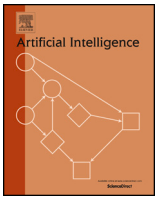
Artificial Intelligence 338 (2025) 104245

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint



Online learning in sequential Bayesian persuasion: Handling unknown priors ☆

Martino Bernasconi *, Matteo Castiglioni, Alberto Marchesi, Nicola Gatti, Francesco Trovò

Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, Italy

ARTICLE INFO

Keywords:
Online learning
Bayesian persuasion
Sequential decision making

ABSTRACT

We study a repeated *information design* problem faced by an informed *sender* who tries to influence the behavior of a self-interested *receiver*, through the provision of payoff-relevant information. We consider settings where the receiver repeatedly faces a *sequential decision making* (SDM) problem. At each round, the sender observes the realizations of random events in the SDM problem, which are only partially observable by the receiver. This begets the challenge of how to incrementally disclose such information to the receiver to *persuade* them to follow (desirable) action recommendations. We study the case in which the sender does *not* know random events probabilities, and, thus, they have to gradually learn them while persuading the receiver. We start by providing a non-trivial polytopal approximation of the set of the sender's persuasive information-revelation structures. This is crucial to design efficient learning algorithms. Next, we prove a negative result which also applies to the non-sequential case: *no learning algorithm can be persuasive in high probability*. Thus, we relax the persuasiveness requirement, studying algorithms that guarantee that the receiver's *regret* in following recommendations *grows sub-linearly*. In the *full-feedback* setting—where the sender observes the realizations of *all* the possible random events—, we provide an algorithm with $\tilde{O}(\sqrt{T})$ regret for both the sender and the receiver. Instead, in the *bandit-feedback* setting—where the sender only observes the realizations of random events actually occurring in the SDM problem—, we design an algorithm that, given an $\alpha \in [1/2, 1]$ as input, guarantees $\tilde{O}(T^{\alpha})$ and $\tilde{O}(T^{\max\{\alpha, 1-\frac{\alpha}{2}\}})$ regrets, for the sender and the receiver respectively. This result is complemented by a lower bound showing that such a regret trade-off is tight for $\alpha \in [1/2, 2/3]$.

1. Introduction

Bayesian persuasion [2] (a.k.a. *information design*) is the problem faced by an informed *sender* who aims to influence the behavior of a self-interested *receiver* via the strategic provision of payoff-relevant information. This captures the problem of "who gets to know what", which is fundamental in all economic interactions. Bayesian persuasion is ubiquitous in real-world problems, such as, *e.g.*, on-

☆ A short version of this article appeared in Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems, 2022 [1].

* Corresponding author.
E-mail addresses: martino.bernasconideluca@polimi.it (M. Bernasconi), matteo.castiglioni@polimi.it (M. Castiglioni), alberto.marchesi@polimi.it (A. Marchesi), nicola.gatti@polimi.it (N. Gatti), francesco1.trovo@polimi.it (F. Trovò).

https://doi.org/10.1016/j.artint.2024.104245
Received 17 July 2023; Received in revised form 1 March 2024; Accepted 3 November 2024

Available online 6 November 2024

line advertising [3–6], voting [7–10], traffic routing [11,12], recommendation systems [13], security [14,15], and marketing [16,17]. Some recent works also address the problem of dealing with uncertain parameters in Bayesian persuasion, boosting its applicability in concrete settings [18–23].

In this paper, we study Bayesian persuasion in settings where the receiver plays in a *sequential decision-making* (SDM) problem. An SDM problem is characterized by a tree structure composed of *decision* nodes, where the receiver makes actions, and *chance* nodes, in which *partially observable* random events occur. The sender perfectly observes the realizations of random events, and their goal is to incrementally disclose the acquired information to induce the receiver towards outcomes that are desirable for them. This is not an easy feat when the sender and receiver have different utilities. To do so, the sender commits to a *signaling scheme* specifying a probability distribution over action recommendations for the receiver at each decision node. Specifically, the sender commits to a *persuasive* signaling scheme, meaning that the receiver is incentivized to follow recommendations. We consider the case of a *farsighted* receiver, meaning that they take into account all the possible future events when deciding whether to deviate or *not* from recommendations at each decision node.

With some notable exceptions (such as, *e.g.*, [24,25]), Bayesian persuasion models available in the literature make the stringent assumption that both the sender and receiver know the *prior*, which, in our setting, is defined by the probabilities associated with random events in the SDM problem. We relax such an assumption by considering an online learning framework in which the sender, without any knowledge of the prior, repeatedly interacts with the receiver and gradually learns the prior while still issuing recommendations.

The model studied in this paper fits many real-world settings. For example, an e-commerce platform that issues customized information about products to its users, a recommendation system in a multi-media platforms that has to decide which content to suggest in order to keep users entertained (while still turning a profit), and navigation systems. In the following, we provide an in depth example on the latter application of a navigation app sending route recommendations to a driver. The app (sender) gets to know information about traffic congestion sequentially, as the driver progresses on their route. This is because traffic congestion might change during the trip, especially for long routes. Moreover, at any point in time, the app can send recommendations to the driver (receiver), who in turn has to take decisions on which routes to choose. The app and the driver might have different utilities. For example, the app may aim to minimize overall traffic congestion, while the driver may aim to minimize their travel time. Moreover, the driver interacts with the app multiple times, and, thus, the application must provide good recommendations (*i.e.*, persuasive) to the receiver, otherwise they would switch to another navigation app.

1.1. Original contributions

Our goal is to design online learning algorithms that are no-regret for the sender while being persuasive for the receiver. In particular, in Section 5, we provide a non-trivial polytopal approximation of the set of the sender's persuasive signaling schemes. This approximation leverages the sequence-form formulation of the SDM problem, which gives a convenient representation of the signaling scheme. Intuitively, in this representation, both the sender's utility and the receiver's increase in utility after a single deviation are linear.

This representation is crucial in designing efficient (*i.e.*, polynomial-time) learning algorithms, and it also shows how a sender-optimal signaling scheme can be found in polynomial time in the offline version of our problem, which may be of independent interest. Next, in Section 7, we prove a negative result which also applies to the non-sequential case: *when the prior is unknown no algorithm can be persuasive at each round with high probability*. This is not surprising, as it is easy to imagine a sender that is uncertain over two instances with different priors in which the sets of persuasive signaling schemes are disjoint. We prove our negative result by providing a pair of such instances. In the attempt to provide algorithms that be applied to any prior, we deal with this impossibility by relaxing the persuasiveness constraint. In particular, we study learning algorithms guaranteeing that the receiver's regret in following recommendations grows sub-linearly while guaranteeing the same for the sender's regret. In particular, we design learning algorithms for two different settings.

First, in Section 8, we study the *full-feedback* case, where the sender observes the realizations of *all* the random events that may potentially happen in the SDM problem. In such a setting, we provide an algorithm with $\tilde{O}(\sqrt{T})$ regret for both the sender and the receiver. The algorithm keeps track of a confidence set of the set of persuasive signaling schemes and chooses signaling schemes optimistically with respect to this set.

Then, in Section 9, we focus on the *bandit-feedback* setting, where the sender only observes the realizations of random events on the path in the tree traversed during the SDM problem. The challenge of this scenario stems from the fact that, when ignoring the prior, neither the sender nor the receiver knows the magnitude of the violation of the persuasiveness constraints. Thus, our algorithm performs a preliminary uniform exploration phase of the tree. The algorithm keeps track of a high-probability set of possible priors. After the exploration phase, the algorithm fixes the set of signaling schemes to use in the remaining steps in which it will act optimistically with respect to both the prior and the signaling scheme. In this case, we design an algorithm that achieves $\tilde{O}(T^{\alpha})$ sender's regret and $\tilde{O}(T^{\max\{\alpha, 1-\frac{\alpha}{2}\}})$ receiver's regret, for any $\alpha \in [1/2, 1]$ given as input. Finally, in Section 9.3, we provide a lower bound showing that the regrets trade-off achieved by our algorithm is tight for $\alpha \in [1/2, 2/3]$. This trade-off nature between utility and constraint violation is non-standard and might be of independent interest to other online learning problems. Moreover, our lower bound essentially proves the need for the exploration phase. Indeed our lower bound is built on instances in which there are portions of the tree that are sub-optimal—not being played in the optimal solution—, yet are needed to be explored to know the set of persuasive signaling schemes.

1.2. Related works

In the following, we provide a survey of works that are somehow related to ours, discussing their main differences.

Work by Zu et al. [24] The work that is most related to ours is the one by Zu et al. [24], who study a (non-sequential) Bayesian persuasion problem in which the sender and the receiver do *not* know the prior and they are involved in repeated interaction. In such a non-sequential model, Zu et al. [24] provide a learning algorithm that achieves $\tilde{O}(\sqrt{T})$ sender's regret with respect to the best-in-hindsight signaling scheme that one can produce when knowing the prior, while at the same time being persuasive across all rounds with high probability. This result requires some regularity assumptions over the prior that may not hold in the general case and that, in the sequential case, appear to be even more severe. In the attempt to design algorithms that can be applied with any prior, we adopt a fundamentally different approach, which involves relaxing the persuasiveness assumption and trading off the sender's and receiver's regret.

Bayesian persuasion in Markov decision processes Recently, some works addressed Bayesian persuasion in settings in which the sender and the receiver are involved in a sequential interaction, focusing on *Markov decision processes* (MDPs). In particular, Gan et al. [26] and Wu et al. [27] show how to efficiently find a sender-optimal signaling policy when the receiver is *myopic* (*i.e.*, they only optimize one-step rewards) in MDPs with infinite and finite horizon, respectively. On the other hand Bernasconi et al. [28] studied the problem of persuading farsighted receivers in MDPs. However, both Gan et al. [26] and Bernasconi et al. [28] assume that the environment is known, and, thus, they only focus on the (offline) sender's optimization problem. On the other hand, Wu et al. [27] also consider an online learning framework in which the sender has to learn the prior, as well as their rewards and the transition kernel of the MDP. They provide a learning algorithm that achieves $\tilde{O}(\sqrt{T})$ sender's regret, while at the same time guaranteeing persuasive recommendations to the (myopic) receiver. These works considerably differ from ours, since our model involves a farsighted receiver and the fact that the latter may have partial observability over realizations of random events. More in particular, we can model any information revelation structure in the sequential problem, like the case in which the receiver can distinguish between some nature's states but not others (see Fig. 1 for an example). Let us remark that Gan et al. [26] also study a variation of their model in which the receiver is farsighted, showing that, in such a setting, the problem of finding a sender-optimal signaling policy is NP-hard to approximate. As a result, Gan et al. [26] do *not* provide any algorithmic result for their model with a farsighted receiver. On the other hand, in our model, we consider farsighted receivers. However, this does not contradict our result, as we have a tree-like structure of the game which allows efficient algorithms. Moreover, even if any MDP could be converted into our sequential decision model framework, this transformation is not efficient and thus we cannot use our algorithm to solve persuasion problems with farsighted receivers in MDPs.

Other works on sequential Bayesian persuasion models It is also worth citing that other works study sequential versions of Bayesian persuasion, though they are *not* related to ours as much as [26] and [27]. Among these works, Celli et al. [29] study Bayesian persuasion with multiple receivers interacting in a sequential game with imperfect information. Differently from ours, their model adopts a different notion of persuasiveness, known as *ex ante* persuasiveness, and it assumes that the prior is known. Very recently, Su et al. [30] and Ni et al. [31] initiate the study of settings where the sender and the receiver repeatedly interact in a one-shot Bayesian persuasion problem, with the sender committing to a sequence of signaling schemes in which, at each round of the interaction, the sender's choice depends on the outcomes of all the previous rounds and is consistent with the receiver's posterior believes at that round. These works assume that the prior is known to both the sender and the receiver, and, thus, they considerably depart from ours.

Bayesian persuasion with unknown receivers' types Over the last years, some works (see, *e.g.*, [18,32,33]) addressed the problem of relaxing Bayesian persuasion assumptions by considering the case in which the sender does *not* know receivers' payoffs. In particular, Castiglioni et al. [18,32] consider an online learning framework in which the sender repeatedly interacts with a stream of receivers whose types are selected beforehand by an adversary. These latter works use techniques inherited from the online learning literature as we do in this paper; however, their adoption is fundamentally different from the one we undertake in this paper.

Online learning in SDM problems with constraints Another line of research, which uses techniques that are similar to those employed in this paper, studies learning in SDM problems in which the learner has to satisfy unknown constraints [34,35]. In particular, Bernasconi et al. [35] study the learning problem faced by a single agent in a repeated SDM problem with both rewards and costs. The authors study a setting in which the goal of the learner is to maximize long-term rewards while guaranteeing that some cost constraints are satisfied in the long term, thus allowing for the constraints to be violated during some rounds. The problem that we study in this paper is similar, but it raises considerable additional challenges due to the structure of the persuasiveness constraints.

2. Preliminaries

2.1. Sequential decision making problems

An instance of an SDM problem is defined by a tree structure, utilities, and random events probabilities. The tree structure has a set of nodes $\mathcal{H} := \mathcal{Z} \cup \mathcal{H}_d \cup \mathcal{H}_c$, where: $\mathcal{Z}$ contains all the *terminal nodes* in which the problem ends (corresponding to the leaves

of the tree), $\mathcal{H}_d$ is the set of *decision nodes* in which the agent acts, while $\mathcal{H}_c$ is the set of *chance nodes* where random events occur. Given any non-terminal node $h \in \mathcal{H} \setminus \mathcal{Z}$, we let $A(h)$ be the set of arcs outgoing from h. If $h \in \mathcal{H}_d$, then $A(h)$ is the set of receiver's actions available at h, while, if $h \in \mathcal{H}_c$, then $A(h)$ encodes the possible outcomes of the random event occurring at h. Furthermore, the utility function $u : \mathcal{Z} \to [0,1]$ defines the agent's payoff $u(z)$ when the problem ends in terminal node $z \in \mathcal{Z}$. Finally, each chance node $h \in \mathcal{H}_c$ is characterized by a probability distribution $\mu_h \in \Delta_{A(h)}$ over the possible outcomes of the corresponding random event, with $\mu_h(a)$ denoting the probability of action $a \in A(h)$.[1]

In an SDM problem, the agent has *imperfect information*, since they do *not* perfectly observe the outcomes of random events. Thus, the set of decision nodes $\mathcal{H}_d$ is partitioned into *information sets* (infosets for short), where an infoset $I \subseteq \mathcal{H}_d$ is a subset of decision nodes that are indistinguishable for the agent. We denote the set of infosets as $\mathcal{I}$. For every infoset $I \in \mathcal{I}$ and pair of nodes $h, h' \in I$, it must be the case that $A(h) = A(h') =: A(I)$, otherwise the agent could distinguish between the two nodes. We assume that the agent has *perfect recall*, which means that they never forget information once acquired. Formally, this is equivalent to assuming that, for every infoset $I \in \mathcal{I}$, all the paths from the root of the tree to a node $h \in I$ identify the same ordered sequence of agent's actions.

2.2. Bayesian persuasion in sequential decision making problems

We study *Bayesian persuasion in SDM* (BPSDM) problems. These extend the Bayesian persuasion framework [2] to SDM problems by introducing an exogenous agent that acts as a *sender* by issuing signals to the decision-making agent (the *receiver*).[2] By following the Bayesian persuasion terminology, the probability distributions μ_h for each chance node h are collectively referred to as the *prior*. Thus, the sender observes the realizations of random events occurring in the SDM problem and can partially disclose information to influence the receiver's behavior. Moreover, the sender's utility function is defined over terminal nodes, denoted as $f : \mathcal{Z} \to [0,1]$, and their goal is to commit to a publicly known *signaling scheme* that maximizes their utility in expectation with respect to the prior, the signaling scheme, and the receiver's strategy.

Formally, a signaling scheme for the sender defines a probability distribution $\phi_h \in \Delta_{S(h)}$ at each decision node $h \in \mathcal{H}_d$, where $S(h)$ is a finite set of signals available at h. During the SDM problem, when the receiver reaches a node $h \in \mathcal{H}_d$ belonging to an infoset $I \in \mathcal{I}$, the sender draws a signal $s \sim \phi_h$ and communicates it to the receiver. Then, based on the history of signals observed from the beginning of the SDM problem (s included), the receiver computes a *posterior* belief over the nodes belonging to the infoset I and plays so as to maximize their expected utility in the SDM sub-problem that starts from I, taking into account the just acquired information.

As customary in these settings, a simple revelation-principle-style argument allows us to focus on signaling schemes that are *direct* and *persuasive* [2,36]. In particular, a signaling scheme is direct if signals correspond to action recommendations, namely $S(h) = A(h)$ for all $h \in \mathcal{H}_d$. A direct signaling scheme is persuasive if the receiver is incentivized to follow action recommendations issued by the sender. Moreover, we assume that, if the receiver does *not* follow action recommendations at some decision node, then the sender stops issuing recommendations at nodes later reached during the SDM problem. This is without loss of generality. We refer to [37,38] for a discussion on a similar problem in the field of correlation in sequential games.

2.3. The sequence-form representation

The *sequence form* is a commonly-used, compact way of representing *(mixed) strategies* in SDM problems [39]. In this work, the sequence-form representation will be employed for the receiver's strategies, and to encode the signaling schemes and priors, as we describe in the following.

Receiver's strategies Given any $h \in \mathcal{H}$, we let $\sigma_r(h)$ be the ordered *sequence* of the receiver's actions on the path from the root of the tree to node h. By the perfect recall assumption, given any infoset $I \in \mathcal{I}$, it holds that $\sigma_r(h) = \sigma_r(h') =: \sigma_r(I)$ for every pair of nodes $h, h' \in I$. Thus, we can identify sequences with infoset-action pairs, with $\sigma = (I, a)$ encoding the sequence of actions obtained by appending action $a \in A(I)$ at the end of $\sigma_r(I)$, for any infoset $I \in \mathcal{I}$. Moreover, $\varnothing$ denotes the *empty sequence*. Hence, the receiver's sequences are $\Sigma_r := \{(I, a) \mid I \in \mathcal{I}, a \in A(I)\} \cup \{\varnothing\}$. In the sequence-form representation, mixed strategies are defined by specifying the probability of playing each sequence of actions. Thus, a receiver's strategy is represented by a vector $\boldsymbol{x} \in [0,1]^{|\Sigma_r|}$, where $\boldsymbol{x}[\sigma]$ encodes the realization probability of sequence $\sigma \in \Sigma_r$. Furthermore, a sequence-form strategy is well-defined if and only if it satisfies the following linear constraints:

$$\boldsymbol{x}[\varnothing] = 1 \quad \text{and} \quad \boldsymbol{x}[\sigma_r(I)] = \sum_{a \in A(I)} \boldsymbol{x}[\sigma_r(I)a] \quad \forall I \in \mathcal{I}.$$

We denote by $\mathcal{X}_r$ the polytope of all receiver's sequence-form strategies. We will also need to work with the sets of receiver's strategies in the SDM sub-problem that starts from an infoset $I \in \mathcal{I}$, formally defined as:

$$\mathcal{X}_{r,I} := \left\{ \boldsymbol{x} \in \mathcal{X}_r \mid \boldsymbol{x}[\sigma_r(I)] = 1 \right\}.$$

[1] For a finite set X we denote with Δ_X the set of probability distributions over X.

[2] In Section 3, we show that BPSDM reduces to classical Bayesian persuasion when there is no sequentiality.

Signaling schemes We represent signaling schemes in sequence form by leveraging the fact that the sender can be thought of as a perfect-information agent who plays at the decision nodes of the SDM problem since their actions correspond to recommendations for the receiver. Thus, since the sender's infosets correspond to decision nodes, their sequences $\Sigma_s := \{(h,a) \mid h \in \mathcal{H}_d, a \in A(h)\} \cup \{\varnothing\}$. Then, we denote the polytope of (sequence-form) signaling schemes as $\Phi \subseteq [0,1]^{|\Sigma_s|}$, where each signaling scheme is represented as a vector $\boldsymbol{\phi} \in [0,1]^{|\Sigma_s|}$ satisfying:

$$\boldsymbol{\phi}[\varnothing] = 1 \quad \text{and} \quad \boldsymbol{\phi}[\sigma_s(h)] = \sum_{a \in A(h)} \boldsymbol{\phi}[\sigma_s(h)a] \quad \forall h \in \mathcal{H}_d,$$

where, similarly to $\sigma_r(h)$ for the receiver, $\sigma_s(h)$ denotes the sender's sequence identified by $h \in \mathcal{H}$. We also define $\Pi := \Phi \cap \{0,1\}^{|\Sigma_s|}$ as the set of *deterministic* signaling schemes, which are those that recommend a single action with probability one at each decision node.

Priors We also encode prior probability distributions μ_h using the sequence form. Indeed, these can be thought of as elements of a fixed strategy played by a (fictitious) perfect-information agent that acts at chance nodes. Thus, for such a chance agent, we define Σ_c, $\mathcal{X}_c$, and $\sigma_c(h)$ as their counterparts previously introduced for the receiver. Moreover, in the following, we denote by $\boldsymbol{\mu}^\star \in \mathcal{X}_c$ the (sequence-form) prior, recursively defined as follows:

$$\boldsymbol{\mu}^\star[\varnothing] := 1 \quad \text{and} \quad \boldsymbol{\mu}^\star[\sigma_c(h)a] := \boldsymbol{\mu}^\star[\sigma_c(h)]\,\mu_h(a) \quad \forall h \in \mathcal{H}_c, \forall a \in A(h).$$

Note the prior distribution over nodes h of an infoset I is obtained by normalizing the sequence-form prior as follows:

$$\frac{\boldsymbol{\mu}^\star[\sigma_c(h)]}{\sum_{h' \in I} \boldsymbol{\mu}^\star[\sigma_c(h')]},$$

which defines a distribution over nodes $h \in I$.

Ordering of sequences For the sake of presentation, we introduce a partial ordering relation among sequences. Given two sequences $\sigma = (I,a) \in \Sigma_r$ and $\sigma' = (J,b) \in \Sigma_r$, we write $\sigma \preceq \sigma'$ (read as σ *precedes* σ'), whenever there exists a path in the tree connecting a node in I to a node in J, and such a path includes action a. We adopt analogous definitions for sequences in Σ_s and Σ_c.

2.4. Additional notation

We introduce some additional notation that will be useful in the proofs. Fig. A.4 summarizes the notation used in the paper.

We denote by $\Pi_r := \mathcal{X}_r \cap \{0,1\}^{|\Sigma_r|}$ the set of *deterministic* sequence-form strategies (a.k.a. *pure strategies*) of the receiver, which are the strategies specifying to play a single action with probability one at each infoset. The set of receiver's deterministic strategies in the SDM sub-problem that starts from an infoset $I \in \mathcal{I}$ is denoted as $\Pi_{r,I} := \mathcal{X}_{r,I} \cap \{0,1\}^{|\Sigma_r|}$. Moreover, we let $\Sigma_{r,I} \subseteq \Sigma_r$ be the set of receiver's sequences in the SDM sub-problem that starts from an infoset $I \in \mathcal{I}$; formally, $\Sigma_{r,I} := \{\sigma \in \Sigma_r \mid \sigma_r(I) \preceq \sigma \wedge \exists z \in \mathcal{Z}(I) : \sigma \preceq \sigma_r(z)\}$.

Given any infoset $I \in \mathcal{I}$, we let $\mathcal{Z}(I) \subset \mathcal{Z}$ be the set of terminal nodes $z \in \mathcal{Z}$ such that the path from the root of the tree to z passes through a node in I. Moreover, given $\sigma = (I,a)$ with $a \in A(I)$, we define $\mathcal{Z}(\sigma) = \mathcal{Z}(I,a) \subset \mathcal{Z}(I)$ as the set of terminal nodes whose corresponding paths include playing action a at a node in I. For every infoset $I \in \mathcal{I}$, we also introduce a function $h_I : Z(I) \to I$ such that $h_I(z)$ defines the unique node $h \in I$ on the path from the root of the tree to z.

Given an infoset $I \in \mathcal{I}$ and an action $a \in A(I)$, we define $C(I,a) \subseteq \mathcal{I}$ as the set of all the infosets which immediately follow infoset I through action a, *i.e.*, those infosets $J \in \mathcal{I}$ such that $\sigma_r(J) = (I,a)$. Moreover, we let $C(I) \subseteq \mathcal{I}$ be the set of all infosets that follow $\mathcal{I}$, *i.e.*, those infosets $J \in \mathcal{I}$ such that there exists $a \in A(I)$ with $\sigma = (I,a)$ such that $\sigma \preceq \sigma_r(J)$.

3. Example of BPSDM problem

Initially, we illustrate a simple instance of the SDM problem to further clarify the notation and the concept introduced in the previous section and its relationship with the BPDSM problem.

Example of BPSDM Fig. 1 shows a tree whose set of chance nodes is $\mathcal{H}_c = \{h_0\}$, while the set of decision nodes is $\mathcal{H}_d = \{h_1, h_2, h_3\}$. The set of terminal nodes is $\mathcal{Z} = \{z_1, \dots, z_6\}$. Moreover, the set of decision nodes $\mathcal{H}_d$ is partitioned into the set partition $\mathcal{I} = \{I, J\}$, which is made by two infosets $I = \{h_1\}$ and $J = \{h_2, h_3\}$.

The sets of sequences are constructed as follows. For the chance, we have that $\Sigma_c = \{(h_0,a),(h_0,b),(h_0,c)\}$, while for the agent we have that $\Sigma_r = \{(I,d),(I,e),(J,f),(J,g)\}$. Let us remark that, since the agent cannot distinguish between nodes h_2 and h_3, it only has 2 sequences originating from such nodes; namely (J,f) and (J,g). In the BPSDM problem, the agent plays the role of the receiver, while the sender can be thought of as a perfect-information agent selecting action recommendations for the receiver at decision nodes. Therefore, the sender's set of sequences is $\Sigma_s = \{(h_1,d),(h_1,e),(h_2,f),(h_2,g),(h_3,f),(h_3,g)\}$.

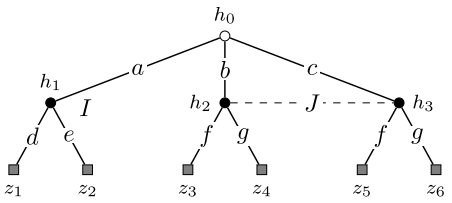


Fig. 1. Example of SDM problem and its sets of sequences Σ_r, Σ_s, and Σ_c.

4. Learning to persuade

In this work, we relax the strong assumption that both the sender and the receiver know the prior $\boldsymbol{\mu}^\star$ by casting the BPSDM problem into an *online learning framework* in which the sender repeatedly interacts with the receiver over a time horizon of length T. At each round $t \in [T]$, the interaction goes as follows[3]: (i) the sender commits to a signaling scheme $\boldsymbol{\phi}_t \in \Phi$; (ii) a vector $\boldsymbol{y}_t \in \{0,1\}^{|\Sigma_c|}$ encoding realizations of random events is drawn according to $\boldsymbol{\mu}^\star$; (iii) the sender and the receiver play an instance of the (one-shot) BPSDM problem (detailed in Section 2.2), in which the sender commits to $\boldsymbol{\phi}_t$, random events at chance nodes are realized as defined by $\boldsymbol{y}_t$, and the receiver sticks to the recommendations issued by the sender; and (iv) the sender observes a *feedback* on the realization of random events at chance nodes, which can be of two types: *full feedback* when the sender observes $\boldsymbol{y}_t$, which specifies the realizations of *all* the random events at chance nodes that are possibly reachable during the SDM problem; *bandit feedback* when the sender observes the terminal node $z_t \in \mathcal{Z}$ reached at the end of the SDM problem. The latter is equivalent to observing the realizations of random events at the chance nodes that are actually reached during the SDM problem, namely $\sigma_c(z_t)$.

By letting $\Phi^\diamond(\boldsymbol{\mu}^\star)$ be the set of persuasive signaling schemes, *i.e.*, such that the receiver is incentivized to follow recommendations (a formal definition is provided in Definition 2), the goal of the sender is to select a sequence of signaling schemes, namely $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_T$, which maximizes their expected utility, while guaranteeing that each signaling scheme $\boldsymbol{\phi}_t$ is persuasive, namely $\boldsymbol{\phi}_t \in \Phi^\diamond(\boldsymbol{\mu}^\star)$.

We measure the performance of a sequence $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_T$ of signaling schemes by comparing it with an optimal (fixed) persuasive signaling scheme. Formally, given a signaling scheme $\boldsymbol{\phi} \in \Phi$, we first define $U(\boldsymbol{\phi}, \boldsymbol{\mu}^\star)$, respectively $F(\boldsymbol{\phi}, \boldsymbol{\mu}^\star)$, as the expected utility achieved by the receiver, respectively the sender, whenever the former follows action recommendations. These can be expressed as linear functions of $\boldsymbol{\phi}$, which, for any $\boldsymbol{\mu} \in \mathcal{X}_c$, are defined as follows:

$$U(\boldsymbol{\phi}, \boldsymbol{\mu}) := \sum_{z \in \mathcal{Z}} \boldsymbol{\mu}[\sigma_c(z)] \boldsymbol{\phi}[\sigma_s(z)] u(z), \quad F(\boldsymbol{\phi}, \boldsymbol{\mu}) := \sum_{z \in \mathcal{Z}} \boldsymbol{\mu}[\sigma_c(z)] \boldsymbol{\phi}[\sigma_s(z)] f(z).$$

Finally, by letting $\boldsymbol{\phi}^\star \in \operatorname{argmax}_{\boldsymbol{\phi} \in \Phi^\diamond(\boldsymbol{\mu}^\star)} F(\boldsymbol{\phi}, \boldsymbol{\mu}^\star)$ be an optimal (fixed) persuasive signaling scheme, the sender' performance over T rounds is measured by the *(cumulative) sender's regret*:

$$R_T := \sum_{t \in [T]} \Big(F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}^\star) - F(\boldsymbol{\phi}_t, \boldsymbol{\mu}^\star) \Big).$$

The goal is to design learning algorithms (for the sender) which select sequences of persuasive signaling schemes such that R_T grows asymptotically sub-linearly in T, namely $R_T = o(T)$.

5. On the characterization of persuasive signaling schemes

5.1. A local decomposition of persuasiveness

In this section, we formally introduce the set of persuasive signaling schemes $\Phi^\diamond(\boldsymbol{\mu}^\star)$ as the set of signaling schemes for which the receiver's expected utility by following recommendations is greater than the one provided by an optimal *deviation policy* (DP).[4] In addition, we show how to decompose any DP into components defined locally at each infoset, which will be crucial in the following Section 5.2. Intuitively, a DP for the receiver is specified by two elements: (i) a set of *deviation points* in which the DP prescribes to stop following action recommendations; and (ii) the *continuation strategies* to be adopted after deviating from recommendations.

We represent deviation points by vectors $\boldsymbol{\omega} \in \{0,1\}^{|\Sigma_r|}$, which are defined so that $\boldsymbol{\omega}[\sigma] = 1$ if and only if the DP prescribes to deviate upon observing the sequence of action recommendations $\sigma \in \Sigma_r$. Moreover, by leveraging the w.l.o.g. assumption that the sender stops issuing recommendations after the receiver deviated from them, we focus on DPs such that each path from the root of the tree to a terminal node involves only one deviation point. This is standard result when dealing with sequential recommendations (see [37] for an example). As a result, the set of all valid vectors $\boldsymbol{\omega} \in \{0,1\}^{|\Sigma_r|}$ is formally defined as

$$\Omega := \left\{ \boldsymbol{\omega} \in \{0,1\}^{|\Sigma_r|} \;\middle|\; \sum_{\sigma \in \Sigma_r : \sigma \preceq \sigma_r(z)} \boldsymbol{\omega}[\sigma] \leq 1 \quad \forall z \in \mathcal{Z} \right\}.$$

[3] Throughout this work, for $n \in \mathbb{N}$, we denote with $[n]$ the set $\{1, \dots, n\}$.

[4] For ease of exposition, all the definitions and results in this section are provided for the prior $\boldsymbol{\mu}^\star$. It is straightforward to generalize them to the case of a generic $\boldsymbol{\mu} \in \mathcal{X}_c$.

We represent the continuation strategies of DPs by introducing the set of *continuation strategy profiles*, denoted as $\mathcal{P} := \bigtimes_{\sigma=(I,a)\in\Sigma_r} \mathcal{X}_{r,I}$. A continuation strategy profile $\boldsymbol{\rho} \in \mathcal{P}$, with $\boldsymbol{\rho} = (\boldsymbol{\rho}_\sigma)_{\sigma\in\Sigma_r}$, defines a strategy $\boldsymbol{\rho}_\sigma \in \mathcal{X}_{r,I}$ for every receiver's sequence $\sigma = (I, a) \in \Sigma_r$. Intuitively, $\boldsymbol{\rho}_\sigma$ is the strategy for the SDM sub-problem starting from infoset I that is used by the receiver after deviating upon observing sequence $\sigma \in \Sigma_r$. As a result, any pair $(\boldsymbol{\omega}, \boldsymbol{\rho}) \in \Omega \times \mathcal{P}$ specifies a valid DP; formally:

Definition 1 *(Deviation policy).* Given a vector $\boldsymbol{\omega} \in \Omega$ and a profile $\boldsymbol{\rho} \in \mathcal{P}$, the $(\boldsymbol{\omega}, \boldsymbol{\rho})$-*DP* prescribes to follow sender's recommendations until action a is recommended at infoset I for some sequence $\sigma = (I, a)$ such that $\boldsymbol{\omega}[\sigma] = 1$; from that point on, it prescribes to play according to strategy $\boldsymbol{\rho}_\sigma$.

We denote by $U^{\boldsymbol{\omega}\to\boldsymbol{\rho}}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star)$ the receiver's expected utility obtained with a $(\boldsymbol{\omega}, \boldsymbol{\rho})$-DP, so that we can state the following formal definition of persuasive signaling schemes.

Definition 2 *(Persuasiveness).* A signaling scheme $\boldsymbol{\phi} \in \Phi$ is ϵ-*persuasive*, namely $\boldsymbol{\phi} \in \Phi^\diamond_\epsilon(\boldsymbol{\mu}^\star)$, if

$$\max_{(\boldsymbol{\omega},\boldsymbol{\rho})\in\Omega\times\mathcal{P}} U^{\boldsymbol{\omega}\to\boldsymbol{\rho}}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) - U(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) \leq \epsilon. \tag{1}$$

Moreover, a signaling scheme $\boldsymbol{\phi} \in \Phi$ is *persuasive*, namely $\boldsymbol{\phi} \in \Phi^\diamond(\boldsymbol{\mu}^\star)$, if it is 0-persuasive.

Intuitively, the above definition states that a signaling scheme is ϵ-persuasive if the receiver's expected utility by following recommendations is at most ϵ less than the one obtained by an optimal DP, which is a DP maximizing receiver's expected utility.

Our local decomposition of DPs is based on suitably defined, simple deviation policies, which we call *single-point DPs* (SPDPs). These are special cases of DPs that stop following the sender's action recommendations only when a specific single infoset is reached and a particular action is recommended therein. SPDPs are formally defined as follows:

Definition 3 *(Single-point deviation strategy).* Given a receiver's sequence $\sigma = (I, a) \in \Sigma_r$ and a receiver's strategy $\boldsymbol{\rho}_\sigma \in \mathcal{X}_{r,I}$ for the SDM sub-problem starting from infoset I, the $(\sigma, \boldsymbol{\rho}_\sigma)$-*SPDP* prescribes to follow sender's recommendations until action a is recommended at infoset I; from that point on, the strategy prescribes to play according to $\boldsymbol{\rho}_\sigma$.

We denote by $U_{\sigma\to\boldsymbol{\rho}_\sigma}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star)$ the receiver's expected utility obtained by following an $(\sigma, \boldsymbol{\rho}_\sigma)$-SPDP. The following theorem provides the key result underlying our decomposition.[5] It shows that the difference between the utility achieved by a $(\boldsymbol{\omega}, \boldsymbol{\rho})$-DP and that obtained by following recommendations can be decomposed into the sum over all the sequences $\sigma \in \Sigma_r$ of analogous differences defined for the $(\sigma, \boldsymbol{\rho}_\sigma)$-SPDPs, where each difference is weighted by $\boldsymbol{\omega}[\sigma]$.

Theorem 1. *Given a signaling scheme $\boldsymbol{\phi} \in \Phi$ and a $(\boldsymbol{\omega}, \boldsymbol{\rho})$-DP, it holds:*

$$U^{\boldsymbol{\omega}\to\boldsymbol{\rho}}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) - U(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) = \sum_{\sigma\in\Sigma_r} \boldsymbol{\omega}[\sigma]\Big(U_{\sigma\to\boldsymbol{\rho}_\sigma}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) - U(\boldsymbol{\phi}, \boldsymbol{\mu}^\star)\Big).$$

5.2. A polytopal approximation of the set of persuasive signaling schemes

In the following, we show how to exploit Theorem 1 to provide an approximate characterization of the set $\Phi^\diamond_\epsilon(\boldsymbol{\mu}^\star)$ using a polynomially-sized polytope. First, we state a corollary of Theorem 1 showing that persuasiveness can be bounded by suitably defined SPDPs. Formally, we have[6]:

Corollary 1. *Given a signaling scheme $\boldsymbol{\phi} \in \Phi$, the following holds:*

$$\max_{(\boldsymbol{\omega},\boldsymbol{\rho})\in\Omega\times\mathcal{P}} U^{\boldsymbol{\omega}\to\boldsymbol{\rho}}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) - U(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) \leq \sum_{\sigma=(I,a)\in\Sigma_r} \left[\max_{\boldsymbol{\rho}_\sigma\in\mathcal{X}_{r,I}} U_{\sigma\to\boldsymbol{\rho}_\sigma}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) - U(\boldsymbol{\phi}, \boldsymbol{\mu}^\star)\right]^+.$$

By exploiting Corollary 1, we introduce the following definition of ϵ-*persuasive polytope* (Lemma 1 justifies the term polytope), as the set of signaling schemes for which there is no $(\sigma, \boldsymbol{\rho}_\sigma)$-SPDP that achieves a receiver's utility that exceeds by more than $\epsilon/|\Sigma_r|$ that achieved by following the recommendations.

Definition 4 *(Persuasive polytope).* The ϵ-*persuasive polytope* is defined as:

$$\Lambda_\epsilon(\boldsymbol{\mu}^\star) := \left\{\boldsymbol{\phi} \in \Phi \;\middle|\; \max_{\boldsymbol{\rho}_\sigma\in\mathcal{X}_{r,I}} U_{\sigma\to\boldsymbol{\rho}_\sigma}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) - U(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) \leq \epsilon/|\Sigma_r| \quad \forall\sigma \in \Sigma_r\right\}.$$

[5] All the proofs are provided in Appendix B, Appendix C, Appendix D, and Appendix E.

[6] Given any $x \in \mathbb{R}$, we let $[x]^+ := \max(x, 0)$.

Moreover, we denote by $\Lambda(\boldsymbol{\mu}^\star)$ the 0-persuasive polytope.

As we show in the following lemma, $\Lambda_\epsilon(\boldsymbol{\mu}^\star)$ is an efficiently-representable polytope.

Lemma 1. *The set $\Lambda_\epsilon(\boldsymbol{\mu}^\star)$ can be described using a polynomial number of linear constraints.*

The following lemma shows that the ϵ-persuasive polytope is contained in $\Phi^\diamond_\epsilon(\boldsymbol{\mu}^\star)$, and that the set of persuasive signaling schemes is contained in the former. Formally:

Lemma 2. *It is always the case that $\Phi^\diamond(\boldsymbol{\mu}^\star) \equiv \Lambda(\boldsymbol{\mu}^\star) \subseteq \Lambda_\epsilon(\boldsymbol{\mu}^\star) \subseteq \Phi^\diamond_\epsilon(\boldsymbol{\mu}^\star)$.*

Lemma 2 also implies that the polytope $\Lambda(\boldsymbol{\mu}^\star)$ exactly characterizes the set of persuasive signaling schemes $\Phi^\diamond(\boldsymbol{\mu}^\star)$. Thus, by adding the maximization of the sender's expected utility $F(\boldsymbol{\phi}, \boldsymbol{\mu}^\star)$ on top of the linear constraints describing $\Lambda(\boldsymbol{\mu}^\star)$, we obtain a polynomially-sized linear program for finding an optimal sender's signaling scheme in any instance of the BPSDM problem in which $\boldsymbol{\mu}^\star$ is known.

Theorem 2. *The BPSDM problem can be solved in polynomial time when the prior $\boldsymbol{\mu}^\star$ is known.*

6. Relation to non-sequential Bayesian persuasion

Now, we clarify the relationship between the BPSDM problem we study in this paper and the classical Bayesian persuasion framework introduced by Kamenica and Gentzkow [2]. In particular, we show that any instance of the classical Bayesian persuasion problem can be mapped to an instance of the BPSDM problem.

A Bayesian persuasion problem instance is defined by a set $\mathcal{A}$ of $k := |\mathcal{A}|$ actions for the receiver, a set $\mathcal{S}$ of signals for the sender, and a set Θ of $d := |\Theta|$ possible outcomes of a (single) random event (called *states of nature* in the classical Bayesian persuasion terminology). The receiver's payoff function is $u^{\mathsf{R}} : \Theta \times \mathcal{A} \to [0,1]$, while the sender's one is $u^{\mathsf{S}} : \Theta \times \mathcal{A} \to [0,1]$. The sender observes the realized state of nature, which is drawn according to a commonly-known prior distribution $\mu \in \Delta_\Theta$. Then, they partially disclose information about the state by committing to a signaling scheme $\varphi : \Theta \to \Delta_{\mathcal{S}}$, which is a randomized mapping from states of nature to signals for the receiver. Thus, the interaction between the sender and the receiver is as follows.

(i) The sender commits to a publicly known signaling scheme φ.
(ii) The sender observes the state of nature $\theta \sim \mu$.
(iii) The sender samples a signal $s \sim \varphi(\theta, \cdot)$ and sends it to the receiver.
(iv) The receiver computes their posterior belief over the states Θ.
(v) The receiver plays an action $a \in \mathcal{A}$ that maximizes their expected payoff.

The posterior beliefs that the receiver computes in step (iv) after observing a signal $s \in \mathcal{S}$ are defined by a probability distribution $\xi_s \in \Delta_\Theta$ such that:

$$\xi_s(\theta) := \frac{\mu(\theta)\varphi(\theta, s)}{\sum_{\theta' \in \Theta} \mu(\theta')\varphi(\theta', s)},$$

and, thus, after observing signal s the receiver plays an action

$$a \in \arg\max_{a' \in \mathcal{A}} \sum_{\theta \in \Theta} \xi_s(\theta) u^{\mathsf{R}}(\theta, a').$$

A revelation-principle-style argument [2] allows the sender to focus on direct and persuasive signaling schemes, where the latter property means that $\mathcal{S} \equiv \mathcal{A}$, with signals corresponding to actions recommendations for the receiver. A persuasive signaling scheme $\varphi : \Theta \to \Delta_{\mathcal{S}}$ is such that the receiver is always incentivized to follow action recommendations; formally:

$$\sum_{\theta \in \Theta} \mu(\theta)\varphi(\theta, a) u^{\mathsf{R}}(\theta, a) \geq \sum_{\theta \in \Theta} \mu(\theta)\varphi(\theta, a) u^{\mathsf{R}}(\theta, a') \quad \forall a, a' \in \mathcal{A}. \tag{2}$$

Mapping between classical, non-sequential Bayesian persuasion to BPSDM Given an instance of the classical Bayesian persuasion problem [2], a corresponding (equivalent) instance of our BPSDM problem can be constructed as follows.

(1) There is a unique chance node h_0 which is the root of the tree defining the SDM problem.
(2) At the chance node, there are d possible outcomes (namely $A(h_0) \equiv \Theta$), each corresponding to a state of nature $\theta \in \Theta$ and having probability $\mu(\theta)$ of occurring, so that with a slight abuse of notation we can write $\boldsymbol{\mu}^\star[\varnothing] = 1$ and $\boldsymbol{\mu}^\star[\theta] = \mu(\theta)$ for all $\theta \in \Theta$.
(3) The receiver has a unique infoset I, which contains one decision node for each possible outcome at the chance node.
(4) At infoset I, the receiver has a set $A(I) \equiv \mathcal{A}$ of available actions.

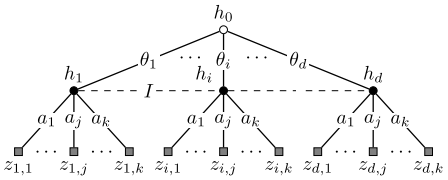


Fig. 2. Instance of BPSDM problem corresponding to a given instance of Bayesian persuasion problem.

(5) Terminal nodes $\mathcal{Z}$ are determined by the state of nature, receiver's action pairs, so that each $\theta_i \in \Theta$ and $a_j \in \mathcal{A}$ define a corresponding terminal node $z_{i,j}$ in the SDM problem.

See Fig. 2 for an example of the mapping. The following theorem formally states that our definition of persuasiveness (Definition 2) instantiated to the BPSDM problem instances described above is equivalent to the definition of persuasiveness for classical Bayesian persuasion problems (Equation (2)). This establishes that our framework encompasses classical Bayesian persuasion problems as a special case.

Theorem 3. *Given any Bayesian persuasion instance, a signaling scheme is persuasive (Equation (2)) if and only if it is persuasive (Definition 2) in the corresponding instance of the BPSDM problem.*

Proof. It is sufficient to prove the equivalence between Equation (1) for $\epsilon = 0$ and Equation (2) applied to the BPSDM problem instance representing the given Bayesian persuasion instance. To do that, we employ Theorem 1 and Lemma 6 in such a BPSDM problem instance, so that, using the notation introduced in this section, it is straightforward to see that Equation (1) reads as follows:

$$\max_{a' \in \mathcal{A}} \sum_{\theta \in \Theta} \varphi(\theta, a)\mu(\theta)u^{\mathsf{R}}(\theta, a') - \sum_{\theta \in \Theta} \varphi(\theta, a)\mu(\theta)u^{\mathsf{R}}(\theta, a) \leq 0 \quad \forall a \in \mathcal{A}.$$

By rearranging the terms, we get Equation (2), which concludes the proof. □

7. Always being persuasive is impossible: a relaxation is needed

In this section, we prove that it is impossible to design an algorithm that returns a sequence of persuasive signaling schemes for a generic BPSDM problem.

Theorem 4 *(Impossibility of persuasiveness). There exists a constant $\gamma \in (0, 1)$ such that no algorithm can guarantee to output a sequence $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_T$ of signaling schemes such that, with probability at least γ, all the signaling schemes $\boldsymbol{\phi}_t$ are persuasive.*

Notice that this result is not in contrast with the work by Zu et al. [24] which provides a no-regret algorithm that outputs sequences of signaling schemes that are guaranteed to be persuasive with high probability. Indeed, Zu et al. [24] require regularity assumptions over the prior, while in our result the prior can be any. Theorem 4 motivates the introduction of a less restrictive requirement on the signaling schemes. In particular, we look for algorithms that output signaling schemes $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_T$, such that the expected utility loss incurred by the receiver by following recommendations rather than playing an optimal DP is small. To capture such a requirement, we introduce the following definition of *(cumulative) receiver's regret*:

$$V_T := \sum_{t \in [T]} \max_{(\omega, \rho) \in \Omega \times \mathcal{P}} U^{\omega \to \rho}(\boldsymbol{\phi}_t, \boldsymbol{\mu}^\star) - \sum_{t \in [T]} U(\boldsymbol{\phi}_t, \boldsymbol{\mu}^\star).$$

Therefore our goal becomes that of designing algorithms guaranteeing that the cumulative receiver's regret grows sub-linearly in T, namely $V_T = o(T)$, while continuing to ensure that $R_T = o(T)$.

In Sections 8 and 9, we design algorithms achieving sub-linear V_T and R_T for the learning problem described in Section 4. The algorithms implement two functions: (i) SELECTSTRATEGY(), which, at each $t \in [T]$, draws a signaling scheme $\boldsymbol{\phi}_t \in \Phi$ on the basis of the internal state of the algorithm; and (ii) UPDATE(o_t), which modifies the internal state based on the observation o_t received as feedback. Each algorithm alternates these two functions as the interaction between the sender and the receiver unfolds as described in Section 4. Specifically, under full feedback the sender observes $\boldsymbol{y}_t$ and calls UPDATE($\boldsymbol{y}_t$), while in the bandit feedback it observes z_t and calls UPDATE(z_t).

8. Learning with full feedback

In this section, we will discuss the online problem faced by the sender that wants to optimize online its utility $F(\boldsymbol{\phi}, \boldsymbol{\mu}^\star)$ while learning the unknown prior $\boldsymbol{\mu}^\star$. We start by providing a learning algorithm (Algorithm 1) working with full feedback, *i.e.*, when the sender observes the realizations of *all* the possible random events. The main idea of the algorithm is to choose signaling schemes $\boldsymbol{\phi}_t$ that belong to suitable sets $\Lambda_{\beta_t}(\widehat{\boldsymbol{\mu}}_t)$ which are designed to be "close" to the set $\Phi^\circ(\boldsymbol{\mu}^\star)$ of persuasive signaling schemes. At each round

Algorithm 1 Full-feedback algorithm.

function SELECTSTRATEGY():
  $\boldsymbol{\phi}_t \leftarrow \arg\max_{\boldsymbol{\phi} \in \Lambda_{\beta_t}(\widehat{\boldsymbol{\mu}}_t)} F(\boldsymbol{\phi}, \widehat{\boldsymbol{\mu}}_t)$
  return $\boldsymbol{\phi}_t$

function UPDATE($\boldsymbol{y}_t$):
  $\widehat{\boldsymbol{\mu}}_{t+1}[\sigma] \leftarrow \sum_{\tau=1}^{t} \boldsymbol{y}_\tau[\sigma]/t \ \forall \sigma \in \Sigma_c$
  $\epsilon_{t+1} \leftarrow \sqrt{\frac{\log(2T|\Sigma_c|/\delta)}{2t}}$
  $\beta_{t+1} \leftarrow 2|\mathcal{Z}||\Sigma_r|\epsilon_{t+1}$

$t \in [T]$, Algorithm 1 defines the desired set as follows. First, it maintains an estimate $\widehat{\boldsymbol{\mu}}_t$ of $\boldsymbol{\mu}^\star$; formally, it defines a radius ϵ_t such that the event $\mathcal{E} := \left\{ \|\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^\star\|_\infty \leq \epsilon_t \ \forall t \in [T] \right\}$ holds with probability at least $1-\delta$. Second, it defines a parameter β_t such that, conditionally to the realization of the event $\mathcal{E}$, the following two conditions hold: (i) the decision space $\Lambda_{\beta_t}(\widehat{\boldsymbol{\mu}}_t)$ contains the optimal signaling scheme $\boldsymbol{\phi}^\star$; (ii) $\Lambda_{2\beta_t}(\boldsymbol{\mu}^\star)$ contains the signaling scheme $\boldsymbol{\phi}_t$. Intuitively, the first condition is needed to have a low sender's regret, while the second one yields approximately persuasive signaling schemes.[7]

The polytopal approximation that we provide in Section 5.2 plays a crucial role in the complexity of Algorithm 1. Specifically, it allows it to select the desired $\boldsymbol{\phi}_t$ in polynomial time by optimizing over the set $\Lambda_{\beta_t}(\widehat{\boldsymbol{\mu}}_t)$, which can be done efficiently. The use of the set $\Lambda_{\beta_t}(\widehat{\boldsymbol{\mu}}_t)$ over $\Phi^\circ_{\beta_t}(\widehat{\boldsymbol{\mu}}_t)$ is necessary due to the fact that the latter is *not* known to admit an efficient representation. Formally:

Theorem 5. *Given any $\delta \in (0,1)$, with probability at least $1-\delta$, Algorithm 1 guarantees:*

$$R_T = \mathcal{O}\left(|\mathcal{Z}| \sqrt{T \log\left(T|\Sigma_c|/\delta\right)} \right), \quad V_T = \mathcal{O}\left(|\Sigma_r||\mathcal{Z}| \sqrt{T \log\left(T|\Sigma_c|/\delta\right)} \right).$$

Moreover, the algorithm runs in polynomial time.

9. Learning with bandit feedback

Algorithm 2 Bandit-feedback algorithm.

function SELECTSTRATEGY():
  if $t \leq N$ **then** ▷ First Phase
    $\sigma = (h, a) \leftarrow \arg\min_{\sigma \in \Sigma_c} C_t[\sigma]$
    $\Sigma_s \ni \sigma' \leftarrow \sigma_s(h)$
    Choose $\boldsymbol{\phi}_t \in \Phi : \boldsymbol{\phi}_t[\sigma'] = 1$
  else ▷ Second Phase
    $\boldsymbol{\phi}_t \leftarrow \arg\max_{\boldsymbol{\phi} \in \Lambda_{\beta_N}(\widehat{\boldsymbol{\mu}}_N)} \max_{\boldsymbol{\mu} \in C_t(\delta)} F(\boldsymbol{\phi}, \boldsymbol{\mu})$
  return $\boldsymbol{\phi}_t$

function UPDATE(z_t):
  Build path $\boldsymbol{p}_t \in \{0,1\}^{|\Sigma_c|}$ from $\sigma_c(z_t)$
  Sample $\boldsymbol{\pi}_t \sim \boldsymbol{\phi}_t$ s.t. $\boldsymbol{p}_t[\sigma] = 1 \Rightarrow \sigma \in \Sigma_1(\boldsymbol{\pi}_t)$
  for $\sigma \in \Sigma_c$ **do**
    if $\sigma \in \Sigma_1(\boldsymbol{\pi}_t))$ **then**
      $\mathcal{T}[\sigma] \leftarrow \mathcal{T}[\sigma] \cup \{t\}$
    $C_{t+1}[\sigma] \leftarrow |\mathcal{T}[\sigma]|$
    $\widehat{\boldsymbol{\mu}}_{t+1}[\sigma] \leftarrow \frac{1}{C_{t+1}[\sigma]} \sum_{\tau \in \mathcal{T}[\sigma]} \boldsymbol{p}_\tau[\sigma]$
    $\epsilon_{t+1}[\sigma] \leftarrow \sqrt{\frac{\log(4T|\Sigma_c|/\delta)}{2C_{t+1}[\sigma]}}$
  $C_{t+1}(\delta) \leftarrow \left\{ \boldsymbol{\mu} \,\middle|\, |\boldsymbol{\mu}[\sigma] - \widehat{\boldsymbol{\mu}}_{t+1}[\sigma]| \leq \epsilon_{t+1}[\sigma] \, \forall \sigma \in \Sigma_c \right\}$
  $\beta_{t+1} \leftarrow 2|\mathcal{Z}||\Sigma_c| \sqrt{\frac{|\Sigma_c| \log(4T|\Sigma_c|/\delta)}{2(t+1)}}$

In this section, we build on Algorithm 1 to deal with bandit feedback, *i.e.*, when at each round $t \in [T]$ the sender only observes the terminal node z_t reached at the end of the SDM problem. The main difficulties of such a setting can be summarized by the following observations. First, the feedback z_t only reveals partial information about the prior, and such information also depends on the selected signaling scheme $\boldsymbol{\phi}_t$. Second, even if the sender plays a signaling scheme $\boldsymbol{\phi} \in \Phi$ for an arbitrarily large number of rounds, there is no guarantee that they collect enough information to tell whether $\boldsymbol{\phi} \in \Phi^\circ_\epsilon(\boldsymbol{\mu}^\star)$ or *not* for some $\epsilon > 0$. Indeed, the persuasiveness of a signaling scheme depends on *all* receiver's utilities in the SDM problem, and some parts of the tree may *not* be reached during a

[7] See Lemmas 9 and 10 in Appendix D for the formal statements of these properties.

sufficiently large number of rounds by committing to $\boldsymbol{\phi}$. Thus, any algorithm for the bandit-feedback setting must guarantee a suitable level of exploration over the entire tree, so as to keep track of the entity of the violation of persuasiveness constraints.

We design a two-phase algorithm, whose pseudo-code is provided in Algorithm 2. The algorithm takes as input the number $N \in [T]$ of rounds devoted to the *first phase* guaranteeing the necessary amount of exploration, as detailed in Section 9.1. During this phase, the SELECTSTRATEGY() procedure implements an efficient deterministic uniform exploration policy, which builds an unbiased estimator $\widehat{\boldsymbol{\mu}}_N$ of $\boldsymbol{\mu}^\star$. This allows us to restrict the space of feasible signaling schemes used in the subsequent phase to those that are approximately persuasive, *i.e.*, those in the set $\Lambda_{\beta_N}(\widehat{\boldsymbol{\mu}}_N)$. In Section 9.2, we discuss the *second phase* of the algorithm, composed by the rounds $t > N$, during which the algorithm focuses on the minimization of sender's regret by exploiting the *optimism in face of uncertainty* principle. Finally, in Section 9.3, we provide a lower bound on the trade-off between the sender's and receiver's regrets, matching the upper bounds achieved by Algorithm 2 for a large portion of the trade-off frontier. This result formally motivates the necessity of the uniform exploration which is performed in the first phase of the algorithm.

9.1. Minimizing the receiver's regret

At each round $t \in [T]$, the sender observes a terminal node $z_t \in \mathcal{Z}$ that uniquely determines a path in the tree defining the SDM problem. We encode such a path using a vector $\boldsymbol{p}_t \in \{0,1\}^{|\Sigma_c|}$ such that $\boldsymbol{p}_t[\sigma] = 1$ if and only if the chance sequence $\sigma \in \Sigma_c$ lies on the path from the root of the tree to z_t, namely $\sigma \preceq \sigma_c(z_t)$. If the sender commits to a signaling scheme $\boldsymbol{\phi}_t \in \Phi$, then it is easy to see that, for every $\sigma = (h,a) \in \Sigma_c$, the element $\boldsymbol{p}_t[\sigma]$ is distributed as a Bernoulli of parameter $\boldsymbol{\phi}_t[\sigma_s(h)]\boldsymbol{\mu}^\star[\sigma]$. The crucial observation behind the design of our estimator is that, if the sender commits to a deterministic signaling scheme $\boldsymbol{\pi}_t \in \Pi$ at some round $t \in [T]$, then for all the chance sequences $\sigma \in \Sigma_c$ that are *compatible* with $\boldsymbol{\pi}_t$, *i.e.*, that can be observed when $\boldsymbol{\pi}_t$ is played, we have that $\boldsymbol{p}_t[\sigma]$ is distributed as a Bernoulli of parameter $\boldsymbol{\mu}^\star[\sigma]$. Formally, a sequence $\sigma \in \Sigma_c$ is compatible with $\boldsymbol{\pi}_t$ if there exists a chance node $h \in \mathcal{H}_c$ and an outcome $a \in A(h)$ satisfying $\sigma = (h,a)$ and $\boldsymbol{\pi}_t[\sigma_s(h)] = 1$. This observation leads to the following result:

Lemma 3. *For every deterministic signaling scheme $\boldsymbol{\pi} \in \Pi$, let*

$$\Sigma_\downarrow(\boldsymbol{\pi}) := \left\{\sigma = (h,a) \in \Sigma_c \mid a \in A(h) \wedge \boldsymbol{\pi}[\sigma_s(h)] = 1\right\}.$$

It holds $\mathbb{E}\left[\boldsymbol{p}_t[\sigma] | \boldsymbol{\pi}_t = \boldsymbol{\pi}\right] = \boldsymbol{\mu}^\star[\sigma]$ for every $\boldsymbol{\pi} \in \Pi$ and $\sigma \in \Sigma_\downarrow(\boldsymbol{\pi})$.

Notice that, since in the rounds $t \leq N$, we only employ deterministic strategies we have that $\mathbb{E}[\boldsymbol{p}_t[\sigma]] = \mu^\star[\sigma]$, where the expectation is only of the random chance events. Thus, during the first phase, Algorithm 2 builds the desired estimator $\widehat{\boldsymbol{\mu}}_N$ of $\boldsymbol{\mu}^\star$ as follows. At each round $t \leq N$, after observing the feedback z_t, the algorithm samples a deterministic signaling scheme $\boldsymbol{\pi}_t \in \Pi$ according to $\boldsymbol{\phi}_t$ (the one selected at t), so that all the sequences $\sigma \in \Sigma_c$ such that $\boldsymbol{p}_t[\sigma] = 1$ (or, equivalently, $\sigma \preceq \sigma_c(z_t)$) belong to $\Sigma_\downarrow(\boldsymbol{\pi}_t)$.[8] Then, for every $\sigma \in \Sigma_\downarrow(\boldsymbol{\pi}_t)$, the algorithm updates the estimator component $\widehat{\boldsymbol{\mu}}_t[\sigma]$ according to $\boldsymbol{p}_t[\sigma]$. Since the probability of visiting a sequence $\sigma \in \Sigma_c$ depends on $\boldsymbol{\phi}_t$ (and, thus, can be arbitrarily small), the first N rounds must be carefully used to ensure that each sequence is explored at least $N/|\Sigma_c|$ times. To explore a specific sequence $\sigma \in \Sigma_c$, we choose a signaling scheme $\boldsymbol{\phi}_t$ such that $\sigma \in \Sigma_\downarrow(\boldsymbol{\pi}_t)$ for every deterministic $\boldsymbol{\pi}_t \sim \boldsymbol{\phi}_t$. The procedure described above is needed for minimizing the receiver's regret, since, in the second phase, the algorithm selects signaling schemes $\boldsymbol{\phi}_t$ from $\Lambda_{\beta_N}(\widehat{\boldsymbol{\mu}}_N)$. In particular, as shown by the following lemma, Algorithm 2 guarantees that the receiver's regret is upper bounded by $2\beta_N$ at each round $t > N$, since it defines $\epsilon_t[\sigma]$ for each sequence $\sigma \in \Sigma_c$ so that the event $\tilde{\mathcal{E}} := \{|\boldsymbol{\mu}^\star[\sigma] - \widehat{\boldsymbol{\mu}}_t[\sigma]| \leq \epsilon_t[\sigma] \ \forall (t,\sigma) \in [T] \times \Sigma_c\}$ holds with probability at least $1 - \delta/2$.

Lemma 4. *Under the event $\tilde{\mathcal{E}}$, Algorithm 2 guarantees that $\boldsymbol{\phi}_t \in \Lambda_{2\beta_N}(\boldsymbol{\mu}^\star)$ at each round $t > N$.*

9.2. Minimizing the sender's regret

Algorithm 2 also needs to guarantee a small sender's regret. To do so, we would like that $\boldsymbol{\phi}^\star$ is a valid pick for the algorithm, *i.e.*, it belongs to $\Lambda_{\beta_N}(\widehat{\boldsymbol{\mu}}_t)$. However, differently from the full-feedback setting, stopping exploration after the first N round does *not* guarantee optimal rates. To fix this issue, in the second phase, the algorithm selects $\boldsymbol{\phi}_t$ optimistically by maximizing the sender's expected utility $F(\boldsymbol{\phi}, \boldsymbol{\mu})$ over both $\boldsymbol{\phi} \in \Lambda_{\beta_N}(\widehat{\boldsymbol{\mu}}_N)$ and $\boldsymbol{\mu} \in C_t(\delta)$, where $C_t(\delta)$ is a suitably-defined confidence set centered around $\widehat{\boldsymbol{\mu}}_t$ such that $\{\boldsymbol{\mu}^\star \in C_t(\delta)\} \equiv \tilde{\mathcal{E}}$, and, thus, it holds with high probability. This guarantees that $\max_{\boldsymbol{\mu} \in C_t(\delta)} F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}) \geq F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}^\star)$. Formally:

Lemma 5. *If the event $\tilde{\mathcal{E}}$ holds, then, for every round $t > N$, it holds that $\boldsymbol{\phi}^\star \in \Lambda_{\beta_N}(\widehat{\boldsymbol{\mu}}_t)$ and $\max_{\boldsymbol{\mu} \in C_t(\delta)} F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}) \geq F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}^\star)$.*

Thus, $F(\boldsymbol{\phi}_t, \boldsymbol{\mu}^\star) \approx F(\boldsymbol{\phi}_t, \widehat{\boldsymbol{\mu}}_t) \geq \max_{\boldsymbol{\mu} \in C_t(\delta)} F(\boldsymbol{\phi}^\star, \widehat{\boldsymbol{\mu}}) \geq F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}^\star)$ holds in the limit, implying that $F(\boldsymbol{\phi}_t, \boldsymbol{\mu}^\star)$ converges to $F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}^\star)$ after sufficiently many rounds. Formally:

[8] The sampling of $\boldsymbol{\pi}_t \in \Pi$ according to $\boldsymbol{\phi}_t$ can be done efficiently by a straightforward modification of the recursive procedure in [40,41].

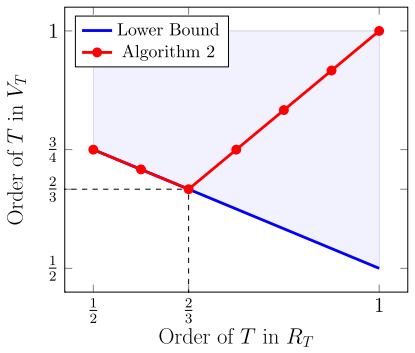


Fig. 3. Trade-off between R_T and V_T in the bandit feedback. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Theorem 6. *Given any $\delta \in (0,1)$ and $N \in [T]$, Algorithm 2 guarantees:*

$$R_T = \mathcal{O}\left(N + \sqrt{\log\left(\frac{T|\Sigma_c|}{\delta}\right)|\Sigma_c|T}\right), \; V_T = \mathcal{O}\left(N + T|\mathcal{Z}||\Sigma_r|\sqrt{\log\left(\frac{T|\Sigma_c|}{\delta}\right)\frac{|\Sigma_c|}{N}}\right),$$

with probability at least $1-\delta$. Moreover, the algorithm runs in polynomial time.

In contrast to the case with full feedback, the optimization problem solved by Algorithm 2 belongs to the class of bilinear problems, which are NP-hard in general [42]. However, in Theorem 6 we prove that our specific problem can be solved in polynomial time. Furthermore, notice that Theorem 6 takes as input the number N of rounds devoted to the first phase. Given an $\alpha \geq 1/2$, by choosing any $N = \lfloor T^\alpha \rfloor$ we get bounds of $R_T = \tilde{\mathcal{O}}(T^\alpha)$ and $V_T = \tilde{\mathcal{O}}(T^{\max\{\alpha, 1-\frac{\alpha}{2}\}})$.

9.3. The lower bound frontier

We conclude by showing that the trade-offs between V_T and R_T achieved by Algorithm 2 are essentially tight. Previously, we provided an intuition as to why the algorithm needs to uniformly explore the entire tree of the SDM problem. Here, we provide a lower bound that corroborates such a statement. In particular, the following theorem shows that, for any $\alpha \in [1/2,\ 1]$, to guarantee a sender's regret of the order of $\mathcal{O}(T^\alpha)$, it is necessary to suffer a receiver's regret of the order of $\Omega(T^{1-\alpha/2})$.[9]

Theorem 7. *For any $\alpha \in [1/2, 1]$, there exists a constant $\gamma \in (0,1)$ such that no algorithm guarantees both $R_T = o(T^\alpha)$ and $V_T = o(T^{1-\alpha/2})$ with probability greater than γ.*

Fig. 3 shows on the horizontal axis the order of the T term in R_T, while, on the vertical axis, it shows the order of the T in V_T. The shaded area over the blue line shows the achievable trade-offs, while the marked red line shows the performances proved in Theorem 6. Thus, we show that Algorithm 2 matches the lower bound for $\alpha \in [1/2, 2/3]$. However, when $\alpha \in [2/3, 1]$, the guarantees proved in Theorem 6 diverge from the ones proved in the lower bound. This is due to the $N = \lfloor T^\alpha \rfloor$ component in the receiver's regret that becomes dominant when $\alpha \geq 2/3$. We conjecture that it is possible to reduce this term to $\sqrt{N}$, hence matching the lower bound of Theorem 7. The reason for such a gap between the lower and upper bounds is that, during the first phase, Algorithm 2 utilizes signaling schemes without taking into account their persuasiveness, thus incurring in large receiver's regret during the first steps. We leave addressing the question of whether it is possible to design exploration strategies by only using approximately-persuasive signaling schemes as future work.

10. Conclusions and future work

In this paper, we address the complex information design problem encountered by an informed sender seeking to influence the behavior of a self-interested receiver through the provision of payoff-relevant information in a sequential game. Our focus is on scenarios where the prior is unknown and needs to be learned online. Previous research in the state-of-the-art has explored specific structures over the prior that enable the design of persuasive no-regret algorithms. However, we prove that when the prior can vary freely, disregarding the prior entirely renders it impossible to guarantee the property of persuasiveness. This finding holds even in non-sequential games. To deal with any prior distribution, we introduce a relaxation of the persuasiveness concept. Specifically, we require that the cumulative violation of the persuasiveness constraint of the receiver grows at a sublinear rate. We present two algorithms

[9] For $\alpha \leq 1/2$, a simple reduction from a standard multi-armed bandit problem provides a lower bound of $\Omega(\sqrt{T})$ on both sender's regret R_T and receiver's regret V_T.

tailored to the full-feedback and bandit-feedback settings, respectively. More precisely, in the full-feedback setting, we provide an algorithm with $\tilde{O}(\sqrt{T})$ regret for both the sender and the receiver, and, in the bandit-feedback setting, we design an algorithm that, given an $\alpha \in [1/2, 1]$ as input, guarantees $\tilde{O}(T^{\alpha})$ and $\tilde{O}(T^{\max\{\alpha, 1-\frac{\alpha}{2}\}})$ regrets, for the sender and the receiver respectively. We also provide a lower bound showing that such a regret trade-off is tight for $\alpha \in [1/2, 2/3]$.

In the future, we will investigate the tightness of our bounds for $\alpha \in (2/3, 1]$. In particular, we conjecture that our lower bound is tight and therefore that the problem admits an algorithm with a regret upper bound better than the algorithm we provide in this paper, for specific values of the trade off between the regret of the sender and the receiver. We will also investigate extensions of our model where, *e.g.*, the number of receivers is larger than one.

CRediT authorship contribution statement

Martino Bernasconi: Investigation, Writing – original draft, Writing – review & editing. **Matteo Castiglioni:** Investigation, Writing – original draft, Writing – review & editing. **Alberto Marchesi:** Investigation, Writing – original draft, Writing – review & editing. **Nicola Gatti:** Supervision, Funding acquisition. **Francesco Trovò:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Italian MIUR PRIN 2022 Project "Targeted Learning Dynamics: Computing Efficient and Fair Equilibria through No-Regret Algorithms", by the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence), and by the EU Horizon project ELIAS (European Lighthouse of AI for Sustainability, No. 101120237).

Appendix A. Notation table

Symbol	Description
Sequential Decision Making	
$\mathcal{Z}, \mathcal{H}_d, \mathcal{H}_c$	Terminal, decision and chance nodes, respectively.
$\mathcal{H} := \mathcal{Z} \cup \mathcal{H}_d \cup \mathcal{H}_c$	Tree nodes.
$u(z)$	Agent's utility at leaf $z \in \mathcal{Z}$
$f(z)$	Sender's utility at leaf $z \in \mathcal{Z}$
$\mathcal{I}$	Receiver's infosets
$A(I)$	Actions from infoset $I \in \mathcal{I}$
Σ_r	Set of receiver's sequences
$\mathcal{X}_r$	Set of receiver's sequence form strategies
$\mathcal{X}_{r,I}$	Set of receiver's sequence form strategies from infoset $I \in \mathcal{I}$
Bayesian Persuasion in Sequential Decision Making	
Φ	Set of sender's signaling schemes
$\Phi^{\circ}_{\epsilon}(\boldsymbol{\mu}^{\star})$	ϵ-persuasive signaling schemes w.r.t. prior $\boldsymbol{\mu}^{\star}$
$U(\boldsymbol{\phi}, \boldsymbol{\mu}^{\star})$	Receiver's utility of following the signaling scheme $\boldsymbol{\phi}$
$F(\boldsymbol{\phi}, \boldsymbol{\mu}^{\star})$	Sender's utility of following the signaling scheme $\boldsymbol{\phi}$
Ω	Receiver's deviation points
$\mathcal{P}$	Set of receiver's continuation policies
$U^{\omega \to \rho}(\boldsymbol{\phi}, \boldsymbol{\mu}^{\star})$	Receiver's utility of (ω, ρ) deviation policy, for $(\omega, \rho) \in \Omega \times \mathcal{P}$
$U_{\sigma \to \rho_\sigma}(\boldsymbol{\phi}, \boldsymbol{\mu}^{\star})$	Receiver's utility of $(\sigma, \boldsymbol{\rho}_\sigma)$ single point deviation policy, for $\sigma = (I, a) \in \Sigma_r$ and $\boldsymbol{\rho}_\sigma \in \mathcal{X}_{r,I}$
$\Lambda_{\epsilon}(\boldsymbol{\mu}^{\star})$	ϵ-persuasive polytope

Fig. A.4. Summary of notation.

Appendix B. Proofs omitted from Section 5

Let us remark that all the results in Section 5 can be straightforwardly generalized to the case of a generic $\boldsymbol{\mu} \in \mathcal{X}_c$, as needed for the proofs of the results in Sections 8 and 9. For ease of exposition, we state and prove the results of Section 5 for the prior $\boldsymbol{\mu}^{\star}$.

First, we prove a preliminary lemma that allows us to express the receiver's expected utility difference between using a $(\sigma, \boldsymbol{\rho}_\sigma)$-SPDP and following action recommendations by only considering the terminal nodes under the infoset in which the SPDP prescribed to deviate. A similar result for the case of correlated strategies can be found in [43, Appendix A].

Lemma 6. *Given $\boldsymbol{\phi} \in \Phi$, for every $(\sigma, \boldsymbol{\rho}_\sigma)$-SPDP with $\sigma = (I, a) \in \Sigma_r$ and $\boldsymbol{\rho}_\sigma \in \mathcal{X}_{r,I}$, it holds:*

$$U_{\sigma \to \boldsymbol{\rho}_\sigma}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) - U(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) = \sum_{z \in \mathcal{Z}(I)} \boldsymbol{\phi}[(h_I(z), a)] \boldsymbol{\rho}_\sigma[\sigma_r(z)] \boldsymbol{\mu}^\star[\sigma_c(z)] u(z) - \sum_{z \in \mathcal{Z}(\sigma)} \boldsymbol{\phi}[\sigma_s(z)] \boldsymbol{\mu}^\star[\sigma_c(z)] u(z).$$

Proof. We define the following three disjoint events for any $(\sigma, \boldsymbol{\rho}_\sigma)$-SPDP, where $\sigma = (I, a)$.

(C1): A terminal node $z \in \mathcal{Z}(\sigma)$ is reached.
(C2): A terminal node $z \in \mathcal{Z}(I, a')$ for some $a' \neq a \in A(I)$ is reached.
(C3): A terminal node $z \in \mathcal{Z} / \mathcal{Z}(I)$ is reached.

Next, under each event, we define the probability $p_{\sigma \to \boldsymbol{\rho}_\sigma}(z)$ of reaching a terminal node z:

(C1): Since $z \in \mathcal{Z}(\sigma)$, the node z is reached by means of the continuation strategy $\boldsymbol{\rho}_\sigma$. Thus:

$$p^{(1)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z) := \boldsymbol{\phi}[(h_I(z), a)] \boldsymbol{\mu}^\star[\sigma_c(z)] \boldsymbol{\rho}_\sigma[\sigma_r(z)].$$

(C2): Since $z \in \mathcal{Z}(I, a')$ for $a' \neq a \in A(I)$, the node z can be reached either by deviating and then committing to the continuation strategy $\boldsymbol{\rho}_\sigma$ or by following recommendations. Moreover, these two cases are exclusive, and, thus, we can write:

$$p^{(2)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z) := \boldsymbol{\phi}[(h_I(z), a)] \boldsymbol{\mu}^\star[\sigma_c(z)] \boldsymbol{\rho}_\sigma[\sigma_r(z)] + \boldsymbol{\phi}[\sigma_s(z)] \boldsymbol{\mu}^\star[\sigma_c(z)].$$

(C3): Since $z \in \mathcal{Z} / \mathcal{Z}(I)$, the node z is reached by following recommendations:

$$p^{(3)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z) := \boldsymbol{\phi}[\sigma_s(z)] \boldsymbol{\mu}^\star[\sigma_c(z)].$$

We observe that $p^{(2)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z) = p^{(1)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z) + p^{(3)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z)$, and, thus, we can write $U_{\sigma \to \boldsymbol{\rho}_\sigma}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star)$ as:

$$\begin{aligned} U_{\sigma \to \boldsymbol{\rho}_\sigma}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) &:= \sum_{z \in \mathcal{Z}(\sigma)} p^{(1)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z) u(z) + \sum_{\substack{z \in \mathcal{Z}(I, a'): \\ a' \neq a \in A(I)}} p^{(2)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z) u(z) + \sum_{z \in \mathcal{Z} / \mathcal{Z}(I)} p^{(3)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z) u(z) \\ &\leq \sum_{z \in \mathcal{Z}(I)} p^{(1)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z) u(z) + \sum_{z \in \mathcal{Z} / \mathcal{Z}(\sigma)} p^{(3)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z) u(z). \end{aligned}$$

Furthermore, by using the definition of $p^{(3)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z)$, we can write $U(\boldsymbol{\phi}, \boldsymbol{\mu}) := \sum_{z \in \mathcal{Z}} u(z) p^{(3)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z)$. Thus:

$$U_{\sigma \to \boldsymbol{\rho}_\sigma}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) - U(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) = \sum_{z \in \mathcal{Z}(I)} p^{(1)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z) u(z) - \sum_{z \in \mathcal{Z}(\sigma)} p^{(3)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z) u(z),$$

which is the statement of the lemma by substituting the definitions of $p^{(1)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z)$ and $p^{(3)}_{\sigma \to \boldsymbol{\rho}_\sigma}(z)$. □

Now, we exploit Lemma 6 to prove the following local decomposition of a DP into SPDPs.

Theorem 1. *Given a signaling scheme $\boldsymbol{\phi} \in \Phi$ and a $(\boldsymbol{\omega}, \boldsymbol{\rho})$-DP, it holds:*

$$U^{\boldsymbol{\omega} \to \boldsymbol{\rho}}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) - U(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) = \sum_{\sigma \in \Sigma_r} \boldsymbol{\omega}[\sigma] \Big(U_{\sigma \to \boldsymbol{\rho}_\sigma}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) - U(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) \Big).$$

Proof. For any terminal node $z \in \mathcal{Z}$, let $p^{\boldsymbol{\omega} \to \boldsymbol{\rho}}(z; \boldsymbol{\phi}, \boldsymbol{\mu}^\star)$ be the probability of reaching node z when the receiver employs the $(\boldsymbol{\omega}, \boldsymbol{\rho})$-DP under the signaling scheme $\boldsymbol{\phi}$ and the prior $\boldsymbol{\mu}^\star$. It holds:

$$\begin{aligned} p^{\boldsymbol{\omega} \to \boldsymbol{\rho}}(z; \boldsymbol{\phi}, \boldsymbol{\mu}^\star) := &\sum_{\sigma = (I, a) \in \Sigma_r : \sigma \leq \sigma_r(z)} \boldsymbol{\omega}[\sigma] \boldsymbol{\phi}[(h_I(z), a)] \boldsymbol{\rho}_\sigma[\sigma_r(z)] \boldsymbol{\mu}^\star[\sigma_c(z)] \\ &+ \boldsymbol{\phi}[\sigma_s(z)] \boldsymbol{\mu}^\star[\sigma_c(z)] \left(1 - \sum_{\sigma \in \Sigma_r : \sigma \leq \sigma_r(z)} \boldsymbol{\omega}[\sigma] \right). \end{aligned}$$

The sum in the first term in the definition of $p^{\boldsymbol{\omega} \to \boldsymbol{\rho}}(z; \boldsymbol{\phi}, \boldsymbol{\mu}^\star)$ accounts for the probabilities of reaching z when the receiver reaches infoset I, is recommended to play action a, and deviates by following the continuation strategy $\boldsymbol{\rho}_\sigma$ thereafter, for all the sequences $\sigma = (I, a)$ that precede the sequence $\sigma_r(z)$ reaching z. Instead, the second term in the definition of $p^{\boldsymbol{\omega} \to \boldsymbol{\rho}}(z; \boldsymbol{\phi}, \boldsymbol{\mu}^\star)$ accounts for the probability of reaching z by following recommendations. Thus, $U^{\boldsymbol{\omega} \to \boldsymbol{\rho}}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) = \sum_{z \in \mathcal{Z}} p^{\boldsymbol{\omega} \to \boldsymbol{\rho}}(z; \boldsymbol{\phi}, \boldsymbol{\mu}^\star) u(z)$.

By rearranging the terms in $U^{\omega\to\rho}(\boldsymbol{\phi},\boldsymbol{\mu}^\star)$, we get to the following result:

$$\begin{aligned}U^{\omega\to\rho}(\boldsymbol{\phi},\boldsymbol{\mu}^\star) &= U(\boldsymbol{\phi},\boldsymbol{\mu}^\star) + \sum_{z\in\mathcal{Z}}\left[\sum_{\sigma=(I,a):\,\sigma\preceq\sigma_r(z)}\boldsymbol{\omega}[\sigma]\boldsymbol{\phi}[(h_I(z),a)]\boldsymbol{\rho}_\sigma[\sigma_r(z)]\boldsymbol{\mu}^\star[\sigma_c(z)]u(z)\right.\\ &\qquad\left. - \sum_{\sigma\in\Sigma_r:\,\sigma\preceq\sigma_r(z)}\boldsymbol{\omega}[\sigma]\boldsymbol{\phi}[\sigma_s(z)]\boldsymbol{\mu}^\star[\sigma_c(z)]u(z)\right]\\ &= U(\boldsymbol{\phi},\boldsymbol{\mu}^\star) - \sum_{\sigma\in\Sigma_r}\boldsymbol{\omega}[\sigma]\sum_{z\in\mathcal{Z}(\sigma)}\boldsymbol{\phi}[\sigma_s(z)]\boldsymbol{\mu}^\star[\sigma_c(z)]u(z)\\ &\quad + \sum_{\sigma\in\Sigma_r}\boldsymbol{\omega}[\sigma]\sum_{z\in\mathcal{Z}(I)}\boldsymbol{\phi}[(h_I(z),a)]\boldsymbol{\rho}_\sigma[\sigma_r(z)]\boldsymbol{\mu}^\star[\sigma_c(z)]u(z).\end{aligned} \tag{B.1}$$

Thus, by combining Lemma 6 with Equation (B.1) we get that:

$$U^{\omega\to\rho}(\boldsymbol{\phi},\boldsymbol{\mu}^\star) - U(\boldsymbol{\phi},\boldsymbol{\mu}^\star) = \sum_{\sigma\in\Sigma_r}\boldsymbol{\omega}[\sigma]\left[U_{\sigma\to\rho_\sigma}(\boldsymbol{\phi},\boldsymbol{\mu}^\star) - U(\boldsymbol{\phi},\boldsymbol{\mu}^\star)\right],$$

which concludes the proof. □

Corollary 1. *Given a signaling scheme $\boldsymbol{\phi}\in\Phi$, the following holds:*

$$\max_{(\boldsymbol{\omega},\boldsymbol{\rho})\in\Omega\times\mathcal{P}} U^{\omega\to\rho}(\boldsymbol{\phi},\boldsymbol{\mu}^\star) - U(\boldsymbol{\phi},\boldsymbol{\mu}^\star) \le \sum_{\sigma=(I,a)\in\Sigma_r}\left[\max_{\rho_\sigma\in\mathcal{X}_{r,I}} U_{\sigma\to\rho_\sigma}(\boldsymbol{\phi},\boldsymbol{\mu}^\star) - U(\boldsymbol{\phi},\boldsymbol{\mu}^\star)\right]^+.$$

Proof. By using Theorem 1, we derive the following:

$$\begin{aligned}\max_{(\boldsymbol{\omega},\boldsymbol{\rho})\in\Omega\times\mathcal{P}} U^{\omega\to\rho}(\boldsymbol{\phi},\boldsymbol{\mu}^\star) - U(\boldsymbol{\phi},\boldsymbol{\mu}^\star) &= \max_{(\boldsymbol{\omega},\boldsymbol{\rho})\in\Omega\times\mathcal{P}}\sum_{\sigma\in\Sigma_r}\boldsymbol{\omega}[\sigma]\left(U_{\sigma\to\rho_\sigma}(\boldsymbol{\phi},\boldsymbol{\mu}^\star) - U(\boldsymbol{\phi},\boldsymbol{\mu}^\star)\right)\\ &\le \max_{(\boldsymbol{\omega},\boldsymbol{\rho})\in\Omega\times\mathcal{P}}\sum_{\sigma\in\Sigma_r}\boldsymbol{\omega}[\sigma]\left[U_{\sigma\to\rho_\sigma}(\boldsymbol{\phi},\boldsymbol{\mu}^\star) - U(\boldsymbol{\phi},\boldsymbol{\mu}^\star)\right]^+\\ &\le \max_{\rho\in\mathcal{P}}\sum_{\sigma\in\Sigma_r}\left[U_{\sigma\to\rho_\sigma}(\boldsymbol{\phi},\boldsymbol{\mu}^\star) - U(\boldsymbol{\phi},\boldsymbol{\mu}^\star)\right]^+\\ &= \sum_{\sigma\in\Sigma_r}\left[\max_{\rho_\sigma\in\mathcal{X}_{r,I}} U_{\sigma\to\rho_\sigma}(\boldsymbol{\phi},\boldsymbol{\mu}^\star) - U(\boldsymbol{\phi},\boldsymbol{\mu}^\star)\right]^+.\end{aligned}$$

This concludes the proof. □

Lemma 1. *The set $\Lambda_\epsilon(\boldsymbol{\mu}^\star)$ can be described using a polynomial number of linear constraints.*

Proof. In order to prove that the set $\Lambda_\epsilon(\boldsymbol{\mu}^\star)$ can be described by means of linear constraints, we employ duality arguments related to the max problem in the definition of $\Lambda_\epsilon(\boldsymbol{\mu}^\star)$ (Definition 4).

By Lemma 6, for every sequence $\sigma=(I,a)\in\Sigma_r$, we can rewrite the expression in the left-hand side of the inequality characterizing $\Lambda_\epsilon(\boldsymbol{\mu}^\star)$ in Definition 4 as follows:

$$\max_{\rho_\sigma\in\mathcal{X}_{r,I}}\left\{\sum_{z\in Z(I)}\boldsymbol{\phi}[(h_I(z),a)]\boldsymbol{\rho}_\sigma[\sigma_r(z)]\boldsymbol{\mu}^\star[\sigma_c(z)]u(z)\right\} - \sum_{z\in Z(\sigma)}\boldsymbol{\phi}[\sigma_s(z)]\boldsymbol{\mu}^\star[\sigma_c(z)]u(z),$$

so that $\Lambda_\epsilon(\boldsymbol{\mu}^\star)$ can be expressed as the set of all $\boldsymbol{\phi}\in\Phi$ such that the above expression has value less than or equal to $\epsilon/|\Sigma_r|$ for every $\sigma\in\Sigma_r$. Observe that the expression in the max operator is a linear function of $\boldsymbol{\rho}_\sigma$, and that the set $\mathcal{X}_{r,I}$ is a polytope by definition. Thus, for every $\sigma=(I,a)\in\Sigma_r$, the maximization above can be equivalently rewritten as the following linear program:

$$\max_{\boldsymbol{x}^{I,a}\ge\mathbf{0}}\ \left(\boldsymbol{x}^{I,a}\right)^\top \boldsymbol{c}(\boldsymbol{\phi},\boldsymbol{\mu}^\star)\quad\text{s.t.} \tag{B.2a}$$

$$\boldsymbol{F}_I\boldsymbol{x}^{I,a} = \boldsymbol{f}_I \tag{B.2b}$$

where $\boldsymbol{x}^{I,a}$ is a vector of variables indexed over sequences $\Sigma_{r,I}\cup\{\sigma_r(I)\}$. Notice that $c(\boldsymbol{\phi},\boldsymbol{\mu}^\star)\in\mathbb{R}^{|\Sigma_{r,I}|}$ is a vector of coefficients such that the component corresponding to each $\sigma'\in\Sigma_{r,I}$ is

$$c(\boldsymbol{\phi},\boldsymbol{\mu}^\star)[\sigma'] := \sum_{z\in\mathcal{Z}(I):\,\sigma_r(z)=\sigma'}\boldsymbol{\phi}[(h_I(z),a)]\boldsymbol{\mu}^\star[\sigma_c(z)]u(z),$$

while $c(\boldsymbol{\phi}, \boldsymbol{\mu}^\star)[\sigma_r(I)] := 0$. Moreover, $\boldsymbol{F}_I \in \{-1, 0, 1\}^{(1+|C(I)|)\times|\Sigma_{r,I}|}$ is a matrix of coefficients whose components are defined as follows: $[\boldsymbol{F}_I]_{I_\varnothing,\sigma_r(I)} := 1$ and $[\boldsymbol{F}_I]_{I_\varnothing,\sigma'} := 0$ for all sequences $\sigma' \in \Sigma_{r,I}$, where $I_\varnothing$ is a fictitious infoset indexing the first row, while, for every infoset $J \in C(I)$ following I (this included) and sequence $\sigma' \in \Sigma_{r,I} \cup \{\sigma_r(I)\}$:

$$[\boldsymbol{F}_I]_{J,\sigma'} := \begin{cases} -1 & \text{if} \quad \sigma' = \sigma_r(J) \\ 1 & \text{if} \quad \sigma' = (J, a') \text{ for some } a' \in A(J) \\ 0 & \text{otherwise} \end{cases}.$$

Finally, $\boldsymbol{f}_I \in \{0, 1\}^{1+|C(I)|}$ is a vector whose components are all zero apart from that one corresponding to the sequence $\sigma_r(I)$, which is one (see also [39]).

The dual linear program of Problem (B.2) reads as:

$$\min_{\boldsymbol{y}^{I,a}} \boldsymbol{y}^{I,a}[I_\varnothing] \quad \text{s.t.} \tag{B.3a}$$

$$\boldsymbol{F}_I^\top \boldsymbol{y}^{I,a} \geq \boldsymbol{c}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star), \tag{B.3b}$$

where $\boldsymbol{y}^{I,a}$ is a vector of dual variables indexed over $C(I) \cup \{I_\varnothing\}$. For ease of notation, we let $\texttt{OPT}_{I,a}$ be the optimal value of Problem (B.3) instantiated for the sequence $\sigma = (I, a)$.

By strong duality, we have that the optimal value of the primal (Problem (B.2)) is equal to the optimal value of the dual (Problem (B.3)), and this allows us to readily rewrite the set $\Lambda_\epsilon(\boldsymbol{\mu}^\star)$ as follows:

$$\Lambda_\epsilon(\boldsymbol{\mu}^\star) = \left\{ \boldsymbol{\phi} \in \Phi \;\middle|\; \texttt{OPT}_{I,a} - \sum_{z \in Z(\sigma)} \boldsymbol{\phi}[\sigma_s(z)]\boldsymbol{\mu}^\star[\sigma_c(z)]u(z) \leq \frac{\epsilon}{|\Sigma_r|} \quad \forall \sigma = (I, a) \in \Sigma_r \right\}. \tag{B.4}$$

Moreover, we can remove $\texttt{OPT}_{I,a}$ in Equation (B.4) since it appears in the right-hand side of a $\leq$ inequality and Problem (B.3) is a min problem. Thus, the set $\Lambda_\epsilon(\boldsymbol{\mu}^\star)$ can be written as follows:

$$\Lambda_\epsilon(\boldsymbol{\mu}^\star) = \left\{ \boldsymbol{\phi} \in \Phi \;\middle|\; \exists \boldsymbol{y}^{I,a} \in \mathbb{R}^{1+|C(I)|} : \boldsymbol{y}^{I,a}[I_\varnothing] - \sum_{z \in Z(\sigma)} \boldsymbol{\phi}[\sigma_s(z)]\boldsymbol{\mu}^\star[\sigma_c(z)]u(z) \leq \frac{\epsilon}{|\Sigma_r|} \right.$$
$$\left. \wedge\, \boldsymbol{F}_I^\top \boldsymbol{y}^{I,a} \geq \boldsymbol{c}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) \quad \forall \sigma = (I, a) \in \Sigma_r \right\}, \tag{B.5}$$

which is comprised of a polynomial number of inequalities and variables, concluding the proof.

Let us also notice that, by expanding the constraints of Problem (B.3), one can easily check that they can be equivalently rewritten recursively, as follows. For every sequence $\sigma' = (J, a') \in \Sigma_{r,I}$, Constraints (B.3b) can be rewritten as:

$$\boldsymbol{y}^{I,a}[J] \geq \sum_{z \in \mathcal{Z}(I) : \sigma_r(z) = (J,a')} \boldsymbol{\phi}[(h_I(z), a)]\boldsymbol{\mu}^\star[\sigma_c(z)]u(z) + \sum_{K \in C(J,a')} \boldsymbol{y}^{I,a}[K], \tag{B.6}$$

while, for sequence $\sigma_r(I)$, Constraint (B.3b) can be written as $\boldsymbol{y}^{I,a}[I_\varnothing] \geq \boldsymbol{y}^{I,a}[I]$. Intuitively, at any optimal solution to Problem (B.3), we can interpret the value of the dual variable $\boldsymbol{y}^{I,a}[I_\varnothing]$ as the receiver's expected utility obtained by playing the best possible continuation strategy after being recommended action a at infoset I. Indeed, the first term in the right-hand-side of Equation (B.6) is the utility immediately obtainable after playing a' at infoset J, while the second term recursively encodes the utilities obtained (non-immediately) following a' at J. □

Lemma 2. *It is always the case that* $\Phi^\diamond(\boldsymbol{\mu}^\star) \equiv \Lambda(\boldsymbol{\mu}^\star) \subseteq \Lambda_\epsilon(\boldsymbol{\mu}^\star) \subseteq \Phi^\diamond_\epsilon(\boldsymbol{\mu}^\star)$.

Proof. First, we prove that $\Phi^\diamond(\boldsymbol{\mu}^\star) \equiv \Lambda(\boldsymbol{\mu}^\star)$. Suppose that $\boldsymbol{\phi} \in \Phi^\diamond(\boldsymbol{\mu}^\star)$. Then, Definition 2 implies

$$U^{\omega \to \rho}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) - U(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) \leq 0,$$

for every $\omega \in \Omega$ and $\rho \in \mathcal{P}$. Thus, by Theorem 1 we have that:

$$\sum_{\sigma \in \Sigma_r} \omega[\sigma] \Big(U_{\sigma \to \rho_\sigma}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) - U(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) \Big) \leq 0,$$

for every $\omega \in \Omega$ and $\rho \in \mathcal{P}$, which implies that:

$$\max_{\rho_\sigma \in \mathcal{X}_{r,I}} U_{\sigma \to \rho_\sigma}(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) - U(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) \leq 0 \quad \forall \sigma \in \Sigma_r,$$

and $\boldsymbol{\phi} \in \Lambda(\boldsymbol{\mu}^\star)$, proving the first part of the statement.

On the other hand, $\Lambda(\boldsymbol{\mu}^\star) \subseteq \Phi^\diamond(\boldsymbol{\mu}^\star)$ is directly implied by Corollary 1. Thus, $\Lambda(\boldsymbol{\mu}^\star) \equiv \Phi^\diamond(\boldsymbol{\mu}^\star)$. Moreover, from Definition 4 it trivially follows that $\Lambda(\boldsymbol{\mu}^\star) \subseteq \Lambda_\epsilon(\boldsymbol{\mu}^\star)$.

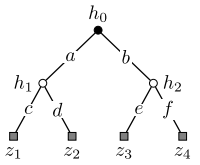


Fig. C.5. Tree structure used in the proof of Theorem 4. Black round nodes are decision nodes $\mathcal{H}_d$. White round nodes are the chance nodes $\mathcal{H}_c$, while gray square nodes are the terminal nodes $\mathcal{Z}$.

Finally, we prove that $\Lambda_\epsilon(\boldsymbol{\mu}^\star) \subseteq \Phi_\epsilon^\diamond(\boldsymbol{\mu}^\star)$. Given $\epsilon > 0$, let $\boldsymbol{\phi} \in \Lambda_\epsilon(\boldsymbol{\mu}^\star)$. By Corollary 1, it holds:

$$\begin{aligned}\max_{(\boldsymbol{\omega},\boldsymbol{\rho})\in\Omega\times\mathcal{P}} U^{\boldsymbol{\omega}\to\boldsymbol{\rho}}(\boldsymbol{\phi},\boldsymbol{\mu}^\star) - U(\boldsymbol{\phi},\boldsymbol{\mu}^\star) &\le \sum_{\sigma=(I,a)\in\Sigma_r} \left[\max_{\rho_\sigma\in\mathcal{X}_{r,I}} U_{\sigma\to\rho_\sigma}(\boldsymbol{\phi},\boldsymbol{\mu}^\star) - U(\boldsymbol{\phi},\boldsymbol{\mu}^\star)\right]^+ \\ &\le \sum_{\sigma=(I,a)\in\Sigma_r} \frac{\epsilon}{|\Sigma_r|} = \epsilon,\end{aligned}$$

which implies that $\boldsymbol{\phi} \in \Phi_\epsilon^\diamond(\boldsymbol{\mu}^\star)$. This concludes the proof. □

Theorem 2. *The BPSDM problem can be solved in polynomial time when the prior $\boldsymbol{\mu}^\star$ is known.*

Proof. It is easy to check that the problem can be written as the following linear program:

$$\max_{\boldsymbol{\phi}\in\Lambda(\boldsymbol{\mu}^\star)} F(\boldsymbol{\phi},\boldsymbol{\mu}^\star),$$

where the objective function is linear and $\Lambda(\boldsymbol{\mu}^\star)$ is a polytope that can be represented by a polynomial number of linear inequalities, by Lemma 1. □

Appendix C. Proofs omitted from Section 7

Theorem 4 *(Impossibility of persuasiveness). There exists a constant $\gamma \in (0,1)$ such that no algorithm can guarantee to output a sequence $\boldsymbol{\phi}_1,\dots,\boldsymbol{\phi}_T$ of signaling schemes such that, with probability at least γ, all the signaling schemes $\boldsymbol{\phi}_t$ are persuasive.*

Proof. We define two instances $\mathfrak{i}$ and $\mathfrak{j}$ of the BPSDM problem based on the tree structure presented in Fig. C.5. In instance $\mathfrak{i}$, respectively $\mathfrak{j}$, the prior is defined as follows:

$$\mathfrak{i} := \begin{cases} \boldsymbol{\mu}^\star[(h_1,c)] = \frac{1}{2} + \epsilon \\ \boldsymbol{\mu}^\star[(h_1,d)] = \frac{1}{2} - \epsilon \\ \boldsymbol{\mu}^\star[(h_2,e)] = \frac{1}{2} - \epsilon \\ \boldsymbol{\mu}^\star[(h_2,f)] = \frac{1}{2} + \epsilon \end{cases},$$

$$\mathfrak{j} := \begin{cases} \boldsymbol{\mu}^\star[(h_1,c)] = \frac{1}{2} - \epsilon \\ \boldsymbol{\mu}^\star[(h_1,d)] = \frac{1}{2} + \epsilon \\ \boldsymbol{\mu}^\star[(h_2,e)] = \frac{1}{2} + \epsilon \\ \boldsymbol{\mu}^\star[(h_2,f)] = \frac{1}{2} - \epsilon \end{cases}.$$

Moreover, for both instances $u(z_1) = u(z_3) = 1$ and $u(z_2) = u(z_4) = 0$. A direct computation shows that, in instance $\mathfrak{i}$, it holds $V_T^{\mathfrak{i}} = 2\epsilon \sum_{t=1}^T \boldsymbol{\phi}_t[(h_0,b)]$, while one can similarly compute that $V_T^{\mathfrak{j}} = 2\epsilon \sum_{t=1}^T \boldsymbol{\phi}_t[(h_0,a)]$. Let $\mathbb{P}^{\mathfrak{i}}$ and $\mathbb{P}^{\mathfrak{j}}$ be the probability measures of instance $\mathfrak{i}$ and $\mathfrak{j}$, respectively. Assume that $\mathbb{P}^{\mathfrak{j}}[V_T^{\mathfrak{j}} \le 0] \ge 1 - \delta$. Then, we know from the Pinsker inequality that:

$$\mathbb{P}^{\mathfrak{i}}\left[\sum_{t=1}^T \boldsymbol{\phi}_t[(h_0,a)] \le 0\right] \ge 1 - \sqrt{\frac{1}{2}\mathcal{K}(\mathfrak{i},\mathfrak{j})} - \delta,$$

where $\mathcal{K}(\mathfrak{i},\mathfrak{j})$ is the Kullback-Leibler divergence between instance $\mathfrak{i}$ and $\mathfrak{j}$. By using the Kullback-Leibler decomposition (see, *e.g.*, [44] for more details), we can state that:

$$\mathcal{K}(\mathfrak{i},\mathfrak{j}) = 2T\mathcal{K}(B_{1/2+\epsilon}, B_{1/2-\epsilon}),$$

where $\mathcal{K}(B_{1/2+\epsilon}, B_{1/2-\epsilon}) \le 16\epsilon^2$ is the Kullback-Leibler divergence between a Bernoulli of parameter $1/2+\epsilon$ and one of parameter $1/2-\epsilon$. Thus:

$$\mathbb{P}^{\mathrm{i}}\left[\sum_{t=1}^{T}\boldsymbol{\phi}_t[(h_0,a)]\leq 0\right]\geq 1-4\epsilon\sqrt{T}-\delta.$$

Moreover, in instance i, we have that $V_T^{\mathrm{i}}=2\epsilon\sum_{t=1}^{T}\boldsymbol{\phi}_t[(h_0,b)]$, which implies:

$$\mathbb{P}^{\mathrm{i}}\left[V_T^{\mathrm{i}}\geq 2\epsilon T\right]\geq 1-4\epsilon\sqrt{T}-\delta.$$

By setting $\epsilon=\frac{1}{16\sqrt{T}}$, we have that:

$$\mathbb{P}^{\mathrm{i}}\left[V_T^{\mathrm{i}}\geq\frac{1}{8}\sqrt{T}\right]\geq 0.75-\delta.$$

Thus, any algorithm that guarantees with high probability $R_T\leq 0$ in instance j fails with high probability in instance i. This proves the claim. □

Appendix D. Proofs omitted from Section 8

Before presenting the proofs of the results in Section 8, we introduce some preliminary lemmas.

Lemma 7. *Given any $\boldsymbol{\phi}\in\Phi$ and $\boldsymbol{\mu},\boldsymbol{\mu}'\in\mathcal{X}_c$, if it is the case that $\boldsymbol{\phi}\in\Lambda_\epsilon(\boldsymbol{\mu})$ and $\|\boldsymbol{\mu}-\boldsymbol{\mu}'\|_\infty\leq\gamma$, then it holds that $\boldsymbol{\phi}\in\Lambda_{\epsilon'}(\boldsymbol{\mu}')$ with $\epsilon'=2|\mathcal{Z}||\Sigma_r|\gamma+\epsilon$.*

Proof. For every $(\sigma,\boldsymbol{\rho}_\sigma)$-SPDP with $\sigma=(I,a)$, the following inequalities hold:

$$\begin{aligned}
&U_{\sigma\to\boldsymbol{\rho}_\sigma}(\boldsymbol{\phi},\boldsymbol{\mu}')-U(\boldsymbol{\phi},\boldsymbol{\mu}')\\
&\quad=\sum_{\mathcal{Z}(I)}\boldsymbol{\phi}[(h_I(z),a)]\boldsymbol{\rho}_\sigma[\sigma_r(z)]\boldsymbol{\mu}'[\sigma_c(z)]u(z)-\sum_{z\in\mathcal{Z}(\sigma)}\boldsymbol{\phi}[\sigma_s(z)]\boldsymbol{\mu}'[\sigma_c(z)]u(z)\\
&\quad\leq\sum_{\mathcal{Z}(I)}\boldsymbol{\phi}[(h_I(z),a)]\boldsymbol{\rho}_\sigma[\sigma_r(z)]\left(\boldsymbol{\mu}'[\sigma_c(z)]-\boldsymbol{\mu}[\sigma_c(z)]\right)u(z)\\
&\qquad-\sum_{z\in\mathcal{Z}(\sigma)}\boldsymbol{\phi}[\sigma_s(z)]\left(\boldsymbol{\mu}'[\sigma_c(z)]-\boldsymbol{\mu}[\sigma_c(z)]\right)u(z)+\frac{\epsilon}{|\Sigma_r|}\\
&\quad\leq 2|\mathcal{Z}|\|\boldsymbol{\mu}-\boldsymbol{\mu}'\|_\infty+\frac{\epsilon}{|\Sigma_r|}\leq 2|\mathcal{Z}|\gamma+\frac{\epsilon}{|\Sigma_r|},
\end{aligned}$$

where in the first inequality we added and subtracted the difference $U_{\sigma\to\boldsymbol{\rho}_\sigma}(\boldsymbol{\phi},\boldsymbol{\mu})-U(\boldsymbol{\phi},\boldsymbol{\mu})$ and used the fact that $\boldsymbol{\phi}\in\Lambda_\epsilon(\boldsymbol{\mu})$, while the second-to-last inequality follows from Hölder's inequality. Since $U_{\sigma\to\boldsymbol{\rho}_\sigma}(\boldsymbol{\phi},\boldsymbol{\mu}')-U(\boldsymbol{\phi},\boldsymbol{\mu}')\leq 2|\mathcal{Z}|\gamma+\frac{\epsilon}{|\Sigma_r|}:=\frac{\epsilon'}{|\Sigma_r|}$ holds for every $(\sigma,\boldsymbol{\rho}_\sigma)$-SPDP, we have that $\boldsymbol{\phi}\in\Lambda_{\epsilon'}(\boldsymbol{\mu}')$ with $\epsilon'=|\mathcal{Z}||\Sigma_r|\gamma+\epsilon$, concluding the proof. □

Lemma 8. *Given any $\delta\in(0,1)$, Algorithm 1 guarantees that $\mathbb{P}[\mathcal{E}]\geq 1-\delta$, where:*

$$\mathcal{E}:=\left\{\|\widehat{\boldsymbol{\mu}}_t-\boldsymbol{\mu}^\star\|_\infty\leq\epsilon_t\quad\forall t\in[T]\right\},$$

and ϵ_t is chosen according to Algorithm 1.

Proof. Let $\mathcal{B}_t(\delta)$ be defined as follows:

$$\mathcal{B}_t(\delta):=\left\{\boldsymbol{\mu}\,\middle|\,|\boldsymbol{\mu}[\sigma]-\widehat{\boldsymbol{\mu}}_t[\sigma]|\leq\sqrt{\frac{\log(2T|\Sigma_c|/\delta)}{2t}}\ \ \forall\sigma\in\Sigma_c\right\}.$$

Clearly, $\mathbb{P}[\mathcal{E}]=\mathbb{P}[\boldsymbol{\mu}^\star\in\mathcal{B}_t(\delta)\,\forall t\in[T]]$. By Hoeffding's inequality, we have that:

$$\mathbb{P}\left(|\boldsymbol{\mu}^\star[\sigma]-\widehat{\boldsymbol{\mu}}_t[\sigma]|\leq\sqrt{\frac{\log(2T|\Sigma_c|/\delta)}{2t}}\right)\geq 1-\frac{\delta}{T|\Sigma_c|}.$$

By a union bound over $\sigma\in\Sigma_c$ and $t\in[T]$, we get that:

$$\mathbb{P}\left(|\boldsymbol{\mu}^\star[\sigma]-\widehat{\boldsymbol{\mu}}_t[\sigma]|\leq\sqrt{\frac{\log(2T|\Sigma_c|/\delta)}{2t}}\ \ \forall\sigma\in\Sigma_c\ \forall t\in[T]\right)\geq 1-\delta.$$

This concludes the proof of the lemma. □

Lemma 9. *If the event $\mathcal{E}$ holds, Algorithm 1 guarantees that $\boldsymbol{\phi}^\star\in\Lambda_{\beta_t}(\widehat{\boldsymbol{\mu}}_t)$ for all $t\in[T]$.*

Proof. By definition, we have that $\boldsymbol{\phi}^\star \in \Lambda(\boldsymbol{\mu}^\star)$. Moreover, since we conditioned on $\mathcal{E}$, we have that:

$$\|\boldsymbol{\mu}^\star - \widehat{\boldsymbol{\mu}}_t\|_\infty \leq \epsilon_t \ \ \forall t \in [T].$$

Thus, we can exploit Lemma 7, which, by letting $\beta_t := 2|\mathcal{Z}||\Sigma_r|\epsilon_t$, gives that $\boldsymbol{\phi}^\star \in \Lambda_{\beta_t}(\widehat{\boldsymbol{\mu}}_t)$. □

Lemma 10. *If the event $\mathcal{E}$ holds, Algorithm 1 guarantees that $\boldsymbol{\phi}_t \in \Lambda_{2\beta_t}(\boldsymbol{\mu}^\star)$ for all $t \in [T]$.*

Proof. Given how Algorithm 1 works, we have that $\boldsymbol{\phi}_t \in \Lambda_{\beta_t}(\widehat{\boldsymbol{\mu}}_t)$. On the other hand, since we conditioned on the event $\mathcal{E}$, it must be the case that $\|\boldsymbol{\mu}^\star - \widehat{\boldsymbol{\mu}}_t\| \leq \epsilon_t$ for all $t \in [T]$. Thus, by Lemma 7 we obtain that $\boldsymbol{\phi}_t \in \Lambda_{2\beta_t}(\boldsymbol{\mu}^\star)$, where β_t is defined as in the proof of Lemma 9. □

Theorem 5. *Given any $\delta \in (0,1)$, with probability at least $1-\delta$, Algorithm 1 guarantees:*

$$R_T = \mathcal{O}\left(|\mathcal{Z}|\sqrt{T\log\left(T|\Sigma_c|/\delta\right)}\right), \quad V_T = \mathcal{O}\left(|\Sigma_r||\mathcal{Z}|\sqrt{T\log\left(T|\Sigma_c|/\delta\right)}\right).$$

Moreover, the algorithm runs in polynomial time.

Proof. First, we bound the computational complexity of the algorithm, then we separately analyze the sender's regret R_T and the receiver's regret V_T.

Complexity. With an argument analogous to the one used for the proof of Theorem 2, we have that the optimization problem solved by SELECTSTRATEGY() in Algorithm 1 is a polynomially-sized linear problem (Lemma 1). Hence, it can be solved in polynomial time.

Sender's regret. If the event $\mathcal{E}$ holds, which happens with probability at least $1-\delta$, then:

$$\boldsymbol{\mu}^\star[\sigma] - \epsilon_t \leq \widehat{\boldsymbol{\mu}}_t[\sigma] \leq \boldsymbol{\mu}^\star[\sigma] + \epsilon_t,$$

for every sequence $\sigma \in \Sigma_c$ and round $t \in [T]$. This implies that, for every $\boldsymbol{\phi} \in \Phi$, we have:

$$F(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) - |\mathcal{Z}|\epsilon_t \leq F(\boldsymbol{\phi}, \widehat{\boldsymbol{\mu}}_t) \leq F(\boldsymbol{\phi}, \boldsymbol{\mu}^\star) + |\mathcal{Z}|\epsilon_t.$$

Moreover, under the event $\mathcal{E}$, we have that $\boldsymbol{\phi}^\star \in \Lambda_{\beta_t}(\widehat{\boldsymbol{\mu}}_t)$ and, thus, $F(\boldsymbol{\phi}^\star, \widehat{\boldsymbol{\mu}}_t) \leq F(\boldsymbol{\phi}_t, \widehat{\boldsymbol{\mu}}_t)$ as $\boldsymbol{\phi}_t$ is computed by optimizing $F(\cdot, \widehat{\boldsymbol{\mu}}_t)$ over $\Lambda_{\beta_t}(\widehat{\boldsymbol{\mu}}_t)$. By putting all the above results together, we get that, under event $\mathcal{E}$, the following holds:

$$F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}^\star) \leq F(\boldsymbol{\phi}^\star, \widehat{\boldsymbol{\mu}}_t) + |\mathcal{Z}|\epsilon_t \leq F(\boldsymbol{\phi}_t, \widehat{\boldsymbol{\mu}}_t) + |\mathcal{Z}|\epsilon_t \leq F(\boldsymbol{\phi}_t, \boldsymbol{\mu}^\star) + 2|\mathcal{Z}|\epsilon_t.$$

By rearranging the terms, taking the sum over $t \in [T]$, and using $\sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$, we get:

$$R_T := \sum_{t=1}^T \left(F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}^\star) - F(\boldsymbol{\phi}_t, \boldsymbol{\mu}^\star)\right) \leq 2|\mathcal{Z}| \sum_{t=1}^T \epsilon_t \leq 2|\mathcal{Z}|\sqrt{2\log(2T|\Sigma_c|/\delta)T},$$

which holds under the event $\mathcal{E}$, and, thus, with probability at least $1-\delta$.

Receiver's regret. If the event $\mathcal{E}$ holds, thanks to Lemma 10 we have that $\boldsymbol{\phi}_t \in \Lambda_{2\beta_t}(\boldsymbol{\mu}^\star)$. Thus, by using Lemma 2, we can conclude that $\boldsymbol{\phi}_t \in \Phi^\diamond_{2\beta_t}(\boldsymbol{\mu}^\star)$. This implies that, with probability at least $1-\delta$, the following holds:

$$V_T \leq 2\sum_{t=1}^T \beta_t \leq 4|\Sigma_r||\mathcal{Z}|\sqrt{2\log(2T|\Sigma_c|/\delta)T},$$

which concludes the proof. □

Appendix E. Proofs omitted from Section 9

Lemma 3. *For every deterministic signaling scheme $\boldsymbol{\pi} \in \Pi$, let*

$$\Sigma_\downarrow(\boldsymbol{\pi}) := \left\{\sigma = (h,a) \in \Sigma_c \mid a \in A(h) \wedge \boldsymbol{\pi}[\sigma_s(h)] = 1\right\}.$$

It holds $\mathbb{E}\left[\boldsymbol{p}_t[\sigma] | \boldsymbol{\pi}_t = \boldsymbol{\pi}\right] = \boldsymbol{\mu}^\star[\sigma]$ for every $\boldsymbol{\pi} \in \Pi$ and $\sigma \in \Sigma_\downarrow(\boldsymbol{\pi})$.

Proof. Let $\boldsymbol{\pi}_t$ be the pure strategy sampled from $\boldsymbol{\phi}_t$. Hence, for any sequence σ we have that the event of reaching node h is a Bernoulli with parameter:

$$\mathbb{E}[\boldsymbol{p}_t[\sigma]|\boldsymbol{\pi}=\boldsymbol{\pi}_t]=\boldsymbol{\pi}_t[\sigma]\boldsymbol{\mu}^\star[\sigma].$$

Thus, by definition, if $\sigma\in\Sigma_\downarrow(\boldsymbol{\pi}_t)$ we have that $\boldsymbol{\pi}_t[\sigma]=1$ and the statement holds. □

Lemma 11. *Given any $\delta\in(0,1)$, Algorithm 2 guarantees that with probability at least $1-\delta/2$:*

$$\sum_{t=N+1}^{T}\sum_{z\in\mathcal{Z}}\epsilon_t[\sigma_c(z)]\boldsymbol{\phi}_t[\sigma_s(z)]\le\sqrt{\log(4T|\Sigma_c|/\delta)|\Sigma_c|T}+|\mathcal{Z}|\sqrt{\log(2/\delta)T},$$

where the terms $\epsilon_t[\sigma]$ for $\sigma\in\Sigma_c$ and $t\in[T]$ are defined according to Algorithm 2.

Proof. First, let us consider the deterministic signaling scheme $\boldsymbol{\pi}_t\in\Pi$ sampled by the algorithm according to $\boldsymbol{\phi}_t$ at round $t\in[T]$. For convenience, in the following we report the definition of $\epsilon_t[\sigma]$ (according to Algorithm 2) for each $\sigma\in\Sigma_c$ and $t\in[T]$:

$$\epsilon_t[\sigma]:=\sqrt{\frac{\log(4T|\Sigma_c|/\delta)}{2C_t[\sigma]}},$$

where $C_t[\sigma]$ represents the number of rounds $t'\le t$ in which it is the case that $\sigma\in\Sigma_\downarrow(\boldsymbol{\pi}_{t'})$. Then, the following chain of inequalities holds:

$$\sum_{t=N+1}^{T}\sum_{z\in\mathcal{Z}}\epsilon_t[\sigma_c(z)]\boldsymbol{\pi}_t[\sigma_s(z)] \tag{E.1a}$$

$$=\sum_{t=N+1}^{T}\sum_{\substack{\sigma\in\Sigma_c:\\ \exists z\in\mathcal{Z}:\sigma=\sigma_c(z)}}\left(\epsilon_t[\sigma]\sum_{\substack{\sigma'\in\Sigma_s:\\ \exists z\in\mathcal{Z}:\sigma=\sigma_c(z)\wedge\sigma'=\sigma_s(z)}}\boldsymbol{\pi}_t[\sigma']\right) \tag{E.1b}$$

$$\le\sum_{t=N+1}^{T}\sum_{\sigma=(h,a)\in\Sigma_c}\epsilon_t[\sigma]\boldsymbol{\pi}_t[\sigma_s(h)] \tag{E.1c}$$

$$=\sum_{\sigma=(h,a)\in\Sigma_c}\sum_{\substack{t\in[T]:\\ t\ge N+1\wedge\boldsymbol{\pi}_t[\sigma_s(h)]=1}}\epsilon_t[\sigma] \tag{E.1d}$$

$$=\sum_{\sigma\in\Sigma_c}\sum_{t=C_{N+1}[\sigma]}^{C_T[\sigma]}\sqrt{\frac{\log(4T|\Sigma_c|/\delta)}{2t}} \tag{E.1e}$$

$$\le\sum_{\sigma\in\Sigma_c}\sqrt{\log(4T|\Sigma_c|/\delta)C_T[\sigma]} \tag{E.1f}$$

$$\le\sqrt{\log(4T|\Sigma_c|/\delta)|\Sigma_c|T}, \tag{E.1g}$$

where Equation (E.1c) follows by the definition of sequence-form signaling scheme of the sender, Equation (E.1d) follows by exchanging the sums over $\sigma\in\Sigma_c$ and $t\in[T]$ and recalling that $\boldsymbol{\pi}_t$ is a deterministic signaling scheme, Equation (E.1e) holds by definition of ϵ, while Equation (E.1f) comes from $\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\le 2\sqrt{T}$. Finally, Equation (E.1g) follows from the Cauchy-Schwarz inequality.

Next, we provide a similar bound on $\sum_{t=N+1}^{T}\sum_{z\in\mathcal{Z}}\epsilon_t[\sigma_c(z)]\boldsymbol{\phi}_t[\sigma_s(z)]$. We do this by exploiting the Azuma-Hoeffding inequality [45]. Indeed, we have that $\mathbb{E}[\boldsymbol{\pi}_t[\sigma]|\mathcal{F}_{t-1}]=\boldsymbol{\phi}_t[\sigma]$, where $\mathcal{F}_{t-1}$ is the filtration generated up to time $t-1$ from the interaction between the algorithm and the BPSDM problem. Thus, with probability at least $1-\delta/2$ the following holds:

$$\sum_{t=N+1}^{T}\sum_{z\in\mathcal{Z}}\epsilon_t[\sigma_c(z)]\boldsymbol{\phi}_t[\sigma_s(z)]\le\sum_{t=N+1}^{T}\sum_{z\in\mathcal{Z}}\epsilon_t[\sigma_c(z)]\boldsymbol{\pi}_t[\sigma_c(z)]+|\mathcal{Z}|\sqrt{\log(2/\delta)T}.$$

By combining the equation above with Equation (E.1f), we obtain:

$$\sum_{t=N+1}^{T}\sum_{z\in\mathcal{Z}}\epsilon_t[\sigma_c(z)]\boldsymbol{\phi}_t[\sigma_s(z)]\le\sqrt{\log(4T|\Sigma_c|/\delta)|\Sigma_c|T}+|\mathcal{Z}|\sqrt{\log(2/\delta)T}.$$

This concludes the proof. □

Lemma 4. *Under the event $\tilde{\mathcal{E}}$, Algorithm 2 guarantees that $\boldsymbol{\phi}_t\in\Lambda_{2\beta_N}(\boldsymbol{\mu}^\star)$ at each round $t>N$.*

Proof. The proof is similar to the one of Lemma 10. If the event $\tilde{\mathcal{E}}$ holds, then we have that:

$$\|\boldsymbol{\mu}^\star - \widehat{\boldsymbol{\mu}}_N\|_\infty \leq \max_{\sigma \in \Sigma_c} \epsilon_t[\sigma] := \epsilon_N.$$

Moreover, $\boldsymbol{\phi}_t \in \Lambda_{\beta_N}(\widehat{\boldsymbol{\mu}}_N)$ and we can use Lemma 7 to conclude that $\boldsymbol{\phi}_t \in \Lambda_{\beta_N + 2\epsilon_N |\Sigma_r||\mathcal{Z}|}(\boldsymbol{\mu}^\star)$ for all $t > N$. The proof follows from $\beta_N \geq 2\epsilon_N |\mathcal{Z}||\Sigma_r|$, since $\epsilon_N \leq \sqrt{\frac{\log(4T|\Sigma_c|/\delta)|\Sigma_c|}{2N}}$. □

Lemma 5. *If the event $\tilde{\mathcal{E}}$ holds, then, for every round $t > N$, it holds that $\boldsymbol{\phi}^\star \in \Lambda_{\beta_N}(\widehat{\boldsymbol{\mu}}_t)$ and $\max_{\boldsymbol{\mu} \in C_t(\delta)} F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}) \geq F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}^\star)$.*

Proof. Since $\boldsymbol{\phi}^\star \in \Lambda(\boldsymbol{\mu}^\star)$ and, under the event $\tilde{\mathcal{E}}$, it holds that:

$$\|\boldsymbol{\mu}^\star - \widehat{\boldsymbol{\mu}}_N\|_\infty \leq \max_{\sigma \in \Sigma_c} \epsilon_t[\sigma] := \epsilon_N,$$

we can use Lemma 7 to conclude that $\boldsymbol{\phi}^\star \in \Lambda_{2|\Sigma_c||\mathcal{Z}|\epsilon_N}(\widehat{\boldsymbol{\mu}}_N)$. The proof of the first statement is concluded by observing that $\beta_N \geq 2|\Sigma_r||\mathcal{Z}|\epsilon_N$, since $\epsilon_N \leq \sqrt{\frac{\log(4T|\Sigma_c|/\delta)|\Sigma_c|}{2N}}$. The second statement directly follows from the observation that, under the event $\tilde{\mathcal{E}}$, it holds $\boldsymbol{\mu}^\star \in C_t(\delta)$. □

Theorem 6. *Given any $\delta \in (0,1)$ and $N \in [T]$, Algorithm 2 guarantees:*

$$R_T = \mathcal{O}\left(N + \sqrt{\log\left(\frac{T|\Sigma_c|}{\delta}\right)|\Sigma_c|T}\right), V_T = \mathcal{O}\left(N + T|\mathcal{Z}||\Sigma_r|\sqrt{\log\left(\frac{T|\Sigma_c|}{\delta}\right)\frac{|\Sigma_c|}{N}}\right),$$

with probability at least $1-\delta$. Moreover, the algorithm runs in polynomial time.

Proof. First, we bound the computational complexity of the algorithm, then we separately analyze the sender's regret R_T and the receiver's regret V_T.

Complexity. First, observe that $F(\boldsymbol{\phi}, \boldsymbol{\mu})$ is a linear function in $\boldsymbol{\mu}$ and it only has positive terms. Thus, for every $\boldsymbol{\phi} \in \Phi$, the maximum over $C_t(\delta)$ in the optimization problem solved during the second phase of the SELECTSTRATEGY() procedure is reached on the boundary of $C_t(\delta)$, so that larger entries of $\boldsymbol{\mu}$ provide larger objective values. Formally, we define:

$$\boldsymbol{\mu}_t \in \arg\max_{\boldsymbol{\mu} \in C_t(\delta)} F(\boldsymbol{\phi}, \boldsymbol{\mu}),$$

which is independent of $\boldsymbol{\phi}$. Then, for every $\sigma \in \Sigma_c$, we have that $\boldsymbol{\mu}_t[\sigma] = \widehat{\boldsymbol{\mu}}_t[\sigma] + \epsilon_t[\sigma]$. Thus, we can compute the signaling scheme $\boldsymbol{\phi}_t$ with a linear program as follows:

$$\boldsymbol{\phi}_t \leftarrow \max_{\boldsymbol{\phi} \in \Lambda_{\beta_t}(\widehat{\boldsymbol{\mu}}_t)} F(\boldsymbol{\phi}, \boldsymbol{\mu}_t), \tag{E.2}$$

and, similarly to the proof of Theorem 5, we have that the optimization problem in Equation (E.2) is a polynomially-sized linear program by Lemma 1. Hence, it can be solved in polynomial time.

Sender's regret. Under the event $\tilde{\mathcal{E}}$, which happens with probability at least $1 - \delta/2$, we have that $|\boldsymbol{\mu}^\star[\sigma] - \widehat{\boldsymbol{\mu}}_t[\sigma]| \leq \epsilon_t[\sigma]$ for all $t > N$. Thus,

$$\|\boldsymbol{\mu}^\star - \widehat{\boldsymbol{\mu}}_t\|_\infty \leq \max_{\sigma \in \Sigma_c} \epsilon_t[\sigma] := \epsilon_N. \tag{E.3}$$

Then, we can conclude that, under event $\tilde{\mathcal{E}}$, it holds $\boldsymbol{\mu}^\star[\sigma] + 2\epsilon_t[\sigma] \geq \boldsymbol{\mu}_t[\sigma]$. This in turn implies:

$$F(\boldsymbol{\phi}_t, \boldsymbol{\mu}_t) \leq F(\boldsymbol{\phi}_t, \boldsymbol{\mu}^\star) + 2\sum_{z \in \mathcal{Z}} \epsilon_t[\sigma_c(z)]\boldsymbol{\phi}_t[\sigma_s(z)].$$

By Lemma 5, we have that, under event $\tilde{\mathcal{E}}$, it holds $\boldsymbol{\phi}^\star \in \Lambda_{\beta_N}(\widehat{\boldsymbol{\mu}}_N)$. Hence, $F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}_t) \leq F(\boldsymbol{\phi}_t, \boldsymbol{\mu}_t)$ as $\boldsymbol{\phi}_t$ is computed as the optimum over $\Lambda_{\beta_N}(\boldsymbol{\mu}_t)$. Moreover, by Lemma 5 we also have that $F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}^\star) \leq F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}_t)$, which implies:

$$F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}^\star) \leq F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}_t) \leq F(\boldsymbol{\phi}_t, \boldsymbol{\mu}_t) \leq F(\boldsymbol{\phi}_t, \boldsymbol{\mu}^\star) + 2\sum_{z \in \mathcal{Z}} \epsilon_t[\sigma_c(z)]\boldsymbol{\phi}_t[\sigma_s(z)].$$

Then, we can decompose the sender's regret as:

$$R_T = \sum_{t=1}^{N}\Big(F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}^\star) - F(\boldsymbol{\phi}_t, \boldsymbol{\mu}^\star)\Big) + \sum_{t=N+1}^{T}\Big(F(\boldsymbol{\phi}^\star, \boldsymbol{\mu}^\star) - F(\boldsymbol{\phi}_t, \boldsymbol{\mu}^\star)\Big)$$

$$\leq N + 2 \sum_{t=N+1}^{T} \sum_{z \in \mathcal{Z}} \epsilon_t[\sigma_c(z)] \boldsymbol{\phi}_t[\sigma_s(z)].$$

By using Lemma 11 and a union bound, we can conclude that with probability at least $1-\delta$:

$$R_T \leq N + 2\left(\sqrt{\log(4T|\Sigma_c|/\delta)|\Sigma_c|T} + |\mathcal{Z}|\sqrt{\log(2/\delta)T}\right).$$

Receiver's regret. By Lemma 4, under the event $\tilde{\mathcal{E}}$, we have that $\boldsymbol{\phi}_t \in \Lambda_{2\beta_N}(\boldsymbol{\mu}^\star)$ for all $t \geq N$. Moreover, by Lemma 2, it holds that $\Lambda_{2\beta_N}(\boldsymbol{\mu}^\star) \subseteq \Phi^\circ_{2\beta_N}(\boldsymbol{\mu}^\star)$. Hence, with probability at least $1-\delta$:

$$V_T \leq N + 2T\beta_N = N + 4T|\mathcal{Z}||\Sigma_r|\sqrt{\frac{|\Sigma_c|\log(4T|\Sigma_c|/\delta)}{2N}}.$$

This concludes the proof. □

Theorem 7. *For any $\alpha \in [1/2, 1]$, there exists a constant $\gamma \in (0,1)$ such that no algorithm guarantees both $R_T = o(T^\alpha)$ and $V_T = o(T^{1-\alpha/2})$ with probability greater than γ.*

Proof. We define two instances i and j of a BPSDM problem whose tree structures are as in Fig. C.5. In both instances, we have that $f(z_1) = f(z_2) = 1$ and $f(z_3) = f(z_4) = 0$ for the sender, while $u(z_1) = u(z_3) = 1$ and $u(z_2) = u(z_4) = 0$ for the receiver. Moreover, in both instances we have that for the chance node h_1 it holds $\boldsymbol{\mu}^\star[(h_1, c)] = \boldsymbol{\mu}^\star[(h_1, d)] = 1/2$. Instead, the two instances differ in the probabilities of chance node h_2, which are defined as follows:

$$\mathrm{i} := \begin{cases} \boldsymbol{\mu}^\star[(h_2, e)] = \frac{1}{2} - \epsilon \\ \boldsymbol{\mu}^\star[(h_2, f)] = \frac{1}{2} + \epsilon \end{cases},$$

$$\mathrm{j} := \begin{cases} \boldsymbol{\mu}^\star[(h_2, e)] = \frac{1}{2} + \epsilon \\ \boldsymbol{\mu}^\star[(h_2, f)] = \frac{1}{2} - \epsilon \end{cases}.$$

Simple calculations show that, in instance j, we have that the regret of the sender is:

$$R_T^{\mathrm{j}} = \sum_{t=1}^{T} \boldsymbol{\phi}_t[(h_0, b)]$$

Hence, if we require that (in high probability with respect to the measure $\mathbb{P}^{\mathrm{j}}$ of instance j) the sender's regret is smaller than a threshold K, then:

$$\mathbb{P}^{\mathrm{j}}\left[\sum_{t=1}^{T} \boldsymbol{\phi}_t[(h_0, b)] \leq K\right] \geq 1 - \delta.$$

The Pinsker's inequality states that:

$$\mathbb{P}^{\mathrm{i}}\left[\sum_{t=1}^{T} \boldsymbol{\phi}_t[(h_0, b)] \leq K\right] \geq 1 - \delta - \sqrt{\frac{1}{2}\mathcal{K}(\mathrm{j}, \mathrm{i})},$$

where $\mathcal{K}(\mathrm{j}, \mathrm{i})$ is the Kullback-Leibler divergence between instance j and instance i. By the well-known decomposition theorem of the divergence, we know that:

$$\mathcal{K}(\mathrm{j}, \mathrm{i}) = \mathbb{E}^{\mathrm{j}}\left[\sum_{t=1}^{T} \boldsymbol{\phi}_t[(h_0, b)]\right] \mathcal{K}(B_{1/2+\epsilon}, B_{1/2-\epsilon}) \leq 16\epsilon^2 \mathbb{E}^{\mathrm{j}}\left[\sum_{t=1}^{T} \boldsymbol{\phi}_t[(h_0, b)]\right],$$

where $\mathcal{K}(B_{1/2+\epsilon}, B_{1/2-\epsilon})$ is the Kullback-Leibler divergence between two Bernoulli random variable with parameter $1/2+\epsilon$ and $1/2-\epsilon$. Now, we can upper bound $\mathbb{E}^{\mathrm{j}}\left[\sum_{t=1}^{T} \boldsymbol{\phi}_t[(h_0, b)]\right]$ in terms of the probability $\mathbb{P}^{\mathrm{j}}$ with the reverse Markov inequality, as follows:

$$\begin{aligned} \mathbb{E}^{\mathrm{j}}\left[\sum_{t=1}^{T} \boldsymbol{\phi}_t[(h_0, b)]\right] &\leq \mathbb{P}^{\mathrm{j}}\left[\sum_{t=1}^{T} \boldsymbol{\phi}_t[(h_0, b)] \geq K\right](T-K) + K \\ &\leq \delta(T-K) + K. \end{aligned}$$

Thus, we can conclude that:

$$\mathbb{P}^{\mathrm{i}}\left[\sum_{t=1}^{T} \boldsymbol{\phi}_t[(h_0, b)] \leq K\right] \geq 1 - \delta - 2\epsilon\sqrt{2(\delta(T-K)+K)}. \tag{E.4}$$

Now, we consider the receiver's regret in instance i, which can be computed as:

$$V_T^{\mathrm{i}} = \epsilon \sum_{t=1}^{T} \boldsymbol{\phi}_t[(h_0, b)].$$

This, together with Equation (E.4), allows us to conclude that:

$$\mathbb{P}^{\mathrm{i}}\left[V_T^{\mathrm{i}} \geq \epsilon(T-K)\right] \geq 1 - \delta - 2\epsilon\sqrt{2(\delta(T-K)+K)}.$$

By setting $K = \frac{T^\alpha}{8}$ and $\epsilon = \frac{T^{-\alpha/2}}{8}$, we can conclude that if

$$\mathbb{P}^{\mathrm{j}}\left[R_T^{\mathrm{j}} \leq \frac{T^\alpha}{8}\right] \geq \mathbb{P}^{\mathrm{j}}\left[\sum_{t=1}^{T} \boldsymbol{\phi}_t[(h_0, a)] \leq \frac{T^\alpha}{8}\right] \geq 1 - \delta,$$

then

$$\mathbb{P}^{\mathrm{i}}\left[V_T^{\mathrm{i}} \geq \frac{T^{1-\alpha/2}}{16}\right] \geq 1 - \frac{\sqrt{2}}{16} - \delta \geq 0.91 - \delta,$$

where we used that $\frac{T^{1-\alpha/2}}{8} - \frac{T^{\alpha/2}}{64} \geq \frac{T^{1-\alpha/2}}{16}$ for $T \geq 1$ and that we can assume $\delta \leq \frac{T^{\alpha-1}}{4}$. □

Data availability

No data was used for the research described in the article.

References

[1] M. Bernasconi, M. Castiglioni, A. Marchesi, N. Gatti, F. Trovò, Sequential information design: learning to persuade in the dark, in: NeurIPS, 2022.
[2] E. Kamenica, M. Gentzkow, Bayesian persuasion, Am. Econ. Rev. 101 (6) (2011) 2590–2615.
[3] P. Bro Miltersen, O. Sheffet, Send mixed signals: earn more, work less, in: EC, 2012, pp. 234–247.
[4] Feldman M. EmekY, I. Gamzu, R. PaesLeme, M. Tennenholtz, Signaling schemes for revenue maximization, TEAC 2 (2) (2014) 1–19.
[5] A. Badanidiyuru, K. Bhawalkar, H. Xu, Targeting and signaling in ad auctions, in: SODA, 2018, pp. 2545–2563.
[6] M. Castiglioni, G. Romano, A. Marchesi, N. Gatti, Signaling in posted price auctions, in: AAAI, 2022.
[7] R. Alonso, O. Câmara, Persuading voters, Am. Econ. Rev. 106 (11) (2016) 3590–3605.
[8] Y. Cheng, H.Y. Cheung, S. Dughmi, E. Emamjomeh-Zadeh, L. Han, S. Teng, Mixture selection, mechanism design, and signaling, in: IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17–20 October, 2015, 2015, pp. 1426–1445.
[9] M. Castiglioni, Gatti N. CelliA, Persuading voters: it's easy to whisper, it's hard to speak loud, in: AAAI, 2020, pp. 1870–1877.
[10] M. Castiglioni, N. Gatti, Persuading voters in district-based elections, in: AAAI, 2021, pp. 5244–5251.
[11] U. Bhaskar, Ko Y.K. ChengY, C. Swamy, Hardness results for signaling in Bayesian zero-sum and network routing games, in: EC, 2016, pp. 479–496.
[12] M. Castiglioni, A. Celli, A. Marchesi, N. Gatti, Signaling in Bayesian network congestion games: the subtle power of symmetry, in: AAAI, 2021, pp. 5252–5259.
[13] Y. Mansour, A. Slivkins, V. Syrgkanis, Z.S. Wu, Bayesian exploration: incentivizing exploration in Bayesian games, in: EC, 2016, p. 661.
[14] Z. Rabinovich, A.X. Jiang, M. Jain, H. Xu, Information disclosure as a means to security, in: AAMAS, 2015, pp. 645–653.
[15] H. Xu, R. Freeman, V. Conitzer, S. Dughmi, M. Tambe, Signaling in Bayesian Stackelberg games, in: AAMAS, 2016, pp. 150–158.
[16] Y. Babichenko, S. Barman, Algorithmic aspects of private Bayesian persuasion, in: ITCS, 2017.
[17] O. Candogan, Persuasion in networks: public signals and k-cores, in: EC, 2019, pp. 133–134.
[18] M. Castiglioni, A. Celli, A. Marchesi, N. Gatti, Online Bayesian persuasion, in: NeurIPS, 2020, pp. 16188–16198.
[19] M. Castiglioni, A. Marchesi, A. Celli, N. Gatti, Multi-receiver online Bayesian persuasion, in: ICML, 2021, pp. 1314–1323.
[20] Y. Zu, K. Iyer, H. Xu, Learning to persuade on the fly: robustness against ignorance, in: EC, 2021, pp. 927–928.
[21] Y. Babichenko, I. Talgam-Cohen, H. Xu, K. Zabarnyi, Regret-minimizing Bayesian persuasion, in: EC, 2021, p. 128.
[22] M. Castiglioni, A. Marchesi, N. Gatti, Bayesian persuasion meets mechanism design: going beyond intractability with type reporting, in: AAMAS, 2022.
[23] M. Bernasconi, M. Castiglioni, A. Celli, A. Marchesi, F. Trovò, N. Gatti, Optimal rates and efficient algorithms for online Bayesian persuasion, in: International Conference on Machine Learning, in: PMLR, 2023, pp. 2164–2183.
[24] Y. Zu, K. Iyer, H. Xu, Learning to persuade on the fly: robustness against ignorance, in: EC, 2021, pp. 927–928.
[25] M.K. Camara, J.D. Hartline, A. Johnsen, Mechanisms for a no-regret agent: beyond the common prior, in: 2020 Ieee 61st Annual Symposium on Foundations of Computer Science (Focs), IEEE, 2020, pp. 259–270.
[26] J. Gan, R. Majumdar, G. Radanovic, A. Singla, Bayesian persuasion in sequential decision-making, in: AAAI, 2022.
[27] J. Wu, Z. Zhang, Z. Feng, Z. Wang, Z. Yang, M.I. Jordan, et al., Sequential information design: Markov persuasion process and its efficient reinforcement learning, arXiv preprint arXiv:2202.10678, 2022.
[28] M. Bernasconi, M. Castiglioni, A. Marchesi, M. Mutti, Persuading farsighted receivers in mdps: the power of honesty, in: NeurIPS, 2023.
[29] A. Celli, S. Coniglio, N. Gatti, Private Bayesian persuasion with sequential games, AAAI 34 (02) (2020) 1886–1893.
[30] S.T. Su, V.G. Subramanian, G. Schoenebeck, Bayesian persuasion in sequential trials, in: Web and Internet Economics: 17th International Conference, WINE 2021, Potsdam, Germany, December 14–17, 2021, Proceedings, Springer, 2022, pp. 22–40.
[31] B. Ni, W. Shen, P. Tang, Sequential persuasion using limited experiments, arXiv preprint arXiv:2303.10619, 2023.
[32] M. Castiglioni, A. Marchesi, A. Celli, N. Gatti, Multi-receiver online Bayesian persuasion, in: ICML, vol. 139, 2021, pp. 1314–1323.
[33] M. Castiglioni, A. Marchesi, N. Gatti, Bayesian persuasion meets mechanism design: going beyond intractability with type reporting, in: AAMAS 2022. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2022, pp. 226–234.
[34] M. Bernasconi-de-Luca, F. Cacciamani, S. Fioravanti, N. Gatti, A. Marchesi, F. Trovò, Exploiting opponents under utility constraints in sequential games, in: NeurIPS 2021, 2021, pp. 13177–13188.
[35] M. Bernasconi, F. Cacciamani, M. Castiglioni, A. Marchesi, N. Gatti, F. Trovò, Safe learning in tree-form sequential decision making: handling hard and soft constraints, in: ICML 2022, vol. 162, in: PMLR, 2022, pp. 1854–1873.

[36] Babichenko Y. ArieliI, Private Bayesian persuasion, J. Econ. Theory 182 (2019) 185–217.
[37] B. Von Stengel, F. Forges, Extensive-form correlated equilibrium: definition and computational complexity, Math. Oper. Res. 33 (4) (2008) 1002–1022.
[38] D. Morrill, R. D'Orazio, R. Sarfati, M. Lanctot, J.R. Wright, A.R. Greenwald, et al., Hindsight and sequential rationality of correlated play, in: AAAI, vol. 35, 2021, pp. 5584–5594.
[39] D. Koller, N. Megiddo, B. Von Stengel, Efficient computation of equilibria for extensive two-person games, Games Econ. Behav. 14 (2) (1996) 247–259.
[40] G. Farina, A. Celli, A. Marchesi, N. Gatti, Simple uncoupled no-regret learning dynamics for extensive-form correlated equilibrium, arXiv preprint arXiv:2104.01520, 2021.
[41] G. Farina, R. Schmucker, T. Sandholm, Bandit linear optimization for sequential decision making and extensive-form games, in: AAAI, vol. 35, 2021, pp. 5372–5380.
[42] C.J. Hillar, L.H. Lim, Most tensor problems are np-hard, J. ACM 60 (6) (2013) 1–39.
[43] A. Celli, A. Marchesi, G. Farina, N. Gatti, No-regret learning dynamics for extensive-form correlated equilibrium, in: NeurIPS, vol. 33, 2020, pp. 7722–7732.
[44] T. Lattimore, C. Szepesvári, Bandit Algorithms, Cambridge University Press, 2020.
[45] N. Cesa-Bianchi, G. Lugosi, Prediction, Learning, and Games, Cambridge University Press, 2006.