ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240735

Identifying the Best Arm in the Presence of Global Environment Shifts

Phurinut Srisawad, Juergen Branke and Long Tran-Thanh

University of Warwick, United Kingdom

Abstract. This paper formulates a new Best-Arm Identification problem in the non-stationary stochastic bandits setting, where the means of all arms are shifted in the same way due to a global influence of the environment. The aim is to identify the unique best arm across environmental change given a fixed total budget. While this setting can be regarded as a special case of Adversarial Bandits or Corrupted Bandits, we demonstrate that existing solutions tailored to those settings do not fully utilise the nature of this global influence, and thus, do not work well in practice (despite their theoretical guarantees). To overcome this issue, in this paper we develop a novel selection policy that is consistent and robust in dealing with global environmental shifts. We then propose an allocation policy, LinLUCB, which exploits information about global shifts across all arms in each environment. Empirical tests depict a significant improvement in our policies against other existing methods.

1 Introduction

A Multi-Armed Bandit (MAB) is an abstract concept of a decision problem, where a decision maker has a choice between different actions (arms), and selecting an action yields a stochastic reward. Best-arm identification (BAI) [31], a sub-problem in MAB, aims at identifying the best among all designs/arms without caring about accumulating regret during the exploration. The standard assumption for BAI is that each arm has an underlying reward distribution that is stationary. However, in practice, the reward distributions may change over time. One possible objective in such a non-stationary setting is to track the best arm or minimise the cumulative regret over time, adapting to the environmental changes, and this has been explored extensively in the literature [20, 3, 13, 11, 21].

In this paper, we consider a different problem of identifying the design that works best in expectation, across environments. We furthermore assume that environmental changes affect the underlying reward distributions of all arms in the same additive way, i.e., the mean of the reward distribution of arm i in environment j can be described as $\mu_{ij} = \mu_i + s_j$. We call this problem Multi-Armed Bandits in the Presence of Global Environment Shifts.

This is motivated by the fact that environmental changes often influence the reward of different actions in the same way. For example, when aiming to identify the best pricing strategy for a taxi application, customer's willingness to pay may differ from day to day, based on weather or specific events such as concerts or football matches, which may similarly influence the achievable profit for all considered pricing strategies. Or consider advertising on social media, where the click-through rate of different adverts may increase and decrease synchronously over time depending on external effects such as Christmas approaching, the product being discussed in a talk show, or a celebrity wearing the product. This is confirmed by recently published examples of daily empirical means from marketing experiments with uniformly collected data [18]. The trend of empirical means of all arms is positively correlated, and their relative gaps are quite well-behaved.

Identifying the best arm under such settings is challenging, because an arm evaluated more often in more favourable environments (positive offset s_j) may appear better than an arm that was evaluated more often in less favourable environments, even though the latter is better according to the underlying (environment-independent) expected reward μ_i. Note that this setting can be considered as a special case of Corrupted Bandits [40] and Adversarial Bandits [1] where the adversary can only corrupt rewards of all arms with the same constant s_j, and the agent can only observe *when* the adversary attacks the bandits, and otherwise just receives the corrupted feedback. As such, in theory, existing robust BAI algorithms designed for adversarial environments can be applied to our setting. However, as we will show later in this paper, those algorithms can be less efficient compared to a round-robin exploration since they do not exploit information about global attacks and the notice of corruption. As such, we pose the question whether one can design efficient algorithms that work well in under such global environment shifts and perform better than the trivial round-robin policy.

Against this background, this paper proposes a novel method that takes advantage of this special setting by estimating the global shift from rewards across different arms and uses it to design a suitable statistic for an algorithm design.

Our contribution and organisation: As far as we are aware, this is the first paper to consider MAB in the presence of global environment shifts. In Section 2, we provide a formal definition of the considered problem, then discuss related work. To address the MAB problem with global environment shifts, we transform it into a regression problem in Section 3 and explain why its solution is a good choice for the best-arm predictor. In Section 4, we propose the LinLUCB allocation policy which applies the confidence bound based on a regression estimator. Numerical experiments in Section 5 are conducted to understand the effectiveness of the proposed shift estimator in different allocation policies and to examine how our proposed LinLUCB algorithm performs in various problem settings. Finally, a summary and ideas for future work will be provided in Section 6

Notation: Vectors are denoted by lowercase boldface letters and matrices by uppercase boldface letters. In general, we use a superscript of t or k to refer to its value at time step t or its k^{th} value, respec-

tively. For any integer K, $[K]$ denotes $\{1, .., K\}$. A standard basis of $\mathbb{R}^d$ is given by $\{\boldsymbol{e}_i(d)$ for $i = 1, .., d\}$ where the i^{th} coordinate of vector $\boldsymbol{e}_i(d)$ is 1, otherwise 0. For a matrix $\boldsymbol{A}$, we denote its transpose by $\boldsymbol{A}'$. An identity matrix with a size of $d \times d$ is denoted by $\boldsymbol{I}_d$. A probability measure is denoted by $\mathbb{P}$. We use $\mathbb{E}[\cdot]$ to refer to the expectation of uni- or multi-variate random variables. $\mathbb{V}[\cdot]$ and $Cov(\cdot, \cdot)$ denote the variance of a random variable and the covariance between two random variables. For a multivariate random variable, $Cov[\cdot]$ denotes its covariance matrix. Denote $\mathbb{1}[E]$ as an indicator function of event E. We denote a discrete uniform distribution and a continuous uniform distribution with parameters of minimum a and maximum b as $\tilde{\mathcal{U}}(a, b)$ and $\mathcal{U}(a, b)$, respectively.

2 Problem Formulation & Related Work

In this section, we formally define the new K-armed stochastic bandit problem in the presence of global environment shifts. We then review literature related to our setting and show how global environmental shifts negatively affect existing algorithms for finding the best arm.

2.1 BAI with Global Environment Shifts

Given a finite discrete set of arms $[K]$, the reward r_{ij} from arm i under the j^{th} environment is an *i.i.d* random variable, consisting of three components:

$$r_{ij} = \mu_i + s_j + \epsilon$$

where $\mu_i \in \mathbb{R}$ is the true quality of arm i, s_j represents a global shift on the reward of all arms that depends on the environment j, and noise is normally distributed, $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

We assume that an agent can only observe two things:

1. the reward r_{ij} if arm i is chosen during environment j, and
2. the time of an environmental change.

Note that no information about the environment is available, in particular we cannot directly observe its shift s_j, nor do we have features that describe the environment and that could be used, e.g., in contextual MAB. In addition, we suppose that shifts and noise are independent and do not make any specific assumptions about the structure of environment shifts.

The best arm is defined as the arm with the largest expected reward, $i^* = \arg\max_i \mu_i$, which is independent of the environment. Every environment j is assumed to remain valid from time $t = cp_{j-1}$ to $t = cp_j$, i.e., the environment is piecewise stationary. The duration during which an environment is valid may be stochastic, and we do not need to assume an underlying distribution for the length of environments. Note, however, that since there is no prior knowledge about s_j for each environment, a single observed reward under an environment cannot provide any statistical information about the arm. Such extremely short environment durations would thus have to be ignored in practice. Instead, we here assume that the length of environment $\Delta cp_j := cp_j - cp_{j-1} \geq 2$ for all j.

In each time step t, we can first observe whether the environment has changed, and then allocate one sample to one of the arms. Our aim is to design an allocation policy that decides which arm to sample next, given all the historical information, and a selection policy that will recommend the best arm at the end of sampling, after having exhausted the available budget of T samples, i.e., we consider a fixed budget setting. A policy π is defined as a mapping from sequences of action-reward information, including the environment ordinal, $I^t := (j^1, i^1, r_{i^1 j^1}, ..., j^{t-1}, i^{t-1}, r_{i^{t-1} j^{t-1}}, j^t)$, to a set of arms $[K]$. Figure 1 illustrates an allocation policy in such a setting where the problem has 5 Gaussian arms with different means and the same variance. We use the probability of incorrect selection (PICS) as a performance measure to assess the efficiency of policies. Since the best arm maintains its rank over environmental changes in our setting, the PICS is simply defined by the expectation of 0-1 loss function $\mathcal{L}_{0,1}(\cdot, \cdot)$ as follows

$$\text{PICS} := \mathbb{E}[\mathcal{L}_{0,1}(\hat{i}, i^*)] = \mathbb{P}\left(\hat{i} \neq i^*\right)$$

where $\hat{i}$ is the arm recommended by the selection policy, and i^* is the true best arm. There also is a so-called fixed-confidence setting aiming to minimise the amount of budget to achieve the specified PICS, but we do not consider such a setting in this paper.

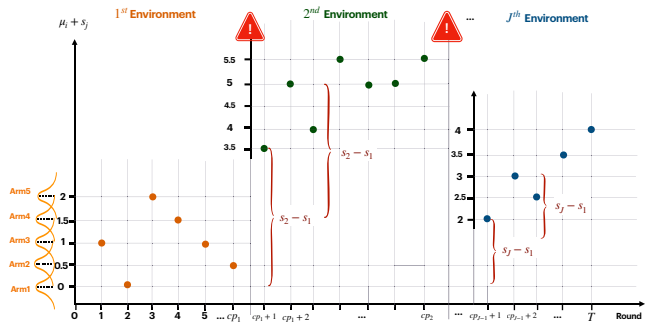


Figure 1: Example of a policy sampling from arms on a BAI problem with global environment shifts.

2.2 Related Work

Our problem formulation shares similarities with (but is different from) many papers on the problem of identifying the best arm in a non-stationary setting. Furthermore, in Section 2.3, we will explain that existing algorithms for stationary settings can identify the best arm in the long run with a high probability under some shift conditions. Therefore, we also review some literature on stationary settings.

BAI in a stationary setting: There is a rich literature on BAI algorithms which assume rewards are *i.i.d*, drawn from a stationary distribution and mostly bounded. In fixed-confidence settings, most algorithms are either elimination-based or confidence-bound-based [24], such as Exponential Gap Elimination [29], LUCB [28], and Lil'UCB [26]. A simple and efficient algorithm for the fixed-budget setting is Sequential Halving (SH), which divides the total budget into multiple phases and halves the number of candidate arms after a sampling phase ends [29]. There are also some variations of the upper confidence bound (UCB) algorithm applied to this task [12, 7].

Ranking&Selection (R&S): R&S [23] is a problem class in the stochastic simulation literature, and it is closely related to BAI in MAB, although usually Gaussian reward distributions are assumed [6]. In a fixed-budget setting, most algorithms are derived either from an equivalent problem of the PICS minimisation with a budget constraint or dynamic programming [35], such as the OCBA procedure [14], 0-1 procedure [15] and Knowledge-Gradient policy [19]. These adaptive policies allow batch sampling in multiple phases and work well in practice.

Adversarial Bandits & Corrupted Bandits: In general, adversarial settings assume that a sequence of rewards $\{r_i^t\}_{t=1}^T$ for each arm is determined by an adversary [9], which results in a reward that is

not a random variable. There are some variants of adversarial bandits which merge a stochastic structure into the problem formulation. For instance, *corrupted bandits* assume reward distributions can be attacked by an adversary which strives to trick an agent by injecting contaminated information [27, 33]. One can treat our setting as a special case of corrupted bandits where an adversary can instead choose a sequence of global shifts to fool a learner in advance. Few papers consider BAI tasks in this formulation. For a fixed-budget setting, [40] assume an agent can only observe corrupted rewards $r_i^t = \mu_i + s_i^t + \epsilon_i^t$ where an adversary has a bounded total corruption budget, $\sum_i^T \max_{i\in[K]} |s_i^t| \leq S$ for some constant S and corrupted rewards are bounded. They propose the PSS algorithm, which is an extension of the SH algorithm [29] with uniform randomisation. Besides, BAI in a corrupted model is studied in a more general way for fixed-confidence settings without any strict assumptions of true reward distribution and contaminated distribution by [5]. In adversarial bandits, the unique best arm over the total time horizon T is possibly undefined without rigorous assumptions. [1] assumes the unique best arm with respect to the highest cumulative rewards exists and studies BAI for the Best-of-Both-world problem. They propose an algorithm P1, in which the probability of sampling each arm p_i^t is generated from a ranking of the inverse-propensity-score (IPS) estimator, and the final recommendation is an arm with the highest IPS estimator. Note that the IPS estimator $\bar{\bar{r}}_i^T := (1/T)\sum_{t=1}^T r_i^t/p_i^t \cdot \mathbb{1}[i^t = i]$ is an unbiased estimator of the average reward up to time T. Another way to define the best arm is by assuming the convergence of the reward sequence or $\lim_{t\to\infty} r_i^t$ exists [25, 34] and the universal best arm is defined by the highest limit. To the best of our knowledge, no BAI study in adversarial bandits considers the global structure of change, and in Section 2.3, we will empirically show that without exploiting such a structure, these algorithms do not work well in our setting.

Piecewise-stationary Bandits: This type of bandit problem is quite relevant to our setting since it allows mean μ_i^t of the reward distribution to remain stationary within a certain time horizon Δcp_j for $j \in [J]$ where J is the number of environmental changes up to time T. Similar to adversarial settings, the task of minimising regret is more natural to study. When environments do not change too frequently, and the change is abrupt, there are three general approaches to tackle this setting [20, 3, 13, 11, 21]:

1. Reset strategy if drift is detected
2. Discounted factors to reduce the importance of rewards received long ago
3. Sliding window to only evaluate rewards from a desirable time window.

Some works also introduced an evolutionary algorithm and an adaptive allocation strategy to track the best arm under abrupt changes [30]. We are aware of only one study of BAI for piecewise stationary bandits [4]. Their setting is a generalisation of adversarial settings where an adversary chooses a sequence of reward distributions instead of a sequence of rewards. Some distributions possibly have zero variances. The best arm is defined by $i^*_{PWS} = \arg\max_i \sum_{t=1}^T \mu_i^t$. They propose the SER3 algorithm that combines a successive elimination mechanism with randomised round-robin sampling, utilising a criterion derived from Hoeffding's inequality to eliminate potentially inferior arms until only one best-predicted arm remains. In our paper, drift detection is not required since we assume the agent knows when the change occurs. Besides, our study is a fixed-budget setting, different to the fixed-confidence setting of [4]. But most importantly, we assume global shifts that affect all arms in the same way, whereas this is not the case in the other publications.

Linear Bandits: In Section 3, our reward model will be vectorised as a linear function of the index of the arm and of the environment, which is closely related to the linear relationship of feature and reward of linear bandits. For BAI in linear bandits setting, each arm i is represented by a known feature vector $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d, |\mathcal{X}| = K$. At time t, a noisy reward r^t is assumed to be a linear function of an unknown model parameter $\boldsymbol{\theta}^t \in \mathbb{R}^d$; $r^t = \boldsymbol{x}^t\boldsymbol{\theta}^{t\prime} + \epsilon^t$. In fixed-budget settings, most of the works assume the unknown parameter is fixed, $\boldsymbol{\theta}^t = \boldsymbol{\theta}^*$ for all t; therefore, the best arm is defined by the highest expected reward mean, $i^*_{LB} = \arg\max_i \boldsymbol{x}_i\boldsymbol{\theta}^{*\prime}$. [10] develops the GSE algorithm for which the total budget is evenly split into multiple phases, and a specified number of arms is eliminated after each phase ends. The GSE algorithm applies an adaptive sampling in each phase and uses the least square estimator of $\boldsymbol{\theta}^*$ to rank the arms for elimination of the worst. [39] proposes the OD-LinBAI algorithm which combines the ideas of the SH algorithm and G-optimal design [31]. [2] propose a variant of the SH algorithm equipped with the least square estimator which is robust to moderate levels of misspecification from the linear bandits model. A recent paper [37] generalises the assumption of a static model parameter to a non-stationary setting. The goal is to find the optimal arm i^* over the average model parameter $\bar{\boldsymbol{\theta}}^T = \sum_{t=1}^T \boldsymbol{\theta}^t/T$ at the specified time horizon T; $i^* = \arg\max_i \boldsymbol{x}_i\bar{\boldsymbol{\theta}}^T$. The authors propose the G-BAI algorithm, which samples the next allocation based on G-optimal design and estimates θ^t from an inverse-propensity score estimator. From the BAI in linear bandits literature, a major difference to linear bandits from our study is that the dimensionality d is fixed, whereas in our setting, the number of dimensions (environments encountered) keeps growing. In order to apply linear bandit algorithms in our setting, since there is no feature about the environment apart from a growing index of environment, a tabular approach and an approach of averaging the model parameter will not be very effective.

2.3 *Effect of Environment Change on Existing Policies*

In our setting, the global shift can affect policies in two major ways:

1. the behaviour of the adaptive allocation policy, and
2. the selection of the best-predicted arm.

We consider a sample mean of reward, which is one of the most commonly used statistics in BAI algorithms such as SH, UCB, and LUCB, including the criteria of the selection policy of round-robin sampling. Denote $\bar{r}_i := \sum_{j=1}^J (\sum_{k=1}^{n_{ij}} r_{ij}^k)/\sum_{j=1}^J n_{ij}$ as the sample mean of arm i where J is the latest environment during sampling, r_{ij}^k is the k^{th} reward or arm i in environment j, and n_{ij} is the number of samples on arm i under the j^{th} environment. Under our setting the difference of sample means between arm i_1 and i_2, $\bar{r}_{i_1} - \bar{r}_{i_2}$ contains the term of $\sum_{j=1}^J s_j \left(n_{i_1 j}/\sum_{j=1}^J n_{i_1 j} - n_{i_2 j}/\sum_{j=1}^J n_{i_2 j}\right)$. From such a calculation, the influence of the environment can lead to biased sample means and biased differences if the numbers of samples of each arm under each environment are different. For example, in the case of only one environment change happening or $J = 2$, suppose an inferior arm i_1 such that $\mu_{i_1} - \mu_{i_2} < 0$ has more samples than a superior arm i_2 in the second environment $n_{i_1 2} \geq n_{i_2 2}$ meanwhile for the first environment they have an equal number of samples $n_{i_1 1} = n_{i_2 1}$. If s_2 is sufficiently larger than s_1 then decision-makers may select an inferior arm due to $\bar{r}_{i_1} - \bar{r}_{i_2} > 0$.

Such a calculation is a main issue for the sample-mean-based final selection if an adaptive allocation policy is used. This phenomenon can also occur in elimination-based algorithms, even when uniform

sampling is used, since the change cannot be controlled. We may deduce that the sample mean is not a suitable statistic for both allocation policy and selection policy if no knowledge about the shift is provided. However, if the shift satisfies the conditions in Corollary 2 of [17], such as shift is a uniform random variable, existing BAI algorithms that sample all arms sufficiently under different environments will be able to identify the best arm with a high probability. The main reason is the shift term in the sample-mean calculation $\sum_{j=1}^{J} s_j \left(n_{i_1 j} / \sum_{j=1}^{J} n_{i_1 j} - n_{i_2 j} / \sum_{j=1}^{J} n_{i_2 j} \right) \to 0$ as $J \to \infty$ for all $i \in [K], j \in [J]$.

Another approach is to use the IPS estimator, which is an unbiased estimator for randomisation-based algorithms in adversarial settings. However, with the same reason as sample mean calculation, insufficient sampling for some arms in some environments can still cause a bias for ranking the IPS estimator since a probability-weighted reward in a favourable environment can be excessive when it is compared to the one in a less favourable environment. Lastly, implementing robust BAI algorithms in contaminated bandits could alleviate the estimator problem, but without exploiting the global shift structure, that algorithm still needs high budgets to identify the best arm.

Figure 2 depicts how different existing policies perform under the presence of global shifts when the shift is relatively big in comparison to the gap between optimal arm and suboptimal arm. On the horizontal axis, the sample average refers to the given budget T for each policy except the SER3 algorithm, where it means the average of the required number of samples to achieve different PICSs. The round-robin sampling is executed as a simple baseline. For UCB-based algorithms, LUCB [24] with the sample-mean-based recommendation and UCB [8] for minimising cumulative regret with the most-frequency-based recommendation are implemented by using the normal confidence bound in [8]. For an algorithm in adversarial settings, the P1 algorithm and the EXP4P algorithm [38] with different final recommendations are executed; one is the sample mean, and another is the IPS estimator. For BAI in contaminated bandits, we apply the PSS(2) algorithm with a slight modification by using a randomised round-robin instead of uniform randomisation to ensure each arm is sampled equally. In addition, we mimic such an idea by testing the Successive Rejects (SR) algorithm [7] with a randomised round-robin sampling. We also implement the 0-1_1 procedure [16] from R&S literature which works well in practice with the Gaussian distribution assumption. The PICS plot of these adaptive policies decreases significantly slower compared to the round-robin sampling when the budget is higher. SR algorithm performs slightly worse than round-robin sampling as sample-mean-based elimination criteria have more risk in this setting. Interestingly, the PICS of the PSS algorithms show a significant difference even if they use the same sampling policy. Two major reasons are that first, eliminating half of the candidate arms in the first phase by using a sample mean has a higher risk of excluding the optimal arm than one-arm elimination, and second the sample mean in the PSS algorithm is computed from rewards in one particular phase which is not sufficient to reduce the influence of shift in the sample mean calculation. The best policy is the SER3 algorithm, which is quite robust to global change, even though the elimination criteria are built on the bounded reward assumption. However, this policy is not quite suitable for use in fixed-budget settings since we need to tune the hyperparameter of the probability of selecting the best to match the limited budget. This result, hence, raises the question of whether there is a better estimator and adaptive policy compared to uniform-exploration-based sampling.

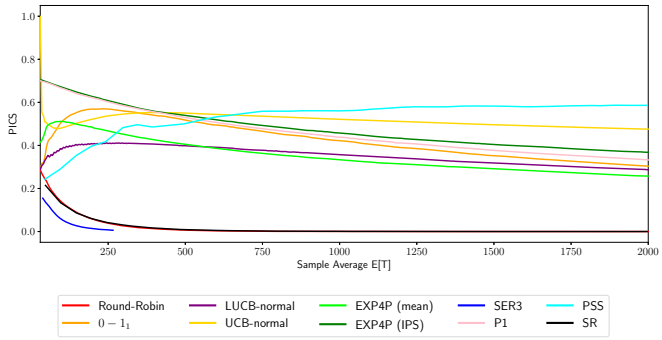


Figure 2: PICS of existing algorithms from 10^5 replications on the Gaussian configuration of 5 arms where the gaps of ordered arms ($\delta = 0.5$) are equally distributed and arms have equal variance ($\sigma = 1$). The lengths of environments j are uniformly distributed, $\Delta cp_j \sim \tilde{\mathcal{U}}(2, 50)$ and the shift is a random variable, $s_j \sim \mathcal{U}(0, 20)$.

3 Linear Regression for The Selection Policy

As explained in Section 2.3, even if s_j is bounded, a sample mean of rewards may not be an appropriate statistic for predicting the best arm since different arms may have been evaluated under different environment shifts. In the following, we derive a point estimate by formulating a regression problem.

3.1 Ordinary Least Square (OLS) Estimator

Since we are only interested in identifying the best arm, without loss of generality, we assume that $s_1 = 0$. Considering the stated problem as a regression model, a reward matrix, given a total number of evaluations N across J environments, can be rewritten in two ways as follows

$$\boldsymbol{r} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{s} + \boldsymbol{\epsilon} \tag{1}$$

$$\boldsymbol{r} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{2}$$

where

- $\boldsymbol{r} := [r^1 \; r^2 \; \dots \; r^N]'$ is a column vector containing all rewards obtained from N evaluations
- $\boldsymbol{\epsilon} := [\epsilon^1 \; \epsilon^2 \; \dots \; \epsilon^N]'$ is a corresponding noise vector where $\boldsymbol{\epsilon} \sim \mathcal{MN}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N)$.
- $\boldsymbol{A}$ is a coefficient matrix in which each row $\boldsymbol{a}_t$ is a vector of the standard basis of $\mathbb{R}^K$ referring to the chosen arm i^t. i.e., $\boldsymbol{a}^t = \boldsymbol{e}'_K(i^t)$
- Similarly, $\boldsymbol{B}$ is a coefficient matrix referring to the environment ordinal j^t, i.e., each row $\boldsymbol{b}_t = \boldsymbol{e}'_{J-1}(j^t - 1)$ for $j \geq 2$.
- $\boldsymbol{\mu} := [\mu_1 \; \dots \; \mu_K]'$ is a $K-$dimensional column vector containing the actual means of all arms.
- $\boldsymbol{s} := [s_2 \; \dots \; s_J]'$ is a $(J-1)-$dimensional column vector containing actual shifts relative to the first environment.
- $\boldsymbol{X} = [\boldsymbol{A} \; \boldsymbol{B}]$ is a coefficient block matrix. Similarly, the $(K+J-1)-$dimensional joint parameter vector $\boldsymbol{\theta} = [\boldsymbol{\mu}' \; \boldsymbol{s}']'$.

Note that the dimensions of $J-1$ and $K+J-1$ are due to the zero-valued shift s_1 assumption; therefore, such a shift will not be estimated. Model (1) is a hybrid linear model similar to the model in [32]. One difference is the dimension of our parameters $\boldsymbol{s}$, which grows by 1 when transitioning to a new environment, but the values of parameters in previous environments are unchanged.

To find the solution to the regression problem, the second model (2) is easier to solve. Based on a least squares method, we can derive a unique solution;

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{r}.$$

Note that such an estimator is unbiased ($\mathbb{E}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$) which means that the estimated mean and shift are also unbiased; $\mathbb{E}[\hat{\mu}_i] = \mu_i, \mathbb{E}[\hat{s}_j] = s_j$. In addition, the distribution of OLS estimators is $\hat{\boldsymbol{\theta}} \sim \mathcal{MN}(\boldsymbol{\theta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$, provided that $\boldsymbol{X}$ is fixed, due to Gaussian noise assumption. By using block matrix inversion, we can separate the solution for each parameter as follows

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= \left(\boldsymbol{A}'(\boldsymbol{I}-\boldsymbol{H_B})\boldsymbol{A}\right)^{-1}\boldsymbol{A}'(\boldsymbol{I}-\boldsymbol{H_B})\boldsymbol{r}\\ \hat{\boldsymbol{s}} &= \left(\boldsymbol{B}'(\boldsymbol{I}-\boldsymbol{H_A})\boldsymbol{B}\right)^{-1}\boldsymbol{B}'(\boldsymbol{I}-\boldsymbol{H_A})\boldsymbol{r}\end{aligned}$$

where $\boldsymbol{H_A} := \boldsymbol{A}\left(\boldsymbol{A}'\boldsymbol{A}\right)^{-1}\boldsymbol{A}'$ and $\boldsymbol{H_B} := \boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B}\right)^{-1}\boldsymbol{B}'$. In addition, the covariance of both estimators can be computed by

$$\begin{aligned}Cov[\hat{\boldsymbol{\mu}}] &= \sigma^2[\boldsymbol{A}'(\boldsymbol{I}-2\boldsymbol{H_B}+\boldsymbol{H_B}\boldsymbol{H_A}\boldsymbol{H_B})\boldsymbol{A}]^{-1}\\ Cov[\hat{\boldsymbol{s}}] &= \sigma^2[\boldsymbol{B}'(\boldsymbol{I}-2\boldsymbol{H_A}+\boldsymbol{H_A}\boldsymbol{H_B}\boldsymbol{H_A})\boldsymbol{B}]^{-1}.\end{aligned}$$

In the case that a common variance σ^2 is not known, an unbiased estimator for such variance can be calculated from the following formula

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{J}\sum_{i=1}^{K}\sum_{k=1}^{n_{ij}}\left(r_{ij}^k - \hat{\mu}_i - \hat{s}_j\right)^2}{N-(K+J-1)}.$$

Consistency of mean estimator

The key challenge of this work is whether the mean estimator can guarantee the correct ranking in the long run since the dimension of the parameter keeps growing. Due to our assumption that $s_1 = 0$, the correlation between estimated parameters is likely not to vanish, leading to an inconsistent mean estimator. Nevertheless, the estimated ranking is more crucial to identify the best arm; we therefore consider the difference between two mean estimators (ranking) instead.

Theorem 1. *For any policy under which the OLS estimator is valid and all arms are sampled infinitely often, or $N_i := \sum_{j=1}^{J} n_{ij} \to \infty$ for all i, assume that $J \to \infty$ and there exist constants v^*, w^* such that $0 < v^* \leq \mathbb{V}[\hat{s}_j], Cov(\hat{s}_j, \hat{s}_m) \leq w^* < \infty$ for all $j \neq m$.*

1) The mean estimator $\hat{\mu}_i$ is not consistent.
2) If $J \in o(N)$ and $N_1, N_2 \in \Theta(N)$, the difference in mean estimators between those two arms, $\hat{\mu}_1 - \hat{\mu}_2$, is consistent.

The above theorem implies that when all arms are sampled infinitely often, the mean estimator does not converge to its true value, but it can be used to identify the best arm since the ranking still converges to the actual one when the environment change grows sublinearly, and the number of samples for each arm grows linearly. In addition, the consistency of difference holds empirically without such additional assumptions. Besides, the mean estimator and the difference estimator benefit from robustness against environmental shifts since their consistency does not depend on the shift magnitude. Due to the page limit, the proof and discussion are provided in supplementary material [36].

3.2 Requirements for Regression

Merely merging an OLS estimator with an allocation policy may lead to an ill-posed problem due to a singularity of matrix $X^T X$. In linear bandits, the singularity problem is alleviated by e.g. adding a regularisation term in the regression loss function [32, 22] or applying a dimensionality reduction technique [39, 10]. However, in our setting, these approaches are not helpful if all arms are not observed in the same environment or the loss function can be partitioned and optimised separately since mean estimators are not comparable. Therefore, in general, any allocation policy that applies regression and is not aware of environmental change cannot be directly implemented. In addition, evaluating only one arm in one environment can lead to unchanged mean estimators since an estimated shift in such an environment can be varied arbitrarily. To ensure the existence and uniqueness of the regression solution, there are a few requirements for allocation policies.

Initialisation for Regression

Disconnected evaluations across different environments can cause an ill-posed optimisation problem. For instance, given a 5-arms setting, if a policy evaluates arms $\{1, 2\}$ under the first environment, arms $\{1, 2, 3\}$ under the second environment and arms $\{4, 5\}$ under the third environment, then parameters of the loss function will be separated into two partitions for arms 1, 2, 3 and arms 4, 5. The information share of regression parameters, in fact, can be represented by a graph where the vertices are arms, and the undirected edge between two vertices exists if the corresponding arms have been sampled under the same environments. From the mentioned example, we can represent it with two sub-graphs where one is $1-2-3-1$, and another is $4-5$ as in the top row of Figure 3. So the estimators of arm 1 and arm 4 are not comparable. The regression approach requires a connected graph connecting all arms to fully share information - if the graph is disconnected, it is impossible to rank solutions from different arms relative to each other. The initialisation phase is crucial for every allocation policy to generate at least a tree structure.

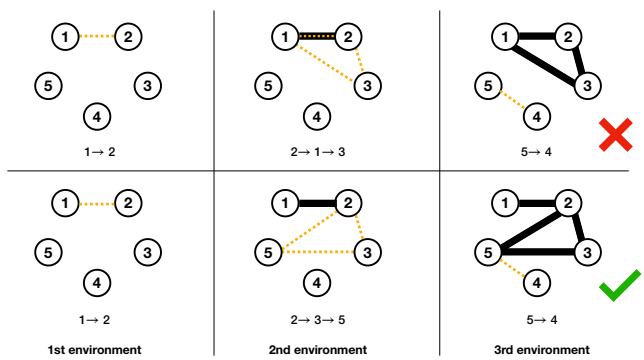


Figure 3: Representative graph structure illustrates how an allocation policy produces the evolution of the graph at the end of each environment.

For the initialisation phase, the *randomised round-robin sampling* is modified to evaluate the last arm chosen under the previous environment at the start of the next environment if all arms cannot be observed within one environment. The pseudocode is provided in Algorithm 1. For example, sequentially evaluate arm $1 \to 2$ (in the 1^{st} environment) $\to 2 \to 3 \to 5$ (in the 2^{nd} environment), and $\to 5 \to 4$ (in the 3^{rd} environment) as in the bottom row of Figure 3. With such initialisation, the estimators of arm 1 and arm 4 can be quantitatively compared through arm 2 and then arm 5.

Evaluating two first distinct arms when the environment changes

Even if an environment length is relatively short, some allocation policies may evaluate only one arm under the same environment. This would not provide any valuable information since the estimated shift under such an environment can be any arbitrary value subject to

Algorithm 1 Randomised round-robin sampling for initialisation

Require: Number of initial samples per arm $n_0 \geq 2$
Set the initial ordinal of environment $j = 1$
Set an initial arm set $S = [K]$ and shuffle.
Set arm i^1 as the first indexed arm in S.
for $t = 1, ..., n_0 K$ **do**
  if EnvChange==True **then**
    Set the ordinal of the environment: $j \leftarrow j + 1$
    if BuildTreeSuccess==False **then**
      Play arm $i^t \leftarrow i^{t-1}$
    else
      Play arm i^t from S in an order following i^{t-1}
    end if
  else
    Play arm i^t from S in an order following i^{t-1}
  end if
  Obtain a reward r^t
  Remove arm i^t from S if its number of samples $N_{i^t} = n_0$
  if The last indexed arm in S is played **then**
    Shuffle S
    BuildTreeSuccess $\leftarrow$ True
  end if
end for

the value of the estimated mean. In other words, there are no updates in estimator values if only one arm is evaluated in one environment. In order to avoid such an issue, evaluating at least two distinct arms once the environment changes is imperative.

4 LinLUCB Allocation Policy

Given a normal distribution of OLS estimator for the actual means $\hat{\boldsymbol{\mu}}$ at time t, the upper confidence bound of the actual mean of arm i can be defined as

$$\text{UCB}_i^t = \hat{\mu}_i + \gamma_i^t \sqrt{\boldsymbol{a}^{t\prime}\sigma^2[\boldsymbol{A}'(\boldsymbol{I} - 2\boldsymbol{H_B} + \boldsymbol{H_B}\boldsymbol{H_A}\boldsymbol{H_B})\boldsymbol{A}]^{-1}\boldsymbol{a}^t}$$

where γ^t is an exploration rate at a time step t. We propose a new variant of the LUCB algorithm modified from [28] for our linear model. The LUCB algorithm was originally designed for a PAC subset selection in a fixed-confidence setting where sampling two arms every time step is allowed. However, we found its potential to be implemented in a fixed-budget setting, especially in our setting. The LUCB algorithm ensures that at least two arms are evaluated in every environment, allows for an adaptive budget T, and is optimal in a two-armed setting with the worst environment length of 2. The algorithm starts by executing Algorithm 1 for the initialisation phase and then alternating samples, the greedy arm and the most potentially best arm from the rest, while guaranteeing that the two first samples in the new environment are distinct. At time t, the greedy arm is indexed based on the highest mean estimator $l^t := \arg\max_{i\in[K]} \hat{\mu}_i$, in which ties are broken arbitrarily, then the rest of arms are filtered to find the highest UCB arm $u^t := \arg\max_{i\in[K]\setminus\{l_t\}} \text{UCB}_i^t$. In this part, since our setting does not allow the sampling of two arms in one time step, we mimic the batch sampling by sequentially selecting l^t and u^t instead of using the interleaving strategy. If there is an environment change and the choice of second sampling in such a new environment is the same as the choice of first sampling, we can swap the sampling order of l^t and u^t to ensure two first choices of sampling are different. Motivated by the UCB1-normal algorithm from [8], we use $\gamma_i^t = \sqrt{16\ln(t)/\sum_{j=1}^{J} n_{ij}}$ as an exploration rate. Finally, the selection policy chooses the highest OLS mean estimator as the best-predicted arm. The pseudocode of LinLUCB policy is provided in supplementary material [36].

5 Empirical Evaluation

In order to understand how environmental change influences different policies on various configurations, we conduct numerical experiments for the proposed algorithm and modified versions of some existing policies. We chose the examined problem settings from [16] since it was a seminal paper developing a policy for PICS minimisation for Gaussian rewards. Two configurations are *monotone decreasing means* (MDM) configuration and *slippage* configuration (SC), with a modification by adding random shifts $s_j \sim \mathcal{U}(0, 20)$. For the MDM configuration, rewards for alternatives $i = 1, ..., K$ are

$$r_{ij} \sim \mathcal{N}\left(\delta(i-1) + s_j, \sigma^2\right),$$

while for the SC configuration, rewards are

$$r_{ij} \sim \mathcal{N}\left(s_j, \sigma^2\right) \text{ for } 1 \leq i < K, \quad r_{Kj} \sim \mathcal{N}\left(\delta + s_j, \sigma^2\right).$$

We use PICS as the performance measure estimated by the fraction of replications selecting the true best alternative correctly. For a fair comparison, all procedures in all time steps share the same set of potential observations by controlling random seeds. The PICS convergence plots below are generated using 10^5 replications. We set the value of parameters in the problem as $\delta = 0.5$ and $\sigma = 1$.

5.1 Comparison against standard policies

From Theorem 1, the uncertainty of the estimated shift plays a vital role in the convergence of the mean OLS estimator. Since the environment length has a significant influence on the shift estimation, we test the performance of our proposed LinUCB policy against other existing policies in the following environmental change scenarios with 5 arms, additional results on different scenarios can be found in the supplementary material [36].

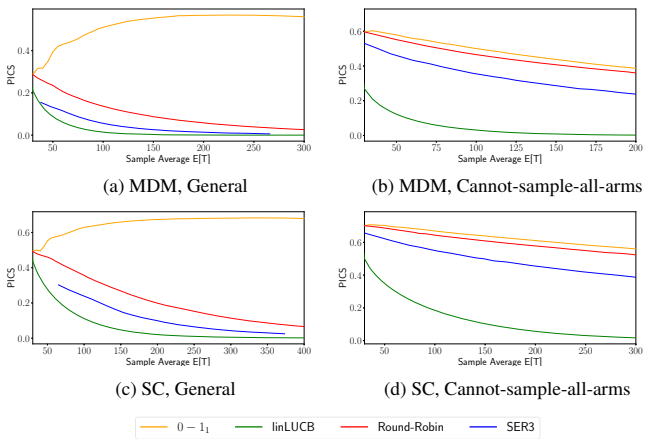
- **General scenario,** where $\Delta cp_j \sim \tilde{\mathcal{U}}(2, 10K)$: The duration of the stationary phase of the environment may vary from very short to relatively long, leading to different challenges in estimating.
- **Cannot-sample-all-arms scenario,** where $\Delta cp_j \sim \tilde{\mathcal{U}}(2, K-1)$: The environment is very short and policies cannot explore all arms in one environment

The following list describes the tested policies:

- **Round-robin**: round-robin sampling with $\bar{r}_i$ as a selection policy
- **0-1**$_1$: procedure proposed in [16] and $\bar{r}_i$ as a selection policy
- **SER3**: the elimination-based algorithm from [4] for fixed-confidence piecewise-stationary bandits where prior knowledge about the optimal gap $\mu_1 - \max_{i\neq 1} \mu_i$ is provided
- **LinLUCB** : Our proposed method (Section 4)

Following [16], 0-1$_1$ and LinLUCB first perform an initialisation phase with $n_0 = 6$ samples with a round-bin sampling and Algorithm 1, respectively. As shown in Figure 4, LinLUCB significantly outperforms other policies in all configurations. With short environment durations (Cannot-sample-all-arms), shift estimation has more uncertainty, and consequently we observe a slower decaying

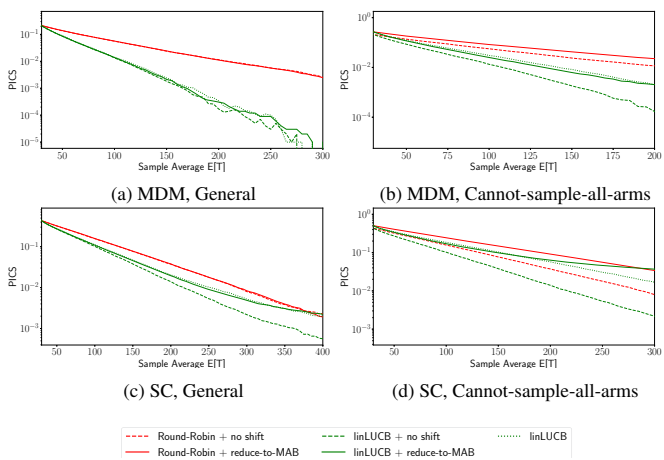
PICS compared to the General setting for both MDM and SC configurations. For the 0-1_1 policy, the General setting seems actually more difficult because an imbalance of samples per environment can strongly bias the sampling strategy and the selection policy. Sampling from several environments can reduce the dominating effect of a few environmental shifts, resulting in better PICS in quickly changing environments (compare Figure 4b with 4a and Figure 4d with 4c including the initial worsening in all cases). But even in the Cannot-sample-all-arms scenario, 0-1_1 performs worse than Round-Robin.



(a) MDM, General (b) MDM, Cannot-sample-all-arms

(c) SC, General (d) SC, Cannot-sample-all-arms

Figure 4: The performance of LinLUCB and benchmark policies



(a) MDM, General (b) MDM, Cannot-sample-all-arms

(c) SC, General (d) SC, Cannot-sample-all-arms

Figure 5: Comparison of Reduce-to-MAB strategies and the corresponding performances in a stationary environment

5.2 Reduce-to-MAB strategy

In order to gauge the benefit of shift estimation, we test an alternative approach by applying any existing policy designed for a stationary environment, and once a change occurs, we simply subtract the OLS shift estimators from the respective rewards ($r_{ij}^{new} = r_{ij} - \hat{s}_j$) to approximately reduce the problem to a standard MAB problem without shifts (*Reduce-to-MAB strategy*). One can suppose that all modified rewards are Gaussian with the expectation of $\mathbb{E}[r_{ij} - \hat{s}_j] = \mu_i$. All requirements for regression, however, are applied, and all required statistics in any such policies are replaced by statistics calculated from all subtracted rewards instead of original rewards. Note that the sample mean of such modified rewards is equivalent to the OLS mean estimator.

We implemented **Round-Robin** and **LinLUCB** with the **Reduce-to-MAB strategy** and compared these policies in our test scenarios and, for comparison, under idealised conditions without any environmental shifts. Note that for LinLUCB in a stationary environment, the corresponding UCB_i becomes $\bar{r}_i + \sqrt{16\ln(t)\tilde{\sigma}^2/(\sum_{j=1}^{J} n_{ij})^2}$ where $\tilde{\sigma}^2 = \sum_{i=1}^{K}\sum_{k=1}^{n_{i1}} \left(r_{i1}^k - \bar{r}_i\right)^2/(N-K)$ is an unbiased estimator for σ^2. Meanwhile, for LinLUCB with the Reduce-to-MAB strategy, the calculation of $\bar{r}_i$ and $\tilde{\sigma}^2$ for UCB_i is instead computed from rewards with the shift estimators subtracted. We also executed the proposed LinLUCB (Section 4) to investigate the benefit of including the uncertainty of the OLS estimator in the UCB computation. The relatively small gaps in Figure 5 between the policies and their respective performance in a stationary environment demonstrate the effectiveness of shift estimation. Not surprisingly, the gaps are smaller in long-duration environments (General) than in short-duration environments (Cannot-sample-all-arms). Round-robin sampling shows the smallest gap between its variants with and without the OLS estimator. This may be because the adaptive sampling strategy of LinLUCB is susceptible to estimation errors of the shifts, whereas round-robin sampling is not affected. Comparing the variants of LinLUCB, a small advantage of Reduce-to-MAB strategy can only be observed for a very small budget, as may be seen in Figure 5b and 5d. This phenomenon occurs because the value of the exploration term in UCB of the Reduce-to-MAB variant drops faster than of the proposed LinLUCB due to the denominator. Moreover, the variance estimator $\tilde{\sigma}^2$ underestimates its true value in a non-stationary environment and leads to less exploration of the Reduce-to-MAB one.

6 Conclusion and Discussion

In this paper, we formulate a new setting for fixed-budget best-arm identification in which an environment can globally shift the rewards of all arms in the same way. A selection policy based on ordinary linear regression is proposed to ensure an unbiased and consistent best-arm predictor where the number of environments keeps increasing. We also propose LinLUCB, an algorithm which integrates an error from the mean and shift estimator into the sample allocation decision, constructing a confidence bound that naturally arises from the covariance matrix of the OLS estimator. Empirically, the LinLUCB algorithm is effective in dealing with piecewise stationary environments with global shifts. Besides, our numerical experiments demonstrate the benefits of exploiting the OLS estimator and how the distribution of the duration of stationary periods of the environment affects the performance of policies. Simply using the shift estimates produced by our OLS estimator to reduce the problem to a standard MAB setting works well if the length of environments is sufficiently long. Still, LinLUCB works at least as good and mostly better in all tested cases.

The paper opens several interesting avenues for future work. For instance, the strong assumption of global environmental influence may be relaxed. Also, in real-world contexts, environmental changes are often continuous with smooth transitions rather than piecewise stationary, making it significantly harder to detect the changing point. Lastly, the extension to heterogeneous noise for different arms and different environments will be useful for a more general study.

Acknowledgements

Phurinut would like to acknowledge the Royal Thai Government scholarship sponsored by The Institute for the Promotion of Teaching Science and Technology.

References

[1] Y. Abbasi-Yadkori, P. Bartlett, V. Gabillon, A. Malek, and M. Valko. Best of both worlds: Stochastic & adversarial best-arm identification. In *Conference on learning theory*, pages 918–949. PMLR, 2018.

[2] A. Alieva, A. Cutkosky, and A. Das. Robust pure exploration in linear bandits with limited budget. In *International Conference on Machine Learning*, pages 187–195. PMLR, 2021.

[3] R. Allesiardo and R. Féraud. Exp3 with drift detection for the switching bandit problem. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–7. IEEE, 2015.

[4] R. Allesiardo, R. Féraud, and O.-A. Maillard. The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, 3:267–283, 2017.

[5] J. Altschuler, V.-E. Brunel, and A. Malek. Best arm identification for contaminated bandits. *Journal of Machine Learning Research*, 20(91): 1–39, 2019.

[6] S. Amaran, N. V. Sahinidis, B. Sharda, and S. J. Bury. Simulation optimization: a review of algorithms and applications. *Annals of Operations Research*, 240:351–380, 2016.

[7] J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53, 2010.

[8] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

[9] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32 (1):48–77, 2002.

[10] M. Azizi, B. Kveton, and M. Ghavamzadeh. Fixed-budget best-arm identification in structured bandits. In L. D. Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2798–2804. International Joint Conferences on Artificial Intelligence Organization, 2022.

[11] L. Besson, E. Kaufmann, O.-A. Maillard, and J. Seznec. Efficient change-point detection for tackling piecewise-stationary bandits. *The Journal of Machine Learning Research*, 23(1):3337–3376, 2022.

[12] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*, pages 23–37. Springer, 2009.

[13] E. Cavenaghi, G. Sottocornola, F. Stella, and M. Zanker. Non stationary multi-armed bandit: Empirical evaluation of a new concept drift-aware algorithm. *Entropy*, 23(3):380, 2021.

[14] C.-H. Chen, J. Lin, E. Yücesan, and S. E. Chick. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10:251–270, 2000.

[15] S. E. Chick and K. Inoue. New two-stage and sequential procedures for selecting the best simulated system. *Operations Research*, 49(5): 732–743, 2001.

[16] S. E. Chick, J. Branke, and C. Schmidt. Sequential sampling to myopically maximize the expected value of information. *INFORMS Journal on Computing*, 22(1):71–80, 2010.

[17] N. Etemadi. Stability of sums of weighted nonnegative random variables. *Journal of Multivariate Analysis*, 13(2):361–365, 1983.

[18] T. Fiez, H. Nassif, Y.-C. Chen, S. Gamez, and L. Jain. Best of three worlds: Adaptive experimentation for digital marketing in practice. In *Proceedings of the ACM on Web Conference 2024*, pages 3586–3597, 2024.

[19] P. I. Frazier, W. B. Powell, and S. Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

[20] A. Garivier and E. Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer, 2011.

[21] C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag. Multi-armed bandit, dynamic environments and meta-bandits. 2006.

[22] M. Hoffman, B. Shahriari, and N. Freitas. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Artificial Intelligence and Statistics*, pages 365–374. PMLR, 2014.

[23] L. J. Hong, W. Fan, and J. Luo. Review on ranking and selection: A new perspective. *Frontiers of Engineering Management*, 8(3):321–343, 2021.

[24] K. Jamieson and R. Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2014.

[25] K. Jamieson and A. Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial intelligence and statistics*, pages 240–248. PMLR, 2016.

[26] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

[27] K.-S. Jun, L. Li, Y. Ma, and J. Zhu. Adversarial attacks on stochastic bandits. *Advances in neural information processing systems*, 31, 2018.

[28] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

[29] Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *International conference on machine learning*, pages 1238–1246. PMLR, 2013.

[30] D. E. Koulouriotis and A. Xanthopoulos. Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems. *Applied Mathematics and Computation*, 196(2):913–922, 2008.

[31] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[32] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[33] T. Lykouris, V. Mirrokni, and R. Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

[34] C. Shen. Universal best arm identification. *IEEE Transactions on Signal Processing*, 67(17):4464–4478, 2019.

[35] H. Shen, L. J. Hong, and X. Zhang. Ranking and selection with covariates for personalized decision making. *INFORMS Journal on Computing*, 33(4):1500–1519, 2021.

[36] P. Srisawad, J. Branke, and L. Tran-Thanh. Identifying the best arm in the presence of global environment shifts. *arXiv preprint arXiv:2408.12581*, 2024. Full version of this paper.

[37] Z. Xiong, R. Camilleri, M. Fazel, L. Jain, and K. Jamieson. A/b testing and best-arm identification for linear bandits with robustness to non-stationarity. In *International Conference on Artificial Intelligence and Statistics*, pages 1585–1593. PMLR, 2024.

[38] M. Xu and D. Klabjan. Regret bounds and reinforcement learning exploration of exp-based algorithms. *arXiv preprint arXiv:2009.09538*, 2020.

[39] J. Yang and V. Tan. Minimax optimal fixed-budget best arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 35:12253–12266, 2022.

[40] Z. Zhong, W. C. Cheung, and V. Tan. Probabilistic sequential shrinking: A best arm identification algorithm for stochastic bandits with corruptions. In *International Conference on Machine Learning*, pages 12772–12781. PMLR, 2021.