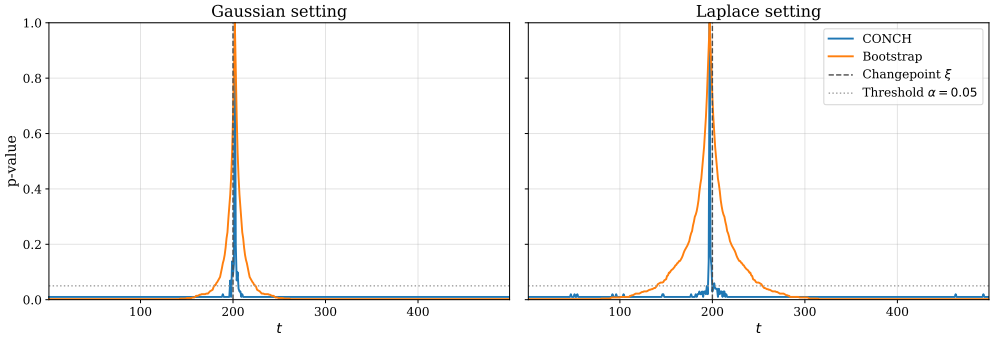

Figure 2: Refinement of bootstrap-based confidence sets using CONCH-CAL under Gaussian and Laplace mean-shift models.

In the Gaussian case, the bootstrap interval $[180, 224]$ is refined by CONCH-CAL to a tight and accurate interval $[197, 205]$. In the Laplace case, the bootstrap interval $[140, 258]$, inflated by heavy-tailed noise, is reduced to $[196, 202]$ after calibration. The bootstrap p-values in the Laplace setting are notably more spread out, while those from CONCH-CAL remain sharply concentrated near the true changepoint, highlighting its robustness and stability across distributional regimes.

7.2 Real data experiments

7.2.1 DomainNet: detecting domain shift

In this experiment, we tackle the problem of detecting a domain shift using the publicly available DomainNet dataset [Peng et al., 2019], which consists of six diverse domains (real, sketch, painting, clipart, infograph, and quickdraw). Among these, we use the *real* and *sketch* domains to construct a changepoint detection setting. Moreover, 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).

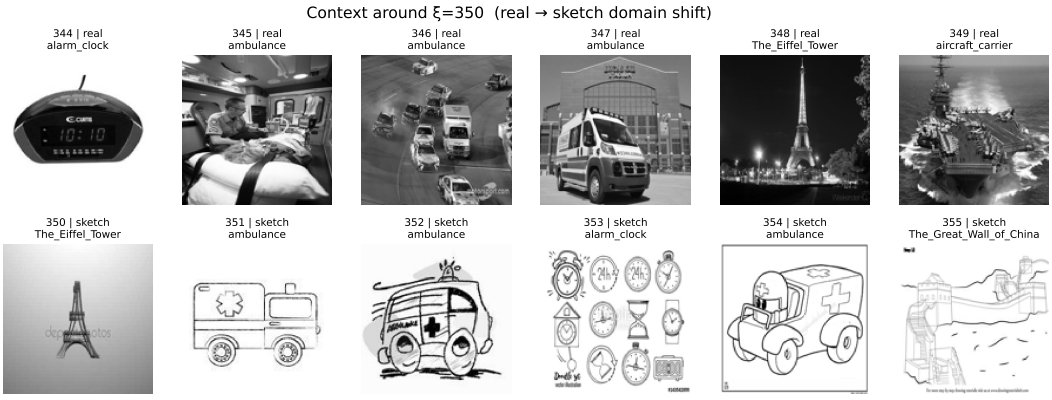

Figure 3: Illustration of the DomainNet changepoint setup: samples switch from the real to the sketch domain at $\xi = 350$ ($n = 800$). Images are drawn from the DomainNet dataset, which was collected via online search; class labels may not perfectly align with visual semantics, making the domain-shift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

7.2.2 SST-2: detecting sentiment change using large 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 labeled with binary sentiment [Socher et al., 2013], we simulate a shift from predominantly

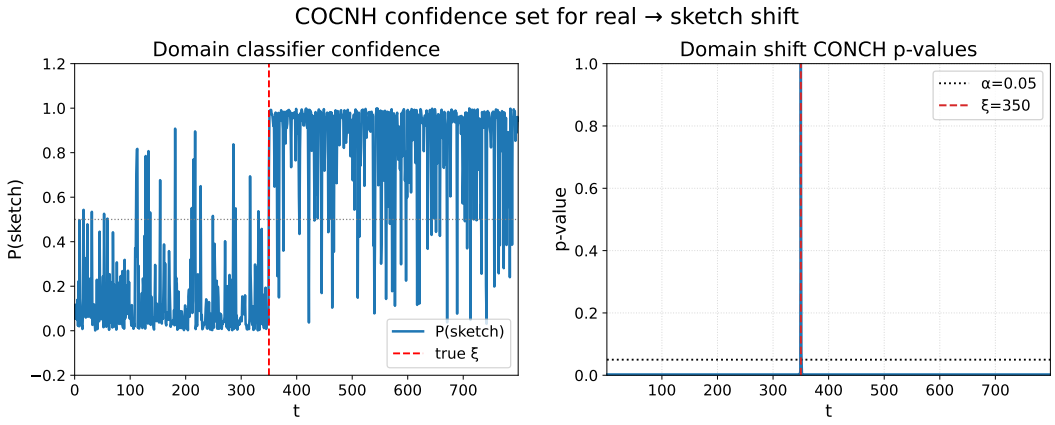

Figure 4: p-values for domain shift detection between real and sketch domains: classifier scores (left) and CONCH p-values (right)

positive to negative sentiment. Such a setup mirrors real-world scenarios, e.g., 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"
- $t = 401$ (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.

Additional experiments Appendix B presents several supplementary experiments covering a range of changepoint detection settings. We begin with a simple two-urn model to illustrate urn-shift detection, followed by detecting a shift in digit class using MNIST handwritten digits, and a class shift on the CIFAR-100 dataset to demonstrate the robustness and flexibility of our approach.

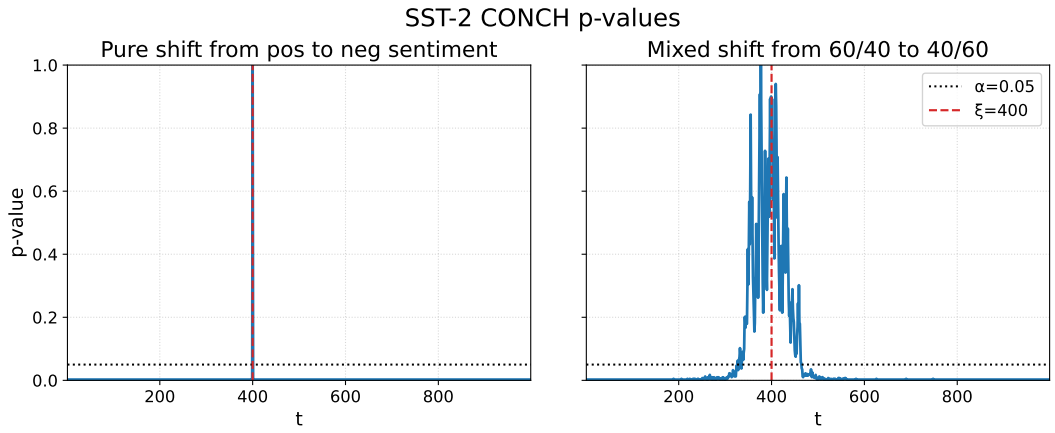

Figure 5: CONCH p-values for sentiment shift in SST-2: from positive to negative reviews at $\xi = 400$ (left), and from 60% positive to 40% positive (right).

8 Conclusion

In this work, we introduced CONCH, a novel 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 several design guidelines, including principled choices of score functions and a Monte Carlo approximation to the full-permutation p-value, which enhance both the power and practicality of the framework. We further established a universality result, positioning CONCH as a canonical method for distribution-free offline changepoint localization, and proposed a simple calibration procedure that can wrap around any localization algorithm to yield valid confidence sets.

While this work has focused on the single-changepoint setting, many real-world problems involve multiple changepoints. Although we do not address this here, we expect that techniques such as wild binary segmentation [Fryzlewicz, 2014] could be adapted to extend CONCH 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.

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.

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.

Jonathan William Duggins. Parametric resampling methods for retrospective changepoint analysis. 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. 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.

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

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. Learning multiple layers of features from tiny images. 2009.

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.

Nicolas Verzelen, Magalie Fromont, Matthieu Lerasle, and Patricia Reynaud-Bouret. Opti-

mal 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 online. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 768–775, 2003.

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. 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.

A Proofs

For notational convenience, throughout this section we write $\mathbf{x} \in \mathcal{X}^n$ to denote the tuple $(x_1, \ldots, x_n)$.

A.1 Proving coverage guarantees of CONCH confidence sets

A.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,

$$\begin{aligned}
\mathbb{P}_{\mathcal{H}_{0t}}\left(p_t(\mathbf{X}) \leq \alpha\right) &= \frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{P}_{\mathcal{H}_{0t}}\left(p_t(\pi(\mathbf{X})) \leq \alpha\right) \\
&= \mathbb{E}_{\mathcal{H}_{0t}}\left[\frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\left\{p_t(\pi(\mathbf{X})) \leq \alpha\right\}\right] \\
&= \mathbb{E}_{\mathcal{H}_{0t}}\left[\frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\left\{\frac{1}{|\Pi_t|} \sum_{\pi' \in \Pi_t} \mathbb{1}\left\{S_t(\pi'(\mathbf{x})) \leq S_t(\pi(\mathbf{x}))\right\} \leq \alpha\right\}\right] \leq \alpha,
\end{aligned}$$

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. □

A.1.2 Proof of Theorem 3.2

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

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

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

Hence, note that with $\pi_{1,t}, \ldots, \pi_{M,t} \overset{iid}{\sim} \text{Unif}(\Pi_t)$, 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 (3.3), $\tilde{p}_t \equiv \tilde{p}_t(\mathbf{X}; \pi_{1,t}, \ldots, \pi_{M,t})$, and therefore,

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

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,

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

This completes the proof. □

A.2 Proving properties of the CPP score and optimality results

A.2.1 Proof of Proposition 4.1

The first part of the result follows immediately by noting that when S_t satisfies the t-wise symmetry assumption in (4.1), then by the definitions of conformal p-values in (3.2) and (3.3), p_t and $\tilde{p}_t$ are identically equal to 1, as required.

For the second part, fix $t \in [n-1]$. We prove the result for CONCH, noting that it holds identically for CONCH-MC. 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 $C_1 \subseteq C_2$.

A.2.2 Proof of Lemma 4.2 (conformal NP lemma)

We are given the following hypothesis testing problem:

$$\mathcal{H}_0' : \mathbf{X} \sim \otimes_{j=1}^{t} P_0 \times \otimes_{j=t+1}^{n} P_1 \quad \text{v.s.} \quad \mathcal{H}_1' : \mathbf{X} \sim \otimes_{j=1}^{\xi} P_0 \times \otimes_{j=\xi+1}^{n} P_1.$$

Given samples $\mathbf{X} \in \mathbb{R}^n$, the Neyman-Pearson lemma states that the most powerful test for $\mathcal{H}_0'$ against $\mathcal{H}_1'$ is given by any test of the form

$$\frac{\mathsf{d}(\mathbb{P}_{0,\xi} \times \mathbb{P}_{1,\xi})}{\mathsf{d}(\mathbb{P}_{0,t} \times \mathbb{P}_{1,t})} = \frac{\prod_{i \le \xi} f_0(X_i) \prod_{i > \xi} f_1(X_i)}{\prod_{i \le t} f_0(X_i) \prod_{i > t} f_1(X_i)} \ge \tau_\alpha$$

for an appropriate threshold $\tau_\alpha \in \mathbb{R}$ that controls Type I error under $\mathcal{H}_0'$ at level α. Now, we will show that the test $\mathbb{1}\{p_t(s^\star) \le \alpha\}$, where p_t is as in (4.2) with s replaced by $s^\star$, admits the same form. This will complete the proof for the optimality of $s^\star$.

We define $\mathbf{X}_\pi = \pi(\mathbf{X})$ where $\pi \sim \mathrm{Unif}(\Pi_t)$, and let $F_{s^\star(\mathbf{X}_\pi)}$ denote the cumulative distribution function of score $s^\star(\mathbf{X}_\pi)$, conditional on $\mathbf{X}$. By definition of p_t in (4.2), we observe that

$$\mathbb{1}\{p_t \le \alpha\} = \mathbb{1}\left\{F_{s^\star(X_\pi)}(s^\star(\mathbf{X})) \le \alpha\right\} \le \mathbb{1}\left\{\frac{\prod_{i \le \xi} f_0(X_i) \prod_{i > \xi} f_1(X_i)}{\prod_{i \le t} f_0(X_i) \prod_{i > t} f_1(X_i)} \ge \tilde{\tau}_\alpha\right\}$$

for an appropriate threshold $\tilde{\tau}_\alpha \in \mathbb{R}$ that controls Type I error under $\mathcal{H}_0'$ at level α. This establishes the equivalence of these two tests, as required. □

A.2.3 Proof of Theorem 4.3

Since only the t-th coordinate of CPP score S_t determines the CONCH p-value p_t defined in (3.2), with the notation laid out in Section 4.1, we can write

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

Finally, noting that for any $j \in [n-1]$, $\mathcal{H}_{0,j} \cap \mathcal{P}_{\mathrm{IID}} = \mathcal{H}_j'$ and recalling that the p-value $p_t(S_t)$ must be valid under $\mathcal{H}_t'$, applying Lemma 4.2, the optimal form of S_t^{OPT} follows readily.

A.3 Proof of Theorem 5.1 (universality theorem)

First, based on the given confidence set C, we define a CPP score by

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

for any $t \in [n-1]$. We will show that the CONCH confidence set based on the score S, denoted $\mathcal{C}_{1-\alpha}^{\mathrm{CONCH}}$, matches exactly with the confidence set C.

We first show that $\mathcal{C}_{1-\alpha}^{\mathrm{CONCH}}(\mathbf{X}) \supseteq C(\mathbf{X})$; 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,

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

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

$$\frac{1}{|\Pi_t|} \sum_{\pi \in \Pi_t} \mathbb{1}\left\{t \in C(\pi(\mathbf{x}))\right\} \geq 1 - \alpha. \tag{A.1}$$

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

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

Therefore, it follows that

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

or equivalently, (A.1) holds.

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

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

where the last step follows from (A.1). This completes the proof. □

B Additional Experiments

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

In the Gaussian mean-shift setting described in Section 7.1.1, 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 even broader than the widest interval obtained by CONCH (using the weighted-mean score). This reflects the conservative nature of their method, a property not exhibited by CONCH.

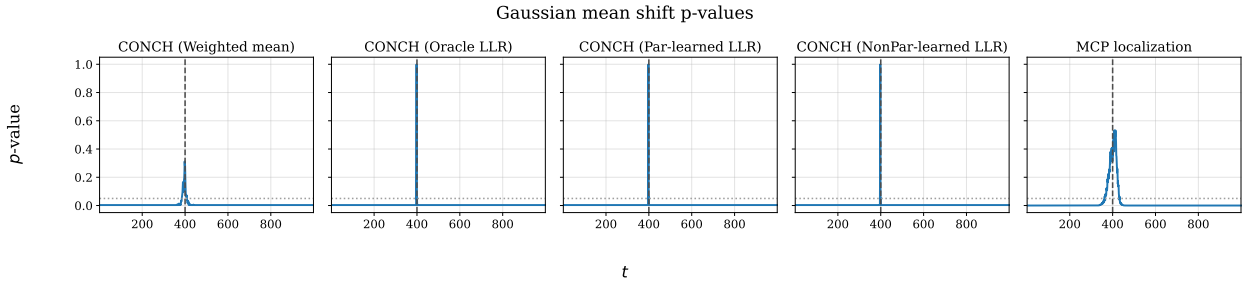

Figure 6: p-value distributions from Dandapanthula and Ramdas [2025] and CONCH under the Gaussian mean-shift model.

B.2 Two urns model: effect of dissimilarity between $\mathcal{P}_{0,\xi}$ and $\mathcal{P}_{1,\xi}$ on confidence set length

Here, we evaluate the performance of the CONCH confidence sets on a two-urn model with finite populations. In particular, we have two urns, each with 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: for the first $\xi = 400$ draws from urn 1, and thereafter from urn 2, yielding a total of $n = 1000$ observations. Our goal is to detect the changepoint ξ. Here, we choose the weighted mean difference as the CPP score, and for each value of $\delta \in \{0.05, 0.10, \ldots, 0.45, 0.50\}$, we compute the CONCH-MC algorithm (Algorithm 2) with $M = 500$ permutations to obtain confidence sets.

When δ is small, the pre-change and post-change distributions are nearly indistinguishable. Therefore, any procedure would fail to sharply localize the change, and so do CONCH confidence sets. With a larger δ, the pre-change and post-change distributions become more distinct. As a result, we observe that the average length of the CONCH confidence sets decreases with δ, as shown in the right panel of Figure 7, where the shaded region indicates one standard error around the mean. Across the whole collection of δ values, the true change-point $\xi = 400$ lies within the reported confidence set, demonstrating the validity of our procedure (left panel of Figure 7).

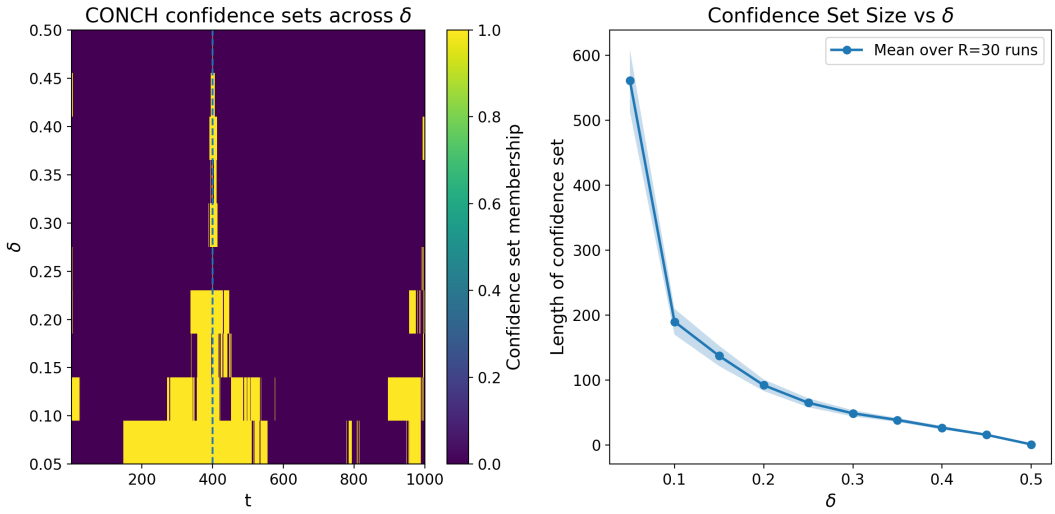

Figure 7: Two-urn changepoint experiment: CONCH confidence sets across δ values. Left: confidence sets always contain $\xi = 400$; right: average confidence set length decreases as dissimilarity δ increases.

B.3 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 8).

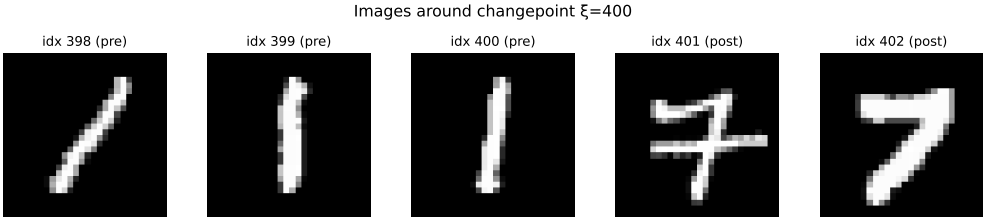

Figure 8: Illustration of MNIST changepoint setup: digit class shifts from '1' to '7' at $\xi = 400$ ($n = 1000$).

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 9). 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.

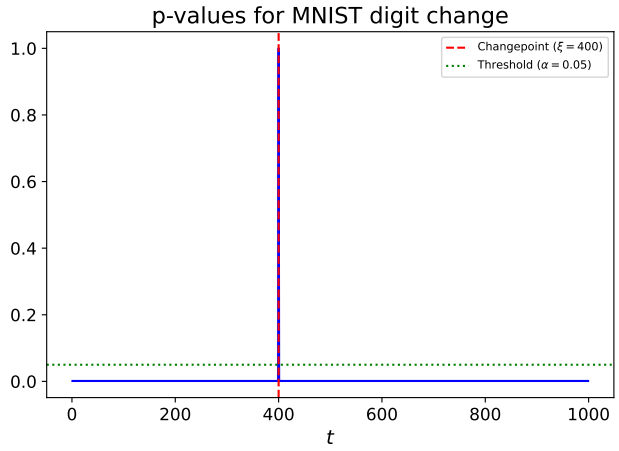

Figure 9: p-values for digit shift detection in MNIST: from digit '1' to digit '7' at $\xi = 400$

B.4 CIFAR100: classifier strength affects power of CONCH

We simulate a class-shift scenario using the CIFAR-100 image dataset [Krizhevsky, 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 10). Because bears and beavers share many visual attributes, accurately localizing the changepoint is a non-trivial task.

Figure 10: Illustration of CIFAR-100 changepoint setup: sequence shifts from bear images to beaver images at $\xi = 400$ ($n = 1000$).

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 11 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 narrower confidence set $[398, 408]$ compared to the weaker model's wider interval $[393, 427]$. This experiment highlights both the sensitivity of CONCH to classifier quality and its ability to localize changepoints even under subtle visual differences between classes.

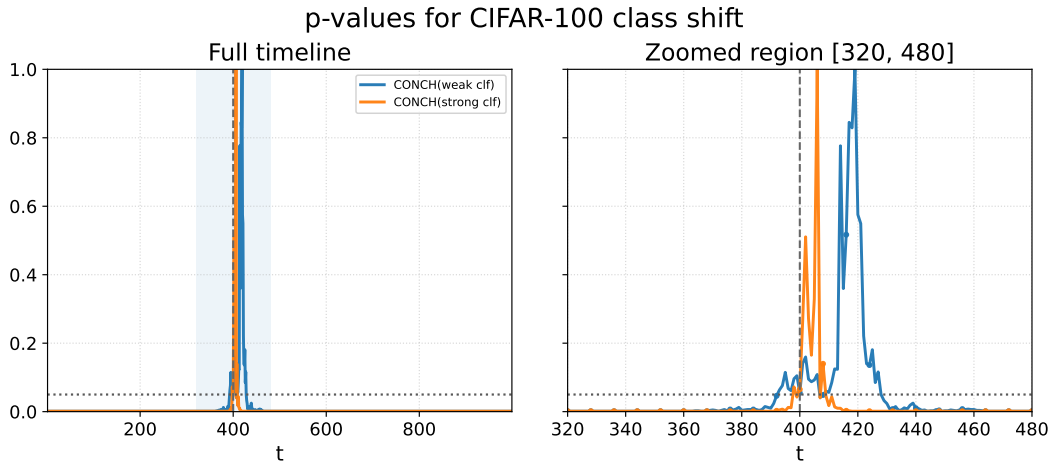

Figure 11: CONCH p-values for CIFAR-100 class shift (bear $\rightarrow$ beaver): weak vs. strong classifiers over the full timeline (left) and a zoomed window around $\xi = 400$ (right)