Course project: Fiduciary Bandits and consideration of uncertainty

Mechanism Design for Data Science
Rotem Ben Zion
rotm@campus.technion.ac.il

1 Introduction

In this project, I consider the paper Fiduciary Bandits (henceforth FB) [1] by Gal Bahar, Omer Ben-Porat, Kevin Leyton-Brown and Moshe Tennenholtz. In FB, the authors define individual guarantees in the Multi-Armed-Bandits (MABs) setting, meant to assert trust in recommender systems. They proceed to develop a recommendation mechanism which holds the individual guarantees, while still optimizing social welfare. On the one hand, when the proposed individual guarantees are met, following the given recommendation yields the maximal expectation in every agent's viewpoint. On the other, maximizing pure expectation is not necessarily the agent's goal in real-life applications; uncertainty may also have a large impact on trust. In this project, I define a new metric for agent satisfaction and use it to compare the base-version algorithm developed in FB (called "FEE") to baselines. Motivated by the idea of variance as a measure of uncertainty, I empirically test the effect of arm variance in the MAB setting on this metric. Finally, I generate alternative individual guarantees tailored to it.
In section 2, I present the motivation behind FB, followed by their definitions and contributions. I then address the course material relevant to FB in section 3. My claim on the importance of uncertainty is presented in section 4, including the alternative metric, individual guarantees and the method I use to analyze them. Section 5 shows experimental results and qualitative analysis.

2 Fiduciary Bandits

2.1 Motivation

In the setting of multi-armed-bandits (MABs), 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. In FB, the arms are assumed to be deterministic, in the sense that after an arm is realized, it will constantly produce the same value for all future agents. The motivation to FB resides in a simplified version of recommender systems, where a recommendation mechanism uses past experience to suggest actions, and updates its belief for future iterations based on the actions taken by sequentially-arriving agents. The recommender seeks to optimize social welfare (sum of all agents' rewards), while individuals wish to maximize their own utility. An example given in FB is navigation applications (such as Waze or Google maps), where agents are drivers and actions (arms) are routes. MABs in general and recommender systems in particular deal with a problematic tradeoff between exploration, which may discover improved actions for future agents, and exploitation, which gives the current best reward. Moreover, in recommender systems, too much exploration has an additional setback: individual agents may be better off avoiding the recommendation. In other words, blindly attempting to maximize social welfare may "sacrifice" some of the users, and thus raise trust issues.
In FB, the authors face this challenge by defining *individual guarantees* - guarantees that the system

should fulfill for each agent independently from other agents and their recommendations. The goal of these guarantees is to balance the total social welfare maximization with individual interests.

2.2 Contributions

FB explores a novel compromise between the two extremes, which they term *ex-ante* individual rationality (EAIR). They additionally define a more demanding concept of *ex-post* individual rationality (EPIR). They analyze the classical MABs case, which is equivalent to agents that always accept the recommendation, as well as the more challenging case where agents are strategic and may choose a different arm than the one suggested. A novel recommendation algorithm called Fiduciary Explore & Exploit (FEE) is developed, and is shown to be EAIR and to obtain the highest possible social welfare by any EAIR mechanism up to an additive factor of $o(\frac{1}{n})$. Finally, they design an asymptotically optimal Incentive Compatible (IC) and EPIR mechanism, and analyze the social welfare cost of adopting either EAIR or EPIR mechanisms. FB is the first to define individual guarantees to agents in this setting.
In this project, I will focus on the basic FEE algorithm with non-strategic agents.

2.3 MABs , Individual Compatibility and social welfare

In this subsection and the next one, I list the definitions used and developed in FB. The notation is taken verbatim from the original paper.
Let $A = \{a_1, \ldots, a_K\}$ be a set of K arms (actions). Rewards are deterministic but initially unknown: the reward of arm a_i is a random variable X_i, and $(X_i)_{i=1}^K$ are mutually independent. We denote by R_i and observed value of X_i. Rewards are realized only once; $X_i = R_i$ for the rest of the execution after a_i is chosen for the first time. Denote by μ_i the expected value of X_i, and assume w.l.o.g. that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. Additionally assume X_i is fully supported on the set $[H]^+ \equiv \{0, 1, \ldots, H\}$. There are n agents, arriving sequentially. Denote by a^l the action of the l'th agent and R^l her reward. Agents are fully aware of the distribution of $(X_i)_{i=1}^K$.
A mechanism is a recommendation engine that interacts with agents. The input for the mechanism at stage l is the sequence of arms pulled and rewards received by the previous $l-1$ agents. The output of the mechanism is a recommended arm for agent l. Formally, a mechanism is a function $M : \bigcup_{l=1}^n (A \times \mathbb{R}_+)^{l-1} \to \Delta(A)$. The mechanism has a global objective, which is to maximize agents' social welfare: $\sum_{l=1}^n R^l(a^l)$.
The first definition refers to the agent scheme in which agents are strategic, and wish to maximize their own reward. In this case, Incentive Compatibility (IC) means that following the mechanism's recommendation is a dominant strategy.

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 $a_r, a_i \in A$,*

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

If agents are non-strategic or the mechanism is IC, we can assume that all agents follow their recommendation. Thus, we define the mechanism's (expected) social welfare by

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

Where $X_{M(h_l)} = \sum_{r=1}^K Pr_{M(h_l)}(a_r)\mathbb{E}(X_r|h_l)$. Note that $X_{M(h_l)}$ depends on the randomness of the rewards and, possibly, the randomness of $M(H_l)$.

2.4 Individual guarantees

A mechanism is *delegate* if for every agent $l \in \{1, \ldots, n\}$, every history $h \in (A \times \mathbb{R})^{l-1}$ and every distribution $\mathbf{p}$ over A, it holds that $\mathbb{E}(X_{M(h)}|h) \geq \sum_{r=1}^K \mathbf{p}(r)\mathbb{E}(X_r|h)$. This is the strongest individual guarantee and leads to maximal exploitation, named GREEDY. Therefore, this mechanism probably leads to low social welfare. On the other far end, we have the FULL-EXPLORATION mechanism, which first explores all arms sequentially, and then exploits the best arm. The latter is

optimal when the number of agents is large enough, but doesn't hold individual guarantees for the first K agents.
The following individual guarantee is defined by FB, and demands that the recommendation generated by the mechanism leads **in expectation** to a reward at least as good as choosing X_1, which has the (a-priory) largest expectation.

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

$$\sum_{r=1}^{K} Pr_{M(h)}(a_r)\mathbb{E}[X_r|h] \geq \mathbb{E}(X_1|h).$$

Notice that EAIR mechanisms guarantee each agent the value of the default arm, but only in expectation. The next definition proposed in FB is a more strict form of individual rationality, *ex-post* individual rationality (EPIR).

Definition 3 (*Ex-Post* Individual Rationality) *A mechanism M is ex-post individually rational (EPIR) if for every agent $l \in \{1, \dots, n\}$, 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 $\mathbb{E}(X_r - X_1|h) \geq 0$.*

2.5 Fiduciary Explore & Exploit

The main technical contribution of FB is an EAIR mechanism that asymptotically achieves optimal social welfare for an EAIR algorithm up to an additive constant of $o(\frac{1}{n})$. The mechanism, which they term Fiduciary Explore & Exploit (FEE), is described as Algorithm 1. It consists of three stages:

1. **Primary exploration.** The mechanism compares the default arm a_1 to whichever other arms are permitted by the individual rationality constraint. 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, or another a_i was found to be superior to a_1.
2. **Secondary exploration.** 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.
3. **Exploitation.** After no more exploration is possible / needed, the final phase is exploitation, where the mechanism recommends the most rewarding arm observed.

For the primary exploration phase, the MAB game is modeled as a Goal Markov Decision Process (GMDP), which is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$, where:

- $\mathcal{S}$ is a finite set of states. Each state s is a pair (O, U), where:
  - $O \subseteq \{(a, c) | a \in A, c \in [H]^+\}$ is the set of arm-reward pairs that have been observed so far, with each a appearing at most once in O (since rewards are deterministic): $\forall a \in A, |\{c | (a, c) \in O\}| \leq 1$.
  - $U \subseteq A$ is the set of arms not yet explored.
  - The initial state is thus $s_0 = (\emptyset, A)$.

  For notational convenience, denote $\alpha(O)$ to be the reward observed for arm a_1, noting that the construction forces arm a_1 to be chosen first, making it well-defined within the algorithm. In addition, denote $\beta(O) = \max_{c:\exists a,(a,c) \in O} c$ to be the maximal reward observed.
- $\mathcal{A} = \bigcup_{s \in \mathcal{S}} \mathcal{A}_s$ is an infinite set of actions. For each $s = (O, U) \in \mathcal{S}$, $\mathcal{A}_s$ is defined as follows:
  1. If $s = s_0$, then $\mathcal{A}_{s_0} = \Delta(\{a_1\})$: a deteministic 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 $\mathbf{p} \in \mathcal{A}_s$ if and only if

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

Denote by $\mathcal{S}_T$ the set of *terminal* states, namely $\mathcal{S}_T = \{s \in \mathcal{S} | \mathcal{A}_s = \emptyset\}$.

- $\mathcal{P}$ is the transition probability function. Let $s = (O, U) \in \mathcal{S}$ 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 $\mathbf{p}$ is defined by $\mathcal{P}(s'|s, \mathbf{p}) = \mathbf{p}(a_i)Pr(X_i = c)$. If these conditions are not met, transition probability will be zero.
- $\mathcal{R} : \mathcal{S}_T \to \mathbb{R}$ is the reward function, defined on terminal states only. For each terminal state $s = (O, U) \in \mathcal{S}_T$,

$$\mathcal{R}(s) = \begin{cases} \alpha(O) & \alpha(O) = \beta(O) \\ \mathbb{E}[\max\{\beta(O), \max_{a_{i'} \in U} X_{i'}\}] & \alpha(O) < \beta(O) \end{cases}$$

A policy is a function from all GMDP histories to an action. However, the authors of FB prove that it is sufficient to consider *stationary* policies, which are functions of the state only, so we can denote $\pi : \mathcal{S} \to \mathcal{A}$. A policy is *valid* if $\forall s \in \mathcal{S} : \pi(s) \in \mathcal{A}_s$. Lastly, they denote by $W(\pi, s)$ the expected reward of π when initialized from s by the recursive formula:

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

In order to find to optimal policy π^*, denote $d(\mu) = |\alpha(O) - \mu|$ and $\bar{i} = r, \bar{r} = i$. They proceed to prove (in slightly different notation) that it must take the following form:

$$\mathbf{p}_{ir}^{\alpha}(a) = \begin{cases} \frac{d(\mu_{\bar{a}})}{d(\mu_i) + d(\mu_r)} & \text{if } a \in \{i, r\} \\ 0 & \text{otherwise.} \end{cases}$$

and $\mathbf{p}_{ii}^{\alpha}(a) = 1$ if and only if $a = i$. Moreover, the optimal policy π^* holds $\pi^*(s_0) = \mathbf{p}_{11}$, and for every non-terminal state $s \neq s_0$, it holds $\pi^* = \mathbf{p}_{i^* r^*}$ such that $(i^*, r^*) \in \mathcal{A}_s$ maximize

$$\left(1 - \frac{I\{i = r\}}{2}\right) \left[\mathbf{p}_{ir}(i) \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s, \mathbf{p}_{ii}) W(\pi^*, s') + \mathbf{p}_{ir}(r) \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s, \mathbf{p}_{rr}) W(\pi^*, s')\right]$$

The optimal policy can therefore be found using dynamic programming.

Theoretical analysis of FEE

FEE satisfies the EAIR condition. Denote by $\text{OPT}_{\textbf{EAIR}}$ the highest welfare attained by any EAIR mechanism. FB shows a lower bound on the social welfare of FEE in Lemma 2.2:

Theorem 2.1 *It holds that $\mathit{OPT}_{\mathit{\textbf{EAIR}}} \leq W(\pi^*, s_0)$.*

Lemma 2.2 *$SW_n(FEE) \geq \mathit{OPT}_{\mathit{\textbf{EAIR}}} - O(\frac{KH^2}{n})$.*

They proceed to characterize $\text{OPT}_{\textbf{EAIR}}$ by comparing it to OPT, the actual optimal social welfare:

Proposition 2.3 *For every $K, H \in \mathbb{N}$, there exists an instance $\langle K, A, (X_i) \rangle$ with*

$$\frac{\mathit{OPT}}{\mathit{OPT}_{\mathit{\textbf{EAIR}}}} \geq H(1 - e^{-\frac{K}{H}}).$$

Let $\text{OPT}_{\textbf{EPIR}}$ and $\text{OPT}_{\textbf{DEL}}$ be the optimal social welfare achievable by an EPIR mechanism and delegate mechanism respectively. These propositions bound the relationships between them:

Proposition 2.4 *For every $K, H \in \mathbb{N}$, there exists an instance $\langle K, A, (X_i) \rangle$ with*

$$\frac{\mathit{OPT}_{\mathit{\textbf{EAIR}}}}{\mathit{OPT}_{\mathit{\textbf{EPIR}}}} \geq \frac{H + 2}{3}(1 - e^{-\frac{K-2}{H}}).$$

Proposition 2.5 *For every $K, H \in \mathbb{N}$, there exists an instance $\langle K, A, (X_i) \rangle$ with*

$$\frac{\mathit{OPT}_{\mathit{\textbf{EPIR}}}}{\mathit{OPT}_{\mathit{\textbf{DEL}}}} \geq \frac{H}{3}(1 - e^{-\frac{K-2}{H}}).$$

Algorithm 1 Fiduciary Explore & Exploit (FEE)

```
Initialize a GMDP instance $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$, and compute $\pi^*$.
Set $s = (O, U) = (\emptyset, A)$.
**while** s is not terminal **do**
    Draw arm $a_i \sim \pi^*(s)$, recommend $a_i$ and observe $R_i$.
    $O \leftarrow O \cup \{(a_i, R_i)\}, \quad U \leftarrow U \setminus \{a_i\}$.
    $s \leftarrow (O, U)$.
**end while**
**if** $\beta(O) > R_1$ **then**
    **while** $U$ is not empty **do**
        Let $a_{\tilde{r}} \in \arg\max_{a_r \in A \setminus U} R_r$.
        Select an arbitrary arm $a_i \in U$.
        **if** $Pr(X_i > R_{\tilde{r}}) = 0$ **then**
            $U \leftarrow U \setminus \{a_i\}$
            **continue**
        **end if**
        Draw $Y \sim Uni[0, 1]$
        **if** $Y \leq \frac{R_{\tilde{r}} - R_1}{R_{\tilde{r}} - \mu_i}$ or $\mu_i \geq R_1$ **then**
            Recommend $a_i$ and observe $R_i$
            $U \leftarrow U \setminus \{a_i\}$
        **end if**
    **end while**
**end if**
Recommend $a_{i^*} \in \arg\max_{a_i \in A \setminus U} R_i$ to all agents.
```

3 Connection to course material

Some of the subjects that we have discussed throughout the course are relevant to the subject of this paper. Recommendation systems are the main motivation behind FB, and can be approached using the definitions and algorithms FB has developed. Since MABs is a general setting, it can be used in many choice-making situations - for example, in On-Line Advertising, agents (customers) arrive sequentially and are given a recommendation (ad) by the mechanism (in the case of a single slot). We could model the CTR as a random variable and get a special case of On-Line Advertising as a MABs setting. Lastly, since FB assumes that the history is available to the mechanism but not to the agents, it can be modeled as a game of incomplete information, where agents play versus the mechanism and their actions are whether or not to accept the recommendation.

4 A new challenge: probability instead of expectation

4.1 Intuition

The battle of exploration versus exploitation is a well studied topic. In the field of Reinforcement Learning, this is often tackled with the concept of *uncertainty*: every exploration move alleviates some uncertainty about the world, which may help exploitation in the future. This approach allows strategies such as the celebrated *optimism in the face of uncertainty* [2]. In MABs, uncertainty lies in the unobserved arms; only once realized, we know their deterministic reward. This raises two crucial notes:

1. When performing exploration, agents do not necessarily want to maximize plain expectation; they may prefer an action that yields a good reward with high certainty (low risk) than an action that yields a lower reward most of the time, and much higher reward with low probability (even if the latter has higher expectation). As a concrete example consider the following MAB setting with free parameter $T > 10$:

$$H = T,\ K = 2,\ X_1 = \begin{cases} 1 & \text{w.p. } \frac{T-10}{T} \\ T & \text{w.p. } \frac{10}{T} \end{cases},\ X_2 = 10 \text{ (w.p. 1)}$$

Although $\mathbb{E}[X_1] = 10 + \frac{T-10}{T} > 10 = \mathbb{E}[X_2]$, as $T \to \infty$[1] it is clear that most humans would prefer to take X_2. Note that recommending X_2 violates EAIR.

2. The uncertainty of a state/action **cannot** be characterized using expectation alone. This will require using deeper moments, such as the variance of the unobserved arm's distribution.

Combining the two notes, we conclude that in some plausible scenarios, an individual guarantee that utilizes only expectation will not accurately model the agents' satisfaction with the recommender mechanism.

4.2 Trust metric & individual guarantees

Given the set of distributions $\{X_i\}_{i=1}^{K}$ and history $h \in (A \times \mathbb{R})^{l-1}$, the individual guarantees in FB demand a lower bound on $\mathbb{E}[X_r|h]$ and $\mathbb{E}_{r\sim M(h)}\left[\mathbb{E}[X_r|h]\right]$ for EPIR and EAIR, respectively. Recall that $\beta(O) \equiv \max_{c:\exists a,(a,c)\in O} c$ was defined to be the maximal reward observed. An alternative option, that uses probability rather than expectation, will be to lower bound the value $Pr(X_r \geq \beta(O)|h)$ and $\mathbb{E}_{r\sim M(h)}\left[Pr(X_r \geq \beta(O)|h)\right]$ correspondingly. In words, we lower bound the probability to receive at least the best reward found so far. This yields the ***Ex-Ante*** **Individual Likely-Exploitation** (δ-EAILE) and ***Ex-Post*** **Individual Likely-Exploitation** (δ-EPILE) individual guarantees:

Definition 4 (*Ex-Ante* Individual Likely-Exploitation) *A mechanism M is δ ex-ante individually likely-exploiting (δ-EAILE) if for every agent $l \in \{1,\ldots,n\}$, and for every history $h \in (A \times \mathbb{R})^{l-1}$,*

$$\sum_{r=1}^{K} Pr_{M(h)}(a_r) Pr(X_r \geq \beta(O)|h) \geq \delta.$$

Definition 5 (*Ex-Post* Individual Likely-Exploitation) *A mechanism M is δ ex-post individually likely-exploiting (δ-EPILE) 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 $Pr(X_r \geq \beta(O)|h) \geq \delta$.*

Just like EPIR is a more strict form of individual rationality than EAIR, EPILE as a more strict version of EAILE.
We can now define what I term the *trust metric*, which evaluates a mechanism by finding the largest δ value for which it is δ-EAILE.

Definition 6 (Trust metric) *Let M be a mechanism. The trust metric is defined as*

$$TM(M) = \sup_{\delta\in[0,1]} \left\{ \delta : \quad \forall l \in \{1,\ldots,n\} \forall h \in (A \times \mathbb{R})^{l-1},\ \sum_{r=1}^{K} Pr_{M(h)}(a_r) Pr(X_r \geq \beta(O)|h) \geq \delta. \right\}$$

A few notes on these definitions:

- The parameter δ controls the tradeoff between social welfare and individual guarantee. When $\delta \to 0$, the criterion is easily met by any chosen arm, except ones that necessarily give sub-optimal rewards given the current knowledge, and social welfare can be maximized more easily. When $\delta \to 1$, meeting the criterion requires exploitation or "safe" exploration (unobserved arms that are likely to benefit a larger reward than the current best), increasing trust in the mechanism.
- Both GREEDY and FULL-EXPLORATION are not δ-EAILE nor δ-EPILE for $\delta > 0$, as shown by the example in the first note in 4.1.
- FEE is 1-EAILE in the exploitation phase and can easily be adapted to be δ-EAILE for any $\delta < 1$ in the secondary-exploration phase, but isn't δ-EAILE for any $\delta > 0$ during primary exploration.
- An algorithm that holds these conditions is likely to explicitly leverage additional moments other than expectation. For example, using Cantelli's inequality [3], we can get a lower bound on the probability of a random variable X to be greater or equal to $\beta(O)$ using its variance:

$$Pr(X - \mathbb{E}[X] \geq \lambda) \leq \frac{\sigma^2}{\sigma^2 + \lambda^2} \;\Rightarrow\; Pr(X \geq \beta(O) \leq \frac{\sigma^2}{\sigma^2 + (\beta(O) - \mathbb{E}[X])^2})$$

[1] We actually end up with Pascal's Wager.

4.3 Method

I test my definitions empirically by simulating the MABs setting. For each of the mechanisms (FEE, GREEDY, FULL-EXPLORATION), I approximate the trust metric by executing the MABs game, and calculating $Pr(X_r \geq \beta(O)|h)$ after every step. I then take the minimum of that value over all steps. For reliability, I perform this process multiple times and report the average. Moreover, I control a parameter named 'meta-variance', which sets the variance of **variance values** of the different arms in the game. In other words, when 'meta-variance' is small, the distributions of the different arms have relatively similar variances, and vice versa. The motivation to investigate this parameter is that a large value of 'meta-variance' is likely to lead to a situation similar to the example in the first note in 4.1, where a mechanism that satisfies EAIR may easily fail at EAILE.

Random Variables

To create the random variables (arms), I chose to create distributions in which at most two entries receive a positive probability. Given two integers H (upper bound on the support set) and μ (expectation), I choose two arms l, u such that $1 \leq l \leq \mu \leq u \leq H$. I then define

$$\mathbf{p}_l = \begin{cases} 1 & \text{if } l = u \\ \frac{u-\mu}{u-l} & \text{if } l \neq u \end{cases}$$

$$\mathbf{p}_u = 1 - \mathbf{p}_l = \begin{cases} 1 & \text{if } l = u \\ \frac{\mu-l}{u-l} & \text{if } l \neq u \end{cases}$$

And the random variable (arm) X will be

$$X = \begin{cases} l & \text{w.p.} & \mathbf{p}_l \\ u & \text{w.p.} & \mathbf{p}_u \end{cases}$$

Notes:

- $\mathbb{E}[X] = l \cdot \mathbf{p}_l + u \cdot \mathbf{p}_u = \mu$
- $\text{Var}[X] = (\mu - l) \cdot \mathbf{p}_l + (u - \mu) \cdot \mathbf{p}_u = 2 \cdot \frac{(\mu-l)(u-\mu)}{u-l}$
- The variance ranges from zero (when $l = u$) to $2(\mu - 1)\frac{H-\mu}{H-1}$ (when $l = 1, u = H$)

Reward calculation in terminal states

By the reward definition in FEE, we need to calculate

$$\mathcal{R}(s) = \begin{cases} \alpha(O) & \alpha(O) = \beta(O) \\ \mathbb{E}[\max\{\beta(O), \max_{a_{i'} \in U} X_{i'}\}] & \alpha(O) < \beta(O) \end{cases}$$

To allow feasible calculation when $\alpha(O) < \beta(O)$, we can write

$$\mathbb{E}[\max\{\beta(O), \max_{a_{i'} \in U} X_{i'}\}] = \sum_{k=\beta+1}^{H} [k \cdot \mathbf{p}_k] + \beta \cdot \left(1 - \sum_{k=\beta+1}^{H} \mathbf{p}_k\right)$$

With $\mathbf{p}_k$ being the probability that $\max\{\beta(O), \max_{a_{i'} \in U} X_{i'}\} = k$, which can be written as

$$\begin{aligned} \mathbf{p}_k &= \mathcal{P}(\forall i \in U : X_i \leq k \ \wedge \ \exists i \in U : X_i = k) \\ &= \mathcal{P}(\forall i \in U : X_i \leq k) - \mathcal{P}(\forall i \in U : X_i \leq k \ \wedge \ \forall i \in U : X_i \neq k) \\ &= \mathcal{P}(\forall i \in U : X_i \leq k) - \mathcal{P}(\forall i \in U : X_i < k) \\ &= \prod_{i \in U} \mathcal{P}(X_i \leq k) - \prod_{i \in U} \mathcal{P}(X_i < k) \\ &= \prod_{i \in U} \mathcal{P}(X_i \leq k) - \prod_{i \in U} \mathcal{P}(X_i \leq k - 1) \end{aligned}$$

5 Experiments and analysis

The fully reproducible code is available on GitHub[2]. In both of the following experiments, game parameters are:

- Upper bound on rewards: $H = 40$.
- Number of arms (actions): $K = 10$.
- Number of agents (turns): $n = 30$.
- Approximation of trust metric is averaged over 50 calculations for FEE and 500 calculations for the baselines.
- The variance of an arm's distribution is sampled $\sim N(7, \text{‘meta-variance‘})$ and then l, u are chosen to create a variance as close as possible to the result.

Experiment 1 - FULL-EXPLORATION performance

To understand the effect of the ‘meta-variance‘ parameter, I test the FULL-EXPLORATION mechanism. The result is shown in figure 1. We can see that as ‘meta-variance‘ increases, both social welfare and the number of EAIR violations rise. Both are expected: social welfare increases since the probability for high maximal rewards grows with added variance, and the number of EAIR violations increases because low-variance arms are more likely to cause a violation in expectation.

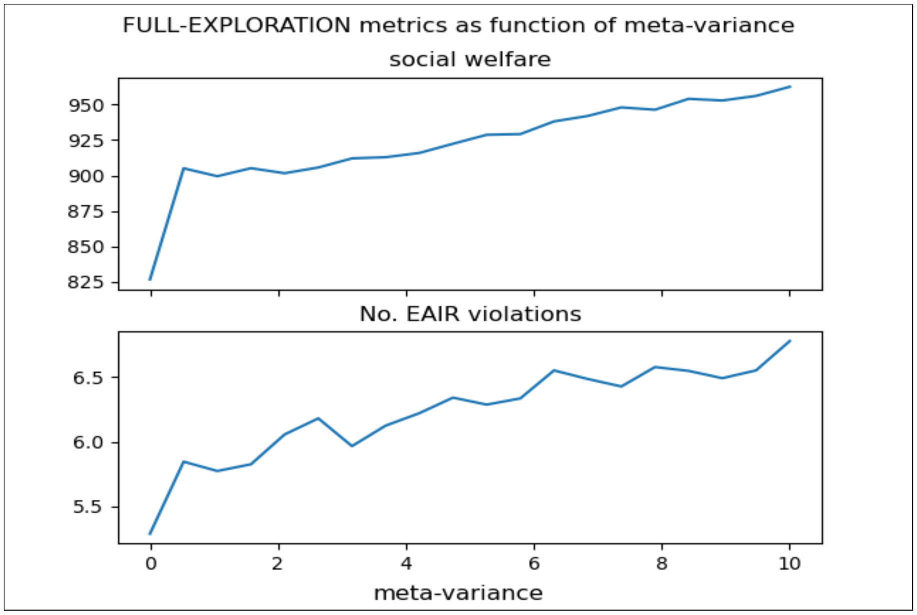

Figure 1: Performance of the FULL-EXPLORATION mechanism. On the top is the social welfare (sum of rewards) and on the bottom is the number of steps in which the recommendation violated the EAIR criterion.

Experiment 2 - Trust metric vs meta-variance

In this experiment, I report the approximated trust metric as explained in section 4. The FULL-EXPLORATION mechanism is not included in this experiment, because it has much smaller values. The result for FEE and GREEDY are reported in figure 2. The GREEDY algorithm has a higher trust metric than FEE, which is intuitive as it is *delegate* (defined in section 2). In addition, note that FEE always has a positive metric, which is the case because of lines 12-14 in algorithm 1 together with the EAIR constraint during primary exploration. We see a general increase of the trust metric with ‘meta-variance‘. I find this result counter-intuitive, because situations where expectation is not a good heuristic to the probability of a random variable being larger than a constant are more likely with variance difference. This result calls for further investigation.

[2] To reproduce the results, run game.py

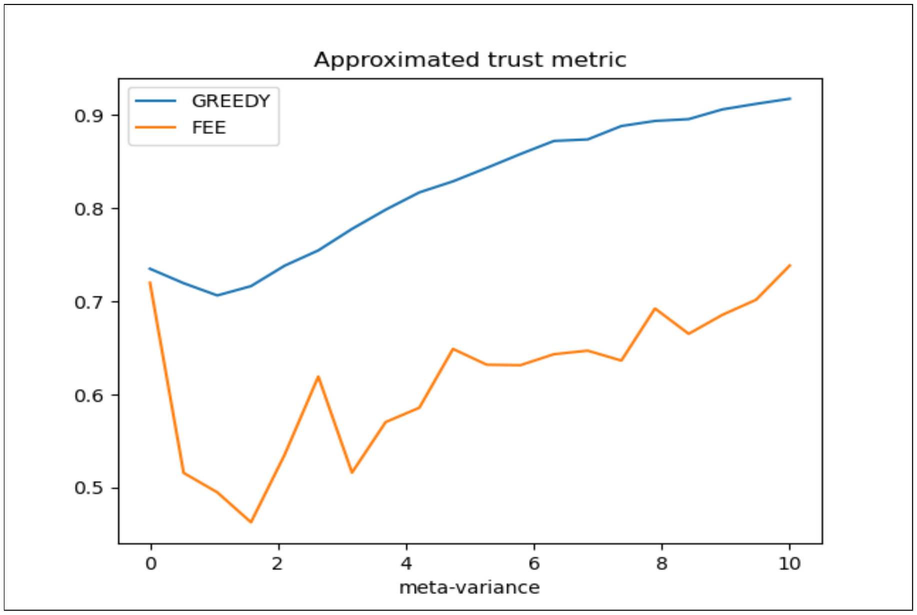

Figure 2: Approximated trust metric for each mechanism as a function of meta-variance.

References

[1] Gal Bahar, Omer Ben-Porat, Kevin Leyton-Brown, and Moshe Tennenholtz. Fiduciary bandits. In *International Conference on Machine Learning*, pages 518–527. PMLR, 2020.

[2] Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

[3] Wikipedia. Cantelli's inequality, 2022.