Sequential Changepoint Detection via Backward Confidence Sequences

Shubhanshu Shekhar and Aaditya Ramdas

Department of Statistics and Data Science

Carnegie Mellon University

I. Sequential Changepoint Detection

- Problem Definition
- Confidence Sequences (CSs)
- Assumptions

Sequential Changepoint Detection

- Stream of independent $\mathcal{X}$-valued observations: $X_1, X_2, \ldots$

- For some $T \in \mathbb{N} \cup \{\infty\}$:
  - $X_t \sim P_0$ for $t \leq T$, and
  - $X_t \sim P_1 \neq P_0$ for $t > T$.

- Mild requirements on the distributions:
  - Both P_0, P_1 are unknown, and
  - $P_0, P_1 \in \mathcal{P}$ for some known class of distributions $\mathcal{P}$.

- Decide between

$$H_0 : T = \infty, \quad \text{versus} \quad H_1 : T < \infty.$$

- **Objective:** Define a stopping time τ to declare a detection, that
  - minimizes false alarms under H_0, and
  - has a small detection delay, $(\tau - T)^+$, under H_1

Performance Measures

When $T = \infty$ (no changepoint)

For an $\alpha \in (0, 1)$, control

- Average Run Length (ARL): $\mathbb{E}_\infty[\tau] \geq \frac{1}{\alpha}$, or
- the Probability of False Alarm (PFA): $\mathbb{P}_\infty(\tau < \infty) \leq \alpha$

Performance Measures

When $T = \infty$ (no changepoint)

For an $\alpha \in (0, 1)$, control

- Average Run Length (ARL): $\mathbb{E}_\infty[\tau] \geq \frac{1}{\alpha}$, or
- the Probability of False Alarm (PFA): $\mathbb{P}_\infty(\tau < \infty) \leq \alpha$

When $T < \infty$

Ensure that

- $(\tau - T)^+$ is small, either in expectation or with high probability.
- The guarantee should hold for worst case choice of T.

Main Technical Tool: Confidence Sequences

- Suppose $X_1, X_2, \ldots \sim P_\theta$ i.i.d. with $\theta \in \Theta$.
- $\{C_t \subset \Theta : t \geq 1\}$ is a level-$(1-\alpha)$ CS for θ, if

$$\mathbb{P}\left(\forall t \geq 1 : \theta \in C_t\right) \geq 1 - \alpha \quad \equiv \quad \mathbb{P}\left(\exists t \geq 1 : \theta \notin C_t\right) \leq \alpha.$$

Main Technical Tool: Confidence Sequences

- Suppose $X_1, X_2, \ldots \sim P_\theta$ i.i.d. with $\theta \in \Theta$.
- $\{C_t \subset \Theta : t \geq 1\}$ is a level-$(1-\alpha)$ CS for θ, if

$$\mathbb{P}(\forall t \geq 1 : \theta \in C_t) \geq 1 - \alpha \quad \equiv \quad \mathbb{P}(\exists t \geq 1 : \theta \notin C_t) \leq \alpha.$$

- The i.i.d. assumption is not necessary.

Main Technical Tool: Confidence Sequences

- Suppose $X_1, X_2, \ldots \sim P_\theta$ i.i.d. with $\theta \in \Theta$.
- $\{C_t \subset \Theta : t \geq 1\}$ is a level-$(1-\alpha)$ CS for θ, if

$$\mathbb{P}\left(\forall t \geq 1 : \theta \in C_t\right) \geq 1-\alpha \quad \equiv \quad \mathbb{P}\left(\exists t \geq 1 : \theta \notin C_t\right) \leq \alpha.$$

- The i.i.d. assumption is not necessary.

- Suppose $X_1, X_2, \ldots$ are independent, with $X_t \sim P_{\theta_t}$.
- $\{C_t \subset \Theta : t \geq 1\}$ is a level-$(1-\alpha)$ CS for $\widetilde{\theta}_t := \frac{1}{t}\sum_{i=1}^{t} \theta_i$, if

$$\mathbb{P}\left(\forall t \geq 1 : \widetilde{\theta}_t \in C_t\right) \geq 1-\alpha \quad \equiv \quad \mathbb{P}\left(\exists t \geq 1 : \widetilde{\theta}_t \notin C_t\right) \leq \alpha.$$

Example: CS for Gaussian mean

- Suppose $X_1, X_2, \ldots \overset{i.i.d.}{\sim} N(\theta, 1)$. Then, we have

$$\mathbb{P}\left(\forall t \geq 1 : \ \theta \in \left[\frac{1}{t}\sum_{i=1}^{t} X_i - w_t, \ \frac{1}{t}\sum_{i=1}^{t} X_i + w_t\right]\right) \geq 1 - \alpha,$$

where

$$w_t = 1.7\sqrt{\log\log(2t) + 0.72\log(10.4/\alpha)} = \mathcal{O}\left(\sqrt{\log\log(t/\alpha)/t}\right).$$

Howard et al. (2021).

Example: CS for Gaussian mean

- Suppose $X_1, X_2, \ldots \overset{i.i.d.}{\sim} N(\theta, 1)$. Then, we have

$$\mathbb{P}\left(\forall t \geq 1 : \ \theta \in \left[\frac{1}{t}\sum_{i=1}^{t} X_i - w_t, \ \frac{1}{t}\sum_{i=1}^{t} X_i + w_t\right]\right) \geq 1 - \alpha,$$

  where

$$w_t = 1.7\sqrt{\log\log(2t) + 0.72\log(10.4/\alpha)} = \mathcal{O}\left(\sqrt{\log\log(t/\alpha)/t}\right).$$

- More generally, suppose $X_t \sim N(\theta_t, 1)$. Then,

$$\mathbb{P}\left(\forall t \geq 1 : \ \frac{1}{t}\sum_{i=1}^{t} \theta_i \in \left[\frac{1}{t}\sum_{i=1}^{t} X_i - w_t, \ \frac{1}{t}\sum_{i=1}^{t} X_i + w_t\right]\right) \geq 1 - \alpha,$$

  with the same w_t.

Howard et al. (2021).

Main Assumptions

- We work with distribution class $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$.
- Possibly infinite dimensional Θ endowed with metric d.
- $P_0 = P_{\theta_0}$ and $P_1 = P_{\theta_1}$ for θ_0, θ_1 such that $d(\theta_0, \theta_1) > 0$.

Assumptions

1. Uniformly decaying width: We can construct a CS $\{C_t(\theta) : t \geq 1\}$ for all $\theta \in \Theta$, satisfying

$$\sup_{\theta \in \Theta} \sup_{\theta', \theta'' \in C_t(\theta)} d(\theta', \theta'') \leq w_t \equiv w_t(\Theta, \alpha),$$

   such that $\lim_{t \to \infty} w_t = 0$.

2. Enough pre-change data: Under H_1, the changepoint T is large enough to ensure $w_T < \Delta := d(\theta_1, \theta_0)$.

Overview of our results

If we can construct a CS for θ $\Longrightarrow$ We can detect changes in θ

Overview of our results

If we can construct a CS for θ	$\Longrightarrow$	We can detect changes in θ

- Scheme 1: Uses a single forward CS (FCS).
  - strong false alarm control (PFA)
  - weak guarantees on detection delay
- Scheme 2: Combines one FCS with a new backward CS (BCS) every round.
  - non-asymptotic guarantees over ARL
  - tight control over the expected detection delay
- Addresses several classical and modern problems in a unified manner.

II. Scheme 1: FCS-Detector

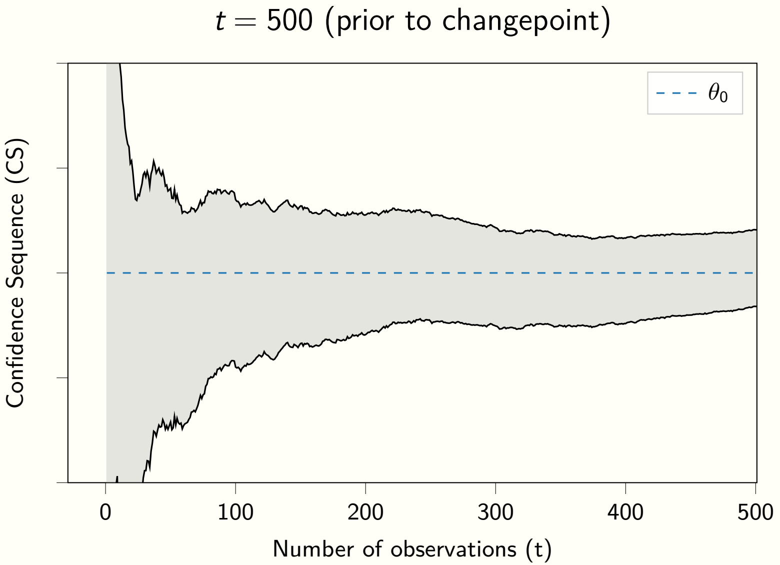
- The general strategy
- Performance Analysis
- Drawbacks

The FCS-Detector

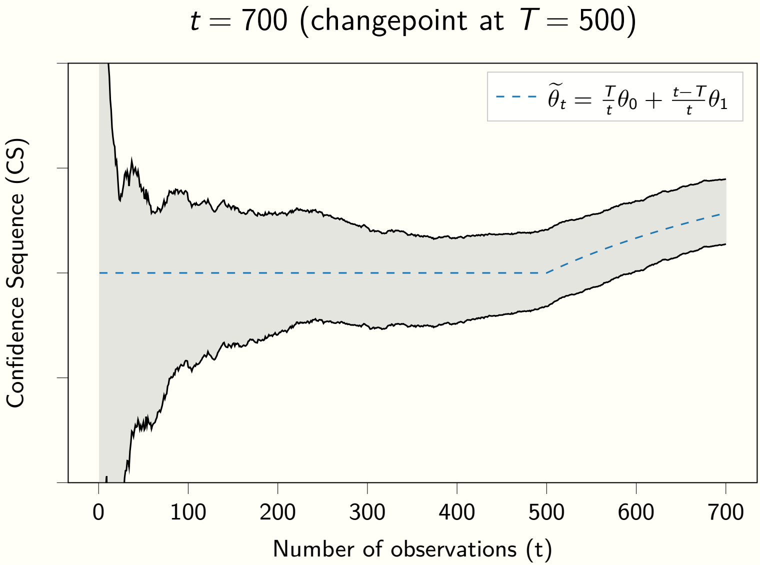
- Observations $X_1, X_2, \ldots$
- Construct one forward CS $\{C_t : t \geq 1\}$ using the observations.
- After changepoint T, the CS tracks $\widetilde{\theta}_t = \frac{T}{t}\theta_0 + \frac{t-T}{t}\theta_1$.
- For $t > T$, the term $\widetilde{\theta}_t$ drifts away from θ_0. Stop as soon as the CS becomes inconsistent.
- Formally, we define

$$\tau = \inf\{n \geq 1 : \cap_{t=1}^{n} C_t = \emptyset\}.$$

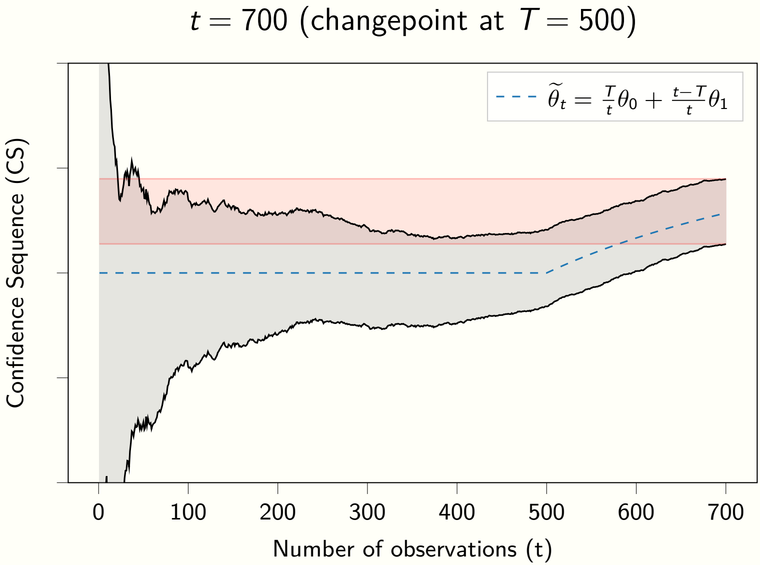
The FCS-Detector



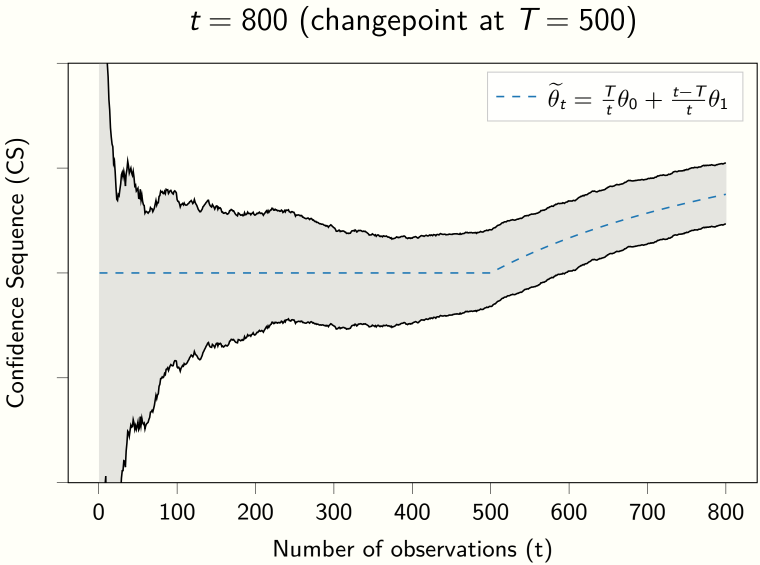
The FCS-Detector



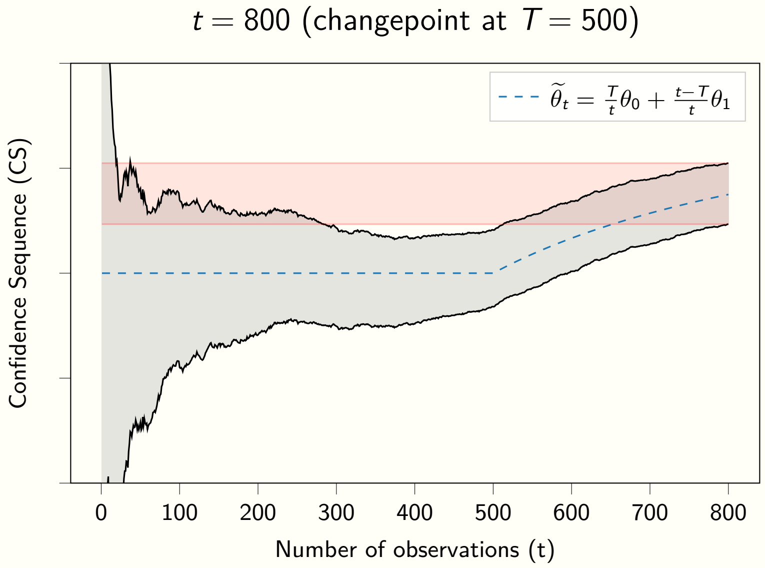
The FCS-Detector



The FCS-Detector



The FCS-Detector



Performance Guarantees

- Control over the probability of false alarm (PFA):

$$\text{Under } H_0 : \ \mathbb{P}\left(\tau < \infty\right) \leq \alpha.$$

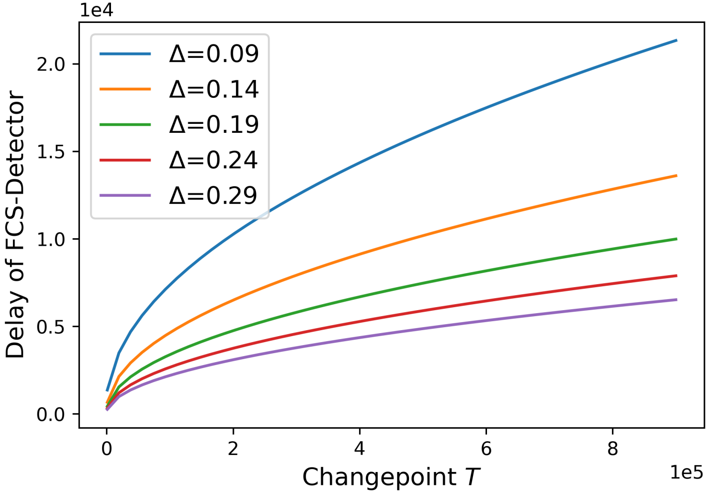
- Control over the detection delay under H_1:
  - If $w_t = \mathcal{O}\left(\sqrt{\log\log t/t}\right)$, then

$$(\tau - T)^+ = \widetilde{\mathcal{O}}\left(\sqrt{T}\right), \quad \text{w.p. } \geq 1 - \alpha.$$

  - The result also generalizes to arbitrary $w_t \to 0$ (backup slides).

Empirical Performance

$$(\tau - T)^+ = \widetilde{\mathcal{O}}\left(\sqrt{T}\right), \quad \text{w.p. } \geq 1 - \alpha.$$



Summary of FCS-Detector

- Pros.
  - Strong control of false alarms.
  - Computationally efficient — usually linear in τ.

- Cons.
  - Weak control over detection delay.
  - Can be made arbitrarily large by increasing T.

Summary of FCS-Detector

- Pros.
  - Strong control of false alarms.
  - Computationally efficient — usually linear in τ.

- Cons.
  - Weak control over detection delay.
  - Can be made arbitrarily large by increasing T.

Our next method achieves a better trade-off between control over false alarms and detection delays.

III. Scheme 2: BCS-Detector

- Backward Confidence Sequences (BCS)
- The general strategy
- Performance Analysis

Backward Confidence Sequences (BCS)

Given $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$, a 'backward CS' for θ is a collection of sets $\{B_t^{(n)} : t \in [n]\}$ satisfying:

- $B_t^{(n)}$ is $\sigma(X_t, X_{t+1}, \ldots, X_n)$ measurable, and
- $\mathbb{P}(\forall t \in [n] : \theta \in B_t^{(n)}) \geq 1 - \alpha$.

Backward Confidence Sequences (BCS)

Given $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$, a 'backward CS' for θ is a collection of sets $\{B_t^{(n)} : t \in [n]\}$ satisfying:

- $B_t^{(n)}$ is $\sigma(X_t, X_{t+1}, \ldots, X_n)$ measurable, and
- $\mathbb{P}(\forall t \in [n] : \theta \in B_t^{(n)}) \geq 1 - \alpha$.

In practice, we can construct a BCS in the following steps:

- Flip the observations:

$$Y_1 = X_n, \ \ldots Y_t = X_{n+1-t}, \ \ldots Y_n = X_1.$$

- Construct a usual (forward) CS $\{C_t : t \in [n]\}$ using $Y_1, \ldots, Y_n$.
- Flip the index of the CS:

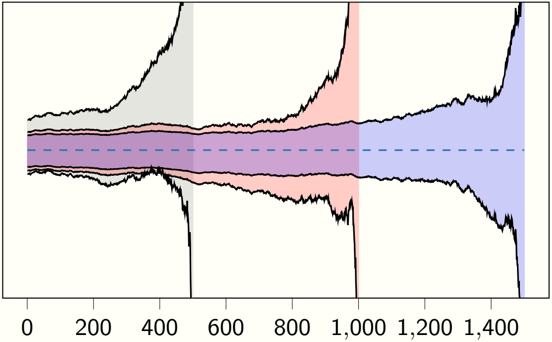
$$B_1^{(n)} = C_n, \ \ldots, B_t^{(n)} = C_{n+1-t}, \ldots \ B_n^{(n)} = C_1.$$

Backward Confidence Sequences (BCS)

Given $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$, a 'backward CS' for θ is a collection of sets $\{B_t^{(n)} : t \in [n]\}$ satisfying:

- $B_t^{(n)}$ is $\sigma(X_t, X_{t+1}, \ldots, X_n)$ measurable, and
- $\mathbb{P}(\forall t \in [n] : \theta \in B_t^{(n)}) \geq 1 - \alpha$.

Backward CSs at $n = 500, 1000, 1500$



The BCS-Detector

- Construct one forward CS.
- Construct a new backward CS every round.
- Stop the first time when the FCS and BCS disagree.

The BCS-Detector

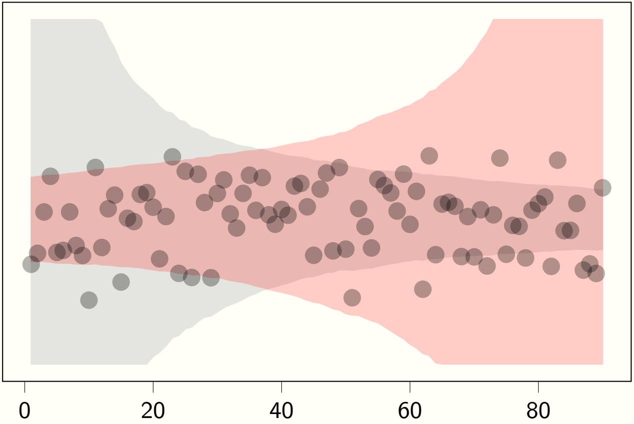
- Construct one forward CS.
- Construct a new backward CS every round.
- Stop the first time when the FCS and BCS disagree.

Formally, we define

$$\tau = \inf \left\{ n \geq 1 : \cap_{i,j=1}^{n} C_i \cap B_j^{(n)} = \emptyset \right\}$$

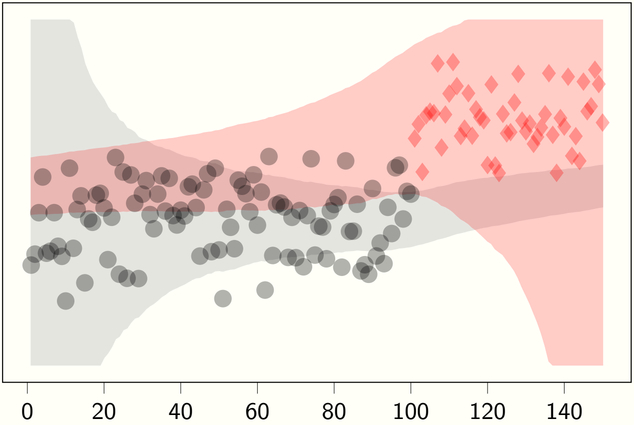
The BCS-Detector

- Construct one forward CS.
- Construct a new backward CS every round.
- Stop the first time when the FCS and BCS disagree.



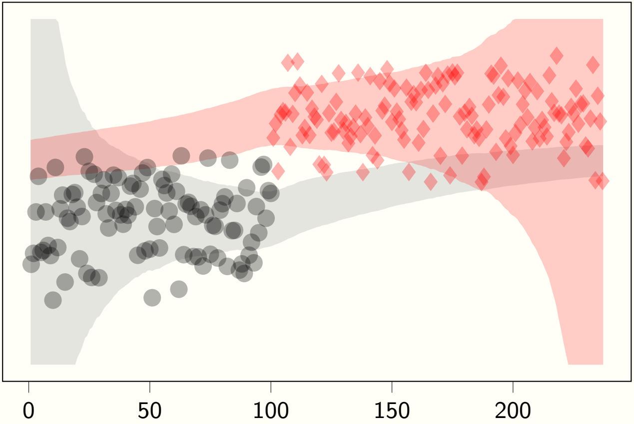
The BCS-Detector

- Construct one forward CS.
- Construct a new backward CS every round.
- Stop the first time when the FCS and BCS disagree.



The BCS-Detector

- Construct one forward CS.
- Construct a new backward CS every round.
- Stop the first time when the FCS and BCS disagree.



Performance Guarantees

- Control over the ARL:

$$\text{Under } H_0 : \; \mathbb{E}[\tau] \; \geq \; 1/2\alpha - 3/2.$$

- Control over the detection delay under H_1:
  - Introduce the "good event": $\mathcal{E} = \{\forall t \leq T : \theta_0 \in C_t\}$.
  - If $w_t = \mathcal{O}\left(\sqrt{\log\log t/t}\right)$, then

$$\mathbb{E}[(\tau - T)^+ | \mathcal{E}] = \mathcal{O}\left(\frac{\log\log(1/\Delta)}{\Delta^2}\right) \text{ where } \Delta = d(\theta_0, \theta_1).$$

  - Can generalize to arbitrary $w_t \to 0$ (next slide).

Detection Delay Analysis: general w_t

For general $w_t \to 0$, we have

$$\mathbb{E}[(\tau - T)^+ | \mathcal{E}] = \mathcal{O}\Big(\inf\{t - T : w_T + w_{t-T} < d(\theta_0, \theta_1)\}\Big).$$

Main Idea

- Recall that τ is the first time at which FCS and BCS disagree:

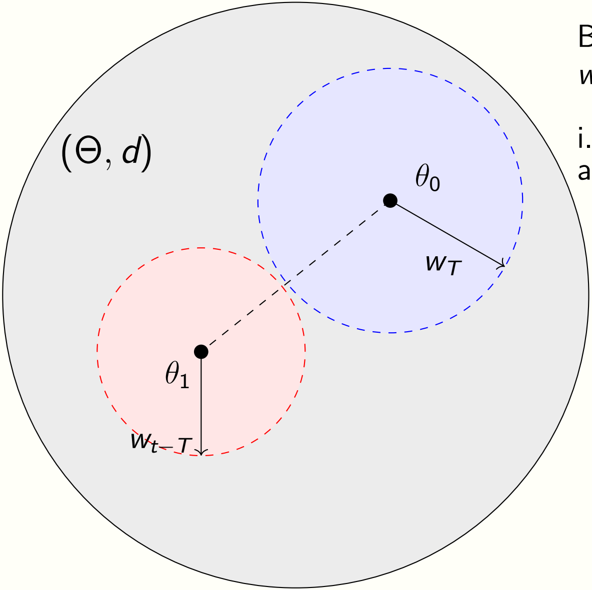
$$\min \ t \geq 1 : \ \bigcap_{i,j=1}^{t} C_i \cap B_j^{(t)} = \emptyset.$$

- Define t^* as the first $t > T$, such that FCS and BCS disagree at T:

$$\min \ t > T : \ \ C_T \cap B_T^{(t)} = \emptyset.$$

- By definition, $(\tau - T)^+ \leq t^* - T$.
  - t^* is easier to bound.

Detection Delay Analysis: general w_t



Bound t^* by the first t s.t. $w_{t-T} + w_T < d(\theta_1, \theta_0)$

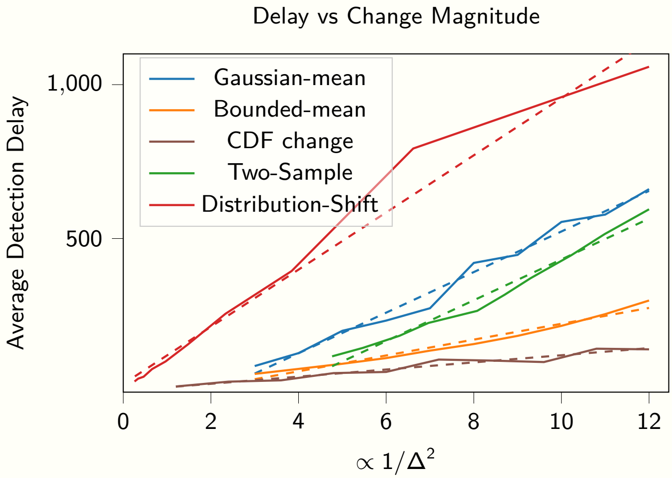
i.e., when the balls centered at θ_0 and θ_1 become disjoint

Applications

- Mean-shift detection with univariate Gaussians
  - $P_{\theta_0} = N(\theta_0, 1)$, and $P_{\theta_1} = N(\theta_1, 1)$, with $\Delta = |\theta_1 - \theta_0|$.
- Mean-shift detection with bounded observations
  - P_{θ_0} and P_{θ_1}, supported on $[0, 1]$ with $\Delta = |\theta_1 - \theta_0|$.
- Changes in CDF
  - $\Delta = d_{KS}(\theta_0, \theta_1)$, with $\theta_i =$ CDF of P_{θ_i}.
- Two-sample changepoint detection
  - $\theta_0 = P \times P$, $\theta_1 = P \times Q$, and $\Delta = \text{MMD}(P, Q)$.
- Several other problems: distribution shifts in ML, nonparametric regression, exponential family.

Applications

$$\mathbb{E}[(\tau - T)^+ | \mathcal{E}] = \mathcal{O}\left(\frac{\log\log(1/\Delta)}{\Delta^2}\right) \text{ where } \Delta = d(\theta_0, \theta_1).$$



Conclusion and Future Work

- We developed two simple SCD schemes based on CSs
- Scheme1: FCS-Detector
  - controls PFA under H_0, poor detection delay under H_1, low computational cost
- Scheme 2: BCS-Detector
  - controls ARL under H_0, tight detection delay under H_1, high computational cost
- Addresses several problems in a unified manner

Future Directions

- Estimating the changepoint T
- Estimating the change magnitude $\Delta = d(\theta_0, \theta_1)$
- Reducing the computational cost of BCS-Detector

Thank You.

Reference: S. Shekhar and A. Ramdas, *"Sequential changepoint detection via backward confidence sequences"*. ICML 2023.

Backup Slides

- T and Δ estimation
- Details of Assumptions
- Detection Delay of FCS-Detector

Changepoint and Change Magnitude Estimation

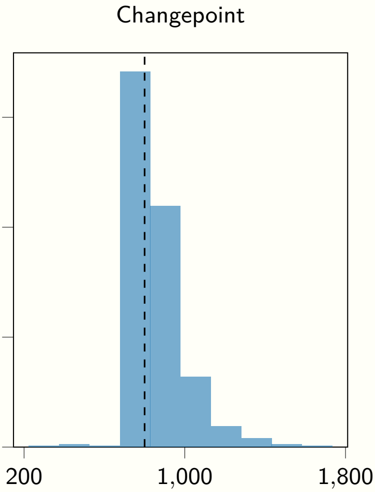
- We can estimate the changepoint as the time at which forward and backward CS disagree the most.

$$\widehat{T} = \max \operatorname*{arg\,max}_{1 \le t \le \tau} d(C_t, B_t^{(\tau)}).$$

- The maximum distance between points in $C_{\widehat{T}}$ and $B_{\widehat{T}}^{(\tau)}$ gives an upper bound on the change magnitude Δ.

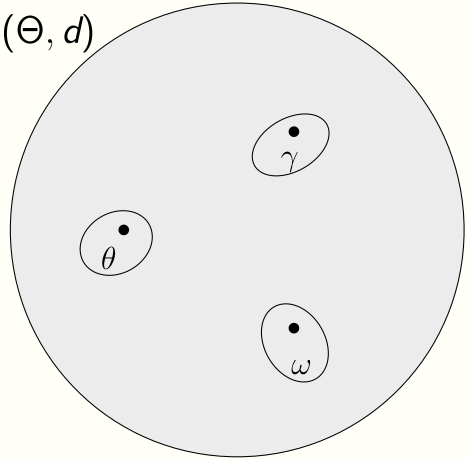
$$\widehat{\Delta} = \max_{\theta \in C_{\widehat{T}}, \theta' \in B_{\widehat{T}}^{(\tau)}} d(\theta, \theta')$$

Changepoint and Change Magnitude Estimation



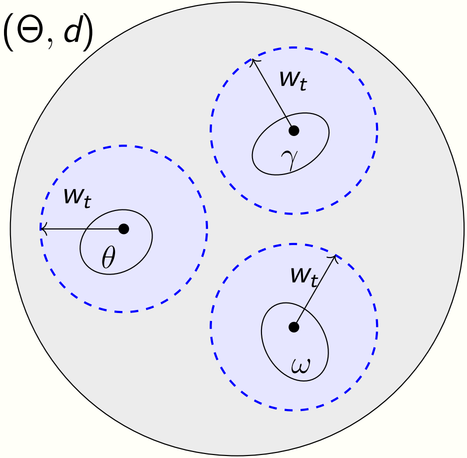
Assumptions

Assumption 1: CS for all $\theta \in \Theta$ after t observations are contained in balls of radius $w_t \equiv w_t(\Theta, \alpha)$.



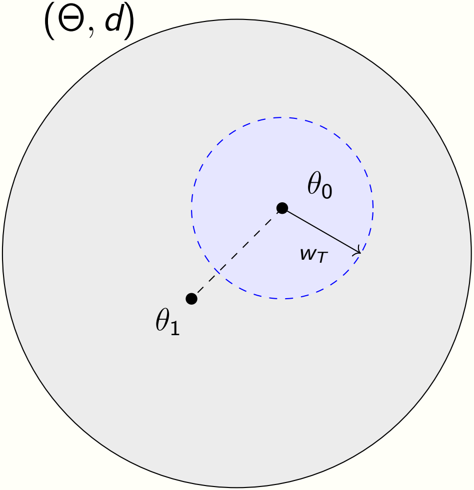
Assumptions

Assumption 1: CS for all $\theta \in \Theta$ after t observations are contained in balls of radius $w_t \equiv w_t(\Theta, \alpha)$.



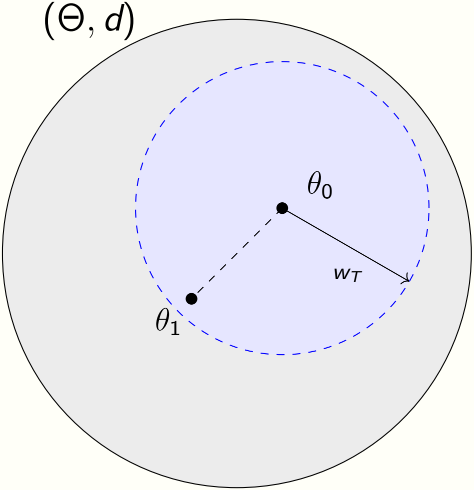
Assumptions

Assumption 2: There are enough pre-change data to ensure $w_T < d(\theta_0, \theta_1)$.



Assumptions

Assumption 2: There are enough pre-change data to ensure $w_T < d(\theta_0, \theta_1)$.



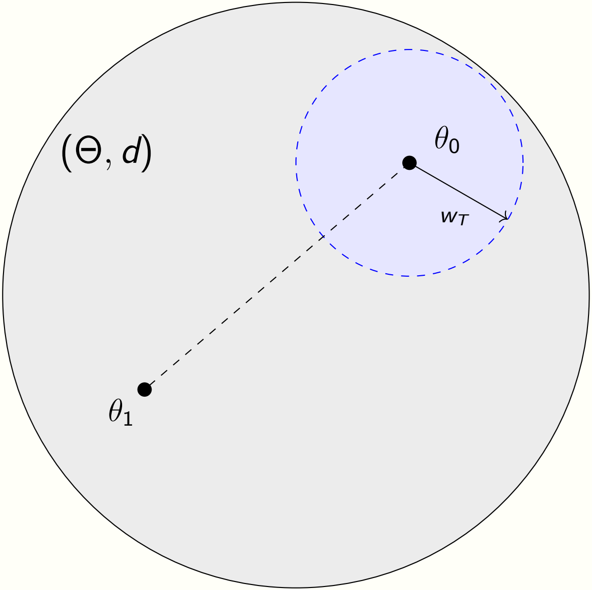
Insufficient pre-change data!

Detection Delay Analysis of FCS-Detector

Main Idea

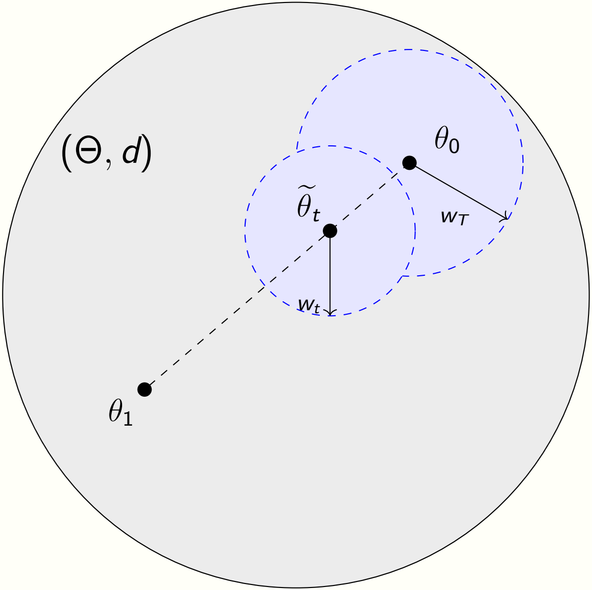
- Recall that τ is the first t, such that $\cap_{i=1}^{n} C_i = \emptyset$.
- Define t^* as the first $t \geq T$, such that $C_t \cap C_T = \emptyset$
- Note that $\tau \leq t^*$
- Bounding t^* is easier

Detection Delay Analysis of FCS-Detector



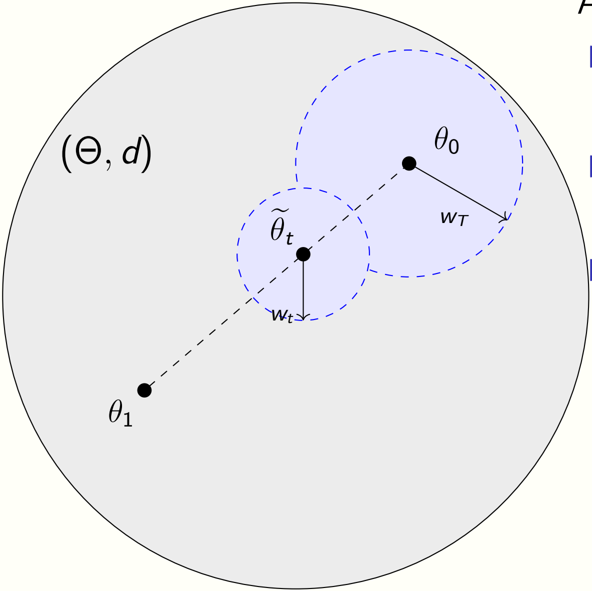
Width of CS at changepoint T

Detection Delay Analysis of FCS-Detector



For t just after T

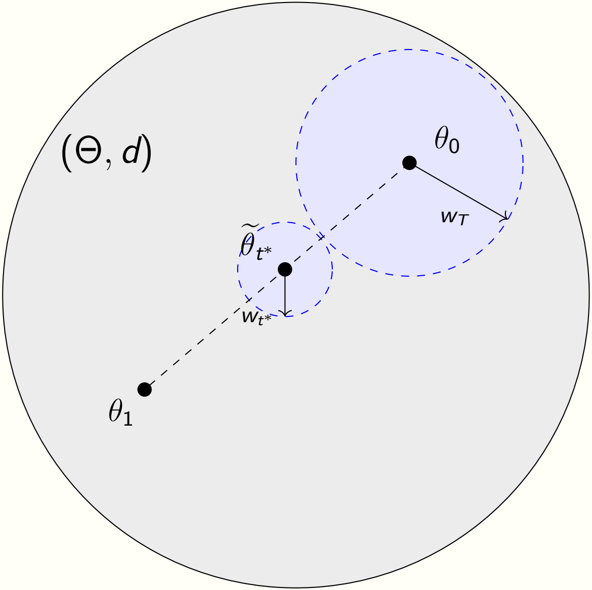
Detection Delay Analysis of FCS-Detector



As t increases:

- FCS tracks $\widetilde{\theta}_t = \frac{T}{t}\theta_0 + \frac{t-T}{t}\theta_1$
- $\widetilde{\theta}_t$ drifts away from θ_0
- w_t decreases

Detection Delay Analysis of FCS-Detector



Stops before t^*:
$w_{t^*} + w_T < d(\widetilde{\theta}_{t^*}, \theta_0)$