TEL AVIV UNIVERSITY אוניברסיטת תל-אביב
Tel Aviv University
Raymond and Beverly Sackler Faculty of Exact Sciences
The Blavatnik School of Computer Science

OPTIMAL ALGORITHM FOR BAYESIAN INCENTIVE-COMPATIBLE EXPLORATION

by

Lee Cohen

under the supervision of
Prof. Yishay Mansour

Thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science

2018

Abstract

We consider a social planner faced with a stream of myopic selfish agents. The goal of the social planner is to maximize the social welfare, however, it is limited to using only information asymmetry (regarding previous outcomes) and cannot use any monetary incentives. The planner recommends actions to agents, but her recommendations need to be Bayesian Incentive Compatible to be followed by the agents.

Our main results is an **optimal** algorithm for the planner, in the case that the actions realizations are deterministic and have a limited support, making significant important progress on this open problem. Our optimal protocol has two interesting features. First, it always completes the exploration of *a priori* more beneficial actions before exploring a priori less beneficial actions. Second, the randomization in the protocol is correlated across agents and actions (and not independent at each decision time).

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor Prof. Yishay Mansour for his helpful guidance throughout the various stages of this thesis. His endless patience and support, together with all the useful comments made me both enjoy and learn a lot during this research. Finally, I would like to thank my family and to my partner Oria who believed in me, encouraged me and supported me during my studies.

Contents

Chapter 1

Introduction

The inherent trade-off between exploration and exploitation is at the core of any reactive learning algorithm. Multi-arm bandit is a simple model which highlights this inherent trade-off. Multi-arm bandits can model a variety of scenarios, including pricing (where the actions are prices), recommendation (e.g., where actions are news articles) and many other settings.

To a large part, multi-arm bandits are viewed as a model for learning and optimization in which the planner can select any available action. However, when we are considering human agents as the entities performing the action, then incentives become a major issue. While a planner can recommend actions to the agents (in order to explore different alternatives), the agents ultimately decide whether to follow the recommendation given. This raises the issue of incentives in addition to the exploration-exploitation trade-off.

The planner can induce explorations in many ways. The simplest is using monetary transfers, paying the agents in order to explore (for example, Frazier et al. [FKKK14]). We are interested in the case when the social planner is unable or prefers to avoid any monetary transfers. (This can be due to regulatory constraints, business model, social norms, or any other reason.) The main advantage of the planner in our model is the information asymmetry, namely, the fact that the planner has much more information than the agents.

As a motivating example for information asymmetry, consider a GPS driving application. The application is recommending to the drivers (agents) the best route to drive, given the changing road delays, and observes the actual road delays when the route is driven. While the application can recommend driving routes, ultimately, the driver decides which route to actually drive. The application needs periodically to send drivers (agents) on exploratory routes, where it has uncertainty regarding the actual delay, in order to observe their delay. The driver (agent) is aware that the application has updated information regarding the current delays on various roads. For this reason, the driver (agent) would be willing to follow the recommendation even if she knows that there is a small probability that she is asked to explore. On the other extreme, if the driver would assume that with high probability a certain recommended route has a higher delay, she might drive an alternate route. This inherent balancing of exploration and exploitation while satisfying agents' incentives, is at the core of our work.

The abstract model that we consider is the following. There is a finite set of actions, and for each action there is a prior distribution on its rewards. A social planner is faced with a sequence of myopic selfish agents, and each agent appears only once. The social planner would like to maximize the social welfare, the sum of the agents' utilities. The social planner recommends to each agent an action, and if the recommendation is Bayesian incentive compatible, the agent will follow the action. This model was presented in Kremer et al. [KMP14] and studied in [MSS15, MSSW16, MSW18]. The work of Kremer et al. [KMP14] presented an optimal algorithm for the social planner in the case of *two* action with deterministic outcome. (Deterministic outcome implies that each time the action is performed we receive the same reward, and the uncertainty is what that value will be, which is govern by the prior distribution.)

Our main focus is to make progress on this important open problem of providing an optimal policy for this setting. For this end, we consider a somewhat more restricted setting, where each action has a finite support. If we assume that there are only two possible values, say $\{-1, +1\}$, then the task becomes trivial. We can simply order the agents according to their expectation, and ask them to explore until we reach an action of value $+1$, and then recommend it forever. This would work even if we provide the agents with the realizations of the previous actions. In this work we take a small, yet significant, step away from this trivial model. We assume that the best a priori action has a larger support. For the most part we analyze the case that the support of the a priori best action is $\{-1, 0, +1\}$, while the other actions have support $\{-1, +1\}$. We do extend our results to handle a more general setting of a continuous distribution with full support on $[-1, 1]$ for the a priori best action.

The simple model has a significant complexity and allows us to draw a few interesting insights. To understand the challenges, consider the case that the actions have a negative expected reward. (For simplicity, we assume that the actions are sorted by their expected reward, where action 1 has the highest expectation.) In such a case, if the realization of action 1 is $+1$, clearly the planner would recommend it for all the following agents. If the realization of action 1 is -1, clearly any other action is superior to it. However, the challenging case that the realization of action 1 is 0. In this case, the selfish agents would prefer to perform action 1 with 0 reward (since other actions have negative expected reward). The challenge to the social planner is to incentivize the agents to explore. The main idea is that of information asymmetry. When the planner recommends action 2, the agent is unsure whether the social planner observed that outcome of action 1 is -1, in which it would like to perform it, or whether the social planner observed that the outcome is 0 and asks the agent to explore. The social planner, by a delicate balancing of the exploration probability, can make the recommendation Bayesian incentive compatible.

Our main result is an **optimal** algorithm for the social planner when faced with k actions, both for support $\{-1, 0, +1\}$ and $[-1, +1]$ for the best apriori action. First, the algorithm makes sure that the BIC constraints are tight, which is a simple intuitive requirement and is clearly required for optimality. However, we need to exhibit much more refine properties to construct an optimal algorithm. An interesting issue regarding the exploration order is whether when we force a tight BIC constraint we might be forced to explore an action j before we know the values of actions $1, \dots, j-1$. We show that this is not the case in the optimal algorithm, namely, the exploration of action j starts only after the social planner knows the realizations of all the better a priori actions, i.e., $1, \dots, j-1$. While this seems like an intuitive outcome, it relies on the very delicate way in which our algorithm performs its randomization. (Recall that the recommendation algorithm uses randomization to balance between exploring and exploiting.) The implementation of the randomization is the second interesting property of our algorithm. In our randomization, we use a correlation between agents and actions. Specifically, the randomization selects for each action a random agent that might potentially explore it (if needed). Special care needs to be taken to make sure that for different actions we always select different agents.

We show that our algorithm does not only maximize the social welfare but in addition minimize the exploration time, the time until the social planner does not need to explore any more. For the most part we assume that the number of agents is large enough that the social planner completes the exploration. We show also how to derive the optimal policy in the case of a limited number of agents.

1.1 Related works

As mentioned, the work of Kremer et al. [KMP14] presented the model and derived the optimal policy for two deterministic actions. Mansour et al. [MSS15] derive tight asymptotic regret bound in the case of stochastic actions as well as a reduction from an arbitrary non-BIC policy to a BIC one. Bahar et al. [BST16] enrich the model by embedding the agents in a social network, and allowing them to observe their neighbors. Mansour et al. [MSSW16] extended the model to allow a multi-agent game in each time step, rather than a single agent. Mansour et al. [MSW18] consider the case of two competing planners.

Frazier et al. [FKKK14] consider a model with monetary transfers, where the social planner can pay agents to explore. Che and Hörner [CH13] consider a setting with two binary-valued actions and continuous information flow and a continuum of agents. Finally, Slivkins [Sli17] has an excellent overview of the topic.

A related topic is that of Bayesian Persuasion by Kamenica and Gentzkow [KG11] where the planner tries to infer a value of an "unobservable" state using interaction with multiple agents. See [DKQ16, DX, DX16] for a more algorithmic perspective of Bayesian Persuasion.

Multi-armed bandits [CBL06, GGW11] is a well-studied model for exploration-exploitation trade-off both in operations research and machine learning. The main focus in learning multi-arm bandits is on designing efficient algorithms that have a guaranteed performance compared to the best single action.

Chapter 2

Model

Let $A := \{1, 2, ..., k\}$ be the set of possible actions. The prior distribution $D = D_1 \times D_2 \times ... \times D_k$ defines random variables X_j for the rewards of actions $j \in A$. The reward of action $j \in A$, denoted by x_j, is sampled from D_j (it is sampled once, and any application of action j yields the same reward x_j).

In this work we focus on the case that the support of distribution of D_1 is $\{-1, 0, +1\}$ (the case of support $[-1, 1]$ appears in Section 5). The support of distribution D_j, for $j \geq 2$, is $\{+1, -1\}$. We denote by $p_j^\alpha := \Pr[X_j = \alpha]$, which implies that the distribution D_j, for $j \geq 2$, has a single parameter, p_j^1 (and $p_j^{-1} = 1 - p_j^1$). W.l.o.g. we assume that $p_j^1 > 0$, otherwise the action has a constant reward of -1.

The interaction between the planner and the agents proceeds as follows. At time t, the t-th agent arrives, and the planner recommends to the t-th agent action $\sigma_t \in A$, which is called *the recommended action*. Given the recommended action σ_t, the t-th agent selects an action a_t, receives a reward x_{a_t}, and leaves. Formally, the t-th agent has a *utility function* u_t and $u_t(a) = x_a$ if action a has been explored, else $\mathbb{E}[u_t(a)] = \mu_a$. A *history* at time t, h_t, contains all the previous chosen actions by the agents, i.e., $a_1, \ldots, a_t$, and their corresponding rewards, $x_{a_1}, \ldots, x_{a_t}$. A strategy for the planner is a *recommendation policy*, π, where $\pi_t(h_{t-1}) = \lambda_t \in \Delta(A)$, where $\Delta(A)$ is the set of distributions over A, i.e., $\Delta(A) = \{\lambda_t \in \mathbb{R}^k | \forall j \in A, \lambda_t[j] \geq 0 \; and \; \Sigma_{j=1}^k \lambda_t[j] = 1\}$. The value of $\lambda_t[j]$ is the probability that $\sigma_t = j$, i.e., $\lambda_t[j] = \Pr[\sigma_t = j]$.

A recommended action, σ_t, is *Bayesian incentive-compatible (BIC)* if for any action $j \in A$, we have $\mathbb{E}[u_t(\sigma_t) - u_t(j)] \geq 0$. Such constrains are called *BIC constrains*. I.e., there is no other action $j \in A$ that can increase agent t's expected reward, based on the prior D, the policy π, the recommended action σ_t, and the agent's place in line t, all of which the agent observe (note that the agent does not observe the history h_{t-1}). A recommendation policy for the planner, π, is BIC if all it's recommendations are BIC. Namely, for any agent t and any history h_{t-1}, the recommendation σ_t is BIC, i.e., for any action $j \in A$, we have $\mathbb{E}[u_t(\sigma_t) - u_t(j) | \sigma_t] \geq 0$.

The *social welfare* is the expected cumulative reward of all the agents. The social welfare of a BIC recommendation policy π is: $SW_T(\pi) := \mathbb{E}[\Sigma_{t=1}^T u_t(\sigma_t)] = \mathbb{E}[\Sigma_{t=1}^T u_t(\pi_t(h_{t-1}))]$,

The Bayesian prior D on the rewards, is a common knowledge to the planner as well as all the agents.

W.l.o.g, we restrict the planner's recommendation policy to be BIC, which assures that the agents follow the recommended actions. Our main goal is to design a BIC algorithm that maximizes social welfare (i.e., the cumulative reward of the agents).

Chapter 3

BIC Algorithm for $\text{supp}(D_1) = \{-1, 0, 1\}$

We start with a simpler case that will have most of the ingredients of the more general case. We restrict the first action to have only three possible values $\{-1, 0, 1\}$, namely, the support of D_1 is $\{-1, 0, 1\}$. The second restriction is that we assume that there are only three actions, i.e., $k = 3$. The terminology is provided for k-actions settings, but some of the intuition and motivation are provided for three actions settings. The proofs appear in Appendix B. The algorithm for the general case of $k > 3$ actions, and some of its proofs are in the appendix A.

Given this special case, we claim that the challenging case is when $0 > \mu_2 > \mu_3$. In the case that $\mu_1 > \mu_2 > \mu_3 > 0$, we can simply recommend to the first agent action 1, i.e., $\sigma_1 = 1$. When we observe x_1, then: (1) If $x_1 = 1$, we recommend to all the agents action 1, i.e., $\sigma_t = 1$. (2) If $x_1 = 0$ or $x_1 = -1$, we recommend to the second agent action 2, i.e., $\sigma_2 = 2$. This is BIC since $\mu_2 > 0 \geq x_1$ in this case. If $x_2 = 1$ we recommend to all the agents $\sigma_t = 2$. Otherwise, $x_2 = -1$, and we recommend to the third agent action 3, i.e., $\sigma_3 = 3$. Again, this is BIC since $\mu_3 > 0 \geq x_1 \geq x_2$. Either way, all the agents after the first three will be performing the optimal action. The above policy maximizes social welfare even if we do not restrict the information flow, and the planner announces to the agents the actions' realizations. In the case that $\mu_1 > \mu_2 > 0 > \mu_3$, we can execute for the first two agents the above strategy, and essentially reduce the number of actions to two, for which the optimal policy was given by Kremer et al. [KMP14]. For this reason, we assume that $0 > \mu_2 > \mu_3$. (And for k actions, we assume $0 > \mu_2 > \cdots > \mu_k$.)

To build intuition we start with a simple example, in order to explain how a BIC policy can give a recommendation $\sigma_t \neq 1$.

Example 1. *Consider a recommendation $\sigma_t = 2$ to agent t. The possible reasons for it is one of the following:*

1. ***Exploitation driven recommendation**: Action 2 is the best action given the history. This can be due to one of the following cases:*
   (a) ***A known reward**: The planner already observed that $x_2 = 1$, which is the maximum possible reward. From that time, the recommended action is $\sigma_t = 2$, as it has the maximum possible*

reward.

(b) ***An unknown reward****: The observed realizations have the minimum possible reward, i.e.,* $x_1 = -1$ *and maybe* $x_3 = -1$*. Given this realization, we know that* $u_t(2) = \mu_2 > x_1$ *(and in case that* $x_3 = -1$*, also* $u_t(2) = \mu_2 \geq x_3$*). This makes action* 2 *the best action to execute, considering the history.*

2. ***Exploration driven recommendation****: The planner has not yet observed an action with the best possible reward (i.e.,* 1*), and observed* $x_1 = 0$*. Since we assume that* $0 > \mu_2 > \mu_3$*, such a recommendation would not benefit for agent* t *(but the planner is recommending it since it might benefit future agents).*

Fortunately, the agents do not know the realizations of the actions' rewards, hence cannot infer the reason for their recommendations. This is where the *information asymmetry* translates into an advantage for the planner, and enable her to maximize social welfare.

3.1 Information States

It would be very useful to partition the histories depending on the information that the planner has, regarding the realized values of the actions. Since we have only three actions, we have at most three realized values, and we can encode them in a vector of length three. We use the $*$ symbol to indicate that a value is still unknown. For example, $\langle 0, -1, * \rangle$ implies that we know that $x_1 = 0$, $x_2 = -1$ and we never explored the value of X_3. Any history of the first $t-1$ agents which is compatible with $\vec{z} = \langle 0, -1, * \rangle$ is assigned to the information state $S_t^{\vec{z}}$. The recommendation to the t-th agent would depend on the planner's information state.

Note that the agents do not know the planner's information state. However, given the recommendation σ_t, and the planner policy π, they can deduce the probabilities of each state, conditioned on the recommendation σ_t they received. Those probabilities allow them to test whether the recommended action is indeed BIC, i.e., maximizes their expected reward given the information they observe.

Going back to example 1, we can now describe it using information states.

Example 2. *Consider a recommendation to agent* t*,* $\sigma_t = 2$*. Every possible reason for it can be one of the following:*

1. ***States that result in exploitation driven recommendation****, action* 2 *either has:*
   (a) ***A known reward****: The planner has already observed action* 2*'s reward and it is the maximum possible reward. I.e., the planner is in one of the following information states:* $S_t^{\langle -1,1,* \rangle}$*,* $S_t^{\langle -1,1,-1 \rangle}$ *or* $S_t^{\langle 0,1,* \rangle}$*.*

   (b) ***An unknown reward****: The only action with a better prior expected reward compared to action* 2*, action* 1*, has been explored and resulted in minimal reward (i.e,* $x_1 = -1$*). Action* 2 *that now has the best utility, has not yet explored. We denote this state with* $S_t^{\langle -1,*,* \rangle}$*. (An additional possible state is* $S_t^{\langle -1,*,-1 \rangle}$ *where the planner also observed that* $x_3 = -1$*.)*

The set of these exploitation states is denoted by Γ_t^{j+}, for the reason that following a recommendation for action j in such states produces higher expected utility for agent t compared to action 1.

2. ***States that may result in exploration driven recommendation**: Action 2 has not been explored yet, whereas $x_1 = 0$. This implies that the planner is either in information state $S_t^{\langle 0,*,* \rangle}$, or in information state $S_t^{\langle 0,*,-1 \rangle}$.*
The set of these exploration states is denoted by Γ_t^{j-}, for the reason that following a recommendation for action j in such states produces lower expected utility for agent t then selecting action 1.

3.2 The optimal BIC recommendation algorithm

Given the information states, we can describe the planner's recommendation policy. The recommendation policy will map the information states to recommended actions. In the case of an "Exploration driven Recommendation" the mapping would be stochastic, to make sure that the incentives are maintained. `Algorithm 3-actions` is described in Table 3.1, defining what recommendation to give in each information state.

`Algorithm 3-actions` uses two functions, $f_t^2(y)$ and $f_t^3(y)$, which control the exploration and are based on a mutual parameter y, which will be selected uniformly at random in $[0, 1]$. The states are marked also as *terminal states* if there is a unique recommendation for all future agents, and *exploration* if the recommended action might not have the highest expected reward ($S_t^{\vec{z}} \in \Gamma_t^{j-}$). States not marked as *exploration* result in a exploitation driven recommendation, and are therefore *exploitation* states ($S_t^{\vec{z}} \in \Gamma_t^{j+}$).

Looking at `Algorithm 3-actions` in Table 3.1 might be intimidating, however, in most of the information states the recommendations are rather straightforward. In the initial information state, i.e., $\langle *, *, * \rangle$, the only BIC recommendation is action 1, since the first agent knows that the planner has no additional information beyond the prior. In any information state in which some $x_i = 1$, the planner recommends that action, the agents get the maximum reward, and the state does not change (i.e., terminal state). In any information state in which all the realized actions are $x_i = -1$, the planner recommends an unexplored action with the highest expected reward, the agents get the maximum expected reward, and after it the state does change to include the new explored action.

The main challenge is in the cases that the realized value of action 1 is $x_1 = 0$ and $0 > \mu_2 > \mu_3$. In such information states we have a tension between the agent incentive, to perform action 1 and maximize her expected reward, and the planner incentive to explore new actions to the benefit of future agents. Indeed we have two information states in which we explore stochastically, balancing between the incentives of the agent and making the recommendation BIC. In information state $\langle 0, *, * \rangle$ the planner explores with some probability action 2, and in information state $\langle 0, -1, * \rangle$ the planner explores with some probability action 3.

We stress that the stochastic exploration is not done in a "independent" way, but rather in a coordinated way through the parameter $y \in [0, 1]$, which is selected initially uniformly at random, and never changes. The property that we will have is that while we are in information state $\langle 0, *, * \rangle$ we eventually have an agent

Recommendation Table. Policy Parameters: (y,t)				
State	Information state	Recommendation (σ_t)	Terminal	Exploration
$S_1^{\langle *,*,*\rangle}$	$X_1 = *$	1		
$S_t^{\langle 1,*,*\rangle}$	$X_1 = 1$	1	✓	
$S_2^{\langle -1,*,*\rangle}$	$X_1 = -1, X_2 = *$	2		
$S_t^{\langle -1,1,*\rangle}$	$X_1 = -1, X_2 = 1$	2	✓	
$S_3^{\langle -1,-1,*\rangle}$	$X_1 = -1, X_2 = -1, X_3 = *$	3		
$S_t^{\langle -1,-1,1\rangle}$	$X_1 = -1, X_2 = -1, X_3 = 1$	3	✓	
$S_t^{\langle -1,-1,-1\rangle}$	$X_1 = -1, X_2 = -1, X_3 = -1$	1	✓	
$S_t^{\langle 0,*,*\rangle}$	$X_1 = 0, X_2 = *$	$f_t^2(y) \in \{1,2\}$		✓
$S_t^{\langle 0,1,*\rangle}$	$X_1 = 0, X_2 = 1$	2	✓	
$S_t^{\langle 0,-1,*\rangle}$	$X_1 = 0, X_2 = -1, X_3 = *$	$f_t^3(y) \in \{1,3\}$		✓
$S_t^{\langle 0,-1,1\rangle}$	$X_1 = 0, X_2 = -1, X_3 = 1$	3	✓	
$S_t^{\langle 0,-1,-1\rangle}$	$X_1 = 0, X_2 = -1, X_3 = -1$	1	✓	
$S_t^{\langle 0,*,-1\rangle}$	$X_1 = 0, X_2 = *, X_3 = -1$	infeasible[corollary 14]		✓
$S_t^{\langle 0,*,1\rangle}$	$X_1 = 0, X_2 = *, X_3 = 1$	infeasible[corollary 14]	✓	
$S_t^{\langle 0,1,-1\rangle}$	$X_1 = 0, X_2 = 1, X_3 = -1$	infeasible[corollary 14]	✓	

Table 3.1: Algorithm 3-Action's recommendation policy

that explores action 2, and its index is $f^2(y)$. Similarly, while we are in information state $\langle 0,-1,*\rangle$ we eventually have an agent that tries action 3, and its index is $f^3(y)$. We need to take special care to make sure that agent $f^2(y)$, which explores action 2, is different than agent $f^3(y)$, which explores action 3. (Clearly, each agent can explore at most one action.) This is why we use a *coordinate sampling* (to be define later).

We also show that some information states are never reachable, namely, $\langle 0,*,1\rangle$, $\langle 0,*,-1\rangle$ and $\langle 0,1,-1\rangle$. This will be due to the fact that for any $y \in [0,1]$ we will show that $f^2(y) < f^3(y)$, which implies that we complete the exploration of action 2 before exploring action 3. As we extend to k actions, we use the same y to coordinate between the stochastic exploration of all the actions. Then again, by showing that for any pair of actions $i < j$, it holds that $f^i(y) < f^j(y)$, we deduce that the order in which the actions are explored is from the a priori highest expected reward to the lowest, i.e., $2, 3, \ldots, k$.

3.3 Exploration Rates

In this section we formalize the exploration rate that the planner can have. A *BIC exploration rate*, denoted by q, measures the probability that a BIC recommendation $\sigma_t = j$ is given when the planner is in some exploration state. Namely, for any BIC recommendation policy π, the BIC exploration rate is $\sum_{S_t^{\vec{z}} \in \Gamma_t^{j-}} \Pr_\pi[S_t^{\vec{z}}, \sigma_t = j]$, where $\Pr_\pi[S_t^{\vec{z}}, \sigma_t = j]$ is the probability that the planner is in $S_t^{\vec{z}}$ at time t, and recommends to explore action j, assuming that all the recommendations until the current agent use π.

Let $\hat{\pi}$ denote a BIC recommendation policy that recommends actions base on Table 3.1 (or Table A.1 for $k > 3$ actions) and uses maximum BIC exploration rates for every agent t and for every $j \in A$. *Maximal BIC exploration rate*, denoted by q_t^j is the maximum probability of exploration, subject to the BIC constraints, and bounded by the probability that the planner is in exploration state at time t with j as a recommended action. I.e., q_t^j is the solution of:

$$\begin{aligned} q_t^j = \max_q \quad & q \\ \text{s.t.} \quad & \sum_{S_t^{\vec{z}} \in \Gamma_t^{j+}} \Pr_{\hat{\pi}}[S_t^{\vec{z}} | \sigma_t = j] \, \mathbb{E}[u_t(j) - u_t(1) | S_t^{\vec{z}}, \sigma_t = j] + \mu_j \frac{q}{\Pr_{\hat{\pi}}[\sigma_t = j]} \geq 0 \\ & 0 \leq q \leq \sum_{S_t^{\vec{z}} \in \Gamma_t^{j-}} \Pr_{\hat{\pi}}[S_t^{\vec{z}}] \end{aligned} \tag{3.1}$$

The first constraint makes sure that σ_t is a BIC recommendation. Its first summand is a summation taken over each exploitation state probability, multiplied by the "gain" from choosing action j instead of action 1 in this state. The second summand is the "loss" of the agent, namely the prior expected reward of action j (i.e., μ_j), multiplied by the exploration rate q and divided by the probability of the event $\sigma_t = j$ (which includes also the exploration probability q). The terms "gain" and "loss" are from the agent's perspective. By looking at Table 3.1, we can see that when $\sigma_t = j$ is given in exploitation state, the expected utility difference is positive, therefore the agent has a "gain" of reward in these states. On the other hand, as we assume that $\mu_j < 0$, the agent has a "loss" of reward in the exploration states (all of which share $x_1 = 0$). When this entire expression is non-negative (i.e., the first constraint holds), it is BIC.

Notice that $\hat{\pi}$ is defined as a BIC policy, and as such every recommendation $\sigma_t = j$ is BIC, i.e., its BIC constraints must be met for every action $i \neq j$. We argue that in $\hat{\pi}$, if the BIC constraint of action j compared to action 1 is satisfied, all the other BIC constraints for agent t are met. Therefore, we only refer to the BIC constraint with respect to action 1 when calculating q_t^j. The reason is that for any pair of actions $a < b$, and for every $y \in [0, 1]$, we show that $f^a(y) < f^b(y)$, i.e., the exploration of action a is done before the exploration of action b. Along with Table 3.1 that represents the recommendations of $\hat{\pi}$, we deduce that whenever a recommendation $\sigma_t = j$ is given, the reward of action j is either unknown (i.e, the expected reward is $\mu_j > -1$) or X_j has been observed and $x_j = 1$. Now, for any action i such that $1 \neq i < j$, $f^i(y) < f^j(y)$ yields that X_i has been observed and $x_i = -1$. As for every action $j < i$, since

$f^j(y) < f^i(y)$ yields that X_i has not to been sampled yet, and from the assumption that $\mu_i < \mu_j$ we know that $\mu_i < \mu_j \le \mathbb{E}[u_t(j)]$. Either way $\mathbb{E}[u_t(\sigma_t) - u_t(i)] \ge 0$.

The second constraint in (3.1) prevents the exploration rate from exceeding the probability that the planner is in exploration state ($S_t^{\vec{z}} \in \Gamma_t^{j-}$). This guarantees that we can actually use of all of q to give an exploration driven recommendation. Namely, $q = \sum_{S_t^{\vec{z}} \in \Gamma_t^{j-}} \Pr_{\hat{\pi}}[S_t^{\vec{z}}, \sigma_t = j] \le \sum_{S_t^{\vec{z}} \in \Gamma_t^{j-}} \Pr_{\hat{\pi}}[S_t^{\vec{z}}]$.
Let n_j denote the index of *last agent* $t \le T$ that might explore action j, i.e., $n_j := argmax_t(q_t^j > 0)$. For convenience, for every agent t and action j, we denote

$$A_t^j := \begin{cases} 0 & t < j \\ \frac{2p_j^1 \Pi_{i<j} p_i^{-1} + p_j^1 \sum_{\tau=j}^{t-1} q_\tau^j}{1-2p_j^1} & t \ge j \end{cases}$$

and

$$B_t^j := \begin{cases} 0 & t < j \\ p_1^0 - \sum_{m=2}^{t-1} q_m^2 & t \ge j = 2 \\ p_{j-1}^{-1} \sum_{\tau=j-1}^{t-1} q_\tau^{j-1} - \sum_{\tau=j}^{t-1} q_\tau^j & t \ge j \ge 3 \end{cases}$$

3.4 Computing the Maximum BIC Exploration Rates

We now calculate the maximum BIC exploration rates. (The next lemma's proof for $k = 3$, namely, q_t^2 and q_t^3, is in Appendix B, and the proof for $k > 3$, i.e., any q_t^j, is in Appendix A.)

Lemma 3. *Given $q_2^2, \dots, q_{t-1}^2$, we have*

$$q_t^2 = \begin{cases} 0 & t = 1 \\ \min(A_t^2, B_t^2) & t \ge 2 \end{cases} \tag{3.2}$$

And for action $j \ge 3$, given q_τ^i for $i \le j-1$ and $\tau \le t-1$, assuming $q_{t-1}^{j-1} = A_{t-1}^{j-1}$ and $t \le n_{j-1}$, we have

$$q_t^j = \begin{cases} 0 & t < j \\ \min(A_t^j, B_t^j) & t \ge j \end{cases} \tag{3.3}$$

In addition we show that $q_t^j \le p_{j-1}^{-1} A_{t-1}^{j-1}$.

The next lemma derives the value of q_t^j (without an assumption on q_{t-1}^{j-1}).

Lemma 4. *For action $j \ge 3$, given q_τ^i for $i \le j-1$ and $\tau \le t-1$, such that $t > n_{j-1}$, we have*

$$q_t^j = \begin{cases} 0 & t < j \\ \min(A_t^j, B_t^j) & t \ge j \end{cases} \tag{3.4}$$

The following are consequences of Lemma 3 and Lemma 4.

Lemma 5. *For every $j > 1$, the exploration rate of agent j for action j is strictly positive, i.e., $q_j^j > 0$.*

The following lemmas relate the exploration rates q_t^j and the parameters A_t^j and B_t^j.

Lemma 6. *For every action j and agent $t > j$ and that $q_{t-1}^j > 0$, it holds that $A_j^j \leq A_{t-1}^j < A_t^j$.*

Lemma 7. *For every action j, for every $n_j \geq t > n_{j-1}$, it holds that*

1. $B_{t-1}^j > B_t^j \geq 0$.
2. $q_t^j > 0$.
3. *If $q_t^j = B_t^j (> 0)$, then it holds that $q_{t+i}^j = B_{t+i}^j = 0$ for every $i \geq 1$, therefore $t = n_j$ and we stop.*

3.5 Properties of the Exploration Rates

In this subsection we show properties regarding the exploration rates which later enable to show that $\hat{\pi}$ is well defined, that $\hat{\pi}$ eventually reaches a terminal state, and finally, that it maximizes expected social welfare.

The following theorem is a corollary to Lemmas 3 - 7, states the exact exploration rates.

Theorem 8. *For action 2 and for agent t we have,*

$$q_t^2 = \begin{cases} 0 & t = 1 \\ \frac{2p_1^{-1}p_2^1 + p_2^1 \sum_{m=2}^{t-1} q_m^2}{1-2p_2^1} & 2 \leq t < n_2 \\ p_1^0 - \sum_{m=2}^{n_2-1} q_m^2 & t = n_2 \\ 0 & t > n_2 \end{cases}$$

For action $j \geq 3$ and agent t we have,

$$q_t^j = \begin{cases} 0 & t < j \\ \frac{2p_j^1 \Pi_{i<j} p_i^{-1} + p_j^1 \sum_{\tau=j}^{t-1} q_\tau^j}{1-2p_j^1} & j \leq t < n_j \\ p_{j-1}^{-1} \sum_{\tau=j-1}^{n_j-1} q_\tau^{j-1} - \sum_{\tau=j}^{n_j-1} q_\tau^j & t = n_j \\ 0 & t > n_j \end{cases}$$

Let $\rho_j = p_1^0 \Pi_{i=2}^{j-1} p_i^{-1}$. We show that ρ_j is the total exploration rate of action j.

Lemma 9. *For $T \geq n_k$, the probability for exploration driven recommendation for any action j is ρ_j, i.e.,*

$$\Pr[\exists t : \sigma_t = j, S_t^{\vec{z}} \in \Gamma_t^{j-}] = \rho_j$$

3.6 Determining the explorers

We now explain how the algorithm chooses which agent will explore each action. Recall that the planner knows the history h_{t-1}, and therefore knows the current state at time t, as defined in Table 3.1. She then sets $\pi_t(h_{t-1}) = \lambda_t$ to be the corresponding recommendation for the current state in Table 3.1. Together with the policy parameters and the functions f^j that we later define in Definitions 10 and 11, respectively, she returns $\sigma_t \sim \lambda_t$ as the recommended action.

Definition 10. *a valid input for our algorithm is a triple $\langle y, t, Q\rangle$ such that:*

1. *$y \in [0,1]$ is a real number that is sampled from a uniform distribution in $[0,1]$.*
2. *t indicates the agent number (the agent for which the algorithm is run).*
3. *$Q = \{q^j | j \in \{2,...,k\}\}$ is a set that contains exploration rates vectors for each action excluding action 1, such that $q^j[t] := q_t^j$ (i.e, the exploration rate for agent t with j as recommended action).*

We now define the functions f^j that determines which agent will explore action j.

Definition 11. *Let $f^j : [0,1] \to \{1,...,n_j\}$ be the function that maps a real number $y \in [0,1]$ to an agent t such that*

$$f^j(y) = \text{argmax}_t(\sum_{\tau=1}^{t-1} q_\tau^j < y\rho_j)$$

Let $f_t^j : [0,1] \to \{1, j\}$ be the function for action j and agent t that maps a real number $y \in [0,1]$ to a recommendation for agent t, i.e., σ_t, and is defined as follows:

$$f_t^j(y) := \begin{cases} j & f^j(y) = t \\ 1 & else \end{cases}$$

The following lemma shows that different actions are explored by different agents, and that better a priori actions are explored always earlier.

Lemma 12. *For every $y \in [0,1]$, and for every action j, $f^j(y) < f^{j+1}(y)$.*

Since $f^j(y) = n_j$, Lemma 12 implies the following corollaries.

Corollary 13. *For every action j, it holds that $n_j + 1 \leq n_{j+1}$*

Corollary 14. *Action j is explored before any action $i > j$, making every state $S_t^{\vec{z}}$ such that $S_t^{\vec{z}}[j] = *$ and $S_t^{\vec{z}}[i] \neq *$ (e.g., $S_t^{\langle 0,*,-1\rangle}$) infeasible for every agent t. Namely,*

$$\Gamma_t^{j-} = \{S_t^{\vec{z}} | S_t^{\vec{z}}[1] = 0, \forall 1 < i < j : S_t^{\vec{z}}[i] = -1, \forall i \geq j : S_t^{\vec{z}}[i] = *]\}$$

Hence, $|\Gamma_t^{j-}| = 1$.

We finish this chapter by showing that $\hat{\pi}$ is well-defined in a sense that every agent gets exactly one action as a recommendation in Theorem 15.

Theorem 15. *Recommendation policy $\hat{\pi}$ is a well-defined recommendation policy, since for every $y \in [0, 1]$, and for pair of actions action $i \neq j$, $f^i(y) \neq f^j(y)$, and there exists $t \in \{1, \ldots, n_j\}$ such that $f^j(y) = t$. This implies that every agent receives a recommendation for exactly one action.*

Chapter 4

Optimality

4.1 Finite exploration

Clearly the flow between the information states is acyclic. From Table 3.1, when the planner is in a non-terminal state, she is exploring, with some probability. This implies that after n_k there is no more exploration. Therefore, $\hat{\pi}$ will eventually reach a terminal state and thus complete the exploration. From this we derive the following theorem:

Theorem 16. *In $\hat{\pi}$, as long as the planner has not observed an action j with $x_j = 1$, she will keep exploring until all actions' rewards are revealed. Therefore $\hat{\pi}$ always reaches a terminal state.*

4.2 Minimum exploration time

Two BIC planners may differ only in their recommendations when they are in the exploration states. We would prefer the one that explores the actions "faster", as it would mean finding the optimal action sooner. For this we define a partial order between policies. We say that a policy is *stochastic dominant* over another if it discovers the realizations of the rewards faster.

Definition 17. *A BIC policy algorithm π_A is stochastic dominant over another BIC policy algorithm π_B if for every prior D and for every agent t, π_A has at least the same probability to observe action j's reward as π_B, and for some action j a strictly higher probability to know it's reward in time t. I.e., for any agent t and action j we have $\Pr_{\pi_A}[S_t^{\vec{z}} \wedge \vec{z}[j] \neq *] \geq \Pr_{\pi_B}[S_t^{\vec{z}} \wedge \vec{z}[j] \neq *]$, and there exists some action j and agent t for which $\Pr_{\pi_A}[S_t^{\vec{z}} \wedge \vec{z}[j] \neq *] > \Pr_{\pi_B}[S_t^{\vec{z}} \wedge \vec{z}[j] \neq *]$.*

The following lemma states that the suggested policy, $\hat{\pi}$ is stochastic dominant over all other BIC policies.

Lemma 18. *Let $\pi_A \neq \hat{\pi}$ be a BIC policy algorithm, with the same recommendations for the exploitation states as in Table 3.1. Then $\hat{\pi}$ is stochastic dominant over π_A.*

From Lemma 18 we easily obtain that $\hat{\pi}$ maximizes exploration rates of each action j and agent t. Due to the use of $y \in [0,1]$ to decide which agent will explore each action, $\hat{\pi}$ manages to maximize exploration rates of all the actions independently. This give us an important result regarding $\hat{\pi}$:

Theorem 19. *$\hat{\pi}$ minimizes the time until terminal state .*

4.3 Maximum expected social welfare

In this chapter we present the main result: The best BIC policy is the one that minimizes exploration time for every action simultaneously.

Theorem 20. *Let π_{opt} be a BIC policy algorithm that maximizes the expected Social welfare. Then for a large number of agents (specifically, $T \geq \lceil \frac{1}{p_k^1} + n_k - 1 \rceil$), it holds that*

$$SW_T(\pi_{opt}) = SW_T(\hat{\pi})$$

Proof. For the sake of contradiction, assume that there exists a prior D, such that

$$SW_T(OPT) = \mathbb{E}_D[\Sigma_{t=1}^T u_t(\pi_{opt}(h_{t-1}))] > \mathbb{E}_D[\Sigma_{t=1}^T u_t(\hat{\pi}(h_{t-1}))] = SW_T(\hat{\pi})$$

π_{opt} maximizes expected social welfare, therefore it is easy to see that π_{opt} must give the same recommendation as in Table 3.1 whenever that social planner is in exploitation state. Since π_{opt} and $\hat{\pi}$ are different, there is at least one agent t, that might not receive the same recommendation from the two policies i.e., $(\lambda_t)_{\pi_{opt}} \neq (\lambda_t)_{\hat{\pi}}$, where $(\lambda_t)_\pi[j]$ is the probability that a policy π recommends action j to agent t. We have already established that this scenario can only happen when the planner is in exploration state. Therefore, it is a result of difference between exploration rates in both policies for at least one action j. Let t_0 and j denote the first indexes of such agent and action, respectively. It is easy to see that if $t_0 > n_j$ then π_{opt} is not in a terminal state and therefore does not maximizes social welfare. So assume $t_0 \leq n_j$. Let $(\psi_t^j)_{\pi_{opt}}$ denote the exploration rate used by π_{opt}, and let ϵ denote the difference between this exploration rate, and in $\hat{\pi}$, i.e., $\epsilon = q_{t_0}^j - (\psi_{t_0}^j)_{\pi_{opt}}$. Let π be a policy identical to π_{opt} that substitutes $(\psi_{t_0}^j)_{\pi_{opt}}$ with $q_{t_0}^j$. Recommendation policy π is a well defined BIC policy as t_0 is the first index of j for which $(\psi_{t_0}^j)_{\pi_{opt}}$ is not the maximum value, $q_{t_0}^j$ is a BIC exploration rate, and from Lemma 18, the rest of the exploration rates can still be used.

Let t_1 be the agent such that $t_1 = \lceil t_0 + \frac{1}{p_j^1} - 1 \rceil$. Since $t_0 \leq n_j \leq n_k$ and $p_j^1 \geq p_k^1$, it holds that $t_1 \leq \lceil n_k + \frac{1}{p_k^1} - 1 \rceil = T$. Now, since

$$SW_T(\pi) - SW_T(\pi_{opt}) \geq \mathbb{E}_D[\Sigma_{t=t_0}^{t_1} u_t(\pi_{opt}(h_{t-1}))] - \mathbb{E}_D[\Sigma_{t=t_0}^{t_1} u_t(\pi(h_{t-1}))]$$

We get

$$SW_T(\pi) - SW_T(\pi_{opt}) \geq \epsilon((t_1 - t_0)p_j^1 + 2p_j^1 - 1) = \epsilon((\frac{1}{p_j^1} - 1)p_j^1 + 2p_j^1 - 1) = \epsilon p_j^1$$

contradicting the optimality of π_{opt}. □

From the the above theorem we deduce the following corollary.

Corollary 21. *Recommendation policy $\hat{\pi}$ maximizes social welfare for every $T \geq \lceil \frac{1}{p_k^1} + n_k - 1 \rceil$*

4.4 Limited number of agents

The planner's goal is to maximize social welfare. If there is a limited number of agents, she cannot rely on the existence of the agent that balances the the loss of social welfare (i.e., agent t_1 in the proof for Theorem 20). Our algorithm must be adjusted for that. A natural solution is to limit the recommendation for exploration, so that the planner must give exploration driven recommendation for action j in round t if the gain for the following agents, $p_j^1(T - t - 1)$ is high enough to cover for the expected loss of the t-th agent, i.e., μ_j. We add the following requirement that must be fulfilled if the algorithm gives an exploration driven recommendations to agent t. Namely

$$(T - t) \cdot p_j^1 + \mu_j \geq 0$$

or alternatively $(T - t + 2) \cdot p_j^1 \geq 1$.

Theorem 20's proof still applies for any pair $\langle j, t \rangle$ that meets the additional requirement. For pairs $\langle j, t \rangle$ that do not meet the requirement, action j is no longer recommended for exploration in round t or afterwards. An exploration driven recommendation for these agents harms the social welfare.

Chapter 5

Continuous distribution for the a priori best action's reward

In this chapter we explore the same model with one significant difference. The prior distribution D_1 is now a continuous distribution that has full support of $[-1, 1]$. (Note that we do not allow mass points.) Consider that the number of agents, T, is large enough so that a social planner must complete the exploration of all the actions.

The different type of recommendation policy algorithm we introduce for this setting is a generalization of the partition policy, originally defined in Kremer et al. [KMP14].

5.1 Partition policy as a recommendation algorithm

The following two definitions are used to define a partition policy (in Definition 24).

Definition 22. *Θ^j is a collection of T **disjoint** sets, $\Theta^j := \{\theta_t^j\}_{t=2}^T$, where $\theta_t^j \subseteq [-1, 1]$.*

Definition 23. *A valid input for any partition policy algorithm is a series $\Theta = (\Theta^j)_{j=2}^k$ s.t. for any pair of actions $i \neq j \in A$, it holds that $\theta_t^j \cap \theta_t^i = \emptyset$ for every agent t.*

Definition 24. *Given a valid input, $(\Theta^j)_{j=2}^k$, and a realization $X_1 = x_1$, a partition policy is a recommendation policy that makes the following recommendations. For agent t we have,*

1. *For $t = 1$ we have $\sigma_1 = 1$.*
2. *If there is an explored action j with a reward of 1 (i.e., it is optimal), then $\sigma_t = j$.*
3. *Else, if $x_1 \in \theta_t^j$ then $\sigma_t = j$. (In this case agent t is the first agent for whom $\sigma_t = j$.).*
4. *$[-1, \mu_j] \subseteq \theta_j^j$*
5. *Else, $\sigma_t = 1$.*

Let us inspect each clause in the above definition with regards to BIC and social welfare.

1. Since action 1 is the a priori better action, any BIC policy must recommend to agent $t = 1$ action $\sigma_1 = 1$ (clause (1)).
2. After finding an explored action with value 1, to maximize the social welfare we must recommend it (clause (2)).
3. Clause (3) is where action j is recommended for the first time. Once the planner will observe the value of the explored action, if it is 1 it will be recommended to all future agents (as indicated in clause 2).
4. Clause (4) deals with the case that the a priori best action has low value. Together with clause 2, it guarantees that agent j performs action j if none of the explored actions has a higher reward.
5. Clause (5) gives an exploitation recommendation. Note that any explored action j, in this case, has $x_1 \geq x_j = -1$.

Notice that a valid input, $(\Theta^j)_{j=2}^k$, insures that every agent $t \geq 2$ receives a recommendation for exactly one action. We can now derive the following lemma:

Lemma 25. *The optimal BIC recommendation policy is a partition policy.*

5.2 The suggested BIC Partition Policy

Recall that agent t finds that recommendation σ_t to be BIC if for any action $i \in A$ we have

$$\mathbb{E}[u_t(j) - u_t(i) | \sigma_t = j] \geq 0$$

Note that this holds if and only if for any action $i \in A$

$$\Pr[\sigma_t = j] \cdot \mathbb{E}[u_t(j) - u_t(i) | \sigma_t = j] \geq 0$$

Namely,

$$\int_{\sigma_t = j} [X_j - X_i] dD \geq 0.$$

We now describe how to extract parameters for the suggested policy iteratively, given a prior D for the problem. Then we continue by showing that these collections can be used as a valid input of a partition policy. Finally, we show that using these parameters produces a BIC recommendation policy.

Definition 26. *The sets $\hat{\Theta}^j = \{\hat{\theta}_t^j\}_{t=2}^{T+1}$ are calculated as follows.*
Let $\hat{\theta}_t^j$ be the ordered interval $(i_t^j, i_{t+1}^j]$, where

1. *For $t < j$ let $\hat{\theta}_t^j = \emptyset$ (this can be done by setting $i_t^j := -1$ for $t \leq j$).*
2. *For $t = j$, recall that $i_j^j = -1$, and let ω_{j+1}^j be the solution to:*

$$\prod_{n=2}^{j-1} (p_n^{-1}) \int_{\mu_j \geq X_1} [\mu_j - X_1] dD_1 = \int_{\mu_j \leq X_1 \leq \omega_{j+1}^j} [X_1 - \mu_j] dD_1 \tag{5.1}$$

3. *For every* $t > j$*, let* ω^j_{t+1} *be the solution to:*

$$p^1_j \int_{-1 \le X_1 \le i^j_t} [1 - X_1] dD_1 = \int_{i^j_t < X_1 \le \omega^j_{t+1}} [X_1 - \mu_j] dD_1 \tag{5.2}$$

4. $i^j_{t+1} = min(1, \omega^j_{t+1})$ *for every* $t \ge j$*. (We have* $\hat{\theta}^j_t = (i^j_t, i^j_{t+1}]$*.).*

Notice that in each step, distribution D, the parameters p^1_j, p^{-1}_n, μ_j and i^j_t are known, therefore one can compute the value of i^j_{t+1}.
In the next lemma, we show that $i^j_t \le i^j_{t+1}$ for every action j and agent t. This will allow us to deduce in Corollary 28 that $\hat{\Theta}_j$ is a collection of disjoint sets, which is required from a valid input for partition policy.

Lemma 27. *For every action* $j \neq 1$ *and agent* $t \in T$*, it holds that* $i^j_t \le i^j_{t+1}$*.*

Proof. Let $j \neq 1$ be some action. The proof is by induction over t.
For the induction base, consider $t \le j - 1$, for which $i^j_t = i^j_{t+1} = -1$.
For the induction step, we assume the induction hypothesis holds for t and prove it for $t + 1$.
The left-hand sides of (5.1) and (5.2) are non-negative since $i^j_t \le 1$, $X_1 < 1$, $0 < p^1_n$ and $p^1_j > 0$. Consequently, the right-hand sides are also non-negative. So from (5.1) we deduce $-1 < \mu_j < i^j_{j+1}$ and from (5.2) and the induction hypothesis we deduce $\omega^j_{t+1} \ge i^j_t$ and as a result, $i^j_{t+1} \ge i^j_t$. □

Corollary 28. *For every action j, since there is no intersection between the ordered intervals* $(i^j_t, i^j_{t+1}]$ *for every agent t,* $\hat{\Theta}_j$ *is a collection of disjoint sets.*

For $(\hat{\Theta}^j)^k_{j=2}$ to be well defined we need to verify that there is no intersection for the same agent t for different actions (as stated in Definition 23). In the next lemma we show that for every agent and action $\langle t, j \rangle$ the right bound of θ^{j+1}_t is smaller than the left bound of θ^j_t.

Lemma 29. *For every* $t \ge j$*, it holds that* $i^j_t \ge i^{j+1}_{t+1}$*, and there is an equality only if* $i^j_t = i^{j+1}_{t+1} = 1$*.*

Proof. Let $j \neq 1$ be some action.The proof is by induction over t.
For the base case, consider $t = j$, for which $i^j_j = \mu_j > \mu_{j+1} = i^{j+1}_{j+1}$.
For the induction step, we assume the induction hypothesis holds for every agent $\le t$.
Consider $t = j$, then according to (5.1) ω^j_{j+1} is the solution to:

$$\prod_{n=2}^{j-1} (p^{-1}_n) \int_{\mu_j \ge X_1} [\mu_j - X_1] dD_1 = \int_{\mu_j \le X_1 \le \omega^j_{j+1}} [X_1 - \mu_j] dD_1 \tag{5.3}$$

Now, since $0 < p^{-1}_j < 1$ and $\mu_j > \mu_{j+1}$,

$$\prod_{n=2}^{j-1} (p^{-1}_n) \int_{\mu_j \ge X_1} [\mu_j - X_1] dD_1 > \prod_{n=2}^{j} (p^{-1}_n) \int_{\mu_{j+1} \ge X_1} [\mu_{j+1} - X_1] dD_1$$

Putting the above inequality with (5.3) for both j and $j+1$ we get,

$$\int_{\mu_j \leq X_1 \leq \omega^j_{t+1}} [X_1 - \mu_j] dD_1 > \int_{\mu_{j+1} \leq X_1 \leq \omega^{j+1}_{t+2}} [X_1 - \mu_{j+1}] dD_1$$

Since $\mu_j > \mu_{j+1}$ (and therefore $-\mu_j < -\mu_{j+1}$) $\omega^{j+1}_{t+2} < \omega^j_{t+1}$ must hold.

For every $t > j$, according to (5.2) ω^j_{t+1} is the solution to:

$$p^1_j \int_{-1 \leq X_1 \leq i^j_t} [1 - X_1] dD_1 = \int_{i^j_t < X_1 \leq \omega^j_{t+1}} [X_1 - \mu_j] dD_1 \tag{5.4}$$

Combining the induction hypothesis ($i^j_t > i^{j+1}_{t+1}$) with $p^1_{j+1} < p^1_j$ we get

$$p^1_j \int_{-1 \leq X_1 \leq i^j_t} [1 - X_1] dD_1 > p^1_{j+1} \int_{-1 \leq X_1 \leq i^{j+1}_{t+1}} [1 - X_1] dD_1$$

Putting the above inequality with (5.4) for both j and $j+1$ we get,

$$\int_{i^j_t < X_1 \leq \omega^j_{t+1}} [X_1 - \mu_j] dD_1 > \int_{i^{j+1}_{t+1} < X_1 \leq \omega^{j+1}_{t+2}} [X_1 - \mu_{j+1}] dD_1$$

Now, since $-\mu_j < -\mu_{j+1}$ and we know that $i^j_t > i^{j+1}_{t+1}$ from the induction hypothesis, we get that $\omega^{j+1}_{t+2} < \omega^j_{t+1}$. Consequently, for every $t \geq$ it holds that $i^{j+1}_{t+2} \leq i^j_{t+1}$. $\square$

Let $\hat{\pi}$ be a partition policy that uses the suggested parameters, $(\hat{\Theta}^j)^k_{j=2}$, as an input.

Corollary 30. *Policy $\hat{\pi}$ is well defined, as for every $x_1 \in [-1, 1]$ there exists exactly one action $j \in A$ such that $\sigma_t = j$.*

We now investigate what is required from a BIC partition policy, in order to show that the suggested partition policy is BIC. Following the same exhaustion demonstrated in Example 1, for every agent $t \geq 2$ a BIC recommendation for action $j \neq 1$ can be either exploration driven or exploitation driven:

1. **Exploitation driven recommendation**: Action $j \neq 1$ is the best action given the history. Once again, one of the following holds:
   (a) **A known reward**- Action j has the best possible realized value (i.e., $x_j = +1$) after one of the previous agents $j \leq \tau < t$ has explored it. Formally, the expected "gain" of agent t from choosing action j over of action 1 in this case is

$$\int_{X_j = 1, X_1 \in \cup_{\tau < t} \theta^\tau_j} [X_j - X_1] dD = p^1_j \int_{X_1 \in \cup_{\tau < t} \theta^\tau_j} [1 - X_1] dD_1$$

   (b) **An unknown reward**- The observed realization x_1 yields lower reward compared to the prior expected reward of action j, i.e., $x_1 < \mu_j (< \cdots < \mu_2)$, and the better a priori actions, $k < j$

have been explored and resulted in minimal reward of -1. Thus action j is better to execute then action 1, and the expected "gain" here is

$$\prod_{n=2}^{j-1}(p_n^{-1})\int_{\mu_j \geq X_1}[\mu_j - X_1]dD_1$$

2. **Exploration driven recommendation**: The planner has not yet observed an action with the best possible reward (i.e., $+1$), and $x_1 > \mu_j$. This recommendation does not benefit with agent t (but might yield higher expected social welfare hence is possible by the planner). The expected "loss" in this case is:

$$\int_{\mu_j < X_1, X_1 \in \theta_j^t}[X_j - X_1]dD = \int_{\mu_j < X_1, X_1 \in \theta_j^t}[\mu_j - X_1]dD_1$$

Taken together, the summation of the expected gains and loss is equivalent to the left hand sides of the integrals in the next lemma.

Lemma 31. *A recommendation $\sigma_t = j \neq 1$ is BIC w.r.t. action 1 if for $t = j$ we have,*

$$\int_{\mu_j < X_1, X_1 \in \theta_t^j}[\mu_j - X_1]dD + \prod_{n=2}^{j-1}(p_n^{-1})\int_{\mu_j \geq X_1}[\mu_j - X_1]dD_1 \geq 0 \tag{5.5}$$

and for $t > j$ we have,

$$\int_{X_1 \in \theta_t^j}[\mu_j - X_1]dD + p_j^1\int_{X_1 \in \cup_{\tau < t}\theta_j^{\tau}}[1 - X_1]dD_1 \geq 0 \tag{5.6}$$

Notice that for $t > j$, $X_1 \in \theta_t^j$ yields $X_1 \geq \mu_j$ since the sets are disjoint and according to clause (4) in definition 24, $[-1, \mu_j] \subseteq \theta_j^j$.

Let us compare the BIC constraints in above lemma to the suggested series, $\hat{\Theta}$.
If $i_t^j = 1$, then from Lemma 27 and the definition of i_t^j, it holds that $i_{t+1}^j = 1$, which yields $\hat{\theta}_t^j = \emptyset$.
Since both parts of (5.2) are zero in this case, so the (exploitation driven) recommendation $\sigma_t = j$ is BIC.
Else, $i_t^j = \omega_t^j$, in which case (5.1) and (5.2) satisfy (5.5) and (5.6) respectively. This allows us to derive the following corollary.

Corollary 32. *The partition policy $\hat{\pi}$ that uses the series $\hat{\Theta}$ as input, outputs BIC recommendations $\sigma_t = j$ w.r.t. action 1 for every agent t.*

Due to exploration of the a priori better actions earlier, it is sufficient for the suggested partition policy, $\hat{\pi}$ to maintain the BIC constraint w.r.t. action 1 alone. I.e., if $\sigma_t = j$ is an exploration driven recommendation, then j has the lowest index of an unexplored action (and therefore has the highest expected value among all unknown rewards.). Agent t would not prefer any other action $i \neq j$ as such a behavior would yield either a lower prior reward for actions $i > j$ or the lowest possible reward (i.e., -1) for actions $i < j$. The following are corollaries of the last lemma.

Theorem 33. *Policy $\hat{\pi}$ is a BIC partition policy.*

Corollary 34. *Policy $\hat{\pi}$ recommends the actions in ascending order, i.e., for every $j < i$, action j is explored before action i.*

5.3 Optimality

We follow the same logic process of the proofs provided in chapter 4 with some tiny adjustments. The abstraction of the information states is used here with one small difference- considering all the BIC policies recommend action 1 to the first agent and observe reward $x_1 \in [-1, 1]$, we update the definition of *stochastic dominant* to be conditioned on x_1 as follows.

Definition 35. *A BIC policy algorithm π_A is stochastic dominant over another BIC policy algorithm π_B if for every prior D and a realization $x_1 \in [-1, 1]$, and for any agent t, π_A has at least the same probability to observe action j's reward as π_B, and for some action j a strictly higher probability to observe it's reward in time t. I.e., for any agent t, action j and realization $x_1 \in [-1, 1]$ we have $\Pr_{\pi_A}[S_t^{\vec{z}} \wedge \vec{z}[j] \neq *|\vec{z}[1] = x_1] \geq \Pr_{\pi_B}[S_t^{\vec{z}} \wedge \vec{z}[j] \neq *|\vec{z}[1] = x_1]$, and there exists some agent t and action j for which $\Pr_{\pi_A}[S_t^{\vec{z}} \wedge \vec{z}[j] \neq *|\vec{z}[1] = x_1] > \Pr_{\pi_B}[S_t^{\vec{z}} \wedge \vec{z}[j] \neq *|\vec{z}[1] = x_1]$.*

The next step is to show the partition policy with the suggested parameters as input yields stochastic dominance. It is done in the following lemma.

Lemma 36. *Let $(\Theta_j)_{j=2}^k$, and the realization $X_1 = x_1$, be the input for a (BIC) partition policy $\pi_A \neq \hat{\pi}$. Then $\hat{\pi}$ is stochastic dominant over π_A.*

Proof. For contradiction, suppose that there exists a prior, D, an action j and time t_1 such that

$$\Pr_{\hat{\pi}}[S_{t_1}^{\vec{z}} \wedge \vec{z}[j] \neq *|\vec{z}[1] = x_1] < \Pr_{\pi_A}[S_{t_1}^{\vec{z}} \wedge \vec{z}[j] \neq *|\vec{z}[1] = x_1] \; , \tag{5.7}$$

where j is the least such action and t_1 is the least such agent for action j. It means that π_A recommends to some agent $j \leq t < t_1$ to explore action j while $\vec{z}[j] = *$ and $(\sigma_t)_{\hat{\pi}} \neq j$. If $(\sigma_t)_{\hat{\pi}} = 1$ it means that $\theta_j^t = \emptyset$, therefore from Lemma 25, π_A is not a BIC policy (as it does not a valid partition policy). So consider exploration driven recommendation (i.e., $S_t^{\vec{z}} \wedge \vec{z}[i] = * \wedge \vec{z}[j] = *$), $(\sigma_t)_{\hat{\pi}} = i < j$ (since $i > j$ contradictions Corollary 34). In such a case,

$$\Pr_{\hat{\pi}}[S_{t+1}^{\vec{z}} \wedge \vec{z}[i] \neq *|\vec{z}[1] = x_1] > \Pr_{\pi_A}[S_{t+1}^{\vec{z}} \wedge \vec{z}[i] \neq *|\vec{z}[1] = x_1] \; , \tag{5.8}$$

Hence π_A is not stochastic dominant over $\hat{\pi}$. □

From Lemma 36 we deduce that $\hat{\pi}$ maximizes exploration for each action j and agent t. Due to the use of disjoint sets in the definition of a partition policy, $\hat{\pi}$ manages to maximize exploration rates independently. This gives us the important result of time minimization:

Theorem 37. *$\hat{\pi}$ minimizes the time until terminal state.*

Finally, we show the main result for this case as well.

Theorem 38. *$\hat{\pi}$ maximizes social welfare for unlimited number of agents.*

Proof. For the sake of contradiction, assume that there exists a prior D and a realization $X_1 = x_1$, such that

$$SW_T(OPT) = \mathbb{E}_D[\Sigma_{t=1}^{T} u_t(\pi_{opt}(h_{t-1}))] > \mathbb{E}_D[\Sigma_{t=1}^{T} u_t(\hat{\pi}(h_{t-1}))] = SW_T(\hat{\pi})$$

π_{opt} maximizes expected social welfare, so according to Lemma 25 it must be a partition policy. By differing from $\hat{\pi}$, there exist an action j and a time t_1 in which $(\sigma_{t_1})_{\hat{\pi}} = j$ and $(\sigma_{t_1})_{\pi_{OPT}} = i \neq j$, where j is the least such action and t_1 is the least such agent for action j. If $i < j$ then action i is already explored by both planners and $x_i = -1$, and the result is a lower SW for OPT. If $i > j$, since $\mu_i < \mu_j$ the result is a lower social welfare for OPT. Overall we get that $SW_T(OPT) < SW_T(\hat{\pi})$. □

Chapter 6

Conclusion

This paper explores the problem of incentivizing exploration via Bayesian persuasion. We consider two different supports for the a priori better action, a discrete version $\{-1, 0, 1\}$, and a continuous version $[-1, 1]$. In both settings, our optimal policy explores the better a priori actions earlier. In addition, it maximizes the exploration, subject to the BIC constraints. This leads to a planner policy that maximize the social welfare.

Our optimal policy also achieves both: (1) minimizing the time until all of the actions are explored, and (2) that all the actions are explored, in case of large enough T. Our optimal policy requires special correlated randomization to guarantee the optimality.

There are few obvious open problems, First, to extend the support of *all* actions to be, for example $[-1, +1]$. Second, to consider stochastic actions, even simple Bernoulli random variables with different probabilities. Third, enabling the agents to receive limited amount of information about the past. (The challenge here is to make the information informative, and still allow the planner to explore all actions, eventually.)

6.1 The case of continuous distribution for all actions' reward

Let us discuss a case in which the support of each action's reward is $[-1, 1]$. In order to tackle such setting, we must first update Definition 24. First, we need to remove clause (2), as there is no action with optimal reward (unless all rewards are known). Second, we need to change clause (5) to $\sigma_t = \text{argmax}_j x_j$, so that every exploitation driven recommendation with known reward will be for the action with the best reward.
The real issue here is that we can no longer assume that if a partition policy is BIC w.r.t. action 1 then it is BIC w.r.t. all other actions. Consequently, we must also update the way we choose which agent should explore which action. For example, if agent 5 receives a recommendation for action 3, i.e., $\sigma_5 = 3$ and there is a very low probability that either x_2 or x_1 are smaller than μ_3 and also μ_3 is very small (slightly bigger than -1). In this case, the agent knows that it is highly unlikely for such a recommendation to benefit her,

and might prefer action 2 over this recommendation.

By updating Definition 26 we can calculate partitions w.r.t. every action as follows. Let $[\theta_t^j]^m$ denote the ordered interval w.r.t. action m such that (5.1) and (5.2) are calculated with X_m and D_m instead of X_1 and D_1. Let us assume a weakening assumption of stochastic dominance of action j_1's reward over every action's $j_2 < j_1$ reward. We can derive that for every $< m_1, m_2, t, j >$ such that for every $t > m_1 > m_2, j$ if both intervals are not empty then $[i_t^j]^{m_1} < [i_t^j]^{m_2}$. An 'easy' suggestion for the planner to handle BIC constraints and still be able to explore new actions is to make sure that $\sigma_t = j$ if $j \leq [i_{t+1}^j]^{m_1}$ for every $m_1 \leq j$ or simply $j \leq [i_{t+1}^j]^{j-1}$. The optimality of this suggestion remains to be uncertainty, as a different planner might find a better way to recommend actions with unknown rewards sooner, thus gaining stochastic dominance over it.

Appendix A

K actions: Algorithm and Missing Proofs

In this appendix, we extend the algorithm for 3 actions to handle any number of actions, k. Recall that when calculating BIC constraints, we consider 3 different reasons for a recommendation, $\sigma_t = j$, as explained in Example 2. To handle with multiple actions' rewards, we abandon the explicit states, e.g., $S_t^{\langle -1,1,*\rangle}$. We intentionally dismiss any states where actions are, as we will see that these states are infeasible by $\hat{\pi}$.

- **Exploitation driven recommendation**, action j either has:
  - **A known reward**- The planner has already observed action j's reward and it is indeed the maximum possible reward, i.e., $x_j = 1$. As we are about to show, such a scenario is possible only when $x_1 \in \{0, 1\}$ and for every action $1 < i < j$, it's reward has been observed, and $x_i = -1$.
  - **An unknown reward**- action j is yet to be explored, and every action $i < j$, has been explored and $x_i = -1$.
- **Exploration driven recommendation** Action j has not been explored yet, every $i < j$ has been explored, and $x_i = -1$ for $i < j$, whereas $x_1 = 0$. This implies that the planner is in the exploration state of action j.

In Table A.1, we are extending the algorithm described in Table 3.1 to all states that may result in a recommendation for action j when using $\hat{\pi}$. We also added the gain of each state compared to action 1 (i.e., $\mathbb{E}[u_t(j) - u_t(1)|S_t^{\vec{z}}, \sigma_t = j]$), as well as the probability that the planner is in these states in round t (i.e., $\Pr[S_t^{\vec{z}}|\sigma_t = j]$).

A BIC exploration rate q_j^t for our algorithm is still the maximum value that satisfy the same constraints as before (i.e., (3.1)).

Theorem 39. *For k actions, given $q_2^2, \ldots, q_{t-1}^2, \ldots, q_{j-1}^{j-1}, \ldots, q_{t-1}^{j-1}$, for $j > 2$,*

$$q_t^2 = \begin{cases} 0 & t = 1 \\ \min(\frac{2p_1^{-1}p_2^1}{1-2p_2^1}, p_1^0) & t = 2 \\ \min(\frac{2p_1^{-1}p_2^1 + p_2^1\sum_{m=2}^{t-1} q_m^2}{1-2p_2^1}, p_1^0 - \sum_{m=2}^{t-1} q_m^2) & t > 2 \end{cases} \tag{A.1}$$

Recommendation Table. Policy Parameters: (y, t)				
X_1	X_j	σ_t	$\mathbb{E}[u_t(j) - u_t(1) \mid S_t^{\vec{z}}, \sigma_t = j]$	$\Pr[S_t^{\vec{z}}]$
-1	$*$	j	$2p_j^1$	$\mathbb{1}[t = j]\Pi_{i<j} p_i^{-1}$
0	1	j	1	$p_j^1 \sum_{\tau=j}^{t-1} q_\tau^j$
-1	1	j	2	$\mathbb{1}[t = j] p_j^1 \Pi_{i<j} p_i^{-1}$
0	$*$	$f_t^j(y) \in \{1, j\}$	$2p_j^1 - 1$	$p_{j-1}^{-1} \sum_{\tau=j-1}^{t-1} q_\tau^{j-1} - \sum_{\tau=j}^{t-1} q_\tau^j$

Table A.1: Extension for states that may recommend on action j

$$q_t^j = \begin{cases} 0 & t < j \\ \min(\frac{2p_j^1 \Pi_{i<j} p_i^{-1}}{1-2p_j^1}, p_{j-1}^{-1} q_{j-1}^{j-1}) & t = j \\ \min(\frac{2p_j^1 \Pi_{i<j} p_i^{-1} + p_j^1 \sum_{\tau=j}^{t-1} q_\tau^j}{1-2p_j^1}, p_{j-1}^{-1} \sum_{\tau=j-1}^{t-1} q_\tau^{j-1} - \sum_{\tau=j}^{t-1} q_\tau^j) & t > j \end{cases} \tag{A.2}$$

In addition we show that

$$q_{t+1}^{j+1} \leq p_j^{-1} q_t^j \tag{A.3}$$

Proof. The proof is done by induction over action j and agent t. The induction base and q_t^2's part are provided by Theorem 3. For the induction step we assume that (A.2) and (A.3) holds for any $t_0 < t$ and $j_0 < j$.

As for q_t^j, by using the induction hypothesis, we know that an exploitation driven recommendation $\sigma_t = j$ can only come from the first three exploitation states described in Table A.1. We also know from it that an exploration driven recommendation can only come from the last state in Table A.1. For this we notice that $\Pr[S_t^{\vec{z}}] = \Pr[S_t^{\vec{z}}, \sigma_t = j]$ for every $S_t^{\vec{z}} \in \Gamma_t^{j+}$ and that $\Pr[S_t^{\vec{z}}, \sigma_t = j] = q_t^j$ Hence, the BIC constraint for any agent $t \geq j$ is

$$\mathbb{E}[u_t(j) - u_t(1) \mid \sigma_t = j] \Pr[\sigma_t = j] = 2p_j^1 \Pi_{i<j} p_i^{-1} + (2p_j^1 - 1) q_t^j + p_j^1 \sum_{\tau=j}^{t-1} q_\tau^j \geq 0$$

Notice that the first and third state in Table A.1 have the same value for

$$\Pr_{\hat{\pi}}[S_t^{\vec{z}} \mid \sigma_t = j] \; \mathbb{E}[u_t(j) - u_t(1) \mid S_t^{\vec{z}}, \sigma_t = j]$$

Therefore we merged them. In order for q_t^j to be a valid, it must also satisfy

$$q_t^j = \Pr[\sigma_t = j, S_t^{\vec{z}} \in \Gamma_t^{j-}] \leq \Pr[S_t^{\vec{z}} \in \Gamma_t^{j-}]$$

And by substituting $S_t^{\vec{z}} \in \Gamma_t^{j-}$ with the probability for the last state in Table A.1 we get

$$q_t^j \leq p_{j-1}^{-1} \sum_{\tau=j-1}^{t-1} q_\tau^{j-1} - \sum_{\tau=j}^{t-1} q_\tau^j$$

Therefore,

$$q_t^j = \min(\frac{2p_j^1 \Pi_{i<j} p_i^{-1} + p_j^1 \sum_{\tau=j}^{t-1} q_\tau^j}{1-2p_j^1}, p_{j-1}^{-1} \sum_{\tau=j-1}^{t-1} q_\tau^{j-1} - \sum_{\tau=j}^{t-1} q_\tau^j)$$

We can upper bound q_t^j as follows

$$q_t^j \leq \frac{2p_j^1 \Pi_{i<j} p_i^{-1} + p_j^1 \sum_{\tau=j}^{t-1} q_\tau^j}{1-2p_j^1} \leq \frac{2p_{j-1}^1 \Pi_{i<j} p_i^{-1} + p_{j-1}^1 \sum_{\tau=j}^{t-1} q_\tau^j}{1-2p_{j-1}^1}$$

the second inequality is correct due to the assumption that $p_j^1 < p_{j-1}^1$. From the induction hypothesis, we know that $q_\tau^j < p_{j-1}^{-1} q_\tau^{j-1}$ for every $\tau \leq t-1$. Hence,

$$q_t^j \leq \frac{2p_{j-1}^1 \Pi_{i<j} p_i^{-1} + p_{j-1}^1 \sum_{\tau=j}^{t-1} q_\tau^j}{1-2p_{j-1}^1} \leq \frac{2p_{j-1}^1 \Pi_{i<j} p_i^{-1} + p_{j-1}^{-1} p_{j-1}^1 \sum_{\tau=j-1}^{t-2} q_\tau^{j-1}}{1-2p_{j-1}^1} = p_{j-1}^{-1} A_{t-1}^{j-1} = p_{j-1}^{-1} q_{t-1}^{j-1} \ .$$

By using the induction hypothesis again, we get $0 < \sum_{\tau=j}^{t-1} q_\tau^j < p_j^{-1} \sum_{\tau=j-1}^{t-2} q_\tau^{j-1}$, thus

$$p_{j-1}^{-1} q_\tau^{j-1} \leq p_{j-1}^{-1} q_\tau^{j-1} + p_{j-1}^{-1} \sum_{\tau=j-1}^{t-2} q_\tau^{j-1} - \sum_{\tau=j}^{t-1} q_\tau^j = p_{j-1}^{-1} \sum_{\tau=j-1}^{t-1} q_\tau^{j-1} - \sum_{\tau=j}^{t-1} q_\tau^j$$

which completes the proof. □

Appendix B

Missing proofs from Sections 3 and 4

Proof of Lemma 3. We prove here lemma for q_t^2 and q_t^3, i.e., $j = 3$, and the full proof, for $j \geq 4$, can be found in Appendix A. The values q_t^2's are stated in the theorem, and we will show that they satisfy the required conditions. In addition we state the values of q_t^3, and later show that they also satisfy the required conditions,

$$q_t^3 = \begin{cases} 0 & t < 3 \\ \min(\frac{2p_3^1 p_1^{-1} p_2^{-1}}{1-2p_3^1}, p_2^{-1} q_2^2) & t = 3 \\ \min(\frac{2p_1^{-1} p_2^{-1} p_3^1 + p_3^1 \sum_{\tau=3}^{t-1} q_\tau^3}{1-2p_3^1}, p_2^{-1} \sum_{\tau=2}^{t-1} q_\tau^2 - \sum_{\tau=3}^{t-1} q_\tau^3) & t > 3 \end{cases} \tag{B.1}$$

Finally we need to show that $q_t^3 \leq p_2^{-1} q_{t-1}^2$.

In the following, we calculate exploration rates for $\hat{\pi}$, and just for notational convenience, we will refer to $\Pr_{\hat{\pi}}[\,]$ simply as $\Pr[\,]$. The first agent, by knowing her place in line, knows that none of the actions have been explored yet. Hence, it is for her best interest to choose action 1, the action with the maximal prior expected reward. Therefore, it is necessary that $q_1^2 = 0$ for $\hat{\pi}$ to be BIC policy. Consequently, the first agent explores action 1, and gains reward of $u_1(1) = x_1$. As for the second agent, the planner can now use her knowledge of x_1 for σ_2. If $x_1 = 1$, (the best possible reward), since she wishes to maximize the social welfare, she must recommend action 1 to the rest of the agents, i.e., $S_t^{\langle 1,*,* \rangle}$ is a terminal state with maximal social welfare. If $x_1 = -1$, action 2 currently has the best expected reward, i.e., $\mu_2 > \mu_3 > -1 = x_1$. It also does not decrease the reward comparing to the known reward of action 1. Therefore, any social BIC planner (including $\hat{\pi}$) must recommend on action 2 in this round, i.e., $\sigma_2 = 2$. If $x_1 = 0$, although action 2 is not the best action for agent 2, a social planner would probably want to recommend it to the second agent, at least with some probability. For that she uses her advantage of knowing the realization of action 1's reward. Note that q_2^2 influences the second agent only when $x_1 = 0$.

To complete the definition of the recommendation for the second agent, we calculate the value of q_2^2. The expected utility that agent 2 for following the recommendation must be at least the expected utility of choosing action 1. Therefore, q_2^2 satisfies the BIC constraint, i.e.,

$$\mathbb{E}[u_2(2) - u_2(1)|\sigma_2 = 2]\Pr[\sigma_2 = 2] = (\mu_2 - (-1))\Pr[S_2^{\langle -1,*,*\rangle}|\sigma_2 = 2]\Pr[\sigma_2 = 2] + \mu_2 q_2^2 \geq 0.$$

Information state $S_2^{\langle -1,*,*\rangle}$ always leads the planner to recommend $\sigma_2 = 2$, therefore

$$\Pr[S_2^{\langle -1,*,*\rangle}|\sigma_2 = 2]\Pr[\sigma_2 = 2] = \Pr[S_2^{\langle -1,*,*\rangle}, \sigma_2 = 2] = \Pr[S_2^{\langle -1,*,*\rangle}]$$

According to Table 3.1, the above is true for any action j and exploitation state $S_t^{\vec{z}} \in \Gamma_t^{j+}$, which all have in common a corresponding recommendation $\sigma_t = j$. For this reason, for $S_t^{\vec{z}} \in \Gamma_t^{j+}$, we can replace $\Pr[S_t^{\vec{z}}|\sigma_t = j]\Pr[\sigma_t = j]$ by $\Pr[S_t^{\vec{z}}]$.

Recall that q_2^2 not only must satisfy the above constraint but also to satisfy the second part of (3.1). Hence,

$$(1 + \mu_2)\Pr[S_2^{\langle -1,*,*\rangle}] + \mu_2 q_2^2 \geq 0 \quad \text{and} \quad q_2^2 \leq \Pr[S_2^{\langle 0,*,*\rangle}]$$

Since, $\mu_2 = 2p_2^1 - 1$, $\Pr[S_2^{\langle 0,*,*\rangle}] = p_1^0$, and $\Pr[S_2^{\langle -1,*,*\rangle}] = p_1^{-1}$, we have

$$q_2^2 = \min(\frac{2p_2^1 p_1^{-1}}{1 - 2p_2^1}, p_1^0) .$$

We proceed to calculate the first positive value of q_t^3. Due to the assumption $\mu_2 > \mu_3$, agent 2 can deduce that action 3 has not been explored yet, consequently she would definitely not follow a recommendation to choose action 3, therefore, $q_2^3 = 0$. Consider recommending action 3 to agent 3, i.e., $\sigma_3 = 3$. This recommendation can occur when the planner is in either exploitation state $\langle -1, -1, *\rangle$, or exploration state $\langle 0, -1, *\rangle$. The BIC constraint is:

$$\mathbb{E}[u_3(3) - u_3(1)|\sigma_3 = 3]\Pr[\sigma_t = 3] = 2p_3^1 \Pr[S_3^{\langle -1,-1,*\rangle}] + \mu_3 q_3^3 \geq 0$$

We again maximize over q_3^3, subject to the the second constraint in (3.1) as well, i.e.,

$$q_3^3 \leq \frac{2p_3^1 p_1^{-1} p_2^{-1}}{1 - 2p_3^1} \quad \text{and} \quad q_3^3 \leq \Pr[S_3^{\langle 0,-1,*\rangle}] + \Pr[S_3^{\langle 0,*,*\rangle}] = p_1^0 - q_2^2 p_2^1$$

Since for $t = 3$ and $j = 3$, we assume that $q_2^2 = A_2^2$, we have

$$q_3^3 \leq A_3^3 = \frac{2p_3^1 p_1^{-1} p_2^{-1}}{1 - 2p_3^1} < \frac{2p_2^1 p_1^{-1} p_2^{-1}}{(1 - 2p_2^1)} = A_2^2 p_2^{-1} = q_2^2 p_2^{-1} = q_2^2(1 - p_2^1) \leq p_1^0 - q_2^2 p_2^1$$

it implies that,

$$q_3^3 = \min(\frac{2p_3^1 p_1^{-1} p_2^{-1}}{1 - 2p_3^1}, q_2^2 p_2^{-1}) \tag{B.2}$$

Since $0 < p_1^{-1}, p_2^{-1}, p_2^1, p_3^1 < 1$, and $\mu_3 = (2p_3^1 - 1) < 0$, we get that both the numerator and the denominator of q_3^3 are positive, therefore

$$q_3^3 > 0 \tag{B.3}$$

We now prove by induction on agent t, the following:

$$q_t^3 \leq p_2^{-1} q_{t-1}^2 \tag{B.4}$$

This assumption assures that every time $\hat{\pi}$ has recommends $\sigma_t = 3$, it is done after action 2 has been explored. (This will be clear after we define $f^2(y)$ and $f^3(y)$, however, observe that the exploration rate of the third action is bounded by the exploration of the second action up to the previous agent times the probability that the second action realization is -1.) This will imply that $\Pr[S_t^{\langle 0,*,-1\rangle}] = \Pr[S_t^{\langle 0,*,1\rangle}] = \Pr[S_t^{\langle 0,1,-1\rangle}] = 0$. This will simplify the derivation, as a recommendation $\sigma_t = 2$ can come from only two exploitation states, $S_t^{\langle -1,1,*\rangle}, S_t^{\langle 0,1,*\rangle}$, or from $S_t^{\langle 0,*,*\rangle}$, the only exploration driven state that may cause recommendation for action 2.

For the induction base, consider agent $t = 3$. Indeed, $q_3^3 \leq q_2^2 p_2^{-1}$ from (B.2). For the induction step, assume that (B.4) holds for every time $t < t_0$. The exploration rates, q_t^2 and q_{t+1}^3 for each agent $t = t_0$ can be derived from the following the constraints in (3.1). Starting with q_t^2.

$$\mathbb{E}[u_t(2) - u_t(1) | \sigma_t = 2] \Pr[\sigma_t = 2] = 2\Pr[S_t^{\langle -1,1,*\rangle}] + \mu_2 q_t^2 + \Pr[S_t^{\langle 0,1,*\rangle}] \geq 0 \tag{B.5}$$

The probabilities of each mentioned state are as follows.

- $\Pr[S_t^{\langle -1,1,*\rangle}] = p_1^{-1} p_2^1$. Also, note that $S_t^{\langle -1,1,*\rangle}$ is a terminal state.
- $\Pr[S_t^{\langle 0,1,*\rangle}]$ is the intersection of the following events:
  - Action 2 has been explored before agent t, i.e. $\sum_{\tau=2}^{t-1} q_\tau^2$, so (implicitly) action 1 has already sampled and $x_1 = 0$.
  - $\Pr[X_2 = 1] = p_2^1$
  - This is a terminal state, so no further events.

Combining all with (B.5),

$$2p_1^{-1} p_2^1 + (2p_2^1 - 1) q_t^2 + p_2^1 \sum_{\tau=2}^{t-1} q_\tau^2 \geq 0$$

and $q_t^2 = \Pr_{\hat{\pi}}[\sigma_t = 2, S_t^{\langle 0,*,*\rangle}] \leq \Pr_{\hat{\pi}}[S_t^{\langle 0,*,*\rangle}] \leq p_1^0 - \sum_{\tau=2}^{t-1} q_\tau^2$. So,

$$q_t^2 = min(\frac{2p_1^{-1} p_2^1 + p_2^1 \sum_{\tau=2}^{t-1} q_\tau^2}{1 - 2p_2^1}, p_1^0 - \sum_{\tau=2}^{t-1} q_\tau^2)$$

As for q_t^3, the recommendation $\sigma_t = 3$ can come from the exploitation states $S_t^{\langle -1,-1,1\rangle}$ and $S_t^{\langle 0,-1,1\rangle}$ (recall that $\Pr[\sigma_3 = 3 | S_3^{\langle -1,-1,*\rangle}] = 1$, which implies that agent 3 will perform action 3, and therefore any agent

$t \geq 4$ has $\Pr[S_t^{\langle -1,-1,* \rangle}] = 0$), or from the only exploration state relevant for agent t, $S_t^{\langle 0,-1,* \rangle}$. Hence, the BIC constraint is

$$\mathbb{E}[u_t(3) - u_t(1)|\sigma_t = 3] \Pr[\sigma_t = 3] = 2 \Pr[S_t^{\langle -1,-1,1 \rangle}] + \mu_3 q_t^3 + \Pr[S_t^{\langle 0,-1,1 \rangle}] \geq 0$$

substituting the probabilities we have

$$2p_1^{-1}p_2^{-1}p_3^1 + (2p_3^1 - 1)q_t^3 + p_3^1 \sum_{\tau=3}^{t-1} q_\tau^3 = 0$$

Note that $\sum_{\tau=3}^{t-1} q_\tau^3$ implicitly states that $x_1 = 0$ and that $x_2 = -1$, or else action 3's reward might has been revealed by exploitation, or it still unknown, but did not revealed by exploration. In order for q_t^3 to be a valid, it must also satisfy

$$q_t^3 = \Pr[\sigma_t = 3, S_t^{\langle 0,-1,* \rangle}] \leq \Pr[S_t^{\langle 0,-1,* \rangle}] = p_2^{-1} \sum_{\tau=2}^{t-1} q_\tau^2 - \sum_{\tau=3}^{t-1} q_\tau^3$$

where $p_2^{-1} \sum_{\tau=2}^{t-1} q_\tau^2$ is the probability that there was an agent before agent t that explored action 2, and $x_2 = -1$. From this probability we subtract the probability that action 3 has been explored, i.e., $\sum_{\tau=3}^{t-1} q_\tau^3$, so that up until agent t the state is $\langle 0, -1, * \rangle$ Therefore,

$$q_t^3 = \min(\frac{2p_1^{-1}p_2^{-1}p_3^1 + p_3^1 \sum_{\tau=3}^{t-1} q_\tau^3}{1 - 2p_3^1}, p_2^{-1} \sum_{\tau=2}^{t-1} q_\tau^2 - \sum_{\tau=3}^{t-1} q_\tau^3)$$

We can upper bound q_t^3 as follows

$$q_t^3 \leq \frac{2p_1^{-1}p_2^{-1}p_3^1 + p_3^1 \sum_{\tau=3}^{t-1} q_\tau^3}{1 - 2p_3^1} \leq \frac{2p_1^{-1}p_2^{-1}p_2^1 + p_2^1 \sum_{\tau=3}^{t-1} q_\tau^3}{1 - 2p_2^1}$$

the second inequality is correct due to the assumption that $p_3^1 < p_2^1$. From the induction hypothesis (B.4) we know that $q_\tau^3 \leq p_2^{-1} q_{t-1}^2$ for every $\tau \leq t-1$. Since, for t and $j = 3$ we assume that $q_{t-1}^2 = A_{t-1}^2$, we have,

$$q_t^3 \leq \frac{2p_1^{-1}p_2^{-1}p_2^1 + p_2^1 \sum_{\tau=3}^{t-1} q_\tau^3}{1 - 2p_2^1} \leq \frac{2p_1^{-1}p_2^{-1}p_2^1 + p_2^1 p_2^{-1} \sum_{\tau=2}^{t-2} q_\tau^2}{1 - 2p_2^1} = p_2^{-1} A_{t-1}^2 = p_2^{-1} q_{t-1}^2 \ .$$

By using the induction hypothesis again, with (B.3), we get $0 < \sum_{\tau=3}^{t-1} q_\tau^3 < p_2^{-1} \sum_{\tau=2}^{t-2} q_\tau^2$, thus

$$q_t^3 \leq p_2^{-1} q_{t-1}^2 \leq p_2^{-1} q_{t-1}^2 + p_2^{-1} \sum_{\tau=2}^{t-2} q_\tau^2 - \sum_{\tau=3}^{t-1} q_\tau^3 = p_2^{-1} \sum_{\tau=2}^{t-1} q_\tau^2 - \sum_{\tau=3}^{t-1} q_\tau^3$$

which completes the proof of (B.4) and the proof of the theorem. □

Proof of Lemma 4. First, we show that if action j is recommended for agent $t > n_{j-1}$, then action $(j-1)$'s reward has been observed. If $t > n_{j-1}$ then $q_t^{j-1} = 0$. From the constraints in (3.1) regarding action $j-1$ for $t > n_{j-1} \geq j$, as the "gain" part is strictly positive and the "loss" part is strictly negative, we get that $\sum_{\Gamma_t^{(j-1)-}} \Pr_{\hat{\pi}}[S_t^{\bar{z}}] = 0$. Therefore, all the states for which action j is explored before action $j-1$ are infeasible for agent t as they were for $t \leq n_{j-1}$ and we get to continue the induction without relying on $q_t^j \leq p_{j-1}^{-1} A_{t-1}^{j-1}$. Hence, we get exactly the same constraints for action j as for the case of $t \leq n_{j-1}$, i.e.,

$$\mathbb{E}[u_t(j) - u_t(1)|\sigma_t = j] \Pr[\sigma_t = j] = 2p_j^1 \Pi_{i<j} p_i^{-1} + (2p_j^1 - 1)q_t^j + p_j^1 \sum_{\tau=j}^{t-1} q_\tau^j \geq 0$$

and also,

$$q_t^j \leq p_{j-1}^{-1} \sum_{\tau=j-1}^{t-1} q_\tau^{j-1} - \sum_{\tau=j}^{t-1} q_\tau^j$$

which completes the proof of the lemma. □

proof of Lemma 6. The proof is by induction over t. For the induction base consider $t = j$. From Lemma 5, we have that $A_j^j \geq q_j^j > 0$, therefore

$$A_j^j = \frac{2p_j^1 \Pi_{i<j} p_i^{-1}}{1 - 2p_j^1} < \frac{2p_j^1 \Pi_{i<j} p_i^{-1} + p_j^1 q_j^j}{1 - 2p_j^1} = A_{j+1}^j$$

For the inductions step, we assume that the induction hypothesis holds for t and prove it for $t+1$. From the inductive hypothesis we have that $A_{t-1}^j < A_t^j$. If $q_t^j > 0$, we get

$$A_t^j = \frac{2p_j^1 \Pi_{i<j} p_i^{-1} + p_j^1 \sum_{\tau=j}^{t-1} q_\tau^j}{1 - 2p_j^1} < \frac{2p_j^1 \Pi_{i<j} p_i^{-1} + p_j^1 \sum_{\tau=j}^{t} q_\tau^j}{1 - 2p_j^1} = A_{t+1}^j$$

which proves the lemma. □

Proof of Lemma 7. We show this lemma by induction over action index j. For the base of the induction, consider action $j = 2$. We prove the case of action 2 by induction over t. For this induction we use the base case of $t = n_1 + 1 = 2$ (since agent 1 always explores action 1). We have

- $B_2^2 \geq q_2^2 > 0 = B_1^2$.
- $0 < q_2^2$, directly from Lemma 5.
- $q_2^1 = 0$.
- If $q_2^2 = B_2^2 = p_1^0 (> 0)$, we show by induction that for every $i \geq 1$ it holds that $q_{2+i}^2 = B_{2+i}^2 = 0$. For base consider $i = 1$. Then by using Lemma 6 we get

$$q_3^2 = min(A_3^2, B_3^2) = min(A_3^2, p_1^0 - q_2^2) = min(A_3^2, p_1^0 - p_1^0) = min(A_3^2, 0) = 0 \ .$$

For the induction step assume this property holds for $i-1$ and show for i. From the induction hypothesis and Lemma 6 we get

$$q^2_{2+i} = min(A^2_{2+i}, B^2_{2+i}) = min(A^2_{2+i}, B^2_{2+i-1} - q^t_{2+i-1}) = min(A^2_{2+i}, 0) = 0$$

For the induction step we assume that hypothesis of the lemma holds for action 2 and for every agent $\leq t-1$, and show it for t.

- From the induction hypothesis $q^2_{t-1} > 0$, so we get

$$B^2_{t-1} = p^0_1 - \sum_{m=2}^{t-2} q^2_m > p^0_1 - \sum_{m=2}^{t-1} q^2_m = B^2_t$$

- Since $B^2_{t-1} > B^2_t$ and that $q^2_{t-1} = A^2_{t-1}$ (or else $t > n_2$), and we know that $A^2_{t-1} > 0$ from Lemma 6. Therefore,

$$q^j_t = \min(A^j_t, B^2_t) < \min(A^j_t, B^2_{t-1}) = \min(A^j_t, q^2_{t-1}) > 0$$

- If $q^j_t = B^2_t (> 0)$, we show by induction that for every $i \geq 1$ it holds that $q^2_{t+i} = B^2_{t+i} = 0$. For base consider $i = 1$. Then by using Lemma 6 we get

$$q^2_{t+1} = min(A^2_{t+1}, B^2_{t+1}) = min(A^2_{t+1}, B^2_t - q^2_t) = min(A^2_t, 0) = min(A^2_t, 0) = 0 \ .$$

  For the induction step consider that this property holds for $i-1$ and show for i. From the induction hypothesis and Lemma 6 we get

$$q^2_{t+i} = min(A^2_{t+i}, B^2_{t+i}) = min(A^2_{t+i}, B^2_{t+i-1} - q^2_{t+i-1}) = min(A^2_{t+i}, 0) = 0$$

So the hypothesis of the lemma holds for action $j = 2$.

We now assume that it holds for every action $\leq j-1$ and show it for action j, again by induction over t. For base, consider $t = j$

- $B^j_j \geq q^j_j > 0 = B^j_{j-1}$.
- $0 < q^j_j$, directly from Lemma 5.
- If $q^j_j = B^j_j (> 0)$, we show by induction that for every $i \geq 1$ it holds that $q^j_{j+i} = B^j_{j+i} = 0$. For the induction base consider $i = 1$. Then by using Lemma 6 and the since we assume that $t = j > n_{j-1}$, it holds that $q^{j-1}_j = 0$ therefore

$$q^j_{j+1} = min(A^j_{j+1}, B^j_{j+1}) = min(A^j_{j+1}, B^j_j + p^{-1}_{j-1} q^{j-1}_j - q^j_j) = min(A^j_{j+1}, p^{-1}_{j-1} q^{j-1}_j) = min(A^j_{j+1}, 0) = 0 \ .$$

  For the induction step consider that this property holds for $i-1$. Since $t = j > n_{j-1}$, it holds that

$q^{j-1}_{j+i-1} = 0$. Using Lemma 6, we get

$$q^j_{j+i} = min(A^j_{j+i}, B^j_{j+i}) = min(A^j_{j+i}, B^j_{j+i-1} - q^j_{j+i-1} + p^{-1}_{j-1} q^j_{j-1+i}) = min(A^j_{j+i}, 0) = 0$$

For the induction step we assume that the hypothesis of the lemma holds for agent $t-1$ and for both actions $j-1$ and j. We now show that it holds for agent t with action j.

- Since $t > n_{j-1}$ we get

$$\sum_{\tau=j-1}^{t} q^{j-1}_\tau = \sum_{\tau=j-1}^{t-1} q^{j-1}_\tau = \sum_{\tau=j-1}^{n_{j-1}} q^{j-1}_\tau$$

  And from the induction hypothesis $0 < q^j_{t-1}$, hence

$$B^j_t = p^{-1}_{j-1} \sum_{\tau=j-1}^{n_{j-1}} q^{j-1}_\tau - \sum_{\tau=j}^{t-1} q^j_\tau > p^{-1}_{j-1} \sum_{\tau=j-1}^{n_{j-1}} q^{j-1}_\tau - \sum_{\tau=j}^{t-2} q^j_\tau = B^j_{t-1}$$

- From the induction hypothesis and $t \leq n_j$ we get $0 < B^j_t$. From Lemma 6, $0 < A^j_t$, therefore $0 < q^j_t$.
- If $q^j_t = B^j_t (> 0)$, we show by induction that for every $i \geq 1$ it holds that $q^j_{t+i} = B^j_{t+i} = 0$. For the induction base consider $i = 1$. Then by using Lemma 6 and the since $t > n_{j-1}$, it holds that $q^{j-1}_t = 0$,therefore,

$$q^j_{t+1} = min(A^j_{t+1}, B^j_{t+1}) = min(A^j_{t+1}, B^j_j + p^{-1}_{j-1} q^{t-1}_j - q^j_t) = min(A^j_{t+1}, p^{-1}_{j-1} q^{t-1}_j) = min(A^j_{t+1}, 0) = 0 \; .$$

  For the induction step consider that this property holds for $i-1$. Since $t = j > n_{j-1}$, it holds that $q^{j-1}_{t+i-1} = 0$. Using Lemma 6, we get

$$q^j_{t+i} = min(A^j_{t+i}, B^j_{t+i}) = min(A^j_{t+i}, B^j_{t+i-1} - q^j_{t+i-1} + p^{-1}_{j-1} q^j_{t-1+i}) = min(A^j_{t+i}, 0) = 0$$

□

Proof of Lemma 9. For every $t_1 < t_2$, it holds that $[\sigma_{t_2} = j, S^{\vec{z}}_{t_2} \in \Gamma^{j-}_{t_2}] \cap [\sigma_{t_1} = j, S^{\vec{z}}_{t_1} \in \Gamma^{j-}_{t_1}] = \emptyset$, since if $\sigma_{t_1} = j$ then $S^{\vec{z}}_{t_2} \notin \Gamma^{j-}_{t_2}$. From Theorem 8, it implies that

$$\Pr[\exists t : \sigma_t = j, S^{\vec{z}}_t \in \Gamma^{j-}_t] = \sum_{t=1}^{T} \Pr[\sigma_t = j, S^{\vec{z}}_t \in \Gamma^{j-}_t] = \sum_{t=j}^{n_j} q^j_t = B^j_{n_j} + \sum_{t=j}^{n_j - 1} q^j_t$$

We now prove by induction on action j the following:

$$\Pr[\exists t : \sigma_t = j, S^{\vec{z}}_t \in \Gamma^{j-}_t] = \rho_j$$

The induction base is done for action $j = 2$ as follows

$$\Pr[\exists t : \sigma_t = 2, S_t^{\vec{z}} \in \Gamma_t^{2-}] = q_{n_2}^2 + \Sigma_{t=1}^{n_2-1} q_t^2 = p_1^0 - \sum_{\tau=2}^{n_2-1} q_\tau^2 + \sum_{t=1}^{n_2-1} q_t^2 = p_1^0$$

Suppose the induction hypothesis is true for action $j-1$. For action j we have,

$$\Pr[\exists t : \sigma_t = j, S_t^{\vec{z}} \in \Gamma_t^{j-}] = q_{n_j}^j + \Sigma_{t=1}^{n_j-1} q_t^j = p_{j-1}^{-1} \sum_{\tau=j-1}^{n_j-1} q_\tau^{j-1} - \sum_{\tau=j}^{n_j-1} q_\tau^j + \sum_{t=1}^{n_j-1} q_t^j = p_{j-1}^{-1} \sum_{\tau=j-1}^{n_j-1} q_\tau^{j-1}$$

From the induction hypothesis we have that

$$\Pr[\exists t : \sigma_t = j-1, S_t^{\vec{z}} \in \Gamma_t^{(j-1)-}] = \sum_{t=j-1}^{n_{j-1}} q_t^{j-1} = \rho_{j-1}$$

We get

$$\Pr[\exists t : \sigma_t = j, S_t^{\vec{z}} \in \Gamma_t^{j-}] = p_{j-1}^{-1} \sum_{\tau=j-1}^{n_j-1} q_\tau^{j-1} = p_{j-1}^{-1}\rho_{j-1} = \rho_j$$

which completes the proof of the lemma. □

Proof of Lemma 12. From the definitions of $f^j(y)$, we get

$$f^j(y) + 1 = \text{argmax}_{t+1}(\sum_{\tau=j}^{t-1} q_\tau^j < y\rho_j) = \text{argmax}_{t+1}(\sum_{\tau=j}^{t-1} p_j^{-1} q_\tau^j < y p_j^{-1}\rho_j)$$

Since $q_{\tau+1}^{j+1} \leq p_j^{-1} q_\tau^j$ for every agent $\tau < t \leq n_j$ from Lemma 3, and $p_j^{-1}\rho_j = \rho_{j+1}$, we get

$$f^j(y) < f^j(y) + 1 \leq \text{argmax}_{t+1}(\sum_{\tau=j}^{t-1} q_{\tau+1}^{j+1} < y\rho_{j+1}) = \text{argmax}_{t+1}(\sum_{\tau=j+1}^{t} q_\tau^{j+1} < y\rho_{j+1}) =$$

$$\text{argmax}_t(\sum_{\tau=j+1}^{t-1} q_\tau^{j+1} < y\rho_{j+1}) = f^{j+1}(y)$$

hence $f^j(y) < f^{j+1}(y)$ for every $y \in [0,1]$. □

Proof of Lemma 15. The first part, $f^i(y) \neq f^j(y)$, is direct consequence of Lemma 12. As for the second part, for every action j and each agent t, $q_t^j \geq 0$ and $\sum_{t=1}^{n_j} q_t^j = \rho_j$, by Lemma 9. We also have monotone increasing exploration between agents- for every $j \leq t < n_j - 1$ and for every $j \in \{2, ..., k-1\}$, $q_t^j < q_{t+1}^j$, as a result of Lemma 6 ($A_t^j < A_{t+1}^j$) and Lemma 8 ($q_t^j = A_t^j$).

Let $y \in [0,1]$. Then $f^j(y) = \text{argmax}_t(\sum_{\tau=1}^{t-1} q_\tau^j < y\rho_j)$. From the above get

$$0 = q_1^j \leq \sum_{\tau=1}^{t-1} q_\tau^j \leq \sum_{\tau=1}^{n_j} q_\tau^j = \rho_j$$

which completes the proof. $\square$

Proof of Lemma 18. For action $j \in \{2, \ldots, k\}$ let $(\psi^j)_{j=1}^k$ denote the exploration rates used in π_A, i.e., $\psi_t^j = \sum_{S_t^{\vec{z}} \in \Gamma_t^{j-}} \Pr_{\pi_A}[S_t^{\vec{z}}, \sigma_t = j]$. Both π_A and $\hat{\pi}$ are BIC policy algorithms, with the same recommendations for exploitation states. This implies that the only difference between the probabilities of π_A and $\hat{\pi}$ to know action j's reward at time t (i.e., the probability $S_t^{\vec{z}}[j] \neq *$) is the difference between the sum of exploration rates of action j until time t. Meaning that for every BIC policy π with the same recommendations for exploitation states like $\hat{\pi}$,

$$\Pr_\pi[S_t^{\vec{z}} \wedge \vec{z}[j] \neq *] = \sum_{\tau=1}^{t-1} \sum_{S_\tau^{\vec{z}} \in \Gamma_\tau^{j-}} \Pr_\pi[S_\tau^{\vec{z}}, \sigma_\tau = j] + \Pr_\pi[\forall i < j : x_i = -1, j < t]$$

The sum $\sum_{\tau=1}^{t-1} \sum_{S_\tau^{\vec{z}} \in \Gamma_\tau^{j-}} \Pr_\pi[S_\tau^{\vec{z}}, \sigma_\tau = j]$ is the sum of all exploration rates of action j until $t-1$, therefore

$$\Pr_{\hat{\pi}}[S_t^{\vec{z}} \wedge \vec{z}[j] \neq *] - \Pr_{\pi_A}[S_t^{\vec{z}} \wedge \vec{z}[j] \neq *] = \Sigma_{\tau=j}^{t-1} q_\tau^j - \Sigma_{\tau=j}^{t-1} \psi_\tau^j$$

for every agent t.

For contradiction, suppose that there exists a prior, D, an action j and time t_1 such that

$$\Pr_{\hat{\pi}}[S_{t_1}^{\vec{z}} \wedge \vec{z}[j] \neq *] < \Pr_{\pi_A}[S_{t_1}^{\vec{z}} \wedge \vec{z}[j] \neq *] \ , \tag{B.6}$$

where j is the least such action and t_1 is the least such agent for action j.

Every q_t^j was calculated inductively so that it would attain a maximum value and maintain the constraints in (3.1), independent of the other actions. Therefore it is not possible that there exists time $t_0 < t_1$ such that $\psi_t^j = q_t^j$ for every $t < t_0$ and $\psi_{t_0}^j > q_{t_0}^j$. If $\psi_t^j = q_t^j$ for every $t < t_1$, then $\psi_{t_1}^j \leq q_{t_1}^j$.

So let $t_0 < t_1$ denote the first time that there is lower exploration rate in π_A for action j rather than in $\hat{\pi}$, i.e., $t_0 = \text{argmin}_{t<t_1} \psi_t^j < q_t^j$ and $\psi_t^j = q_t^j$ for every $t < t_0$. Hence, the probability that π_A is in exploitation state at time $t = t_0 + 1$ w.r.t. $\hat{\pi}$ is lower, i.e.,

$$\sum_{S_t^{\vec{z}} \in \Gamma_t^{j+}} \Pr_{\pi_A}[S_t^{\vec{z}} | \sigma_t = j] = \sum_{S_t^{\vec{z}} \in \Gamma_t^{j+}} \Pr_{\pi_A}[S_t^{\vec{z}}] < \sum_{S_t^{\vec{z}} \in \Gamma_t^{j+}} \Pr_{\hat{\pi}}[S_t^{\vec{z}}] = \sum_{S_t^{\vec{z}} \in \Gamma_t^{j+}} \Pr_{\hat{\pi}}[S_t^{\vec{z}} | \sigma_t = j] \ ,$$

and the probability that π_A is in exploration state at time $t = t_0 + 1$ w.r.t. $\hat{\pi}$ is higher, i.e.,

$$\sum_{S_t^{\vec{z}} \in \Gamma_t^{j-}} \Pr_{\hat{\pi}}[S_t^{\vec{z}}] < \sum_{S_t^{\vec{z}} \in \Gamma_t^{j-}} \Pr_{\pi_A}[S_t^{\vec{z}}] \ .$$

Since $\mathbb{E}[u_t(j) - u_t(1) | S_t^{\vec{z}}, \sigma_t = j]$ and μ_j depend only on the prior D, their value remain the same. As a result of $\psi_{t_0}^j < q_t^j$, the exploration rate of action j in π_A at the next time, $t_0 + 1$ that maintains the constraint is smaller than it could have been while using $\hat{\pi}$, and for every time $t > t_0$ it holds that $\psi_t^j \leq q_t^j$. This is true for every time $t_0 < t_1$ such that $\psi_{t_0}^j < q_{t_0}^j$ and therefore contradicts (B.6). Therefore, $\psi_t^j \leq q_t^j$ for every action j and agent t.

The difference between the policies indicates that there is an action j agent t with $\psi_t^j \neq q_t^j$. Since $\psi_t^j \leq q_t^j$, it implies that $\psi_t^j < q_t^j$ and we get

$$\Pr_{\hat{\pi}}[S_{t+1}^{\vec{z}} \wedge \vec{z}[j] \neq *] < \Pr_{\pi_A}[S_{t+1}^{\vec{z}} \wedge \vec{z}[j] \neq *] \ , \tag{B.7}$$

Which completes the proof. □

Bibliography

[BST16] Gal Bahar, Rann Smorodinsky, and Moshe Tennenholtz. Economic recommendation systems: One page abstract. In *Proceedings of the 2016 ACM Conference on Economics and Computation, EC*, page 757, 2016.

[CBL06] Nicolò Cesa-Bianchi and Gabor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[CH13] Yeon-Koo Che and Johannes Hörner. Optimal design for social learning. 2013.

[DKQ16] Shaddin Dughmi, David Kempe, and Ruixin Qiang. Persuasion with limited communication. In *Proceedings of the ACM Conference on Economics and Computation, EC, Maastricht, The Netherlands, July 24-28*, pages 663–680, 2016.

[DX] Shaddin Dughmi and Haifeng Xu. Algorithmic persuasion with no externalities. In *Proceedings of the 2017 ACM Conference on Economics and Computation, EC, Cambridge, MA, USA, June 26-30, 2017*, pages 351–368.

[DX16] Shaddin Dughmi and Haifeng Xu. Algorithmic bayesian persuasion. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC*, pages 412–425, 2016.

[FKKK14] Peter Frazier, David Kempe, Jon M. Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *ACM Conference on Economics and Computation, EC '14, Stanford , CA, USA, June 8-12, 2014*, pages 5–22, 2014.

[GGW11] John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, 2011.

[KG11] Emir Kamenica and Matthew Gentzkow. Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

[KMP14] Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the ”wisdom of the crowd”. *J. of Political Economy*, 122:988–1012, October 2014. Preliminary version appeared in *ACM Conf. on Economics and Computation*, 2014.

[MSS15] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. *CoRR*, abs/1502.04147, 2015.

[MSSW16] Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, EC '16, pages 661–661, New York, NY, USA, 2016. ACM.

[MSW18] Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. Competing Bandits: Learning Under Competition. In *Innovations in Theoretical Computer Science Conference (ITCS 2018)*, 2018.

[Sli17] Aleksandrs Slivkins. Incentivizing exploration via information asymmetry. *ACM Crossroads*, 24(1):38–41, 2017.

תקציר

אנחנו בוחנים מקרה שבו מתכננת עבור הרווחה החברתית ניצבת מול זרם של סוכנים אנוכיים. המטרה של המתכננת היא למקסם את הרווחה החברתית (Social Welfare) באמצעות האסימטריה של האינפורמציה (לגבי תוצאות קודמות) וללא שימוש בתמריצים כספיים. המתכננת ממליצה לסוכנים על פעולות, ועל מנת שהסוכנים יעקבו אחריה ההמלצות צריכות להיות בעלות תאימות תמריץ בייסיאנית (Bayesian Incentive Compatible).

התוצאה העיקרית שלנו הינה אלגוריתם **אופטימלי** למתכננת, במקרה שבו הראליזציות הינן דטרמיניסטיות ובעלות תומך מוגבל, ובכך מהווה התקדמות משמעותית וחשובה בחקר הבעיה הפתוחה הזו. לפרוטוקול האופטימלי שלנו יש שתי תכונות מעניינות. הראשונה, הוא תמיד משלים את הגישוש (exploration) של הפעולות לפי סדר הרווחיות הא-פריורית שלהם. השניה, הרנדומיזציה בפרוטוקול מתואמת (correlated) בין סוכנים ובין פעולות (ולא בלתי תלויה ביניהם).

הפקולטה למדעים מדויקים ע״ש ריימונד ובברלי סאקלר
בית הספר למדעי המחשב ע״ש בלבטניק

אלגוריתם אופטימלי לגישוש בעל תאימות תמריץ בייסיאנית

חיבור זה הוגש כחלק מהדרישות לקבלת התואר

"מוסמך האוניברסיטה (M.Sc.)"

על ידי

לי כהן

עבודת המחקר בוצעה בהנחייתו של

פרופ' ישי מנצור