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

doi:10.3233/FAIA240723

Merit-Based Fair Combinatorial Semi-Bandit with Unrestricted Feedback Delays

Ziqun Chen[1], **Kechao Cai**[1,*], **Zhuoyue Chen**[1], **Jinbei Zhang**[1] **and John C.S. Lui**[2]

[1]Sun Yat-sen University, Shenzhen, China
[2]The Chinese University of Hong Kong, Hong Kong, China

Abstract. We study the stochastic combinatorial semi-bandit problem with unrestricted feedback delays under merit-based fairness constraints. This is motivated by applications such as crowdsourcing, and online advertising, where immediate feedback is not immediately available and fairness among different choices (or arms) is crucial. We consider two types of unrestricted feedback delays: reward-independent delays where the feedback delays are independent of the rewards, and reward-dependent delays where the feedback delays are correlated with the rewards. Furthermore, we introduce merit-based fairness constraints to ensure a fair selection of the arms. We define the reward regret and the fairness regret and present new bandit algorithms to select arms under unrestricted feedback delays based on their merits. We prove that our algorithms all achieve sublinear expected reward regret and expected fairness regret, with a dependence on the quantiles of the delay distribution. We also conduct extensive experiments using synthetic and real-world data and show that our algorithms can fairly select arms with different feedback delays.

1 Introduction

In the stochastic combinatorial multi-armed bandit (CMAB) problem with semi-bandit feedback, a learner can select more than one arm at each round and can receive feedback from each selected arm. However, in practice, the feedback is not readily available in many real-world applications. For example, consider the task assignment problem in a crowdsourcing platform where arms represent the workers and feedback (reward) represents the payoff of selecting a worker. Each completed task yields a payoff based on the quality of the worker. The payoff may be delayed since each task requires a certain amount of time to complete. This differs from the typical bandit settings where the learner can receive the feedback immediately after selecting an arm. As another example, in online advertising, the customers usually take hours or even days to make a purchase after clicking an ad [2].

In general, the feedback delays in the bandit problems may be *unrestricted* with unbounded support or expectations. Previous studies on stochastic delayed bandit problems relied on various assumptions regarding the delay distribution such as bounded expectation [13, 19], identical delay distribution across arms [28], and the prior knowledge of delay distribution [8], and none of them can address unrestricted delays. In this paper, we consider two different unrestricted delay settings, depending on the relationship between delays and rewards. The first is the *reward-independent delay* setting, where the delay of the feedback from an arm is independent of the reward of the arm. The second is the *reward-dependent delay* setting, where the delay of the feedback of each arm is correlated with the reward of the arm. The reward-dependent delay is motivated by the applications mentioned earlier: in crowdsourcing, the time the worker takes to complete the assigned task is tied to the payoff as tasks with more payoff can take longer to finish; in online advertising, the delay after collecting the revenue from an ad click often depends on the purchase price paid by the customer. Such a setting is challenging as the feedback would provide a biased estimation of the expected reward of an arm. Take an arm with a Bernoulli reward as an example. If the feedback delay associated with reward 1 is smaller than the feedback delay associated with reward 0, the learner would receive reward 1 earlier and more frequently than reward 0. As a result, the observed empirical average reward of the arm would deviate from the actual mean reward and bias towards reward 1. In some cases, the directions of such deviations may be opposite between different arms. When the fraction of unobserved feedback is large, the observed empirical average reward of the good arm may be much smaller than that of the bad arm, which adds another dimension of complexity to the problem.

In addition, ensuring fairness among the arms is another critical concern in many bandit problems. While existing works mainly focus on maximizing the cumulative rewards, there is a growing recognition that such a unitary consideration can be problematic as it ignores the interests of arms, resulting in an unfair selection of arms [20]. Consider a bandit algorithm that tries to maximize the reward by assigning tasks to workers in a crowdsourcing platform, the algorithm will learn which worker has the highest quality and constantly assign the task to that worker, even if other workers are almost equally good. This will result in a winner-takes-all allocation where many skillful workers will not receive sufficient tasks, and therefore lose interest in the platform. Thus, to build a sustainable platform, a good algorithm must ensure fairness among workers and guarantee that workers with similar skill levels have similar probabilities of receiving tasks. Similarly, in online advertising, the ad publishers wish to ensure fairness among ads and guarantee that all ads have some opportunities to be displayed. This approach not only enhances the platform's appeal to advertisers but also sustains a diverse range of content on the website.

Main contributions. In this paper, we formulate a combinatorial semi-bandit problem to maximize the cumulative reward while ensuring merit-based fairness among arms with unrestricted feedback delays.

* Corresponding Author. Email: caikch3@mail.sysu.edu.cn

We define the *merit* of an arm as a function of its expected reward and impose *merit-based fairness* constraints to ensure each arm is selected with a probability proportional to its merit under feedback delays. In particular, we do not make any assumptions on the delay distributions and allow for unbounded support and expectation of delays. We propose four different fair algorithms for both reward-independent and reward-dependent delay settings and define reward regret and fairness regret to measure their performance. Specifically, in the reward-independent delay setting, we propose an algorithm (FCUCB-D) based on Upper Confidence Bound (UCB) and a computation-efficient algorithm (FCTS-D) based on Thompson Sampling (TS) to ensure merit-based fairness among arms. In the more challenging reward-dependent delay setting, we propose OP-FCUCB-D and OP-FCTS-D algorithms using both *optimistic and pessimistic estimates* of the delayed unobserved rewards to accommodate the estimation biases.

We prove that our proposed algorithms all achieve sublinear upper bounds for both expected fairness regret and expected reward regret, scaling with the quantile of delay distributions. We further conduct experiments using synthetic and real-world data. Our experiment results show that our algorithms outperform other algorithms by fairly selecting arms according to the merits of the arms while maximizing the cumulative reward under different types of feedback delays.

2 Related work

The CMAB problems have been extensively studied [24, 16]. Many works extend the combinatorial semi-bandit to various settings, such as general nonlinear reward [4], probabilistically triggered arms [5, 30], etc. Their algorithmic designs either follow the principle of optimism in the face of uncertainty such as the UCB algorithm [1], or posterior sampling such as the TS algorithm [27].

Delayed bandit Delayed feedback has drawn lots of attention since Dudik et al. [7] first introduced it in stochastic bandit problems. Most studies make various assumptions on the delay distributions. For instance, Joulani et al. [13] explore the impact of delay in both the stochastic and adversarial settings under the assumption that the expectations of the delays are bounded. Mandel et al. [19] develop a bandit model with bounded delays. Besides, Vernade et al. [28] study the delayed bandit with partially observable feedback, where the learner cannot differentiate between the non-received reward and the zero reward. They assume that delays are the same for all arms and have a bounded expectation. Gael et al. [8] also consider partially observed feedback and study heavy-tailed delay distributions which might have infinite expectations. Nevertheless, they assume the parameter of delay distributions is known to the learner. There has also been an emerging interest in bandit problems with unrestricted delays. The recent work [31] proves a sublinear regret upper bound for the TS algorithm with arbitrary delay distributions. Lancewicki et al. [15] introduce reward-dependent feedback delays and design algorithms based on successive elimination with no fairness concerns.

Bandit with fairness constraints Joseph et al. [11, 12] study fairness learning in bandit problems, introducing the notion of meritocratic fairness, where a better arm is always no less likely to be selected than a worse arm. However, their approach favors the arm with the highest expected reward and ignores the merits of other arms. Schumann et al. [23] partition arms into groups based on specific features. They introduce a group fairness notion, preventing the learner from favoring one arm over another based on group information. Other studies [17, 21, 25] investigate fairness guarantees in bandit problems to ensure that each arm must be selected for a predetermined required fraction over all rounds. Liu et al. [18] impose a smoothness constraint to achieve calibrated fairness where the probability of selecting an arm equals the probability of it having the highest reward. Our model subsumes their setting by introducing a more general merit function, with the objective of guaranteeing that each arm receives a selection fraction proportional to its merit. This concept of merit-based fairness has been explored in the single-play bandit [29] and combinatorial contextual bandit [10] where the goal is to ensure that similar arms obtain comparable treatment.

Our work differs from previous studies by considering two types of unrestricted feedback delays, namely, reward-independent delays and reward-dependent delays, in combinatorial semi-bandit bandit problems. Moreover, our algorithms not only ensure the maximization of cumulative reward but also guarantee the selection of each arm with a probability proportional to its merit, all without assuming any specific delay distributions.

3 Fair CMAB with General Feedback Delays

Let $[K] := \{1, 2, ..., K\}$ denote the set of K arms and $[T] := \{1, 2, ..., T\}$. A learner will interact with the arms sequentially over T rounds. At each round $t \in [T]$, each arm $a \in [K]$ is associated with: (i) a reward $R_{t,a} \in [0, 1]$ that follows an unknown distribution ν_a with mean μ_a; (ii) an unknown delay $D_{t,a} \in \mathbb{N}$ such that the reward of arm a can only be revealed to the learner at the end of the round $t + D_{t,a}$. At round t, the learner selects a subset A_t of L ($L \leq K$) arms from $[K]$ receives possibly delayed feedback $Y_{t,a}$ from each arm $a \in [K]$. Essentially, $Y_{t,a}$ is the aggregated rewards from arm a in previous rounds and can be expressed as follows:

$$Y_{t,a} = \sum_{s=1}^{t} R_{s,a} \mathbb{1}_{\{D_{s,a}=t-s\}} \mathbb{1}_{\{a \in A_s\}}, \tag{1}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function. The term $\mathbb{1}_{\{D_{s,a}=t-s\}}$ in (1) takes account of the delays $D_{s,a}$ for $s \leq t$. We note that neither the delay $D_{s,a}$ nor the round number s (the original time of the reward) can be deduced from the feedback $Y_{t,a}$. Let $N_{t,a} = \sum_{s:s<t} \mathbb{1}_{\{a \in A_s\}}$ denote the number of rounds that arm a has been selected up to round $t-1$, and $M_{t,a} = \sum_{s:s+D_{s,a}<t} \mathbb{1}_{\{a \in A_s\}}$ denote the number of delayed feedbacks that the learner can receive from arm a up to round $t-1$. As the feedback may be delayed, we have $M_{t,a} \leq N_{t,a}$. Thus, at the beginning of round t, the empirical average reward of arm a can be expressed as: $\hat{\mu}_{t,a} = \frac{1}{M_{t,a} \vee 1} \sum_{s':s'<t} Y_{s',a}$, where $m \vee n = \max\{m, n\}$. Note that we do not assume that the delays follow any particular distribution and even allow $D_{t,a}$ being infinite, in which case the reward from arm a would never be received. Specifically, we introduce a quantile function to describe the distribution of the delays for each arm. For an arm a with a delay D_a, we define the quantile function $d_a(q)$ as

$$d_a(q) = \min\{\zeta \in \mathbb{N} \mid \mathbb{P}[D_a \leq \zeta] \geq q\}, \tag{2}$$

where the quantile $q \in (0, 1]$ and $d^*(q) = \max_a d_a(q)$.

Finally, we consider a merit function $f(\cdot) > 0$ that maps the expected reward of an arm to a positive merit value. We have two assumptions on the merit function $f(\cdot)$.

Assumption 1. *The merit of each arm is bounded such that (i)* $\exists \lambda > 0$ *and* $\min_\mu f(\mu) \geq \lambda$, *(ii)* $\forall \mu_1, \mu_2 \in [0, 1]$, $\frac{f(\mu_1)}{f(\mu_2)} \leq \frac{K-1}{L-1}$ *for* $L > 1$.

Assumption 2. *The merit function* f *is* M*-Lipschitz continuous, i.e., there exists a positive constant* $M > 0$, *such that* $\forall \mu_1, \mu_2 \in [0, 1], |f(\mu_1) - f(\mu_2)| \leq M |\mu_1 - \mu_2|$.

To ensure merit-based fairness among the arms, we enforce a constraint that the probability p_a of selecting arm a is proportional to its merit $f(\mu_a)$. Formally, we have

$$\frac{p_a}{f(\mu_a)} = \frac{p_{a'}}{f(\mu_{a'})}, \quad \forall a \neq a', a, a' \in [K]. \tag{3}$$

Fairness criteria in various applications can be tailored by selecting different $f(\cdot)$. For instance, setting $f(\cdot)$ as a threshold function would grant higher merits to arms whose expected rewards exceed a predefined threshold.

We now show that there is a unique optimal fair policy that fulfills the fairness constraints in (3) in the following theorem.

Theorem 3. *For any $\mu_a, a \in [K]$ and any choice of merit function $f(\cdot) > 0$, there exist a unique optimal fair policy $\boldsymbol{p}^* = \{p_1^*, p_2^*, ..., p_K^*\}$ such that*

$$p_a^* = \frac{Lf(\mu_a)}{\sum_{a'=1}^{K} f(\mu_{a'})}, \quad \forall a \in [K], \tag{4}$$

that satisfies the merit-based fairness constraints in (3).

We refer the interested readers to Appendix A in the full version of the paper [6] for the proofs of all the theorems. Theorem 3 implies that the optimal fair policy is no longer selecting a fixed optimal set of L arms as in classical bandit problems, but a probability distribution on all the possible sets $A_t \subseteq [K]$, $|A_t| = L$. To be more specific, we characterize an arm selection algorithm with a probabilistic selection vector $\boldsymbol{p}_t = \{p_{t,1}, p_{t,2}, ..., p_{t,K}\}$ where $p_{t,a} \in [0,1]$ is the probability of selecting arm $a \in [K]$ at round t, and $\sum_{a=1}^{K} p_{t,a} = L$ since only L arms can be selected at each round. To measure the gap of cumulative reward between the optimal fair policy and a bandit algorithm, we define the *reward regret* of an algorithm as follows:

$$\mathrm{RR}_T = \sum_{t=1}^{T} \max\left\{\sum_{a=1}^{K} p_a^* \mu_a - \sum_{a=1}^{K} p_{t,a}\mu_a, 0\right\}. \tag{5}$$

We use the reward regret to quantify the speed of reward optimization of an algorithm. Specifically, we only consider the non-negative part at each round in (5) as a less fair algorithm could yield a larger reward than the optimal fair policy and cause a negative reward gap. Moreover, we also require a measure to quantify its fairness guarantee. In this work, we define the *fairness regret* that measures the cumulative 1-norm distance between the optimal fair policy $\boldsymbol{p}^*$ and the selection vector $\boldsymbol{p}_t$ of an algorithm as follows:

$$\mathrm{FR}_T = \sum_{t=1}^{T}\sum_{a=1}^{K} |p_a^* - p_{t,a}|. \tag{6}$$

The fairness regret measures the overall violation of the merit-based fairness constraints. Our objective is to design algorithms that have both *sublinear expected reward regret* and *sublinear expected fairness regret* with respect to the number of rounds T, where the expectations are taken over the randomness in both the arm selections and the rewards. By doing so, we can approach the optimal fair policy and maximize the cumulative reward while ensuring merit-based fairness among all the arms in the long run.

It is important to point out that both Assumption 1 and Assumption 2 are necessary for designing bandit algorithms as stated in the following theorem and remark.

Theorem 4. *For any bandit algorithm, if either Assumption 1 (i) or Assumption 2 does not hold, the lower bound of the fairness regret is linear; in other words, there exists a CMAB instance with linear expected fairness regret $O(T)$.*

Remark 5. *Assumption 1 (ii) ensures that the selection probability $p_{t,a}$ in the form of $Lf(\cdot)/\sum_{a=1}^{K} f(\cdot)$ is constrained in $[0,1]$.*

In the following sections, we introduce fair bandit algorithms under two types of feedback delays, *reward-independent feedback delays* and *reward-dependent feedback delays*.

4 Algorithms for Reward-independent Delays

In this section, we first consider that the feedback delays are independent of the rewards of arms. We design two bandit algorithms to ensure merit-based fairness under the reward-independent delays.

4.1 FCUCB-D Algorithm

Algorithm 1 shows the details of our *Fair CUCB with reward-independent feedback Delays* (FCUCB-D) algorithm, which follows the principle of optimism in the face of uncertainty without requiring any prior knowledge of delay distributions. At each round t, we first calculate the average rewards of all arms based on the received delayed feedback. With the average rewards, we construct a confidence region $\mathcal{C}_t$ (see Line 8) using both UCB estimates $U_{t,a}$ and LCB (Lower Confidence Bound) estimates $B_{t,a}$ of all arms, where the vector $\tilde{\boldsymbol{\mu}} := (\tilde{\mu}_a)_{a\in[K]}$ and $c_{t,a}$ denotes the confidence radius of each arm a. We clip $U_{t,a}$ to 1 and $B_{t,a}$ to 0 since the rewards have support on $[0,1]$. Then we find a vector $\tilde{\boldsymbol{\mu}}_t$ in the confidence region $\mathcal{C}_t$ that maximizes the expected reward of a fair policy as shown in Line 9. Specifically, according to Theorem 3, we construct the probability of selecting arm a as $\frac{Lf(\tilde{\mu}_a)}{\sum_{a'=1}^{K} f(\tilde{\mu}_{a'})}$ to satisfy the merit-based fairness constraints, which is limited to the interval $[0,1]$ under Assumption 1 (ii). Different from the conventional bandit algorithms such as CUCB [3] which deterministically selects L arms at each round, our algorithm selects L arms stochastically with the selection vector $\boldsymbol{p}_t$ to ensure fairness. In particular, we incorporate a randomized rounding scheme (RRS) from [9]. RRS takes a probabilistic selection vector $\boldsymbol{p}_t$ ($\sum_{a=1}^{K} p_{t,a} = L$) as input and generates a set of arms A_t such that $\mathbb{E}[\mathbb{1}_{\{a\in A_t\}}] = p_{t,a}$. Finally, we receive delayed feedback from all arms.

We present the expected fairness regret and reward regret upper bounds of FCUCB-D in the following theorem.

Theorem 6. *Suppose that $\forall t > \lceil K/L \rceil, a \in [K], R_{t,a} \in [0,1]$ and feedback delays are reward-independent. Set $c_{t,a} = \sqrt{\frac{\log(4LKT)}{M_{t,a}\vee 1}}$. When $T > K$, the expected fairness regret of FCUCB-D is upper bounded as:*

$$\mathbb{E}\left[\mathrm{FR}_T\right] = \widetilde{O}\left(\min_{q\in(0,1]}\left\{\frac{ML}{\lambda}\left(\frac{K}{q}\sqrt{T} + Ld^*(q)\right)\right\}\right),$$

and the expected reward regret of FCUCB-D is upper bounded as:

$$\mathbb{E}\left[\mathrm{RR}_T\right] = \widetilde{O}\left(\min_{q\in(0,1]}\left\{\frac{K}{q}\sqrt{T} + Ld^*(q)\right\}\right),$$

where $\widetilde{O}$ hides the polylogarithmic factors in T.

In Theorem 6, the factor $\frac{ML}{\lambda}$ in $\mathbb{E}\left[\mathrm{FR}_T\right]$ comes from Assumption 1 and Assumption 2 on the merit function $f(\cdot)$. Note that both upper bounds are valid for any quantile $q \in (0,1]$, and one can optimize the bounds by selecting the optimal q.

Our FCUCB-D algorithm differs from the single-played FairX-UCB algorithm [29] as it addresses a more challenging combinatorial semi-bandit problem involving feedback delays. Moreover, our

Algorithm 1 **F**air **CUCB** with reward-independent feedback **D**elays (FCUCB-D)

Input: $f(\cdot)$, T, L, K
Init: Select each arm in $[K]$ once with $\lceil K/L \rceil$ rounds.
1: **for** $t = \lceil K/L \rceil + 1$ to T **do**
2: **for** $a \in [K]$ **do**
3: $M_{t,a} = \sum_{s:s+D_{s,a}<t} \mathbb{1}_{\{a \in A_s\}}$
4: $\hat{\mu}_{t,a} = \frac{1}{M_{t,a} \vee 1} \sum_{s':s'<t} Y_{s',a}$
5: $U_{t,a} = \min\{\hat{\mu}_{t,a} + c_{t,a}, 1\}$
6: $B_{t,a} = \max\{\hat{\mu}_{t,a} - c_{t,a}, 0\}$
7: **end for**
8: $\mathcal{C}_t = \{\tilde{\boldsymbol{\mu}} | \forall a \in [K],\ \tilde{\mu}_a \in [B_{t,a}, U_{t,a}]\}$
9: $\tilde{\boldsymbol{\mu}}_t = \arg\max_{\tilde{\boldsymbol{\mu}} \in \mathcal{C}_t} \sum_{a=1}^{K} \frac{Lf(\tilde{\mu}_a)}{\sum_{a'=1}^{K} f(\tilde{\mu}_{a'})} \tilde{\mu}_a$
10: Compute $p_{t,a} = \frac{Lf(\tilde{\mu}_{t,a})}{\sum_{a'=1}^{K} f(\tilde{\mu}_{t,a'})}$ for $a \in [K]$
11: Select arms in $A_t = \mathrm{RRS}(L, \boldsymbol{p}_t)$
12: Receive delayed feedback $Y_{t,a}$ from $a \in [K]$
13: **end for**

theoretical results accommodate unbounded delays since the upper bounds depend on the quantiles of the delay distribution instead of the expectation of the delays as in [13, 22, 25].

4.2 FCTS-D Algorithm

The computational complexity of FCUCB-D may be high for a large K. In particular, Line 9 in Algorithm 1 could involve a non-convex constrained optimization problem, which requires a complex optimization solver for finding the optimal solution.

To tackle this problem, we incorporate a TS-based method in our algorithm design and propose the *Fair CTS with reward-independent feedback Delays* (FCTS-D) algorithm without invoking an optimization solver. The details of FCTS-D are described in Algorithm 2. Initially, the algorithm starts with a prior distribution $\boldsymbol{\mathcal{Q}}_1 := (\mathcal{Q}_{1,a})_{a \in [K]}$ where $\mathcal{Q}_{1,a}$ represents the learner's prior belief about the reward of arm a. At each round t, for each arm a, we generate a sample $\tilde{\mu}_{t,a}$ as the reward estimate from the posterior distribution $\mathcal{Q}_{t,a}$ (see Line 3) and compute the selection probability $p_{t,a}$. Then we select L arms using the selection probability distribution $\boldsymbol{p}_t$ via the RRS described in Algorithm 1. Finally, we update the posterior distribution $\boldsymbol{\mathcal{Q}}_t := (Q_{t,a})_{a \in [K]}$ using the received delayed feedback $\boldsymbol{Y}_t := (Y_{t,a})_{a \in [K]}$ at Line 8.

Based on the Bayesian setting and given the prior reward distributions, we derive the following theorem on the expected fairness/reward regret of FCTS-D.

Theorem 7. *$\forall a \in [K]$, given a uniform prior on μ_a and suppose that $\forall t \in [T]$, $R_{t,a}$ is Bernoulli distributed and the feedback delays are reward-independent. When $T > K$, the expected fairness regret of FCTS-D is upper bounded as:*

$$\mathbb{E}\left[\mathrm{FR}_T\right] = \widetilde{O}\left(\min_{q \in (0,1]} \left\{ \frac{ML}{\lambda} \left(\frac{K}{q}\sqrt{T} + Ld^*(q) \right) \right\}\right),$$

and the expected reward regret of FCTS-D is upper bounded as:

$$\mathbb{E}\left[\mathrm{RR}_T\right] = \widetilde{O}\left(\min_{q \in (0,1]} \left\{ \frac{K}{q}\sqrt{T} + Ld^*(q) \right\}\right),$$

where $\widetilde{O}$ hides the polylogarithmic factors in T.

Algorithm 2 **F**air **CTS** with reward-independent feedback **D**elays (FCTS-D)

Input: $f(\cdot)$, T, L, K, $\boldsymbol{\mathcal{Q}}_1$
1: **for** $t = 1$ to T **do**
2: **for** $a \in [K]$ **do**
3: Sample $\tilde{\mu}_{t,a}$ from posterior $\mathcal{Q}_{t,a}$
4: Compute $p_{t,a} = \frac{Lf(\tilde{\mu}_{t,a})}{\sum_{a'=1}^{K} f(\tilde{\mu}_{t,a'})}$
5: **end for**
6: Select arms in $A_t = \mathrm{RRS}(L, \boldsymbol{p}_t)$
7: Receive delayed feedback $Y_{t,a}$ from $a \in [K]$
8: $\boldsymbol{\mathcal{Q}}_{t+1} = \mathrm{Update}(\boldsymbol{\mathcal{Q}}_t, \boldsymbol{Y}_t)$
9: **end for**

Note that the expected fairness regret and reward regret upper bounds of the FCTS-D are in the same order as the expected fairness/reward regret of the FCUCB-D. They are also dependent on the quantiles of the delay distribution. Nevertheless, FCTS-D avoids solving the optimization problem by using the Bayesian posterior sampling method, thus it is more computationally efficient than FCUCB-D.

5 Algorithms for Reward-dependent Delays

We now consider a more challenging reward-dependent delay setting where the feedback delay of each arm is correlated with the received reward at the same round. In other words, the two random variables are drawn from a joint distribution over delays and rewards. Then we propose another two bandit algorithms to maximize the cumulative reward and ensure merit-based fairness among arms under the reward-dependent feedback delays.

5.1 OP-FCUCB-D Algorithm

In the reward-dependent delay setting, the key challenge arises as the empirical average reward of each arm is no longer an unbiased estimator of the expected reward. This issue occurs when the feedback delays associated with high rewards distribute differently from the feedback delays associated with low rewards. Thus, the empirical average rewards would be quite different from the actual expected rewards, given that high rewards and low rewards are received with differently distributed delays. In this context, our previous algorithms, FCUCB-D and FCTS-D, that require unbiased reward estimates, are no longer applicable.

To address such biases in the delayed feedback, we introduce a novel variant of FCUCB-D, named *Optimistic-Pessimistic Fair CUCB with reward-dependent feedback Delays* (OP-FCUCB-D), detailed in Algorithm 3. We leverage both the observed rewards and the optimistic-pessimistic estimates of delayed unobserved rewards. Specifically, in calculating the UCB of an arm, we adopt optimistic estimates, assuming all delayed unobserved rewards attain the maximal value of 1 at Line 5. Conversely, in calculating the LCB, we adopt pessimistic estimates, presuming all the delayed unobserved rewards are the minimal value of 0 at Line 6. Subsequently, we construct an expanded confidence region $\mathcal{C}_t^{\pm}$ using the optimistic UCB $U_{t,a}^{+}$ and pessimistic LCB $B_{t,a}^{-}$ of all arms at Line 10. This approach ensures that the actual expected reward of an arm falls within the expanded confidence region with high probability.

We present the upper bounds on the expected fairness regret and reward regret of OP-FCUCB-D in the following theorem.

Algorithm 3 **O**ptimistic-**P**essimistic **F**air **CUCB** with reward-dependent feedback **D**elays (OP-FCUCB-D)

Input: $f(\cdot), T, L, K$
Init: Select each arm in $[K]$ once with $\lceil K/L \rceil$ rounds.
1: **for** $t = \lceil K/L \rceil + 1$ to T **do**
2: **for** $a \in [K]$ **do**
3: $M_{t,a} = \sum_{s:s+D_{s,a}<t} \mathbb{1}_{\{a \in A_s\}}$
4: $N_{t,a} = \sum_{s:s<t} \mathbb{1}_{\{a \in A_s\}}$
5: $\hat{\mu}^+_{t,a} = \frac{N_{t,a}-M_{t,a}}{N_{t,a}} + \frac{1}{N_{t,a}} \sum_{s':s'<t} Y_{s',a}$
6: $\hat{\mu}^-_{t,a} = \frac{1}{N_{t,a}} \sum_{s':s'<t} Y_{s',a}$
7: $U^+_{t,a} = \min\{\hat{\mu}^+_{t,a} + c_{t,a}, 1\}$
8: $B^-_{t,a} = \max\{\hat{\mu}^-_{t,a} - c_{t,a}, 0\}$
9: **end for**
10: $\mathcal{C}^{\pm}_t = \{\tilde{\boldsymbol{\mu}} | \forall a \in [K], \tilde{\mu}_a \in [B^-_{i,t}, U^+_{i,t}\}$
11: $\tilde{\boldsymbol{\mu}}_t = \arg\max_{\tilde{\boldsymbol{\mu}} \in \mathcal{C}^{\pm}_t} \sum_{a=1}^{K} \frac{Lf(\tilde{\mu}_a)}{\sum_{a'=1}^{K} f(\tilde{\mu}_{a'})} \tilde{\mu}_a$
12: Compute $p_{t,a} = \frac{Lf(\tilde{\mu}_{t,a})}{\sum_{a'=1}^{K} f(\tilde{\mu}_{t,a'})}$ for $a \in [K]$
13: Select arms in $A_t = \mathrm{RRS}(L, \boldsymbol{p}_t)$
14: Receive delayed feedback $Y_{t,a}$ from $a \in [K]$
15: **end for**

Theorem 8. *Suppose that $\forall t > \lceil K/L \rceil, a \in [K]$, $R_{t,a} \in [0,1]$ and feedback delays are reward-dependent. For any $\delta \in (0,1)$, set $c_{t,a} = \sqrt{\frac{\log(6LKT)}{N_{t,a}}}$. When $T > K$, the expected fairness regret of OP-FCUCB-D is upper bounded as:*

$$\mathbb{E}\left[\mathrm{FR}_T\right] = \widetilde{O}\left(\min_{q \in (0,1]} \left\{ \frac{MLK}{\lambda} \left((1-q)T + d^*(q)\sqrt{T} \right) \right\} \right),$$

and the expected reward regret of OP-FCUCB-D is upper bounded as:

$$\mathbb{E}\left[\mathrm{RR}_T\right] = \widetilde{O}\left(\min_{q \in (0,1]} \left\{ L(1-q)T + Kd^*(q)\sqrt{T} \right\} \right),$$

where $\widetilde{O}$ hides the polylogarithmic factors in T.

Compared to FCUCB-D, the regret analysis for OP-FCUCB-D is more challenging since we must consider the entire feedback rather than just the observed ones. Moreover, OP-FCUCB-D could have biased estimates of the actual expected reward using the optimistic-pessimistic estimates, while FCUCB-D always has the unbiased ones. Therefore, it would be reasonable to expect that OP-FCUCB-D has larger reward regret and fairness regret than FCUCB-D. In Theorem 8, we show the two regret upper bounds minimized over the quantile $q \in (0,1]$. In particular, OP-FCUCB-D achieves sublinear expected reward regret and expected fairness regret upper bounds $O(T^{\kappa})$ by setting the quantile $q \geq 1 - T^{\kappa-1}$, $\kappa < 1$.

5.2 OP-FCTS-D Algorithm

To avoid solving the computationally expensive optimization problem in OP-FCUCB-D at Line 11 in Algorithm 3, we further propose a TS-based algorithm, named *Optimistic-Pessimistic Fair CTS with reward-dependent feedback Delays* (OP-FCTS-D), described in Algorithm 4.

In OP-FCTS-D, we also consider both the observed rewards and the delayed unobserved rewards by constructing an optimistic posterior distribution $\boldsymbol{\mathcal{Q}}^+_t := (\mathcal{Q}^+_{t,a})_{a \in [K]}$ and a pessimistic posterior distribution $\boldsymbol{\mathcal{Q}}^-_t := (\mathcal{Q}^-_{t,a})_{a \in [K]}$. When updating the optimistic posterior, all the delayed unobserved rewards are treated as the maximal value of 1. In updating the pessimistic posterior, all the delayed unobserved rewards are considered as the minimal value of 0. At each round t, we sample an optimistic estimate $\tilde{\mu}^+_{t,a}$ from the optimistic posterior $\mathcal{Q}^+_{t,a}$, and a pessimistic estimate $\tilde{\mu}^-_{t,a}$ from the pessimistic posterior $\mathcal{Q}^-_{t,a}$ for $a \in [K]$. Using the average of $\tilde{\mu}^+_{t,a}$ and $\tilde{\mu}^-_{t,a}$, we can then compute the selection probability for each arm a at Line 5. This equal weighting of the optimistic and pessimistic estimates facilitates the analysis of the gap between the optimal fair policy and OP-FCTS-D.

Algorithm 4 **O**ptimistic-**P**essimistic **F**air **CTS** with reward-dependent feedback **D**elays (OP-FCTS-D)

Input: $f(\cdot), T, L, K, \boldsymbol{\mathcal{Q}}_1$
1: **for** $t = 1$ to T **do**
2: **for** $a \in [K]$ **do**
3: Sample $\tilde{\mu}^+_{t,a}$ from optimistic posterior $\mathcal{Q}^+_{t,a}$
4: Sample $\tilde{\mu}^-_{t,a}$ from pessimistic posterior $\mathcal{Q}^-_{t,a}$
5: Compute $p_{t,a} = \frac{Lf\left((\tilde{\mu}^+_{t,a} + \tilde{\mu}^-_{t,a})/2\right)}{\sum_{a'=1}^{K} f\left((\tilde{\mu}^+_{t,a'} + \tilde{\mu}^-_{t,a'})/2\right)}$
6: **end for**
7: Select arms in $A_t = \mathrm{RRS}(L, \boldsymbol{p}_t)$
8: Receive delayed feedback $Y_{t,a}$ from $a \in [K]$
9: $\boldsymbol{\mathcal{Q}}^+_{t+1} = \mathrm{Update}(\boldsymbol{\mathcal{Q}}_t, \boldsymbol{Y}_t)^+$
10: $\boldsymbol{\mathcal{Q}}^-_{t+1} = \mathrm{Update}(\boldsymbol{\mathcal{Q}}_t, \boldsymbol{Y}_t)^-$
11: **end for**

We use the expected fairness regret and the expected reward regret to measure the performance of OP-FCTS-D. We prove the upper bounds of the regrets in the following theorem.

Theorem 9. *$\forall a \in [K]$, given a uniform prior on μ_a and suppose that $\forall t \in [T]$, $R_{t,a}$ is Bernoulli distributed and feedback delays are reward-dependent. The expected fairness regret of OP-FCTS-D is upper bounded as:*

$$\mathbb{E}\left[\mathrm{FR}_T\right] = \widetilde{O}\left(\min_{q \in (0,1]} \left\{ \frac{MLK}{\lambda} \left((1-q)T + \frac{d^*(q)}{q}\sqrt{T} \right) \right\} \right),$$

and the expected reward regret of OP-FCTS-D is upper bounded as:

$$\mathbb{E}\left[\mathrm{RR}_T\right] = \widetilde{O}\left(\min_{q \in (0,1]} \left\{ \frac{MLK}{\lambda} \left((1-q)T + \frac{d^*(q)}{q}\sqrt{T} \right) \right\} \right),$$

where $\widetilde{O}$ hides the polylogarithmic factors in T.

According to Theorem 9, OP-FCTS-D achieves sublinear expected reward regret and expected fairness regret upper bounds $O(T^{\kappa})$ if the quantile $q \geq 1 - T^{\kappa-1}$, $\kappa < 1$. Compared to FCTS-D, both the expected reward regret and the expected fairness regret of OP-FCTS-D depend on the constants λ and M described in Assumption 1 and Assumption 2. This is because OP-FCTS-D does not have the accurate posterior distribution of the rewards due to the reward-dependent feedback delays, and we derive its expected reward regret from its expected fairness regret using Assumption 1 and Assumption 2.

Remark 10. *While our OP-FCUCB-D and OP-FCTS-D algorithms are primarily tailored for the reward-dependent delay setting, they are versatile enough to be applied to CMAB problems with reward-independent delays. However, this application may lead to potentially larger reward regret and fairness regret due to the biases in the optimistic-pessimistic estimates.*

6 Experiments

Here, we conduct experiments[1] using both synthetic and real-world data to demonstrate the effectiveness of our algorithms. We also discuss several interesting observations derived from the experiment results.

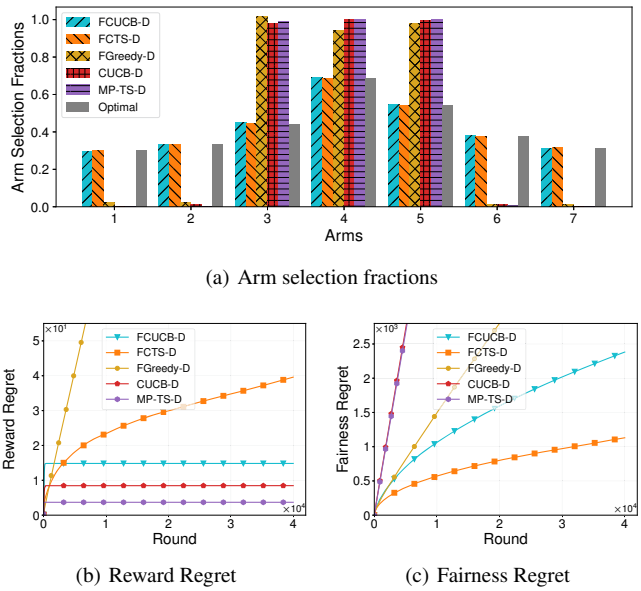
Experiments using synthetic data. We consider a CMAB problem with $K = 7$ arms where the learner selects $L = 3$ arms at each round. The rewards of each arm follow a Bernoulli distribution with mean in $\boldsymbol{\mu} = \{0.3, 0.5, 0.7, 0.9, 0.8, 0.6, 0.4\}$. We examine the impact of delays on the expected fairness/reward regret of our algorithms with several feedback delay settings (i.e., fixed delays [7], geometric delays [28], α-Pareto delays [8], packet-loss delays and biased delays [15]) considered in prior work. We use the merit function $f(\mu) = 1 + 2\mu^c$ to calculate the merit of an arm with expected reward μ under Assumption 1 and Assumption 2, where the parameter c controls the gradients of the merit function. We set $c = 4$ in the following and conduct additional experiments using merit functions with different c in Appendix B in the supplementary material, where we find that the regret gap between different bandit algorithms widens as the parameter c increases. Moreover, in Appendix B, we provide the running time for the same number of rounds of FCUCB-D and FCTS-D (and their corresponding OP versions) to demonstrate the effectiveness of TS-type algorithms in avoiding solving optimization problems. All results are averaged over 100 runs.

We first examine the fairness of different algorithms under the geometric delays with the success probability parameter equal to 0.05. In this case, the feedback delays can be arbitrarily long but the expectation of the delays is finite.

For comparison, we implement three other CMAB algorithms, CUCB-D, MP-TS-D, and FGreedy-D which are adapted from CUCB [3], MP-TS [14] and ϵ-Greedy, respectively, to account for feedback delays. In particular, FGreedy-D selects L arms uniformly at random in the exploration phase and selects L arms with probability $\frac{Lf(\hat{\mu}_{t,a})}{\sum_{a'=1}^{K} f(\hat{\mu}_{t,a'})}$ via RRS in the exploitation phase. FairX-UCB and FairX-TS proposed in [29] are limited in applicability to our setting since they can only select a single arm at each round without accounting for feedback delays. Additionally, other fair bandit algorithms use different fairness metrics, making them unsuitable for direct comparison with our algorithms.

Figure 1(a) illustrates the average arm selection fractions of CUCB-D, MP-TS-D, FGreedy-D, the optimal fair policy, and our FCUCB-D and FCTS-D. Each bar corresponds to the fraction of times an arm is chosen over $T = 4 \times 10^4$ rounds by a specific algorithm. As shown in Figure 1(a), CUCB-D and MP-TS-D are unfair by mainly selecting the arms (arm $3, 4, 5$) with high rewards, neglecting the potential merits of other arms. FGreedy-D tends to select arms uniformly randomly since it randomly explores the arms in the exploration phase. In contrast, both FCUCB-D and FCTS-D can converge to the optimally fair policy. This observation shows the effectiveness of our algorithms in achieving merit-based fairness, ensuring that each arm receives a selection allocation proportional to its merit.

In Figure 1(b) and 1(c), it is evident that CUCB-D and MP-TS-D consistently exhibit smaller reward regret and larger fairness regret when compared to FCUCB-D and FCTS-D. This observation suggests that CUCB-D and MP-TS-D attain high rewards but substantially violate the merit-based fairness constraints. Moreover, the



(a) Arm selection fractions

(b) Reward Regret

(c) Fairness Regret

Figure 1. Comparison of different bandit algorithms under geometric feedback delays.

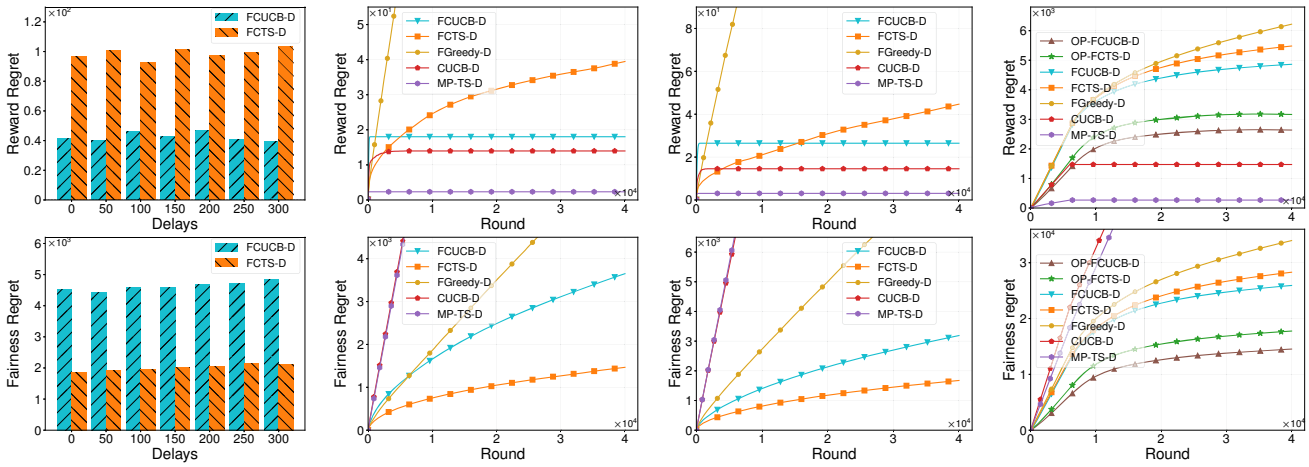
reward/fairness regrets of our algorithms are smaller than FGreedy-D and increase sublinearly in T, aligning with the bounds we derived in Theorem 6 and Theorem 7. In particular, FCUCB-D outperforms FCTS-D in reward regret. However, this advantage comes at the cost of incurring higher fairness regret.

Then we evaluate the performance of different algorithms under different feedback delay settings by changing the delay distributions. Figure 2(a) shows the reward regret and fairness regret of FCUCB-D and FCTS-D under different fixed delays after $T = 10^5$ rounds. As shown in Figure 2(a), the reward regrets and the fairness regrets of our algorithms are quite close under different fixed delays. The reason is that the algorithms can increase exploration of the merits of all the arms by receiving the possible delayed rewards from the wrongly selected arms at each round, and thus they do not incur much regret. This indicates that our algorithms are not sensitive to fixed delays that are not excessively large.

Next, we show the fairness/reward regret of different bandit algorithms under α-Pareto delays and packet-loss delays in Figure 2(b) and Figure 2(c), respectively. These types of delays pose additional challenges as their expected values may be infinite. In the case of α-Pareto delays, the delays of each arm a follow the Pareto Type I distribution with the tail index α_a. A smaller α_a indicates a heavier tail of the delay distribution, and when $\alpha_a \leq 1$, the delays have an infinite expectation. We uniformly sample α_a from the interval $(0, 1]$ for each arm a to model delays with infinite expectations. In the packet-loss delays, the delay is 0 with probability p and infinite otherwise. We uniformly sample the probabilities p from interval $(0.3, 0.8]$ for each arm. Remarkably, compared to other algorithms, FCUCB-D and FCTS-D can achieve both sublinear fairness and reward regret upper bounds across various delay distributions with infinite expectations.

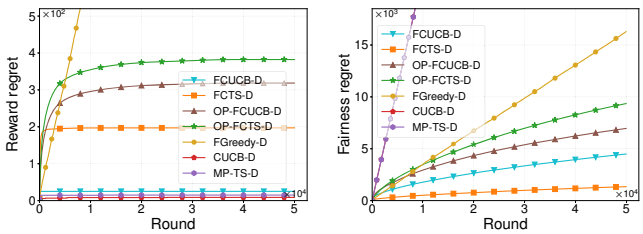
Finally, we examine the performance of different algorithms under the reward-dependent (biased) delays. We note that OPSE [15] is also tailored to handle reward-dependent delays; however, we refrain from comparing it with our algorithms as it eliminates the bad arms, resulting in substantial fairness regret. We set the reward-dependent delays as follows: the good arms (arm $3, 4, 5$) have a fixed delay of $6,000$ rounds for reward 1 and 0 round for reward 0, and the bad

[1] Source code available at https://github.com/MLCL-SYSU/FairCMAB-Delays (Accessed 29-July-2024)



(a) Fixed delays (b) α-Pareto delays (c) Packet-loss delays (d) Biased delays

Figure 2. Experiment results of different bandit algorithms under different types of feedback delays.



(a) Reward Regret (b) Fairness Regret

Figure 3. Experiment results using the real-world conversion log dataset.

arms (arm $1, 2, 6, 7$) have a fixed delay of $6,000$ rounds for reward 0 and 0 round for reward 1. In this setting, as the reward 1 from a bad arm could be received earlier than the reward 1 from a good arm, the empirical average reward of a bad arm would be larger than that of a good arm at the beginning. In Figure 2(d), we observe that OP-FCUCB-D and OP-FCTS-D significantly outperform FCUCB-D and FCTS-D in both reward regret and fairness regret since FCUCB-D and FCTS-D are not aware of the biases in the empirical average rewards. This shows the effectiveness of the optimistic-pessimistic estimates in OP-FCUCB-D and OP-FCTS-D.

Experiments using real-world data. We conduct additional experiments on our algorithms using the conversion log dataset [26] that contains data on users' interactions with a small sample of ads. Each row in the dataset corresponds to a user clicking on an ad, including a conversion indicator denoting whether the user makes a purchase after clicking the ad, as well as the time between the click and the purchase.

We select the top-10 ($K = 10$) clicked ads from the dataset and allocate them to three regions for ad placement ($L = 3$). Each ad is treated as an arm, where a user's click and purchase represent the reward of an arm, and the delay between the click and the purchase serves as the feedback delay. We determine the conversion rate of each ad by normalizing the number of conversions using min-max scaling. Then we generate the reward for each arm using a Bernoulli distribution, with the mean given by the corresponding conversion rate. Since the dataset lacks information on the click rate of ads, we assumed a click rate of 5% for each ad. We compute the time between page visits based on this assumed click rate and the number of ad clicks in the last week provided in the dataset. Then we can derive the delay (in page visits) of the purchase by dividing the time between the click and the purchase by the time between page visits. We use the merit function of the form $f(\mu) = 1 + 3.5\mu^c$ with parameter $c = 4$ and run simulations for $T = 5 \times 10^4$ page visits. All results are averaged over 100 runs.

Figure 3 shows the experiment results on fairness/reward regret for different algorithms. We observe that our algorithms achieve sublinear bounds on fairness/reward regret and exhibit a better tradeoff between reward and fairness on the conversion log dataset, in comparison to other algorithms. In particular, FCUCB-D and FCTS-D outperform OP-FCUCB-D and OP-FCTS-D in terms of reward regret and fairness regret. The rationale behind this is that: the dataset only provides the delay of ads with successful conversions (click and then purchase, reward 1); For an ad with no successful conversion (click without purchase, reward 0), we determine its feedback delay by randomly sampling from the delays of the ads with successful conversions. This approach makes the ads' feedback delays independent of the ads' rewards. Thus, in such a reward-independent delay setting, OP-FCUCB-D and OP-FCTS-D still take the unobserved feedback of the ads into account and incur larger reward regret and fairness than FCUCB-D and FCTS-D.

7 Conclusion & Future Work

In this paper, we propose a novel combinatorial semi-bandit setting with merit-based fairness constraints and two types of unrestricted feedback delays: reward-independent delays and reward-dependent delays. We employ UCB, Thompson Sampling, and optimistic-pessimistic estimates and design novel algorithms that achieve both sublinear expected reward regret and sublinear expected fairness regret. Our extensive simulation results using both synthetic dataset and real-world dataset show that our algorithms fairly select arms according to the merits of the arms under different feedback delays.

For future research, it is interesting to eliminate the assumption that the learner is aware of the independence/dependence between rewards and delays. The goal would be to design a single algorithm capable of accommodating both reward-independent and reward-dependent delays. Another interesting direction is to derive the matching lower bounds of reward regret and fairness regret for our algorithms.

Acknowledgements

This work was supported in part by National Key R&D Program of China under Grant 2022YFB2902700, NSF China (Grant No. 62202508, 62071501), and Shenzhen Science and Technology Program (Grant 20220817094427001, JCYJ20220818102011023, ZDSYS20210623091807023).

References

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

[2] O. Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1097–1105, 2014.

[3] W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159. PMLR, 2013.

[4] W. Chen, W. Hu, F. Li, J. Li, Y. Liu, and P. Lu. Combinatorial multi-armed bandit with general reward functions. *Advances in Neural Information Processing Systems*, 29, 2016.

[5] W. Chen, Y. Wang, Y. Yuan, and Q. Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778, 2016.

[6] Z. Chen, K. Cai, Z. Chen, J. Zhang, and J. C. S. Lui. Merit-based fair combinatorial semi-bandit with unrestricted feedback delays, 2024. URL https://arxiv.org/abs/2407.15439. Full version of this paper.

[7] M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, UAI 2011*, page 169, 2011.

[8] M. A. Gael, C. Vernade, A. Carpentier, and M. Valko. Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pages 3348–3356. PMLR, 2020.

[9] R. Gandhi, S. Khuller, S. Parthasarathy, and A. Srinivasan. Dependent rounding and its applications to approximation algorithms. *Journal of the ACM (JACM)*, 53(3):324–360, 2006.

[10] O. Jeunen and B. Goethals. Top-k contextual bandits with equity of exposure. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 310–320, 2021.

[11] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. *Advances in Neural Information Processing Systems*, 29, 2016.

[12] M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 158–163, 2018.

[13] P. Joulani, A. Gyorgy, and C. Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461. PMLR, 2013.

[14] J. Komiyama, J. Honda, and H. Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *International Conference on Machine Learning*, pages 1152–1161. PMLR, 2015.

[15] T. Lancewicki, S. Segal, T. Koren, and Y. Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In *International Conference on Machine Learning*, pages 5969–5978. PMLR, 2021.

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

[17] F. Li, J. Liu, and B. Ji. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7 (3):1799–1813, 2019.

[18] Y. Liu, G. Radanovic, C. Dimitrakakis, D. Mandal, and D. C. Parkes. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*, 2017.

[19] T. Mandel, Y.-E. Liu, E. Brunskill, and Z. Popović. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[20] M. Mansoury, B. Mobasher, and H. van Hoof. Exposure-aware recommendation using contextual bandits. In *5th FAccTRec Workshop: Responsible Recommendation*. Association for Computing Machinery (ACM), 2022.

[21] V. Patil, G. Ghalme, V. Nair, and Y. Narahari. Achieving fairness in the stochastic multi-armed bandit problem. *The Journal of Machine Learning Research*, 22(1):7885–7915, 2021.

[22] C. Pike-Burke, S. Agrawal, C. Szepesvari, and S. Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113. PMLR, 2018.

[23] C. Schumann, Z. Lang, N. Mattei, and J. P. Dickerson. Group fairness in bandits with biased feedback. In *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022*, 2022.

[24] A. Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

[25] J. Steiger, B. Li, and N. Lu. Learning from delayed semi-bandit feedback under strong fairness guarantees. In *IEEE Conference on Computer Communications (IEEE INFOCOM)*, pages 1379–1388. IEEE, 2022.

[26] M. Tallis and P. Yadav. Reacting to variations in product demand: An application for conversion rate (cr) prediction in sponsored search. In *IEEE International Conference on Big Data (Big Data)*, pages 1856–1864. IEEE, 2018.

[27] W. 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.

[28] C. Vernade, O. Cappé, and V. Perchet. Stochastic bandit models for delayed conversions. In *Conference on Uncertainty in Artificial Intelligence*, 2017.

[29] L. Wang, Y. Bai, W. Sun, and T. Joachims. Fairness of exposure in stochastic bandits. In *International Conference on Machine Learning*, pages 10686–10696. PMLR, 2021.

[30] Q. Wang and W. Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. *Advances in Neural Information Processing Systems*, 30, 2017.

[31] H. Wu and S. Wager. Thompson sampling with unrestricted delays. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 937–955, 2022.