Fixed-Confidence Multiple Change Point Identification under Bandit Feedback

Joseph Lazzaro[1] Ciara Pike-Burke[1]

Abstract

Piecewise constant functions describe a variety of real-world phenomena in domains ranging from chemistry to manufacturing. In practice, it is often required to confidently identify the locations of the abrupt changes in these functions as quickly as possible. For this, we introduce a fixed-confidence piecewise constant bandit problem. Here, we sequentially query points in the domain and receive noisy evaluations of the function under bandit feedback. We provide instance-dependent lower bounds for the complexity of change point identification in this problem. These lower bounds illustrate that an optimal method should focus its sampling efforts adjacent to each of the change points, and the number of samples around each change point should be inversely proportional to the magnitude of the change. Building on this, we devise a simple and computationally efficient variant of Track-and-Stop and prove that it is asymptotically optimal in many regimes. We support our theoretical findings with experimental results in synthetic environments demonstrating the efficiency of our method.

1. Introduction

There are a variety of settings where it is necessary to confidently identify the location of change points in a piecewise constant function via sequential queries. For example, in communication and computer system engineering, the system can become overloaded and we need to determine the point at which this happens (Lan et al., 2009); during material development it is essential to find the experimental conditions under which the behavior of a new material changes abruptly between phases (Park et al., 2021; 2023); or in Oceanology we may wish to map out the edge of a cliff on the seafloor (Hayashi et al., 2019). In all these settings, data is expensive to collect due to the cost, time, or compute of each experiment. Moreover, there are often health or safety constraints which mean that we need to have high confidence in our results. This motivates us to develop sequential methods which can learn about the locations of these change points efficiently and with high confidence.

In this paper we consider a multi-armed bandit setting in which the expected rewards are piecewise constant across an action space $\mathcal{A}$, see Figure 1 for an example. In each round $t = 1, 2, \dots$ we select an action in our action space and observe the mean reward at that action plus some random noise (see Section 3 for more details). We assume the mean reward function is stationary across time, but piecewise constant across $\mathcal{A}$. The objective of the learner is to confidently identify the locations of the N change points/jumps in the mean reward function with the fewest number of samples possible. While there exist methods for locating the minimum of a smooth or convex function under bandit feedback (e.g. Kleinberg et al., 2008; Srinivas et al., 2010; Bubeck et al., 2011), the abrupt changes in the mean reward function across the action space make these methods inapplicable to our setting. There has been work on identification of a single change point across the action space with bandit feedback, under a fixed-budget constraint (Hall & Molchanov, 2003; Lan et al., 2009; Lazzaro & Pike-Burke, 2025). However, in practice it is often necessary to continue sampling until we are sufficiently confident we have identified $N \geq 1$ change points correctly, at which point we stop playing. For example, in material development we need to identify the phase transitions of an unfamiliar new material with high confidence for the sake of both production and safety precautions (Park et al., 2021; 2023). It is therefore important to study this fixed-confidence variant of the problem, which we will refer to as the fixed-confidence piecewise constant bandits problem.

In this paper we provide a comprehensive study of the fixed-confidence piecewise constant bandits problem. We firstly study the complexity of searching for N change points, when the true number of change points is known to be exactly N (see Definition 5.1). However, in practice, there may be an unknown number of additional changes present and it

[1] Department of Mathematics, Imperial College London, London, England. Correspondence to: Joseph Lazzaro <joseph.lazzaro18@imperial.ac.uk>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

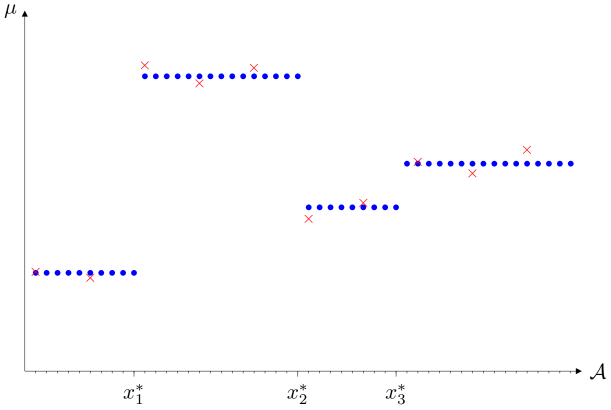

Figure 1. Example of an environment with piecewise constant expected rewards μ across the action space $\mathcal{A}$. Here the expected rewards for different actions in the action space are represented by blue dots. There are three change points in this environment at x_1^*, x_2^*, and x_3^*. We additionally plot noisy rewards obtained from 10 arbitrary actions that have been played in $\mathcal{A}$ and represent these with red crosses.

is important to develop methods robust to this. For example, when looking for a cliff on the sea floor, there may be other smaller changes in the sea-depth elsewhere. Hence, we also study the complexity of finding any N change points out of an unknown number of changes $m \geq N$ (see Definition 5.3). In both cases we prove a lower bound on the expected sample complexity (Theorems 5.2, 5.5), which provides some key insights. Firstly, it tells us that an optimal method should focus most of its sampling efforts on actions immediately adjacent to each of the change points. Secondly, the number of samples adjacent to each change should be inversely proportional to its magnitude. We use these insights to construct the Multiple Change Point Identification (MCPI) policy (Algorithm 2), a variant of Track-and-Stop (Garivier & Kaufmann, 2016), which we show to be asymptotically optimal (Theorem 5.7, for both objectives). Unlike some other variants of Track-and-Stop (Garivier & Kaufmann, 2016; Juneja & Krishnasamy, 2018), MCPI is simple and computationally inexpensive. In Section 6 we conduct experiments in synthetic environments to illustrate MCPI is optimal and that it outperforms existing related works in Clustering Bandits (Yang et al., 2022).

2. Related Work

Fixed Confidence Pure Exploration The most well-studied fixed-confidence pure exploration problem is best-arm identification where the learner aims to identify the action with the largest mean reward. While some works have focused on minimax optimal sample complexity (see Jamieson & Nowak, 2014), these methods are often instance-dependent asymptotically suboptimal. More recent works have focused asymptotic optimality (e.g. Track-and-Stop and Top-Two methods by Garivier & Kaufmann, 2016; Jourdan et al., 2022, resp.). Our method is a variant of the Track-and-Stop approach (Garivier & Kaufmann, 2016). Normally, when using Track-and-stop methods, the actions played are chosen after solving an optimisation problem in each round. As these optimisation problems often have to be solved numerically, the choice of action in each round can be somewhat unintuitive and potentially computationally expensive. However, in the proposed change point identification setting, we can explicitly solve the corresponding optimisation problems to provide a computationally efficient and intuitive variant of Track-and-Stop. This is discussed in more detail in Section 4.

There has also been recent work on fixed confidence Clustering Bandits (Yang et al., 2022; Thuot et al., 2024; Yavas et al., 2025), where there are exactly N distinct mean rewards which any arm in the action space can have. The objective is to partition the action space into N sets of arms, such that each set contains arms with the same mean reward. Unlike the problem presented in this paper, in the clustering bandits problem there is no (piecewise constant) structure of the mean rewards across the action space. Hence these methods are significantly suboptimal when applied to the fixed-confidence piecewise constant bandit problem. Furthermore we additionally consider settings where we do not know the number of change points/clusters N. We discuss this further in Section 6.

Change points & Non-Stationary Bandits There is a vast literature in statistics devoted to the detection of change points in a given data set (for survey see Aminikhanghahi & Cook, 2017). Change points have appeared in the bandit literature in the context of non-stationary environments where the expected rewards from the actions change *over time* (e.g. Garivier & Moulines, 2011; Auer et al., 2019; Chen et al., 2019; Gopalan et al., 2021; Hou et al., 2024). We emphasize that this is distinct from the setting considered in this paper where the environment is stationary across time with abrupt changes in the expected rewards of the actions *across the action space*.

Change Points Across The Action Spaces Anytime methods have been proposed to identify change points or actively learn an entire piecewise smooth function under bandit feedback (Gramacy & Lee, 2008; Park et al., 2021; 2023; Hayashi et al., 2019). These methods use variants of Bayesian Optimization, but have only been assessed empirically and are not accompanied by theoretical analysis. In this paper we focus on deriving computational efficient policies that are also theoretically optimal.

Other works have also studied settings with change points across the action space under a fixed budget assumption, where the total number of samples the learner can make is bounded by a fixed $T > 0$. For example, Castro et al. (2005)

proposed an (instance-independent) nearly minimax optimal two-stage method for active learning of a whole piecewise constant function in $\mathbb{R}^d$. Similarly, Lan et al. (2009); Hall & Molchanov (2003) provided multistage-methods with asymptotic guarantees (as $T \to \infty$) for single change point (or change surface) identification in the fixed budget setting. Recently, Lazzaro & Pike-Burke (2025) provided non-asymptotic minimax optimal binary search methods for identifying a single change point in the fixed budget setting. In contrast, we consider *fixed-confidence* methods for identifying change points. Moreover, most prior work has focused on settings with exactly one change point (or surface), while we consider identifying $N \geq 1$ change points.

We also point out the concurrent work by Bacchiocchi et al. (2025), which studies a regret minimization variant of our problem setting in which the learner tries to maximize cumulative rewards received when the underlying mean rewards are piecewise linear across the action space. In this paper, however, we specifically focus on optimally identifying the location of the change points under a fixed-confidence assumption.

3. Problem Setting

We consider a multi-armed bandit setting where there are K arms ($\mathcal{A} = [K]$)[1] with mean rewards $(\mu_i)_{i=1}^K$. Note that this is equivalent to a discretization of the continuous action analogue of the problem, see (Lazzaro & Pike-Burke, 2025). We assume that the mean rewards are piecewise constant. Let $\underline{x}^* \subset [K-1]$ be the set of N change points with $|\underline{x}^*| = N$. We use the convention that if there is a change point at $i \in \underline{x}^*$, then the mean reward changes between the ith and $(i+1)$st arm, i.e. $\mu_i \neq \mu_{i+1}$. Conversely, for any arm that is not a change point, i.e. $i \in [K-1] \backslash \underline{x}^*$, we have $\mu_i = \mu_{i+1}$. In each round $t = 1, 2, \ldots$ we play an action $a_t \in [K]$ and observe a reward $y_t = \mu_{a_t} + \epsilon_t$, where the noise $\epsilon_t \sim N(0, \sigma^2)$ is i.i.d. Gaussian with mean zero and variance $\sigma^2 \in \mathbb{R}_+$. We assume that the noise variance is known and $\sigma^2 = 1$ for simplicity. The assumption of a known variance (proxy) is common in the fixed-confidence pure exploration literature (e.g. Chen et al., 2014; Jamieson et al., 2014; Barrier et al., 2022).

We define $V_{K,N}$ as the set of environments with piecewise constant mean rewards when there is a total of K actions and N change points. In particular, for any $v \in V_{K,N}$ we denote the mean rewards in v as $(\mu_{i,v})_{i=1}^K$ and the set of change points in v as

$$\underline{x}_v^* = \{j \in [K-1] : \mu_{v,j} \neq \mu_{v,j+1}\}.$$

We denote the elements of the set of change points as

$$\underline{x}_v^* = \{x_{v,1}^*, \ldots, x_{v,N}^*\}$$

such that they are indexed from left to right, namely $x_{v,1}^* < \cdots < x_{v,N}^*$. Furthermore, we assume that the change points are separated by at least one action, namely

$$\forall i \in [K-1], \quad x_{v,i}^* + 1 < x_{v,i+1}^*.$$

This is reasonable when considering our finite action space as a discretization of a continuous one which is sufficiently fine. We use $v(i)$ to denote the reward distribution of arm i in environment v, namely $N(\mu_{v,i}, 1)$, and $D(\cdot, \cdot)$ to denote the KL-divergence between two distributions. We additionally denote the size of the change in mean at the jth change point in environment v, $x_{v,j}^*$, as

$$\Delta_{v,j} = \left| \mu_{v,x_{v,j}^*} - \mu_{v,x_{v,j}^*+1} \right|$$

Finally, we denote $x_{v,(i)}^*$ as the change point with the i'th largest magnitude and denote its magnitude by $\Delta_{v,(i)}$. Hence, $\Delta_{v,(1)} \geq \cdots \geq \Delta_{v,(N)}$. We will drop the v notation when it is clear which environment we are considering.

In the fixed confidence setting we consider, the learner will be given a small fixed **confidence level** $\delta \in (0,1)$ and an **objective** for their change point estimates to be "correct". There are several natural objectives we can consider in this problem, as outlined in Sections 4 and 5. The goal of the leaner is to satisfy this objective with probability greater than $1 - \delta$, while minimizing the number of samples required to do so. In particular, we propose methods which sequentially select actions to play in each round until they are confident of the location of the change points, at which point they return the estimated change points. More formally, we define a *policy* π to consist of three rules:

1. **Sampling Rule:** A procedure to determine which action to play in each round, given the sequence of actions and rewards observed so far.
2. **Stopping Rule:** A rule defining a stopping time τ at which point we determine we have collected sufficient data and the policy stops.
3. **Recommendation Rule:** A rule to return our final estimate for the set of change points. After stopping time τ, we denote the estimate as $\hat{\underline{x}}_\tau$.

A good policy will return an estimate $\hat{\underline{x}}_\tau$ that satisfies our objective with a small expected stopping time (i.e. sample complexity), $\mathbb{E}_{\pi,v}[\tau]$, and stops almost surely in finite time, $\mathbb{P}_{\pi,v}(\tau < \infty) = 1$. Here, we define $\mathbb{P}_{\pi,v}$ to be the measure induced by all interactions between a policy π and an environment v, dropping the subscripts when it is clear which policy and environment we refer to.

4. Warm-Up: Exactly One Change Point

Although our aim is to develop methods to confidently identify N change points, we begin by considering the case

[1] Here we denote $[K] = \{1, \ldots, K\}$.

where there is a single change point. This allows us to present the key insights that we will build upon in the multiple change point setting in Section 5. In this section, we additionally assume that the learner knows there is exactly one change point. In Section 5, we will show that a single algorithm can be optimal in this setting as well when as there are potentially multiple change points, but we just want to identify one.

4.1. Instance Dependent Lower Bound

Suppose that we are in an environment with exactly one change point and the learner is given this information. Namely, it is known that $|\underline{x}^*| = 1$. In this simple setting, we want our policy to return an estimate for the change point (after stopping) which is equal to the true change point with probability greater than $1 - \delta$. This should happen after a finite number of samples almost surely. We refer to a policy which satisfies this objective as an Exact-$(1, \delta)$ policy.

Definition 4.1. Exact-$(1, \delta)$ Policy: For any $v \in V_{K,1}$, an Exact-$(1, \delta)$ policy π with stopping time τ returns an estimate $\hat{x}_\tau \in [K-1]$ satisfying

$$\mathbb{P}_{\pi,v}(\hat{x}_\tau = x^*_{v,1}) > 1 - \delta, \tag{1}$$

$$\mathbb{P}_{\pi,v}(\tau < \infty) = 1. \tag{2}$$

For any Exact-$(1, \delta)$ policy, we provide the following non-explicit, instance-dependent lower bound on its expected stopping time, which is a modification of the general lower bound first described by Garivier & Kaufmann (2016). See Appendix B for the proof.

Theorem 4.2. *For any Exact-$(1, \delta)$ policy π with stopping time τ in environment $v \in V_{K,1}$ we have*

$$\mathbb{E}_{\pi,v}[\tau] \geq c^*(v) \log\left(\frac{1}{4\delta}\right),$$

where we define

$$c^*(v)^{-1} = \sup_{\alpha \in \mathcal{P}_K} \inf_{v' \in V^{alt}_{K,1}(x^*_v)} \sum_{i=1}^{K} \alpha_i D\left(v(i), v'(i)\right). \tag{3}$$

Here $V^{alt}_{K,1}(x^*_v)$ denotes the set of environments in $V_{K,1}$ which have change point not equal to x^*_v and $\mathcal{P}_K$ is the standard K-dimensional simplex.

From (Garivier & Kaufmann, 2016) we know that solving the optimization problem in (3) can be informative both for understanding the complexity of the problem as well as for understanding what an optimal strategy would be. Let $\alpha^*_i(v)$ be the weights which solve the optimization problem (3), namely

$$c^*(v)^{-1} = \inf_{v' \in V^{alt}_{K,1}(x^*_v)} \sum_{i=1}^{k} \alpha^*_i(v) D(v_i, v'_i).$$

These $\alpha^*_i(v)$ determine the optimal proportion of samples we should allocate to each action, $i \in [K]$, in environment ν. While this optimization problem (3) is typically expensive to solve (Degenne et al., 2019), we are able to solve (3) explicitly to show the unique optimal weights are

$$\alpha^*_i(v) = \begin{cases} 1/2 & \text{if } i = x^*_v \text{ or } x^*_v + 1 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

This leads to

$$c^*(v)^{-1} = \Delta_1^2 / 8\sigma^2.$$

This proves Corollary 4.3 below. This suggests that an optimal Exact-$(1, \delta)$ policy will asymptotically focus all of the samples immediately either side of the change point, and will play these two actions equally often. This observation motivates the algorithms we develop in subsequent sections. See Appendix C for the proof of Corollary 4.3.

Corollary 4.3. *For any Exact-$(1, \delta)$ policy π with stopping time τ, a lower bound for the expected stopping time in environment $v \in V_{K,1}$ with a single change in mean of magnitude Δ is*

$$\mathbb{E}_{\pi,v}[\tau] \geq \frac{8\sigma^2}{\Delta^2} \log\left(\frac{1}{4\delta}\right).$$

4.2. Track-and-Stop Approach and Upper Bounds

In other fixed confidence bandits settings, there are two popular methods to attain asymptotic optimality. The first are Track-and-Stop methods (e.g. Garivier & Kaufmann, 2016; Degenne et al., 2019; Degenne & Koolen, 2019; Juneja & Krishnasamy, 2018) which focus on first developing lower bounds and then playing actions according to approximations of the optimal proportions α^* (which appear in the lower bounds). Secondly, in Top-Two methods (e.g. Russo, 2016; Qin et al., 2017; Jourdan et al., 2022) the learner alternates between playing a 'leader' action and a 'challenger' action which represent potential positions of the target/objective in the action space. Since we have shown in Corollary 4.3 that for our problem the optimal proportions α^* have a simple closed form expression, it is natural to consider Track-and-Stop approaches rather than Top-Two methods. Furthermore, Track-and-Stop methods have been studied in settings with multiple correct answers (Degenne & Koolen, 2019; Chen et al., 2025), leading us to believe they will be more suited to the multiple change point setting than Top-Two methods which have only been studied in settings where it is required to return one action. Finally, the main downside to Track-and-Stop methods in other settings is their potentially very large computational complexity. This is not an issue in our setting since we can explicitly solve the optimization problem in Corollary 4.3. We now show how to adapt Track-and-Stop methods, using Corollary 4.3, to construct our policy called Single Change Point

Identification (CPI) which is Exact-$(1, \delta)$ and has optimal sample complexity. CPI is stated explicitly in Algorithm 1.

4.2.1. Sampling Rule

Suppose we knew we were in environment v. Then from our lower bound in Corollary 4.3 and equation (4), we would optimally play half of our actions adjacently either side of the change point x_v^* (i.e., we would play actions x_v^* and $x_v^* + 1$ each half of the time and not play any other actions). However, in practice we do not know the environment v and therefore we do not know which actions x_v^* and $x_v^* + 1$ to play. Hence, we propose the following tracking and forced exploration rules, which are of the general form of D-Tracking introduced by Garivier & Kaufmann (2016) for best-arm identification. We note that there are alternative tracking rules in the Track-and-Stop literature such as C-tracking and proportional tracking, however these have been shown to be less effective empirically (Garivier & Kaufmann, 2016; Degenne et al., 2019).

Tracking For the tracking component, we first estimate the change point in each round t, $\hat{x}_t$. We propose to use the following simple estimator for the change point which selects the action corresponding to the largest empirical change in mean, namely

$$\hat{x}_t(S) = \text{argmax}_{a \in S} |\hat{\mu}_a(t) - \hat{\mu}_{a+1}(t)|. \tag{5}$$

Here S is the set of potential change points, which in this case is $[K-1]$, but may change when we consider multiple change points in Section 5. Here $\hat{\mu}_a(t)$ is the empirical mean of the rewards obtained from playing action a up to time t. We choose this estimator due to its simplicity, but we expect alternative estimators, such as least squares, could also lead to similar results.

Once we have an estimate for the change point in round t, $\hat{x}_t$, we can establish the tracking criteria by noting that if $\hat{x}_t$ were the true change point then we would want to play equally either side of this change in mean. Consequently, we estimate the optimal proportions for each action at time t as $\hat{\alpha}_i(t) = 1/2$ if $i \in \{\hat{x}_t, \hat{x}_t + 1\}$, and 0 otherwise. Using these estimates, our tracking criteria is to play the action with the proportion of plays furthest below the estimated optimal proportions in round t. Let $T_i(t)$ be the total number of times we have played action i up to round t. Then in round t, we play

$$a_t = \text{argmin}_{i \in \{1,\ldots,K\}} \left(\frac{T_i(t)}{t} - \hat{\alpha}_i(t) \right). \tag{6}$$

This is equivalent to playing whichever arm has been played least out of $\{\hat{x}_t, \hat{x}_t + 1\}$, namely $a_t = \text{argmin}_{i \in \{\hat{x}_t, \hat{x}_t+1\}} T_i(t)$, as in Algorithm 1 line 8.

Forced Exploration If we only do the above tracking, the quality of our estimates $\hat{x}_t$ may improve slowly over time (if at all). Hence we additionally include some forced exploration across the action space, similar to previous works (e.g. Garivier et al., 2018; Yang et al., 2022). For this, if at any time t, we have an action $i \in [K]$ that has been played less than $\sqrt{t}$ times, then we play that action (breaking ties arbitrarily) as seen in Algorithm 1 lines 5-6. The idea is that this forced exploration is large enough to force our estimates for the change point and optimal proportions to be close to their true values of x^* and α^*. Simultaneously, this forced exploration should not be so large that our policy becomes overly exploratory and suboptimal. We will show in Theorem 4.5 that this combination of tracking and forced exploration is sufficient to achieve asymptotic optimality.

4.2.2. Stopping time

We now present the second component of CPI, the stopping rule. We give an explicit definition for the stopping time and show that a policy using it satisfies the first condition (equation (1)) of Definition 4.1 for an Exact-$(1, \delta)$ policy. We emphasize that, unlike in many prior track-and-stop works, the condition for the stopping time can also be checked explicitly and directly, rather than needing to solve the optimization problem as in (10) (e.g. Yang et al., 2022; Garivier & Kaufmann, 2016). In particular, for our proposed stopping time defined in Proposition 4.4, we only need to check if the threshold has been exceeded in (7). No numerical optimization is required for checking this in each round, making it computationally efficient. This makes our proposed approach simple and further computationally efficient.

Proposition 4.4. *Define the stopping time in Algorithm 1 as*

$$\tau_\delta = \min \left\{ t : \frac{T_{\hat{x}_t}(t) T_{\hat{x}_t+1}(t)}{2(T_{\hat{x}_t}(t) + T_{\hat{x}_t+1}(t))} \hat{\Delta}^2_{\hat{x}_t} \geq \beta(t, \delta) \right\} \tag{7}$$

with

$$\hat{\Delta}^2_{\hat{x}_t} := |\hat{\mu}_{\hat{x}_t}(t) - \hat{\mu}_{\hat{x}_t+1}(t)|^2.$$

We define the threshold

$$\beta(t, \delta) = \log(t\gamma(K-1)/\delta) + 8 \log\log(t\gamma(K-1)/\delta) \tag{8}$$

with $\gamma = 2e^3 9^6 / \log(3)$. *Then in any environment* $v \in V_{K,N}$ *where* $N \geq 1$ *and any policy* π *with stopping time* τ_δ*, we have*

$$\mathbb{P}_{\pi,v}(\hat{x}_{\tau_\delta} \notin \underline{x}_v^*) \leq \delta. \tag{9}$$

Proof. For a full proof see Appendix D. This stopping time is motivated by Chernoff stopping times (Kaufmann & Koolen, 2021) of the general form

$$\min \left\{ t : \inf_{v' \in V^{alt}_{K,1}(\hat{x})} \sum_{i=1}^{K} T_i(t) D(v'(i), \hat{v}(i)) \geq \beta(t, \delta) \right\}. \tag{10}$$

Algorithm 1 Single Change Point Identification (CPI)

```
Input: confidence $\delta \in (0,1)$
Play each action once
while $Z(t) < \beta(t,\delta)$ do
    Update $\hat{\mu}_i(t)$, $\hat{\alpha}_t$, $\hat{x}_t$
    if $\min_{i\in\{1,\dots,K\}} T_i(t) < \sqrt{t}$ then
        Play action $a_t = \text{argmin}_{i\in\{1,\dots,K\}} T_i(t)$
    else
        Play $a_t = \text{argmin}_{i\in\{\hat{x}_t,\hat{x}_t+1\}} T_i(t)$
    end if
end while
Return: $\hat{x}_\tau$
```

Since it is optimal to only play either side of the change point, we demonstrate that it is sufficient to only consider the sum in (10) over actions $i \in \{\hat{x}_t, \hat{x}_t+1\}$. □

Interestingly, the form of the stopping time in (7) is reminiscent of common tests in offline change point analysis literature (e.g. see the GLR test statistic CUSUM seen in Chen & Gupta, 2012; Verzelen et al., 2020). Our algorithm CPI is given explictly in Algorithm 1, where we denote

$$Z(t) = \frac{T_{\hat{x}_t}(t)T_{\hat{x}_t+1}(t)}{2(T_{\hat{x}_t}(t)+T_{\hat{x}_t+1}(t))}\hat{\Delta}^2_{\hat{x}_t}. \tag{11}$$

4.3. Asymptotic Upper Bound

We prove an asymptotic upper bound for the expected sample complexity of CPI (Algorithm 1). The broad idea for the analysis is to firstly prove that: if the empirical mean reward from each action is well concentrated, then, due to the forced exploration, our estimates for the change point and optimal proportions $(\hat{x}_t, \hat{\alpha}_t)$ should get closer to their true values (x^*, α^*) over time. Hence, due to the tracking, the number of times we have played actions x^* and x^*+1 adjacent to the change point should quickly increase and get closer to $t/2$. Subsequently, the values of $Z(t)$ in our stopping time should increase quickly, passing the threshold β and stopping the algorithm. The analysis for our upper bounds focuses on understanding the rate at which this occurs. We state the upper bound in Theorem 4.5 with the proof detailed in Appendix E.

Theorem 4.5. *For any $v \in V_{K,1}$, using algorithm π described in Algorithm 1, we can upper bound the expected sample complexity as*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\pi,v}[\tau_\delta]}{\log(1/\delta)} \le c^*(v).$$

This demonstrates that CPI (Algorithm 1) is an asymptotically optimal Exact-$(1,\delta)$ policy according to the lower bound in Corollary 4.3. We emphasize that this optimality is tight in terms of the constants that appear in both the asymptotic upper bound and the lower bounds. In Section 5 we will illustrate that, due to the explicit solutions found for the optimization problems (3) and (10), we are able to construct simple and interpretable non-asymptotic upper bounds for the sample complexity of our methods as well.

5. Identifying Multiple Change Points

While identifying a single change point is important, in practice we may need to confidently identify N change points. Hence, we extend our results and methods from Section 4 to consider multiple change point identification.

5.1. Known number of changes

Suppose that we are in an environment with exactly N change points and the learner is given this information. In this case, we want to return an estimate for the set of N change points which is equal to the true set of change points with probability greater than $1-\delta$. We refer to a policy that does this as an Exact-(N,δ) policy.

Definition 5.1. Exact-(N,δ) policy For any $v \in V_{K,N}$, an Exact-(N,δ) policy π with stopping time τ returns an estimate $\hat{x}_\tau$ satisfying

$$\mathbb{P}_{\pi,v}(\underline{\hat{x}}_\tau = \underline{x}^*_v) > 1-\delta,$$
$$\mathbb{P}_{\pi,v}(\tau < \infty) = 1.$$

By starting with a general bound (as in Theorem 4.2), we can prove the following lower bound on the expected sample complexity of any Exact-(N,δ) policy.

Theorem 5.2. *For any Exact-(N,δ) policy π with stopping time τ, a lower bound for the expected stopping time in environment $v \in V_{K,N}$ is*

$$\mathbb{E}_{\pi,v}[\tau] \ge c_2^*(v)\log\left(\frac{1}{4\delta}\right) \tag{12}$$

$$\ge 4\sigma^2 \log\left(\frac{1}{4\delta}\right)\left(\sum_{i=1}^{N}\frac{1}{\Delta_i^2}\right), \tag{13}$$

where we define

$$c_2^*(v)^{-1} = \sup_{\alpha\in\mathcal{P}_K}\ \inf_{v'\in V^{alt}_{K,N}(\underline{x}^*_v)} \sum_{i=1}^{K}\alpha_i D(v_i, v_i'). \tag{14}$$

Here $V^{alt}_{K,N}(x^*_v)$ denotes the set of environments in $V_{K,N}$ whose set of change points are not equal to $\underline{x}^*_v$ and $\mathcal{P}_K$ is the standard K-dimensional simplex.

If the learner knows that there is exactly N change points, then knowledge regarding the location of one change point can be informative for the location of an adjacent change point. Incorporating this information means that the optimal proportion of actions we should play near one change point no longer just depends on that change point, but also on adjacent change points. Because of this, the optimization problem in equation (14) becomes more complicated and finding a general closed form solution becomes more challenging than the Exact-$(1, \delta)$ case in (3) where we do not have this coupled effect. However, we are able to lower bound $c_2^*(v)$ to provide the final lower bound (13) in Theorem 5.2. Note that the lower bound in (13) is the sum of the complexities of finding a single change point (shown in Theorem 4.3) in N different settings, up to a factor of 2. Furthermore as we will see from the rest of Section 5, compared to settings where we do know the true number of change points (Theorem 5.2), the cost of not knowing the true number of change points is asymptotically at most a factor of two in the expected stopping time (Theorems 5.4, 5.5, and 5.7).

5.2. Unknown number of changes

When we want to confidently identify N change points in our environment, it is important to provide methods which are robust to the presence of additional change points. In particular, we would like to develop policies which are able to identify N change points out of an *unknown number of change points*. In this case, we want to return an estimate for a set of N change points which is in the set of true change points $\underline{x}_v^*$, where $|\underline{x}_v^*| = m \geq N$, with probability greater than $1 - \delta$. We refer to a policy that achieves this as an Any-(N, δ) policy.

Definition 5.3. Any-(N, δ) Policy For any $v \in V_{K,m}$ where $m \geq N$, an Any-(N, δ) policy π with stopping time τ returns an estimate $\hat{\underline{x}}_\tau$ of size N satisfying

$$\mathbb{P}_{\pi,v}(\hat{\underline{x}}_\tau \subseteq \underline{x}_v^*) > 1 - \delta, \tag{15}$$

$$\mathbb{P}_{\pi,v}(\tau < \infty) = 1. \tag{16}$$

Suppose that we had an Any-(N, δ) policy π and, for now, additionally suppose that there are exactly N change points in the environment (note that this is different to the Exact-(N, δ) setting considered in Section 5.1 where the learner knows there are exactly N change points). In this case, we can show that the optimal proportions become

$$\alpha_j^* = \alpha_{j+1}^* = \frac{\frac{1}{\Delta_j^2}}{2\sum_{i=1}^N \frac{1}{\Delta_i^2}} \tag{17}$$

for $j \in \{x_{v,1}^*, \dots, x_{v,N}^*\}$, and zero for all other α_j^* values. This intuitively suggests that constructing an optimal policy requires (asymptotically) sampling most of our actions adjacent to the changes in mean. The proportion of samples around each change in mean should be inversely proportional to the size of the change squared. This allows us to prove the lower bound in Theorem 5.4 below, detailed in Appendix G.

Theorem 5.4. *For any Any-(N, δ) policy π with stopping time τ, a lower bound for the expected stopping time in environment $v \in V_{K,N}$ is*

$$\mathbb{E}_{\pi,v}[\tau] \geq 8\sigma^2 \log\left(\frac{1}{4\delta}\right)\left(\sum_{i=1}^N \frac{1}{\Delta_i^2}\right).$$

Now, suppose our aim is to confidently identify N change points out of a set of $m \geq N$ change points. In this case, intuitively the complexity of finding any N change points is lower bounded by the complexity of finding the N most easily identifiable change points, $\{x_{(1)}^*, \dots, x_{(N)}^*\}$. This is formalized in Theorem 5.5 where the first sum in (18) is over the N largest changes in mean. See Appendix H for a proof.

Theorem 5.5. *For any Any-(N, δ) policy π with stopping time τ, the expected stopping time in environment $v \in V_{K,m}$ for $N \leq m < K$ is lower bounded by*

$$\begin{aligned}\mathbb{E}_{\pi,v}[\tau] \geq & 8\sigma^2(1-\delta)\log\left(\frac{1}{4\delta}\right)\left(\sum_{i=1}^N \frac{1}{\Delta_{(i)}^2}\right) \\ & - \log(2)\left(\sum_{i=1}^m \frac{1}{\Delta_i^2}\right).\end{aligned} \tag{18}$$

Note that, as $\delta \to 0$ the lower bound in Theorem 5.5 becomes similar to the lower bound for the Exact-(N, δ) policies in Theorem 5.2 up to a constant factor of 2. Furthermore, the lower bound in Theorem 5.5 becomes similar to Theorem 5.4, except we are summing over the N largest change points rather than all of them. In Section 5.3 we provide a policy which is simultaneously Exact-(N, δ) and Any-(N, δ). This policy is asymptotically optimal when the true number of change points is unknown ($m \geq N$) and optimal up to a factor of 2 when the learner is given the exact number of true change points in the environment.

5.3. Sequential Approach and Upper Bounds

Motivated by Binary Segmentation methods from offline change point analysis (e.g. Fryzlewicz, 2014; Scott & Knott, 1974) and and the single change point setting studied in Section 4, we propose the Multiple Change Point Identification algorithm (MCPI, Algorithm 2). MCPI works by sequentially identifying one change point at a time until we have found N. In particular we repeatedly run the loop in CPI (Algorithm 1 lines 3-10), which identifies one change point, a total of N times. After each loop, we will have stopped and identified one change point. We then remove this action from the set of potential change points and add this action to the set of change points we will return at the end (Algorithm 2 line 17). After this we repeat the process until we have identified N change points in total.

Algorithm 2 Multiple Change Point Identification (MCPI)

```
1: Input: confidence $\delta \in (0,1)$
2: Input: number of changes to look for, $N$
3: Initialize $S \leftarrow [K-1]$, $\underline{\hat{x}}_\tau \leftarrow \emptyset$
4: Play each action once
5: for phase $j$ in $\{1,\dots,N\}$ do
6:   while $Z(t) < \beta(t, \delta/N)$ do
7:     Update $\hat{\mu}_i(t)$, $\hat{\alpha}_t$
8:     if $\exists i, j \in [K-1]$ s.t. (19) holds then
9:       Update $\hat{x}_t = \hat{x}_t(S)$
10:    end if
11:    if $\min_{i \in \{1,\dots,K\}} T_i(t) < \sqrt{t}$ then
12:      Play action $a_t = \text{argmin}_{i \in \{1,\dots,K\}} T_i(t)$
13:    else
14:      Play $a_t = \text{argmin}_{i \in \{\hat{x}_t, \hat{x}_t+1\}} T_i(t)$
15:    end if
16:  end while
17:  $\underline{\hat{x}}_\tau \leftarrow \underline{\hat{x}}_\tau \cup \hat{x}_t$ , $S \leftarrow S \backslash \{\hat{x}_t\}$
18: end for
19: Return: $\underline{\hat{x}}_\tau$
```

Edge Case (Multiple Equally Sized Changes) Suppose we had an Any-$(1,\delta)$ policy, which confidently identifies one change point, and there are multiple equally sized changes in the environment. For example $\Delta_{v,(1)} = \Delta_{v,(2)}$. We see from Theorem 5.5 that the lower bound only includes the complexity from finding the position of any single *one* of the largest changes. Hence, an optimal policy would simply focus its sampling efforts around any *one* of the largest changes. By simply running CPI in this edge case, there is a risk that our estimate $\hat{x}_t$ will fluctuate between different change points of equal size and therefore our actions will not be focused around just one change point. To avoid this, in MCPI we include a condition that we only update our estimate $\hat{x}_t$ when one empirical change in mean reward is sufficiently larger than another. In particular, when there exists $i, j \in [K-1]$ such that

$$|\hat{\mu}_i(t) - \hat{\mu}_{i+1}(t)| > |\hat{\mu}_j(t) - \hat{\mu}_{j+1}(t)| + r(t) \qquad (19)$$

holds (with $r(t)$ defined in (20)), we update our estimated change point to be $\hat{x}_t(S)$. If (19) does not hold in round t then we do not update our estimate, namely $\hat{x}_t \leftarrow \hat{x}_{t-1}$. This is included in lines 8-9 of MCPI (Algorithm 2). Note however, that if we know that there is exactly one change point or that there are no changes of equal size, then we do not need to check (19) before updating our estimate for the change point and this can be omitted.

We now provide a non-asymptotic upper bound for the expected stopping time of MCPI. To do so, we define

$$r(t) = \sqrt{\frac{4\log(t) + 2\log(2\log(t)) + 1/2}{(t^{1/4} - K)_+}} \qquad (20)$$

$$T_0'(\delta) = \min\left\{T \in \mathbb{N} : T - 2KT^{\frac{1}{2}} \geq \sum_{i=1}^{N} \frac{8\beta(T, \delta/N)}{(\Delta_{(i)} - 2r(T))^2}\right\}$$

$$T_1'(v) = \min\left\{T \in \mathbb{N} : r(T) < \frac{\Delta_{(N)} - \Delta_{(\ell)}}{4}\right\}.$$

Here β is the threshold function defined in equation (8). If there are exactly m change points in the environment we define $\Delta_{(m+1)} = 0$. We also let $\ell = \text{argmax}_{i>N}\{\Delta_{(i)} : \Delta_{(N)} > \Delta_{(i)}\}$ be such that $\Delta_{(\ell)}$ is the next largest change in mean *strictly* smaller than $\Delta_{(N)}$. Intuitively, $T_1'(v)$ represents the time taken for forced exploration to provide a good estimate for the positions of the change points in environment v. Additionally, $T_0'(\delta)$ represents the time taken for our tracking to put enough samples around the true change point for us to be confident in its position and stop. Using these definitions we provide the following non-asymptotic upper bound for the expected stopping time of MCPI.

Proposition 5.6. *For any environment $v \in V_{K,m}$ for $m \geq N$, the expected sample complexity for Algorithm 2 is bounded by*

$$\mathbb{E}[\tau_\delta] \leq T_0'(\delta) + T_1'(v) + 2eK.$$

Subsequently, by considering $\delta \to 0$, we provide an asymptotic upper bound for the sample complexity of MCPI and show that it is simultaneously Exact-(N,δ) and Any-(N,δ).

Theorem 5.7. *MCPI (Algorithm 2) is an Exact-(N,δ) policy and an Any-(N,δ) policy. For any $v \in V_{K,m}$ where $m \geq N$, we can upper bound the expected sample complexity of MCPI as*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\pi,v}[\tau_\delta]}{\log(1/\delta)} \leq 8\sigma^2 \left(\sum_{i=1}^{N} \frac{1}{\Delta_{(i)}^2}\right).$$

We see that the sample complexity for MCPI asymptotically matches the lower bound for any Any-(N,δ) policy (Theorem 5.5) with the correct constants. Furthermore we see that MCPI simultaneously matches the lower bound for any Exact-(N,δ) policy (Theorem 5.2) asymptotically up to (at most) a constant factor of 2. Note that we could remove this constant by directly tracking and numerically approximating the optimal proportions from (14), however this could be computationally very expensive and would not be robust to an unknown number of changes, unlike MCPI. In the special case where there is exactly $N = 1$ change point and this is known to the learner, MCPI is also optimal with correct constants (Theorem 4.3).

We note that another alternative approach to identifying any N change points could be to extend CPI by directly tracking (as in (6)) with respect to an estimate of our optimal proportions (17). We suspect this alternative approach will

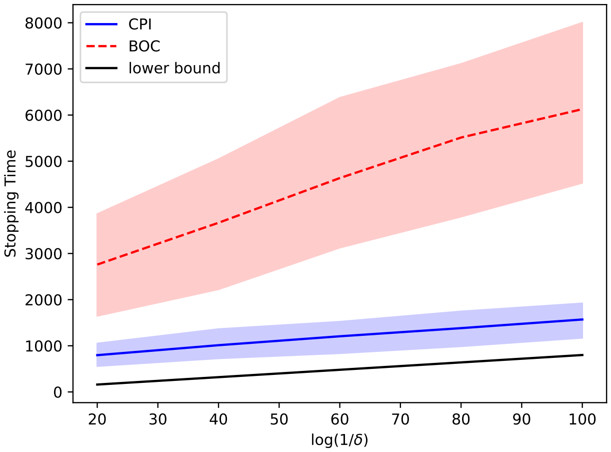

Figure 2. We run the MCPI and BOC algorithm at a range of values for δ in environment v_1. At each confidence level we repeat 100 runs and plot the average stopping time with 90 percent confidence intervals. We also plot the lower bound in Theorem 5.5.

also attain an optimal upper bound asymptotically. However, this will lead to identifying all of the change points at the same rate. Whereas if there is one change point that is significantly larger than the rest, it may be advantageous for a practitioner to quickly identify this change first. This is what the proposed MCPI algorithm does.

6. Experiments

To complement our theoretical results we conduct experiments to test our proposed algorithms in synthetic environments. In Figure 2, we consider an environment v_1 with only one change point with means $\mu = (2, 2, 2, 2, 2, 2, 1, 1, 1)$. In this case, we run our proposed policy MCPI (setting $N = 1$) and a range of choices for δ. At each of these choices for δ we run MCPI 100 times and plot the average stopping time. We also plot our lower bound from Theorem 5.5 on the same axes. From Figure 2 we see that the our expected stopping time increases parallel to the lower bound, supporting our theoretical demonstration of the asymptotic optimality of MCPI. On the same figure we plot the stopping times when, instead of tracking with respect to our estimated optimal proportions (4), we sample with respect to the oracle optimal proportions of the Bandit Online Clustering (BOC) algorithm (Yang et al., 2022). In particular, we provide BOC with the true number of change points (required as an input of BOC) as well as the oracle optimal proportions α^* described in (Yang et al., 2022) for clustering in v_1. Despite the additional information provided to BOC, it does not take advantage of the piecewise constant structure of v_1. Hence we see that BOC's sample complexity increases at a much faster rate than MCPI and the lower bound, emphasizing the sub optimality of BOC in the fixed confidence piecewise constant bandits setting. See Appendix A for further simulations with multiple change points.

7. Discussion and future work

In this paper we have studied the fixed confidence piecewise constant bandits problem. We proved multiple instance-dependent lower bounds on the expected stopping time of policies with different objectives, illustrating how the magnitude of the change points affect the complexity of our problem. Additionally, we constructed the computationally efficient MCPI algorithm to sequentially locate N change points with fixed confidence. By proving non-asymptotic and asymptotic upper bounds we showed that MCPI is asymptotically optimal under different objectives.

A related problem is the task of identifying *all* changes greater than some $\epsilon > 0$. This would be relevant in settings where there is no domain-specific or contextual knowledge of an appropriate number of change points to search for. We expect that it would be possible to extend MCPI to this setting, but we leave this to future work. Another direction for future work is to extend our setting to try to identify change points between piecewise smooth regions of a continuous action space (e.g. $\mathcal{A} = [0, 1]$). We could also consider different objectives such as minimizing the simple regret (i.e. the bias of our change point estimates) or maximizing the cumulative rewards in this piecewise constant setting. For now, we have filled a gap in the change point identification literature by providing an in-depth study of the fixed-confidence multiple change point identification problem.

Acknowledgments

The authors would like to thank Professor Niall Adams for support and insightful discussions about this problem. The authors would also like to thank the anonymous reviewers and area chairs for their feedback and suggestions during the review process. Joseph Lazzaro was supported by a Roth Scholarship from the Department of Mathematics, Imperial College London.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Aminikhanghahi, S. and Cook, D. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 2017.

Auer, P., Gajane, P., and Ortner, R. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Proceedings of the Thirty-Second Conference on Learning Theory*, 2019.

Bacchiocchi, F., Castiglioni, M., Marchesi, A., and Gatti, N. Regret minimization for piecewise linear rewards: Contracts, auctions, and beyond. In *Proceedings of the 26th ACM Conference on Economics and Computation*, 2025.

Barrier, A., Garivier, A., and Kocák, T. A non-asymptotic approach to best-arm identification for gaussian bandits. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

Bauer, H. Minimalstellen von funktionen und extremalpunkte. *Archiv der Mathematik*, 9:389–393, 1958.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. X-armed bandits. *Journal of Machine Learning Research*, 2011.

Castro, R. M., Willett, R. M., and Nowak, R. D. Faster rates in regression via active learning. In *Neural Information Processing Systems*, 2005.

Chen, J. and Gupta, A. K. *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance; 2nd ed.* Springer, 2012.

Chen, S., Lin, T., King, I., Lyu, M. R., and Chen, W. Combinatorial pure exploration of multi-armed bandits. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, 2014.

Chen, Y., Lee, C.-W., Luo, H., and Wei, C.-Y. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Proceedings of the Thirty-Second Conference on Learning Theory*, 2019.

Chen, Z., Karthik, P., Chee, Y. M., and Tan, V. Y. Optimal multi-objective best arm identification with fixed confidence. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Degenne, R. and Koolen, W. M. Pure exploration with multiple correct answers. In *Neural Information Processing Systems*, 2019.

Degenne, R., Koolen, W. M., and Ménard, P. Non-asymptotic pure exploration by solving games. In *Neural Information Processing Systems*, 2019.

Fryzlewicz, P. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42, December 2014.

Garivier, A. and Kaufmann, E. Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, 2016.

Garivier, A. and Moulines, E. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory*, 2011.

Garivier, A., Ménard, P., and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 2016.

Garivier, A., Ménard, P., Rossi, L., and Menard, P. Thresholding bandit for dose-ranging: The impact of monotonicity, 2018. URL https://arxiv.org/abs/1711.04454.

Gopalan, A., Lakshminarayanan, B., and Saligrama, V. Bandit quickest changepoint detection. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

Gramacy, R. and Lee, H. Adaptive design and analysis of supercomputer experiments. *Technometrics*, 51, 2008.

Hall, P. and Molchanov, I. Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces. *The Annals of Statistics*, 2003.

Hayashi, S., Kawahara, Y., and Kashima, H. Active change-point detection. In *Proceedings of The Eleventh Asian Conference on Machine Learning*, 2019.

Hou, Y., Tan, V. Y. F., and Zhong, Z. Almost minimax optimal best arm identification in piecewise stationary linear bandits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Jamieson, K. and Nowak, R. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, 2014.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. lil' ucb : An optimal exploration algorithm for multi-armed bandits. In *Proceedings of The 27th Conference on Learning Theory*, 2014.

Jourdan, M., Degenne, R., Baudry, D., de Heide, R., and Kaufmann, E. Top two algorithms revisited. *Advances in Neural Information Processing Systems*, 2022.

Juneja, S. and Krishnasamy, S. Sample complexity of partition identification using multi-armed bandits. In *Annual Conference Computational Learning Theory*, 2018.

Kaufmann, E. and Koolen, W. M. Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals. *Journal of Machine Learning Research*, 2021.

Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, 2008.

Lan, Y., Banerjee, M., and Michailidis, G. Change-point estimation under adaptive sampling. *The Annals of Statistics*, 2009.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

Lazzaro, J. and Pike-Burke, C. Fixed-budget change point identification in piecewise constant bandits. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Magureanu, S., Combes, R., and Proutiere, A. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Proceedings of The 27th Conference on Learning Theory*, 2014.

Park, C., Qiu, P., Carpena-Núñez, J., Rao, R., Susner, M., and Maruyama, B. Sequential adaptive design for jump regression estimation. In *ArXiv Preprint ArXiv:1904.01648*, 2021.

Park, C., Waelder, R., Kang, B., Maruyama, B., Hong, S., and Gramacy, R. Active learning of piecewise gaussian process surrogates. In *ArXiv Preprint ArXiv:2301.08789*, 2023.

Qin, C., Klabjan, D., and Russo, D. Improving the expected improvement algorithm. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Russo, D. Simple bayesian algorithms for best arm identification. In *29th Annual Conference on Learning Theory*, 2016.

Scott, A. J. and Knott, M. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, (3):507–512, 1974.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. In *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, 2010.

Thuot, V., Carpentier, A., Giraud, C., and Verzelen, N. Active clustering with bandit feedback, 2024. URL https://arxiv.org/abs/2406.11485.

Verzelen, N., Fromont, M., Lerasle, M., and Reynaud-Bouret, P. Optimal change-point detection and localization. *The Annals of Statistics*, 2020.

Yang, J., Zhong, Z., and Tan, V. Y. F. Optimal clustering with bandit feedback. *J. Mach. Learn. Res.*, 2022.

Yavas, R. C., Huang, Y., Tan, V. Y., and Scarlett, J. A general framework for clustering and distribution matching with bandit feedback. *IEEE Transactions on Information Theory*, 2025.

Appendix

A. Additional Experiments

In addition to the experiments shown in Section 6, we simulate our methods in environments with more than one change point. For example, in Figure 3 we consider and environment v_2 with $K = 19$ actions and $N = 2$ change points with mean rewards $\mu_{v_2} = (2, 2, 2, 2, 2, 2, 4, 4, 4, 4, 4, 4, 4, 0, 0, 0, 0, 0, 0)$. In Figure 3 we again plot the average stopping time when running the MPCI algorithm inputting $N = 2$ 100 times at different values for $\log(1/\delta)$. We also run BOC with oracle weights. We again see that the stopping time for BOC increases at a much faster rate than MCPI (which stays approximately parallel to the lower bound from Theorem 5.5). We see similar results when running the same experiment but using environment v_3 $K = 9$ actions and with $N = 3$ change points. Here the mean rewards are $\mu_{v_3} = (2, 2, 3, 3, 3, 3, 1, 1, 4)$. The results from simulations on this environment are in Figure 4.

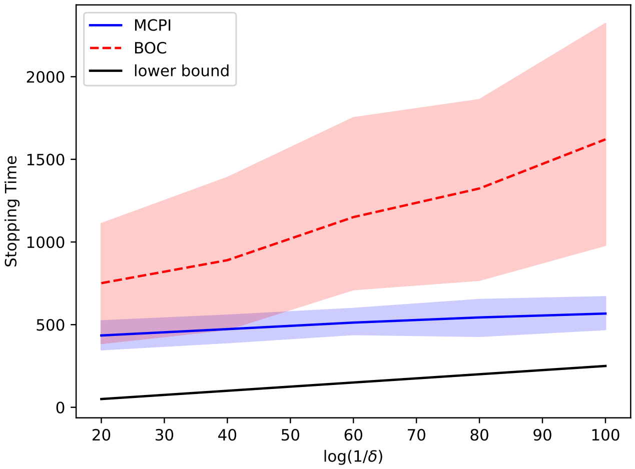

Figure 3. We run the MCPI and BOC algorithm at a range of values for δ in environment v_2. At each confidence level we repeat 100 runs and plot the average stopping time with 90 percent confidence intervals. We also plot the lower bound in Theorem 5.5.

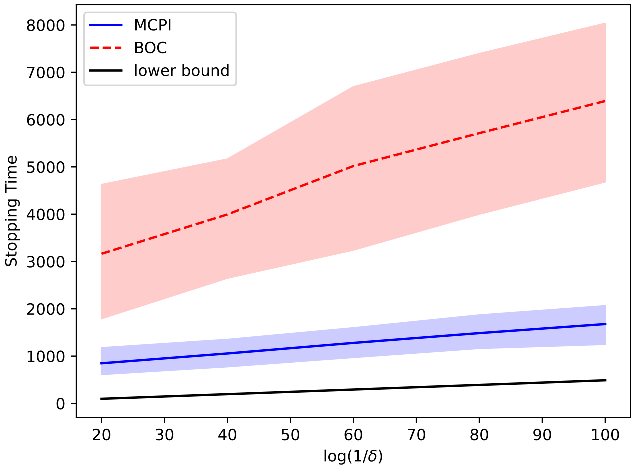

Figure 4. We run the MCPI and BOC algorithm at a range of values for δ in environment v_3. At each confidence level we repeat 100 runs and plot the average stopping time with 90 percent confidence intervals. We also plot the lower bound in Theorem 5.5.

Finally, we also run MCPI in an environment v_4 with $m = 5$ change points $\mu_{v_4} = (2, 2, 2.5, 2.5, 3, 3, 2, 2, 1.5, 1.5, 1.5, 1.5, 1.25, 1.25)$. In this case, however we only input $N = 1$ into MCPI such that MCPI will search for only one change point. In this case, where there are an additional 4 change points present, we simulate the performance of MCPI 1000 ties at each of a range of δ values. We then plot our average observed stopping time and compare this with the lower bound in Theorem 5.5 with $N = 1$. We again see that the average stopping time of MCPI is approximately parallel to the lower bound, supporting our theoretical results that MCPI is an asymptotically optimal Any-(N, δ) policy which is robust to an unknown number of additional change points present in the environment.

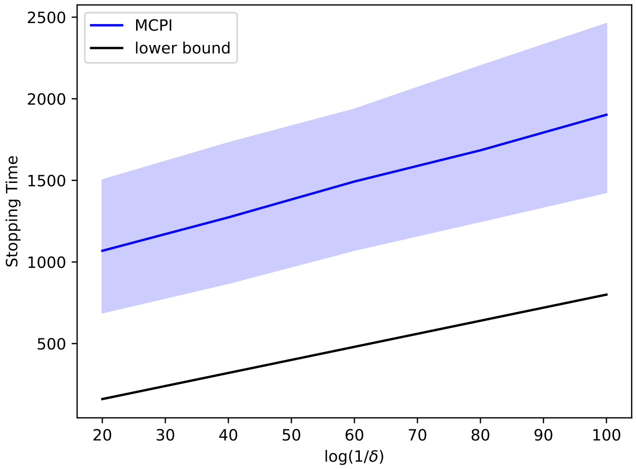

Figure 5. We run the MCPI algorithm with $N = 1$ at a range of values for δ in environment v_4. At each confidence level we repeat 100 runs and plot the average stopping time with 90 percent confidence intervals. We also plot the lower bound in Theorem 5.5.

Proofs for Section 4

B. Proof Sketch of Theorem 4.2

The proof of this Theorem can be seen in both Theorem 1 of (Garivier & Kaufmann, 2016) and Theorem 33.5 of (Lattimore & Szepesvári, 2020), where the difference is that we include additional structural assumptions on the mean rewards in our environments and we have a different objective from best arm identification. In particular, we reiterate the infimum in Theorem 4.2 is over the 'alternative' set of piecewise constant environments v' which have exactly one change point which is not equal to x_v^*.

We repeat the two important steps here for clarity and motivation. Firstly, using a change of measure argument and considering any $v' \in V_{K,1}^{alt}(x_v^*)$, they show that

$$\sum_{i=1}^{K} \mathbb{E}_{\pi,v}[T_i(t)]D(v_i, v_i') \geq \log(1/4\delta) \tag{21}$$

Secondly, they use (21) and definition of $c^*(v)$ to show the following steps

$$\begin{aligned}\frac{\mathbb{E}_{\pi,v}[\tau_\delta]}{c^*(v)} &= \mathbb{E}_{\pi,v}[\tau_\delta] \sup_{\alpha \in \mathcal{P}_K} \inf_{v' \in V^{alt}(x_v^*)} \sum_{i=1}^{K} \alpha_i D(v_i, v_i') \\ &\geq \mathbb{E}_{\pi,v}[\tau_\delta] \inf_{v' \in V^{alt}(x_v^*)} \sum_{i=1}^{K} \frac{\mathbb{E}_{\pi,v}[T_i(\tau_\delta)]}{\mathbb{E}_{\pi,v}[\tau_\delta]} D(v_i, v_i') \\ &= \inf_{v' \in V^{alt}(x_v^*)} \sum_{i=1}^{K} \mathbb{E}_{\pi,v}[T_i(\tau_\delta)]D(v_i, v_i') \\ &\geq \log(1/4\delta)\end{aligned} \tag{22}$$

The proof is then complete. □

Recall the definition of α^* as the solution of the optimisation problem in equation (4.1). Then we observe that the only way to achieve equality in equation (22) is for $\alpha_i^* = \frac{\mathbb{E}_{\pi,v}[T_i(\tau_\delta)]}{\mathbb{E}_{\pi,v}[\tau_\delta]}$ to hold. Hence, to asymptotically achieve the lower bound presented in Theorem 4.2, we want to have the proportion of times we play each action equal to α^*. Hence, we often refer to α^* as the vector of optimal proportions.

C. Proof of Theorem 4.3

We show that we can explicitly solve the following optimization problem for $c^*(v)^{-1}$.

$$c^*(v)^{-1} = \sup_{\alpha \in \mathcal{P}_K} \inf_{v' \in V^{alt}_{K,1}(x^*_v)} \sum_{i=1}^{K} \alpha_i D\left(v(i), v'(i)\right).$$

First, suppose we fix some $\alpha \in \mathcal{P}_K$. Then we try to to choose a $v' \in V^{alt}_{K,1}(x^*_v)$ to minimize

$$\sum_{i=1}^{K} \alpha_i D(v_i, v'_i). \tag{23}$$

Denote the mean rewards in environment v as $(\mu_i)_{i=1}^{K}$ and the mean rewards in environment v as $(\mu'_i)_{i=1}^{K}$.

Case 1: Consider some v' such that $x^*_{v'} > x^*_v$. Then, we can rewrite the sum in (23) as

$$\begin{aligned}\sum_{i=1}^{K} \alpha_i D(v_i, v'_i) &= \sum_{i=1}^{x^*_v} \alpha_i D(v_i, v'_i) + \sum_{i=x^*_v+1}^{x^*_{v'}} \alpha_i D(v_i, v'_i) + \sum_{i=x^*_{v'}+1}^{K} \alpha_i D(v_i, v'_i) \\ &= \frac{1}{2\sigma^2}\left(\sum_{i=1}^{x^*_v} \alpha_i(\mu_1 - \mu'_1)^2 + \sum_{i=x^*_v+1}^{x^*_{v'}} \alpha_i(\mu_K - \mu'_1)^2 + \sum_{i=x^*_{v'}+1}^{K} \alpha_i(\mu_K - \mu'_K)^2 \right) \end{aligned} \tag{24}$$

Where (24) comes from the definition of the KL divergence between two Gaussian distributions. In order to minimize the third term in (24), we choose $\mu'_K = \mu_K$. In order to minimize the first two terms, we choose

$$\mu'_1 = \mu_1 + (\mu_K - \mu_1)\frac{\sum_{i=x^*_v+1}^{x^*_{v'}} \alpha_i}{\sum_{i=1}^{x^*_v} \alpha_i + \sum_{i=x^*_v+1}^{x^*_{v'}} \alpha_i}$$

We plug these choices for μ'_1, μ'_K into (24) to get

$$\sum_{i=1}^{K} \alpha_i D(v_i, v'_i) = \frac{\Delta^2 \left(\sum_{i=1}^{x^*_v} \alpha_i\right)\left(\sum_{i=x^*_v+1}^{x^*_{v'}} \alpha_i\right)}{2\sigma^2 \left(\sum_{i=1}^{x^*_v} \alpha_i + \sum_{i=x^*_v+1}^{x^*_{v'}} \alpha_i\right)} \tag{25}$$

which is minimised when $x^*_{v'} = x^*_v + 1$.

Case 2: Now, consider some v' such that $x^*_{v'} < x^*_v$. Similar to case 1, we can minimise the sum in (23) by setting

$$\begin{aligned} \mu'_1 &= \mu_1 \\ \mu'_K &= \mu_1 + (\mu_K - \mu_1)\frac{\sum_{i=x^*_{v'}+1}^{x^*_v} \alpha_i}{\sum_{i=x^*_{v'}+1}^{x^*_v} \alpha_i + \sum_{i=x^*_v+1}^{K} \alpha_i} \\ x^*_{v'} &= x^*_v - 1 \end{aligned}$$

to attain

$$\sum_{i=1}^{K} \alpha_i D(v_i, v_i') = \frac{\Delta^2 \left(\sum_{i=x^*_{v'}+1}^{x^*_v} \alpha_i\right) \left(\sum_{i=x^*_v+1}^{K} \alpha_i\right)}{2\sigma^2 \left(\sum_{i=x^*_{v'}+1}^{x^*_v} \alpha_i + \sum_{i=x^*_v+1}^{K} \alpha_i\right)} \tag{26}$$

Combining Cases To Find α^*: The above arguments show that the 'closest' environments to v are when we shift the change point by 1 to either the left or the right. We therefore have that the unique maximizing choice for α^* will minimize the minimum of (25) and (26).

$$\begin{aligned}\underset{\alpha \in \mathcal{P}_K}{\text{argmax}} \inf_{v' \in V^{alt}(x^*_v)} \sum_{i=1}^{k} \alpha_i D(v_i, v_i') &= \underset{\alpha \in \mathcal{P}_{K-1}}{\text{argmax}} \min \left\{ \frac{\Delta^2 \left(\sum_{i=1}^{x^*_v} \alpha_i\right) \left(\alpha_{x^*_v+1}\right)}{2\sigma^2 \left(\sum_{i=1}^{x^*_v} \alpha_i + \alpha_{x^*_v+1}\right)}, \frac{\Delta^2 \left(\alpha_{x^*_v}\right) \left(\sum_{i=x^*_v+1}^{K} \alpha_i\right)}{2\sigma^2 \left(\alpha_{x^*_v} + \sum_{i=x^*_v+1}^{K} \alpha_i\right)} \right\} \\ &= \begin{cases} 1/2 & \text{if } i \in \{x^*_v, x^*_v + 1\} \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

Hence, we have attained (4). This also proves the Theorem by plugging this back into (23) and subsequently our calculation for $c^*(v)$ into Theorem 4.2. □

D. Proof of Proposition 4.4

Before starting the proof, we note that in previous works using Track-and-Stop methods, a Chernoff stopping time was used (Garivier & Kaufmann, 2016). This stopping time stops the algorithm when an infimum of a self-normalized sum (Kaufmann & Koolen, 2021; Lattimore & Szepesvári, 2020) crosses some threshold β. The general form of these stopping times are

$$\min\left\{t : \inf_{v' \in V^{alt}_{K,1}(\hat{x})} \sum_{i=1}^{K} T_i(t) D\left(v'(i), \hat{v}(i)\right) \geq \beta(t,\delta)\right\}. \tag{27}$$

where $\hat{v}$ is an estimate for the environment. However, in our setting since it is optimal to only play either side of the change point, we demonstrate that it is sufficient to only consider the self-normalized sum over actions $i \in \{\hat{x}_t, \hat{x}_t + 1\}$ since we expect the contribution from these terms to dominate. Namely, we will define our stopping time to be

$$\tau_\delta = \min\left\{t : \inf_{v' \in V^{alt}_{K,1}(\hat{x})} \sum_{i=\hat{x}_t}^{\hat{x}_t+1} T_i(t) D\left(v'(i), \hat{v}(i)\right) \geq \beta(t,\delta)\right\}. \tag{28}$$

We now prove that (28) is an appropriate stopping time. Fix an environment $v \in V_{K,1}$ and a policy π with the stated stopping time (28). Denote $\mu \in \mathbb{R}^K$ as the mean reward vector in environment v. Use any estimate for the change point. We then have the following sequence of inequalities, where the first steps are motivated by the Track-and-Stop analysis for best arm identification in Lattimore & Szepesvári (2020).

$$\begin{aligned}
\mathbb{P}_{\pi,v}(\hat{x}_\tau \neq x_v^*) &= \mathbb{P}_{\pi,v}(v \in V^{alt}(\hat{x}_\tau)) \\
&\leq \mathbb{P}_{\pi,v}\left(\frac{1}{2}\left[T_{\hat{x}_\tau}(\tau)(\hat{\mu}_{\hat{x}_\tau}(\tau) - \mu_{\hat{x}_\tau})^2 + T_{\hat{x}_\tau+1}(\tau)(\hat{\mu}_{\hat{x}_\tau+1}(\tau) - \mu_{\hat{x}_\tau+1})^2\right] \geq \beta(\tau,\delta)\right) \qquad (29) \\
&\leq \mathbb{P}_{\pi,v}\left(\exists s \in \mathbb{N}^+, \exists k \in [K-1] : \frac{1}{2}\left[T_k(s)(\hat{\mu}_k(s) - \mu_k)^2 + T_{k+1}(s)(\hat{\mu}_{k+1}(s) - \mu_{k+1})^2\right] \geq \beta(s,\delta)\right) \\
&\leq \sum_{k=1}^{K-1} \sum_{s=1}^{\infty} \mathbb{P}_{\pi,v}\left(\frac{1}{2}\left[T_k(s)(\hat{\mu}_k(s) - \mu_k)^2 + T_{k+1}(s)(\hat{\mu}_{k+1}(s) - \mu_{k+1})^2\right] \geq \beta(s,\delta)\right) \\
&\leq \sum_{k=1}^{K-1} \sum_{s=1}^{\infty} \exp(3)\left(\frac{\beta(s,\delta)\left(\beta(s,\delta)\log(s) + 1\right)}{2}\right)^2 \exp(-\beta(s,\delta)) \qquad (30)
\end{aligned}$$

Where (29) comes form the definition of our stopping time (28) and the final inequality (30) comes from Theorem 2 of (Magureanu et al., 2014). Now, we will make similar arguments to Garivier et al. (2018) to complete the proof. We first point out some helpful observations.

Lemma D.1. *The following inequalities hold for all* $t \geq 1$ *and* $\delta \in (0,1)$

(i) $\beta(t,\delta) \leq 9\log\left(\frac{t\gamma(K-1)}{\delta}\right)$

(ii) $\log(t) \leq \beta(t,\delta)$

(iii) $\beta(t,\delta) \geq 1$

Proof. These results follow from the definition of our threshold β, defined in equation (8) and the fact that $\gamma(K-1)/\delta \geq 3$. □

Now, by plugging in our definition for $\beta(t, \delta)$ into equation (30), we get the following sequence of inequalities.

$$\mathbb{P}_{\pi,v}(\hat{x}_{\tau_\delta} \neq x_v^*) \leq \sum_{k=1}^{K-1} \sum_{s=1}^{\infty} \frac{\exp(3)\delta}{4s\gamma(K-1)} \frac{\left(\beta(s,\delta)\left(\beta(s,\delta)\log(s)+1\right)\right)^2}{\log^8(s\gamma(K-1)/\delta)}$$

$$\leq \sum_{k=1}^{K-1} \sum_{s=1}^{\infty} \frac{\exp(3)\delta}{4s\gamma(K-1)} \frac{4\left(\beta(s,\delta)\right)^6}{\log^8(s\gamma(K-1)/\delta)} \tag{31}$$

$$\leq \sum_{k=1}^{K-1} \sum_{s=1}^{\infty} \frac{\exp(3)\delta}{4s\gamma(K-1)} \frac{4\left(9\log\left(\frac{t\gamma(K-1)}{\delta}\right)\right)^6}{\log^8(s\gamma(K-1)/\delta)} \tag{32}$$

$$= \sum_{k=1}^{K-1} \sum_{s=1}^{\infty} \frac{\exp(3)\delta}{s\gamma(K-1)} \frac{9^6}{\log^2(s\gamma(K-1)/\delta)}$$

$$= \sum_{k=1}^{K-1} \frac{9^6\exp(3)\delta}{\gamma(K-1)} \sum_{s=1}^{\infty} \frac{1}{s\log^2(s\gamma(K-1)/\delta)}$$

$$\leq \sum_{k=1}^{K-1} \frac{9^6\exp(3)\delta}{\gamma(K-1)} \sum_{s=1}^{\infty} \frac{1}{s\log^2(3s)} \tag{33}$$

$$\leq \sum_{k=1}^{K-1} \frac{9^6\exp(3)\delta}{\gamma(K-1)} \frac{2}{\log(3)} \tag{34}$$

$$= \delta$$

Where equations (31) and (32) come from using the facts from Lemma D.1. Furthermore, the inequality in (33) is true since we have $\gamma(K-1)/\delta \geq 3$ from our definition of γ. Inequality (34) is true by using an upper bounding integral.

$$\sum_{s=1}^{\infty} \frac{1}{s\log^2(3s)} \leq \frac{1}{log^2(3)} + \int_{s=1}^{\infty} \frac{1}{s\log^2(3s)} = \frac{1}{log^2(3)} + \frac{1}{3log(3)} \leq \frac{2}{\log(3)}$$

The final equation comes from our definition of γ. We have therefore proven that the stopping time (28) is an appropriate stopping time.

Finally, by using similar steps as in the proof of Theorem 4.3 in Appendix C we can prove the following lemma.

Lemma D.2. *The solution to the optimization problem within stopping time is* (28) *is*

$$\inf_{v' \in V_{K,1}^{alt}(\hat{x})} \sum_{i=\hat{x}_t}^{\hat{x}_t+1} T_i(t) D\left(v'(i), \hat{v}(i)\right) = \frac{T_{\hat{x}_t}(t)T_{\hat{x}_t+1}(t)}{2(T_{\hat{x}_t}(t)+T_{\hat{x}_t+1}(t))}\hat{\Delta}^2_{\hat{x}_t}$$

By plugging this into equation (28), recover the form of the stopping time in Proposition 4.4 and the proof is complete. □

E. Proof of Proposition 4.5

For this asymptotic upper bound our will stay somewhat close the structure of the original track and stop analysis from (Garivier & Kaufmann, 2016). The following Lemma is a consequence from Lemma 6 in (Degenne et al., 2019), the idea being that we define the good event E_T to be when all of the empirical means for all the arms are well behaved at all times up to time T.

Lemma E.1. *Define the event*

$$E_t = \left\{ \forall s \leq t, \forall a \in [K] \quad |\hat{\mu}_a(s) - \mu_a|^2 \leq \frac{4\log(t) + 2\log(2\log(t)) + 1}{T_a(s)} \right\} \tag{35}$$

Then

$$\forall t \geq 3, \quad \mathbb{P}_\mu(\mathcal{E}_t^c) \leq 2eK\frac{\log t}{t^2}, \quad \sum_{t=3}^{+\infty} \mathbb{P}_\mu(\mathcal{E}_t^c) \leq 2eK.$$

and

$$\sum_{t=3}^{+\infty} \mathbb{P}_\mu(\mathcal{E}_t^c) \leq 2eK \sum_{t=3}^{+\infty} \frac{\log t}{t^2} \leq 2eK \int_{x=1}^{+\infty} \frac{\log x}{x^2} dx = 2eK.$$

Now, if the good event E_t holds and our empirical mean rewards for all the arms are well behaved, then we know that even our simple estimator $\hat{x}_t$ for the change point will eventually always be correct after some time T_1. We write this formally in the following proposition.

Now, by defining

$$T_1(v) = \min\left\{ T : \frac{4\log(T) + 2\log(2\log(T)) + 1}{(\sqrt{T} - K)_+} < \Delta^2/4 \right\},$$

we get the following useful proposition.

Proposition E.2. *For any $T \geq T_1$, on the event E_T, we have that our estimate is correct. $\hat{x}_T = x^*$*

Proof. Under event E_T and for $T > T_1(v)$ we have that the empirical means at time T are well concentrated $\forall a \in [K] \quad |\hat{\mu}_a(T) - \mu_a|^2 \leq \Delta^2/4$. This is true since our forced exploration will enforce that $\forall s \in \mathbb{N} \quad T_a(s) \geq (\sqrt{s} - K)_+$. Hence, by our definition for our simple estimator (5), we know that the estimate for the change point at time T will be correct. □

Hence, if we are always correct in our estimate for the change point after time $T_1(v)$, then our estimate for the optimal proportions $\hat{\alpha}$ will be correct after time $T_1(v)$. Therefore, due to our tracking procedure, the proportions of samples we make for each action will get closer to the optimal proportions at the following rate.

Proposition E.3. *There exists a constant $T_1(v) \in \mathbb{N}_+$ (which depends only on the environment) such that for any $T \geq T_1$, on the event E_T, we have*

$$\forall t > \sqrt{T}, \max_i |T_i(t)/t - \alpha_i^*| \leq \frac{1 + (K-2)T_1(v)}{2t}$$

Now, if the proportion of plays for each action is getting closer to the optimal proportions as described in the above lemma, then we know that the number of times we play actions x^* and $x^* + 1$ will get closer to $t/2$ and therefore the value of Z_t (defined in (11) for our stopping time in equation (7)) will get larger. Hence, we are able to provide a lower bound for Z_t. In order to do so, we first make the following definition.

Definition E.4. Let

$$l(t) = \frac{1 + (K-2)T_1(v)}{2t}$$
$$r'(t) = \frac{4\log(t) + 2\log(2\log(t)) + 1}{(\sqrt{t} - K)_+}$$

Then we define $\tilde{c}(t)$ as the following quantity.

$$\tilde{c}(t) := 0.5\frac{(1/2 - l(t))^2}{1 + 2l(t)}(\Delta - r'(t))^2$$

Noting that as $t \to \infty$, we have $\tilde{c}(t) \to \frac{\Delta^2}{8} = c^*(v)^{-1}$ which was the optimal coefficient defined for the lower bound Theorem 4.3 defined in (3).

Then, as a concequence of our true proportion of plays approaching the optimal ones, we can write the following lower bound for Z_t.

Proposition E.5. *Under event $E_T(\epsilon)$ we have for all $t > \sqrt{T}$ when $T \geq 2T_1(v)$ we have*

$$Z_t \geq t\tilde{c}(t).$$

Proof. Recall the definition of Z_t, namely

$$Z_t = \frac{T_{\hat{x}_t}(t)T_{\hat{x}_t+1}(t)}{T_{\hat{x}_t}(t) + T_{\hat{x}_t+1}(t)}\hat{\Delta}^2.$$

where we define $\hat{\Delta}^2 := |\hat{\mu}_{\hat{x}_t}(t) - \hat{\mu}_{\hat{x}_t+1}(t)|^2$. The result then comes from our well behaved proportions under our well behaved event $E_T(\epsilon)$ from Proposition E.3 as well as the concentration assumptions from event $E_T(\epsilon)$ directly to lower bound $\hat{\Delta}^2$. □

Now, using similar steps to the original track and stop analysis of (Garivier & Kaufmann, 2016), we have:

For $T \geq 2T_1(v)$ and assuming E_T holds:

$$\begin{aligned}
\min\{\tau_\delta, T\} &\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}\{\tau_\delta > t\} \\
&\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}\{Z_t \leq \beta(t,\delta)\} \\
&\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}\{t\tilde{c}(t) \leq \beta(t,\delta)\} \\
&\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}\{t\tilde{c}(\sqrt{T}) \leq \beta(T,\delta)\} \\
&\leq \sqrt{T} + \frac{\beta(T,\delta)}{\tilde{c}(\sqrt{T})}
\end{aligned}$$

Hence, by defining the quantity

$$T_0(\delta) = \min\left\{T \in \mathbb{N}^+ : \sqrt{T} + \frac{\beta(T,\delta)}{\tilde{c}(\sqrt{T})} \leq T\right\}$$

we have that for any $T > \max\{T_0(\delta), 2T_1(v)\}$, then we have $E_T \Rightarrow \{\tau_\delta \leq T\}$, and therefore

$$\mathbb{P}(\tau_\delta > T) \leq 2eK\frac{\log(T)}{T^2}.$$

Subsequently, we have

$$\begin{aligned}\mathbb{E}[\tau] &= \sum_{t=0}^{\infty} \mathbb{P}(\tau_\delta > t) \\ &\leq T_0(\delta) + 2T_1(v) + \sum_{t=T_0(\delta)+T_1}^{\infty} \mathbb{P}(\tau_\delta > t) \\ &\leq T_0(\delta) + 2T_1(v) + \sum_{t=T_0(\delta)+2T_1(v)}^{\infty} 2eK\frac{\log(t)}{t^2} \\ &\leq T_0(\delta) + 2T_1(v) + 2eK \end{aligned} \tag{36}$$

From this upper bound on the expected stopping time, we not that the only dependence on δ is in $T_0(\delta)$ hence we focus on studying this since the other two terms will vanish when we divide through by $\log(1/\delta)$ (asymptotically). We can now use some arguments from (Garivier et al., 2018), where we restate their useful Lemma E.6 below, and by using our own definition for the threshold β we can carry out the following steps to upper bound $T_0(\delta)$. First we define the two events

$$\begin{aligned} H_1(\epsilon) &= \min\left\{T \in \mathbb{N}^+ : T - \sqrt{T} \geq T/(1+\epsilon)\right\}, \\ H_2(\epsilon) &= \min\left\{T \in \mathbb{N}^+ : \tilde{c}(\sqrt{T}) \in \left[\frac{\Delta^2}{8} - \epsilon, \frac{\Delta^2}{8} + \epsilon\right]\right\}, \end{aligned}$$

which allows us to initially simplify the definition for $T_0(\delta)$ as follows where we denote $C_\epsilon^* = \frac{\Delta^2}{8} - \epsilon$

$$\begin{aligned} T_0(\delta) &= \min\left\{T \in \mathbb{N}^+ : \sqrt{T} + \frac{\beta(T,\delta)}{\tilde{c}(\sqrt{T})} \leq T\right\} \\ &\leq H_1(\epsilon) + \min\left\{T \in \mathbb{N}^+ : \beta(T,\delta) \leq \frac{\tilde{c}(\sqrt{T})T}{1+\epsilon}\right\} \\ &\leq H_1(\epsilon) + H_2(\epsilon) + \min\left\{T \in \mathbb{N}^+ : \beta(T,\delta) \leq \frac{C_\epsilon^* T}{1+\epsilon}\right\} \\ &\leq H_1(\epsilon) + H_2(\epsilon) + \min\left\{T \in \mathbb{N}^+ : \log\left(T\gamma(K-1)/\delta\right) + 8\log\log(T\gamma(K-1)/\delta) \leq \frac{C_\epsilon^* T}{1+\epsilon}\right\} \\ &\leq H_1(\epsilon) + H_2(\epsilon) + \max\left(\frac{(1+\epsilon)^2}{C_\epsilon^*}\log\left(\frac{e(1+\epsilon)^2\gamma(K-1)}{C_\epsilon^*\delta}\log\left(\frac{(1+\epsilon)^2\gamma(K-1)}{C_\epsilon^*\delta}\right)\right), \frac{\delta}{\gamma(K-1)}\exp\left(g(\epsilon/8)\right)\right) \end{aligned} \tag{37}$$

Where for equation (37) we have used Lemma E.6. Hence, by plugging in the upper bound from (37) into (36), and observing the dependence of the upper bound on δ we attain the asymptotic upper bound stated in the theorem. Hence CPI has asymptotically optimal expected sample complexity.

□

Helpful Lemma

We state the following helpful lemma from (Garivier et al., 2018). For $0 < y \leq 1/e$, let g be the function defined by

$$g(y) := \frac{1}{y}\log\left(\frac{e}{y}\log\left(\frac{1}{y}\right)\right).$$

From this definition they state the following inequality in Lemma 10 (Garivier et al., 2018).

Lemma E.6. *Let $A, B > 0$, then for all $\epsilon \in (0,1)$ such that $(1+\epsilon)/A < e$ and $B/\epsilon > e$, for all $x \geq \max\left(g(A/(1+\epsilon), \exp(g(\epsilon/B))\right))$ we have*

$$\log(x) + B\log\log(x) \leq Ax$$

Proofs for Section 5

F. Proof of Theorem 5.2

We first note that the proof of the first inequality (12) is similar to Theorem 4.2 in Appendix B, except we now restrict to the set of environments with exactly N change points (i.e. $V_{K,N}$). In order to prove the second inequality (13) we simplify and only consider the set of alternative environments in which we shift one of the change points in v by one to the left or the right.

Consider the following two sets of alternatives $\{v_j'\}_{j=1}^N, \{v_j''\}_{j=1}^N \subset V_{K,N}$ and we shift only the left/right mean reward adjacent to each of the change points in v. In particular, we define

$$\mu_{v',i} = \begin{cases} \mu_{v,i+1} & \text{for} \quad i = \underline{x}^*_{v,j} \\ \mu_{v,i} & \text{otherwise} \end{cases}$$

$$\mu_{v'',i} = \begin{cases} \mu_{v,i-1} & \text{for} \quad i = \underline{x}^*_{v,j} + 1 \\ \mu_{v,i} & \text{otherwise} \end{cases}$$

Now, by denoting $V'' = \{v_j'\}_{j=1}^N \cup \{v_j''\}_{j=1}^N \subset V^{alt}_{K,N}(\underline{x}^*_v)$. We have that from the definition of $c_2^*(v)^{-1}$ that

$$\begin{aligned} c_2^*(v)^{-1} &= \sup_{\alpha \in \mathcal{P}_K} \inf_{v' \in V^{alt}_{K,N}(\underline{x}^*_v)} \sum_{i=1}^K \alpha_i D(v_i, v_i') \\ &\leq \sup_{\alpha \in \mathcal{P}_K} \inf_{v' \in V''} \sum_{i=1}^K \alpha_i D(v_i, v_i') \\ &= \sup_{\alpha \in \mathcal{P}_K} \min_j \left\{ \min \left\{ \frac{\Delta_j^2}{2\sigma^2} \alpha_{\underline{x}^*_{v,j}}, \frac{\Delta_j^2}{2\sigma^2} \alpha_{\underline{x}^*_{v,j}+1} \right\} \right\} \end{aligned} \tag{38}$$

Hence, by substituting (38) into equation (12) of Theorem 5.2 we get the following.

$$\begin{aligned} \log\left(\frac{1}{4\delta}\right) &\leq \mathbb{E}_{\pi,v}[\tau] \sup_{\alpha \in \mathcal{P}_K} \min_j \left\{ \min \left\{ \frac{\Delta_j^2}{2\sigma^2} \alpha_{\underline{x}^*_{v,j}}, \frac{\Delta_j^2}{2\sigma^2} \alpha_{\underline{x}^*_{v,j}+1} \right\} \right\} \\ &= \mathbb{E}_{\pi,v}[\tau] \frac{1}{4\sigma^2 \left(\sum_{i=1}^m \frac{1}{\Delta_i^2} \right)} \\ \implies \mathbb{E}_{\pi,v}[\tau] &\geq 4\sigma^2 \log\left(\frac{1}{4\delta}\right) \left(\sum_{i=1}^m \frac{1}{\Delta_i^2} \right) \end{aligned}$$

As required. □

G. Proof of Theorem 5.4

Consider any alternative environment $v' \in V_{K,m}$ where $m \geq N$ such that the change points in v are not a subset of the change points in v', namely $\underline{x}^*_v \not\subseteq \underline{x}^*_{v'}$. Then, since π is an **Any-**(N, δ) policy and equation (15), we have that $\underline{x}^*_v \not\subseteq \underline{x}^*_{v'}$ implies that $\hat{\underline{x}}_\tau = \underline{x}^*_v$ is a failure event in environment v' occurring with small probability less than

$$\delta \geq \mathbb{P}_{\pi,v'}(\hat{\underline{x}}_\tau = \underline{x}^*_v). \tag{39}$$

Furthermore, again due to equation (15), we have that under environment v we have $\hat{\underline{x}}_\tau \neq \underline{x}^*_v$ occurring with small probability less than

$$\delta \geq \mathbb{P}_{\pi,v}(\hat{\underline{x}}_\tau \neq \underline{x}^*_v) \tag{40}$$

Combining equation (39) and (40) gives the following, where we apply the Bretagnolle-Huber Inequality (Theorem 14.2 Lattimore & Szepesvári, 2020) in equation (41).

$$\begin{aligned}
2\delta &\geq \mathbb{P}_{\pi,v}(\underline{\hat{x}}_\tau \neq \underline{x}^*_v) + \mathbb{P}_{\pi,v'}(\underline{\hat{x}}_\tau = \underline{x}^*_v) \\
&\geq \frac{1}{2}\exp\left(-D(\mathbb{P}_{\pi,v}, \mathbb{P}_{\pi,v'})\right) && (41) \\
&\geq \frac{1}{2}\exp\left(-\sum_{i=1}^{K} \mathbb{E}_{v,\pi}[T_i(\tau)] D\left(v(i), v'(i)\right)\right) \\
\implies \log\left(\frac{1}{4\delta}\right) &\leq \sum_{i=1}^{K} \mathbb{E}_{v,\pi}[T_i(\tau)] D\left(v(i), v'(i)\right) && (42)
\end{aligned}$$

Now, consider a set of N alternative environments $\{v'_j\}_{j=1}^N \subset V_{K,m}$ where $m > N$. Let environment v'_j have mean rewards equal to v everywhere except for actions $a_j := \underline{x}^*_{v,j}$ and $b_j := \underline{x}^*_{v,j} + 1$ such that there is no longer a change point there. In particular, for some $\ell_j \in [\mu_{v,a_j}, \mu_{v,b_j}]$ we set

$$\mu_{v',i} = \begin{cases} \ell_j & \text{when} \quad i = a_j \quad \text{or} \quad i = b_j \\ \mu_{v,i} & \text{otherwise} \end{cases}$$

Now, since equation (42) holds for any of the alternative environments in the set $\{v'_j\}_{j=1}^m$, we have that the following set of inequalities.

$$\begin{aligned}
\log\left(\frac{1}{4\delta}\right) &\leq \min_j \left\{ \inf_{\ell_j} \sum_{i=1}^{K} \mathbb{E}_{v,\pi}[T_i(\tau)] D\left(v(i), v'_j(i)\right) \right\} \\
&\leq \mathbb{E}_{\pi,v}[\tau] \sup_{\alpha \in \mathcal{P}_K} \min_j \left\{ \inf_{\ell_j} \sum_{i=1}^{K} \alpha_i D\left(v(i), v'_j(i)\right) \right\} \\
&= \mathbb{E}_{\pi,v}[\tau] \sup_{\alpha \in \mathcal{P}_K} \min_j \left\{ \frac{\Delta_j^2}{2\sigma^2} \frac{\alpha_{\underline{x}^*_{v,j}} \alpha_{\underline{x}^*_{v,j}+1}}{\alpha_{\underline{x}^*_{v,j}} + \alpha_{\underline{x}^*_{v,j}+1}} \right\} && (43) \\
&= \mathbb{E}_{\pi,v}[\tau] \frac{1}{8\sigma^2 \left(\sum_{i=1}^m \frac{1}{\Delta_i^2}\right)} \\
\implies \mathbb{E}_{\pi,v}[\tau] &\geq 8\sigma^2 \log\left(\frac{1}{4\delta}\right) \left(\sum_{i=1}^m \frac{1}{\Delta_i^2}\right)
\end{aligned}$$

Where in equation (43) we have chosen the minimizing choice for l_j in each environment v'_j. Furthermore, we can show that the choice for α which attains the supremum in equation (43) is

$$\alpha^*_j = \alpha^*_{j+1} = \frac{\frac{1}{\Delta_j^2}}{2\sum_{i=1}^N \frac{1}{\Delta_i^2}}$$

for $j \in \{x^*_{v,1}, \ldots, x^*_{v,N}\}$, and zero for all other α^*_j values. for $j \in \{1, \ldots, m\}$, and zero for all other α^* values. The proof is then complete. □

H. Proof of Theorem 5.5

We will use a similar approach to the proof of Theorem 5.4 in Appendix G, except we will firstly use a slightly different change of measure argument instead of the Bretagnolle-Huber inequality. Recall from Lemma 1 in (Garivier et al., 2016) we have that

$$D(\mathbb{P}_1, \mathbb{P}_2) \geq kl(\mathbb{P}_1(A), \mathbb{P}_2(A)) \geq 0 \tag{44}$$

for some event A. Where $kl(x, y)$ is the KL-divergence between two Bernoulli distributions with parameters x and y. Recall also that

$$kl(x, y) \geq x \log(1/y) - \log(2) \tag{45}$$

Let $v \in V_{K,m}$ where $m > N$ and let π be an Any-(N, δ) policy. Now, for $j \in [m]$ define

$$A_j = \{x^*_{(j)} \in \hat{\underline{x}}_\tau\}.$$

Furthermore, let environment $v'_j \in V_{K,m+1}$ be equal to v everywhere except in $\{x^*_{v,(j)}, x^*_{v,(j)+1}\}$ where the mean reward is equal to l_j, namely

$$\mu_{v'_j,i} = \begin{cases} l_j & \text{for} \quad i \in \{x^*_{v,(j)}, x^*_{v,(j)+1}\} \\ \mu_{v,i} & \text{otherwise} \end{cases}$$

Using the transportation lemma (44), we have for any $j \in [m]$ and for any $l_j \in \mathbb{R}$ that

$$\begin{aligned} kl(\mathbb{P}_{\pi,v}(A_j), \mathbb{P}_{\pi,v'_j}(A_j)) &\leq D(\mathbb{P}_{\pi,v}, \mathbb{P}_{\pi,v'_j}) \\ &= \sum_{i=1}^{K} \mathbb{E}_{v,\pi}[T_i(\tau)] D\left(v(i), v'_j(i)\right) \end{aligned} \tag{46}$$

However, since (46) applies for any choice of l_j in environment v'_j, we can take the following infimum over the RHS.

$$\begin{aligned} kl(\mathbb{P}_{\pi,v}(A_j), \mathbb{P}_{\pi,v'_j}(A_j)) &\leq \inf_{l_j \in \mathbb{R}} \sum_{i=1}^{K} \mathbb{E}_{v,\pi}[T_i(\tau)] D\left(v(i), v'_j(i)\right) \\ &= \mathbb{E}_{\pi,v}\left[T_{x^*_{v,(j)}}(\tau) + T_{x^*_{v,(j)+1}}(\tau)\right] \frac{\Delta^2_{(j)}}{8\sigma^2} \end{aligned} \tag{47}$$

Now, (47) holds for any $j \in [m]$. Hence, if we define $k_j = kl(\mathbb{P}_{\pi,v}(A_j), \mathbb{P}_{\pi,v'_j}(A_j))$ and $t_j = \mathbb{E}_{\pi,v}\left[T_{x^*_{v,(j)}}(\tau) + T_{x^*_{v,(j)+1}}(\tau)\right]$, then from (47) gives us

$$\begin{aligned} 1 &\leq \min_{j \in [m]} \left\{ \frac{t_j \Delta^2_{(j)}}{8\sigma^2 k_j} \right\} \\ &\leq \sup_{t_j : \sum_{j=1}^{m} t_j \leq \mathbb{E}_{\pi,v}[\tau]} \min_{j \in [m]} \left\{ \frac{t_j \Delta^2_{(j)}}{8\sigma^2 k_j} \right\} && (48) \\ &= \frac{\mathbb{E}_{\pi,v}[\tau]}{\sum_{j=1}^{m} \frac{8\sigma^2 k_j}{\Delta^2_{(j)}}} && (49) \end{aligned}$$

Here (48) comes from taking a supremum over the potential values for t_j, taking into account that their sum is upper bounded by $\mathbb{E}_{\pi,v}[\tau]$ and (49) comes from finding the explicit solution to this supremum (similar to other lower bound proofs in this paper). Now, we try to lower bound the denominator of (49).

$$\sum_{j=1}^{m}\frac{8\sigma^2 k_j}{\Delta_{(j)}^2}=\sum_{j=1}^{m}\frac{8\sigma^2\, kl(\mathbb{P}_{\pi,v}(A_j),\mathbb{P}_{\pi,v_j'}(A_j))}{\Delta_{(j)}^2} \tag{50}$$

$$\geq\sum_{j=1}^{m}\frac{8\sigma^2}{\Delta_{(j)}^2}\Big(\mathbb{P}_{\pi,v}(A_j)\log(1/\mathbb{P}_{\pi,v_j'}(A_j))-\log(2)\Big) \tag{51}$$

$$\geq\sum_{j=1}^{m}\frac{8\sigma^2}{\Delta_{(j)}^2}\left(\mathbb{P}_{\pi,v}(A_j)\log(1/\delta)-\log(2)\right) \tag{52}$$

$$=\sum_{j=1}^{m}\frac{8\sigma^2}{\Delta_{(j)}^2}\mathbb{P}_{\pi,v}(A_j)\log(1/\delta)-\sum_{j=1}^{m}\frac{8\sigma^2}{\Delta_{(j)}^2}\log(2) \tag{53}$$

Here (50) comes from the definition of k_j, (51) comes from using the trick in (45), and (52) comes from the Any-(N,δ) assumption on π.

Now, by defining $p_S=\mathbb{P}_{\pi,v}(\hat{x}_\tau=S)$, we note that

$$\mathbb{P}_{\pi,v}(A_j)=\mathbb{P}_{\pi,v}(x_{(j)}^*\in\underline{\hat{x}}_\tau)=\sum_{S\subset[K]}p_S\mathbb{I}\{x_{(j)}^*\in S,|S|=N\}\geq\sum_{S\subset[K]}p_S\mathbb{I}\{x_{(j)}^*\in S,|S|=N,S\subset\underline{x}_v^*\}$$

Hence we can lower bound the first sum on the right hand side of (53) by a linear function of the variables $\{p_S:|S|=N,S\subset\underline{x}_v^*\}$, namely

$$\sum_{j=1}^{m}\frac{8\sigma^2}{\Delta_{(j)}^2}\mathbb{P}_{\pi,v}(A_j)\log(1/\delta)\geq\sum_{j=1}^{m}\frac{8\sigma^2}{\Delta_{(j)}^2}\log(1/\delta)\sum_{S\subset[K]}p_S\mathbb{I}\{x_{(j)}^*\in S,|S|=N,S\subset\underline{x}_v^*\}. \tag{54}$$

Since the right hand side of (54) is a linear function of the p_S variables and by the fact that π is Any-(N,δ) we have $1-\delta\leq\sum p_S\mathbb{I}\{|S|=N,S\subset\underline{x}_v^*\}\leq 1$. Hence, by Bauer's Maximum Principle (Bauer, 1958), the right-hand-side of (54) is minimized at an extrema of the set of potential p_S variables. Namely $p_S=1-\delta$ for some $S=S^*\subset\underline{x}_v^*$ and $p_S=0$ otherwise. In particular we see that a minimiser is with $S^*=\{x_{v,(1)}^*,\ldots,x_{v,(N)}^*\}$ such that we get

$$\sum_{j=1}^{m}\frac{8\sigma^2}{\Delta_{(j)}^2}\mathbb{P}_{\pi,v}(A_j)\log(1/\delta)\geq\min_{S\subset\underline{x}_v^*,|S|=N}\sum_{j=1}^{m}\frac{8\sigma^2}{\Delta_{(j)}^2}\log(1/\delta)\sum_{S\subset[K]}p_S\mathbb{I}\{x_{(j)}^*\in S,|S|=N,S\subset\underline{x}_v^*\}$$

$$=(1-\delta)\sum_{j=1}^{N}\frac{8\sigma^2}{\Delta_{(j)}^2}\log(1/\delta). \tag{55}$$

The, by plugging (55) back into (53). Then plugging (53) back into (49) we get our lower bound.

I. Proof of Proposition 5.6

We have the following concentration inequality as a consequence of Lemma 6 in (Degenne et al., 2019).

Lemma I.1. *Define the event*

$$\mathcal{E}_t = \left\{ \forall s \leq t, \forall a \in [K] \quad |\hat{\mu}_a(s) - \mu_a|^2 \leq \frac{4\log(t) + 2\log(2\log(t)) + 1}{T_a(s)} \right\} \tag{56}$$

Then

$$\forall t \geq 3, \quad \mathbb{P}_\mu(\mathcal{E}_t^c) \leq 2eK\frac{\log t}{t^2}, \quad \sum_{t=3}^{+\infty} \mathbb{P}_\mu(\mathcal{E}_t^c) \leq 2eK.$$

and

$$\sum_{t=3}^{+\infty} \mathbb{P}_\mu(\mathcal{E}_t^c) \leq 2eK \sum_{t=3}^{+\infty} \frac{\log t}{t^2} \leq 2eK \int_{x=1}^{+\infty} \frac{\log x}{x^2} dx = 2eK.$$

Now, if we define

$$r(t) = \sqrt{\frac{4\log(t) + 2\log(2\log(t)) + 1/2}{(t^{1/4} - K)_+}},$$

then we have that

$$\mathcal{E}_t \Longrightarrow \mathcal{E}_t' = \left\{ \forall s \in [t^{1/2}, t], \forall a \in [K] \quad |\hat{\mu}_a(s) - \mu_a| \leq r(t) \right\}.$$

This is true since, for $s \in [t^{1/2}, t]$, our forced exploration in MCPI means that $T_a(s) \geq (s^{1/2} - K)_+ \geq (t^{1/4} - K)_+$. We now state a useful Lemma which is a consequence of the definition of our simple estimator (5) and the event $\mathcal{E}_t'$.

Lemma I.2. *Define*

$$T_1'(v) = \min\left\{ T \in \mathbb{N}^+ : r(T) < \frac{\Delta_{(N)} - \Delta_{(\ell)}}{4} \right\}.$$

Then, for $t \geq T_1'(v)$*, we have that*

$$\mathcal{E}_t' \Longrightarrow B_t = \left\{ \forall s \in [t^{1/2}, t], \hat{x}_s \in \{x_{(1)}, \dots, x_{(N)}\} \right\}.$$

Note that event B_t implies that all tracking actions (see line 14 Algorithm 2) that are played in rounds $s \in [t^{1/2}, t]$ will be in the set

$$a_s \in \{x^*_{(1)}, x^*_{(1)} + 1, \dots, x^*_{(N)}, x^*_{(N)} + 1\}$$

We can now present the following Lemma.

Lemma I.3. *The total number of tracking actions MCPI will play in rounds* $s \in [t^{1/2}, t]$ *before stopping, under event* $\mathcal{E}_t'$*, is at most*

$$\sum_{j=1}^{N} \frac{8(t, \delta/N)}{(\Delta_{(j)} - 2r(t))^2}.$$

Proof. Let us denote n_j to be the number of tracking actions in which we have played action j between rounds $t^{1/4}$ and t. Also let

$$N_j = \min\{n_{x^*_{(j)}}, n_{x^*_{(j)}+1}\}$$

be the smallest number of tracking actions played by $x^*_{(j)}$ or $x^*_{(j)} + 1$.

Now, let $s \in [t^{1/2}, t]$. If we then define

$$\tilde{Z}_j(s) = \frac{T_{x^*_{(j)}}(s) T_{x^*_{(j)}+1}(s)}{2(T_{x^*_{(j)}}(s) + T_{x^*_{(j)}+1}(s))} \hat{\Delta}^2_{x^*_{(j)}}(s).$$

Then since $T_{x^*_{(j)}}(s), T_{x^*_{(j)}+1}(s) \geq N_j$, we have

$$\tilde{Z}_j(s) \geq \frac{N_j N_j}{2(N_j + N_j)} \hat{\Delta}^2_{x^*_{(j)}}(s).$$

Therefore, under our event $\mathcal{E}'_t$,

$$\tilde{Z}_j(s) \geq \frac{N_j}{4}(\Delta_{(j)} - 2r(t))^2. \tag{57}$$

Note, by the definition of our stopping time (see (7) and line 6 of Algorithm 2) and since $\beta(t, \delta/N) > \beta(s, \delta/N)$; when

$$\tilde{Z}_j(s) \geq \beta(t, \delta/N)$$

occurs, we will never play actions $x^*_{(j)}$ or $x^*_{(j)} + 1$ when tracking again. Hence, from (57), if

$$\frac{N_j}{4}(\Delta_(j) - 2r(t))^2 \tag{58}$$

holds then we will never play actions $x^*_{(j)}$ or $x^*_{(j)} + 1$ when tracking again. By isolating the N_j values in (58) and summing through $j \in [N]$ (i.e. summing over the total number of tracking action we can mae before stopping), we get the restul in the lemma.

□

Now, in rounds $s \in [1, t]$ there will at most $K\sqrt{t}$ forced exploration actions. Hence in rounds $s \in [t^{1/2}, t]$ there will be at least $(t - 2K\sqrt{t})_+$ tracking actions. We therefore get the following lemma.

Lemma I.4. *If $t > T_1(v)$ and $t > T_0(\delta) = \min\left\{T \in \mathbb{N}^+ : T - 2KT^{\frac{1}{2}} \geq \sum_{i=1}^N \frac{8\beta(T,\delta/N)}{(\Delta_{(i)} - 2r(T))^2}\right\}$. Then*

$$\mathcal{E}'_t \implies \tau \leq t$$

Now, putting everything together using Lemma I.4 and Lemma I.1 gives us the following high probability upper bound for the stopping time for MCPI.

Theorem I.5. *Let $t > \max\{T_1(v), T_0(\delta)\}$ then*

$$\mathbb{P}(\tau < t) \leq \mathbb{P}(\mathcal{E}_t^C) \leq \frac{2eK\log(t)}{t^2}$$

Then, as a consequence of Theorem I.5

$$\begin{aligned}
\mathbb{E}[\tau] &= \sum_{t=0}^{\infty} \mathbb{P}(\tau_\delta > t) \\
&\leq T_0(\delta) + T_1 + \sum_{t=T_0(\delta)+T_1}^{\infty} \mathbb{P}(\tau_\delta > t) \\
&\leq T_0(\delta) + T_1 + \sum_{t=T_0(\delta)+T_1}^{\infty} 2eK\frac{\log(t)}{t^2} \\
&\leq T_0(\delta) + T_1 + 2eK
\end{aligned}$$

As required.

□

J. Proof of Theorem 5.7

Now, to prove the asymptotic upper bound on the expectation. Note that the final two terms of the non-asymptotic upper bound (Theorem 5.6) do not depend on δ and hence they will vanish when diving through by $\log(1/\delta)$. Hence, we focus on studying the first term $T_0(\delta)$. To do so, define the following, where $\epsilon_2, \epsilon > 0$ are any small real number.

$$T_2(\epsilon_2) = \min\left\{T \in \mathbb{N}^+ : T - 2KT^{\frac{1}{2}} \geq T(1-\epsilon_2)\right\}$$
$$T_3(\epsilon_3) = \min\left\{T \in \mathbb{N}^+ : \sum_{i=1}^{N} \frac{1}{(\Delta_{(i)} - 2r(T))^2} \leq \sum_{i=1}^{N} \frac{1}{\Delta_{(i)}^2} + \epsilon_3\right\}$$

We can then write

$$\begin{aligned} T_0(\delta) &\leq T_2(\epsilon_2) + T_3(\epsilon_3) + \min\left\{T \in \mathbb{N}^+ : T(1-\epsilon_2) \geq 8\beta(T, \delta/N)\left(\sum_{i=1}^{N} \frac{1}{\Delta_{(i)}^2} + \epsilon_3\right)\right\} \\ &= T_2(\epsilon_2) + T_3(\epsilon_3) + \min\left\{T \in \mathbb{N}^+ : T(1-\epsilon_2) \geq \log(1/\delta)\frac{8\beta(T, \delta/N)}{\log(1/\delta)}\left(\sum_{i=1}^{N} \frac{1}{\Delta_{(i)}^2} + \epsilon_3\right)\right\} \end{aligned}$$

Hence, we have

$$\begin{aligned} \limsup_{\delta \to 0} \frac{\mathbb{E}[\tau]}{\log(1/\delta)} &\leq \frac{1}{\log(1/\delta)} \min\left\{T \in \mathbb{N}^+ : T(1-\epsilon_2) \geq 8\log(1/\delta)\left(\sum_{i=1}^{N} \frac{1}{\Delta_{(i)}^2} + \epsilon_3\right)\right\} \\ &= \frac{8\log(1/\delta)}{\log(1/\delta)(1-\epsilon_2)}\left(\sum_{i=1}^{N} \frac{1}{\Delta_{(i)}^2} + \epsilon_3\right) \\ &= \frac{8}{(1-\epsilon_2)}\left(\sum_{i=1}^{N} \frac{1}{\Delta_{(i)}^2} + \epsilon_3\right) \end{aligned}$$

However, ϵ_2 and ϵ_3 were arbitrary hence we can send them to zero $\epsilon_2, \epsilon_3 \to 0$ to attain the upper bound. □

K. Additional Discussion

K.1. Complexity Comparison With Best Arm Identification

In order to compare the complexity of best arm identification and change point identification we construct a simple example in which there is a unique 'best arm' (with largest mean) and exactly one change point as follows. Suppose we are in an environment with K arms (σ^2-Gaussian) with mean rewards $(\mu, \mu, \dots, \mu, \mu + \Delta)$ for some $\mu \in \mathbb{R}$. From (Garivier & Kaufmann, 2016), the complexity of the best arm identification problem in this environment will asymptotically be of order

$$K \frac{\sigma^2}{\Delta^2} \log\left(\frac{1}{\delta}\right).$$

Now, from our Corollary 4.3 the change point identification problem complexity of order

$$\frac{\sigma^2}{\Delta^2} \log\left(\frac{1}{\delta}\right).$$

We no longer have a linear dependence on K in the change point identification problem since we can use our piecewise constant structure and information from across our action space to confidently locate the change in mean. On the other hand, in the best arm identification problem, we have to sample each of the individual K arms sufficiently to confidently identify the arm with the highest mean reward.