Implementing the "Wisdom of the Crowd"*

Ilan Kremer† Yishay Mansour‡ Motty Perry§

25 May 2013

Abstract

We study a novel mechanism design model in which agents each arrive sequentially and choose one action from a set of actions with unknown rewards. The information revealed by the principal affects the incentives of the agents to explore and generate new information. We characterize the optimal disclosure policy of a planner whose goal is to maximize social welfare. One interpretation of our result is the implementation of what is known as the "wisdom of the crowd". This topic has become increasingly relevant with the rapid spread of the Internet over the past decade.

*We wish to thank Michael Borns for his invaluable editorial work.

†Ilan Kremer: Stanford University and the Hebrew University of Jerusalem, ikremer@stanford.edu.

‡Yishay Mansour: Tel Aviv University, mansour@tau.ac.il. This research was supported in part by the Google Inter-university Center for Electronic Markets and Auctions, the Israeli Centers of Research Excellence program, the Israel Science Foundation, the United States-Israel Binational Science Foundation, and by The Israeli Centers of Research Excellence (I-CORE) program, (Center No. 4/11).

§Motty Perry: University of Warwick, motty@huji.ac.il. This research was supported by Israel Science Foundation grant 032-1852.

1 Introduction

The Internet has proven to be a powerful channel for sharing information among agents. As such it has become a critical element in implementing what is known as the "wisdom of the crowd." Hence it is not that surprising that one of the most important recent trends in the new Internet economy is the rise of online reputation systems that collect, maintain, and disseminate reputations. There are now reputation systems for such things as high schools, restaurants, doctors, travel destinations, and even religious gurus. A naive view is that perfect-information sharing through the Internet allows for favorable learning and supports the optimal outcome. We argue that this is not the case because one of the important characteristics of these new markets is the *feedback effect* where users are consumers as well as generators of information. Information that is revealed today affects the choices of future agents and as a result affects the new information that will be generated. A policy that ignores this effect and simply provides the most accurate current recommendation will lead in the long run to insufficient exploration and hence a suboptimal outcome. In this paper, we take a first step toward characterizing an optimal policy of information disclosure when agents behave strategically and, unlike the planner, are myopic.

To this end, we study a novel mechanism design problem in which agents arrive sequentially and each in turn chooses one action from a fixed set of actions with unknown rewards. The agent's goal is to maximize his expected rewards given the information he possesses at the time of arrival. Only the principal, whose interest is to maximize social welfare, observes all past outcomes and can affect the agent's choices by revealing some or all of his information. The principal's challenge is to choose an optimal disclosure/recommendation policy while taking into account that agents are self-interested and myopic. Clearly, a policy not to reveal any information would cause all agents to select the a priori better action, and hence would lead to an inefficient outcome. Nevertheless, a policy of full transparency is not

optimal either because it does not address the incentives of selfish agents, and hence does not allow for enough exploration. Information is a public good and as such one needs to be careful to provide proper incentives to an agent to explore and produce new information. Note that contrary to what is commonly assumed, in our setup the principal is the one who possesses the information, which he reveals optimally through the chosen mechanism.

The new "Internet Economy" provides several related examples for which our model is relevant, and perhaps the first to come to mind is TripAdvisor. TripAdvisor operates within the travel industry, one of the world's largest industries accounting for 11.7% of world GDP and 8% of employment. As its name suggests, TripAdvisor is a website that offers travel advice to its users. It does so by soliciting reviews from users and providing rankings of hotels and restaurants around the world. The company's rankings are based on their own metric called "the Popularity Index," which is a proprietary algorithm. Note that while the individual reviews are also available to users, it is obvious to anyone familiar with TripAdvisor that they are of secondary importance to the rankings, simply because of the overwhelming numbers of properties and reviews. A typical user, then, mostly relies on the rankings and reads only a few reviews to refine his search.

The company is by far the dominant source in the hospitality space, with more than 75 million reviews generated by some 40 million visitors per month.[1] Indeed, the data speaks for itself: the closer a property is to a number-one ranking in its given market, the more numerous its direct online bookings. For example, a property ranked #1 sees 11% more booking per month than the one ranked #2.[2] This difference is striking given that in

[1] See Jeacle and Carter (2011).

[2] The information breaks down as follows: Properties ranked #20 in their market see 10% more booking per month than those ranked #40; properties ranked #10 in their market see 10% more booking per month than those ranked #20; properties ranked #5 in their market see 9% more booking per month than those ranked #10; properties ranked #2 in their market see 7% more booking per month than those ranked #5; properties ranked #1 in their market see 11% more booking per month than those ranked #2 (see

most cases, the difference between similarly ranked hotels is minor.

TripAdvisor's revenue is generated through advertising, and as a result the company's main concern is the volume of visitors to its site. We note, however, that high volume is achieved when the utility of the average customer is maximized. It follows that the company's goal is akin to that of a benevolent social planner. TripAdvisor's Popularity Index is a company secret, yet it is apparent that its exact strategy differs from just a simple aggregation. In this paper, we point to one important aspect of optimality that the company needs to consider.

Another interesting example is a company called Waze-Mobile, which developed a GPS navigation software based on the wisdom of the crowd. Waze's popularity in the West coast is second only to Google Maps, whereas in developing markets such as Brazil, Uruguay, and Indonesia it has surpassed Google by far.[3] Waze is a social mobile application that provides free turn-by-turn navigation based on real-time traffic conditions as reported users. The greater the number of drivers who use this software, the more beneficial it is to its customers. When a customer logs in to Waze with his smartphone, he continuously sends information to Waze about his speed and location and this information, together with information sent by others, enables Waze to recommend to this driver as well as all other drivers an optimal route to their destination. However, in order to provide good recommendations, Waze must have drivers on every possible route. Indeed, as Waze's own president and cofounder admitted,[4] Waze sometimes recommends a particular route to a driver despite (indeed, because of) the absence of information on that route. The information transmitted by this driver is then used to better serve future

Digital Compass by MICROS eCommerce on April 25, 2013).

A similar study about the Travelocity website illustrates that if a hotel increases its score by 1 point on a 5-point scale, the hotel can raise its price by 11.2 percent and still maintain the same occupancy or market share. See Anderson (2012).

[3]Waze, with a user base above 45 millions, was recently bought by Google for roughly $1.1 billion.

[4]http://www.ustream.tv/recorded/21445754

drivers. However, in order not to deter drivers from using the system, Waze must be very careful about how often they "sacrifice" drivers to improve the experience of others. Our model sheds some light on exactly this trade-off.

TripAdvisor and Waze are just two fascinating examples of the rapid growth in the number of rankings and league tables published in recent years and they may well be the face of things to come. Internet evaluations of goods and services are now commonplace. Influential websites provide ratings for activities as diverse as the relative merit of various books and CDs and the teaching prowess of university professors. As we argue in this paper, the managers of these Web sites are facing a non-trivial task as there is a conflict between gathering information from users and making good recommendations to the same users.

Our model also relates to the recent controversy over the health care report-card system. This system entails a public disclosure of patient health outcomes at the level of the individual physician. Supporters argue that the system creates powerful incentives for physicians to improve the quality of health care and also provides patients with important information. Skeptics counter that report cards may encourage physicians to "game" the system by avoiding sick patients, accepting healthy patients, or both. We look at this problem from a different angle by asking how the information available can be optimally revealed to maximize social welfare while taking account of the users' incentives.[5]

With no pretensions to providing a comprehensive solution to this problem, the present paper should be seen as a first small step in this direction. Indeed, the model presented in Section 2 is the simplest one possible that allows us to study the interaction between an informed planner and agents, as described above. In the model the set of actions contains only two *de-*

[5] A striking example is the recent Israeli court order that the government reveal the performance of child liver transplants in public hospitals. Although the evidences was far from statistically significant, parents overwhelmingly preferred to seek the operation abroad and the activity in Israel was virtually stopped.

terministic actions with unknown rewards. We first assume that agents are fully informed about their place in line. For this case the principal's optimal policy is characterized in Section 3. In the optimal policy agent one always chooses the action with the higher mean and we denote his reward by r_1. If $r_1 \in I^t$ then agent t is the first agent to whom the principal recommends trying the other action, while for all agents $t' > t$ the recommendation is the better of the two actions. We show that the sets $\{I^t\}_{t\in T}$ are given by an increasing sequence of thresholds $\{i^t\}_{t\in T}$ where $I^t = (i^{t-1}, i^t)$, and that the number of agents who choose a suboptimal action is bounded independently of T. Consequently, as the number of agents increases, the social welfare converges to the first-best welfare in the unconstrained mechanism.

The informational assumption is relaxed in Section 4, where we assume that agents know only the block to which they belong (say, before, during, or after rush hour) and show that the optimal policy is also a threshold policy. The coarser the partition of the blocks is, the closer the outcome is to the first best, which is obtained in the extreme when all agents belong to the same block.

It is worth noting that in the blocks model, agents have incentives to spend resources to obtain information about their location. If this is a relevant concern, a planner may choose to implement the policy that assumes that agents know their exact location so as to eliminate agents' incentives to waste resources on finding their location. Thus, in such a situation one is led to consider the problem in which the agents know their exact location in line.

In Section 5, we describe a model where the realized payoff of each action is stochastic. We show that our mechanism naturally extends to this case and yields a near optimal performance. Solving for the first-best mechanism in the stochastic setting is difficult and remains an open problem.

We conclude in Section 6 by arguing that a planner who can use monetary transfers will make the best use of his resources by spending it all on relaxing

the IC constraint of the second agent so as to keep the mechanism intact for all other agents.

1.1 Related Literature

The literature on informational cascades that originated with the work of Bikhchandani, Hirshleifer, and Welch (1992) is probably the closest to the model presented here. An informational cascade occurs when it is optimal for an individual who has observed the actions of those ahead of him to follow their behavior without regard to his own information. Our problem is different as agents are not endowed with private signals. Instead we examine a social planner who can control the information received by each individual while implementing the optimal informational policy.

The agents in the model considered here must choose from a set of two-armed bandits (see the classical work of Rothschild 1974). But unlike the vast early work on single-agent decision-making, our work considers strategic experimentation where several agents are involved, along the lines of more recent work by Bolton and Harris (1999) and Keller, Rady, and Cripps (2005), to name just a few. The major departure from the single-agent problem is that an agent in a multi-agent setting can learn from experimentation by other agents. Information is therefore a public good, and a free-rider problem in experimentation naturally arises. It is shown that because of free-riding, there is typically an inefficiently low level of experimentation in equilibrium in these models. In contrast, in our model, free-riding is not a problem as agents have only one chance to act, namely, when it is their turn to move. Our contribution is in approaching the problem from a normative, mechanism design point of view.

Another related paper is Manso (2012) which studies an optimal contract design in a principal-agent setting in which the contract motivates the agent to choose optimally from a set of two-armed bandits. Yet, while in Manso's setup there is one agent who works for two periods, in our setup there are

multiple agents who choose sequentially.

Mechanism design without monetary transfers has been with us from the early days when the focus of interest was the design of optimal voting procedures. One such model that shares the sequential feature of our model is that of Gershkov and Szentes (2009) who analyze a voting model in which there is no conflict of interest between voters and information acquisition is costly. In the optimal mechanism the social planner asks voters randomly and one at a time to invest in information and to report the resulting signal. In recent years, the interest in this type of exercise has gone far beyond voting, as for example in the paper of Martimort and Aggey (2006) which considers the problem of communication between a principal and a privately informed agent when monetary incentives are not available.

Also relevant and closely related to our work are the papers by Kamenica and Gentzkow (2011) and Rayo and Segal (2010). These two papers consider optimal disclosure policies where a principal wishes to influence the choice of an agent by sending the right message. A version of our model with two agents only, is very similar to what they consider. Our contribution is in our consideration of the dynamic aspects of the problem, the real action beginning from the third agent onward.

Finally, two recent papers that examine disclosure of information in a dynamic setup that is very different from ours are Ely, Frankel, and Kamenica (2013) and Horner and Skrzypacz (2012). Ely, Frankel, and Kamenica (2013) consider the entertainment value of information in the media. They examine how a newspaper may release information so as to maximize the utility that readers derive from surprises and suspense. Horner and Skrzypacz (2012) examine a dynamic model in which an agent sells information over time.

2 Model

We consider a binary set of actions $A = \{a_1, a_2\}$. The reward R_i of action a_i is deterministic but ex ante unknown. We assume that each R_i is drawn independently from a continuous distribution π_i that has full support and is common knowledge, and we let π be the joint distribution. Let $\mu_i = E_{R_i \sim \pi_i}[R_i]$ and assume without loss of generality that $\mu_1 \geq \mu_2$.

There are T agents who arrive one by one, choose an action, and realize their payoff. Agents do not observe prior actions and payoffs. We start by assuming that agents know their exact place in line. In Section 4 we show that the main ingredients of the optimal policy remain the same when this assumption is relaxed and agents receive only a noisy signal about their position. The planner, on the other hand, observes the entire history, which consists of his recommendations to the agents as well as their choices and rewards. Let h^t denote a particular history of length t where H^t stands for the set of histories of length t. The planner *commits* to a message (disclosure) policy, which in the general setup is a sequence of functions $\{\tilde{M}^t\}_{t=1,\dots,T}$ where $\tilde{M}^t : H^{t-1} \to M^t$ is a mapping from the set of histories H^{t-1} to the set of possible messages to agent t.[6] Finally, a strategy for agent t is a function $\sigma^t : M^t \to A$.

The goal of agent t is to maximize his expected payoff conditional on his information, while the goal of the planner is to maximize the expected average reward, i.e., $E[\frac{1}{T}\sum_{t=1}^{T} R^t]$. An alternative objective for the planner would be to maximize the discounted payoff, $E[\sum_{t=1}^{T} \gamma^t R^t]$, for some discounting factor $\gamma \in (0,1)$. We focus on the average payoff as it is more suitable to our setup, but a similar result holds if the planner wishes to maximize the discounted payoff.

[6] Restricting the planner to pure strategies is done for the sake of simplicity only. It is easy to see that each of the arguments in the following sections holds true when the planner is also allowed to use mixed strategies, and that the resulting optimal strategy of the planner is pure.

Before we proceed to characterize the optimal solution we note that one can generalize our model so that the distribution of payoffs does not have full support. The distribution does not even need to be continuous. These assumptions are made to simplify the exposition. However, it is important that when $\mu_1 \geq \mu_2$ there be a positive probability that the first action's payoff is lower than μ_2, that is, that $\Pr(R_1 < \mu_2) > 0$ holds when we assume full support. If, however, $\Pr(R_1 < \mu_2) = 0$, then all the agents will choose the first action regardless of any recommendation policy. This follows as every agent knows that everyone before him chose the first action simply because any payoff of the first action exceeds the mean of the second action. In such a setup a planner will find it impossible to convince agents to explore.

3 The Optimal Mechanism

Let us first give an overview of the mechanism and the proof. We start by providing a simple example that illustrates the main properties of the optimal mechanism. Then in Section 3.2 we present some basic properties of incentive-compatible mechanisms. In particular, we establish a revelation principle and show that without loss of generality, we can concentrate on recommendation mechanisms that specify for each agent which action to perform (Lemma 1). We show that once both actions are sampled, the mechanism recommends the better action and stays incentive compatible (Lemma 2). In Section 3.3 we explore the incentive-compatible constraint of the agents.

Section 3.4 develops the optimal mechanism. We first show that initially the optimal mechanism explores as much as possible (Lemma 4). We then show that any value of the better a priori action that is lower than the expectation of the other action causes the second agent to undertake an exploration (Lemma 5). The main ingredient in our proof is that the lower realizations are better incetives for exploration than the higher realizations (Lemma 6). Finally, there is some value of the better action that realizations

above it deter the principal from undertaking any exploration.

This result implies that the optimal incentive-compatible mechanism is rather simple. The principal explores as much as he can (given the incentive-compatible mechanism) up to a certain value (depending on T) for which he does not perform any exploration.

3.1 Example

To gain a better intuition of what follows, consider an example in which the payoff of the first alternative, R_1, is distributed uniformly on $[-1, 5]$ while the payoff of the second alternative, R_2, is distributed uniformly on $[-5, 5]$. For simplicity, suppose that the principal wishes to explore both alternatives as soon as possible.[7]

Consider first what would happen in the case of full transparency. The first agent will choose the first action. The second agent will choose the second alternative only if the payoff of the first alternative is negative, $R_1 \leq 0$. Otherwise, he and all the agents after him will choose the first alternative, an outcome that is suboptimal.

Now consider a planner who does not disclose R_1 but instead recommends the second alternative to the second agent whenever $R_1 \leq 1$. The agent will follow the recommendation because he concludes that the expected value of the first alternative is zero, which is equal to the expected value of the second alternative. This implies that the outcome under this policy allows more exploration as compared to the policy under full transparency. Hence, we can already conclude that full transparency is suboptimal.

But we can do even better. Consider the more interesting case, the recommendation for agent three. Suppose that the planner's policy is such that he recommends that agent three use the second alternative if one of the following two cases obtains: (I) the second agent has been recommended to

[7] The decision to explore depends on both the realization of R_1 and of T. However for large T the planner would like to explore for almost all values of R_1.

test the second action ($R_1 \leq 1$) and based on the experience of the second agent the planner knows that $R_2 > R_1$, or (II) the third agent is the first to be recommended to test the second alternative because $1 < R_1 \leq 1 + x$ (to be derived below). Note that conditional on (I) the agent *strictly* prefers to follow the recommendation, while conditional on (II) he prefers not to, and the higher x is, the less attractive the recommendation is. In the appendix we show that for $x = 2.23$ the agent is just indifferent.

The computation for the fourth agent is similar, and here we get that this agent will explore (i.e., be the first to test R_2) for the remaining values of $R_1 < 5$. The better of the two actions is recommended to all the remaining agents.

The rest of the paper is devoted to showing that this logic can be extended to form the optimal policy and that the number of exploring agents is bounded.

3.2 Preliminary

We start the analysis with two simple lemmas that, taken together, establish that it is possible without loss of generality to restrict attention to a special class of mechanisms in which the principal recommends an action to the agents, and once both actions are sampled, the better of the two is recommended thereafter. The first lemma is a version of the well-known *Revelation Principle*.

Definition 1 *A recommendation policy is a mechanism in which at time t, the planner recommends an action $a^t \in A$ that is incentive compatible. That is, $E[R_j - R_i | a^t = a_j] \geq 0$ for each $a_j \in A$. We denote by $\hat{M}$ the set of recommendation policies.*

Note that the above expectation $E[R_j - R_i | a^t = a_j]$ implicitly assumes that the agent knows the mechanism. Hence, from now on, whenever we refer

to a mechanism as incentive compatible, we assume that the agent knows the mechanism and takes it as given.

Lemma 1 *For any mechanism M, there exists a recommendation mechanism that yields the same expected average reward.*

The above lemma is a special case of Myerson (1988) and consequently the proof is omitted.

Thus, we can restrict our attention to recommendation policies only. The next lemma allows us to focus the discussion further by restricting attention to the set of partition policies. A partition policy has two restrictions. The first is that the principal recommends action a_1 to the first agent. This is an essential condition for the policy to be *IC*. The second restriction is that once both actions are sampled, the policy recommends the better one.

Definition 2 *A partition policy is a recommendation policy that is described by a collection of disjoint sets $\{I_t\}_{t=2}^{T+1}$. If $r_1 \in I^t$ for $t \leq T$, then agent t is the first agent for whom $a^t = a_2$ and for all $t' > t$ we have $a^{t'} = \max\{a_1, a_2\}$. If $r_1 \in I^{T+1}$, then no agent explores. If $I^t = \emptyset$, then agent t never explores.*

Lemma 2 *The optimal recommendation mechanism, is a partition mechanism.*

Proof: Note first that since $\mu_1 \geq \mu_2$, the first agent will always choose the first action. Also, since the principal wishes to maximize the average reward, $E[\frac{1}{T}\sum_{t=1}^{T} R^t]$, it will always be optimal for him to recommend the better action once he has sampled both actions. Clearly, recommending the better of the two actions will only strengthen the *IC* of the agent to follow the recommendation. Hence, for each agent $j \geq 2$ we need to describe the realizations of R_1 that will lead the planner to choose agent j to be the first agent to try the second action. □

We next show that the optimal partition is a threshold policy.

3.3 Incentive-Compatibility (IC) Constraints

Agent t finds the recommendation $a^t = a_2$ incentive compatible if and only if

$$E(R_2 - R_1 | a^t = a_2) \geq 0 \ .$$

Note that this holds if and only if

$$\Pr(a^t = a_2) * E(R_2 - R_1 | a^t = a_2) \geq 0 \ .$$

We use the latter constraint, since it has a nice intuitive interpretation regarding the distribution, namely,

$$\int\limits_{a^t = a_2} [R_2 - R_1] d\pi \geq 0 \ .$$

For a partition policy the above constraint can be written as

$$\int\limits_{R_1 \in \cup_{\tau<t} I^\tau, R_2 > R_1} [R_2 - R_1] d\pi + \int\limits_{R_1 \in I^t} [\mu_2 - R_1] d\pi \geq 0 \ . \tag{1}$$

The first integral represents *exploitation*, which is defined as the benefit for the agent in the event that the principal is informed about both actions, i.e., $R_1 \in \cup_{\tau<t} I^t$. Obviously this integrand is positive. The second integral, the *exploration* part, represents the loss in the case where the principal wishes to explore and agent t is the first agent to try the second action. We show that in the optimal mechanism this integrand is negative. Alternatively (1) can be expressed as

$$\int\limits_{R_1 \in \cup_{\tau<t} I^\tau, R_2 > R_1} [R_2 - R_1] d\pi \geq \int\limits_{R_1 \in I^t} [R_1 - \mu_2] d\pi \ .$$

The following lemma shows that it is sufficient to consider the IC of action a_2.

Lemma 3 *Assume that the recommendation $a^t = a_2$ to agent t is IC. Then the recommendation $a^t = a_1$ is also IC.*

Proof: Let $K^t = \{(R_1, R_2)|a^t = a_2\}$ be the event in which the recommendation to agent t is $a^t = a_2$. If $K^t = \emptyset$, then the lemma follows since $E[R_1 - R_2] > 0$. Otherwise $K^t \neq \emptyset$ and because the recommendation $a^t = a_2$ is *IC* we must have $E[R_2 - R_1|K^t] \geq 0$. Recall, however, that by assumption $E[R_2 - R_1] \leq 0$.

Now, since

$$E[R_2 - R_1] = E[R_2 - R_1|K^t]\Pr[K^t] + E[R_2 - R_1|\neg K^t]\Pr[\neg K^t] \leq 0,$$

it follows that $E[R_2 - R_1|\neg K^t] \leq 0$, which in particular implies that recommending $a_t^t = a_1$ is *IC* in the case of $\neg K^t$. □

3.4 Optimality of the Threshold Policy

Definition 3 *A threshold policy is a partition policy in which the sets I^t are ordered intervals. Formally, $I^2 = (-\infty, i^2]$, $I^t = (i^{t-1}, i^t]$.*

Note that if $i^{t-1} = i^t$ then $I^t = \emptyset$ and agent t never explores.

The following simple claim establishes that in every period, the planner will undertake as much exploration as the *IC* condition allows.

Lemma 4 *Let M^* be an optimal partition policy and assume that in M^* agent $t + 1 \geq 3$ explores with some positive probability (i.e., $\Pr[I^{t+1}] > 0$). Then agent t has a tight IC constraint.*

Proof: Assume by way of contradiction that agent t does not have a tight *IC* constraint. Then we can "move" part of the exploration of agent $t + 1$ to agent t, and still satisfy the *IC* constraint. The average reward will only increase, since agent $t + 1$, rather than exploring in this set of realizations of

R_1, will choose the better of the two actions. To be precise, assume that the IC condition for agent t does not hold with equality. That is,

$$\int_{R_1 \in I^t} [R_1 - \mu_2] d\pi \quad < \int_{R_1 \in \cup_{\tau<t} I^\tau, R_2 > R_1} [R_2 - R_1] d\pi \tag{2}$$

Recall that I^t consists of those values r_1 for which agent t is the first to explore action a_2 when $R_1 = r_1$. By assumption we have $\Pr[I^{t+1}] > 0$. Note that the RHS of (2) does not depend on I^t. Therefore, we can find a subset $\hat{I} \subset I^{t+1}$ where $\Pr[\hat{I}] > 0$ and then replace the set I^t with $I'^t = I^t \cup \hat{I}$ and the set I^{t+1} with $I'^{t+1} = I^{t+1} - \hat{I}$ and still keep the IC constraint. The only change is in the expected rewards of agents t and $t+1$.

Before the change, the expected sum of rewards of agents t and $t+1$, conditional on $R_1 \in \hat{I}$, was $E[R_1 | R_1 \in \hat{I}] + \mu_2$, while the new sum of expected rewards (again conditional on $R_1 \in \hat{I}$,) is $\mu_2 + E[\max\{R_1, R_2\} | R_1 \in \hat{I}]$, which is strictly larger (since the prior is continuous). The IC constraint of agent t holds by construction of the set $\hat{I}$, while the constraint for agent $t+1$ holds because his payoffs from following the recommendation increased since we removed only exploration. None of the other agents is affected by this modification. Therefore, we have reached a contradiction to the claim that the policy is optimal. □

Lemma 5 *In the optimal partition policy, agent 2 explores for all values $r_1 \leq \mu_2$. Formally*

$$I^2 \supseteq \{r_1 : r_1 \leq \mu_2\}.$$

Proof: Assume that policy M' is a partition policy and let B include the values of the first action that are below the expectation of the second action, and are not in I^2, i.e.,[8]

$$B = \{r_1 : r_1 \leq \mu_2,\ r_1 \notin I^2\}.$$

[8]Recall that we assume that $\Pr[R_1 < \mu_2] > 0$.

If $\Pr[B] > 0$ then a policy M'', which is similar to M' except that now $I'^2 = B \cup I^2$ and $I'^t = I^t - B$ for $t \geq 3$, is a recommendation policy with a higher expected average reward. Consider the policy M' and let $B^t = B \cap I^t$ for $t \geq 3$. Because M' is a recommendation policy, agent t finds it optimal to follow the recommendations and in particular to use action a_2 when recommended. Next consider the policy M'' and observe that the incentives of agent t to follow the recommendation to use action a_2 are stronger now because for $R_1 \in B^t$ his payoff in M' is R_2 while in M'' it is $\max\{R_1, R_2\}$. The agents t between 3 and T have a stronger incentive to follow the recommendation, since now in the event of $R_1 \in B^t$ we recommend the better of the two actions rather than a_1. Because $R_1 < \mu_2$ it is immediate that expected average rewards in M'' are higher than in M'. For agent 2 we have only increased the *IC*, since $E[R_2 - R_1 | R_1 \in B] \geq 0$. □

The discussion so far allows us to restrict attention to partition policies in which: (i) once both R_1 and R_2 are observed, the policy recommends the better action, (ii) the *IC* constraint is always tight, and (iii) the set $I^2 \supseteq \quad (-\infty, \mu_2]$. Next, we argue that we should also require the policy to be a threshold policy. Note that if a partition policy $\{I^j\}_{j=2}^{T+1}$ is not a threshold policy (up to measure zero events) then there exist indexes $t^2 > t^1$ and sets $B^1 \subseteq I^{t_1}$and $B^2 \subset I^{t_2}$ such that: $\sup B^2 < \inf B^1$ and $\Pr[B^1], \Pr[B^2] > 0$.

A useful tool in our proof is an operation we call *swap,* which changes a policy M' to a policy M''.

Definition 4 *A swap operation is a modification of a partition policy. Given two agents* t_1 *and* $t_2 > t_1$ *and subsets* $B^1 \subset I^{t_1}$, $B^2 \subset I^{t_2}$ *where* $\sup B^2 < \inf B^1$, *swap constructs a new partition policy such that* $I'^{t_1} = I^{t_1} \cup B^2 - B^1$ *and* $I'^{t_2} = I^{t_2} \cup B^1 - B^2$, *while other sets are unchanged, i.e.,* $I'^t = I^t$ *for* $t \notin \{t_1, t_2\}$.

Definition 5 *We say that a swap is proper if* [9]

$$\int_{R_1 \in B^1} [\mu_2 - R_1] d\pi \quad = \int_{R_1 \in B^2} [\mu_2 - R_1] d\pi.$$

Lemma 6 *The optimal recommendation policy is a threshold policy.*

Proof: Let M be a recommendation policy that is not a threshold policy. Following the discussion above one can construct a proper swap. Let M' be the resulting recommendation policy. Consider a proper swap operation. First we show that the swap does not change the expected reward of agent t_1 conditional on a recommendation to choose action a_2. From the perspective of agent t_1, the change is that in the case where $r_1 \in B^1$ the action recommended to him at M' is a_1 rather than the action a_2 recommended to him at M, and in the case where $r_1 \in B^2$ it is a_2 (at M') rather than a_1 (at M). Since the swap operation is proper, his IC constraint at M' can be written as

$$\begin{aligned} &\int_{R_1 \in \cup_{\tau<t_1} I^\tau, R_2 > R_1} [R_2 - R_1] d\pi + \int_{R_1 \in I^{t_1}} [\mu_2 - R_1] d\pi \qquad (3) \\ &+ \int_{R_1 \in B^2} [\mu_2 - R_1] d\pi - \int_{R_1 \in B^1} [\mu_2 - R_1] d\pi \\ &= \int_{R_1 \in \cup_{\tau<t_1} I^\tau, R_2 > R_1} [R_2 - R_1] d\pi + \int_{R_1 \in I^{t_1}} [\mu_2 - R_1] d\pi \geq 0. \end{aligned}$$

Therefore the swap does not change the expected reward of agent t_1 and M' satisfies IC for this agent.

Next consider all agents except agents t_2 and t_1. Observe first that all agents $t < t_1$ and $t > t_2$ do not observe any change in their incentives (and rewards) and we are left with agents t where $t_1 < t < t_2$. The expected rewards of these agents can only increase because the effect of the swap is

[9] A proper swap always exists when a swap operation exists due to our assumption on no mass points.

only on the first integral $\int\limits_{R_1 \in \cup_{\tau<t} I^\tau, R_2 > R_1} [R_2 - R_1] d\pi$ of the IC constraint (see (3)) which increases as a result of the swap because instead of the set $\cup_{\tau<t} I^\tau$ we now have $\cup_{\tau<t} I^\tau \cup B^2 - B^1$ and $\sup B^2 < \inf B^1$.

Thus, it is left for us to analyze the incentives and rewards of agent t_2 (and only when $t_2 \leq \bar{T}$) to follow the recommendation to choose action a_2. First observe that if $r_1 \notin B^1 \cup B^2$ then M and M' are identical, and hence the only case to consider is when $r_1 \in B^1 \cup B^2$. The expected reward under M conditional on $r_1 \in B^1 \cup B^2$ is

$$\frac{1}{\Pr[B^1 \cup B^2]} \left[\int\limits_{R_1 \in B^1, R_2 > R_1} [R_2 - R_1] d\pi + \int\limits_{R_1 \in B^2} [\mu_2 - R_1] d\pi \right],$$

and the expected reward under M' is

$$\frac{1}{\Pr[B^1 \cup B^2]} \left[\int\limits_{R_1 \in B^2, R_2 > R_1} [R_2 - R_1] d\pi + \int\limits_{R_1 \in B^1} [\mu_2 - R_1] d\pi \right],$$

We would like to show that

$$\int\limits_{R_1 \in B^1, R_2 > R_1} [R_2 - R_1] d\pi + \int\limits_{R_1 \in B^2} [\mu_2 - R_1] d\pi \quad < \quad \int\limits_{R_1 \in B^2, R_2 > R_1} [R_2 - R_1] d\pi + \int\limits_{R_1 \in B^1} [\mu_2 - R_1] d\pi,$$

which is equivalent to showing that (recall that the swap is proper)

$$\int\limits_{R_1 \in B^1} \int\limits_{R_2 > R_1} [R_2 - R_1] d\pi \quad < \int\limits_{R_1 \in B^2} \int\limits_{R_2 > R_1} [R_2 - R_1] d\pi.$$

Since $(-\infty, \mu_2] \subseteq I^2$ and $\inf B^1 > \sup B^2$ we conclude that $\Pr[B^2] > \Pr[B^1]$, which implies the last inequality. This again implies that the IC constraint is satisfied for this agent and that the swap operation increases his rewards.

We now show that the proper swap operation increases the expected payoff. First consider agent t_1. His net change in expected payoff is

$$\int_{R_1 \in B^2} [R_2 - R_1] d\pi \; + \int_{R_1 \in B^1} [R_1 - R_2] d\pi \; = \; 0,$$

where the equality follows since it is a proper swap. Next consider agents t where $t_1 < t < t_2$. The net change in expected payoff of agent t is

$$\begin{aligned} \int_{R_1 \in B^2} \int_{R_2} \max\{R_2, R_1\} - R_1 d\pi \; - \int_{R_1 \in B^1} \int_{R_2} \max\{R_2, R_1\} - R_1 d\pi \; = \\ \int_{R_1 \in B^2, R_2 > R_1} [R_2 - R_1] \, d\pi \; - \int_{R_1 \in B^2, R_2 > R_1} [R_2 - R_1] \, d\pi \; \geq \; 0. \end{aligned}$$

The last inequality, similar to (3), follows from the fact that $\Pr(B^2) > \Pr(B^1)$ and that $\sup(R_1 | R_1 \in B^2) < \inf(R_1 | R_1 \in B^1)$.

Finally, we consider agent t_2, where the net change in expected payoffs is,

$$\begin{aligned} \int_{R_1 \in B^2} \int_{R_2} \max\{R_2, R_1\} - R_2 d\pi \; - \int_{R_1 \in B^1} \int_{R_2} \max\{R_2, R_1\} - R_2 d\pi \; = \\ \int_{R_1 \in B^2} \int_{R_2} \max\{R_2, R_1\} - R_1 d\pi \; - \int_{R_1 \in B^1} \int_{R_2} \max\{R_2, R_1\} - R_1 d\pi \; = \\ \int_{R_1 \in B^2, R_2 > R_1} [R_2 - R_1] \, d\pi \; - \int_{R_1 \in B^2, R_2 > R_1} [R_2 - R_1] \, d\pi \; \geq \; 0. \end{aligned}$$

where the first equality follows from the fact that it is a proper swap, and the inequality follows as in (3). □

Lemma 6 implies that an optimal policy must be a threshold policy. That is, the sets $\{I^t\}_{t \in \bar{T}}$ are restricted to being sets of intervals. Moreover, the *IC* constraint is tight for any agent $t \leq \bar{T}$ provided that there is a positive

probability that agent $t+1$ will be asked to explore.

Note that with a finite number of agents there always exist high enough realizations of R_1 after which exploration is suboptimal. The next section solves for the optimal policy that accounts for this effect.

3.5 The Optimal Threshold Policy

Consider first the case where T is infinite. In this case exploration is maximized as the planner wishes to explore for any realized value of the first action, r_1. The optimal policy is defined by an increasing sequence of thresholds $i^{2,\infty} < i^{3,\infty}$, where for $t = 2$,

$$\int_{R_1=-\infty}^{i^{2,\infty}} [R_1 - \mu_2] d\pi = 0.$$

For $t > 2$, as long as $i^{t,\infty} < \infty$, we have

$$i^{t+1,\infty} = \sup \left\{ i | \int_{R_1 \leq i^t, R_2 > R_1} [R_2 - R_1] d\pi \geq \int_{R_1=i^{t,\infty}}^{i} [R_1 - \mu_2] d\pi \right\}.$$

If $i^{t,\infty} = \infty$ then we define $i^{t',\infty} = \infty$ for all $t' \geq t$. Note that if $i^{t+1,\infty} < \infty$ then the above supremum can be replaced with the following equality:

$$\int_{R_1 \leq i^t, R_2 > R_1} [R_2 - R_1] d\pi = \int_{R_1=i^{t,\infty}}^{i^{t+1,\infty}} [R_1 - \mu_2] d\pi. \tag{4}$$

Consider the case where T is finite. As we shall see, the planner will ask fewer agents to explore. Consider the t-th agent. The RHS is the expected loss due to exploration by the current agent. The expected gain in exploitation, if we explore, is $(T-t)E[\max\{R_2 - r_1, 0\}]$. We set the threshold θ_t for agent

t to be the maximum r_1 for which it is beneficial to explore. Let θ_t be the solution to

$$(T-t)E[\max\{R_2-\theta_t,0\}]=\theta_t-\mu_2.$$

When considering agent t there are $T-t+1$ agents left; then θ_t is the highest value for which it is still optimal to explore. Note that θ_t is increasing in t. Our main result is:

Theorem 7 *The optimal policy, M^{opt}, is defined by the sequence of thresholds*

$$i^{t,T}=\min\{i^{t,\infty},\theta_\tau\},$$

where τ is the minimal index for which $i^{t,\infty}>\theta_t$.

Next we argue that even when T is arbitrarily high, exploration is limited to a bounded number of agents where the bound doesn't depend on either the number of agents or the realizations of R_1 and R_2. This implies that the memory required by the planner to implement the optimal policy is bounded by a constant.

Theorem 8 *Let $t^*=\min\{t|i^t=\infty\}$; then $t^*\leq\frac{\mu_1-\mu_2}{\alpha}$ where*

$$\begin{aligned}\alpha=&\int\limits_{R_1\leq i^2,R_2>R_1}[R_2-R_1]d\pi\\&\geq\Pr[R_2\geq\mu_2]\cdot\Pr[R_1<\mu_2]\cdot\left(E[R_2|R_2\geq\mu_2]-E[R_1|R_1<\mu_2]\right).\end{aligned}$$

Since t^* is finite, the principal is able to explore both actions after t^* agents.

The proof appears in the appendix but we can provide the intuition here. Consider (4): the LHS represents the gain agent t expects to receive by following the recommendation of the principal who has already tested both alternatives. It is an increasing sequence as the planner becomes better informed as t increases. This implies that these terms can be bounded from

below when we consider agent $t = 2$. The RHS represents the expected loss the agent expects to experience when he is the first agent to try the second alternative. The sum of the RHS over all t is $\mu_1 - \mu_2$. The proof is based on these two observations when we sum the LHS and the RHS.

The above theorem has important implications. Consider the first-best outcome in which the principal can force agents to choose an action. The above theorem implies that for any T the aggregate loss of the optimal mechanism as compared to the first-best outcome is bounded by $\frac{(\mu_1-\mu_2)^2}{\alpha}$. As a result we conclude that:

Corollary 9 *As T goes to infinity the average loss per agent as compared to the first-best outcome converges to zero at a rate of $1/T$. Apart from a finite number of agents, t^*, all other agents are guaranteed to follow the optimal action.*

4 Imperfect Information about Location

In this section we relax the assumption that agents are perfectly informed about their location in line and study the consequences of this uncertainty. Indeed, if agents have no information about their location and assign equal probability to every possible position, then it is easy to see that the planner can implement the first-best outcome. This is simply because there is no conflict of interests between the agent and the planner who wishes to maximize the utility of the average agent. In what follows we examine an intermediate case in which agents do not know their exact location but know to which group of agents they belong location-wise. For example, in the context of the real-time navigation problem, it is reasonable to assume that while drivers are not perfectly aware of their exact place in order, they do know whether it is before, during, or after the rush hour.

Thus, consider a sequence of integers $1 = \tau^1 < \tau^2 < ...\tau^k = T + 1$ such that if $\tau^j \leq t \leq \tau^{j+1} - 1$, then agent t believes that his location is uniformly

distributed between $\tau^j$and $\tau^{j+1}-1$. To simplify the exposition we assume that the first agent knows his location (i.e., $\tau_2 = 2$) and therefore always chooses action one. We assume that this sequence is commonly known and refer to the set of agents $\tau^j \leq t \leq \tau^{j+1}-1$ as *block* j. Note that when the number of blocks is T we are in the model of Section 2 while when there is only one block the agents are uninformed and the first best is implementable.

We first argue that our main result of Section 3 also holds in this model and the planner's optimal strategy is a recommendation-based threshold policy. Indeed, the steps leading to the conclusion that the optimal policy is a partition policy are *exactly* the same as in Section 3. Therefore, it suffices to show that the swap operation, which is the key step in our proof of the optimality of a threshold policy, is still valid.

Assume that the planner follows a partition policy. Given the information agents have about their position, their IC constraint now becomes:

$$\frac{1}{\tau^{j+1}-\tau^j}\sum\nolimits_{t=\tau^j}^{\tau^{j+1}-1}\left[\int\limits_{R_1\in\cup_{\tau<t}I^\tau, R_2>R_1}[R_2-R_1]d\pi+\int\limits_{R_1\in I^t}[\mu_2-R_1]d\pi\right]\geq 0. \tag{5}$$

As before, consider a non-threshold partition policy and recall that if a partition policy $\{I^j\}_{j=2}^{T+1}$ is not a threshold policy then there exist indexes $t^2 > t^1$ and sets $B^1 \subseteq I^{t_1}$and $B^2 \subset I^{t_2}$ such that

$$\sup B^2 < \inf B^1 \quad \textit{and} \quad \Pr[B^1], \Pr[B^2] > 0$$

and

$$\int\limits_{R_1\in B^1}[\mu_2-R_1]d\pi \quad = \int\limits_{R_1\in B^2}[\mu_2-R_1]d\pi$$

and we can construct a new partition policy M' such that $\hat{I}^{t_1} = I^{t_1}\cup B^2 - B^1$ and $\hat{I}^{t_2} = I^{t_2}\cup B^1 - B^2$, while the other sets are unchanged, i.e., $\hat{I}^t = I^t$ for $t \notin \{t_1, t_2\}$. Recall from the proof in Section 3.4 that following a proper

swap, the terms

$$\int\limits_{R_1 \in \cup_{\tau<t} I^\tau, R_2 > R_1} [R_2 - R_1] d\pi + \int\limits_{R_1 \in I^t} [\mu_2 - R_1] d\pi$$

weakly increase for *all* agents t. This implies that also the term

$$\frac{1}{\tau^{j+1} - \tau^j} \sum\nolimits_{t=\tau_j}^{\tau^{j+1}-1} \left[\int\limits_{R_1 \in \cup_{\tau<t} I^\tau, R_2 > R_1} [R_2 - R_1] d\pi + \int\limits_{R_1 \in I^t} [\mu_2 - R_1] d\pi \right]$$

weakly increases. We conclude that the *IC* constraint remains the same for some agents and becomes stronger for others and that, following a proper swap, the sum of agents' payoffs strictly increases. We thereby conclude that the optimal policy is a threshold policy, that is, a policy in which the sets I^t are ordered intervals $I^t = (i^{t-1}, i^t]$. Next we argue that in a given block only the first agent explores.

Lemma 10 *In the optimal policy, for every block $j = 1, ..., k$, we have $I^{\tau^j} = (i^{\tau^j - 1}, i^{\tau^j}]$ and $I^t = \varnothing$ for $\tau^j < t < \tau^{j+1}$.*

Proof: Consider an arbitrary threshold policy and a specific block j. Suppose we ask only the first agent in the block to explore, and only when $r_1 \in (i^{\tau^j - 1}, i^{\tau^{j+1} - 1}]$, i.e., whenever someone in the block explores in the original policy. Then the aggregate loss from exploration in the *IC* constraint (see (5)) remains the same for everyone in the block. However, we improve the expected payoff from exploitation for all agents. Hence, the *IC* becomes stronger and the expected welfare higher. □

Note that in the above lemma we may have slack in the *IC* constraint and the planner can even induce more exploration from the first agent in the block. Specifically, we can calculate the optimal threshold $i^{\tau^{j+1}}$ by replacing (4) with

$$\int_{R_1 \leq i^{\tau^{j-1}}, R_2 > R_1} [R_2 - R_1] d\pi + \left(\tau^{j+1} - \tau^J - 1\right) \int_{R_1 \leq i^{\tau^j}, R_2 > R_1} [R_2 - R_1] d\pi$$

$$= \int_{R_1 = i^{\tau^{j-1}}}^{i^{\tau^j}} [R_1 - \mu_2] d\pi.$$

The next theorem summarizes the discussion above.

Theorem 11 *The optimal policy in the blocks model is given by a sequence of thresholds* $\{\theta_i\}$ *such that only the first agent in block* j *explores when* $r_1 \in (\theta_{j-1}, \theta_j]$. *That is action* a_2 *is recommended to all the other agents only when it is known that* $R_2 > R_1$.

Finally, we argue that as the information that agents have about their location becomes coarser, the policy that the planner can implement is closer to the first best. We define a block structure to be *coarser* if it is constructed by joining adjacent blocks. As in the proof of the lemma above, in the optimal policy only the first agent in this new block explores and he explores for a bigger set of realizations. Clearly, this results in a more efficient outcome.

Theorem 12 *If block structure* $\mathcal{B}_1$ *is coarser than block structure* $\mathcal{B}_2$, *then the optimal policy in* $\mathcal{B}_1$ *is more efficient.*

5 The Stochastic Case

Our main goal in this section is to show that we can essentially extend the optimal mechanism in the deterministic case to the stochastic model, and achieve a near optimal expected average reward. The optimal stochastic mechanism will have thresholds like the optimal deterministic mechanism, and at the high level we keep the same structure.

5.1 Model

The stochastic model, like the deterministic model, has a binary set of actions $A = \{a_1, a_2\}$. There is a prior distribution π_i over all possible distributions of payoffs for action a_i, which is common knowledge. From π_i a distribution D_i is drawn before the process starts, and is unknown. The reward R_i of action a_i is drawn independently from the distribution D_i for each agent, and we denote by R_i^t the reward of action a_i to agent t. The a priori expected reward of action a_i is

$$\mu_i = E_{\pi_i}(E_{D_i}[R_i]),$$

and as before we assume w.l.o.g. that $\mu_1 > \mu_2$ and that $\Pr[E[R_1] < \mu_2] > 0$; otherwise exploration is impossible. For simplicity, we assume that the range of any realized R_i is $[0, 1]$. (However, the result can be extended to many other settings.)

Note that in the stochastic model there are two sources of uncertainty. One is the distribution D_i that is selected from the prior π_i. The second is due to the variance in the realizations of R_i^t that are drawn from the distribution D_i.

5.2 Threshold Algorithm for the Sstochastic Model

We define a mechanism S for the planner that guarantees near optimal performance. The parameter of mechanism S is a sequence of thresholds $(\theta_1, \theta_2, \ldots)$. We partition the agents to T/m blocks of m agents each, where the i-th block includes agents $(i-1)m + 1$ until im. All the agents in each block will receive an identical recommendation.

To the agents in block 1, the first m agents, S recommends action a_1. Given the realizations, it computes

$$\hat{\mu}_1 = \frac{1}{m} \sum_{t=1}^{m} R_1^t.$$

Note that $\hat{\mu}_1$ is fixed, and *never changes* and does not necessarily reflect all the information that is available to the planner.

For blocks $i \geq 2$, mechanism S does the following:

1. If $\hat{\mu}_1 \in (\theta_{i-1}, \theta_i]$, then S recommends action a_2. The agents in block i will be the first to explore action a_2. Given the realizations of the rewards, we set $\hat{\mu}_2 = \frac{1}{m} \sum_{t=(i-1)m+1}^{im} R_2^t$ and define $a_{best} = a_1$ if $\hat{\mu}_1 \geq \hat{\mu}_2$ and otherwise $a_{best} = a_2$.

2. If $\hat{\mu}_1 \leq \theta_{i-1}$ then S recommends action a_{best}.

3. If $\hat{\mu}_1 > \theta_i$ then S recommends action a_1.

5.3 Setting the Thresholds

As before, the planner needs to balance exploration and exploitation to guarantee the IC constraint. First, we set $\theta_{2,\infty}$ for block $i = 2$, as the solution to the following equality:

$$0 = E[R_2 - R_1 | \hat{\mu}_1 \leq \theta_{2,\infty}],$$

Then, consider the expected loss, assuming that block $i \geq 3$ was the first to explore action a_2.

$$Loss(\theta_{i-1}, \theta_i) = E[R_1 - R_2 | \hat{\mu}_1 \in (\theta_{i-1}, \theta_i]] \Pr[\hat{\mu}_1 \in (\theta_{i-1}, \theta_i]].$$

Next we consider the expected gain, assuming that action a_2 was already sampled and that $a_{best} = a_2$.

$$Gain(\theta_{i-1}) = E[R_2 - R_1 | \hat{\mu}_1 \leq \theta_{i-1}, a_{best} = a_2] \Pr[\hat{\mu}_1 \leq \theta_{i-1}, a_{best} = a_2].$$

We set $\theta_{i,\infty}$ inductively. After we set $\theta_{j,\infty}$ for $j < i$, we set $\theta_{i,\infty}$ such that $Gain(\theta_{i-1,\infty}) = Loss(\theta_{i-1,\infty}\theta_{i,\infty})$. Let ℓ_i be the solution to

$$(T - mi)E_{D_2}[\max\{E[R_2] - \ell_i, 0\}] = (\ell_i - \mu_2)m.$$

We set the threshold to be $\theta_i = \min\{\ell_i, \theta_{i,\infty}\}$.

5.4 Analysis

The following theorem establishes that as the number of agents T increases, the average loss per agent goes to zero, as compared to the case where the planner knows the distributions of payoffs. Note that this represents a better performance than that of a planner who is not subject to the *IC* constraint as there is no need for explorations. The following theorem establishes the near optimal performance of S.

Theorem 13 *The mechanism S is* IC *and, when we set $m = T^{2/3}\ln T$, it has an average expected reward of at least*

$$E_{D_1,D_2}[\max\{E_{R_1\sim D_1}[R_1], E_{R_2\sim D_2}[R_2]\}] - C\frac{\ln T}{T^{1/3}},$$

where the constant C depends only on π_1 and π_2.

The theorem follows from the observation that, by the Hoeffding inequality, $|\hat{\mu}_i - E[R_i]| \geq \lambda$ with probability at most $\delta \leq 2e^{-2\lambda^2 m}$. For $\lambda = T^{-1/3}$, since $m = T^{2/3}\ln T$, we have $\delta \leq 2T^{-2}$. This implies that we have three sources of loss as compared with always playing the better actiom. The first source of loss is due to the exploration, which spans a constant number of β blocks, and depend only on the priors. Since each block is of size m this contributes at most $\beta m/T$ to the average loss. The second source of loss is the fact that $E[R_i]$ is only approximated; this loss is at most $\lambda = T^{-1/3}$ per agent. Finally, there is a small probability $\delta \leq 2T^{-2}$ that our estimates

are incorrect which contributes at most T^{-1} to the expected loss per agent. Summing the three sources of loss yields the above theorem.

6 Concluding Remarks: Monetary Transfers

We have focused on mechanisms in which the planner is not allowed to use monetary transfers. An interesting extension is to consider the case where the planner can use cash to provide incentives to agents to explore. It is straightforward that in our setup the planner will exercise this option only with the second agent and leave the mechanism intact for all other agents. Thus, if the planner has a large enough budget, then he can obtain the first-best by convincing the second agent (or even the first agent) to explore whenever this is required by the first-best. Otherwise, then all the planner's resources should be used to increase the set I^2 in which agent 2 explores. This also holds in the more realistic case where the budget is raised through taxation and taxation distorts efficiency.

References

[1] Anderson C. (2012): "The Impact of Social Media on Lodging Performance," Cornell Hospitality Report Vol. 12, No. 15. Cornell University.

[2] Bikhchandani S., D. Hirshleifer and I. Welch (1992): "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades," *Journal of Political Economy*, 100, No. 5.

[3] Bolton P. and C. Harris (1999): "Strategic Experimentation,", *Econometrica*, 67, 349–374.

[4] Ely J. A. Frankel and E. Kamenica (2013) "Suspense and Surprise" working paper.

[5] Gershkov A. and B. Szentes (2009): "Optimal Voting Schemes with Costly Information Acquisition," *Journal of Economic Theory*, 144, 36–68.

[6] Ingrid J. and C. Carter (2011): "In TripAdvisor We Trust: Rankings, Calculative Regimes and Abstract Systems," *Accounting, Organizations and Society,* 36, 293–309.

[7] Gustavo and Manso (2012): "Motivating Innovation," *Journal of Finance* (forthcoming).

[8] Kamenica E. and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review,* 101, 2590–2615.

[9] Keller G., S. Rady and M. W. Cripps (2005): "Strategic Experimentation with Exponential Bandits," *Econometrica*, 73, 39–68.

[10] Martimort, D. and S. Aggey (2006): "Continuity in Mechanism Design without Transfers," *Economics Letters*, 93, 182–189.

[11] Myerson R. (1986): "Multistage Games with Communication," *Econometrica* 54, 323-358.

[12] Rayo L. and I. Segal (2010) "Optimal Information Disclosure," *Journal of Political Economy,* 118(5), 949–987.

[13] Rothschild M. (1974): "A Two-Armed Bandit Theory of Market Pricing," *Journal of Economic Theory*, 9, 185–202.

[14] Skrzypacz A.and J. Horner (2012): "Selling Information" working paper.

A Appendix

Detailed calculation of the example: When calculating the benefit from choosing the second alternative, agent three considers two cases:

I: $R_1 \leq 1, R_2 > R_1$: In this case the third agent is certain that the second alternative has already been tested by the second agent and was found to be optimal; this implies that $R_2 > -1$. When computing the expected gain conditional on this event, one can divide it into two sub-cases: I_a : $R_2 > 1, I_b : R_2 \in [-1, 1]$. The probability of these two events (conditional on case I) are

$$\begin{aligned}\Pr(I_a|I) &= \frac{\Pr(R_2 > 1, R_1 \leq 1, R_2 > R_1)}{\Pr(R_2 > 1, R_1 \leq 1, R_2 > R_1) + \Pr(R_2 \in [-1, 1], R_1 \leq 1, R_2 > R_1)} \\ &= \frac{0.4 * 1/3}{0.4 * 1/3 + 0.2 * 1/3 * 1/2} = 0.8\end{aligned}$$

$$\Pr(I_b|I) = 1 - \Pr(I_a|I) = 0.2.$$

The gain conditional on (I_a) is: $E(R_2 - R_1|I_a) = E(R_2|R_2 > 1) - E(R_1|R_1 < 1) = 3 - 0 = 3$. The gain conditional on I_b is $E(R_2 - R_1|I_b) = E(R_2 - R_1|R_1, R_2 \in [-1, 1], R_2 > R_1) = 2/3$. Hence, the gain conditional on I is given by:

$$E(R_2 - R_1|I) = \frac{0.8 * 3 + 0.2 * 2/3}{0.8 + 0.2} = \frac{38}{15}.$$

The relative gain from following the recommendation when we multiply by the probability of I is

$$\Pr(I) * E(R_2 - R_1|I) = \frac{2}{2 + x} * \frac{38}{15}.$$

$II : 1 < R_1 \leq 1 + x$: Conditional on this case our agent is the first to test

the second alternative. The expected loss conditional on this event is

$$E(R_1 - R_2|II) = E\left[R_1|R_1 \in [1, 1+x])\right] - E(R_2) = \frac{1+(1+x)}{2} - 0 = \frac{2+x}{2}\ .$$

When we multiply this by the probability of this event we get

$$\Pr(II) * E(R_2 - R_1|II) = \frac{x}{2+x} * \frac{2+x}{2} = \frac{x}{2}.$$

Equating the gain and the loss yields $x = 2.23$. This implies that if the second action is recommended to agent $t = 3$ when $I : R_1 \leq 1$ and the planner has learned that the second action is optimal or when $II : 1 < R_1 \leq 3.23$, then agent $t = 3$ will follow the recommendation.

Proof of Theorem 8: Given our characterization it is sufficient to focus on the case where $T = \infty$. Consider the summation of the RHS in (4):

$$\sum_{t=2}^{\infty} \int_{R_1=i_{t,\infty}}^{i_{t+1,\infty}} [R_1 - \mu_2]d\pi = \lim_{t\to\infty} \int_{R_1=i_{2,\infty}}^{i_{t,\infty}} [R_1 - \mu_2]d\pi.$$

Since $\int_{R_1=-\infty}^{i_{2,\infty}} [R_1 - \mu_2]d\pi = 0$ and since $\int_{R_1\leq x} [R_1 - \mu_2]d\pi$ is increasing in x we conclude that

$$\sum_{t=2}^{\infty} \int_{R_1=i_{t,\infty}}^{i_{t+1,\infty}} [R_1 - \mu_2]d\pi \leq \lim_{x\to\infty} \int_{R_1\leq x} [R_1 - \mu_2]d\pi = \mu_1 - \mu_2.$$

Looking at the summation of the LHS

$$\sum_{t=2}^{\infty} \int_{R_1\leq i_t, R_2>R_1} [R_2 - R_1]d\pi$$

we note that $\int_{R_1\leq x, R_2>R_1} [R_2 - R_1]d\pi$ is increasing in x. The fact that i_t is

increasing in t implies that if we let

$$\alpha \equiv \int\limits_{R_1 \leq i_2, R_2 > R_1} [R_2 - R_1] d\pi$$

we then have

$$\alpha \leq \int\limits_{R_1 \leq i_t, R_2 > R_1} [R_2 - R_1] d\pi.$$

Hence, this sum can be bounded from below by $t^* \alpha$, which implies the claim. □

B Proof of Theorem 13

Let r_1 and r_2 be the realized values of R_1 and R_2, respectively. First, assume that we have that $|r_1 - r_2| \leq 2T^{-1/3}$. Since $\max\{r_1, r_2\} \leq \min\{r_1, r_2\} + 2T^{-1/3}$, then the average reward of any policy of selecting the actions would be at most $2T^{-1/3}$ from the reward of always selecting the best action.

Second, assume that we have that $|r_1 - r_2| > 2T^{-1/3}$. Using the Hoeffding inequality, we have that $|\hat{\mu}_i - E[R_i]| \geq \lambda$ with probability at most $\delta \leq 2e^{-2\lambda^2 m}$. For $\lambda = T^{-1/3}$, since $m = T^{2/3} \ln T$, we have $\delta \leq 2T^{-2}$. Assume that the event that $|\hat{\mu}_i - E[R_i]| < \lambda$ for both a_1 and a_2 holds (this happens with probability at least $1 - 2\delta \geq 1 - 4T^{-2}$). Given that the event holds, we have that $\hat{\mu}_1 > \hat{\mu}_2$ iff $\mu_1 > \mu_2$, namely, a_{best} is the optimal action. In this case we have a loss do to exploring the worse action. If the worse action is a_2, then we have a loss of m (one block) and therefore the average loss, per action, compared to always performing the optimal action is at most $m/T = T^{-1/3} \ln T$.

If the worse action is action a_1, then we claim that we explore action a_2 after only a finite number of β blocks. Note that $Gain(\theta)$ is an increasing function in θ. Since $\theta_{2,\infty}$ is at least μ_2, we can claim that $Gain(\theta_{i,\infty}) \geq Gain(\mu_2)$ for any $i \geq 2$. Recall that $Gain(\theta_{i-1,\infty}) = Loss(\theta_{i-1,\infty}, \theta_{i,\infty})$, and note that $\sum_i Loss(\theta_{i-1,\infty}, \theta_{i,\infty}) = E[R_1 - R_2]$. This implies that the number of is needed is at most $\lceil E[R_1 - R_2]/Gain(\mu_2) \rceil \geq \beta$. Note that β is a constant that depends only on the distributions D_1 and D_2 and is independent from T. In this case the loss would be at most $\beta m/T$.

The third case is when our assumption does not hold, namely, either $|\hat{\mu}_1 - \mu_1| > T^{-1/3}$ or $|\hat{\mu}_2 - \mu_2| > T^{-1/3}$. This occurs with probability at most $4T^{-2}$, and therefore adds to the loss of each agent, in expectation, at most $4T^{-2}$.

Therefore the difference between using always the better action, and using

the mechanism S is at most

$$\max\{T^{-1/3}\ln T, 4T^{-2} + T^{-1/3}\ln T, 4T^{-2} + \beta T^{-1/3}\ln T\} \leq CT^{-1/3}\ln T$$

for some constant $C > 0$. $\square$