The Fair Contextual Multi-Armed Bandit

Extended Abstract

Yifang Chen
University of Southern California
Los Angeles, CA
yifang@usc.edu

Alex Cuellar
University of Southern California
Los Angeles, CA
alexcuel@mit.edu

Haipeng Luo
University of Southern California
Los Angeles, CA
haipengl@usc.edu

Jignesh Modi
University of Southern California
Los Angeles, CA
jigneshm@usc.edu

Heramb Nemlekar
University of Southern California
Los Angeles, CA
nemlekar@usc.edu

Stefanos Nikolaidis
University of Southern California
Los Angeles, CA
nikolaid@usc.edu

ABSTRACT

When an AI system interacts with multiple users, it frequently needs to make allocation decisions. For instance, a virtual agent decides whom to pay attention to in a group setting, or a factory robot selects a worker to deliver a part. Demonstrating *fairness* in decision making is essential for such systems to be broadly accepted. We introduce a Multi-Armed Bandit algorithm with fairness constraints, where fairness is defined as a minimum rate that a task or a resource is assigned to a user. The proposed algorithm uses *contextual* information about the users and the task and makes no assumptions on how the losses capturing the performance of different users are generated. We view this as an exciting step towards including fairness constraints in resource allocation decisions.

ACM Reference Format:
Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. 2020. The Fair Contextual Multi-Armed Bandit. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION

We focus on the problem of an AI system assigning tasks or distributing resources to multiple humans, one at a time, while maximizing a given performance metric. For instance, a virtual agent decides whom to pay attention to in a group setting, or a factory robot selects a worker to deliver a part.

If there is clearly a user who outperforms everyone else, the solution to this optimization problem would result in the agent constantly selecting that user. This approach, however, fails to account that this may be perceived as unfair by others, which in turn may affect their acceptance of the system.

How can we integrate *fairness* in the agent's decisions? The aim of our work is to address this question. Recent works [5–7] have proposed multi-armed bandit algorithms for *fair* task allocation, where fairness is defined as a constraint on the minimum rate of arm selection. A user study on an online Tetris game, where the computer (player) selects users (arms) based on their score, has shown that users' trust is significantly improved when a fairness constraint is satisfied [5].

These works, however, have assumed that the performance of each user, observed in the form of a loss vector by the agent, follows a fixed distribution that is specific to that particular user. It thus fails to account that people may have different task-related skills. For instance, when making a pin, one worker may be specialized in cutting the wire, while another worker in measuring it. It also fails to account for cases where we can not make statistical assumptions about the generation of losses, for instance in an adversarial domain.

We generalize this work by proposing a fair multi-armed bandit algorithm that accounts for different *contexts* in task allocation. The algorithm also does not make any assumption on how the loss vector is generated, allowing for applications in non-stationary and even adversarial settings.

We provide theoretical guarantees on performance that show that the algorithm achieves regret equivalent to classic Follow The Regularized Leader (FTRL) algorithms [1].

2 PROBLEM DEFINITION

We study the online learning problem of contextual bandits (CB) with fairness constraints. We assume M possible contexts and K available actions (arms), and use the notation $[M]$ and $[K]$ to denote the set $\{1, \ldots, M\}$ and $\{1, \ldots, K\}$. For each time step $t = 1, ..., T$:

(1) The environment first decides the context $j_t \in [M]$ and the loss vector $l_t \in [0, 1]^K$.
(2) The learner observes the context $j_t \in [M]$ and selects the action $i_t \in [K]$.
(3) The learner suffers the loss $l_t(i_t)$.

We assume that the contexts $j_1, \ldots, j_T$ are i.i.d. samples of a fixed distribution $q \in \Delta_M$ which is known to the learner. However, we make no assumption on how the loss vectors $l_1, \ldots, l_T$ are generated, and in general l_t could depend on the entire history before round t, which is a key difference compared to previous work [5].

Let Δ_K be the set of distributions over K arms. Given the history up to the beginning of round t and that context j_t is j, we let $p_t^j \in \Delta_K$ be the conditional distributions of the player's selected arm i_t, for $j = 1, \ldots, M$. We require the following *fairness* constraint

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

parameterized by $v \in (0, 1/K)$:

$$\sum_{j=1}^{M} q(j) p_t^j(i) \geq v, \ \forall t, i, \tag{1}$$

that is, the marginal probability of each arm being pulled is at least v for each time.

For notational convenience, we denote a collection of M distributions over arms by $P = (p^1, ..., p^M)$ and the feasible set of these collections in terms of the above constraint by:

$$\Omega = \left\{ P = (p^1, ..., p^M) \middle| \begin{array}{l} p^1, ..., p^M \in \Delta_K \text{ and} \\ \sum_{j=1}^{M} q(j) p^j(i) \geq v, \forall i \in [K] \end{array} \right\}, \tag{2}$$

which is clearly a convex set and is non-empty since the uniform distribution (for all contexts) is always in the set.

The learner's goal is to minimize her regret, defined as the difference between her total loss and the loss of the best fixed distribution satisfying the fairness constraint:

$$\text{Reg} = \max_{P_* \in \Omega} \mathbb{E}\left[\sum_{t=1}^{T} \left\langle p_t^{j_t} - p_*^{j_t}, l_t \right\rangle \right].$$

Achieving sublinear regret $\text{Reg} = o(T)$ thus implies that in the long run the average performance of the learner is arbitrarily close to the best fixed distribution in hindsight.

3 ALGORITHM

Without the fairness constraint, there is no connection among the contexts and the optimal algorithm is just to run M instances of any standard MAB algorithm separately for each possible context. For example, classic FTRL algorithm would compute for each context $j \in [M]$:

$$p_t^j = \arg\min_{p \in \Delta_K} \sum_{s: j_s = j} \left\langle p, \hat{l}_s \right\rangle + \frac{1}{\eta} \sum_{i=1}^{K} \psi(p(i)) \tag{3}$$

at the beginning of round t, where $\psi : [0, 1] \to \mathbb{R}$ is some regularizer, $\eta > 0$ is some learning rate, and $\hat{l}$ is the standard unbiased importance-weighted estimator with:

$$\hat{l}_s(i) = \frac{l_s(i)}{p_s^{j_s}(i)} \mathbf{1}\{i_s = i\}, \ \forall i \in [K].$$

Upon observing the actual context j_t for round t, the algorithm then samples i_t from $p_t^{j_t}$. Standard results [3] show that the j-th instance of FTRL suffers regret $O(\sqrt{|\{t : j_t = j\}|K})$, and thus the total regret is $\sum_{j=1}^{M} O(\sqrt{|\{t : j_t = j\}|K}) = O(\sqrt{TMK})$ via the Cauchy-Schwarz inequality.

With the fairness constraint, however, we can no longer treat each context separately. A natural idea is to optimize jointly over the feasible set Ω defined in Eq. (2), that is, to find $P_t = (p_t^1, \cdots, p_t^M)$ at round t such that:

$$P_t = \arg\min_{P \in \Omega} \sum_{s=1}^{t-1} \left\langle p^{j_s}, \hat{l}_s \right\rangle + \frac{1}{\eta} \sum_{j=1}^{M} \sum_{i=1}^{K} \psi(p^j(i)).$$

It is clear that when $v = 0$ (that is, no fairness constraint), the feasible set Ω simply becomes $\Delta_K \times \cdots \times \Delta_K$ and the joint optimization above decomposes over j so that the algorithm degenerates to that

Algorithm 1 Fair CB with Known Context Distribution

Input: learning rate $\eta > 0$, fairness constraint parameter v
Define: $\Psi(P) = \frac{1}{\eta} \sum_{j=1}^{M} \sum_{i=1}^{K} \psi(p^j(i))$ where $\psi(p) = p \ln p$
for $t = 1, \ldots, T$ **do**
  Compute $P_t = \arg\min_{P \in \Omega} \sum_{s=1}^{t-1} \left\langle p^{j_s}, \hat{l}_s \right\rangle + \Psi(P)$
  Observe j_t and play $i_t \sim p_t^{j_t}$
  Construct loss estimator $\hat{l}_t(i) = \frac{l_t(i)}{p_t^{j_t}(i)} \mathbf{1}\{i_t = i\}, \ \forall i \in [K]$
end for

described in Eq. (3). When $v \neq 0$, the algorithm satisfies the fairness constraint automatically and can be seen as an instance of FTRL over a more complicated decision set Ω.

We deploy the standard entropy regularizer $\psi(p) = p \ln p$, used in the classic Exp3 algorithm [2] for MAB. See Algorithm 1 for the complete pseudocode. We remark that even though unlike Exp3, there is no closed form for computing P_t, one can apply any standard convex optimization toolbox to find P_t when implementing the algorithm.

We prove the following regret guarantee of our algorithm, which is essentially the same as the aforementioned bound for $v = 0$. The proof of the algorithm

Theorem 3.1. *With learning rate* $\eta = \sqrt{\frac{M \ln K}{TK}}$, *Algorithm 1 achieves*

$$Reg = O\left(\sqrt{TMK \ln K}\right).$$

The proof follows standard techniques (such as [1]) once we rewrite our algorithm as FTRL in the space of $\mathbb{R}^{MK}$. We provide the proof, as well as empirical results and user studies that show the effectiveness of the proposed algorithm, in an extended arxived version of this work [4].

4 DISCUSSION

Theoretically, we show how the classic FTRL framework can be naturally generalized to ensure fairness and we rigorously analyze the performance of the proposed algorithm in terms of regret guarantees. Designing AI systems that ensure and demonstrate fairness when interacting with people is critical to their acceptance. Beyond the theoretical results, we are excited to establish experimental foundations for fair decision making systems, which is still an under-served aspect in Human-AI Interaction.

REFERENCES

[1] Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. 2015. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems*. 2197–2205.
[2] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. 2002. The nonstochastic multiarmed bandit problem. *SIAM journal on computing* 32, 1 (2002), 48–77.
[3] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5, 1 (2012), 1–122.
[4] Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. 2019. Fair Contextual Multi-Armed Bandits: Theory and Experiments. *arXiv preprint arXiv:1912.08055* (2019).
[5] Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. 2019. Reinforcement Learning with Fairness Constraints for Resource Distribution in Human-Robot Teams. *arXiv preprint arXiv:1907.00313* (2019).

[6] Fengjiao Li, Jia Liu, and Bo Ji. 2019. Combinatorial sleeping bandits with fairness constraints. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 1702–1710.
[7] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y Narahari. 2019. Achieving Fairness in the Stochastic Multi-armed Bandit Problem. *arXiv preprint arXiv:1907.10516* (2019).