Bandits with Concave Aggregated Reward

Yingqi Yu[1] , **Sijia Zhang**[1] , **Shaoang Li**[1] , **Lan Zhang**[1,2] , **Wei Xie**[1] , **Xiang-Yang Li**[1]
[1]University of Science and Technology of China, Hefei, China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{yu971207, sxzsj, lishaoa}@mail.ustc.edu.cn, {zhanglan, xxieww, xiangyangli}@ustc.edu.cn

Abstract

Multi-armed bandit is a simple but powerful algorithmic framework, and many effective algorithms have been proposed for various online models. In numerous applications, the decision-maker faces diminishing marginal utility. With non-linear aggregations, those algorithms often have poor regret bounds. Motivated by this, we study a bandit problem with diminishing marginal utility, which we termed the bandits with concave aggregated reward(BCAR). To tackle this problem, we propose two algorithms SW-BCAR and SWUCB-BCAR. Through theoretical analysis, we establish the effectiveness of these algorithms in addressing the BCAR issue. Extensive simulations demonstrate that our algorithms achieve better results than the most advanced bandit algorithms.

1 Introduction

Multi-Armed Bandit (MAB) is one of the most classical frameworks for making decisions sequentially under uncertainty. In this setting, an agent chooses an action and receives a reward in each round. The agent aims to maximize her/his cumulative reward during the game by capturing the exploration-exploitation tradeoff. Based on the basic bandits model, many studies have proposed a number of variants, and a large number of algorithms have emerged to tackle related problems [Gittins, 1979; Agrawal *et al.*, 1988; Auer *et al.*, 2002a; Maillard *et al.*, 2011; Li *et al.*, 2023]. These bandit models have been applied across various domains, such as advertisements [Li *et al.*, 2010; Li *et al.*, 2011], dynamic procurement [Badanidiyuru *et al.*, 2012; Badanidiyuru *et al.*, 2013] and crowdsourcing platforms [Slivkins and Vaughan, 2014].

The classical bandit problems often assume that the pulls of arms are independent and the cumulative reward of the pulled arms is a linear function of all actions over the time horizons T. However, due to the diminishing marginal utility, in many application domains, the agent collects a *concave aggregated* reward: 1) the marginal utility of each arm gradually decreases, 2) the value generated by a selected arm in each round will impact the subsequent rounds' rewards. For example, an advertiser wants to promote a new product. There are several different advertisements to choose from, and different advertisements have different effects on the product's popularity. The advertiser has to decide which advertisement to use on each day. In this scenario, each arm corresponds to an advertisement, and the total reward corresponds to the product's popularity or the number of registered users. Since the decay of Click-Through-Rate(CTR) is largely due to the amount of repeat exposure[Agarwal *et al.*, 2009], the longer the advertising time, the more people learn about the product, and the diminishing marginal utility leads to a decline in the growth rate of the product's popularity. The advertiser aims to the expose product to more people, i.e., to maximize the aggregated reward over the total T rounds. In this problem, due to diminishing marginal rewards, even if the increase in product popularity on the first day is more significant than that on the second day, it cannot be concluded that the advertisement quality of the first day is higher than that of the second day. The quality of the advertisement is a hidden "value" and the increase in product popularity is an observable "reward". The agent's goal is to select options with higher "values" based on observable "rewards" to maximize his total reward. Besides advertising, the establishment of datasets for large-scale models and the employment of personnel also exhibit the property of diminishing marginal utility.

In this work, we consider the situation where an agent faces a concave aggregated reward function. We define it as the bandits with concave aggregated reward(BCAR) problem. In this model, the reward of pulling any arm decays over time. In our setting, there are agent-unknown random *values* $v(t)$ and real marginal *rewards* $r(t)$. The agent-unknown random values $v(t)$ are similar to the rewards in the classic MAB model and sampled in the range $(0, 1]$. These two values are correlated by an increasing and concave function $f(\cdot)$, i.e., $r(t) = f(\sum_{j\le t} v(j)) - f(\sum_{j\le t-1} v(j))$. The objective of the agent is to maximize his expected total reward $\mathbb{E}\left[f\left(\sum_{j=1}^{T} v_{a_j}(j)\right)\right]$ in a time horizon T by carefully pulling arms in each round.

Challenges. (1) Potential sublinear reward: The cumulative reward may not increase linearly with T, so the algorithm with a sublinear regret upper bound may not achieve a reward of $(1 - o(1))$OPT. (2) Indirect connection between $r(t)$ and $v(t)$: The thorniest one stems from the particular aggregated reward function, namely that the rewards $r(t)$ obtained by

sampling cannot reflect the value $v(t)$ of the arms. (3) Non-independence among selections: The agent's past selections will impact the rewards of future selected arms due to the implicit dependence between arms. As the total reward is not a simple sum of all arm's values in each round, the selection of arms in different rounds is no longer independent.

Contributions. (1) To the best of our knowledge, this is the first work to investigate the bandits with concave aggregated reward. This model formulates the bandit problem associated with diminishing marginal utility. Our algorithms achieve rewards of $(1 - o(1))$OPT, which existing approaches cannot attain. (2) To tackle this problem, we propose a sliding window based algorithm SW-BCAR with provable regret bound. We prove that the algorithm SW-BCAR attains a regret as $\tilde{O}(K^{1/3}T^{-1/3})$OPT, where K, T and OPT are the number of arms, the number of rounds and the best-arm benchmark. (3) We also study an extended case of the BCAR, where the agent knows that the maximum of the arms' mean values μ^* is in the range $[\frac{1}{\sigma}, 1]$ and $\sigma > 1$. We call this case the parameter case. We propose a UCB based algorithm SWUCB-BCAR with a provable regret guarantee $\tilde{O}(K^{1/2}T^{-1/2})$OPT. (4) We validate the performance of our algorithms via numerical simulations.

2 Problem Formulation and Preliminaries

We formally define the bandits with concave aggregated reward(BCAR) as follows. There is an unknown monotonically increasing concave reward function $f : \mathbb{R}^+ \to \mathbb{R}^+$. Let $\mathcal{K}$ be the set of alternative arms and $\mathcal{K} = \{1, 2, ..., K\}$. Each arm a is associated with a value distribution $\mathcal{D}_a$. Let $\mu(a) = \mathbb{E}_{X\sim\mathcal{D}_a}[X]$. All the value distributions are supported in $(0, 1]$. The game is played T rounds. When the agent chooses one of the arms, the arm will generate a value sampled from its distribution. Formally, the problem protocol at each round $t = 1, ..., T$ is as follows:

• The agent chooses one arm $a_t \in \mathcal{K}$ and the arm generates a *value* $v_t = v_{a_t}(t)$ which is unknown to the agent.

• The agent receives a *reward* $r_t = r_{a_t}(t) = f\left(\sum_{j=1}^{t} v_{a_j}(j)\right) - f\left(\sum_{j=1}^{t-1} v_{a_j}(j)\right)$ for this selection.

The objective of the agent is to maximize his expected total reward $\mathbb{E}\left[f\left(\sum_{j=1}^{T} v_{a_j}(j)\right)\right]$. Figure 1 illustrates the model of BCAR.

Regret. An algorithm is optimal if it maximizes the expected total reward. We define the optimal expected total reward as $\text{OPT} = \max_{a_1,...,a_T} \mathbb{E}\left[f\left(\sum_{j=1}^{T} v_{a_j}(j)\right)\right]$.

The suboptimal algorithm A are evaluated via the expected total regret: $\mathbb{E}[\mathcal{R}_A(T)] = \text{OPT} - \mathbb{E}_A\left[f\left(\sum_{j=1}^{T} v_{a_j}(j)\right)\right]$.

Since the reward function may not increase linearly with T in this problem, if an algorithm B attains a sublinear regret $\mathbb{E}[\mathcal{R}_B(T)]$, $\mathbb{E}[\mathcal{R}_B(T)]$ may be $\Theta(f(T))$ and the agent may attain a reward of $\Theta(1)$OPT, which signifies a considerably undesirable outcome. For example, $f(x) = \log x$ and $\mathbb{E}[\mathcal{R}_B(T)] = O(\sqrt{T})$. To eliminate the effects of the concave aggregated reward function, we consider the ratio of the algorithm's regret and the best-arm benchmark OPT, i.e., $\mathbb{E}[\mathcal{R}_A(T)]/\text{OPT}$, to measure the algorithm performance. For example, if $f(x) = \log x$ and $\mathbb{E}[\mathcal{R}_A(T)]/\text{OPT} = O(\sqrt{T}/T)$, the algorithm A attains a regret of $O(T^{-1/2}\log T)$. And the algorithm A is better than an algorithm B that attains a regret of $\mathbb{E}[\mathcal{R}_B(T)] = O(\sqrt{T})$.

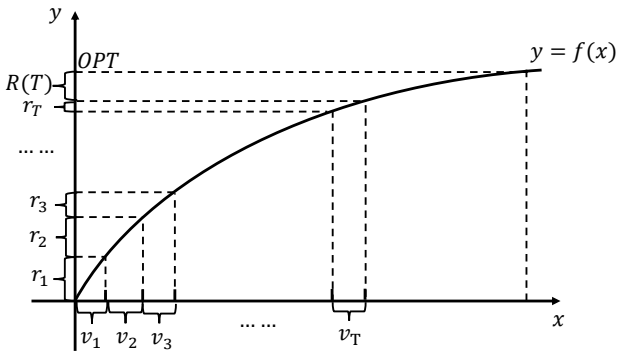


Figure 1: Graphical representation of the BCAR model.

Notation	Description
$f(\cdot)$	the unknown reward function.
a_t	the arm chosen by at time t, $a_t \in \{1, ..., K\}$.
$\mu(a)$	the expectation of arm a's value distribution.
v_t	the value generated by the arm a_t at time t.
r_t	the reward generated by the arm a_t at time t.
a^*	the arm with the largest mean value.
f'_{end}	$f'_{end} = f'(\sum_{j=1}^{T} v_{a_j}(j))$.
$\mathcal{R}_A(T)$	the regret of the algorithm A.
OPT	the best arm benchmark.

Table 1: Description of commonly-used notations.

In [Bubeck and Cesa-Bianchi, 2012], Bubeck *et al.* proved that in the Stochastic MAB problem, fix round number T and arm number K, for any bandit algorithm, there exists a problem instance such that $\mathbb{E}[\mathcal{R}(T)] \geq \Omega(\sqrt{KT})$ and $\frac{\mathbb{E}[\mathcal{R}(T)]}{\text{OPT}} = \frac{\mathbb{E}[\mathcal{R}(T)]}{O(T)} \geq \Omega\left(K^{1/2}T^{-1/2}\right)$. Because the stochastic MAB is equivalent to the case where $f(x) = x$ in the BCAR problem, it is a sub-problem of our model. The algorithm designed for the BCAR problem should also apply to the stochastic MAB. We can prove the following theorem:

Theorem 1. *In the BCAR problem, fix round number T and arm number K, for any bandit algorithm, there exists a problem instance such that*

$$\mathbb{E}[\mathcal{R}(T)]/\text{OPT} \geq \Omega\left(K^{1/2}T^{-1/2}\right). \tag{1}$$

We summarize key notations used throughout this paper in Table 1.

3 Fundamental Case

The model is a variant of the classical stochastic multi-armed bandit problem [Lai *et al.*, 1985], motivated by diminishing rewards. In this problem, the learner cannot observe the *value* sampled from the selected arm's distribution. However, the observed diminishing *reward* is a concave function of the value of the action currently chosen and the values generated by the past actions. The learner's goal is to design a policy to find the optimal arm quickly from observed rewards to maximize the aggregated reward or minimize the regret.

In this problem, even if the aggregated reward function is a Lipschitz function, i.e., $\exists L \in \mathbb{R}$ s.t. $|f(x+y)-f(x)| \leq L \cdot |y|$, $\forall x, y \in \mathbb{R}$, existing algorithms, for example UCB1[Auer *et al.*, 2002a], can only attain a regret as $\tilde{O}\left(\sqrt{KTL}\right)$. However, the reward function does not correlate linearly with the round number T. So, if an algorithm attains the regret sublinear in T, the algorithm may achieve a regret of $\Theta(1)$OPT, which signifies a considerably undesirable outcome.

In order to address the BCAR problem, we propose the SW-BCAR algorithm. Our algorithm aims to detect arms with small mean values through a period of exploration, after which these "bad" arms will be abandoned and excluded from further selection. The arms that have not been abandoned are referred to as reserved arms. For further convenience, we introduce the following definitions:

Definition 1. *Let $\mathcal{K}_{end}$ be the set of arms reserved at the end and a^*_{end} be the arm with the maximum mean value in $\mathcal{K}_{end}$. Let $\mu^*_{end} = \mu(a^*_{end})$. We define bad arms to be the arms whose mean values are less than μ^*_{end} and good arms to be the arms whose mean values are not less than μ^*_{end}.*

Definition 2. *Let weight w_t in round t be the ratio between the reward obtained in round t and the value generated by the chosen arm in round t, i.e. $w_t = \frac{r_{a_t}(t)}{v_{a_t}(t)}$.*

Key idea. Initially, we use a round-robin approach to select arms from the arm set, and a sliding window approach to reduce the impact of the concave aggregated reward function on our assessment of arm mean values. When an arm is selected, the reward is determined by both the value generated by the chosen arm and the weight determined by the aggregated function. As the value of the aggregated function increases, the slope may gradually decrease, indicating that the weight may decrease as the selection process progresses. If the weights of an arm at different rounds vary significantly, our evaluation of the arm's mean value may be imprecise. To address this, we employ a sliding window-based approach to ignore the rewards in the early rounds, thereby reducing the impact of weight differences on arm assessment. We assume that there are two arms a and a', where $\mu(a) > \mu(a')$. If the agent is unable to determine which arm has a lower mean value, he can deduce that the lower bound of the change rate of weights is linearly related to $\mu(a) - \mu(a')$. In this way, the agent can ensure that the regret is bounded even when more rounds of exploration are undertaken.

To minimize the regret upper bound according to its expression, we choose an appropriate length of the sliding windows denoted as m. We can prove that when $m = \lceil (T/K)^{2/3} \log^{1/3} T \rceil$, we can minimize the regret upper bound.

Algorithm 1: SW-BCAR

Input: K, T

1 Initialize the set of arms $\mathcal{K}$; $m \leftarrow \lceil (T/K)^{2/3} \log^{1/3} T \rceil$;
2 Build a circular linked list B on A, a vector $C \in \mathbb{N}^{K \times K}$, count vectors $D, E, N \in \mathbb{N}^K$;
3 Initialize all items in C and D, E, N to 0, $a_1 = 1$;
4 **for** *round* $t = 1, 2, ..., Km$ **do**
5 pick arm a_t in B, receive r_t; $N_{a_t} \leftarrow N_{a_t} + 1$; $r^{N_{a_t}}_{a_t} \leftarrow r_t$; $a_{t+1} \leftarrow next(a_t)$ in B;
6 **end**
7 **for** *arm* $j \in [1, K]$ **do**
8 $\overline{r}_j \leftarrow \frac{1}{m} \sum_{n=N_j-m+1}^{N_j} r^n_j$;
9 **end**
10 **for** *round* $t = Km+1, Km+2, ..., T$ **do**
11 pick arm a_t in B, receive r_t; $N_{a_t} \leftarrow N_{a_t} + 1$; $r^{N_{a_t}}_{a_t} \leftarrow r_t$; $\overline{r}_{a_t} \leftarrow \frac{1}{m} \sum_{n=N_{a_t}-m+1}^{N_{a_t}} r^n_{a_t}$;
12 **for** *arm* $j \in B$ **do**
13 **if** $C_{a_t,j} \neq m$ *and* $\overline{r}_{a_t} > \overline{r}_j$ **then**
14 $C_{a_t,j} \leftarrow C_{a_t,j} + 1$;
15 **if** $C_{a_t,j} = m$ **then**
16 $D_{a_t} \leftarrow D_{a_t} + 1$; $E_j \leftarrow E_j + 1$; Remove-Arms(D,E,B);
17 **end**
18 **end**
19 **if** $C_{a_t,j} \neq m$ *and* $\overline{r}_{a_t} \leq \overline{r}_j$ **then**
20 $C_{a_t,j} \leftarrow 0$;
21 **end**
22 **end**
23 $a_{t+1} \leftarrow next(a_t)$ in B;
24 **end**

Arm a and a' are two arms in the arm set $\mathcal{K}$. For all $i \in \{1, ..., T\}$, we define that a' is *i-better* than a if the following requirements are met: 1) for all $j \in \{1, ..., i\}$, we denote the weight multiplied by the value of arm a in it's j-th latest selection as $\hat{w}_j$. We denote the weight multiplied by the value of arm a' in it's j-th latest selection as $\hat{w}'_j$. For all $j \in \{1, ..., i\}$, $\hat{w}_j \geq \hat{w}'_j$; 2) when the event in 1) occurs, we compare the mean reward of the latest i selections of arm a and that of arm a'. We call these comparisons as i-mean value comparisons. For these comparisons, there are i consecutive comparisons satisfy that the mean reward of the latest i selections of arm a is lower than that of arm a'. For arms a and a', we define that a is *i-worse* than a' if a' is *i-better* than a. When an arm a is m-better than half of the other reserved arms, the agent will abandon all the arms that are m-worse than a; when an arm is m-worse than more than half of the other reserved arms, the agent will abandon the arm a.

The details of SW-BCAR are shown in Algorithm 1. It is composed of three phases:

- **Initial phase** (line 1-3): Let the parameter $m =$

Algorithm 2: Remove-Arms

Input: D, E, B

$flag \leftarrow 1$;
while $flag = 1$ **do**
  $flag \leftarrow 0$;
  for *arm* $j, a \in B$ **do**
    if $D_j \geq |B|/2$ *and* $C_{j,a} = m$ **then**
      remove a from circular linked list B;
      Update D, E; $flag \leftarrow 1$;
    end
    if $E_j \geq |B|/2$ **then**
      remove j from circular linked list B;
      Update D, E; $flag \leftarrow 1$;
    end
  end
end

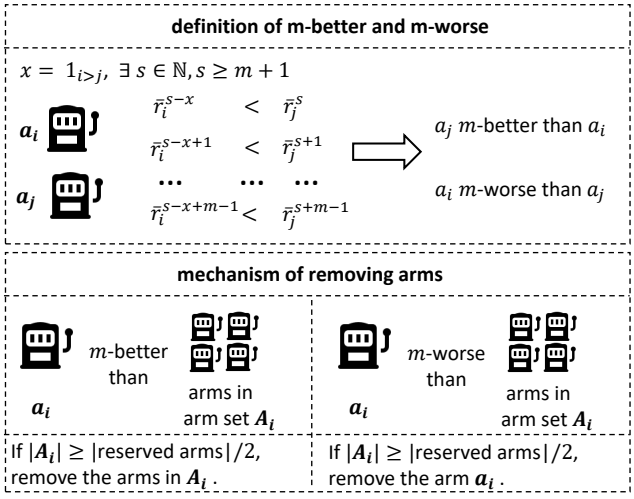


Figure 2: An illustration of SW-BCAR.

$\lceil (T/K)^{2/3} \log^{1/3} T \rceil$. The agent maintains a count vector $C \in \mathbb{N}^{K \times K}$, where $C_{a,a'}$ is used to record the number of consecutive comparisons satisfy that the mean reward of the latest m selections of arm a is larger than that of arm a' for the m-mean value comparison. When $C_{a,a'} = m$, it will no longer change. In addition to this, the agent maintains three other vectors $D, E, N \in \mathbb{N}^K$. D_a, where $a \in \mathcal{K}$, is the number of the reserved arms which is m-worse than arm a. E_a, where $a \in \mathcal{K}$, is the number of the reserved arms which is m-better than arm a. N_a, where $a \in \mathcal{K}$ is the number of arm a has been pulled. The agent initializes all items in all vectors to 0. B is a circular linked list of reserved arms.

• **Exploration phase** (line 4-9): The agent uses round robin to select each arm m times and calculates the mean reward for each arm in this phase.

• **Sliding window phase** (line 10-24): The agent puts all the arms that have not been abandoned yet into the reserved arm set and then uses round robin to select arms from this set in future. When the agent selects an arm a_t, he will calculate the mean reward of the arm a_t's latest m selections and then compare it with the other reserved arms' mean rewards of their latest m selections and update $C_{a_t,a'}$, D_{a_t} and $E_{a'}$ for all a' in the reserved arms, where $a' \neq a_t$. When an arm a is m-better than half of the other reserved arms, the agent will abandon all the arms that are m-worse than a; when an arm is m-worse than more than half of the other reserved arms, the agent will abandon the arm a. By removing the arms in this way, we can have a good performance guarantee for μ^*_{end} and limit regret due to abandoning the largest mean value among the reserved arms.

In the rest of this section, we will prove the regret bound of Algorithm 1. Let $v_a(t)$ and $r_a(t)$ be the value and reward generated by the arm $a \in \mathcal{K}$ when it is selected for the t-th time. Let $w_a(t) = r_a(t)/v_a(t)$ and $w_t = w_{a_t}(t)$. We can know that $\forall t_1 < t_2, t_1, t_2 \in \{1, \cdots, T\}, w_{t_1} \geq w_{t_2}$.

Let $\overline{r}_a(t)$, where $t \geq m$, be the mean reward generated by the latest m choices of the arm a (from its $(t-m+1)$-th choice to its t-th choice), i.e. $\overline{r}_a(t) = \frac{1}{m}\sum_{i=t-m+1}^{t} r_a(i)$. Let $\overline{v}_a(t)$, where $t \geq m$, be the mean value generated by the latest m choices of the arm a, i.e. $\overline{v}_a(t) = \frac{1}{m}\sum_{i=t-m+1}^{t} v_a(i)$. We then estimate the deviation of the average from the true expectation. By defining the confidence radius $\hat{\delta}_t(a) = 2w_a(t)\sqrt{\frac{2\log T}{m}}$, we have $\forall i \in \{t-m+1, ..., t\}, r_a(i) \leq w_a(i) \leq w_a(t-m+1)$, i.e. $r_a(i) \in (0, w_a(t-m+1)]$. From Azuma-Hoeffding's inequality, we can show that

$$\mathbb{P}\left[\left|\overline{r}_a^{N_a} - \mathbb{E}\left[\overline{r}_a^{N_a}\right]\right| \leq \hat{\delta}_{N_a}(a)\right] \geq 1 - \frac{2}{T^4}. \tag{2}$$

Similarly, let $\delta' = \sqrt{\frac{2\log T}{m}}$, from Hoeffding's inequality, we have

$$\mathbb{P}\left[\left|\overline{v}_a^{N_a} - \mathbb{E}\left[\overline{v}_a^{N_a}\right]\right| \leq \delta'\right] \geq 1 - \frac{2}{T^4}. \tag{3}$$

Definition 3. *We define the good event to be the event that* (2) *and* (3) *hold for all arms and all rounds simultaneously. Let "bad event" be the complement of the good event.*

We will analyze the regret for the good event and the bad event separately. We note that if a lousy arm is preserved, the slope of the reward function will change fast. Then, we have the following result:

Theorem 2. *There are K alternative arms and T rounds. For the bandits with concave aggregated reward, SW-BCAR algorithm achieves regret for the BCAR bounded by*

$$\frac{\mathbb{E}\left[\mathcal{R}_{SW-BCAR}(T)\right]}{OPT} = O\left(\frac{(K\log T)^{1/3}\log K}{T^{1/3}}\right). \tag{4}$$

Proof Sketch. From the properties of concave functions, we can infer that $f(\mu^* T) \geq$ OPT. We first analyze the difference between the total revenue and $f(\mu^* T)$, and then convert the result to a bound on the regret.

Although the proof for Theorem 2 is complicated, the key is to divide the regrets generated by SW-BCAR into three categories and analyze them separately: 1) In good events, the regrets generated by the algorithm as the agent selects the bad arms, denoted as $\hat{\mathcal{R}}_1$; 2) In good events, the regrets generated by the algorithm as the agent removes the good arms,

denoted as $\hat{\mathcal{R}}_2$; 3) In bad events, the regrets generated by the algorithm, denoted as $\hat{\mathcal{R}}_3$.

First, we give the upper bound of $\mathbb{E}[\hat{\mathcal{R}}_1]$. We prove that if a lousy arm is not abandoned on time, $\frac{\mathbb{E}[r_{a_t}(t)]}{\mu(a_t)}$ changes fast as selections progress. Then we prove that the regret bound for BCAR generated by the "bad arms", $\mathbb{E}\left[\hat{\mathcal{R}}_1\right] = O\left(T^{-1/3}(K\log T)^{1/3}\log K\right) f(\mu^* T)$. Second, we give the upper bound of $\mathbb{E}[\hat{\mathcal{R}}_2]$. We prove that in good events $\mu^* - \mu^*_{end} < O(\delta' \log K)$. Then, we prove that the regret bound for BCAR generated by the "good arms", $\mathbb{E}\left[\hat{\mathcal{R}}_2\right] = O\left(T^{-1/3}(K\log T)^{1/3}\log K\right) f(\mu^* T)$. Then, we give the upper bound of $\mathbb{E}[\hat{\mathcal{R}}_3]$. Since the probability of bad events happening is $O(\frac{1}{T^2})$, we can prove that the regret bound for BCAR generated by the "bad events", $\mathbb{E}\left[\hat{\mathcal{R}}_3\right] = O\left(\frac{1}{T}\right)$.

By summing up the above results and converting them to an analysis of the regret, we are able to complete the proof. □

4 Parametric Case

In the previous section, we discuss the fundamental case. Since both the reward function and the value distribution of each arm are unknown, it is difficult for us to analyze the slope range of the function and the actual distribution of arms. In many applications, although we do not know the mean value of the optimal arm, we know that it has a constant lower bound.

In this section, we consider the parametric case. In the BCAR model, assume that there is a parameter $\sigma > 1$. We define the optimal arm a^* as the arm with the largest mean value and $\mu^* = \mu(a^*)$, where μ^* is in the range of $[\frac{1}{\sigma}, 1]$.

We propose a variant of Algorithm 1 named SWUCB-BCAR to solve the problem in this situation. This algorithm also uses the upper/lower confidence bound method and attains a regret as $\tilde{O}(K^{1/2}T^{1/2})\frac{f(\mu^* T)}{\mu^* T}$. Since $\mathbb{E}[\hat{\mathcal{R}}(T)] \geq \Omega(\sqrt{KT})$, Algorithm 3 is asymptotically optimal. The pseudo algorithm for this method is given by Algorithm 3.

Similar to the proof of Theorem 2, we can prove the following theorem:

Theorem 3. *For the bandits with concave aggregated reward in the parametric case, SWUCB-BCAR algorithm achieves regret for the BCAR bounded by*

$$\frac{\mathbb{E}\left[\mathcal{R}_{SWUCB-BCAR}(T)\right]}{OPT} = O\left(\frac{K^{1/2}\sigma\log T}{T^{1/2}}\right). \quad (5)$$

Proof Sketch. For each arm $a \in \mathcal{K}$, let $\Delta(a) = \frac{\mu^* - \mu(a)}{\mu^*}$ be the difference between the mean value of the optimal arm a^* and the mean value of the arm $a (a \neq a^*)$. Let $m_a = \lceil \frac{8\log T\sigma^2}{\Delta(a)^2} \rceil$. For each arm $a \in \mathcal{K}$, let $\mathcal{R}^a(T)$ be the regret generated by the difference between the arm a and a^*. Let f_a be the reward generated by arm a. Similar to the proof of Theorem 2, we have $\mathbb{E}[\mathcal{R}^a(T)] \leq O\left(\frac{\sigma^2\mu^*\log^2 T}{\Delta(a)}\right)\frac{f(\mu^* T)}{\mu^* T}$. Let us fix some $\epsilon > 0$; then regret consists of two parts: 1) All arms a with $\Delta(a) \leq \epsilon$ contributes at most a total of $\epsilon\mu^* T f'_{end}$; 2) each arm a with $\Delta(a) > \epsilon$ contributes at most a total of $O\left(\frac{\mu^* K\sigma^2\log^2 T}{\epsilon}\right) \cdot \frac{f(\mu^* T)}{\mu^* T}$. Combining these two parts, in good events we have $\mathbb{E}[\mathcal{R}(T)] \leq O(\epsilon\mu^* T + \frac{\mu^* K\sigma^2\log^2 T}{\epsilon})\frac{f(\mu^* T)}{\mu^* T}$. Let $\epsilon = \sqrt{TK}\sigma\log T$. By summing up the above results, we can complete the proof. □

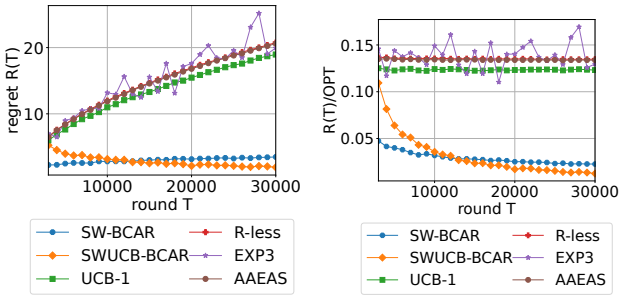
Algorithm 3: SWUCB-BCAR

Input: K, T
Initialize the set of arms $\mathcal{K}$;
Build a circular linked list B on A, a count vector $C \in \mathbb{N}^{K\times K\times T}$, a count vector $N \in \mathbb{N}^K$; Initialize all items in C and N to 0, $a_1 = 1, l_{min} = 1$;
for *round* $l = 1, 2, ..., T$ **do**
  $\delta'_l \leftarrow \sqrt{\frac{2\log T}{l}}\sigma$;
  if $\delta'_l \geq 1$ **then**
    $l_{min} = l + 1$;
  end
end
for *round* $t = 1, 2, ..., T$ **do**
  pick arm a_t in B, receive r_t;
  $N_{a_t} \leftarrow N_{a_t} + 1$; $r_{a_t}^{N_{a_t}} \leftarrow r_t$;
  for $l = l_{min}, l_{min} + 1, ..., N_{a_t}$ **do**
    $\overline{r}_{a_t,l} \leftarrow \frac{1}{l}\sum_{n=N_{a_t}-l+1}^{N_{a_t}} r_{a_t}^n$;
    for *arm* $j \in B$ **do**
      if $N_j \geq l$ **then**
        if $\overline{r}_{a_t,l} > \overline{r}_{j,l} \cdot \frac{1+\delta'_l}{1-\delta'_l}$ **then**
          $C_{a_t,j,l} \leftarrow C_{a_t,j,l} + 1$;
          if $C_{a_t,j,l} = l$ **then**
            remove j from circular linked list B;
          end
        end
        if $\overline{r}_{a_t,l} \leq \overline{r}_{j,l} \cdot \frac{1+\delta'_l}{1-\delta'_l}$ **then**
          $C_{a_t,j,l} \leftarrow 0$;
        end
      end
    end
  end
  $a_{t+1} \leftarrow next(a_t)$ in B;
end

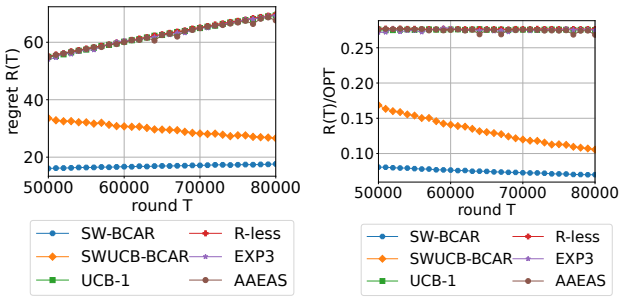
5 Evaluation By Simulation

In this section, we evaluate our algorithms' effectiveness in terms of regrets. We compare the performance of SW-BCAR and SWUCB-BCAR with benchmark algorithms.

We use UCB1 [Auer *et al.*, 2002a] and exp3 [Auer *et al.*, 2002b] for comparison. Both of them are the most classic bandits algorithms. Besides, we use Rless [Metelli *et al.*, 2022] and AAEAS [Lykouris *et al.*, 2020] for comparison. Both of them are the algorithms for the non-stationary bandits. The algorithm AAEAS is used to tackle the bandits with adversarial scaling. The algorithm Rless is used

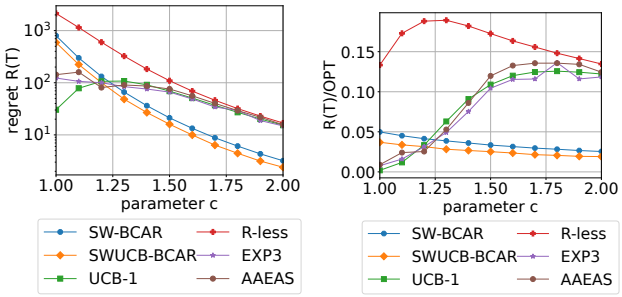


(a) $R(T) : K = 2$ (b) $R(T)/\text{OPT} : K = 2$



(c) $R(T) : K = 20$ (d) $R(T)/\text{OPT} : K = 20$

Figure 3: Regret with different round numbers.



(a) $R(T) : f_{3,c}(x)$ (b) $R(T)/\text{OPT} : f_{3,c}(x)$

Figure 4: Regret with different functions.

to tackle the stochastic rising bandits. These bandit models are similar to the BCAR model. All of the four algorithms require a range of rewards in each round. Although our algorithms do not require such a setting, to facilitate the comparison, we use the aggregated reward functions satisfying the Lipschtiz assumption. Specifically, we use the following reward functions with different parameters: (1)$f_{1,c}(x) = c - c \cdot e^{(-x/c)}, c > 0$; (2)$f_{2,c}(x) = [\log(1 + cx)]/c, c > 0$; (3)$f_{3,c}(x) = (1 + x)^{1/c} - 1, c \geq 1$. Here we mainly consider the truncated normal distributions and the reward function $f_{3,c}(\cdot)$ to evaluate the algorithms' performance.

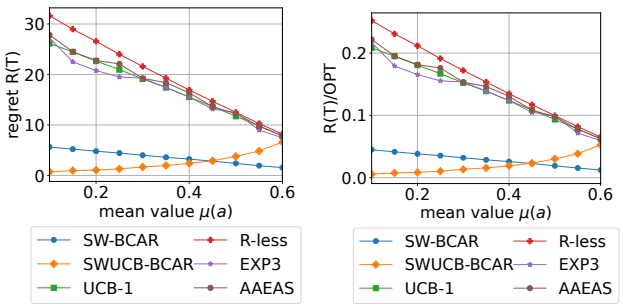
We set the variance of all truncated normal distributions in the experiment to 0.2, and designed several different sets of experiments to study the effect of variables on the algorithms. All results are the averages over 20 runs. In the experiments, variables other than specified separately were fixed as follows: 1) the round number $T = 20000$; the arm number $K = 2$; 2) the optimal arm's mean value $\mu^* = 0.8$; the suboptimal arms' mean values $\mu(a) = 0.4$; 3) the aggregated reward function $f(x) = \sqrt{1 + x} - 1$; 4) the parameter for the value range $\sigma = 2$.

Impact of Round Number. Figure 3 presents the impact of round number on algorithms' regrets. As shown in Figure 3, SW-BCAR and SWUCB-BCAR can always obtain better results than the benchmark algorithms in the considered situations. As T increases, the regrets of other methods will increase faster than that of our algorithms. In Figure 3(b) and Figure 3(d), as T grows, the ratio between the regrets attained by our algorithms and the optimal reward gradually decreases. At the same time, the ratio between the regrets attained by the benchmark algorithms and the optimal reward is almost unchanged. We calculate that the attained total regrets of our algorithms are at least 79.38% lower than those of the benchmark algorithms, when $T \geq 20000$ and $K = 2$. In Figure 3(a) and Figure 3(c), the regret for SWUCB-BCAR decreases with T. In our problem, the reward function is a monotonically increasing concave function. Although $\mu^* T - \sum_{t=1}^{T} v_{a(t)}(t)$ increases as T increases, $f(\mu^* T) - f(\sum_{t=1}^{T} v_{a(t)}(t))$ does not necessarily increase as T increases. An illustrative example is $\sqrt{1.21} - \sqrt{1} \geq \sqrt{4.31} - \sqrt{4}$. We can establish that the experimental results align with the theoretical analysis.

Impact of Reward Function. Figure 4 presents the impact of different aggregated reward functions on algorithms' regrets. The larger the function parameter c is, the faster the slope of the function changes as its input changes. When $c = 1$, i.e. $f(x) = x$, this problem degenerates into the Stochastic MAB Problem. The algorithm UCB1 and exp3 which are designed specifically for this situation have lower regret than SW-BCAR and SWUCB-BCAR. The results are within our expectations. When the function's slope varies rapidly with its input, SW-BCAR and SWUCB-BCAR attain a smaller regret than other algorithms. Furthermore, our algorithms can provide an excellent performance guarantee no matter how the reward function changes. These experimental results also verify our algorithm's effectiveness on the BCAR. We calculate that the attained total regrets of our algorithms are at least 67.96% lower than those of the benchmark algorithms, when $c \geq 1.4$.

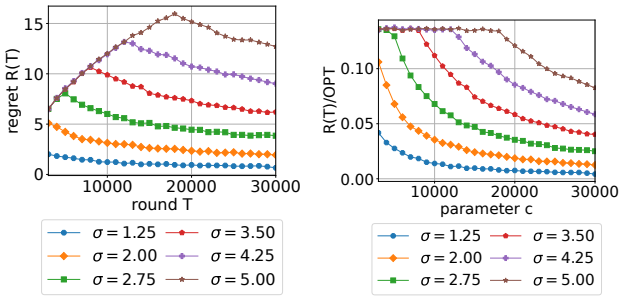
Impact of Mean Value. Figure 5 presents the impact of the suboptimal arm's mean value on algorithms' regret. In these simulations, we find that SW-BCAR and SWUCB-BCAR can always obtain better result than benchmark algorithms. The cost of selecting the wrong arm decreases as the mean of the suboptimal arm increases. Consequently, various algorithms can yield satisfactory results in this scenario. Conversely, when the suboptimal arm mean is small, SW-BCAR and SWUCB-BCAR surpass other algorithms by achieving significantly better rewards.

Impact of Parameter for the Value Range. Figure 6 presents the impact of the parameter for the value range on algorithms' regret. As shown in Figure 6, for parameters that meet the requirements, the closer they are to $\frac{1}{\mu^*}$, the smaller



(a) $R(T)$: mean values (b) $R(T)/\text{OPT}$: mean values

Figure 5: Regret with different mean values.



(a) $R(T)$: parameters (b) $R(T)/\text{OPT}$: parameters

Figure 6: Regret with different parameters.

regrets SWUCB-BCAR will obtain. As the parameter becomes larger, SWUCB-BCAR requires more rounds to distinguish the optimal arm. A small round number is not enough for the agent to distinguish which arm is better, thus the algorithm cannot obtain a satisfactory result. When T is large enough, $R(T)/\text{OPT}$ will gradually decrease as T increases.

6 Related Work

The Bandits with Concave Reward Functions. Some models have been proposed for formulating bandits with concave reward functions [Agrawal and Devanur, 2014; Xu *et al.*, 2020]. However, in these models the arm values and the concave functions are assumed to be known, which significantly decreases the complexity of the problem.

The Stochastic MAB Problem. In the stochastic MAB problem [Gittins, 1979], when the agent chooses an arm a, he will receive a stochastic reward, which is sampled independently from an unknown distribution. The two most representative solutions for this problem are UCB [Gittins, 1979] method and Thompson sampling [Thompson, 1933]. Both methods give a regret bound of $\mathbb{E}[R(T)] \leq \sqrt{KT \log T}$. In our model, the concave aggregated reward function introduces additional complexity into the problem and requires development of different strategies to effectively solve it.

The Adversarial MAB Problem. In the adversarial MAB problem [Auer *et al.*, 2002b; Mohri and Yang, 2016; Saha and Tewari, 2011; Scaman *et al.*, 2017; van der Hoeven *et al.*, 2020; Lykouris *et al.*, 2020], an adversary possesses the ability to manipulate the reward associated with each arm[Hazan and Kale, 2011]. One of the most widely employed algorithms to address this problem is the exp3 algorithm[Auer *et al.*, 2002b], which achieves a regret of $O(\sqrt{KT \log K})$ in this model. However, the regret fails to completely meet the agent's requirements in our model.

The Non-stationary Stochastic MAB problem. In the non-stationary stochastic MAB problem[Russac *et al.*, 2019; Besbes *et al.*, 2014; Trovò *et al.*, 2020; Garivier and Moulines, 2008], the mean reward of each arm changes over time. In the restless bandits[Whittle, 1988], the reward distributions change in each round, but the agent is aware of these changes. In the rested bandits[Tekin and Liu, 2012], an arm's reward distribution changes only when it is chosen by the agent. In the rotting bandits[Levine *et al.*, 2017], the reward of each arm decays with the number of times it is selected, and the algorithm SWA achieves a regret as $\tilde{O}(K^{1/3}T^{2/3})$. In the bandits with adversarial scaling[Lykouris *et al.*, 2020], rewards consist of two components: the random values generated by the arms and the "scaling" determined by an adversary. The algorithm AAEAS attains a regret as $O\left(\sum_{a \neq a^*} \frac{k \log(kT)}{\Delta(a)}\right)$ in this model. In the stochastic rising bandits[Metelli *et al.*, 2022], the expected rewards of arms are determined by the number of times they have been pulled or the current round number. In this model, The algorithm Rless attains a regret as $\tilde{O}(T^{2/3})$.

In our model, the aggregated reward function introduces two novel challenges that must be addressed: 1) the rewards obtained from the agent's previous selections have an impact on his future rewards, and 2) the cumulative reward may not increase linearly with the round number T. Even if the agent employs existing methodologies and attains a regret that is sublinear with respect to T, we are unable to provide a theoretical proof regarding the agent's ability to attain a reward of $(1 - o(1))OPT$. However, by utilizing our proposed methods, we are able to establish a theoretical proof that the agent can achieve a reward of $(1 - o(1))OPT$ in this problem.

7 Conclusion

In this study, we proposed a novel bandit framework named the Bandits with Concave Aggregated Reward. To address this problem, we developed a sliding-window type algorithm SW-BCAR, and proved that the regret is upper bounded by $\tilde{O}(K^{1/3}T^{-1/3})\text{OPT}$. Additionally, for the parameter case, we proposed another algorithm SWUCB-BCAR, and we also proved that the regret is upper bounded by $\tilde{O}(K^{1/2}T^{-1/2})\text{OPT}$, which matches the optimal regret bound for this problem. Our framework and results provide a starting point for further exploration of the BCAR problem. There are several intriguing questions that remain for further research. One is to establish a lower bound on the regret for the fundamental case. Another question is to study a general MAB under additional constraints such as budget limits and combinatorial bandits. Furthermore, extending the optional range from an arm set to a d-dimensional space is also an inspiring and significant problem that awaits to a solution.

Acknowledgements

The research is partially supported by National Key R&D Program of China under Grant No. 2021ZD0110400, Innovation Program for Quantum Science and Technology 2021ZD0302900 and China National Natural Science Foundation with No. 62132018, 62231015, "Pioneer" and "Leading Goose" R&D Program of Zhejiang, 2023C01029, and 2023C01143, Fundamental Research Funds for the Central Universities WK2150110024, and Key Laboratory of Internet and industrial integration and innovation (CAICT), MIIT.

References

[Agarwal *et al.*, 2009] Deepak Agarwal, Bee-Chung Chen, and Pradheep Elango. Spatio-temporal models for estimating click-through rate. In *Proceedings of the 18th international conference on World wide web*, pages 21–30, 2009.

[Agrawal and Devanur, 2014] Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006, 2014.

[Agrawal *et al.*, 1988] Rajeev Agrawal, MV Hedge, and Demosthenis Teneketzis. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. *IEEE Transactions on Automatic Control*, 33(10):899–906, 1988.

[Auer *et al.*, 2002a] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[Auer *et al.*, 2002b] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[Badanidiyuru *et al.*, 2012] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Yaron Singer. Learning on a budget: posted price mechanisms for online procurement. In *Proceedings of the 13th ACM conference on electronic commerce*, pages 128–145, 2012.

[Badanidiyuru *et al.*, 2013] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216. IEEE, 2013.

[Besbes *et al.*, 2014] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27, 2014.

[Bubeck and Cesa-Bianchi, 2012] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

[Garivier and Moulines, 2008] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

[Gittins, 1979] John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

[Hazan and Kale, 2011] Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(4), 2011.

[Lai *et al.*, 1985] Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[Levine *et al.*, 2017] Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. *arXiv preprint arXiv:1702.07274*, 2017.

[Li *et al.*, 2010] Lihong Li, Wei Chu, John Langford, and Robert 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.

[Li *et al.*, 2011] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306, 2011.

[Li *et al.*, 2023] Shaoang Li, Lan Zhang, Yingqi Yu, and Xiangyang Li. Optimal arms identification with knapsacks. In *International Conference on Machine Learning*, pages 20529–20555. PMLR, 2023.

[Lykouris *et al.*, 2020] Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Bandits with adversarial scaling. In *International Conference on Machine Learning*, pages 6511–6521. PMLR, 2020.

[Maillard *et al.*, 2011] Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 497–514. JMLR Workshop and Conference Proceedings, 2011.

[Metelli *et al.*, 2022] Alberto Maria Metelli, Francesco Trovo, Matteo Pirola, and Marcello Restelli. Stochastic rising bandits. In *International Conference on Machine Learning*, pages 15421–15457. PMLR, 2022.

[Mohri and Yang, 2016] Mehryar Mohri and Scott Yang. Optimistic bandit convex optimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2297–2305, 2016.

[Russac *et al.*, 2019] Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. *arXiv preprint arXiv:1909.09146*, 2019.

[Saha and Tewari, 2011] Ankan Saha and Ambuj Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 636–642. JMLR Workshop and Conference Proceedings, 2011.

[Scaman *et al.*, 2017] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *international conference on machine learning*, pages 3027–3036. PMLR, 2017.

[Slivkins and Vaughan, 2014] Aleksandrs Slivkins and Jennifer Wortman Vaughan. Online decision making in crowdsourcing markets: Theoretical challenges. *ACM SIGecom Exchanges*, 12(2):4–23, 2014.

[Tekin and Liu, 2012] Cem Tekin and Mingyan Liu. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8):5588–5611, 2012.

[Thompson, 1933] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[Trovò *et al.*, 2020] Francesco Trovò, Stefano Paladino, Marcello Restelli, and Nicola Gatti. Sliding-window thompson sampling for non-stationary settings. *J. Artif. Intell. Res.*, 68:311–364, 2020.

[van der Hoeven *et al.*, 2020] Dirk van der Hoeven, Ashok Cutkosky, and Haipeng Luo. Comparator-adaptive convex bandits. *arXiv preprint arXiv:2007.08448*, 2020.

[Whittle, 1988] Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.

[Xu *et al.*, 2020] Huanle Xu, Yang Liu, Wing Cheong Lau, and Rui Li. Combinatorial multi-armed bandits with concave rewards and fairness constraints. In *IJCAI*, pages 2554–2560, 2020.