# Good Advice Costs Nothing and it's Worth the Price: Incentive Compatible Recommendation Mechanisms for Exploring Unknown Options

A thesis presented by

Perry Green

to

Computer Science

in partial fulfillment of the honors requirements

for the degree of Bachelor of Arts

Harvard College

Cambridge, Massachusetts

April 1, 2014

# Abstract

Recommender systems are valuable to their users to the extent that they have unique information about which options are best. One way that such a system can gain this knowledge is by recommending that a user explore an option whose value is unknown, and receiving the feedback of the user. If this is done too often, though, the quality of the recommendations provided may suffer to the point where users begin ignoring the system altogether. Therefore, I study the mechanism design problem of how a recommender can quickly learn the values of unknown options, within the constraint that it still be in agents' interests to follow the recommendations. The main conceptual contribution is a simplifying abstraction that transforms the problem from one of making decisions based on the total set of possible histories, into an acquisition problem where purchases made at one time affect the budget available in the future. I also characterize the optimal policy for exploring all options in a class of special cases, and prove that the recommender can decrease the time necessary to explore a particular target option by introducing new options.

# Chapter 1

## 1.1 Introduction

The ease of modern communication has opened opportunities for a plethora of services that collect information about the quality of items from its users, aggregate and analyze this information, and then use it to make recommendations or provide items directly. For example, Spotify is a music streaming application which, based on a users listening history, populates a 'Discover' page which recommends artists, albums, or songs that a user has not listened to before. Netflix, a television and movie streaming website, also recommends content based on the ratings and viewing behavior of its users. TripAdvisor and Yelp aggregate reviews and provide recommendations for hotels and restaurants respectively. Examples are practically limitless; there are services of this form making recommendations on everything from doctors[1] to college professors[2].

This structure can also be seen in companies that provide physical products themselves. Birchbox is a subscription service where people pay to receive monthly packages of items selected by Birchbox (mostly samples of beauty products). Customers can earn reward-points, redeemable for full-size products, by reviewing the samples sent to them. The value

[1] http://www.healthgrades.com/
[2] http://www.ratemyprofessors.com/

that Birchbox provides is therefore in part the result of the exploration by earlier customers who tried new products, and then reported back valuable information that Birchbox can use when constructing future packages. Birchbox writes of their feedback program (Corliss [2012]):

> You [Birchbox subscribers] aren't afraid to shake off your beauty shackles, try new things and, most importantly, let us know what you think! [...]
>
> And what happens to your feedback once we receive it? Well, we read it, of course! It gives us an idea of trending products, fuels ideas for future boxes, and helps us improve with each month!

A common thread throughout all of these services is a misalignment of interests between the center (e.g. Spotify, Birchbox) and the agents (subscribers). The agents primarily want the center to provide or recommend the option that, according to the current knowledge of the center, is the best. From the agents perspective the costs of exploring some new, and likely worse, choice are completely internalized. In contrast, the benefit of the exploration, the possibility that the information gained might improve future recommendations, is spread out over all agents, and so is overwhelmingly externalized. Agents will therefore desire to perform a socially sub-optimal amount of exploration of new options.

Conversely, from the center's perspective, the more information it has about the products it recommends, the greater the value of its service. The center cares about the recommendations given to all agents and so may have an interest in recommending possibly suboptimal options to some agents, with the hope that the information gained may lead to more valuable recommendations later on. If Spotify wants to be on the cutting edge of discovering new popular artists, it may need to recommend newly released albums even when it doesn't believe that they are the most likely to be the best recommendations for a given user. Birchbox may have more popular packages in the long run if it occasionally

takes a risk by sending a product with uncertain popularity, instead of a product with known but likely greater popularity, since the rare successes of the uncertain option can lead to better packages for many subsequent months.

Of course, the center does not have free reign to compel agents to explore unknown options. If agents do not believe that the recommendations given to them are in their best interest, they always have the option to ignore the recommendations completely, and act according to whatever prior beliefs they held about which option is best.

I therefore investigate the mechanism design question of what recommendation policy a center in this setting should adopt. I assume that there are a finite number of options, each with unknown value but a commonly known prior. The center first publicly adopts some policy that determines which option it will recommend to each agent in every circumstance. Each agent in an infinite sequence then has a choice either to receive the option recommended by the center, or to pick any option without knowing the recommendation. The agent then reports the value of the option they received back to the center. The center has one of several possible objectives that it is trying to achieve, related to gaining information about the values of the options.

The restriction that agents must follow a recommendation if they receive one is appropriate in settings where the center is actually providing the item being recommended. In these circumstances, the choice to receive a recommendation is frequently synonymous with receiving the recommended item itself. For example, subscribing with Birchbox is simultaneously choosing to receive a recommendation for beauty products, as well as contracting to purchase whatever beauty products are recommended.

Alternatively, this same restriction can be motivated under a slightly different setting: suppose that all agents must always follow the recommendation of the center, and do not have the option to instead choose an option independently. If the center enacts a recommendation policy as if the agents could opt to choose any option instead of receiving

a recommendation, it will ensure that no agent is made worse off under this policy than if instead there were no center to aggregate information, and each agent individually chose any option that they wanted. In effect, it ensures that no agent can complain that the existence of the center has made them worse off.

In Chapter 2 I formally define the model. In Chapter 3 I present the main conceptual contribution — notions of surplus and cost that can transform the problem of what option to recommend given a history of observations, into a simpler problem of how to best purchase items sequentially when past purchases influence future budgets. Using the surplus/cost abstraction, I demonstrate a variety of results about this setting, including the guarantee that it is always possible for the center to learn the values of all the options. This abstraction is also used throughout the remainder of the paper, and is a crucial tool for proving the subsequent results.

I then consider two possible goals for the center. In Chapter 4, I analyze a setting in which the center's objective is to learn the values of all the options within as few agents as possible, and attempt to find the optimal recommendation policy. The main result is a full characterization of the optimal policy the center can adopt for a particular class of distributions for the values. One natural member of this class is when the prior distributions for the values of all but the (ex-ante) best option are the same.

In Chapter 5 I consider a setting in which the center's objective is to instead learn the value of one particular target option. I ask whether the center can decrease number of agents necessary to learn the value of some specific target option by introducing new, additional options. This introduction of new options could correspond to a service like Netflix adding to its library of movies, or Birchbox assembling its packages by drawing from a larger pool of potential items. I demonstrate that, in fact, introducing 'dummy' options in this way does reduce the expected number of agents needed to learn the value for a target option. This implies, perhaps counter-intuitively, that recommendation systems

that wish to more efficiently gain information about some existing array of choices should broaden, rather than restrict, the possible options they might recommend. Chapter 6 is the conclusion, and presents ideas for future work in this area.

## 1.2 Related Work

There is a large body of existing literature on designing effective recommender systems. The problem is generally approached as a machine learning problem of how to leverage existing data about users and items in order to make accurate predictions about their future likes or dislikes. One common technique is nearest-neighbor models, which involve making recommendations based on known information about similar users/items. Another is latent factor models, which assume that the available high-dimensional data can be explained by comparatively low-dimensional hidden properties of the users/items. By learning these properties, one can make predictions about how much users will like some new item. For an overview of these techniques and others, see Lu et al. [2012]. In contrast with these methods, I analyze the missing information as a mechanism design problem, and not a statistical inference problem. Instead of the recommender making predictions based on the information it has, in my setting the recommender uses the recommendations themselves to acquire missing information.

In Golbandi et al. [2011] a recommender must determine the best questions to ask of a new user in order to elicit information for improving future recommendations. This is similar to my model, in that it is concerned with information elicitation, rather than inference. However, in their setting questions are asked as part of an interview process before any recommendations are made, whereas in my setting all information must be gained endogenously to the recommendation system itself.

Outside of recommender systems, the model here is reminiscent of the multi-armed

bandit (MAB) problem. In MAB, there is a set of possible actions, each of which has a reward non-deterministically generated by some process. An agent must sequentially decide which action to take. In doing so, the agent must balance exploration, trying new options to gain information about the likely reward, with exploitation, choosing actions that the agent has learned are likely to have high rewards. While typically there is only one agent in MAB, in Bolton and Harris [1999] they consider a case in which there are many agents, each of which can benefit from the exploration of earlier agents, as in this setting.

In my setting there is similarly an array of options from which agents choose one and receive a reward. Exploring new options is costly, but exploiting information about these options in the future can be beneficial. However, here each agent only acts once, and so only has reason to exploit. It is only because agents receive information about past actions exclusively through the center, who determines a policy for selectively revealing information in the form of a recommendation, that exploration becomes possible.

The idea of selectively revealing information in order to induce a particular action in others (as in the center giving a recommendation to an agent to induce exploration), is also present in Kamenica and Gentzkow [2011]. They consider a simpler two agent setting. The first agent is the sender, who has information and the ability to send a signal to the receiver. The second is the receiver, who must choose from a set of actions, and whose utility is a function of the information possessed by the sender. The policy for how the sender determines the signal from the information is known to the receiver, but the information itself is not. The sender wants to determine a policy for sending signals that will maximize the probability that the receiver takes some desired action.

While this is similar to the model for this paper, there are two primary differences. First, while in their model the receiver can decide any action after he has received the signal, in my model the agents choose to either not receive a recommendation, or to contract

with the center and always follow the recommendation. Second, the recommendation setting is sequential, in that the information available to the center depends on previous recommendations that have been made. So, the center must consider both the overall objective, as well as how recommendations for one agent may allow or disallow future recommendations to other agents.

Finally, the model presented here is closely based on the one in Kremer et al. [2013], and should be seen as expanding on and generalizing the setting that they present. Kremer et al. originally conceived of a center adopting a policy for recommending options to a sequence of agents, constrained by the requirement that it be in each agent's interest to follow the recommendation. There are three major alterations between their model and the one presented here. First, and most importantly, their model assumes that there are exactly two options that can be recommended, where no such assumption is made here. The generalization to more than two options vastly increases the complexity of possible policies that the center can adopt, which implies a need for more robust abstractions to succinctly reason about them. Second, like in Kamenica and Gentzkow [2011], in Kremer et al. [2013] agents can always receive a recommendation and then choose any option, whereas under the model presented here, before learning the recommended option agents must contract with the center and agree to follow the recommendation. Third, Kremer et al. [2013] assumes that the objective of the center is to maximize the average utility of the agents, to which exploring new options is merely an instrumental goal. In contrast, while the model presented here is general enough to accommodate an arbitrary objective for the center, both of the objectives considered explicitly in the paper are to gain information, and not increase the utility of the agents.

# Chapter 2

## 2.1 Notation and Model

There are $N \geq 2$ *options*. These represent the set of items that the center can recommend. Each of the options has a *value*, which are defined by continuous, independent, random variables $\mathcal{O}_1, \ldots \mathcal{O}_N$ respectively. These correspond to the quality of each option, and the utility that each option provides to an agent. Note that a value is a random variable, not a distribution, and so a single realization for the value of an option will hold for all agents. This implies that the that the utility for receiving any given option is the same for all agents. The vector $(\mathcal{O}_1, \ldots \mathcal{O}_N)$ is denoted $\boldsymbol{\mathcal{O}}$, and has continuous probability density function $f$. For all $j$, the distribution for $\mathcal{O}_j$ has probability density function $f_j$, and $\mu_j$ is defined to be $E(\mathcal{O}_j)$. All of the distributions are commonly known to all agents and the center. All of the realizations of the values are initially unknown.

In order to make learning the values of the options non-trivial for the center, I assume that there is a single 'best' option ex-ante — if not for the recommendation mechanism provided by the center, all agents would maximize their expected utility by choosing this single option, and so the center would not be able to learn the values of the other options. So, let $\mu_1 > \mu_2 \geq \cdots \geq \mu_N$. The strict inequality corresponds to requiring that there be a single option which has higher expected value than all others. The weak inequalities are

without loss of generality.

Additionally, I assume that:

$$P(\mathcal{O}_1 \leq \mu_2) > 0 \qquad \text{(exploration possibility condition)}$$

Below, I show that this condition is necessary for the center to learn values other than $\mathcal{O}_1$. At the end of Chapter 3, I use the surplus/cost abstraction to show that it is also sufficient.

There is an infinite sequence of unique *agents* who arrive in order, choose an option $j$, receive utility $\mathcal{O}_j$, and then report the value $\mathcal{O}_j$ back to the center. Agents are risk-neutral utility maximizers. They are aware of their location in the sequence, but do not know the options chosen by previous agents, or the values that previous agents have received. Conversely, the center is always aware of the history at any point in time, which includes both which option each agent chose, as well as the realized values for those options. If an agent is the first in the sequence to choose option $j$, then they are said to have *explored* $\mathcal{O}_j$.

The center publicly adopts a *recommendation policy* $\pi$, known to all agents, which is a function from histories to $[N]$ (where $[N]$ is the set $\{1, 2, \ldots N\}$). $\pi(h) = j$ denotes that if the center observes history $h$, the next option it will recommend is option $j$ (abusing notation, I will write 'recommending $\mathcal{O}_j$' to mean recommending option $j$ from now on). Before choosing an option on their own, agents may first contract with the center. If they do, then they learn the recommendation of the center as defined by the recommendation policy, and must then choose this option.

Without a recommendation, the highest expected utility an agent can receive is by picking option 1, receiving $\mu_1$. Therefore, in order to incentivize agents to follow the recommendation policy, the center must ensure that the expected value for the recommended option is greater than $\mu_1$. Let $rec_j^i$ be the event that a given policy recommends $\mathcal{O}_j$ to

agent $i$. A policy $\pi$ is *ex-ante incentive compatible* (IC) if, for all $i$:

$$\sum_{j\in[N]} E(O_j|rec_j^i)P(rec_j^i) \geq \mu_1 \qquad \text{(ex-ante IC condition)}$$

Intuitively, this means that each agent is weakly better off always following the recommendation than they would be if they instead always picked $\mathcal{O}_1$, the best option to choose without having a recommendation. The use of 'ex-ante' in this context refers to the fact that it is in the agent's interest to follow the recommendation prior to knowing what option is recommended.[1]

The IC condition motivates the 'exploration possibility condition' stated above. If there is no possibility that $\mathcal{O}_1 \leq \mu_2$, then there is no circumstance under which it would be IC for an agent to explore an option other than $\mathcal{O}_1$.

**Theorem 2.1.1.** *If $P(\mathcal{O}_1 \leq \mu_2) = 0$, then the unique IC policy is to always recommend option 1 to all agents.*

*Proof.* Let agent $k$ be the first agent such that $P(rec_j^k) > 0$ for some $j \neq 1$. Then for all $j \neq 1$, $E(\mathcal{O}_j|rec_j^k) = \mu_j \leq \mu_2$, since whether the center recommends option $j$ cannot depend $\mathcal{O}_j$ before the center learns $\mathcal{O}_j$, and the center cannot learn $\mathcal{O}_j$ before agent $k$ by choice of $k$. Since $P(\mathcal{O}_1 \leq \mu_2) = 0$, $E(\mathcal{O}_1|rec_j^k) > \mu_2$ for all $j$. But then, $\sum_{j\in[N]} E(\mathcal{O}_j|rec_j^k)P(rec_j^i) < \sum_{j\in[N]} E(\mathcal{O}_1|rec_j^k)P(rec_j^i) = \mu_1$ violating the IC condition. So, all agents must always be recommended option 1. □

Throughout the paper, I will consider two different objectives that the center might be trying to achieve. In both cases, an IC recommendation policy is *optimal* if no other IC policy performs strictly better according to the objective.

[1] In Kremer et al., the authors consider a slightly modified setting in which agents can choose any option even after learning the center's recommendation. Intuitively, this would imply a stricter IC condition. However, they also restrict the setting to the special case where there are only 2 options. When there are only 2 options, the two IC conditions are equivalent. See Appendix A for details.

# Chapter 3

## 3.1 Example of Reasoning about Policies

A short example of reasoning about recommendation policies will build intuitions and make subsequent arguments much easier to follow. Suppose there are 3 options ($N = 3$). $\mathcal{O}_1 \sim Unif(-2, 6)$, $\mathcal{O}_2 \sim Unif(-3, 5)$, and $\mathcal{O}_3 \sim Unif(-2, 4)$. First, the center must determine which option to recommend to agent 1. Observe that only recommending option 1 will be IC — if the policy recommended either of the other two options, the agent would prefer choose option 1 instead of receiving a recommendation. In fact, any IC policy will always recommend option 1 to agent 1 (not just in this example), since by assumption $\mu_1 > \mu_j$ for $j \neq 1$.

Next, consider what recommendation the center can give to agent 2. It is easy to verify that again always recommending option 1 would be IC, and always recommending either option 2 or 3 would again not be IC. However, now $\mathcal{O}_1$ has been explored, and so the center knows $\mathcal{O}_1$, and can condition the recommendation on this value. So, the center could define the policy as follows: if $\mathcal{O}_1 < 4$, then recommend option 2. Otherwise, recommend option 1. The expected value for agent 2 when he follows this policy is:

$$P(\mathcal{O}_1 < 4)E(\mathcal{O}_2) + P(\mathcal{O}_1 \geq 4)E(\mathcal{O}_1|\mathcal{O}_1 > 4) = \frac{3}{4} * 1 + \frac{1}{4} * 5 = 2 = \mu_1$$

So, this is IC. This is clearly not the only possibility. One can easily verify that, e.g., recommending option 2 when $\mathcal{O}_1 < 1$ and option 3 when $\mathcal{O}_1 > 5$ would also be IC.

Observe that in both of these examples, there are realizations for which following the recommendation is beneficial to the agent (when $\mathcal{O}_1 < 1$), and realizations for which following the recommendation is detrimental to the agent ($1 < \mathcal{O}_1 < 4$ in the former case, $5 < \mathcal{O}_1$ in the latter). By balancing the utilities associated with these two cases, the center has managed to explore other options within the constraints of the IC condition, even for high realizations of $\mathcal{O}_1$.

Next, let's consider the policy for agent 3 (assuming agent 2 explores $\mathcal{O}_2$ for $\mathcal{O}_1 < 4$). The center always knows $\mathcal{O}_1$. Also, if $\mathcal{O}_1 < 4$, then the center knows $\mathcal{O}_2$. The center can exploit all of this information to increase the utility of agent 3 for some realizations of $\boldsymbol{\mathcal{O}}$, which enables the center to make agent 3 explore in other realizations. To do this most effectively, for all realization of $\boldsymbol{\mathcal{O}}$ for which the center does not recommend that agent 3 explore, the center should recommend the *maximum* of the known options.

For example, consider recommending $\mathcal{O}_3$ for $\mathcal{O}_1 > 5$, $\mathcal{O}_1$ for $4 \leq \mathcal{O}_1 \leq 5$ and $\max(\mathcal{O}_1, \mathcal{O}_2)$ for $\mathcal{O}_1 < 4$. The utility for agent 3 would then be:

$$
\begin{aligned}
&P(\mathcal{O}_1 > 5)E(\mathcal{O}_3) + P(4 \leq \mathcal{O}_1 \leq 5)E(\mathcal{O}_1 | 4 \leq \mathcal{O}_1 \leq 5) \\
&+ P(\mathcal{O}_1 < 4 \wedge -2 < \mathcal{O}_2 < 4)E(\max(\mathcal{O}_1, \mathcal{O}_2) | \mathcal{O}_1 < 4 \wedge -2 < \mathcal{O}_2 < 4) \\
&+ P(\mathcal{O}_1 < 4 \wedge \mathcal{O}_2 < -2)E(\mathcal{O}_1 | \mathcal{O}_1 < 4) + P(\mathcal{O}_1 < 4 \wedge \mathcal{O}_2 > 4)E(\mathcal{O}_2 | \mathcal{O}_2 > 4) \\
&= \tfrac{1}{8}1 + \tfrac{1}{8}\tfrac{9}{2} + \tfrac{3}{4}\tfrac{3}{4}2 + \tfrac{3}{4}\tfrac{1}{8}1 + \tfrac{3}{4}\tfrac{1}{8}\tfrac{9}{2} = \tfrac{149}{64} > \mu_1
\end{aligned}
$$

So, this policy would be IC for agent 3 even though, unlike agent 2, agent 3 doesn't explore for any realizations which are beneficial. Instead, agent 3 is able to benefit from the exploration of earlier agents, which enables the center to recommend the maximum of $\mathcal{O}_1$ and $\mathcal{O}_2$ for some realizations.

The key takeaway is that exploring can be enabled in two ways. First, by having an agent explore $\mathcal{O}_j$ for realizations $\mathcal{O}_1 < \mu_j$. Second, by exploiting the exploration of earlier

agents. The surplus/cost abstraction formalizes and quantifies this insight, reframing the IC condition in terms of the costs associated with exploring a new option for some set of realizations, and the surplus that exploration from previous agents has provided.

## 3.2 Surplus/Cost Abstraction

### 3.2.1 Surplus

Surplus represents the increase in utility that exploration by previous agents provides. Let $supp(\boldsymbol{\mathcal{O}}) = \{\boldsymbol{\mathcal{O}} : f(\boldsymbol{\mathcal{O}}) > 0\}$ be the support of $\boldsymbol{\mathcal{O}}$, and $EXPL$ be the set of functions $expl : supp(\boldsymbol{\mathcal{O}}) \mapsto 2^{[N]}$. Members of $EXPL$ are referred to as exploration states. Intuitively, an exploration state $expl \in EXPL$ denotes which options are explored for each possible realization of $\boldsymbol{\mathcal{O}}$ at a given time.

Define the *surplus* function $S : EXPL \mapsto \mathbb{R}$ by:

$$S(expl) = \int_{supp(\boldsymbol{\mathcal{O}})} f(\boldsymbol{\mathcal{O}}) \left( \max_{k \in expl(\boldsymbol{\mathcal{O}})} (\mathcal{O}_k) - \mathcal{O}_1 \right) d\boldsymbol{\mathcal{O}} \qquad \text{(surplus function)}$$

$S(expl)$ is the increase in utility that an agent would get if he was always recommended the maximum of the previously explored options (according to $expl$), as opposed to not participating in the mechanism (and so always choosing $\mathcal{O}_1$). The *marginal surplus* for a realization, denoted $S|\boldsymbol{\mathcal{O}}$ is simply the integrand evaluated at that realization.

$$S|\boldsymbol{\mathcal{O}}(expl) = f(\boldsymbol{\mathcal{O}}) \left( \max_{k \in expl(\boldsymbol{\mathcal{O}})} (\mathcal{O}_k) - \mathcal{O}_1 \right) \qquad \text{(marginal surplus)}$$

The surplus function for a region (set of realizations) $R$, $S|R$, is similarly defined by:

$$S|R(expl) = \int_{R} f(\boldsymbol{\mathcal{O}}) \left( \max_{k \in expl(\boldsymbol{\mathcal{O}})} (\mathcal{O}_k) - \mathcal{O}_1 \right) d\boldsymbol{\mathcal{O}} \qquad \text{(surplus for a region)}$$

The interpretations for these functions is straightforward: the marginal surplus is the instantaneous increase in utility at $\boldsymbol{\mathcal{O}}$ from being recommended the maximum explored option, instead of always $\mathcal{O}_1$, given exploration state $expl$. The surplus over a region is the increase in utility for an agent due to exploration over that region, which is just the result of integrating over the marginal surplus. The total surplus can then just be seen as the surplus over all possible realizations.

For all IC policies, before agent 1 no options are explored, and agent 1 always explores $\mathcal{O}_1$. For agents $i > 1$, each policy defines an exploration state $expl_i$, where $expl_i(\boldsymbol{\mathcal{O}}) = A$ denotes that given a realization $\boldsymbol{\mathcal{O}}$, the options in $A$ have been explored before agent $i$. Given a fixed policy, the surplus for an agent $i$ is defined by $S_i = S(expl_i)$, and similarly for $S_i|\boldsymbol{\mathcal{O}}$ and $S_i|R$

### 3.2.2 Gain

Gain represents how much the exploration by an agent according to a policy will increase the surplus for the subsequent agent. Let $single(A) = \{\{a\} : a \in A\}$ be the set of singleton subsets of $A$, and $E$ be the set of functions $e : supp(\boldsymbol{\mathcal{O}}) \mapsto single([N]) \cup \{\emptyset\}$. Intuitively, a function $e \in E$ represents exploration performed by a single agent, mapping each realization either to the set containing the option he explores for that realization, or to the empty set if he does not explore for that realization (and so is recommended the maximum known value).[1]

For $expl \in EXPL$ and $e \in E$, let $expl \cup e \in EXPL$ be the exploration state defined by $expl \cup e(\boldsymbol{\mathcal{O}}) = expl(\boldsymbol{\mathcal{O}}) \cup e(\boldsymbol{\mathcal{O}})$. Intuitively, $expl \cup e$ is the exploration state if you have previously explored according to $expl$, and then explore new options according to $e$.

From this it is possible to define the gain in surplus that exploring new options for some

[1]For readers having trouble remembering the distinction between $e/E$ and $expl/EXPL$, note that $e/E$ have a single letter, and so represent exploration taken by only a single agent, while $expl/EXPL$ have multiple letters, and so represent the previous exploration of multiple agents.

realizations will provide. Define the gain function $Gain : EXPL \times E \mapsto \mathbb{R}$ by:

$$Gain(expl, e) = S(expl \cup e) - S(expl) \qquad \text{(gain function)}$$

If previously the exploration state is $expl$, and new options are explored according to $e$, then $Gain(expl, e)$ represents the increase in surplus provided by that new exploration. Define the marginal gain and the gain for a region by $Gain|\boldsymbol{\mathcal{O}} = S|\boldsymbol{\mathcal{O}}(expl \cup e) - S|\boldsymbol{\mathcal{O}}(expl)$ and $Gain|R = S|R(expl \cup e) - S|R(expl)$ respectively. These have analogous interpretations to the marginal surplus and surplus for a region.

A policy defines for each agent $i$ the function $e_i$, the exploration performed by that agent, including both the realizations for which the agent explores, and the option explored for each of these realizations. These definitions allow the expression of basic identities, such as $expl_i \cup e_i = expl_{i+1}$ (the exploration performed before agent $i+1$ is exactly the exploration by agent $i$, combined with the exploration performed by agent $i$), and $S_i + Gain(S_i, e_i) = S_{i+1}$ (the surplus for agent $i+1$ is the surplus for agent $i$ plus the gain in surplus from agent $i$'s exploration).

### 3.2.3 Cost

Cost represents how much exploration by an agent according to a policy will decrease the utility for that agent. Abusing notation, for $e \in E$ let $supp(e) = \{\boldsymbol{\mathcal{O}} : e(\boldsymbol{\mathcal{O}}) \neq \emptyset\}$ be the set of realizations for which $e$ implies that an agent is actually exploring.

Define the cost function $Cost : EXPL \times E \mapsto \mathbb{R}$ as:

$$Cost(expl, e) = \int_{supp(e)} f(\boldsymbol{\mathcal{O}}) \left(\mathcal{O}_1 - \mathcal{O}_{e(\boldsymbol{\mathcal{O}})}\right) d\boldsymbol{\mathcal{O}} + S|supp(e)(expl) \qquad \text{(cost function)}$$

The cost is the loss in utility from being recommended to explore according to $e$, instead

of always being recommended the maximum known values. The first term is the difference in utility between being recommended $\mathcal{O}_1$ and being recommended according to $e$. The second is the additional loss of the surplus the agent would have received if he had been recommended not just $\mathcal{O}_1$, but the maximum of the known options. When they need to be referred to separately, I will refer to the first term as the *base cost*, and the second as the *opportunity cost*. The marginal cost (for values in the support of $e$) and cost for a region $R \subseteq supp(e)$ are defined $Cost|\boldsymbol{\mathcal{O}}(expl, e) = f(\boldsymbol{\mathcal{O}})\left(\mathcal{O}_1 - \mathcal{O}_{e(\boldsymbol{\mathcal{O}})}\right) + S|\boldsymbol{\mathcal{O}}(expl)$ and $Cost|R(expl, e) = \int_R f(\boldsymbol{\mathcal{O}})\left(\mathcal{O}_1 - \mathcal{O}_{e(\boldsymbol{\mathcal{O}})}\right) d\boldsymbol{\mathcal{O}} + S|R(expl)$ respectively.

It will frequently be necessary to refer to the ratio of gain to cost. The *gain-to-cost ratio* for exploring an option at a point with positive cost is simply the marginal gain at that point divided by the marginal cost at that point. Similarly, for a region that has positive marginal cost everywhere, the gain-to-cost ratio for exploring is just the gain from exploring divided by the cost from exploring.

If a point has negative or 0 cost, then its gain-to-cost ratio is defined to be $\infty$. This matches the intuitive idea that gain-to-cost should be a measure of how 'efficient' a purchase is, and purchasing something with 0 or negative cost is more efficient than any purchase with positive cost.

### 3.2.4 Reframing the IC Condition

The IC condition can then be rewritten the surplus/cost notation. Intuitively, since the surplus represents the increase in utility from recommending the maximum value explored instead of $\mathcal{O}_1$, and the cost represents a decrease in utility from exploring instead of recommending the maximum known value, a recommendation should have higher expected utility than $\mu_1$ iff the cost is less than the surplus. I confirm this intuition below.

**Theorem 3.2.1.** *A policy is IC iff agent 1 explores $\mathcal{O}_1$, and for each agent $i > 1$:*

$$Cost(expl_i, e_i) \leq S_i \qquad \text{(cost/surplus inequality)}$$

*Proof.* As observed earlier, all IC policies must recommend $\mathcal{O}_1$ to agent 1. So, consider agent $i > 1$. The utility for agent $i$ under the recommendation policy can be written as:

$$\int_{supp(e_i)} f(\boldsymbol{\mathcal{O}})\mathcal{O}_{e_i(\boldsymbol{\mathcal{O}})}d\boldsymbol{\mathcal{O}} + \int_{supp(\boldsymbol{\mathcal{O}})-supp(e_i)} f(\boldsymbol{\mathcal{O}})\mathcal{O}_1 d\boldsymbol{\mathcal{O}} + S_i|(supp(\boldsymbol{\mathcal{O}}) - supp(e_i))$$

To see this, note that the utility is just the value recommended at $\boldsymbol{\mathcal{O}}$ times $f(\boldsymbol{\mathcal{O}})$ integrated over $supp(\boldsymbol{\mathcal{O}})$. The first term in the sum is that integral over the region $supp(e_i)$. The sum of the second and third terms is the integral over $supp(\boldsymbol{\mathcal{O}}) - supp(e_i)$. It follows that the utility of agent $i$ is:

$$\int_{supp(e_i)} f(\boldsymbol{\mathcal{O}})\mathcal{O}_{e_i(\boldsymbol{\mathcal{O}})}d\boldsymbol{\mathcal{O}} - \int_{supp(e_i)} f(\boldsymbol{\mathcal{O}})\mathcal{O}_1 d\boldsymbol{\mathcal{O}} + \int_{supp(\boldsymbol{\mathcal{O}})} f(\boldsymbol{\mathcal{O}})\mathcal{O}_1 d\boldsymbol{\mathcal{O}} + S_i - S_i|supp(e_i)$$

$$= \int_{supp(e_i)} f(\boldsymbol{\mathcal{O}})\left(\mathcal{O}_{e_i(\boldsymbol{\mathcal{O}})} - \mathcal{O}_1\right) d\boldsymbol{\mathcal{O}} - S_i + \mu_1 + S_i = -Cost(expl_i, e_i) + \mu_1 + S_i$$

So, invoking the definition of IC, the policy will be IC so long as:

$$-Cost(expl_i, e_i) + \mu_1 + S_i \geq \mu_1$$

$$S_i \geq Cost(expl_i, e_i)$$

□

From now on I use the cost/surplus inequality interchangeably with the IC condition.

## 3.3 Policies using Surplus/Cost

### 3.3.1 Example cont.

Returning to the earlier example, we can see how to use cost/surplus inequality to show the policy from the beginning of the chapter is IC for the first 3 agents. The first agent explores $\mathcal{O}_1$, as required. The second agent has a surplus of 0 ($S_2 = 0$), and explores $\mathcal{O}_2$ when $\mathcal{O}_1 < 4$. This means exploring option 2 over the region $R_2 = [-2, 4] \times [-3, 5] \times [-2, 4]$. To find the cost we calculate:

$$S|R_2(expl_2) = \int_R f(\boldsymbol{\mathcal{O}}) \left(\mathcal{O}_1 - \mathcal{O}_1\right) d\boldsymbol{\mathcal{O}} = 0$$

$$\int_{R_2} f(\boldsymbol{\mathcal{O}})(\mathcal{O}_1 - \mathcal{O}_2)d\boldsymbol{\mathcal{O}} = P(\boldsymbol{\mathcal{O}} \in R_2) \left(E(\mathcal{O}_1|\mathcal{O}_1 \in [-2, 4]) - E(\mathcal{O}_2|\mathcal{O}_2 \in [-3, 5])\right) = 0$$

So, the $cost(expl_2, e_2) = 0 \leq 0 = S_2$ as desired. Observe that the region for which it was beneficial to explore $\mathcal{O}_2$, when $\mathcal{O}_1 < 1$, had a negative cost, and so allowed exploration even when there was no surplus. In fact, it is now possible to reinterpret the exploration possibility condition (Theorem 2.1.1) as proving that no exploration is possible unless there is some region with negative cost.

For agent 3, we first calculate $S_3$:

$$S_3 = \int_{supp(\boldsymbol{\mathcal{O}})} f(\boldsymbol{\mathcal{O}}) \left(\max_{k \in expl_3(\boldsymbol{\mathcal{O}})} (\mathcal{O}_k) - \mathcal{O}_1\right) d\boldsymbol{\mathcal{O}}$$

$$= \int_{R_2} f(\boldsymbol{\mathcal{O}}) \left(\max(\mathcal{O}_2, \mathcal{O}_1) - \mathcal{O}_1\right) d\boldsymbol{\mathcal{O}} + \int_{supp(\boldsymbol{\mathcal{O}}) - R_2} f(\boldsymbol{\mathcal{O}}) \left(\max(\mathcal{O}_1) - \mathcal{O}_1\right) d\boldsymbol{\mathcal{O}}$$

$$= P(\boldsymbol{\mathcal{O}} \in R_2)(E(max(\mathcal{O}_1, \mathcal{O}_2)|\boldsymbol{\mathcal{O}} \in R_2) - E(\mathcal{O}_1|\boldsymbol{\mathcal{O}} \in R_2)) + 0$$

$$= \left(\frac{3}{4}\right)\left(\frac{1}{8} * 1 + \frac{1}{8}\frac{9}{2} + \frac{3}{4}2 - 1\right) = \frac{57}{64}$$

Agent 3 explores $\mathcal{O}_3$ when $\mathcal{O}_1 > 5$, so for the region $R_3 = [5,6] \times [-3,5] \times [-2,4]$. $S|R_3(expl_3) = 0$ (the work is identical to above), and:

$$\int_{R_3} f(\boldsymbol{\mathcal{O}})(\mathcal{O}_1 - \mathcal{O}_3)d\boldsymbol{\mathcal{O}} = P(\boldsymbol{\mathcal{O}} \in R_3)\left(E(\mathcal{O}_1|\mathcal{O}_1 \in [5,6]) - E(\mathcal{O}_3|\mathcal{O}_3 \in [-2,4])\right)$$

$$= \frac{1}{8}\left(\frac{11}{2} - 1\right) = \frac{36}{64}$$

As desired $cost(expl_3, e_3) = \frac{36}{64} \leq \frac{57}{64} = S_3$, so again it is IC. Observe that, even though there was no region explored that had negative cost, exploration was possible because of the surplus generated by agent 2.

### 3.3.2 Possible Policies and Distinguishability

The example above demonstrates how to confirm that a given policy is IC using surplus/cost. However, we do not yet have the tools to properly describe new policies with this language in a way that ensures that the result validly describes a policy that is possible. For example, continuing the example above, it would be possible for a policy to recommend that agent 4 should explore $\mathcal{O}_3$ for the region $R_4 = [-2,4] \times [-3,4] \times [-2,4]$. This is because agent 1 will have already explored $\mathcal{O}_1$ for all realizations $\boldsymbol{\mathcal{O}}$, and agent 2 will have explored $\mathcal{O}_2$ when $\mathcal{O}_1 < 4$. So, even though the center will not know the complete realization $\boldsymbol{\mathcal{O}}$ by agent 4, the center will be able to know with certainty whether or not $\boldsymbol{\mathcal{O}} \in R_4$.

In contrast, it would not describe a possible policy to say that agent 4 should explore $\mathcal{O}_2$ for (exactly) the region $R_4' = [4,5] \times [-3,5] \times [-2,0]$, because when recommending for agent 4, if $\mathcal{O}_1 \in [4,5]$, the center will not be able to know based on the history of options

explored whether or not $\mathcal{O}_3 \in [-2, 0]$. So, it is not possible for a policy to recommend $\mathcal{O}_3$ to agent 4 for only $R'_4$, given the information the center will have at the time. It would also obviously not describe a possible policy to say that both $\mathcal{O}_2$ and $\mathcal{O}_3$ should be recommended to agent 4 over the same region.

To allow for concise descriptions of policies using surplus/cost, but ensuring that these problems do not arise, I introduce the notion of distinguishability. For a given policy and agent $i$, two realizations $\boldsymbol{\mathcal{O}} = (\mathcal{O}_1, \ldots \mathcal{O}_N)$ and $\boldsymbol{\mathcal{O}'} = (\mathcal{O}'_1, \ldots \mathcal{O}'_N)$ are *distinguishable* to that agent if there exists $j \in [N]$ such that $\mathcal{O}_j \neq \mathcal{O}'_j$, and $j \in expl_i(\boldsymbol{\mathcal{O}}) \cap expl_i(\boldsymbol{\mathcal{O}'})$. In words, two realizations are distinguishable to the agent if they differ on some coordinate $j$, and exploration before agent $i$ would allow the center to know $\mathcal{O}_j$ for both realizations, and so allow the center to rule out either $\boldsymbol{\mathcal{O}}$ or $\boldsymbol{\mathcal{O}'}$ as the actual realization. Region $R$ is said to be *closed under indistinguishability* (or CUI) for agent $i$ if for every point $\boldsymbol{\mathcal{O}} \in R$, if $\boldsymbol{\mathcal{O}}$ and $\boldsymbol{\mathcal{O}'}$ are not distinguishable by agent $i$ (and $\boldsymbol{\mathcal{O}'} \in supp(\boldsymbol{\mathcal{O}})$), then $\boldsymbol{\mathcal{O}'} \in R$.

It is now possible to describe an IC policy simply by listing region/option pairs for each agent. Each list denotes that the policy will have the agent explore each option for the corresponding region, and explore the maximum of the known options everywhere else. A policy which recommends $(R_1, \mathcal{O}_{i_1}), (R_2, \mathcal{O}_{i_2}), \ldots (R_\ell, \mathcal{O}_{i_\ell})$ to agent $k$ is possible and IC if it satisfies the cost/surplus inequality, all pairs of regions $R_i \neq R_j$ are disjoint, and all regions $R_i$ are CUI to agent $k$. The CUI condition ensures that, based only on the observed history, the center will be able to tell apart any two realizations for which the recommendation differs, and so that the policy described is possible.

The following aids in reasoning about CUI regions, and so which regions it is possible for an agent to explore:

**Lemma 3.3.1.** *Suppose that $R$ and $R'$ are both CUI for agent $k$. Then $R \cap R'$, $R \cup R'$ and $R - R'$ are also CUI for agent $k$.*

*Proof.* Suppose $\boldsymbol{\mathcal{O}} \in R \cap R'$, and $\boldsymbol{\mathcal{O}}$ and $\boldsymbol{\mathcal{O}}'$ are not distinguishable by agent $k$. Since $R$ is CUI, $\boldsymbol{\mathcal{O}}' \in R$. Similarly, $\boldsymbol{\mathcal{O}}' \in R'$. So, $R \cap R'$ is CUI.

If $\boldsymbol{\mathcal{O}} \in R \cup R'$, then either $\boldsymbol{\mathcal{O}} \in R$ or $\boldsymbol{\mathcal{O}} \in R'$. If $\boldsymbol{\mathcal{O}}'$ is not distinguishable from $\boldsymbol{\mathcal{O}}$ by agent $k$, then, either $\boldsymbol{\mathcal{O}}' \in R$ or $\boldsymbol{\mathcal{O}}' \in R'$, and so $\boldsymbol{\mathcal{O}}' \in R \cup R'$. So, $R \cup R'$ is CUI.

Suppose $\boldsymbol{\mathcal{O}} \in R - R'$, and $\boldsymbol{\mathcal{O}}$ and $\boldsymbol{\mathcal{O}}'$ are not distinguishable by agent $k$. Since $\boldsymbol{\mathcal{O}} \in R$ and $R$ is CUI, $\boldsymbol{\mathcal{O}}' \in R$. So, $\boldsymbol{\mathcal{O}}' \in R - R'$ or $\boldsymbol{\mathcal{O}}' \in R \cap R'$. If $\boldsymbol{\mathcal{O}}' \in R \cap R'$, this contradicts that $R \cap R'$ is CUI, since $\boldsymbol{\mathcal{O}}$ and $\boldsymbol{\mathcal{O}}'$ are not distinguishable. So, $\boldsymbol{\mathcal{O}}' \in R - R'$, and $R - R'$ is CUI. □

## 3.4 Exploration using Surplus/Cost

Using the surplus/cost abstraction, it is possible to prove general properties about exploration independent of the center's objective. For example, while the gain from exploring $\mathcal{O}_i$ over $R$ may depend on the full distribution for $\mathcal{O}_i$, the cost for exploring depends only on $\mu_i$.

**Theorem 3.4.1.** *If region $R$ is CUI for agent $k$, then the cost for exploring $\mathcal{O}_i$ over $R$ is: $P(\boldsymbol{\mathcal{O}} \in R)(E(\mathcal{O}_1|\boldsymbol{\mathcal{O}} \in R) - \mu_i) + S_k|R$.*

*Proof.* The cost is:

$$\int_R f(\boldsymbol{\mathcal{O}})(\mathcal{O}_1 - \mathcal{O}_i)d\boldsymbol{\mathcal{O}} + S_k|R = P(\boldsymbol{\mathcal{O}} \in R)(E(\mathcal{O}_1|\boldsymbol{\mathcal{O}} \in R) - E(\mathcal{O}_i|\boldsymbol{\mathcal{O}} \in R)) + S_k|R$$

So, it suffices to show that $E(\mathcal{O}_i|\boldsymbol{\mathcal{O}} \in R) = \mu_i$.

Suppose $R$ includes $\boldsymbol{\mathcal{O}} = (\mathcal{O}_1, \ldots \mathcal{O}_i, \ldots \mathcal{O}_N)$. Since $R$ is CUI, and $\mathcal{O}_i$ is not previously explored, it includes the line $\ell = \{(\mathcal{O}_1, \ldots \mathcal{O}'_i, \ldots \mathcal{O}_N) : \mathcal{O}'_i \in \mathit{supp}(\mathcal{O}_i)\}$. For any such line $\ell$, consider $E(\mathcal{O}_i|\boldsymbol{\mathcal{O}} \in \ell)$:

$$E(\mathcal{O}_i|\boldsymbol{\mathcal{O}} \in \ell) = \int_{supp(\boldsymbol{\mathcal{O}})} \mathcal{O}_i f(\boldsymbol{\mathcal{O}}|\boldsymbol{\mathcal{O}} \in \ell) d\boldsymbol{\mathcal{O}} = \int_{\ell} \mathcal{O}_i \frac{\prod_j f_j(\mathcal{O}_j)}{\prod_{j \neq i} f_j(\mathcal{O}_j)} d\boldsymbol{\mathcal{O}}$$

$$= \int_{\ell} \mathcal{O}_i f_i(\mathcal{O}_i) d\boldsymbol{\mathcal{O}} = \mu_i$$

Partition $R$ into equivalence classes, where two points $\boldsymbol{\mathcal{O}}$ and $\boldsymbol{\mathcal{O}}'$ are equivalent if they differ only on the value $\boldsymbol{\mathcal{O}}_i$. By above, each of these classes must be a complete line. Since the expected value of $\boldsymbol{\mathcal{O}}_i$ within each line is $\mu_i$, the expected value across all lines, and so all of $R$, must also be $\mu_i$. □

It trivially follows that the costs for exploring the options over any fixed region $R$ are ordered in the same way as the means of the options.

**Corollary 3.4.2.** *If region $R$ is CUI for agent $k$, then the cost for exploring $\mathcal{O}_i$ over $R$ is less than the cost of exploring $\mathcal{O}_j$ over $R$ iff $\mu_i > \mu_j$.*

Intuitively, the gain-to-cost ratio seems like a good measure of what regions you should explore; if the surplus for an agent is viewed as a budget that enables exploration, then exploring regions with a high gain-to-cost ratio is analogous to spending your fixed budget in the most efficient way possible for increasing the surplus of the subsequent agent. The surplus/cost abstraction shows that the gain-to-cost ratio for exploring a region is independent of the probability density over that region, but depends both on the maximum explored value for each point in the region and on the full distribution of the option being explored.

**Theorem 3.4.3.** *Let $\ell = \{(\mathcal{O}_1, \ldots, \mathcal{O}'_i \ldots \mathcal{O}_n) : \mathcal{O}'_i \in supp(\mathcal{O}_i)\}$ be a line, $\mathcal{O}_i$ be an unexplored option over $\ell$, and $V$ be the max of the explored values in $\mathcal{O}_1 \ldots \mathcal{O}_{i-1}, \mathcal{O}_{i+1}, \ldots \mathcal{O}_N$. Then the marginal gain for exploring $\mathcal{O}_i$ over $\ell$ is $f(\ell)(E(\max(\mathcal{O}_i, V)) - V)$, and the*

*marginal cost for the line is $f(\ell)(V - \mu_i)$. So, the gain-to-cost ratio over $\ell$ is:*

$$\frac{E(\max(\mathcal{O}_i, V)) - V}{V - \mu_i}$$

*Since the numerator is non-increasing in $V$, and the denominator is increasing in $V$, it follows that the gain-to-cost ratio is decreasing in $V$.*

*Proof.* By the definition of *Gain*:

$$Gain|\ell = S|\ell(expl \cup e) - S|\ell(expl)$$

$$= f(\ell)\left(E(\max(\mathcal{O}_i, V)) - \mathcal{O}_1\right) - f(\ell)\left(V - \mathcal{O}_1\right)$$

$$= f(\ell)\left(E(\max(\mathcal{O}_i, V)) - V\right)$$

By Theorem 3.4.1:

$$Cost|\ell(expl, e) = f(\ell)\left(E(\mathcal{O}_1 - \mathcal{O}_i)\right) + S|\ell(expl)$$

$$= f(\ell)(\mathcal{O}_1 - \mu_i) + f(\ell)(V - \mathcal{O}_1) = f(\ell)(V - \mu_i)$$

□

Note the use of $V$, rather than $\mathcal{O}_j$, in the statement of the lemma above. This is meant to emphasize that it depends only on the actual *numerical value* of the maximum explored option, and not even what option that value came from.

In addition to allowing for a quantification of the gain-to-cost ratio, the surplus/cost abstraction demonstrates that early exploration cannot allow for new regions with better gain-to-cost ratios to become available; there are increasing costs, and decreasing gains, for continuing to explore different options over the same region.

**Lemma 3.4.4.** *The cost of exploring $\mathcal{O}_i$ over a region $R$ is non-decreasing, and the gain of exploring $\mathcal{O}_i$ over region $R$ is non-increasing, as more options other than $\mathcal{O}_i$ are explored over $R$.*

*Proof.* The marginal cost of exploring $\mathcal{O}_i$ for realization $\boldsymbol{\mathcal{O}}$ depends only on $\mathcal{O}_i$, $f(\boldsymbol{\mathcal{O}})$, and the maximum of the previously explored options at $\boldsymbol{\mathcal{O}}$. Exploring a new option $\mathcal{O}_j$ at $\boldsymbol{\mathcal{O}}$ cannot change $\mathcal{O}_i$ or $f(\boldsymbol{\mathcal{O}})$, but may increase that maximum of the explored options at $\boldsymbol{\mathcal{O}}$, and so increase the marginal cost. Since exploring never decreases the marginal cost at a point, it also cannot decrease the cost for a region, since that is just the integral of the marginal costs over the region. Analogous reasoning shows that gain is non-increasing. □

The above lemma, however, is far too weak. It is likely obvious that exploring a region will only increase the cost of that region (by increasing the opportunity cost), and cannot increase the gain from exploring some other option over that same region. However, exploring has the benefit of making more points distinguishable, and so allows for new regions to be explored that previously could not. So, the above lemma is not sufficient to show that agent $k$ exploring cannot make a new region $R$ CUI for agent $k+1$, such that exploring some other option over $R$ has higher gain-to-cost for than any CUI region agent $k$ could have explored.

The following theorem proves this stronger claim. If there is some region that is CUI for an agent, then if that agent had explored strictly fewer options over that same region, then that agent could have explored some other overlapping CUI region for higher gain and lower cost. It follows that exploring cannot make new regions with higher gain-to-cost CUI.

**Theorem 3.4.5.** *Let $A$ be the set of options previously explored for CUI region $R$ under policy $\pi$ for agent $k$. Also let agent $k$ exploring $\mathcal{O}_i \notin A$ for $R$ have cost $C$ and gain $G$.*

*Suppose that under policy $\pi'$, agent $k'$ has explored options $A' \subset A$ (where $A'$ includes at least $\mathcal{O}_1$) for region $R$. Then there exists some CUI region $R'$ where $R \cap R' \neq \emptyset$ such that agent $k'$ can explore $\mathcal{O}_i$ over $R'$ for cost $C' \leq C$ and gain $G' \geq G$.*

*Proof.* Let $S \supseteq R$ be the transitive closure of $R$ under indistinguishability for agent $k'$. Since all the added points are indistinguishable to some point in $R$ for agent $k'$, the set of explored options for all added points must also be $A'$, since two indistinguishable points must be explored for the same set of options. Note that $S$ is now CUI for agent $k'$.

If the cost for agent $k'$ to explore $\mathcal{O}_i$ over $S$ is no more than $C$, then let $R' = S$. Otherwise, define $F : \mathbb{R} \mapsto 2^S$ as follows: $F(n) = \{\boldsymbol{\mathcal{O}} : \boldsymbol{\mathcal{O}} \in S \wedge \max_{j \in A'}(\mathcal{O}_j) \leq n\}$. Let $c(n)$ be the cost to agent $k'$ of exploring $\mathcal{O}_i$ over $F(n)$. Choose $n'$ to satisfy $c(n') = C$, and let $R' = F(n')$.[2] In other words, if agent $k$ can explore $S$ for cost less than $C$, $R' = S$. Otherwise, $R'$ is the subset of $S$ with cost $C$ made by including points with the lowest maximum values explored before agent $k'$.

It immediately follows that $C'$, the cost for $R'$, is no more than $C$. I claim that $R'$ is CUI for agent $k'$, overlaps with $R$, and that exploring $\mathcal{O}_i$ over $R'$ has gain $G' \geq G$.

$R'$ is CUI: Suppose $\boldsymbol{\mathcal{O}} \in R'$, and $\boldsymbol{\mathcal{O}'}$ does not differ from $\boldsymbol{\mathcal{O}}$ on any option in $A'$. Since $S$ is CUI and $\boldsymbol{\mathcal{O}} \in R' \subseteq S$, $\boldsymbol{\mathcal{O}'} \in S$. Also, the maximum explored value of $\boldsymbol{\mathcal{O}}$ must be the same as that for $\boldsymbol{\mathcal{O}'}$, since they are the same for all explored options. So, by construction of $R'$, since $\boldsymbol{\mathcal{O}} \in R'$, $\boldsymbol{\mathcal{O}'} \in R'$.

$R$ overlaps with $R'$: Consider any point $\boldsymbol{\mathcal{O}} \in R'$. Since $R' \subseteq S$, $\boldsymbol{\mathcal{O}} \in S$. This means that either $\boldsymbol{\mathcal{O}} \in R$, or $\boldsymbol{\mathcal{O}}$ is indistinguishable from a point $\boldsymbol{\mathcal{O}'} \in R$, by construction of $S$. If the former, then there is an overlap between $R$ and $R'$. If the latter, then since $R'$ is CUI, $\boldsymbol{\mathcal{O}'} \in R'$, and so again there is an overlap between $R$ and $R'$.

$G' \geq G$: If $R' = S$, then since $S \supseteq R$, $R' \supseteq R$. This means agent $k'$ explores for the

[2]Such an $n'$ is guaranteed to exist because $\lim_{n \to \infty} c(n) > C$, $\lim_{n \to -\infty} c(n) = 0$, and continuity of cost as points are continuously added.

entire region that agent $k$ does. By Lemma 3.4.4, then, he also gets at least as large of a gain for that region, and so $G' \geq G$ as desired.

If $R' \neq S$, then by construction of $R'$, $C' = C$. Consider the regions $R' \cap R$, $R' - R$, and $R - R'$. Agent $k'$ explores $\mathcal{O}_i$ for $R' \cap R$ and $R' - R$, while agent $k$ explores $\mathcal{O}_i$ for $R' \cap R$ and $R - R'$. I show that agent $k'$ gets no less gain from $R' \cap R$ than agent $k$ does for the same region, and also that agent $k'$ gets no less gain from $R' - R$ than agent $k$ does from $R - R'$. Together, these show $G' \geq G$.

First, consider $R' \cap R$: by Lemma 3.4.4, agent $k'$ gets no less gain from this region than agent $k$ does, because the options explored before agent $k'$ are a subset of those explored before agent $k$. Additionally note that, also by Lemma 3.4.4, the cost over this region is less for agent $k'$ than agent $k$.

Since the cost is less for agent $k'$ to explore $R' \cap R$, but the total costs for agent $k$ exploring $R$ and agent $k'$ exploring $R'$ are equal ($C = C'$), it must be that the cost for agent $k'$ to explore $R' - R$ is greater than the cost for agent $k$ to explore $R - R'$.

Consider $\boldsymbol{\mathcal{O}} \in R \cap R'$, and $\boldsymbol{\mathcal{O}'}$ that differs from $\boldsymbol{\mathcal{O}}$ only on the coordinate $\mathcal{O}_i$. Since $R$ is CUI for agent $k$, $\boldsymbol{\mathcal{O}'} \in R$. Similarly, since $R'$ is CUI for agent $k'$, $\boldsymbol{\mathcal{O}'} \in R'$. So, $\boldsymbol{\mathcal{O}'} \in R \cap R'$. It follows that for every point $\boldsymbol{\mathcal{O}} = (\mathcal{O}_1, \ldots \mathcal{O}_i, \ldots \mathcal{O}_N)$ in $R \cap R'$, the entire line $\ell = \{(\mathcal{O}_1, \ldots \mathcal{O}'_i, \ldots \mathcal{O}_N) : \mathcal{O}'_i \in supp(\mathcal{O}_i)\}$ is contained in $R \cap R'$.

So, consider any point $\boldsymbol{\mathcal{O}} \in R - R'$, and a point $\boldsymbol{\mathcal{O}'}$ which differs from $\boldsymbol{\mathcal{O}}$ only on the coordinate $\mathcal{O}_i$. Since $\boldsymbol{\mathcal{O}} \in R$ and $R$ is CUI for agent $k$, $\boldsymbol{\mathcal{O}'} \in R$. If $\boldsymbol{\mathcal{O}'} \in R \cap R'$, then by above the entire corresponding line is in $R \cap R'$, contradicting that $\boldsymbol{\mathcal{O}} \in R - R'$. So, $\boldsymbol{\mathcal{O}'} \in R - R'$. It follows that for every point $\boldsymbol{\mathcal{O}} = (\mathcal{O}_1, \ldots \mathcal{O}_i, \ldots \mathcal{O}_N)$ in $R - R'$, the entire line $\ell = \{(\mathcal{O}_1, \ldots \mathcal{O}'_i, \ldots \mathcal{O}_N) : \mathcal{O}'_i \in supp(\mathcal{O}_i)\}$ is contained in $R - R'$.

By the same reasoning, for every point $\boldsymbol{\mathcal{O}} = (\mathcal{O}_1, \ldots \mathcal{O}_i, \ldots \mathcal{O}_N)$ in $R' - R$, the entire line $\ell = \{(\mathcal{O}_1, \ldots \mathcal{O}'_i, \ldots \mathcal{O}_N) : \mathcal{O}'_i \in supp(\mathcal{O}_i)\}$ is contained in $R' - R$.

So, partition $R' - R$ and $R - R'$ into the lines (varying only on $\mathcal{O}_i$) that make them up.

I claim that the every line in $R' - R$ has higher gain-to-cost ratio for agent $k'$ than every line in $R - R'$ does for agent $k$. Combined with the observation that the cost for $R' - R$ is higher for agent $k'$ than $R - R'$ is for agent $k$, this implies that the gain from $R' - R$ for agent $k'$ is higher than the gain from $R - R'$ for agent $k$, completing the proof.

So, consider $\boldsymbol{\mathcal{O}} \in R - R'$. Since $\boldsymbol{\mathcal{O}} \in R$, $\boldsymbol{\mathcal{O}} \in S$. $\boldsymbol{\mathcal{O}} \in S$ and $\boldsymbol{\mathcal{O}} \notin R'$, means that $\max_{j \in A'}(\mathcal{O}_j)$ is higher than the maximum explored value for every point in $R'$, and so every point in $R' - R$, by the construction of $R'$. $\max_{j \in A}(\mathcal{O}_j) \geq \max_{j \in A'}(\mathcal{O}_j)$ because $A' \subseteq A$, meaning that the highest value in $\boldsymbol{\mathcal{O}}$ explored before agent $k$ is higher than the highest explored value for agent $k'$ for every point in $R' - R$. By Lemma 3.4.3, then, the line containing $\boldsymbol{\mathcal{O}}$ has lower gain-to-cost ratio for agent $k$ than every line in $R' - R$ does for agent $k'$. □

The preceding theorem is complicated, but constructive. So, it may be helpful to see an example of the procedure it describes using the distributions from the continued example.

Suppose under some policy $\pi$, before agent $k$, $\mathcal{O}_1$ and $\mathcal{O}_3$ have been explored for the region $R = [-2, 6] \times [-3, 5] \times [-2, 2]$. In another policy $\pi'$, before agent $k'$, only $\mathcal{O}_1$ has been explored for $R$. Lemma 3.4.5 shows that if agent $k$ can explore $\mathcal{O}_2$ for $R$ with cost $C$ and gain $G$, then there must be some other region $R'$ such that agent $k'$ can explore $\mathcal{O}_2$ over $R'$ for cost $C' \leq C$ and gain $G' \geq G$. Let's find $R'$:

The transitive closure of $R$ under indistinguishability by agent $k'$ is $S = [-2, 6] \times [-3, 5] \times [-2, 6]$, because agent $k'$ has not yet explored $\mathcal{O}_3$. Then, calculating $C$ gives:

$$C = P(\boldsymbol{\mathcal{O}} \in R)(E(\max(\mathcal{O}_1, \mathcal{O}_3) | \boldsymbol{\mathcal{O}} \in R) - \mu_2) = \frac{2}{3}(\frac{1}{2}\frac{2}{3} + \frac{1}{2}4 - 1) = \frac{8}{9}$$

The cost for agent $k'$ to explore $\mathcal{O}_2$ over $S$ is 1, greater than $C = \frac{8}{9}$. So, to construct $R'$, one needs to calculate the value $n'$ such that exploring all points in $S$ with $\mathcal{O}_1 \leq n'$

(since $\mathcal{O}_1$ is the only explored option for agent $k'$) will yield a region with cost $C$:

$$\frac{8}{9} = P(\boldsymbol{\mathcal{O}} \in R' \wedge \mathcal{O}_1 \leq n')(E(\mathcal{O}_1 | \mathcal{O}_1 \in R' \wedge \mathcal{O}_1 \leq n') - \mu_2)$$

$$\frac{8}{9} = \frac{n'+2}{8}\left(\frac{n'-2}{2} - 1\right)$$

$$n' = 1 + \sqrt{209}/3$$

So, $R' = [-2, 1+\sqrt{209}/3] \times [-3, 5] \times [-2, 6]$. Continuing with the argument in Theorem 3.4.5, one can demonstrate that $R'$ has a larger gain for agent $k'$ than $R$ does for agent $k$. Exploring the overlapping region, $R \cap R' = [-2, 1+\sqrt{209}/3] \times [-3, 5] \times [-2, 2]$, gives more gain at less cost for agent $k'$ than it does for agent $k$, by Lemma 3.4.4. Then, compare $R' - R$ to $R - R'$:

$$R' - R = [-2, 1+\sqrt{209}/3] \times [-3, 5] \times [2, 6]$$

$$R - R' = [1+\sqrt{209}/3, 6] \times [-3, 5] \times [-2, 2]$$

The maximum explored option for agent $k'$ in $R' - R$ is never more than $1 + \sqrt{209}/3$. Conversely, the maximum explored option for agent $k$ in $R - R'$ is never less than $1 + \sqrt{209}/3$. So, by Lemma 3.4.3, every line in $R' - R$ has higher gain-to-cost ratio for agent $k'$ than every line in $R - R'$ does for agent $k$. Since the cost of $R' - R$ must be higher for agent $k'$ than $R - R'$ is for agent $k$, it follows that the gain from $R' - R$ is higher for agent $k'$ than $R - R'$ is for agent $k$.

Lastly, the surplus/cost abstraction can easily show that the exploration possibility condition from section 2 is sufficient for all options to eventually be explored.

**Theorem 3.4.6.** *If it is possible to explore an option other than $\mathcal{O}_1$ for any realization $\boldsymbol{\mathcal{O}}$, then it is possible to always explore all options over supp($\boldsymbol{\mathcal{O}}$) within a fixed finite number of agents.*

*Proof.* By Lemma 2.1.1, if it is possible to ever explore an option other than $\mathcal{O}_1$, then the exploration possibility condition must hold. Namely, $P(\mathcal{O}_1 < \mu_2) > 0$. So, the region defined by $\mathcal{O}_1 < \mu_2$ has non-zero mass, and by Theorem 3.4.1 has negative cost. So, agent 2 can explore $\mathcal{O}_2$ over this region for a cost of $-c$, as well as $\mathcal{O}_2$ over another region with cost $c$, while still being IC. Note that the total cost for exploring $\mathcal{O}_2$ for all remaining realizations does not change before and after agent 2 explores, since the net cost spent by agent 2 is 0.

Also, $P(\mathcal{O}_1 < \mathcal{O}_2 | \mathcal{O}_1 < \mu_2) > 0$, and so this negative cost region must have positive surplus. So, $S_3 > 0$. Since surplus can never decrease from exploration, $S_i \geq S_3 > 0$ for all $i \geq 3$.

It is then possible to just explore each option, in order. By above, the total cost for exploring $\mathcal{O}_2$ over remaining realizations of $\boldsymbol{\mathcal{O}}$ after agent 2 is the same as exploring $\mathcal{O}_2$ over $supp(\boldsymbol{\mathcal{O}})$ before agent 2. By Theorem 3.4.1, since only $\mathcal{O}_1$ is previously explored, this is $\mu_1 - \mu_2$. It will then require no more than $\left\lceil \frac{\mu_1 - \mu_2}{S_3} \right\rceil$ agents after agent 2 to explore $\mathcal{O}_2$ for all of these realizations, since each agent can spend a surplus of at least $S_3$, and once a total cost of $\mu_1 - \mu_2$ for exploring $\mathcal{O}_2$ is paid across all agents, $\mathcal{O}_2$ will have been explored for all realizations.

More generally, by Theorem 3.4.1, to explore $\mathcal{O}_i$ over $supp(\boldsymbol{\mathcal{O}})$ after options $\mathcal{O}_1 \dots \mathcal{O}_{i-1}$ have been explored has cost $E(\max_{j \in [i-1]}(\mathcal{O}_j)) - \mu_i$. So, after options $\mathcal{O}_1 \dots \mathcal{O}_{i-1}$ have can be explored, it requires no more than $\left\lceil \frac{E(\max_{j \in [i-1]}(\mathcal{O}_j)) - \mu_i}{S_3} \right\rceil$ agents to explore $\mathcal{O}_i$.

So, all options can be explored with no more than $2 + \sum_{i \in [2,N]} \left\lceil \frac{E(\max_{j \in [i-1]}(\mathcal{O}_j)) - \mu_i}{S_3} \right\rceil$ agents.

□

This result is particularly surprising, since it only relies on $P(\mathcal{O}_1 < \mu_2) > 0$. It still

holds if for $i > 2$, $P(\mathcal{O}_1 < \mu_i) = 0$ (it is never negative cost for an agent to explore $\mathcal{O}_i$). In fact, it even holds if for $i > 2$, $P(\mathcal{O}_1 < \mathcal{O}_i) = 0$ (there is never positive gain from exploring $\mathcal{O}_i$)! If any option can ever be explored, then for all realizations, all options, no matter how bad, can always be explored.[3]

[3] A similar proof is carried out in Kremer et al. in the special case of 2 options (and so without the surprising implications detailed here), and without surplus/cost abstraction hiding the details of various integrals and inequalities. As a result, however, there are several mistakes in their proof. While under Theorem 3.4.6 there is a bound of $2 + \left\lceil \frac{\mu_1 - \mu_2}{S_3} \right\rceil$ before both options are explored, in Kremer et al. the bound is given as simply $\frac{\mu_1 - \mu_2}{S_3}$. The missing 2 is because the authors forget that the $S_3$ surplus is only available starting with agent 3, and the missing $\lceil \rceil$ is because the authors forget that, even if an agent does not need to spend their full surplus, that agent will still need to explore (e.g. if $\mu_1 - \mu_2 = 3$ and $S_3 = 2$, it may require 2 agents before the total cost of 3 is spent).

While these mistakes seem clear in this context, without the surplus/cost abstraction they correspond to subtle mistakes in manipulating complex inequalities between sums of integrals. This provides some evidence that, aside from the technical results it enables, the surplus/cost abstraction is useful for providing semantic context for the important integrals in this setting, and so making it easier to reason precisely and correctly.

# Chapter 4

In this chapter, I apply the surplus/cost abstraction to the setting in which the center's objective is to explore all options within the fewest number of agents in the worst-case. So, a policy is optimal if it is IC, and there is no other IC policy that explores all options within fewer agents in the worst case. I succeed in characterizing the optimal policy in this setting under the conditions that the options are orderable, and that in an optimal policy the options are explored in the corresponding order. I first define orderability, and find an optimal policy under these conditions. Then I analyze in what circumstances these conditions hold, and present a natural example of a case in which they do.

## 4.1 Introducing Orderability

Recall that the values $\mathcal{O}_1 \dots \mathcal{O}_N$ were numbered such that $\mu_1 > \mu_2 \geq \cdots \geq \mu_N$. Values $\mathcal{O}_2 \dots \mathcal{O}_N$ are *orderable* if they additionally satisfy for all $m \in \mathbb{R}$:

$$E(\max(\mathcal{O}_2, m)) \geq \cdots \geq E(\max(\mathcal{O}_N, m)) \qquad \text{(orderability condition)}$$

If the values are orderable, then $\mathcal{O}_i$ is said to precede $\mathcal{O}_j$, written $\mathcal{O}_i \prec \mathcal{O}_j$, if $i < j$. Note that if two different option's values have the same distribution, then there might be more than one valid way the options could have been ordered, just as there could have

been more than one way to index the options initially. Nevertheless, a single ordering is arbitrarily chosen by breaking ties based on the index of the option. $\prec$ is therefore a total order over the options.

Intuitively, orderability can be thought of as a strong condition ensuring that exploring earlier options is 'better' than exploring later ones. Just $\mu_i > \mu_j$, is not strong enough of a condition to ensure that there is greater gain from exploring $\mathcal{O}_i$ than there is from exploring $\mathcal{O}_j$ over some region. For example, consider a region over which there is a previously explored option with value at least $V$. Then if $P(\mathcal{O}_i > V) = 0$, exploring $\mathcal{O}_i$ over that region will have 0 gain. It is possible that another option $\mathcal{O}_j$ will have positive gain over that region ($P(\mathcal{O}_j > V) > 0$), even though $\mu_j < \mu_i$. $\mathcal{O}_i \prec \mathcal{O}_j$, however, is strong enough to rule out these situations.

**Theorem 4.1.1.** *If $R$ is CUI and $\mathcal{O}_i \prec \mathcal{O}_j$, then the gain from exploring $\mathcal{O}_i$ over $R$ is no less than that of exploring $\mathcal{O}_j$. Additionally, the cost for exploring $\mathcal{O}_i$ is no greater, and so the gain-to-cost ratio is no less.*

*Proof.* $R$ is CUI and neither $\mathcal{O}_i$ nor $\mathcal{O}_j$ has been previously explored over it. So, for any point $\boldsymbol{\mathcal{O}} = (\mathcal{O}_1, \mathcal{O}_2, \ldots \mathcal{O}_N) \in R$, $R$ also contains the entire plane defined by $p = \{(\mathcal{O}_1, \ldots \mathcal{O}'_i, \ldots \mathcal{O}'_j, \ldots \mathcal{O}_N) : \mathcal{O}'_i \in \mathit{supp}(\mathcal{O}_i), \mathcal{O}'_j \in \mathit{supp}(\mathcal{O}_j)\}$. So, $R$ can be partitioned into planes of this form.

Fix one such plane $p$, and Let $V$ be the maximum of the explored options over $p$. Then by the same reasoning as in Theorem 3.4.3, the gain from exploring $\mathcal{O}_i$ over $p$ is $f(p)(E(\max(\mathcal{O}_i, V)) - V)$, and the gain from exploring $\mathcal{O}_j$ is $f(p)(E(\max(\mathcal{O}_j, V)) - V)$. Since $\mathcal{O}_i \prec \mathcal{O}_j$, the first quantity is greater, so the gain from exploring $\mathcal{O}_i$ is greater. $\mathcal{O}_i \prec \mathcal{O}_j$, so $\mu_i \geq \mu_j$, and by Corollary 3.4.2 the cost for exploring $\mathcal{O}_i$ is no greater than the cost from exploring $\mathcal{O}_j$.

Since the marginal gain from every plane is no less for $\mathcal{O}_i$, and the marginal cost from

every plane is no greater for $\mathcal{O}_i$, the result follows. □

If the options are orderable, then they are said to be *explored in order* if whenever $\mathcal{O}_i$ is explored for $\boldsymbol{\mathcal{O}}$, if $\mathcal{O}_j \prec \mathcal{O}_i$, then $\mathcal{O}_j$ has already been explored for $\boldsymbol{\mathcal{O}}$. This can be stated equivalently using the notion of a layer—exploration over a region is said to occur in *layer* $n$ if, after exploring, there are $n$ explored options for every point in that region. So, for example, the initial exploration of $\mathcal{O}_1$ by the first agent occurs in layer 1, since afterwards there is exactly 1 options explored at every point. Agent 2 will then explore some regions for which $\mathcal{O}_1$ is already explored, and so will explore layer 2 regions. Agent 3 might explore some regions where both $\mathcal{O}_1$ and $\mathcal{O}_2$ have been previously explored (layer 3), or regions over which only $\mathcal{O}_1$ has been explored (layer 2).

A region $R$ is said to be in layer $n$ if currently there are $n-1$ options explored over $R$. In other words, saying that exploration occurs in layer $n$ is equivalent to saying that exploration occurred over a region in layer $n$. With $N$ options, the possible layers are 1 to $N$. Using this terminology, options are explored in order iff for all $i$, $\mathcal{O}_i$ is explored only in layer $i$ ($\mathcal{O}_i$ is the $i$th option to be explored for any realization).

I claim that if the options are orderable, and there is an optimal policy that explores options in order, the optimal policy is greedy, having each agent explore regions with the highest possible gain-to-cost ratio.

## 4.2 Greediness

Assume that the options are orderable. A policy is *greedy* if it always explores regions with the highest gain-to-cost ratio that are available to each agent—when an agent explores $\mathcal{O}_i$ over CUI region $R$, then there is no CUI region $R'$ and option $\mathcal{O}_i'$ which has a higher gain-to-cost ratio, but which the agent does not explore—and every agent spends as much

of the surplus as possible.[1] Observe that by theorem 4.1.1, a greedy policy explores the options in order, since for any region it explores, there is higher gain-to-cost by exploring an earlier option over a later one.

To build an intuition for what a greedy policy looks like, it is helpful to characterize the way that a greedy policy determines what to explore within a given layer, and between different layers. Specifically, a greedy policy is *intra-layer greedy* (ILG) and *between-layer greedy* (BLG).[2] As the names suggest, these respectively correspond to requirements that agents explore higher gain-to-cost regions within each layer first, and that agents explore higher gain-to-cost regions across layers first.

Formally, a policy is ILG if, after every agent, the region explored within layer $i$ can be expressed as $R = \{\boldsymbol{\mathcal{O}}' : \boldsymbol{\mathcal{O}}' \in \mathit{supp}(\boldsymbol{\mathcal{O}}) \wedge \max_{j \in [i-1]}(\mathcal{O}'_j) \leq V\}$ for some $V$. Note that by Theorem 3.4.3, the gain-to-cost ratio is decreasing in $V$, the maximum previously explored value. Exploring when the maximum previously explored value is small will both increase the gain, and decrease the cost. So, within layer $i$, the greatest gain-to-cost for exploring $\mathcal{O}_i$ is where the maximum previously explored value, $V$, is the smallest. If within layer $i$ a policy explores in increasing values of $V$ (that is, small values of $V$ first), the resulting region will always be defined by a set of the form above. So, any policy which is greedy, and so explores regions in increasing order of gain-to-cost, will be ILG.

A policy is BLG if whenever an agent explores a region $R$ in layer $i$, there is no CUI region in layer $j \neq i$ with a higher gain-to-cost ratio that the agent does not explore. Where the ILG condition ensures that regions within a given layer are explored greedily, the BLG condition ensures that, between different layers, regions with higher gain-to-cost ratio are explored earlier.

[1] And, so that the greedy policy is unambiguously defined, if there are multiple regions with 0-gain the greedy policy explores lower cost regions first. See Appendix B for an example where this occurs.

[2] While inter-layer greediness seems a more fitting counterpart, it unfortunately does not have a unique initialism.

Together, these conditions are enough to straightforwardly *implement* a greedy policy, meaning that the greedy policy can be effectively determined by the center. This is done at the end of the chapter.

Note that a greedy policy does not, between agents, necessarily explore regions in strictly increasing order of gain-to-cost. For example, agent 2 will explore some region with negative cost (where $\mathcal{O}_1 < \mu_2$), as well as some region with positive cost. Agent 3 might then also explore a region with negative cost (where $\max(\mathcal{O}_1, \mathcal{O}_2) < \mu_3$), and some region with positive cost. Both might have explored greedily, but agent 3 still explored a region with higher gain-to-cost ratio than a region explored by agent 2 (the negative cost region for agent 3, compared to the positive cost region for agent 2). Observe, though, that in these cases the reason why agent 2 does not explore for this negative cost region is that the region was in a layer higher than it was possible to explore — agent 2 cannot explore in layer 3, since exploration must occur on layer 2 first. If it were not 'blocked' in this way, the greedy policy would have explored the higher gain-to-cost region first. So, it is possible to make a more restricted claim about the order of exploration between agents: in a greedy policy, when a region $R$ in layer $i$ with higher gain-to-cost is explored after another region in a different layer with lower gain-to-cost, $R$ is explored in layer $i$ by the first agent who could do so while still exploring in order. This observation is sufficient to show that the greedy policy is optimal.

**Theorem 4.2.1.** *If the options are orderable, and an optimal policy explores in order, then greedy is an optimal policy.*

*Proof.* Fix an optimal policy that explores in order. Let the surplus for each agent $i$ under this policy be $S_i$. Let $S_i'$ be the surplus for agent $i$ under the optimal policy. I claim that for all $i$, $S_i' \geq S_i$.

Let the surplus after all options have been explored over $supp(\boldsymbol{\mathcal{O}})$ be $S^*$. This means

that if $S_k = S^*$, and so all options are explored in the optimal policy before agent $k$, then $S'_k = S^*$, and so all options are explored before agent $k$ in the greedy policy. It follows that the greedy policy does not take more agents in the worst-case to explore all options, and so that the greedy is an optimal policy.

First, both the greedy and the optimal policy have agent 1 explore $\mathcal{O}_1$, since all IC policies do so. This means that, $S_2 = 0 = S'_2$.

For the sake of contradiction, consider the first agent $i$ for which $S'_i < S_i$. Since $i$ is selected to be the first agent where this condition holds, it must be that the total cost spent across all agents $1 \dots i-1$ in the greedy policy is no smaller than the total cost for those agents in the optimal policy. But, since $S'_i < S_i$, the gain across those agents must have been greater in the optimal policy. It follows that, before agent $i$, the optimal policy has higher gain-to-cost aggregated over all regions/options it has explored than greedy does.

Since both greedy and optimal explore in order, whenever either explores $\mathcal{O}_j$ over $R$, both have the same gain and cost for doing so (both have explored exactly $\mathcal{O}_1 \dots \mathcal{O}_{j-1}$ previously over that region). So, over the (region, option) pairs explored by both greedy and optimal, they both received the gain at the same cost. It follows that in order for optimal to have higher gain-to-cost than greedy, there must be some pair $(R, \mathcal{O}_j)$ explored in optimal but not greedy before agent $i$ which has higher gain-to-cost than some region $(R', \mathcal{O}'_j)$ explored in greedy but not in optimal before agent $i$. Otherwise, it would be impossible for the gain-to-cost to be higher in optimal than greedy before agent $i$.

But, since greedy eventually explores $(R, \mathcal{O}_j)$, and this has higher gain-to-cost than $(R', \mathcal{O}'_j)$ which is explored earlier, it follows that $(R, \mathcal{O}_j)$ is explored by the first agent who can do so while still exploring options in order. Thus, that the optimal policy explores this earlier contradicts the assumption that the optimal policy explores in order.

So, $S'_i \geq S_i$ for all $i$, completing the proof.

□

## 4.3 Investigating Orderability

Since the result above relies on the orderability of the options, it makes sense to consider how natural of an assumption it is. So, I find general characterizations of pairs of random variables $X$ and $Y$ that satisfy $(\forall m)E(\max(X,m)) \geq E(\max(Y,m))$. For this section, say that $X \prec Y$ iff $(\forall m)E(\max(X,m)) \geq E(\max(Y,m))$. By chaining together pairs of this form, it is possible to construct natural sets of distributions which are orderable. That is, if $X \prec Y$ and $Y \prec Z$, then by transitivity $\{X, Y, Z\}$ is also orderable.

First, say that $X$ is a *translation up* of $Y$ if $X$ is distributed the same as $Y + c$ for some constant $c > 0$.

**Theorem 4.3.1.** *If $X$ is a translation up of $Y$, then $X \prec Y$.*

*Proof.* Since $c > 0$, for all $m$:

$$E(\max(X,m)) = E(\max(Y+c,m)) \geq E(\max(Y,m))$$

□

Next, let $\mu_Y = E(Y)$, and let $Y$ be a symmetric distribution. Say that $X$ is a *stretch* of $Y$ if $X$ is distributed the same as $a(Y - \mu_Y) + \mu_Y$ for $a > 1$. Intuitively, this corresponds to centering $Y$ at 0, multiplying by a constant to 'stretch' the PDF for $Y$, and then re-centering $Y$ at the correct mean.

**Theorem 4.3.2.** *If $X$ is a stretch of $Y$, then $X \prec Y$.*

*Proof.* First, observe that if for all $m$: $E(\max(X - \mu_Y, m)) \geq E(Y - \mu_Y, m)$, then $E(\max(X, m+\mu_Y)) - \mu_Y \geq E(Y, m+\mu_Y) - \mu_Y$, and so for all $m$, $E(\max(X,m)) \geq E(Y,m)$. So, let $Y' = Y - \mu_Y$ and $X' = X - \mu_Y$, and I show that $E(\max(X',m)) \geq E(Y',m)$. Ob-

serve that now, $X'$ and $Y'$ are both symmetric and centered at 0, and $X'$ is distributed the same as $aY'$ for $a > 1$.

Fix any $m \geq 0$. Then:

$$E(\max(X', m))$$

$$= E(\max(X', m)|X' \leq am)P(X' \leq am) + E(X'|X' > am)P(X' > am)$$

$$= E(\max(aY', m)|Y' \leq m)P(Y' \leq m) + E(aY'|Y' > m)P(Y' > m)$$

$$\geq mP(Y' \leq m) + E(Y'|Y' > m)P(Y' > m) = E(\max(Y', m))$$

Fix any $m < 0$. Then:

$$E(\max(X', m))$$

$$= E(X'|m \leq X' \leq -m)P(m \leq X' \leq -m) + mP(X' < m) + E(X'|X' > -m)P(X' > -m)$$

$$= mP(aY' < m) + E(aY'|aY' > -m)P(aY' > -m)$$

$$\geq mP(Y' < m) + E(Y'|Y' > -m)P(Y' > -m)$$

$$= E(Y'|m \leq Y' \leq -m)P(m \leq Y' \leq -m) + mP(Y' < m) + E(Y'|Y' > -m)P(Y' > -m)$$

$$= E(\max(Y', m))$$

Where the first inequality comes from observing that $P(aY' < m) \geq P(Y' < m)$ when $m < 0$, and that:

$$E(aY'|aY' > -m)P(aY' > -m) = \int_{\frac{-m}{a}}^{\infty} f_Y(y)(ay)dy$$

$$\geq \int_{-m}^{\infty} f_Y(y)(ay)dy \geq \int_{-m}^{\infty} f_Y(y)(y)dy = E(Y'|Y' > -m)P(Y' > -m)$$

So, for all $m$, $E(\max(X', m)) \geq E(Y', m)$ as desired. □

Using just the notions of translating and stretching, it is possible to create many natural families of orderable distributions. For example, suppose the options were distributed as follows: $\mathcal{O}_1 \sim Unif(1, 6)$, $\mathcal{O}_2 \sim Unif(-2, 6)$, $\mathcal{O}_3 \sim Unif(-3, 5)$, and $\mathcal{O}_4 \sim Unif(-2, 4)$. $\mathcal{O}_2 \prec \mathcal{O}_3$, since $\mathcal{O}_2$ is just $\mathcal{O}_3$ translated up. And, $\mathcal{O}_3 \prec \mathcal{O}_4$, since $\mathcal{O}_3$ is just a stretch of $\mathcal{O}_4$. So, this set of distributions is orderable.

In addition to uniform distributions, normal distributions also fall nicely into this pattern. If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then if $\mu_1 \geq \mu_2$ and $\sigma_1 \geq \sigma_2$, then $X_1 \prec X_2$. To see this, note that increasing the mean is just a translation, and increasing the variance is just a stretch. So, increasing both is just a translation followed by a stretch, which by transitivity of $\prec$ over the intermediary state ensures that $X_1 \prec X_2$.

However, this result is limited by the fact that orderability was only one of the requirements from the greediness result — it also required that all options be explored in order. None of above responds to the question of under what circumstances an optimal policy will explore options in order. This is an important area of further research.

Fortunately, there is a single case in which it can be easily deduced that options are orderable, and they are explored in order in an optimal policy. Namely, if $\mathcal{O}_2 \ldots \mathcal{O}_N$ all have the same distribution. When this is the case, the options trivially are orderable. And, by symmetry, the gain, cost, and surplus is identical no matter which option is explored when. So, any optimal policy can be modified so that the options are explored in order, without changing the surplus after any agent, and so resulting in another optimal policy. Therefore, when $\mathcal{O}_2 \ldots \mathcal{O}_N$ are all distributed the same way, the greedy policy is optimal. This is not a particularly unnatural setting — it represents a circumstance in which there is

a single option which has some positive information known about it (a higher expectation), and then a set of remaining options for which there is no information, and so symmetrically they have the same prior distribution.

## 4.4 Implementing an Optimal Greedy Policy

Since the greedy policy is optimal when $\mathcal{O}_2 \dots \mathcal{O}_N$ are all IID, it might be enlightening to begin calculating a concrete example of an optimal policy. In particular, it will show how ILG and BLG can be used to find the regions with the highest gain-to-cost. So, let $\mathcal{O}_1 \sim Unif(0, 2)$, and $\mathcal{O}_2, \mathcal{O}_3 \sim Unif(-1, 2)$.

Agent 1: agent 1 always explores $\mathcal{O}_1$ over $supp(\boldsymbol{\mathcal{O}})$. So, agent 1 explores $\mathcal{O}_1$ over $[0, 2] \times [-1, 2] \times [-1, 2]$.

Agent 2: $S_2 = 0$, and since exploring in order, $\mathcal{O}_2$ is explored for layer 2. So, agent 2 explores $\mathcal{O}_2$ as much as possible for a total cost of 0. By ILG, this is exploring for $\mathcal{O}_1 \leq V$ for some $V$. By Theorem 3.4.1, exploring for $0 \leq \mathcal{O}_1 \leq 1$ has cost 0. So, agent 2 explores $\mathcal{O}_2$ over $[0, 1] \times [-1, 2] \times [-1, 2]$.

Agent 3: By Theorem 3.4.3, since $\mathcal{O}_2$ and $\mathcal{O}_3$ have the same distribution, the highest gain-to-cost across both layers is for the lowest values of $V$, the maximum previously explored option. So, to satisfy BLG, no region in layer 2 can be explored before all layer 3 regions are explored for $V \leq 1$. So, agent 3 explores $\mathcal{O}_3$ for at least $[0, 1] \times [-1, 1] \times [-1, 2]$. This has a base cost of:

$$P(\mathcal{O}_1 \leq 1)E(\mathcal{O}_1 - \mu_3 | \mathcal{O}_1 \leq 1) = \frac{1}{2}(\frac{1}{2} - \frac{1}{2}) = 0$$

So, instead of finding the total cost, I just calculate the remaining surplus excluding this

region. The surplus over the region $R = [0,1] \times [1,2] \times [-1,2]$ is:

$$P(\boldsymbol{\mathcal{O}} \in R)E(\max(\mathcal{O}_1, \mathcal{O}_2) - \mathcal{O}_1 | \boldsymbol{\mathcal{O}} \in R) = \left(\frac{1}{2}\frac{1}{3}\right)\left(\frac{3}{2} - \frac{1}{2}\right) = \frac{1}{6}$$

By BLG, this surplus will be spent exploring in both layers 2 and 3 for all regions with maximum explored values less than $V$ for some $V$. And by ILG and Theorem 3.4.3, this will be the same $V$ for both. To calculate this $V$, I find total cost across both layers as a function of $V$, and set it equal to $\frac{1}{6}$, using Theorem 3.4.1:

$$\frac{1}{6} = C(V) = cost\_layer2(V) + cost\_layer3(V)$$

$$= P(1 \leq \mathcal{O}_1 \leq V)(E(\mathcal{O}_1 | 1 \leq \mathcal{O}_1 \leq V) - \mu_2) + P(\mathcal{O}_1 \leq 1 \wedge 1 \leq \mathcal{O}_2 \leq V)(E(\mathcal{O}_2 | 1 \leq \mathcal{O}_2 \leq V) - \mu_3)$$

$$= \frac{V-1}{2}\left(\frac{V+1}{2} - \frac{1}{2}\right) + \left(\frac{1}{2}\frac{V-1}{3}\right)\left(\frac{V+1}{2} - \frac{1}{2}\right)$$

$$= \frac{V^2 - V}{3}$$

So, $V = \frac{1+\sqrt{3}}{2}$. This means that, (including the exploration for $\mathcal{O}_3$ detailed first), agent 3 explores $\mathcal{O}_2$ for $[1, \frac{1+\sqrt{3}}{2}] \times [-1,2] \times [-1,2]$, and $\mathcal{O}_3$ for $[0,1] \times [-1, \frac{1+\sqrt{3}}{2}] \times [-1,2]$.

Agent 4, etc.: Calculating subsequent agents is much the same as agent 3. First, by BLG, agent 4 will definitely explore $\mathcal{O}_3$ for $[1, \frac{1+\sqrt{3}}{2}] \times [-1, \frac{1+\sqrt{3}}{2}] \times [-1,2]$. The cost for this region is calculated. Then, the new $V$ must be calculated, by finding the cost as a function of $V$ and setting it equal to the remaining surplus after $[1, \frac{1+\sqrt{3}}{2}] \times [-1, \frac{1+\sqrt{3}}{2}] \times [-1,2]$ has been explored. Agent 4 then explores explores $\mathcal{O}_2$ for $[\frac{1+\sqrt{3}}{2}, V] \times [-1,2] \times [-1,2]$, and $\mathcal{O}_3$ for $[1, \frac{1+\sqrt{3}}{2}] \times [\frac{1+\sqrt{3}}{2}, V] \times [-1,2]$.

Another example of an implemented optimal policy is included in Appendix B.

# Chapter 5

## 5.1 Setting

In this chapter, I apply the surplus/cost abstraction to the setting in which the center's objective is to explore a particular option within the fewest number of agents in expectation. Rather than attempting to find the optimal policy, I ask whether the center can improve upon the optimal policy by introducing a new option, referred to as the dummy. I find that under mild conditions introducing a dummy does improve upon the optimal policy. It simplifies the details of these conditions greatly to fix the number of options, so I assume that there are only three options: the target option, one other option with a higher expectation than the target option, and (possibly) the dummy. However, it will be clear from the technique used that it can be generalized to larger cases.

Formally, $\mathcal{O}_t$ (the target option), $\mathcal{O}_1$ (the better option), and $\mathcal{O}_d$ (the dummy option) are the options. As in the original setting, I assume that $\mu_1 > \mu_t$ (so that the optimal policy is not recommending agent 1 explore the target) and that $P(\mathcal{O}_1 \leq \mu_t) > 0$, ensuring that it is possible to explore the target even without the dummy. However, unlike in the earlier setting where all options are always present, in this setting the center can decide as part of the policy whether to allow $\mathcal{O}_d$ to be chosen by agents. This might correspond to the center broadening its array of choices (e.g. Yelp adding new restaurants to their website),

or even bringing a new option to market (e.g. Netflix commissioning a new original TV series). Since these actions tend to be costly, it is important to investigate when these costs can be offset by the added value of exploring unknown options. The question is whether $\mathcal{O}_d$ is introduced in the optimal policy.

When $\mathcal{O}_d$ is not present, the IC condition is the same as in the original case. When $\mathcal{O}_d$ is added, though, there is the possibility that the best option in expectation, and so the option that agents would choose if they didn't follow the recommendation policy, is now $\mathcal{O}_d$. In this case the IC condition (and the cost/surplus equations) would need to be modified so that all references to $\mathcal{O}_1$ are instead to $\mathcal{O}_d$. This will not present any difficulty, though, since the result in the following section will assume that $\mu_d < \mu_1$.

## 5.2 Dummy as Insurance

The intuition behind why the dummy option can easily be shown to improve upon the optimal policy using just $\mathcal{O}_1$ and $\mathcal{O}_t$, is to observe that the requirements for there eventually being a region for which it is negative cost to explore $\mathcal{O}_d$ are very limited. So long as there is some probability that $\mu_d$ is greater than $\mathcal{O}_1$ and $\mathcal{O}_t$, exploring $\mathcal{O}_d$ over the region where $\mathcal{O}_1 < \mu_d$ and $\mathcal{O}_t < \mu_d$ will have negative cost, and positive gain. This negative cost and positive gain translates into being able to have the corresponding agents explore $\mathcal{O}_t$ over more regions. I demonstrate this formally, including several caveats to the above intuition, below.

**Theorem 5.2.1.** *If $\mu_d < \mu_1$, $P(\mathcal{O}_1 < \mu_d \wedge \mathcal{O}_t < \mu_d) > 0$, and $\mathcal{O}_t$ is not explored over supp($\boldsymbol{\mathcal{O}}$) before agent 4, then any optimal policy will include introducing $\mathcal{O}_d$.*

*Proof.* Suppose there were some optimal policy $\pi$ that didn't introduce $\mathcal{O}_d$. I show that this policy can be improved by introducing $\mathcal{O}_d$, contradicting that this is an optimal policy.

First, observe that since $\mu_d < \mu_1$, it would be IC to introduce $\mathcal{O}_d$ and make no changes to what recommendations were made to each agent. So, it is sufficient to find a single modification that improves the policy, and then leave the rest of the policy unchanged.

Next, I confirm that all exploration over regions with negative cost is done by agent 2. Suppose $\mathcal{O}_t$ is explored for some region $R$ by an agent $i > 2$ in $\pi$, and that $R$ has negative cost. Observe that the surplus from exploring this region is greater than the negation of the cost:

$$-Cost|R(expl_i, e) = -\int_R f(\boldsymbol{\mathcal{O}})(\mathcal{O}_1 - \mathcal{O}_t)d\boldsymbol{\mathcal{O}} = \int_R f(\boldsymbol{\mathcal{O}})(\mathcal{O}_t - \mathcal{O}_1)d\boldsymbol{\mathcal{O}}$$

$$\leq \int_R f(\boldsymbol{\mathcal{O}})(\max(\mathcal{O}_1, \mathcal{O}_t) - \mathcal{O}_1)d\boldsymbol{\mathcal{O}} = Surplus|R(expl_{i+1})$$

So, suppose instead of agent $i$ exploring $R$, agent 2 explored $R$. This would be IC for agent 2, since it would only lower his total cost. All agents $2 < j < i$ would have greater surplus and no greater cost. Agent $i$ would have greater cost, but by above the increase in surplus is greater than this increase in cost, so it would be IC for agent $i$. And, all agents $j > i$ would have their IC condition unaffected. So, this is an IC policy. And, for some realizations $\mathcal{O}_t$ is explored earlier, so this is better for the objective, contradicting that $\pi$ is an optimal policy. Therefore, there can be no such region $R$, and all regions with negative cost are explored by agent 2.

So, agent 2 must explore $\mathcal{O}_t$ for the region defined by $\mathcal{O}_1 < \min(\mu_t, \mu_d)$, since it has negative cost. So the region defined by $\max(\mathcal{O}_1, \mathcal{O}_t) < \min(\mu_t, \mu_d)$ is CUI for agent 3. Agent 3 can explore $\mathcal{O}_d$ over this region, which has negative cost $-c$. Since $\mathcal{O}_t$ is not explored over $supp(\boldsymbol{\mathcal{O}})$ before agent 4, some agent $j > 3$ must explore $\mathcal{O}_t$ for some realizations of $\mathcal{O}_1$. Since now the IC condition for agent 3 is not tight, some region explored by agent $j$ can instead be explored by agent 3 (by making the region small enough, it will have cost

less than $c$). This improves upon $\pi$, since $\mathcal{O}_t$ is explored no later under any realization, and for some realizations it is explored by agent 3 instead of agent $j > 3$, contradicting that $\pi$ is optimal. Therefore, any optimal policy must introduce $\mathcal{O}_d$. □

One way to think about the construction in the proof for how to improve upon a policy by using a dummy, is that the dummy item is acting as insurance. Exploring over that negative cost region corresponds to assuring agents that, in the case that the realizations for $\mathcal{O}_1$ and $\mathcal{O}_t$ are especially poor, they can instead receive this new option. The additional utility corresponding to having insurance for these realizations allows the center to recommend that agents explore $\mathcal{O}_t$ earlier in other realizations.

One can also see how the idea behind this result could be generalized to cases when there are more options initially, or when more than one dummy item can be added: So long as a dummy isn't the new best option ex-ante, introducing it won't change the IC condition. And, so long as there is some probability that all other options will be less than the mean for a new dummy item, there will be a region for which the dummy will have negative cost (act as insurance). Exploring the dummy over that negative cost region can therefore be used to explore the target earlier for some realizations.

Lastly, suppose the center could choose the expectation of the dummy by translating its distribution. Raising the expectation lowers the cost and increases the surplus for exploring the dummy, and so allows for more exploration of $\mathcal{O}_t$ (i.e. it is better insurance). However, once the expectation surpasses $\mu_1$, it ceases to act as insurance, since now it is the best option ex-ante and will be explored first by agent 1. Increasing the expectation of the dummy even more past $\mu_1$ raises the cost and lowers the surplus for exploring both $\mathcal{O}_t$ and $\mathcal{O}_1$, without any corresponding benefit. This implies that if the center wants to introduce a new option in order to improve their ability to gain information, they should aim to introduce an option as good as the currently best option *but no better*.

# Chapter 6

## 6.1 Future Work

The mechanisms designed in this paper are highly unlikely to be used as-is by any recommender system in the real world. The IC condition implies that the optimal policy under one of these mechanisms should make an agent indifferent between using the recommendation service, and abandoning it completely. Competition from other recommenders, the necessity to provide a good user experience, and a slew of other un-modeled considerations make this solution impractical.

However, there is a strong possibility of effectively including mechanisms similar to these in otherwise statistically-oriented recommender systems. Some options may be positioned such that gaining information about their quality allows for a broad range of inferences about other options. For example, if two clusters of movies are known to be internally similar, but the relationship between the clusters is unknown, learning that a movie in one cluster is similar to another movie in the other cluster could transitively imply similarities between all movies in both clusters. If key unknown options of this form can be identified, a mechanism design approach could be used to elicit specific information in a targeted fashion.

There are many technical results presented here which can be improved or expanded

upon. Characterizing those distributions which are both orderable *and* explored in order in an optimal policy would greatly increase the applicability of the optimal policy in Chapter 4. Chapter 5 only demonstrates that introducing new options will be part of an optimal policy for exploring a target option, and does not begin to fully define the optimal policy. However, advancements in either of these directions are likely to be less impactful than designing a protocol for incorporating the existing results into a realistic machine learning style recommender.

# Appendix A

Suppose, as in Kremer et al., that agents could choose any option even after hearing the recommendation of the center. Then, in order for it to be in an agents interest to follow the recommendation, the following must hold for all agents $i$ and options $j$:

$$E(\mathcal{O}_j|rec_j^i) \geq \max_{k \neq j}(E(\mathcal{O}_k|rec_j^i)) \qquad \text{(ex-interim IC condition)}$$

Intuitively, this means that after being recommended $\mathcal{O}_j$, $\mathcal{O}_j$ must have the highest expected value. The use of 'ex-interim' refers to the fact that it is IC for the agent after having heard the recommendation, but before learning the realized value of the recommended option.

This is no-weaker of a requirement than the ex-ante IC condition; it is trivial to demonstrate that if a policy is ex-interim IC, then it is also ex-ante IC. In the special case where there are exactly 2 options, as in Kremer et al., the two conditions are actually equivalent.

**Theorem A.0.1.** *When there are exactly 2 options ($N = 2$), a recommendation policy is ex-ante IC iff it is ex-interim IC.*

*Proof.* Suppose a recommendation policy is ex-ante IC for some agent $i$. Then:

$$E(\mathcal{O}_1|rec_1^i)P(rec_1^i) + E(\mathcal{O}_2|rec_2^i)P(rec_2^i) \geq E(\mathcal{O}_1)$$

$$\iff E(\mathcal{O}_2|rec_2^i) \geq \frac{E(\mathcal{O}_1) - E(R_1|rec_1^i)P(rec_1^i)}{P(rec_2^i)}$$

$$\iff E(\mathcal{O}_2|rec_2^i) \geq \frac{E(\mathcal{O}_1|rec_2^i)P(rec_2)}{P(rec_2^i)}$$

$$\iff E(\mathcal{O}_2|rec_2^i) \geq E(\mathcal{O}_1|rec_2^i)$$

So, recommending $\mathcal{O}_2$ is ex-interim IC. It is also proven in Kremer et al. [2013] that if recommending $\mathcal{O}_2$ is ex-interim IC for an agent, then so is recommending $\mathcal{O}_1$. For completeness, I reproduce an equivalent proof below:

**Theorem A.0.2.** *If $N = 2$ and recommending $\mathcal{O}_2$ to agent $i$ is ex-interim IC, then recommending $\mathcal{O}_1$ to agent $i$ is IC.*

$$E(\mathcal{O}_2|rec_1^i)P(rec_1^i) + E(\mathcal{O}_2|rec_2^i)P(rec_2^i) = E(\mathcal{O}_2)$$

$$\leq E(\mathcal{O}_1) = E(\mathcal{O}_1|rec_1^i)P(rec_1^i) + E(\mathcal{O}_1|rec_2^i)P(rec_2^i)$$

$$\leq E(\mathcal{O}_1|rec_1^i)P(rec_1^i) + E(\mathcal{O}_2|rec_2^i)P(rec_2^i)$$

So, $E(\mathcal{O}_2|rec_1^i)P(rec_1^i) \leq E(\mathcal{O}_1|rec_1^i)P(rec_1^i)$, and $E(\mathcal{O}_2|rec_1^i) \leq E(\mathcal{O}_1|rec_1^i)$ as desired.

□

# Appendix B

Let $\mathcal{O}_1 \sim Unif(0,2)$, and $\mathcal{O}_2, \mathcal{O}_3 \sim Unif(0,1)$

Agent 1: agent 1 always explores $\mathcal{O}_1$ over $supp(\boldsymbol{\mathcal{O}})$. So, agent 1 explores $\mathcal{O}_1$ over $[0,2] \times [0,1] \times [0,1]$.

Agent 2: $S_2 = 0$, and since exploring in order, $\mathcal{O}_2$ is explored for layer 2. So, agent 2 explores $\mathcal{O}_2$ as much as possible for a total cost of 0. By ILG, this is exploring for $\mathcal{O}_1 \leq V$ for some $V$. By Theorem 3.4.1, exploring for $0 \leq \mathcal{O}_1 \leq 1$ has cost 0. So, agent 2 explores $\mathcal{O}_2$ over $[0,1] \times [0,1] \times [0,1]$.

Agent 3: By Theorem 3.4.3, the highest gain-to-cost is for the lowest values of $V$, the maximum previously explored option. So, to satisfy ILG, no region in layer 2 can be explored before all layer 3 regions are explored for $V \leq 1$. So, agent 3 explores $\mathcal{O}_3$ for $[0,1] \times [0,1] \times [0,1]$. The opportunity cost for this region is exactly the surplus, and the base cost is 0, so this uses the entire surplus exactly. So, no other region is explored.

At this point, there are no regions with positive gain remaining. So, whatever the surplus is for agent 4, will be the surplus for all subsequent agents. So:

$$S_4 = P(\mathcal{O}_1 \leq 1)(E(max(\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3)|\mathcal{O}_1 \leq 1) - E(\mathcal{O}_1|\mathcal{O}_1 \leq 1)) = \frac{1}{2}(\frac{3}{4} - \frac{1}{2}) = \frac{1}{8}$$

To calculate how many more agents are required, calculate the total cost for exploring the rest of layer 2 followed by layer 3, and divide by $\frac{1}{8}$:

The total remaining cost for layer 2 of $[1,2] \times [0,1] \times [0,1]$ is:

$$P(\mathcal{O}_1 > 1)(E(\mathcal{O}_1|\mathcal{O}_1 > 1) - \mu_2) = \frac{1}{2}(\frac{3}{2} - \frac{1}{2}) = \frac{1}{2}$$

The total remaining cost for layer 3 is the same, since $\max(\mathcal{O}_1, \mathcal{O}_2) = \mathcal{O}_2$ over this region. So, in the greedy policy, agents 4-7 explore the rest of layer 2, and agents 8-11 explore the rest of layer 3.

# Bibliography

Patrick Bolton and Christopher Harris. Strategic experimentation. *Econometrica*, 67(2): 349–374, March 1999.

Ally Corliss. Birchbox feedback: A refresher course, 2012. URL `http://blog.birchbox.com/post/13554242164/birchbox-feedback-a-refresher-course`.

Nadav Golbandi, Yehuda Koren, and Ronny Lempel. Adaptive bootstrapping of recommender systems using decision trees. In *Proceedings of the fourth ACM international conference on Web search and data mining*, page 595604. ACM, 2011. URL `http://dl.acm.org/citation.cfm?id=1935910`.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, October 2011. ISSN 0002-8282. doi: 10.1257/aer.101.6.2590. URL `http://pubs.aeaweb.org/doi/abs/10.1257/aer.101.6.2590`.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "Wisdom of the crowd". In *14th ACM Conference on Electronic Commerce*. ACM, June 2013. ISBN 978-1-4503-1961-1.

Linyuan Lu, Mat Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. Recommender systems. *Physics Reports*, 519(1):1–49, Octo-

ber 2012. ISSN 03701573. doi: 10.1016/j.physrep.2012.02.006. URL http://linkinghub.elsevier.com/retrieve/pii/S0370157312000828.