Eliciting Decision Advice with Good Incentives and Decision-Maker Freedom

Della Penna and Balduzzi

Friday 29th March, 2019

1. INTRODUCTION

This chapter considers a subject facing a decision, who wishes to incentivize multiple experts in providing advice so as to pick a decision that maximizes the rewards the subject receives, while maintaining the freedom of the subject. Experts do not have intrinsic interest in the action the subject chooses or actually takes, nor do they face any costs in acquiring their signals.

Preserves subjects freedom is a inherently desirable practical design criterion as it enables no-regret exploration on the part of the subjects. Subject remains at all point in control of the decision, and so trying the mechanism cannot reduce their choices or their expected welfare.[1].

The main contribution of this chapter is casting this setting as one of a single unit efficient allocation with interdependent valuations ([**??????**]). This allows us to leverage existing results in that setting, which provide both necessary conditions for efficient allocations (single crossing conditions on the signal structure) and impossibility results when those are not met.

The key conceptual contribution that allows this is to consider the allocation not over the decisions, but instead over the experts providing the advice. The item to be allocated is the right to observe the signal reports, provide the advice and a (linear) share of the reward obtained by the subject. We term these mechanisms *advice auctions*. The term advice is chosen to highlight that the decision is not ultimately determined by the market, thus preserving the subjects freedom. The term auction highlights that this procedure does not produce a sequence of prices through time. It is this simultaneous nature that allows us to side-step the negative (form the perspective of freedom) results that constraint sequential mechanism to having full support over actions in order to provide incentives.

Sharing rewards after choosing what decision to advice is neither a pure private value (since the optimal choice conditional on information makes the value the same for everyone) nor pure common value (since the ability the select the optimal choice given access to the other signals might vary across agents).

A advantage of this setting to having a mechanism that directly outputs a chosen action is that it allows for the expert making the recommendation to have different influences on subject (i.e. some experts may be more persuasive). If the mechanism output was a choice to advice the subject directly, we would have to restrict the expected reward the subject receives not to depend on which expert submitted which report, this can be seen as a limit on subject freedom (since it it is an external constraint). Further, by having the mechanism select the expert instead of the choice, the natural extension to the more practical mechanisms that do not require the mecha-

[1] One can set a reserve price of the advice auction to the expected value of the action the agent would have picked if they did not participate in the mechanism. As long as the reserve price is set apriori or alternative reports form the non-winning bidders are the only thing used to set it, the incentives do not change from those analized here.

nism to have access to the valuation functions, but instead rely on the highest bidder to be able to aggregate signals effectively.

1.1. Limits to Subject Freedom in Sequential Proper Scoring Rule Based Decision Markets

One way to incentivize them is by applying the machinery of prediction markets based on sequentially shared proper scoring rules to the expected reward conditional on the action. A challenge that presents itself is how to settle the markets for the reward conditional on the action which is not taken. One natural approach is to void the trades in the markets for these actions, this being the originally proposed mechanism in this line of work [Hanson 2002], and only settling the markets where actions are taken. While seemingly natural, this is not incentive compatible for the experts, even in the weak myopic sense, as shown in [Othman and Sandholm 2010].

To understand why this is the case, consider a last trader facing the prediction market (sequential proper scoring rule) where the price is correct (matches the expected reward) for the optimal action but there is some other action that is misprinted. The profit maximizing move for this trader is to lower the price of the optimal action bellow the true price of the previously misprinted action, and correct the mispriced action to its true price. The utility maximizing subject would then carry out the suboptimal action, the expert would be rewarded for correctly predicting it and would receive no punishment for the error they introduced into the reward of the optimal action. The mechanism proposed in [Hanson 2002] is not BNIC for the experts who provide advice, as witnessed by the example above (and shown in [Chen et al. 2014; Othman and Sandholm 2010]). More generally, any sequential proper scoring rule based mechanism that is incentive compatible for the experts is incompatible with maintaining the subject's freedom to select the action that appears optimal ex-post ([Chen et al. 2014]).

1.2. Summary and Outline

The rest of the chapter is structured as follows. We first introduce a formal model and notation. We then present advice auctions as a direct mechanism, show when they are truthful and their limits. We then consider two practical indirect variations of the procedure, which removes the need for the mechanism to have any knowledge of valuations, and a consider sufficient conditions for their efficiency and truthfulness.

2. MODEL

As before, the subject seeks advice on a decision they will take form some finite set of alternatives A, let c be the choice that is given as advice to the subject and a the decision that the subject actually takes. The rest of our model and notation largely follow that of [**?**].Each expert $i \in \{1, \dots, n\}$ receives a single signal s_i which is known only to expert i. Potential signals for bidder $1 \leq i \leq n$ form a discrete signal space S_i. Let $\vec{s} = (s_1, s_2, \dots, s_n)$ be a signal profile. The reward r that the subject receives depends on their chosen action a and the underlying state of the world as determined by the signal profile $\vec{s}$. Let $\vec{s}_{-i}$ denote all signals but s_i, and let $(s'_i, \vec{s}_{-i})$ denote the profile $\vec{s}$ but where s_i has been replaced with s'_i. Since r does not depend on the choice of c by the expert, and that at the point the mechanism is run a has not been selected, we use the following reduced form value function for the value of the rights bundle to expert i: $v_i : \times_i S_i \to \mathbb{R}_{\geq 0}$, which maps every signal profile to the expected value of linear share of the reward αr.

Each expert reports a signal b_i, and the vector of reported signals is $\vec{b} = (b_1, b_2, \dots, b_n)$. Each possible signal profile $\vec{s}$ corresponds to a underlying state of the world, this includes both inherent physical properties of the subject and the actions, as well as the subjects probability of choice of a in response to different choices of c.

The valuation functions for all bidders i are monotone non-decreasing in every signal s_j for all j.

Mechanisms are a pair (x, p), where x is a set of allocation functions $x = \{x_1(\vec{b}), \ldots, x_n(\vec{b})\}$ satisfying $\sum_i x_i(\vec{s}) \leq 1$ for all possible $\vec{b}$, and p a set of payment functions $p = \{p_1(\vec{b}), \ldots, p_n(\vec{b})\}$. An allocation function $x_i : \times_j S_j \to [0, 1]$ maps every bid profile $\vec{b}$ to the probability that expert i gets allocated. A payment rule $p_i : \times_j S_j \to$ maps the reported signals $\vec{b}$ to the expected payment from bidder i. Experts are risk neutral, so their expected utility is quasilinear, given in the reduced form by $x_i(\vec{b}) \cdot v_i(\vec{s}) - p_i(\vec{b})$ where $\vec{s}$ is the true signal profile of the experts.

$$x_i(\vec{s}) \cdot v_i(\vec{s}) - p_i(\vec{s}) \geq x_i(b_i, \vec{s}_{-i}) \cdot v_i(\vec{s}) - p_i(b_i, \vec{s}_{-i}) \qquad \forall \vec{s} \in \times_j S_j, b_i \in S_i. \qquad \text{[IC]}$$

3. A DIRECT REWARD SHARING MECHANISM

The simplest class of mechanisms to incentivize advice is based on sharing a fraction of the rewards with the experts. For the single expert case this is mentioned by [Othman and Sandholm 2010]. Here the idea is extended to the multiple experts case. We do this by instantiating our notion of advice auctions with a simple mechanism, the generalized VCG mechanism proposed by [**?**]. This mechanism is *direct* in the standard sense that agents report their signals.

The core of the mechanism is simple. Since there is knowledge by the mechanism over the value function for a given vector of signals, it can use the reported signals to select the highest value expert. The net payment to that expert is then just his share of the reward minus his value at the lowest signal he could have misreported and still obtained the allocation give the other reports. More formally:

MECHANISM 1. *[Direct Reward Share VCG] Then mechanism gives the rights bundle to the expert $i*$ with the highest valuation under the reported signals. That is, the allocation rule is:*

$$x(\vec{s}) = i \qquad \textit{when} \qquad x_j(\vec{s}) = \begin{cases} 1 & \textit{if } j = i \\ 0 & \textit{otherwise.} \end{cases}$$

This lets the expert $i*$ observe $\vec{b}$ and then select c. Then subject observes c and $\vec{b}$, takes their action a and receives reward r, which the mechanism observes.

The non selected experts receive no payment, while the selected expert $i*$ receives their share α of the reward r minus the value of the share of the reward at the lowest b'_i (the critical signal) that would have still resulted in expert i being selected. More formally, given signals for agents $\neq i$, $\vec{s}_{-i}$, the critical signal for i is: if there exists some b_i such that $x_i(s_{-i}, b_i) = 1$ then set $b^*_i = \min_{b_i} x_i(s_{-i}, b_i) = 1$, otherwise there is no critical signal for i. Thus, the payment rule is:

$p_i = \alpha r - v_i(\vec{s}_{-i}, b^*_i)$

An allocation function x_i is called *deterministic* if $x_i(\vec{b}) \in \{0, 1\}$ for all i and all $\vec{b}$. The generalized direct VCG mechanism is deterministic and prior-free.

3.1. Truthfulness with Single Crossing Signals

DEFINITION 1 (MONOTONICITY). *An allocation function x_i is said to be monotone if for every $\vec{b}_{-i}$, $x_i(\vec{b}_{-i}, b_i)$ is monotone non-decreasing in the signal b_i.*

Truthful mechanisms can be characterized as follows [**?**].

PROPOSITION 1. *Monotonicity is a necessary and sufficient condition for allocation functions x to be* implementable*, i.e., there exist payment functions p such that the mechanism (x, p) is truthful. Moreover, an analogue of Myerson's payment identity holds, so the payment is uniquely determined by the allocation function.*

It follows that constructing a truthful mechanism is equivalent to constructing a monotone allocation function. For deterministic truthful mechanisms, the payment identity of **?** implies the following about the cost charged to chosen expert (from [**?**]).

PROPOSITION 2. *Let agent i be the allocated winner at report profile $\vec{s}$ in a deterministic truthful mechanism. Then their cost is their value at the critical report.*

A single-crossing condition captures the idea that bidder i's signal has a greater effect on experts i's value than on any other experts's value. We follow the definition in [**?**]:

For $s_i = 1, \ldots, k_i$, define $\frac{\partial v_j(s_i, \vec{s}_{-i})}{\partial s_i} = v_j(s_i, \vec{s}_{-i}) - v_j(s_i - 1, \vec{s}_{-i})$.

DEFINITION 2 (SINGLE-CROSSING). *A valuation profile is said to satisfy the single-crossing condition if for every expert i, for any set of other experts signals $\vec{s}_{-i}$, and for every expert j,*

$$\frac{\partial v_i(s_i, \vec{s}_{-i})}{\partial s_i} \geq \frac{\partial v_j(s_i, \vec{s}_{-i})}{\partial s_i}.$$

THEOREM 3. *There is a truthful and efficient ex-post Nash equilibrium of the DRSA mechanism when signals satisfy the single crossing property.*

PROOF. Allocating to the bidder with the highest value is a monotone allocation rule, and therefore, according to Proposition 1 it is implementable. The cost for the rights bundle of the chosen expert is then just their value at their critical signal, which is the corresponding payment. □

Further, one cannot do better than this as per Proposition 1, monotonicity of the allocation rule is necessary for a efficient and truthful mechanism with interdepedent values. Hence, without single-crossing, it is impossible to have a truthful advice auctions in general.

This procedure for a direct advice elicitation mechanism based on the advice auction procedure was here instantiated using [**?**] as the underlying auction mechanism. But the procedure is generic. It could be, for example, instantiated instead with the randomized mechanism of [**?**], and would obtain the approximation properties that algorithm provides in auctions in our advice setting.

4. PRACTICAL MECHANISMS: ADVICE AUCTIONS

The assumptions used in [**?**] that are used for the above result are extremely minimal relative to the existing literature in most of decision markets and auctions with interdependent values.

However, the mechanism having access to the value functions seems highly impractical in most potential applications. Given the practical settings that motivate this work, we do not assume access by the mechanism to a the valuation functions, and consider two practical alternatives.

4.1. A Bid and Signal Reward Sharing Mechanism

The first modification one can do to make the payment of the highest bidder depend on the bid of the second highest, this removes the dependency of the payments function on the

The challenge this faces is that the signals reports being submitted would not matter, so any signal report is in equilibrium. To correct this and create strict incentives again, we can add a further payment received by all experts that is also a linear share of the reward. This makes the truthful signaling equilibrium potentially strict.

MECHANISM 2. *[Allocation with Bids, Reward Share with Signals (ABRSS)] Experts report both a bid and a signal, we slightly abuse notation and use $vec(s)$ to denote the reported signals, note their only use is to be displayed to the expert allocated to make the choice.. Then mechanism gives the rights bundle to the expert $i*$ with the highest bid:*

$$x(\vec{b}) = i \qquad \textit{when} \qquad x_j(\vec{b}) = \begin{cases} 1 & \textit{if } j = i \\ 0 & \textit{otherwise.} \end{cases}$$

This lets the expert $i*$ observe reported signals $\vec{s}$ and then select c. Then subject observes c and $\vec{b}$ and $\vec{s}$, takes their action a and receives reward r, which the mechanism observes.

The non selected experts receive payment βr, while the selected expert $i*$ receives their shares $(\alpha + \beta)r$ of the reward minus the second highest bid, b_{i^*-1}.

Thus, the payment rule is:

$p_i^* = (\alpha + \beta)r-, b_{i^*-1})$ and $p_{j \neq i*} = \beta r$

For this allocation rule to reach efficient allocation achieved by the direct mechanism, a stronger assumption on signal structure is needed. If we only require single crossing then the identity of the highest valuation expert can depend on interaction of their signals and those of other agents. Without access to the value function the mechanism cannot in general hope to achieve this allocation. Thus beyond single crossing, we also require that the highest valuation expert must be so for any possible set of signals other than their own. In this case the allocation of this mechanism is efficient, since it coincides with the direct mechanisms allocation.

DEFINITION 3 (SINGLE SIGNAL MAX VALUE). *A valuation profile is said to satisfy the single-signal max value condition if highest value expert $i*$ knows he is the highest value when given their signal, and for any set of other experts signals $\vec{s}_{-i}$, and for every expert j,*

$$v_i(s_i, \vec{s}_{-i}) \geq v_j(s_i, \vec{s}_{-i})$$

Note this property is not as strong as it may at first seem. If there is a expert who has the trust of the subject, understands what they find persuasive, and is sufficiently competent at evaluating the reported signals of others, it may be best able to select c to maximize r no matter what the state of the world. Being able to see other experts signals may however substantially raise the reward and hence the value of the highest valued expert.

THEOREM 4. *There is a truthful and efficient ex-post Nash equilibrium of the ABRSS mechanism when signals satisfy the single signal max value property.*

PROOF. The highest value bidder bids his worst possible value given the other bids or anything higher (knowing he will win and be charged according to the second highest price makes him indifferent between these when others are truthful. Allocating to the bidder with the highest value is a monotone allocation rule, and therefore, according to Proposition 1 it is implementable. The cost for the rights bundle of the chosen expert is then the second highest bid, which is the corresponding payment. □

For contrast consider the identical mechanism but without the experts submitting their signals and receiving share β.

MECHANISM 3. *[Bid Only Advice Auction] Each expert observes their signal and then report only a bid b_i. The mechanism gives the rights bundle to the expert $i*$ with the highest bid:*

$$x(\vec{b}) = i \qquad \textit{when} \qquad x_j(\vec{b}) = \begin{cases} 1 & \textit{if } j = i \\ 0 & \textit{otherwise.} \end{cases}$$

This lets the expert $i*$ observe reported bids $\vec{b}$ and then select c. Then the subject observes c and $\vec{b}$, takes their action a and receives reward r, which the mechanism observes.

The non-selected experts receive payment βr, while the selected expert $i*$ receives their shares $(\alpha + \beta)r$ of the reward minus the second highest bid, b_{i^*-1}.

Thus, the payment rule is:
$p_i^* = (\alpha + \beta)r-, b_{i^*-1})$ *and* $p_{j \neq i*} = \beta r$

For some very limited information structures the bidding mechanism still aggregates information efficiently. For example, if each expert knows the expected outcome conditional on one action don't know what happens under other actions, and there is at least one expert who is informed about each action. These information structures correspond to the standard private values settings in the original VCG mechanism [**?**].

This kind of indirect mechanism is inherently limited when the value from the signals of experts are interelated (as we proved before). Their bid, cannot encode the full information contained in the signals, and thus this limits any mechanism that relies solely on single rounds of bids to aggregate information.

To illustrate this consider a setting that satisfies the single max bidder assumption. Denote by i^* the experts whose valuation is higher than all other experts in every state of the world and who knows the subject trust them completely so if he wins $c = a$. Have a second expert whose value is 0 in all state of the world[2] but has a binary independent signal s_treat that determines which choice is best for the subject, without knowing s_treat the two best choices have equal expected reward $r_uniform$, while knowing the value of the signal allows to select the appropriate choice and results in double the reward $2r_uniform$. Since his value is always is always 0 and so is his bid in equilibrium. There is no prior-free way for the unpersuasive expert to encode his signal into his bid (even through he is incentivize to reveal his signal to get a higher $\beta payment).Noticethatthereisnolimitinthesizeofthegapintherewardsbetweenthetwoallocationsingeneral, asinthee$

5. CONCLUSION

This chapter shows how to use the bundle of rights perspective on advisers to recast the incentivizes for decision elicitation from multiple experts into the thriving literature on VCG with interdependent valuations. We then use Proposition 1 of [**?**] to show that using the generalized VCG mechanism of [**?**] to allocate the rights results in a direct incentive compatible and efficient mechanism when signals have a single crossing property. We then explore two practical variations of the mechanism that relax the assumption the mechanism can access the value functions and that signals can be transmitted between experts. We introduce a notion that shows when mechanisms that

[2] For example the expert might know the subject is biased and refuses to hear the expert due to color of their skin.

5.1. Introduction

This chapter differs from the previous ones in it's relationship to the thesis. While previous we seek to extend the understanding of previously presented settings (bandit algorithms and decision markets), this chapter seeks to introduce a new setting that contains these two.

Our motivating applications in medicine are suggestive of a sequence of similar decisions faced by a sequence of agents in order, all of whom face an individual choice on their own course of action. Every day new patients perceive their symptoms, and they seek diagnoses and treatments from medical providers. A corporation faces new investment opportunities with regularity, and similar opportunities appear to many firms that might not be competitive (for example vertical divisions in a conglomerate might be offered similar projects to automate part of their workflows and must choose to attempt them or not, and might be able to ask its inhouse experts for advice on the right course of action. Scenarios such as these that motivate optimal decision elicitation are in a sense naturally cast not as one-shot interactions, but as repeated games with many experts and a sequence of subjects who seek advice before making a decision which only affects them.

This combines the central aspects of bandits with compliance awareness (a sequence of choices and learning from past experience, where the actions of subjects are not bound to follow the algorithm choice) as well as elicitation of information from experts to enable optimal decisions without the advice being binding.

The study of decision markets so far, including the previous chapter, have focused on a setting with a single decision and multiple advisers ([Chen et al. 2014; Hanson 2002; Othman and Sandholm 2010]). This chapter poses a novel and natural generalization of this setting that also captures the compliance aware bandit setting and the advice auctions as special cases. A sequence of T subjects (patients in the medical motivation), and a fixed set of K advisersors (experts) with access to signals about different patients expected rewards r under different advice c and actual courses of action a. Bounded regret algorithms with compliance awareness can be seen as addressing the special case where the experts' signals are known apriori to be uninformative, so $K = 0$ effectively, and thus only the experience can be learned from. Our one-subject mechanism in chapter 5 is the special case for $T = 1$, thus there is no role for exploration or learning from experience, since there are no future decisions to help inform. A situation where experts always report their signals truthfully and have no knowledge over how to aggregate them beyond that possessed by the mechanism is equivalent to a compliance-aware contextual bandit problem. When contexts are constant across all time steps, the situation further reduces to a bandit problem with compliance awareness. When the subject always follows the mechanism or a cannot be observed, it reduces further to the standard multi-armed bandit problem.

In contrast to the previous chapter's motivation in the literature, in this chapter our focus is first an foremost on constructing a practical mechanism. The motivation for this switch is that the setting is natural, and no mechanisms (or the setting itself) have been previously proposed to the best of our knowledge.

The most conceptually interesting possibility when moving to a sequence of T agents is that by pooling the risk of having with their action across agents, it can be ex-post incentive compatible to take the exploratory actions for subjects, by linking them to suitably large transfers. The size of the payoff required to make the choosing agent change actions provide a estimate of its ex-post value on the actions.

We build up to the main practical design by analyzing two simplified models that illustrate the two key characteristics of our mechanism. First, the need for incentives to motivate exploratory choices. For this, the rewards from the choice of action must be linked not just to the reward during the period the action is taken, but to the full sequence of subsequent future rewards. Second, to aggregate signals when single crossing conditions and their approximations are violated, we propose using an off-line contextual bandit algorithm to evaluate the counter-factual (marginal) value of the signals each expert provides. We present a mechanism that combines both ideas, and explore some of its limitations.

6. MODEL

The game occurs over T steps, at each step:

(1) A new subject t arrives and each i of K experts receives a signal $s_{t,i}$ for that subject. The mechanism randomly allocates a exploration transfer of l to one of th e actions available to the subject, we denote this action by l_t.
(2) Each expert i reports to the mechanism $b_{t,i}$, and after all reports are received the mechanism selects a expert i^*
(3) The subject observes c_t, a'_t and $b_{t,i}$, picks an action a_t, and receives a reward r_t.
(4) The mechanism provides feedback about s_t, c_t, a_t, and r_t to experts.

At the end of the final period the mechanism makes payments to the experts p_i.

6.1. Subjects' Beliefs and Incentives

The previous work on incentive compatible bandits [Kremer et al. 2014; Mansour et al. 2015] has shown that there is a distribution of rewards if all agents were rational and this common knowledge, then some actions can never be explored (assuming only information revelation and no transfers can be used by the mechanism). Actions that apriori have lower expected rewards than all others no matter what is revealed by previous instances of other actions cannot be explored. The logic behind this is that knowing no previous signal could persuade an agent to take the action, an agent told to take the action knows that in expectation they can do better otherwise. That literature has largely been focused on finding information revelation strategies that are optimal, subject to the incentive constraints. Our lottery payments us to side step these impossibility results, by providing a reason for exploration for a individual subject.

7. A SEQUENCE OF REPEATED ONE-SHOT-EFFICIENT MECHANISMS IS INEFFICIENT

Even when signal structures satisfy the single crosing property running the direct mechanism of chapter 5 repeatedly, once for each subject, results in choosing the arm with the maximum posterior expected reward at each step t and using the payment rule:

$$\pi_i = \sum_{1}^{T} \begin{cases} \alpha(r - \mathbb{E}[\hat{r}_{t,-i}]), & \text{if } \hat{c}_{t,-i} \neq c_t \\ 0, & \text{otherwise} \end{cases}$$

The repeated use of single-subject-efficient mechanisms thus creates incentives for a greedy policy in the presence of multiple experts. This is immediate from the definition of the single subject direct mechanism: it selects the arm that maximizes the rewards for that period given the reports. If the reports are truthful this is the highest expected reward arm on that period.

EXAMPLE 1 (TWO SIGNALS WITH TWO REGIMES). *We consider 2 experts and 3 arms with T sequential subjects. The first arm is a safe arm with no variance and a known reward of $1/2$. The other arms have a-priori a lower expected value, of $1/3$, but conditional on both agents' signals, one arm has an expected value of $2/3$ and the other* *of* 0*. Each agent receives a binary signal. The optimal arm is the parity (XOR) of both agents signals.*

In this example the greedy policy always plays the safe arm and has an expected regret of $(2/3 - 1/2)T$ relative to the optimal (over all signals) contextual policy in hindsight. Note that the optimal policy with exploration only requires 1 exploration step to identify the mapping to the best arms, thus the regret of the mechanism choice relative to the optimal policy with exploration is $(2/3 - 1/2)(T - 1) - (1/2 - 1/3)$.

Note that the example weakly satisfies a single crossing signal structure on a single round, since experts values are unchanged by their signals.

DEFINITION 4 (FULL DISCLOSURE). *We say a decision elicitation mechanism has* full disclosure *if all experts receive feedback about the value of c_t, a_t, and r_t in every period.*

Under full disclosure, a repeated direct reward sharing mechanism (DRSM) in Example 1, there is a NE of the repeated single subject mechanism which results in the greedy policy.

Given that there is no winner's curse due to the signal structure[3], both agents bid their valuations. If the winner of the auction does not choose the safe arm, and instead explores in that period, they receive a lower payoff in expectation in that period. In future periods their bid, and by symmetry and under full disclosure the other agents' bids, are higher, since they can both now deduce the higher payoff arm and that is their new expected value. Thus given the second price mechanism their payoffs are no higher in later periods. Thus exploration is not in equilibrium.

On possible attempt to fix this would be to only reveal the outcome to the winning bidder, thus allowing them to internalize the informational advantage in future rounds payoffs, in other words by not having full disclosure. This internalizes the benefits of explorations, yet it prevents the other experts from learning in those rounds when they do not win, severely limiting the situations in which the mechanism can be efficient.

8. A SIMPLE BIDDING MECHANISM WITH EXPLORATION

To overcome the exploration limitation of the repeated one shot mechanism, a mechanism must internalize for the decision making expert the informational benefits of exploration steps on the rewards of future periods. This naturally motivates a mechanism that generalizes the expert bidding mechanism, by providing the expert with rewards proportional to all future periods when it wins the auction.

MECHANISM 4 (BIDDING FOR OWNERSHIP OF CHOICE MECHANISM (BOCM)). *An expert i is the* owner *at a given time period t if they have won the last auction that had a winner (if no bids in a auction meet the reserve price the owner remains unchanged). Denote by $o_{i,t}$ an indicator variable encoding with a value of 1 if the agent i was the* owner *of the choice at time t.*

$$\pi_i = \sum_1^T \begin{cases} \alpha r_t, & \textit{if } o_{i,t} = 1 \\ 0, & \textit{otherwise} \end{cases} + \sum_1^T \begin{cases} -b_{\hat{2},t}, & \textit{if } o_{i,t} = 0 \land o_{i,t+1} = 1 \\ b_{\hat{2},t}, & \textit{if } o_{i,t} = 1 \land o_{i,t+1} = 0 \\ 0, & \textit{otherwise} \end{cases}$$

The first part of the payments sums over the rewards for all periods during which an agent owns the rights. The second part determines the payments when a new agent i

[3] that is, the winner of the auction who bids their value without conditioning that value on having won the auction (which implies having the highest signal) gets the same payoff as if they do condition.

becomes the owner; they pay out the second highest bid of that period. When another agent takes over them as the owner, they are paid the second highest bid in that period. Note that the reserve price can be encoded in the owner's bid in this notation, since when it wins there is no change in owner and no further payments are made. This linking of payments addresses the incentive problem by internalizing the positive inter-temporal information externality created by selecting actions that have not previously been selected.

PROPOSITION 5. *There is a expost NE under which the BOCM results in sublinear regret in Example 1.*

The optimal contextual policy with exploration has payoff of $2/3T(T-1)+1/3$

PROOF. The optimal contextual policy with exploration has payoff of $2/3T(T-1)+1/3$. The value of the choice for an agent who controls the full sequence and observes the full set of signals is thus $\alpha(2/3T(T-1)+1/3)$, and given the second price mechanism this can be their initial bid in a NE. The agent explores in the first choice, and exploits in all subsequent choices. If the agent does not explore in the first choice they obtain a lower payoff. If the agent makes a lower bid they do not improve their payoff since they never win. □

9. CHOICE INCENTIVE LOTTERIES; USING TRANSFERABLE UTILITY AS A SOURCE OF UNBIASED VARIATION

MECHANISM 5 (LOTTERY FOR EXPLORATORY CHOICE (LEC) MECHANISM). *Inputs:*

At the start of the game before the first subject a vector of payments Γ is chosen. In each time period t a new subject arrives and agents receive their signals s_t and then send their reports $\hat{s_{t,i}}$. A one-shot encoding of the reports is used as context in A to select an arm c_t which lead to choice a_t is made and then reward observed r_t. At the end of the last time period, for each expert i, estimate the loss that would be obtained by the contextual bandit algorithm without using that expert's report in its context: denote it $E(\hat{s_{-i}}, A)$.

The payment rule for each expert i is as follows:

$$\pi_i = \alpha(\sum_{1}^{T} r_t - E(\hat{s_{-i}}, A))$$

The payment rule for each subject t is as follows:

$$\pi_t = \Gamma_{t,a}))$$

The key observation is that by making Γ have payments that are sufficiently large in magnitude, it can encourage Since the payments are completely exogenous to the signals and preferences, they are a ideal instrumental variable, which can be used to estimate the rewards of different underlying actions. This avoids the problem of needing to force subjects to take the proposed action of the mechanism we had in th contextual bandit driven policy, while still providing a way of estimating the full counterfactual.

10. A BID AND SIGNAL MECHANISM WITHOUT PRIORS

The above signal-only mechanism can be potentially inefficient when there are experts who know how to map the signals to actions, and thus can help the subjects avoid some of the regret in the learning. More broadly, experts can have additional information relative to the mechanisms that helps them aggregate the signals better but requires signals by other experts to be reported to them.

It is worth emphasizing the crucial role played in the reward function by the unbiased nature of the estimator. Alternatively to the contextual bandit, when exploration is not required or compliance not assured, the same randomness can be inserted into the mechanism through a lottery, as sketched in the previous section.

MECHANISM 6. *[] Inputs: A contextual bandit algorithm A and an unbiased offline evaluation algorithm E.*

A lottery Γ for each action and each subject is drawn, the resultant payment rule is announced. In each period: all experts report signals and bids to the mechanism, the mechanism displays the other experts' reported signals (for all previous periods) to the winner of the bidding, the winner selects the chosen action c_t, and this is displayed to the subject, who takes action a_t and receives reward r_t.

At the end of the last time period, for each expert i, estimate the loss that would be obtained by the contextual bandit algorithm without using that expert's report in its context: denote this by $E(s\hat{}_{-i}, A)$.

The payment for expert i rule is:

$$\pi_i = \alpha \sum_1^T r_t - \mathbb{E}[\sum_1^T \hat{r}_{-i,t}] + \sum_1^T \begin{cases} \beta r_t, & \textit{if } o_{i,t} = 1 \\ 0, & \textit{otherwise} \end{cases} + \sum_1^T \begin{cases} -b_{\hat{2},t}, & \textit{if } o_{i,t} = 0 \wedge o_{i,t+1} = 1 \\ b_{\hat{2},t}, & \textit{if } o_{i,t} = 1 \wedge o_{i,t+1} = 0 \\ 0, & \textit{otherwise} \end{cases}$$

Where α and β are set ex ante.

The payment rule for each subject t is as follows:

$$\pi_t = \Gamma_{t,a}))$$

The condition that must be satisfied to make the payments from the mechanism smaller than the surplus it brings collectively to the subjects is $\alpha + \beta < 1/2NT$.

The above algorithm is far from perfect. The dynamic nature of the market creates a major concern that an expert would not reveal their signal truthfully and lose out on that part of the reward if they can benefit more from being the *owner*. By withholding their signal they can suppress the bids of other experts who are thus at a disadvantage; this is a particular concern since the other experts may be able to achieve higher rewards.

Consider a setting where all experts signals are symmetric and perfect complements to each other. For example, the value of the reward depends on their product. All signals are equally valuable in the counter-factual sense used to establish rewards. To the extent the second highest bidders value is close to the first, there is almost no net expected value from being the owner. On the other hand, if a bidder does not report his signal truthfully, then the other bidders valuation for being the owner are 0, and the misreporting bidder can appropriate the full value of the $alpha$ part of the rewards. Thus α ¡ β for incentive compatibility.

Note that the choice of lottery payments Γ is restricted to those which generate full support so that the estimator of the signal rewards can be fully evaluate. If the rewards are not IID the full support induced by the lottery must be maintained throughout all time periods. Thus the mechanism is inefficient in so far as the owner who knows a priori the correct policy given signals cannot fully implement it.

It is not clear how to prove when there is a efficient full revelation mechanism for the above mechanism, since the interaction between the owners information about how to aggregate and learn over the signals complicates the dynamic VCG styles of analysis.

11. CONCLUSION

We introduced a new and natural setting, that generalizes advice auctions and compliance aware bandit problems. Building on these, we proposed a mechanism that is plausibly practical but hard to analyze. The broad setting appears very likely to be fruitful, and

REFERENCES

Yiling Chen, Ian A Kash, Michael Ruberry, and Victor Shnayder. 2014. Eliciting predictions and recommendations for decision making. *ACM Transactions on Economics and Computation* 2, 2 (2014), 6.

Robin Hanson. 2002. Decision markets. *Entrepreneurial Economics: Bright Ideas from the Dismal Science* (2002), 79–85.

Ilan Kremer, Yishay Mansour, and Motty Perry. 2014. Implementing the Wisdom of the Crowd. *Journal of Political Economy* 122, 5 (2014), 988–1012.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 297–306.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. 2015. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. ACM, 565–582.

Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. 2016. Bayesian Exploration: Incentivizing Exploration in Bayesian Games. *arXiv preprint arXiv:1602.07570* (2016).

Abraham Othman and Tuomas Sandholm. 2010. Decision rules and decision markets. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 625–632.

Vasilis Syrgkanis, Akshay Krishnamurthy, and Robert E Schapire. 2016. Efficient Algorithms for Adversarial Contextual Learning. *arXiv preprint arXiv:1602.02454* (2016).