Fiduciary Bandits

Gal Bahar 1  Omer Ben-Porat 1  Kevin Leyton-Brown 2  Moshe Tennenholtz 1

Abstract

Recommendation systems often face exploration-exploitation tradeoffs: the system can only learn about the desirability of new options by recommending them to some user. Such systems can thus be modeled as multi-armed bandit settings; however, users are self-interested and cannot be made to follow recommendations. We ask whether exploration can nevertheless be performed in a way that scrupulously respects agents’ interests—i.e., by a system that acts as a fiduciary.

More formally, we introduce a model in which a recommendation system faces an exploration-exploitation tradeoff under the constraint that it can never recommend any action that it knows yields lower reward in expectation than an agent would achieve if it acted alone. Our main contribution is a positive result: an asymptotically optimal, incentive compatible, and ex-ante individually rational recommendation algorithm.

1. Introduction

Multi-armed bandits (henceforth MABs) (Bubeck et al., 2012; Cesa-Bianchi & Lugosi, 2006) is a well-studied problem domain in online learning. In that setting, several arms (i.e., actions) are available to a planner; each arm is associated with an unknown reward distribution, from which rewards are sampled independently each time the arm is pulled. The planner selects arms sequentially, aiming to maximize her sum of rewards. This often involves a trade-off between exploiting arms that have been observed to yield good rewards and exploring arms that could yield even higher rewards. Many variations of this model exist, including stochastic (Abbasi-Yadkori et al., 2011; Karin et al., 2013), Bayesian (Chapelle & Li, 2011; Agrawal & Goyal, 2012), contextual (Chu et al., 2011; Slivkins, 2014), adversarial (Auer et al., 1995) and non-stationary (Besbes et al., 2014; Levine et al., 2017) bandits.

This paper considers a setting motivated by recommender systems. Such systems suggest actions to agents based on a set of current beliefs and assess agents’ experiences to update these beliefs. For instance, in navigation applications (e.g., Waze; Google maps) the system recommends routes to drivers based on beliefs about current traffic congestion. The planner’s objective is to minimize users’ average travel time. The system cannot be sure of the congestion on a road segment that no agents have recently traversed; thus, navigation systems offer the best known route most of the time and explore occasionally. Of course, users are not eager to perform such exploration; they are self-interested in the sense that they care more about minimizing their own travel times than they do about conducting surveillance about traffic conditions for the system.

A recent line of work (Kremer et al., 2014; Mansour et al., 2015), inspired by the viewpoint of algorithmic mechanism design (Nisan & Ronen, 1999; Nisan et al., 2007), deals with that challenge by incentivizing exploration—that is, setting up the system in such a way that no user would ever rationally choose to decline an action that was recommended to him. The key reason that it is possible to achieve this property while still performing a sufficient amount of exploration is that the planner has more information than the agents. At each point in time, each agent holds beliefs about the arms’ reward distributions; the planner has the same information, but also knows about all of the arms previously pulled and the rewards obtained in each case. More specifically, Kremer et al. (2014) consider a restricted setting and devise an MAB algorithm that is incentive compatible (IC), meaning that whenever the algorithm recommends arm i to an agent, the best response of the agent is to select arm i.

Although this approach explicitly reasons about agents’ incentives, it does not treat agents fairly: agents who are asked to explore receive lower expected rewards. More precisely, in their attempt to reach optimality (in the static setting) or minimize regret (in the online setting), these IC MAB algorithms are intentionally providing (a priori) sub-optimal recommendations to some of the agents. In particular, some of the agents could be better off by not using the system and follow their default arm—the a priori superior arm, which...
would be every agent’s rational choice in the absence both of knowledge of other agents’ experiences and of a trusted recommendation. Thus, it would be natural for agents to see the recommendations of such IC MAB algorithms as a betrayal of trust; they might ask “why should I trust a recommender that occasionally gives out recommendations it has every reason to believe could make me worse off?”

In this work, we explicitly suggest that a social welfare maximization standpoint might raise societal issues, harming the trust agents put in recommender systems. The central premise of this paper is that explore-and-exploit AI systems should satisfy individual guarantees—guarantees that the system should fulfill for each agent independently from the other agents and their recommendations. At the one end of the spectrum are current MAB algorithms—successful in maximizing welfare, but do not offer the slightest individual guarantee. At the other end is the fiduciary duty: borrowed from law applications, it requires that the mechanism acts in the interest of its clients with all its knowledge. This is the strictest, and strongest, individual guarantee the system could provide. However, if we applied this standard, we would be left only with the mechanism that greedily picks the apparently best arm in each iteration. In some settings, perhaps this is the best that can be achieved; however, note that this mechanism is rarely able to learn anything. It is therefore natural to ask for an approach that enjoys both worlds—maximizing welfare while satisfying individual guarantees.

Our contribution We explore a novel compromise between these two extreme points, which we call ex-ante individual rationality (EAIR). To motivate it, we consider the benchmark reward of each agent to be that of the default arm: the reward agents would get if the recommender system is unavailable. A mechanism is EAIR if the reward of every recommendation it makes beats that benchmark in expectation, per the mechanism’s knowledge. More technically, a mechanism is EAIR if any probability distribution over arms that it selects has expected reward that is always at least as great as the reward of the default arm, both calculated based on the mechanism’s knowledge (which is more extensive than that of agents). While it is possible for the mechanism to sample a recommendation from a distribution that is a priori inferior to the (realization of the) default arm, the agent receiving the recommendation is nevertheless guaranteed to realize expected reward weakly greater than that offered by the default arm. Satisfying this requirement makes a MAB algorithm more appealing to agents; we foresee that in some domains, such a requirement might be imposed as fairness constraints by authorities.

Algorithmically, we focus on constructing optimal EAIR mechanisms. Our model is a bandit model with $K \geq 2$ arms and $n$ agents (rounds). Similarly to Kremer et al. (2014), we assume that rewards are fixed but initially unknown.

We consider two agent schemes. In the first part of the paper, we assume that agents follow recommendations, as in the classical MAB literature. This is the case if, e.g., agents are oblivious to some of the actions’ desirability, unaware of the entire set of alternatives, or if the cognitive overload of computing expectations is high. The main technical contribution of this paper is an EAIR mechanism, which obtains the highest possible social welfare by any EAIR mechanism up to an additive factor of $o(\frac{1}{\epsilon})$. Due to our static setting (rewards are realized only once), following the wrong exploration policy for even one agent has detrimental effect on social welfare. The optimality of our mechanism, which we term Fiduciary Explore & Exploit (FEE) and outline as Algorithm 1, follows from a careful construction of the exploration phase. Our analysis uses an intrinsic property of the setting, which is further elaborated in Theorem 1.

Later on, in Section 4, we adopt a different agent scheme, which is fully aligned with the incentivizing exploration literature. We assume that agents are strategic and have (the same) Bayesian prior over the rewards of the arms. In this context, a mechanism is incentive compatible (IC) if each agent’s expected reward is maximized by the recommended action. We provide a positive result in this challenging case as well. Our second technical contribution is Incentive Compatible Fiduciary Explore & Exploit (IC-FEE), which uses FEE as a black box, and is IC, EAIR and asymptotically optimal.

To complement this analysis, we also propose the more demanding concept of ex-post individual rationality (EPIR). The EPIR condition requires that a recommended arm must never be a priori inferior to the default arm given the planner’s knowledge. The EAIR and EPIR requirements differ in the guarantees that they provide to agents and correspondingly allow the system different degrees of freedom in performing exploration. We design an asymptotically optimal IC and EPIR mechanism. Finally, we analyze the social welfare cost of adopting either EAIR or EPIR mechanisms.

Related work Background on MABs can be found in Cesa-Bianchi & Lugosi (2006) and a recent survey (Bubeck et al., 2012). Despite that many works address MAB rounds as interacting agents, Kremer et al. (2014) is the first work of which we are aware that suggests that vanilla algorithms should be modified to deal with agents due to human nature and incentives. The authors considered two deterministic arms, a prior known both to the agents and the planner, and an arrival order that is common knowledge among all agents, and presented an optimal IC mechanism. Cohen & Mansour (2019) extended this optimality result to several arms under further assumptions. This setting has also been extended to regret minimization (Mansour et al., 2015), so-
Fiduciary Bandits

cial networks (Bahar et al., 2016; 2019), and heterogeneous agents (Chen et al., 2018; Immorlica et al., 2019). All of this literature disallows paying agents; monetary incentives for exploration are discussed in e.g., (Chen et al., 2018; Frazier et al., 2014). None of this work considers the orthogonal, societal consideration of individual rationality constraint as we do here.

Our work also contributes to the growing body of work on fairness in Machine Learning (Ben-Porat & Tennenholtz, 2018; Dwork et al., 2012; Hardt et al., 2016; Liu et al., 2018). In the context of MABs, some recent work focuses on fairness in the sense of treating arms fairly. In particular, Liu et al. (2017) aim at treating similar arms similarly and Joseph et al. (2016) demand that a worse arm is never favored over a better one despite a learning algorithm’s uncertainty over the true payoffs. Finally, we note that the EAIR requirement we impose—that agents be guaranteed certainty over the true payoffs. Finally, we note that the mechanism makes action recommendations, but cannot compel agents to follow these recommendations. In this scheme, we say that a mechanism is incentive compatible (IC) when following its recommendations is a dominant strategy: that is, when given a recommendation, an agent’s best response is to follow her own recommendation. Formally,

**Definition 1** (Incentive Compatibility). A mechanism $M$ is incentive compatible (IC) if $\forall l \in \{1, \ldots, n\}$, for every history $h \in (A \times \mathbb{R}_+)^{l-1}$ and for all actions $a_r, a_i \in A$,

$$
\mathbb{E}(R^l(a_r) - R^l(a_i) | M(h) = a_r) \geq 0.
$$

(1)

Unless stated otherwise, we address the non-strategic agents scheme. We handle the other agent scheme in Section 4.

When agents follow the mechanism, we can represent the mechanism’s (expected) social welfare by

$$
SW(M) = \mathbb{E} \left[ \frac{1}{n} \sum_{l=1}^{n} X_{M(h_l)} \right],
$$

(2)

where $X_{M(h_l)} = \sum_{r=1}^{K} \Pr_{M(h_l)}(a_r) \mathbb{E}(X_r | h_l)$ is the reward agent $l$ receives. Notice that $X_{M(h_l)}$ depends on the randomness of the rewards and, possibly, the randomness of $M(h_l)$.

Denote the highest possible social welfare under non-strategic agents by OPT. A mechanism $M^*$ is said to be optimal if $SW(M^*) = OPT$. A mechanism $M^*$ is asymptotically optimal if, for every “large enough” number of agents $n$ greater than some number $n'$, it holds that $SW(M^*) \geq OPT - \alpha(\frac{1}{n})$. This definition of approximation is equivalent to sub-linear regret in the MAB literature.

### 2.1. Individual Guarantees

An individual guarantee is a guarantee that a mechanism can provide to the agents it interacts with, independently of the other agents. In this subsection, we present our main conceptual contribution: a meaningful individual guarantee that allows exploration.

To put our guarantee in the right context, we first present the strictest and the strongest guarantee that could be provided. A mechanism is a delegate if it acts as the agent would have acted had it revealed the information it has with her. Formally, a mechanism $M$ is a delegate if for every agent $l \in \{1, \ldots, n\}$, every history $h \in (A \times \mathbb{R}_+)^{l-1}$ and every distribution $p$ over $A$, it holds that $\mathbb{E}(X_{M(h_l)} | h) \geq$

We consider two agent schemes. The first is non-strategic agents, i.e., agents always follow the recommendation. An underlying assumption of classical MAB algorithms, such behavior could be explicit in case the mechanism makes decisions for the agents; or implicit, e.g., agents are unaware of the entire set of alternatives or their desirability, or high cognitive overload is required to compute it. The second agent scheme is strategic agents: the mechanism makes action recommendations, but cannot compel agents to follow these recommendations. In this scheme, we say that a mechanism is incentive compatible (IC) when following its recommendations is a dominant strategy: that is, when given a recommendation, an agent’s best response is to follow her own recommendation. Formally,

**Definition 1** (Incentive Compatibility). A mechanism $M$ is incentive compatible (IC) if $\forall l \in \{1, \ldots, n\}$, for every history $h \in (A \times \mathbb{R}_+)^{l-1}$ and for all actions $a_r, a_i \in A$,

$$
\mathbb{E}(R^l(a_r) - R^l(a_i) | M(h) = a_r) \geq 0.
$$

(1)

Unless stated otherwise, we address the non-strategic agents scheme. We handle the other agent scheme in Section 4.

When agents follow the mechanism, we can represent the mechanism’s (expected) social welfare by

$$
SW(M) = \mathbb{E} \left[ \frac{1}{n} \sum_{l=1}^{n} X_{M(h_l)} \right],
$$

(2)

where $X_{M(h_l)} = \sum_{r=1}^{K} \Pr_{M(h_l)}(a_r) \mathbb{E}(X_r | h_l)$ is the reward agent $l$ receives. Notice that $X_{M(h_l)}$ depends on the randomness of the rewards and, possibly, the randomness of $M(h_l)$.

Denote the highest possible social welfare under non-strategic agents by OPT. A mechanism $M^*$ is said to be optimal if $SW(M^*) = OPT$. A mechanism $M^*$ is asymptotically optimal if, for every “large enough” number of agents $n$ greater than some number $n'$, it holds that $SW(M^*) \geq OPT - \alpha(\frac{1}{n})$. This definition of approximation is equivalent to sub-linear regret in the MAB literature.

### 2.1. Individual Guarantees

An individual guarantee is a guarantee that a mechanism can provide to the agents it interacts with, independently of the other agents. In this subsection, we present our main conceptual contribution: a meaningful individual guarantee that allows exploration.

To put our guarantee in the right context, we first present the strictest and the strongest guarantee that could be provided. A mechanism is a delegate if it acts as the agent would have acted had it revealed the information it has with her. Formally, a mechanism $M$ is a delegate if for every agent $l \in \{1, \ldots, n\}$, every history $h \in (A \times \mathbb{R}_+)^{l-1}$ and every distribution $p$ over $A$, it holds that $\mathbb{E}(X_{M(h_l)} | h) \geq$

We consider two agent schemes. The first is non-strategic agents, i.e., agents always follow the recommendation. An underlying assumption of classical MAB algorithms, such behavior could be explicit in case the mechanism makes decisions for the agents; or implicit, e.g., agents are unaware of the entire set of alternatives or their desirability, or high cognitive overload is required to compute it. The second agent scheme is strategic agents: the mechanism makes action recommendations, but cannot compel agents to follow these recommendations. In this scheme, we say that a mechanism is incentive compatible (IC) when following its recommendations is a dominant strategy: that is, when given a recommendation, an agent’s best response is to follow her own recommendation. Formally,

**Definition 1** (Incentive Compatibility). A mechanism $M$ is incentive compatible (IC) if $\forall l \in \{1, \ldots, n\}$, for every history $h \in (A \times \mathbb{R}_+)^{l-1}$ and for all actions $a_r, a_i \in A$,

$$
\mathbb{E}(R^l(a_r) - R^l(a_i) | M(h) = a_r) \geq 0.
$$

(1)

Unless stated otherwise, we address the non-strategic agents scheme. We handle the other agent scheme in Section 4.
We compare the two options: picking the default arm or ward distribution) guaranteed by the mechanism. Due welfare. While sometimes relaxing this strong guarantee is impossible (e.g., banking or health-care), in many situations the planner is willing to relax individual guarantees to favor better social welfare.

The other extreme is to adopt a policy that we term FULL-EXPLORATION. FULL-EXPLORATION is the mechanism that first explores all arms sequentially, and then exploit the best arm. Clearly, at least for the non-strategic agent scheme, FULL-EXPLORATION is optimal when the number of agents is large enough. Nevertheless, with very high probability, it picks sub-optimal arms for the first \( K \) agents, which can be a highly undesired property.

Our guarantee builds on the popular economic concept of individual rationality. To introduce it, we propose the following thought experiment. Assume that agents have to make decisions without the mechanism. The agents know that \( \mu_1 = \max_i, \mu_i; \) hence, we shall assume that every agent’s default action is \( a_1 \).\(^1\) The default action is the action each agent would have selected if she did not use the mechanism. We compare the two options: picking the default arm or following the mechanism’s action. If a mechanism guarantees that the latter is higher in expectation according to its knowledge, agents are better off using the mechanism. As a result, an individually rational mechanism should guarantee each agent at least the reward obtained by her default action. The next definition relies on this reasoning.

**Definition 2 (Ex-Ante Individual Rationality).** A mechanism \( M \) is ex-ante individually rational (EAIR) if for every agent \( l \in \{1, \ldots, n\} \), and every history \( h \in (A \times \mathbb{R}_+)^{l-1} \),

\[
\sum_{r=1}^{K} \Pr_{M,h}(a_r) E(X_r \mid h) \geq E(X_1 \mid h). \tag{3}
\]

The EAIR definition is conditioned on histories, i.e., the mechanism’s knowledge. The right hand side is what an agent would get, given the knowledge of the mechanism, if she follows the default arm (which is optimal according to her knowledge). The left hand side is the expected value (over lotteries selected by the mechanism and reward distribution) guaranteed by the mechanism. Due to the mutual independence assumption, we must have \( E(X_r \mid h) = R_e \) if arm \( a_r \) was observed under the history \( h \) and \( E(X_r \mid h) = \mu_e \) otherwise. An EAIR mechanism must select a portfolio of arms with expected reward never inferior to the reward of the default arm \( a_1 \).

\(^1\) As it will become apparent later, if agents have different default arms the social welfare can only increase since more arms could be explored.

**Example.** We now give an example to illustrate our setting and to familiarize the reader with our notation. Consider \( K = 3 \) arms, \( H = 30 \) and \( X_1 \sim \text{Uni}(0, \ldots, 30) \), \( X_2 \sim \text{Uni}(0, \ldots, 20) \), \( X_3 \sim \text{Uni}(0, \ldots, 10) \); thus \( \mu_1 = 15, \mu_2 = 10, \) and \( \mu_3 = 5 \). As always, \( a_1 \) is the default arm. To satisfy EAIR, a mechanism should recommend \( a_1 \) to the first agent, since EAIR requires that the expected value of any recommendation should weakly exceed \( R_1 \). Let \( h_1 = (a_1, R_1) \) be the history after the first agent. Now, we have three different cases. First, if \( R_1 > \mu_2 = 10 \), we know that \( E(X_2 \mid h_1) < R_1 \) and \( E(X_3 \mid h_1) < R_1 \); therefore, an EAIR mechanism can never explore any other arm, since any distribution over \( \{a_2, a_3\} \) would violate Inequality (3). Second, if \( R_1 \leq \mu_3 = 5 \), then \( E(X_2 \mid h_1) \geq R_1 \) and \( E(X_3 \mid h_1) \geq R_1 \), and hence an EAIR mechanism can explore both \( a_2 \) and \( a_3 \).

The third and most interesting case is where \( \mu_3 < R_1 \leq \mu_2 \), as when \( R_1 = 8 \). In this case, arm \( a_3 \) could only be recommended through a portfolio. An EAIR mechanism could select any distribution over \( \{a_2, a_3\} \) that satisfies Inequality (3): any \( p \in [0, 1] \) such that \( p \cdot \mu_2 + (1 - p) \cdot \mu_3 \geq R_1 \). This means that an EAIR mechanism can potentially explore arm \( a_3 \), yielding higher expected social welfare overall than simply recommending a non-inferior arm deterministically.

### 3. Asymptotically Optimal EAIR Mechanism

In this section, we consider the case of non-strategic agents. We present the main technical contribution of this paper: a mechanism that asymptotically optimally balances the explore-exploit tradeoff while satisfying the EAIR property. The mechanism, which we term Fiduciary Explore & Exploit (FEE), is described as Algorithm 1. FEE is an event-based protocol that triggers every time an agent arrives. We now give an overview of FEE, focusing on the case where all agents adopt the recommendation of the mediator (we treat the other case in Section 4). We explain the algorithm’s exploration phase in Subsection 3.1, describe the overall algorithm in Subsection 3.2, and prove the algorithm’s formal guarantees in Subsection 3.3. We provide a comprehensive example of the way FEE operates in Section F.

FEE is composed of three phases: primary exploration (Lines 1–6), secondary exploration (Lines 7–18), and exploitation (Line 19). During the primary exploration phase, the mechanism compares the default arm \( a_1 \) to whichever other arms are permitted by the individual rationality constraint. This turns out to be challenging for two reasons. First, the order in which arms are explored matters; tackling them in the wrong order can reduce the set of arms that can be considered overall. Second, it is nontrivial to search in the continuous space of probability distributions over arms. To address this latter issue, we present a key lemma that allows us to use dynamic programming and find the optimal
Fiduciary Bandits

exploration policy in time $O(2^K K^2 H^2)$. Because we expect $K$ either to be fixed or to be significantly smaller than $n$, $H$, this policy is computationally efficient. Moreover, we note that the optimal exploration policy can be computed offline prior to the agents' arrival.

The primary exploration phase terminates in one of two scenarios: either the reward $R_1$ of arm $a_1$ is the best that was observed and thus no other arm could be explored (as in our example when $R_1 > 10$, or when $R_1 = 8$ and exploring $a_2$ yielded $R_2 < R_1$ and thus $a_3$ could not be explored), or another arm $a_i$ was found to be superior to $a_1$: i.e., an arm $a_i$ was observed for which $R_i > R_1$. In the latter case, the mechanism gains the option of conducting a secondary exploration, using arm $a_i$ to investigate all the arms that were not explored in the primary exploration phase. The third and final phase—to which we must proceed directly after the primary exploration phase if that phase does not identify an arm superior to the default arm—is to exploit the most rewarding arm observed.

Remark. In this section we assume that agents are non-strategic and follow the mechanism's recommendation.

3.1. Primary Exploration Phase

Performing primary exploration optimally requires solving a planning problem; it is a challenging one, because it involves a continuous action space and a number of states exponential in $K$ and $H$. We approach this task as a Goal Markov Decision Process (GMDP) (see, e.g., (Barto et al., 1995)) that abstracts everything but pure exploration. In our GMDP encoding, all terminal states fall into one of two categories. The first category is histories that lead to pure exploration, using arm $a_i$ to investigate all the arms that were not explored in the primary exploration phase. The second is histories in which an arm superior to $a_1$ was found. Non-terminal states thus represent histories in which it is still permissible for some arms to be explored. The set of actions in each non-terminal state is the set of distributions over the non-observed arms (i.e., portfolios) corresponding to the history represented in that state, which satisfy the EAIR condition. The transition probabilities encode the probability of choosing each candidate arm from a portfolio; observe that the rewards of each arm are fixed, so this is not a source of additional randomness in our model.

GMDP rewards are given in terminal nodes only: either the observed $R_1$ if no superior arm was found or the expected value of the maximum between the superior reward discovered and the maximal reward of all unobserved arms (since in this case, as we show later on, the mechanism is able to explore all arms w.h.p. during the secondary exploration phase).

Formally, the GMDP is a tuple $(S, A, P, R)$, where

- $S$ is a finite set of states. Each state $s$ is a pair $(O, U)$, where $O \subseteq \{(a, c) \mid a \in A, c \in H\}$ is the set of arm-reward pairs that have been observed so far, with each $a$ appearing at most once in $O$ (since rewards from the arms are deterministic): for every $(O, U)$ and every $a \in A$, $|\{(c \mid (a, c) \in O)\}| \leq 1$. $U \subseteq A$ is the set of arms not yet explored. The initial state is thus $s_0 = (\emptyset, A)$. For every non-empty\(^2\) set of pairs $O$ we define $\alpha(O)$ to be the reward observed for arm $a_1$, and $\beta(O) = \max_{c \in \Delta a, (a, c) \in O} c$ to be the maximal reward observed.

- $A = \bigcup_{a \in S} \mathcal{A}_s$ is an infinite set of actions. For each $s = (O, U) \in S$, $\mathcal{A}_s$ is defined as follows:
  1. If $s = s_0$, then $\mathcal{A}_{s_0} = \Delta(\{a_1\})$: i.e., a deterministic selection of $a_1$.
  2. Else, if $\alpha(O) < \beta(O)$, then $\mathcal{A}_s = \emptyset$. This condition implies that we can move to secondary exploration.
  3. Otherwise, $\mathcal{A}_s$ is a subset of $\Delta(U)$, such that $p \in \mathcal{A}_s$ if and only if

$$
\sum_{a_i \in U} p(a_i) \mu_{a_i} \geq \alpha(O).
$$

Notice that this resembles the EAIR condition given in Inequality (3). Moreover, the case where none of the remaining arms have strong enough priors to allow exploration falls here as a vacuous case of the above inequality.

We denote by $S_T$ the set of terminal states, namely $S_T = \{s \in S \mid \mathcal{A}_s = \emptyset\}$.

- $P$ is the transition probability function. Let $s = (O, U) \in S$, and let $s' = (O', U')$ such that $O' = O \cup \{(a_i, c)\}$ and $U' = U \setminus \{a_i\}$ for some $a_i \in U, c \in [H]^+$. Then, the transition probability from $s$ to $s'$ given an action $p$ is defined by $P(s'|s, p) = p(a_i) \Pr(X_i = c)$. If $s'$ is some other state that does not meet the conditions above, then let $P(s'|s, p) = 0$ for every $p \in \mathcal{A}_s$.

- $R : S_T \rightarrow \mathbb{R}$ is the reward function, defined on terminal states only. For each terminal state $s = (O, U) \in S_T$,

$$
R(s) = \begin{cases} 
\alpha(O) & \text{if } \alpha(O) = \beta(O), \\
\mathbb{E}[\max\{\beta(O), \max_{a_i, c \in U} X_i\}] & \text{otherwise.}
\end{cases}
$$

That is, when $a_1$ was the highest-reward arm observed, the reward of a terminal state is $\alpha(O)$; otherwise, it is the expectation of the maximum between $\beta(O)$ and the highest reward of all unobserved arms. The reward depends on unobserved arms since the secondary exploration phase allows us to explore all these arms; hence, their values are also taken into account.

\(^2\)Due to the construction, every non-empty $O$ must contain $(a_1, c)$ for some $c \in [H]^+$. 

A policy \(\pi : (S \times A)^* \times S \rightarrow A\) is a function from all GMDP histories (sequences of states and actions) and a current state to an action. A policy \(\pi\) is valid if for every history \(h\) and every non-terminal state \(s, \pi(h, s) \in A_s\). A policy \(\pi\) is stationary if for every two histories \(h, h'\) and a state \(s, \pi(h, s) = \pi(h', s)\). When discussing a stationary policy, we thus neglect its dependency on \(h\), writing \(\pi(s)\).

Given a policy \(\pi\) and a state \(s\), we denote by \(W(\pi, s)\) the expected reward of \(\pi\) when initialized from \(s\), which is defined recursively from the terminal states:

\[
W(\pi, s) = \begin{cases} R(s) & \text{if } s \in S_T \\ \sum_{s' \in S} P(s'|s, \pi(s))W(\pi, s') & \text{otherwise.} \end{cases}
\]

We now turn to our technical results. The following lemma shows that we can safely focus on stationary policies that effectively operate on a significantly reduced state space.

**Lemma 1.** For every policy \(\pi\) there exists a stationary policy \(\pi'\) such that (1) \(\pi'(s) = \pi'(s')\) for every pair of states \(s = (O, U)\) and \(s' = (O', U)\) with \(\alpha(O) = \alpha(O')\) and \(\beta(O) = \beta(O')\); and (2) for every state \(s\), \(W(\pi', s) \geq W(\pi, s)\).

Lemma 1 tells us that there exists an optimal, stationary policy that selects the same action in every pair of states that share the same unobserved set \(U\) and values \(\alpha(O)\) and \(\beta(O)\), but are distinguished in the \(O\) component. Thus, we do not need a set of states whose size depends on the number of possible arm-reward observation histories: all we need to record is \(U\) and a real value for either \(\alpha(O)\) and \(\beta(O)\), reducing the number of states to \(O(2^KH)\).

We still have one more challenge to overcome: the set of actions \(A_s\), available in each state \(s\) is infinite. Despite that \(A_s\) is a convex polytope and thus we can apply Linear Programming, our approach is much more computationally efficient and interpretable. We prove that there exists an optimal “simple” policy, which we denote \(\pi^*\). Given two indices \(i, r \in \{2, \ldots, K\}\), we denote by \(p^i_r\) (for \(i \neq r\) and by \(p^i_i\) (for \(i = r\)) the distributions over \(\{a_1, \ldots, a_K\}\) such that

\[
p^i_r(a) = \begin{cases} \frac{|\alpha - \mu_r|}{|\alpha - \mu_i| + |\alpha - \mu_r|} & \text{if } a = a_i \\ \frac{|\alpha - \mu_i|}{|\alpha - \mu_i| + |\alpha - \mu_r|} & \text{if } a = a_r \\ 0 & \text{otherwise} \end{cases}
\]

and \(p^i_i(a) = 1\) if and only if \(a = a_i\). When \(\alpha = \alpha(O)\) is clear from context, we omit it from the superscript.

We are now ready to describe the policy \(\pi^*\), which we later prove to be optimal. For the initial state \(s_0\), \(\pi^*(s_0) = p^1_1\). For every non-terminal state \(s = (O, U) \in S\) with \(s \neq s_0\), \(\pi^*(s) = p^i_r\), such that \((i^*, r^*) \in A_s\) maximize

\[
\left(1 - \frac{1}{2^r}\right) \left[ p^i_r(i) \sum_{s' \in S} P(s'|s, p^i_i)W(\pi^*, s') + p^i_r(r) \sum_{s' \in S} P(s'|s, p^i_r)W(\pi^*, s') \right].
\]

The optimality of \(\pi^*\) follows from a property that is formally proven in Theorem 1: any policy \(\pi\) that satisfies the conditions of Lemma 1 can be presented as a mixture of policies that solely take actions of the form \((p^i_r), i, r\). As a result, we can improve \(\pi\) by taking the best such policy from that mixture. We derive \(\pi^*\) via dynamic programming, where the base cases are the set of terminal states. For any other state, \(\pi^*(s)\) is the best action of the form \(p^i_r\), as defined above, considering all states that are reachable from \(s\). While any policy \(\pi^*\) can be encoded as a weighted sum over such “simple” policies, \(\pi^*\) is the best one, and hence is optimal.

**Theorem 1.** For every valid policy \(\pi\) and every state \(s\), it holds that \(W(\pi^*, s) \geq W(\pi, s)\).

Since our compressed state representation consists of \(O(2^K H)\) states, the computation of \(\pi^*\) in each stage requires us to consider \(O(K^2)\) candidate actions, each of which involves summation of at most \(H + 1\) summands; thus, \(\pi^*\) can be computed in \(O(2^K K^2 H^2)\) time.

### 3.2. Intuitive Description of FEE

We now present the FEE algorithm, stated formally as Algorithm 1. The primary exploration phase (Lines 1–6) is based on the GMDP from the previous subsection. It is composed of computing \(\pi^*\) and then producing recommendations according to its actions, each of which defines a distribution over (at most) two actions. Let \((U, O)\) denote the terminal state reached by \(\pi^*\) (the primary exploration selects a fresh arm in each stage; hence such a state is reached after at most \(K\) agents).

We then enter the secondary exploration phase. If \(\beta(O) = R_1\) then this phase is vacuous: no distribution over the unobserved arms can satisfy the EAIR condition and/or all the observed arms are inferior to arm \(a_1\). On the other hand, if \(\beta(O) > R_1\) (Line 7), we found an arm \(a_1\) with a reward superior to \(R_1\), and can use it to explore all the remaining arms. For every \(a_i \in U\), the mechanism operates as follows. If the probability of \(a_i\) yielding a reward greater than \(a_1\) is zero, we neglect it (Lines 11–13). Else, if \(\mu_i \geq R_1\), we recommend \(a_i\). This is manifested in the second condition in Line 15. Otherwise, \(\mu_i < R_1\). In this case, we select a distribution over \(\{a_1, a_i\}\) that satisfies the EAIR condition and explore \(a_i\) with the maximal possible probability, which is \(p^1_i(i)\). As we show formally in the proof of Lemma 2, the probability of exploring \(a_i\) in this case is at least \(1 - \frac{1}{2^r}\).
We begin by arguing that FEE is indeed EAIR.

Algorithm 1 Fiduciary Explore & Exploit (FEE)

1: Initialize a GMDP instance \((S, A, P, R)\), and compute \(\pi^*\).
2: Set \(s = (O, U) = (\emptyset, A)\).
3: \textbf{while} \(s\) is not terminal \textbf{do}
4: \hspace{1em} Draw arm \(a_i \sim \pi^*(s)\), recommend \(a_i\) and observe \(R_i\).
5: \hspace{1em} \(O \leftarrow O \cup \{(a_i, R_i)\}\), \(U \leftarrow U \setminus \{a_i\}\).
6: \hspace{1em} \(s \leftarrow (O, U)\).
7: \hspace{1em} \textbf{if} \(\beta(O) > R_1\) \textbf{then}
8: \hspace{2em} \textbf{while} \(U\) is not empty \textbf{do}
9: \hspace{3em} Let \(a_{\pi} \text{ s.t. } a_{\pi} \in \arg\max_{a_i \in A \setminus U} R_i\).
10: \hspace{3em} Select an arbitrary arm \(a_i \in U\).
11: \hspace{3em} \textbf{if} \(\Pr(X_i > R_{\pi}) = 0\) \textbf{then}
12: \hspace{4em} \(U \leftarrow U \setminus \{a_i\}\).
13: \hspace{4em} \textbf{continue}.
14: \hspace{3em} Draw \(Y \sim U \eta[0,1]\).
15: \hspace{3em} \textbf{if} \(Y \leq \frac{R_0 - R_1}{R_0 - \mu_1}\) \textbf{or} \(\mu_i \geq R_1\) \textbf{then}
16: \hspace{4em} Recommend \(a_i\) and observe \(R_i\).
17: \hspace{4em} \(U \leftarrow U \setminus \{a_i\}\).
18: \hspace{4em} \textbf{else}, recommend \(a_\pi\).
19: \hspace{1em} \textbf{end} \textbf{while} \(U\) is not empty \textbf{do}
20: \hspace{2em} \textbf{ Recommend} \(a_{\pi} \in \arg\max_{a_i \in A \setminus U} R_i\) to all agents.

implying that after \(H\) tries in expectation the algorithm would succeed to explore \(a_i\).

Ultimately (Line 19), FEE recommends the best observed arm to all the remaining agents.

3.3. Algorithmic Guarantees

We begin by arguing that FEE is indeed EAIR.

**Proposition 1.** FEE satisfies the EAIR condition.

The proof of Proposition 1 is highly intuitive: the reward of every recommendation FEE makes always exceed \(R_1\) in expectation. We now move on to consider the social welfare of FEE. Let \(\text{OPT}_{\text{EAIR}}\) denote the highest welfare attained by any EAIR mechanism. First, we show that the expected value of \(\pi^*\) at \(s_0\), denoted by \(W(\pi^*, s_0)\), upper bounds the social welfare of any EAIR mechanism.

**Theorem 2.** It holds that \(\text{OPT}_{\text{EAIR}} \leq W(\pi^*, s_0)\).

The proof proceeds by contradiction: given an EAIR mechanism \(M\), we construct a series of progressively-easier-to-analyze EAIR mechanisms with non-decreasing social welfare; we modify the final mechanism by granting it oracular capabilities, making it violate the EAIR property and yet preserving reducibility to a policy for the GMDP of Subsection 3.1. We then argue via the optimality of \(\pi^*\) that the oracle mechanism cannot obtain a social welfare greater than \(W(\pi^*, s_0)\). Next, we lower bound the social welfare of FEE.

**Lemma 2.** \(SW_n(\text{FEE}) \geq \text{OPT}_{\text{EAIR}} - O\left(\frac{KH^2}{n}\right)\).

The proof relies mainly on an argument that the primary and secondary explorations will not be too long on average: after \((K+1)H\) agents the mechanism is likely to begin exploiting. Noting that the lower bound of Lemma 2 asymptotically approaches the upper bound of Theorem 2, we conclude that FEE is asymptotically optimal.

4. Incentive Compatibility

In this section, we consider the second and more challenging agent scheme: strategic agents. Our main goal is to show that FEE, which we developed in Section 3.3, can be modified to satisfy IC as well.\(^3\)We remark that there are cases that an IC mechanism cannot explore all arms, regardless of individual rationality constraints. To illustrate, assume that \(\Pr(X_i \geq \mu_2) = 1\), i.e., the reward of arm \(a_1\) is always greater or equal to the expected reward of arm \(a_2\). In this case, no agent will ever follow a recommendation for arm \(a_2\). Consequently, we shall make the following standard assumption (see, e.g., (Mansour et al., 2015))

**Assumption 1.** For every \(i, j\) such that \(1 \leq i < j \leq K\), it holds that \(\Pr(X_i < \mu_j) > 0\).

If Assumption 1 does not hold for some pair \((i, j)\), arm \(a_j\) would never be explored; hence, we can remove such arms from \(A\). We shall also make the simplifying assumption that \(\mu_1 > \mu_2 \geq \cdots \geq \mu_K\), as otherwise the problem becomes easier to solve.

Among other factors, the expectation in Inequality (1) is taken over agents’ information on the arrival order. On the one extreme, the arrival order could be uniform, i.e., each agent \(l\) is entirely oblivious about her “place in line.” In this case, as we show in Section E, FEE satisfies IC as is assuming that there are sufficiently many agents. On the other extreme, which is the more popular in prior work (Kremer et al., 2014; Mansour et al., 2015), agents have complete information about their rounds. Namely, the agent arriving at time \(l\) knows that she is the \(l\)th agent. The complete information case is the more demanding one, and an IC mechanism for this case will also be IC under any distributional assumption on the arrival order. Nevertheless, as we demonstrate shortly, it requires more technical work.

We build on the techniques of Mansour et al. (2015) and use phases: each phase contains one round of exploration (that is, following FEE) and the other rounds are either exploitation via GREEDY (defined in Subsection 2.1) or

\(^3\)For simplicity, we formulated IC-FEE to satisfy IC in the best response sense: given that all other agents follow their recommendations, it is an agent’s best response to adopt the recommendation as well. However, IC-FEE can be easily modified to offer dominant strategies to agents.
recommendation of arm $a_1$. An IC version of FEE, which we term IC-FEE, is outlined as Algorithm 2.

IC-FEE works as follows. It initializes an instance of FEE, and uses it seldom in the earlier rounds, and regularly after-ward (every time IC-FEE makes a recommendation, it updates FEE). In Line 2, it recommends $a_1$ to the first agent. Recall that $\pi^*$ employed by FEE is only allowed to pick $a_1$ w.p. 1 in the first round; hence, FEE and IC-FEE coincide with the first recommendation. Then, depending on the value of $R_i$, it recommends agents $2, \ldots, K$ either greedily (maximizing the reward in each round, Line 3) or arm $a_1$ (Line 4). Later, in Line 5, it splits the remaining rounds to phases of size $B$ (to be determined later). In each such phase $k$, we first ask whether FEE is exploring or exploiting (Line 7). If FEE exploits (Line 9 in Algorithm 1), every agent of every phase from here on will be recommended by FEE. If that is not the case (see the else block starting at Line 10), IC-FEE picks one agent from the $B$ agents of this phase uniformly at random, denoted $l(k)$. Then, agent $l(k)$ gets the recommendation from FEE. The recommendation policy for the rest of the agents in this phase depends on the observed arms. If IC-FEE already discovered an arm $a_i$ with $R_i > R_1$ (Line 11), we let agent $l$ exploit using GREEDY. Otherwise (Line 12), IC-FEE recommends $a_1$.

Lines 11 and 12 are also where our mechanism departs from the principles of prior work. For example, in the work of Mansour et al. (2015), each phase contains one round of exploration and the rest are exploitation rounds. In their setting, agents that are not exploring might still not exploit. The distinction between Lines 11 and 12 is crucial: exploiting unobserved arms might lead to sub-optimal welfare, since they are the chance to explore arms with expected reward below $R_1$. We elaborate more in the proof of Theorem 3.

To determine the phase length $B$, we introduce the following quantities $\xi$ and $\gamma$. Due to Assumption 1, there exist $\xi > 0$ and $\gamma > 0$ such that for all $i \in [K]$, it holds that $\Pr(\forall i' \in [K] \setminus \{i\}: \mu_i - X_{i'} > \xi) > \gamma$. In words, it says that the reward of every arm $i$ is greater than all other arms by at least $\xi$, w.p. of at least $\gamma$. The following Theorem 3 summarizes the properties of IC-FEE.

**Theorem 3.** Let the phase length be $B = \left\lceil \frac{H}{\gamma^2} \right\rceil + 1$. Under Assumption 1, IC-FEE satisfies EAIR and IC. In addition, $SW_n(\text{IC-FEE}) \geq OPT_{\text{EAIR}} - O\left(\frac{KH^2}{n\gamma^2}\right)$.

5. Further Analysis

Notice that EAIR mechanisms guarantee each agent the value of the default arm, but only in expectation. We now propose a more strict form of individual rationality, ex-post individual rationality (EPIR).

**Algorithm 2** IC Fiduciary Explore & Exploit (IC-FEE)

1: Initialize an instance of FEE and update it after every recommendation.
2: Recommend $a_1$ to the first agent, observe $R_1$.
3: if $R_1 < \mu_K$ then recommend as GREEDY to agents $2, \ldots, K$.
4: else, recommend $a_1$ to agents $2, \ldots, K$.
5: Split the remaining rounds into consecutive phases of $B$ rounds each.
6: for phase $k = 1, \ldots, B$ do
7: if FEE exploits (Line 19 in FEE) then
8: follow FEE.
9: else
10: Pick one agent $l(k)$ from the $B$ agents in this phase uniformly at random, and recommend her according to FEE.
11: As for the rest of the agents,
12: if an arm $a_i$ with $R_i > R_1$ was revealed then recommend as GREEDY.
13: else, recommend $a_1$.

**Definition 3** (Ex-Post Individual Rationality). A mechanism $M$ is ex-post individually rational (EPIR) if for every agent $l \in \{1, \ldots, n\}$, every history $h \in (A \times \mathbb{R}_+)^{l-1}$, and every arm $a_r$ such that $\Pr_M(h)(a_r) > 0$, it holds that $E(X_r - X_1 | h) \geq 0$.

Satisfying EPIR means that the mechanism never recommends an arm that is a priori inferior to arm $a_1$ given the mechanism’s knowledge. It is immediate to see that every EPIR mechanism is also EAIR. EPIR mechanisms are quite conservative, since they can only explore arms that yield expected rewards of at least the value $R_1$ obtained for $a_1$. We develop an optimal IC/EPIR mechanism in Section D.1.

5.1. Social Welfare Analysis

We now analyze the loss in social welfare due to individual rationality constraints. For simplicity, we consider the case of non-strategic agents. Recall that OPT is the highest possible social welfare, and OPT_{\text{EAIR}} is its counterpart after imposing EAIR. In addition, let OPT_{\text{EPIR}} and OPT_{\text{DEL}} denote the best asymptotic social welfare (w.r.t. some instance $\langle K, A, (X_i) \rangle$) and infinitely many agents) achievable by an EPIR and a delegate mechanisms, respectively. Noticeably, for every instance $\langle K, A, (X_i) \rangle$, it holds that OPT $\geq$ OPT_{\text{EPIR}} $\geq$ OPT_{\text{EAIR}} $\geq$ OPT_{\text{DEL}}. In the rest of this subsection, we analyze the ratio of two subsequent optimal welfares. We begin by showing that individual guarantees can deteriorate welfare even for the most flexible notion, EAIR.

**Proposition 2.** For every $K, H \in \mathbb{N}$, there exists an instance $\langle K, A, (X_i) \rangle$ with $\frac{\text{OPT}_{\text{EAIR}}}{\text{OPT}_{\text{DEL}}} \geq H \left(1 - e^{-\frac{H}{n}}\right)$.
We thank the participants of the Computational Data Science seminar at Technion – Israel Institute of Technology and the participants of Young Researcher Workshop on Economics and Computation for their comments and suggestions. Additionally, we thank ICML 2020 anonymous reviewers who provided comments that improved the manuscript. The work of G. Bahar, O. Ben-Porat and M. Tennenholtz is funded by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement n° 740435). The work of K. Leyton-Brown is funded by the NSERC Discovery Grants program, DND/NSERC Discovery Grant Supplement, Facebook Research and Canada CIFAR AI Chair Amii. Part of this work was done while K. Leyton-Brown was a visiting researcher at Technion – Israel Institute of Science and was partially funded by the European Union’s Horizon 2020 research and innovation programme (grant agreement n° 740435).

6. Conclusions and Discussion

This paper introduces a model in which a recommender system must manage an exploration-exploitation tradeoff under the constraint that it may never knowingly make a recommendation that will yield lower reward than any individual agent would achieve if he/she acted without relying on the system.

We see considerable scope for follow-up work. First, from a technical point of view, our algorithmic results are limited to discrete reward distributions. One possible future direction would be to present an algorithm for the continuous case. More conceptually, we see natural extensions of EPIR and EAIR to stochastic settings, either by assuming a prior and requiring the conditions w.r.t. the posterior distribution or by requiring the conditions to hold with high probability. Moreover, we are intrigued by non-stationary settings—where e.g., rewards follow a Markov process—since the planner would be able to sample a priori inferior arms with high probability assuming the rewards change fast enough, thereby reducing regret.

Acknowledgements

We thank the participants of the Computational Data Science seminar at Technion – Israel Institute of Technology and the participants of Young Researcher Workshop on Economics and Computation for their comments and suggestions. Additionally, we thank ICML 2020 anonymous reviewers who provided comments that improved the manuscript. The work of G. Bahar, O. Ben-Porat and M. Tennenholtz is funded by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement n° 740435). The work of K. Leyton-Brown is funded by the NSERC Discovery Grants program, DND/NSERC Discovery Grant Supplement, Facebook Research and Canada CIFAR AI Chair Amii. Part of this work was done while K. Leyton-Brown was a visiting researcher at Technion – Israel Institute of Science and was partially funded by the European Union’s Horizon 2020 research and innovation programme (grant agreement n° 740435).

References


A. Omitted Proofs from Subsection 3.1

Proof of Lemma 1. The proof follows from Propositions 5 and 6 below.

Proposition 5. For every non-stationary policy $\pi$, there exists a stationary policy $\pi'$ such that for every state $s \in S$, $W(\pi, s) \leq W(\pi', s)$.

Moreover, the following Proposition 6 implies that we can substantially reduce the state space by disregarding the observed part $O$ and

Proposition 6. For every stationary policy $\pi$ there exists a stationary policy $\pi'$ such that:

1. $\pi'(s) = \pi'(s')$ for every pair of states $s = (O, U), s' = (O', U)$ with $\alpha(O) = \alpha(O')$ and $\beta(O) = \beta(O')$.
2. for every state $s$, $W(\pi', s) \geq W(\pi, s)$.

Proof of Proposition 5. Fix an arbitrary non-stationary policy $\pi$. We prove the claim by iterating over all states in an increasing order of the number of elements in $U$. We use induction to show that the constructed $\pi'$ indeed satisfies the assertion.

For every $s = (O, U) \in S$ such that $|U| = 1$, i.e., $U = \{a\}$. If $s$ is terminal, then $A_s = \emptyset$ and $W(\pi, s) = W(\pi', s) = \alpha(O)$. Otherwise, the unique element in $A_s$ is the action that assigns probability 1 to $a$, and by setting $\pi'(s) = \pi(s)$ we get $W(\pi', s) = W(\pi, s)$.

Assume that the assertion holds for every $|U| \leq j$; namely, that $W(\pi', s) \geq W(\pi, s)$ for all $s = (O, U) \in S$ with $|U| \leq j$. We now prove the assertion for $s = (O, U) \in S$ with $|U| = j + 1$. If $s$ is a terminal state, then we are done. Else, since $U$ and the support of each arm are finite, there exists a finite number of possible histories that lead from $s_0$ to $s$ that we will mark as $h_1, \ldots, h_w$. For every possible history $h \in \{h_1, \ldots, h_w\}$, $\pi$ assigns an action $p_h \in A_s$ that (can) depend on the history $h$. Let

$$p^* \in \arg \max_{p_h, h \in \{h_1, \ldots, h_w\}} \left\{ \sum_{a_i \in U} p_h(a_i) \sum_{s' \in S} \mathcal{P}(s' \mid s, p_{ii}) W(\pi, s') \right\},$$

(5)

breaking ties arbitrarily. We set $\pi'(s) = p^*$. Hence we get:

$$W(\pi, s) = \sum_{s' \in S, h \in \{h_1, \ldots, h_w\}} \Pr(h) \mathcal{P}(s' \mid s, \pi, h) W(\pi, s')$$

$$= \sum_{a_i \in U, h \in \{h_1, \ldots, h_w\}} \Pr(h) p_h(a_i) \sum_{s' \in S} \mathcal{P}(s' \mid s, p_{ii}) W(\pi, s')$$

$$\leq \sum_{h \in \{h_1, \ldots, h_w\}} \Pr(h) \arg \max_{h \in \{h_1, \ldots, h_w\}} \left\{ \left( \sum_{a_i \in U} p_h(a_i) \sum_{s' \in S} \mathcal{P}(s' \mid s, p_{ii}) W(\pi, s') \right) \right\}$$

$$= 1 \arg \max_{h \in \{h_1, \ldots, h_w\}} \left\{ \sum_{a_i \in U} p_h(a_i) \sum_{s' \in S} \mathcal{P}(s' \mid s, p_{ii}) W(\pi, s') \right\}$$

$$= \arg \max_{h \in \{h_1, \ldots, h_w\}} \left\{ \sum_{a_i \in U} p_h(a_i) \sum_{s' \in S} \mathcal{P}(s' \mid s, p_{ii}) W(\pi', s') \right\}$$

$$= \sum_{a_i \in U} p^*(a_i) \sum_{s' \in S} \mathcal{P}(s' \mid s, p_{ii}) W(\pi', s')$$

$$= W(\pi', s);$$

hence, $W(\pi, s) \leq W(\pi', s)$. This concludes the proof.

Proof of Proposition 6. The proof is similar to the proof of Proposition 5, and is given for completeness. Fix an arbitrary stationary policy $\pi$. We prove the claim by iterating over all states in an increasing order of the number of elements in $U$. We use induction to show that the constructed $\pi'$ indeed satisfies the assertion.
Algorithm 3 Optimal Policy $\pi^*$ for the GMDP

**Input:** an instance $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$.

**Output:** an optimal policy $\pi^*$.

1: **for** every non-terminal state $s = (O, U) \in \mathcal{S}$ **do**
2:     **if** $s = s_0$ **then**
3:         $\pi^*(s) \leftarrow p_{11}$.
4:     **else**
5:         $\pi^*(s) \leftarrow p_{i^*,r^*}$ such that

\[
(i^*, r^*) \in \arg\max_{(i,r) \in U \times U, p_{i,r} \in \mathcal{A}_s} \left\{ \left( 1 - \frac{1_{i^*=r^*}}{2} \right) \left( p_{i,r}(i) \sum_{s' \in \mathcal{S}} \mathcal{P}(s', p_{i,r})W(\pi^*, s') + p_{i,r}(r) \sum_{s' \in \mathcal{S}} \mathcal{P}(s', p_{i,r})W(\pi^*, s') \right) \right\}.
\] (7)

For every $s = (O, U) \in \mathcal{S}$ such that $|U| = 1$, i.e., $U = \{a\}$, if $s$ is terminal then $W(\pi, s) = W(\pi', s) = O$. Otherwise, the unique element in $\mathcal{A}_s$ is the action that assigns probability 1 to $a$; hence, by setting $\pi'(s) = \pi(s)$ we get $W(\pi', s) = W(\pi, s)$.

Assume the assertion holds for every $|U| \leq j$; namely, that $W(\pi', s) \geq W(\pi, s)$ for all $s = (O, U) \in \mathcal{S}$ with $|U| \leq j$. Next, we prove the assertion for $s = (O, U) \in \mathcal{S}$ with $|U| = j + 1$. If $s$ is a terminal state, then we are done. Else, since the size of $O$ and the support of each arm are finite, there exists only a finite number of states with the same $U$ and $\alpha$, which we mark as $s = s_0 = (O, U), s_1 = (O^1, U), \ldots, s_w = (O^w, U)$. For every state $s_j = (O^j, U)$, $\pi$ assigns an action $p_{s_j} \in \mathcal{A}_s$.

Let

\[
p^* \in \arg\max_{p_{s_j}, j \in \{0,1,\ldots,w\}} \left\{ \sum_{a_i \in U} p_{s_j}(a_i) \sum_{s' \in \mathcal{S}} \mathcal{P}(s' \mid s, p_{i})W(\pi', s') \right\},
\] (6)

breaking ties arbitrarily. Next, set $\pi'(s) = p^*$. We have that

\[
W(\pi, s) = \sum_{s' \in \mathcal{S}} \mathcal{P}(s' \mid s, \pi)W(\pi, s')
\]
\[
= \sum_{a_i \in U} p_{s_j}(a_i) \sum_{s' \in \mathcal{S}} \mathcal{P}(s' \mid s, p_{i})W(\pi, s')
\]
\[
\leq \arg\max_{s_j \in \{s_0,s_1,\ldots,s_w\}} \left\{ \left( \sum_{a_i \in U} p_{s_j}(a_i) \sum_{s' \in \mathcal{S}} \mathcal{P}(s' \mid s, p_{i})W(\pi, s') \right) \right\}
\]
\[
\leq \arg\max_{s_j \in \{s_0,s_1,\ldots,s_w\}} \left\{ \left( \sum_{a_i \in U} p_{s_j}(a_i) \sum_{s' \in \mathcal{S}} \mathcal{P}(s' \mid s, p_{i})W(\pi', s') \right) \right\}
\]
\[
= W(\pi', s).
\]

**Proof of Theorem 1.** Fix an arbitrary policy $\pi$. We prove the claim by iterating over all states in an increasing order of the number of elements of $U$. We use induction to show that the constructed $\pi^*$ indeed satisfies the assertion. For convenience, we restate $\pi^*$ elaborately in Algorithm 3.

For every $s = (O, U) \in \mathcal{S}$ such that $|U| = 1$, the claim holds trivially. To see this, recall that if $s$ is terminal, $\mathcal{A}_s = \emptyset$; otherwise, the unique element in $\mathcal{A}_s$ is the action that assigns probability 1 to the sole element in $U$. Either way, $W(\pi^*, s) = W(\pi, s)$.

Assume the assertion holds for every $|U| \leq j$; namely, that $W(\pi^*, s) \geq W(\pi, s)$ for all $s = (O, U) \in \mathcal{S}$ with $|U| \leq j$. If $s$ is a terminal state, then we are done. Else, we shall make use of the following claim, which shows that every action in $\mathcal{A}_s$ can be viewed as a weighted sum over the elements of $\{p_{i,r} \in \mathcal{A}_s\}$.
Claim 1. For any \( s \in S \) and \( p \in A_s \), there exist coefficients \((z_{i,r})_{(a_i,a_r) \in U \times U}\) such that

- \( z_{i,r} \geq 0 \),
- \( \sum_{(a_i,a_r) \in U \times U} z_{i,r} = 1 \), and
- \( p = \sum_{p_{i,r} \in A_s} z_{i,r} p_{i,r} \).

The proof of the claim appears below this proof. In particular, Claim 1 suggests that \( \pi(s) \), which is valid and thus \( \pi(s) \in A_s \) w.p. 1, can be presented as a weighted sum over all pairs \( p_{i,r} \in A_s \). Finally,

\[
W(\pi, s) = \sum_{s' \in S} \mathcal{P}(s' | s, \pi) W(\pi, s')
= \sum_{a_i \in U} \sum_{s' \in S} \pi(s) (a_i) \mathcal{P}(s' | s, p_{i}) W(\pi, s')
= \sum_{a_i \in U} \sum_{a_r \in U} \sum_{p_{i,r} \in A_s} (1 - \frac{1}{2}) (z_{i,r} \pi, s_{i} + z_{r,i} \pi, s_{r}) \sum_{s' \in S} \mathcal{P}(s' | s, p_{i}) W(\pi, s')
\leq \arg \max_{a_i, a_r \in U, p_{i,r} \in A_s} \left\{ \left( 1 - \frac{1}{2} \right) \left( \mathcal{P}(s' | s, p_{i}) \sum_{s' \in S} \mathcal{P}(s' | s, r) W(\pi, s') \right) \right\},
= W(\pi^*, s),
\]

where the last equality follows since \( \pi^*(s) = p_{i^*, r^*} \) and by the definition of \((i^*, r^*)\) given in Equation (7). To sum, the constructed \( \pi^* \) satisfies \( W(\pi, s) \leq W(\pi^*, s) \) for every state \( s \).

Proof of Claim 1. To ease readability, we shall use the notation \( \alpha = \alpha(O) \) and \( d_i = |\alpha - \mu_i| \) in this proof. Let \( s \) be an arbitrary state and \( p \in A_s \) be an arbitrary action. Notice that \( p \) could be described as

\[
p = \sum_{a_i \in U} v_i \cdot p_{i} + \sum_{p_{i,r} \in A_s} z_{i,r} p_{i,r}, \tag{8}
\]

where \( v_i = p(i) \) and \( z_{i,r} = 0 \) for every \( a_i, a_r \in U \) such that \( p_{i,r} \in A_s \). We now describe a procedure that shifts mass from the set \((v_i)\) to \((z_{i,r})_{i,r}\), while still satisfying the equality in Equation (8). Each time we apply this procedure we decrease the value of one or more elements from \((v_i)\), and increase one or more elements from \((z_{i,r})_{i,r}\) by the same quantity. As a result, when it converges (assuming that it does), namely when \( \sum_i v_i = 0 \), we are guaranteed that all the conditions of the claim hold. Importantly, throughout the course of this procedure, the following inequalities hold

\[
\sum_{a_i \in U} v_i \cdot \mu_i \geq \alpha \sum_{a_i \in U} v_i, \tag{9}
\]

\[
\sum_{a_i \in U} v_i + \sum_{p_{i,r} \in A_s} z_{i,r} = 1. \tag{10}
\]

For the initial set of \((v_i)\), Equations (9)-(10) trivially hold due to the way we initialize \((v_i)\) and since \( p \in A_s \) implies that

\[
\sum_{a_i \in U} p(i) \mu_i \geq \alpha. \tag{11}
\]

In each step of the procedure, we use the prime notation to denote the coefficients in the end of that step. The procedure operates as follows:

- If \( v_i = 0 \) for every \( a_i \in U \), the claim holds.
• Else, if for every \( i \) such that \( v_i > 0, \mu_i \geq \alpha \), then for every \( i \) with \( v_i > 0 \) set \( z'_{i,i} = z_{i,i} + v_i \) and set \( v'_i = 0 \). Notice that after this change Equations (8)–(10) still hold.

• There exists \( i \) with \( \mu_i < \alpha \) and \( v_i > 0 \). Consequently, since Equation (9) holds, there must exist \( r \) such that \( \mu_r > \alpha \) and \( v_r > 0 \). We divide the analysis into three sub-cases, depending on the relation between \( \frac{dv}{dr} \) and \( \frac{dv}{dr} \).

1. \( \frac{dv}{dr} > \frac{dv}{dr} \): we replace \( v_i, v_r \) and \( z_{i,r} \) with \( v'_i, v'_r \) and \( z'_{i,r} \) such that \( v'_i = 0 \), \( v'_r = v_r - v_i \frac{dv}{dr} = v_r + v_i - \frac{v_i}{p_r(i)} \) and \( z'_{i,r} = z_{i,r} + v_i \frac{dv}{dr} = z_{i,r} + \frac{v_i}{p_r(i)} \). Clearly, after this modification the new coefficients are non-negative. To show that Equation (8) still holds, we need to show that \( p(i), p(r) \) can be decomposed using the new coefficients. Notice that

\[
p(i) = v_i + \sum_{j, p_{i,j} \in A_s} z_{i,j} p_{i,j}(i) + \sum_{j, p_{i,j} \in A_s} z_{j,i} p_{j,i}(i)
= v'_i + v_i + z_{i,r} p_{i,r}(i) + \sum_{j, j \neq r, p_{i,j} \in A_s} z_{i,j} p_{i,j}(i) + \sum_{j, p_{j,i} \in A_s} z_{j,i} p_{j,i}(i)
= v'_i + v_i \frac{p_{i,r}(i)}{p_r(i)} + z_{i,r} p_{i,r}(i) + \sum_{j, j \neq r, p_{i,j} \in A_s} z'_{i,j} p_{i,j}(i) + \sum_{j, p_{j,i} \in A_s} z'_{j,i} p_{j,i}(i)
= v'_i + \left( v_i \frac{1 - p_{i,r}(i)}{p_r(i)} + z_{i,r} p_{i,r}(i) \right) + \sum_{j, j \neq r, p_{i,j} \in A_s} z'_{r,j} p_{r,j}(r) + \sum_{j, p_{j,i} \in A_s} z'_{j,r} p_{j,r}(r)
= v'_i + v_i \frac{p_{r,i}(i)}{p_r(i)} + z_{i,r} p_{i,r}(i) + \sum_{j, j \neq i, p_{r,j} \in A_s} z'_r p_{r,j}(r) + \sum_{j, p_{j,i} \in A_s} z'_{j,r} p_{j,r}(r)
= v'_i + \left( v_i \frac{1 - p_{r,i}(i)}{p_r(i)} + z_{i,r} p_{i,r}(i) \right) + \sum_{j, j \neq i, p_{r,j} \in A_s} z'_{r,j} p_{r,j}(r) + \sum_{j, p_{j,i} \in A_s} z'_{j,r} p_{j,r}(r)
= v'_i + \sum_{j, p_{r,j} \in A} z'_{r,j} p_{r,j}(r) + \sum_{j, p_{j,i} \in A_s} z'_{j,r} p_{j,r}(r).
\]

Similarly,
As a result, Equation (8) holds. As for Equation (9), observe that

\[
\sum_{a_j \in U} v_j^l : \mu_j = v_i^l : \mu_i + v_i^r : \mu_i + \sum_{j \notin \{i,r\}, a_j \in U} v_j : \mu_j
\]

\[
= v_r : \mu_r - v_i : \mu_i \frac{d_i}{d_r} + \sum_{j \notin \{i,r\}, a_j \in U} v_j : \mu_j
\]

\[
= v_r : \mu_r + v_i : \mu_i - v_i : \mu_i \frac{d_i}{d_r} + \sum_{j \notin \{i,r\}, a_j \in U} v_j : \mu_j
\]

\[
= -v_i : \mu_i - v_i : \mu_i \frac{d_i}{d_r} + \sum_{a_j \in U} v_j : \mu_j
\]

\[
\geq -v_i : \mu_i - v_i : \mu_i \frac{d_i}{d_r} + \alpha \sum_{a_j \in U} v_j
\]

\[
= -v_i : \mu_i - v_i : \mu_i \frac{d_i}{d_r} + \alpha \cdot v_i + \alpha \cdot v_r + \alpha \sum_{j \notin \{i,r\}, a_j \in U} v_j
\]

\[
= -v_i : \mu_i - v_i : \mu_i \frac{d_i}{d_r} + \alpha \cdot v_i + \left(\alpha \cdot v_r + \alpha \cdot v_i \frac{d_i}{d_r}\right) + \alpha \cdot v_i + \alpha \sum_{j \notin \{i,r\}, a_j \in U} v_j
\]

\[
= -v_i : \mu_i - v_i : \mu_i \frac{d_i}{d_r} + \alpha \cdot v_i + \alpha \cdot v_i \frac{d_i}{d_r} + \alpha \sum_{a_j \in U} v_j
\]

\[
= v_i \left(-\mu_i - \mu_i \frac{d_i}{d_r} + \alpha + \alpha \frac{d_i}{d_r}\right) + \alpha \sum_{a_j \in U} v_j
\]

\[
= v_i \left(d_i - d_r \frac{d_i}{d_r}\right) + \alpha \sum_{a_j \in U} v_j
\]

\[
= \alpha \sum_{a_j \in U} v_j^l;
\]

hence, Equation (9) holds. Finally, \(v_i + v_r + z_{i,r} = v_i^l + v_r^l + z_{i,r}^l\) while all other coefficients are left unchanged; thus Equation (10) holds as well.

2. \(\frac{\alpha}{\pi_i} < \frac{\alpha}{\pi_r}\); the analysis is similar to the previous case and hence omitted.

3. \(\frac{d_r}{d_i} = \frac{v_i}{v_r}\); the analysis is similar to the first case and hence omitted.

This concludes the proof.

\[\square\]

### B. Omitted Proofs from Subsection 3.3

**Proof of Proposition 1.** We need to show that Inequality (3) holds for every history \(h\). Since FEE operates in phases, it would be convenient to divide the arguments into these three phases, according to which phase \(h\) belongs.

- **Exploration phase:** the recommendation is based on the action of \(\pi^*\), the optimal policy of the GMDP in Subsection 3.1. If \(h\) is the empty history, then it is translated to \(s_0\), and \(\pi^*\) selects \(a_1\) w.p. 1. Otherwise, due to Equation (4) the action space of the GMDP is restricted to distributions over the unobserved arms with expectation greater or equal to the observed value \(R_1\). As a result, in both cases Inequality (3) holds.

- **Experience phase:** in this phase, FEE \((h)\) is a distribution over two arms, \(\tilde{r}\) and \(i\), with \(R_{\tilde{r}}\) greater than the obtained value \(R_1\) of arm \(a_1\). Further, \(X_i > R_{\tilde{r}}\) with positive probability, or otherwise arm \(a_i\) would have been discarded (Lines 11–13). If, in addition, \(\mu_i \geq R_1\), then the If sentence in Line 15 would select arm \(a_i\) with probability 1, satisfying Inequality (3). On the other hand, if \(\mu_i < R_1\), then FEE selects arm \(a_i\) w.p. \(\frac{R_i - R_{\tilde{r}}}{R_{\tilde{r}} - \mu_i}\), and \(a_{\tilde{r}}\) with the remaining
probability (Lines 14–19); hence expected value of $\text{FEE} \ (h)$ is

$$
\mu_i \cdot \frac{R_{\bar{\tau}} - R_1}{R_{\bar{\tau}} - \mu_i} + R_{\bar{\tau}} \left(1 - \frac{R_{\bar{\tau}} - R_1}{R_{\bar{\tau}} - \mu_i}\right),
$$

which is greater or equal to $R_1$.

- Exploit phase: in this phase $\text{FEE} \ (h)$ is a deterministic selection of one arm — the most rewarding one. Since the value of arm $a_1$, $R_1$ was observed before (as mentioned for the exploration phase), the arm $a_{i^*}$ selected in Line 19, satisfies $R_i > R_i^*$.

\[\square\]

\section*{B.1. Optimality}

\textbf{Proof of Theorem 2.} To facilitate the proof, we introduce the following definitions: given a mechanism $M$ and a history $h$, we say that $M$ is fruitless w.r.t. $h$ if $M(h)$ gives a positive probability to at least one observed arm $a_i, i \neq 1$, with $R_i \leq R_1$, i.e., reward that is at most $R_1$ (notice that it implies that $a_1$ and $a_i$ were observed). In addition, we say that a history $h$ is auspicious if an action with reward greater than that of $a_1$ is observed under $h$.

We are ready to begin the proof. Let $M$ be an arbitrary mechanism, and for the sake of the proof fix the number of agents, and only consider histories of length of at most $n$. The proof contains three steps. In Step 1 we slightly modify $M$, resulting in a new mechanism $M^{(1)}$ that attains a social welfare at least as high as that of $M$, and is still EAIR. In Step 2, we modify $M^{(1)}$ to use an oracle whenever it reaches an auspicious history. As we show, the resulting mechanism, $M^{(2)}$ has an improved social welfare, $\text{SW}(M^{(2)}) \geq \text{SW}(M^{(1)})$. Finally, in Step 3 we show that the social welfare of $M^{(2)}$ is at most $W(\pi^*, s_0)$.

\textbf{Step 1:} In this step we construct a modification of $M$ with at least the same social welfare, which is not fruitless on any history $h$. We define a mechanism $M^{(1)}$ that receives $M$ as a black box and uses it for recommendations. $M^{(1)}$ is defined as follows:

1. Let $\tilde{h}$ be the empty history. Act as $M(\tilde{h})$ and update $\tilde{h}$ accordingly.

2. While the length of $\tilde{h}$ is less than $n$:
   
   2.1 Draw $a_i \sim M(\tilde{h})$. If the reward of $a_i$ was already observed and $R_i \leq R_1$, recommend $a_1$ and set $\tilde{h} = \tilde{h} \oplus (a_i, R_i)$. Else, act as $M(\tilde{h})$ and update $\tilde{h}$ accordingly.

It is straightforward to see that $M^{(1)}$ satisfies the EAIR condition, and that $\text{SW}(M^{(1)}) \geq \text{SW}(M)$.

\textbf{Step 2:} In this step, we present a non-feasible mediator $M^{(2)}$ that modifies the way $M^{(1)}$ operates on auspicious histories. $M^{(2)}$ uses an oracle that hints the best arm.

More concretely, $M^{(2)}$ is defined as follows:

1. Let $\tilde{h}$ be the empty history. Act as $M^{(1)}(\tilde{h})$ and update $\tilde{h}$ accordingly.

2. While $\tilde{h}$ is not auspicious:
   
   2.1 Act as $M^{(1)}(\tilde{h})$ and update $\tilde{h}$ accordingly.

3. If $\tilde{h}$ is auspicious:
   
   3.1 Use an oracle to reveal the best arm, $a^*$. From here on, recommend $a^*$ to all users.

Notice that $M^{(2)}$ is EAIR for every non-auspicious history, but not EAIR in general; for this reason, it is not feasible. Moreover, it holds that $\text{SW}(M^{(2)}) \geq \text{SW}(M^{(1)})$. 

\[\square\]
The proof of the lemma is identical to the proof of Lemma 1 and hence omitted. Lemma 3 suggests that we can focus on strategies that distinguish between states based on \( U, \alpha(O) \) and \( \beta(O) \) solely. The reduced state space does allow self loop by selecting \( a_1 \), without having any effect on the reward. It is thus straightforward to see that an optimal strategy that ignores \( a_1 \) exists, with a reward of exactly \( W(\pi^*, s_0) \).

\[ \text{Step 3: The final step is to claim that the resulting mechanism } M^{(2)} \text{ cannot get more than the optimal value of the GMDP in Section 3. However, the GMDP does not allow selecting } a_1, \text{ so we have to have some minor modifications.} \]

This step is structured as follows. First, formally define a modified version of the GMDP presented in Section 3, with minor modifications. We call the new GMDP \textit{Repeated GMDP}, or R-GMDP for abbreviation to distinguish between the two. Then, we show that the best achievable value in the R-GMDP is exactly \( W(\pi^*, s_0) \). The final step is mapping \( M^{(2)} \) obtained in Step 2 to a non-stationary strategy in the R-GMDP, which achieves at least as as the social welfare of \( M^{(2)} \), that is \( SW(M^{(2)}) \). The claim then follows since the policy constructed using \( M^{(2)} \) cannot obtain more that \( W(\pi^*, s_0) \).

Consider the following R-GMDP:

\[ S = \text{a finite set of states. Each state } s \text{ is a pair } (O, U), \text{ where } O \subseteq \{(a, c) \mid a \in A, c \in H\} \text{ is the set of arm–reward pairs that have been observed so far. } U \subseteq A \text{ is the set of arms not yet explored. The initial state is thus } s_0 = (\emptyset, A). \text{ For every non-empty set of pairs } O \text{ we define } \alpha(O) \text{ to be the reward observed for arm } a_1 \text{ (that can be obtained several times, as we explain shortly), and } \beta(O) = \max_{c: \exists a_1(a, c) \in O} c \text{ to be the maximal reward observed.} \]

\[ A = \bigcup_{s \in S} A_s \text{ is an infinite set of actions. For each } s = (O, U) \in S, A_s \text{ is defined as follows:} \]

1. If \( s = s_0, \text{ then } A_{s_0} = \Delta(\{a_1\}) \); i.e., a deterministic selection of \( a_1 \).
2. Else, if \( \alpha(O) < \beta(O), \text{ then } A_s = \emptyset. \)
3. Otherwise, \( A_s \) is a subset of \( \Delta(U \cup \{a_1\}) \), such that \( p \in A_s \) if and only if

\[ \sum_{a_i \in U \cup \{a_1\}} p(a_i) \mu_{a_i} \geq \alpha(O). \]

We denote by \( S_T \) the set of terminal states, namely \( S_T = \{s \in S \mid A_s = \emptyset\}. \)

\[ P \text{ is the transition probability function. Let } s = (O, U) \in S, \text{ and let } s' = (O', U') \text{ such that } O' = O \cup \{a_i\} \text{ and } U' = U \setminus \{a_i\} \text{ for some } a_i \in U \cup \{a_1\}, c \in [H]^+. \]

Then, the transition probability from \( s \) to \( s' \) given an action \( p \) is defined by

\[ P(s'|s, p) = \begin{cases} p(a_i) \Pr(x_i = c) & a_i \in U, a_i = a_1, \\ 1 & c = \alpha(O). \end{cases} \]

If \( s' \) is some other state that does not meet the conditions above, then let \( P(s'|s, p) = 0 \) for every \( p \in A_s \).

\[ R : S_T \to \mathbb{R} \text{ is the reward function, defined on terminal states only. For each terminal state } s = (O, U) \in S_T, \]

\[ R(s) = \begin{cases} \alpha(O) & \alpha(O) = \beta(O), \\ \mathbb{E} \left[ \max \{\beta(O), \max_{a_i \in U} X_i\} \right] & \alpha(O) < \beta(O). \end{cases} \]

Next, we prove that there exists an optimal policy for the R-GMDP with a significantly reduced support.

\[ \text{Lemma 3. For every policy } \pi \text{ for R-GMDP, there exists a stationary policy } \pi' \text{ such that} \]

1. \( \pi'(s) = \pi'(s') \) for every pair of states \( s = (O, U) \) and \( s' = (O', U') \) with \( \alpha(O) = \alpha(O') \) and \( \beta(O) = \beta(O'). \)
2. For every state \( s \), \( W(\pi', s) \geq W(\pi, s) \).

The proof of the lemma is identical to the proof of Lemma 1 and hence omitted. Lemma 3 suggests that we can focus on strategies that distinguish between states based on \( U, \alpha(O) \) and \( \beta(O) \) solely. The reduced state space does allows self loop by selecting \( a_1 \), without having any effect on the reward. It is thus straightforward to see that an optimal strategy that ignores \( a_1 \) exists, with a reward of exactly \( W(\pi^*, s_0) \).

\[ ^4\text{The crucial difference between R-GMDP and GMDP is in the action space and the transition probabilities, colored in red for readability.} \]
Notice that $M^{(2)}$ defines a non-stationary policy $\pi$ for the R-GMDP, by mimicking the actions (distributions) $\pi$ selects. When $M^{(2)}$ gets to an auspicious history or could not explore anymore, the policy $\pi$ gets to a terminal state and obtains a reward. Each time $M^{(2)}$ directs an agent, that agent gets at most the maximal reward $M^{(2)}$ discovered; hence, $SW(M^{(2)})$ is less or equal to the reward obtained by that non-stationary policy $\pi$, which is at most $W(\pi^*, s_0)$.

This completes the proof of the theorem. \hfill \Box

**Proof of Lemma 2.** Let $N_1, N_2$ denote the r.v. representing the number of agents in the explore and experience phases, respectively. Notice that the definition of social welfare given in Equation 2 can be interpreted as

$$SW(FEE) = \frac{1}{n} \left( E \left( \sum_{l=1}^{N_1+N_2} X_{M(h_l)} \right) + E \left( \sum_{l'=N_1+N_2+1}^{n} X_{M(h_{l'})} \right) \right).$$

(12)

Observe that every agent in the explore and experience phases obtains the reward of arm $a_1$ in expectation. Moreover, every agent in the exploit phase obtains $W(\pi^*, s_0)$ in expectation; hence, Equation (12) can be rearranged as

$$SW(FEE) = \frac{1}{n} \left( E \left( \sum_{l'=1}^{N_1+N_2} X_1 \right) + E \left( \sum_{l'=N_1+N_2+1}^{n} W(\pi^*, s_0) \right) \right)$$

$$= \frac{1}{n} \left( \mu_1 E (N_1 + N_2) + W(\pi^*, s_0) E(n - N_1 - N_2) \right)$$

$$= W(\pi^*, s_0) - \frac{1}{n} E (N_1 + N_2) (W(\pi^*, s_0) - \mu_1).$$

To finalize the proof, recall that $N_1 \leq K$ almost surely since there are $K$ arms that could be explored, and on every step in the exploration phase exactly one arm gets explored. Moreover, due to Observation 3 it holds that $E(N_2) \leq KH$; hence,

$$SW(FEE) \geq W(\pi^*, s_0) - \frac{1}{n} (K + KH) (W(\pi^*, s_0) - \mu_1).$$

\hfill \Box

**C. Omitted Proofs from Section 4**

**Proof of Theorem 3.** It is immediate to see that IC-FEE is EAIR. Satisfying EAIR follows from mixing FEE, which is EAIR, with GREEDY, which satisfies the delegate property and hence also the EAIR constraint, and recommendations of $a_1$.

Moreover, IC-FEE is asymptotically optimal since, after finitely many agents, its recommendations will coincide with those of FEE, and FEE is asymptotically optimal. While IC-FEE is not exploiting (Line 7), its recommendations coincide with those of FEE at least once per phase. Since the expected exploration time of FEE is $O\left(\frac{KH^2}{n}\right)$ (see Lemma 2), IC-FEE explores for $O\left(\frac{BK^2}{n}\right)$ rounds in expectation.

Showing that IC-FEE satisfies IC is trickier. We divide the analysis to several parts:

- The first agent gets $a_1$, which is the a-priori best action.
- Agents $2, \ldots, K$ either get recommendations from GREEDY (Line 3) or are recommended $a_1$ (Line 4). In the former, agents get the best arm known to the mechanism. In the latter, the only new information agents could learn is that $X_1 \geq \mu_K$; thus, for every $a_j \neq a_1$ it holds that

$$\mathbb{E}(X_1 - X_j | X_1 \geq \mu_K) \overset{iid}{\geq} \mathbb{E}(X_1 - X_j) \geq 0.$$

Agents cannot know if they are being recommended by Line 3 or Line 4, but in both cases they are better off with accepting the recommendation; hence, IC holds for agents $2, \ldots, K$. 


Agents $K + 1, \ldots, n$, and the recommended arm is $a_1$. This case might be trivial at first glance, but it is not as innocent. IC-FEE can recommend $a_1$ via Lines 8 and 12. In both cases, we know that $a_1$ is the best among all the explored arms. Nevertheless, there could still be unexplored arms with an expected value greater than $R_1$. One such scenario is when $R_1$ revealed by the first agent yielded $\mu_2 \leq R_1 \leq \mu_3$. In this case, IC-FEE recommends agent $K + 1$, assuming that she was not selected to be the exploring agent, arm $a_1$. Nevertheless, according to IC-FEE’s information at that point, arm $a_2$ is the best arm. Recommending $a_2$ greedily might disallow the mechanism to explore more arms using the mixture FEE employs, which leads to sub-optimal social welfare.

Nevertheless, we will show that IC holds in this case as well. Fix an agent $l$ and some phase $k$, and assume IC-FEE recommended agent $l$ arm $a_1$. Let $E_O^l$ denote the event indicating that $O \subseteq A$ arms where observed just before agent $l$ arrives, and $X_1 \geq X_i$ for every $a_i \in O$. Clearly, agent $l$ does not know whether $E_O^l$ occurs or not, but she can compute the occurrence probability. We have that

$$
\mathbb{E}(X_1 - X_j | M=a_1) = \mathbb{E}(X_1 - X_j | X_1 > \mu_2) \Pr(X_1 > \mu_2) + \sum_{O \subseteq A} \mathbb{E}(X_1 - X_j | X_1 \leq \mu_2, E_O^l) \Pr(X_1 \leq \mu_2, E_O^l).
$$

In addition,

$$
\mathbb{E}(X_1 - X_j \mid X_1 \leq \mu_2, E_O^l) \geq \mathbb{E}(X_1 - X_j \mid X_1 \leq \mu_2).
$$

This inequality follows immediately if $a_j \in O$. Otherwise, if $a_j \notin O$, due to the i.i.d. assumption, $X_1 - X_j$ could only increase conditioning on $E_O^l$; hence,

$$
\text{Eq.}(13) \geq \mathbb{E}(X_1 - X_j | X_1 > \mu_2) \Pr(X_1 > \mu_2) + \mathbb{E}(X_1 - X_j | X_1 \leq \mu_2) \Pr(X_1 \leq \mu_2).
$$

We conclude that agents $K + 1, \ldots, n$ follow IC-FEE when it recommends $a_1$.

- Agents $K + 1, \ldots, n$, and the recommended arm is $a_i \neq a_1$. Fix an agent $l$ and some phase $k$, and assume IC-FEE recommended agent $l$ arm $a_i \neq a_1$. We need to show that for every $a_j$, it holds that $\mathbb{E}[X_i - X_j \mid M = a_i] \geq 0$. Due to Assumption 1, there exists $\xi > 0, \gamma > 0$ such that

$$
\forall i \in [K]: \quad \Pr(\forall i' \in [K] \setminus \{i\} : \mu_i - X_i' > \xi) > \gamma.
$$

In words, Assumption 1 guarantees that with positive probability $\gamma$, all arms but $i$ have a reward that is less than $\mu_i$ by at least $\xi$. Denote this event by $\mathbb{1}_{a_i}$. If $\mathbb{1}_{a_i}$ occurs, we are guaranteed that arm $a_i$ will be explored in Line 3. Moreover, denote by $\mathbb{1}_{l,\text{exp}}$ the event that agent $l$ is the agent selected by IC-FEE to explore in Line 10. We have that

$$
\mathbb{E}[X_i - X_j | M = a_i] = \mathbb{E}[X_i - X_j | M = a_i, \mathbb{1}_{l,\text{exp}}] \Pr(\mathbb{1}_{l,\text{exp}}) + \mathbb{E}[X_i - X_j | M = a_i, \mathbb{1}_{l,\text{exp}}^\perp] \Pr(\mathbb{1}_{l,\text{exp}}^\perp)
$$

$$
\geq \frac{-H}{B} + \mathbb{E}[X_i - X_j | M = a_i, \mathbb{1}_{l,\text{exp}}, \mathbb{1}_{a_i}] \Pr(\mathbb{1}_{l,\text{exp}}, \mathbb{1}_{a_i}) + \mathbb{E}[X_i - X_j | M = a_i, \mathbb{1}_{l,\text{exp}}, \mathbb{1}_{a_i}^\perp] \Pr(\mathbb{1}_{l,\text{exp}}, \mathbb{1}_{a_i}^\perp)
$$

$$
\geq \frac{-H}{B} + \frac{\xi \gamma (B - 1)}{B},
$$

and the latter is non-negative if $B \geq \frac{2H}{\xi \gamma} + 1$.

Overall, we showed that every agent is better off by accepting IC-FEE’s recommendation; hence, IC-FEE is IC.

D. Omitted Proofs and Claims from Section 5

D.1. Ex-Post Individual Rationality

Notice that EAIR mechanisms guarantee each agent the value of the default arm, but only in expectation. We now propose a more strict form of individual rationality, ex-post individual rationality (EPIR).

Definition 4 (Ex-Post Individual Rationality). A mechanism $M$ is ex-post individually rational (EPIR) if for every agent $l \in \{1, \ldots, n\}$, every value $R_1$ in the support of $X_1$, every history $h = (h_1, \ldots, h_{l-1}) \in (A \times \mathbb{R}_+)^{l-1}$, and every arm $a_r$ such that $\Pr(M(h) = r) > 0$, it holds that $\mathbb{E}(X_r \mid h) \geq R_1$. □
Algorithm 4 IC EPIR Explore & Exploit (IC-EP-FEE)

1: Initialize an instance of $M_{\text{EPIR}}$ and update it after every recommendation.
2: Recommend as $\text{GREEDY}$ to agents $1, 2, \ldots, K$.
3: Split the remaining rounds into consecutive phases of $B$ rounds each.
4: for phase $k = 1, \ldots$ do
5: if $M_{\text{EPIR}}$ exploits then
6: follow $M_{\text{EPIR}}$
7: else
8: Pick an agent $l(k)$ from the $B$ agents in this phase uniformly at random.
9: Every agent in this phase is recommended as $\text{GREEDY}$, except agent $l(k)$ who is recommended according to $M_{\text{EPIR}}$.

Satisfying EPIR means that the mechanism never recommends an arm that is a priori inferior to arm $a_1$. Noticeably, every EPIR mechanism is also EAIR, yet EPIR mechanisms are quite conservative, since they can only explore arms that yield expected rewards of at least the value $R_1$ obtained for $a_1$.

An optimal EPIR mechanism is immediate in case of non-strategic agents; we denote by $M_{\text{EPIR}}$ this intuitive mechanism. First, explore arm $a_1$, and observe $R_1$. Then, remove all arms $a_r$ with $\mu_r < R_1$, and name the obtained set $A'$. Then, proceed with FULL-EXPLORATION on $A' \cup \{a_1\}$.

For the case of strategic agents, $M_{\text{EPIR}}$ is not enough: agents might be reluctant to explore arms with a-priori low rewards. We propose IC-EP-FEE, which is an asymptotically optimal IC and EPIR mechanism. IC-EP-FEE relies on the same technique we use in Section 4 and is outline in Algorithm 4.

**Theorem 4.** Let the phase length be $B = \left\lceil \frac{H}{\xi \gamma} \right\rceil + 1$. Under Assumption 1, IC-EP-FEE satisfies EPIR and IC. In addition, $SW_n(\text{IC-FEE}) \geq \text{OPT}_{\text{EPIR}} - O \left( \frac{K H^3}{K \xi \gamma} \right)$.

The proof of Theorem 4 is similar to that of Theorem 3, and is hence omitted.

D.2. Omitted Proofs from Subsection 5.1

**Proof of Proposition 2.** Let $X_1$ be such that $\Pr(X_1 = 1) = 1$, and for every $i$ such that $2 \leq i \leq K$ let

$$X_i = \begin{cases} 0 & \text{w.p. } 1 - \frac{1}{H} + \epsilon \\ H & \text{w.p. } \frac{1}{H} - \epsilon \end{cases},$$

for a small positive constant $\epsilon$. Clearly, $\mu_1 = 1$ while $\mu_i < 1$ for $2 \leq i \leq K$; hence, $\text{OPT}_{\text{EAIR}} = 1$. On the other hand,

$$\text{OPT} = \mathbb{E}(\max_{1 \leq i \leq K} X_i) = \Pr(\max_{2 \leq i \leq K} X_i = H)H + \Pr(\max_{2 \leq i \leq K} X_i = 0) \cdot 1 = \left( 1 - (1 - \frac{1}{H} + \epsilon)^{K-1} \right)H + (1 - \frac{1}{H} + \epsilon)^{K-1}.$$ 

Taking $\epsilon$ to zero, we get that $\text{OPT}$ is arbitrarily close to

$$\left( 1 - (1 - \frac{1}{H})^{K-1} \right)H + (1 - \frac{1}{H})^{K-1} = H \left( 1 - (1 - \frac{1}{H})^K \right).$$

Finally, we use the fact that $e^{-x} \geq (1 - \frac{x}{n})^n$ whenever $|x| \leq n$. By setting $n = K$ and $x = \frac{K}{H}$, we conclude that $e^{-\frac{K}{H}} \geq (1 - \frac{1}{H})^K$; therefore,

$$\frac{\text{OPT}}{\text{OPT}_{\text{EAIR}}} \geq H \left( 1 - e^{-\frac{H}{n}} \right).$$
Proof of Proposition 3. Let $X_1, X_2, \ldots, X_K$ such that
\[ X_1 = \begin{cases} 1 & \text{w.p. } \frac{1}{H} - \frac{1}{n-1} - \epsilon, \\ 0 & \text{w.p. } 1 - \frac{1}{n-1} + \epsilon \end{cases}, \quad X_2 = \begin{cases} 2 & \text{w.p. } 1, \\ \forall 3 \leq i \leq K: \quad X_i = \begin{cases} 0 & \text{w.p. } \frac{1}{n} + \epsilon, \\ H & \text{w.p. } \frac{1}{n} - \epsilon \end{cases}. \]

It holds that
\[ \text{OPT}_{\text{EPIR}} = H \left( \frac{1}{H-1} + \epsilon \right) + 2 \left( 1 - \frac{1}{H-1} - \epsilon \right). \]

On the other hand,
\[ \text{OPT}_{\text{EAIR}} = H \left( \frac{1}{H-1} + \epsilon \right) + 2 \left( 1 - \frac{1}{H} + \epsilon \right)^{K-2} + H \left( 1 - \left( 1 - \frac{1}{H} + \epsilon \right)^{K-2} \right). \]

Taking $\epsilon$ to zero, we get
\[ \frac{\text{OPT}_{\text{EAIR}}}{\text{OPT}_{\text{EPIR}}} \geq \frac{H + 2}{3} \left( 1 - e^{-\frac{K-2}{H}} \right). \]

Proof of Proposition 4. Let $X_1, X_2, \ldots, X_K$ such that
\[ X_1 = \begin{cases} 0 & \text{w.p. } \frac{1}{2} - \epsilon, \\ 2 & \text{w.p. } 1 \end{cases}, \quad X_2 = \begin{cases} 1 & \text{w.p. } 1, \\ \forall 3 \leq i \leq K: \quad X_i = \begin{cases} 0 & \text{w.p. } 1 - \frac{1}{n} + \epsilon, \\ H & \text{w.p. } 1 - \frac{1}{n} - \epsilon \end{cases}. \]

For $\epsilon \to 0$. It holds that $\text{OPT}_{\text{DEL}} = \frac{1}{2} \cdot 2 + \frac{1}{2} \cdot 1 = 1.5$. On the other hand,
\[ \text{OPT}_{\text{EPIR}} = \frac{1}{2} \cdot 2 + \frac{1}{2} \left( \left( 1 - \frac{1}{H} \right)^{K-2} + H \cdot \left( 1 - \left( 1 - \frac{1}{H} \right)^{K-2} \right) \right) \]
\[ \leq 1 + \frac{H}{2} \left( 1 - e^{-\frac{K-2}{H}} \right); \]

thus, $\frac{\text{OPT}_{\text{EPIR}}}{\text{OPT}_{\text{DEL}}} \geq \frac{H}{3} \left( 1 - e^{-\frac{K-2}{H}} \right). \]

Proposition 7. Fix $K, H \in \mathbb{N}$. Let $X_1 \sim \text{Uni}[H]^+$, and let $X_1 = \begin{cases} \text{Uni}[H]^+ & \text{w.p. } 1 - \epsilon, \\ H & \text{w.p. } \epsilon \end{cases}$ for arbitrarily small $\epsilon > 0$. It holds that $\frac{\text{OPT}_{\text{EAIR}}}{\text{OPT}_{\text{EPIR}}} \leq \frac{8}{7} + O(\epsilon)$.

Proof of Proposition 7. Assume for simplicity that $H$ is even. First, by simple probability tricks one can show that
\[ \text{OPT} = \mathbb{E}(\max_{1 \leq i \leq K} X_i) = (1 - \epsilon) \frac{K}{K+1} H + \epsilon H = \frac{K}{K+1} H + O(\epsilon). \]

Second, since $\mathbb{E}(X_i) > \mathbb{E}(X_i)$ for every $i \in \{2, \ldots, K\}$, any EAIR mechanism must explore $X_1$ first. Notice that $\max_{2 \leq i \leq K} \mu_i = \frac{H}{2}$. Thus,
\[ \text{OPT}_{\text{EAIR}} = \Pr(X_1 > \frac{H}{2}) \mathbb{E}(X_1 | X_1 > \frac{H}{2}) + \Pr(X_1 \leq \frac{H}{2}) \mathbb{E}(\max_{1 \leq i \leq K} X_i | X_1 \leq \frac{H}{2}) \]
\[ \geq \frac{1}{2} - \epsilon \frac{3H}{2} + \epsilon H + \frac{1}{2} \mathbb{E}(\max_{2 \leq i \leq K} X_i) \]
\[ = \frac{3H}{8} + \frac{1}{2} \frac{K-1}{K} H + O(\epsilon). \]
By taking $\epsilon$ to zero and applying standard manipulations, we obtain

$$\frac{\text{OPT}}{\text{OPT}_{\text{EAIR}}} \leq \frac{K^2}{\frac{7}{8} K^2 + \frac{3}{8} K - \frac{1}{2}}.$$  

This term attains $\frac{16}{15}$ for $K = 2$ and is monotonically increasing for $K \geq 3$; hence, the claim is proven by taking $K$ to infinity.

**E. Incentive Compatible Mechanism for Strategic Agents and Uniform Arrival**

In this section, we consider strategic agents and uniform arrival. Formally, we assume that the $n$ agents arrive in a random order, $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$, where $\sigma$ is selected uniformly at random from the set of all permutations. We show that FEE satisfies IC as is, assuming that there are sufficiently many agents. We introduce the following quantity $\sigma_i = \Pr(\forall i' \in [K] \setminus \{i\} : X_i > X_{i'})$. In words, $\sigma_i$ is the probability that arm $a_i$ is superior to all other arms. Clearly, if Assumption 1 holds, $\sigma_i > 0$ for every arm $i \in [K]$. In addition, let $\delta = \min_{i \in [K]} \sigma_i$. Lemma 4 implies that if there are $\text{poly}(H, K, \frac{1}{\delta})$ agents, then FEE is IC.

**Lemma 4.** Under Assumption 1 and uniform arrival, if $n \geq \frac{24H^2}{\delta} \max\{K, H \ln \frac{4H}{\delta}\}$, then FEE is IC.

**Proof of Lemma 4.** To prove the statement, we need to show that whenever an agent is recommended arm $a_r$, her best response is to select arm $a_r$. We focus on an arbitrary agent, and present the analysis from her point of view. In addition, if $r = 1$, either she is the first agent to arrive at the system or no better arm was discovered, resulting in $a_1$ being a best response. Otherwise, $r \neq 1$. We define the following events: let $E_{\text{rec}}$ be the event indicating that FEE recommends arm $a_r$ to the agent; $E_{\text{open}}^r$ indicates whether arm $a_r$ was recommended to some agent; and $E_{\text{opt}}^r$ indicates whether $a_r$ is an optimal arm. All of these events are defined w.r.t. the distribution over histories and the agent arrival distribution. Due to the uniform arrival distribution, the probability of $E_{\text{rec}}^r$ matches the proportion of agents who are recommended arm $a_r$. We proceed by analyzing the odds of being recommended $a_r$. Due to the definition of $\epsilon$ and the way FEE works when it observed a superior arm,

$$\Pr(E_{\text{opt}}^r \mid E_{\text{open}}^r) \geq \delta, \quad \Pr(\overline{E_{\text{opt}}^r} \mid E_{\text{open}}^r) \leq 1 - \delta. \quad (15)$$

Next, we present a lemma that gives a large deviation bound on the number of agents needed for the experience phase.

**Lemma 5.** Let $Q(\epsilon) = \max\{2KH, 2H^2 \ln \frac{1}{\epsilon}\}$. The experience phase terminates after $Q(\epsilon)$ agents w.p. of at least $1 - \epsilon$.

The proof of Lemma 5 and other claims we use in this lemma appear just after the end of this proof. For simplicity, denote $Q = Q(\epsilon)$. Conditioning on $E_{\text{open}}^r$, arm $a_r$ is either recommended exactly once (in case its reward is observed to be inferior to another arm during the execution), or several times. The latter can only happen if $R_r > R_1$ and arm $a_r$ is used by FEE to explore other, unobserved arms. In this case, Lemma 5 implies that would not happen more than $Q$ times, w.h.p. As a result,

$$\Pr(E_{\text{rec}}^r \mid E_{\text{opt}}^r, E_{\text{open}}^r) \leq (1 - \epsilon) \frac{Q + n}{n} + \epsilon \frac{n - Q}{n} \leq \frac{Q + 1 + \epsilon n}{n}.$$  

Moreover,

**Observation 1.** For every history $h$ such that $E_{\text{opt}}^r, E_{\text{open}}^r$ occur, if FEE already reached the exploit phase (Line 19) under $h$, then $\Pr(\text{FEE}(h) = a_r) = 1$.

Due to Observation 1, we also conclude that

$$\Pr(E_{\text{rec}}^r \mid E_{\text{opt}}^r, E_{\text{open}}^r) \geq (1 - \epsilon) \frac{n - Q - K}{n}. \quad (17)$$

We now analyze the ratio between the probability of arm $a_r$ being optimal and the probability that it is not, given $E_{\text{rec}}^r$. We
Applying the bounds from Equations (15), (16) and (17) to Equation (18), we get

\[
\frac{\Pr(E_{opt}^r \mid E_{rec}^r, E_{open}^r)}{\Pr(E_{opt}^r \mid E_{rec}^r, E_{open}^r)} = \frac{\Pr(E_{rec}^r, E_{opt}^r, E_{open}^r)}{\Pr(E_{rec}^r, E_{open}^r)} \frac{\Pr(E_{rec}^r, E_{opt}^r, E_{open}^r)}{\Pr(E_{rec}^r, E_{open}^r)} = \frac{\Pr(E_{rec}^r, E_{opt}^r, E_{open}^r)}{\Pr(E_{rec}^r, E_{opt}^r, E_{open}^r)} \frac{\Pr(E_{opt}^r \mid E_{rec}^r, E_{open}^r) \Pr(E_{rec}^r \mid E_{open}^r) \Pr(E_{opt}^r \mid E_{rec}^r, E_{open}^r)}{\Pr(E_{opt}^r) \Pr(E_{rec}^r \mid E_{open}^r) \Pr(E_{opt}^r \mid E_{rec}^r, E_{open}^r)}.
\]

Applying the bounds from Equations (15), (16) and (17) to Equation (18), we get

\[
\frac{\Pr(E_{opt}^r \mid E_{rec}^r, E_{open}^r)}{\Pr(E_{opt}^r \mid E_{rec}^r, E_{open}^r)} \geq \frac{\delta(1 - \epsilon)(n - Q - K)}{(1 - \delta)(Q + 1 + \epsilon n)}.
\]

and by rearranging we obtain

\[
\Pr(E_{opt}^r \mid E_{rec}^r, E_{open}^r) \geq \Pr(E_{opt}^r \mid E_{rec}^r, E_{open}^r) \delta(1 - \epsilon)(n - Q - K) / (1 - \delta)(Q + 1 + \epsilon n).
\]

Next, we bound the expected difference between the reward of arm \(a_r\) and that of an arbitrary arm \(a_i\), with \(i \neq r\). We have

\[
E(X_r - X_i \mid E_{rec}^r) = E(X_r - X_i \mid E_{rec}^r, E_{open}^r) = E(X_r - X_i \mid E_{rec}^r, E_{open}^r) \Pr(E_{opt}^r \mid E_{rec}^r, E_{open}^r) + E(X_r - X_i \mid E_{rec}^r, E_{open}^r, E_{opt}^r) \Pr(E_{opt}^r \mid E_{rec}^r, E_{open}^r) \geq 1 \cdot \Pr(E_{opt}^r \mid E_{rec}^r, E_{open}^r) - H \cdot \Pr(E_{opt}^r \mid E_{rec}^r, E_{open}^r).
\]

By plugging in the bound obtained in Equation (19) to Equation (20) we get

\[
E(X_r - X_i \mid E_{rec}^r) \geq \Pr(E_{opt}^r \mid E_{rec}^r, E_{open}^r) \left( \frac{\delta(1 - \epsilon)(n - Q - K)}{(1 - \delta)(Q + 1 + \epsilon n)} - H \right).
\]

Ultimately, since

**Observation 2.** Let \(\epsilon = \frac{\delta}{4H}\) and \(Q = \max\{2KH, 2H^2 \ln \frac{4H}{\delta}\}\). If \(n \geq \frac{6HQ}{\delta}\), it holds that

\[
\frac{\delta(1 - \epsilon)(n - Q - K)}{(1 - \delta)(Q + 1 + \epsilon n)} \geq H.
\]

The proof is completed by combining Observation 2 with Equation (21) to show that \(E(X_r - X_i \mid E_{rec}^r) \geq 0\) for every arm \(a_i\).

\(\square\)

**Proof of Lemma 5.** Let \(z = \frac{\delta}{4H}\) denote the number of agents receiving recommendations in the experience phase (Lines 16 and 18). The proof is based on two observations: first, we show that \(Z\) is first-order stochastically dominated by an easy-to-analyze random variable. Then, we use a concentration bound to complete the proof.

**Observation 3.** For every \(z \in \mathbb{N}\),

\[
\Pr \left( \text{NBin}(K, \frac{1}{H}) \geq z \right) \geq \Pr(Z \geq z).
\]

Moreover, using Hoeffding’s inequality we have

**Claim 2.** Let \(\epsilon > 0\), \(K, H \in \mathbb{N}\), and let \(Q = \max\{2KH, 2H^2 \ln \frac{1}{\epsilon}\}\). It holds that

\[
\Pr \left( \text{NBin}(K, \frac{1}{H}) \geq Q \right) \leq \epsilon.
\]
By combining Observation 3 and Claim 2, we get

\[ \Pr(Z \geq Q) \leq \Pr(\text{NBin}(K, \frac{1}{H}) \geq Q) \leq \epsilon \]

This completes the proof of this lemma.

**Proof of Observation 1.** To see why Observation 1 holds, recall that if \( E_{\text{open}} \) occurs, then FEE revealed \( R_r \). Moreover, reaching Line 19 suggests that the experience phase is over; therefore, the rewards of all arms are revealed. Finally, since \( E_{\text{opt}} \) holds, FEE will pick it with probability 1.

**Proof of Observation 2.** First, notice that \( \epsilon < \frac{1}{2} \) and \( Q > K \); thus,

\[ \frac{\delta(1 - \epsilon)(n - Q - K)}{(1 - \delta)(Q + 1 + \epsilon n)} \geq \frac{\delta(n - 2Q)}{(2Q + \epsilon n)}. \]  

(22)

It suffices to show that the right-hand side of Equation (22) is greater or equal to \( H \). Now,

\[ \frac{\delta(n - 2Q)}{(2Q + \epsilon n)} \geq H \iff \frac{\delta}{2}(n - 2Q) \geq H(2Q + \epsilon n) \iff \frac{\delta n}{2} - H \geq 2HQ + \epsilon Hn \]

\[ \iff \frac{\delta n}{2} - \epsilon Hn \geq 2HQ + \delta Q \iff n \left( \frac{\delta}{2} - \epsilon H \right) \geq 2HQ + \delta Q \iff n \geq \frac{Q(2H + \delta)}{(\frac{\delta}{2} - \epsilon H)}. \]  

(23)

Inserting the values of \( \epsilon \) and \( Q \), we argue that the statement holds as long as

\[ n \geq \max\left\{ 2KH, 2H^2 \ln \frac{4H}{\epsilon} \right\} \frac{(2H + \delta)}{(\frac{\delta}{2} - \epsilon H)}. \]  

(24)

To conclude the proof, recall that \( n \geq \frac{12HQ}{\delta} \); hence

\[ n \geq \frac{12H \max\left\{ 2KH, 2H^2 \ln \frac{4H}{\epsilon} \right\}}{\delta} \geq \frac{4 \max\left\{ 2KH, 2H^2 \ln \frac{4H}{\epsilon} \right\} (2H + \delta)}{\delta}; \]

thus, Equation (24) holds.

**Proof of Claim 2.** First, observe that

\[ \Pr\left( \text{NBin}(K, \frac{1}{H}) \geq Q \right) = \Pr\left( \text{Bin}(Q, \frac{1}{H}) \leq K \right). \]  

(25)

Next, notice that \( k \leq \frac{Q}{2H} \); thus,

\[ \text{Eq. (25)} \leq \Pr\left( \text{Bin}(Q, \frac{1}{H}) \leq \frac{Q}{2H} \right). \]  

(26)

By using the multiplicative version of the Chernoff Bound, we get that

\[ \text{Eq. (26)} \leq e^{-\frac{Q}{2H^2}}. \]  

(27)

Recall that \( Q \geq 2H^2 \ln \frac{1}{\epsilon} \); therefore,

\[ e^{-\frac{Q}{2H^2}} \leq e^{-\frac{2H^2 \ln \frac{1}{\epsilon}}{2H^2}} = \epsilon. \]

(28)

**Proof of Observation 3.** The exploration phase of FEE is based on \( \pi^* \). Once \( \pi^* \) reaches a terminal state, there are two options:
The terminate state exhibits $\beta = R_1$. In this case, the statement of the If sentence in Line 7 is false, and there is no need for experience. Consequently, $Z = 0$ w.p. 1 and the statement holds.

The terminate state exhibits $\beta > R_1$. In this case, FEE enters the While loop in Line 8. In each iteration of the While loop, either the size of $U$ decreases by 1 (Lines 12 and 16), or stays the same (Line 18). The statement in Line 18 will only execute if the arm $a_i$ selected in Line 10 satisfies $\mu_i < R_1$, otherwise the If condition in Line 15 would execute; hence, the probability of executing Line 18 is bounded by

$$\Pr \left( \text{Uni}(0, 1) \geq \frac{R_i - R_1}{R_i - \mu_i} \right) \leq \Pr \left( \text{Uni}(0, 1) \geq \frac{1}{H} \right) = 1 - \frac{1}{H}.$$

This applies for every iteration of the While loop. Recall that there are at most $K - 2$ arms needed to be explored, and hence the statement holds.

\qed

### E.1. The Full Exploration Mechanism

**Proposition 8.** Under Assumption 1 and uniform arrival, FULL-EXPLORATION is IC and asymptotically optimal.

**Proof of Proposition 8.** Asymptotic optimality is straightforward. The proof of being IC goes along the lines of Theorem 4 and hence omitted.

### F. Elaborated Example of FEE

In this section, we provide an elaborated example of the way FEE operates. Consider $K = 4$ arms, $H = 40$ and $X_1 \sim \text{Uni}(0, 40), X_2 \sim \text{Uni}(0, 30), X_3 \sim \text{Uni}(0, 20), X_4 \sim \text{Uni}(0, 10)$; thus $\mu_1 = 20, \mu_2 = 15, \mu_3 = 10, \mu_4 = 5$. As always, $a_1$ is the default arm. Let us assume for the sake of this example that $X = (X_1, X_2, X_3, X_4) = (6, 3, 7, 2)$, but these values are not known to the algorithm. We illustrate $\pi^*$ in Figure 1, obtained from a simple Python program.

Nodes with a square frame are associated with states of the GMDP. The leaves are terminal states, and the intermediate nodes are non-terminal. Blue circled nodes are auxiliary, and separate between values the newly observed arm can take. The outgoing edges from each non-terminal white node are the transition probabilities. For instance, in $v_1$, the outgoing edges are $p_{2,4}(2)$ and $p_{2,4}(4)$, hinting that the action taken in $v_1$ is $p_{2,4}$.

The colored leaves represent terminal states. Green leaves are terminal states where the policy revealed an arm with a value greater than $R_1$, i.e., $\beta > R_1$ (see Line 7 in FEE). Yellow leaves are terminal states in which $\pi^*$ reveals all the rewards, but those are less or equal to $R_1$. And the red node, $v_3$ refers to the terminal state in which $a_2, a_3$ were explored and $R_2, R_3$ were less or equal to $R_1$, and $a_4$ was not explored. Notice that $v_5$ and $v_{11}$ are associated with the same state, and since $\pi^*$ is stationary, their sub-trees are identical. Per our assumption on $X$, the GMDP will reach one of the leaves in $\{v_4, v_7, v_{13}\}$, depending on the coin flips. To illustrate, we assume that $\pi^*$ reached $v_4$ and explain the trajectory.

The root of the tree, $v_0$, denotes the initial state $s_0$. Due to the construction of the optimal policy $\pi^*$, it will always explore $a_1$ in the first round (level 0 of the tree in Figure 1); thus, FEE recommends the first agent $a_1$, and observes that $R_1 = 6$ (recall we assume the rewards are according to $X$). The GMDP then transitions to $v_1$. At $v_1$, $\pi^*$ picks $p_{2,4}$. FEE then draw coins (Line 4), which realized with $a_2$ (since we assume the leaf $v_4$ was realized eventually), and selects $a_2$ for the second agent. The value of $R_2 = 3$ is then observed, and the GMDP moves $v_2$. $\pi^*$ picks $p_{3,4}$, FEE draw coins (Line 4), which realized with $a_3$, and selects $a_2$ for the second agent. The value of $R_3 = 7$ is realized, and the GMDP reaches $v_4$, which is a terminal state. FEE exists the while loop in Line 3. FEE then enters the if statement of Line 7, since it observe that $\beta = R_3 > R_1$. At this point, the set of unobserved arms $U$ is $\{a_4\}$, and so FEE enters the while loop in Line 8. In Line 9, it sets $a_\tau = a_3$, following by setting $a_i = a_4$ in the subsequent line. Since there is a positive probability that $X_4 > R_3$, FEE skips the if block in Line 11.

Then, in Line 14, FEE draws $Y \sim \text{Uni}[0, 1]$. Since $\mu_4 = 5 \leq 6 = R_1$, the second condition of the if block in Line 15 does not hold; hence, the only way to enter the if block in Line 15 is by having $Y \leq \frac{R_3 - R_1}{R_3 - \mu_4} = \frac{2}{3} = 0.5$. If $Y > 0.5$, FEE moves to Line 18, recommends $a_3$ to the fourth agent, and starts another iteration of the while loop in Line 8. With
probability 1, after finitely many agents, FEE will draw $Y \leq 0.5$. Then, it will recommend $a_4$ in Line 16, and observe $R_4$. In Line 17, $U$ becomes the empty set. FEE will then exit the while loop in Line 8, move to Line 19, and every subsequent agent will exploit—FEE will recommend $a_3$ from then on.