Eliciting Decision Advice with Good Incentives and Decision-Maker Freedom

Della Penna and Balduzzi

Tuesday 26th 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. 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.

Our main practical contribution is a two step procedure, which we term *advice auctions*, that is consistent with our normative freedom requirement and effectively reduces it to a well studied question (a single unit auction) in mechanism design with interrelated valuations. The first step is a sealed bid auction for the right to both observe all reported signals and choose the action to advice, while the second step simply allows the highest bidder on the first step their choice. This closely follows work on efficiency allocation with interdependent values [**??**], and our mechanism has the same two steps: first agents submit their signals so as to be able to determine the probabilities over states of the world, and the results of that are used to pick an efficient allocation.

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.

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]). A mechanism in which the ex-post optimal action (allocation) is the one a utility optimizing subject would pick is said to be ex-post incentive compatible for the subject.

1.2. Mechanism Design with Interdepedent Values

Sharing the 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).

Our mechanism is very similar to that of [**?**], we seek to first induce agents to truthfully report their signals (in our case by sharing the reward with the agent, while in their case by having agents be paid for signal reports that agree with those of other agents).

1.3. Summary and Outline

The rest of the chapter is structured as follows. We first introduce a formal model and notation.We then present an idealized procedure for advice elicitation as a direct mechanism with reward sharing

This idealized procedure requires the experts and the mechanism both having access to a common prior over the joint probabilities of the rewards, actions and signals, which is impractical. We then consider two practical variation of the procedure, which removes the need for a common prior, and a consider sufficient conditions for their efficiency and truthfulness.

2. MODEL

Our model and notation largely follow that of [**?**], but we use signal instead of type to avoid overloading the T in the thesis. That is, the set of signals S in our work corresponds to the set of types T in [**?**].

Let $\Omega = \{\omega_1, \omega_2 ... \omega_M\}$ represent the finite set of states of nature. Each represents a complete description of all relevant aspects of the world for the decision. This includes both inherent physical properties of the susbject and the treatments, as well as the subjects likely choice of a in response to different choices of c.

Let C denote the set of available choices, that the mechanism outputs one of them c as the to advice for the subject. For each state of the world ω and action c that is chosen as advice to the subject, there is an associated expected reward r the subject receives which depends on ω and a. That is is the effect of c on the rewards is entirely through it's influence on the choice of the subject a.

Let S_i be a finite set of possible types of expert i. As in [**?**] an experts information may be of two qualitatively different kinds: information about the objective characteristics of the subject and the effects of their alternative decisions and how generally persuasive they are to find the recommended action.

Idiosyncratic information about the agent himself: their ability to aggregate the signals efficiently, and how persuasive the would result as the advisors who selected c in the second stage of the game. The former is of interest to other agentsand consequently is the cause of the interdependence of agents valueswhile the latter is irrelevant to other agents in calculating their values.

A direct mechanism announces a payment rule $p_i(\hat{s}, c, a, r)$, with p_i being the payment to expert i, with $\hat{s}$ being the reported signals. Each expert i receives their signals s_i and makes a report $\hat{s}_i$ to the mechanism, which then outputs a chosen action based on those reports. The subject observes the action chosen by the mechanism c, and takes

a (potentially different) action a. A reward r is observed, and the mechanism makes the corresponding payments.

We denote by c_s^* the reward maximizing choice given the true signals s. An expert i's report is truthful when $\hat{s}_i = s_i$. A mechanism is incentive compatible when truthful reports maximize the payment experts receive and this holds for all experts. This can occur in dominant strategies or in a Bayesian Nash Equilibrium (BNE).

2.1. Compliance

To account for the possibility that the subject does not follows what the common prior plus signals posit as the choice c we can introduce as before a actual action variable a.

For each state of the world and action that is given as advice to the subject, there is an associated reward the subject receives this can be marginalized out,

As long as the common prior of the experts includes both c and a then the mechanism will output the optimal c to advice the subject, which might differ from the optimal action for the subject to take (a) if a is likely to result in non compliance (to a lower reward alternative than c) but c is not.

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. The core insight is simple, if there is a common prior and the mechanism can access it, it can aggregate the signals and pick optimally. For a given experts, if all other experts are being truthful, truthfulness maximizes the reward.

MECHANISM 1. *[Direct Reward Sharing Mechanism (DRSM)] All agents report their signals to the mechanism, and the mechanism selects the action c that maximizes expected rewards given the reported signals, the agent takes their action a and receives reward r, and all agents receive a linear share of the reward.*

More formally the allocation rule is the "max decision", and the payment rule is invariant to the experts reported types and only depends on the rewards received by the subject.

3.1. Efficient and Incentive Compatible

In a Bayes-Nash equilibrium the mechanism is efficient and truthful: if all other agents are reporting truthfully then a truthful report i results the maximization over reports to be equal to the true maximization given the true signals, and by construction one cannot do better.

PROPOSITION 1. *The DRSM has a* efficient *BNE*

PROOF. TODO □

PROPOSITION 2. *The efficient BNE of the DRSM has the highest payoffs to experts.*

PROOF. This is immediate from

□

It is not possible to have in dominant strategies for the experts, the intuition for that is that when one agent misreports another one (if possible) is better off by making a opposite misreport so as to have the mechanism select the correct action.

This reward sharing mechanism is two important limitations. First, it requires that the mechanism has access to the common prior, which is for many applications highly unrealistic. Second, its requirement that signals be sent makes it more complicated

than the standard direct mechanisms. The next two sections explore variations on the idea to address these.

4. PRACTICAL MECHANISMS: ADVICE AUCTIONS

Given the practical settings that motivate this work, we do not assume acess by the mechanism to a common prior, we now consider two practical alternatives

MECHANISM 2. *[Bid and Signal Advice Auction] All agents report their signals and a bid to the mechanism, The mechanism selects the highest bidder, reveals the other reported signals to them, the highest bidder selects the action c, the subject take their prefered action, and .*

Payments are

4.1. Efficient Aggregators and Limits to free entry

if we can limit entry into the mechansim to such that all participants satisfy the efficient aggregator condition (for example by having a test with simulated signals and known cases), then this reduces to an auction of a common value good.

5. A SIMPLE ADVICE AUCTION WITHOUT SIGNAL PASSING

In many practical applications of interest, direct signal mechanisms might be impractical. It might be that the signals can't be practically reported: for example, because while the expert knows something when they see it, they do not have a vocabulary to unambiguously express it. If we replace the reports in the direct mechanism by bids, and we allow the winner of the auction to pick the action (after observing the losing bids), we obtain a mechanism that doesn't make any strong assumptions beyond that the agents understand the mechanism and are profit driven.

MECHANISM 3. *[]*

Each agent observes their signal and report a bid b_i. The agent with the highest bid in the auction, the owner *of the choice, observes the full set of bids and then chooses an action c; the subject then selects the action a and the reward r is received.*

Denote by o_i a indicator variable encoding with a value of 1 if the agent i was owner *of the choice, and let $b_{\hat{2}}$ denote the value of the second highest bid. If both agents' bids are equal, select the owner between them uniformly at random.*

$$\pi_i = \begin{cases} r - b_{\hat{2}}, & \textit{if } o_i = 1 \\ 0, & \textit{otherwise} \end{cases}$$

This kind of indirect mechanism is inherently limited when the signals of the agent are multi-dimensional. Their bid, being scalar valued, cannot encode the full information contained in the signals, and thus this limits any mechanism that relies solely on simple bids to aggregate information.

For some information structures the bidding mechanism still aggregates information efficiently, for example:

DEFINITION 1 (FULL SET OF SPECIALIST EXPERTS). *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.*

Note that this situation fails to satisfy the unilaterally sufficient expert requirement. The bidding mechanism is able to meaningfully aggregate information in at least some information structures; note no single expert knows the optimal action.

6. CONCLUSION

6.1. 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.

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 extension of this setting motivated by the bandit setting: a sequence of subjects (patients in the medical motivation) as in that setting, and a fixed set of advisers (doctors) with access to side information about different patients expected values under different courses of action.

We present a simple mechanism that preserves subjects freedom, argue that this is a inherently desirable practical design criterion as it enables no-regret exploration (since the subject remains at all point in control of the decision, and so trying the mechanism cannot reduce their choices).

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. 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. Scenarios such as these that motivate optimal decision elicitation are more 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. Bounded regret algorithms with compliance awareness can be seen as addressing the special case where the experts' signals are known a priori to be uninformative, and thus only the experience can be learned from.

Our one-subject mechanism 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.

Given the focus of this thesis on algorithms and mechanisms that preserve freedom, we do not wish to use forced randomization of the action taken which would allow for an unbiased estimator to be constructed and then used to set rewards. We can instead

In contrast to the previous chapter's motivation in the literature, in this chapter our focus is 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. In the two special cases of compliance aware bandits and single subject decision markets, the proposed mechanism reduces to the

previously proposed mechanism. There are, however, (infinite) other mechanisms that share this property.

A key limitation of the one shot case is the need to randomize the agents choice to evaluate the counterfactual, and when there is a single agent these exploration (of the sub-optimal relative to the information reported) steps are not ex-post (signal reports) incentive compatible. This is a different way to frame the result from the previous chapter on paying agents that bring uninformative signals: since evaluating the counterfactual value of the signal is not possible while preserving freedom of the agent taking the choice. In the terminology that motivates this thesis, they are not implementable while preserving the agents freedom. The most conceptually interesting change when moving to a sequence of T agents is that by pooling the risk of having with their action across agents, we can make it ex-post incentive compatible to take the exploratory actions, by linking them to suitably large random payoffs; 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 for motivating 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 without having the mechanism or without all experts having access to a good common prior, 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.

7. 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.
(2) Each expert i reports to the mechanism $\hat{s}_{t,i}$, and after all reports are received the mechanism selects a chosen action c_t
(3) The subject observes c_t, 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 $p_i(\cup\hat{s}_t, c_t, r_t, a_t \forall t \in T)$.

It is worth noting that having the signals represented a vector is without loss of generality. For example, it can be a one-shot encoding of the underlying signal. This leads to some abuse of notation, as we still denote by $s_{t,i}$ the subset of the variables encoding experts i's signal in period t. It does, however, lead to a better mapping with the contextual bandit literature, tools from which will be crucial to our results.

7.1. 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 a priori 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. We assume that subjects do not

have direct access to the previous outcomes or recommendations, and that their beliefs over the actions of other subjects are consistent with any potential action conditional on knowledge of such actions. That is, we do not assume common knowledge of a prior and subject rationality. This allows us to side step the impossibility results outlined above.

A closely related assumption (for the setting without outside experts providing advice, but with other agents also participating in a game with joint payoffs [Mansour et al. 2016]) are *explorable actions*: the actions which some incentive-compatible policy can recommend with non-zero probability. Note that by making beliefs that there are other agents who are likely to take any actions sufficiently often enough to reveal if the rewards of that action are of the highest value, all actions can become explorable.

The motivation behind this choice is that in many settings it is natural that the experts are playing the game repeatedly and for profit, thus rationality can be naturally achieved and sustained; this is much less likely to be the case for the subjects. [1] It is precisely because subjects are in an unfamiliar situation that they seek out the help of experts in making their decisions. The model's main concern is thus in contrast to ([Kremer et al. 2014; Mansour et al. 2015]), who take the subjects as rational and the mechanism as the social planner which learns from experience, without relying on information from rational experts. While the assumption of rationality and common knowledge in that case enable the use of information revelation structures for incentive exploration, here we take the exploration for granted and focus on the incentives of the experts.

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

Even with access to the prior by the mechanism (where the VCG-style mechanism provides efficiency) or in situations in which the bidding mechanism is efficient in the one-shot case, repeated use of such mechanisms fails to achieve the efficient outcome in the sequential setting. Running the mechanisms 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:

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

As usual with $\alpha < 1/N$ so as to preserve rational entry by the subjects by limiting the payments. 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 VCG-style 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 agents and 3 arms with T time periods. 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.*

[1] Ample evidence from experimental economics shows that while humans in unfamiliar environments can be far from rational, experienced professionals faced with similar real world tasks are much likelier to be rational, and this common knowledge. A vivid illustration of this is provided by experiments where chess players are faced with a centipede game.

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's BNE choice sequence relative to the optimal policy with exploration is $(2/3 - 1/2)(T - 1) - (1/2 - 1/3)$.

DEFINITION 2 (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 and a repeated one shot VCG-like mechanism in Example 1, there is a BNE of the repeated single subject VCG-style which results in the greedy policy. The same inefficiency occurs in the repeated second price auction-style mechanism. Given that there is no winner's curse due to the signal structure[2], 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; 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.

A different approach would be to seek a direct mechanism which internalizes exploration: a dynamic VCG-style mechanism. However, the requirements that there be a common prior over all possible sequences of signals, actions and outcomes, and that this be known to the mechanism, making this approach impractical. From a conceptual perspective, such an approach does not shed any new light relative to what was explored in the previous chapter.

9. 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 (BOC) MECHANISM). *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.*

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

[2]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.

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 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 externalities created by selecting actions that have not previously been selected.

PROPOSITION 3. *There is a BNE under which the BOC mechanism 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 BNE. 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 (both result in 0 payoff). □

One problematic feature of the above is that it relies on agents' signals being truthfully reported without there being any benefit from doing so. Formally, this can be sustained in BNE because deviations do not benefit the individual making them, holding the other agent actions constant. Notice, however, that this is dependent on there being no cost at all to reporting truthfully. Conceptually this is lacking, in that the motivation for the mechanism is precisely to provide incentives for experts to be truthful.

10. PAYING ONLY FOR USEFUL SIGNALS WITHOUT COMMON PRIORS

In the previous chapter we showed that in the one-shot case without a common prior it was not possible to reward only experts who provided valuable signals, since there was no way to evaluate the counterfactual reward obtained by selecting an action without their reports.

Interestingly, the sequential version of the problem opens up aggregation in the repeated version, where doing so is not possible in the one shot-case without paying useless experts in expectation. The reason why we need the prior in the one shot case to be able to evaluate the contribution the signal makes to picking an action with good rewards is that since we observe a single action there is no way to estimate this without using a prior. We now consider a setting where the experts submit their signals, however, without the mechanism having access to the prior.

While we saw that there was a truthful equilibrium, the mechanism failed the property of paying for useless experts. We show that the repeated nature of the problem can be used to overcome payments to useless experts. We do this in two parts. First, reducing the problem to a contextual bandit algorithm where the context is the signals reported by all experts. Second, using the randomization used by such algorithms to construct unbiased estimators of the gain in rewards brought about by a given agent's signals, and making their reward proportional to this. The first part allows the mechanism to learn how to use the experts' signals. The second shows how much (marginal) value they provide.

Assume that the subjects arrive IID. Use an unbiased contextual bandit algorithm evaluation strategy such as [Li et al. 2011] Algorithm 2 and a contextual bandit algorithm (such as [Syrgkanis et al. 2016]) to learn to use the reported signals.

MECHANISM 5. *[Signals without Priors] Inputs: A contextual bandit algorithm A and an unbiassed offline evaluation algorithm E.*

In each time period t a new subject arrives and experts 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 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 is as follows:

$$\pi_i = \alpha(\sum_1^T r_t - E(\hat{s_{-i}}, A))$$

Setting $\alpha < 1/N$ ex ante bounds the total payments to the experts below the benefits of having the signals. When a agent i doesn't change the exploration policy (in expectation), then $\mathbb{E}[r_t] - E(\hat{s_{-i}}, A)) = 0$.

10.1. Limitations

To preserve the tractability of the analysis we did not include the agent that makes the choice and the possibility that they might not select an action the algorithm. The contextual bandit algorithm could be replaced with a compliance aware version by using the hierarchical construct from Chapter 2. However, the freedom preserving version would be incompatible the unbiased assumption, since the subject making the choice could introduce arbitrary bias between the proposal and observed distribution, for example by the distribution of actual actions not having full support.

Note contextual bandit based mechanism implies learning how to interpret the signals; this, especially in the presence of multiple experts with conflicting priors, appears to be the most practical application. In other words, we are not too worried about what we lose from having to randomize the choice, since this is needed for learning when the experts disagree due to different priors.

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

To maintain subjects freedom we do not wish, as in the example above, to conflate the choice of the algorithm with that of the subject.

MECHANISM 6 (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.

12. 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.

In the previous chapter on the one-shot setting, we saw that payments for signals without a common prior that allow us to evaluate the counterfactual are problematic, in that we cannot reward useful experts more than useless ones. In the previous mechanism of this chapter, we saw that in the repeated setting this is not the case. We can use an unbiased estimator of rewards to estimate this counterfactual without requiring a prior. 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.

This mechanism is the composition of the bid mechanism and signal mechanism.

MECHANISM 7. *[] 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 agents 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 $E(\hat{s_{-i}}, A)$.

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 disadvan-

tage; 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.

13. CONCLUSION

We introduced a new and natural setting, that generalizes decision markets and bandit problems. We showed a series of natural but ultimately flawed mechanisms that allow us to refine what is needed for a mechanism not to be flawed. Building on these, we proposed a mechanism that is plausibly practical but hard to analyze.

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).