Bayesian Persuasion in Sequential Decision-Making

Jiarui Gan, Rupak Majumdar, Goran Radanovic, Adish Singla

Max Planck Institute for Software Systems
{jrgan,rupak,gradanovic,adishs}@mpi-sws.org

Abstract

We study a dynamic model of Bayesian persuasion in sequential decision-making settings. An informed principal observes an external parameter of the world and advises an uninformed agent about actions to take over time. The agent takes actions in each time step based on the current state, the principal's advice/signal, and beliefs about the external parameter. The action of the agent updates the state according to a stochastic process. The model arises naturally in many applications, e.g., an app (the principal) can advice the user (the agent) on possible choices between actions based on additional real-time information the app has. We study the problem of designing a signaling strategy from the principal's point of view. We show that the principal has an optimal strategy against a myopic agent, who only optimizes their rewards locally, and the optimal strategy can be computed in polynomial time. In contrast, it is NP-hard to approximate an optimal policy against a far-sighted agent. Further, if the principal has the power to threaten the agent by not providing future signals, then we can efficiently compute a threat-based strategy. This strategy guarantees the principal's payoff as if playing against an agent who is far-sighted but myopic to future signals.

1   Introduction

Uncertainty is prevalent in models of sequential decision making. Usually, an agent relies on prior knowledge and Bayesian updates as a basic approach to dealing with uncertainties. In many scenarios, a knowledgeable *principal* has direct access to external information and can reveal it to influence the agent's behavior. For example, a navigation app (the principal) normally knows about the global traffic conditions and can inform a user (the agent), who then decides a particular route based on the app's advice. The additional information can help improve the quality of the agent's decision-making. Meanwhile, by strategically revealing the external information, the principal can also persuade the agent to act in a way beneficial to the principal.

We study the related persuasion problem in a dynamic environment. In a static setting, the interaction between the principal and the agent is modeled by *Bayesian persuasion* (Kamenica and Gentzkow 2011), where the principal uses their information advantage to influence the agent's strategy in a one-shot game, by way of signaling. In this paper,

we extend this setting to include interaction in an infinite-horizon Markov decision process (MDP), where rewards incurred depend on the state of the environment, the action performed, as well as an external parameter sampled from a known prior distribution at each step. The principal, who cannot directly influence the state, observes the realization of this external parameter and signals the agent about their observation. The agent chooses to perform an action based on the state and the signal, and the action updates the state according to a stochastic transition function. Both the principal and the agent aim to optimize their own rewards received in the course of the play.

If the objectives of the principal and the agent are completely aligned, the principal should reveal true information about the external parameter, so the more interesting case is when they are misaligned. For example, a user of an navigation app only wants to optimize their commute times but the app may want to incentivize the user to upgrade to a better service, or to increase traffic throughput when the app is provided by a social planner. We consider two major types of agents—*myopic* and *far-sighted*—and investigate the problem of optimal signaling strategy design against them. A myopic agent optimizes their payoff locally: in each step, they take an action that will give them the highest immediate reward. It can model a large number of "short-lived" agents each appearing instantly in a system (e.g., users of a ride-sharing app or an E-commerce website). A far-sighted agent, on the other hand, optimizes their long-run cumulative reward and considers future information disclosure.

We show that, in the myopic setting, an optimal signaling strategy for the principal can be computed in polynomial time through a reduction to linear programming. On the other hand, in the case of a far-sighted agent, optimal signaling strategy design becomes computationally intractable: if P≠NP, there exists no polynomial time approximation scheme. Our proof of computational intractability is quite general, and extends to showing the hardness of similar principal-agent problems in dynamic settings.

To work around the computational barrier, we focus on a special type of far-sighted agents who are *advice-myopic*. An advice-myopic agent optimizes their cumulative reward over time based on the history of information disclosures, but does not assume that the principal will continue to provide information in the future. We expect such behavior to

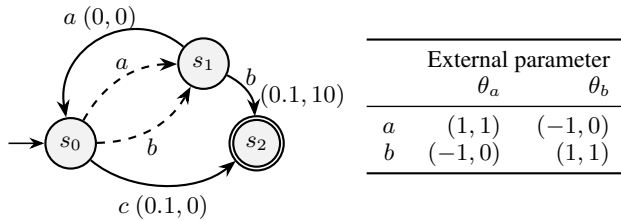|	External parameter	
	θ_a	θ_b
a	$(1,1)$	$(-1,0)$
b	$(-1,0)$	$(1,1)$

Figure 1: A simple example: a principal wishes to reach s_2 while maximizing rewards. All transitions are deterministic, and every edge is labeled with the corresponding action and (in the brackets) rewards for the agent and the principal, respectively. The rewards for state-action pairs (s_0, a) and (s_0, b) (dashed edges) also depend on the 2-valued external parameter, as specified in the table; the value of the parameter is sampled uniformly at random at each step. Assume uniform discounting with discount $\frac{1}{2}$. Without signaling, the agent will always take action c in s_0, whereby the principal obtains payoff 0. The principal can reveal information about the external parameter to attract the agent to move to s_1. If the agent is myopic, the principal can reveal full information, which leads to the agent moving to s_1, taking action b, and ending in s_2; the principal obtains payoff 6 as a result. However, if the agent is far-sighted, this will not work: the agent will end up in a loop in s_0 and s_1, resulting in overall payoff $4/3$ for the principal. To improve, the principal can use a less informative strategy in s_0: e.g., advising the agent to take the more profitable action 10% of the time and a uniformly sampled action in $\{a, b\}$ the remaining 90% of the time. The agent will be incentivized to move to s_1 then. Alternatively, the principal can also use a threat-based strategy, which yields an even higher payoff in this instance: always reveal the true information in s_0, advise the agent to take b in s_1, and stop providing any information if the agent does not follow the advice. The outcome of this strategy coincides with how an advice-myopic agent behaves: they will choose b at s_1 as future disclosures are not considered.

be a natural heuristic in the real world when agents resort to prior knowledge, but not future information disclosure, to estimate future rewards. We then show that optimal signaling strategies can again be computed in polynomial-time. More interestingly, the solution can be used to design a threat-based signaling strategy against a far-sighted agent. We show that this threat-based strategy induces the same reaction from a far-sighted agent as from an advice-myopic one. Hence, it guarantees the principal the same payoff obtained against an advice-myopic agent, when the agent is actually far-sighted. Figure 1 shows the subtleties of optimal signaling strategies in the dynamic setting.

Related Work   Our starting point is the work on Bayesian persuasion (Kamenica and Gentzkow 2011), which looks at optimal signaling under incomplete information in the *static* case. Many variants of this model have been proposed and studied ever since, with applications in security, voting, advertising, finance, etc. (e.g., Rabinovich et al. 2015; Xu et al. 2015; Goldstein and Leitner 2018; Badani-

diyuru, Bhawalkar, and Xu 2018; Castiglioni, Celli, and Gatti 2020); also see the comprehensive surveys (Kamenica 2019; Dughmi 2017). Dynamic models of Bayesian persuasion were studied recently (Ely 2017; Renault, Solan, and Vieille 2017), and some more recent works focused on algorithmic problems from several dynamic models, such as a model built on extensive-form games (EFGs) (Celli, Coniglio, and Gatti 2020) and an online persuasion model (Castiglioni et al. 2020, 2021). These models are sufficiently different from ours. In the EFG model, in particular, an EFG parameterized by the state of nature (akin to our external parameter) is instantiated before the play, and a group of receivers then engage in the EFG and they infer the EFG being played according to signals from a sender. Hence, information exchange happens only once in this model, whereas it happens in every step in ours. Such one-off persuasions also appeared in several other works on Bayesian persuasion and, more broadly, on non-cooperative IRL (inverse reinforcement learning) and incentive exploration (Zhang et al. 2019; Mansour et al. 2021; Simchowitz and Slivkins 2021).

Reversing the roles of the players in terms of who has the power to commit leads to a dual problem of Bayesian persuasion, which is often known as automated mechanism design (Conitzer and Sandholm 2002, 2004). In such problems, the signal receiver commits to a mechanism that specifies the action they will take upon receiving each signal, and the signal sender sends signals optimally in response. A very recent work by Zhang and Conitzer (2021) considered automated mechanism design in a dynamic setting similar to ours, and offered a complementary view to our work. In their work, the primary consideration is a finite-horizon setting and history-based strategies. In contrast, we focus primarily on unbounded horizons and memory-less strategies.

The interaction between the principal and the agent can be viewed as a stochastic game (Shapley 1953) where one player (i.e., the principal) has the power to make a strategy commitment (Letchford and Conitzer 2010; Letchford et al. 2012). Games where multiple agents jointly take actions in a dynamic environment have been widely studied in the literature on multi-agent reinforcement learning, but usually in settings without strategy commitment (Littman 1994; Buşoniu, Babuška, and De Schutter 2010).

More broadly, our work also relates to the advice-based interaction framework (e.g., Torrey and Taylor 2013; Amir et al. 2016), where the principal's goal is to communicate advice to an agent on how to act in the world. This advice-based framework is also in close relationship to the machine teaching literature (Goldman and Kearns 1995; Singla et al. 2014; Doliwa et al. 2014; Zhu et al. 2018; Ng and Russell 2000; Hadfield-Menell et al. 2016) where the principal (i.e., the teacher) seeks to find an optimal training sequence to steer the agent (i.e., the learner) towards the desired goal. Similarly, in environment design, the principal modifies the rewards or transitions to steer the behavior of the agent. The objective may be obtaining fast convergence (Ng, Harada, and Russell 1999; Mataric 1994), or inducing a target policy of the agent (Zhang and Parkes 2008; Zhang, Parkes, and Chen 2009; Ma et al. 2019; Rakhsha et al. 2020b; Huang and Zhu 2019; Rakhsha et al. 2020a). These problem set-

tings are similar to ours in that the principal cannot directly act in the environment but can influence the agent's actions via learning signals. We see our setting and techniques as complementary to these studies; in particular, our hardness results can be extended there as well.

2 The Model

Our formal model is an MDP with reward uncertainties, given by a tuple $\mathcal{M} = \left\langle S, A, P, \Theta, (\mu_s)_{s \in S}, R, \widetilde{R} \right\rangle$ and involving two players: a *principal* and an *agent*. Similar to a standard MDP, S is a finite state space of the environment; A is a finite action space *for the agent*; $P : S \times A \times S \to [0, 1]$ is the transition dynamics of the state. When the environment is in state s and the agent takes action a, the state transitions to s' with probability $P(s, a, s')$; both the principal and the agent are aware of the state throughout. Meanwhile, rewards are generated for both the principal and the agent, and are specified by the reward functions $R : S \times \Theta \times A \to \mathbb{R}$ and $\widetilde{R} : S \times \Theta \times A \to \mathbb{R}$, respectively. Hence, unlike in a standard MDP, here the rewards also depend on an external parameter $\theta \in \Theta$. This parameter captures an additional layer of uncertainty of the environment; it follows a distribution $\mu_s \in \Delta(\Theta)$ and is drawn anew every time the state changes. For all $s \in S$, μ_s is common prior knowledge shared between the principal and the agent; however, only the principal has access to the realization of θ.

Crucially, since the actions are taken only by the agent, the principal cannot directly influence the state. Instead, the principal can use their information advantage about the external parameter to persuade the agent to take certain actions, by way of signaling.

Signaling and Belief Update   Let G be a space of *signals*. A signaling strategy of the principal generates a distribution over G. Our primary consideration in this paper is Markovian signaling strategies, whereby signals to send only depend on the current state (independent of the history). Formally, a signaling strategy $\pi = (\pi_s)_{s \in S}$ of the principal consists of a function $\pi_s : \Theta \to \Delta(G)$ for each state $s \in S$. Upon observing an external parameter θ, the principal will send a signal sampled from $\pi_s(\theta)$ when the current state is s; we denote by $\pi_s(\theta, g)$ the probability of $g \in G$ in this distribution.

The signal space is broadly construed. For example, one simple signaling strategy is to always reveal the true information, which always sends a deterministic signal g_θ associated with the observed external parameter $\theta \in \Theta$ (i.e., a message saying "The current external state is θ"); formally, we write $\pi_s(\theta) = \hat{\mathbf{e}}_{g_\theta}$.[1] In contrast, if the same signal is sent irrespective of the external parameter, i.e., $\pi_s(\theta) = \pi_s(\theta')$ for all $\theta, \theta' \in \Theta$, then the signaling strategy is completely uninformative. Without loss of generality, we assume that signals in G are distinct from each other from the agent's point of view.

Upon receiving a signal g, the agent updates their posterior belief about the (distribution of) the external parameter:

the conditional probability of the parameter being θ is

$$\Pr(\theta | g, \pi_s) = \frac{\mu_s(\theta) \cdot \pi_s(\theta, g)}{\sum_{\theta' \in \Theta} \mu_s(\theta') \cdot \pi_s(\theta', g)}. \quad (1)$$

To derive the above posterior also relies on knowledge about the principal's signaling strategy π. Indeed, we follow the Bayesian persuasion framework, whereby the principal commits to a signaling strategy π at the beginning of the game and announces it to the agent.

Signaling Strategy Optimization   We take the principal's point of view and investigate the problem of optimal signaling strategy design: given $\mathcal{M}$, find a signaling strategy π that maximizes the principal's (discounted) cumulative reward $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \theta_t, a_t) | \mathbf{z}, \pi\right]$, where $\mathbf{z} = (z_s)_{s \in S}$ is the distribution of the starting state, $\gamma \in [0, 1)$ is a discount factor, and the expectation is taken over the trajectory $(s_t, \theta_t, a_t)_{t=0}^{\infty}$ induced by the signaling strategy π. To completely specify this task requires a behavioral model for the agent. We will consider two major types of agents—*myopic* and *far-sighted*—and will define them separately in the next two sections; a myopic agent only cares about their instant reward in each step, whereas a far-sighted agent considers the cumulative reward with respect to a discount factor $\tilde{\gamma} > 0$ (which need not be equal to γ).

In summary, the game proceeds as follows. At the beginning, the principal commits to a signaling strategy π and announces it to the agent. Then in each step, if the environment is in state s, an external parameter $\theta \sim \mu_s$ is drawn (by nature); the principal observes θ, samples a signal $g \sim \pi_s(\theta)$, and sends g to the agent. The agent receives g, updates their belief about θ (according to (1)), and decides an action $a \in A$ to take accordingly. The state then transitions to $s' \sim P(s, a, \cdot)$.

3 When Agent is Myopic

We first consider a myopic agent. A myopic agent aims to maximize their reward in each individual step. Upon receiving a signal g in state s, the agent will take a best action $a \in A$, which maximizes $\mathbb{E}_{\theta \sim \Pr(\cdot | g, \pi_s)} \widetilde{R}(s, \theta, a)$. We study the problem of computing an optimal signaling strategy against a myopic agent, termed OPTSIG-MYOP.

Action Advice   According to a standard argument via the revelation principle, it is often without loss of generality to consider signaling strategies in the form of action advice. This also holds in our model. Specifically, for any signaling strategy, there exists an equivalent strategy π which uses only a finite set $G_A := \{g_a : a \in A\}$ of signals, and each signal g_a corresponds to an action $a \in A$; moreover, π is *incentive compatible* (IC), which means that the agent is also incentivized to take the corresponding action a upon receiving g_a, i.e., we have $\mathbb{E}_{\theta \sim \Pr(\cdot | g_a, \pi_s)} \widetilde{R}(s, \theta, a) \geq \mathbb{E}_{\theta \sim \Pr(\cdot | g_a, \pi_s)} \widetilde{R}(s, \theta, a')$ for all $a' \in A$,[2] or equivalently:

$$\sum_{\theta \in \Theta} \Pr(\theta | g_a, \pi_s) \cdot \left(\widetilde{R}(s, \theta, a) - \widetilde{R}(s, \theta, a') \right) \geq 0 \ \forall a'. \quad (2)$$

---

[1] We let $\hat{\mathbf{e}}_i$ denote a unit vector, of which the i-th element is 1.

[2] By convention, we assume that the agent breaks ties by taking the advised action when there are multiple optimal actions.

In other words, π signals which action the agent should take and it is designed in a way such that the agent cannot be better off deviating from the advised action with respect to the posterior belief. 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 (2). We refer the reader to the full version of this paper for more details about the generality of IC action advices in our model.[3]

We can easily characterize the outcome of an IC action advice π: at each state s, since the agent is incentivized to follow the advice, with probability $\phi_s^\pi(\theta, a) := \mu_s(\theta) \cdot \pi_s(\theta, g_a)$ they will take action a while the realized external parameter is θ; hence, ϕ_s^π is a distribution over $\Theta \times A$. We can then define the following set $\mathcal{A}_s \subseteq \Delta(\Theta \times A)$, which contains all such distributions that can be induced by some π:

$$\mathcal{A}_s = \{\phi_s^\pi : \pi \text{ is an IC action advice}\}.$$

It would now be convenient to view the problem facing the principal as an (single-agent) MDP $\mathcal{M}^* = \langle S, (\mathcal{A}_s)_{s \in S}, P^*, R^* \rangle$, 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 $P^* : S \times \Delta(\Theta \times A) \times S \to [0, 1]$ and reward function $R^* : S \times \Delta(\Theta \times A) \to \mathbb{R}$ are such that

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

Namely, $\mathcal{M}^*$ is defined as if the principal can choose actions (which are (θ, 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 $\mathcal{M}^*$. We show next that we can exploit a standard approach to compute an optimal policy but we need to address a key challenge as the action space of $\mathcal{M}^*$ may contain infinitely many actions.

LP Formulation 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:

$$V(s) = \max_{\mathbf{x} \in \mathcal{A}_s} \left[R^*(s, \mathbf{x}) + \gamma \cdot \sum_{s' \in S} P^*(s, \mathbf{x}, s') \cdot V(s') \right] \forall s.$$
$$(3)$$

It is well-known that there exists a unique solution to the above system of equations, from which an optimal policy can be extracted. In particular, one approach to computing this unique solution is by using the following LP (linear program) formulation, where $V(s)$ are the variables; The optimal value of this LP directly gives the cumulative reward of optimal policies under a given initial state distribution $\mathbf{z}$.

$$\min_V \quad \sum_{s \in S} z_s \cdot V(s) \tag{4}$$
$$\text{s.t.} \quad V(s) \geq R^*(s, \mathbf{x}) + \gamma \cdot \sum_{s' \in S} P^*(s, \mathbf{x}, s') \cdot V(s')$$
$$\forall s \in S, \mathbf{x} \in \mathcal{A}_s \tag{4a}$$

The issue with this LP formulation is that there may be infinitely many constraints as (4a) must hold for all $\mathbf{x} \in \mathcal{A}_s$. This differs from MDPs with a finite action space, in which case the LP formulation can be reduced to one with a finite set of constraints, where each constraint corresponds to an action. We address this issue by using the *ellipsoid method* as sketched below. More practically, we can also derive a concise LP formulation by exploiting the duality principle (see the full version of the paper).

Theorem 1. OPTSIG-MYOP *is solvable in polynomial time.*

Proof sketch. We show that LP (4) can 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 our problem values of $V(s)$), the oracle should decide correctly whether all the constraints of the LP are satisfied and, if not, output a violated one.

To implement the separation oracle for our problem amounts to solving the following optimization for all $s \in S$:

$$\max_{\mathbf{x} \in \mathcal{A}_s} \quad R^*(s, \mathbf{x}) + \gamma \cdot \sum_{s' \in S} P^*(s, \mathbf{x}, s') \cdot V(s') - V(s).$$

By checking if the above maximum value is positive, we can identify if (4a) is violated for some $\mathbf{x} \in \mathcal{A}_s$. Indeed, the set of IC action advices can be characterized by the constraints in (2), which are linear constraints if we expand $\Pr(\theta|g_a, \pi_s)$ according to (1) and eliminate the denominator (where we also treat $\pi_s(\theta, g_a)$ as the additional variables and add the constraint $x(\theta, a) = \mu_s(\theta) \cdot \pi_s(\theta, g_a)$ for every θ and a). $\quad \square$

4 When Agent is Far-sighted

A far-sighted (FS) agent looks beyond the immediate reward and considers the cumulative reward discounted by $\tilde{\gamma}$. We now study signaling strategy design against an FS agent.

4.1 Optimal Signaling against FS Agent

When facing an FS agent, we cannot define an inducible set $\mathcal{A}_s$ independently for each state. The principal needs to take a global view and aim to induce the agent to use a *policy* that benefits the principal. We term the problem of optimal signaling strategy design against an FS agent OPTSIG-FS.

Best Response of FS Agent We first investigate an FS agent's best response problem. When the principal commits to a signaling strategy π, the best response problem facing the agent can be formulated as an MDP $\mathcal{M}^\pi = \left\langle S \times G, A, P^\pi, \widetilde{R}^\pi \right\rangle$. In each step the agent observes the state $s \in S$ of $\mathcal{M}$ along with a signal $g \in G$ from the principal; the tuple (s, g) constitutes a state in $\mathcal{M}^\pi$, and we call it a *meta-state* to distinguish it from states in $\mathcal{M}$. From the agent's perspective, after they take action a, the meta-state transitions to (s', g') with probability

$$P^\pi((s, g), a, (s', g')) = P(s, a, s') \cdot \mathbb{E}_{\theta \sim \mu_{s'}} \pi_{s'}(\theta, g'). \tag{5}$$

Namely, a next state s' of $\mathcal{M}$ is sampled from $P(s, a, \cdot)$, then a new external parameter θ is sampled from $\mu_{s'}$ and the principal sends a signal $g \sim \pi_{s'}(\theta)$. Meanwhile, the following reward is yielded for the agent:

$$\widetilde{R}^\pi((s, g), a) = \mathbb{E}_{\theta \sim \Pr(\cdot|g, \pi_s)} \widetilde{R}(s, \theta, a), \tag{6}$$

where the posterior belief $\Pr(\theta|g, \pi_s)$ is defined in (1).

Hence, an optimal policy $\sigma : S \times G \to A$ for $\mathcal{M}^\pi$ defines a best response of the agent against π. An optimal signaling strategy of the principal maximizes the cumulative reward against the agent's best response.

Inapproximability We show that OPTSIG-FS is highly intractable: even to find an approximate solution to OPTSIG-FS requires solving an NP-hard problem. Hence, it is unlikely that there exists any efficient approximation algorithm for this task, assuming that P=NP is unlikely.

Theorem 2. *Assuming that* $P \neq NP$, *then* OPTSIG-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 is via a reduction from the MAXIMUM INDEPENDENT SET problem, which is known to be NP-hard to approximate (Zuckerman 2006). The result may also be of independent interest: It can be easily adapted to show the inapproximability of similar principal-agent problems in dynamic settings. This hardness result also indicates a "phase transition" between the cases where $\tilde\gamma = 0$ and $\tilde\gamma > 0$ given the tractability of OPTSIG-MYOP showed in Section 3.

4.2 Advice-myopic Agent

The intractability of OPTSIG-FS motivates us to consider *advice-myopic (AM)* agents, who account for their future rewards like an FS agent does, but behave myopically and ignore the principal's future signals. In other words, they always assume that the principal will disappear in the next step and rely only on their prior knowledge to estimate the future payoff. We refer to the optimal signaling strategy problem against an AM agent as OPTSIG-AM.

Equivalence to the Myopic Setting Since an AM agent does not consider future signals, their future reward is independent of the principal's signaling strategy. This allows us to define a set of inducible distributions of (θ, a) independently for each state, similarly to our approach to dealing with a myopic agent. In other words, an AM agent is equivalent to a myopic agent who adds a fixed value to their reward function, and this fixed value is the best future reward they can achieve without the help of any signals. This value is independent of the signaling strategy and can be calculated beforehand. Let $\widetilde{R}^+ : S \times \Theta \times A \to \mathbb{R}$ be the reward function of this equivalent myopic agent. We have

$$\widetilde{R}^+(s, \theta, a) = \widetilde{R}(s, \theta, a) + \tilde\gamma \cdot \mathbb{E}_{s' \sim P(s,a,\cdot)} \overline{V}(s', g_0), \quad (7)$$

where $\overline{V}$ is the optimal value function of the agent when completely uninformative signals are given. In more detail, let $\bot : \Theta \to \Delta(G)$ be a completely uninformative signaling strategy, with $\bot(\theta) = \hat{\mathbf{e}}_{g_0}$ for all θ (i.e., it always sends the same signal g_0). Then $\overline{V}$ is the optimal value function for the MDP $\mathcal{M}^\bot = \left\langle S \times \{g_0\}, A, P^\bot, \widetilde{R}^\bot \right\rangle$, defined the

same way as $\mathcal{M}^\pi$ in Section 4.1, with $\pi = \bot$. Hence, the Bellman equation gives

$$\overline{V}(s, g_0)$$
$$= \max_{a \in A} \left(\widetilde{R}^\bot(s, \theta, a) + \tilde\gamma \cdot \mathbb{E}_{(s', g_0) \sim P^\bot((s, g_0), a, \cdot)} \overline{V}(s', g_0) \right)$$
$$= \max_{a \in A} \left(\mathbb{E}_{\theta \sim \mu_s} \widetilde{R}(s, \theta, a) + \tilde\gamma \cdot \mathbb{E}_{s' \sim P(s, a, \cdot)} \overline{V}(s', g_0) \right) \quad (8)$$

for all $s \in S$, where the second transition follows by (5) and (6) and we also use the facts that the posterior $\Pr(\cdot|g, \bot)$ degenerates to the prior $\mu_s(\cdot)$ as $\bot$ is uninformative, and that $P^\bot((s, g_0), a, (s', g_0)) = P(s, a, s')$ as the meta-state only transitions among the ones in the form (s, g_0).

We can compute $\overline{V}$ efficiently by solving the above Bellman equation. (A standard LP approach suffices given that $\mathcal{M}^\bot$ has a finite action space.) Then we obtain $\widetilde{R}^+$ according to (7), with which we can construct an equivalent OPTSIG-MYOP instance and solve it using our algorithm in Section 3. The solution is also optimal to the original OPTSIG-AM instance as we argued above; we state this result in the theorem below and omit the proof.

Theorem 3. OPTSIG-AM *is solvable in polynomial time.*

4.3 Threat-based Action Advice against FS Agent

Now that we can efficiently solve OPTSIG-AM, we will show that we can use a solution to OPTSIG-AM to efficiently design a signaling strategy against an FS agent. Interestingly, we can prove that this strategy guarantees the principal the payoff as if they are playing against an AM agent, when the agent is actually FS. The idea is to add a threat in the action advice: if the agent does not take the advised action, then the principal will stop providing any information in future steps (equivalently, switching to strategy $\bot$). Essentially, this amounts to a one-memory strategy, denoted $\varpi = (\varpi_s)_{s \in S}$, where each $\varpi_s : S \times \Theta \times G \times A \to \Delta(A)$ also depends on the signal and the action taken in the previous step (i.e., whether the action follows the signal).

More formally, suppose that $\pi = (\pi_s)_{s \in S}$ is a solution to OPTSIG-AM and without loss of generality it is an IC action advice. We construct a one-memory strategy:

$$\varpi_s((s, \theta), g, a) = \begin{cases} \pi_s(\theta), & \text{if } g \in \{g_a, null\} \\ \bot(\theta) = \hat{\mathbf{e}}_{g_0}, & \text{otherwise} \end{cases} \quad (9)$$

where g and a are the signal and action taken in the previous step (assume that g is initialized to *null* in the first step); each signal g_a advises the agent to take the corresponding action a, and g_0 is a signal that does not correspond to any action.

Our key finding is that, via this simple threat-based mechanism, the strategy ϖ we design is persuasive for an FS agent: the threat it makes effectively incentivizes the FS agent to take advised actions. To show this, we first analyze the problem facing the agent when the principal commits to ϖ.

Best Response to ϖ From an FS agent's perspective, the principal committing to ϖ results in an MDP $\mathcal{M}^\varpi = \left\langle S \times G, A, P^\varpi, \widetilde{R}^\varpi \right\rangle$. We have $G = \{g_0\} \cup G_A$, so each

meta-state (s, g) in $\mathcal{M}^\varpi$ consists of a state of $\mathcal{M}$ and a signal from the principal. The transition dynamics depend on whether the signal sent in the current state is g_0 or not (i.e., whether the principal has switched to the threat-mode):

- For all $(s, g_a) \in S \times G_A$, the agent following the advised action a results in transition probabilities:

$$P^\varpi((s, g_a), a, \cdot) = P^\pi((s, g_a), a, \cdot); \quad (10a)$$

Otherwise, i.e., if any action $b \neq a$ is taken, the principal will fulfill the threat and send g_0 in the next step. Hence,

$$P^\varpi((s, g_a), b, (s', g))$$

$$= \begin{cases} \sum_{g' \in G} P^\pi((s, g_a), a, (s', g')), & \text{if } g = g_0 \\ 0, & \text{otherwise} \end{cases} \quad (10b)$$

- For all $(s, g_0) \in S \times \{g_0\}$, the threat is activated in these meta-states, we have:

$$P^\varpi((s, g_0), a, (s', g')) = P^\perp((s, g_0), a, (s', g'))$$

$$= \begin{cases} P(s, a, s'), & \text{if } g' = g_0 \\ 0, & \text{otherwise} \end{cases} \quad (10c)$$

Similarly, the reward function differs in meta-states (s, g_a) and (s, g_0). We have

$$\widetilde{R}^\varpi((s, g), \cdot) = \begin{cases} \widetilde{R}^\pi((s, g), \cdot), & \text{if } g \in G_A \\ \widetilde{R}^\perp((s, g_0), \cdot), & \text{otherwise} \end{cases} \quad (11)$$

Persuasiveness of ϖ To show the persuasiveness of ϖ, we argue that the following policy $\sigma : S \times G \to A$ of the agent, in which the agent always takes the advised action, is optimal in response to ϖ. For all $s \in S$, we define

$$\sigma(s, g) = \begin{cases} a, & \text{if } g = g_a \in G_A \\ \bar{\sigma}(s, g_0), & \text{if } g = g_0 \end{cases} \quad (12)$$

where $\bar{\sigma}$ is an optimal policy against $\perp$, the value function $\overline{V} : S \times G \to \mathbb{R}$ of which (as from the agent's perspective) is already defined in Section 4.2 and satisfies (8).

Theorem 4 shows the optimality of σ. Intuitively, we show that the value function of σ is at least as large as $\overline{V}$, so the agent has no incentive to provoke the threat.

Theorem 4. *The policy σ defined in* (12) *is an optimal response of an FS agent to ϖ. (Hence, ϖ incentivizes an FS agent to take the advised action.)*

The direct consequence of Theorem 4 is that ϖ guarantees the principal the best payoff they can obtain when facing an AM agent, even though the agent is FS (Corollary 5). Hence, ϖ serves as an alternative approach to deal with an FS agent. Note that this threat-based strategy may not be an optimal one-memory strategy. Indeed, with minor changes to our proof of Theorem 2, we can show that for any positive integer k the problem of computing an optimal k-memory strategy is inapproximable (see the full version of the paper). In contrast, in the myopic and advice-myopic settings, since the agent's behavior is Markovian, the optimal signaling strategies we designed remain optimal even when we are allowed to use memory-based strategies.

Corollary 5. *By using ϖ against an FS agent, the principal's cumulative reward is the same as the highest cumulative reward they can obtain against an AM agent.*

5 Experiments

We empirically evaluate signaling strategies obtained with our algorithms. The goal is to compare the payoffs yielded for the principal. We use Python (v3.9) to implement our algorithms and Gurobi (v9.1.2) to solve all the LPs. All results were obtained on a platform with a 2 GHz Quad-Core CPU and 16 GB memory, and each is averaged over at least 20 instances. We conduct experiments on (i) general instances without any specific underlying structure, and (ii) instances generated based on a road navigation application.

General Instances The first set of instances are generated as follows. The transition probabilities and the initial state distribution are generated uniformly at random (and normalized to ensure that they sum up to 1). We also set an integer parameter n^*, and change n^* states to terminal states. The reward values are first generated uniformly at random from the range $[0, 1]$. Then, we tune the agent's rewards according to a parameter $\beta \in [-1, 1]$, resetting $\widetilde{R}(s, \theta, a) \leftarrow (1 - |\beta|) \cdot \widetilde{R}(s, \theta, a) + \beta \cdot R(s, \theta, a)$. Hence, when $\beta = 0$, the agent's rewards are independent of the principal's; when $\beta = 1$, they are completely aligned; and when $\beta = -1$, they are zero-sum.

We evaluate optimal signaling strategies against a myopic and an AM agent; the latter is equivalent to our threat-based strategy against an FS agent (THREAT-FS). We use two benchmarks, which are by nature also the lower and upper bounds of payoffs of other strategies: i) when the principal cannot send any signal and the agent operates with only the prior knowledge (NOSIG-MYOP and NOSIG-AM/FS; AM and FS agents have the same behavior in this case); and ii) when the principal has full control over the agent (FULLCONTROL). For ease of comparison, all results are shown as their ratios to results of FULLCONTROL.

Figure 2 summarizes the results. It is clearly seen that OPTSIG improves significantly upon NOSIG in all figures. The gap appears to increase with β, and when $\beta \geq 0$ (when the agent's rewards are positively correlated to that of the principal), OPTSIG is very closed to FULLCONTROL. It is also noted that differences between results obtained in the myopic setting and in the FS/AM setting are very small (e.g., compare (a) and (b)). This is mainly due to the random nature of the instances: in expectation, future rewards of all actions are the same. Hence, in the remaining figures we only present results obtained in the myopic setting. As shown in these figures, payoff improvement offered by the optimal strategies increases slowly with the number of actions and the number of external parameters. Intuitively, as these two numbers increase, the agent's decision making in each state becomes more reliant on advice from the principal. Nevertheless, the results do not appear to vary insignificantly with other parameters, such as the number of states or the number of terminal states as shown in (e) and (f) (also see the full version of the paper for additional experiment results).

Road Navigation Instances In the navigation application, the agent wants to travel from a starting node to a destination node on a road network, and is free to choose any path. In each step, the agent picks a road at the current node and trav-
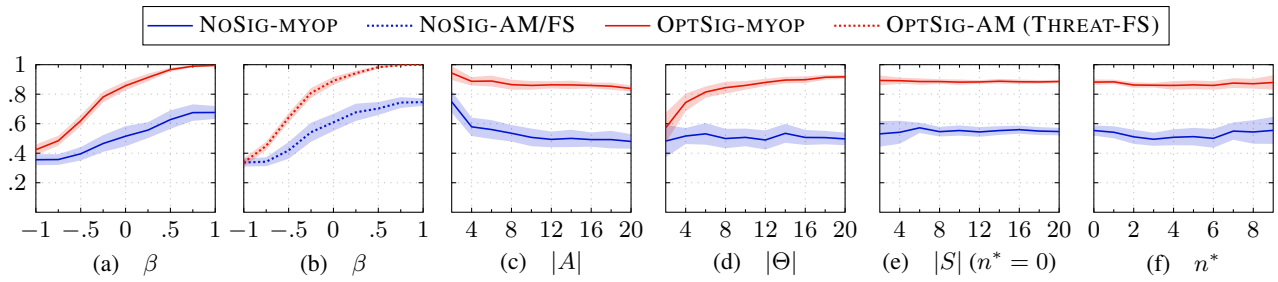
Figure 2: Comparison of signaling strategies: all results are shown as ratios to FULLCONTROL on the y-axes. Meanings of x-axes are noted in the captions. Shaded areas represent standard deviations (mean $\pm$ standard deviation). In all figures, we fix $|S| = |\Theta| = |A| = 10$, $\gamma = \tilde{\gamma} = 0.8$, $n^* = 5$, and $\beta = 0$ unless they are variables.
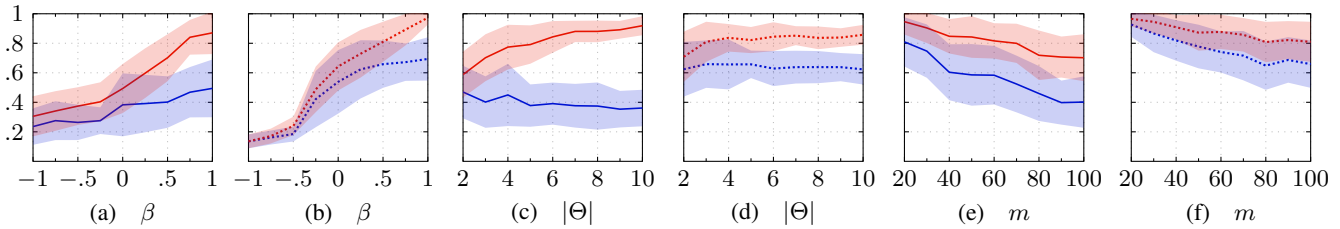

Figure 3: Comparison of signaling strategies in a navigation application: all results are shown as the ratios of FULLCONTROL to them on the y-axes (now that rewards are costs). All curves and axes have the same meanings as in Figure 2. All results are obtained on instances with $n = 20$ and $m = 100$ (i.e., numbers of nodes and edges in the network), where we also fix $|\Theta| = 3$, $\gamma = \tilde{\gamma} = 0.8$, and $\beta = 0.5$ unless they are variables.

els through it. The reward the agent receives at each step is a *cost* representing the travel time through the chosen road, which depends on the congestion level represented by the external parameter. The principal, as a social planner, has a preference over the path the agent picks (e.g., in consideration of the overall congestion or noise levels across the city), and this is encoded in a reward function for the principal: whenever the agent picks a road, the principal also receives a cost according to this reward function. Naturally, the agent's position (the node the agent is on) defines the state of the MDP. For simplicity, we assume that the road network is a directed acyclic graph (DAG), so the agent always reaches the destination in a finite number of steps.

To generate an instance, we first generate a random DAG with specified numbers of nodes and edges (roads). Let these numbers be n and m, respectively ($n \leq m \leq \frac{n(n-1)}{2}$). We sample a Prüfer sequence of length $n - 2$ uniformly at random and then convert it into the corresponding tree. We index the nodes according to their order in a breadth-first search. The node with the smallest/largest index is chosen as the start/destination. Then we add an edge between a pair of nodes chosen uniformly at random, from the node with the smaller index to the node with the larger index, until there are m edges on the graph. In the case that some node has no outgoing edge and it is not the destination, we also add an edge linking this node to the destination, so the graph generated may actually have more than m edges. In this way the graph generated is always a DAG.

The results are presented in Figure 3. The results exhibit similar patterns to their counterparts in Figure 2. Nevertheless, the gaps between different strategies appear to be narrower in the FS/AM setting than those in the myopic setting, which is not obvious in Figure 2.

6 Conclusion

We described and studied a dynamic model of persuasion in infinite horizon Markov processes. Our main results characterize the nature and computational complexity of optimal signaling against different types of agents. A limitation of the current model is that it requires common knowledge of transitions and rewards; hence, studying online versions of our problem (Castiglioni et al. 2020) is an immediate future step. While we focus on the algorithmic aspects of persuasion, our results indicate how a social planner might influence agents optimally. In particular implementations, the planner's incentives may not be aligned with societal benefits. In these cases, a careful analysis of the persuasion mechanisms and their moral legitimacy must be considered.

Acknowledgments

References

Amir, O.; Kamar, E.; Kolobov, A.; and Grosz, B. J. 2016. Interactive Teaching Strategies for Agent Training. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*, 804–811.

Badanidiyuru, A.; Bhawalkar, K.; and Xu, H. 2018. Targeting and signaling in ad auctions. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2545–2563. SIAM.

Buşoniu, L.; Babuška, R.; and De Schutter, B. 2010. Multi-Agent Reinforcement Learning: An Overview. *Innovations in Multi-Agent Systems and Applications-1*, 183–221.

Castiglioni, M.; Celli, A.; and Gatti, N. 2020. Persuading Voters: It's Easy to Whisper, It's Hard to Speak Loud. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'20)*, volume 34, 1870–1877.

Castiglioni, M.; Celli, A.; Marchesi, A.; and Gatti, N. 2020. Online Bayesian Persuasion. In *Advances in Neural Information Processing Systems (NeurIPS'20)*, volume 33, 16188–16198.

Castiglioni, M.; Marchesi, A.; Celli, A.; and Gatti, N. 2021. Multi-receiver online bayesian persuasion. In *Proceedings of 38th International Conference on Machine Learning (ICML'21)*, 1314–1323.

Celli, A.; Coniglio, S.; and Gatti, N. 2020. Private Bayesian persuasion with sequential games. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'20)*, volume 34, 1886–1893.

Conitzer, V.; and Sandholm, T. 2002. Complexity of Mechanism Design. In Darwiche, A.; and Friedman, N., eds., *Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence (UAI'02)*, 103–110. Morgan Kaufmann.

Conitzer, V.; and Sandholm, T. 2004. Self-interested automated mechanism design and implications for optimal combinatorial auctions. In *Proceedings of the 5th ACM Conference on Electronic Commerce (EC'04)*, 132–141.

Doliwa, T.; Fan, G.; Simon, H. U.; and Zilles, S. 2014. Recursive Teaching Dimension, VC-dimension and Sample Compression. *Journal of Machine Learning Research*, 15(1): 3107–3131.

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

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

Goldman, S. A.; and Kearns, M. J. 1995. On the Complexity of Teaching. *Journal of Computer and System Sciences*, 50(1): 20–31.

Goldstein, I.; and Leitner, Y. 2018. Stress tests and information disclosure. *Journal of Economic Theory*, 177: 34–69.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS'16)*.

Huang, Y.; and Zhu, Q. 2019. Deceptive Reinforcement Learning Under Adversarial Manipulations on Cost Signals.

In Alpcan, T.; Vorobeychik, Y.; Baras, J. S.; and Dán, G., eds., *Decision and Game Theory for Security (GameSec'19)*, 217–237.

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

Kamenica, E.; and Gentzkow, M. 2011. Bayesian persuasion. *American Economic Review*, 101(6): 2590–2615.

Letchford, J.; and Conitzer, V. 2010. Computing optimal strategies to commit to in extensive-form games. In *Proceedings of the 11th ACM conference on Electronic commerce (EC'10)*, 83–92.

Letchford, J.; MacDermed, L.; Conitzer, V.; Parr, R.; and Isbell, C. L. 2012. Computing Optimal Strategies to Commit to in Stochastic Games. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*, 1380–1386.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, 157–163. Elsevier.

Ma, Y.; Zhang, X.; Sun, W.; and Zhu, X. 2019. Policy Poisoning in Batch Reinforcement Learning and Control. In *Advances in Neural Information Processing Systems (NeurIPS'19)*, 14543–14553.

Mansour, Y.; Slivkins, A.; Syrgkanis, V.; and Wu, Z. S. 2021. Bayesian Exploration: Incentivizing Exploration in Bayesian Games. *Operations Research*.

Mataric, M. J. 1994. Reward Functions for Accelerated Learning. In *Proceedings of the 11th International Conference on International Conference on Machine Learning (ICML'94)*, 181–189.

Ng, A. Y.; Harada, D.; and Russell, S. J. 1999. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In *Proceedings of the 16th International Conference on Machine Learning (ICML'99)*, 278–287.

Ng, A. Y.; and Russell, S. J. 2000. Algorithms for Inverse Reinforcement Learning. In *ICML*.

Rabinovich, Z.; Jiang, A. X.; Jain, M.; and Xu, H. 2015. Information disclosure as a means to security. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS'15)*, 645–653.

Rakhsha, A.; Radanovic, G.; Devidze, R.; Zhu, X.; and Singla, A. 2020a. Policy Teaching in Reinforcement Learning via Environment Poisoning Attacks. *CoRR*, abs/2011.10824.

Rakhsha, A.; Radanovic, G.; Devidze, R.; Zhu, X.; and Singla, A. 2020b. Policy Teaching via Environment Poisoning: Training-time Adversarial Attacks against Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, volume 119, 7974–7984.

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

Shapley, L. S. 1953. Stochastic Games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100.

Simchowitz, M.; and Slivkins, A. 2021. Exploration and Incentives in Reinforcement Learning. arXiv:2103.00360.

Singla, A.; Bogunovic, I.; Bartók, G.; Karbasi, A.; and Krause, A. 2014. Near-Optimally Teaching the Crowd to Classify. In *Proceedings of the 31st International Conference on Machine Learning (ICML'14)*, volume 32, 154–162.

Torrey, L.; and Taylor, M. 2013. Teaching on a Budget: Agents Advising Agents in Reinforcement Learning. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems (AAMAS'13)*, 1053–1060.

Xu, H.; Rabinovich, Z.; Dughmi, S.; and Tambe, M. 2015. Exploring information asymmetry in two-stage security games. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15)*, volume 29.

Zhang, H.; and Conitzer, V. 2021. Automated Dynamic Mechanism Design. *Advances in Neural Information Processing Systems (NeurIPS'21)*, 34.

Zhang, H.; and Parkes, D. C. 2008. Value-Based Policy Teaching with Active Indirect Elicitation. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI'08)*, 208–214.

Zhang, H.; Parkes, D. C.; and Chen, Y. 2009. Policy Teaching through Reward Function Learning. In *Proceedings of the 10th ACM conference on Electronic Commerce (EC'09)*, 295–304.

Zhang, X.; Zhang, K.; Miehling, E.; and Basar, T. 2019. Non-Cooperative Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS'19)*, volume 32.

Zhu, X.; Singla, A.; Zilles, S.; and Rafferty, A. N. 2018. An Overview of Machine Teaching. *CoRR*, abs/1801.05927.

Zuckerman, D. 2006. 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)*, 681–690. Association for Computing Machinery.