Fair Algorithms with Probing for Multi-Agent Multi-Armed Bandits

Tianyi Xu[1], **Jiaxin Liu**[2], **Nicholas Mattei**[1], **Zizhan Zheng**[1],

[1]Department of Computer Science, Tulane University, New Orleans, LA, USA
[2]Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign, Urbana, IL, USA
{txu9, nsmattei, zzheng3}@tulane.edu, jiaxin26@illinois.edu

Abstract

We propose a multi-agent multi-armed bandit (MA-MAB) framework to ensure fair outcomes across agents while maximizing overall system performance. For example, in a ridesharing setting where a central dispatcher assigns drivers to distinct geographic regions, utilitarian welfare (the sum of driver earnings) can be highly skewed—some drivers may receive no rides. We instead measure fairness by Nash social welfare, i.e., the product of individual rewards. A key challenge in this setting is decision-making under limited information about arm rewards (geographic regions). To address this, we introduce a novel probing mechanism that strategically gathers information about selected arms before assignment. In the offline setting, where reward distributions are known, we exploit submodularity to design a greedy probing algorithm with a constant-factor approximation guarantee. In the online setting, we develop a probing-based algorithm that achieves sublinear regret while preserving Nash social welfare. Extensive experiments on synthetic and real-world datasets demonstrate that our approach outperforms baseline methods in both fairness and efficiency.

Introduction

The multi-agent multi-armed bandit (MA-MAB) framework models a scenario where M agents compete for A arms over discrete rounds. In each round, a centralized decision-maker assigns each agent to an arm and observes the individual reward returned by each agent–arm pair, which can then be aggregated to measure total system performance. Typically this aggregation is measured by the utilitarian welfare, or sum of total rewards. One such application is ridesharing, where a central dispatcher assigns drivers (agents) to geographic regions (arms) based on estimated demand and driver availability. The dispatcher then observes the individual reward obtained by each driver in its assigned region.

Maximizing total expected reward is a common objective in the MA-MAB framework, but it can lead to inequalities in practice (Joseph et al. 2016). Optimizing for aggregate performance often results in a concentration of profitable arms among a few agents, disadvantaging others (Kleinberg et al. 2018; Agarwal et al. 2014). This is especially concerning in applications where fair resource assignment is essential, such as in ridesharing platforms, where drivers need equitable access to profitable areas (Bubeck, Cesa-Bianchi et al. 2012; Lattimore and Szepesvári 2020), and in content recommendation systems, where creator exposure should not be monopolized (Abdollahpouri et al. 2020).

Many MA-MAB studies aim to improve fairness, often by altering assignment strategies or adding constraints to naive utilitarian objectives (Joseph et al. 2018; Patil et al. 2021). A natural baseline is to maximize the sum of agent rewards (utilitarian welfare), but this can produce highly imbalanced assignments: profitable arms tend to be concentrated on a few agents while others receive little or nothing. This effect—where some agents are effectively "starved" of reward despite high aggregate system utility—is well documented in assignment and scheduling literature and is referred to as starvation. Maximizing total reward without regard to its distribution can therefore systematically disadvantage certain agents even when the system as a whole appears efficient. To address this, a line of work replaces or augments the sum objective with fairness-aware criteria such as Nash Social Welfare (NSW), which takes the product (equivalently the geometric mean) of individual utilities, thereby discouraging assignments that leave any agent with very low reward and yielding a balanced equity-efficiency trade-off (Heidari et al. 2018; Zhang and Conitzer 2021). In particular, Jones, Nguyen, and Nguyen (2023) demonstrated that optimizing NSW in MA-MAB settings can prevent persistent exclusion and achieve simultaneous improvements in fairness and overall utility.

However, a common limitation of existing approaches is their dependence on instantaneous reward feedback for updating estimates and steering assignment policies, even though in many real-world scenarios key information is either unobserved or noisy (Li, Karatzoglou, and Gentile 2016). For instance, in ridesharing platforms, uncertainty about passenger demand and fluctuating road conditions can corrupt the platform's estimates of driver-region rewards. Those corrupted estimates then propagate into the assignment step, potentially producing assignments that are unfair—for example, by repeatedly starving certain drivers who are misestimated as low-reward. To mitigate such risks, probing can be used to actively gather targeted information before making assignments (Oh and Iyengar 2019). Originating in economics (Weitzman 1979) as a cost-bearing

method to reduce uncertainty in sequential decision-making, probing has been adapted in domains such as database query optimization (Deshpande, Hellerstein, and Kletenik 2016; Liu et al. 2008), real-time traffic-aware vehicle routing (Bhaskara et al. 2020; Xu et al. 2025a), and wireless network scheduling (Xu et al. 2021; Xu, Zhang, and Zheng 2023).

Our work extends the NSW-based multi-agent multi-armed bandit (MA-MAB) framework by incorporating a probing mechanism to gather extra information, refining reward estimates and improving the exploration-exploitation balance while ensuring fairness. The decision-maker first probes a subset of arms for detailed reward data, then assigns agents fairly according to an objective that incorporates a fairness measure. In the offline setting, where reward distributions are known, we develop a greedy probing algorithm with a provable performance bound, leveraging the submodular structure of our objective. For the online setting, we propose a combinatorial bandit algorithm with a derived regret bound.

Integrating probing into MA-MAB poses the challenge of balancing exploration and exploitation while maintaining fairness. Related probing problems have been shown to be NP-hard (Goel, Guha, and Munagala 2006), and while previous works (Zuo, Zhang, and Joe-Wong 2020), have explored probing strategies in MA-MAB, they have simplified assumptions, such as limiting rewards to Bernoulli distributions and ignoring fair assignment. Our framework overcomes these limitations by considering general reward distributions, ensuring fairness, and introducing a probing budget to optimize performance under exploration constraints.

The primary contributions of our work are as follows:

(1) We extend the multi-agent MAB framework with a novel probing mechanism that tests selected arms before assignment. This approach ensures fairness through Nash Social Welfare optimization, departing from previous work that focuses solely on the sum of rewards (Zuo, Zhang, and Joe-Wong 2020).

(2) For known reward patterns in offline setting, we develop a simple yet effective greedy probing strategy with provable performance guarantees, while maintaining fairness across agents.

(3) For the online setting where rewards are unknown, we propose an algorithm that balances exploration and fairness, proving that probing and fair assignment do not compromise asymptotic performance.

(4) Experiments on both synthetic and real-world data demonstrate that our method achieves superior performance compared to baselines, validating the effectiveness of the probing strategy and our algorithm.

Related Work

The multi-armed bandit (MAB) framework has been key for sequential decision-making under uncertainty (Lai and Robbins 1985; Garivier and Cappé 2011; Kumar and Kleinberg 2000; Zhang et al. 2020). While early MAB models involve a single decision-maker choosing one arm per round, many real-world problems involve multiple agents acting simultaneously (Martínez-Rubio, Kanade, and Rebeschini 2019), each potentially choosing different arms (Hossain, Micha, and Shah 2021). Most existing multi-agent MAB methods focus on maximizing the total sum of rewards, which can unfairly favor some agents. To address this, researchers have proposed fairness-aware approaches (Joseph et al. 2016), with Nash Social Welfare (NSW) (Kaneko and Nakamura 1979) proving effective because it maximizes the product of agents' utilities.

In practice, key aspects of reward distributions are often unknown (Slivkins 2019; Lattimore and Szepesvári 2020). Existing fair MA-MAB methods generally rely on passive feedback (Liu and Zhao 2010; Gai and Krishnamachari 2014). Probing, or "active exploration," seeks extra information by testing a subset of arms before committing (Chen et al. 2015; Amin, Rostamizadeh, and Syed 2014), and is valuable when exploring poorly understood arms is risky (Golovin and Krause 2011; Bhaskara et al. 2020).

Recent single-agent studies formally introduce probing costs: Aaron D. Tucker et al. (2023) analyze bandits with costly reward observations, giving matching $\Theta(c^{1/3}T^{2/3})$ bounds; Eray Can Elumar, Cem Tekin, and Osman Yağan (2024) allow paying to probe one arm per round and achieve $\tilde{\mathcal{O}}(\sqrt{KT})$ regret; and Observe-Before-Play bandits (Zuo, Zhang, and Joe-Wong 2020) permit a limited number of pre-observations each round. Offline work has also examined submodular probing for routing problems (Bhaskara et al. 2020). However, the bulk of probing research still optimizes aggregate metrics—coverage, latency, or total reward—without incorporating inter-agent equity. This leaves a gap between fair multi-agent MAB methods that rely on passive feedback and probing strategies that ignore fairness. We close this gap by coupling cost-aware, submodular probe selection with NSW-oriented assignment, so information gained through active exploration translates directly into fair outcomes for all agents.

Problem Formulation

In this section, we define the fair multi-agent multi-armed bandit (MA-MAB) problem by extending the classical multi-armed bandit framework to incorporate fairness, multi-agent interactions, and the effect of probing decisions. The goal is to optimize both fairness and utility while accounting for probing overhead.

Agents, Arms, and Rewards

We consider a set of M *agents*, indexed by $[M] = \{1, \ldots, M\}$, and a set of A *arms*, indexed by $[A] = \{1, \ldots, A\}$. For every pair (j, a) with $j \in [M]$ and $a \in [A]$, let $D_{j,a}$ denote an unknown reward distribution with cumulative distribution function $F_{j,a}$ and mean $\mu_{j,a} \in [0, 1]$.

At each round $t \in [T]$, the decision-maker first selects a *probing set* $S_t \subseteq [A]$, receiving for each $j \in [M]$ and $a \in S_t$ a fresh reward $R_{j,a,t}$ drawn i.i.d. from $D_{j,a}$. If $a \notin S_t$, it instead relies on the current estimate $\mu_{j,a}$ (true mean in offline analysis). Using these observed rewards and estimates, the decision-maker then assigns each agent j to an arm $a_{j,t} \in [A]$.

Let $R_t = \{R_{j,a,t} \mid j \in [M], a \in S_t\}$ denote the set of rewards revealed by probing in round t. All rewards lie in $[0, 1]$.

Illustrative Mappings. (a) Ridesharing. Agents j are drivers; arms a are pickup zones obtained by a $0.01° \times 0.01°$ city grid. Before dispatching, the platform may probe a handful of zones (querying live app pings) to refine demand estimates, then assign drivers under a fairness objective. **(b) 60 GHz WLAN Scheduling.** Here agents are client devices, arms are access-points, and probing corresponds to brief beam-sounding measurements before scheduling transmissions.

Fairness Objective: Nash Social Welfare

To balance efficiency and equity, we maximize the (expected) *Nash Social Welfare* (NSW). An *assignment policy* at round t is a matrix $\pi_t = [\pi_{j,a,t}] \in [0, 1]^{M \times A}$ where $\pi_{j,a,t}$ is the probability that agent j receives arm a. Given S_t, the realized rewards R_t, and the mean matrix $\boldsymbol{\mu} = [\mu_{j,a}]$, we define the instantaneous NSW as

$$\mathrm{NSW}(S_t, R_t, \boldsymbol{\mu}, \pi_t) = \prod_{j \in [M]} \left(\sum_{a \in S_t} \pi_{j,a,t}\, R_{j,a,t} + \sum_{a \notin S_t} \pi_{j,a,t}\, \mu_{j,a} \right).$$

so each agent's utility contributes multiplicatively, discouraging assignments that leave any agent with a very low expected reward (e.g., a driver stranded without passengers).

The policy must satisfy

$$\sum_a \pi_{j,a,t} = 1 \quad (\forall\, j), \qquad \sum_j \pi_{j,a,t} \le 1 \quad (\forall\, a),$$

ensuring, in expectation, one arm per agent and no arm overbooked. As in Jones, Nguyen, and Nguyen (2023), $\pi_{j,a,t}$ may be fractional, reflecting randomized assignment commonly used in online platforms.

Why Nash Social Welfare? We adopt *Nash Social Welfare* (NSW) as our fairness objective for three key reasons: (i) **Pareto efficiency** —maximizing the geometric mean never sacrifices total reward when a Pareto-improvement is possible; (ii) **Scale invariance** — multiplying all utilities by the same constant leaves the maximizer unchanged, preventing bias due to units of measurement; (iii) **Balanced equity-efficiency trade-off** — NSW penalizes inequality more gently than max–min fairness while still discouraging highly skewed assignments, offering a smooth continuum between purely utilitarian and purely egalitarian objectives (Caragiannis et al. 2019; Thomson 2011).

Probing Overhead

Probing incurs a cost that increases with the size of the probing set. Since probing provides additional reward realizations but also consumes resources, we model the effective (instantaneous) reward at time t as follows:

$$\mathcal{R}_t^{\mathrm{total}} = \Big(1 - \alpha(|S_t|)\Big)\, \mathbb{E}_{R_t}\Big[\mathrm{NSW}(S_t, R_t, \boldsymbol{\mu}, \pi_t)\Big],$$

where $\alpha : \{0, 1, \ldots, I\} \to [0, 1]$ is a non-decreasing overhead function satisfying $\alpha(0) = 0$ and $\alpha(I) = 1$. Here, the expectation $\mathbb{E}_{R_t}[\cdot]$ is taken over the reward realizations R_t (with rewards drawn i.i.d. from the corresponding $D_{j,a}$).

This formulation accounts for the trade-off between exploration and exploitation. When more arms are probed, the decision-maker obtains more accurate information about reward distributions, leading to better assignment decisions. However, probing incurs costs, such as time delays, energy consumption, or computational overhead, which reduce the net benefit. For instance, in a wireless scheduling scenario, probing more channels provides better estimates of channel conditions but increases latency, reducing the system's effective throughput (Xu et al. 2021). This formulation is appropriate because the probing impact often scales with the system's overall performance, and similar scaling effects have been modeled in prior work (Xu et al. 2021, 2025a). The function $\alpha(|S_t|)$ captures this diminishing return, ensuring that excessive probing is discouraged.

Decision Problem

At each round t, the decision-maker proceeds in two stages:

Probing Stage: Select a probing set $S_t \subseteq [A]$ that balances exploration (gathering new information) and exploitation (leveraging current knowledge). Upon selecting S_t, the system probes the corresponding arms and obtains their reward realizations. That is, for each $a \in S_t$ and each agent $j \in [M]$, a reward $R_{j,a,t}$ is sampled i.i.d. from the distribution $D_{j,a}$.

Assignment Stage: Based on the observed rewards R_t for arms in S_t (and the known mean rewards $\mu_{j,a}$ for arms not in S_t), choose a probabilistic assignment policy $\pi_t \in \Delta^A$ to assign arms to agents.

The decision-maker's goal is to select (S_t, π_t) in each round so as to maximize $\mathcal{R}_t^{\mathrm{total}}$ and thereby achieve sublinear cumulative regret.

Regret and Performance Measure

To evaluate the performance of our online learning approach, we define regret by comparing the achieved reward with the optimal reward obtained in an offline setting.

In the offline setting, the decision-maker has full knowledge of $\boldsymbol{\mu}$ and can directly compute the optimal probing set and assignment policy (S_t^*, π_t^*) that maximizes the expected NSW objective:

$$(S_t^*, \pi_t^*) = \arg\max_{S, \pi} \Big(1 - \alpha(|S|)\Big)\, \mathbb{E}_{R_t}\Big[\mathrm{NSW}(S, R_t, \boldsymbol{\mu}, \pi)\Big].$$

This serves as a performance benchmark.

In contrast, the online setting requires the decision-maker to learn $\boldsymbol{\mu}$ over time while making sequential decisions based on observed rewards. The cumulative regret measures the performance gap between the online strategy and the offline optimal policy:

$$\mathcal{R}_{\mathrm{regret}}(T) = \sum_{t=1}^{T} \Big[\Big(1 - \alpha\big(|S_t^*|\big)\Big) \cdot\, \mathbb{E}_{R_t}\Big[\mathrm{NSW}(S_t^*, R_t, \boldsymbol{\mu}, \pi_t^*)\Big] \;-\; \mathcal{R}_t^{\mathrm{total}} \Big]. \tag{1}$$

Efficient algorithms aim to ensure that $\mathcal{R}_{\text{regret}}(T)$ grows sublinearly with T, thereby balancing fairness, overall utility, and the cost of probing.

The Offline Setting

In the offline setting, all reward distributions are known in advance, reducing the problem to a static optimization over the probing set S without the time index t. Given any fixed S, the optimal assignment policy π (Cole and Gkatzelis 2015) can be computed, allowing the effective reward to be expressed as a function of S.

This optimization is computationally challenging, as similar problems have been shown to be NP-hard (Goel, Guha, and Munagala 2006). To address this, we develop a greedy algorithm based on submodular maximization techniques to obtain an approximate solution while accounting for probing costs.

Optimization Objective and Probing Utility

Since the offline setting is static, we drop the time index t. The decision-maker is now tasked with selecting a probing set $S \subseteq [A]$ from the A available arms. For each agent $j \in [M]$ and each arm $a \in [A]$, let $D_{j,a}$ denote the known reward distribution with CDF $F_{j,a}$ and mean $\mu_{j,a}$. When an arm a is probed (i.e. $a \in S$), a reward realization $R_{j,a}$ is observed (drawn i.i.d. from $D_{j,a}$); otherwise, the decision-maker relies on the mean reward $\mu_{j,a}$.

Given a probing set S, an assignment policy $\pi = [\pi_{j,a}]$ can be applied to assign arms to agents. For any fixed S, one may compute the optimal assignment policy that maximizes the expected Nash Social Welfare (Cole and Gkatzelis 2015). Hence, we define the effective objective as

$$\mathcal{R}(S) = \Big(1 - \alpha(|S|)\Big) \cdot \mathbb{E}_R\Big[\text{NSW}(S, R, \boldsymbol{\mu}, \pi^*(S))\Big],$$

$\pi^*(S)$ denotes the optimal assignment policy given S, and $\alpha(|S|)$ is the probing overhead function.

Directly optimizing $\mathcal{R}(S)$ is challenging due to both the combinatorial nature of the set S and the multiplicative form of NSW. To address this, we decompose $\mathcal{R}(S)$ into two components.

Defining a Simplified Utility $g(S)$. To isolate the contribution of probed arms and simplify the multiplicative structure, we define

$$g(S) \;=\; \max_{\pi \in \Delta^A_S} \mathbb{E}\Bigg[\prod_{j\in[M]} \Big(\sum_{a\in S} \pi_{j,a}\, R_{j,a}\Big)\Bigg],$$

where

$$\Delta^A_S = \Big\{\pi \in \mathbb{R}_+^{M\times A} \,|\pi_{j,a} = 0, \quad \forall a \notin S,\ \forall j \in [M], \sum_{a\in S} \pi_{j,a} \le 1, \quad \forall j \in [M]\Big\}.$$

Since only arms in S yield random rewards, this formulation ensures that $g(S)$ is naturally monotonic in S.

Defining the Non-probing Utility $h(S)$. Complementarily, we define the non-probing utility as

$$h(S) \;=\; \max_{\pi \in \Delta^A_{[A]\setminus S}} \prod_{j\in[M]} \Big(\sum_{a\notin S} \pi_{j,a}\, \mu_{j,a}\Big),$$

where

$$\Delta^A_{[A]\setminus S} = \Big\{\pi \in \mathbb{R}_+^{M\times A} \,|\pi_{j,a} = 0, \quad \forall a \in S,\ \forall j \in [M], \sum_{a\in [A]\setminus S} \pi_{j,a} \le 1, \quad \forall j \in [M]\Big\}.$$

This formulation captures the baseline utility achievable by assigning exclusively among the unprobed arms. Similar to $g(S)$, the assignment policy ensures that each agent's total assignment probability does not exceed one, allowing partial assignments across multiple arms.

Log Transformation and Piecewise-Linear Approximation. Taking logarithms converts the product $\prod_j(\dots)$ into the sum $\log g(S)$. Unfortunately, the resulting set function is still *non-additive*—its marginal gain depends on the current value $g(S)$—and generally non-submodular, so classical greedy guarantees no longer apply. Even a much simpler problem is already intractable: selecting at most k items to maximise a strictly concave, increasing function of their total weight is NP-hard to approximate within any constant factor (Ahmed and Atamtürk 2011). Consequently, directly optimising $\log g(S)$ is computationally challenging. To regain tractability, we adopt the classical idea of piecewise-linear upper envelopes for concave functions(Salhi 1994; Tawarmalani and Sahinidis 2013). Specifically, we construct a piecewise-linear, concave, and non-decreasing function $\phi : [0, x_{\max}] \to \mathbb{R}$ such that for all $x > 0$, we have $\phi(x) \ge \log(x)$. We then define $f_{\text{upper}}(S) = \phi\big(g(S)\big)$. This construction yields an objective that (i) upper-bounds $\log\big(g(S)\big)$ and (ii) exhibits diminishing returns, as explained next.

Piecewise-Linear Upper Bound Construction

To approximate $\log\big(g(S)\big)$ in a tractable manner, we construct ϕ as follows:

- **Upper Bound Assumption:** Assume that $x_{\max}$ is an upper bound on $g(S)$, i.e., for all S, $g(S) \le x_{\max}$.
- **Partitioning:** Partition the interval $[0, x_{\max}]$ into segments with breakpoints $\tau_0, \tau_1, \dots, \tau_L$ (with $0 < \tau_0 < \tau_1 < \cdots < \tau_L = x_{\max}$). Moreover, choose the partition sufficiently fine so that there exists a constant $\eta > 0$ (with η sufficiently small) such that $\tau_{i+1} - \tau_i \le \eta$ for all i, and, importantly, for any $S \subseteq [A]$ and any arm $a \notin S$, $g(S \cup \{a\}) - g(S) \le \eta$. This guarantees that for every S and a, both $g(S)$ and $g(S \cup \{a\})$ lie within the same linear segment of ϕ.
- **Tangent Lines:** For each breakpoint τ_i, define the tangent line $T_{\tau_i}(z) = \log(\tau_i) + \frac{1}{\tau_i}(z - \tau_i)$. By the concavity of $\log(\cdot)$, for every $z \in [\tau_i, \tau_{i+1}]$ we have $\log(z) \le T_{\tau_i}(z)$.

We then define $\phi(z) = \max_{0 \le i < L} T_{\tau_i}(z)$, so that for all $z > 0$, $\phi(z) \ge \log(z)$. By construction, ϕ is concave and non-decreasing. In particular, if $x \in [\tau_i, \tau_{i+1}]$ and $y \in [\tau_j, \tau_{j+1}]$ with $x < y$ (so that $\tau_i \le \tau_j$), the slopes on these segments are given by $1/\tau_i$ and $1/\tau_j$, respectively. Since $\tau_i < \tau_j$, we have $\frac{1}{\tau_i} \ge \frac{1}{\tau_j}$. Thus, for any $0 < x < y \le x_{\max}$ and any increments $d, d' \ge 0$ (with $d, d' \le \eta$ so that $x + d$ and $y + d'$ lie within single linear segments), it holds that

$$\frac{\phi(x+d) - \phi(x)}{d} = \frac{1}{\tau_i} \ge \frac{1}{\tau_j} = \frac{\phi(y+d') - \phi(y)}{d'}.$$

Additional Assumption. In order to handle the case when the increments differ (i.e., when $d < d'$), we assume (see Appendix for a detailed proof) that the partition is sufficiently fine so that if $x \in [\tau_i, \tau_{i+1}]$ and $y \in [\tau_j, \tau_{j+1}]$ with $\tau_i \le \tau_j$, then $\frac{d}{d'} \ge \frac{\tau_i}{\tau_j}$. In addition, for analytical convenience, we view the offline benchmark through the allocation classes associated with probed and unprobed arms.

Finally, define the composed function $f_{\text{upper}}(S) = \phi(g(S))$. In Lemma 3, we show that $f_{\text{upper}}(S)$ is submodular, which enables the approximation guarantee of our greedy probing strategy. The proof is provided in the Appendix in the full version (Xu et al. 2025b).

Submodularity Properties

Utilizing the construction above, we establish the following key properties (proofs are in the Appendix). These lemmas are fundamental for proving Theorem 1, as they help establish the submodularity and approximation guarantees of Algorithm 1.

Lemma 1 (Monotonicity of $g(S)$)**.** *For any $S \subseteq T \subseteq [A]$, we have $g(S) \le g(T)$.*

Lemma 2 (Monotonicity)**.** *For any $S \subseteq T \subseteq [A]$, we have $f_{upper}(S) \le f_{upper}(T)$.*

Lemma 3 (Submodularity)**.** *For any $S \subseteq T \subseteq [A]$ and $a \notin T$,*

$$f_{upper}(S \cup \{a\}) - f_{upper}(S) \;\ge\; f_{upper}(T \cup \{a\}) - f_{upper}(T).$$

Lemma 4 (Monotonicity of $h(S)$)**.** *For any $S \subseteq T \subseteq [A]$, we have $h(S) \ge h(T)$.*

Lemma 5. *For any probing set S, we have:*

$$\begin{aligned}\mathcal{R}(S) &= (1 - \alpha(|S|))\mathbb{E}_R\Big[\mathrm{NSW}(S, R, \boldsymbol{\mu}, \pi^*(S))\Big] \\ &\le (1 - \alpha(|S|))(g(S) + h(S)).\end{aligned}$$

Greedy Algorithm and Approximation

Finally, we can greedily pick arms one by one to maximize the incremental gain in $f_{\text{upper}}(S)$, subject to budget I. By standard results on submodular maximization with cardinality constraints (Iyer and Bilmes 2013), this yields a $(1-1/e)$-approximation for maximizing f_{upper}. Moreover, we can combine it with the overhead term $(1 - \alpha(|S|))$ to trade off between total payoff and probing cost.

Algorithm 1 first initializes $I+1$ empty probing sets $S_0, S_1, \ldots, S_I$ (lines 1–2). In each iteration i (lines 3–5), it selects the arm a that maximizes the marginal gain $f_{\text{upper}}(S_{i-1} \cup \{a\}) - f_{\text{upper}}(S_{i-1})$ and updates S_i. The candidate sets S_j are then sorted by the adjusted reward $(1 - \alpha(|S_j|))f_{\text{upper}}(S_j)$ (line 7). The algorithm iterates through these sets (lines 8–14), returning an empty set if the highest-ranked S_j does not exceed $h(\emptyset)$ (lines 9–10). Otherwise, it computes the expected reward $\mathcal{R}(S_j)$ (line 11) and proceeds to the next candidate if $(1 - \alpha(|S_j|))f_{\text{upper}}(S_j)$ exceeds $\zeta\mathcal{R}(S_j)$ (lines 11–12). If all conditions are met, S_j is returned as the final probing set S^{pr} (lines 14).

Algorithm 1: Offline Greedy Probing

Input: $\{F_{m,a}\}_{m \in [M],\, a \in [A]},\ \alpha(\cdot),\ I,\ \zeta \ge 1.$
Output: S^{pr}

1: **for** $i = 0$ to I **do**
2: $\quad S_i \leftarrow \emptyset$
3: **for** $i = 1$ to I **do**
4: $\quad a \leftarrow \arg\max_{a \in [A] \setminus S_{i-1}} \begin{bmatrix} f_{\text{upper}}(S_{i-1} \cup \{a\}) \\ - f_{\text{upper}}(S_{i-1}) \end{bmatrix}$
5: $\quad S_i \leftarrow S_{i-1} \cup \{a\}$
6: $\Pi \leftarrow \{0, 1, \ldots, I\}$
7: **sort** Π so that if i precedes j, then $(1 - \alpha(|S_i|))\, f_{\text{upper}}(S_i) \ge (1 - \alpha(|S_j|))\, f_{\text{upper}}(S_j)$
8: **for** each j in Π (largest to smallest upper-bound) **do**
9: $\quad$ **if** $(1 - \alpha(|S_j|))\, f_{\text{upper}}(S_j) < h(\emptyset)$ **then**
10: $\quad\quad S^{\text{pr}} \leftarrow \emptyset$; **return** S^{pr}
11: $\quad$ **if** $(1 - \alpha(|S_j|))\, f_{\text{upper}}(S_j) > \zeta\mathcal{R}(S_j)$ **then**
12: $\quad\quad$ **continue**
13: $\quad$ **else**
14: $\quad\quad S^{\text{pr}} \leftarrow S_j$; **return** S^{pr}

Theorem 1. *Let S^* be an optimal subset maximizing $\mathcal{R}(S)$. Then the set S^{pr} returned by Algorithm 1, for any $\zeta \ge 1$, satisfies*

$$\mathcal{R}(S^{\text{pr}}) \;\ge\; \frac{e-1}{2e-1}\,\frac{1}{\zeta}\mathcal{R}(S^*).$$

Theorem 1 provides the main theoretical guarantee of our work, with a detailed proof in the Appendix. Our algorithm effectively balances exploring additional arms and probing costs, ensuring a near-optimal reward in the offline setting.

The Online Setting

In the online setting, we consider a system with M agents and A arms over T rounds. Unlike the offline setting where rewards are known, here they are unknown, requiring a balance between exploration and exploitation.

The Online Fair Multi-Agent UCB with Probing (OFMUP) algorithm (Algorithm 2) maintains empirical statistics for each agent–arm pair (j, a), including an empirical CDF estimate $\widehat{F}_{j,a}$, and constructs upper confidence bounds (UCBs). Our **key contribution** in the online setting is integrating Algorithm 1 into the online framework and designing a **novel UCB-based** strategy. This enables efficient learning while ensuring fairness across agents. The procedure executes the following steps:

1. Initialization (Lines 1–4). Each agent–arm pair (j, a) starts with: $\widehat{F}_{j,a,t} \leftarrow 1$, $N_{j,a,t} \leftarrow 0$, $\hat{\mu}_{j,a,t} \leftarrow 0$, $w_{j,a,t} \leftarrow 0$. These serve as optimistic estimates before data collection. The confidence bound $U_{j,a,t}$ is defined later.

2. Warm-Start Rounds (Lines 5–10). For the first MA rounds, each agent–arm pair is explored at least once under assignment constraints. The selection follows:

$$(\mathbf{e}_t)_{j,a} = \begin{cases} 1, & \text{if } j = j_t \text{ and } a = a_t, \\ 0, & \text{otherwise,} \end{cases}$$

This ensures each agent samples all arms and each arm is probed multiple times.

3. Main Loop (Lines 11–16). For $t > MA$, the algorithm iterates as follows:

a. Probe Set Selection (Line 12).

$$S_t \leftarrow \textsc{Algorithm 1}\big(\widehat{F}_{j,a,t},\, \alpha(\cdot),\, I, \zeta\big),$$

where I is the probing budget and $\alpha(\cdot)$ the overhead function. This subroutine greedily selects $S_t \subseteq [A]$ based on $\widehat{F}_{j,a,t}$.

b. Probing and Updates (Line 13). Each arm in S_t is probed, revealing rewards $R_{j,a,t}$, and updating $\widehat{F}_{j,a,t}$, $N_{j,a,t}$, $\hat{\mu}_{j,a,t}$, and $w_{j,a,t}$. The confidence bound is:

$$U_{j,a,t} = \min(\hat{\mu}_{j,a,t} + w_{j,a,t}, 1).$$

c. Policy Optimization (Line 14). The optimal policy is:

$$\pi_t \leftarrow \arg\max_{\pi_t \in \Delta^A} (1 - \alpha(|S_t|)) \cdot \mathbb{E}_{R_t}[\mathrm{NSW}(S_t, R_t, U_t, \pi_t)],$$

where $\mathrm{NSW}(\cdot)$ is the Nash social welfare objective.

d. Arm Pulls and Final Updates (Lines 15–16). Each agent j pulls $a_{j,t} \sim \pi_t$, observes $R_{j,a_{j,t},t}$, and updates $\widehat{F}_{j,a_{j,t}}$, $N_{j,a_{j,t}}$, $\hat{\mu}_{j,a_{j,t}}$, $w_{j,a_{j,t}}$, and $U_{j,a_{j,t}}$.

Analysis

In this section, we present our regret analysis for the proposed OFMUP algorithm. Detailed proofs and Lemma 8 are deferred to the Appendix.

Smoothness of the NSW Objective

Lemma 6 (Smoothness of the NSW Objective)**.** *Let $\mu, \nu \in [0,1]^{M \times A}$ be two reward matrices. For any probing set $S \subseteq [A]$ and observed rewards R, we have*

$$\Big|\mathrm{NSW}(S, R, \mu, \pi) - \mathrm{NSW}(S, R, \nu, \pi)\Big| \leq \sum_{j=1}^{M} \sum_{a=1}^{A} \pi_{j,a} \Big|\mu_{j,a} - \nu_{j,a}\Big|.$$

Concentration of Reward Estimates

Lemma 7 (Concentration of Reward Estimates)**.** *Let $\delta \in (0,1)$. Then with probability at least $1 - \frac{\delta}{2}$, for all $t > A$, $a \in [A]$, and $j \in [M]$, we have*

$$\Big|\mu_{j,a} - \hat{\mu}_{j,a,t}\Big| \leq \sqrt{\frac{2(\hat{\mu}_{j,a,t} - \hat{\mu}^2_{j,a,t}) \ln\!\big(\frac{2MAT}{\delta}\big)}{N_{j,a,t}}} + \frac{\ln\!\Big(\frac{2\,M\,A\,T}{\delta}\Big)}{3\,N_{j,a,t}} = w_{j,a,t}$$

Algorithm 2: Online Fair Multi-Agent UCB with Probing (OFMUP)

Input: $A, M, T, I, c, \alpha(\cdot), \Delta$ etc.

1: **Initialize**:
2: **for** $j = 1 \to M,\ a = 1 \to A$ **do**
3: $\quad \widehat{F}_{j,a,t} \leftarrow 1, \quad N_{j,a,t} \leftarrow 0, \quad \hat{\mu}_{j,a,t} \leftarrow 0,$
4: $\quad w_{j,a,t} \leftarrow 0$
5: **for** $t = 1 \to MA$ **do**
6: $\quad j_t \leftarrow ((t-1) \mod M) + 1$
7: $\quad a_t \leftarrow ((t-1)/M) + 1$
8: $\quad \pi_t \leftarrow \mathbf{e}_t, \quad S_t \leftarrow \{a_t\}$
9: $\quad$ **agent** j_t **pulls arm** a_t
10: $\quad$ **observe** $R_{j_t,a_t,t}$, **update** $\widehat{F}_{j_t,a_t,t}$, $\hat{\mu}_{j_t,a_t,t}$, $w_{j_t,a_t,t}$, $N_{j_t,a_t,t}$
11: **for** $t = MA + 1 \to T$ **do**
12: $\quad S_t \leftarrow \textsc{Algorithm 1}\big(\widehat{F}_{j,a,t}, \alpha(\cdot), I, \zeta\big)$
13: $\quad$ **probe each arm in** S_t, **observe** $R_{j,a,t}$, **update** $\widehat{F}_{j,a,t}, N_{j,a,t}, \hat{\mu}_{j,a,t}, w_{j,a,t}$, $U_{j,a,t} = \min(\hat{\mu}_{j,a,t} + w_{j,a,t}, 1)$
14: $\quad \pi_t \leftarrow \arg\max_{\pi_t \in \Delta^A} \Big[(1 - \alpha(|S_t|)) \cdot \mathbb{E}_{R_t}\big[\mathrm{NSW}(S_t, R_t, U_t, \pi_t)\big]\Big]$
15: $\quad$ **each agent** j **pulls** $a_{j,t} \sim \pi_t$, **observe** $R_{j,a_{j,t},t}$
16: $\quad$ **update:** $\widehat{F}_{j,a_{j,t}}, N_{j,a_{j,t},t}, \hat{\mu}_{j,a_{j,t},t}, w_{j,a_{j,t},t}, U_{j,a_{j,t},t}$

Main Regret Guarantee

Combining the smoothness of the NSW objective, the concentration bounds, and the auxiliary technical results, we obtain our main regret guarantee.

Theorem 2. *For any $\delta \in (0,1)$, with probability at least $1 - \delta$, the cumulative regret $\mathcal{R}_{\mathrm{regret}}(T)$ of the Online Fair multi-Agent UCB with Probing algorithm (Algorithm 2) satisfies*

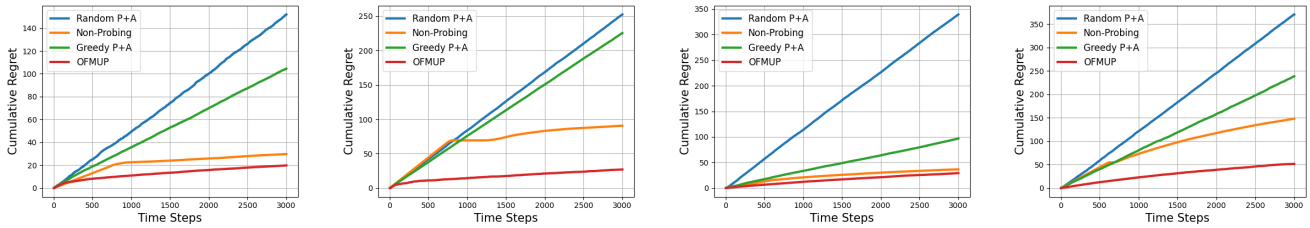
$$\mathcal{R}_{\mathrm{regret}}(T) \;=\; O\Big(\zeta\,\big(\sqrt{M\,A\,T} \;+\; M\,A\big)\,\ln^{c}\!\Big(\tfrac{M\,A\,T}{\delta}\Big)\Big),$$

for some constant $c > 0$.

Theorem 2 shows that our probing-based algorithm achieves sublinear regret with a provable constant-factor improvement over the non-probing baseline. This gain arises from faster elimination of suboptimal arms and more informed assignment decisions. Experimental results confirm that probing consistently lowers regret across all time horizons. Full proofs of Lemmas 6–8 and Theorem 2 are provided in the Appendix.

Experiments

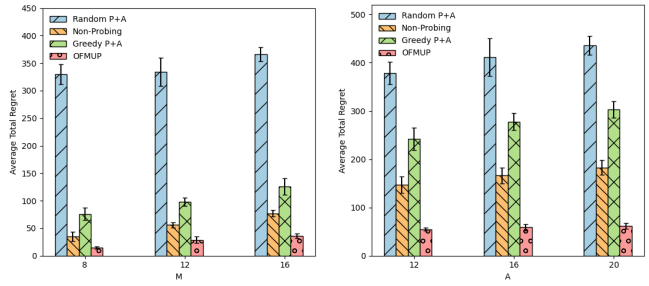
We evaluate our framework using controlled simulations and a real-world ridesharing case study. In our simulated environment, we consider a multi-agent multi-armed bandit (MA-MAB) setting with M agents and A arms (e.g., $M = 12, A = 8$ for small-scale and $M = 20, A = 10$ for large-scale scenarios). For each agent–arm pair (j, a), rewards are generated as i.i.d. samples from a fixed distribution $D_{j,a}$ with mean $\mu_{j,a}$. In the real-world setting, we apply our framework to the NYYellowTaxi 2016



(a) (b) (c) (d)

Figure 1: (a): Agents number $M = 12$, arms number $A = 8$, Bernoulli distribution for reward. (b): Agents number $M = 20$, arms number $A = 10$, Bernoulli distribution for reward. (c): Agents number $M = 12$, arms number $A = 8$, General distribution for reward. (d): Agents number $M = 20$, arms number $A = 10$, General distribution for reward. Data from NYYellowTaxi 2016.

dataset (Shah, Lowalekar, and Varakantham 2020), treating vehicles as agents and discretized pickup locations (binned into 0.01° grids) as arms. Rewards are determined by the normalized Manhattan distance between vehicles and pickup points—closer distances yield higher rewards. Vehicle locations are randomly pre-sampled within city bounds and remain fixed, underscoring the practical effectiveness of our approach.



(a) (b)

Figure 2: Scalability analysis across two dimensions: (a) Fixed arms number $A = 8$ with varying agents; (b) Fixed agents number $M = 20$ with varying arms.

To test different conditions, we consider two reward distributions: a Bernoulli distribution (with rewards 0 or 1 and with mean rewards μ in $[0.3, 0.8]$) and a discrete distribution (with rewards sampled from $\{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$). Cumulative regret is computed as in Equation (1), with the optimal reward determined via exhaustive search. For numerical stability, cumulative Nash social welfare is aggregated using the geometric mean of per-agent rewards. We further verify that our offline objective $f_{\text{upper}}(S)$ closely approximates $\log(g(S))$ (difference 0.03), confirming the validity of our formulation. All experimental settings satisfy the assumptions outlined in our paper.

Baselines

Three algorithms serve as our comparison baselines: **Non-Probing**, a fair MAB algorithm from Jones, Nguyen, and Nguyen (2023) without probing capability, focusing only on optimal assignment with current information; **Random Probing with Random assignment (Random P+A)**, which randomly selects a fixed number of arms for probing and then assigns the arms randomly; and **Greedy Probing with Random assignment (Greedy P+A)**, which uses the same greedy probing strategy as our algorithm but performs random assignment after probing.

Results

Figure 1 shows OFMUP's performance across different tests. In small-scale tests (M = 12, A = 8), OFMUP reduces regret by 85% vs. Random P+A and 60% vs. Greedy P+A after 3000 steps. The advantage is greater in medium-scale tests (M = 20, A = 10), where regret is 88% lower than Random P+A and 80% lower than Greedy P+A at T=3000, showing improved scalability. For discrete rewards, OFMUP continues to outperform. The gap with other methods grows as it learns reward patterns, reducing regret by 85% vs. Random P+A and 65% vs. Non-Probing at T=3000. Notably, Random P+A's regret is slightly higher in small-scale tests due to fairness computation via the geometric mean and increased variability with smaller samples. These results validate our theoretical analysis—OFMUP effectively balances exploration and exploitation while ensuring fairness. The experiments further confirm the crucial role of probing in gathering information and guiding assignment. We further examine scalability by independently varying agent numbers (Figure 2). OFMUP performs well in all our tests, and its advantage over the other methods grows as the problems become more complex.

Conclusion

We propose a fair MA-MAB framework combining selective probing with Nash Social Welfare objectives for efficient information gathering and equitable assignments. Our offline greedy algorithm attains a constant-factor approximation bound and the online extension guarantees sublinear regret. Experiments on synthetic and real-world ridesharing data show refined reward estimates, improved exploration–exploitation balance, and superior performance over baselines, demonstrating practical applicability.

Acknowledgments

Mattei was supported in part by NSF Awards IIS-III-2107505, IIS-RI-2134857, IIS-RI-2339880, and CNS-SCC-2427237, as well as the Harold L. and Heather E. Jurist Center of Excellence for Artificial Intelligence at Tulane University and the Tulane University Center of Excellence for Community-Engaged Artificial Intelligence (CEAI). Zheng was supported in part by NSF Award CNS-2146548 and Louisiana BOR-RCS award 088A-25.

References

Aaron D. Tucker; Caleb Biddulph; Claire Wang; and Thorsten Joachims. 2023. Bandits with Costly Reward Observations. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2147–2156. PMLR.

Abdollahpouri, H.; Adomavicius, G.; Burke, R.; Guy, I.; Jannach, D.; Kamishima, T.; Krasnodebski, J.; and Pizzato, L. 2020. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, 30(1): 127–158.

Agarwal, A.; Hsu, D.; Kale, S.; Langford, J.; Li, L.; and Schapire, R. 2014. Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits. In *Proceedings of the 31st International Conference on Machine Learning*, 1638–1646.

Ahmed, S.; and Atamtürk, A. 2011. Maximizing a Class of Submodular Utility Functions. *Mathematical Programming*, 128(1): 149–169.

Amin, K.; Rostamizadeh, A.; and Syed, U. 2014. Repeated Contextual Auctions with Strategic Buyers. In *Advances in Neural Information Processing Systems*.

Bhaskara, A.; Gollapudi, S.; Kollias, K.; and Munagala, K. 2020. Adaptive probing policies for shortest path routing. In *Advances in Neural Information Processing Systems*.

Bubeck, S.; Cesa-Bianchi, N.; et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1): 1–122.

Caragiannis, I.; Kurokawa, D.; Moulin, H.; Procaccia, A. D.; Shah, N.; and Wang, J. 2019. The unreasonable fairness of maximum Nash welfare. *ACM Transactions on Economics and Computation (TEAC)*, 7(3): 1–32.

Chen, Y.; Javdani, S.; Karbasi, A.; Bagnell, J.; Srinivasa, S.; and Krause, A. 2015. Submodular Surrogates for Value of Information. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Cole, R.; and Gkatzelis, V. 2015. Approximating the Nash social welfare with indivisible items. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, 371–380.

Deshpande, A.; Hellerstein, L.; and Kletenik, D. 2016. Approximation algorithms for stochastic submodular set cover with applications to boolean function evaluation and min-knapsack. *ACM Transactions on Algorithms (TALG)*, 12(3): 1–28.

Eray Can Elumar; Cem Tekin; and Osman Yağan. 2024. Multi-Armed Bandits with Costly Probes. *IEEE Transactions on Information Theory*. Early Access.

Gai, Y.; and Krishnamachari, B. 2014. Distributed Stochastic Online Learning Policies for Opportunistic Spectrum Access. *IEEE Transactions on Signal Processing*, 6184–6193.

Garivier, A.; and Cappé, O. 2011. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, 359–376. JMLR Workshop and Conference Proceedings.

Goel, A.; Guha, S.; and Munagala, K. 2006. Asking the right questions: Model-driven optimization using probes. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 203–212.

Golovin, D.; and Krause, A. 2011. Adaptive Submodularity: Theory and Applications in Active Learning and Stochastic Optimization. *Journal of Artificial Intelligence Research*, 427–486.

Heidari, H.; Ferrari, C.; Gummadi, K.; and Krause, A. 2018. Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. In *Advances in Neural Information Processing Systems*.

Hossain, S.; Micha, E.; and Shah, N. 2021. Fair Algorithms for Multi-Agent Multi-Armed Bandits. In *Proceedings of Advances in Neural Information Processing Systems*, 24005–24017.

Iyer, R. K.; and Bilmes, J. A. 2013. Submodular optimization with submodular cover and submodular knapsack constraints. *Advances in neural information processing systems*.

Jones, M.; Nguyen, H.; and Nguyen, T. 2023. An efficient algorithm for fair multi-agent multi-armed bandit with low regret. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8159–8167.

Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2018. Meritocratic Fairness for Infinite and Contextual Bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 158–163.

Joseph, M.; Kearns, M.; Morgenstern, J. H.; and Roth, A. 2016. Fairness in Learning: Classic and Contextual Bandits. In *Proceedings of Advances in Neural Information Processing Systems*.

Kaneko, M.; and Nakamura, K. 1979. The Nash Social Welfare Function. *Econometrica*, 423–435.

Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Rambachan, A. 2018. Algorithmic Fairness. *AEA Papers and Proceedings*.

Kumar, A.; and Kleinberg, J. 2000. Fairness measures for resource allocation. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, 75–85.

Lai, T. L.; and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 4–22.

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

Li, S.; Karatzoglou, A.; and Gentile, C. 2016. Collaborative Filtering Bandits. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 539–548.

Liu, K.; and Zhao, Q. 2010. Distributed Learning in Multi-Armed Bandit With Multiple Players. *IEEE Transactions on Signal Processing*, 5667–5681.

Liu, Z.; Parthasarathy, S.; Ranganathan, A.; and Yang, H. 2008. Near-optimal algorithms for shared filter evaluation in data stream systems. In *ACM SIGMOD*.

Martínez-Rubio, D.; Kanade, V.; and Rebeschini, P. 2019. Decentralized Cooperative Stochastic Bandits. In *Advances in Neural Information Processing Systems*.

Oh, M.-h.; and Iyengar, G. 2019. Thompson Sampling for Multinomial Logit Contextual Bandits. In *Advances in Neural Information Processing Systems*.

Patil, V.; Ghalme, G.; Nair, V.; and Narahari, Y. 2021. Achieving fairness in the stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 22(174): 1–31.

Salhi, A. 1994. Global optimization: Deterministic approaches. *Journal of the Operational Research Society*, 45(5): 595–597.

Shah, S.; Lowalekar, M.; and Varakantham, P. 2020. Neural approximate dynamic programming for on-demand ride-pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Slivkins, A. 2019. *Introduction to Multi-Armed Bandits*. Now Publishers Inc.

Tawarmalani, M.; and Sahinidis, N. V. 2013. *Convexification and global optimization in continuous and mixed-integer nonlinear programming: theory, algorithms, software, and applications*, volume 65. Springer Science & Business Media.

Thomson, W. 2011. Chapter Twenty-One - Fair Allocation Rules. In *Handbook of Social Choice and Welfare*, volume 2 of *Handbook of Social Choice and Welfare*, 393–506. Elsevier.

Weitzman, M. L. 1979. Optimal Search for the Best Alternative. *Econometrica*, 47(3): 641–654.

Xu, T.; Chen, Y.; Li, H.; Bian, Z.; Dall'Anese, E.; and Zheng, Z. 2025a. Online Learning with Probing for Sequential User-Centric Selection. *arXiv preprint arXiv:2507.20112*.

Xu, T.; Liu, J.; Mattei, N.; and Zheng, Z. 2025b. Fair Algorithms with Probing for Multi-Agent Multi-Armed Bandits. *arXiv preprint arXiv:2506.14988*. Full version of this paper.

Xu, T.; Zhang, D.; Pathak, P. H.; and Zheng, Z. 2021. Joint AP Probing and Scheduling: A Contextual Bandit Approach. *IEEE Military Communications Conference (MILCOM)*.

Xu, T.; Zhang, D.; and Zheng, Z. 2023. Online Learning for Adaptive Probing and Scheduling in Dense WLANs. *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*.

Zhang, H.; and Conitzer, V. 2021. Incentive-Aware PAC Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 5797–5804.

Zhang, X.; Xie, H.; Li, H.; and C.S. Lui, J. 2020. Conversational Contextual Bandit: Algorithm and Application. In *Proceedings of The Web Conference 2020*, 662–672.

Zuo, J.; Zhang, X.; and Joe-Wong, C. 2020. Observe Before Play: Multi-Armed Bandit with Pre-Observations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 7023–7030.