Online Bayesian Persuasion

Matteo Castiglioni
Politecnico di Milano
matteo.castiglioni@polimi.it

Andrea Celli*
Facebook Core Data Science
andreacelli@fb.com

Alberto Marchesi
Politecnico di Milano
alberto.marchesi@polimi.it

Nicola Gatti
Politecnico di Milano
nicola.gatti@polimi.it

Abstract

In Bayesian persuasion, an informed sender has to design a signaling scheme that discloses the right amount of information so as to influence the behavior of a self-interested receiver. This kind of strategic interaction is ubiquitous in real-world economic scenarios. However, the seminal model by Kamenica and Gentzkow makes some stringent assumptions that limit its applicability in practice. One of the most limiting assumptions is, arguably, that the sender is required to know the receiver's utility function to compute an optimal signaling scheme. We relax this assumption through an *online learning* framework in which the sender repeatedly faces a receiver whose type is unknown and chosen adversarially at each round from a finite set of possible types. We are interested in *no-regret* algorithms prescribing a signaling scheme at each round of the repeated interaction with performances close to that of a best-in-hindsight signaling scheme. First, we prove a hardness result on the per-round running time required to achieve no-α-regret for any $\alpha < 1$. Then, we provide algorithms for the *full* and *partial feedback* models with regret bounds sublinear in the number of rounds and polynomial in the size of the instance.

1 Introduction

Bayesian persuasion was first introduced by Kamenica and Gentzkow [23] as the problem faced by an informed *sender* trying to influence the behavior of a self-interested *receiver* via the strategic provision of payoff-relevant information. In Bayesian persuasion, the agents' beliefs are influenced only by controlling 'who gets to know what'. This 'sweet talk' is ubiquitous among all sorts of economic activities, and it was famously attributed to a quarter of the GDP in the United States by McCloskey and Klamer [28]. [2] The computational study of Bayesian persuasion has been largely driven by its application in domains such as auctions and online advertisement [7, 19, 11], voting [1, 14, 16], traffic routing [9, 32], recommendation systems [26], security [30, 34], and product marketing [6, 13].

In the model by Kamenica and Gentzkow [23], the sender's and receiver's payoffs are determined by the receiver's action and a set of parameters collectively termed the *state of nature*. Unlike the receiver, the sender observes the realized state of nature drawn from a shared prior distribution. The sender uses this private information to determine a signal for the receiver according to a publicly known *signaling scheme*, *i.e.*, a mapping from states of nature to probability distributions over signals.

In this paper, we focus on arguably one of the most severe limitations of the basic model: the sender must know exactly the receiver's utility function to compute an optimal signaling scheme.

*The work was conducted while the author was a postdoc at Politecnico di Milano.

[2] A more recent estimate by Antioch and others [2] places this figure at 30%.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Our model and results We deal with uncertainty about the receiver's type by framing the Bayesian persuasion problem in an online learning framework. We study a repeated Bayesian persuasion problem where, at each round, the receiver's type is adversarially chosen from a finite set of types. Our goal is the design of an online algorithm that recommends a signaling scheme at each round, guaranteeing an expected utility for the sender close to that of the best-in-hindsight signaling scheme. We study this problem under two models of feedback: in the *full information* model, the sender selects a signaling scheme and later observes the type of the best-responding receiver; in the *partial information* model, the sender only observes the actions taken by the receiver.

First, in Section 4, we provide a negative result that rules out, even in the full information setting, the possibility of designing a no-regret algorithm with polynomial per-round running time. Furthermore, the same hardness result holds when adopting the notion of no-α-regret (in the additive sense) for any $\alpha < 1$. Then, we focus on the problem of designing no-regret algorithms by relaxing the running time constraint. We show that it is possible to achieve a regret polynomial in the size of the problem instance and sublinear in the number of rounds T under both full (with $O(T^{-1/2})$) and partial feedback (with $O(T^{-1/5})$). We present these results in Sections 5 and 6, respectively.

Related works The closest line of research to ours is the one studying online learning problems in *Stackelberg games*. In these games, a *leader* commits to a probability distribution over a set of actions, and a *follower* plays an action maximizing her/his utility given the leader's commitment [33]. In this setting, Letchford *et al.* [25] and Blum *et al.* [10] study the problem of computing the best leader's strategy against an unknown follower using a polynomial number of best-response queries. Marecki *et al.* [27] study the problem with a single follower with type drawn from a Bayesian prior.

Balcan *et al.* [8] study how to minimize the leader's regret in an online setting in which the follower's type is unknown and chosen adversarially from a finite set. Although the problem is conceptually similar to ours, the Bayesian persuasion framework presents a number of additional challenges: the solution to a Stackelberg game consists of a point in a finite-dimensional simplex, while the solution to a Bayesian persuasion problem is a probability distribution with potentially infinite support size. This probability distribution is subject to additional consistency constraints, which (under partial feedback) rule out the possibility of exploiting unbiased estimators of the sender's expected utility.

Finally, it is worth mentioning that known online learning algorithms (for either the full or partial feedback setting) do not provide any guarantee in the case of Bayesian persuasion. Indeed, the regret bounds of those algorithms depend linearly or sublinearly in the number of actions, but the action space in Bayesian persuasion is infinite. A large body of previous works in other fields resolves the issue of dealing with an infinite action space by requiring specific assumptions (*e.g.*, linear or convex utility function [4, 12, 22, 35]). However, in the online Bayesian persuasion setting, these assumptions do not hold as the sender's utility depends on the receiver's best response, which yields a function that is not linear nor convex (or even continuous in the space of signaling schemes).

2 Preliminaries

The receiver has a finite set of m actions $\mathcal{A} := \{a_i\}_{i=1}^m$ and a set of n possible types $\mathcal{K} := \{k_i\}_{i=1}^n$. For each type $k \in \mathcal{K}$, the receiver's payoff function is $u^k : \mathcal{A} \times \Theta \to [0,1]$, where $\Theta := \{\theta_i\}_{i=1}^d$ is a finite set of d states of nature. For notational convenience, we denote by $u_\theta^k(a) \in [0,1]$ the utility observed by the receiver of type $k \in \mathcal{K}$ when the realized state of nature is $\theta \in \Theta$ and she/he plays action $a \in \mathcal{A}$. The sender's utility when the state of nature is $\theta \in \Theta$ is described by the function $u_\theta^{\mathsf{s}} : \mathcal{A} \to [0,1]$. As it is customary in Bayesian persuasion, we assume that the state of nature is drawn from a common prior distribution $\boldsymbol{\mu} \in \text{int}(\Delta_\Theta)$, which is explicitly known to both the sender and the receiver. [3] Moreover, the sender can commit to a *signaling scheme* ϕ, which is a randomized mapping from states of nature to *signals* for the receiver. Formally $\phi : \Theta \to \Delta_\mathcal{S}$, where $\mathcal{S}$ is a finite set of signals. We denote by ϕ_θ the probability distribution employed by the sender when the state of nature is $\theta \in \Theta$, with $\phi_\theta(s)$ being the probability of sending signal $s \in \mathcal{S}$.

A *one-shot* interaction between the sender and the receiver goes on as follows: (i) the sender commits to a publicly known signaling scheme ϕ and the receiver observes the commitment; (ii) the sender

[3] $\text{int}(X)$ is the *interior* of set X and Δ_X is the set of all probability distributions over X. Vectors are denoted by bold symbols. For any vector $\mathbf{x}$, the value of its i-th component is denoted by x_i.

observes the realized state of nature $\theta \sim \boldsymbol{\mu}$; (iii) the sender draws a signal $s \sim \phi_\theta$ and communicates it to the receiver; (iv) the receiver observes s and rationally updates her/his prior beliefs over Θ according to the *Bayes rule*; (v) the receiver selects an action maximizing her/his expected utility.

Let $\Xi := \Delta_\Theta$ be the set of receiver's posterior beliefs over the states of nature. In step (iv), after observing $s \in \mathcal{S}$, the receiver performs a Bayesian update and infers a posterior belief $\boldsymbol{\xi} \in \Xi$ over the states of nature such that the component of $\boldsymbol{\xi}$ corresponding to state of nature $\theta \in \Theta$ is:

$$\xi_\theta := \frac{\mu_\theta \, \phi_\theta(s)}{\sum_{\theta' \in \Theta} \mu_{\theta'} \, \phi_{\theta'}(s)}. \tag{1}$$

After computing $\boldsymbol{\xi}$, the receiver solves a decision problem to find an action maximizing her/his expected utility given the current posterior. Letting $a \in \mathcal{A}$ be the receiver's choice, the receiver observes payoff $u^k_\theta(a)$, where $k \in \mathcal{K}$ is the receiver's type, while the sender observes payoff $u^{\mathsf{s}}_\theta(a)$.

2.1 Working in the space of posterior distributions

It is oftentimes useful to represent signaling schemes as convex combinations of posterior beliefs they can induce. First, we describe such interpretation (see [24] for further details). Then, we define the receiver's best response given an arbitrary posterior belief.

Representing signaling schemes Given a signaling scheme ϕ, each signal realization $s \in \mathcal{S}$ leads to a posterior belief $\boldsymbol{\xi}^s \in \Xi$, whose components are defined as in Equation (1). Accordingly, each signaling scheme leads to a distribution over posterior beliefs. We denote a distribution over posteriors by $\mathbf{w} \in \Delta_\Xi$. We say that a signaling scheme $\phi : \Theta \to \Delta_\mathcal{S}$ *induces* $\mathbf{w} \in \Delta_\Xi$ if, for every $\boldsymbol{\xi} \in \Xi$, the component of $\mathbf{w}$ corresponding to $\boldsymbol{\xi}$ is defined as follows:

$$w_{\boldsymbol{\xi}} := \sum_{s \in \mathcal{S} : \boldsymbol{\xi}^s = \boldsymbol{\xi}} \sum_{\theta \in \Theta} \mu_\theta \, \phi_\theta(s). \tag{2}$$

Intuitively, if ϕ induces $\mathbf{w}$, then $w_{\boldsymbol{\xi}}$ represents the probability that ϕ induces the posterior $\boldsymbol{\xi} \in \Xi$. We let $\text{supp}(\mathbf{w}) := \{\boldsymbol{\xi} \in \Xi \mid w_{\boldsymbol{\xi}} > 0\}$ be the set of posteriors induced with strictly positive probability. We say that a distribution over posteriors $\mathbf{w} \in \Delta_\Xi$ is *consistent* (*i.e.*, intuitively, there exists a valid signaling scheme ϕ inducing $\mathbf{w}$) if the following hods:

$$\sum_{\boldsymbol{\xi} \in \text{supp}(\mathbf{w})} w_{\boldsymbol{\xi}} \, \xi_\theta = \mu_\theta, \quad \text{for all } \theta \in \Theta. \tag{3}$$

We let $W \subseteq \Delta_\Xi$ be the set of distributions over posteriors that are consistent according to Equation (3). In the remainder of the paper, we equivalently employ ϕ or $\mathbf{w}$ to denote an arbitrary signaling scheme.

Receiver's best-response set After observing a signal $s \in \mathcal{S}$ that induces a posterior $\boldsymbol{\xi} \in \Xi$, the receiver best responds by choosing an action that maximizes her/his expected utility (step (v)). The set of actions maximizing the receiver's expected utility given posterior $\boldsymbol{\xi}$ is defined as follows:

Definition 1 (BR-set)**.** *Given posterior $\boldsymbol{\xi} \in \Xi$ and type $k \in \mathcal{K}$, the* best-response set (BR-set) *is:*

$$\mathcal{B}^k_{\boldsymbol{\xi}} := \arg\max_{a \in \mathcal{A}} \sum_{\theta \in \Theta} \xi_\theta \, u^k_\theta(a).$$

We denote by $b^k_{\boldsymbol{\xi}}$ the action belonging to the BR-set $\mathcal{B}^k_{\boldsymbol{\xi}}$ played by the receiver. When the receiver is indifferent among multiple actions for a given posterior $\boldsymbol{\xi}$, we assume that the receiver breaks ties in favor of the sender, *i.e.*, she/he chooses an action $b^k_{\boldsymbol{\xi}} \in \arg\max_{a \in \mathcal{B}^k_{\boldsymbol{\xi}}} \sum_\theta \xi_\theta \, u^{\mathsf{s}}_\theta(a)$. [4]

We conclude the section by introducing some additional notation. We denote by $u^{\mathsf{s}}(\boldsymbol{\xi}, k) := \sum_\theta \xi_\theta \, u^{\mathsf{s}}_\theta(b^k_{\boldsymbol{\xi}})$ the sender's expected utility when she/he induces a posterior $\boldsymbol{\xi} \in \Xi$ and the receiver is of type $k \in \mathcal{K}$. Moreover, we use $u^{\mathsf{s}}(\phi, k)$ to denote the sender's expected utility achieved with the signaling scheme ϕ. Formally, $u^{\mathsf{s}}(\phi, k) := \sum_{\boldsymbol{\xi} \in \text{supp}(\mathbf{w})} w_{\boldsymbol{\xi}} \, u^{\mathsf{s}}(\boldsymbol{\xi}, k)$, where $\mathbf{w} \in \Delta_\Xi$ is the distribution over posteriors induced by ϕ. Analogously, we write $u^{\mathsf{s}}(\mathbf{w}, k)$.

Finally, letting OPT be the sender's optimal expected utility, we say that a signaling scheme is α*-optimal* (in the additive sense) if it provides the sender with a utility at least as large as $OPT - \alpha$.

[4] This assumption is customary in settings involving commitments, such as Stackelberg games [17, 18, 29].

		State G ($\mu_{\mathrm{G}} = .3$)		State I ($\mu_{\mathrm{I}} = .7$)	
$\mathcal{A}$	A	0	0	0	1
	C	1	1	1	0

		Realized state State G	State I
$\mathcal{S}$	s_1	0	$4/7$
	s_2	1	$3/7$

		State of nature State G	State I	$\mathbf{w}^\star$
supp($\mathbf{w}^\star$)	$\boldsymbol{\xi}_1$	0	1	$2/5$
	$\boldsymbol{\xi}_2$	$1/2$	$1/2$	$3/5$

Figure 1: **Left:** The prosecutor/judge game. Rows represent the judge's actions. For each possible state of nature $\{\mathrm{G}, \mathrm{I}\}$, the first column is the prosecutor's payoff while the second is the judge's payoff. **Center:** The optimal signaling scheme $\phi^\star$. Each column describes the probability with which the two signals are drawn given the realized state of nature. **Right**: Representation of $\phi^\star$ as the convex combination of posteriors $\mathbf{w}^\star$.

2.2 Example

We illustrate the key notion of signaling scheme in a simple example with a single receiver type (*i.e.,* $|\mathcal{K}| = 1$) inspired by Kamenica and Gentzkow [23]: a prosecutor (the sender) is trying to convince a rational judge (the receiver) that a defendant is guilty. The judge has two available actions: to *acquit* or to *convict* the defendant (denoted by A and C, respectively). There are two possible states of nature: the defendant is either *guilty* (denoted by G) or *innocent* (denoted by I). The prosecutor and the judge share a prior belief $\mu_{\mathrm{G}} = .3$. Moreover, the prosecutor gets utility 1 if the judge convicts the defendant and 0 otherwise, regardless of the state of nature. The prosecutor gets to observe the realized state of nature (*i.e.,* whether the defendant is guilty or innocent). The she/he can exploit this information to select a signal from set $\mathcal{S} = \{s_1, s_2\}$ and send it to the judge. The judge has a unique type and she/he gets utility 1 for choosing the just action (convict when guilty and acquit when innocent) and utility 0 for choosing the unjust action (see Figure 1-Left).

Figure 1-Center depicts a sender-optimal signaling scheme $\phi^\star$ obtained via the following LP:

$$\arg\max_{\phi \geq 0} u^{\mathsf{s}}(\phi, k) \quad \text{s.t.} \quad \sum_{s \in \mathcal{S}} \phi_\theta(s) = 1 \quad \forall \theta \in \Theta,$$

where k is the unique type of the judge. When the sender acts according to $\phi^\star$, signal s_1 (resp., s_2) originates posterior $\boldsymbol{\xi}_1$ (resp., $\boldsymbol{\xi}_2$; see Figure 1-Right). Applying Equation (3) yields the equivalent representation of $\phi^\star$ as a convex combination of posteriors, *i.e.*, $w^\star_{\boldsymbol{\xi}_1} = 2/5$ and $w^\star_{\boldsymbol{\xi}_2} = 3/5$.

By unpacking the objective function of the above LP (and dropping the dependency on k) we have: $\mathcal{B}_{\boldsymbol{\xi}_1} = \{\mathrm{A}\}$ and $\mathcal{B}_{\boldsymbol{\xi}_2} = \{\mathrm{A}, \mathrm{C}\}$. Therefore, if the posterior is $\boldsymbol{\xi}_1$, the judge will acquit the defendant, *i.e.,* $b_{\boldsymbol{\xi}_1} = \mathrm{A}$. Otherwise, if the posterior is $\boldsymbol{\xi}_2$, we have $b_{\boldsymbol{\xi}_2} = \mathrm{C}$ since the receiver breaks ties in favor of the sender. This highlights an intuitive interpretation of the signaling problem: the two signals may be interpreted as action recommendations. Signal s_1 (resp., s_2) is interpreted by the judge as a recommendation to play A (resp., C). Then, our definition of best-response set (Definition 1) implies that it is in the receiver's best interest to follow the action recommendations. The best-response conditions can be formulated in terms of linear constraints on ϕ_θ as follows:

$$\sum_{\theta \in \Theta} \mu_\theta \, \phi_\theta(s_1) \Big(u_\theta(\mathrm{A}) - u_\theta(\hat{a}) \Big) \geq 0 \quad \text{and} \quad \sum_{\theta \in \Theta} \mu_\theta \, \phi_\theta(s_2) \Big(u_\theta(\mathrm{C}) - u_\theta(\hat{a}) \Big) \geq 0 \quad \forall \hat{a} \in \{\mathrm{A}, \mathrm{C}\}.$$

3 The online Bayesian persuasion framework

We consider the following online setting. The sender plays a repeated game in which, at each round $t \in [T]$, she/he commits to a signaling scheme ϕ^t, observes a state of nature $\theta^t \sim \boldsymbol{\mu}$, and she/he sends signal $s^t \sim \phi^t_{\theta^t}$ to the receiver. [5] Then, a receiver of unknown type updates her/his prior distribution and selects an action a^t maximizing her/his expected reward (in the *one-shot* interaction at round t). We focus on the problem in which the sequence of receiver's types $\mathbf{k} := \{k^t\}_{t \in [T]}$ is selected beforehand by an adversary. After the receiver plays a^t, the sender receives a *feedback* on her/his choice at round t. In the *full information* feedback setting, the sender observes the receiver's type k^t. Therefore, the sender can compute the expected payoff for any signaling scheme she/he could have chosen other than ϕ^t. Instead, in the *partial information* feedback setting, the sender only observes the action a^t played by the receiver at round t.

[5] Throughout the paper, the set $\{1, \ldots, x\}$ is denoted by $[x]$.

We are interested in algorithms computing ϕ^t at each round t. The performance of one such algorithm is measured using the average per-round *regret* computed with respect to the best signaling scheme in hindsight. Formally:

$$R^T := \max_{\phi} \left\{ \frac{1}{T} \sum_{t=1}^{T} \left(u^{\mathsf{s}}(\phi, k^t) - \mathbb{E}\left[u^{\mathsf{s}}(\phi^t, k^t) \right] \right) \right\},$$

where the expectation is on the randomness of the online algorithm (*i.e.*, the probability distribution which is used by the sender to draw the signaling scheme at round t) and T is the number of rounds. Ideally, we would like to find an algorithm that generates a sequence $\{\phi^t\}_{t \in [T]}$ with the following properties: (i) the regret is polynomial in the size of the problem instance, *i.e.*, $\mathsf{poly}(n, m, d)$, and goes to zero as a polynomial of T; (ii) the per-round running time is $\mathsf{poly}(t, n, m, d)$. An algorithm satisfying property (i) is usually called a *no-regret* algorithm.

In the case in which requiring no-regret is too limiting, we use the following relaxed notion of regret. An algorithm has *no-α-regret* if there exists a constant $c > 0$ such that: $R^T \leq \alpha + \frac{1}{T^c} \mathsf{poly}(n, m, d)$. The idea of no-$\alpha$-regret is that the regret approaches α after a sufficiently large number of rounds (polynomial in the size of the game).

4 Hardness of sub-linear regret

Our first result is negative: for any $\alpha < 1$, it is unlikely (*i.e.*, technically, it is not the case unless $\mathsf{NP} \subseteq \mathsf{RP}$) that there exists a no-α-regret algorithm for the online Bayesian persuasion problem requiring a per-round running time polynomial in the size of the instance. In order to prove the result, we provide an intermediate step, showing that the problem of approximating an optimal signaling scheme is computationally intractable even in the *offline* Bayesian persuasion problem in which the sender knows the probability distribution over the receiver's types (see Theorem 1 below).

Definition 2 (OPT-SIGNAL)**.** *Given an offline Bayesian persuasion problem in which the distribution over the receiver's types $\rho \in \Delta_{\mathcal{K}}$ is uniform, i.e., $\rho_k = \frac{1}{n}$ for all $k \in \mathcal{K}$, we call* OPT-SIGNAL *the problem of finding an optimal signaling scheme $\phi : \Theta \to \Delta_{\mathcal{S}}$, i.e., one maximizing the sender's expected utility $\frac{1}{n} \sum_{k \in \mathcal{K}} u^{\mathsf{s}}(\phi, k)$.*

Then, we can prove the following result (the omitted proofs can be found in Appendix B).

Theorem 1. *For every $0 \leq \alpha < 1$, it is NP-hard to compute an α-optimal solution to* OPT-SIGNAL.

Now, we use the approximation-hardness of the offline Bayesian persuasion problem to provide lower bounds on the α-regret in the online setting. In order to do this, we employ a set of techniques introduced by Roughgarden and Wang [31], which lead to the following result. [6].

Theorem 2. *For every $\alpha < 1$, there is no polynomial-time algorithm for the online Bayesian persuasion problem providing no-α-regret, unless $\mathsf{NP} \subseteq \mathsf{RP}$.*

5 Full information feedback setting

The negative result of the previous section (Theorem 2) rules out the possibility of designing an algorithm which satisfies the no-regret property and requires a $\mathsf{poly}(t, n, m, d)$ per-round running time. A natural question is whether it is possible to devise a no-regret algorithm for the online Bayesian persuasion problem by relaxing the running-time constraint. This is not a trivial problem because, at every round t, the sender has to choose a signaling scheme among an infinite number of alternatives and her/his utility depends on the receiver's best response, which yields a function that is not linear nor convex (or even continuous in the space of the signaling schemes). However, we show that it is possible to provide a no-regret algorithm for the full information setting by restricting the sender's action space to a finite set of posteriors. All the omitted proofs are in Appendix C.

First, we show that it is always possible to design a sender-optimal signaling scheme defined as a convex combination of a specific finite set of posteriors. For each type $k \in \mathcal{K}$ and action $a \in \mathcal{A}$, we define $\Xi_a^k \subseteq \Delta_\Theta$ as the set of posterior beliefs in which a is a receiver's best response. Formally,

[6] Theorem 2 can be obtained as a corollary of Theorem 6.2 by Roughgarden and Wang [31].

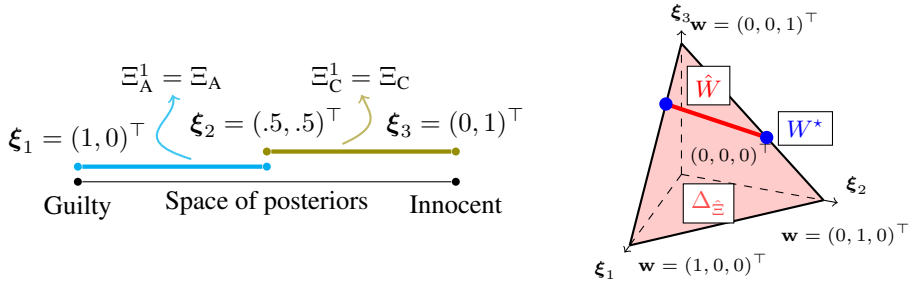

Figure 2: ***Left***: Subdivision of the space of posteriors Ξ in the two best-response regions. If $\boldsymbol{\xi} \in \Xi_{\text{A}}$ (resp., $\boldsymbol{\xi} \in \Xi_{\text{C}}$) then the judge's best response under $\boldsymbol{\xi}$ is acquitting (resp., convicting) the defendant. When $\boldsymbol{\xi} = \boldsymbol{\xi}_2$, the judge is indifferent among her/his available actions. We have $\hat{\Xi} = \{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3\}$. ***Right***: Visual depiction of $\Delta_{\hat{\Xi}}$, $\hat{W} \subseteq \Delta_{\hat{\Xi}}$, and $W^\star = V(\hat{W})$. The set $\hat{W}$ comprises of the distributions over posteriors in $\hat{\Xi}$ consistent with the prior $\boldsymbol{\mu} = (.3,.7)^\top$ and it is obtained by intersecting $\Delta_{\hat{\Xi}}$ with $[\boldsymbol{\xi}_1 \mid \boldsymbol{\xi}_2 \mid \boldsymbol{\xi}_3] \cdot \mathbf{w} \geq \boldsymbol{\mu}$. As a result, we obtain $\hat{W} = \text{conv}\{(.3,0,.7)^\top, (0,.6,.4)^\top\}$. Finally, $W^\star = V(\hat{W}) = \{(.3,0,.7)^\top, (0,.6,.4)^\top\}$.

$\Xi^k_a := \left\{\boldsymbol{\xi} \in \Xi \mid a \in \mathcal{B}^k_{\boldsymbol{\xi}}\right\}$. Let $\mathbf{a} = (a^k)_{k\in\mathcal{K}} \in \bigtimes_{k\in\mathcal{K}} \mathcal{A}$ be a tuple specifying one action for each receiver's type k. Then, for each tuple $\mathbf{a}$, let $\Xi_{\mathbf{a}} \subseteq \Delta_\Theta$ be the (potentially empty) polytope such that each action a^k is optimal for the corresponding type k, *i.e.,* $\Xi_{\mathbf{a}} := \bigcap_{k\in\mathcal{K}} \Xi^k_{a^k}$. The polytope $\Xi_{\mathbf{a}}$ has a simple interpretation: a probability distribution over posteriors in $\Xi_{\mathbf{a}}$ yields a signaling scheme such that, for every type k, the receiver has no interest in deviating from a^k in the induced posteriors $\Xi_{\mathbf{a}}$ (*i.e.,* the constraints analogous to those of the example in Section 2.2 are satisfied).

Then, let $\hat{\Xi} \subseteq \Xi$ be the set of posteriors defined as $\hat{\Xi} := \bigcup_{\mathbf{a}\in\bigtimes_{k\in\mathcal{K}}\mathcal{A}} V(\Xi_{\mathbf{a}})$. [7] Finally, we define the following set of consistent (according to Equation (3)) distributions over posteriors in $\hat{\Xi}$:

$$\hat{W} := \left\{\mathbf{w} \in \Delta_{\hat{\Xi}} \mid \sum_{\boldsymbol{\xi}\in\hat{\Xi}} w_\theta \xi_\theta = \mu_\theta, \ \forall \theta \in \Theta\right\}. \tag{4}$$

By letting M be a suitably defined $|\Theta| \times |\hat{\Xi}|$-dimensional matrix with one column for each $\xi \in \hat{\Xi}$, then the affine hyperplanes defined by Equation (3) are in the form $M \cdot \mathbf{w} = \boldsymbol{\mu}$. Since $\mathbf{w} \in \Delta_{\hat{\Xi}}$, we can safely rewrite the consistency constraints as $M \cdot \mathbf{w} \geq \boldsymbol{\mu}$ (see the example below for a better intuition). Then, $\hat{W}$ can be seen as the intersection between the simplex $\Delta_{\hat{\Xi}}$ and a finite number of half-spaces. Therefore, $\hat{W}$ is a convex polytope, whose vertices compose the finite action space that will be employed by the no-regret algorithm. Specifically, let

$$W^\star := V(\hat{W}). \tag{5}$$

Example Consider the game of Section 2.2 (see Figure 1–Left) where the receiver has a single type (*type 1*). We obtain $\hat{\Xi}$ by partitioning the space of posteriors in different best response regions and by taking the vertices of the resulting polytopes (see Figure 2–Left). Then, we provide a visual depiction of $\hat{W}$ and $W^\star$, which are obtained, respectively, by intersecting $\Delta_{\hat{\Xi}}$ with the hyperplanes corresponding to consistency constraints (see Equation (4)), and then taking the vertices of the resulting polytope (see Figure 2–Right). Another example, with two receiver's types, is provided in Appendix A.

For an arbitrary sequence of receiver's types, we show that there exists $\mathbf{w}^\star \in W^\star$ guaranteeing to the sender an expected utility that is equal to the best-in-hindsight signaling scheme.

Lemma 1. *For every sequence of receiver's types* $\boldsymbol{k} = \{k^t\}_{t\in[T]}$*, it holds*

$$\max_{\mathbf{w}\in W} \sum_{t=1}^{T} u^{\text{s}}(\mathbf{w}, k^t) = \max_{\mathbf{w}^\star\in W^\star} \sum_{t=1}^{T} u^{\text{s}}(\mathbf{w}^\star, k^t).$$

[7] $V(X)$ denotes the set of vertices of polytope X.

The size of the sender's finite action space grows exponentially in the number of states of nature d.

Lemma 2. *The size of $W^\star$ is $|W^\star| \in O\left((n\,m^2+d)^d\right)$.*

Now, by letting $\eta \in [0,1]$ be the maximum absolute payoff value, we can employ any algorithm satisfying $R^T \leq O\left(\eta\sqrt{\log|A|/T}\right)$ as a black box (see, *e.g.*, *Polynomial Weights* [15] and *Follow the Lazy Leader* [22]). By taking $W^\star$ as the sender action space, we obtain the following.

Theorem 3. *Given an online Bayesian persuasion problem with full information feedback, there exists an online algorithm such that, for every sequence of receiver's types $\boldsymbol{k} = \{k^t\}_{t\in[T]}$:*

$$R^T \leq O\left(\sqrt{\frac{d\,\log(n\,m^2+d)}{T}}\right).$$

Notice that any no-regret algorithm working on $W^\star$ requires a per-round running time polynomial in n, m and exponential in d (see the bound in Lemma 2). This shows that the source of the hardness result in Theorem 2 is the number of states of nature d, while achieving no-regret in polynomial time is possible when the parameter d is fixed.

6 Partial information feedback setting

In this setting, at every round t, the sender can only observe the action a^t played by the receiver. Therefore, the sender has no information on the utility $u^{\mathsf{s}}(\mathbf{w}, k^t)$ that she/he would have obtained by choosing any signaling scheme $\mathbf{w} \in W^\star$ other than $\mathbf{w}^t$. We show how to design no-regret algorithms with regret bounds that depend polynomially in the size of the problem instance by exploiting a reduction from the partial information setting to the full information one. [8] The main idea is to use a full-information no-regret algorithm in combination with a mechanism to estimate the sender's utilities corresponding to signaling schemes different from the one recommended by the algorithm. In particular, the overall time horizon T is split into a given number of equally-sized blocks, each corresponding to a window of time simulating a single round of a full information setting. During this window, the strategy suggested by the full-information algorithm is played in most of the rounds (exploitation phase), while only few rounds are chosen uniformly at random and used by the mechanism that estimates the utilities provided by other signaling schemes (exploration phase). Algorithm 1 provides a sketch of the overall procedure, where Z (Line 1) denotes the number of blocks, which are the intervals of consecutive rounds $\{I_\tau\}_{\tau\in[Z]}$ defined in Line 4. The FULL-INFORMATION$(\cdot)$ sub-procedure is a black box representing a no-regret algorithm for the full information setting, working on a subset $W^\circ \subseteq W^\star$ of signaling schemes. After the execution of all the rounds of each block $\tau \in [Z]$, it takes as input the utility estimates computed during I_τ and returns a recommended strategy $\mathbf{q}^{\tau+1} \in \Delta_{W^\circ}$ for the next block $I_{\tau+1}$ (see Line 14).

Algorithm 1 ONLINE BAYESIAN PERSUASION WITH PARTIAL INFORMATION FEEDBACK

Input: Full-information no-regret algorithm FULL-INFORMATION$(\cdot)$ working on $W^\circ \subseteq W^\star$; subset of signaling schemes $W^\circledcirc \subseteq W^\star$ used for exploration ▷ See Appendix D.2 for the definitions of W° and $W^\circledcirc$

1: Let Z be defined as in Theorem 3
2: Let $\mathbf{q}^1 \in \Delta_{W^\circ}$ be the uniform distribution over W°
3: **for** $\tau = 1, \dots, Z$ **do**
4: $\quad I_\tau \leftarrow \left\{(\tau-1)\frac{T}{Z}+1, \dots, \tau\frac{T}{Z}\right\}$
5: $\quad$ Choose a random permutation $\pi : [|W^\circledcirc|] \to W^\circledcirc$ and $t_1, \dots, t_{|W^\circledcirc|}$ rounds at random from I_τ
6: $\quad$ **for** $t = (\tau-1)\frac{T}{Z}+1, \dots, \tau\frac{T}{Z}$ **do**
7: $\quad\quad$ **if** $t = t_j$ for some $j \in [|W^\circledcirc|]$ **then**
8: $\quad\quad\quad \mathbf{q}^t \leftarrow \mathbf{q} \in \Delta_{W^\star}$ such that $q_{\mathbf{w}} = 1$ for the signaling scheme $\mathbf{w} = \pi(j)$ ▷ Exploration phase
9: $\quad\quad$ **else**
10: $\quad\quad\quad \mathbf{q}^t \leftarrow \mathbf{q}^\tau$ ▷ Exploitation phase
11: $\quad\quad$ Play a signaling scheme $\mathbf{w}^t \in W^\star$ randomly drawn from $\mathbf{q}^t$
12: $\quad\quad$ Observe sender's utility $u^{\mathsf{s}}(\mathbf{w}^t, k^t)$ and receiver's action $a^t \in \mathcal{A}$
13: $\quad$ Compute estimators $\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})$ of $u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) \coloneqq \frac{1}{|I_\tau|}\sum_{t\in[T]:t\in I} u^{\mathsf{s}}(\mathbf{w}, k^t)$ for all $\mathbf{w} \in W^\circ$
14: $\quad \mathbf{q}^{\tau+1} \leftarrow$ FULL-INFORMATION$\left(\left\{\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})\right\}_{\mathbf{w}\in W^\circ}\right)$

[8] The reduction is an extension of those proposed by Balcan *et al.* [8] and Awerbuch and Mansour [5].

During each block I_τ with $\tau \in [Z]$, Algorithm 1 alternates between two tasks: (i) *exploration* (Line 8), trying all the signaling schemes in a subset $W^{\circledcirc} \subseteq W^\star$ given as input, so as to compute the required estimates of the sender's expected utilities; and (ii) *exploitation* (Line 10), playing strategy $\mathbf{q}^\tau$ recommend by FULL-INFORMATION$(\cdot)$ for I_τ.

Our main result is the proof that Algorithm 1 achieves the no-regret property. Formally:

Theorem 4. *Given an online Bayesian persuasion problem with partial feedback, there exist $W^\circ \subseteq W^\star$, $W^{\circledcirc} \subseteq W^\star$, and estimators $\tilde{u}^{\mathtt{s}}_{I_\tau}(\mathbf{w})$ such that Algorithm 1 provides the following regret bound:*

$$R^T \leq O\left(\frac{nm^{2/3}d\log^{1/3}(mn+d)}{T^{1/5}}\right).$$

In order to prove this result, we show that Algorithm 1 provides a regret bound that depends on the number $|W^{\circledcirc}|$ of signaling schemes used for exploration, the logarithm of $|W^\circ|$, and the range and bias of the estimators $\tilde{u}^{\mathtt{s}}_{I_\tau}(\mathbf{w})$. To do this, we extend a result shown by Balcan *et al.* [8, Lemma 6.2] to the more general case in which only *biased* utility estimators are available, rather than unbiased ones. This result can be generalized to any partial information setting (beyond online Bayesian persuasion).

In any block I_τ with $\tau \in [Z]$, for every $\mathbf{w} \in W^\circ$, we assume that Algorithm 1 has access to an estimator $\tilde{u}^{\mathtt{s}}_{I_\tau}(\mathbf{w})$ of the sender's average utility $u^{\mathtt{s}}_{I_\tau}(\mathbf{w}) = \frac{1}{|I_\tau|}\sum_{t\in[T]:t\in I} u^{\mathtt{s}}(\mathbf{w}, k^t)$ obtained by committing to $\mathbf{w}$ during the block I_τ, with the following properties:

(i) the *bias is bounded* by a given constant $\iota \in (0,1)$, *i.e.*, it holds $\left|u^{\mathtt{s}}_{I_\tau}(\mathbf{w}) - \mathbb{E}\left[\tilde{u}^{\mathtt{s}}_{I_\tau}(\mathbf{w})\right]\right| \leq \iota$;

(ii) the *range is limited*, *i.e.*, there exists a $\eta \in \mathbb{R}$ such that $\tilde{u}^{\mathtt{s}}_{I_\tau}(\mathbf{w}) \in [-\eta, +\eta]$.

Lemma 3. *Suppose that Algorithm 1 has access to estimators $\tilde{u}^{\mathtt{s}}_{I_\tau}(\mathbf{w})$ with properties (i) and (ii) for some constants $\iota \in (0,1)$ and $\eta \in \mathbb{R}$, for every signaling scheme $\mathbf{w} \in W^\circ$ and block I_τ with $\tau \in [Z]$. Moreover, let $Z \coloneqq T^{2/3}|W^{\circledcirc}|^{-2/3}\eta^{2/3}\log^{1/3}|W^\circ|$. Then, Algorithm 1 guarantees regret:*

$$R^T \leq O\left(\frac{|W^{\circledcirc}|^{1/3}\eta^{2/3}\log^{1/3}|W^\circ|}{T^{1/3}}\right) + O(\iota).$$

Lemma 3 shows that even if utility estimators have small bias, we can still hope for a no-regret algorithm. However, we have to guarantee that $W^{\circledcirc}$ has a polynomial size, and that the estimator has a limited range. These requirements can be achieved by estimating sender's utilities indirectly by means of other related estimates, at the cost of giving up on the unbiasedness of the estimators.

The key observation that allows to get the desired estimators $\tilde{u}^{\mathtt{s}}_{I_\tau}(\mathbf{w})$ by only exploring a polynomially-sized set $W^{\circledcirc}$ is that the utilities $u^{\mathtt{s}}_{I_\tau}(\mathbf{w})$ that we wish to estimate are *not* independent, but they all depend on the frequency of each receiver's type during block I_τ. Thus, only these (polynomially many) quantities need to be estimated. In order to do so, we use the concept of *barycentric spanners* [4] (see Appendix D.2 for the details). A direct application of barycentric spanners to our setting would require being able to induce *any* receiver's posterior during the exploration phase. Unfortunately, this is not possible as the sender is forced to play consistent signaling schemes (see Equation (2)), which could prevent her from inducing certain posteriors. We achieve the goal of keeping the bias and the range of the estimators small by adopting the following two technical caveats:

(i) we focus on posteriors that can be induced by a signaling scheme with at least some ('not too small') probability, which ensures that the resulting estimators have a limited range; and

(ii) we restrict the full-information algorithm to signaling schemes $W^\circ \subseteq W^\star$ inducing a small number of posteriors, which guarantees to have estimators with a small bias.

We provide our complete technical results in Appendix D.

7 Discussion and future works

We proposed the online Bayesian persuasion framework as a natural extension of the original model by Kamenica and Gentzkow [23]. This is, to the best of our knowledge, the first work relaxing the

assumption that the sender has a perfect knowledge of the receiver's utility function. We proved that any no-regret algorithm for this setting has to require an exponential per-round running time, and we designed no-regret algorithms for the partial and full information feedback settings with adversarially chosen sequences of types. In the future, it would be interesting to study what happens if the receiver can play, at each round, an approximate best response (ϵ-best response) to the sender signal. We conjecture that in this case it should be possible to build a no-regret algorithm with quasi-polynomial per-round running time.

Broader Impact

Bayesian persuasion is a fascinating model that suffers from some limiting assumptions, which prevented a widespread use of the framework in practical applications. This work tries to amend one of such limitations, by relaxing the constraint that the sender has to have a perfect knowledge of the payoff structure of the game. This goes in the direction of developing a complete theory of *Bayesian persuasion from data* as a framework based solely on sender's and receiver's historical observations. In the future, an application of this framework at scale (*e.g.*, on large social platforms) could raise some societal challenges (see, *e.g.*, recent works on Bayesian persuasion as an election-manipulation tool). Therefore, future research in this direction should prioritize the study of how to protect receivers from excessive information garbling, and how to incentivize senders to work towards a socially-acceptable outcome.

Acknowledgments and Disclosure of Funding

This work has been partially supported by the Italian MIUR PRIN 2017 Project ALGADIMAR "Algorithms, Games, and Digital Market".

References

[1] Ricardo Alonso and Odilon Câmara. Persuading voters. *American Economic Review*, 106(11):3590–3605, 2016.

[2] Gerry Antioch et al. Persuasion is now 30 per cent of US GDP: Revisiting McCloskey and Klamer after a quarter of a century. *Economic Round-up*, (1):1, 2013.

[3] Benjamin Assarf, Ewgenij Gawrilow, Katrin Herr, Michael Joswig, Benjamin Lorenz, Andreas Paffenholz, and Thomas Rehn. Computing convex hulls and counting integer points with `polymake`. *Mathematical Programming Computation*, 9(1):1–38, 2017.

[4] Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

[5] Baruch Awerbuch and Yishay Mansour. Adapting to a reliable network path. In *Proceedings of the twenty-second annual symposium on Principles of distributed computing*, pages 360–367, 2003.

[6] Yakov Babichenko and Siddharth Barman. Algorithmic aspects of private Bayesian persuasion. In *Innovations in Theoretical Computer Science Conference*, 2017.

[7] Ashwinkumar Badanidiyuru, Kshipra Bhawalkar, and Haifeng Xu. Targeting and signaling in ad auctions. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2545–2563, 2018.

[8] Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D. Procaccia. Commitment without regrets: Online learning in stackelberg security games. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, page 61–78, 2015.

[9] Umang Bhaskar, Yu Cheng, Young Kun Ko, and Chaitanya Swamy. Hardness results for signaling in bayesian zero-sum and network routing games. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 479–496, 2016.

[10] Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. Learning optimal commitment to overcome insecurity. In *Advances in Neural Information Processing Systems*, pages 1826–1834. 2014.

[11] Peter Bro Miltersen and Or Sheffet. Send mixed signals: earn more, work less. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 234–247, 2012.

[12] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[13] Ozan Candogan. Persuasion in networks: Public signals and k-cores. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 133–134, 2019.

[14] Matteo Castiglioni, Andrea Celli, and Nicola Gatti. Persuading voters: It's easy to whisper, it's hard to speak loud. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 1870–1877, 2020.

[15] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[16] Yu Cheng, Ho Yee Cheung, Shaddin Dughmi, Ehsan Emamjomeh-Zadeh, Li Han, and Shang-Hua Teng. Mixture selection, mechanism design, and signaling. In *56th Annual Symposium on Foundations of Computer Science*, pages 1426–1445, 2015.

[17] Vincent Conitzer and Dmytro Korzhyk. Commitment to correlated strategies. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, page 632–637, 2011.

[18] Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, page 82–90, 2006.

[19] Yuval Emek, Michal Feldman, Iftah Gamzu, Renato PaesLeme, and Moshe Tennenholtz. Signaling schemes for revenue maximization. *ACM Transactions on Economics and Computation*, 2(2):1–19, 2014.

[20] Ewgenij Gawrilow and Michael Joswig. `polymake`: a framework for analyzing convex polytopes. In *Polytopes—combinatorics and computation (Oberwolfach, 1997)*, volume 29 of *DMV Sem.*, pages 43–73. Birkhäuser, Basel, 2000.

[21] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

[22] Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

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

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

[25] Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. Learning and approximating the optimal strategy to commit to. In *International Symposium on Algorithmic Game Theory*, pages 250–262, 2009.

[26] Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 661–661, 2016.

[27] Janusz Marecki, Gerry Tesauro, and Richard Segal. Playing repeated stackelberg games with unknown opponents. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, page 821–828, 2012.

[28] Donald McCloskey and Arjo Klamer. One quarter of GDP is persuasion. *The American Economic Review*, 85(2):191–195, 1995.

[29] Praveen Paruchuri, Jonathan P. Pearce, Janusz Marecki, Milind Tambe, Fernando Ordonez, and Sarit Kraus. Playing games for security: An efficient exact algorithm for solving bayesian stackelberg games. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, page 895–902, 2008.

[30] Zinovi Rabinovich, Albert Xin Jiang, Manish Jain, and Haifeng Xu. Information disclosure as a means to security. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 645–653, 2015.

[31] Tim Roughgarden and Joshua R. Wang. Minimizing regret with multiple reserves. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, page 601–616, 2016.

[32] Shoshana Vasserman, Michal Feldman, and Avinatan Hassidim. Implementing the wisdom of waze. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 660–666, 2015.

[33] Bernhard Von Stengel and Shmuel Zamir. Leadership games with convex strategy sets. *Games and Economic Behavior*, 69(2):446–457, 2010.

[34] Haifeng Xu, Rupert Freeman, Vincent Conitzer, Shaddin Dughmi, and Milind Tambe. Signaling in bayesian stackelberg games. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, pages 150–158, 2016.

[35] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 928–936, 2003.

A Additional example

This example (see Figure 3) builds on the classical prosecutor/judge game by Kamenica and Gentzkow [23] described in Section 2.2. Here, the judge has two possible types. A judge of *type 1* gets payoff 1 for a just decision, and 0 otherwise. A judge of *type 2* has a worse perception of acquitting a guilty defendant, for which she gets -1. In this case, the computation of best-response regions is more involved because different judge's types yield different boundaries on the space of posteriors. Specifically, by Equation (4), $\hat{W}$ is the result of the intersection between the simplex $\Delta_{\hat{\Xi}}$ and the closed half-spaces specified by $[\boldsymbol{\xi}_1|\boldsymbol{\xi}_2|\boldsymbol{\xi}_3|\boldsymbol{\xi}_4] \cdot \mathbf{w} \geq \boldsymbol{\mu}$. The vertices of the resulting polytope are $\mathbf{w}_1 = (3/10, 0, 0, 7/10)^\top$, $\mathbf{w}_2 = (0, 9/10, 0, 1/10)^\top$, and $\mathbf{w}_3 = (0, 0, 3/5, 2/5)^\top$. Then, the new sender's action space can be restricted to $W^\star = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3\}$. [9]

Type 1					
		State G ($\mu_{\text{G}} = .3$)		State I ($\mu_{\text{I}} = .7$)	
$\mathcal{A}$	A	0	0	0	1
	C	1	1	1	0

Type 2					
		State G ($\mu_{\text{G}} = .3$)		State I ($\mu_{\text{I}} = .7$)	
$\mathcal{A}$	A	0	-1	0	1
	C	1	1	1	0

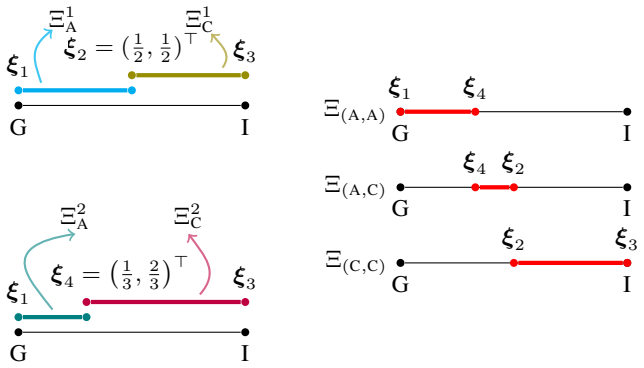

Figure 3: ***Left***: A prosecutor/judge game with two types. When the judge is of type 2 she has a worse perception of acquitting a guilty defendant. ***Center***: A visual depiction of Ξ^k_{A} and Ξ^k_{C} for each possible type $k \in \{1, 2\}$. When $k = 2$, the judge is less inclined towards acquitting and, therefore, the best-response boundary is $\boldsymbol{\xi}_4$. When $k = 1$ (resp., $k = 2$) and the posterior is $\boldsymbol{\xi}_2$ (resp., $\boldsymbol{\xi}_4$), the judge is indifferent between acquitting and convicting the defendant. ***Right***: Best-response regions for the possible joint actions. When $\mathbf{a} = (\text{C}, \text{A})$ we have $\Xi_{\mathbf{a}} = \varnothing$ because there is no posterior for which A is a best response for a receiver of type 1, and C is a best response for a receiver of type 2. We have $\hat{\Xi} = \{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3, \boldsymbol{\xi}_4\}$.

B Proofs omitted from Section 4

Theorem 1. *For every $0 \leq \alpha < 1$, it is* NP*-hard to compute an α-optimal solution to* Opt-Signal.

Proof. In order to prove Theorem 1, we resort to a result by Guruswami and Raghavendra [21] (see Theorem 5 below), which is about the following *promise problem* related to the satisfiability of a fraction of linear equations with rational coefficients and variables restricted to the hypercube.

Definition 3 (Lineq-Ma$(1-\zeta, \delta)$ by Guruswami and Raghavendra [21])**.** *For any two constants $\zeta, \delta \in \mathbb{R}$ satisfying $0 \leq \delta \leq 1 - \zeta \leq 1$,* Lineq-Ma$(1-\zeta, \delta)$ *is the following promise problem: Given a set of linear equations $\mathbf{A}\mathbf{x} = \mathbf{c}$ over variables $\mathbf{x} \in \mathbb{Q}^{n_{\text{var}}}$, with coefficients $\mathbf{A} \in \mathbb{Q}^{n_{\text{eq}} \times n_{\text{var}}}$ and $\mathbf{c} \in \mathbb{Q}^{n_{\text{var}}}$, distinguish between the following two cases:*

- *there exists a vector $\hat{\mathbf{x}} \in \{0, 1\}^{n_{\text{var}}}$ that satisfies at least a fraction $1 - \zeta$ of the equations;*
- *every possible vector $\mathbf{x} \in \mathbb{Q}^{n_{\text{var}}}$ satisfies less than a fraction δ of the equations.*

Theorem 5 (Guruswami and Raghavendra [21])**.** *For all valid $\zeta, \delta > 0$,* Lineq-Ma$(1-\zeta, \delta)$ *is* NP*-hard.*

We introduce a reduction from Lineq-Ma$(1-\zeta, \delta)$ to Opt-Signal, showing the following:

- *Completeness*: If an instance of Lineq-Ma$(1-\zeta, \delta)$ admits a $1-\zeta$ fraction of satisfiable equations when variables are restricted to lie the hypercube $\{0, 1\}^{n_{\text{var}}}$, then an optimal solution to Opt-Signal provides the sender with an expected utility at least of $1 - 2\zeta$;

[9] The polytopes were computed using `Polymake`, a tool for computational polyhedral geometry [3, 20].

- *Soundness*: If at most a δ fraction of the equations can be satisfied, then an optimal solution to OPT-SIGNAL has sender's expected utility at most δ.

Since ζ and δ can be arbitrary (with $0 \leq \delta \leq 1 - \zeta \leq 1$), the two properties above immediately prove the result. In the rest of the proof, given a vector of variables $\mathbf{x} \in \mathbb{Q}^{n_{\text{var}}}$, for $i \in [n_{\text{var}}]$, we denote with x_i the component corresponding to the i-th variable. Similarly, for $j \in [n_{\text{eq}}]$, c_j is the j-th component of the vector $\mathbf{c}$, whereas, for $i \in [n_{\text{var}}]$ and $j \in [n_{\text{eq}}]$, the (j, i)-entry of $\mathbf{A}$ is denoted by A_{ji}.

Reduction As a preliminary step, we normalize the coefficients by letting $\bar{\mathbf{A}} := \frac{1}{\tau}\mathbf{A}$ and $\bar{\mathbf{c}} := \frac{1}{\tau}\mathbf{c}$, where we let $\tau := 2 \max\left\{\max_{i \in [n_{\text{var}}], j \in [n_{\text{eq}}]} A_{ji}, \max_{j \in [n_{\text{eq}}]} c_j, n_{\text{var}}^2\right\}$. It is easy to see that the normalization preserves the number of satisfiable equations. Formally, the number of satisfied equations of $\mathbf{A}\mathbf{x} = \mathbf{c}$ is equal to the number of satisfied equations of $\bar{\mathbf{A}}\bar{\mathbf{x}} = \bar{\mathbf{c}}$, where $\bar{\mathbf{x}} = \frac{1}{\tau}\mathbf{x}$. For every variable $i \in [n_{\text{var}}]$, we define a state of nature $\theta_i \in \Theta$. Moreover, we introduce an additional state $\theta_0 \in \Theta$. The prior distribution $\mu \in \text{int}(\Delta_\Theta)$ is defined in such a way that $\mu_{\theta_i} = \frac{1}{n_{\text{var}}^2}$ for every $i \in [n_{\text{var}}]$, while $\mu_{\theta_0} = \frac{n_{\text{var}} - 1}{n_{\text{var}}}$ (notice that $\sum_{\theta \in \Theta} \mu_\theta = 1$). We define a receiver's type $k_j \in \mathcal{K}$ for each equation $j \in [n_{\text{eq}}]$ (recall that the distribution over receiver's types $\rho \in \Delta_\mathcal{K}$ is uniform by definition of OPT-SIGNAL). The receiver has three actions available, namely $\mathcal{A} := \{a_0, a_1, a_2\}$, whereas, for every $k_j \in \mathcal{K}$, the utilities of type k_j are $u_{\theta_i}^{k_j}(a_0) = \frac{1}{2}$, $u_{\theta_i}^{k_j}(a_1) = \frac{1}{2} - \bar{A}_{ji} + \bar{c}_j$, and $u_{\theta_i}^{k_j}(a_2) = \frac{1}{2} + \bar{A}_{ji} - \bar{c}_j$ for every $i \in [n_{\text{var}}]$, while $u_{\theta_0}^{k_j}(a_0) = \frac{1}{2}$, $u_{\theta_0}^{k_j}(a_1) = \frac{1}{2} + \bar{c}_j$, and $u_{\theta_0}^{k_j}(a_2) = \frac{1}{2} - \bar{c}_j$. Finally, the sender's utility is 1 when the receiver plays a_0, while it is 0 otherwise, independently of the state of nature. Formally, $u_\theta^{\mathsf{s}}(a_0) = 1$ and $u_\theta^{\mathsf{s}}(a_1) = u_\theta^{\mathsf{s}}(a_2) = 0$ for every $\theta \in \Theta$.

Completeness Suppose there exists a vector $\hat{\mathbf{x}} \in \{0, 1\}^{n_{\text{var}}}$ such that at least a fraction $1 - \zeta$ of the equations in $\mathbf{A}\hat{\mathbf{x}} = \mathbf{c}$ are satisfied. Let $X^1 \subseteq [n_{\text{var}}]$ be the set of variables $i \in [n_{\text{var}}]$ with $x_i = 1$, while $X^0 := [n_{\text{var}}] \setminus X^1$. Given the definition of $\bar{\mathbf{A}}$ and $\bar{\mathbf{c}}$, there exists a vector $\bar{\mathbf{x}} \in \{0, \frac{1}{\tau}\}^{n_{\text{var}}}$ such that at least a fraction $1 - \zeta$ of the equations in $\bar{\mathbf{A}}\bar{\mathbf{x}} = \bar{\mathbf{c}}$ are satisfied, and, additionally, $\bar{x}_i = \frac{1}{\tau}$ for all the variables in $i \in X^1$, while $\bar{x}_i = 0$ whenever $i \in X^0$. Let us consider an (indirect) signaling scheme $\phi : \Theta \to \Delta_\mathcal{S}$ defined for the set of signals $\mathcal{S} := \{s_1, s_2\}$. Let $q := \frac{n_{\text{var}}(n_{\text{var}} - 1)}{\tau - |X^1|}$. For every $i \in [n_{\text{var}}]$, we define $\phi_{\theta_i}(s_1) = q$ and $\phi_{\theta_i}(s_2) = 1 - q$ if $i \in X^1$, while $\phi_{\theta_i}(s_1) = 0$ and $\phi_{\theta_i}(s_2) = 1$ otherwise. Moreover, we let $\phi_{\theta_0}(s_1) = 1$ and $\phi_{\theta_0}(s_2) = 0$. Now, let us take the receiver's posterior $\boldsymbol{\xi}^1 \in \Delta_\Theta$ induced by signal s_1. Let $h := \frac{\frac{q}{n_{\text{var}}^2}}{\sum_{i \in X^1} \frac{q}{n_{\text{var}}^2} + \frac{n_{\text{var}} - 1}{n_{\text{var}}}}$. Then, using the definition of $\boldsymbol{\xi}^1$, it is easy to check that $\xi_{\theta_i}^1 = h$ for every $i \in X^1$, $\xi_{\theta_i}^1 = 0$ for every $i \in X^0$, while $\xi_{\theta_0}^1 = \frac{\frac{n_{\text{var}} - 1}{n_{\text{var}}}}{\sum_{i \in X^1} \frac{q}{n_{\text{var}}^2} + \frac{n_{\text{var}} - 1}{n_{\text{var}}}} = 1 - h\left|X^1\right|$. Next, we prove that given the posterior $\boldsymbol{\xi}^1$ at least a fraction $1 - \zeta$ of the receiver's types has action a_0 as a best response, implying that the expected utility of the sender is equal to $\frac{1}{n}\sum_{k \in \mathcal{K}} u^{\mathsf{s}}(\phi, k) \geq \frac{n-1}{n}(1 - \zeta) \geq 1 - 2\zeta$, which holds for n large enough. For each satisfied equality $j \in [n_{\text{eq}}]$ in $\bar{\mathbf{A}}\bar{\mathbf{x}} = \bar{\mathbf{c}}$, the receiver of type $k_j \in \mathcal{K}$ experiences a utility of $\sum_{\theta \in \Theta} \xi_\theta^1 u_\theta^{k_j}(a_0) = \frac{1}{2}$ by playing action a_0. Instead, the utility she gets by playing a_1 is defined as follows:

$$\begin{aligned}
\sum_{\theta \in \Theta} \xi_\theta^1 u_\theta^{k_j}(a_1) &= \sum_{i \in X^1} h\left(\frac{1}{2} - \bar{A}_{ji} + \bar{c}_j\right) + \xi_{\theta_0}^1\left(\frac{1}{2} + \bar{c}_j\right) = \\
&= h\left|X^1\right|\left(\frac{1}{2} + \bar{c}_j\right) - h\sum_{i \in X^1} \bar{A}_{ji} + \left(1 - h\left|X^1\right|\right)\left(\frac{1}{2} + \bar{c}_j\right) = \\
&= \frac{1}{2} + \bar{c}_j - h\sum_{i \in X^1} \bar{A}_{ji} = \frac{1}{2} + \bar{c}_j - \frac{1}{\tau}\sum_{i \in X^1} \bar{A}_{ji} = \frac{1}{2},
\end{aligned}$$

where the second to last equality holds since $h = \frac{1}{\tau}$ (by definition of h and q), while the last equality follows from the fact that the j-th equation is satisfied, and, thus, $\frac{1}{\tau}\sum_{i \in X^1} \bar{A}_{ji} = \bar{c}_j$ (recall that $\bar{x}_i = \frac{1}{\tau}$ for all $i \in X^1$). Using similar arguments, we can write $\sum_{\theta \in \Theta} \xi_\theta^1 u_\theta^{k_j}(a_2) = \frac{1}{2}$, which concludes the completeness proof.

Soundness Suppose, by contradiction, that there exists a signaling scheme $\phi : \Theta \to \Delta_\mathcal{S}$ providing the sender with an expected utility greater than δ. This implies, by an averaging argument, that there exists a signal inducing a posterior $\boldsymbol{\xi} \in \Delta_\Theta$ in which at least a fraction δ of the receiver's types best responds by playing action a_0. Let $\mathcal{K}^1 \subseteq \mathcal{K}$ be the set of such reviver's types. For every receiver's type $k_j \in \mathcal{K}$, it holds $\sum_{\theta \in \Theta} \xi_\theta u_\theta^{k_j}(a_0) = \frac{1}{2}$.

Moreover, it is the case that:

$$\sum_{\theta\in\Theta}\xi_\theta u_\theta^{k_j}(a_1)=\sum_{i\in[n_{\text{var}}]}\xi_{\theta_i}\left(\frac{1}{2}-\bar{A}_{ji}+\bar{c}_j\right)+\xi_{\theta_0}\left(\frac{1}{2}+\bar{c}_j\right)=\frac{1}{2}+\bar{c}_j-\sum_{i\in[n_{\text{var}}]}\xi_{\theta_i}\bar{A}_{ji}.$$

Similarly, it holds:

$$\sum_{\theta\in\Theta}\xi_\theta u_\theta^{k_j}(a_2)=\frac{1}{2}-\bar{c}_j+\sum_{i\in[n_{\text{var}}]}\xi_{\theta_i}\bar{A}_{ji}.$$

By assumption, for every type $k_j\in\mathcal{K}^1$, it is the case that $\sum_{\theta\in\Theta}\xi_\theta u_\theta^{k_j}(a_0)\geq\sum_{\theta\in\Theta}\xi_\theta u_\theta^{k_j}(a_1)$, which implies that $\bar{c}_j-\sum_{i\in[n_{\text{var}}]}\xi_{\theta_i}\bar{A}_{ji}\leq 0$, whereas $\sum_{\theta\in\Theta}\xi_\theta u_\theta^{k_j}(a_0)\geq\sum_{\theta\in\Theta}\xi_\theta u_\theta^{k_j}(a_2)$, implying $-\bar{c}_j+\sum_{i\in[n_{\text{var}}]}\xi_{\theta_i}\bar{A}_{ji}\leq 0$. Thus, $\sum_{i\in[n_{\text{var}}]}\xi_{\theta_i}\bar{A}_{ji}=\bar{c}_j$ for every $j\in[n_{\text{eq}}]$ such that $k_j\in\mathcal{K}^1$ and the vector $\hat{\mathbf{x}}\in\mathbb{Q}^{n_{\text{var}}}$ with $\hat{x}_i=\xi_{\theta_i}$ for all $i\in[n_{\text{var}}]$ satisfies at least a fraction δ of the equations, reaching a contradiction. □

C Proofs omitted from Section 5

Lemma 1. *For every sequence of receiver's types $\boldsymbol{k}=\{k^t\}_{t\in[T]}$, it holds*

$$\max_{\mathbf{w}\in W}\sum_{t=1}^{T}u^{\mathsf{s}}(\mathbf{w},k^t)=\max_{\mathbf{w}^\star\in W^\star}\sum_{t=1}^{T}u^{\mathsf{s}}(\mathbf{w}^\star,k^t).$$

Proof. The idea to prove the lemma is the following: any posterior distribution $\boldsymbol{\xi}$ in $\text{supp}(\mathbf{w})$ can be represented as the convex combination of elements of $\hat{\Xi}$. We denote such convex combination by $\mathbf{w}^{\boldsymbol{\xi}}\in\Delta_{\hat{\Xi}}$. We define a new signaling scheme $\mathbf{w}^\star\in\Delta_{\hat{\Xi}}$ as follows:

$$w^\star_{\boldsymbol{\xi}'}:=\sum_{\substack{\boldsymbol{\xi}\in\text{supp}(\mathbf{w}):\\ \boldsymbol{\xi}'\in\text{supp}(\mathbf{w}^{\boldsymbol{\xi}})}}w_{\boldsymbol{\xi}}w^{\boldsymbol{\xi}}_{\boldsymbol{\xi}'}\quad\text{for each}\quad\boldsymbol{\xi}'\in\hat{\Xi}.\tag{6}$$

Since $\mathbf{w}$ is consistent (*i.e.,* $\mathbf{w}\in W$) we have by construction that $\mathbf{w}^\star$ is consistent, and therefore $\mathbf{w}^\star\in\hat{W}$. Finally, we show that $\mathbf{w}^\star$ guarantees to the sender an expected utility which is greater than or equal to that achieved via $\mathbf{w}$. The crucial point here is showing that whenever the decomposition over $\hat{\Xi}$ involves a vertex (*i.e.,* a posterior) where the receiver is indifferent between two or more actions, her/his choice does not damage the sender. This happens at the boundaries of best-response regions (see, *e.g.*, what happens at $\boldsymbol{\xi}_2$ and $\boldsymbol{\xi}_4$ in the example of Figure 3). The sender's expected utility is a linear function of the signaling scheme $\mathbf{w}^\star$. Therefore, the sender can limit her attention to $W^\star$, since her/his maximum expected utility is attained at one of the vertices of $\hat{W}$.

Consider a posterior $\boldsymbol{\xi}\in\Xi$ and let $\mathbf{a}=\{b^k_{\boldsymbol{\xi}}\}_{k\in\mathcal{K}}$ (*i.e.*, $\mathbf{a}$ is the tuple specifying the best-response action under posterior $\boldsymbol{\xi}$ for each receiver's type k). Tuple $\mathbf{a}$ defines polytope $\Xi_{\mathbf{a}}\subseteq\Xi$. By Carathéodory's theorem, any $\boldsymbol{\xi}\in\Xi_{\mathbf{a}}$ is the convex combination of a finite number of points in $\Xi_{\mathbf{a}}$. Specifically, there exists $\mathbf{w}^{\boldsymbol{\xi}}\in\Delta_{V(\Xi_{\mathbf{a}})}$ such that, for each $\theta\in\Theta$, $\sum_{\boldsymbol{\xi}'\in V(\Xi_{\mathbf{a}})}w^{\boldsymbol{\xi}}_{\boldsymbol{\xi}'}\xi'_\theta=\xi_\theta$.

Let $\mathbf{w}\in\hat{W}$ (*i.e.,* $\mathbf{w}$ is consistent). By following Equation (6), we define a distribution $\mathbf{w}^\star$ such that, for each $\boldsymbol{\xi}'\in\hat{\Xi}$,

$$w^\star_{\boldsymbol{\xi}'}:=\sum_{\substack{\boldsymbol{\xi}\in\text{supp}(\mathbf{w}):\\ \boldsymbol{\xi}'\in\text{supp}(\mathbf{w}^{\boldsymbol{\xi}})}}w_{\boldsymbol{\xi}}w^{\boldsymbol{\xi}}_{\boldsymbol{\xi}'}.$$

By construction, $\mathbf{w}^\star$ is a well-defined convex combination of elements of $\hat{\Xi}$. Moreover, since $\mathbf{w}$ is consistent, the same holds true for $\mathbf{w}^\star$, which implies $\mathbf{w}^\star\in\hat{W}$.

Fix a type $k\in\mathcal{K}$ and a posterior $\boldsymbol{\xi}\in\Xi$, and let $\mathbf{a}$ be defined as the tuple specifying the best response under $\boldsymbol{\xi}$ for each k. At each posterior $\boldsymbol{\xi}'\in V(\Xi_{\mathbf{a}})$, the receiver plays $b^k_{\boldsymbol{\xi}'}$. The following holds:

$$b^k_{\boldsymbol{\xi}'}\in\arg\max_{a'\in\mathcal{B}^k_{\boldsymbol{\xi}'}}\sum_{\theta\in\Theta}\xi'_\theta u^{\mathsf{s}}_\theta(a')\geq\sum_{\theta\in\Theta}\xi'_\theta u^{\mathsf{s}}_\theta(b^k_{\boldsymbol{\xi}}),\tag{7}$$

where the inequality holds because, by construction, $b^k_{\boldsymbol{\xi}} \in \mathcal{B}^k_{\boldsymbol{\xi}'}$. Therefore, we can show that the sender's expected utility when decomposing $\boldsymbol{\xi}$ as $\mathbf{w}^{\boldsymbol{\xi}} \in \Delta_{V(\Xi_{\mathbf{a}})}$ is guaranteed to be greater than or equal to the expected utility under $\boldsymbol{\xi}$. Specifically,

$$\begin{aligned}
\sum_{\boldsymbol{\xi}' \in V(\Xi_{\mathbf{a}})} w^{\boldsymbol{\xi}}_{\boldsymbol{\xi}'} u^{\mathsf{s}}(\boldsymbol{\xi}', k) &= \sum_{\boldsymbol{\xi}' \in V(\Xi_{\mathbf{a}})} w^{\boldsymbol{\xi}}_{\boldsymbol{\xi}'} \sum_{\theta \in \Theta} \xi'_\theta u^{\mathsf{s}}_\theta(b^k_{\boldsymbol{\xi}'}) \\
&\geq \sum_{\boldsymbol{\xi}' \in V(\Xi_{\mathbf{a}})} w^{\boldsymbol{\xi}}_{\boldsymbol{\xi}'} \sum_{\theta \in \Theta} \xi'_\theta u^{\mathsf{s}}_\theta(b^k_{\boldsymbol{\xi}}) && \text{(By Equation (7))} \\
&= \sum_{\theta \in \Theta} \xi_\theta u^{\mathsf{s}}_\theta(b^k_{\boldsymbol{\xi}}) && \text{(By definition of } \mathbf{w}^{\boldsymbol{\xi}}) \\
&= u^{\mathsf{s}}(\boldsymbol{\xi}, k).
\end{aligned}$$

Let $\mathbf{w} \in W$ be the best-in-hindsight signaling scheme. We show that, for any sequence of receiver's types $\mathbf{k} = \{k^t\}_{t \in [T]}$, the sender's expected utility achieved via $\mathbf{w}$ is matched by the expected utility guaranteed by $\mathbf{w}^\star \in \hat{W}$ defined as in Equation (6). We have

$$\begin{aligned}
\sum_{t \in [T]} \sum_{\boldsymbol{\xi} \in \text{supp}(\mathbf{w}^\star)} w^\star_{\boldsymbol{\xi}} u^{\mathsf{s}}(\boldsymbol{\xi}, k^t) &= \sum_{t \in [T]} \sum_{\boldsymbol{\xi} \in \text{supp}(\mathbf{w}^\star)} \sum_{\substack{\boldsymbol{\xi}' \in \text{supp}(\mathbf{w}): \\ \boldsymbol{\xi} \in \text{supp}(\mathbf{w}^{\boldsymbol{\xi}'})}} w_{\boldsymbol{\xi}'} w^{\boldsymbol{\xi}'}_{\boldsymbol{\xi}} u^{\mathsf{s}}(\boldsymbol{\xi}, k^t) \\
&= \sum_{t \in [T]} \sum_{\boldsymbol{\xi}' \in \text{supp}(\mathbf{w})} w_{\boldsymbol{\xi}'} \sum_{\boldsymbol{\xi} \in \text{supp}(\mathbf{w}^{\boldsymbol{\xi}'})} w^{\boldsymbol{\xi}'}_{\boldsymbol{\xi}} u^{\mathsf{s}}(\boldsymbol{\xi}, k^t) \\
&\geq \sum_{t \in [T]} \sum_{\boldsymbol{\xi}' \in \text{supp}(\mathbf{w})} u^{\mathsf{s}}(\boldsymbol{\xi}', k^t) \\
&= \sum_{t \in [T]} u^{\mathsf{s}}(\mathbf{w}, k^t).
\end{aligned}$$

Finally, since $\sum_{t \in [T]} u^{\mathsf{s}}(\mathbf{w}^\star, k^t) = \sum_{t \in [T]} \sum_{\boldsymbol{\xi} \in \text{supp}(\mathbf{w}^\star)} w^\star_{\boldsymbol{\xi}} u^{\mathsf{s}}(\boldsymbol{\xi}, k^t)$ is a linear function in the signaling scheme $\mathbf{w}^\star$, its maximum is attained at a vertex of $\hat{W}$. This concludes the proof. □

Lemma 2. *The size of $W^\star$ is $|W^\star| \in O\left((n\,m^2 + d)^d\right)$.*

Proof. By definition, for any $\mathbf{a} = (a^k)_{k \in \mathcal{K}}$, $W_{\mathbf{a}} \subseteq \Xi$. Then, each $\mathbf{w} \in V(W_{\mathbf{a}})$ is an extreme point of a $(d-1)$-dimensional convex polytope, and therefore the point lies at the intersection of $(d-1)$ linearly independent defining half-spaces of the polytope. Now, to provide a bound for $|\hat{\Xi}|$ we first compute the number of half-spaces separating best-response regions corresponding to different actions. For each type $k \in \mathcal{K}$, there are at most $\binom{m}{2}$ half-spaces each separating W^k_a and $W^k_{a'}$ for two actions $a \neq a'$. Then, in order to take all the incentive constraints into account, we have to sum over all possible reveiver's types, obtaining $O(n\,m^2)$ half-spaces. The set $\hat{\Xi}$ is the result of the intersection between the region defined by such half-spaces, and the d constraints defining the simplex. Each extreme point of the polytope defined by points in $\hat{\Xi}$ lies at the intersection of $d-1$ half-spaces. Therefore, there are at most $\binom{nm^2+d}{d-1} \in O\left((nm^2+d)^d\right)$ such extreme points. The convex polytope $\hat{W}$ is the result of the intersection between the simplex defined over $\hat{\Xi}$, which has $O\left((nm^2+d)^d\right)$ extreme points, and d half-spaces defining consistency constraints. Then, $\hat{W}$ has a number of extreme points which is less than or equal to $O\left((nm^2+d)^d\right)$. □

Theorem 3. *Given an online Bayesian persuasion problem with full information feedback, there exists an online algorithm such that, for every sequence of receiver's types $\mathbf{k} = \{k^t\}_{t \in [T]}$:*

$$R^T \leq O\left(\sqrt{\frac{d\,\log(n\,m^2+d)}{T}}\right).$$

Proof. We employ an arbitrary algorithm satisfying $R^T \leq O\Big(\eta\sqrt{\log|A|/T}\Big)$ with action set $A = W^\star$. Let $\mathbf{w}^* \in W$ be the sender-optimal signaling scheme in hindsight. Then,

$$\begin{aligned}\sum_{t\in[T]} \mathbb{E}[u^{\mathsf{s}}(\mathbf{w}^t, k^t)] &\geq \sum_{t\in[T]} u^{\mathsf{s}}(\mathbf{w}^*, k^t) - O\left(\sqrt{T\log|W^\star|}\right) \\ &\geq \sum_{t\in[T]} u^{\mathsf{s}}(\mathbf{w}^*, k^t) - O\left(\sqrt{T\log\left(nm^2+d\right)^d}\right) \qquad \text{(By Lemma 2)} \\ &= \sum_{t\in[T]} u^{\mathsf{s}}(\mathbf{w}^*, k^t) - O\left(\sqrt{Td\log\left(nm^2+d\right)}\right).\end{aligned}$$

This completes the proof. □

D Additional results on the partial information feedback setting

Appendix D.1 reports the proof of Lemma 3, which shows a regret bound for the reduction from partial information to full information that exploits biased estimators. Appendix D.2 provides a detailed treatment on how Algorithm 1 computes the required sender's utility estimates. Finally, Appendix D.3 concludes with the proof of Theorem 4.

D.1 Proof of Lemma 3

Lemma 3. *Suppose that Algorithm 1 has access to estimators $\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})$ with properties (i) and (ii) for some constants $\iota \in (0,1)$ and $\eta \in \mathbb{R}$, for every signaling scheme $\mathbf{w} \in W^\circ$ and block I_τ with $\tau \in [Z]$. Moreover, let $Z \coloneqq T^{2/3}|W^\circledcirc|^{-2/3}\eta^{2/3}\log^{1/3}|W^\circ|$. Then, Algorithm 1 guarantees regret:*

$$R^T \leq O\left(\frac{|W^\circledcirc|^{1/3}\eta^{2/3}\log^{1/3}|W^\circ|}{T^{1/3}}\right) + O(\iota).$$

Proof. In order to prove the desired regret bound for Algorithm 1, we rely on two crucial observations:

- during the exploration phase of each block I_τ with $\tau \in [Z]$, *i.e.*, the iterations $t_1, \ldots, t_{|W^\circledcirc|}$, the algorithm plays a strategy $\mathbf{q}^t \neq \mathbf{q}^\tau$, where $\mathbf{q}^\tau$ is the last strategy recommended by FULL-INFORMATION$(\cdot)$, resulting in a corresponding utility loss that can be as large as -1 (since the utilities are in the range $[0,1]$);
- running the full-information no-regret algorithm (*i.e.*, the sub-procedure FULL-INFORMATION$(\cdot)$) using biased estimates of the sender's utilities (rather than their real values) results in the regret bound being worsened by only a term that is proportional to the bias ι of the adopted estimators.

In the following, we denote with R^Z_{full} the cumulative regret achieved by FULL-INFORMATION$(\cdot)$, where we remark the fact that each block I_τ simulates a single iteration of the full information setting, and, thus, the number of iterations for the full-information algorithm is Z rather than T. Formally, we have the following definition:

$$R^Z_{\text{full}} \coloneqq \max_{\mathbf{w}\in W^\circ} \sum_{\tau\in[Z]} \tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w}) - \sum_{\tau\in[Z]} \sum_{\mathbf{w}\in W^\circ} q^\tau_{\mathbf{w}} \tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w}),$$

where we notice that the regret is computed with respect to the estimates $\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})$ of the sender's average utilities $u^{\mathsf{s}}_{I_\tau}(\mathbf{w})$ experienced in each block I_τ, defined as $u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) = \frac{1}{|I_\tau|}\sum_{t\in I_\tau} u^{\mathsf{s}}(\mathbf{w}, k^t)$ for every $\mathbf{w} \in W^\circ$. We also remark that the full-information algorithm is run on a subset $W^\circ \subseteq W^\star$ of signaling schemes, and, thus, the regret R^Z_{full} is defined with respect to them. Moreover, from Section 5, we know that there exists an algorithm satisfying the regret bound $R^Z_{\text{full}} \leq O\left(\eta\sqrt{Z\log|W^\circ|}\right)$, where η is the range of the utility values observed by the algorithm that, in our case, corresponds to the range of the estimates observed by the algorithm, which is limited thanks to property (ii) of the estimators.

In order to prove the result, we also need the following relation, which holds for every $\tau \in [Z]$ and signaling scheme $\mathbf{w} \in W^\circ$:

$$\sum_{t\in I_\tau} u^{\mathsf{s}}(\mathbf{w}, k^t) = |I_\tau| u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) \geq |I_\tau|\Big(\mathbb{E}[\tilde{u}^{\mathsf{s}}_{I_\tau}] - \iota\Big) = \frac{T}{Z}\Big(\mathbb{E}[\tilde{u}^{\mathsf{s}}_{I_\tau}] - \iota\Big), \tag{8}$$

where the first equality holds by definition, the inequality holds thanks to property (i) of the estimators, while the last equality is given by $|I_\tau| = \frac{T}{Z}$.

Letting U be the sender's expected utility achieved by playing according to Algorithm 1, the following relations hold:

$$
\begin{aligned}
\frac{1}{T}U &\coloneqq \frac{1}{T}\sum_{\tau\in[Z]}\sum_{t\in I_\tau}\sum_{\mathbf{w}\in W^\circ} q^t_{\mathbf{w}} u^{\mathsf{s}}(\mathbf{w}, k^t) \\
&\geq \frac{1}{T}\sum_{\tau\in[Z]}\sum_{\mathbf{w}\in W^\circ} q^\tau_{\mathbf{w}}\sum_{t\in I_\tau} u^{\mathsf{s}}(\mathbf{w}, k^t) - \frac{|W^\circledcirc|Z}{T} && (\mathbf{q}^t \neq \mathbf{q}^\tau \text{ in } |W^\circledcirc| \text{ iterations and max. loss} = -1) \\
&\geq \frac{1}{T}\sum_{\tau\in[Z]}\sum_{\mathbf{w}\in W^\circ} q^\tau_{\mathbf{w}}\frac{T}{Z}\Big(\mathbb{E}\left[\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})\right] - \iota\Big) - \frac{|W^\circledcirc|Z}{T} && \text{(By Equation (8))} \\
&= \frac{1}{Z}\sum_{\tau\in[Z]}\sum_{\mathbf{w}\in W^\circ} q^\tau_{\mathbf{w}}\Big(\mathbb{E}\left[\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})\right] - \iota\Big) - \frac{|W^\circledcirc|Z}{T} \\
&= \frac{1}{Z}\sum_{\tau\in[Z]}\sum_{\mathbf{w}\in W^\circ} q^\tau_{\mathbf{w}}\mathbb{E}\left[\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})\right] - \iota - \frac{|W^\circledcirc|Z}{T} && \left(\text{Since } \sum_{\tau\in[Z]}\sum_{\mathbf{w}\in W^\circ} q^\tau_{\mathbf{w}} = Z, \text{ being } \mathbf{q}^\tau \in \Delta_{W^\circ}\right) \\
&= \frac{1}{Z}\mathbb{E}\left[\sum_{\tau\in[Z]}\sum_{\mathbf{w}\in W^\circ} q^\tau_{\mathbf{w}}\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})\right] - \iota - \frac{|W^\circledcirc|Z}{T} \\
&= \frac{1}{Z}\mathbb{E}\left[\max_{\mathbf{w}\in W^\circ}\sum_{\tau\in Z}\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w}) - R^Z_{\text{full}}\right] - \iota - \frac{|W^\circledcirc|Z}{T} && \text{(Definition of } R^Z_{\text{full}}) \\
&\geq \frac{1}{Z}\max_{\mathbf{w}\in W^\circ}\sum_{\tau\in[Z]}\mathbb{E}\left[\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})\right] - \frac{1}{Z}R^Z_{\text{full}} - \iota - \frac{|W^\circledcirc|Z}{T} && \text{(Jensen's inequality)} \\
&\geq \frac{1}{Z}\max_{\mathbf{w}\in W^\circ}\sum_{\tau\in[Z]}\left(u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) - \iota\right) - \frac{1}{Z}R^Z_{\text{full}} - \iota - \frac{|W^\circledcirc|Z}{T} && \text{(By property (i))} \\
&= \frac{1}{Z}\max_{\mathbf{w}\in W^\circ}\sum_{\tau\in[Z]}u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) - \iota - \frac{1}{Z}R^Z_{\text{full}} - \iota - \frac{|W^\circledcirc|Z}{T} \\
&= \frac{1}{Z}\max_{\mathbf{w}\in W^\circ}\frac{Z}{T}\sum_{\tau\in[Z]}\sum_{t\in I_\tau}u^{\mathsf{s}}(\mathbf{w}, k^t) - \frac{1}{Z}R^Z_{\text{full}} - 2\iota - \frac{|W^\circledcirc|Z}{T} && \left(\text{By def. of } u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) \text{ and } |I_\tau| = \frac{T}{Z}\right) \\
&= \frac{1}{T}\max_{\mathbf{w}\in W^\circ}\sum_{\tau\in[Z]}\sum_{t\in I_\tau}u^{\mathsf{s}}(\mathbf{w}, k^t) - \frac{1}{Z}R^Z_{\text{full}} - 2\iota - \frac{|W^\circledcirc|Z}{T} \\
&= \frac{1}{T}\max_{\mathbf{w}\in W^\circ}\sum_{t\in[T]}u^{\mathsf{s}}(\mathbf{w}, k^t) - \frac{1}{Z}R^Z_{\text{full}} - 2\iota - \frac{|W^\circledcirc|Z}{T} = \\
&\geq \frac{1}{T}\max_{\mathbf{w}\in W^\circ}\sum_{t\in[T]}u^{\mathsf{s}}(\mathbf{w}, k^t) - \frac{1}{Z}O\left(\eta\sqrt{Z\log|W^\circ|}\right) - 2\iota - \frac{|W^\circledcirc|Z}{T} \\
&\geq \frac{1}{T}\max_{\mathbf{w}\in W^\circ}\sum_{t\in[T]}u^{\mathsf{s}}(\mathbf{w}, k^t) - O\left(\frac{|W^\circledcirc|^{1/3}\eta^{2/3}\log^{1/3}|W^\circ|}{T^{1/3}}\right) - 2\iota - \frac{|W^\circledcirc|^{1/3}\eta^{2/3}\log^{1/3}|W^\circ|}{T^{1/3}} \\
&\geq \frac{1}{T}\max_{\mathbf{w}\in W^\circ}\sum_{t\in[T]}u^{\mathsf{s}}(\mathbf{w}, k^t) - O\left(\frac{|W^\circledcirc|^{1/3}\eta^{2/3}\log^{1/3}|W^\circ|}{T^{1/3}}\right) - O(\iota)
\end{aligned}
$$

By using the definition of the regret R^T of Algorithm 1, we get the statement. □

D.2 Details on sender's average utilities estimation

In the following, we show in details how to compute the estimates needed by Algorithm 1 by using random samples from a polynomially-sized set $W^{\circledcirc} \subseteq W^{\star}$. Let us recall that, during each block I_τ with $\tau \in [Z]$, Algorithm 1 needs to compute the estimators $\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})$ of $u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) = \frac{1}{|I_\tau|}\sum_{t \in I_\tau} u^{\mathsf{s}}(\mathbf{w}, k^t)$ for all the signaling schemes $\mathbf{w} \in W^{\circ}$ (Line 13). Notice that the set $W^{\circ} \subseteq W^{\star}$ is defined (as shown in Lemma 6) in order to be able to build estimators with the desired properties (i) and (ii).

As discussed in Section 6, the key insight that allows us to get the required estimates by using only a polynomial number of random samples is that the utilities to be estimated are *not* independent. This is because they depend on the frequencies of the receiver's actions during bock I_τ, which depend, in turn, on the frequencies of the receiver's types. Thus, the goal is to devise estimators for the frequencies of the receiver's types during each block I_τ. As an intuition, imagine that the sender commits to a signaling scheme such that each receiver's type best responds by playing a different action. Then, by observing the receiver's action, the sender gets to know the receiver's type with certainty. In general, for a given signaling scheme, there might be many different receiver's types that are better off playing the same action. In order to handle this problem and build the required estimates of the frequencies of the receiver's types, we use insights from the *bandit linear optimization* literature, and, in particular, we use the concept of *barycentric spanner* introduced by Awerbuch and Kleinberg [4].

For every block I_τ with $\tau \in [Z]$, we let $f_\tau : [0,1]^n \to \mathbb{R}$ be a function that, given a vector $\mathbf{x} = [x_1, \dots, x_n] \in [0,1]^n$, returns the sum of the number of times the receiver's types in $\mathcal{K}$ were active during block I_τ, weighted by the coefficients defined by the vector $\mathbf{x}$. Formally, the following definition holds:

$$f_\tau(\mathbf{x}) \coloneqq \sum_{k \in \mathcal{K}} x_k \sum_{t \in B_\tau} \mathbb{I}\{k^t = k\},$$

where $\mathbb{I}\{k^t = k\}$ is an indicator function that is equal to 1 if and only if it is the case that $k^t = k$, while it is 0 otherwise. Notice that, for a given $\tau \in [Z]$ and $k \in \mathcal{K}$, the term $\sum_{t \in B_\tau} \mathbb{I}\{k^t = k\}$ is a constant, and, thus, the function f_τ is linear. Intuitively, f_τ is the key element that allows us to connect the utilities that we need to estimate with the actual quantities we can estimate through the use of barycentric spanners.

The first crucial step is to restrict the attention to posteriors that can be induced with at least some ('not too small') probability. This ensures that our estimators have a limited range. Given a probability threshold $\sigma \in (0,1)$, we denote with $\Xi^{\circledcirc} \subseteq \Xi$ the set of posteriors that can be induced with probability at least σ by some signaling scheme. We can verify whether a given posterior $\boldsymbol{\xi} \in \Xi$ belongs to $\Xi^{\circledcirc}$ by solving an LP. Formally, $\boldsymbol{\xi} \in \Xi^{\circledcirc}$ if and only if the following set of linear equations admits a feasible solution $\mathbf{w} \in \Delta_\Xi$:

$$w_{\boldsymbol{\xi}} \geq \sigma \tag{10a}$$

$$\sum_{\boldsymbol{\xi} \in \Xi} w_{\boldsymbol{\xi}} \boldsymbol{\xi}_\theta = \mu_\theta \qquad \forall \theta \in \Theta. \tag{10b}$$

We define $\mathcal{R}$ as the set of all the tuples $\mathbf{a} = (a^k)_{k \in \mathcal{K}} \in \bigtimes_{k \in \mathcal{K}} \mathcal{A}$ for which there exists a posterior $\boldsymbol{\xi} \in \Xi^{\circledcirc}$ such that, for every receiver's type $k \in \mathcal{K}$, the action a^k specified by the tuple is a best response to $\boldsymbol{\xi}$ for type k. Formally:

$$\mathcal{R} \coloneqq \bigcup_{\boldsymbol{\xi} \in \Xi^{\circledcirc}} \left(b^1_{\boldsymbol{\xi}}, \dots, b^n_{\boldsymbol{\xi}}\right),$$

where we recall that $b^k_{\boldsymbol{\xi}}$ denotes the best response of type $k \in \mathcal{K}$ under posterior $\boldsymbol{\xi}$. Intuitively, $\mathcal{R}$ is the set of tuples of receiver's best responses which result from the posteriors that the sender can induce with probability at least σ. [10]

Given a tuple $\mathbf{a} = (a^k)_{k \in \mathcal{K}} \in \mathcal{R}$ and a receiver's action $a \in \mathcal{A}$, we denote with $\mathbb{I}_{(\mathbf{a}=a)} \in \{0,1\}^n$ an indicator vector whose k-th component is equal to 1 if and only if type $k \in \mathcal{K}$ plays action a in $\mathbf{a}$, *i.e.*, it holds $a^k = a$. Moreover, we define $\mathcal{X}$ as the set of all the indicators vectors; formally, $\mathcal{X} \coloneqq \left\{\mathbb{I}_{(\mathbf{a}=a)} \mid \mathbf{a} \in \mathcal{R}, a \in \mathcal{A}\right\}$.

Since the set $\mathcal{X}$ is a finite (and hence compact) subset of the Euclidean space $\mathbb{R}^n$, we can use the following proposition due to Awerbuch and Kleinberg [4] to introduce the *barycentric spanner* of $\mathcal{X}$.

Proposition 1 ([4], Proposition 2.2)**.** *If $\mathcal{X}$ is a compact subset of an n-dimensional vector space $\mathcal{V}$, then there exists a set $\mathcal{H} = \{\mathbf{h}^1, ..., \mathbf{h}^n\} \subseteq \mathcal{X}$ such that for all $\mathbf{x} \in \mathcal{X}$, $\mathbf{x}$ may be expressed as a linear combination of elements of $\mathcal{H}$ using coefficients in $[-1,1]$. That is, for all $\mathbf{x} \in \mathcal{X}$, there exists a vector of coefficients $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_n] \in [-1,1]^n$ such that $\mathbf{x} = \sum_{i \in [n]} \lambda_i \mathbf{h}^i$. The set $\mathcal{H}$ is called* barycentric spanner *of $\mathcal{X}$.*

[10] Let us remark that the sets $\Xi^{\circledcirc}$ and $\mathcal{R}$ depend on the given threshold $\sigma \in (0,1)$. In the following, for the ease of notation, we omit such dependence, as the actual value of σ that the two sets refer to will be clear from context.

In the following, we denote with $\mathcal{H} \coloneqq \{\mathbf{h}^1, ..., \mathbf{h}^n\} \subseteq \mathcal{X}$ a barycentric spanner of $\mathcal{X}$. Notice that, since each element $\mathbf{h} \in \mathcal{H}$ of the barycentric spanner belongs to $\mathcal{X}$ by definition, there exist a tuple $\mathbf{a} \in \mathcal{R}$ and a receiver's action $a \in \mathcal{A}$ such that $\mathbf{h}$ is equal to the indicator vector $\mathbb{I}_{(\mathbf{a}=a)}$. Moreover, by definition of $\mathcal{R}$, there exists a posterior $\boldsymbol{\xi} \in \Xi^{\circledcirc}$ such that the tuple of best responses $\left(b^1_{\boldsymbol{\xi}}, \dots, b^n_{\boldsymbol{\xi}}\right)$ coincides with $\mathbf{a}$.

Next, we describe how Algorithm 1 computes the required estmates. During the exploration phase of block I_τ with $\tau \in [Z]$, one iteration is devoted to each element $\mathbf{h} \in \mathcal{H}$ of the barycentric spanner, so as to get an estimate of $f_\tau(\mathbf{h})$. During such iteration, the algorithm plays a signaling scheme $\mathbf{w} \in \Delta_\Xi$ that is feasible for the LP defined by Constraints (10) where the posterior $\boldsymbol{\xi} \in \Xi^{\circledcirc}$ is that associated to $\mathbf{h}$. As a result, the set of all such signaling schemes can be used as $W^{\circledcirc}$ in Algorithm 1. Moreover, when the induced receiver's posterior is $\boldsymbol{\xi}$ and the receiver responds by playing action a, the algorithm sets a variable $p_\tau(\mathbf{h})$ to the value $\frac{1}{w_{\boldsymbol{\xi}}}$, otherwise $p_\tau(\mathbf{h})$ is set to 0.

The following lemma shows that the variables $p_\tau(\mathbf{h})$ computed by the algorithm during each block I_τ with $\tau \in [Z]$ are unbiased estimates of the values $f_\tau(\mathbf{h})$.

Lemma 4. *For any $\tau \in [Z]$ and $\mathbf{h} \in \mathcal{H}$, it holds $\mathbb{E}\left[p_\tau(\mathbf{h}) \cdot |I_\tau|\right] = f_\tau(\mathbf{h})$.*

Proof. First, recall that $p_\tau(\mathbf{h}) = \frac{1}{w_{\boldsymbol{\xi}}}$ if and only if during the iteration of exploration devoted to $\mathbf{h}$, the induced receiver's posterior is $\boldsymbol{\xi}$ and she/he best responds by playing a (otherwise, $p_\tau(\mathbf{h}) = 0$). Since the iteration is selected uniformly at random over the block I_τ and the sequence of receiver's types $\mathbf{k} = \{k^t\}_{t \in [T]}$ is chosen adversarially before the beginning of the game, we can conclude that also the receiver's type for that iteration is picked uniformly at random. Thus, $\mathbb{E}\left[p_\tau(\mathbf{h})\right] = \frac{1}{w_{\boldsymbol{\xi}}} \cdot w_{\boldsymbol{\xi}} \cdot \mathbb{P}\Big\{\text{randomly chosen type from } I_\tau \text{ best responds to } \boldsymbol{\xi} \text{ consistently with } \mathbf{h}\Big\}$, where by best responding consistently we mean that the type $k \in \mathcal{K}$ is such that $h_k = 1$, *i.e.*, she plays action a in $\mathbf{a}$. By using the definition of $f_\tau(\mathbf{h})$, we can write the following:

$$\mathbb{E}\left[p_\tau(\mathbf{h})\right] = \frac{\sum_{k \in \mathcal{K}: h_k = 1} f_\tau(\mathbf{e}^k)}{|I_\tau|} = \frac{f_\tau(\mathbf{h})}{|I_\tau|},$$

where $\mathbf{e}^k \in \mathbb{R}^n$ denotes an n-dimensional vector whose k-th component is 1, while others components are 0. □

For any $\mathbf{x} \in \mathcal{X}$, we let $\boldsymbol{\lambda}(\mathbf{x}) = [\lambda_1(\mathbf{x}), \dots, \lambda_n(\mathbf{x})] \in [-1, 1]^n$ be the vector of coefficients representing $\mathbf{x}$ with respect to basis $\mathcal{H}$. Formally, we can write $\mathbf{x} = \sum_{i \in [n]} \lambda_i(\mathbf{x}) \mathbf{h}^i$.

For any posterior $\boldsymbol{\xi} \in \Xi^{\circledcirc}$, let $\mathbf{a}[\boldsymbol{\xi}] \in \mathcal{R}$ be such that $\mathbf{a}[\boldsymbol{\xi}] = \left(b^1_{\boldsymbol{\xi}}, \dots, b^n_{\boldsymbol{\xi}}\right)$. Then, for each $\tau \in [Z]$, let us define

$$\tilde{u}^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi}) \coloneqq \sum_{a \in \mathcal{A}} \sum_{k \in \mathcal{K}} \lambda_k \left(\mathbb{I}_{\mathbf{a}[\boldsymbol{\xi}]=a}\right) p_\tau\left(\mathbf{h}^k\right) \sum_{\theta \in \Theta} \xi_\theta u^{\mathsf{s}}_\theta(a).$$

Letting $u^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi}) \coloneqq \frac{1}{|I_\tau|} \sum_{t \in \tau} u^{\mathsf{s}}(\boldsymbol{\xi}, k^t)$ be the sender's average utility achieved by inducing the receiver's posterior $\boldsymbol{\xi} \in \Xi^{\circledcirc}$ during each iteration of block I_τ with $\tau \in [Z]$, the following lemma shows that $\tilde{u}^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi})$ is an unbiased estimator of $u^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi})$, and, additionally, the range in which the estimator values lie is not to large.

Lemma 5. *For any posterior $\boldsymbol{\xi} \in \Xi^{\circledcirc}$ and $\tau \in [Z]$, it holds $\mathbb{E}\left[\tilde{u}^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi})\right] = u^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi})$. Moreover, $\tilde{u}^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi}) \in [-\frac{mn}{\sigma}, \frac{mn}{\sigma}]$.*

Proof. The first statement follows from the following relations:

$$\begin{aligned}
\mathbb{E}\left[\tilde{u}^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi})\right] &= \mathbb{E}\left[\sum_{a \in \mathcal{A}} \sum_{k \in \mathcal{K}} \lambda_k \left(\mathbb{I}_{\mathbf{a}[\boldsymbol{\xi}]=a}\right) p_\tau\left(\mathbf{h}^k\right) \sum_{\theta \in \Theta} \xi_\theta u^{\mathsf{s}}_\theta(a)\right] \\
&= \sum_{a \in \mathcal{A}} \sum_{k \in \mathcal{K}} \lambda_k \left(\mathbb{I}_{\mathbf{a}[\boldsymbol{\xi}]=a}\right) \mathbb{E}\left[p_\tau\left(\mathbf{h}^k\right)\right] \sum_{\theta \in \Theta} \xi_\theta u^{\mathsf{s}}_\theta(a) \\
&= \sum_{a \in \mathcal{A}} \sum_{\theta \in \Theta} \xi_\theta u^{\mathsf{s}}_\theta(a) \sum_{k \in \mathcal{K}} \lambda_k \left(\mathbb{I}_{\mathbf{a}[\boldsymbol{\xi}]=a}\right) \mathbb{E}\left[p_\tau\left(\mathbf{h}^k\right)\right] \\
&= \sum_{a \in \mathcal{A}} \sum_{\theta \in \Theta} \xi_\theta u^{\mathsf{s}}_\theta(a) \sum_{k \in \mathcal{K}} \lambda_k \left(\mathbb{I}_{\mathbf{a}[\boldsymbol{\xi}]=a}\right) \frac{f_\tau\left(\mathbf{h}^k\right)}{|I_\tau|} && \text{(By Lemma 4)} \\
&= \sum_{a \in \mathcal{A}} \sum_{\theta \in \Theta} \xi_\theta u^{\mathsf{s}}_\theta(a) \sum_{k \in \mathcal{K}} \frac{f_\tau\left(\mathbb{I}_{\mathbf{a}[\boldsymbol{\xi}]=a}\right)}{|I_\tau|} && \text{(By definition of } f_\tau) \\
&= u^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi}),
\end{aligned}$$

where the last equality holds by using again the definition of f_τ and re-arranging the terms.

As for the second statement, since $\lambda_k\left(\mathbb{I}_{\mathbf{a}[\boldsymbol{\xi}]=a}\right) \in [-1,1]$, $\sum_{\theta\in\Theta}\xi_\theta u^{\mathsf{s}}_\theta(a) \in [0,1]$, and $p_\tau\left(\mathbf{h}^k\right) \in \left[0,\frac{1}{\sigma}\right]$, it is easy to show that $\tilde{u}^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi}) \in [-\frac{mn}{\sigma},\frac{mn}{\sigma}]$. □

In the next lemma, we show that there always exists a best-in-hindsight signaling scheme that uses (*i.e.*, induces with positive probability) only a small number of posteriors. This is the final step needed to show that the estimators $\tilde{u}^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi})$ allow to compute slightly biased estimates of the utilities needed by the full-information algorithm.

Lemma 6. *Given a sequence of receiver's types $\mathbf{k} = \{k^t\}_{t\in[T]}$, there always exists a best-in-hindsight signaling scheme $\mathbf{w}^\star \in W^\star$ such that the set of posteriors it induces with positive probability $\left\{\boldsymbol{\xi}\in\Xi \mid w^\star_{\boldsymbol{\xi}} > 0\right\}$ has cardinality at most the number of states d.*

Proof. Notice that a best-in-hindsight signaling scheme $\mathbf{w}^\star \in W^\star$ can be computed by solving the following LP:

$$\begin{aligned}\max_{\mathbf{w}\in\Delta_\Xi} \quad & \sum_{t\in[T]}\sum_{\boldsymbol{\xi}\in\Xi} w_{\boldsymbol{\xi}} u^{\mathsf{s}}(\mathbf{w},k^t)\\ \text{s.t.} \quad & \sum_{\boldsymbol{\xi}\in\Xi} w_{\boldsymbol{\xi}}\xi_\theta = \mu_\theta \qquad\qquad \forall\theta\in\Theta.\end{aligned}$$

Since the LP has d equalities, it always admits an optimal basic feasible solution in which at most d variables $w_{\boldsymbol{\xi}}$ are greater than 0. This concludes the proof. □

Then, we define the W° used by Algorithm 1 as the set of signaling schemes $\mathbf{w}\in W^\star$ whose support is at most d, *i.e.*, it is the case that $|\{\boldsymbol{\xi}\in\Xi \mid w_{\boldsymbol{\xi}} > 0\}| \le d$. By definition of $W^\star$ and Lemma 6, it is easy to see that a best-in-hindsight signaling scheme is always guaranteed to be in the set W°.

Letting $\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w}) \coloneqq \sum_{\boldsymbol{\xi}\in\Xi^\circledcirc} w_{\boldsymbol{\xi}}\tilde{u}^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi})$ for every $\mathbf{w}\in W^\circ$ and $\tau\in[Z]$, the following lemma shows that each $\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})$ is a biased estimator of the sender's average utility $u^{\mathsf{s}}_{I_\tau}(\mathbf{w})$ in block I_τ, while also providing bounds on the bias and the range of the estimators. This final result allows us to effectively use the estimators $\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})$ defined above in Algorithm 1.

Lemma 7. *For any signaling scheme $\mathbf{w}\in W^\circ$ and $\tau\in[Z]$, it holds $u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) \ge \mathbb{E}\left[\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})\right] \ge u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) - d\sigma$. Moreover, it is the case that $\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w}) \in [-\frac{mn}{\sigma},\frac{mn}{\sigma}]$.*

Proof. By using Lemma 5, it is easy to check that the left inequality in the first statement holds:

$$u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) = \sum_{\boldsymbol{\xi}\in\Xi} w_{\boldsymbol{\xi}} u^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi}) \ge \sum_{\boldsymbol{\xi}\in\Xi^\circledcirc} w_{\boldsymbol{\xi}} u^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi}) = \sum_{\boldsymbol{\xi}\in\Xi^\circledcirc} w_{\boldsymbol{\xi}}\mathbb{E}\left[\tilde{u}^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi})\right] = \mathbb{E}\left[\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})\right].$$

Moreover, it is the case that:

$$\begin{aligned}\mathbb{E}\left[\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})\right] &= \sum_{\boldsymbol{\xi}\in\Xi^\circledcirc} w_{\boldsymbol{\xi}}\mathbb{E}\left[\tilde{u}^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi})\right]\\ &= \sum_{\boldsymbol{\xi}\in\Xi^\circledcirc} w_{\boldsymbol{\xi}} u^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi}) && \text{(By Lemma 5)}\\ &= u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) - \sum_{\boldsymbol{\xi}\in\Xi\setminus\Xi^\circledcirc} w_{\boldsymbol{\xi}} u^{\mathsf{s}}_{I_\tau}(\boldsymbol{\xi}) && \text{(By definition of } u^{\mathsf{s}}_{I_\tau}(\mathbf{w}))\\ &\ge u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) - \sum_{\boldsymbol{\xi}\in\Xi\setminus\Xi^\circledcirc} w_{\boldsymbol{\xi}} && \text{(Since } u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) \le 1)\\ &\ge u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) - \sum_{\boldsymbol{\xi}\in\Xi\setminus\Xi^\circledcirc} \sigma && \text{(By definition of } \Xi^\circledcirc\text{, it must be } w_{\boldsymbol{\xi}} < \sigma)\\ &\ge u^{\mathsf{s}}_{I_\tau}(\mathbf{w}) - d\sigma && \text{(Since } \mathbf{w}\in W^\circ)\end{aligned}$$

Finally, $\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w}) \in [-\frac{mn}{\sigma},\frac{mn}{\sigma}]$ follows from the fact that, by definition, $\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})$ is the weighted sum of quantities within the range $[-\frac{mn}{\sigma},\frac{mn}{\sigma}]$, with the weights sum being at most 1. □

D.3 Proof of Theorem 4

Theorem 4. *Given an online Bayesian persuasion problem with partial feedback, there exist $W^\circ \subseteq W^\star$, $W^{\circledcirc} \subseteq W^\star$, and estimators $\tilde{u}^{\mathsf{s}}_{I_\tau}(\mathbf{w})$ such that Algorithm 1 provides the following regret bound:*

$$R^T \leq O\left(\frac{nm^{2/3}d\log^{1/3}(mn+d)}{T^{1/5}}\right).$$

Proof. By setting $\sigma \coloneqq d^{-2/5}T^{-1/5}$, it is sufficient to run Algorithm 1 with estimators $u^{\mathsf{s}}_{I_\tau}(\mathbf{w})$ for every $\mathbf{w} \in W^\circ$ computed as previously described in this section. Thus, it holds $|W^{\circledcirc}| = n$ and $\eta = mnd^{2/5}T^{1/5}$. By Theorem 3, the following holds:

$$\begin{aligned}
R^T &\leq O\left(\frac{|W^{\circledcirc}|^{1/3}\eta^{2/3}log^{1/3}|W^\circ|}{T^{1/3}}\right) + O(\iota) \\
&= O\left(\frac{n^{1/3}\left(mnd^{2/5}T^{1/5}\right)^{2/3}\log^{1/3}|W^\circ|}{T^{1/3}}\right) + O\left(\frac{d}{d^{2/5}T^{1/5}}\right) \\
&= O\left(\frac{nm^{2/3}d^{4/15}\left(d\log\left(m^2n+d\right)\right)^{1/3}}{T^{1/5}}\right) + O\left(\frac{d^{3/5}}{T^{1/5}}\right) \\
&= O\left(\frac{nm^{2/3}d^{3/5}\log^{1/3}(mn+d)}{T^{1/5}}\right).
\end{aligned}$$

This concludes the proof. □