Sequential Decision Making With Information Asymmetry

Jiarui Gan
University of Oxford, UK

Rupak Majumdar
Max Planck Institute for Software Systems (MPI-SWS), Kaiserslautern, Germany

Goran Radanovic
Max Planck Institute for Software Systems (MPI-SWS), Saarbrücken, Germany

Adish Singla
Max Planck Institute for Software Systems (MPI-SWS), Saarbrücken, Germany

Abstract

We survey some recent results in sequential decision making under uncertainty, where there is an information asymmetry among the decision-makers. We consider two versions of the problem: persuasion and mechanism design. In persuasion, a more-informed principal influences the actions of a less-informed agent by signaling information. In mechanism design, a less-informed principal incentivizes a more-informed agent to reveal information by committing to a mechanism, so that the principal can make more informed decisions. We define Markov persuasion processes and Markov mechanism processes that model persuasion and mechanism design into dynamic models. Then we survey results on optimal persuasion and optimal mechanism design on myopic and far-sighted agents. These problems are solvable in polynomial time for myopic agents but hard for far-sighted agents.

2012 ACM Subject Classification Theory of computation → Models of computation

Keywords and phrases Bayesian persuasion, Automated mechanism design, Markov persuasion processes, Markov mechanism processes, Myopic agents

Digital Object Identifier 10.4230/LIPIcs.CONCUR.2022.4

Category Invited Talk

Funding *Rupak Majumdar*: This research was funded in part by the Deutsche Forschungsgemeinschaft project 389792660 TRR 248–CPEC.

1 Introduction

Sequential decision making under uncertainty is a fundamental problem in modeling and analysis of systems. In concurrency theory and formal verification, many such models have been studied extensively. In *Markov decision processes* (MDPs), a single agent observes the state of the world, picks an action, and the new state of the world is determined by an uncertain transition relation. The goal of the agent is to find a policy that optimizes her expected utility, usually over an infinite horizon. In *partially observable* MDPs (POMDPs), the state is no longer perfectly observed; the agent gets a signal about the state of the world and has to find a policy with partial information about the world. Finally, in *stochastic games*, multiple agents play against each other. The objectives of the agents can be zero-sum (the two player, purely adversarial situation) or non-zero sum. The complexity landscape of these models have been studied extensively. Broadly, full information settings (MDPs) are polynomial time solvable [14], partial observation settings are undecidable [20, 4], and games are intermediate in complexity [6, 13, 5].

© Jiarui Gan, Rupak Majumdar, Goran Radanovic, and Adish Singla;
licensed under Creative Commons License CC-BY 4.0

33rd International Conference on Concurrency Theory (CONCUR 2022).
Editors: Bartek Klin, Sławomir Lasota, and Anca Muscholl; Article No. 4; pp. 4:1–4:18
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

There are a number of applications of sequential decision making where the interaction between agents and the world involve *information asymmetry*. These are games of imperfect information on one side, in which one agent influences the behavior of another by selectively signaling additional information about the state of the world, or incentivizes the other to provide accurate information about the world. These models have been largely studied in the economics and artificial intelligence literature, as problems of *persuasion* or of *mechanism design*, but have not received attention in the concurrency theory literature.

In persuasion (also called information design), a knowledgeable principal knows some aspects of the state of the world and interacts with an agent who does not. However, only the agent has the capacity to take an action. Since the objectives of the principal and the agent may be misaligned, the agent may not do the principal's bidding. The goal of the principal is to strategically reveal information about the world, through a process of *signaling*, so that the agent's actions optimize the principal's own interests.

In mechanism design, one or more agents know the state of the world; the principal can take an action based on the report from the agents. Again, it is possible that the agent misrepresents the state of the world to optimize their own payoff. The goal of the principal is to design incentive mechanisms to elicit the agent's private information about the state of the world, so as to make more informed decisions.

If the principal and the agent are completely aligned in their utilities, the signals or the mechanisms involve revealing the unknown information; the more interesting case is when the objectives are misaligned. Persuasion and mechanism design problems in the sequential setting involve partial information and strategic interaction but have not been considered in the concurrency theory literature. The goals of this paper are to provide an introduction to these models, describe some basic results and pointers to the literature, and to point out open problems in the domain.

Persuasion. Kamenica and Gentzkow [17] introduced a fundamental and very influential model of *Bayesian persuasion* as a formal model for persuasion problems. They consider a two player game between a principal and an agent. The players share a common prior on the state of the world, but only the principal observes the realization. The principal commits to a signaling strategy before the game starts. On observing the realization, the principal signals the agent and the agent picks an action based on the signal. They each receive a payoff dependent on the realized state of the world and the action. Kamenica and Gentzkow characterize the optimal signaling strategy of the principal.

Since the publication of this work [17], Bayesian persuasion has seen many applications in the field of economics and algorithmic game theory. The basic model has also been extended in many ways. We refer the reader to the comprehensive survey [16] for pointers to the literature. Our focus in this survey is on algorithmic problems in dynamic models, where persuasion is performed repeatedly over time. Work in this direction is fairly new [12, 23, 15, 26].

Mechanism Design. In automated mechanism design, we consider models where the roles of the players are reversed: now, the principal is the receiver of information, and commits to a mechanism that specifies the action they will take upon receiving each signal. The agent is the signal sender and, knowing the principal's mechanism, sends signals optimally in response. Intuitively, to design a good mechanism requires balancing between the goals of eliciting more information from the agent and of acting optimally based on the elicited information. The principal aims to find a mechanism that maximizes their overall utility from the interaction.

The model follows the line of work on automated mechanism design, initiated by Conitzer and Sandholm [7, 8]. It is shown in their work that the problem of computing an optimal mechanism is NP-hard in general, in settings that allow restrictions to be placed on what signals can be sent given the true state of the world. We consider models without such restrictions, which are less expressive in this regard but arguably also captures a wide range of applications. Following the seminal work of Conitzer and Sandholm, variants of their model have been proposed and studied [24, 18, 19, 27, 28]. A recent work of Zhang and Conitzer [28] introduces a dynamic model of automated mechanism design, and studies some fundamental algorithmic questions for this model. There is a broader literature on various forms of dynamic mechanism design in economics. We refer the reader to the comprehensive surveys [22, 2].

Dynamic Models. Most problems in persuasion and mechanism design were studied in the one-shot setting. More recently, *dynamic* versions of these models have been introduced to capture persuasion and mechanism design in sequential decision making [12, 23, 26, 3, 15, 28]. Dynamic models generalize MDPs from a single agent to settings in which a principal and an agent interact, with an information asymmetry between them. The game is played over a state space. In addition, there is an external parameter, chosen from a known prior distribution, that is the source of information asymmetry. In a *Markov persuasion process* (MPP), in each step, the principal observes the realizations of the external parameters and signals the agent to elicit a favorable action. The agent picks the action based on the current state of the MPP and the signal, both the principal and the agent receive a reward, and the game moves to the next state based on a probabilistic transition relation. In a *Markov mechanism process* (MMP), in each step, the agent observes the realizations of the external parameters. The agent is incentivized by the principal to provide true information by a mechanism – a precommitment to act in a certain way. The agent reports the external parameters as a best response to the precommitment, and the principal chooses an action based on this information. Both principal and agent receive a reward, and the game moves to a new state based on the current state and the chosen action.

Dynamic models of persuasion and mechanism design are special cases of stochastic games of incomplete information [1, 25] and many fundamental insights in characterizing optimal strategies carry over. By focusing on the subclass of games with persuasion and mechanism design as the central aspects, we are able to provide specialized algorithmic results that are applicable to many problems of practical interest.

Myopic and Far-sighted Agents. A new aspect in the study of dynamic persuasion and mechanism design problems is the nature of the agent. In models of concurrency, we usually assume that all players are long-lived, that is, survive throughout the game. In MPPs and MMPs, we distinguish between *far-sighted* and *myopic* agents. A far-sighted agent is long lived and optimizes their expected utility in the long run – it is the "usual case" we study in concurrent games.

In contrast, a myopic agent is short-lived, and only interested in optimizing the payoff in the current stage of the game. In a game with myopic agents, the long-lived principal interacts with a sequence of independent myopic agents, one for each time step. As we shall see, decision problems often become easier when we deal with myopic agents.

There is good motivation for studying myopic agents in both persuasion and mechanism design problems. As an example of a dynamic persuasion problem with myopic agents, consider a ride-sharing app, where the application developer is the long-running principal, and users of the app can be seen as myopic agents. The users are interested in optimizing their current commute times. The application developer may have a different goal, that of minimizing congestion. The application developer may provide a noisy signal about the status of roads to persuade the commuters to choose routes that minimize overall congestion.

As an example of a dynamic mechanism design problem with myopic agents, consider a firm that consults with a research organization to decide upon a product strategy [28]. Each year, the research organization presents its market research. The firm decides to invest in certain directions based on the reports. The goal of the firm is to have a strong long term business while keeping costs low. On the other hand, the research organization's goal can be myopic – to generate as much revenue from the firm each year, by possibly misrepresenting market conditions. A mechanism in this case is a compensation strategy of the firm that ensures each research report truthfully represents market conditions.

Current Status. In this article, we summarize some recent decidability and complexity results for MPPs and MMPs [15, 28, 26]. We shall see that the principal's optimal signaling strategy and optimal mechanism design problems can be solved in polynomial-time in the infinite horizon setting, against myopic agents. In contrast, we can only show some intractability for these problems against far-sighted agents but a complete characterization remains open.

We have collected the basic results of persuasion and mechanism design in this article and we hope it can serve as the starting point for investigating the specification and verification of dynamic models with information asymmetry in the context of concurrency theory.

2 Persuasion: Principal Observes, Agents Act

2.1 One-shot Bayesian Persuasion

The basic persuasion model by Kamenica and Gentzkow [17] considers two agents: Sender and Receiver (who are the principal and agent, respectively). Receiver has a utility function $u(a, \omega)$ that depends on her action a from a fixed set A of available actions, as well as a state of the world ω from a set Ω (chosen by nature). Sender has a utility function $v(a, \omega)$, that also depends on the receiver's action a and ω. Both players share a common prior μ_0 on Ω. Sender does not influence the world by picking an action himself, but influences Receiver by transmitting a *signal*.

A signal, broadly construed, is some information about the state of the world that Sender can transmit to Receiver. Let G be a sufficiently large space of *signal realizations*. A signaling strategy $\pi : \Omega \to \Delta(G)$ of Sender is a map that associates each realization of the state of the world to a distribution over G. Using π, Sender will send a signal g to Receiver with probability $\pi(\omega, g)$ whenever ω is observed. Intuitively, the strategy specifies a statistical relationship between the state of the world and Receiver's observed data.

For example, one simple signaling strategy is to always reveal the true information, which always sends a deterministic signal g_ω associated with the observed ω (i.e., g_ω is a message saying "The current state of the world is ω.", and $\pi(\omega)$ is a Dirac delta distribution at g_ω). In contrast, if the same signal is sent irrespective of the realized ω, i.e., $\pi(\omega) = \pi(\omega')$ for all $\omega, \omega' \in \Omega$, then the signaling strategy is completely uninformative: observing the signal gives Receiver no information about the current realization of ω.

The steps of Bayesian persuasion are as follows.

1. Sender and Receiver share a prior μ_0.
2. Sender picks a signaling strategy $\pi : \Omega \to \Delta(G)$ and commits to it; Receiver observes π.

3. Nature picks $\omega \sim \mu_0$ and reveals it to Sender.
4. Sender picks $g \sim \pi(\omega)$ according to his commitment.
5. Receiver observes the realized g, and takes some action $a \in A$ (we describe below how the action is chosen).
6. Sender receives utility $v(a, \omega)$ and Receiver receives $u(a, \omega)$.

Upon receiving a signal g, Receiver updates her posterior belief about the state of the world using the Bayes' rule, whereby the following conditional probability is derived:

$$\Pr(\omega \mid g, \pi) = \frac{\mu_0(\omega) \cdot \pi(\omega, g)}{\sum_{\omega' \in \Omega} \mu_0(\omega') \cdot \pi(\omega', g)}. \tag{1}$$

Receiver picks an action $a^*(\Pr(\cdot \mid g, \pi))$ that maximizes $\mathbb{E}_{\omega \sim \Pr(\cdot|g,\pi)}[u(a, \omega)]$. By convention, we assume that Receiver breaks ties in favor of Sender when there are multiple optimal actions. Given the choice of Receiver, Sender solves

$$\max_{\pi \in \Pi} \mathbb{E}_{\omega \sim \mu_0} \mathbb{E}_{g \sim \pi(\omega)} v(a^*(\Pr(\cdot \mid g, \pi)), \omega) \tag{2}$$

to optimize her expected utility, where Π is the set of all signaling strategies.

The optimization problem seems complicated at a first glance, since the space G of signals can be arbitrary, and the choice of π influences the utility of Sender both by influencing how the signal realizations are distributed and by influencing the action that Receiver picks based on the signal realization. However, we shall show that the problem can be reduced to an optimization problem of a simpler form.

2.2 The Revelation Principle and Action Advice

According to a standard argument via the revelation principle [21, 17], we can restrict attention to signaling strategies in the form of *action advice* without any loss of generality. Specifically, for any signaling strategy in an arbitrary space of signals, there exists an equivalent strategy π that uses only a finite set $G_A := \{g_a : a \in A\}$ of signal realizations, where each signal g_a corresponds to an action $a \in A$. With the signal g_a, Sender "advises" Receiver to play a. Moreover, we can additionally ensure that π is *incentive compatible* (IC), which means that Receiver is indeed incentivized to take the corresponding action a upon receiving g_a. Formally, π ensures that

$$\mathbb{E}_{\omega \sim \Pr(\cdot|g_a,\pi)} u(a, \omega) \geq \mathbb{E}_{\omega \sim \Pr(\cdot|g_a,\pi)} u(a', \omega)$$

for all $a' \in A$, or equivalently:

$$\sum_{\omega \in \Omega} \Pr(\omega \mid g_a, \pi) \cdot (u(a, \omega) - u(a', \omega)) \geq 0 \quad \text{for all } a' \in A. \tag{3}$$

In other words, π signals which action Receiver should take and it is designed in a way such that Receiver cannot be better off deviating from the advised action with respect to the posterior belief. (Again, we assume that Receiver breaks ties in favor of Sender, which means following the advice in this case.) We call a signaling strategy that only uses signals in G_A an *action advice*, and call it an IC action advice if it also satisfies (3).

In case A and Ω are finite sets, we can write Sender's optimization problem as a linear program (LP) with variables $\{\pi(\omega, g_a) \mid \omega \in \Omega, a \in A\}$ (see, e.g., [11, 10]):

$$\begin{aligned}
\max \quad & \sum_{\omega\in\Omega}\sum_{a\in A} \mu_0(\omega)\cdot\pi(\omega,g_a)\cdot v(a,\omega) && && (4)\\
\text{subject to} \quad & \sum_{\omega\in\Omega} \mu_0(\omega)\cdot\pi(\omega,g_a)\cdot(u(a,\omega)-u(a',\omega)) \geq 0, && \text{for } a,a' \in A && (5)\\
& \sum_{a\in A}\pi(\omega,g_a) = 1, && \text{for } \omega\in\Omega && (6)\\
& \pi(\omega,g_a) \geq 0, && \text{for } \omega\in\Omega, a\in A && (7)
\end{aligned}$$

The variable $\pi(\omega, g_a)$ denotes the conditional probability of recommending action a when the state of the world is ω. The LP maximizes the expected utility of Sender over the joint distribution of ω and a, subject to incentive compatibility (i.e., (5), where $\Pr(\omega \mid g_a, \pi)$ in (3) is replaced by $\mu_0(\omega) \cdot \pi(\omega, g_a)$ according to (1)). Since linear programming can be solved in polynomial time, the above formulation shows that one-shot persuasion can be solved in polynomial time when the actions and the external parameters are given explicitly.

▶ **Theorem 2.1** [11]. *Sender's optimization problem can be solved in polynomial time in $|A|$ and $|\Omega|$.*

More generally, Kamenica and Gentzkow showed a characterization of the optimal function for compact action spaces and payoff functions that are continuous in the action [17].

Given a signal, each signal realization g_a induces a posterior belief $\mu_a \in \Delta(\Omega)$. The marginal probability of signal realization g_a is $\Pr[g_a] = \sum_{\omega\in\Omega} \mu_0(\omega)\cdot\pi(\omega, a)$ and the posterior distribution $\Pr(\omega \mid g_a, \pi) = \frac{\mu_0(\omega)\cdot\pi(\omega,g_a)}{\Pr[g_a]}$.

Thus, we can think of a feasible solution of the LP as a distribution over posteriors (an element of $\Delta(\Delta(\Omega))$), one per signal realization, whose expectation equals the prior μ_0 (such a distribution of posteriors is called *Bayes plausible*). Thus, if μ_0 is represented as a point in the simplex $\Delta(\Omega)$, then the signal corresponds to writing μ_0 as a convex combination of posterior distributions in $\Delta(\Omega)$. The incentive compatibility constraints ensure that action a is preferred by Receiver on the posterior distribution on Ω induced by a.

Each posterior distribution $\mu \in \Delta(\Omega)$ is associated with a preferred action $a^*(\mu)$ for Receiver, i.e., the action that maximizes $\mathbb{E}_{\omega\sim\mu}u(a, \omega)$. We can plot Sender's utility as a function $V : \Delta(\Omega) \to \mathbb{R}$ of the posterior: $V(\mu) = \mathbb{E}_{\omega\sim\mu}v(a^*(\mu), \omega)$. Define $\text{cav}(V)$ as the *concavification* of V: the pointwise smallest concave function that is an upper bound for V. Equivalently,

$$\text{cav}(V)(\mu) = \sup\{z : (\mu, z) \in \text{co}(V)\} \tag{8}$$

where $\text{co}(V)$ is the convex hull of the graph of V. The convex hull $\text{co}(V)$ is the set of pairs (μ, z) such that if the prior is μ, there exists a signal with value z. Thus, $\text{cav}(V)(\mu_0)$ is the optimal utility that Sender can achieve when the prior is μ_0.

This is a very general result, holding also for compact spaces of actions and continuous reward functions. It also follows from an older result on games of imperfect information studied by Aumann and Maschler [1].

Note that if V is already concave, then Sender reveals no information. For example, in the zero-sum case when the utility functions of Sender and Receiver sum to zero, V is concave. On the other hand, if the Sender and Receiver have completely aligned utility functions, V is convex and Sender reveals all information.

In general, we do not know how to compute the concavification of an arbitrary function $V : \Delta(\Omega) \to \mathbb{R}$. If the graph of V is semi-algebraic (defined by a Boolean combination of polynomial inequalities), we can use techniques from the theory of reals, using the characterization that the concavification of V evaluated at μ is $\sup\{z \mid (\mu, z) \in \mathrm{co}(V)\}$ and that $\mathrm{co}(V)$ is a semi-algebraic set if the graph of V is semi-algebraic.

The above LP assumes that the world is given explicitly. In case the world is given symbolically, as valuations to a set of variables, it still works if we assume that the prior has small (polynomial-size in the size of the problem) support. The optimization problem can sometimes be solved even when this assumption is not true. Consider the case in which $u(a, \omega)$ and $v(a, \omega)$ are real-valued random variables that can be arbitrarily correlated. We say actions are independent if $u(a) = u(a, \omega)$ and $u(a') = u(a', \omega)$ are independent random variables for distinct actions $a \neq a'$, and the same is true for $v(a) = v(a, \omega)$ and $v(a') = v(a', \omega)$. Then, the distribution μ_0 is fully specified by the marginal distribution of the pair $(u(a), v(a))$ for each action a. We assume that each action's marginal distribution has finite support, and refer to each element of the support as a *type*.

Dughmi and Xu [11] show that in case $u(a)$ and $u(a')$ are independent and identically distributed (IID) for $a \neq a'$, and $v(a)$ and $v(a')$ are also IID, Sender's optimization problem can be solved in polynomial time in the number of actions n and the number of types m. This is non-trivial, since the above LP has exponentially many (m^n) states of the world. On the other hand, the problem becomes #P-hard if the distributions are arbitrary.

2.3 Examples

Prosecution. Kamenica and Gentzkow [17] give an example of Bayesian persuasion in a courtroom setting. A prosecutor (Sender) is trying to convince a judge (Receiver) that a defendant is guity. When the defendant is guilty, revealing all the evidence will help the prosecutor, but when the defendant is innocent, revealing all the evidence will likely hurt the prosecutor's case. Kamenica and Gentzkow show that when the prosecutor and the judge are rational Bayesian, a prosecutor can organize their argument to increase the probability of conviction.

Concretely, assume that the judge has two actions: *acquit* or *convict*. The states of the world correspond to the defendant's status: *guilty* or *innocent*. The judge gets a utility of 1 for choosing the just action (convict the guilty and acquit the innocent) and utility 0 for the unjust action. The prosecutor gets a utility of 1 if the judge convicts and 0 otherwise – regardless of the defendant's status. Assume that the prior $\Pr[\textit{guilty}] = 0.3$ is common knowledge.

We model the prosecutor's possible investigations into the case as distributions $\pi(\cdot \mid \textit{guilty})$ and $\pi(\cdot \mid \textit{innocent})$. The prosecutor has to pick π and truthfully report the realization to the judge (the *commitment* step). (It is required by law that the prosecutor cannot hide evidence, even it makes a conviction unlikely.)

If there is no communication, e.g., if the investigation is completely uninformative, the judge always acquits, since innocence is more likely than guilt according to the prior. If the investigation is fully informative, i.e., reveals the defendant's status with probability 1, then the judge convicts 30% of the time. However, suppose that the prosecutor picks an investigation as follows:

$$\pi(\textit{acquit} \mid \textit{innocent}) = \frac{4}{7} \qquad \pi(\textit{acquit} \mid \textit{guilty}) = 0$$

$$\pi(\textit{convict} \mid \textit{innocent}) = \frac{3}{7} \qquad \pi(\textit{convict} \mid \textit{guilty}) = 1$$

This constitutes an IC action advice for the judge. Notice that the judge convicts with probability 60% (Bayes' rule!). This is true even though the judge knows that 70% of defendants are innocent and even though the judge is fully aware that the prosecutor's advice (the signal) is designed to maximize the probability of conviction!

Traffic Control. Das et al. [9] describe a simple example of persuasion to improve congestion in uncertain traffic conditions. Imagine a traffic network with two paths between a source and an origin. Travel time on Path I is independent of the number of agents using it, but depends on an uncertain state of nature (e.g., Path I is a highway that is prone to repair). Travel time on Path II depends on the number of agents taking the path: the more agents take the path, the more time it takes. The goal of Sender (a social planner) is to signal the state of Path I to the agents so that the congestion on Path II is reduced to a social optimum. Hence, each agent is an individual Receiver, and they are modeled as non-atomic players, who individually is a zero-measure and have negligible influence to the system (but collectively their influence integrates).

Let us be more precise. There are two paths P_1 and P_2, and the state of the world is $\omega \in \{0,1\}$, both states are equally likely. The travel times are given by $c(P_1) = \omega$ and $c(P_2) = \frac{1}{3} + 2s$. Agents seek to minimize their travel costs.

If Sender can mandate how everyone drives, the socially optimum cost is calculated as follows. If $\omega = 0$, everyone uses P_1 and the total cost is zero. If $\omega = 1$, the socially optimum move is to send $\frac{1}{6}$ of the agents to P_2 so that the aggregate cost is $\frac{17}{18}$. Thus, the expected aggregate travel cost is $\frac{17}{36}$.

Suppose Sender provides exact information. Then, when $\omega = 1$, agents will crowd P_2 until the costs of the two paths are equalized: $\frac{1}{3} + 2s = 1$, or $s = \frac{1}{3}$. The aggregate cost is 1 and therefore the expected aggregate cost is $\frac{1}{2}$, which is worse than the optimum.

Now consider the following signaling strategy.

$$\pi(\text{take } P_1 \mid \omega = 0) = 1 \qquad \pi(\text{take } P_2 \mid \omega = 0) = 0$$
$$\pi(\text{take } P_1 \mid \omega = 1) = \frac{5}{6} \qquad \pi(\text{take } P_2 \mid \omega = 1) = \frac{1}{6}$$

(Namely, when $\omega = 1$, we send the message "take P_1" to 5/6 of the agents and "take P_2" to the rest.) Then, when $\omega = 0$, everyone takes P_1 and the cost is zero. When $\omega = 1$, we expect $\frac{1}{6}$ fraction to go on P_2. The overall expected cost is the same as the optimal: $\frac{17}{36}$. Thus, the social planner persuades some fraction of people to take P_1.

We observe that the signal is incentive compatible. Upon seeing the advice "take P_1" the expectation of the cost of P_1 is

$$\Pr[\omega = 1 \mid \text{take } P_1] \cdot 1 = \frac{\frac{5}{6}}{\frac{5}{6} + 1} = \frac{5}{11}$$

(where $\Pr[\omega = 1 \mid \text{take } P_1]$ is the posterior belief given π) and the expectation of the cost of P_2 is

$$\frac{1}{3} + 2\left(\Pr[\omega = 0 \mid \text{take } P_1] \cdot 0 + \Pr[\omega = 1 \mid \text{take } P_1] \cdot \frac{1}{6}\right) = \frac{16}{33} > \frac{5}{11}$$

Thus, the agent should pick P_1. Similarly, on seeing "take P_2", the expectation of P_1 is 1 and the expectation of P_2 is $\frac{2}{3} < 1$. Thus, the agent should again pick P_2.

2.4 Markov Persuasion Processes

We now extend the model of Bayesian persuasion to the sequential setting. Our formal model, called *Markov persuasion processes* (MPP),[1] is an MDP with reward uncertainties, given by a tuple

$$\mathcal{M} = \langle S, A, P, \Omega, (\mu_s)_{s\in S}, u, v \rangle \tag{9}$$

that represents the repeated interaction between Sender and Receiver.

Similar to a standard MDP, S is a finite state space; A is a finite action space available to Receiver; $P : S \times A \times S \to [0, 1]$ is the transition dynamics of the state. When the environment is in state s and Receiver takes action a, the state transitions to s' with probability $P(s, a, s')$; both Sender and Receiver are aware of the state throughout. Meanwhile, rewards are generated for both Sender and Receiver, and are specified by the reward functions $u : S \times \Omega \times A \to \mathbb{R}$ and $v : S \times \Omega \times A \to \mathbb{R}$, respectively. That is, unlike in a standard MDP, the rewards in our setting also depend on an external parameter $\omega \in \Omega$ (akin to the state of the world in the basic model). This parameter captures an additional layer of uncertainty of the environment. At each state $s \in S$, we assume that the parameter follows a distribution $\mu_s \in \Delta(\Omega)$ and is drawn anew every time the state changes. μ_s is common prior knowledge shared between Sender and Receiver, but only Sender has access to the realization of ω.

Since the actions are taken only by Receiver, Sender does not directly influence the state. As in Bayesian persuasion, Sender influences Receiver's action by signaling. We only consider *Markovian signaling strategies*, whereby signals only depend on the current state (independent of the history). As in the one-shot case, a revelation theorem argument shows that Sender only needs to consider IC action advice at each state.

Formally, a signaling strategy $\pi = (\pi_s)_{s\in S}$ of Sender consists of a function $\pi_s : \Omega \to \Delta(G_A)$ for each state $s \in S$. Sender will commit to a strategy before the start of play. In every step, upon observing the realization of the external parameter ω, Sender will send an action advice sampled from $\pi_s(\omega)$ when the current state is s.

2.5 Optimal Signaling Problem

Similarly to the one-shot setting, we take Sender's point of view and investigate the problem of optimal signaling strategy design: given $\mathcal{M}$, find a signaling strategy π that maximizes Sender's (discounted) cumulative reward. The cumulative reward is defined as

$$\mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \cdot v(s_t, a_t, \omega_t) \,\middle|\, \mathbf{z}, \pi, P\right], \tag{10}$$

where $\mathbf{z} = (z_s)_{s\in S}$ is the distribution of the starting state, $\gamma \in [0, 1)$ is a discount factor, T is a given horizon, and the expectation is taken over the trajectory $(s_t, \omega_t, a_t)_{t=0}^{T}$ induced by $\mathbf{z}$, the signaling strategy π, and the dynamics P. If T is finite, we call the problem the *finite horizon* setting, and if T is infinite, we call the setting *infinite horizon*.

Finally, we introduce a behavioral model for Receiver. We will consider two major types of Receivers – *myopic* and *far-sighted*. A myopic Receiver only cares about their instant reward in each step, whereas a far-sighted Receiver considers the cumulative reward with respect to a discount factor $\tilde{\gamma} > 0$ (which need not be equal to γ).

[1] The nomenclature comes from [26].

In summary, the game proceeds as follows. At the beginning, Sender commits to a signaling strategy π and announces it to Receiver. Then in each step, an external parameter $\omega \sim \mu_s$ is drawn (by nature) according to the state $s \in S$ of the MPP; Sender observes $\omega \in \omega$, samples an action advice $g \sim \pi_s(\omega)$, and sends g to Receiver. Receiver receives g, updates their belief about ω and decides an action $a \in A$ to take. Sender receives $v(s, \omega, a)$ and Receiver receives $u(s, \omega, a)$. The state then transitions to $s' \sim P(s, a, \cdot)$, which both players observe. The game proceeds until the horizon T (or forever, if $T = \infty$).

2.6 Solving the Optimal Signaling Problem

2.6.1 Myopic Receiver

We first consider the case where Receiver is myopic. In this case, Receiver aims to maximize her reward in each individual step. Upon receiving a signal g in state s, Receiver takes a best action $a \in A$, which maximizes the immediate expected reward $\mathbb{E}_{\omega \sim \Pr(\cdot|g,\pi_s)} u(s, a, \omega)$. Think of a myopic Receiver as a sequence of "short-lived" Receivers, one for each time step. Receiver in step t plays a one-shot Bayesian persuasion game with Sender, collects their reward, and disappears.

We consider the problem of computing an optimal signaling strategy in an infinite-horizon MPP ($T = \infty$) with a myopic Receiver. We call this problem $\textsc{OptimalSignaling}_\infty$-$\textsc{myopic}$.

▶ **Theorem 2.2** [15]. $\textsc{OptimalSignaling}_\infty$-$\textsc{myopic}$ *can be solved in polynomial time.*

The proof of Theorem 2.2 is via a reduction from the problem to linear programming. The approach is as follows.

We can easily characterize the outcome of an IC action advice π: at each state s, since Receiver is incentivized to follow the advice, with probability $\phi_s^\pi(\omega, a) := \mu_s(\omega) \cdot \pi_s(\omega, g_a)$ they will take action a when the realized external parameter is ω. Thus, ϕ_s^π is a distribution over $\Omega \times A$.

We then define the following set $\mathcal{A}_s \subseteq \Delta(\Theta \times A)$, which contains all such distributions that can be induced by some IC action advice:

$$\mathcal{A}_s = \{\phi_s^\pi \in \Delta(\Omega \times A) : \pi \text{ is an IC action advice}\}.$$

We can now view the problem facing Sender as an (single-agent) MDP

$$\mathcal{M}^* = \langle S, (\mathcal{A}_s)_{s \in S}, P^*, v^* \rangle,$$

where S is the same state space in $\mathcal{M}$; $\mathcal{A}_s$ defines an (possibly infinite) action space for each s; the transition dynamics $P^* : S \times \Delta(\Omega \times A) \times S \to [0, 1]$ and reward function $v^* : S \times \Delta(\Omega \times A) \to \mathbb{R}$ are such that

$$P^*(s, \mathbf{x}, s') = \mathbb{E}_{(\omega,a) \sim \mathbf{x}} P(s, a, s') \quad \text{and} \quad v^*(s, \mathbf{x}) \quad = \mathbb{E}_{(\omega,a) \sim \mathbf{x}} v(s, a, \omega)$$

for any $\mathbf{x} \in \mathcal{A}_s$. Namely, $\mathcal{M}^*$ is defined as if Sender can choose actions (which are (ω, a) pairs) freely from $\mathcal{A}_s$, whereas the choice is actually realized through persuasion. A policy σ for $\mathcal{M}^*$ maps each state s to an action $\mathbf{x} \in \mathcal{A}_s$, and it corresponds to an IC action advice π in $\mathcal{M}$, with $\phi_s^\pi = \sigma(s)$ for all s. The problem of designing an optimal action advice then translates to computing an optimal policy for $\mathcal{M}^*$.

The standard approach to computing an optimal policy for an MDP is to compute a value function $V : S \to \mathbb{R}$ that satisfies the Bellman equation:

$$V(s) = \max_{\mathbf{x} \in \mathcal{A}_s} \left[v^*(s, \mathbf{x}) + \gamma \cdot \sum_{s' \in S} P^*(s, \mathbf{x}, s') \cdot V(s') \right] \quad \text{for all } s \in S.$$

There exists a unique solution to the above system of constraints, from which an optimal policy can be extracted. The solution is posed as the following linear program over variables $\{V(s) : s \in S\}$:

$$\begin{aligned} \min \quad & \sum_{s\in S} z_s \cdot V(s) && && (11) \\ \text{subject to} \quad & V(s) \geq v^*(s,\mathbf{x}) + \gamma \cdot \sum_{s'\in S} P^*(s,\mathbf{x},s') \cdot V(s') && \text{for all } s \in S, \mathbf{x} \in \mathcal{A}_s && (12) \end{aligned}$$

The optimal value of this LP directly gives the cumulative reward of optimal policies under a given initial state distribution $\mathbf{z}$.

The issue with this LP formulation is that there may be infinitely many constraints as (12) must hold for all $\mathbf{x} \in \mathcal{A}_s$. This is unlike MDPs with a finite action space, where there are a finite number of constraints, one for each action.

Gan et al. [15] show that LP (11) can nevertheless be solved in polynomial time by using the ellipsoid method. The key to this approach is to implement the *separation oracle* in polynomial time. For any given value assignment of the variables (in the above LP, values of $V(s)$), the oracle should decide correctly whether all the constraints of the LP are satisfied or not and, if not, output a violated one.

Implementing the separation oracle for the LP requires solving $\max_{\mathbf{x}\in\mathcal{A}_s} v^*(s,\mathbf{x}) + \gamma \cdot \sum_{s'\in S} P^*(s,\mathbf{x},s') \cdot V(s') - V(s)$ for all $s \in S$: by checking if the maximum value is positive, we can identify if (12) is violated for some $\mathbf{x} \in \mathcal{A}_s$. Indeed, the set of IC action advice can be characterized by (5)–(7). Hence, we obtain the following LP implementation of the separation oracle, where $\{x(\omega,a) : \omega \in \Omega, a \in A\}$ and $\{\pi_s(\omega,g_a) : \omega \in \Omega, a \in A\}$ are the variables.

$$\begin{aligned} \max \quad & v^*(s,\mathbf{x}) + \gamma \cdot \sum_{s'\in S} P^*(s,\mathbf{x},s') \cdot V(s') - V(s) \\ \text{s.t.} \quad & x(\omega,a) = \mu_s(\omega) \cdot \pi_s(\omega,g_a) && \text{for all } \omega \in \Omega, a \in A, s \in S \\ & \sum_{\omega\in\Omega} \mu_s(\omega) \cdot \pi_s(\omega,g_a) \cdot (u(s,a,\omega) - u(s,a',\omega)) \geq 0, && \text{for } a, a' \in A, s \in S \\ & \sum_{a\in A} \pi_s(\omega,g_a) = 1, && \text{for } \omega \in \Omega, s \in S \\ & \pi_s(\omega,g_a) \geq 0, && \text{for } \omega \in \Omega, a \in A, s \in S \end{aligned}$$

Since the ellipsoid method runs in polynomial time, the tractability of OPTIMALSIGNALING$_\infty$-MYOPIC follows immediately. By exploiting the duality of linear programming, one can provide a different, "direct" encoding into a linear programming problem as well (see [15]).

▶ Remark 2.3 Finite Horizon. When the horizon is finite, one can set up the Bellman equation and evaluate it by backward induction. Each step in the process solves a one-shot persuasion problem using the linear programming formulation. This gives a polynomial time algorithm when the time horizon is given in unary. Wu et al. [26] study several variants of this problem, as well as the setting of reinforcement learning.

▶ Remark 2.4. In the *reachability problem* for Markov persuasion processes, there is a subset of marked states and Sender receives a unit reward if and only if one of these states is reached along a trajectory. The reachability problem asks what is the expected probability that the subset is reached. The above linear programming formulation can be used to solve the reachability problem against myopic Receivers. Since the reachability problem is at the core of model checking logics on MDPs, we should be able to build up a logic on Markov persuasion processes and obtain efficient model checking algorithms in case of myopic agents. We leave the design of appropriate logics and model checking, as well as the computation of optimal signals for omega-regular properties, as future work.

2.6.2 Far-sighted Receiver

A far-sighted (FS) Receiver looks beyond the immediate reward and optimizes the cumulative reward

$$\mathbb{E}\left[\sum_{t=0}^{T} \tilde{\gamma}^t \cdot u(s_t, a_t, \omega_t) \,\middle|\, \mathbf{z}, \pi, P\right], \tag{13}$$

where, as in (10), $\mathbf{z} = (z_s)_{s \in S}$ is the distribution of the starting state, $\tilde{\gamma} \in [0, 1)$ is a discount factor possibly different from Sender's discount factor, T is the horizon, and the expectation is taken over the trajectory $(s_t, a_t, \omega_t)_{t=0}^T$ induced by the initial distribution $\mathbf{z}$, the signaling strategy π, and the dynamics P.

When facing an FS Receiver, we cannot define a set $\mathcal{A}_s$ independently for each state. Sender needs to take a global view and aim to induce Receiver to use a *policy* that benefits Sender. We consider the problem of optimal signaling strategy design in an infinite horizon setting against an FS Receiver, called OptimalSignaling$_\infty$-FS.

At this point, we know very little about the decidability and complexity of this problem or a characterization of optimal strategies. For example, we know that Sender can do better with history-dependent signaling. We also know that the problem is hard.

▶ **Theorem 2.5** [15]. *Assuming that* P $\neq$ NP, OptimalSignaling$_\infty$-FS *does not admit any polynomial-time* $\frac{1}{\lambda^{1-\epsilon}}$*-approximation algorithm for any constant* $\epsilon > 0$*, where* λ *is the number of states* $s \in S$ *in which the prior distribution* μ_s *is non-deterministic (i.e., supported on at least two external parameters). This holds even when* $|\Theta| = 2$ *and the discount factors* $\gamma, \tilde{\gamma} \in (0, 1)$ *are fixed.*

The proof of Theorem 2.5 is via a reduction from the Maximum Independent Set problem, which is known to be NP-hard to approximate [29].

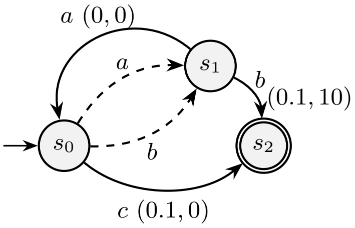
2.6.3 Advice-myopic Receiver

Between the tractable (myopic) Receivers and the intractable (FS) Receivers lie the *advice-myopic* Receivers. An advice-myopic (AM) Receiver accounts for the cumulative future rewards just as an FS Receiver, but behaves myopically in ignoring the future signals of Sender. In other words, an AM Receiver always assumes that Sender will disappear in the next step and relies only on their own prior knowledge to estimate any future payoff.

▶ **Theorem 2.6** [15]. OptimalSignaling$_\infty$-AM *is solvable in polynomial time.*

The idea is that, since an AM Receiver does not consider future signals, their future reward is independent of Sender's signaling strategy. One can compute the future payoff in polynomial time by fixing the uninformative signal for Sender and solving the resulting MDP. This payoff is added to the reward function of the AM Receiver, but now we can consider Receiver to be myopic since the future payoffs have been taken into account.

The interest in AM Receiver is that an optimal signaling policy of Sender assuming an AM Receiver can be used to define a strategy against an arbitrary FS Receiver. The idea is to provide a threat: if Receiver ever deviates from the action advice, Sender will forever provide only uninformative signals. One can show that this threat strategy enables Sender to get an expected payoff that is at least as much against any AM Receiver.



	External parameter	
	ω_a	ω_b
a	$(1, 1)$	$(-1, 0)$
b	$(-1, 0)$	$(1, 1)$

Figure 1 A simple example from [15].

The threat strategy uses one bit of memory (to remember if Receiver had deviated from the advice). However, this threat-based strategy may not be an optimal one-memory strategy. Indeed, for any positive integer k, the problem of computing an optimal k-memory strategy against FS Receivers is inapproximable (via an adapted version of the reduction for proving Theorem 2.5). In contrast, in the myopic and advice-myopic settings, since Receiver's behavior is Markovian, the optimal signaling strategies we designed remain optimal even when we are allowed to use memory-based strategies.

2.7 Example

Figure 1 shows a simple example to distinguish myopic, far-sighted, and advice-myopic Receivers. In the MPP, Sender wishes to reach s_2 while maximizing rewards. Transitions are deterministic. Each edge is labeled with the corresponding action and (in the brackets) rewards for Receiver and Sender, respectively. The rewards for state-action pairs (s_0, a) and (s_0, b) (dashed edges) also depend on the 2-valued state of the world $\{\omega_a, \omega_b\}$, as specified in the table. The state of the world is sampled uniformly at random at each step. Assume discount factor $\frac{1}{2}$ both for Sender and for Receivers.

With no signaling, Receiver will always take action c in s_0, so Sender will obtain payoff 0. Sender can reveal information about the external parameter to attract Receiver to move to s_1. If Receiver is myopic, Sender can reveal full information, which leads to Receiver moving to s_1, taking action b, and ending in s_2. As a result, Sender obtains payoff 6.

However, if Receiver is FS, this strategy will not work. Receiver will loop between s_0 and s_1, resulting in overall payoff $4/3$ for Sender. To improve, Sender can choose to be less informative in s_0, e.g., advising Receiver to take the more profitable action 10% of the time and a uniformly sampled action in $\{a, b\}$ the remaining 90% of the time. Receiver will move to s_1 under this signaling, breaking ties in favor of Sender. Sender's expected payoff is 5.55.

Alternatively, Sender can also use the following threat-based strategy, which again yields a payoff of 6. Sender always reveals the true information in s_0, advises Receiver to take b in s_1, and threatens to stop providing any information if Receiver does not follow the advice. The outcome of this strategy coincides with how an advice-myopic Receiver behaves. Such a Receiver will choose b at s_1 as future disclosures are not considered.

2.8 Extensions to the Model

In our model of MPPs thus far, the external parameter ω is picked independently at each step. We can envision a more general model, in which the external parameter also evolves according to a stochastic process. For example, we can assume that the external parameter evolves according to a Markov chain. Such extensions have been studied [12, 23], but we do not know of any general algorithmic results.

One can show that against myopic Receivers, the optimal value can be calculated on a Markov process on the space of distributions in $S \times \Delta(\Omega)$; the initial belief is the initial distribution of the state of the world and the value function maps beliefs to values and is the fixpoint of a functional mapping beliefs to beliefs. The functional is a contraction map on a suitable topological space, and therefore the fixpoint exists and is unique. While one can approximately evaluate the fixpoint numerically, we do not know how to characterize the complexity of the decision problem. Since the belief space $\Delta(\Omega)$ is infinite, we can no longer set up a (finite) linear programming problem nor argue about termination of the iterations.

3 Mechanism: Agent Observes, Principal Acts

A dual scenario of persuasion is one where Receiver is the principal and Sender is the agent. In this case Receiver can commit to a mechanism to influence Sender's signaling behavior. A mechanism $\sigma : G \to \Delta(A)$ is a map from Sender's signal space G to a distribution over the action space A, which specifies how Receiver will act, upon receiving each signal from Sender.

3.1 One-shot Mechanism Design

In the one-shot setting, the steps in this scenario are as follows.

1. Sender and Receiver share a prior μ_0.
2. Receiver picks a mechanism $\sigma : G \to \Delta(A)$ and commits to it; Sender observes σ.
3. Nature picks $\omega \sim \mu_0$ and reveals it to Sender.
4. Sender observes ω and sends a signal $g \in G$ (we describe below how this signal is chosen).
5. Receiver observes g and takes an action $a \sim \sigma(g)$ according to her commitment.
6. Sender receives utility $v(a, \omega)$ and Receiver receives $u(a, \omega)$.

In Step 4, as a rational player, Sender best-responds to the mechanism σ, sending a signal so that the action taken by Receiver in Step 5 maximizes Sender's payoff in expectation. Namely, the following signal is sent:

$$g \in \arg\max_{g \in G} \mathbb{E}_{a \sim \sigma(g)} v(a, \omega). \tag{14}$$

Here, one subtlety, similar to the one in the persuasion setting, is that there is actually no predefined signal space or one that is agreed upon between the two players, so the mechanism is not well-defined if Sender picks a signal outside of G. The revelation principle then comes in again, which now says that it is without loss of generality to consider *direct mechanisms*, whereby the signal space is restricted to a finite set $G_\Omega := \{g_\omega : \omega \in \Omega\}$; each signal $g_\omega \in G_\Omega$ corresponds to a realization of the state of the world. In other words, the interaction in Step 4 can be viewed as an information elicitation process, where Receiver asks Sender: what is the realization of the external parameter? Sender answers ω by sending the corresponding signal g_ω.

Specifically, given an arbitrary mechanism $\sigma : G \to \Delta(A)$, an equivalent mechanism $\varsigma : G_\Omega \to \Delta(A)$ can be constructed by letting $\varsigma(g_\omega) = \sigma\left(f(\omega)\right)$ for all $\omega \in \Omega$, where $f : \Omega \to G$ is a map defined by (14) (by fixing an arbitrary tie-breaking rule to select g in case there are multiple optimal signals). It is not hard to see that ς induces an equivalent signaling behavior of Sender and the same payoffs in Step 6. Moreover, it also elicits truthful information from Sender, incentivizing Sender to send g_ω whenever the realization is ω.

In summary, the revelation principle indicates that it is without loss of generality to consider mechanisms that are both direct and IC. Given this result, the problem of computing an optimal mechanism for Receiver can be formulated as the following LP with variables $\{\sigma(g_\omega, a) : \omega \in \Omega, a \in A\}$, i.e., $\sigma(g_\omega, a)$ is the probability of Receiver taking action a upon receiving g_ω.

$$\begin{aligned} \max \quad & \sum_{\omega\in\Omega}\sum_{a\in A} \mu_0(\omega)\cdot\sigma(g_\omega, a)\cdot u(a,\omega) && && (15)\\ \text{subject to} \quad & \sum_{a\in A}\sigma(g_\omega, a)\cdot v(a,\omega) \geq \sum_{a\in A}\sigma(g_{\omega'}, a)\cdot v(a,\omega), && \text{for } \omega,\omega' \in A && (16)\\ & \sum_{a\in A}\sigma(g_\omega, a) = 1, && \text{for } a \in A && (17)\\ & \sigma(g_\omega, a) \geq 0 && \text{for } \omega\in\Omega, a\in A && (18) \end{aligned}$$

The formulation takes a form symmetric to LP (4). The first constraint requires σ to be IC.

3.2 Markov Mechanism Process

Moving to the dynamic setting, we consider the same MDP $\mathcal{M} = \langle S, A, P, \Omega, (\mu_s)_{s\in S}, u, v\rangle$ as in (9). Receiver commits to a state-dependent mechanism $\sigma_s : G_\Omega \to \Delta(A)$. At every step, both players observes the state s of $\mathcal{M}$, and nature samples an external parameter $\omega \sim \mu_s$. Sender observes ω and sends a signal g to Receiver. Receiver plays an action $a \sim \sigma_s(g)$ according to a pre-committed state-dependent mechanism. Consequently, rewards $v(s, a, \omega)$ and $u(s, a, \omega)$ are generated for the players, and $\mathcal{M}$ transitions to a next state $s' \sim P(s, a, \cdot)$. We ask the infinite-horizon optimal mechanism design problem from Receiver's prospective. In what follows we present a polynomial-time algorithm for this problem when Sender is myopic. The approach is similar to the LP-based algorithm for OptimalSignaling$_\infty$-myopic.

3.3 Optimal Mechanism Design for Myopic Sender

Call the optimal mechanism design problem OptimalMechanism$_\infty$-myopic when Sender is myopic.

▶ **Theorem 3.1.** OptimalMechanism$_\infty$-myopic *can be solved in polynomial time.*

The proof is similar to that of Theorem 2.2. We reduce the problem to linear programming and use the ellipsoid method. We define the set of possible outcomes of a direct IC mechanism σ as follows:

$$\mathcal{A}_s = \{\phi_s^\sigma \in \Delta(\Omega \times A) : \sigma \text{ is a direct IC mechanism}\},$$

where ϕ_s^σ is a distribution with $\phi_s^\sigma(\omega, a) := \mu_s(\omega)\cdot\sigma_s(g_\omega, a)$ being the probability that Receiver takes action a while the realized external parameter is ω. The problem facing Receiver then reduces to an (single-agent) MDP $\mathcal{M}^* = \langle S, (\mathcal{A}_s)_{s\in S}, P^*, u^*\rangle$, where the transition dynamics P^* and reward function u^* are such that $P^*(s, \mathbf{x}, s') = \mathbb{E}_{(\omega,a)\sim\mathbf{x}} P(s, a, s')$, and $u^*(s, \mathbf{x}) = \mathbb{E}_{(\omega,a)\sim\mathbf{x}} u(s, \omega, a)$ for any $\mathbf{x} \in \mathcal{A}_s$. The follwoing LP, similar to LP (11), is then devised to compute an optimal mechanism (with variables $\{V(s) : s \in S\}$).

$$\begin{aligned} \min \quad & \sum_{s\in S} z_s\cdot V(s) && && (19)\\ \text{subject to} \quad & V(s) \geq u^*(s,\mathbf{x}) + \gamma\cdot\sum_{s'\in S} P^*(s,\mathbf{x},s')\cdot V(s') && \text{for } s\in S, \mathbf{x}\in\mathcal{A}_s && (20) \end{aligned}$$

The separation oracle of this LP can further be implemented by solving the following LP for all $s \in S$, where $\{x(\omega, a) : \omega \in \Omega, a \in A\}$ and $\{\sigma_s(g_\omega, a) : \omega \in \Omega, a \in A\}$ are the variables.

$$\begin{aligned}
\max \quad & u^*(s, \mathbf{x}) + \gamma \cdot \sum_{s' \in S} P^*(s, \mathbf{x}, s') \cdot V(s') - V(s) & \\
\text{subject to} \quad & x(\omega, a) = \mu_s(\omega) \cdot \sigma_s(g_\omega, a) & \text{for } s \in S, \omega \in \Omega, a \in A \\
& \sum_{a \in A} \sigma_s(g_\omega, a) \cdot v(s, a, \omega) \geq \sum_{a \in A} \sigma_s(g_{\omega'}, a) \cdot v(s, a, \omega), & \text{for } \omega, \omega' \in A, s \in S \\
& \sum_{a \in A} \sigma_s(g_\omega, a) = 1, & \text{for } a \in A, s \in S \\
& \sigma_s(g_\omega, a) \geq 0 & \text{for } \omega \in \Omega, a \in A, s \in S
\end{aligned}$$

▶ Remark 3.2. Zhang and Conitzer [28] studied a more general model in the finite-horizon case and consider history-dependent mechanisms. In their model, Receiver cannot observe the state of the MDP and has to rely on Sender to make observations; essentially, the state is equivalent to the external parameter in our model but follows a stochastic process. They show that the problem is polynomial time solvable in the finite horizon case when Sender is myopic, but NP-hard to approximate when Sender is FS. They also characterize optimal mechanisms and show that the optimal mechanism against an FS sender depends on the history of state-action trajectories, as well as the current state. Note that the NP-hardness does not imply the hardness of the optimal mechanism design problem we defined against an FS Sender, where the goal is to compute an optimal Markov mechanism for an infinite horizon, whereas the external parameter is sampled independently in each step. We leave the complexity of this problem open for future work.

4 Conclusion

We have described some basic results in the theory of Markov decision processes with information asymmetry. We show that in the two settings we study, persuasion and mechanism design, one can obtain optimal signaling policy and optimal mechanism design in polynomial time against myopic agents. As we point out throughout the article, many algorithmic questions in these domains remain open. While the models have been applied to many problems in economics and game theory, their applications to system design have not been explored so far. We hope our article can act as a starting point for studying these models and their algorithmic properties, in the context of concurrency theory and system design.

References

1 Robert J. Aumann and Michael B. Maschler. *Repeated Games with Incomplete Information*. MIT Press, 1995.

2 Dirk Bergemann and Juuso Välimäki. Dynamic mechanism design: An introduction. *Journal of Economic Literature*, 57(2):235–74, 2019.

3 Andrea Celli, Stefano Coniglio, and Nicola Gatti. Private Bayesian persuasion with sequential games. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'20)*, pages 1886–1893, 2020.

4 Krishnendu Chatterjee, Martin Chmelik, and Mathieu Tracol. What is decidable about partially observable markov decision processes with ω-regular objectives. *J. Comput. Syst. Sci.*, 82(5):878–911, 2016. doi:10.1016/j.jcss.2016.02.009.

5 Krishnendu Chatterjee and Thomas A. Henzinger. A survey of stochastic ω-regular games. *J. Comput. Syst. Sci.*, 78(2):394–413, 2012. doi:10.1016/j.jcss.2011.05.002.

6 Anne Condon. The complexity of stochastic games. *Inf. Comput.*, 96(2):203–224, 1992. doi:10.1016/0890-5401(92)90048-K.

7 Vincent Conitzer and Tuomas Sandholm. Complexity of mechanism design. In Adnan Darwiche and Nir Friedman, editors, *Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence (UAI'02)*, pages 103–110. Morgan Kaufmann, 2002.

8 Vincent Conitzer and Tuomas Sandholm. Self-interested automated mechanism design and implications for optimal combinatorial auctions. In *Proceedings of the 5th ACM Conference on Electronic Commerce (EC'04)*, pages 132–141, 2004.

9 Sanmay Das, Emir Kamenica, and Renee Mirka. Reducing congestion through information design. In *Proceedings of the 55th Allerton Conference on Communication, Control, and Computing*, pages 1279–1284, 2017.

10 S. Dughmi. Algorithmic information structure design. *ACM SIGecom Exch.*, 15(2):2–24, 2017.

11 Shaddin Dughmi and Haifeng Xu. Algorithmic Bayesian persuasion. In Daniel Wichs and Yishay Mansour, editors, *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 412–425. ACM, 2016. doi:10.1145/2897518.2897583.

12 J. Ely. Beeps. *American Economic Review*, 107(1):31–53, 2017.

13 Kousha Etessami and Mihalis Yannakakis. On the complexity of Nash equilibria and other fixed points. *SIAM J. Comput.*, 39(6):2531–2597, 2010. doi:10.1137/080720826.

14 J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Springer-Verlag, 1997.

15 Jiarui Gan, Rupak Majumdar, Goran Radanovic, and Adish Singla. Bayesian persuasion in sequential decision-making. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, (AAAI'22)*. AAAI Press, 2022.

16 Emir Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, 11:249–272, 2019.

17 Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

18 Andrew Kephart and Vincent Conitzer. Complexity of mechanism design with signaling costs. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS'15)*, pages 357–365, 2015.

19 Andrew Kephart and Vincent Conitzer. The revelation principle for mechanism design with reporting costs. In *Proceedings of the 2016 ACM Conference on Economics and Computation (EC'16)*, pages 85–102, 2016.

20 Omid Madani, Steve Hanks, and Anne Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artif. Intell.*, 147(1-2):5–34, 2003. doi:10.1016/S0004-3702(02)00378-8.

21 Roger B. Myerson. Incentive compatibility and the bargaining problem. *Econometrica*, 47(1):61–73, 1979.

22 Alessandro Pavan. Dynamic mechanism design: Robustness and endogenous types. In *Advances in Economics and Econometrics: Eleventh World Congress*, volume 1, pages 1–62, 2017.

23 J. Renault, E. Solan, and N. Vieille. Optimal dynamic information provision. *Games and Economic Behavior*, 104:329–349, 2017.

24 Tuomas Sandholm, Vincent Conitzer, and Craig Boutilier. Automated design of multistage mechanisms. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*, volume 7, pages 1500–1506, 2007.

25 Sylvain Sorin. *A First Course on Zero-Sum Repeated Games*. Springer, 2008.

26 Jibang Wu, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Zhuoran Yang, Michael I. Jordan, and Haifeng Xu. Sequential information design: Markov persuasion process and its efficient reinforcement learning. *CoRR*, abs/2202.10678, 2022. arXiv:2202.10678.

27 Hanrui Zhang, Yu Cheng, and Vincent Conitzer. Automated mechanism design for classification with partial verification. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI'21)*, volume 35(6), pages 5789–5796, 2021.

28 Hanrui Zhang and Vincent Conitzer. Automated dynamic mechanism design. *Advances in Neural Information Processing Systems (NeurIPS'21)*, 34, 2021.

29 David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing (STOC'06)*, pages 681–690. Association for Computing Machinery, 2006.