Decisions, Learning and Games:
You've Got To Have Freedom.

Nicolás Della Penna

A thesis submitted for the degree of
Doctor of Philosophy
The Australian National University

August 2019

Except where otherwise indicated, this thesis is my own original work. In particular, Chapter 3 contains material jointly written with David Balduzzi.

Nicolás Della Penna
27 August 2019

a Lucia, Laura & Claudio, por todo.

Acknowledgments

Mark D. Reid provided maximal freedom as a supervisor. Bob Williamson was incredibly patient as chairman. He and three anonymous reviewers who provided constructive comments that greatly improved this work.

David Balduzzi was an excellent friend and productive collaborator. Yiling Chen gave generous advice when I was starting to understand the predictions literature. Scott Sanner was always thoughtful and methodical in engaging with even my wildest questions and frustrations.

Leo Celi and Rob Stretch where generous in helping me in my attempts to understand decision making in medical practice. Ben Moran and Adam Kelly, for the friendship and the many conversations that helped me understand the constructive side of contingent claims.

Rafael Frongillo,Brendan van Rooyen, and Garrett Johnson, for their steady and reasonable friendships. Daniel Reusche and Raphael Kabo for both their friendship and their help proof reading this. Alice Hutchings, Brendan Loo Gee, Karina Vennonen, Mitchell Jon Brown, Carolina Brum, Nathan Trachimowicz, Mako Hill, Mika Matsuzaki, Blaine Cook and Natasha Carolan for their incredibly generous hospitality and friendship over these yeas.

Avi Ruderman was an ideal co-conspirator and academic sibling. Maja Tampe, Tamar Regev, and Jacinta Torres, Darío García García, Carsten Riechelmmann and Erik Garrison, for the many decisions that they made better, and the countless others were they graciously accepted my unreasonableness.

Abstract

Maintaining a subject's freedom to decide imposes structure and constraints on learning systems that aim to guide those decisions. Two natural sources from which subjects can learn to make good decisions are past experiences and advice from others. Both are affected by the subject's freedom to ultimately act as they wish, giving rise to learning theoretic and game theoretic repercussions respectively.

To study the effect of past experiences, we extend the standard bandit setting: after the algorithm chooses an action, the subject may actually carry out a different action. This is then observed along with the reward. Algorithms whose choice of action is mediated by the subject can gain from awareness of the subject's actual actions, which we term compliance awareness. We present algorithms that take advantage of compliance awareness, while maintaining worst case regret bounds up to multiplicative constants. We study their empirical finite sample performance on synthetic data and simulations using real data from clinical trials.

To study the effect of advice of others, we consider the literature on incentives for multiple experts by a decision maker that will take an action and receive a reward about which the experts may have information. Existing mechanisms for multiple experts are known not to be truthful, even in the limited sense of myopic incentive compatibility, unless the decision maker renounces their ability to always take on the best ex-post action and commits to a randomized strategy with full support. We present a new class of mechanisms based on second price auctions that maintain the subject's freedom. Experts submit their private information, and the algorithm auctions off the rights to a share of the reward of the subject, who then has freedom to pick the action they desire after observing the submitted information. We show several situations in which existing mechanisms fail and this one succeeds. We also consider strategic limitations of this mechanism beyond the myopic setting that arise due to complementary information between experts, and practical considerations in its implementation in real institutions.

We conclude by considering a natural hybrid setting, where a sequence of subjects make decisions and each can receive advice from a fixed set of experts that the mechanism seeks to incentivize. The model for this setting is extremely general, having as special cases standard, compliance aware and contextual bandits, as well as decision markets. We present a novel practical market structure for this setting that incentivizes exploration, information revelation, and aggregation with selfish experts.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

That thing is called free, which exists solely by the necessity of its own nature, and of which the action is determined by itself alone. On the other hand, that thing is necessary, or rather constrained, which is determined by something external to itself to a fixed and definite method of existence or action.– Spinoza, Ethics, Part I, Definition VII

1.1 Thesis Statement

A decision maker's freedom has both positive and normative implications for the design of learning algorithms and mechanisms that seek to improve decisions. Positively, incorporating awareness of subject freedom can improve the performance of learning algorithms for decision problems, relative to those which do not take it into account. Normatively, it motivates maintaining subject freedom as a design criterion in the design of mechanisms for decision making.

1.2 Problem Statement

One can learn to decide from experience or from the advice of others. Consider the following two situations:

1. An algorithm seeks to help a doctor facing a sequence of patients for which there is an established and a novel treatment.
2. Patients seek to elicit information from experts to select the optimal treatment for their condition.

In both situations it is natural to assume that the patients have the last say on what treatment they take. In a more abstract sense the *subject* who takes the action and lives through its consequences retains their freedom; their actions are not externally determined by the system which helps to inform them. *Maintaining freedom* for the subject in decision support thus implies that the actions the system suggests need not be those the subject takes.

These two motivating interaction patterns, ignoring considerations on the subject's freedom, are found reflected in two previously separate parts of the literature: the first is the classic bandit setting [Thompson, 1933]; the second is the more recent and relatively obscure literature on decision markets [Berg and Rietz, 2003; Hanson, 2002; Othman and Sandholm, 2010; Boutilier, 2012; Chen et al., 2014]. In both it is widely assumed that the action selected by the algorithm or mechanism is the one carried out by the subject.

This is implicit in most of the bandit literature, where no variable encodes the potential distinction between the algorithm's and the subject's choices of actions; rarely consideration is given to the possibility that they can differ. Incentive-compatible bandits [Kremer et al., 2014; Mansour et al., 2015, 2016] are a noteworthy exception.

The subject's follow-through of the algorithm's or mechanism's selected action is explicit in the decision market literature. Those mechanisms based on sequential proper scoring rules contingent on the action taken (voiding the markets contingent on the actions not taken) require not only that the subject follow the mechanism's choice, but select ex-post dominated actions with positive probability to incentivize experts.

Operationally, in the bandit setting, our notion of the subject's freedom can be captured by considering, in addition to the usual variable which encodes the action that the algorithm or mechanism selects, a second variable for the action that the subject actually takes. Naively using such a variable and simply replacing the chosen with the observed action (in a standard worst case sub-linear regret algorithm) leads to linear regret in the worst case.

In the mechanism design problem of expert elicitation for decision making, maintaining freedom rules out classes of mechanisms that rely on the subject taking dominated actions with positive probability. Previously, no mechanisms that are incentive compatible with many experts where the subject retains its freedom (is not knowingly required to take an ex-post dominated action with positive probability) were known [Othman and Sandholm, 2010; Chen et al., 2014].

1.3 Freedom: Subject as Principal for Decisions with no Externality

This thesis takes freedom of subjects as a design criterion and seeks to further the understanding of how to incorporate it into the algorithms and mechanisms where it is relevant. The natural setting where this is a good design criterion is actions that only affect a single agent, the *subject*, who both carries out the action and receives the reward. Motivated by Spinoza Ethics' Definition VII that opens this chapter, we define the subject as free if the subject's action is not determined by the output of the algorithm or mechanism.

In the bandit setting, preserving the freedom of the subject requires that the algorithm does not directly control the action taken. This opens the possibility that the algorithm's choice of action for a round is different from the action the principal carries out.

In the work on the normative implications for incentive schemes used in part II we assume a utility maximizing principal; mechanisms that retain their freedom allow them to pick the action that maximizes their utility. In the decision market mechanisms previously proposed in the literature, incentive compatibility of the max decision rule for the experts necessitates [Othman and Sandholm, 2010; Chen et al., 2014] violating the freedom of the subject by dictating that the distribution of actions they take have full support. This rules out subjects that always take their optimal decision.

It is worth noting this design criterion clashes with other desirable ones, most notably the utilitarian objective of maximizing social welfare, where the principal is an abstract social planner who aims to maximize the sum over all agents' utility. This social welfare objective means that optimal mechanisms there [Kremer et al., 2014; Mansour et al., 2015, 2016] constrain the information set revealed to the subjects (e.g. the patient who is or is not taking the treatment at a given point in time for the clinical case).

1.4 Decisions

Decision making, as understood in this thesis, is concerned with selecting an action so as to achieve a favorable outcome. Examples of such decision making problems are:

1. prescribing a treatment to a patient so as to maximize their quality adjusted life years.
2. selecting which ad to display to a web user so as to maximize the probability the user will click on the ad.
3. advising a company in which of some competing projects to invest in so as to maximize their profits.

The literature on bandit algorithms was originally motivated by the first, and this also is the motivating application in this thesis. More recently, work within computer science has often had some variation of the second as the motivating application. The third has been the motivating application in the decision markets literature.

Decision problems can be contrasted with prediction problems. In a prediction problem, the canonical example being weather forecasting, the performance of any strategy can be directly evaluated once the event of interest is realized. In a decision problem, the performance of strategies that take actions different from those that were used is inherently counter-factual.

In the settings with a sequence of decisions we assume that a decision does not directly affect future decisions. That is, while the underlying state of the system may be changing, the decisions do not affect its evolution.

In the expert elicitation for decision setting, we assume no inherent interest of experts on actions, nor any cost to them in acquiring their signals. For example, the expert doctors offering advice have no conflict of interest and would not profit more from carrying out a specific treatment.

1.5 Learning

We focus on two distinct sources of learning and their interaction. First, as has been the focus in the machine learning literature on online learning, we consider learning from experience in a setting where a choice from a finite set of K possible actions is sequentially repeated T times. Second, as is the focus in the decisions market literature, we consider learning from recommendations of a set of N *experts* who may have information about which of the K actions is best in a given situation.

Taking into account the subject's freedom can make learning possible in settings where it is not without doing so. A particularly relevant class of learning settings where this can be true and which arises naturally in personalized medicine and lifestyle interventions is when $K/T > 1$.

On the other hand, providing freedom to the subject can render mechanisms infeasible that seek to create the right incentives to learn from experts, by dictating the distribution over the K actions that will take place. In particular, all past incentive compatible mechanisms for $N > 1$ experts have required that the distribution over the K actions has full support.

1.6 Games

Algorithms for bandit problems have been much analysed within game theory. This has largely focused on giving worst case guarantees that result from minimax analysis of zero sum games against an adversary. Game theory plays an even more fundamental role in mechanisms for optimal decision elicitation such as decision markets, since equilibrium considerations and not just worst case concerns are inherent to the setting. Our focus in the equilibrium based analysis is on the strategic aspects of the experts offering the advice, and we consider mechanisms.

1.7 Thesis Contributions

We first turn to the purely learning theoretic implications of subject freedom, in learning from sequences of decisions taken by such free subjects. This addresses the positive aspect of our thesis statement, by showing that awareness of the consequences of subject's freedom can improve learning. If subjects have freedom, we should not assume that the actions an algorithm selects are those that are carried out in the world. Valuable information can be learned from observing when that is the case, and what happens when it is not. Formally, this is done by extending the bandit setting and to incorporate compliance information while preserving regret bounds. We present bandit algorithms that use compliance awareness and empirically outperform their standard variant, while preserving worst case regret guarantees up to multiplicative constants. We then present empirical results from simulations using implementations of these algorithms.

We then turn to purely strategic considerations, focusing on the incentive structure for the elicitation of advice on an optimal action from multiple experts. We take a normative stance, proposing preserving freedom for the subject as a first order design criterion for the mechanism. This implies that the mechanism can't have the subject take dominated actions. We present mechanisms that can elicit decision information from multiple experts without committing to taking dominated actions with positive probability. We show sufficient conditions on the signal structure of the experts for incentive compatibility and efficiency. The crucial conceptual contribution which enables this is a reduction to an auction with interdependent signals and valuations.

Finally, we consider a natural setting that emerges from the combination of the above. A sequence of subjects make decisions, and each can receive advice from a fixed set of experts that the mechanism seeks to incentivize. The model for this setting is extremely general, having as special cases standard, compliance aware and contextual bandits, as well as decision markets. We show that in natural information structures the repeated sequential use of the single-agent multi-expert mechanism fails to explore or aggregate information efficiently. We present a simple and practical market structure that incentivizes exploration, information revelation and aggregation with selfish experts, while maintaining subject freedom. We then briefly consider some of the limitations of this simple mechanism.

1.8 Scope

When we focus on incentive compatibility, we do so for the experts, not the subject. Assuming a utility maximizing subject – one that uses the max decision rule – restricts the freedom of that subject. For example, having unstable preferences that will change once the mechanism commences brings both limits and possibilities. While it limits the richness of the mechanics we can use (since we need to account

Setting	Subjects	Information	Solution Concept
Forecasting	T	past rewards for all actions	Minimax
Bandit	T	past rewards	Minimax
Peer Prediction	1	N strategic reports	BNE
Prediction Market	1	N reports, full reward vector	NE
Decision Market	1	N reports, reward for action	NE
Advice Auction	1	N reports, reward for action, taken action	NE
Compliance Aware Bandit	T	reward for action and compliance	minimax
N sided Advice Markets	T	N reports and rewards for actions	NE

Table 1.1: Relation between learning settings in the thesis and literature.

for a subject that may or may not respond to incentives), it also liberates the analysis from constraints created by assuming all subjects are rational. For example, in bayesian exploration [Mansour et al., 2015] there are priors over arms rewards where some arms are never explored, even through they may be optimal with positive probability [1]. The possibility of some share of agents not being utility maximizers means the mechanism can explore such arms.

While the direct decision elicitation mechanism we propose sidesteps the main problems of previously proposed mechanisms, it makes very strong use of a common prior assumption that extends over both compliance probabilities of subjects and a common prior probability distribution accross experts over their joint signals. This creates a tension with the canonical concern of Wilson (1987):

> Game theory has a great advantage in explicitly analyzing the consequences of trading rules that presumably are really common knowledge; it is deficient to the extent it assumes other features to be common knowledge, such as one agents probability assessment about another's preferences or information. I foresee the progress of game theory as depending on successive reductions in the base of common knowledge required to conduct useful analyses of practical problems. Only by repeated weakening of common knowledge assumptions will the theory approximate reality.

This motivates our second mechanism, which retains the structure of the direct mechanism but replaces signals with bids. We analyze different information structures to understand when information can still aggregate appropriately in this setting.

The relation between the different settings considered in this thesis and in the literature is summarized in the table bellow.

[1] They might even be optimal with a probability of almost one half

1.9 Publications and Collaborations

Most of chapters 3 and 4 on compliance aware bandits appears in [Della Penna et al., 2016b]. The work in Chapters 5 and 6 on decision elicitation from multiple experts has benefited from feedback of David Balduzzi.

During the course of the PhD I also collaborated on related publications in, prediction markets [Frongillo et al., 2012], market making [Kinathil et al., 2014, 2016], crowdsourcing [Della Penna and Reid, 2012] and medical applications [Della Penna et al., 2016a].

1.10 Thesis Outline

In Chapter 2, we provide background about the two settings that this thesis makes contributions to. In Chapter 3, we present two novel classes of algorithms and associated regret guarantees that take into account the underlying freedom not to comply with an algorithm's chosen treatment. We then study the empirical performance of these algorithms based on both synthetic and real data in Chapter 4. In Chapter 5, we turn our attention to eliciting an optimal action, and offer the first incentive compatible algorithm for elicitation from multiple experts that does not restrict the agent's freedom. We show it to be optimal while exploring some of its practical limitations from its extensive use of a common prior, as well as what is lost when we move to a simpler mechanism that relies on bids instead of signals. In Chapter 6, we present a novel setting with both multiple experts and multiple subjects that arrive sequentially, which we term *two sided decision markets*. We propose an extension of the simple mechanism based on a sequence of second price auctions that internalizes the benefits of exploration, while rewarding only valuable experts.

Chapter 2

Background and Related Work

> Desvarío laborioso y empobrecedor el de componer vastos libros; el de explayar en quinientas páginas una idea cuya perfecta exposición oral cabe en pocos minutos.
>
> Jorge Luis Borges, Prólogo de Ficciones.

Learning what action to take to maximize a reward is a fundamental problem in decision theory. We first provide an overview of the game theoretical background that underpins all aspects of this work.

We then present the building blocks from the two branches, bounded regret bandit algorithms and mechanism design, that are used in our contributions. We finalize the chapter by giving an overview of related work that we do not build upon, but which nonetheless informs or motivates our analysis, most notably the analysis of results from randomized controlled trials in the medical literature.

2.1 Notation and Conventions

The notation used for this work is a compromise to accommodate to as great an extent as possible the conventions of both the bandit algorithms and mechanism design literature, while adding the distinction between the action the algorithm or mechanism chooses and the one that the subject actually carries out in the world. We refer to the former throughout as the *chosen* action and notate it as c, and we refer to the action that the subject takes in the world as the *actual* action, which we notate as a.

We follow the bandit literature and refer to rewards (notated as r throughout) as directly observable after the actual action is taken. This is skipping the mapping of actions to outcomes, and the utility functions which map those outcomes to rewards, which is standard in the mechanism design literature. The reader wishing to move the analysis more explicitly towards the mechanism design tradition can replace the observed rewards with the von Neumann Morgensten utility function of the agent over the realized outcome. We treat experts' signals in the manner of features in a contextual bandit problem. To translate this to the formalism of the mechanism design literature, we can consider the cross-product of the agents' signals as defining a set of partitions, with one partition for each value. When we speak of two agents' signals being identical,

All models in the thesis use finite action spaces. In the strategic setting this guarantees the existence of a Nash Equilibrium. Since actions involve reports of signals s, this constrains all signals to also be discrete; note that this does not constrain the underlying latent state of the world u to be discrete.

2.2 Game Theory

Statistical learning algorithms can be understood game theoretically as a game between a forecaster and nature. This is particularly natural in the sequential (online) setting, and a framework termed Learning with Expert Advice and (Cesa-Bianchi and Lugosi [2006]) provides a unified treatment from a worst case game theoretic perspective of many such learning settings and algorithms. This literature largely considers the underlying structures to be zero-sum and thus uses a adversarial model of nature to construct strategies that have good worst case properties. In other words, this is game theory in the style of Von Neumann and Morgensten 1948. The framework was applied to the setting of sequential experiments initially by Wald, and the bandit formalism was introduced by (Robbins [1952]).

When there are multiple agents beyond nature interacting, as in the case of elicitation from multiple experts for decision making, there are severe limits to what worst case analysis alone can yield. In particular, the notions of the Nash Equilibrium(Nash et al. [1950]) and common knowledge (Aumann [1976])provide a useful starting point to thinking about such settings, though they leave us with an embarrassment of riches in terms of the potential equilibrium set.

2.3 Online Learning

The central object of analysis in the online learning framework, also known as the learning with expert advice framework, is the *regret* of an algorithm (the forecaster). This is defined as the difference in the cumulative reward between the reward the algorithm gets and the reward that would have been obtained by some benchmark. The most common benchmark is that of best fixed action in hindsight, and this is termed *static regret*. When we use the term regret without further modifiers in this thesis, we are referring to this notion.

Two main settings appear in the literature for the play of the environment that is carried out: in the stochastic setting an adversary picks a distribution over actions at the start of the game from which i.i.d. draws are later made; in the non-stochastic (oblivious) adversary setting they select a specific sequence of play before the game begins. Both choices are made with knowledge of the strategy of the learner, which is thus necessarily randomized in the non-stochastic adversarial case.

The basic structure of an online learning game is as follows.

For each round:

1. The environment chooses an action without revealing it
2. The algorithm chooses a probability distribution over the set of N actions and draws
3. The algorithm observes the reward which depends on its realized action and the realized action of the environment

A crucial aspect of online learning is the feedback model. Two fundamental extremes are full feedback and the bandit setting. In the first, after realizing a reward, the reward that would have been obtained for any other choice of action by the algorithm is also revealed. In the second, only the reward of the chosen action is revealed. More generally, prediction with partial monitoring Cesa-Bianchi et al. [2006] generalizes this as follows:

1. The environment chooses an action without revealing it
2. The algorithm chooses a probability distribution over the set of N actions and draws
3. The algorithm receives the reward which depends on its realized action and the realized action of the environment
4. The feedback is revealed to the forecaster

The feedback and loss matrices are known. In the full information setting, the feedback exactly pins down the value of the reward of the algorithm for any possible choice of the algorithm's action on that period; in the bandit setting the feedback pins down the reward only for the taken action.

2.3.1 Bandit Algorithms

Our conceptual contributions are largely agnostic about the underlying algorithm that is used. Three main families of algorithms with theoretical guarantees exist in the literature, and are based on Bayesian, upper confidence bounds and exponential weights. We also explore a heuristic algorithm, epsilon greedy, which has been observed to often have good empirical performance Kuleshov and Precup [2014].

Thompson Sampling is a very natural bayesian algorithm with good practical performance, first proposed in the literature by Thompson [1933]: it plays each action with probability equal to its posterior probability of being the best action, given the rewards observed up to that point.

A second family of algorithms encountered in the literature are based on Upper Confidence Bounds (UCB) and originate in (Lai and Robbins [1985]; Katehakis and Robbins [1995]; Agrawal [1995]). They play the action with the highest upper bound on its expected value. This embodies the principle of optimism in the face of uncertainty. A finite time analysis of the regret was presented in (Auer et al. [2002a]).

A third family, based on exponential weights, offers maximally robust guarantees, in the sense that it has close optimal (minimax) performance with regard to arbitrary non-stochastic underlying sequences of rewards. This becomes useful when we wish to create hierarchical bandits when the original sequence may not be independent and identically distributed (IID); the sequence that results from a bandit algorithm's choices will not be by construction.

A thorough analysis of bandit algorithms with adversarial regret guarantees can be found in(Bubeck [2012]; Lattimore and Szepesvari [2016]).

2.4 Mechanism Design

The central question of mechanism design is how to structure a game so as to incentivize self-interested agents to achieve some objective. The two central objectives are *efficiency* – that the sum of utilities be as great as possible – and *revenue optimal* – that the principal which runs the mechanism receives maximal net payment. In our setting, we are interested in efficiency, that is, allocating the right choice of action for the agents. The rest of this section and the later section thus focus on mechanisms in relation to that objective.

Each agent's information is characterized by their *signal*, which allows the agent to narrow down which realization of possible states of the world they are in. In the literature this is often also called an agent's *type*, particularly when it describes a private valuation of a good by that agent. We say a mechanism is *incentive compatible* when agents reach maximal utility if they report their true signals to the mechanism. Without any further assumptions (i.e. without a probability distribution over said states of the world), signals are of limited use beyond situations with a dominant strategy equilibrium.

Ideally one would like to search for mechanisms that are *strictly dominant strategy* incentive compatible. For many objectives of interest, such mechanisms do not exist, and optimal decision elicitation will turn out to be one of them. It is worth noting that a *weak* dominant strategy mechanism for optimal decision elicitation is trivial: if the payment to the experts is 0 for all possible states of the world, then any action is weakly dominant, including truthfulness. For this reason in the substantive chapters we will simply use the term *dominant strategy* and focus on strictly dominant mechanisms.

A canonical problem in mechanism design is one where each agent has a quasi-linear utility function that depends on the chosen social alternative, on their private signal, and on monetary transfers, but not on the information available to other agents. This is known as *private values*. A class of mechanisms known as Vickrey-Clarke-Groves (VCG) (Vickrey [1961]; Clarke [1971]; Groves [1973]) guarantee that truthful revelation of private information is the dominant strategy for each agent; that is, the mechanisms are dominant *incentive compatible*, and the *efficient* decision is taken. This holds for arbitrary dimensions and distributions of signals. Under independence of signal draws between agents, (Jehiel and Moldovanu [2001]) provide an efficient mechanism for the case where the quasi-linear utility function of an agent can depend on all agents' private signals. To maximize the future social welfare in a dynamic setting, variations of the efficient (VCG) mechanism exist for relatively general dynamic settings (Bergemann and Välimäki [2010]; Parkes and Singh [2003]; Athey and Segal [2007]).

We focus on models where expert information is endowed to the agents and has no cost of acquisition.

Many settings of interest, including ours, do not in general have dominant strategy mechanisms. The literature in microeconomics has largely dealt with this by using a probability distribution over signals, and then treats the problem as one of *bayesian* mechanism design. The designer then seeks to optimize the objective in expectation over this distribution. The agents' incentives are relaxed relative to dominant strategies, to ensure that their actions are a best response in expectation to the distribution of actions of other agents.

In a bayesian game there are three stages of knowledge possessed by the agents:

- *Ex ante*: before values are drawn from the distribution, the agents know this distribution but not their own types (or those of others).
- *Interim*: after the agents learn their types, but before playing in the game, the agents know their distribution and know that the other agents' types are drawn from the prior distribution conditioned on their own type.
- *Ex post*: the game has been played and the actions of all agents are known.

A simple but fundamental result in mechanism design is the *Revelation Principle*. For any mechanism and equilibrium of the mechanism, there exists an incentive compatible mechanism with the same equilibrium. The reason is that one can wrap the original non-incentive compatible mechanism with a mechanism that takes a report, assumed truthful, and simulates its optimal play in the original mechanism to pick its payments and allocations, thus achieving the same equilibrium but from the truthful reports. This holds in a vast range of situations in both the Bayes-Nash and the Dominant Strategy sense of equilibrium. It however fails to hold in natural settings of optimal decision elicitation, when agents only learn their types over time or when the mechanism designer does not know the prior (and thus can't simulate). The learning of types over time is inevitable in the learning setting where there is a sequence of subjects, while in one-off markets a common prior over the signal distribution seems almost impossible.

2.4.1 Bandit Algorithms as Mechanisms

A recent and notable exception to assuming that the algorithm in a bandit setting is able to implement any choice it desires is in the mechanism design literature around bandits (Kremer et al. [2014]; Mansour et al. [2015]). In this setting the principal is a social planner, considered to be optimizing the welfare across a sequence of agents, and strategically reveals information about past outcomes to incentivize agents to explore.

A closely related literature to the work of this thesis is focused on the incentive properties of bandit algorithms for their subjects. The study of this problem was initiated in (Kremer et al. [2014]). At each step of the bandit problem, a new agent (subject) must select which arm to pull. The incentive offered by the social planner is the recommended action. The planner does not offer payments for following the recommendation. This setting is studied in (Mansour et al. [2015]), who provide a generic black box reduction from bandit algorithms with arbitrary context and (extra) feedback to incentive compatible mechanisms.

The setting where payments are offered to the agents at each step is considered in (Frazier et al. [2014]). These works assume the central algorithm embodies a benevolent social planner that attempts to maximize social welfare, and focus on the incentives of the subjects. In contrast to these works, we abstain from subject incentive considerations, and instead focus on how to incentivize those providing the advice of which decision to take.

Another literature that studies bandit problems in a mechanism design framework is called *Strategic bandit models* and focuses on several players facing (identical) copies of the same set of arms. Players can observe not only their own outcome but also that of their neighbors. A good review of this literature, and the broader literature on the interaction between learning and strategic considerations, is in (Hörner and Skrzypacz [2016]).

2.5 Information Aggregation and Incentives

A literature in economics and particularly mechanism design is centered on when and how information can be aggregated from multiple agents that receive signals about the state of the world, and have various degrees of strategic sophistication in their actions. A definitive article on the topic with respect to the common prediction market and Arrow Debreu general equilibrium models is (Ostrovsky [2012]), which also provides an excellent overview of the historical literature within economics. Ostrovsky studies information aggregation in dynamic markets with a finite number of partially informed strategic traders. Trading takes place in a bounded time interval and in every equilibrium; as time approaches the end of the interval, the market price of a *separable* security converges in probability to its expected value conditional on the traders' pooled information. If the security is non-separable, then there exists a common prior over the states of the world and an equilibrium such that information does not get aggregated.

In these models the fact that securities are settled unambiguously implies that the state of the world is eventually observed. A largely separate literature, motivated by crowdsourcing, considers how to create incentives to elicit information when the underlying state of the world is not observable to the mechanism. In the initial mechanism (Prelec [2004]; Miller et al. [2005]), truth-telling is a strict Bayesian Nash Equilibrium. These mechanisms typically have many other non-truthful equilibria as well, and some of them may pay better than the truth telling equilibrium, motivating agents to coordinate on non-informative equilibria. A knowledge-free peer prediction mechanism that does not require knowledge of the information structure and can truthfully elicit private information for a set of information structures slightly smaller than the maximal set is proposed in (Zhang and Chen [2014]). (Kong and Schoenebeck [2016]) present a framework for information elicitation mechanisms where truth-telling is the highest paid equilibrium, even when the mechanism does not know the common prior.

2.6 Prediction Markets

The closest contact point between the online learning and information elicitation literature is in the fully supervised case. That is, the information the market is attempting to aggregate is a forecast of the future state of the world that is not contingent on the actions that the market can influence. Thus, at the time of the realization of the event, we can judge not only the forecast obtained but also any other potential forecast that could have been received. This contrasts with the bandit setting, where instead of a forecast of the state of the world we seek an action that results in a state of the world that is maximally beneficial. The equivalence between trading shares and eliciting beliefs from a single agent by the means of scoring rules goes back at least to (Savage [1971]).

Initiating with the equivalence between market scoring rules and regularized follow-the-leader algorithms in (Chen and Vaughan [2010]), a series of follow up works (Abernethy et al. [2013]; Frongillo et al. [2012]; Hu et al. [2014]; Frongillo and Reid [2015]) map prediction markets to learning algorithms. How the wealth (and thus the accuracy of prices) is concentrated between informed trades in a sequence of markets, using a natural if highly specific trader model (Kelly bettors, equivalently log utility maximizers), is studied in (Beygelzimer et al. [2012]).

The subject's freedom makes no difference in the analysis of the fully supervised setting; since there is no action to take, there is no sense in which a subject may not follow along. To the degree the information these prediction markets surface is not being used by participants in the world in a way that affects it (as the models assume), we can perfectly evaluate how accurate they are regardless of the other agents' reports.

2.7 Decision Markets

Can markets be used not just to understand the underlying distribution over future states of the world, but to select which action to take so as to induce the best distribution? In the vivid image of (MacKenzie [2008]), a market as *an engine, not a camera*. The idea of using prediction markets for decision support originates in (Berg and Rietz [2003]; Hanson [2002]). These mechanisms rely on running a prediction market for the outcome variable of interest for each possible action that the decision maker can take, and voiding those markets for the action not taken.

In (Othman and Sandholm [2010]), the authors argue that corporate prediction markets do not capture the right problem for their clients. In particular, by focusing on eliciting probabilities about what their effects will be after decisions have been made, they cannot be used to inform those decisions. It then considers the manipulability of a decision market where (in our terminology) the subject seeks to maximize their utility by always selecting the action that the market prices indicate is best, hence following the *max decision rule*. It is then shown that there are no incentive compatible market scoring rules (and thus by equivalence cost function) markets under this decision rule with multiple experts. The intuition is elementary: the last expert to trade with the market can force which of the conditional markets will be settled, so they maximize their profits by changing the price that is most incorrect, and lowering the price of all other actions bellow that. These results are formally generalized in (Chen et al. [2014]), to show that the subject must use a decision rule with full support to create the right incentives in conditional prediction markets that are used for decision support.

The above works all assume that the participants in the markets are only motivated by the payments they receive on the market (as does this thesis). A related line of work in (Boutilier [2012]) considers the case when the expert has an inherent interest in the decision.

Part I

Positive Implication for Bandit Learning: Compliance Awareness

Chapter 3

Model and Algorithms

3.1 Introduction

Bandit problems are concerned with optimal repeated decision-making in the presence of uncertainty. The main challenge is to trade-off exploration and exploitation, so as to collect enough samples to estimate the rewards from different strategies whilst also strongly biasing samples towards those actions most likely to yield high rewards.

Our running example is an algorithm that recommends treatments to patients. For concreteness, consider a mobile app that encourages patients who have recently suffered a stroke to carry out various low intensity interventions that may be beneficial in preventing future strokes. These could be as simple as meditating, going for a walk or taking an aspirin. The effects of the interventions on the probability of a future stroke may be small. The social benefits of collectively choosing the most effective interventions, however, may be large.

People often don't do as they are told. Approximately 50% of patients suffering from chronic illness do not take prescribed medications (Sabaté [2003]). It is safe to assume that the rate at which patients or doctors will follow the recommendations provided by an algorithm will fall well short of 100%. There are other settings in which compliance information is available. For example, an algorithm could recommend treatments to doctors. Whether or not the doctor then prescribes the recommended treatment to the patient is extremely informative, since the doctor may make observations and have access to background knowledge that is not available to the algorithm.

A quite different setting is online advertising, where bandit algorithms are extensively applied to recommend which ad to display (Graepel et al. [2010]; McMahan et al. [2013]). In practice, the recommendations provided by the bandit may not be followed. For example, sales teams often have hand-written rules that override the bandit in certain situations. Alternatively, the algorithm may assign a user to a treatment on their laptop, and when the user is not logged in, expose him to a different treatment on their mobile. Clearly, the bandit algorithm should be able to learn more efficiently if it is provided with information about which ads were actually shown.

Unfortunately, despite its importance in medical applications (Vrijens et al. [2012]; Hugtenburg et al. [2013]), compliance has not been analyzed in the bandit literature. In this chapter, we introduce compliance awareness into the bandit setting. In the classic multi-armed bandit setting, the player chooses one of K arms on each round and receives a reward(Auer et al. [2002b]; Auer [2002]). The player is not told what the reward would have been had they chosen a different arm. The goal is to minimize the cumulative regret over a series of T rounds. In the more general compliance setting, the action chosen by the algorithm is not necessarily the action that is finally carried out, see section 3.2.1. Instead, a compliance process mediates between the algorithm's recommendation and the action that is actually taken. Importantly, the compliance process may depend on latent characteristics of the subject of the decision. We focus on the case where the outcome of the compliance process is observable.

Unfortunately, compliance information is a two-edged sword. There are settings where it is useful; but it can also lead to linear regret. We present sub-linear regret algorithms that incorporate compliance information and provide worst case regret guarantees that match the standard ones for multi-armed bandits (which we term the chosen strategy) up to multiplicative constants. We also show stylized example situations where compliance aware algorithms have bounded regret and standard ones do not.

3.1.1 Outline

Section 3.2 introduces the formal compliance setting and introduces three protocols for incorporating compliance information into bandit algorithms. Each protocol has strengths and weaknesses.

The simplest protocol ignores compliance information – yielding the classical setting where standard regret bounds hold. If instead of attending to its recommendations the algorithm attends to whether the subject actually follows the recommendation, it is possible to learn faster than without compliance information. On the other hand, there are no guarantees on convergence when an algorithm attends purely to the compliance of subjects and ignores its own prior recommendations – examples of linear regret are provided in section **??**. A natural goal is thus to simultaneously incorporate compliance information whilst preserving the no-regret guarantees of the classical setting. We present two hybrid algorithms that do both.

The first, `HierarchicalBandit`, is a two-level bandit algorithm. The bottom-level learns three experts that specialize on different kinds of compliance information. The top-level is another bandit that learns which expert performs optimally. The algorithm thus has no-regret against both the treatments and two natural reward protocols that incorporate compliance information.

The second algorithm, `ThompsonBandit`, rapidly converges to Thompson sampling with standard guarantees. However, when Thompson sampling is unsure about which arm to pull, the algorithm takes advantage of the uncertainty to introduce arm-pulls sampled from `HB`.

Empirically, `TB` achieves a surplus of 8.9 extra survivals (that is, human lives) relative to the randomized baseline. The `HB` algorithm with `Epsilon Greedy` as the base algorithm achieves a surplus of 9.2. In contrast, the best performing strategy that is not compliance aware is Thompson sampling, which yields 7.9 extra survivals.

3.1.2 Comparison with other bandit settings

It is useful to compare noncompliance with other bandit settings. Partial monitoring is concerned with situations where the player only partially observes their loss Alon et al. [2015]. Our setting is an extension of the bandit setting, where additional compliance information is provided. Whether or not a patient complies is a form of side-information. However, in contrast to the side-information available to contextual bandits, compliance is only observed *after* an arm is pulled. An interesting question, left for future work, is how contextual and compliance information can both be incorporated into bandit algorithms simultaneously.

Hybrid algorithms were previously proposed in the best-of-both-worlds scenario (Bubeck and Slivkins [2012]; Seldin and Slivkins [2014]), where the goal is to construct a bandit that plays optimally in both stochastic and adversarial environments. Vapnik introduced a related notion of side-information into the supervised setting with his learning under privileged information framework (Vapnik and Vashist [2009]).

An important point of comparison is the bandits with unobserved confounders model introduced in (Bareinboim et al. [2015]). That paper was motivated using an extended example involving two subpopulations (drunk and sober) gambling in a casino. Since we are primarily interested in clinical applications, we map their example onto two subpopulations of patients, rich and poor. Suppose that rich patients always take the treatment (since they can afford it) and that they are also healthier in general. Poor patients only take the treatment when prescribed by a doctor.

In Bareinboim et al. [2015]) they observe that the question "what is the patient's expected reward when taking the treatment?" is confounded by the latent variable `wealth`. Estimating the effect of the treatment – which may differ between poor and rich patients – requires more refined questions. In our notation: "what is the patient's expected reward when taking the treatment, given she is wealthy?" and "what is the patient's expected reward when taking the treatment, given she is poor?", see example **??**. The solution proposed in (Bareinboim et al. [2015]) is based on the regret decision criterion (RDC), which estimates the optimal action to

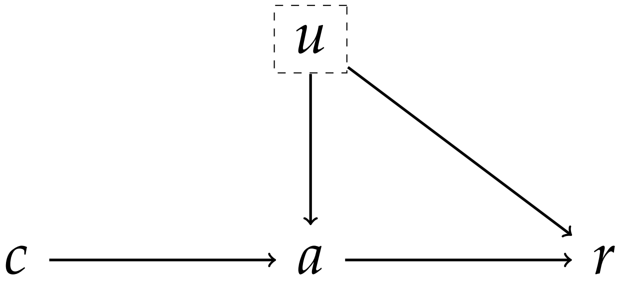

Figure 3.1: Bandit with Compliance Awareness DAG

be the one maximizing the expected reward given the patient's inclination, where the action chosen may *differ* from the patient's latent inclination. Essentially, computing the RDC requires imposing interventions. However, overruling a patient or doctor's decision is often impossible and/or unethical in clinical settings. The counterfactual information required to compute the RDC may therefore not be available in practice. Compliance information does not act as a direct substitute for imposition of interventions. However, compliance information is often readily available and, as we show below, can be used to ameliorate the effect of confounders by giving a partial view into the latent structure of the population that the bandit is interacting with.

3.2 Model

This section introduces a formal setting for bandit algorithms with noncompliance and introduces protocols that prescribe how to make use of compliance information. Before diving into the formalism, let us discuss informally how compliance information can be useful.

First, suppose that the patient population is homogeneous in their response to the treatment, and that patients take the treatment with probability p if prescribed and probability $1 - p$ otherwise, where $p < 0.5$. In this setting, it is clear that a bandit algorithm will learn faster by rewarding arms according to whether the treatment was *taken* by the patient, rather than whether it was *recommended* to the patient.

As a second example, consider *corrective compliance* where patients who benefit from a treatment are more likely to take it, since they have access to information that the algorithm does not. The algorithm clearly benefits by learning from the information expressed in the behavior of the patients. Learning from the treatment actually taken is therefore more efficient than learning from the algorithm's recommendations. Further examples are provided in section 3.2.1.

3.2.1 Formal setting

We consider a sequential decision making problem where a process mediates between the actions chosen by the algorithm and the action carried out in the world. Let $\mathcal{A} = [k] = \{1, \dots, k\}$ be the set of possible actions, and let T be the number of observed time steps. The general game is as follows:

Definition 1 (bandit with compliance information)**.**
At each time step $t \in [T]$, the player selects an action $c^{(t)} \in \mathcal{A}$ (the chosen action). The environment responds by carrying out an action $a^{(t)} \in \mathcal{A}$ (the actual action) and providing reward $r^{(t)} \in [0, 1]$.

The standard bandit setting is when $a^{(t)}$ is either unobserved or $a^{(t)} = c^{(t)}$ for all $t \in [T]$.

The set of compliance behaviors is the set of functions $\mathcal{C} = \{\nu : \mathcal{A} \to \mathcal{A}\}$ from advice to taken treatment Koller and Friedman [2009].

Definition 2 (model assumptions)**.**
We make the following assumptions:

1. *Compliance $\nu(u) \in \mathcal{C}$ depends on a latent variable sampled i.i.d. for each time step from unknown distribution $\mathbf{P}(U)$ over a set U.*
2. *Outcomes $r(\nu(u), a, u)$ depend on compliance behavior, treatment taken and the latent u. That is, outcomes are a fixed function $r : \mathcal{C} \times \mathcal{A} \times U \to [0, 1]$*

When $|\mathcal{A}| = k = 2$ (e.g., control and treatment), we can list the compliance-behaviors explicitly.

Definition 3 (compliance behaviors)**.**
For $k = 2$, the following four subpopulations capture all deterministic compliance-behaviors:

- *never-takers* $\mathfrak{N} : \Big(0 \mapsto 0, 1 \mapsto 0\Big)$
- *always-takers* $\mathfrak{A} : \Big(0 \mapsto 1, 1 \mapsto 1\Big)$
- *compliers* $\mathfrak{C} : \Big(0 \mapsto 0, 1 \mapsto 1\Big)$
- *defiers* $\mathfrak{D} : \Big(0 \mapsto 1, 1 \mapsto 0\Big)$

Unfortunately, the subpopulations cannot be distinguished from observations. For example, a patient that takes a prescribed treatment may be a complier or an always-taker. Nevertheless, observing compliance-behavior provides potentially useful side-information. The setting can be contrasted from contextual bandits because the side-information is only available *after* the bandit algorithm chooses an arm.

Definition 4 (stochastic reward model)**.**
The expected reward given subpopulation s and the actual treatment $a \in \mathcal{A}$ is

$$r_{s,a} := \mathop{\mathbb{E}}_{u \sim \mathbf{P}(U)} \left[r(\nu(u), a, u) \,\middle|\, \nu(u) = s \right] \quad \textit{for } s \in \{\mathfrak{N}, \mathfrak{A}, \mathfrak{C}, \mathfrak{D}\}.$$

The goal is to maximize the cumulative reward received, i.e. choose a sequence of actions $(c^{(t)})_{t \in [T]}$ that maximizes

$$\mathop{\mathbb{E}}_{u \sim \mathbf{P}(U)} \left[\sum_{t \in [T]} r(\nu(u), \nu(u)(c^{(t)}), u) \right]$$

In the non-compliance setting there is additional information available to the algorithm. Ignoring the compliance-information (`Chosen`) reduces to the standard bandit setting. However, it should be possible to improve performance by taking advantage of observations about when treatments are *actually* applied. Using compliance information is not trivial, since bandit algorithms that rely purely on treatments (`Actual`) or purely on compliance (`Comply`) can have linear regret.

This section proposes two hybrid algorithms that take advantage of compliance information, have bounded regret, and empirically outperform algorithms running the `Chosen` protocol.

Consider the regret not relative to a best fixed action as usual, but relative to an algorithm that has access to compliance information. We show that this regret scales $O(T)$ in the non-stationary setting if the regime from which losses are drawn changes frequently enough. We also show that within each regime the compliance awareness helps converge faster, for example due to very high rates of noncompliance of subjects that don't have different underlying characteristics, making `Actual` perform well, or due to subjects having information that helps them switch to the best arm.

3.3 A Hierarchical Algorithm

A natural idea is to use the three protocols to train three experts and, simultaneously, learn which expert to apply. The resulting hierarchical bandit (specified in 1) integrates compliance-information in a way that ensures the algorithm (i) has no-regret, because one of the base algorithms uses `Chosen`, and therefore has no regret; and (ii) benefits from the compliance-information if it turns out to be useful.

The general construction is as follows. At the bottom-level are three bandit algorithms implementing the protocols `Chosen`, `Actual` and `Comply`. On the top-level is a fourth bandit algorithm whose arms are the three bottom-level algorithms. The top-level bandit learns which protocol is optimal. Note the top-level bandit is *not* in an i.i.d. environment even when the external environment is i.i.d, since the low-level bandits are learning.

Algorithm 1 `HierarchicalBandit (HB)`

```
Input: Bandits $\mathcal{B}_i$ running NoRegretAlgorithm on Comply Chosen, and Actual
for $i = \{1,2,3\}$ respectively, with arms corresponding to treatments
Input: Bandit $\mathcal{H}$ running NoRegretAlgorithm compatible with adaptive environments,
with arms corresponding to $\mathcal{B}_i$ above
for $t = 1$ to $T$ do
  Draw bandit $i^{(t)} \in \{1,2,3\}$ from $\mathcal{H}$ and arm $j^{(t)}$ from $\mathcal{B}_{i^{(t)}}$
  Pull arm $j^{(t)}$ of $\mathcal{B}_{i^{(t)}}$; observe loss $\ell = \ell^{(t)}_{i^{(t)},j^{(t)}}$; observe compliance
  Update $\mathcal{H}$ with loss $\ell$ applied to bandit-arm $i^{(t)}$
  if $i^{(t)} = 1$ then
    Update $\mathcal{B}_1$ with loss $\ell$ applied to treatment-arm $j^{(t)}$
  end if
  Update $\mathcal{B}_{2/3}$ with loss $\ell$ according to protocols Chosen and Actual respectively
end for
```

3.3.1 Regret analysis

Definition 5 (Regret)**.** *The* regret *of an online learning algorithm* A *is*

$$\mathtt{Regret}_{\mathtt{A}}(T) = \sum_{t\in[T]} \ell^{(t)}_{j^{(t)}} - \min_j \sum_{t\in[T]} \ell^{(t)}_j$$

where $j^{(t)}$ is the action chosen by the algorithm in time step t, and $\min_j \sum_{t\in[T]} \ell^{(t)}_j$ *is the minimal accumulated loss one can obtain when fixing an action and choosing that same fixed action each step.*

This section shows that constructing a hierarchical bandit with `Exp3` (Algorithm 2) as the top-level bandit algorithm yields a no-regret algorithm. The result is straightforward; we include it for completeness. A similar result was shown in Chang and Kaelbling [2005].

The `Exp3` Algorithm (Auer et al. [2002b]) is a bandit algorithm whose worst case regret bound is robust to adaptive environments.

To obtain a hierarchical algorithm using `Exp3` as top level, we first construct a hierarchical version of `Hedge` [**??**], Algorithm 3, which is applicable in the full-information variant of our setting (the model where the counterfactual value of actions can be observed). We then modify it using the principle from `EXP3` to make it work for bandit feedback settings.

`Hedge` (Chang and Kaelbling [2005]) is an algorithm with bounded regret in the expert setting. On the bottom-level of our hierarchical version (Algorithm 3), there are M instantiations of `Hedge`. Instantiation i, for $i \in [M]$, plays an N-dimensional weight vector and receives N-dimensional loss vector $\boldsymbol{\ell}^{(t)}_i$ on round t. We impose the assumption that all instantiations play N-vectors for notational convenience. The top-level is another instantiation of `Hedge`, which plays a

Algorithm 2 `Exp3`

```
**Input:** $\gamma \in [0,1]$
Initialize weight vector $w_i^{(1)} = 1$ for $i \in [N]$ where $N$ is the number of arms;
**for** $t = 1$ **to** $T$ **do**
    Define probabilities $x_i^{(t)} = (1-\gamma)\frac{w_i^{(t)}}{\sum_j w_j^{(t)}} + \gamma\frac{1}{N}$
    Draw an arm $i^{(t)} \sim \mathbf{x}^{(t)}$
    Pull arm $i^{(t)}$
    Incur loss $\ell^{(t)}$
    Update:
        $w_i^{(t+1)} = \begin{cases} w_i^{(t)} \cdot \exp(-\gamma\frac{\ell^{(t)}}{N \cdot x_{i^{(t)}}}) & \text{if } i = i^{(t)} \\ w_i^{(t)} & \text{else} \end{cases}$
**end for**
```

weighted combination of the bottom-level instantiations.

We have the following lemma:

Lemma 1. *Introduce compound loss vectors $\tilde{\boldsymbol{\ell}}^{(t)}$ with*

$$\tilde{\ell}_i^{(t)} := \sum_{j=1}^{N} \ell_{i,j}^{(t)} \cdot y_{i,j}^{(t)}$$

Then ρ can be chosen in `HHedge` *such that for all $i \in [M]$:*

$$\sum_{t=1}^{T} \langle \mathbf{x}^{(t)}, \tilde{\boldsymbol{\ell}}^{(t)} \rangle \leq \sum_{t=1}^{T} \tilde{\ell}_i^{(t)} + O(\sqrt{T \log M})$$

Moreover, ρ and η can be chosen such that for all $i \in [M]$ and all $j \in [N]$,

$$\sum_{t=1}^{T} \langle \mathbf{x}^{(t)}, \tilde{\boldsymbol{\ell}}^{(t)} \rangle \leq \sum_{t=1}^{T} \ell_{i,j}^{(t)} + O(\sqrt{T \log M} + \sqrt{T \log N}).$$

Proof. From Theorem 5 in [**?**] we have that the loss for `Hedge` with M actions and loss $\tilde{\boldsymbol{\ell}}_i^{(t)}$:

$$\sum_{t=1}^{T} \tilde{\boldsymbol{\ell}}_{i^{(t)}}^{(t)} \leq \min_{i \in [M]} \sum_{t=1}^{T} \tilde{\ell}_i^{(t)} + O(\sqrt{T \log M})$$

The upper level of `HHedge` is `Hedge` with loss $\tilde{\boldsymbol{\ell}}_{i^{(t)}}^{(t)}$, hence:

Algorithm 3 `Hierarchical Hedge (HHedge)`

```
Input: $\eta, \rho > 0$
$v_i^{(1)} = 1$ for $i \in [M]$
$w_{i,j}^{(1)} = 1$ for $(i,j) \in [M] \times [N]$
for $t = 1$ to $T$ do
    Set $\mathbf{x}^{(t)} = \frac{\mathbf{v}^{(t)}}{\sum_{i \in [M]} v_i^{(t)}}$
    Set $\mathbf{y}_i^{(t)} = \frac{\mathbf{w}_i^{(t)}}{\sum_{j \in [N]} w_{i,j}^{(t)}}$ for $i \in [M]$.
    Receive feedback $\boldsymbol{\ell}^{(t)} \in [0,1]^{M \times N}$
    Incur loss $\sum_{i=1}^{M} x_i^{(t)} \cdot \sum_{j=1}^{N} \ell_{i,j}^{(t)} \cdot y_{i,j}^{(t)}$
    Update weights for all $i, j$:
        $v_i^{(t+1)} = v_i^{(t)} \cdot \exp\big(-\eta \sum_{j=1}^{N} \ell_{i,j}^{(t)} \cdot y_{i,j}^{(t)}\big)$
        $w_{i,j}^{(t+1)} = w_{i,j}^{(t)} \cdot \exp\big(-\rho \cdot \ell_{i,j}^{(t)}\big)$
end for
```

$$\sum_{t=1}^{T} \langle \mathbf{x}^{(t)}, \tilde{\boldsymbol{\ell}}^{(t)} \rangle = \sum_{t=1}^{T} \sum_{k=1}^{M} \mathbf{x}_k^{(t)} \tilde{\boldsymbol{\ell}}_k^{(t)} = \sum_{t=1}^{T} \tilde{\boldsymbol{\ell}}_{i^{(t)}}^{(t)} \leq \sum_{t=1}^{T} \tilde{\boldsymbol{\ell}}_i^{(t)} + O(\sqrt{T \log M}) \forall i \in [M]$$

The lower level of `HHedge` is `Hedge` with loss $\tilde{\boldsymbol{\ell}}_{i,j^{(t)}}^{(t)}$ for i-th instantiation, hence:

$$\sum_{t=1}^{T} \tilde{\boldsymbol{\ell}}^{(t)} = \sum_{t=1}^{T} \sum_{j=1}^{N} \mathbf{y}_{i,j}^{(t)} \tilde{\boldsymbol{\ell}}_{i,j}^{(t)} = \sum_{t=1}^{T} \tilde{\boldsymbol{\ell}}_{i,j^{(t)}}^{(t)} \leq \sum_{t=1}^{T} \tilde{\boldsymbol{\ell}}_{i,j}^{(t)} + O(\sqrt{T \log M}) \forall i \in [M] \prod [N]$$

Combining the two gives the lemma □

Lemma 1 says, firstly, that `HHedge` has bounded regret relative to the bottom-level instantiations and, secondly, that it has bounded regret relative to any of the $M \times N$ experts on the bottom-level.

Algorithm 4 modifies `HHedge` so that it is suitable for bandit feedback, yielding `HExp3`. A corresponding no-regret bound follows immediately:

Lemma 2. *Define $\tilde{\boldsymbol{\ell}}^{(t)}$ as in Lemma 1 to obtain the expected loss for the upper-level `Exp3` instances:*

$$\tilde{\ell}_i^{(t)} := \sum_{j=1}^{N} \ell_{i,j}^{(t)} \cdot y_{i,j}^{(t)} = \mathbb{E}\left[\ell_{i,j^{(t)}}^{(t)}\right]$$

Then ρ can be chosen in `HExp3` *such that for all $i \in [M]$*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{i^{(t)},j^{(t)}}^{(t)}\right] \leq \sum_{t=1}^{T} \tilde{\ell}_i^{(t)} + O(\sqrt{TM \log M})$$

Moreover, ρ and η can be chosen such that for all $i \in [M]$ and $j \in [N]$

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{i^{(t)},j^{(t)}}^{(t)}\right] \leq \sum_{t=1}^{T} \ell_{i,j}^{(t)} + O(\sqrt{TM \log M} + \sqrt{TN \log N})$$

Proof. The bound for `Exp3` with M actions and loss $\boldsymbol{\ell}_i^{(t)}$(Corollary 3.2 in Auer et al. [2002b]) is:

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{i^{(t)},j^{(t)}}^{(t)}\right] \leq \min_{i \in [M]} \sum_{t=1}^{T} \ell_i^{(t)} + O(\sqrt{TM \log M})$$

Note that:

$$\tilde{\boldsymbol{\ell}}^{(t)} = \sum_{j=1}^{N} \mathbf{y}_{i,j}^{(t)} \boldsymbol{\ell}_{i,j}^{(t)} = \mathbb{E}\left[\boldsymbol{\ell}_{i,j^{(t)}}^{(t)}\right]$$

Upper level is `Exp3` with loss $\boldsymbol{\ell}_{i^{(t)},j^{(t)}}^{(t)}$, hence:

$$\mathbb{E}_i\left[\sum_{t=1}^{T} \ell_{i^{(t)},j^{(t)}}^{(t)}\right] \leq \sum_{t=1}^{T} \ell_{i,j^{(t)}}^{(t)} + O(\sqrt{TM \log M}) \forall i \in [M]$$

Hence:

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{i^{(t)},j^{(t)}}^{(t)}\right] \leq \sum_{t=1}^{T} \mathbb{E}_j\left[\ell_{i,j^{(t)}}^{(t)}\right] + O(\sqrt{TM \log M}) = \sum_{t=1}^{T} \tilde{\ell}^{(t)}$$

Lower level is `Exp3` with loss $\boldsymbol{\ell}_{i,j^{(t)}}^{(t)}$, hence:

$$\sum_{t=1}^{T} \ell_i^{(t)} = \mathbb{E}_j\left[\sum_{t=1}^{T} \ell_{i,j^{(t)}}^{(t)}\right] \leq \sum_{t=1}^{T} \ell_{i,j}^{(t)} + O(\sqrt{TM \log M}) \forall j \in [N]$$

Combining these yields the lemma.

□

Theorem 1 (No-regret with respect to `Actual`, `Comply` and individual treatment advice)**.**

Algorithm 4 `Hierarchical Exp3 (HExp3)`

```
Input: $\eta, \rho \in [0,1]$
$v_i^{(1)} = 1$ for $i \in [M]$
$w_{i,j}^{(1)} = 1$ for $(i,j) \in [M] \times [N]$
for $t = 1$ to $T$ do
    Set $\mathbf{x}^{(t)} = (1-\eta)\frac{\mathbf{v}^{(t)}}{\sum_{i \in [M]} v_i^{(t)}} + \eta \frac{1}{M}$
    Set $\mathbf{y}_i^{(t)} = (1-\rho)\frac{\mathbf{w}_i^{(t)}}{\sum_{j \in [N]} w_{i,j}^{(t)}} + \rho \frac{1}{N}$ for $i \in [M]$
    Draw bandit $i^{(t)} \sim \mathbf{x}^{(t)}$ and arm $j^{(t)} \sim \mathbf{y}_{i^{(t)}}^{(t)}$
    Pull arm $j^{(t)}$ on bandit $i^{(t)}$
    Incur loss $\ell = \ell_{i^{(t)},j^{(t)}}^{(t)} \in [0,1]$
    Update:
        $v_i^{(t+1)} = \begin{cases} v_i^{(t)} \cdot \exp\big(-\eta \frac{\ell}{M \cdot x_i^{(t)}}\big) & \text{if } i = i^{(t)} \\ v_i^{(t)} & \text{else} \end{cases}$
        $w_{i,j}^{(t+1)} = \begin{cases} w_{i,j}^{(t)} \cdot \exp\big(-\rho \frac{\ell}{N \cdot x_i^{(t)} \cdot y_{i,j}^{(t)}}\big) & \text{if } (i,j) = (i^{(t)}, j^{(t)}) \\ w_{i,j}^{(t)} & \text{else} \end{cases}$
end for
```

Let `Exp3` *be the no-regret algorithm used in Algorithm 1 for both the bottom and top-level bandits, with suitable choice of learning rate. Then,* `HB` *satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{a^{(t)}}^{(t)}\right] \leq \sum_{t=1}^{T} \tilde{\ell}_{\texttt{Actual/Comply}}^{(t)} + O(\sqrt{T})$$

where $\tilde{\ell}_{\texttt{Actual/Comply}}^{(t)}$ *denotes the expected loss vector of* `Exp3` *under the respective protocol on round* t*. Furthermore, the regret against individual treatments* $j \in [K]$ *is bounded by*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{i,j^{(t)}}^{(t)}\right] \leq \sum_{t=1}^{T} \ell_{i,j}^{(t)} + O(\sqrt{TK \log K})$$

Proof. Apply Lemma 2 to `HierarchicalBandit`. □

3.4 Compliance Awareness with i.i.d. Rewards

In an i.i.d. setting the previous strategy achieves a sub-optimal bound. Here we consider a different strategy to guarantee low regret specialized for i.i.d. settings which achieves the optimal bound up to multiplicative factors.

Algorithm 5 `ThompsonBounded (TB)`

```
Input: Bandit algorithm $\mathcal{H}$
Input: Thompson sampler under Chosen protocol
for $t = 1$ to $T$ do
    Sample $t$ and $t'$ from Thompson
    if $t = t'$ then
        Pull arm sampled from Thompson
    else
        Pull arm chosen by $\mathcal{H}$
    end if
    Incur loss, update algorithm used to pull arm
end for
```

The strategy starts from the observation that Thompson sampling often outperforms other bandit algorithms in stochastic settings (Thompson [1933]; Chapelle and Li [2011]) and has logarithmic regret (Agrawal and Goyal [2012]; Kaufmann et al. [2012]). A natural goal is to design an algorithm that performs like Thompson sampling under the `Chosen` protocol in the long run – since Thompson sampling under `Chosen` is guaranteed to match the best action in hindsight in $O(\log T)$ time – but also takes advantage of compliance information when Thompson sampling has *not* converged onto sampling a single arm with high probability. Note that under the i.i.d. setting it is not possible to obtain a stronger expected regret bound than static regret.

The proposed algorithm, `TB`, uses an algorithm that does not have guarantees (is not certified) and a Thompson Sampling based algorithm. Unlike `HB`, it does not stack them, but instead uses the Thomson algorithm to bound the behaviour of the uncertified algorithm.The Thompson sampler is initially unbiased between arms; as it learns, the probabilities it assigns to arms become increasingly concentrated. `TB` takes advantage of Thompson sampling's uncertainty about which arm to pull in early rounds to safely introduce compliance information. `TB` draws two samples. If they agree, it plays a third Thompson sample. If they disagree, it plays the arm chosen by the hierarchical bandit. Intuitively, if Thompson sampling is uncertain, `TB` tends to use the `uncertified` bandit. As the sampler's confidence increases, `TB` is more likely to follow its advice. The next theorem shows that initially mixing in side information has no qualitative effect on the algorithm's regret, which grows as $\log(T)$.

Theorem 2. *The regret of* `TB` *is bounded by*

$$\texttt{Regret}_{TB}(T) \leq O(\log T).$$

Proof. Suppose without loss of generality that arm 1 yields a higher average payoff. Let p_j be the probability that `Thompson` assigns to arm j on round t, so that $p_F = \sum_{j=2}^{k} p_j$ is the probability that Thompson sampling does *not* pull arm 1. The

Algorithm 6 `Base Thompson Sampler (BTS)`

```
Input: Probability $p$ that BTS is called by top-bandit
Set $\tilde{1} \leftarrow 1/p$
For each arm $i$ sample $\theta_i \sim \beta(S_i + \tilde{1}, F_i + \tilde{1})$
Play arm $i^{(t)} := \text{argmax}_i\, \theta_i$ and observe reward $r^{(t)}$
Sample $b$ from Bernoulli with success probability $r^{(t)}$
If $b = 1$ then $S_i \leftarrow S_i + \tilde{1}$ else $F_i \leftarrow F_i + \tilde{1}$
```

probability that `TB` follows the uncertified algorithm then is

$$1 - \sum_{j=1}^{k} p_j^2 = 1 - (1 - p_F)^2 - \sum_{j=2}^{k-1} p_j^2 \leq 2p_F.$$

The additional expected regret from deviating from Thompson sampling is therefore at most twice the regret `Thompson` incurs by pulling suboptimal arms. Finally, it was shown in Agrawal and Goyal [2012]; Kaufmann et al. [2012] that Thompson sampling has logarithmic regret. □

The algorithm can be generalized beyond the use of while preserving the bound. Note that the above proof makes no use of any property of `HB`, thus we can replace it with any other bandit algorithm, including ones that do not have regret bounds. One natural variation is to use `Actual` or `Comply` depending on whether a priori the expected effects on noncompliance include noncompliance conditional on unobservable heterogeneity of patients (in which case `Comply` would make sense) or selection towards more effective treatments (in which) that have homogeneous effects across subjects.

While for the case of Bernoulli rewards and no context, pulling the arm twice is unnecessary. We could instead use the expected probability of a pull and square it, and use the resulting probability. It does have the advantage of black boxing the details for the underlying Thompson sampling implementation, which enables the use of beliefs where deriving precise probabilities is expensive but sampling is not.

3.4.1 Hierarchical Bandit with Thompson sampler base

Using `EXP3` at the top-level and a `ThompsonSampler` as the bottom-level also yields a no-regret algorithm.

Algorithm 6 (`BTS`) shows how to modify the Thompson sampler for use as a bottom-level algorithm in `HB`. The modification applies the importance weighting trick: replace 1 in Thompson sampling with $\tilde{1} = 1/p$, where p is the probability that the top-level bandit calls `BTS` on the given round.

3.5 Data-efficiency

As described, the hybrid algorithms are data-inefficient, since despite the i.i.d. assumption on the patient population, the certified strategies only learn when they are executed. We describe a *recycling trick* to improve the efficiency of the certified strategies.

A naive approach to increase data-efficiency is to reward the certified strategy on rounds where the executed strategy selects the same action as the certified strategy. However, this introduces a systematic bias. For example, consider two strategies: the first always picks arm 1, the second picks arms 1 and 2 with equal probability. Running a top-level algorithm that picks both with equal probability results in a mixed distribution biased towards arm 1.

The recycling trick stores actions and subsequent rewards by non-certified strategies in a cache. When there is at least one of each action in the cache, the certified strategy is rewarded on rounds where it was not executed by sampling, without replacement, from the cache. Sampling without replacement is important in our setting since it prevents early unrepresentative samples introducing a bias into the behavior of the certified strategy through repeated sampling. A related trick, referred to as "experience replay", was introduced in reinforcement learning in (Mnih et al. [2015]).

Chapter 4

Empirical Evaluation

4.1 Introduction

The previous chapter introduced a pair of new algorithms that incorporate compliance information and proved that they preserve the worst case performance guarantees up to multiplicative factors. This leaves open the question whether these new algorithms can in fact outperform their standard counterparts in practical settings. This chapter uses simulations to assess that. We first consider simulations based on data from a clinical trial. Given that there is a single suitable dataset, and to explore settings were the data generating process is exactly known, we also consider several stylized models of patient compliance and simulate them.

The full source code and electronic versions of the simulation results can be found at https://github.com/nikete/thesis/blob/master/Simulations.ipynb.

4.2 International Stroke Trial Simulation

The simulation data is taken from the International Stroke Trial (IST) database, a randomized trial where patients believed to have acute ischemic stroke are treated with aspirin, subcutaneous heparin, both, or neither (Group [1997]). It contains complete compliance and mortality data at 14 days for each of 19,422 patients.

To the best of our knowledge, this is the only publicly available clinical trial with compliance data that is suitable for simulations of compliance aware algorithms. An extensive search failed to find other suitable open randomized clinical trials datasets that included compliance. A systematic review by (Ebrahim et al. [2014]) identified only 37 reanalyses of patient-level data from previously published randomized control trials; only five were performed by entirely independent authors. Data from drug abuse clinical trials is used in (Kuleshov and Precup [2014]). However, non-compliance is coded as failure, so this source, and drug dependence treatments more generally, cannot be used in our setting.

Given there is substantial loss of follow up at the 6 month measure we focus on the 14 day outcome.

4.2.1 Construction of Actual Actions from Compliance Variables

The main sources of non-compliance in the dataset are the initial event not being a stroke, clinical decision, administration problems and the patient missing out more than 3 doses. A detailed table and counts of these are included in the dataset's open access article (Sandercock et al. [2011]). While these might initially seem like reasons to discard the patients from the dataset, non-compliance is not necessarily random. Discarding these patients could cause algorithms to have unbounded regret (since the loss we care about is over all patients). In particular, misdiagnoses, administrative problems, not taking doses and other sources of noncompliance can be confounded with a patient's socio-economic status, age, and overall health.

To construct our "actual arm" variable, we assume that non-compliance entails taking the opposite treatment. This is well-defined in the Aspirin case, which only has two arms, and thus non-compliance with placebo is likely to be taking the treatment. Assigning an actual arm pulled in the heparin part of the trial is less clear cut, as it has three arms: no, low and medium dosage. We construct the actual arm variable by combining assignment and non-compliance. Non-compliance to low and medium assigned treatments is coded as not-takers, while non-compliance by a patient prescribed "none" is coded as low.

4.2.2 Results

We simulated 10,000 patients per run, which allowed us to not oversample the data in any single simulation. 2000 runs were performed, all algorithms were tested against the same draw of the run to minimize unnecessary sampling variation.

The `EXP3` gamma parameter was set ahead of time to 0.085, a choice determined by the regret-bounds for $T = 10000$ and $K = 2$ or $K = 3$. Epsilon-Greedy used a standard annealing schedule. No data dependent parameter tuning was used. The simulation was carried out by creating a "counter-factual patient": one patient was sampled i.i.d. from each treatment and control groups in the clinical trial. If the algorithm selected the treatment, it received the reward and observed the action taken by the subject sampled form the treatment group, and vice versa for the control.

Empirically, `TB` achieved a surplus of 8.9 extra survivals (that is, human lives) with 95% confidence interval $[8.1, 9.7]$, relative to the randomized baseline. `HB` with `Epsilon Greedy` as the base algorithm achieved a surplus of 9.2 (CI: $[8.3, 10.0]$) In contrast, the best performing strategy that was not compliance aware was Thompson sampling, which yielded 7.9 extra survivals (CI: $[7.2, 8.7]$).

The gains were largely concentrated in the Aspirin trial, which is consistent with the lack of benefits or severe ill effects found in the original study Group [1997] for heparin, and with the small but beneficial effect found for aspirin. If the underlying treatment has no positive or negative effect, side-information after the fact alone cannot be helpful. Note that `Actual`, and to a lesser extent `Comply`, performed

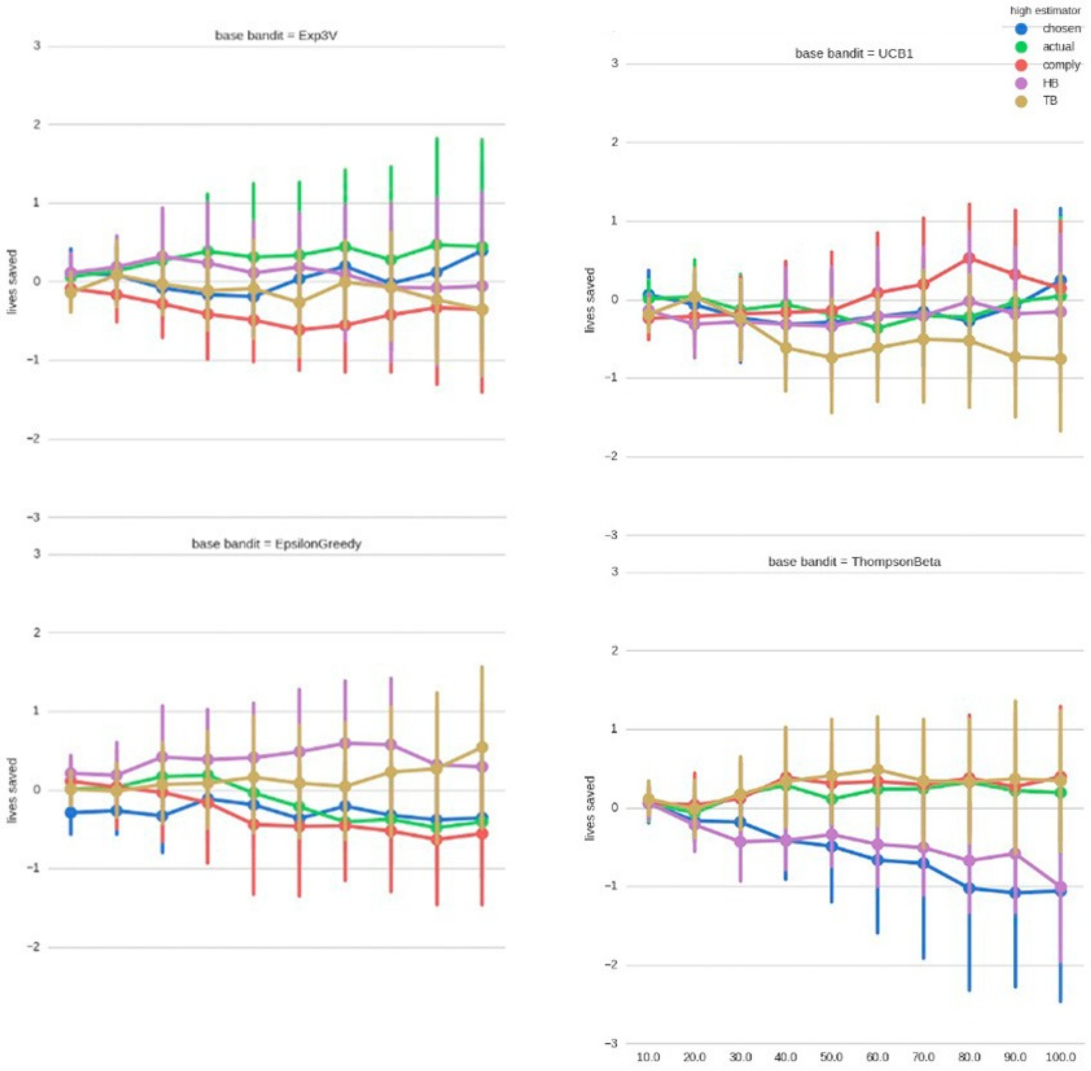

Figure 4.1: **14 Day survivals:** average lives saved over uniform random policy per 10,000 patients in 10,000 simulated trials of both Aspirin and Heparin combined.

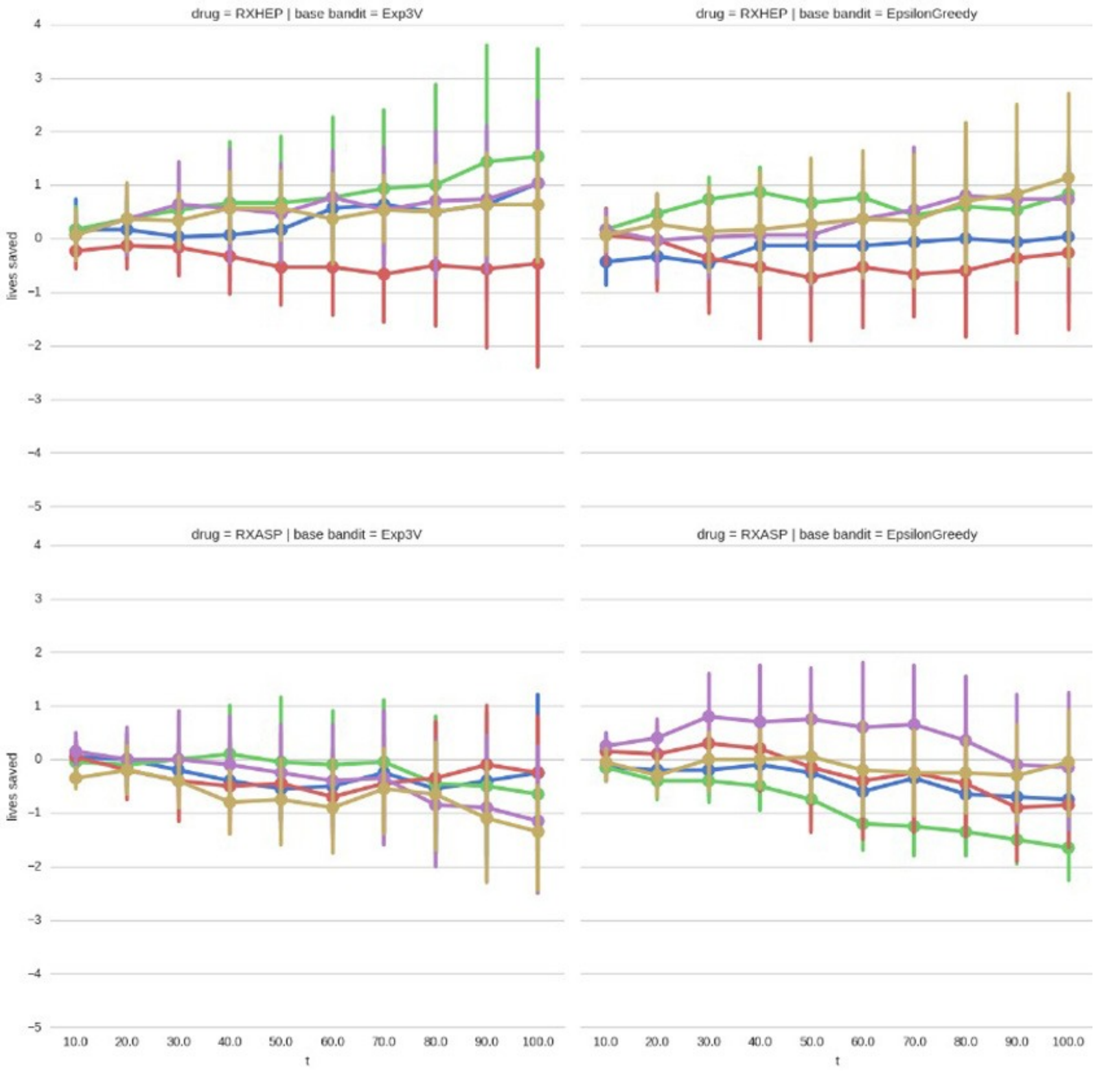

Figure 4.2: **14 Day survivals:** average lives saved over uniform random policy per 10,000 patients in 10,000 simulated trials of Aspirin and Heparin individually.

better than either Chosen or the hybrid algorithms. However, these are problematic to use directly since no guarantees apply to their worst case performance. The performance of the hybrids benefits from the information encoded in Actual and Comply whilst keeping the guarantees of Chosen.

A striking secondary empirical observation from the experiments is the strong interaction between the base learning algorithm and the nature of the feedback used. In particular, our naive implementation of EXP3 performed extremely badly under both Actual and Comply, in both the aspirin and heparin simulations. In contrast, EXP3 performs well when used as a top-level algorithm in HB in both trials, and other algorithms on the same data with the same protocol are able to learn better than Chosen, while EXP3 does worse than random (note the EXP3 guarantees are only for the Chosen protocol, so they do not apply here).

4.3 Synthetic Data

To better understand the behavior of the algorithms in a more varied range of settings, we present results of simulations with synthetic data.

4.3.1 Selection of treatment on unobservables

The first simulation illustrates Example (**??**), supposing there are two equally sized subpopulations of rich, healthy patients who always take the treatment, and poor, less healthy patients who only take the treatment if prescribed. Suppose the treatment reduces the probability for survival by 0.25. Assume that the rich patients would all do well (receive reward of 1) if they didn't take the treatment, but they all take it and so face only a 0.75 chance of survival. Poor patients who face a baseline survival of 0.75 only take the treatment if instructed, which brings their survival probability down to 0.5.

For comparison, T was kept at 10,000, and we simulated the binary outcome case. We assigned half the patients to rich and half to poor randomly. Fig. (4.3) shows that the performance of Actual and Comply is much worse than Chosen and the hybrid algorithms. Results are for 100 samples.

4.3.2 Small T

The second simulation concerns small T. A motivation for very small T adaptive clinical trials is provided by rare diseases. The overall size of the patient population is very restricted in this setting. The priors for the mechanisms of action are also often poorly understood, so potential alternative treatments can have radically different probabilities of success. We simulated a $T = 12$ adaptive trial, a not uncommon size of clinical trials in rare diseases or neonatal populations. We used binary outcomes, with two treatments and expected rewards drawn uniformly from

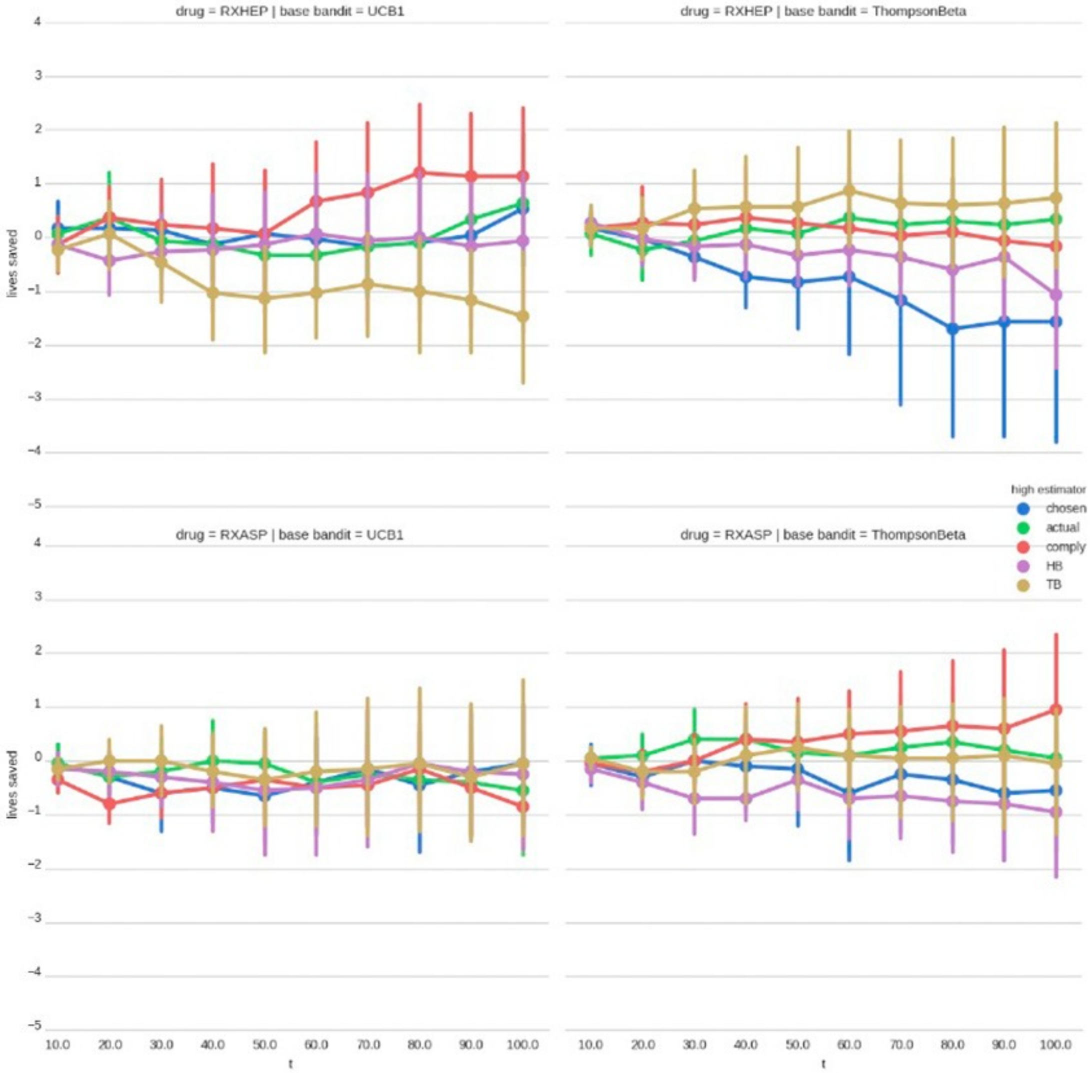

Figure 4.3: Worse than random regret across estimators with naive uses of compliance awareness with simulated data from 100 simulations sampled from a model of a harmful treatment that is profound by selection into treatment.

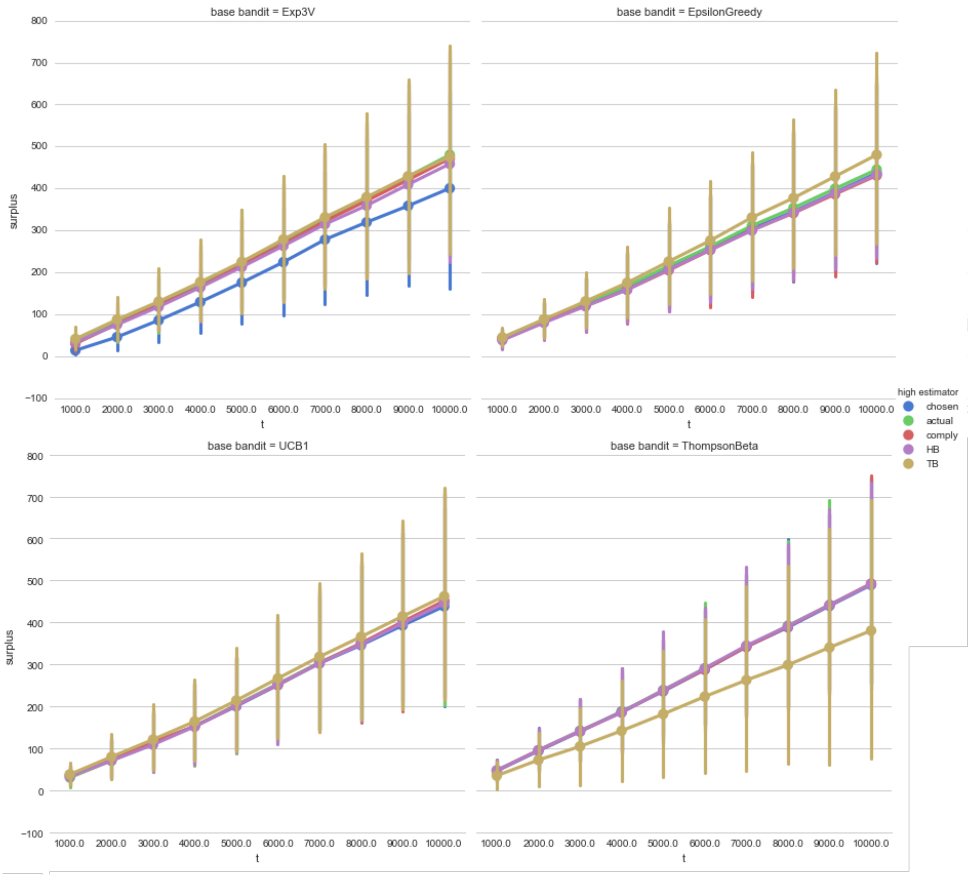

Figure 4.4: Results from 1000 simulations with $T = 12$ and synthetic data:with two treatments and expected rewards drawn uniformly from the unit interval, and compliance uniformly at random.

the unit interval, and compliance drawn uniformly at random. We sampled 1,000 such simulations; results can be seen in Fig. (4.4). While our bounds are vacuous in this setting, it is interesting that there is, on average, an improvement when taking the noncompliance information into account.

4.3.3 Noncompliance for Best Arm

A natural scenario for noncompliance, and one that offers substantial potential, is when the subject is better informed than the algorithm and realizes they know a better alternative. This provides potentially huge practical advantages, especially in situations with very large numbers of a priori low expectation but high variance actions. They allow later subjects to benefit from the information that previous subjects bring to the mechanism, while current compliance unaware algorithms not only waste this information but hurt later subjects by unnecessarily raising the apparent variance of the rewards in the arms (since the chosen arm may indeed be very bad relative to the actual arm).

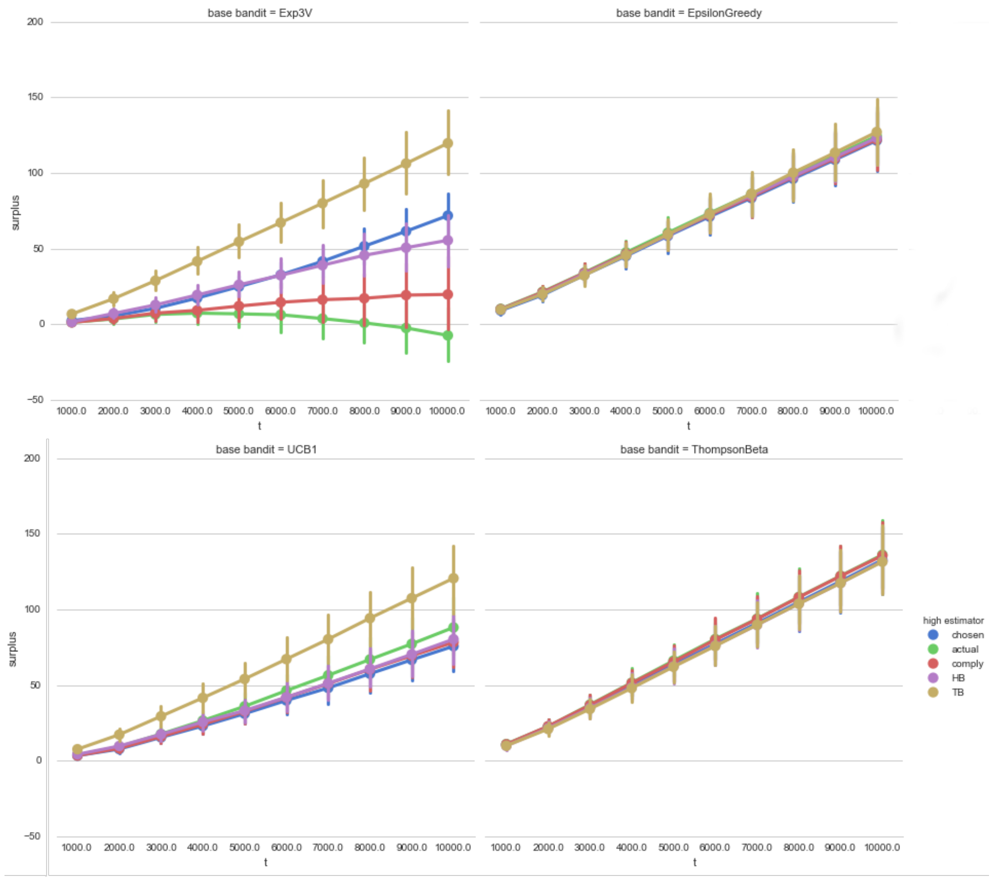

Figure 4.5: Noncompliance for best arm: 100 simulations from synthetic data of $T = 10,000$ with noncompliance proportional to how much better the best arm is than the algorithm selection.

Part II

Normative Implications for Mechanism Design: Advise Auctions

Give the people what they want
when they want it
and they wants it all the time

Parliament
Supergroovalisticprosifunkstication

Chapter 5

Advice Auctions: Subject Freedom and Adviser Incentives for a Single Decision

5.1 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 reward 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. This model is closely related to that studied in *decision markets*, but varies by requiring that the subject of the decision maintain autonomy over what action is taken. This effectively rules out those mechanisms for decision markets that constraint the action taken.

Preserving subject autonomy is an inherently desirable practical design criterion, as it enables no-regret exploration on the part of the subject. The subject remains in control of the decision at all points, so trying using the mechanism cannot restrict their choices or reduce their expected welfare.[1].

To highlight this contrast we term them *advice markets* for the model in general, and in particular propose a *advice auction* mechanism, which auctions off the right to a share of the expected value to the agent of the received advice. This framing, the key contribution of this chapter, allows casting this setting as one of a single unit efficient allocation with interdependent valuations [Milgrom and Weber, 1982; Maskin, 1992; Ausubel et al., 1999; McLean and Postlewaite, 2004; Roughgarden and Talgam-Cohen, 2016; Eden et al., 2018]. 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

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

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 obtain 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 subject's 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 (from the perspective of freedom) results that constrain sequential mechanisms to having full support over actions in order to provide incentives.

Sharing rewards after choosing what decision to advise 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 framing as opposed 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 the subject (i.e. some experts may be more persuasive). If the mechanism output was a choice to advise 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. By having the mechanism select the expert instead of the choice, it also becomes natural to extend to the more practical setting that does not require the mechanism to have access to the valuation functions, but instead rely on the highest bidder to be able to aggregate reported signals effectively.

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

One way to incentivize the experts 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 mispriced. The profit maximizing move for this trader is to lower the price of the optimal action below the true price of the previously mispriced 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 she introduced into the reward of the optimal action. A mechanism is called Bayes Nash Incentive Compatible (BNIC) if there is a equilibrium were every agent reports their signal truthfully and this maximizes their reward in expectation (over the state of the world). The mechanism proposed in [Hanson, 2002] is not BNIC for the experts who provide advice, as witnessed by the example above, and shown in [Othman and Sandholm, 2010; Chen et al., 2014]. 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].

5.1.2 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 as well as their limits. We then consider two practical indirect variations of the procedure, which remove the need for the mechanism to have any knowledge of valuations, and consider sufficient conditions for their efficiency and truthfulness.

5.2 Model

As before, the subject seeks advice on a decision they will take from some finite set of alternatives $\mathcal{A}$. Let $c \in \mathcal{A}$ be the choice that is given as advice to the subject and $a \in \mathcal{A}$ the decision that the subject actually takes. The rest of our model and notation largely follow that of [Eden et al., 2018].

Potential signals for expert $i \in \{1, \ldots, n\}$ form a discrete signal space S_i. Each expert receives a single signal $s_i \in S_i$ which is known only to expert i. Denote a signal profile as $\vec{s} = (s_1, s_2, \ldots, s_n)$. Let $\vec{s}_{-i}$ denote all signals but s_i, and let $(s_i', \vec{s}_{-i})$ denote the profile $\vec{s}$ where s_i has been replaced with s_i'. Each possible signal profile $\vec{s}$ corresponds to an underlying state of the world; this includes inherent physical properties of both the subject and the actions, as well as the subject's probability for choosing a in response to different choices of c.

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 $\vec{s}$. Since r does not depend on the choice of c by the expert, and since at the point the mechanism is run a has not been selected, we use a reduced form value function $v_i : \times_i S_i \to \mathbb{R}_{\geq 0}$ for the value of the rights bundle to expert i, which maps every signal profile to the expected value of a linear share of the reward αr.

Each expert reports a signal $b_i \in S_i$, and the vector of reported signals is denoted $\vec{b} = (b_1, b_2, \ldots, b_n)$.

Mechanisms are pairs (x, p), where $x = (x_1, x_2, \ldots, x_n)$ is a set of allocation functions and $p = (p_1, p_2, \ldots, p_n)$ is a set of payment functions. The allocation functions $x_i : \times_j S_j \to [0, 1]$ map a bid profile $\vec{b}$ to the probability that expert i gets allocated. They hence satisfy $\sum_i x_i(\vec{b}) \leq 1$ for all possible $\vec{b}$. The payment rules $p_i : \times_j S_j \to \mathbb{R}$ map 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.

One cannot hope for truth-telling to be a dominant strategy for the experts. One expert's misreport can cause other experts to also misreport to compensate. Thus the strongest incentive-compatibility (IC) notion that we can hope for in this setting is that truthfulness is an ex-post Nash Equilibrium. That is, it is in every agent i's best interest to report her true signal s_i given that all other agents reported the true signal profile $\vec{s}_{-i}$. Formally, for all $\vec{s} \in \times_j S_j, b_i \in S_i$ we have

$$x_i(\vec{s}) \cdot v_i(\vec{s}) - p_i(\vec{s}) \geq x_i(b_i, \vec{s}_{-i}) \cdot v_i(\vec{s}) - p_i(b_i, \vec{s}_{-i})$$

5.3 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 [Maskin, 1992]. 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 her share of the reward minus her value at the lowest signal she could have misreported and still obtained the allocation give the other reports. More formally:

Mechanism 1. *[Direct Reward Share VCG (DRSA)] The mechanism gives the rights bundle to the expert $i^* = argmax_j\{v_j(\vec{b})\}$ with the highest valuation under the reported signals (a randomly picked one of them, if there are several). That is, the allocation rule is*

$$x_i(\vec{b}) = \begin{cases} 1 & \text{if } i = i^* \\ 0 & \text{otherwise.} \end{cases}$$

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

The experts that were not selected receive no payment, while the selected expert i^* receives her share α of the reward r minus her valuation of the lowest bid $b^*_{i^*}$ (the critical signal) that would have still resulted in expert i^* being selected.

More formally, given $\vec{b}_{-i}$ (the bids for all agents except i), the critical signal *for i is*

$$b_i^* = \min\{b \in S_i \mid x_i(b, \vec{b}_{-i}) = 1\}$$

if this minimum exists (otherwise there is no critical signal for i). The payment rule then is

$$p_i(\vec{b}) = \begin{cases} \alpha r - v_i(b_i^*, \vec{b}_{-i}) & \text{if } i = i^* \\ 0 & \text{otherwise.} \end{cases}$$

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.

5.3.1 Truthfulness with Single Crossing Signals

Definition 6 (Monotonicity). *An allocation function x_i is said to be* monotone *if for every $\vec{b}_{-i}$, $x_i(b_i, \vec{b}_{-i})$ is monotone non-decreasing in b_i.*

Truthful mechanisms can be characterized as follows [Roughgarden and Talgam-Cohen, 2016].

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 [Roughgarden and Talgam-Cohen, 2016] implies the following about the cost charged to a chosen expert [Eden et al., 2018].

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

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 expert's value. We follow the definition in [Eden et al., 2018]:

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 7 (Single-Crossing). *A valuation profile is said to satisfy the single-crossing condition if for every expert i, for any set of other expert's 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 an efficient and truthful mechanism with interdependent values. Hence, without single-crossing, it is impossible to have a truthful advice auction in general.

This procedure for a direct advice elicitation mechanism based on the advice auction procedure was here instantiated using [Maskin, 1992] as the underlying auction mechanism, but the procedure is generic. It could be, for example, instantiated instead with the randomized mechanism of [Eden et al., 2018], and would obtain the approximation properties that algorithm provides in auctions in our advice setting.

5.4 Practical mechanisms: Advice Auctions

The assumptions in [Roughgarden and Talgam-Cohen, 2016] 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 to the valuation functions by the mechanism, and consider two practical alternatives.

5.4.1 A Bid and Signal Reward Sharing Mechanism

The first modification one can make is 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 valuation functions of experts. The challenge this faces is that the signal reports being submitted would not matter, so any signal report is in equilibrium. To correct this and restore strict incentives, we can add a further payment received by all experts, that is also a linear share of the reward (denoted by β). 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 the mechanism gives the rights bundle to the expert i^* with the highest bid:*

$$x_i(\vec{b}) = \begin{cases} 1 & \text{if } i = i^* \\ 0 & \text{otherwise.} \end{cases}$$

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

The experts that were not selected 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:

$$p_i(\vec{b}) = \begin{cases} ((\alpha + \beta)r - b_{i^*-1}) & \text{if } i = i^* \\ \beta r & \text{otherwise.} \end{cases}$$

For this allocation rule to reach efficient allocation, as achieved by the direct mechanism, a stronger assumption on signal structure is needed. If we only require single crossing, the identity of the highest valuation expert can depend on interaction of her signals with those of other experts. 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 mechanism's allocation.

Definition 8 (Single Signal Max Value)**.** *A valuation profile is said to satisfy the* single signal max value property *if highest value expert i^* knows she is the highest value when given their signal, and for any set of other experts' signals $\vec{s}_{-i^*}$ and 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 an expert who has the trust of the subject, understands what they find persuasive, and is sufficiently competent at evaluating the reported signals of others, she may be best able to select c to maximize r no matter what the state of the world is. 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 her worst possible value given the other bids, or anything higher (knowing she will win and be charged according to the second highest price makes her 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 then is 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:*

$$x_i(\vec{b}) = \begin{cases} 1 & \textit{if } i = i^* \\ 0 & \textit{otherwise.} \end{cases}$$

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

The experts that were not selected 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:

$$p_i(\vec{b}) = \begin{cases} ((\alpha + \beta)r - b_{i^*-1}) & \textit{if } i = i^* \\ \beta r & \textit{otherwise.} \end{cases}$$

For some very limited information structures the bidding mechanism still aggregates information efficiently. These information structures correspond to the standard private values settings in which the Vickrey auction is efficient [Vickrey, 1961]. Private values occur when other experts' signals are not informative of the value for an expert of being assigned the rights bundle to make the choice.An example situation is when each expert's signal is only informative about the expected outcome conditional on one action, and there is one expert who is informed about each action.

This kind of indirect mechanism is inherently limited when the values from signals of experts are interrelated (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 expert whose valuation is higher than all other experts' in every state of the world and who knows the subject trust them completely, so if she wins it will be $c = a$. Have a second expert j whose value is 0 in all states of the world[2], but that 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, while knowing the value of s_{treat} allows i^* to select the appropriate choice and results in double the reward $2r$. Since j's value for the rights bundle is always 0, her bid is in equilibrium. There is no prior-free way for the unpersuasive expert to encode her signal into her bid (even through she is incentivized to reveal their signal to get a higher β payment). Notice that there is no limit in the size of the gap in the rewards between the two allocations in general, as in the example above we can replace 2 by any number.

5.5 Conclusion

This chapter shows how to use the bundle of rights perspective on advisers to recast the incentives for decision elicitation from multiple experts into the thriving literature on VCG with interdependent valuations. We then use a result of [Roughgarden and Talgam-Cohen, 2016] for the generalized VCG mechanism of [Maskin, 1992] 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 that the mechanism cannot access the value functions of experts and that signals cannot be transmitted between experts. We give sufficient conditions on the structure of the signals so that the variations of the mechanisms preserve truthfulness and efficiency.

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

Chapter 6

A General Setting and a First Approach

6.0.1 Introduction

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

Our motivating applications in medicine suggest 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. Other applications are conceivable: 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 whether to attempt them or not. They might be able to ask their 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, has focused on a setting with a single decision and multiple advisers (Hanson [2002]; Othman and Sandholm [2010]; Chen et al. [2014]). 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. We consider a sequence of T subjects (patients in the medical motivation), and a fixed set of K advisers (experts) with access to signals about different patients' expected rewards r under different advice c and actual courses of action a. The bounded regret algorithms with compliance awareness introduced in Chapter 3 can be seen as addressing the special case where the experts' signals are known a priori to be uninformative, so $K = 0$ effectively, and thus only the experience can be learned from. 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. 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.

In contrast to the previous chapters' motivation in the literature, in this chapter our focus is first and foremost on constructing a practical mechanism. The reason for this switch is that the setting is natural, and no mechanisms (nor 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 it can be ex-post incentive compatible to take the exploratory actions for subjects, by linking them to suitably large transfers. By introducing randomness into which subjects and which actions have these transfers attached, and into their magnitude, it becomes possible to estimate the underlying causal effect of actions on rewards.

We build up to the main practical design by analyzing two simplified models that illustrate the two key characteristics of our mechanism. The first is 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 in which 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.

6.1 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 contingent transfer payment of $\vec{\gamma}_t$ for each of the possible actions facing the subject.
2. Each expert i reports $b_{t,i}$ to the mechanism. After all reports are received, the mechanism selects an expert i^* to suggest an action c_t.
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 p_i to the experts.

6.1.1 Subjects' Beliefs and Incentives

Previous work on incentive compatible bandits Kremer et al. [2014]; Mansour et al. [2015] has shown that there are distributions of rewards where if all agents were rational and this was common knowledge, some actions can never be explored (assuming only information revelation and no transfers can be used by the mechanism). Namely, 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. That literature has largely been focused on finding information revelation strategies that are optimal, subject to the incentive constraints. Our lottery payments allow us to side step these impossibility results, by providing a reason for exploration for a individual subject.

6.2 A sequence of repeated one-shot-efficient mechanisms is inefficient

One could attempt running the direct mechanism of chapter 5 repeatedly, once for each subject, with the allocation rule of choosing the arm with the maximum posterior expected reward at each step t and using the following payment rule:

$$\pi_i = \sum_1^T \begin{cases} \alpha(r - \mathbb{E}[r^{\hat{(t)}}{}_{-i}]) & \text{if } \hat{c}_{-i}^{(t)} \neq c^{(t)} \\ 0 & \text{otherwise} \end{cases}$$

where as before α parametrizes how much of the reward is shared.

Even when signal structures satisfy the single crossing property, this will not lead to efficient outcomes. 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* $\frac{1}{2}$*. The other arms a priori have a lower expected value of* $\frac{1}{3}$*, but conditional on both agents' signals, one arm has an expected value of* $\frac{2}{3}$ *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 $\left(\frac{2}{3} - \frac{1}{2}\right) T$ relative to the optimal (over all signals) contextual policy in hindsight. Note that the optimal policy with exploration only requires one 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 $\left(\frac{2}{3} - \frac{1}{2}\right)(T - 1) - \left(\frac{1}{2} - \frac{1}{3}\right)$.

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

Definition 9 (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 version of the Chapter 5 Direct Reward Sharing Mechanism (DRSM) in Example 1 has a NE results in the greedy policy. Given that there is no winner's curse due to the signal structure[1], both agents bid their valuations. If the winner of the auction does not choose the safe arm, and instead explores in that period, she receives 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 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. Exploration is not in equilibrium.

One possible attempt to fix this would be to only reveal the outcome to the winning bidder, thus allowing them to use the informational advantage in future rounds' payoffs; in other words, by not having full disclosure. This internalizes the benefits of exploration, 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.

6.3 A Simple Bidding Mechanism with Exploration

To overcome the exploration limitation of the repeated one shot mechanism, a mechanism must enable the decision making expert to exploit 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 she has won the last auction that had a winner (if no bids in an auction meet the reserve price, the owner remains unchanged). Denote by $o_i^{(t)}$ an indicator variable that takes value* 1 *if the agent i*

[1] that is, the winner of the auction who bids her 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.

was the owner *of the choice at time t, and* 0 *otherwise. Further, let $\check{b}^{(t)}$ denote the second highest bid that was placed in round t. The payment rule of this mechanism is*

$$p_i(\vec{b}) = \sum_{t=1}^{T} \begin{cases} \alpha r^{(t)} & \textit{if } o_i^{(t)} = 1 \\ 0 & \textit{otherwise} \end{cases} + \sum_{t=1}^{T} \begin{cases} -\check{b}^{(t)} & \textit{if } o_i^{(t)} = 0 \wedge o_i^{(t+1)} = 1 \\ \check{b}^{(t)} & \textit{if } o_i^{(t)} = 1 \wedge o_i^{(t+1)} = 0 \\ 0 & \textit{otherwise} \end{cases}$$

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 agent i newly becomes the owner; they pay out the second highest bid of that period. When another agent takes over from them as 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 ownership, 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 3. *There is a ex-post Nash Equilibrium under which the BOCM results in sublinear regret in Example 1.*

The optimal contextual policy with exploration has payoff of $\frac{2}{3}T(T-1)+\frac{1}{3}$. The value of the choice for an agent who controls the full sequence and observes the full set of signals thus is $\alpha(\frac{2}{3}T(T-1)+\frac{1}{3})$, 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.

6.4 Choice Incentive Lotteries; Using Transferable Utility as a Source of Unbiased Variation

Mechanism 5 (Lottery for Exploratory Choice (LEC) Mechanism)**.** *At the start of the game, before the first subject arrives, a vector Γ of payments is chosen. In each time period t a new subject arrives, agents receive their signals $\vec{s}_t$ and then send their reports $\vec{b}^{(t)}$. The contingent lottery payments of the subject $\gamma^{(t)}t$ are announced. A one-shot encoding of the reports is used as context in A to select an arm c_t, which leads to a choice a_t being made and a reward r_t being observed.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 $\mathbb{E}(\vec{b}_{-i}, A)$.*

The payment rule for each expert i is as follows:

$$p_i(\vec{b}) = \alpha \left(\sum_{t=1}^{T} r^{(t)} - \mathbb{E}(\vec{b}_{-i}, A) \right)$$

Further, each subject t recieves their lottery payment $\Gamma^{(t)}_{a^{(t)}}$ based on the action $a^{(t)}$ the subject carried out.

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

6.5 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 mechanism's 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 by the unbiased nature of the estimator in the reward function. 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 $\mathbb{A}$ and an unbiased offline evaluation algorithm $\mathbb{E}$.*

A lottery Γ for each action and each subject is drawn, the resulting payment rule is announced. In each period t, all experts report signals $\vec{s}^{(t)}$ and bids $\vec{b}^{(t)}$ 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 $\mathbb{E}(\vec{b}_{-i}, A)$.

The payment for expert i rule is:

$$\begin{aligned} p_i(\vec{b}) =& \alpha \sum_{t=1}^{T} r^{(t)} - \mathbb{E}\left[\sum_{1}^{T} \hat{r}_{-i,t}\right] \\ &+ \sum_{t=1}^{T} \begin{cases} \beta r^{(t)}, & \mathit{if}\, o_i^{(t)} = 1 \\ 0, & \mathit{otherwise} \end{cases} \\ &+ \sum_{t=1}^{T} \begin{cases} -\check{b}^{(t)} & \mathit{if}\, o_i^{(t)} = 0 \land o_i^{(t+1)} = 1 \\ \check{b}^{(t)} & \mathit{if}\, o_i^{(t)} = 1 \land o_i^{(t+1)} = 0 \\ 0 & \mathit{otherwise} \end{cases} \end{aligned}$$

Where α and β are set ex-ante.

Further, each subject t recieves their lottery payment $\Gamma^{(t)}_{a^{(t)}}$ based on the action $a^{(t)}$ the subject carried out.

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

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 when 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 her signal truthfully, the other bidders valuations for being the owner are 0, and the misreporting bidder can appropriate the full value of the α part of the rewards. Thus $\alpha < \beta$ 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 evaluated. If the rewards are not i.i.d., 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 the correct policy given signals a priori cannot fully implement it, since the lottery induces extra variance. This suggests allowing the experts to partially buy out most of the lottery, to reduce the inefficiency it induces when they already have the information required. It is not clear how to prove when there is an 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 already tricky dynamic VCG analysis.

6.6 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, if difficult to analyze.

Chapter 7

Conclusion

We provide a summary of thesis and discuss future directions in which to extend this work in in Section 7.1.

The first part of the thesis introduced compliance information into the bandit setting. Compliance information reflects the choice actually taken by the subject, rather than the algorithm's recommendation. In many cases compliance information can be used to accelerate learning. Further, for situations where the number of arms is large relative to the number of time steps, and the subject's non-compliance is towards the higher reward arms, compliance awareness can make learning possible in problems where otherwise it is not. However, naively incorporating compliance information leads to algorithms with linear regret, as seen in Example **??**. We have therefore developed hybrid strategies that are the first algorithms that incorporate compliance information while maintaining a worst-case guarantee.

The second part of the thesis studied the elicitation of information from self interested experts for decision making. For the single decision case, which had previously been analyzed in the decisions market literature, we introduce the first mechanism for multiple experts with a good ex-post Nash equilibrium that preserves freedom. It achieves incentive compatibility for the experts without requiring randomized strategies with full support from the subject, as previous mechanisms do. Further, entry into the mechanism has positive expectation for both useful experts and subjects. We also introduced the first model for the repeated case, which generalizes both one shot multi-expert elicitation and contextual bandit models.

7.1 Future Work

7.1.1 More Practical Mechanisms for the one shot setting

While many decision do repeat themselves with new subjects, including those that motivate most papers in the one shot decision market literature , others are more unique. The crucial problem is to estimate the counter-factual reward obtained had a different action been taken. Some form of peer elicitation seems inevitable, but it is unclear how to combine this with the signal aggregation aspects of optimal decision advice that are embodied by the bidding mechanism.

7.1.2 Generalization

The final model explored in the thesis has the settings described in the previous two chapters as special cases. The bandit model has a vast range of extensions and special cases in which more specialized algorithms can make substantial advantage. It is interesting to consider the self-interested experts variations of those settings and see if more specialized mechanisms can also do better.

Natural directions for further generalization beyond our final model are:

1. Settings with more supervision. Prediction markets and learning from expert advice are tightly connected. Bandit algorithms and repeated advice elicitation can be seen as two potentially complementary sources of information to aid decisions. However, the two connections are quite different. It could be valuable to extend feedback graphs and other notions that interpolate between the bandit and full supervision settings to take into account incentivized experts as in our model.
2. Incorporating general sources of information that are observed after the action has been chosen by the algorithm. Compliance has a very specific structural relation to the arms selected. It is interesting to explore what can be said generically about what structure the information has to have to be potentially useful to incorporate even when it arrives along with the reward.
3. Costly signal acquisition. It is natural to consider a situation in which experts' signals are costly for the experts to acquire. How can the scale of the rewards that are shared be optimally chosen?
4. Learning valuation functions from bid and signal data. Our mechanism in chapter 5 faces a severe limitation in what signal structures support truthfulness when it does not know the value function of experts. Is it possible to learn the value functions from previous bid and signal reports?

Bibliography

ABERNETHY, J.; CHEN, Y.; AND VAUGHAN, J. W., 2013. Efficient market making via convex optimization, and a connection to online learning. *ACM Transactions on Economics and Computation*, 1, 2 (2013), 12. (cited on page 15)

AGRAWAL, R., 1995. Sample mean based index policies with o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, (1995), 1054–1078. (cited on page 12)

AGRAWAL, S. AND GOYAL, N., 2012. Analysis of Thompson sampling for the multi-armed bandit problem. In *Computational Learning Theory (COLT)*. (cited on pages 30 and 31)

ALON, N.; CESA-BIANCHI, N.; DEKEL, O.; AND KOREN, T., 2015. Online Learning with Feedback Graphs: Beyond Bandits. In *Computational Learning Theory (COLT)*. (cited on page 21)

ATHEY, S. AND SEGAL, I., 2007. Designing efficient mechanisms for dynamic bilateral trading games. *The American economic review*, 97, 2 (2007), 131–136. (cited on page 13)

AUER, P., 2002. Using confidence bounds for exploitation-exploration trade-offs. *JMLR*, 3 (2002), 397–422. (cited on page 20)

AUER, P.; CESA-BIANCHI, N.; AND FISCHER, P., 2002a. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47, 2-3 (2002), 235–256. (cited on page 12)

AUER, P.; CESA-BIANCHI, N.; FREUND, Y.; AND SCHAPIRE, R., 2002b. The non-stochastic multi-armed bandit problem. *SIAM J. Computing*, 32, 1 (2002), 48–77. (cited on pages 20, 25, and 28)

AUMANN, R. J., 1976. Agreeing to disagree. *The annals of statistics*, (1976), 1236–1239. (cited on page 10)

AUSUBEL, L. M. ET AL., 1999. A generalized vickrey auction. *Econo0 metrica*, (1999). (cited on page 45)

BAREINBOIM, E.; FORNEY, A.; AND PEARL, J., 2015. Bandits with Unobserved Confounders: A Causal Approach. In *Adv in Neural Information Processing Systems (NIPS)*. (cited on page 21)

BERG, J. E. AND RIETZ, T. A., 2003. Prediction markets as decision support systems. *Information systems frontiers*, 5, 1 (2003), 79–93. (cited on pages 2 and 16)

BERGEMANN, D. AND VÄLIMÄKI, J., 2010. The dynamic pivot mechanism. *Econometrica*, 78, 2 (2010), 771–789. (cited on page 13)

BEYGELZIMER, A.; LANGFORD, J.; AND PENNOCK, D. M., 2012. Learning performance of prediction markets with kelly bettors. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*, 1317–1318. International Foundation for Autonomous Agents and Multiagent Systems. (cited on page 15)

BOUTILIER, C., 2012. Eliciting forecasts from self-interested experts: scoring rules for decision makers. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, 737–744. International Foundation for Autonomous Agents and Multiagent Systems. (cited on pages 2 and 16)

BUBECK, S., 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5, 1 (2012), 1–122. (cited on page 12)

BUBECK, S. AND SLIVKINS, A., 2012. The best of both worlds: stochastic and adversarial bandits. In *Computational Learning Theory (COLT)*. (cited on page 21)

CESA-BIANCHI, N. AND LUGOSI, G., 2006. *Prediction, learning, and games*. Cambridge university press. (cited on page 10)

CESA-BIANCHI, N.; LUGOSI, G.; AND STOLTZ, G., 2006. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31, 3 (2006), 562–580. (cited on page 11)

CHANG, Y.-H. AND KAELBLING, L. P., 2005. Hedged learning: Regret-minimization with learning experts. In *ICML*. (cited on page 25)

CHAPELLE, O. AND LI, L., 2011. An Empirical Evaluation of Thompson Sampling. In *Adv in Neural Information Processing Systems (NIPS)*. (cited on page 30)

CHEN, Y.; KASH, I. A.; RUBERRY, M.; AND SHNAYDER, V., 2014. Eliciting predictions and recommendations for decision making. *ACM Transactions on Economics and Computation*, 2, 2 (2014), 6. (cited on pages 2, 3, 16, 47, and 55)

CHEN, Y. AND VAUGHAN, J. W., 2010. A new understanding of prediction markets via no-regret learning. In *Proceedings of the 11th ACM conference on Electronic commerce*, 189–198. ACM. (cited on page 15)

Clarke, E. H., 1971. Multipart pricing of public goods. *Public choice*, 11, 1 (1971), 17–33. (cited on page 13)

Della Penna, N.; Celi, L.; and Stretch, R., 2016a. Out of sight: Lack of geographic co-localization of intensive care unit patients and care teams is associated with increased mortality. In *D16. CRITICAL CARE: REDUCING VARIATION IN USUAL CARE*, A6439–A6439. Am Thoracic Soc. (cited on page 7)

Della Penna, N. and Reid, M. D., 2012. Crowd & prejudice: An impossibility theorem for crowd labelling without a gold standard. *Collective Inteligence 2012, arXiv preprint arXiv:1204.3511*, (2012). (cited on page 7)

Della Penna, N.; Reid, M. D.; and Balduzzi, D., 2016b. Compliance-aware bandits. *arXiv preprint arXiv:1602.02852*, (2016). (cited on page 7)

Ebrahim, S.; Sohani, Z. N.; Montoya, L.; Agarwal, A.; Thorlund, K.; Mills, E. J.; and Ioannidis, J. P., 2014. Reanalyses of randomized clinical trial data. *Jama*, 312, 10 (2014), 1024–1032. (cited on page 33)

Eden, A.; Feldman, M.; Fiat, A.; and Goldner, K., 2018. Interdependent values without single-crossing. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 369–369. ACM. (cited on pages 45, 47, 49, and 50)

Frazier, P.; Kempe, D.; Kleinberg, J.; and Kleinberg, R., 2014. Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*, 5–22. ACM. (cited on page 14)

Frongillo, R. and Reid, M. D., 2015. Convergence analysis of prediction markets via randomized subspace descent. In *Advances in Neural Information Processing Systems*, 3034–3042. (cited on page 15)

Frongillo, R. M.; Della Penna, N.; and Reid, M. D., 2012. Interpreting prediction markets: a stochastic approach. In *Advances in Neural Information Processing Systems*, 3266–3274. (cited on pages 7 and 15)

Graepel, T.; Quionero-Candela, J.; Borchert, T.; and Herbrich, R., 2010. Web-scale Bayesian click-through rate prediction for sponsored search and advertising in Microsoft's Bing engine. In *ICML*. (cited on page 19)

Group, I. S. T. C., 1997. The international stroke trial (ist): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. *The Lancet*, 349, 9065 (1997), 1569–1581. (cited on pages 33 and 34)

Groves, T., 1973. Incentives in teams. *Econometrica: Journal of the Econometric Society*, (1973), 617–631. (cited on page 13)

HANSON, R., 2002. Decision markets. *Entrepreneurial Economics: Bright Ideas from the Dismal Science*, (2002), 79–85. (cited on pages 2, 16, 46, 47, and 55)

HÖRNER, J. AND SKRZYPACZ, A., 2016. Learning, experimentation and information design. Technical report, mimeo Yale University. (cited on page 14)

HU, J.; STORKEY, A. J.; ET AL., 2014. Multi-period trading prediction markets with connections to machine learning. In *ICML*, 1773–1781. (cited on page 15)

HUGTENBURG, J. G.; TIMMERS, L.; ELDERS, P.; VERVLOET, M.; AND VAN DIJK, L., 2013. Definitions, variants, and causes of nonadherence with medication: a challenge for tailored interventions. *Patient Prefer Adherence*, 7 (2013), 675–682. (cited on page 20)

JEHIEL, P. AND MOLDOVANU, B., 2001. Efficient design with interdependent valuations. *Econometrica*, 69, 5 (2001), 1237–1259. (cited on page 13)

KATEHAKIS, M. N. AND ROBBINS, H., 1995. Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92, 19 (1995), 8584. (cited on page 12)

KAUFMANN, E.; KORDA, N.; AND MUNOS, R., 2012. Thompson sampling: An asymptotically optimal finite-time analysis. In *ALT*. (cited on pages 30 and 31)

KINATHIL, S.; SANNER, S.; DAS, S.; AND DELLA PENNA, N., 2016. A symbolic closed-form solution to sequential market making with inventory. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. (cited on page 7)

KINATHIL, S.; SANNER, S.; AND DELLA PENNA, N., 2014. Closed-form solutions to a subclass of continuous stochastic games via symbolic dynamic programming. In *Conference on Uncertainty in Artificial Intelligence (UAI)*. (cited on page 7)

KOLLER, D. AND FRIEDMAN, N., 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press. (cited on page 23)

KONG, Y. AND SCHOENEBECK, G., 2016. A framework for designing information elicitation mechanisms that reward truth-telling. *arXiv preprint arXiv:1605.01021*, (2016). (cited on page 15)

KREMER, I.; MANSOUR, Y.; AND PERRY, M., 2014. Implementing the “wisdom of the crowd”. *Journal of Political Economy*, 122, 5 (2014), 988–1012. (cited on pages 2, 3, 14, and 57)

KULESHOV, V. AND PRECUP, D., 2014. Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028*, (2014). (cited on pages 11 and 33)

LAI, T. L. AND ROBBINS, H., 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6 (1985), 4–22. (cited on page 12)

LATTIMORE, T. AND SZEPESVARI, C., 2016. Bandit algorithms. http://banditalgs.com/. (cited on page 12)

MACKENZIE, D., 2008. *An engine, not a camera: How financial models shape markets*. MIT Press. (cited on page 16)

MANSOUR, Y.; SLIVKINS, A.; AND SYRGKANIS, V., 2015. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 565–582. ACM. (cited on pages 2, 3, 6, 14, and 57)

MANSOUR, Y.; SLIVKINS, A.; SYRGKANIS, V.; AND WU, Z. S., 2016. Bayesian exploration: Incentivizing exploration in bayesian games. *arXiv preprint arXiv:1602.07570*, (2016). (cited on pages 2 and 3)

MASKIN, E., 1992. Auctions and privatization. *Privatization*, (1992). (cited on pages 45, 48, 50, and 53)

MCLEAN, R. AND POSTLEWAITE, A., 2004. Informational size and efficient auctions. *The Review of Economic Studies*, 71, 3 (2004), 809–827. (cited on page 45)

MCMAHAN, H. B.; HOLT, G.; AND SCULLEY, D. E., 2013. Ad click prediction: A view from the trenches. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA. (cited on page 19)

MILGROM, P. R. AND WEBER, R. J., 1982. A theory of auctions and competitive bidding. *Econometrica: Journal of the Econometric Society*, (1982), 1089–1122. (cited on page 45)

MILLER, N.; RESNICK, P.; AND ZECKHAUSER, R., 2005. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51, 9 (2005), 1359–1373. (cited on page 15)

MNIH, V.; KAVUKCUOGLU, K.; AND *et al*, 2015. Human-level control through deep reinforcement learning. *Nature*, 518, 7540 (02 2015), 529–533. (cited on page 32)

NASH, J. F. ET AL., 1950. Equilibrium points in n-person games. *Proc. Nat. Acad. Sci. USA*, 36, 1 (1950), 48–49. (cited on page 10)

OSTROVSKY, M., 2012. Information aggregation in dynamic markets with strategic traders. *Econometrica*, 80, 6 (2012), 2595–2647. (cited on page 14)

OTHMAN, A. AND SANDHOLM, T., 2010. Decision rules and decision markets. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, 625–632. International Foundation for Autonomous Agents and Multiagent Systems. (cited on pages 2, 3, 16, 46, 47, 48, and 55)

PARKES, D. C. AND SINGH, S., 2003. An mdp based approach to online mechanism design. In *V Proceedings of 17th Annual Conference on Neural Information Processing Systems (NIPS 03)*. (cited on page 13)

PRELEC, D., 2004. A bayesian truth serum for subjective data. *science*, 306, 5695 (2004), 462–466. (cited on page 15)

ROBBINS, H., 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58, 5 (1952), 527–535. (cited on page 10)

ROUGHGARDEN, T. AND TALGAM-COHEN, I., 2016. Optimal and robust mechanism design with interdependent values. *ACM Transactions on Economics and Computation (TEAC)*, 4, 3 (2016), 18. (cited on pages 45, 49, 50, and 53)

SABATÉ, E., 2003. *Adherence to long-term therapies: evidence for action*. World Health Organization. (cited on page 19)

SANDERCOCK, P. A.; NIEWADA, M.; CZŁONKOWSKA, A.; ET AL., 2011. The international stroke trial database. *Trials*, 12, 1 (2011), 1–7. (cited on page 34)

SAVAGE, L. J., 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66, 336 (1971), 783–801. (cited on page 15)

SELDIN, Y. AND SLIVKINS, A., 2014. One Practical Algorithm for Both Stochastic and Adversarial Bandits. In *ICML*. (cited on page 21)

THOMPSON, W. R., 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25, 3/4 (1933), 285–294. (cited on pages 2, 11, and 30)

VAPNIK, V. AND VASHIST, A., 2009. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22 (2009), 544–557. (cited on page 21)

VICKREY, W., 1961. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16, 1 (1961), 8–37. (cited on pages 13 and 52)

VRIJENS, B.; DE GEEST, S.; HUGHES, D. A.; PRZEMYSLAW, K.; DEMONCEAU, J.; RUPPAR, T.; DOBBELS, F.; FARGHER, E.; MORRISON, V.; LEWEK, P.; ET AL., 2012. A new taxonomy for describing and defining adherence to medications. *British journal of clinical pharmacology*, 73, 5 (2012), 691–705. (cited on page 20)

ZHANG, P. AND CHEN, Y., 2014. Elicitability and knowledge-free elicitation with peer prediction. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 245–252. International Foundation for Autonomous Agents and Multiagent Systems. (cited on page 15)