Learning to Design Information*

Ce Li† Tao Lin‡

October 21, 2025

Job Market Paper

Latest version available here.

Abstract

Information designers, such as large language models and online platforms, often do not know the subjective beliefs of their receivers or users. We construct learning algorithms enabling the designer to learn the receiver's belief through repeated interactions. Our learning algorithms are robust to the receiver's strategic manipulation of the learning process of the designer. We study regret relative to two benchmarks to measure the performance of the algorithms. The static benchmark is T times the single-period optimum for the designer under a known belief. The dynamic benchmark, which is stronger, characterizes global dynamic optimality for the designer under a known belief. Our learning algorithms allow the designer to achieve no regret against both benchmarks at fast convergence speeds of $O(\frac{\log^2 T}{T})$, significantly faster than other approaches in the literature.

*We are especially grateful to Bart Lipman for his detailed comments on this paper. We also thank Gerdus Benadè, Modibo Camara, Yiling Chen, Krishna Dasaratha, Maryam Farboodi, Kevin He, Hu Fu, Nicole Immorlica, Xudong Li, Yingkai Li, Chiara Margaria, Jawwad Noor, Juan Ortner, Roi Orzach, Pengyu Qian, Eitan Sapiro-Gheiler, David Simchi-Levi, Zhihao Tang, Haifeng Xu, Kai Hao Yang, and Jiawei Zhang for valuable discussion. This paper also benefits from audiences at EC'25 Workshop: Information Economics and Large Language Models, the World Congress of the Econometric Society, and the INFORMS Annual Meeting.

†Department of Economics, Boston University, email: celi@bu.edu

‡Microsoft Research, New England and the School of Engineering and Applied Sciences, Harvard University, email: tlin@g.harvard.edu

CONTENTS

1 INTRODUCTION

An information designer (the designer; she) sends signals about an unknown state to steer the action of a receiver (he) in her favor. Designing the optimal signaling strategy requires the designer to know the receiver's prior belief over states. Canonical economic models (e.g., Kamenica and Gentzkow, 2011) assume the parties share a common prior. In many applications, however, the designer has limited knowledge of the receiver's belief. In this paper, we design learning algorithms for the designer such that the designer learns the belief of a strategic receiver through repeated interactions and designs approximately optimal signaling schemes.

For example, consider a large language model (LLM) API platform (the designer) that provides a business user (the receiver) with access to LLMs through Application Programming Interfaces (APIs). The LLM API platform charges the user based on the amount of tokens consumed to handle the user's request. The request may be solvable or unsolvable (the state), which can only be observed by the platform. After the user submits a request, the platform observes its condition and decides how to signal to the user, e.g., recommending the user to proceed for more details or admitting its inability to solve the request. Upon receiving the signal, the user updates his belief about whether his request is solvable, and then takes an action to maximize his payoff, such as obeying the platform's recommendation to proceed or stopping the interaction. The platform always prefers that the user proceed to gain payment from the user for the token consumption, regardless of the state. On the user's side, his payoff is the highest when he pays for tokens that truly solve his request and the lowest when he pays for tokens that generate unwanted content, such as hallucinations.

The user has a *subjective* prior belief about the states, but the platform may not know that belief. As a result, the limited knowledge of the user's belief fundamentally *constrains* the platform from designing the optimal signaling scheme. In addition, the user may choose deceptive actions in some periods on purpose to *manipulate* the platform's learning of his prior belief and further the platform's design of the signaling scheme in favor of the user himself.

Motivated by such scenarios, we study an information design problem where the prior belief of the receiver is unknown to the designer and the receiver is strategic. Standard economic models, when tackling the problem of an unknown belief, would assume that the designer has a belief over possible prior beliefs of the receiver (e.g., Alonso and Câmara, 2016; Kolotilin et al., 2017). However, such an approach suffers from several concerns, such as computational hardness. A detailed discussion of the drawbacks of such an approach is presented in the literature review in detail.

Are there near-optimal signaling strategies for the designer when she does not know the prior belief of a *strategic* receiver? Our main contribution is to construct *approximately optimal* sig-

naling schemes which enable the designer to learn that subjective prior from repeated interactions with a strategic receiver. Moreover, our designer achieves such approximate optimality at efficient speeds. In addition, the designer's learning is robust to the receiver's strategic manipulation. Finally, such fast learning speeds are achieved against both a static benchmark and a dynamic benchmark.

Specifically, we study a *learning* model where an information designer (e.g., the LLM platform) learns to design signaling schemes over T periods based on historical interactions with a long-lived receiver (e.g., a business user). The designer does not know the belief of the receiver. In each period, the designer chooses a signaling scheme, which maps a state i.i.d. sampled from a true distribution over the states to a randomized signal. Upon receiving the signal, the receiver updates his belief about the state and takes an action. The payoffs of the two players are then realized. We design learning algorithms for the designer to improve the signaling schemes over time, such that (1) the designer's time-averaged payoff converges to *optimum benchmarks* and that (2) the convergences are *fast*, while (3) being robust to *strategic manipulation* by the receiver.

The rate at which the designer's payoff converges to the optimal objective depends on the learning algorithm she employs. We measure the performance of the learning algorithm by a general notion of *regret*: the difference between the designer's time-cumulative actual payoffs and an optimality benchmark, based on what the designer could achieve if she knew the receiver's belief. We first characterize different regret benchmarks for the problem of an unknown prior belief in a learning model. The *static* benchmark is T times the single-period optimum for the designer had the designer known the receiver's belief. The *static regret* is defined against the static benchmark. This allows us to relate our results to the results of the canonical information design models in Kamenica and Gentzkow (2011).

In addition, we consider a *dynamic* benchmark that measures the designer's global dynamic optimality under a known belief, which is stronger than the static benchmark. The *dynamic regret* is then defined against the dynamic benchmark. A learning algorithm is called *no-regret* if its regret grows sublinearly with time (alternatively, its time-averaged regret converges to zero). The rate at which the time-averaged regret declines reflects the learning algorithm's efficiency. We aim to design learning algorithms that not only achieve no regret against the two benchmarks, but do so as efficiently as possible, yielding fast convergence to optimality under the unknown belief while being robust to the receiver's strategic manipulation.

Our first main result is an $O(\log^2 T)$ upper bound on the *static regret*. Our second main result is also an $O(\log^2 T)$ regret upper bound on the *dynamic regret*. We design two learning algorithms to ensure that, for any problem instance, the designer's regrets for T periods, relative to the *static benchmark* and *the dynamic benchmark*, respectively, are both no more than a constant times $\log^2 T$ for all T. Both regret upper bounds imply convergence rates of $O(\frac{\log^2 T}{T})$ for the designer's regret to

converge to zero, which are significantly faster than other approaches in the literature, as discussed later.

Both of our two learning algorithms contain two phases. The first phase is the learning of the receiver's belief, which has the same learning logic under two regret benchmarks. The learning of the belief is based on *binary search*. The key idea is to learn the unknown belief from the receivers' actions. To see the intuition, consider two states and two actions, where the receiver prefers different actions in the two states. For the moment, suppose the receiver is myopic in the sense that he does not consider the future in choosing his action. Then, the designer can compute a cutoff belief such that the receiver is indifferent between the two actions. The actual action chosen by the receiver then reveals whether the receiver's belief is below or above the cutoff. By observing the receiver's action, the designer gradually narrows down the range of possible values for the prior. That process does not stop until the designer obtains a sufficiently accurate estimate of the prior belief. Our formal analysis shows how to generalize that intuition to multiple states and multiple actions.

How does the designer do this when the receiver is not myopic? The algorithm repeats the same signaling scheme that sends the designated signal for a fixed number of periods. First, the repetition and the receiver's discount factor ensure that the receiver cannot gain more by acting deceptively in a single round. The same signaling scheme will be repeatedly used multiple times, which reduces the incentive for the receiver to misbehave. Second, the algorithm only cares about the action taken by the receiver when he first observes that certain signal. In that way, the update for the signaling scheme used in the next series of repetition will be influenced by the receiver's actions in the current series of repetition to the least extent. Thus, the information collected about the receiver's belief stays reliable to the largest extent, helping the designer learn the receiver's belief effectively.

The second phase, where our learning algorithms for the two benchmarks differ, is *robustification*. The signaling scheme that is optimal for the designer under the belief estimate is converted into another signaling scheme that is approximately optimal under the receiver's belief. The motivation is that even if two prior beliefs are very close, the signaling scheme optimal under one belief may perform arbitrarily badly under the other, leaving no guarantee on the designer's payoff. However, the robustification varies across benchmarks. Our first learning algorithm uses *static* robustification, where the robustification process is performed on the signaling scheme optimal for the belief estimate *once* to obtain a robustified signaling scheme for being repeatedly used for future periods.

Our second learning algorithm applies a *dynamic robustification* to the designer's globally optimal strategy, where the signaling schemes in different periods are carefully chosen, so that the receiver at any period is still incentivized to participate in the future. The challenge is that the

static robustification may backfire: the static robustification only increases the *persuasiveness* of the signaling scheme, namely, increasing the positive gap of the expected payoff from the recommended action relative to that from any other actions. However, the static robustification may lower the receiver's *ex ante* (before a recommendation is sent) expected payoff from obeying the recommendations in the future. Thus, after an action is recommended by a statically robustified signaling scheme, the receiver's incentive to obey that current recommendation becomes stronger, but his *ex ante* payoff from obeying recommendations in *future periods* may get worse, potentially violating the receiver's participation constraint.

To overcome this, we propose a new type of robustification, *payoff-improving robustification*, which weakly increases the receiver's *ex ante* payoff from being obedient. We then take the designer's optimal dynamic signaling strategy, given her belief estimate, and apply the payoff-improving robustification procedure to each period's signaling scheme in that dynamic strategy. We show that the resulting dynamic strategy is persuasive for the receiver and approximately optimal for the designer.

The superiority of both of our learning algorithms is reflected by the logarithmic upper bounds on their regrets: $O(\log^2 T)$. That means that the designer's average payoff converges to the optimality benchmarks at a fast speed, in the order of $O(\frac{\log^2 T}{T})$. In contrast, an intuitive method is to use an *empirical estimation*: counting the state revealed at the end of each period for obtaining an empirical estimate of the state distribution. The empirical estimation does not work in our problem, since the true distribution over the states is not the same as the receiver's subjective belief. Even if the empirical estimation could be applied, it would suffer a regret of at least $O(\sqrt{T})$ due to sampling error, which is exponentially worse than the $O(\log^2 T)$ regret bound of our learning algorithms. Our algorithms circumvent the limitation of empirical estimation and achieve an exponential improvement by actively learning the receiver's belief from his actions, instead of learning from sampled states passively.

1.1 Related Work

First, this paper contributes to the literature on robust information design under uncertainty, especially when the designer lacks knowledge of the receiver's prior belief. Classical models such as Bayesian persuasion (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2016) and cheap talk (Crawford and Sobel, 1982) typically assume a common prior shared by the designer and the receiver. Recent work relaxes this by introducing uncertainty over the prior. One approach assumes the designer holds a belief over the unknown prior (e.g., Kolotilin et al., 2017), but computing the optimal signaling scheme is generally intractable (Hossain et al., 2024). Another approach, rooted in robust mechanism design, has the designer optimize against a worst-case prior in a given set (e.g., Hu and Weng, 2021; Dworczak and Pavan, 2022; Kosterina, 2022; Dworczak and Kolotilin,

2024; Ball and Kattwinkel, 2025). These papers typically study a single-shot game, limiting the designer's ability to learn over time. Of course, when the designer has the opportunity to learn, one expects she may achieve a much higher payoff. In contrast, we design learning algorithms that infer the prior, achieving near optimality quickly while avoiding computational hardness.

Departing from the standard approach for addressing the robustness, Li and Lin (2025) study how to *learn* the belief of a sequence of *short-lived* receivers from their actions. However, if a receiver interacts with the platform repeatedly, thus becoming long-lived, then he will be incentivized to be strategic. Thus, he may take actions to convey deceptive information about his belief, manipulating the platform's learning in a direction favorable to them but unfavorable to the platform.

There are also papers that address robustness from other perspectives. Mathevet, Perego, and Taneva (2020), Morris, Oyama, and Takahashi (2024), and Ziegler (2020) study the designer's strategy from an adversarial perspective in one-shot games. Li and Norman (2021) analyze sequential games from the adversarial perspective but maintain a common prior. In these papers, either a common prior is assumed or the model is one-shot, whereas our paper tackles both a learning process and an unknown prior. Ball (2023) consider a dynamic information design problem in which the beliefs of the designer and the receiver are different. However, their model uses persistent states while we consider i.i.d. states. More importantly, they consider both players to be equally patient, while we consider a patient designer and an impatient receiver. The different patience levels in our paper have crucial influence on their dynamic strategies and the designer's regret benchmarks. Besides, they assume more structure on the dynamic process and the players' payoffs, while our work keep the dynamic process and the payoff functions general without further structures.

Our work also contributes to the literature on online Bayesian persuasion, which studies learning when some game parameters are unknown. Castiglioni et al. (2020), Castiglioni et al. (2021), and Feng, Tang, and Xu (2022) consider unknown receiver utilities; Zu, Iyer, and Xu (2021) and Wu et al. (2022) focus on unknown priors; and Bacchiocchi et al. (2024) allow both utilities and priors to be unknown. Previous approaches to unknown priors rely on empirical estimation and incur $\Omega(\sqrt{T})$ regret due to sampling error. This option is not available here as the designer needs to learn the receiver's belief, not the true distribution. We bypass this by using best responses to infer the prior more efficiently, achieving $O(\log^2 T)$ regret. Camara, Hartline, and Johnsen (2020) also study information design with learning agents but drop distributional assumptions and allow adversarial state sequences; their regret notion therefore differs from ours, which assumes i.i.d. states from an unknown objective prior.

Our work also connects to the line of works on misspecified learning (Ba, 2023; Bohren and Hauser, 2021; Frick, Iijima, and Ishii, 2023; Esponda, Pouzo, and Yamamoto, 2021; Fudenberg,

Lanzani, and Strack, 2021; Fudenberg, Romanyuk, and Strack, 2017) and on the long-run outcomes of misspecified Bayesian agents who cannot obtain the true distribution over states (e.g., Tversky and Kahneman (1973), Rabin and Schrag (1999), Fudenberg and Lanzani (2023), and He and Libgober (2021)). Those works study belief convergence and actions in general learning environments. However, we consider a receiver whose belief is already the convergence outcome of his misspecified learning. We neglect the process by which such a belief is formed, focusing instead on how the designer interacts with a receiver whose belief has already converged.

2 Models and Preliminaries

2.1 Models

Single-period game. We first define a single-period game of information design (Bayesian persuasion) with subjective prior beliefs. There is a finite set of states of the world Ω. The information designer and the receiver have subjective prior beliefs $\mu_D, \mu_R \in \Delta(\Omega)$ over the states, respectively, where $\Delta(\Omega)$ denotes the set of probability distributions over Ω. We think of μ_D as the true distribution over the states and μ_R as a subjective prior belief over the states, which may not be the same as the true distribution. The designer is better informed about the true distribution from access to sufficient state observation data, so the designer is assumed to know the true distribution. The receiver's prior μ_R can be different from the true distribution, which may be the convergence outcome of the learning process of a misspecified belief (e.g., Fudenberg, Romanyuk, and Strack (2017)).

The receiver has a finite set of actions A. The designer and the receiver have payoff functions $u, v : A \times \Omega \to [0, 1]$, respectively. This single-period game is summarized by an *instance* $\mathcal{I} = (\Omega, A, u, v, \mu_D, \mu_R)$.

At the beginning of the game, the designer announces a *signaling scheme* $\pi : \Omega \to \Delta(S)$, which maps each state to a probability distribution over a finite set of signals S. Let $\Pi = \Omega \to \Delta(S)$ be the set of signaling schemes. The receiver believes that the state is drawn according to the distribution μ_R. Thus, after receiving a signal $s \sim \pi(\cdot|\omega)$ sampled according to state ω and the signaling scheme π, the receiver obtains his posterior belief $\frac{\mu_R(\omega)\pi(s|\omega)}{\sum_{\omega' \in \Omega} \mu_R(\omega')\pi(s|\omega')}, \forall \omega \in \Omega$, and takes an action in response. Let $\rho : S \to \Delta(A)$ be a possibly randomized single-period strategy of the receiver. Under a pair of strategies (π, ρ), the expected payoffs of the designer and the receiver in

the single-period game are, respectively,

$$U(\pi, \rho; \mu_D) = \sum_{\omega \in \Omega} \mu_D(\omega) \sum_{s \in S} \pi(s|\omega) \sum_{a \in A} \rho(a|s) u(a, \omega), \tag{1}$$

$$V(\pi, \rho; \mu_R) = \sum_{\omega \in \Omega} \mu_R(\omega) \sum_{s \in S} \pi(s|\omega) \sum_{a \in A} \rho(a|s) v(a, \omega). \tag{2}$$

We say the receiver *myopically best responds*[1] if he maximizes his single-period payoff by choosing a strategy

$$\rho^* \in \arg\max_{\rho: S \to \Delta(A)} V(\pi, \rho)$$

or, equivalently, an action $\rho^*(s) \in \arg\max_{a \in A} \mu_R(\omega)\pi(s|\omega)v(a, \omega)$ given each signal $s \in S$ based on his posterior belief. With a myopically best-responding receiver, the optimal utility of the designer is

$$U^*_{\mathrm{BP}}(\mathcal{I}) := \max_{\pi \in \Pi} \max_{\rho \in \arg\max_{S \to \Delta(A)} V(\pi, \rho; \mu_R)} U(\pi, \rho; \mu_D). \tag{3}$$

We call $U^*_{\mathrm{BP}}(\mathcal{I})$ the *Bayesian persuasion benchmark*.

T-period learning process. We consider the T-period repetitions of the above single-period game with an instance $\mathcal{I} = (\Omega, A, u, v, \mu_D, \mu_R)$. Importantly, the designer now does not know the receiver's prior belief μ_R but she knows other parameters u, v, μ_D. In particular, the designer *learns* to design signaling schemes using an *algorithm* $\mathcal{G}$ which, at every period t, maps the history $H^{(t-1)} \in \Pi^{(t-1)} \times S^{(t-1)} \times A^{(t-1)}$ to a signaling scheme $\pi^{(t)} \in \Pi$ for the current period. In response to the learning algorithm $\mathcal{G}$, the receiver chooses a T-period strategy $\phi = (\phi^{(t)})_{t=1}^T$ where each $\phi^{(t)}$ maps the history $H^{(t-1)} \in \Pi^{(t-1)} \times S^{(t-1)} \times A^{(t-1)}$ and the current period's signaling scheme $\pi^{(t)} \in \Pi$ to a strategy $\rho^{(t)} : S \to \Delta(A)$ that specifies a randomized action for each possible signal. Then, the pair of strategies $(\mathcal{G}, \phi)$ will generate a stochastic sequence of $(\pi^{(t)}, \rho^{(t)})_{t=1}^T$ of the designer's signaling schemes and the receiver's single-period strategies. We assume that the receiver is *impatient* in the sense that he has a discount factor $\gamma \in (0, 1)$, so his total discounted expected payoff is

$$\mathcal{V}_T(\mathcal{G}, \phi; \mu_R) \;=\; \mathbb{E}_{(\pi^{(t)}, \rho^{(t)})_{t=1}^T \sim (\mathcal{G}, \phi)} \left[\sum_{t=1}^{T} \gamma^t \cdot V(\pi^{(t)}, \rho^{(t)}; \mu_R) \right]. \tag{4}$$

[1] We use "myopic best response" to distinguish from the best response of a strategic receiver in the T-period game.

The designer, on the other hand, is *patient*, that is, having a discount factor of one. The designer aims to maximize her total expected payoff

$$\mathcal{U}_T(\mathcal{G}, \phi; \mu_D) \;=\; \mathbb{E}_{(\pi^{(t)}, \rho^{(t)})_{t=1}^T \sim (\mathcal{G}, \phi)} \left[\sum_{t=1}^{T} U(\pi^{(t)}, \rho^{(t)}; \mu_D) \right]. \tag{5}$$

We consider a *strategic* (or long-lived) receiver who best responds to the designer's learning algorithm $\mathcal{G}$. Formally, given the algorithm $\mathcal{G}$ and the belief μ_R, the receiver chooses a globally optimal strategy

$$\phi^*(\mathcal{G}, \mu_R) = \arg\max_{\phi} \mathcal{V}_T(\mathcal{G}, \phi; \mu_R).$$

Then, the designer obtains her total expected payoff

$$\mathcal{U}_T(\mathcal{G}, \phi^*(\mathcal{G}, \mu_R); \mu_D).$$

Remark 1. *The assumption that the designer is more patient than the receiver is commonly seen in real-world problems. As a high-level intuition, the designer represents a principal, a platform, or a social planner, etc. Thus, the designer is long-lived and patient. The receiver may be an agent or a user, who can be long-lived but impatient. We discuss how the results change if the designer is impatient in Section 6. Technically, if the designer is impatient (with a discount factor less than one), then she will only care about the early periods instead of the whole learning process. In that way, the designer will obtain a total payoff sublinear in T, thus making her achieve a regret sublinear in T (to be formally defined below) trivially. What's even worse, the designer may have already started to neglect her continuing payoffs before finishing learning the receiver's belief. If the receiver is patient (with his discount factor $\gamma = 1$), then the designer cannot achieve sublinear regret regardless of the learning algorithm, as we show in Section 6. If the receiver is patient, then their payoffs under the designer's dynamic benchmark will not oscillate over time, as shown in Section 4.*

Benchmarks and regrets. The *regret* of the designer under instance $\mathcal{I} = \{\Omega, A, u, v, \mu_D, \mu_R\}$ is the difference between a benchmark $\mathcal{U}_T^*(\mathcal{I})$ and the actual payoff $\mathcal{U}_T$ that the designer obtains by using the algorithm $\mathcal{G}$ over the T periods. In particular, we will consider two benchmarks for defining the regret.

The first benchmark is the **static benchmark** denoted by $U_{\mathrm{BP}}^*(\mathcal{I})$, which is T times the single-period Bayesian persuasion benchmark at a problem instance $\mathcal{I}$.

$$\mathcal{U}_T^*(\mathcal{I}) := T \cdot U_{\mathrm{BP}}^*(\mathcal{I}).$$

The static benchmark is a natural benchmark to consider, especially for measuring how good the performance of our learning algorithm is compared to the Bayesian persuasion optimality in the canonical work (Kamenica and Gentzkow, 2011), in which the designer knows the receiver's prior μ_R.

We call the regret against the static benchmark **static regret**:

$$\mathrm{Reg}^{\mathrm{static}}(T;\mathcal{G},\mathcal{I}) \;=\; \mathcal{U}_T^*(\mathcal{I}) \;-\; \mathcal{U}_T(\mathcal{G},\phi^*(\mathcal{G},\mu_R);\mu_D). \tag{6}$$

The second benchmark, the **dynamic benchmark**, is the maximal T-period payoff the designer could obtain had the designer known the receiver's belief μ_R and the receiver best responded over the T periods. We denote this by $U_T^{**}(\mathcal{I})$ at a problem instance $\mathcal{I}$.

$$\mathcal{U}_T^{**}(\mathcal{I}) := \max_{T\text{-period strategy } \sigma} \mathcal{U}_T(\sigma,\phi^*(\sigma,\mu_R);\mu_D),$$

where σ is a T-period strategy of the designer that designs signaling schemes based on historical information and the receiver's belief μ_R.[2]

We call the regret against the dynamic benchmark **dynamic regret**:

$$\mathrm{Reg}^{\mathrm{dynamic}}(T;\mathcal{G},\mathcal{I}) \;=\; \mathcal{U}_T^{**}(\mathcal{I}) \;-\; \mathcal{U}_T(\mathcal{G},\phi^*(\mathcal{G},\mu_R);\mu_D). \tag{7}$$

As we will show in Section 4, the dynamic benchmark $\mathcal{U}_T^{**}(\mathcal{I})$ is always weakly larger than the static benchmark $\mathcal{U}_T^*(\mathcal{I})$. Thus, achieving no regret against the dynamic benchmark automatically implies no regret against the static benchmark.

A regret *upper bound* is a performance guarantee of an algorithm $\mathcal{G}$ in the following way: for any problem instance $\mathcal{I}$, that is, for any unknown belief μ_R, the regret incurred by the algorithm $\mathcal{G}$ grows no faster than some function of T, i.e.,

$$\mathrm{Reg}(T;\mathcal{G},\mathcal{I}) \;\le\; C_{\mathcal{I}} \cdot f(T) \;=\; O(f(T)), \quad \forall T,\ \forall \mathcal{I},$$

where $C_{\mathcal{I}}$ is a constant depending on $\mathcal{I}$ but not on T. Here, $f(T)$ represents the *worst-case* rate at which the regret can grow. An algorithm $\mathcal{G}$ with regret upper bound $O(f(T))$ is *no-regret* if $f(T)$ grows at a rate that is sublinear with T, that is,

$$\lim_{T\to\infty} \frac{f(T)}{T} \to 0.$$

[2]We assume tie-breaking in favor of the designer if the receiver has multiple best responses. However, our results guarantee that obedience is the unique best response for the receiver.

For example, the worst-case rate $f(T)$ can be $f(T) = \sqrt{T}$ or $f(T) = \log T$.

2.2 Regularity Assumptions

We present two regularity assumptions used throughout the paper. First, we assume that the beliefs μ_D, μ_R have full support and the designer knows the minimum probability $p_0 > 0$ of any state. This probability p_0 will be used by the designer for designing the algorithms later.

Assumption 1 (Full-support priors)**.** *Both players' prior beliefs $\mu_D, \mu_R \in \Delta(\Omega)$ have full support and there exists $p_0 \in (0, 1)$ such that the designer knows*

$$\min_{\omega \in \Omega} \mu_R(\omega) \geq p_0 > 0.$$

Assumption 2 (Regularity of receiver's utility)**.** *There is no weakly dominated action for the receiver.*

Assumption 2 implies that there exists a positive number $D > 0$ such that, for every action $a \in A$ of the receiver, there exists a belief $\eta_a \in \Delta(\Omega)$ on which action a is *strictly* better than any other action by a margin of D:

$$\mathbb{E}_{\omega \sim \eta_a}[v(a, \omega)] \;\geq\; \mathbb{E}_{\omega \sim \eta_a}[v(a', \omega)] + D, \quad \forall a' \in A \setminus \{a\}. \tag{8}$$

The designer knows D since she knows the receiver's utility function v.

3 Learning to Achieve No Regret Against the Static Benchmark

We will design a learning algorithm for the designer to achieve a regret upper bounded by $O(\log^2 T)$ against the static benchmark. The learning algorithm contains two parts: one is the learning of the receiver's belief, which will be used again in the later section for dynamic regret. The other is a static robustification process for obtaining a near-optimal signaling scheme from the learned belief.

3.1 Efficient Learning of the Strategic Receiver's Belief

The information designer learns the receiver's belief μ_R through observing his actions.

Denote by $a^*_\omega = \arg\max_{a \in A} v(a, \omega)$ the receiver's optimal action in state $\omega \in \Omega$. We present a natural and mild assumption on the receiver's utility function v as follows.

Assumption 3 (Unique optimal action)**.** *The optimal action a^*_ω for the receiver in each state $\omega \in \Omega$ is unique.*

Since the number of states is finite, Assumption 3 immediately implies that there exists a positive constant $G > 0$ such that

$$\forall \omega \in \Omega, \forall a \in A \setminus \{a^*_\omega\}, v(a^*_\omega, \omega) - v(a, \omega) > G > 0.$$

Since the designer knows the utility functions, she knows this constant G.

We learn the receiver's belief $\mu_R \in \Delta(\Omega)$ by learning the ratio $\frac{\mu_R(\omega_i)}{\mu_R(\omega_j)}$ of μ_R for *any* pair of states $\omega_i, \omega_j \in \Omega$. First, we show how to learn that ratio when the receiver has different optimal actions $a^*_{\omega_i}, a^*_{\omega_j}$ under the pair of states ω_i, ω_j, respectively. We call such a pair of states *distinguishable*, and assume that there exists at least one such pair of states.

Assumption 4 (Distinguishable states)**.** *There exist at least one pair of states $\omega_i, \omega_j \in \Omega$ whose corresponding receiver-optimal actions are different: $a^*_{\omega_i} \neq a^*_{\omega_j}$.*

Assumption 4 makes the designer's learning problem non-trivial. If Assumption 4 does not hold, then the receiver would have the same optimal action a^* across all states. Thus, regardless of the signaling scheme and the belief, the receiver would always take action a^*, and the expected payoff of the designer would be a constant, leaving the designer zero regret trivially.

Algorithm 1 shows how to estimate the receiver's prior belief ratio $\frac{\mu_R(\omega_i)}{\mu_R(\omega_j)}$ for any pair of distinguishable states ω_1 and ω_2 based on *binary search*. The designer uses a signaling scheme π that sends a certain signal s_0 only under states ω_i and ω_j, that is, $\pi(s_0|\omega) > 0$ for $\omega = \omega_i, \omega_j$ and $\pi(s_0|\omega) = 0$ for $\omega \neq \omega_i, \omega_j$. Then the receiver will believe that the state must be either ω_i or ω_j upon receiving signal s_0. In particular, the ratio of the receiver's posterior belief about those two states is

$$\frac{\mu(\omega_j|s_0)}{\mu(\omega_i|s_0)} = \frac{\mu_R(\omega_j)}{\mu_R(\omega_i)} \frac{\pi(s_0|\omega_j)}{\pi(s_0|\omega_i)}, \tag{9}$$

which depends on the receiver's prior belief μ_R and the signaling scheme π used by the designer. The receiver takes an action based on this posterior belief to maximize his total payoff from the current and future periods. Based on the receiver's action, the designer then updates the signaling scheme for the next period. That process iterates until the designer obtains an estimate of the receiver's prior belief ratio $\frac{\mu_R(\omega_i)}{\mu_R(\omega_j)}$ with a desired accuracy.

To learn the belief ratio of a strategic receiver, it is useful to first understand how to learn that ratio if the receiver best responds to the signaling scheme π *myopically*, that is, if the receiver maximizes his payoff only in the current period. If the receiver's prior belief ratio $\frac{\mu_R(\omega_j)}{\mu_R(\omega_i)}$ is sufficiently small, then his posterior belief ratio (9) will also be small. Thus, the receiver will believe that the state is ω_i with high probability and take the corresponding optimal action $a^*_{\omega_i}$. On the other hand, if the receiver's prior belief ratio $\frac{\mu_R(\omega_j)}{\mu_R(\omega_i)}$ is very large, then he will believe that the state is ω_j with

high probability and take the corresponding optimal action $a^*_{\omega_j}$. In between, there is a range for the ratio $\frac{\mu_R(\omega_j)}{\mu_R(\omega_i)}$, where the receiver might take neither action but a completely different action.

Therefore, given the signaling scheme π and a pair of distinguishable states, there exists a *cutoff threshold* $c(\pi, \omega_i, \omega_j)$ on the receiver's prior belief ratio such that the receiver is indifferent in taking action $a^*_{\omega_i}$ versus some other action $\tilde{a}_{\omega_j,\omega_i}$, which might or might not be $a^*_{\omega_j}$. Thus, the receiver takes action $\tilde{a}_{\omega_j,\omega_i}$ when his prior belief ratio $\frac{\mu_R(\omega_j)}{\mu_R(\omega_i)}$ is above the cutoff threshold $c(\pi, \omega_i, \omega_j)$. As explained above, $\tilde{a}_{\omega_j,\omega_i}$ is not necessarily $a^*_{\omega_j}$. We call such an action $\tilde{a}_{\omega_j,\omega_i}$ the *first indifferent action* for the receiver under the signaling scheme π and a pair of distinguishable states ω_i, ω_j. The cutoff threshold $c(\pi, \omega_i, \omega_j)$ is the prior belief ratio $\frac{\mu(\omega_j)}{\mu(\omega_i)}$ that solves the following indifference equation under the experimented signaling scheme π:

$$\mu(\omega_j)\pi(s_0|\omega_j)\Big(v(a^*_{\omega_i}, \omega_j) - v(\tilde{a}_{\omega_j,\omega_i}, \omega_j)\Big) + \mu(\omega_i)\pi(s_0|\omega_i)\Big(v(a^*_{\omega_i}, \omega_i) - v(\tilde{a}_{\omega_j,\omega_i}, \omega_i)\Big) = 0$$

$$\implies \quad c(\pi, \omega_i, \omega_j) = \frac{\mu(\omega_j)}{\mu(\omega_i)} = \frac{\pi(s_0|\omega_i)}{\pi(s_0|\omega_j)} \cdot \frac{v(a^*_{\omega_i}, \omega_i) - v(\tilde{a}_{\omega_j,\omega_i}, \omega_i)}{v(\tilde{a}_{\omega_j,\omega_i}, \omega_j) - v(a^*_{\omega_i}, \omega_j)}.$$

The designer knows the threshold $c(\cdot, \cdot, \cdot)$ based on her knowledge about the receiver's payoff function v.

That enables the designer to use a *binary search* algorithm to identify a precise range containing the receiver's prior belief ratio $\frac{\mu_R(\omega_j)}{\mu_R(\omega_i)}$. In particular, by observing whether the receiver takes action $a^*_{\omega_i}$ or not, the designer is able to tell whether the receiver's actual prior belief ratio $\frac{\mu_R(\omega_j)}{\mu_R(\omega_i)}$ is below or above the cutoff threshold $c(\pi, \omega_i, \omega_j)$. That enables the designer to adjust the signaling scheme and the corresponding cutoff threshold through iterations over time. By multiple iterations, the designer narrows down the range of $\frac{\mu_R(\omega_j)}{\mu_R(\omega_i)}$ to length ε in $O(\log \frac{1}{\varepsilon})$ steps.

We then study the case when the receiver is *strategic*. Being long-lived incentivizes the receiver to take deceptive actions to manipulate the designer's learning of his belief μ_R. We have two crucial steps to make the designer's learning *robust* to such strategic manipulation. First, our binary search algorithm repeats the same signaling scheme $\pi^{(k)}$ in each iteration k for an additional number of periods. With the discount factor, the repetition makes the receiver not gain much by acting deceptively in one round, since the same signaling scheme will recur, and the effect of deception will not occur immediately, but will occur only after some periods. Thus, any future benefit from manipulation by deviating from the best response will be too small compared to the immediate loss due to not best responding myopically. In that way, the repetition of the same signaling scheme together with the influence of the discount factor incentivizes the receiver to act truthfully.

Second, we record the receiver's action only the first time when the signal s_0 appears and ignore his actions in later repetitions. If we were to take into account the receiver's actions in later repetitions, in particular the repetitions near the end of the current iteration of binary search, then

the receiver would have an increasingly strong incentive to manipulate, because there are fewer and fewer repetitions left in the current iteration and the receiver's actions in those tail periods will have immediate effects on the designer's update of the signaling scheme for the next iteration, say $k+1$. Recording the receiver's action only in the first relevant repetition reduces the receiver's manipulation incentives and ensures the correctness of our prior learning algorithm.

Algorithm 1, based on the idea of binary search, formally shows how to learn the receiver's prior belief ratio $\frac{\mu_R(\omega_i)}{\mu_R(\omega_j)}$ for any pair of distinguishable states ω_i, ω_j when the receiver acts strategically.

Algorithm 1: Estimating Prior Ratio for Distinguishable States by Binary Search

Input : two states ω_i, ω_j whose receiver-optimal actions are different.

Parameter: a termination criterion $\tau > 0$, and an integer $m \geq 1$.

Output : an estimation $\hat{\rho}$ of the ratio $\frac{\mu_R(\omega_i)}{\mu_R(\omega_j)}$.

Let $k = 0$, $\ell^{(0)} = 0$, $r^{(0)} = \frac{1}{Gp_0}$.

Let $a^*_{\omega_i} = \arg\max_{a \in A} v(a, \omega_i)$. Let $\tilde{a}$ be the first indifferent action $\tilde{a}_{\omega_j, \omega_i}$.

while $r^{(k)} - \ell^{(k)} > \tau$ **do**

  Let $q = \frac{\ell^{(k)} + r^{(k)}}{2}$.

  Let s_0 be an arbitrary signal in S; let $\pi^{(k)}$ be a signaling scheme that satisfies $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} = q$:
  - if $q \leq 1$, let $\pi^{(k)}(s_0|\omega_j) = q$, $\pi^{(k)}(s_0|\omega_i) = 1$;
  - if $q > 1$, let $\pi^{(k)}(s_0|\omega_j) = 1$, $\pi^{(k)}(s_0|\omega_i) = 1/q$.
  For $\omega \in \Omega \setminus \{\omega_i, \omega_j\}$, $\pi^{(k)}(s_0|\omega) = 0$. For $s \in S \setminus \{s_0\}$, $\pi^{(k)}(s|\omega)$ can be arbitrary.

  Repeat $\pi^{(k)}$ for multiple periods until signal s_0 is realized. Let $a^{(k)}$ be the action taken by the receiver when s_0 is realized.

  Repeat $\pi^{(k)}$ for additional m periods, while ignoring the receiver's actions.

  if $a^{(k)} = a^*_{\omega_i}$ **then**

    Let $\ell^{(k+1)} = q, r^{(k+1)} = r^{(k)}$.

  else

    Let $r^{(k+1)} = q, \ell^{(k+1)} = \ell^{(k)}$.

  end

  $k = k + 1$.

end

Output $\hat{\rho} = \ell^{(k)} \cdot \frac{v(\tilde{a}, \omega_j) - v(a^*_1, \omega_j)}{v(a^*_{\omega_i}, \omega_i) - v(\tilde{a}, \omega_i)}$.

The performance of Algorithm 1 is characterized by Lemma 1, which shows that a good estimate of $\frac{\mu_R(\omega_i)}{\mu_R(\omega_j)}$ can be obtained within a very short amount of time: Algorithm 1 requires at most $O(\log T)$ periods to reach a precision $\tau = O(\frac{1}{T})$.

Lemma 1. *The output of Algorithm 1, $\hat{\rho}$, satisfies $\left|\hat{\rho} - \frac{\mu_R(\omega_i)}{\mu_R(\omega_j)}\right| \leq \frac{1}{G}\left(\tau + \frac{\gamma^m}{(1-\gamma)G^3 p_0^3}\right) \leq \frac{1}{Gp_0}\left(\tau + \frac{\gamma^m}{(1-\gamma)G^3 p_0^3}\right)\frac{\mu_R(\omega_i)}{\mu_R(\omega_j)}$, and Algorithm 1 terminates in at most $(\frac{1}{p_0} + m)\log_2 \frac{1}{Gp_0\tau}$ periods in expectation.*

Proof. See Appendix B.1. □

A key part in the proof of Lemma 1 is the following Lemma 2, which shows that the strategic receiver is incentivized to perform consistently with single-period best responses if the signaling scheme ratio estimate $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)}$ differs from the targeted ratio $\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$ by a positive distance $\iota = O(\gamma^m)$.

Lemma 2. *Choose $\iota = \frac{\gamma^m}{(1-\gamma)G^2 p_0^2}$. In Line 6 of Algorithm 1, when signal s_0 is sent,*

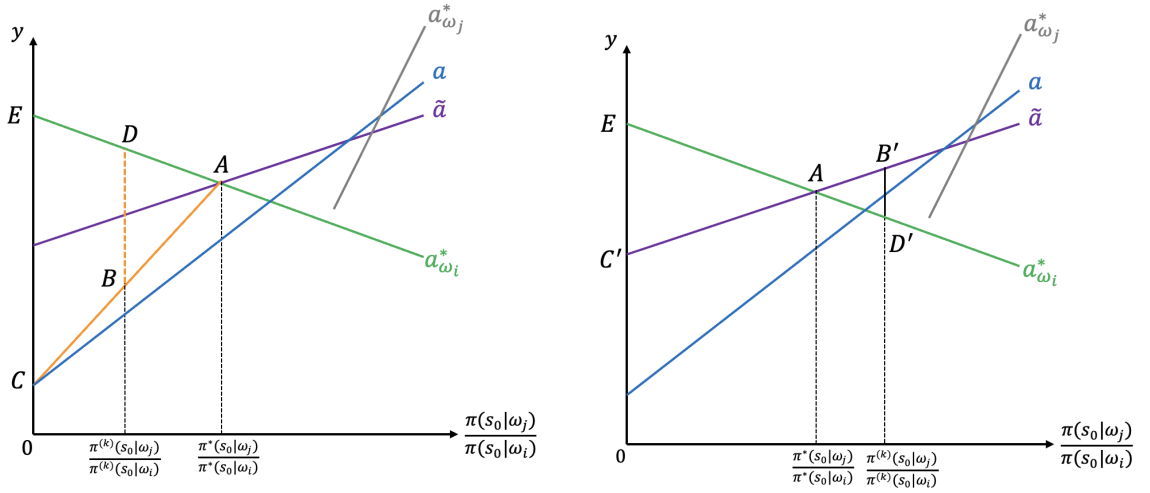
- *if $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} < (1-\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$, then the receiver will take action $a^*_{\omega_i}$;*
- *if $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} > (1+\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$, then the receiver will not take action $a^*_{\omega_i}$.*

To see the intuition of Lemma 2, note that when recommended by a signal s_0 according to Algorithm 1, the strategic receiver is incentivized to take actions truthfully because his loss from deviation is larger than his gain from deviation. Regarding the gain, the receiver's accumulated gain from deviating is upper bounded by a positive number due to his discounted factor. Regarding the loss, the receiver's loss is lower bounded by a positive number that depends on how close our ratio estimate $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)}$ is from the targeted ratio $\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$, as well as the positive margin G by the unique optimal action assumption (Assumption 3). Then, at a controlled distance of ratio estimate $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)}$ from ratio $\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$, the receiver's loss lower bound is strictly larger than his gain upper bound, making him incentivized not to deviate from the recommendation.

As an illustration, Figure 4a shows the case of $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} < (1-\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$, in which the strategic receiver is incentivized to take action $a^*_{\omega_i}$. The payoff brought by action $a^*_{\omega_i}$ is the highest for that case. Figure 4b shows the case of $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} > (1-\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$, in which the strategic receiver is incentivized not to take action $a^*_{\omega_i}$. The payoff brought by action $a^*_{\omega_i}$ is no longer the highest.

Proof. See Appendix B.2. □

What if the receiver-optimal actions of a pair of states are the same? Algorithm 2 deals with such a case, allowing the information designer to estimate the prior ratio between any pair of states. The idea of Algorithm 2 is that, if states ω_i and ω_j have the same receiver-optimal actions, then both of them must be distinguishable from one state in a distinguishable pair of states, say $\omega_k \in \Omega$, by Assumption 4. Thus, we obtain the estimate of $\frac{\mu_R(\omega_i)}{\mu_R(\omega_j)}$ by estimating the ratios $\frac{\mu_R(\omega_i)}{\mu_R(\omega_k)}$ and $\frac{\mu_R(\omega_j)}{\mu_R(\omega_k)}$ separately.



(a) The case of $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} < (1-\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$ (b) The case of $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} > (1+\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$

Figure 1: Normalized utility of the receiver for different actions. The x-coordinate is the signaling ratio $\frac{\pi(s_0|\omega_j)}{\pi(s_0|\omega_i)}$. Each line corresponds to an action a, with equation $y_a(x) = v(a,\omega_i) + \frac{\mu_R(\omega_j)}{\mu_R(\omega_i)} \cdot x \cdot v(a,\omega_j)$.

Algorithm 2: Estimating Belief Ratio for any Pair of States (ω_i, ω_j)

Parameter: $\tau > 0$ and $m \geq 1$

1 If ω_i and ω_j are distinguishable, i.e., $a^*_{\omega_i} \neq a^*_{\omega_j}$, then run Algorithm 1 on ω_i and ω_j.

2 Otherwise, i.e., $a^*_{\omega_i} = a^*_{\omega_j}$, find $\omega_k \in \Omega$ such that $a^*_{\omega_k} \neq a^*_{\omega_i} = a^*_{\omega_j}$. Run Algorithm 1 to obtain an estimate $\hat{\rho}_{ik}$ for $\frac{\mu_R(\omega_i)}{\mu_R(\omega_k)}$ and an estimate $\hat{\rho}_{jk}$ for $\frac{\mu_R(\omega_j)}{\mu_R(\omega_k)}$. Return $\hat{\rho}_{ij} = \frac{\hat{\rho}_{ik}}{\hat{\rho}_{jk}}$.

Full algorithm for learning the belief μ_R. After showing how to learn the belief ratio for any pair of states, we now present the full algorithm in Algorithm 3 for the designer to learn the receiver's belief μ_R. Algorithm 3 first uses Algorithm 2 to learn the ratio $\frac{\mu_R(\omega_i)}{\mu_R(\omega_1)}$ for each state ω_i, $i = 2, \ldots, |\Omega|$ and state ω_1, and then reconstructs the estimate of the unknown belief from those ratio estimates.

Algorithm 3: Learning the Belief of the Strategic Receiver

Parameter: desired estimation accuracy $\varepsilon > 0$

1 Let $\tau = \frac{Gp_0^3\varepsilon}{24|\Omega|}$, $m = \left\lceil \frac{1}{\ln(1/\gamma)} \ln\left(\frac{24|\Omega|}{(1-\gamma)G^4 p_0^6 \varepsilon}\right) \right\rceil$.

2 For every state $\omega_i \in \Omega$, $i = 2, \ldots, |\Omega|$, apply Algorithm 2 to state pair (ω_i, ω_1) with parameters τ and m to learn the ratio $\frac{\mu_R(\omega_i)}{\mu_R(\omega_1)}$, obtaining ratio estimate $\hat{\rho}_{i1}$.

3 Return belief estimate $\hat{\mu} \in \Delta(\Omega)$ by $\begin{cases} \hat{\mu}(\omega_1) = 1/\left(1 + \sum_{i=2}^{|\Omega|} \hat{\rho}_{i1}\right); \\ \hat{\mu}(\omega_i) = \hat{\rho}_{i1}\hat{\mu}(\omega_1), \text{ for } i = 2, \ldots, |\Omega|. \end{cases}$

Algorithm 3 gives a belief estimate $\hat{\mu}$ with an ℓ_1 distance precision ε at $\|\hat{\mu} - \mu_R\|_1 \leq \varepsilon$ within a short running time of $O(\log^2 \frac{1}{\varepsilon})$ periods.

Lemma 3. *Given* $0 < \varepsilon \leq 1$,

- *the belief estimate* $\hat{\mu}$ *from Algorithm 3 satisfies* $\|\hat{\mu} - \mu_R\|_1 \leq \varepsilon$*;*
- *the expected running time of Algorithm 3 is at most* $2|\Omega|\left(\frac{1}{p_0} + \frac{1}{1-\gamma}\ln\left(\frac{24|\Omega|}{(1-\gamma)G^4 p_0^6 \varepsilon}\right)\right)\log_2 \frac{24|\Omega|}{G^2 p_0^4 \varepsilon}$ *periods.*

Proof. See Appendix B.3. □

3.2 Static Robustification of Signaling Schemes

Computing the signaling scheme optimal for a belief. In general, given any prior belief $\mu_0 \in \Delta(\Omega)$ of the receiver, assumed to be known to the designer, the optimal signaling scheme in the static Bayesian persuasion problem can be computed easily.

In particular, given any prior belief $\mu_0 \in \Delta(\Omega)$ known to be the receiver's belief, there exists a signaling scheme π_0 that is *direct* and *persuasive* (Kamenica and Gentzkow, 2011) under μ_0: a signaling scheme $\pi : \Omega \to \Delta(A)$ is *direct* if it maps each state to a probability distribution over *actions*; a direct signaling scheme π_0 is *persuasive* if all actions recommended by π_0 are optimal for the receiver under prior μ_0: $\sum_\omega \mu_0(\omega)\pi_0(a|\omega)\big[v(a,\omega) - v(a',\omega)\big] \geq 0, \forall a, a' \in A$. Denote by $\mathrm{Pers}(\mu_0)$ the set of persuasive signaling schemes given the receiver's belief $\mu_0 \in \Delta(\Omega)$:

$$\mathrm{Pers}(\mu_0) \coloneqq \Big\{\pi : \Omega \to \Delta(A) \;\Big|\; \sum_{\omega\in\Omega} \mu_0(\omega)\pi(a|\omega)\big[v(a,\omega) - v(a',\omega)\big] \geq 0,\ \forall a, a' \in A\Big\}, \quad (10)$$

where

$$\sum_{\omega\in\Omega} \mu_0(\omega)\pi_0(a|\omega)\big[v(a,\omega) - v(a',\omega)\big]\ \forall a, a' \in A$$

is the *persuasiveness* of the signaling scheme π under the receiver's belief μ_0.

The signaling scheme optimal for the designer with μ_D under her knowledge of the receiver's belief μ_0 is solved tractably from the following linear program

$$\pi^*(\mu_0) = \operatorname*{arg\,max}_{\pi\in\mathrm{Pers}(\mu_0)} \sum_{\omega\in\Omega} \mu_D(\omega) \sum_{a\in A} \pi(s|\omega)u(a,\omega). \quad (11)$$

Thus, if the designer knows the receiver's belief μ_R, then the designer will obtain her optimal signaling scheme π^* for μ_R exactly by the linear program 11. However, after the learning phase, the designer has a learned belief $\hat{\mu}$, which is only an approximation for the receiver's prior belief μ_R.

Static robustification. After the first phase of learning, we have obtained an estimate $\hat{\mu}$ of the receiver's prior belief μ_R. However, even when the estimate $\hat{\mu}$ is very close to the prior belief μ_R, the signaling scheme $\hat{\pi}$ optimal for the designer under the estimate $\hat{\mu}$ may not be persuasive or even perform arbitrarily badly under μ_R. The reason is that even though a signal s may induce two close posterior beliefs from two close prior beliefs $\hat{\mu}$ and μ_R, respectively, the optimal actions for the receiver under those two posterior beliefs may be different, which may lead to an arbitrarily low payoff for the designer under the receiver's prior belief μ_R.

We propose designing a signaling scheme π using *static robustification*. In general, we convert any *direct* signaling scheme π under $\hat{\mu}$ to another signaling scheme $\tilde{\pi}$, such that $\tilde{\pi}$, when applied to a targeted prior belief (e.g., the receiver's prior μ_R), will have a weakly higher persuasiveness compared to the case when that typical signaling scheme π is applied to the belief estimate $\hat{\mu}$. Recall that the persuasiveness of a direct signaling scheme under a prior belief measures how any recommended action from that signaling scheme is weakly better than any other unrecommended action. That is, here we have for any $a, a' \in A$,

$$\sum_{\omega} \mu_R(\omega)\tilde{\pi}(a|\omega)[v(a,\omega) - v(a',\omega)] \geq \sum_{\omega} \hat{\mu}(\omega)\pi(a|\omega)[v(a,\omega) - v(a',\omega)].$$

Moreover, the static robustification does not have much influence on the designer's payoff. In particular, if what is used for the static robustification is not a typical signaling scheme π but the signaling scheme $\hat{\pi} = \pi^*(\hat{\pi})$ that is optimal for the designer under the belief estimate $\hat{\mu}$, then $\hat{\pi}$ will be converted to another signaling scheme that guarantees approximate optimality for the designer. By being static, the robustification process is applied to the signaling scheme $\hat{\pi}$ once, and the same statically robustified signaling scheme $\tilde{\pi}$ is used repeatedly in each period for the robustification phase till the end.

The effect of our static robustification is summarized in Lemma 4. We construct the static robustification signaling scheme $\tilde{\pi}$ by taking a mixture of three signaling schemes: the direct signaling scheme π under $\hat{\mu}$, the signaling scheme that induces buffer beliefs η_a for each $a \in A$ (by Assumption 2), and the signaling scheme that fully reveals states. Let $\delta \in (0, \frac{1}{2})$ be the *robustification strength*, which measures how strongly "good" signaling schemes (the signaling scheme that induces buffer beliefs η_a for each $a \in A$ and the full revelation signaling schemes) influence the robustification. In particular, the weights of the mixture are roughly $1 - 2\delta$, δ, and δ, respectively. The full-revelation signaling scheme in the mixture ensures *Bayesian plausibility*.

Lemma 4 (Static Robustification)**.** *Assume Assumptions 1 and 2. Choose $\varepsilon \leq \frac{p_0^2 D}{4}$. Given the belief estimate $\hat{\mu}$ and the robustification strength $\delta \in [\frac{2\varepsilon}{p_0 D}, \frac{p_0}{2}]$, any direct signaling scheme π can be converted into another direct signaling scheme $\tilde{\pi}$ such that*

- *(improved persuasiveness for the receiver) for any prior beliefs μ and $\hat{\mu}$ with $\|\mu - \hat{\mu}\|_1 \leq \varepsilon$,*

for any recommendation $a \in A$ from $\tilde{\pi}$, the persuasiveness of $\tilde{\pi}$ and π under the resulted posterior beliefs $\mu(\cdot|a,\tilde{\pi})$, $\hat{\mu}(\cdot|a,\tilde{\pi})$ are close:

$$\mathbb{E}_{\omega\sim\mu(\cdot|a,\tilde{\pi})}\big[v(a,\omega)-v(a',\omega)\big] \geq \min\big\{\mathbb{E}_{\omega\sim\hat{\mu}(\cdot|a,\pi)}\big[v(a,\omega)-v(a',\omega)\big]+\delta D-\tfrac{2\varepsilon}{p_0},\ 0\big\},\ \forall a' \in A\backslash\{a\},$$

and, in particular, if the direct signaling scheme π is persuasive for $\hat{\mu}$, then its robustification result $\tilde{\pi}$ is persuasive for μ;

- *(tiny influence for the designer) when the receiver obeys the recommendation, the designer's payoff from the robustified signaling scheme $\tilde{\pi}$, compared to that from the direct signaling scheme π, satisfies*
$$U(\tilde{\pi},\rho_{\text{ob}};\mu_D) \geq U(\pi,\rho_{\text{ob}};\mu_D) - \frac{3\delta}{p_0},$$
where ρ_{ob} is the receiver's obedience strategy that puts weight one on the recommended action;

- *(close signaling schemes) the robustification result $\tilde{\pi}$ and the direction signaling scheme π are close:*
$$\|\tilde{\pi}(\cdot|\omega) - \pi(\cdot|\omega)\|_1 \leq \frac{4\delta}{p_0^2},\ \forall\omega\in\Omega.$$

To see the high-level idea, by the distance between $\hat{\mu}$ and μ_R at $\|\hat{\mu}-\mu_R\|_1 \leq \varepsilon$, the designer's optimal signaling scheme $\hat{\pi}$ under $\hat{\mu}$ will *not* be persuasive under the receiver's belief μ_R by a margin that is only at most $O(\frac{\varepsilon}{p_0})$. The signaling scheme that induces the belief η_a for each $a \in A$ is strictly persuasive for the receiver by a margin of D by Assumption 2. The full revelation signaling scheme is weakly persuasive. Thus, the persuasiveness of the mixture, that is, the signaling scheme $\tilde{\pi}$, is

$$D\cdot\delta + 0\cdot\delta - (1-2\delta)\cdot O(\frac{\varepsilon}{p_0}) \geq 0$$

by a choice of the robustification strength $\delta = O(\frac{\varepsilon}{p_0 D})$. As a result, the loss of the designer is at most

$$2\delta\cdot O(\frac{1}{p_0}) = O(\frac{\varepsilon}{p_0^2 D}).$$

In contrast to previous works that apply a robustifying process to the signaling scheme optimal for the belief estimate (Zu, Iyer, and Xu, 2021) or to signaling schemes persuasive for the belief estimate (Li and Lin, 2025), our static robustification is applied to *any direct* signaling schemes, thus being the most general. If we apply Lemma 4 to the designer's optimal signaling scheme $\hat{\pi}$ under the belief estimate $\hat{\mu}$ with the robustification strength $\delta = \frac{2\varepsilon}{p_0^2 D}$, then we will obtain an approximately optimal signaling scheme for the actual prior μ_R of the receiver. This specific application is summarized in Corollary 1.

Corollary 1 (Robustification for the optimal signaling scheme)**.** *With* $\|\hat{\mu} - \mu_R\|_1 \le \varepsilon \le \frac{p_0^2 D}{4}$, *the designer's optimal signaling scheme* $\hat{\pi}$ *under belief* $\hat{\mu}$ *can be converted into another signaling scheme* $\tilde{\pi}$ *that is persuasive under belief* μ_R *and* $\frac{6\varepsilon}{p_0^2 D}$*-optimal for the designer.*

See details of Lemma 4 in Appendix B.4.

3.3 Full Learning Algorithm with $O(\log^2 T)$ Regret

Finally, we combine the prior-learning algorithm and the static robustification to obtain a signaling scheme $\tilde{\pi}$ that is $O(\varepsilon)$-approximately optimal for the designer and persuasive for the receiver under belief μ_R. Using the robustified signaling scheme $\tilde{\pi}$ in each of the remaining periods thus incurs a small regret. The total regret of Algorithm 4 is formally characterized in Theorem 1.

Algorithm 4: Learning Algorithm for the Static Benchmark

Parameter: $\varepsilon > 0$

1 Run Algorithm 3 with parameter ε to obtain an estimate $\hat{\mu}$ of receiver's belief. Let T_0 be the number of periods taken.

2 Apply Corollary 1 to $\hat{\mu}$ with parameter ε to obtain a signaling scheme $\tilde{\pi}$. Use $\tilde{\pi}$ for the remaining $T - T_0$ periods.

Theorem 1. *Assume Assumptions 1, 2, 3, 4. Choose* $\varepsilon = \frac{p_0^2 D}{4T}$*. The static regret of Algorithm 4 is at most*

$$\mathrm{Reg}^{\mathrm{static}}(T; \textit{Algorithm 4}, \mathcal{I}) \le O\Big(\frac{1}{1-\gamma}\log^2 T\Big). \tag{12}$$

Proof. By Lemma 3, we have $\|\hat{\mu}-\mu_R\|_1 \le \varepsilon$ and $\mathbb{E}[T_0] \le 2|\Omega|\big(\frac{1}{p_0}+\frac{1}{1-\gamma}\ln(\frac{24|\Omega|}{(1-\gamma)G^4 p_0^6 \varepsilon})\big)\log_2 \frac{24|\Omega|}{G^2 p_0^4 \varepsilon}$. According to Corollary 1, the constructed signaling scheme $\tilde{\pi}$ is persuasive for μ_R and is $\frac{6\varepsilon}{p_0^2 D}$-approximately optimal for the designer. Thus, the regret of Algorithm 4 satisfies

$$\begin{aligned}
\mathrm{Reg}^{\mathrm{static}}(T; \mathcal{G}, \mathcal{I}) &\le \underbrace{\mathbb{E}[T_0 \cdot 1]}_{\text{regret in the first } T_0 \text{ periods}} + \underbrace{(T - \mathbb{E}[T_0])\frac{6\varepsilon}{p_0^2 D}}_{\text{regret in the remaining periods}} \\
&\le 2|\Omega|\Big(\frac{1}{p_0} + \frac{1}{1-\gamma}\ln\Big(\frac{24|\Omega|}{(1-\gamma)G^4 p_0^6 \varepsilon}\Big)\Big)\log_2 \frac{24|\Omega|}{G^2 p_0^4 \varepsilon} + T\frac{6\varepsilon}{p_0^2 D}
\end{aligned}$$

Choosing $\varepsilon = \frac{p_0^2 D}{4T}$ (which satisfies the condition in Corollary 1), we obtain

$$\begin{aligned}
\mathrm{Reg}^{\mathrm{static}}(T; \mathcal{G}, \mathcal{I}) &\le 2|\Omega|\Big(\frac{1}{p_0} + \frac{1}{1-\gamma}\ln\Big(\frac{96|\Omega|T}{(1-\gamma)G^4 p_0^8 D}\Big)\Big)\log_2 \frac{96|\Omega|T}{G^2 p_0^6 D} + \frac{3}{2} \\
&= O\Big(\frac{1}{1-\gamma}\log^2 T\Big).
\end{aligned}$$

□

To see the intuition of Theorem 1, recall the algorithm has two phases: (i) learning the receiver's belief μ_R for an accuracy ε, and (ii) robustification, which repeatedly uses a robustified signaling scheme that is guaranteed persuasive for the receiver's belief μ_R and approximately optimal for the designer. Phase (i) runs binary search algorithms to keep updating the experimented signaling scheme, which takes at most $O(\log(\frac{1}{\varepsilon}))$ distinct experimented signaling schemes. Further, for each distinct experimented signaling scheme, making the obedience incentive strong enough for a strategic receiver requires an additional repetition sub-phase of length $m = O\big(\frac{\log(\frac{1}{\varepsilon})}{1-\gamma}\big)$, contributing another $\log(\frac{1}{\varepsilon})$. Thus, we get the $O\big(\frac{1}{1-\gamma}\log^2(\frac{1}{\varepsilon})\big)$ exploration cost formalized in Lemma 3. Phase (ii) robustifies the designer's optimal signaling scheme under the estimate $\hat{\mu}$. That preserves persuasiveness for all beliefs within a distance of ε of $\hat{\mu}$, thus also for the receiver's belief μ_R, while degrading the designer's payoff by only $O(\frac{\varepsilon}{p_0^2 D})$ per period (Corollary 1). Choosing $\varepsilon = \frac{p_0^2 D}{4T}$ results in a loss of $T \cdot O(\frac{\varepsilon}{p_0^2 D}) = O(1)$ for phase (ii) (exploitation), so the regret is dominated by phase (i) (exploitation), yielding $\mathrm{Reg}^{\mathrm{static}}(T) = O\big(\frac{1}{1-\gamma}\log^2 T\big)$ as claimed in Theorem 1.

We add a final brief remark on Algorithm 4 and Theorem 1. In the first phase, the designed signaling scheme used in binary searches may be *not* persuasive. The designer's primary goal is to gather informative responses from the receiver's actions to accurately estimate the prior through binary search during that phase. Thus, achieving persuasiveness and optimality at each step is secondary to this learning objective.

4 Characterization of the Dynamic Benchmark

This section characterizes the dynamic benchmark $\mathcal{U}_T^{**}(\mathcal{I})$, which is the optimal payoff obtainable by the designer if she knows the receiver's belief μ_R. That is, it is the designer's optimal payoff in a T-period repeated Bayesian persuasion problem with full commitment power. In Section 4.1, we show that a certain class of grim-trigger-style strategies are optimal for the designer, that the dynamic benchmark is weakly larger than the static benchmark, and that the designer's optimal strategy can be obtained through dynamic programming. In Section 4.2, we present examples to illustrate the dynamic benchmark and the optimal strategy.

4.1 Optimal Dynamically Direct + Punishment (DDP) Strategy

Optimal DDP strategy: dynamic benchmark characterization We show that the designer's dynamic benchmark $\mathcal{U}_T^{**}(\mathcal{I})$ can be achieved by a class of T-period strategies with a *grim-trigger* property: the designer recommends actions to the receiver until the receiver deviates from the recommendation, and provides no information if the receiver deviates. We call such a class of T-period strategies "dynamically direct + punishment (DDP)" strategies.

Definition 1 (DDP strategy)**.** *A T-period strategy σ of the designer is a* dynamically direct + punishment (DDP) strategy *if σ*

- *(dynamically direct) uses a history-dependent direct signaling scheme if the receiver has never deviated from past recommendations,*
- *(punishment) and reveals no information (i.e., always sending one signal regardless of the state) if the receiver has deviated in the past.*

We now introduce two important properties of DDP strategies. The first property is *history independence*, which means that the signaling scheme in each period does not depend on what happened in the past, as long as the receiver has been obedient since the beginning.

Definition 2 (History Independence)**.** *A DDP strategy σ is* history-independent *if the direct signaling scheme $\pi_\sigma^{(t)}$ for each period t depends on the period number t but not on history, as long as the receiver has never deviated from past recommendations.*

The second property is *dynamic persuasiveness*.

Definition 3 (Dynamic Persuasiveness)**.** *A DDP strategy $\sigma = (\pi_\sigma^{(t)})_{t=1}^T$ is* dynamically persuasive *for the receiver if the receiver's T-period obedient strategy*

$$\phi_{\text{ob}} = \textit{obey the action recommendation in every period } t \in \{1, \dots, T\} \textit{ given any history}$$

is a best response to σ.

If the direct signaling scheme in each period of a DDP strategy is all persuasive in the per-period Bayesian persuasion game, then obedience is clearly a best response for the receiver, and the DDP strategy is dynamically persuasive. However, a dynamically persuasive DDP strategy may include some direct signaling schemes that are *not* persuasive in the static game. We will provide such examples in Section 4.2.

The first result of this subsection shows that the designer's dynamic benchmark $\mathcal{U}_T^{**}(\mathcal{I})$ can be achieved by a history-independent and dynamically persuasive DDP strategy σ^*:

Proposition 1 (Optimality of DDP Strategy)**.** *The dynamic benchmark $\mathcal{U}_T^{**}(\mathcal{I})$ can be achieved by a pair of strategies $(\sigma^*, \phi_{\text{ob}})$, where $\sigma^* = (\pi_{\sigma^*}^{(t)})_{t=1}^T$ is a history-independent and dynamically persuasive DDP strategy of the designer, and ϕ_{ob} is the T-period obedient strategy of the receiver.*

Proof.

The proof of Proposition 1 has two steps. First, we characterize the *dynamic revelation principle* (in a similar spirit to Myerson, 1986) in the following Lemma 5, which shows that the designer's $T-$period (global) optimal utility $\mathcal{U}_T^{**}(\mathcal{I})$ can always be achieved by a history-dependent dynamically persuasive DDP strategy. See Appendix C.1 for the proof of Lemma 5.

Lemma 5 (Dynamic revelation principle)**.** *There exists an optimal strategy $\sigma^* = (\pi^{(t)}_{\sigma^*})_{t=1}^T$ for the designer that is a history-dependent dynamically persuasive DDP strategy, that is, $\mathcal{U}_T^{**}(\mathcal{I}) = \mathcal{U}_T(\sigma^*, \phi_{\text{ob}}; \mu_D)$ and $\phi_{\text{ob}} = \phi^*(\sigma^*, \mu_R)$.*

Next, we show in Lemma 6 that history-dependence is not needed. At a high level, this is because the Bayesian persuasion problems at different periods are independent, as the states are i.i.d. sampled. As long as the receiver has not deviated in the past, his actions will not contain information that is valuable for the designer. Thus, the designer does not need to adjust signaling schemes based on the irrelevant history. The formal proof of Lemma 6 is in Appendix C.2.

Lemma 6. *There exists an optimal dynamically persuasive DDP strategy that is history-independent.*

The above two lemmas together prove Proposition 1. □

Computing the optimal DDP strategy (dynamic benchmark) Proposition 1 significantly simplifies the designer's optimization problem. We can write the dynamic benchmark $\mathcal{U}_T^{**}(\mathcal{I})$ as an optimization problem over history-independent DDP strategies subject to the dynamic persuasiveness constraint.

We now study the dynamic persuasiveness constraint in detail. For a history-independent DDP strategy $\sigma = (\pi^{(t)}_\sigma)_{t=1}^T$, *dynamic persuasiveness* requires that for each period t, the sequence of ongoing (period t included) direct signaling schemes $(\pi^{(t)}_\sigma, \dots, \pi^{(T)}_\sigma)$ in $\sigma = (\pi^{(t)}_\sigma)_{t=1}^T$ should satisfy the following: for any action a recommended by $\pi^{(t)}_\sigma$ with positive probability $P_{\mu_R, \pi^{(t)}_\sigma}(a) = \sum_{\omega\in\Omega} \mu_R(\omega)\, \pi^{(t)}_\sigma(a \mid \omega) > 0$, for any $a' \in A$,

$$\begin{aligned} &\gamma^t \sum_{\omega\in\Omega} \frac{\mu_R(\omega)\pi^{(t)}_\sigma(a|\omega)}{P_{\mu_R,\pi^{(t)}_\sigma}(a)} v(a,\omega) + \sum_{t'=t+1}^{T} \gamma^{t'} V(\pi^{(t')}_\sigma, \rho_{\text{ob}}; \mu_R) \\ \geq\ &\gamma^t \sum_{\omega\in\Omega} \frac{\mu_R(\omega)\pi^{(t)}_\sigma(a|\omega)}{P_{\mu_R,\pi^{(t)}_\sigma}(a)} v(a',\omega) + \sum_{t'=t+1}^{T} \gamma^{t'} V_{\text{uninformed}}(\mu_R), \end{aligned} \tag{13}$$

where $V_{\text{uninformed}}(\mu_R) = \max_{a\in A} \sum_{\omega\in\Omega} \mu_R(\omega) v(a,\omega)$ is the receiver's payoff from being punished. The left-hand side of Eq. (13) is the receiver's payoff of obedience from period t to T given the action a recommended at period t, while the second line is the receiver's payoff from period t to T by deviating to a' at period t and then being punished in the future.

An equivalent way to write the dynamic persuasiveness constraint uses the receiver's posterior belief. In particular, denote by $\mu_R(\omega|a, \pi^{(t)}_\sigma) = \frac{\mu_R(\omega)\pi^{(t)}_\sigma(a|\omega)}{P_{\mu_R,\pi^{(t)}_\sigma}(a)}$ the posterior belief of state ω given action a sent by signaling scheme $\pi^{(t)}$ under belief μ_R, Then, Eq. (13) can be rewritten as follows.

For each period t, $\forall a, a' \in A$,

$$\gamma^t \mathbb{E}_{\omega \sim \mu_R(\cdot|a,\pi_\sigma^{(t)})}\big[v(a,\omega) - v(a',\omega)\big] + \sum_{t'=t+1}^{T} \gamma^{t'} \big(V(\pi_\sigma^{(t')}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R)\big) \geq 0. \quad (14)$$

Formally, the *optimization problem* for the dynamic benchmark $\mathcal{U}_T^{**}(\mathcal{I})$ is

$$\begin{aligned} \mathcal{U}_T^{**}(\mathcal{I}) \;=\; & \max_{\text{history-independent DDP strategy } \sigma = (\pi_\sigma^{(t)})_{t=1}^T} && \mathcal{U}_T(\sigma; \phi_{\text{ob}}, \mu_D) \;=\; \sum_{t=1}^{T} U(\pi_\sigma^{(t)}, \rho_{\text{ob}}; \mu_D) \\ & \text{subject to} && \sigma \text{ satisfies (13) or (14).} \end{aligned} \quad (15)$$

Lemma 7 shows that the dynamic benchmark is always weakly larger than the static benchmark. Moreover, the dynamic benchmark is strictly larger than the static benchmark in some instances, as we will show in Section 4.2.

Lemma 7. *Dynamic benchmark is weakly larger than static benchmark:* $\mathcal{U}_T^{**}(\mathcal{I}) \geq \mathcal{U}_T^{*}(\mathcal{I})$.

Proof. Let π_{BP}^* be an optimal signaling scheme in the static Bayesian persuasion problem. From Section 3.2, π_{BP}^* is persuasive, where the obedient strategy ρ_{ob} of the receiver is a single-period best response. Thus, a T times repetition of π_{BP}^*, denoted by $\sigma^{\text{rep}} = (\pi_{\text{BP}}^{*(1)}, ..., \pi_{\text{BP}}^{*(T)})$, is a feasible strategy for the T-period game. The receiver's global best response to σ^{rep} consists of obedience in each period. Thus, the designer's T-period payoff is $T \cdot U_{\text{BP}}^*(\mathcal{I})$. By the optimality of the dynamic benchmark, we immediately have

$$\mathcal{U}_T^{**}(\mathcal{I}) \;\geq\; \mathcal{U}_T(\sigma^{\text{rep}}, \phi_{\text{ob}}; \mu_D) \;=\; T \cdot U_{\text{BP}}^*(\mathcal{I}) = \mathcal{U}_T^{*}(\mathcal{I}).$$

□

A dynamic program to solve for $\mathcal{U}_T^{**}(\mathcal{I})$. Having shown that the designer's dynamic benchmark $\mathcal{U}_T^{**}(\mathcal{I})$ can be achieved by a history-independent dynamically persuasive DDP strategy, we present a *dynamic programming* algorithm for obtaining such an optimal DDP strategy.

Define a value function $F(t, g)$ as the designer's optimal payoff *from period* t to period T under a history-independent dynamically persuasive DDP strategy, such that the receiver's continuation payoff from being obedient (from period t to T) is weakly larger than the continuation payoff from being punished by receiving no information plus a non-negative margin of $g \geq 0$. Then the

designer's optimization problem is, for each period $t = \{T, T-1, ..., 1\}$,

$$\begin{aligned} F(t,g) \;=\; & \max_{\pi^{(t)},\ldots,\pi^{(T)}} \; \sum_{t'=t}^{T} U(\pi^{(t')}, \rho_{\text{ob}}; \mu_D) \\ & \text{s.t.} \quad \sum_{t'=t}^{T} \gamma^{t'-t} \cdot V(\pi^{(t')}, \rho_{\text{ob}}; \mu_R) \;\geq\; \sum_{t'=t}^{T} \gamma^{t'-t} \cdot V_{\text{uninformed}}(\mu_R) \;+\; g \\ & \quad (\pi^{(t)}, \ldots, \pi^{(T)}) \text{ is dynamically persuasive.} \end{aligned} \tag{16}$$

Essentially, the designer promises the receiver that he will obtain a payoff of at least $\sum_{t'=t}^{T} \gamma^{t'-t} \cdot V_{\text{uninformed}}(\mu_R) + g$ if he keeps being obedient from period t to T. Set $F(t,g) = -\infty$ if the optimization problem (16) is infeasible, which leads to the dynamic benchmark obtained at $\mathcal{U}_T^{**}(\mathcal{I}) = F(1,0)$.

We compute the value function $F(t,g)$ recursively from the last period. That is, we first solve the optimization (17) for the last period $t = T$ by solving

$$\begin{aligned} F(T,g) \;=\; & \max_{\pi^{(T)}} \; U\big(\pi^{(T)}, \rho_{\text{ob}}; \mu_D\big) \\ & \text{s.t.} \quad V\big(\pi^{(T)}, \rho_{\text{ob}}; \mu_R\big) - V_{\text{uninformed}}(\mu_R) \geq g, \\ & \quad \pi^{(T)} \text{ is persuasive.} \end{aligned} \tag{17a}$$

Then, for each $t \in \{T-1, T-2, \ldots, 1\}$, $g \geq 0$, solve the following optimization problem, which makes use of the value function $F(t+1, g')$ for the next period $t+1$:

$$\begin{aligned} F(t,g) \;=\; & \max_{\pi^{(t)}, g' \geq 0} \; U(\pi^{(t)}, \rho_{\text{ob}}; \mu_D) + F(t+1, g') \\ & \text{s.t.} \quad V(\pi^{(t)}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R) + \gamma \cdot g' \;\geq\; g \\ & \quad \sum_{\omega \in \Omega} \mu_R(\omega) \pi^{(t)}(a|\omega) \Big(v(a,\omega) - v(a',\omega) + \gamma \cdot g' \Big) \;\geq\; 0 \quad \forall a, a' \in A. \end{aligned} \tag{18}$$

To see the intuition, we note that $F(t,g)$ is the designer's best total payoff from period t onward, while promising the receiver that his obedient payoff will be larger than his deviating payoff by at least a margin of $g \geq 0$, *starting from the current period* t to the end. At the current period t, the designer chooses the current signaling scheme $\pi^{(t)}$ and the next promised payoff g' that the designer must deliver starting from the next period $t+1$. The current signaling scheme $\pi^{(t)}$ and the next promised payoff g' should satisfy the first constraints in the above optimization problems, which ensure that the receiver's obedient payoff is larger than the deviating payoff by a margin of g starting from the *beginning* of period t. The choices of $\pi^{(t)}$ and g' should also satisfy the second

constraints above, such that the receiver, *after* receiving the current recommendation a, will indeed find the recommendation persuasive, considering the gain from deviating from a to a' and the loss of promised future utility g' after deviation.

The following Lemma 8 shows that the dynamic program above computes the dynamic benchmark correctly, with its full proof in Appendix C.3.

Lemma 8 (Dynamic program for dynamic benchmark)**.** *The dynamic program* (17)-(18) *computes an optimal solution* $(\sigma^*, \phi_{\text{ob}})$ *that reaches the dynamic benchmark* $\mathcal{U}_T^{**}(\mathcal{I})$*, where* $\sigma^* = (\pi_{\sigma^*}^{(t)})_{t=1}^T$ *is an optimal DDP strategy that is history-independent and dynamically persuasive.*

4.2 Examples of Optimal DDP Strategies

In this subsection, we illustrate the structure of the history-independent dynamically persuasive DDP strategy, which achieves the dynamic benchmark $\mathcal{U}_T^{**}$. We consider an example with three states $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and three actions $A = \{a_1, a_2, a_3\}$. The payoffs of the designer and receiver are as follows (rows for actions, columns for states):

designer's u	ω_1	ω_2	ω_3
a_1	1	1	1
a_2	1/2	1/2	1/2
a_3	0	0	0

receiver's v	ω_1	ω_2	ω_3
a_1	1	0	0
a_2	0	1	0
a_3	0	0	1

In words, the designer prefers a_1 the most, a_2 the second, a_3 the least, regardless of the state, while the receiver wants to match the state. The designer's belief μ_D is a uniform distribution over the states, while the receiver's prior belief μ_R is not:

$$\mu_D = (1/3, 1/3, 1/3), \quad \mu_R = (1/5, 3/10, 1/2).$$

The receiver believes that ω_3 is the most likely under belief μ_R, so he will take a_3 by default. We will compare the optimal DDP strategy against the optimal signaling scheme π_{BP}^* in the static Bayesian persuasion problem:

$\pi_{\text{BP}}^* =$

	ω_1	ω_2	ω_3
a_1	1	2/3	2/5
a_2	0	1/3	1/5
a_3	0	0	2/5

which provides the designer and the receiver with expected payoffs

$$U^*_{\text{BP}} = U(\pi^*_{\text{BP}}, \rho_{\text{ob}}; \mu_D) = 0.778, \qquad V(\pi^*_{\text{BP}}, \rho_{\text{ob}}; \mu_R) = 0.5 = V_{\text{uninformed}}(\mu_R), \tag{19}$$

respectively. For different choices of T and γ, we compute an optimal DDP strategy $\sigma^* = (\pi^{(t)}_{\sigma^*})_{t=1}^T$ for the designer by using the dynamic programming in Lemma 8 and present the results below.

4.2.1 Optimal DDP strategy for $T = 2$ periods

We start with the simple case where the total number of periods is $T = 2$ and the receiver has discount factor $\gamma = 1$ (the receiver is patient). In this case, the optimal DDP strategy $\sigma^* = (\pi^{(1)}_{\sigma^*}, \pi^{(2)}_{\sigma^*})$ for the designer is the following:

$$\sigma = \left(\pi^{(1)}: \begin{array}{c|ccc} & \omega_1 & \omega_2 & \omega_3 \\ \hline a_1 & 1 & 1 & 1 \\ a_2 & 0 & 0 & 0 \\ a_3 & 0 & 0 & 0 \end{array}, \quad \pi^{(2)}: \begin{array}{c|ccc} & \omega_1 & \omega_2 & \omega_3 \\ \hline a_1 & 1 & 0 & 2/5 \\ a_2 & 0 & 1 & 0 \\ a_3 & 0 & 0 & 3/5 \end{array} \right)$$

With an obedient receiver, the designer's expected payoffs under the two signaling schemes are

$$U(\pi^{(1)}_{\sigma^*}, \rho_{\text{ob}}; \mu_D) = 1, \qquad U(\pi^{(2)}_{\sigma^*}, \rho_{\text{ob}}; \mu_D) = 0.6333.$$

The (average) dynamic benchmark is given by the average payoff

$$\frac{1}{T}\mathcal{U}^{**}_T = \frac{1}{T}\mathcal{U}_T(\sigma^*, \phi_{\text{ob}}; \mu_D) = 0.8167, \tag{20}$$

which is larger than the static BP benchmark (19). The obedient receiver's expected payoffs under the two signaling schemes are

$$V(\pi^{(1)}_{\sigma^*}, \rho_{\text{ob}}; \mu_R) = 0.2, \qquad V(\pi^{(2)}_{\sigma^*}, \rho_{\text{ob}}; \mu_D) = 0.8,$$

with the undiscounted average being

$$\frac{1}{T}\mathcal{V}_T(\sigma^*, \phi_{\text{ob}}; \mu_D) = 0.5 = V_{\text{uninformed}}(\mu_R).$$

We note some interesting observations about the optimal DDP strategy:

- *The signaling scheme $\pi^{(1)}_{\sigma^*}$ in the first period is not persuasive*: it always recommends action a_1, which is in favor of the designer but not the receiver. An obedient receiver's payoff under $\pi^{(1)}_{\sigma^*}$, 0.2, is even lower than the uninformed payoff $V_{\text{uninformed}}(\mu_R) = 0.5$.

- The signaling scheme $\pi^{(2)}_{\sigma^*}$ in the second period is *persuasive* and provides the receiver with a *high ex-ante payoff* of $V(\pi^{(2)}_{\sigma^*}, \rho_{\text{ob}}; \mu_R) = 0.8$, which is significantly larger than the receiver's uninformed payoff $V_{\text{uninformed}}(\mu_R) = 0.5$.

- Due to the high ex-ante payoff promised by $\pi^{(2)}_{\sigma^*}$, the receiver is willing to follow the non-persuasive recommendation from $\pi^{(1)}_{\sigma^*}$ in the first period, because otherwise the receiver will suffer a large loss in the second period. *Therefore, the designer's two-period strategy* $\sigma^* = (\pi^{(1)}_{\sigma^*}, \pi^{(2)}_{\sigma^*})$ *satisfies dynamic persuasiveness.*

- The *average* direct signaling scheme

$$\overline{\pi}_{\sigma^*} = \frac{1}{2}\left(\pi^{(1)}_{\sigma^*} + \pi^{(2)}_{\sigma^*}\right) = \begin{array}{c|ccc} & \omega_1 & \omega_2 & \omega_3 \\ \hline a_1 & 1 & 0.5 & 0.7 \\ a_2 & 0 & 0.5 & 0 \\ a_3 & 0 & 0 & 0.3 \end{array}$$

  is *not persuasive* in the static Bayesian persuasion game. In particular, when $\overline{\pi}_{\sigma^*}$ recommends action a_1, the receiver's posterior belief is $(0.2857, 0.2143, 0.5)$, under which the receiver's best action should be a_3.

In summary, the optimal DDP strategy $\sigma^* = (\pi^{(1)}_{\sigma^*}, \pi^{(2)}_{\sigma^*})$ for $T = 2$ periods exhibits a *first-exploit-then-compensate* structure. It is dynamically persuasive even though the individual signaling schemes are not necessarily persuasive in the static Bayesian persuasion game. By exploiting the receiver early and compensating him later, the designer obtains a higher average payoff than the static Bayesian persuasion benchmark. Such a strategy is feasible only for a sufficiently forward-looking receiver: if the receiver's discount factor γ is too small, then the receiver will not accept later compensation for early exploitation, so the problem reduces to the repetition of $T = 2$ Bayesian persuasion problems, and the dynamic benchmark collapses to the static benchmark. We will discuss the effect of γ in more detail in Section 4.2.3.

4.2.2 Optimal DDP strategy for $T = 30$ periods

We next consider the case with $T = 30$ periods and the receiver has discount factor $\gamma = 0.8$. In Figure 2, we plot the designer's and receiver's per-period payoffs $U(\pi^{(t)}_{\sigma^*}, \rho_{\text{ob}}; \mu_D)$ and $V(\pi^{(t)}_{\sigma^*}, \rho_{\text{ob}}; \mu_R)$ under the direct signaling schemes of the optimal DDP strategy $\sigma^* = (\pi^{(t)}_{\sigma^*})_{t=1}^{T}$; they are the blue and red solid curves in Figure 2. The average payoff of the designer under the optimal DDP

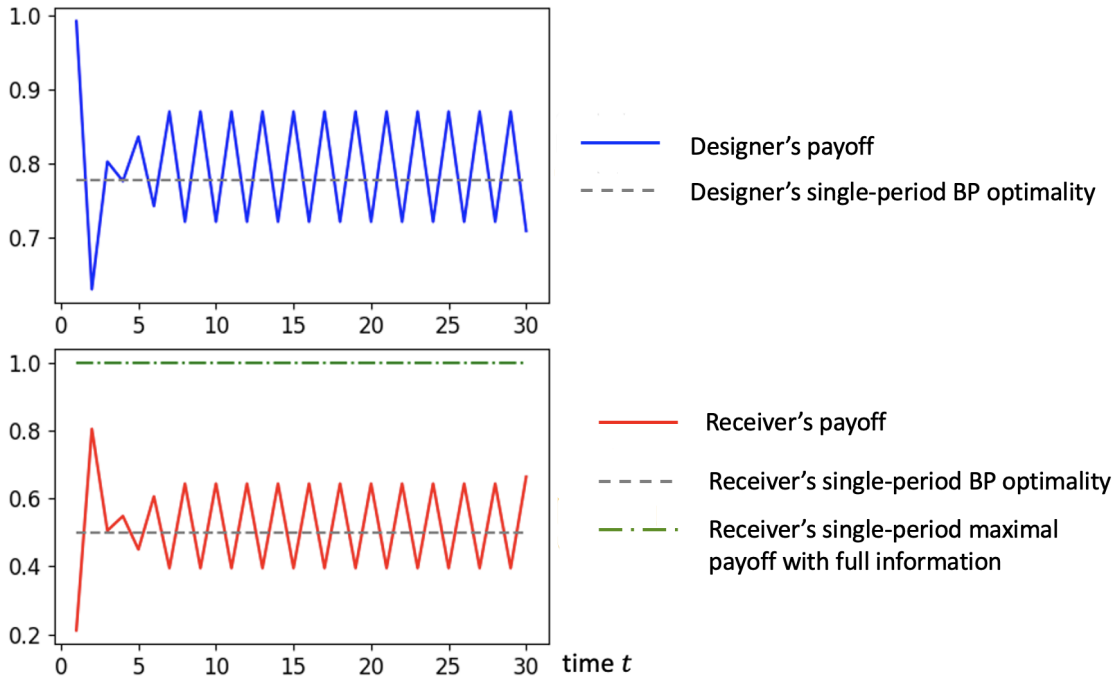


Figure 2: Illustration of the optimal DDP strategy in Section 4.2.2.

strategy is

$$\frac{1}{T}\mathcal{U}_T^{**} = \frac{1}{T}\sum_{t=1}^{T} U(\pi_{\sigma^*}^{(t)}, \rho_{\text{ob}}; \mu_D) = 0.7957, \tag{21}$$

which is larger than the static BP benchmark $U_{\text{BP}}^* = 0.778$.

We then observe *oscillations* in the blue and red curves, so the optimal DDP strategy's influences on the designer's payoff and the receiver's payoff are not monotone. The oscillations can be explained by the first-exploit-then-compensate structure of the optimal 2-period strategy and the fact that the designer is more patient than the receiver. First, to see one oscillation, as in the 2-period case, the designer's optimal strategy is to exploit first to obtain a desired payoff and then to compensate the receiver to "pay for the debt."

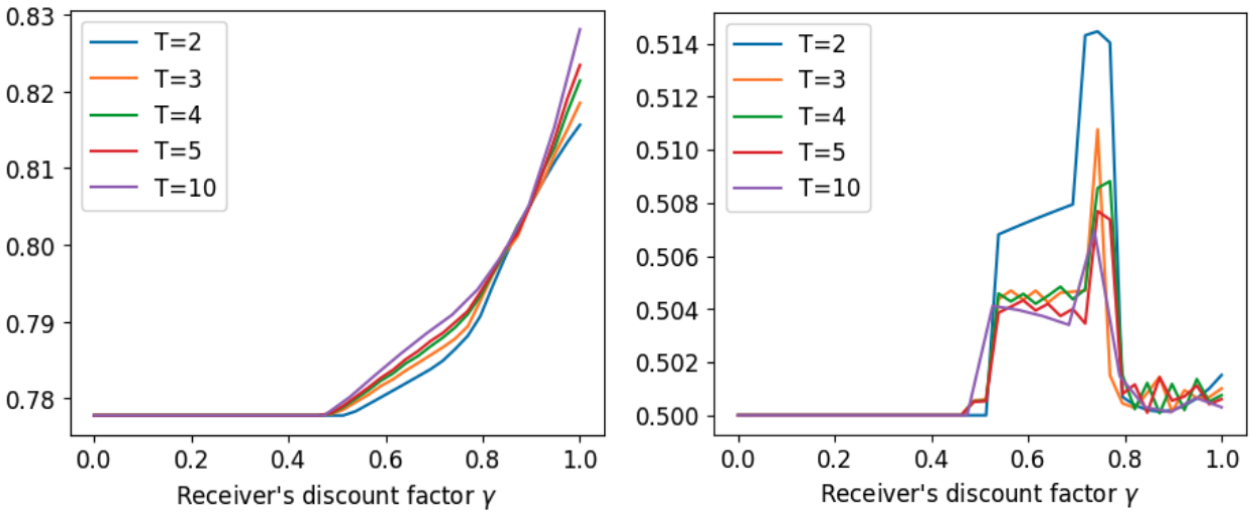
Then, to see the repeated oscillations, the periodic occurrence of the oscillations is explained by the designer's patience and the receiver's impatience. Intuitively, pushing all "repayment" to the end is too costly when the receiver is impatient but the designer is patient. To keep the receiver obedient during an early exploit phase, the designer must promise future utility whose discounted value to the receiver covers the interim shortfall. If that promise is delayed T' periods, then its face value must be scaled up by about $\gamma^{-T'}$ to matter to the receiver. But the designer values future payoffs without discounting, so such compensation is too costly for the designer. The way to satisfy dynamic persuasiveness is therefore to interleave short "exploit" bursts with timely

"compensation" bursts, keeping the receiver's continuation value near the participation threshold while avoiding a large back-loaded debt. That thus yields the observed periodic pattern in both players' payoffs.

Finally, we remark that the oscillation occurs only for a sufficiently forward-looking yet impatient receiver. If the receiver is very impatient with a very small value of γ, then the designer's optimal strategy will be to repeat her optimal strategy of the single-period Bayesian persuasion game, leaving no oscillations for their payoffs.

4.2.3 Effect of γ on Dynamic Benchmark

We illustrate the effect of the receiver's discount factor γ on the designer's dynamic benchmark $\mathcal{U}_T^{**}$. Figure 3 plots the designer's average payoff $\frac{1}{T}\mathcal{U}_T^{**} = \frac{1}{T}\sum_{t=1}^{T} U(\pi_{\sigma^*}^{(t)}, \rho_{\text{ob}}; \mu_D)$ and the receiver's normalized payoff $\frac{1}{\sum_{t=1}^{T}\gamma^t}\sum_{t=1}^{T}\gamma^t V(\pi_{\sigma^*}^{(t)}, \rho_{\text{ob}}; \mu_R)$ under the optimal DDP strategy as a function of γ. We observe that the designer's payoff is increasing with γ, which implies that a more patient receiver leads to a higher dynamic benchmark for the designer. The receiver's normalized payoff, on the other hand, is not monotone with respect to γ. When γ is small (smaller than 0.5 in this example), the dynamic benchmark $\frac{1}{T}\mathcal{U}_T^{**}$ coincides with the static Bayesian persuasion benchmark U_{BP}^*.



(a) Designer's average payoff

(b) Receiver's normalized payoff

Figure 3: Designer's and receiver's payoffs under the optimal DDP strategy, as a function of receiver's discount factor γ.

The observation that the designer's dynamic benchmark $\mathcal{U}_T^{**}$ is increasing with the receiver's discount factor γ is not limited to this example. Note that the optimal DDP strategy σ_γ^* for γ, which is dynamically persuasive for a receiver with discount factor γ, is also dynamically persuasive for

a receiver with a larger discount factor $\gamma' > \gamma$. With a more patient receiver, the better-than-uninformed total utility promised by the DDP strategy in the future becomes more attractive to the receiver, so the receiver is more willing to be obedient. Since the DDP strategy σ^*_γ is dynamically persuasive for $\gamma' > \gamma$, σ_γ is a feasible solution to the designer's dynamic programming problem under γ' and thus is weakly worse than the optimal strategy $\sigma^*_{\gamma'}$ that achieves the dynamic benchmark for the designer under γ'.

Proposition 2. *For any instance $\mathcal{I}$ and any $T \geq 1$, the designer's dynamic benchmark $\mathcal{U}^{**}_T(\mathcal{I})$ is weakly increasing with the receiver's discount factor γ.*

Proof. The full proof is given in Appendix C.4. □

Though the dynamic benchmark $\mathcal{U}^{**}_T(\mathcal{I})$ is increasing in γ, *the designer* is not always better off in the learning problem with an unknown receiver's belief μ_R. The learning cost of $O(\frac{1}{1-\gamma}\log^2\frac{1}{\varepsilon})$ shown in Lemma 3 is also increasing in γ, as a more patient receiver has a larger incentive to manipulate the learning process of the designer. In fact, we will show in Section 6 that the designer has to suffer a linear regret of $\Omega(T)$ when the receiver is fully patient ($\gamma = 1$).

The designer benefits in the dynamic benchmark compared to that in the static benchmark as T goes to infinity. To see that, in Figure 3a, the x-axis intersects the designer's average payoff at her single-period Bayesian persuasion optimality, while the highest horizontal line intersects the designer's average payoff at her single-period optimality under IR. When $T \to \infty$, when the receiver converges to being fully patient $\gamma \to 1$, then the time-averaged dynamic benchmark converges to the designer's single-period optimality under the the receiver's individual rationality constraint (for simplicity, it is named as the designer's optimality under IR), while the designer's optimality under IR is weakly larger than the designer's single-period Bayesian persuasion optimality (i.e., the designer's optimality under the constraint of the signaling scheme being persuasive for the receiver). To see why, the solution to the designer's single-period Bayesian persuasion problem is a feasible solution to her single-period optimization under the receiver's individual rationality constraint. When the receiver is not patient with $\gamma \to 0$, the time-averaged dynamic benchmark converges to the designer's single-period Bayesian persuasion optimality. The idea is formally characterized in Proposition 3.

Proposition 3. *With Assumptions 1 and 3, the designer's dynamic benchmark satisfies*

$$\lim_{T\to\infty}\lim_{\gamma\to 1}\frac{1}{T}\mathcal{U}^{**}_T(\mathcal{I};\gamma) = U^*_{\mathrm{IR}},$$

where U^*_{IR} is the designer's single-period optimality under the constraint of the receiver's individ-

ual rationality:

$$\max_{\pi:\Omega\to\Delta(A)} U(\pi,\rho_{\text{ob}};\mu_D) \text{ s.t. } V(\pi,\rho_{\text{ob}};\mu_R)\geq V_{\text{uninformed}}(\mu_R). \tag{22}$$

Proof. See Appendix C.5. □

5 Learning to Achieve No Regret Against the Dynamic Benchmark

For the dynamic benchmark, the learning algorithm also consists of two phases. The first is an exploration phase for learning the receiver's prior belief, which is based on the same algorithm as for the static benchmark but with different parameters. For the second phase, we propose a new robustification process, *dynamic robustification*, to robustify a designer-optimal strategy computed from the learned prior, which ensures approximate optimality for the designer under the receiver's belief μ_R. In this section, we first study the dynamic robustification based on the estimated belief. Then, we show the full algorithm to achieve no regret for the dynamic benchmark.

5.1 Dynamic Robustification for DDP Strategy

Denote by $\hat{\sigma} = (\pi_{\hat{\sigma}}^{(1)},\dots,\pi_{\hat{\sigma}}^{(T)})$ the designer's optimal DDP strategy, which achieves the dynamic benchmark $\mathcal{U}_T^{**}(\Omega,A,u,v,\mu_D,\hat{\mu})$ under belief $\hat{\mu}$. Proposition 1 implies that the strategy $\hat{\sigma}$ is history-independent and dynamically persuasive for $\hat{\mu}$. We show that $\hat{\sigma}$ can be *dynamically robustified* into another DDP strategy $\tilde{\sigma} = (\pi_{\tilde{\sigma}}^{(1)},\dots,\pi_{\tilde{\sigma}}^{(T)})$ that is history-independent and dynamically persuasive for a strategic receiver with any belief μ_R satisfying $\|\mu_R-\hat{\mu}\|_1\leq\varepsilon$.

We first show how to convert any signaling scheme $\pi_{\hat{\sigma}}^{(t)}$ for any period t from the designer's optimal DDP strategy $\hat{\sigma}$ under belief $\hat{\mu}$ to a corresponding signaling scheme $\pi_{\tilde{\sigma}}^{(t)}$.

A natural idea is to apply static robustification (Lemma 4) to the signaling scheme $\pi_{\hat{\sigma}}^{(t)}$ in each period t in the DDP strategy $\hat{\sigma}$. Recall that static robustification mixes over signaling schemes and thereby increases the *persuasiveness* of the statically robustified signaling scheme under the receiver's prior belief μ_R, i.e.,

$$\mathbb{E}_{\omega\sim\mu_R(\cdot|a,\tilde{\pi}^{(t)})}\big[v(a,\omega)-v(a',\omega)\big]\geq\min\big\{\mathbb{E}_{\omega\sim\hat{\mu}(\cdot|a,\pi_{\hat{\sigma}}^{(t)})}\big[v(a,\omega)-v(a',\omega)\big]+O(\delta),0\big\},\ \forall a'\in A\backslash\{a\},$$

where $\tilde{\pi}^{(t)}$ is the signaling scheme statically robustified from $\pi_{\hat{\sigma}}^{(t)}$. Thus, the persuasiveness of the statically robustified signaling scheme improves only after the receiver receives a recommendation. *Before* the receiver receives a recommendation, however, static robustification may lower the receiver's *ex ante* payoff from obedience, i.e., $V(\tilde{\pi}^{(t')},\rho_{\text{ob}};\mu_R)<V(\pi_{\hat{\sigma}}^{(t')},\rho_{\text{ob}};\mu_R)$ for any $t'>t$. In period t, after a recommendation is sent by the statically robustified signaling scheme, the re-

ceiver's incentive to obey the recommendation in the current period may become stronger. But the receiver's *ex ante* payoff from obeying the recommendations in *future periods onward* may get worse. Thus, on balance, he may refuse to obey the current recommendation and leave the game. In that way, a DDP strategy obtained by statically robustifying the signaling scheme in each period is not guaranteed to be dynamically persuasive.

Hence, we propose a new type of robustification, *payoff-improving robustification,* which not only increases the persuasiveness of an unrobustified signaling scheme but also weakly increases the receiver's *ex ante* payoff from being obedient.

Given a typical signaling scheme $\pi_{\hat{\sigma}}$ in a DDP strategy $\hat{\sigma}$, the main idea of the payoff-improving robustification is as follows. We first compute the signaling scheme $\pi_{\tilde{\sigma}}$ that is obtained from static robustification (Lemma 4) on $\pi_{\hat{\sigma}}$. Then, we mix $\pi_{\tilde{\sigma}}$ with the full-revelation signaling scheme π_{full}, which always recommends the optimal action for the receiver at each state. In particular, if $\pi_{\tilde{\sigma}}$ is already close to π_{full} (with distance less than $O(\sqrt{\delta})$), we directly change $\pi_{\tilde{\sigma}}$ to π_{full}. If $\pi_{\tilde{\sigma}}$ is far from π_{full} (with distance more than $O(\sqrt{\delta})$), we take a mixture $(1-O(\sqrt{\delta}))\pi_{\tilde{\sigma}}+O(\sqrt{\delta})\pi_{\text{full}}$. Such a mixture weakly improves the receiver's utility (compared to $\pi_{\hat{\sigma}}$), improves the persuasiveness, while hurting the designer's utility by at most $O(\sqrt{\delta})$.

Lemma 9 shows the result of any signaling scheme from the dynamic robustification in a DDP strategy $\hat{\sigma}$. We omit the time index of the signaling scheme for simplicity. See Appendix D.1 for the full proof of Lemma 9.[3]

Lemma 9 (Payoff-improving robustification)**.** *Assume Assumptions 1, 2, and 3. For any signaling scheme $\pi_{\hat{\sigma}}$ of the designer's optimal DDP strategy under belief $\hat{\mu}$, with any belief μ satisfying $\|\mu-\hat{\mu}\|_1 \le \varepsilon \le \frac{p_0^4 DG}{16}$, with a dynamic robustification parameter $\delta \in [\frac{2\varepsilon}{p_0 D}, \frac{p_0^3 G}{8}]$, the signaling scheme $\pi_{\hat{\sigma}}$ can be converted into another signaling scheme $\pi_{\tilde{\sigma}}$ such that:*

- *(improved persuasiveness) the persuasiveness of $\pi_{\tilde{\sigma}}$ and $\pi_{\hat{\sigma}}$ for the receiver under beliefs μ and $\hat{\mu}$, respectively, are close:*

$$\mathbb{E}_{\omega\sim\mu(\cdot|a,\pi_{\tilde{\sigma}})}\big[v(a,\omega)-v(a',\omega)\big] \ge \min\Big\{\mathbb{E}_{\omega\sim\hat{\mu}(\cdot|a,\pi_{\hat{\sigma}})}\big[v(a,\omega)-v(a',\omega)\big]+\delta D-\tfrac{2\varepsilon}{p_0},\ 0\Big\},$$

  for any action $a \in A$ recommended by $\pi_{\tilde{\sigma}}$ and $a' \in A \setminus \{a\}$;

- *(weakly improving the payoff of the receiver) the receiver's payoff from obedience is weakly increased: $V(\pi_{\tilde{\sigma}}, \rho_{\text{ob}}; \mu) \ge V(\pi_{\hat{\sigma}}, \rho_{\text{ob}}; \mu)$;*

[3]Lemma 9 can also be used to perform robustification to achieve no-regret against the static benchmark, as Lemma 4 does. However, Lemma 9 is derived from and conceptually more complex than Lemma 4. Lemma 9 gives worse (up to constant factors) results for static robustification than those given by Lemma 4. Thus, we keep using Lemma 4 for static robustification and use Lemma 9 for dynamic robustification.

- *(small influence on the payoff of the designer) with an obedient receiver, the designer's payoff has a small decrease:* $U(\pi_{\tilde{\sigma}}, \rho_{\text{ob}}; \mu_D) \geq U(\pi_{\hat{\sigma}}, \rho_{\text{ob}}; \mu_D) - 4\sqrt{\frac{\delta}{p_0^3 G}}$.

We then apply the above payoff-improving robustification procedure to every signaling scheme $\pi_{\hat{\sigma}}^{(t)}$ in the designer-optimal DDP strategy $\hat{\sigma}$ for belief $\hat{\mu}$, a process we call *dynamic robustification*. The result of the complete dynamic robustification is given in Theorem 2.

Theorem 2 (Dynamic robustification)**.** *Assume Assumptions 1, 2, and 3. Given the designer's optimal DDP strategy* $\hat{\sigma}$ *under belief* $\hat{\mu}$*, with any belief* μ_R *satisfying* $\|\mu_R - \hat{\mu}\|_1 \leq \varepsilon$ *for* $\varepsilon \leq \frac{p_0^4}{1+p_0\gamma/(1-\gamma)} \frac{DG}{16}$*, there exists another history-independent DDP strategy* $\tilde{\sigma} = (\pi_{\tilde{\sigma}}^{(1)}, \dots, \pi_{\tilde{\sigma}}^{(T)})$ *such that*

- $\tilde{\sigma}$ *is dynamically persuasive under belief* μ_R*;*
- $\tilde{\sigma}$ *is approximately optimal for the designer with*

$$\mathcal{U}_T(\tilde{\sigma}, \phi_{\text{ob}}; \mu_D) \;\geq\; \mathcal{U}_T(\hat{\sigma}, \phi_{\text{ob}}; \mu_D) - 4T\sqrt{\left(\tfrac{1}{p_0} + \tfrac{\gamma}{1-\gamma}\right)\tfrac{2\varepsilon}{p_0^3 GD}}.$$

We now give the intuition for Theorem 2. The high-level idea of our dynamic robustification procedure is to robustify the signaling scheme of each period t in the DDP strategy $\hat{\sigma}$ backward from T. In particular, for each period $t = T, T-1, \dots, 1$, we convert $\pi_{\hat{\sigma}}^{(t)}$ to another signaling scheme $\pi_{\tilde{\sigma}}^{(t)}$ by Lemma 9 while treating the dynamically robustified tail $(\pi_{\tilde{\sigma}}^{(t+1)}, \dots, \pi_{\tilde{\sigma}}^{(T)})$ as fixed.

Given a recommended action a at time t and any deviation a', Lemma 9 ensures the persuasiveness of $\pi_{\tilde{\sigma}}^{(t)}$ under the belief μ_R after robustification compared to the persuasiveness of $\pi_{\hat{\sigma}}^{(t)}$ under the estimate $\hat{\mu}$ by

$$\mathbb{E}_{\omega\sim\mu_R(\cdot|a,\pi_{\tilde{\sigma}}^{(t)})}\big[v(a,\omega) - v(a',\omega)\big] \;\geq\; \min\Big\{\mathbb{E}_{\omega\sim\hat{\mu}(\cdot|a,\pi_{\hat{\sigma}}^{(t)})}\big[v(a,\omega) - v(a',\omega)\big] + \delta_t D - \tfrac{2\varepsilon}{p_0},\, 0\Big\},$$

where a *buffer* $\delta_t D$ and a loss from the belief estimate $\frac{2\varepsilon}{p_0}$ are exerted on the persuasiveness of $\pi_{\hat{\sigma}}^{(t)}$ under the estimate $\hat{\mu}$.

But that is not enough. Then, we consider the dynamic persuasiveness constraint to secure the receiver's obedience in period t. The dynamic persuasiveness in period t is

$$\underbrace{\gamma^t \mathbb{E}_{\omega\sim\mu_R(\cdot|a,\pi_{\tilde{\sigma}}^{(t)})}[v(a,\omega) - v(a',\omega)]}_{\text{today's incentive}} + \underbrace{\sum_{t'=t+1}^{T} \gamma^{t'}\big(V(\pi_{\tilde{\sigma}}^{(t')}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R)\big)}_{\text{discounted future payoffs from obedience}} \;\geq\; 0. \quad (\text{DP}_t)$$

There are two intuitive cases. First, when the persuasiveness of $\pi_{\tilde{\sigma}}^{(t)}$ under μ_R is already non-negative, then the (already robustified) tail is dynamically persuasive, so the future accumulated

payoffs from obedience is non-negative as well. Thus, the dynamic persuasiveness constraint (DP_t) holds immediately.

The second case is that the persuasiveness of $\pi_{\tilde{\sigma}}^{(t)}$ under μ_R is already negative. Then we consider the help from the already robustified tail, which comes from two sources: (i) robustifying the tail cannot hurt the receiver (Lemma 9), and (ii) $V(\cdot, \cdot; \cdot)$ is 1-Lipschitz in the belief (Lemma 13), so the gap from $\hat{\mu}$ to μ_R costs at most 2ε per period. Putting the contribution from those two sources of the tail, the left-hand side of the dynamic persuasiveness constraint (DP_t) is at least $\gamma^t\big(\delta_t D - \frac{2\varepsilon}{p_0}\big) - \frac{2\varepsilon\gamma^{t+1}}{1-\gamma}$. Choosing the dynamic robustification parameter $\delta_t = \big(\frac{1}{p_0} + \frac{\gamma}{1-\gamma}\big)\frac{2\varepsilon}{D}$, we obtain a non-negative right-hand side of (DP_t), so the dynamic persuasiveness constraint (DP_t) holds.

Starting from $t = T$ and moving backward, the same argument secures (DP_t) for each period t. Thus the completely dynamically robustified strategy $\tilde{\sigma}$ is dynamically persuasive for every μ_R with $\|\mu_R - \hat{\mu}\|_1 \le \varepsilon$.

By the upper bound of the cost from each robustification at each period t on the designer, $4\sqrt{\frac{\delta_t}{p_0^3 G}}$ (Lemma 9), as well as our choice of δ_t above, we have

$$\mathcal{U}_T(\tilde{\sigma}, \phi_{\text{ob}}; \mu_D) \;\ge\; \mathcal{U}_T(\hat{\sigma}, \phi_{\text{ob}}; \mu_D) - 4T\sqrt{\Big(\frac{1}{p_0} + \frac{\gamma}{1-\gamma}\Big)\frac{2\varepsilon}{p_0^3 G D}}.$$

See the full proof in Appendix D.2.

5.2 Full Learning Algorithm with $O(\log^2 T)$ Regret

The full algorithm for the learning performance relative to the dynamic benchmark is shown in Algorithm 5. Algorithm 5 has two phases: a short *learn-the-prior* phase and a long *exploit* phase. The first phase is learning the belief μ_R, which is the same as in Algorithm 4, but with different choices of parameters. The second phase is the dynamic robustification.

Algorithm 5: Learning Algorithm for the Dynamic Benchmark

Parameter: $\varepsilon > 0$

1. Run Algorithm 3 with parameter ε to obtain an estimate $\hat{\mu}$ of the receiver's belief. Let T_0 be the number of periods taken.
2. Compute the optimal DDP strategy $\hat{\sigma}$ for the $(T - T_0)$-period game assuming that the receiver has a prior belief $\hat{\mu}$ (by using, e.g., the dynamic program in Lemma 8).
3. Apply dynamic robustification (Theorem 2) to $\hat{\sigma}$ to obtain DDP strategy $\tilde{\sigma}$. Use $\tilde{\sigma}$ for the remaining $T - T_0$ periods.

Theorem 3 shows the regret incurred by Algorithm 5 is at most in the order of $O(\log^2 T)$. We first provide a brief intuition for proof here. Algorithm 5 has two phases: a short *learn-the-*

prior phase and a long *exploit* phase. In the learning phase we run the estimation routine until the receiver's belief is identified up to ℓ_1–error ε, which (even against a strategic receiver) takes

$$T_0 \;=\; O\!\left(\frac{1}{1-\gamma}\,\log^2\!\frac{1}{\varepsilon}\right)$$

rounds. In the exploitation phase we compute the optimal DDP plan for $\hat{\mu}$ and then apply *dynamic robustification* so that obedience remains optimal for any belief within ε of $\hat{\mu}$. Dynamic robustification preserves persuasiveness while incurring at most a $T \cdot O(\sqrt{\varepsilon})$ loss in the designer's value. Choosing ε in the order of T^{-2} makes this second-phase loss negligible relative to T_0, so the total dynamic regret is driven by the learning time, yielding

$$O\!\left(\frac{1}{1-\gamma}\,\log^2 T\right).$$

Theorem 3. *Assume Assumptions 1, 2, 3, 4. Choose* $\varepsilon = \frac{p_0^4}{1+p_0\gamma/(1-\gamma)}\frac{DG}{16T^2}$ *with* $\gamma \in (0,1)$. *The dynamic regret of Algorithm 5 is at most*

$$\mathrm{Reg}^{\mathrm{dynamic}}(T; \textit{Algorithm 5}, \mathcal{I}) \;\le\; O\!\left(\frac{1}{1-\gamma}\log^2 T\right). \tag{23}$$

Proof. The proof will use the following Lemma 10 (with proof given in Appendix D.3) to compare the dynamic benchmarks for T-period game and $(T-T_0)$-period game.

Lemma 10. *For any integers* $T \ge T_0 \ge 0$ *and beliefs* $\mu_D, \hat{\mu}$,

$$\mathcal{U}^{**}_{T-T_0}(\mu_D, \hat{\mu}) \ge \mathcal{U}^{**}_{T}(\mu_D, \hat{\mu}) - T_0.$$

By the dynamic robustification result (Theorem 2), $\tilde{\sigma}$ is dynamically persuasive for a receiver with prior μ_R and approximately optimal for the designer in the $(T-T_0)$-period game under a prior $\hat{\mu}$:

$$\begin{aligned}\mathcal{U}_{T-T_0}(\tilde{\sigma}, \phi_{\mathrm{ob}}; \mu_D) \;&\ge\; \mathcal{U}_{T-T_0}(\hat{\sigma}, \phi_{\mathrm{ob}}; \mu_D) - 4(T-T_0)\sqrt{\left(\tfrac{1}{p_0}+\tfrac{\gamma}{1-\gamma}\right)\tfrac{2\varepsilon}{p_0^3 GD}} \\ &=\; \mathcal{U}^{**}_{T-T_0}(\mu_D, \hat{\mu}) - 4(T-T_0)\sqrt{\left(\tfrac{1}{p_0}+\tfrac{\gamma}{1-\gamma}\right)\tfrac{2\varepsilon}{p_0^3 GD}}\end{aligned}$$

where the equality is because $\hat{\sigma}$ is optimal for the $(T-T_0)$-period game under receiver prior $\hat{\mu}$. Using Lemma 10, the total utility of the designer in T periods is at least

$$\mathcal{U}_T \;\ge\; 0 + \mathcal{U}_{T-T_0}(\tilde{\sigma}, \phi_{\mathrm{ob}}; \mu_D) \;\ge\; \mathcal{U}^{**}_T(\mu_D, \hat{\mu}) - T_0 - 4(T-T_0)\sqrt{\left(\tfrac{1}{p_0}+\tfrac{\gamma}{1-\gamma}\right)\tfrac{2\varepsilon}{p_0^3 GD}}.$$

Applying dynamic robustification (Theorem 2) to the optimal DDP strategy under receiver prior μ_R, we obtain an approximately optimal DDP strategy under receiver prior $\hat{\mu}$, so $\mathcal{U}_T^{**}(\mu_D, \hat{\mu}) \geq \mathcal{U}_T^{**}(\mu_D, \mu_R) - 4T\sqrt{\left(\frac{1}{p_0} + \frac{\gamma}{1-\gamma}\right)\frac{2\varepsilon}{p_0^3 GD}}$. Thus, we have

$$\begin{aligned}\mathcal{U}_T \;&\geq\; \mathcal{U}_T^{**}(\mu_D, \mu_R) - 4T\sqrt{\left(\tfrac{1}{p_0} + \tfrac{\gamma}{1-\gamma}\right)\tfrac{2\varepsilon}{p_0^3 GD}} - T_0 - 4(T - T_0)\sqrt{\left(\tfrac{1}{p_0} + \tfrac{\gamma}{1-\gamma}\right)\tfrac{2\varepsilon}{p_0^3 GD}} \\ &\geq\; \mathcal{U}_T^{**}(\mu_D, \mu_R) - T_0 - 8T\sqrt{\left(\tfrac{1}{p_0} + \tfrac{\gamma}{1-\gamma}\right)\tfrac{2\varepsilon}{p_0^3 GD}}.\end{aligned}$$

Choose $\varepsilon = \frac{p_0^4}{1+p_0\gamma/(1-\gamma)}\frac{DG}{16T^2}$, which satisfies the requirement for ε in Theorem 2. By Lemma 3, $T_0 \leq 2|\Omega|\left(\frac{1}{p_0} + \frac{1}{1-\gamma}\ln\left(\frac{24|\Omega|}{(1-\gamma)G^4 p_0^6 \varepsilon}\right)\right)\log_2 \frac{24|\Omega|}{G^2 p_0^4 \varepsilon}$. So the designer's regret is at most

$$\begin{aligned}&\mathrm{Reg}^{\mathrm{dynamic}}(T; \text{Algorithm 5}, \mathcal{I}) \;=\; T_0 + 8T\sqrt{\left(\tfrac{1}{p_0} + \tfrac{\gamma}{1-\gamma}\right)\tfrac{2\varepsilon}{p_0^3 GD}} \\ &\leq\; 2|\Omega|\left(\tfrac{1}{p_0} + \tfrac{1}{1-\gamma}\ln\left(\tfrac{24|\Omega|}{(1-\gamma)G^4 p_0^6 \varepsilon}\right)\right)\log_2 \tfrac{24|\Omega|}{G^2 p_0^4 \varepsilon} + 8T\sqrt{\left(\tfrac{1}{p_0} + \tfrac{\gamma}{1-\gamma}\right)\tfrac{2\varepsilon}{p_0^3 GD}} \\ &=\; O\Big(\frac{1}{1-\gamma}\log^2 T\Big).\end{aligned}$$

□

Interestingly, the regret bounds of our two learning algorithms, Algorithms 4 and 5 (against the static benchmark and the dynamic benchmark, respectively) are both logarithmic in time horizon: $O(\log^2 T)$. While empirical estimation by observing the frequency distribution of states is not relevant here, as the receiver's belief is subjective, we note that our algorithm gives faster learning. Learning a distribution via empirical estimation leads to regret at most $O(\sqrt{T})$, which is exponentially worse than our regret of $O(\log^2 T)$.

6 Necessity of Receiver's Discount: Lower Bound on Regrets

In this section, we prove that the assumption that the receiver has a $\gamma < 1$ discount factor is necessary for the designer to achieve no regret asymptotically (or, sub-linear regret). In particular, if the receiver has a discount factor $\gamma = 1$, then no matter what learning algorithm is used by the designer, there always exists an information design instance such that the designer's regret is at least a constant times T.

We now introduce the regret *lower bound*. A regret *lower bound* means that *for any algorithm,* there exists some instance $\mathcal{I}$, i.e., for some unknown prior μ_R, such that

$$\mathrm{Reg}(T; \mathcal{I}) \;\geq\; c_{\mathcal{I}} \cdot g(T) \;=\; \Omega(g(T)), \quad \forall T,$$

for some instance-dependent constant $c_{\mathcal{I}} > 0$. The function $g(T)$ captures the unavoidable cost of

not knowing μ_R no matter what learning algorithm is used, which fundamentally shows how hard the problem of information design with an unknown prior is.

Denote by $T_\gamma = \sum_{t=1}^{T} \gamma_t$ the receiver's *effective horizon*. Theorem 4 shows that the worst-case regret of any learning algorithm $\mathcal{G}$ for the designer must be *linear* in the receiver's effective horizon, with respect to both the static benchmark and the dynamic benchmark.

Theorem 4 (Regret lower bound for both benchmarks)**.** *For any learning algorithm $\mathcal{G}$, there exists an instance $\mathcal{I}$ such that $T \cdot U^*_{\mathrm{BP}}(\mathcal{I}) = \mathcal{U}^{**}_T(\mathcal{I})$ and the designer's regrets with respect to both benchmarks are at least $\Omega(T_\gamma)$:*

- *for $\gamma \in [0, 1 - \frac{1}{T})$, the designer's regret is at least $\Omega(\frac{1}{1-\gamma})$.*
- *for $\gamma \in [1 - \frac{1}{T}, 1]$, the designer's regret is at least $\Omega(T)$.*

Proof. See Appendix E.3 for the full proof. □

We remark that the instance above satisfies the regularity assumptions in Theorem 1, which implies that the designer's regrets come from the receiver's strategic decision-making instead of irregularities. Theorem 4 implies that the designer cannot obtain sub-linear regret against a patient receiver. This shows the necessity of an impatient receiver in our model.

6.1 Overview of the Proof for the Regret Lower Bound Against Both Benchmarks

The rest of this section gives an overview of the proof of Theorem 4. We construct a family of single-period information design instances that capture a single-period auction problem, where the receiver's subjective belief corresponds to the buyer's private value of the item in the auction. Using the fact that an auctioneer who does not know the buyer's private value suffers a constant revenue loss (Amin, Rostamizadeh, and Syed, 2013), we prove that the designer, not knowing the receiver's belief, must suffer a constant regret in the single-period game. We then *reduce* the T-period game to the single-period game to show that the designer's total regret is at least $\Omega(T_\gamma)$.

Single-period information design problem with menu To prove Theorem 4, we construct a family of instances of single-period information design problems with menus. These instances capture single-buyer single-seller auctions. Suppose there are two states $\Omega = \{\omega_0, \omega_1\}$ and two actions $A = \{a_0, a_1\}$. The designer always prefers the receiver to take a_1 regardless of the state, i.e.,

$$u(a_1, \omega) = 1, \quad u(a_0, \omega) = 0, \quad \forall \omega \in \Omega. \tag{24}$$

The receiver wants to match the state , i.e.,[4]

$$v(a_0, \omega_0) = 0, \quad v(a_0, \omega_1) = 0, \quad v(a_1, \omega_0) = -1, \quad v(a_1, \omega_1) = 1. \tag{25}$$

The designer's belief μ_D is as follows.

$$\mu_D(\omega_0) = 1 - p_0, \quad \mu_D(\omega_1) = p_0, \quad \text{with } 0 < p_0 < 1/2. \tag{26}$$

The receiver's belief μ_R is parameterized by a number $\tilde{\mu}_R \in [0, 1]$ that is unknown to the designer. In particular, set

$$\mu_R(\omega_0) = \tfrac{1}{1+\tilde{\mu}_R}, \quad \mu_R(\omega_1) = \tfrac{\tilde{\mu}_R}{1+\tilde{\mu}_R}. \tag{27}$$

Consider a single-period game where the designer first elicits the receiver's belief and then uses a direct signaling scheme to recommend an action to the receiver. Formally, the designer designs a *menu*, which is a family of direct signaling schemes $\mathcal{M} = \{\pi_\mu\}$. If the receiver reports belief μ and obeys the recommendation from the signaling scheme π_μ, then he gets single-period payoff at

$$\begin{aligned} V^{\text{sp}}(\mathcal{M}, \mu_R, \mu) &= \sum_{\omega \in \Omega} \mu_R(\omega) \sum_{a \in A} \pi_\mu(a|\omega) v(a, \omega) \\ &= -\mu_R(\omega_0)\pi_\mu(a_1|\omega_0) + \mu_R(\omega_1)\pi_\mu(a_1|\omega_1) \\ &= \tfrac{1}{1+\tilde{\mu}_R}\Big(\tilde{\mu}_R \pi_\mu(a_1|\omega_1) - \pi_\mu(a_1|\omega_0)\Big). \end{aligned} \tag{28}$$

Eq. (28) captures the payoff of a buyer in a corresponding auction setting. Specifically, the receiver's belief $\tilde{\mu}_R$ corresponds to the buyer's private value of an item; $\pi_\mu(a_1|\omega_1)$ captures the probability that the receiver gets the item; $\pi_\mu(a_1|\omega_0)$ is the receiver's payment. We say the menu $\mathcal{M} = \{\pi_\mu\}$ is

- *incentive compatible* if the receiver is willing to report his belief (private value) truthfully: $\forall \mu_R, \forall \mu$, $V^{\text{sp}}(\mathcal{M}, \mu_R, \mu_R) \geq V^{\text{sp}}(\mathcal{M}, \mu_R, \mu)$;
- *individually rational* if the receiver is weakly better off with the menu compared to receiving no information: $\forall \mu_R$, $V^{\text{sp}}(\mathcal{M}, \mu_R, \mu_R) \geq V_{\text{uninformed}}(\mu_R) = \max_{a \in A} \sum_{\omega \in \Omega} \mu_R(\omega) v(a, \omega)$.

Benchmarks If the designer knows the receiver's belief μ_R, then the designer's optimal single-period signaling scheme will be

$$\pi^*_{\mu_R}(a_1|\omega_1) = 1, \quad \pi^*_{\mu_R}(a_0|\omega_1) = 0, \quad \pi^*_{\mu_R}(a_1|\omega_0) = \tilde{\mu}_R, \quad \pi^*_{\mu_R}(a_0|\omega_0) = 1 - \tilde{\mu}_R \tag{29}$$

[4] We allow negative utility $v(a_1, \omega_0) = -1$ for convenience. One can define an equivalent instance with positive utility by adding a constant.

where $\pi^*_{\mu_R}(a_1|\omega_1) = 1$ means always allocating the item to the buyer, and $\pi^*_{\mu_R}(a_1|\omega_0) = \tilde{\mu}_R$ means setting the payment equal to the buyer's value $\tilde{\mu}_R$ (see Appendix E.1 for the formal proof). The optimal signaling scheme gives the designer an expected payoff of

$$U^*(\mu_R) = \mu_D(\omega_0)\pi^*_{\mu_R}(a_1|\omega_0) + \mu_D(\omega_1)\pi^*_{\mu_R}(a_1|\omega_1) = (1-p_0)\tilde{\mu}_R + p_0, \tag{30}$$

where $U^*(\mu_R)$ is exactly the Bayesian persuasion benchmark $U^*_{\mathrm{BP}}(\mathcal{I})$

$$U^*(\mu_R) = U^*_{\mathrm{BP}}(\mathcal{I}). \tag{31}$$

Moreover, the Bayesian persuasion benchmark turns out to be equal to the global dynamic benchmark in this instance, so the designer's regret with respect to the two benchmarks will be the same.

Lemma 11. *In the above information design instance, $T \cdot U^*_{\mathrm{BP}}(\mathcal{I}) = \mathcal{U}^{**}_T(\mathcal{I})$.*

Proof. See Appendix E.1. □

Single-period regret When the designer does not know the receiver's belief μ_R and uses a certain incentive compatible and individually rational menu $\mathcal{M}$, she gets a single-period payoff at

$$U^{\mathrm{sp}}(\mathcal{M}, \mu_R) = (1-p_0)\pi_{\mu_R}(a_1|\omega_0) + p_0\pi_{\mu_R}(a_1|\omega_1). \tag{32}$$

Compared to the optimal utility (30), the designer's single-period regret is

$$\mathrm{Reg}^{\mathrm{sp}}(\mathcal{M}, \mu_R) = (1-p_0)\tilde{\mu}_R + p_0 - U^{\mathrm{sp}}(\mathcal{M}, \mu_R). \tag{33}$$

Lemma 12. *For any incentive compatible and individually rational menu $\mathcal{M}$, there exists an instance with $p_0 = \frac{1}{8}$ and $\tilde{\mu}_R \in [\frac{3}{8}, 1]$ (which corresponds to a belief μ_R satisfying the regularity assumption 1) such that $\mathrm{Reg}^{\mathrm{sp}}(\mathcal{M}, \mu_R) \geq \frac{1}{16}$.*

Proof. See Appendix E.2. □

Reduction from T-period game to single-period game We then reduce the T-period information design game to a single-period game. Let $\mathcal{G}$ be any learning algorithm of the designer in the T-period game. Recall that a strategic receiver with prior μ_R best responds by using T-period strategy $\phi^*(\mathcal{G}, \mu_R) = (\phi^{(t)})_{t=1}^T$. The pair $(\mathcal{G}, \phi^*(\mathcal{G}, \mu_R))$ generates a sequence $(\pi^{(t)}, \rho^{(t)})_{t=1}^T$ of the designer's signaling schemes and the receiver's single-period strategies. We consider the average probability $\sigma_{\mu_R}(a|\omega)$ that the receiver takes action a conditioning on state ω when the receiver uses

$\phi^*(\mathcal{G}, \mu_R)$ to interact with the designer's algorithm $\mathcal{G}$, weighted by the receiver's discount factor:

$$\sigma_{\mu_R}(a|\omega) := \mathbb{E}_{(\pi^{(t)},\rho^{(t)})\sim\mathcal{G},\phi^*(\mathcal{G},\mu_R)}\left[\frac{1}{T_\gamma}\sum_{t=1}^{T}\gamma^t\sum_{s\in S}\pi^{(t)}(s|\omega)\rho^{(t)}(a|s)\right]. \tag{34}$$

$\sigma_{\mu_R}(\cdot|\omega)$ is a distribution over A, so σ_{μ_R} can be regarded as a direct signaling scheme. Consider the menu $\mathcal{M} = \{\sigma_{\mu_R}\}_{\mu_R\in\Delta(\Omega)}$ of the direct signaling scheme σ_{μ_R} associated with belief μ_R. Because the receiver is best-responding, the menu $\mathcal{M}$ constructed in this way must be incentive-compatible and individually rational for the receiver; this is reminiscent to a "revelation principle" and we formally prove this in Appendix E.3. Then, according to Lemma 12, for any such menu there exists an instance such that the single-period regret of the designer is at least $\frac{1}{16}$. That is in fact the designer's average regret during the periods when she has not fully learned the receiver's belief μ_R due to receiver's manipulation. Because the receiver's manipulation horizon is T_γ, the designer's total regret is at least $\frac{1}{16}T_\gamma$, which proves Theorem 4. We provide a formal argument in Appendix E.3.

7 Discussion

We study the information design problem when the information designer does not know the belief of a strategic receiver. We design learning algorithms so that the designer can learn the receiver's belief through repeated interactions. Our learning algorithms are robust to the receiver's strategic manipulation of the learning process of the designer.

We define regret relative to two benchmarks to measure the performance of the learning algorithms. The static benchmark is T times the single-period optimum for the designer for the known belief. The dynamic benchmark is the global optimality for the designer for the known belief. Our learning algorithms achieve no regret against both benchmarks at fast speeds of $O(\log^2 T)$. Our work thus provides a robust learning foundation for the information design problems with unknown subjective beliefs of strategic receivers.

We offer a brief discussion of several natural extensions. First, while we consider the receiver's belief as a convergence outcome of his misspecified learning, it would be interesting to extend the analysis to consider a receiver who is learning about the distribution. Will the two players' learning converge to the true distribution of the states? Or will their learning converge to a belief that is neither the true distribution of the states nor the receiver's belief (μ_R in the paper) formed by his own misspecified learning? Will any factors in the game, such as their patience comparisons, determine which learning outcome they will reach?

Second, it is interesting to extend our results to consider a designer with a discount factor close to but strictly less than 1. As our model shows in the paper, the designer cannot be more impatient

than the receiver. But what if the designer is close to being patient, i.e., γ_D close to 1? And what if the designer is only more patient than the receiver but is still impatient, i.e., $\gamma < \gamma_D < 1$? Such questions may have interesting implications for real-world problems. For example, the online platform may have limited patience with a certain user because interactions require time, but time is valuable.

Finally, our learning algorithms indeed enable the designer to achieve no regret at fast speeds. But our learning algorithms rely on using the receiver's actions, which may contain the receiver's privacy and security information. Can the designer still learn the receiver's belief at a fast speed without learning the receiver's privacy? And if learning the privacy cannot be avoided, can the designer use algorithms that unlearn the receiver's privacy to make the learning reliable and trustworthy for the receiver? We leave those as open questions for future work.

REFERENCES

Alonso, Ricardo and Odilon Câmara (Sept. 2016). "Bayesian persuasion with heterogeneous priors". en. In: *Journal of Economic Theory* 165, pp. 672–706. ISSN: 00220531. DOI: 10.1016/j.jet.2016.07.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S0022053116300497 (visited on 12/22/2023).

Amin, Kareem, Afshin Rostamizadeh, and Umar Syed (2013). "Learning Prices for Repeated Auctions with Strategic Buyers". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'13. event-place: Lake Tahoe, Nevada. Red Hook, NY, USA: Curran Associates Inc., pp. 1169–1177.

Ba, Cuimin (2023). *Robust Misspecified Models and Paradigm Shifts*. arXiv: 2106.12727 [econ.TH]. URL: https://arxiv.org/abs/2106.12727.

Bacchiocchi, Francesco et al. (2024). *Markov Persuasion Processes: Learning to Persuade from Scratch*. URL: https://arxiv.org/abs/2402.03077.

Ball, Ian (2023). "Dynamic Information Provision: Rewarding the Past and Guiding the Future". In: *Econometrica* 91.4, pp. 1363–1391. DOI: https://doi.org/10.3982/ECTA17345. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA17345. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA17345.

Ball, Ian and Deniz Kattwinkel (2025). "Robust Robustness". In: *Working Paper*.

Bergemann, Dirk and Stephen Morris (May 2016). "Information Design, Bayesian Persuasion, and Bayes Correlated Equilibrium". In: *American Economic Review* 106.5, pp. 586–91. DOI: 10.1257/aer.p20161046. URL: https://www.aeaweb.org/articles?id=10.1257/aer.p20161046.

Bohren, J. Aislinn and Daniel N. Hauser (2021). "LEARNING WITH HETEROGENEOUS MISSPECIFIED MODELS: CHARACTERIZATION AND ROBUSTNESS". In: *Econometrica*

89.6, pp. 3025–3077. ISSN: 00129682, 14680262. URL: https://www.jstor.org/stable/48636220 (visited on 09/26/2025).

Camara, Modibo K., Jason D. Hartline, and Aleck C. Johnsen (2020). "Mechanisms for a No-Regret Agent: Beyond the Common Prior". In: *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 259–270. URL: https://api.semanticscholar.org/CorpusID:221640554.

Castiglioni, Matteo et al. (2020). "Online Bayesian Persuasion". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 16188–16198. URL: https://proceedings.neurips.cc/paper/2020/file/ba5451d3c91a0f982f103cdbe249bc78-Paper.pdf.

Castiglioni, Matteo et al. (July 2021). "Multi-Receiver Online Bayesian Persuasion". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 1314–1323. URL: https://proceedings.mlr.press/v139/castiglioni21a.html.

Crawford, Vincent P. and Joel Sobel (Nov. 1982). "Strategic Information Transmission". In: *Econometrica* 50.6, p. 1431. ISSN: 00129682. DOI: 10.2307/1913390. URL: https://www.jstor.org/stable/1913390?origin=crossref (visited on 11/21/2023).

Dworczak, Piotr and Anton Kolotilin (Nov. 2024). "The persuasion duality". In: *Theoretical Economics* 19.4, None. DOI: None. URL: https://ideas.repec.org/a/the/publsh/5900.html.

Dworczak, Piotr and Alessandro Pavan (2022). "Preparing for the Worst but Hoping for the Best: Robust (Bayesian) Persuasion". In: *Econometrica* 90.5, pp. 2017–2051. DOI: https://doi.org/10.3982/ECTA19107. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA19107. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA19107.

Esponda, Ignacio, Demian Pouzo, and Yuichi Yamamoto (2021). "Asymptotic behavior of Bayesian learners with misspecified models". In: *Journal of Economic Theory* 195, p. 105260. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2021.105260. URL: https://www.sciencedirect.com/science/article/pii/S0022053121000776.

Feng, Yiding, Wei Tang, and Haifeng Xu (July 2022). "Online Bayesian Recommendation with No Regret". en. In: *Proceedings of the 23rd ACM Conference on Economics and Computation*. Boulder CO USA: ACM, pp. 818–819. ISBN: 978-1-4503-9150-4. DOI: 10.1145/3490486.3538327. URL: https://dl.acm.org/doi/10.1145/3490486.3538327 (visited on 10/17/2022).

Frick, Mira, Ryota Iijima, and Yuhta Ishii (None 2023). "Belief Convergence under Misspecified Learning: A Martingale Approach". In: *The Review of Economic Studies* 90.2, pp. 781–814. DOI: None. URL: https://ideas.repec.org/a/oup/restud/v90y2023i2p781-814..html.

Fudenberg, Drew and Giacomo Lanzani (2023). "Which misspecifications persist?" In: *Theoretical Economics* 18.3, pp. 1271–1315. DOI: https://doi.org/10.3982/TE5298. eprint: https://

onlinelibrary.wiley.com/doi/pdf/10.3982/TE5298. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/TE5298.

Fudenberg, Drew, Giacomo Lanzani, and Philipp Strack (May 2021). "Limit Points of Endogenous Misspecified Learning". In: *Econometrica* 89.3, pp. 1065–1098. DOI: 10.3982/ECTA18508. URL: https://ideas.repec.org/a/wly/emetrp/v89y2021i3p1065-1098.html.

Fudenberg, Drew, Gleb Romanyuk, and Philipp Strack (Sept. 2017). "Active learning with a misspecified prior". In: *Theoretical Economics* 12.3, None. DOI: None. URL: https://ideas.repec.org/a/the/publsh/2480.html.

He, Kevin and Jonathan Libgober (2021). "Evolutionarily Stable (Mis)specifications: Theory and Applications". In: *Proceedings of the 22nd ACM Conference on Economics and Computation*. EC '21. Budapest, Hungary: Association for Computing Machinery, p. 587. ISBN: 9781450385541. DOI: 10.1145/3465456.3467528. URL: https://doi.org/10.1145/3465456.3467528.

Hossain, Safwan et al. (July 2024). "Multi-Sender Persuasion: A Computational Perspective". In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 18944–18971. URL: https://proceedings.mlr.press/v235/hossain24c.html.

Hu, Ju and Xi Weng (Oct. 2021). "Robust persuasion of a privately informed receiver". en. In: *Economic Theory* 72.3, pp. 909–953. ISSN: 0938-2259, 1432-0479. DOI: 10.1007/s00199-020-01299-5. URL: https://link.springer.com/10.1007/s00199-020-01299-5 (visited on 07/04/2023).

Kamenica, Emir and Matthew Gentzkow (Oct. 2011). "Bayesian Persuasion". en. In: *American Economic Review* 101.6, pp. 2590–2615. ISSN: 0002-8282. DOI: 10.1257/aer.101.6.2590. URL: https://pubs.aeaweb.org/doi/10.1257/aer.101.6.2590 (visited on 10/22/2022).

Kolotilin, Anton et al. (2017). "Persuasion of a Privately Informed Receiver". en. In: *Econometrica* 85.6, pp. 1949–1964. ISSN: 0012-9682. DOI: 10.3982/ECTA13251. URL: https://www.econometricsociety.org/doi/10.3982/ECTA13251 (visited on 12/22/2023).

Kosterina, Svetlana (July 2022). "Persuasion with unknown beliefs". In: *Theoretical Economics* 17.3, None. DOI: None. URL: https://ideas.repec.org/a/the/publsh/4742.html.

Li, Ce and Tao Lin (2025). "Information Design with Unknown Prior (Extended Abstract)". In: *Working Paper*.

Li, Fei and Peter Norman (May 2021). "Sequential persuasion". In: *Theoretical Economics* 16.2, None. DOI: None. URL: https://ideas.repec.org/a/the/publsh/3474.html.

Mathevet, Laurent, Jacopo Perego, and Ina Taneva (None 2020). "On Information Design in Games". In: *Journal of Political Economy* 128.4, pp. 1370–1404. DOI: 10.1086/705332. URL: https://ideas.repec.org/a/ucp/jpolec/doi10.1086-705332.html.

Morris, Stephen, Daisuke Oyama, and Satoru Takahashi (2024). "Implementation via Information Design in Binary-Action Supermodular Games". In: *Econometrica* 92.3, pp. 775–813. DOI: https://doi.org/10.3982/ECTA19149. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA19149. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA19149.

Myerson, Roger B. (1986). "Multistage Games with Communication". In: *Econometrica* 54.2, pp. 323–358. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1913154 (visited on 07/01/2025).

Rabin, Matthew and Joel L. Schrag (1999). "First Impressions Matter: A Model of Confirmatory Bias". In: *The Quarterly Journal of Economics* 114.1, pp. 37–82. ISSN: 00335533, 15314650. URL: http://www.jstor.org/stable/2586947 (visited on 09/27/2025).

Tversky, Amos and Daniel Kahneman (1973). "Availability: A heuristic for judging frequency and probability". In: *Cognitive Psychology* 5.2, pp. 207–232. ISSN: 0010-0285. DOI: https://doi.org/10.1016/0010-0285(73)90033-9. URL: https://www.sciencedirect.com/science/article/pii/0010028573900339.

Wu, Jibang et al. (2022). "Sequential Information Design: Markov Persuasion Process and Its Efficient Reinforcement Learning". In: *Proceedings of the 23rd ACM Conference on Economics and Computation*. EC '22. Boulder, CO, USA: Association for Computing Machinery, pp. 471–472. ISBN: 9781450391504. DOI: 10.1145/3490486.3538313. URL: https://doi.org/10.1145/3490486.3538313.

Ziegler, Gabriel (Mar. 2020). *Adversarial bilateral information design*. English. WorkingPaper, pp. 1–63.

Zu, You, Krishnamurthy Iyer, and Haifeng Xu (2021). "Learning to Persuade on the Fly: Robustness Against Ignorance". In: *Proceedings of the 22nd ACM Conference on Economics and Computation*. EC '21. Budapest, Hungary: Association for Computing Machinery, pp. 927–928. ISBN: 9781450385541. DOI: 10.1145/3465456.3467593. URL: https://doi.org/10.1145/3465456.3467593.

A USEFUL LEMMAS

Lemma 13 (Lipschitz continuity of utility)**.** *The designer's and receiver's expected utility functions,* $U(\pi, \rho; \mu_D), V(\pi, \rho; \mu_R)$*, are:*

- 1*-Lipschitz continuous with respect to priors under the* ℓ_1*-norm:*

$$|U(\pi, \rho; \mu_1) - U(\pi, \rho; \mu_2)| \;\leq\; \|\mu_1 - \mu_2\|_1.$$

- 1*-Lipschitz continuous with respect to* π *under the matrix* ℓ_1*-norm:*

$$|U(\pi_1, \rho; \mu_D) - U(\pi_2, \rho; \mu_D)| \;\leq\; \max_{\omega \in \Omega} \sum_{s \in S} \big|\pi_1(s|\omega) - \pi_2(s|\omega)\big|.$$

Analogous inequalities hold for $V(\cdot, \rho; \cdot)$*.*

Proof. Write

$$U(\pi, \rho; \mu_D) = \sum_{\omega} \mu_D(\omega) \sum_{s} \pi(s \mid \omega) \sum_{a} \rho(a \mid s) u(a, \omega).$$

(Lipschitz in the prior). For fixed (π, ρ), the inner expectation lies in $[0, 1]$, hence

$$\big|U(\pi, \rho; \mu_1) - U(\pi, \rho; \mu_2)\big| = \Big| \sum_{\omega} (\mu_1 - \mu_2)(\omega) \cdot \underbrace{\sum_{s,a} \pi(s \mid \omega) \rho(a \mid s) u(a, \omega)}_{\in [0,1]} \Big| \leq \sum_{\omega} |\mu_1 - \mu_2|(\omega) = \|\mu_1 - \mu_2\|_1.$$

(Lipschitz in the signaling scheme). For fixed (ρ, μ_D), set $\bar{u}_s(\omega) := \sum_a \rho(a \mid s) u(a, \omega) \in [0, 1]$. Then

$$U(\pi_1, \rho; \mu_D) - U(\pi_2, \rho; \mu_D) = \sum_{\omega} \mu_D(\omega) \sum_{s} \big(\pi_1 - \pi_2\big)(s \mid \omega)\, \bar{u}_s(\omega).$$

For each ω, the inner term is the difference of expectations of a $[0, 1]$-valued function under two distributions on S, so by the dual characterization of total variation,

$$\Big| \sum_{s} \big(\pi_1 - \pi_2\big)(s \mid \omega)\, \bar{u}_s(\omega) \Big| \leq \mathrm{TV}\big(\pi_1(\cdot \mid \omega), \pi_2(\cdot \mid \omega)\big).$$

Averaging over ω (since μ_D is a probability measure) gives the stated bound. Finally, $\mathrm{TV}(p, q) = \frac{1}{2} \sum_s |p(s) - q(s)| \leq \frac{|S|}{2} \max_s |p(s) - q(s)|$ yields the matrix ℓ_1–norm version. The proof for V is identical with u replaced by v and μ_D by μ_R. □

Lemma 14 (Lipschitz continuity of posterior)**.** *Let* $\pi : \Omega \to \Delta(S)$ *be any signaling scheme. Let* $\mu, \mu' \in \Delta(\Omega)$ *be two priors. Let* μ_s*,* μ'_s *be the posterior belief induced by signal* s *under* π *and prior* μ*,* μ' *respectively. Suppose* $\min_{\omega \in \Omega} \mu(\omega) \geq p_0 > 0$*. Then* $\|\mu_s - \mu'_s\|_1 \leq \frac{2}{p_0} \|\mu - \mu'\|_1$*.*

Proof. Let $\pi(s) = \sum_{\omega\in\Omega} \mu(\omega)\pi(s|\omega)$ and $\pi'(s) = \sum_{\omega\in\Omega} \mu'(\omega)\pi(s|\omega)$ be the probability of signal s under prior μ and μ' respectively. By the definition of μ_s, μ'_s and by triangle inequality,

$$\begin{aligned}\|\mu_s - \mu'_s\|_1 &= \sum_{\omega\in\Omega} \left|\tfrac{\mu(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)}\right| \\ &\le \sum_{\omega\in\Omega} \left|\tfrac{\mu(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi(s)}\right| + \sum_{\omega\in\Omega} \left|\tfrac{\mu'(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)}\right|.\end{aligned}$$

For the first term above, $\sum_{\omega\in\Omega} \left|\tfrac{\mu(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi(s)}\right| = \sum_{\omega\in\Omega} \tfrac{\pi(s|\omega)}{\pi(s)}|\mu(\omega) - \mu'(\omega)|$. We note that, $\forall\omega \in \Omega$,

$$\frac{\pi(s|\omega)}{\pi(s)} = \frac{\pi(s|\omega)}{\sum_{\omega'\in\Omega}\mu(\omega')\pi(s|\omega')} \le \frac{\pi(s|\omega)}{p_0\sum_{\omega'\in\Omega}\pi(s|\omega')} \le \frac{1}{p_0}. \tag{35}$$

Thus, $\sum_{\omega\in\Omega} \left|\tfrac{\mu(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi(s)}\right| \le \sum_{\omega\in\Omega} \tfrac{1}{p_0}|\mu(\omega) - \mu'(\omega)| = \tfrac{1}{p_0}\|\mu - \mu'\|_1$.

For the second term,

$$\begin{aligned}\sum_{\omega\in\Omega} \left|\tfrac{\mu'(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)}\right| &= \sum_{\omega\in\Omega} \mu'(\omega)\pi(s|\omega)\left|\tfrac{\pi'(s)-\pi(s)}{\pi(s)\pi'(s)}\right| \\ &= \sum_{\omega\in\Omega} \mu'(\omega)\pi(s|\omega)\left|\tfrac{\sum_{\omega'\in\Omega}(\mu'(\omega')-\mu(\omega'))\pi(s|\omega')}{\pi(s)\pi'(s)}\right| \\ &\le \sum_{\omega\in\Omega} \mu'(\omega)\pi(s|\omega)\tfrac{\sum_{\omega'\in\Omega}|\mu'(\omega')-\mu(\omega')|\cdot\max_{\omega'\in\Omega}\pi(s|\omega')}{\pi(s)\pi'(s)} \\ &= \|\mu' - \mu\|_1 \sum_{\omega\in\Omega} \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)} \tfrac{\max_{\omega'\in\Omega}\pi(s|\omega')}{\pi(s)} \\ \text{by (35)}\ &\le \|\mu' - \mu\|_1 \sum_{\omega\in\Omega} \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)} \tfrac{1}{p_0} \\ &= \tfrac{1}{p_0}\|\mu' - \mu\|_1.\end{aligned}$$

Therefore, we obtain $\|\mu_s - \mu'_s\|_1 \le \frac{2}{p_0}\|\mu' - \mu\|_1$. □

Lemma 15 (coalescing signals)**.** *Let $s_1, s_2 \in S$ be two signals from a signaling scheme π, to which the receiver responds by the same action $\rho(s_1) = \rho(s_2) \in A$. Then, after coalescing signals s_1 and s_2 into s by letting $\pi^c(s|\omega) = \pi(s_1|\omega) + \pi(s_2|\omega)$, we have:*

- *Assuming that the receiver also responds by action $\rho(s) = \rho(s_1) = \rho(s_2)$ for the coalesced signal, the information designer's and the receiver's expected utilities do not change after coalescing.*
- *For any actions $a, a' \in A$, if the receiver's posterior μ_{s_1} under s_1 satisfies $\mathbb{E}_{\omega\sim\mu_{s_1}}[v(a,\omega) - v(a',\omega)] \ge X_1$ and the posterior μ_{s_2} under s_2 satisfies $\mathbb{E}_{\omega\sim\mu_{s_2}}[v(a,\omega) - v(a',\omega)] \ge X_2$,*

then the receiver's posterior under s satisfies

$$\mathbb{E}_{\omega\sim\mu_s}[v(a,\omega)-v(a',\omega)]\geq\min\{X_1,X_2\}.$$

Proof. The expected utilities of the two players clearly do not change after coalescing. We then verify the second claim. By definition, the receiver's posterior μ_s satisfies

$$\begin{aligned}
&\mathbb{E}_{\omega\sim\mu_s}\Big[v(a,\omega)-v(a',\omega)\Big]\\
&=\sum_{\omega\in\Omega}\frac{\hat{\mu}(\omega)\pi^c(s|\omega)}{\sum_{\omega'\in\Omega}\hat{\mu}(\omega')\pi^c(s|\omega')}\Big[v(a,\omega)-v(a',\omega)\Big]\\
&=\sum_{\omega\in\Omega}\frac{\hat{\mu}(\omega)\pi(s_1|\omega)}{\sum_{\omega'\in\Omega}\hat{\mu}(\omega')\pi^c(s|\omega')}\Big[v(a,\omega)-v(a',\omega)\Big]+\sum_{\omega\in\Omega}\frac{\hat{\mu}(\omega)\pi(s_2|\omega)}{\sum_{\omega'\in\Omega}\hat{\mu}(\omega')\pi^c(s|\omega')}\Big[v(a,\omega)-v(a',\omega)\Big]\\
&=\frac{\sum_{\omega'\in\Omega}\hat{\mu}(\omega')\pi(s_1|\omega')}{\sum_{\omega'\in\Omega}\hat{\mu}(\omega')\pi^c(s|\omega')}\sum_{\omega\in\Omega}\frac{\hat{\mu}(\omega)\pi(s_1|\omega)}{\sum_{\omega'\in\Omega}\hat{\mu}(\omega')\pi(s_1|\omega')}\Big[v(a,\omega)-v(a',\omega)\Big]\\
&\quad+\frac{\sum_{\omega'\in\Omega}\hat{\mu}(\omega')\pi(s_2|\omega')}{\sum_{\omega'\in\Omega}\hat{\mu}(\omega')\pi^c(s|\omega')}\sum_{\omega\in\Omega}\frac{\hat{\mu}(\omega)\pi(s_2|\omega)}{\sum_{\omega'\in\Omega}\hat{\mu}(\omega')\pi(s_2|\omega')}\Big[v(a,\omega)-v(a',\omega)\Big]\\
&=\lambda\cdot\mathbb{E}_{\omega\sim\mu_{s_1}}\Big[v(a,\omega)-v(a',\omega)\Big]+(1-\lambda)\cdot\mathbb{E}_{\omega\sim\mu_{s_2}}\Big[v(a,\omega)-v(a',\omega)\Big]\\
&\geq\min\{X_1,X_2\},
\end{aligned}$$

where $\lambda=\frac{\sum_{\omega'\in\Omega}\hat{\mu}(\omega')\pi(s_1|\omega')}{\sum_{\omega'\in\Omega}\hat{\mu}(\omega')\pi^c(s|\omega')}\in[0,1]$. □

B Omitted Proofs in Section 3

B.1 Proof of Lemma 1

Claim 1. *During the binary search algorithm (Algorithm 1), the left endpoint always satisfies $\ell^{(k)}\leq(1+\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$, and the right endpoint always satisfies $r^{(k)}\geq(1-\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$. Additionally, $\ell^{(k)}\leq r^{(k)}$. (This claim includes the initial step where $k=0$.)*

Proof. First, we verify that the claim holds at the initial step $k=0$. By definition, we have $\ell^{(0)}=0\leq(1+\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$, and $r^{(0)}=\frac{1}{Gp_0}$. Because by definition $\tilde{a}_{\omega_j,\omega_i}$ is the first indifference action between states ω_i and ω_j, $\tilde{a}_{\omega_j,\omega_i}$ is weakly better than $a^*_{\omega_j}$ when the signaling ratio is $\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$ for a myopic receiver:

$$v(\tilde{a}_{\omega_j,\omega_i},\omega_i)-v(a^*_{\omega_j},\omega_i)+\frac{\mu_R(\omega_j)}{\mu_R(\omega_i)}\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}\Big(v(\tilde{a}_{\omega_j,\omega_i},\omega_j)-v(a^*_{\omega_j},\omega_j)\Big)\ \geq\ 0.$$

Note that $v(\tilde{a}_{\omega_j,\omega_i},\omega_i) - v(a^*_{\omega_j},\omega_i) \le 1$ and $v(\tilde{a}_{\omega_j,\omega_i},\omega_j) - v(a^*_{\omega_j},\omega_j) \le -G < 0$, so

$$1 - \frac{\mu_R(\omega_j)}{\mu_R(\omega_i)}\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)} \cdot G \;\ge\; 0,$$

which implies

$$\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)} \;\le\; \frac{\mu_R(\omega_i)}{\mu_R(\omega_j)G} \;\le\; \frac{1}{p_0 G}. \tag{36}$$

Thus, $r^{(0)} = \frac{1}{p_0 G} \ge (1-\iota)\frac{\mu_R(\omega_i)}{\mu_R(\omega_j)G}$.

Then, consider step $k \ge 0$. The algorithm chooses the midpoint $q = \frac{\ell^{(k)}+r^{(k)}}{2}$ and uses a signaling scheme satisfying $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} = q$. When signal s_0 is realized:

- if the receiver takes action $a^*_{\omega_i}$, then by the second item of the Lemma 2, $q \le (1+\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$. The algorithm sets $\ell^{(k+1)}$ to be q, so $\ell^{(k+1)}$ satisfies the claim;
- if the receiver does not take action $a^*_{\omega_i}$, then by the first item of Lemma 2, $q \ge (1-\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$. The algorithm sets $r^{(k+1)}$ to be q, so $r^{(k+1)}$ satisfies the claim.

□

Proof of Lemma 1. When Algorithm 1 ends, we have $|r^{(k)} - \ell^{(k)}| \le \tau$. Then, using the above claim, we have

$$\left|\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)} - \ell^{(k)}\right| \;\le\; \tau + \iota\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)},$$

which implies

$$\begin{aligned}
\left|\frac{\mu_R(\omega_i)}{\mu_R(\omega_j)} - \hat{\rho}\right| \;&=\; \left|\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)} - \ell^{(k)}\right| \cdot \frac{v(\tilde{a}_{\omega_j,\omega_i},\omega_j) - v(a^*_{\omega_i},\omega_2)}{v(a^*_{\omega_i},\omega_i) - v(\tilde{a}_{\omega_j,\omega_i},\omega_i)} \\
&\le\; \left(\tau + \iota\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}\right)\cdot\frac{1}{G} \\
&\le\; \left(\tau + \frac{\delta}{Gp_0}\right)\cdot\frac{1}{G} && \text{by (36)} \\
&\le\; \left(\tau + \frac{\gamma^m}{(1-\gamma)G^3p_0^3}\right)\cdot\frac{1}{G} && \text{by Lemma 2} \\
&\le\; \left(\tau + \frac{\gamma^m}{(1-\gamma)G^3p_0^3}\right)\cdot\frac{1}{Gp_0}\cdot\frac{\mu_R(\omega_i)}{\mu_R(\omega_j)}.
\end{aligned}$$

where the last step is because $\frac{\mu_R(\omega_i)}{\mu_R(\omega_j)} \ge p_0$.

Then, we consider the expected running time of Algorithm 1. Because the difference $r^{(k)} - \ell^{(k)}$

shrinks by a half after each while loop, the number of while loops, k, is at most

$$k \;\leq\; \log_2 \frac{r^{(0)} - \ell^{(0)}}{\tau} \;=\; \log_2 \frac{1}{Gp_0\tau}. \tag{37}$$

Then, we consider the number of periods needed in each while loop. By definition, this is equal to the number of periods until signal s_0 is realized, plus m. Since one of $\pi^{(k)}(s_0|\omega_j)$ and $\pi^{(k)}(s_0|\omega_i)$ is 1 (see Algorithm 1), the probability (according to the information designer's prior belief) that s_0 is realized in each period is at least:

$$\pi^{(k)}(s_0) \;=\; \mu_D(\omega_i)\pi^{(k)}(s_0|\omega_i) + \mu_D(\omega_j)\pi^{(k)}(s_0|\omega_j) \;\geq\; \min\{\mu_D(\omega_i), \mu_D(\omega_j)\} \;\geq\; p_0.$$

So by the property of the geometric random variable, the expected number of periods until a signal s_0 is realized is at most

$$\frac{1}{\pi^{(k)}(s_0)} \;\leq\; \frac{1}{p_0},$$

and the total expected number of periods over k while loops is at most

$$k \cdot \left(\frac{1}{p_0} + m\right) \;\leq\; \left(\frac{1}{p_0} + m\right) \log_2 \frac{1}{Gp_0\tau}.$$

□

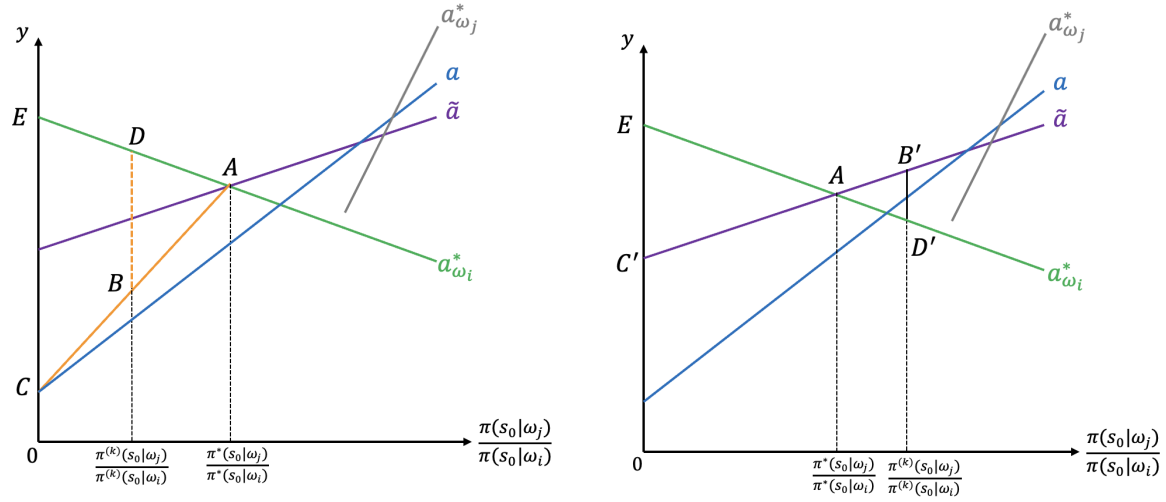
B.2 Proof of Lemma 2

Proof. Case (i): $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} < (1-\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$. If the receiver best responds in that period, then his optimal action will be $a^*_{\omega_i}$, since he is indifferent between $a^*_{\omega_i}$ and the first indifferent action $\tilde{a}_{\omega_j,\omega_i}$ at $\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$ by definition, and also due to the case with $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} < \frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$. Suppose the strategic receiver deviates to an action $a \neq a^*_{\omega_i}$ instead. Then, he will suffer a loss in the current period, and will potentially obtain some gain in the future.

Regarding the gain, the signaling schemes used in the next additional m periods are independent of the receiver's current action, and the receiver's payoff is bounded in $[0, 1]$. Thus, the receiver's total gain in the future is at most

$$\text{gain} \;\leq\; 0 + \sum_{t=m+1}^{T} \gamma^t \cdot 1 \;\leq\; \frac{\gamma^{m+1}}{1-\gamma}.$$

Then, we compute his loss, denoted by $\text{loss}(a^*_{\omega_i}, a)$, which is the difference between the receiver's



(a) The case of $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} < (1-\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$ (b) The case of $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} > (1+\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$

Figure 4: Normalized utility of the receiver for different actions. The x-coordinate is the signaling ratio $\frac{\pi(s_0|\omega_j)}{\pi(s_0|\omega_i)}$. Each line corresponds to an action a, with equation $y_a(x) = v(a,\omega_i) + \frac{\mu_R(\omega_j)}{\mu_R(\omega_i)} \cdot x \cdot v(a,\omega_j)$.

payoff of taking action $a^*_{\omega_i}$ versus the deviation action a in the current round.

$$\begin{aligned}\mathrm{loss}(a^*_{\omega_i}, a) &= \sum_{\omega\in\Omega} \Pr_{\mu_R,\pi^{(k)}}[\omega|s_0]\Big(v(a^*_{\omega_i},\omega) - v(a,\omega)\Big) \\ &= \frac{\mu_R(\omega_i)\pi^{(k)}(s_0|\omega_i)}{\pi^{(k)}(s_0)}\Big(v(a^*_{\omega_i},\omega_i) - v(a,\omega_i)\Big) + \frac{\mu_R(\omega_j)\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0)}\Big(v(a^*_{\omega_i},\omega_j) - v(a,\omega_j)\Big).\end{aligned}$$

Claim 2. $\mathrm{loss}(a^*_{\omega_i}, a) > \iota G^2 p_0^2$.

Proof. Consider the graph (Figure 4a) where the horizontal axis is the ratio $\frac{\pi(s_0|\omega_j)}{\pi(s_0|\omega_i)}$. For each action $a \in A$, we plot the line representing the receiver's single-period utility of taking action a (after a normalization):

$$y_a(x) = v(a,\omega_i) + \frac{\mu_R(\omega_j)}{\mu_R(\omega_i)} \cdot x \cdot v(a,\omega_j), \quad \text{where} \quad x = \frac{\pi(s_0|\omega_j)}{\pi(s_0|\omega_i)}.$$

By definition, point A is the indifference point between actions $a^*_{\omega_i}$ and $\tilde{a}_{\omega_j,\omega_i}$, whose x-coordinate $x_A = \frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$ and y-coordinate $y_A = y_{a^*_{\omega_i}}(x_A) = y_{\tilde{a}_{\omega_j,\omega_i}}(x_A)$. Let $d_x(a^*_{\omega_i}, a) = y_{a^*_{\omega_i}}(x) - y_a(x)$ be the vertical distance between lines $y_{a^*_{\omega_i}}$ and y_a at x-coordinate x. Because A is the first indifference point, line y_a is below line $y_{a^*_{\omega_i}}$ in the $[0, x_A]$ segment: namely, $d_0(a^*_{\omega_i}, a) > 0$ and $d_{x_A}(a^*_{\omega_i}, a) \geq 0$. Consider the line connecting $C = (0, y_a(0) = v(a,\omega_i))$ to A, with B being

the point on this line at x-coordinate

$$x_B = \frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} \leq (1-\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)} = (1-\iota)x_A.$$

Since B is between line y_a and line $y_{a^*_{\omega_i}}$, $d_{x_B}(a^*_{\omega_i}, a) \geq BD$. Triangle ABD is similar to triangle ACE, so

$$BD = CE \cdot \frac{AB}{AC} = CE \cdot \frac{x_A - x_B}{x_A - 0} = \Big(v(a^*_{\omega_i}, \omega_i) - v(a, \omega_i)\Big) \cdot \frac{\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)} - \frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)}}{\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}} > G \cdot \iota.$$

Therefore,

$$G\iota < d_{x_B}(a^*_{\omega_i}, a) = v(a^*_{\omega_i}, \omega_i) + \frac{\mu_R(\omega_j)}{\mu_R(\omega_i)}\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)}v(a^*_{\omega_i}, \omega_j) - v(a, \omega_i) - \frac{\mu_R(\omega_j)}{\mu_R(\omega_i)}\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)}v(a, \omega_j).$$

Multiplying both sides by $\frac{\mu_R(\omega_i)\pi^{(k)}(s_0|\omega_i)}{\pi^{(k)}(s_0)}$,

$$\begin{aligned} G\iota\frac{\mu_R(\omega_i)\pi^{(k)}(s_0|\omega_i)}{\pi^{(k)}(s_0)} &< \frac{\mu_R(\omega_i)\pi^{(k)}(s_0|\omega_i)}{\pi^{(k)}(s_0)}(v(a^*_{\omega_i}, \omega_i) - v(a, \omega_i)) + \frac{\mu_R(\omega_j)\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0)}(v(a^*_{\omega_i}, \omega_j) - v(a^*_{\omega_i}, \omega_j)) \\ &= \text{loss}(a^*_{\omega_i}, a). \end{aligned}$$

We note that $\frac{\mu_R(\omega_i)\pi^{(k)}(s_0|\omega_i)}{\pi^{(k)}(s_0)} \geq Gp_0^2$, because $\mu_R(\omega_i) \geq p_0$, $\pi^{(k)}(s_0) \leq 1$, $\pi^{(k)}(s_0|\omega_i) = \min\{1, \frac{1}{q}\} \geq Gp_0$ from Algorithm 1. Thus, we obtain $\text{loss}(a^*_{\omega_i}, a) \geq \iota G^2 p_0^2$, which proves the claim. □

Comparing gain and loss, given $\iota = \frac{\gamma^m}{(1-\gamma)G^2p_0^2}$, we have

$$\text{loss}(a^*_{\omega_i}, a) \;>\; \iota G^2 p_0^2 \;>\; \frac{\gamma^{m+1}}{1-\gamma} \;\geq\; \text{gain}.$$

Thus, the strategic receiver should not take a for any $a \neq a^*_{\omega_i}$, and thus takes $a^*_{\omega_i}$.

Case (ii): $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} > (1+\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$. See Figure 4b for an illustration. In this case, the optimal action for a myopic receiver is not $a^*_{\omega_i}$, because $\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} > (1+\iota)\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}$. If a strategic receiver takes action $a^*_{\omega_i}$ in the current period, then he suffers a loss in the current period and potentially obtains some gain in the future. As in the previous case, the gain is at most $\frac{\gamma^{m+1}}{1-\gamma}$. Then, consider

the loss. We lower bound the loss by the utility difference between $\tilde{a}$ and $a^*_{\omega_i}$ in the current period:

$$\begin{aligned}
\text{loss}(a^*_{\omega_i}) &\geq \sum_{\omega\in\Omega} \Pr_{\mu_R,\pi^{(k)}}[\omega|s_0]\Big(v(\tilde{a}_{\omega_j,\omega_i},\omega) - v(a^*_{\omega_i},\omega)\Big) \\
&= \frac{\mu_R(\omega_i)\pi^{(k)}(s_0|\omega_i)}{\pi^{(k)}(s_0)}\Big(v(\tilde{a},\omega_1) - v(a^*_{\omega_i},\omega_1)\Big) + \frac{\mu_R(\omega_2)\pi^{(k)}(s_0|\omega_2)}{\pi^{(k)}(s_0)}\Big(v(\tilde{a}_{\omega_j,\omega_i},\omega_j) - v(a^*_{\omega_i},\omega_j)\Big) \\
&= \frac{\mu_R(\omega_i)\pi^{(k)}(s_0|\omega_i)}{\pi^{(k)}(s_0)}\left(v(\tilde{a},\omega_i) - v(a^*_{\omega_i},\omega_i) + \frac{\mu_R(\omega_j)}{\mu_R(\omega_i)}\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)}\Big(v(\tilde{a}_{\omega_j,\omega_i},\omega_j) - v(a^*_{\omega_i},\omega_j)\Big)\right) \\
&= \frac{\mu_R(\omega_i)\pi^{(k)}(s_0|\omega_i)}{\pi^{(k)}(s_0)}B'D' \qquad \text{(see Figure 4b).}
\end{aligned}$$

Because triangle $AB'D'$ is similar to triangle $AC'E$, we have

$$B'D' = C'E\cdot\frac{AB'}{AC'} = \Big(v(a^*_{\omega_i},\omega_i) - v(a,\omega_i)\Big)\cdot\frac{\frac{\pi^{(k)}(s_0|\omega_j)}{\pi^{(k)}(s_0|\omega_i)} - \frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}}{\frac{\pi^*(s_0|\omega_j)}{\pi^*(s_0|\omega_i)}} > G\iota.$$

Same as the previous case, $\frac{\mu_R(\omega_i)\pi^{(k)}(s_0|\omega_i)}{\pi^{(k)}(s_0)} \geq Gp_0^2$. Therefore, $\text{loss}(a^*_{\omega_i}) > G^2p_0^2\iota > \frac{\gamma^{m+1}}{1-\gamma} \geq \text{gain}$, given $\iota = \frac{\gamma^m}{(1-\gamma)G^2p_0^2}$. So, the strategic receiver should not take $a^*_{\omega_i}$. □

B.3 Proof of Lemma 3

Lemma 16. *Let $\varepsilon' = \frac{1}{Gp_0}\big(\tau + \frac{\gamma^m}{(1-\gamma)G^3p_0^3}\big) \leq \frac{1}{2}$. For any two states $\omega_i,\omega_j\in\Omega$, the output $\hat{\rho}_{ij}$ of Algorithm 2 satisfies $\hat{\rho}_{ij}\in(1\pm4\varepsilon')\frac{\mu_R(\omega_i)}{\mu_R(\omega_j)}$. Algorithm 2 terminates in at most $2(\frac{1}{p_0}+m)\log_2\frac{1}{Gp_0\tau}$ periods in expectation.*

Proof. The proof uses Lemma 1 twice. See details below. Let $\varepsilon' = \frac{1}{Gp_0}\big(\tau + \frac{\gamma^m}{(1-\gamma)G^3p_0^3}\big)$. If (ω_i,ω_j) is a pair of distinguishable states, then Lemma 1 shows that the estimate $\hat{\rho}_{ij}$ returned by Algorithm 1 satisfies $\hat{\rho}_{ij}\in(1\pm\varepsilon')\frac{\mu_R(\omega_i)}{\mu_R(\omega_j)}$.

If (ω_i,ω_j) is not a pair of distinguishable states, then by Lemma 1, we have the ratio estimates $\hat{\rho}_{ik}\in(1\pm\varepsilon')\frac{\mu_R(\omega_i)}{\mu_R(\omega_k)}$ and $\hat{\rho}_{jk}\in(1\pm\varepsilon')\frac{\mu_R(\omega_j)}{\mu_R(\omega_k)}$. So,

$$\hat{\rho}_{ij} = \frac{\hat{\rho}_{ik}}{\hat{\rho}_{jk}} \geq \frac{\frac{\mu_R(\omega_i)}{\mu_R(\omega_k)}(1-\varepsilon')}{\frac{\mu_R(\omega_j)}{\mu_R(\omega_k)}(1+\varepsilon')} \geq \frac{\mu_R(\omega_i)}{\mu_R(\omega_j)}(1-2\varepsilon')$$

and

$$\hat{\rho}_{ij} = \frac{\hat{\rho}_{ik}}{\hat{\rho}_{jk}} \leq \frac{\frac{\mu_R(\omega_i)}{\mu_R(\omega_k)}(1+\varepsilon')}{\frac{\mu_R(\omega_j)}{\mu_R(\omega_k)}(1-\varepsilon')} \leq \frac{\mu_R(\omega_i)}{\mu_R(\omega_j)}(1+4\varepsilon')$$

given $\varepsilon' \le \frac{1}{2}$. The multiplicative error of $\hat{\rho}_{ij}$ is thus at most $4\varepsilon'$.

Algorithm 2 needs twice the running time of Algorithm 1, which is at most $2(\frac{1}{p_0}+m)\log_2 \frac{1}{Gp_0\tau}$ periods in expectation by Lemma 1. □

Lemma 17. *Suppose* $\frac{1}{Gp_0}\big(\tau + \frac{\gamma^m}{(1-\gamma)G^3p_0^3}\big) \le \frac{1}{2}$.

- *The estimation* $\hat{\mu}$ *returned by Algorithm 3 satisfies* $\|\hat{\mu} - \mu_R\|_1 \le \frac{12|\Omega|}{Gp_0^3}\big(\tau + \frac{\gamma^m}{(1-\gamma)G^3p_0^3}\big)$.
- *The expected running time of Algorithm 3 is at most* $2(|\Omega|-1)(\frac{1}{p_0}+m)\log_2\frac{1}{Gp_0\tau}$ *periods.*

Proof. By Lemma 16, the estimation $\hat{\rho}_{i1}$ satisfies $\hat{\rho}_{i1} \in (1 \pm 4\varepsilon')\frac{\mu_R(\omega_i)}{\mu_R(\omega_1)}$ with $\varepsilon' = \frac{1}{Gp_0}\big(\tau + \frac{\gamma^m}{(1-\gamma)G^3p_0^3}\big)$. Because $\frac{\mu_R(\omega_i)}{\mu_R(\omega_1)} \le \frac{1}{p_0}$, we have

$$\left|\hat{\rho}_{i1} - \frac{\mu_R(\omega_i)}{\mu_R(\omega_1)}\right| \le \frac{4\varepsilon'}{p_0}.$$

Since μ_R is a probability distribution, we have $1 = \sum_{\omega\in\Omega}\mu_R(\omega) = \mu_R(\omega_1) + \mu_R(\omega_1)\sum_{i=2}^{|\Omega|}\frac{\mu_R(\omega_i)}{\mu_R(\omega_1)}$, so

$$\mu_R(\omega_1) = \frac{1}{1+\sum_{i=2}^{|\Omega|}\frac{\mu_R(\omega_i)}{\mu_R(\omega_1)}}.$$

Because the function $f(x) = \frac{1}{1+x}$ is 1-Lipschitz ($|f'(x)| = \frac{1}{(1+x)^2} \le 1$), we have

$$|\hat{\mu}(\omega_1) - \mu_R(\omega_1)| = \left|\frac{1}{1+\sum_{i=2}^{|\Omega|}\hat{\rho}_{i1}} - \frac{1}{1+\sum_{i=2}^{|\Omega|}\frac{\mu_R(\omega_i)}{\mu_R(\omega_1)}}\right| \le \left|\sum_{i=2}^{|\Omega|}\hat{\rho}_{i1} - \sum_{i=2}^{|\Omega|}\frac{\mu_R(\omega_i)}{\mu_R(\omega_1)}\right| \le |\Omega|\frac{4\varepsilon'}{p_0}.$$

Then, for every $i = 2, \dots, |\Omega|$.

$$\begin{aligned}
|\hat{\mu}(\omega_i) - \mu_R(\omega_i)| &= \left|\hat{\rho}_{i1}\hat{\mu}(\omega_1) - \frac{\mu_R(\omega_i)}{\mu_R(\omega_1)}\mu_R(\omega_1)\right| \\
&\le \left|\hat{\rho}_{i1} - \frac{\mu_R(\omega_i)}{\mu_R(\omega_1)}\right|\hat{\mu}(\omega_1) \;+\; \frac{\mu_R(\omega_i)}{\mu_R(\omega_1)}\left|\hat{\mu}(\omega_1) - \mu_R(\omega_1)\right| \\
&\le \frac{4\varepsilon'}{p_0}\hat{\mu}(\omega_1) \;+\; \frac{\mu_R(\omega_i)}{\mu_R(\omega_1)}|\Omega|\frac{4\varepsilon'}{p_0}.
\end{aligned}$$

Therefore,

$$\begin{aligned}\|\hat{\mu}-\mu_R\|_1 &= |\hat{\mu}(\omega_1)-\mu_R(\omega_1)| + \sum_{i=2}^{|\Omega|} |\hat{\mu}(\omega_i)-\mu_R(\omega_i)| \\ &\le |\Omega|\tfrac{4\varepsilon'}{p_0} + |\Omega|\tfrac{4\varepsilon'}{p_0}\hat{\mu}(\omega_1) + |\Omega|\tfrac{4\varepsilon'}{p_0}\sum_{i=2}^{|\Omega|}\tfrac{\mu_R(\omega_i)}{\mu_R(\omega_1)} \\ &\le |\Omega|\tfrac{8\varepsilon'}{p_0} + |\Omega|\tfrac{4\varepsilon'}{p_0}\tfrac{1}{p_0} \\ &\le \tfrac{12|\Omega|\varepsilon'}{p_0^2} \\ &= \tfrac{12|\Omega|}{Gp_0^3}\big(\tau + \tfrac{\gamma^m}{(1-\gamma)G^3p_0^3}\big),\end{aligned}$$

which proves the first claim of Lemma 17.

Because Algorithm 3 runs Algorithm 2 for $|\Omega|-1$ times and the expected running time of Algorithm 2 is at most $2(\frac{1}{p_0}+m)\log_2\frac{1}{Gp_0\tau}$ periods (by Lemma 16), the expected running time of Algorithm 2 is at most

$$2(|\Omega|-1)(\tfrac{1}{p_0}+m)\log_2\tfrac{1}{Gp_0\tau}.$$

periods. □

Proof of Lemma 3. Lemma 3 follows from Lemma 17 by plugging in

$$\tau = \frac{Gp_0^3\varepsilon}{24|\Omega|}, \quad m = \Big\lceil \frac{1}{\ln(1/\gamma)}\ln\big(\frac{24|\Omega|}{(1-\gamma)G^4p_0^6\varepsilon}\big)\Big\rceil.$$

□

B.4 Proof of Lemma 4

Let $\hat{\mu}\in\Delta(\Omega)$ be a receiver prior satisfying $\min_{\omega\in\Omega}\hat{\mu}(\omega)\ge p_0>0$. Let $B_1(\hat{\mu},\varepsilon)=\{\mu : \|\mu-\hat{\mu}\|_1\le\varepsilon\}$ be the set of priors within distance ε to $\hat{\mu}$. Suppose $\varepsilon\le\frac{p_0^2D}{4}$. Let π be a direct signaling scheme. We want to construct another direct signaling scheme $\tilde{\pi}$ in two steps: (1) First, construct a non-direct signaling scheme $\tilde{\pi}^\circ$. (2) Then, convert $\tilde{\pi}^\circ$ into a direct signaling scheme $\tilde{\pi}$ that still satisfies the requirements of the lemma.

Step (1): Construct a non-direct signaling scheme $\tilde{\pi}^\circ$. Let δ satisfy

$$\tfrac{2\varepsilon}{p_0D}\le\delta\le\tfrac{p_0}{2}.$$

Let $P_{\hat{\mu},\pi}(a) = \sum_{\omega\in\Omega} \hat{\mu}(a)\pi(a|\omega)$ be the unconditional probability that π sends signal a under prior $\hat{\mu}$. Let $\hat{\mu}_{a,\pi} \in \Delta(\Omega)$ be the posterior belief induced by signal a under signaling scheme π and prior $\hat{\mu}$:

$$\hat{\mu}_{a,\pi}(\omega) = \frac{\hat{\mu}(\omega)\pi(a|\omega)}{P_{\hat{\mu},\pi}(a)}, \quad \forall \omega \in \Omega.$$

According to Assumption 2, there exists a belief $\eta_a \in \Delta(\Omega)$ for which $\mathbb{E}_{\omega\sim\eta_a}[v(a,\omega) - v(a',\omega)] \geq D, \forall a' \neq a$. Consider the convex combination of $\hat{\mu}_{a,\pi}$ and η_a with coefficients $1-\delta, \delta$:

$$\xi_a = (1-\delta)\hat{\mu}_{a,\pi} + \delta\eta_a. \tag{38}$$

By the linearity of expectation, we have

$$\begin{aligned}\mathbb{E}_{\omega\sim\xi_a}[v(a,\omega) - v(a',\omega)] &= (1-\delta)\mathbb{E}_{\omega\sim\hat{\mu}_{a,\pi}}[v(a,\omega) - v(a',\omega)] + \delta\mathbb{E}_{\omega\sim\eta_a}[v(a,\omega) - v(a',\omega)] \\ &\geq (1-\delta)\mathbb{E}_{\omega\sim\hat{\mu}_{a,\pi}}[v(a,\omega) - v(a',\omega)] + \delta D. \end{aligned} \tag{39}$$

Let $\xi = \sum_{a\in A} P_{\hat{\mu},\pi}(a)\xi_a \in \Delta(\Omega)$, and write $\hat{\mu}$ as the convex combination of ξ and another belief $\chi \in \Delta(\Omega)$, with some coefficient $y \in [0,1]$:

$$\hat{\mu} \;=\; (1-y)\xi + y\chi \;=\; \sum_{a\in A}(1-y)P_{\hat{\mu},\pi}(a)\xi_a \;+\; y\chi. \tag{40}$$

Lemma 18 (Proposition 1 of Zu, Iyer, and Xu, 2021)**.** *If $\delta \leq \frac{p_0}{2}$, then there exist χ on the boundary of $\Delta(\Omega)$ and $y \leq \frac{\delta}{p_0} \leq \frac{1}{2}$ that satisfy* (40)*.*

Since (40) is a convex decomposition of the prior $\hat{\mu}$, according to Kamenica and Gentzkow, 2011, there exists a signaling scheme $\tilde{\pi}^\circ$ that induces posterior ξ_a with total probability $(1-y)P_{\hat{\mu},\pi}(a)$, $\forall a \in A$, and induces the posterior that puts all probability on ω with total probability $y\chi(\omega)$, $\forall \omega \in \Omega$. Namely, $\tilde{\pi}^\circ$ has signal space $S = A \cup \Omega$ and conditional probability

$$\tilde{\pi}^\circ(s|\omega) = \begin{cases} \frac{(1-y)P_{\hat{\mu},\pi}(a)\xi_a(\omega)}{\hat{\mu}(\omega)} & \text{for } s = a \in A; \\ \frac{y\chi(\omega)}{\hat{\mu}(\omega)} & \text{for } s = \omega \in \Omega; \\ 0 & \text{otherwise.} \end{cases} \tag{41}$$

It is not hard to verify that, under prior $\hat{\mu}$ and signaling scheme $\tilde{\pi}^\circ$, the posterior induced by signal $a \in A$ is equal to ξ_a, and the posterior induced by signal ω is the deterministic distribution on ω.

Claim 3. *For any prior of the receiver $\mu \in B_1(\hat{\mu},\varepsilon)$, given any action recommendation $a \in A$*

from $\tilde{\pi}^\circ$, the receiver's posterior $\mu_{a,\tilde{\pi}^\circ}$ satisfies

$$\mathbb{E}_{\omega\sim\mu_{a,\tilde{\pi}^\circ}}[v(a,\omega)-v(a',\omega)] \geq (1-\delta)\mathbb{E}_{\omega\sim\hat{\mu}_{a,\pi}}[v(a,\omega)-v(a',\omega)] + \delta D - \tfrac{2\varepsilon}{p_0},\ \forall a' \in A\setminus\{a\}.$$

And when $\tilde{\pi}^\circ$ sends signal $\omega\in\Omega$, the receiver will believe that the state is ω deterministically.

Proof. Under signaling scheme $\tilde{\pi}^\circ$, let $\mu_{a,\tilde{\pi}^\circ}$, ξ_a be the posteriors induced by signal a from priors μ, $\hat{\mu}$. By the continuity of posterior (Lemma 14), we have $\|\mu_{a,\tilde{\pi}^\circ}-\xi_a\|_1 \leq \frac{2}{p_0}\|\mu-\hat{\mu}\|_1 \leq \frac{2\varepsilon}{p_0}$. Then, since the receiver's utility is in $[0,1]$, for any action $a'\neq a$, we have

$$\begin{aligned}\mathbb{E}_{\omega\sim\mu_{a,\tilde{\pi}^\circ}}[v(a,\omega)-v(a',\omega)] \ &\geq\ \mathbb{E}_{\omega\sim\xi_a}[v(a,\omega)-v(a',\omega)] - \tfrac{2\varepsilon}{p_0}\\ \text{by (39)}\ &\geq\ (1-\delta)\mathbb{E}_{\omega\sim\hat{\mu}_{a,\pi}}[v(a,\omega)-v(a',\omega)] + \delta D - \tfrac{2\varepsilon}{p_0}.\end{aligned}$$

□

We show that the recommendation part of signaling scheme $\tilde{\pi}^\circ$ is close to π, in the following sense:

Claim 4. *For every $\omega\in\Omega$, the ℓ_1 distance $\|\tilde{\pi}^\circ(\cdot|\omega)-\pi(\cdot|\omega)\|_1 = \sum_{a\in A}\left|\tilde{\pi}^\circ(a|\omega)-\pi(a|\omega)\right| \leq \frac{3\delta}{p_0}$.*

Proof. By definition,

$$\begin{aligned}\sum_{a\in A}\left|\tilde{\pi}^\circ(a|\omega)-\pi(a|\omega)\right| \ &=\ \sum_{a\in A}\left|\frac{(1-y)P_{\hat{\mu},\pi}(a)\xi_a(\omega)}{\hat{\mu}(\omega)} - \frac{P_{\hat{\mu},\pi}(a)\hat{\mu}_a(\omega)}{\hat{\mu}(\omega)}\right|\\ &\leq\ \sum_{a\in A}\left(\frac{P_{\hat{\mu},\pi}(a)}{\hat{\mu}(\omega)}\cdot\left|\xi_a(\omega)-\hat{\mu}_a(\omega)\right| \ +\ y\cdot\frac{P_{\hat{\mu},\pi}(a)\xi_a(\omega)}{\hat{\mu}(\omega)}\right)\\ \text{by (38) and } y\leq\tfrac{1}{2}\ &\leq\ \sum_{a\in A}\left(\frac{P_{\hat{\mu},\pi}(a)}{\hat{\mu}(\omega)}\cdot\delta \ +\ 2y\cdot\frac{(1-y)P_{\hat{\mu},\pi}(a)\xi_a(\omega)}{\hat{\mu}(\omega)}\right).\end{aligned}$$

We have $\sum_{a\in A}P_{\hat{\mu},\pi}(a) = 1$ and $\sum_{a\in A}(1-y)P_{\hat{\mu},\pi}(a)\xi_a(\omega) \leq \hat{\mu}(\omega)$ by (40). So, the above is at most

$$\leq\ \frac{1}{\hat{\mu}(\omega)}\cdot\delta + 2y\cdot 1\ \leq\ \frac{\delta}{p_0}+2\frac{\delta}{p_0}\ =\ \frac{3\delta}{p_0},$$

given $y\leq\frac{\delta}{p_0}$. □

Then, we show that the information designer's utility under scheme $\tilde{\pi}^\circ$ is close to her utility under π:

Claim 5. *The designer's utility $U(\tilde{\pi}^\circ,\rho_{\text{ob}};\mu_D) \geq U(\pi,\rho_{\text{ob}};\mu_D) - \frac{3\delta}{p_0}$.*

Proof. For the non-direct signaling scheme $\tilde{\pi}^\circ$, the notation ρ_{ob} means: when $\tilde{\pi}^\circ$ recommends action, the receiver takes the recommended action, and when $\tilde{\pi}^\circ$ reveals the state, the receiver takes an optimal action for that state. The designer's expected utility is thus

$$\begin{aligned} U(\tilde{\pi}^\circ, \rho_{\text{ob}}; \mu_D) &= \sum_{\omega\in\Omega} \mu_D(\omega)\Big(\sum_{a\in A} \tilde{\pi}^\circ(a|\omega)u(a,\omega) \;+\; \tilde{\pi}^\circ(\omega|\omega)u(a^*_\omega,\omega)\Big) \\ &\geq \sum_{\omega\in\Omega} \mu_D(\omega)\sum_{a\in A} \tilde{\pi}^\circ(a|\omega)u(a,\omega) \\ &\geq U(\pi, \rho_{\text{ob}}; \mu_D) - \tfrac{3\delta}{p_0} \quad \text{by Claim 4 and } U \text{ is 1-Lipschitz in } \pi \text{ (Lemma 13).} \end{aligned}$$

□

Step (2): Convert $\tilde{\pi}^\circ$ into a direct signaling scheme $\tilde{\pi}$. Then, we convert $\tilde{\pi}^\circ$ (whose signal space is $A \cup \Omega$) into a direct signaling scheme $\tilde{\pi}$ (whose signal space is A) by coalescing the signals. Specifically, for each state $\omega \in \Omega$, and action $a \in A$, let $\tilde{\pi}$ send signal a when $\tilde{\pi}^\circ$ sends signal a and when $\tilde{\pi}^\circ$ sends signal ω if the receiver's optimal action $a^*_\omega \in \arg\max_{a\in A} v(a,\omega)$ for state ω equals a:

$$\tilde{\pi}(a|\omega) = \tilde{\pi}^\circ(a|\omega) + \mathbb{1}[a = a^*_\omega] \cdot \tilde{\pi}^\circ(\omega|\omega). \tag{42}$$

By the coalescing lemma 15, when $\tilde{\pi}$ recommends action a, we have: for a receiver with any prior $\mu \in B_1(\hat{\mu}, \varepsilon)$, the posterior $\mu_{a,\tilde{\pi}}$ satisfies

$$\begin{aligned} \mathbb{E}_{\omega\sim\mu_{a,\tilde{\pi}}}\big[v(a,\omega) - v(a',\omega)\big] \;&\geq\; \min\Big\{\mathbb{E}_{\omega\sim\mu_{a,\tilde{\pi}^\circ}}\big[v(a,\omega) - v(a',\omega)\big],\; 0\Big\} \\ \text{(by Claim 3)} \;&\geq\; \min\Big\{(1-\delta)\mathbb{E}_{\omega\sim\hat{\mu}_{a,\pi}}\big[v(a,\omega) - v(a',\omega)\big] + \delta D - \tfrac{2\varepsilon}{p_0},\; 0\Big\}. \end{aligned}$$

We verify that $\tilde{\pi}$ satisfies the condition for the receiver's posterior in Lemma 4. There are two cases:

- If $\mathbb{E}_{\omega\sim\hat{\mu}_{a,\pi}}\big[v(a,\omega) - v(a',\omega)\big] \geq 0$, then because $\delta D - \frac{2\varepsilon}{p_0} \geq 0$ by our condition on δ, we have $\min\big\{(1-\delta)\mathbb{E}_{\omega\sim\hat{\mu}_{a,\pi}}\big[v(a,\omega) - v(a',\omega)\big] + \delta D - \frac{2\varepsilon}{p_0},\; 0\big\} = 0$, so

$$\mathbb{E}_{\omega\sim\mu_{a,\tilde{\pi}}}\big[v(a,\omega) - v(a',\omega)\big] \;\geq\; 0 \;=\; \min\Big\{\mathbb{E}_{\omega\sim\hat{\mu}_{a,\pi}}\big[v(a,\omega) - v(a',\omega)\big] + \delta D - \tfrac{2\varepsilon}{p_0},\; 0\Big\}.$$

- If $\mathbb{E}_{\omega\sim\hat{\mu}_{a,\pi}}\big[v(a,\omega) - v(a',\omega)\big] < 0$, then we have $0 > (1-\delta)\mathbb{E}_{\omega\sim\hat{\mu}_{a,\pi}}\big[v(a,\omega) - v(a',\omega)\big] \geq$

$\mathbb{E}_{\omega\sim\hat{\mu}_{a,\pi}}\big[v(a,\omega)-v(a',\omega)\big]$. So,

$$\begin{aligned}\mathbb{E}_{\omega\sim\mu_{a,\tilde{\pi}}}\big[v(a,\omega)-v(a',\omega)\big] &\geq \min\Big\{(1-\delta)\mathbb{E}_{\omega\sim\hat{\mu}_{a,\pi}}\big[v(a,\omega)-v(a',\omega)\big]+\delta D-\tfrac{2\varepsilon}{p_0},\ 0\Big\}\\ &\geq \min\Big\{\mathbb{E}_{\omega\sim\hat{\mu}_{a,\pi}}\big[v(a,\omega)-v(a',\omega)\big]+\delta D-\tfrac{2\varepsilon}{p_0},\ 0\Big\}.\end{aligned}$$

In both cases, $\tilde{\pi}$ satisfies the condition.

We then verify the condition for the designer's utility, which is $U(\tilde{\pi},\rho_{\mathrm{ob}};\mu_D)\geq U(\pi,\rho_{\mathrm{ob}};\mu_D)-\frac{3\delta}{p_0}$. It directly follows from Claim 5, which shows $U(\tilde{\pi}^\circ,\rho_{\mathrm{ob}};\mu_D)\geq U(\pi,\rho_{\mathrm{ob}};\mu_D)-\frac{3\delta}{p_0}$, and Lemma 15, which shows $U(\tilde{\pi}^\circ,\rho_{\mathrm{ob}};\mu_D)=U(\tilde{\pi},\rho_{\mathrm{ob}};\mu_D)$.

Finally, we verify the closeness condition between $\tilde{\pi}$ and π. For every $\omega\in\Omega$,

$$\begin{aligned}\sum_{a\in A}\big|\tilde{\pi}(a|\omega)-\pi(a|\omega)\big| &\leq \sum_{a\in A}\Big(\big|\tilde{\pi}^\circ(a|\omega)-\pi(a|\omega)\big|+\mathbb{1}[a=a^*_\omega]\cdot\tilde{\pi}^\circ(\omega|\omega)\Big) && \text{by (42)}\\ &\leq \tfrac{3\delta}{p_0}+y\tfrac{\chi(\omega)}{\hat{\mu}(\omega)} && \text{by Claim 4 and Equation (41)}\\ &\leq \tfrac{3\delta}{p_0}+\tfrac{\delta}{p_0^2} && \text{because } y\leq\tfrac{\delta}{p_0} \text{ and } \hat{\mu}(\omega)\geq p_0\\ &\leq \tfrac{4\delta}{p_0^2}.\end{aligned}$$

C Omitted Proofs in Section 4

C.1 Proof of Lemma 5

Fix any feasible designer strategy $\sigma=(\pi_\sigma^{(t)})_{t=1}^T$ and let the receiver play a (history–dependent) best response $\phi^*(\sigma,\mu_R)$. For each period t and realized public history h_t (signals and actions up to $t-1$), $\pi_\sigma^{(t)}$ induces a posterior $\mu_t(\cdot\mid h_t,s_t)$ after signal $s_t\sim\pi_\sigma^{(t)}(\cdot\mid\omega,h_t)$ is sent under μ_R. Let $\mathcal{B}_t(h_t,s_t)\subseteq A$ be the nonempty set of myopic best replies at (h_t,s_t):

$$\mathcal{B}_t(h_t,s_t)\ :=\ \arg\max_{a\in A}\ V_t(a;\mu_t(\cdot\mid h_t,s_t)),$$

and fix a measurable tie–breaking rule τ_t so that $b_t(h_t,s_t):=\tau_t(\mathcal{B}_t(h_t,s_t))\in\mathcal{B}_t(h_t,s_t)$.

Direct, dynamically persuasive transformation. Define a *direct* signal in period t by post–processing the original signal via

$$g_t:\ s_t\mapsto a'_t=b_t(h_t,s_t)\in A,$$

and set $\pi_{\sigma^d}^{(t)}(\cdot\mid\omega,h_t):=\big(g_t\circ\pi_\sigma^{(t)}(\cdot\mid\omega,h_t)\big)$, where $\sigma^d=(\pi_{\sigma^d}^{(t)})_{t=1}^T$ denotes the resulting strategy. Because g_t is a (history–dependent) garbling of s_t, the posterior belief $\mu_t(\cdot\mid h_t,s_t)$ induced by $\pi_\sigma^{(t)}$ is the same posterior belief the receiver holds about states upon receiving the recommendation $a'_t=$

$g_t(s_t)$ under strategy σ^d (Bayes plausibility is preserved by post–processing). By construction, a'_t maximizes $V_t(\cdot;\mu_t)$, so obeying the recommendation is optimal for the receiver at each (h_t, a'_t). Therefore there is a best response ϕ_{ob} under σ^d that *obeys* the recommendation every period. That is, $\phi_{\text{ob}} = \phi^*(\sigma^d, \mu_R)$ and σ^d is dynamically persuasive.

Outcome equivalence and optimality. Consider the outcome path generated by $(\sigma, \phi^*(\sigma, \mu_R))$. At each (h_t, s_t), the action actually taken is some $a_t \in \mathcal{B}_t(h_t, s_t)$. Under $(\sigma^d, \phi_{\text{ob}})$, the action taken at the same (h_t, s_t) is exactly $a'_t = b_t(h_t, s_t) \in \mathcal{B}_t(h_t, s_t)$. Hence, conditional on (ω, h_t, s_t), the action under σ^d is a selection from the same set of best replies as under σ; in particular, the receiver payoff is weakly higher and the distribution over next–period histories (and thus all future posteriors) is preserved. Therefore the designer's expected payoff is unchanged:

$$\mathcal{U}_T(\sigma^d, \phi_{\text{ob}}; \mu_D) \;=\; \mathcal{U}_T(\sigma, \phi^*(\sigma, \mu_R); \mu_D).$$

Applying that transformation to an optimal σ yields a history–dependent, dynamically persuasive direct strategy σ^* that attains the dynamic optimum:

$$\mathcal{U}_T^{**}(\mathcal{I}) \;=\; \mathcal{U}_T(\sigma^*, \phi_{\text{ob}}; \mu_D) \quad \text{with} \quad \phi_{\text{ob}} = \phi^*(\sigma^*, \mu_R).$$

C.2 Proof of Lemma 6

Fix an optimal DDP strategy σ^* (possibly history-dependent) with an obedient receiver strategy ϕ_{ob}, over a finite horizon $t = 1, \ldots, T$. Let the state process $\{\omega_t\}_{t=1}^T$ be i.i.d. with (true) marginal $\lambda \in \Delta(\Omega)$, independent of past play. For each period t and action $a \in A$, define the on-path joint and conditional probabilities induced by $(\sigma, \phi_{\text{ob}})$:

$$Q_t(\omega, a) \;:=\; \Pr_{\sigma, \phi_{\text{ob}}}\big(\omega_t = \omega,\; a_t = a\big), \qquad P_t(a \mid \omega) \;:=\; \Pr_{\sigma, \phi_{\text{ob}}}\big(a_t = a \mid \omega_t = \omega\big),$$

where a_t denotes the action recommended (and, on-path, taken) at t. By the law of total probability, $Q_t(\omega, a) = \lambda(\omega)\, P_t(a \mid \omega)$.

For each t, define a *history-independent* direct scheme $\pi_{\tilde{\sigma}}^{(t)} : \Omega \to \Delta(A)$ by

$$\tilde{\pi}^{(t)}(a \mid \omega) \;:=\; P_t(a \mid \omega) \qquad \text{(choose arbitrarily on } \{\omega : \lambda(\omega) = 0\}).$$

Let $\tilde{\sigma} = (\pi_{\tilde{\sigma}}^{(1)}, \ldots, \pi_{\tilde{\sigma}}^{(T)})$, and keep the same off-path continuation rule as under σ^*: following any deviation, switch to the no-information punishment policy as in σ^* for the remainder of the game.

By construction,

$$\Pr_{\tilde{\sigma},\phi_{\mathrm{ob}}}\big(\omega_t=\omega,\ a_t=a\big) \;=\; \lambda(\omega)\,\pi^{(t)}_{\tilde{\sigma}}(a\mid\omega) \;=\; \lambda(\omega)\,P_t(a\mid\omega) \;=\; Q_t(\omega,a),$$

so the joint law of (ω_t, a_t) is identical under $(\sigma^*, \phi_{\mathrm{ob}})$ and $(\tilde{\sigma}, \phi_{\mathrm{ob}})$ for every t. Because payoffs are additively separable across periods, the designer's total expected payoff is the same under $\tilde{\sigma}$ as under σ^*. Thus, $\tilde{\sigma}$ is optimal if σ is.

Fix a period t and a recommended action a with $\Pr(a_t = a) > 0$. Let μ_R be the receiver's prior. Under $\tilde{\sigma}$, Bayes' rule gives the posterior

$$\mu_t^{\tilde{\sigma}}(\omega\mid a) \;=\; \frac{\mu_R(\omega)\,\pi^{(t)}_{\tilde{\sigma}}(a\mid\omega)}{\sum_{\omega'}\mu_R(\omega')\,\pi^{(t)}_{\tilde{\sigma}}(a\mid\omega')} \;=\; \frac{\mu_R(\omega)\,P_t(a\mid\omega)}{\sum_{\omega'}\mu_R(\omega')\,P_t(a\mid\omega')}.$$

The right-hand side is exactly the *average* of the posteriors induced by $(\sigma^*, \phi_{\mathrm{ob}})$ at all on-path histories that yield recommendation a, weighted by their probability conditional on a (law of iterated expectations). So we have

$$\mu_t^{\tilde{\sigma}}(\cdot\mid a) \;=\; \overline{\mu}_t^{\sigma^*}(\cdot\mid a)$$

where $\overline{\mu}_t^{\sigma^*}(\cdot\mid a)$ is the averaged posterior faced by the receiver under $(\sigma^*, \phi_{\mathrm{ob}})$ when recommendation a is observed.

Let W^g_{t+1} (resp. W^p_{t+1}) denote the receiver's continuation utility from $t{+}1$ onward on the obedience (resp. punishment) path. By our construction of off-path behavior, these continuations are the same under $\tilde{\sigma}$ as under σ:

$$W^{g,\tilde{\sigma}}_{t+1} = W^{g,\sigma^*}_{t+1}, \qquad W^{p,\tilde{\sigma}}_{t+1} = W^{p,\sigma^*}_{t+1}.$$

Therefore, for every deviation $a' \neq a$, the period-t one-shot deviation inequality under $\tilde{\sigma}$,

$$\mathbb{E}_{\omega\sim\mu_t^{\tilde{\sigma}}(\cdot|a)}\big[v(a,\omega)-v(a',\omega)\big] \;+\; \big(W^{g,\tilde{\sigma}}_{t+1}-W^{p,\tilde{\sigma}}_{t+1}\big) \;\geq\; 0,$$

coincides with the corresponding inequality under $(\sigma^*, \phi_{\mathrm{ob}})$, which holds because σ is dynamically persuasive. By the one-shot deviation principle, obedience is optimal at all on-path information sets under $\tilde{\sigma}$ as well.

Combining the above, $\tilde{\sigma}$ is an *optimal* DDP strategy and is *history-independent* by construction.

C.3 Proof of Lemma 8

To begin, we consider the restriction to history-independent DDP.

By Proposition 1, the benchmark

$$\mathcal{U}_T^{**}(\mathcal{I}) = \max_{\sigma} \mathcal{U}_T(\sigma; \phi_{\text{ob}}, \mu_D)$$

is attained by some *history-independent, dynamically persuasive DDP* strategy $\sigma = (\pi_\sigma^{(1)}, \dots, \pi_\sigma^{(T)})$. So we may restrict attention to this class and impose the dynamic persuasiveness constraint in its *posterior* form (14).

We then consider a promise state and two primitive constraints.

For any such σ, define the *promised payoff above the uninformed utility from period t to T* as

$$g_t(\sigma) \;:=\; \sum_{t'=t}^{T} \gamma^{\,t'-t} \Big(V(\pi_\sigma^{(t')}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R) \Big) \quad (\geq 0).$$

Dynamic persuasiveness (14) at period t recommendation a is equivalent to

$$\mathbb{E}_{\omega \sim \mu_R(\cdot|a, \pi_\sigma^{(t)})} \big[v(a, \omega) - v(a', \omega) \big] \;+\; \gamma \cdot g_{t+1}(\sigma) \;\geq\; 0 \qquad \forall a' \in A \text{ and on-path } a, \tag{43}$$

i.e., a *posterior* single-period deviation inequality with continuation buffer $\gamma \cdot g_{t+1}$. Moreover, $g_t(\sigma)$ satisfies the constraint of the receiver's *ex ante* obedience

$$\big(V(\pi_\sigma^{(t)}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R) \big) \;+\; \gamma \cdot g_{t+1}(\sigma) \;\geq\; g_t(\sigma). \tag{44}$$

Then we show that the dynamic programming solution upper bounds the benchmark.

We show by backward induction on t that for every feasible history-independent dynamically persuasive σ,

$$\sum_{t'=t}^{T} U(\pi_\sigma^{(t')}, \rho_{\text{ob}}; \mu_D) \;\leq\; F\big(t,\, g_t(\sigma)\big). \tag{45}$$

For $t = T$, feasibility requires (43) (which reduces to static persuasiveness) and (44) with g_T, which is exactly the constraint of (17a). Thus, we have $U(\pi_\sigma^{(T)}, \rho_{\text{ob}}; \mu_D) \leq F(T, g_T(\sigma))$.

Assume (45) holds for $t{+}1$. At period t, the pair $\big(\pi_\sigma^{(t)},\, g_{t+1}(\sigma)\big)$ satisfies the constraints of the Bellman step (18) with $g = g_t(\sigma)$ by (43) and (44). Therefore,

$$\begin{aligned}
\sum_{t'=t}^{T} U(\pi_\sigma^{(t')}, \rho_{\text{ob}}; \mu_D) \;&=\; U(\pi_\sigma^{(t)}, \rho_{\text{ob}}; \mu_D) \;+\; \sum_{t'=t+1}^{T} U(\pi_\sigma^{(t')}, \rho_{\text{ob}}; \mu_D) \\
&\leq\; U(\pi_\sigma^{(t)}, \rho_{\text{ob}}; \mu_D) \;+\; F\big(t{+}1,\, g_{t+1}(\sigma)\big) \\
&\leq\; F\big(t,\, g_t(\sigma)\big),
\end{aligned}$$

where the last inequality uses the maximization in (18). Taking $t = 1$ and using that $F(1, g)$ is nonincreasing in g (tightening the promise cannot increase designer value), we get

$$\mathcal{U}_T(\sigma; \phi_{\text{ob}}, \mu_D) \;\le\; F(1, g_1(\sigma)) \;\le\; F(1, 0).$$

Maximizing over feasible σ yields $\mathcal{U}_T^{**}(\mathcal{I}) \le F(1, 0)$.

We finally show that the dynamic programming solution is tight.

Let $\{(\pi_{\sigma^*}^{(t)}, g_{t+1}^*)\}_{t=1}^T$ be maximizers in (17a) and (18) along an optimal path starting at $F(1, 0)$ (set $g_1^* := 0$). By definition of $F(t+1, g_{t+1}^*)$, there exists a tail strategy $(\pi_{\sigma^*}^{(t+1)}, \dots, \pi_{\sigma^*}^{(T)})$ delivering the promise g_{t+1}^* and satisfying dynamic persuasiveness from $t+1$ onward. Concatenating these choices yields a history-independent DDP strategy $\sigma^* = (\pi_{\sigma^*}^{(t)})_{t=1}^T$.

We check the feasibility of σ^*: the pair $(\pi^{(t)*}, g_{t+1}^*)$ satisfies (18):

$$V(\pi_{\sigma^*}^{(t)}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R) + \gamma \cdot g_{t+1}^* \ge g_t^*, \tag{46}$$

$$\mathbb{E}_{\omega \sim \mu_R(\cdot|a, \pi_{\sigma^*}^{(t)})}\big[v(a, \omega) - v(a', \omega)\big] + \gamma \cdot g_{t+1}^* \ge 0 \quad \forall a' \in A,\ \forall \text{ on-path } a. \tag{47}$$

Thus promised utilities are kept (46). For dynamic persuasiveness, fix t and an on-path recommendation a. The deviation at period t changes only the *current* action. Then, the DDP strategy switches to the (uninformative) punishment, whose continuation equals the uninformed payoff, while obedience secures an extra $\gamma \cdot g_{t+1}^*$ above that baseline by construction. Thus, the receiver's *obedience minus deviation* payoff is

$$\underbrace{\mathbb{E}[v(a, \omega) - v(a', \omega) \mid a]}_{\text{current gap}} + \underbrace{\gamma\, g_{t+1}^*}_{\text{continuation gap}} \;\ge\; 0 \quad \text{(by (47))},$$

so obedience is optimal at every on-path history, and thus σ^* is dynamically persuasive.

Finally, by the Bellman equalities,

$$\mathcal{U}_T(\sigma^*; \phi_{\text{ob}}, \mu_D) = \sum_{t=1}^{T} U(\pi_{\sigma^*}^{(t)}, \rho_{\text{ob}}; \mu_D) = F(1, 0).$$

Combining with the dynamic programming solution upper bounding the benchmark, we obtain $\mathcal{U}_T^{**}(\mathcal{I}) = F(1, 0)$, attained by the history-independent, dynamically persuasive DDP strategy σ^*.

C.4 Proof of Proposition 4

Proposition 4. *For any instance $\mathcal{I}$ and any $T \ge 1$, the designer's dynamic benchmark $\mathcal{U}_T^{**}(\mathcal{I})$ is weakly increasing in the receiver's discount factor $\gamma \in [0, 1)$.*

Proof. Fix $\gamma' < \gamma''$. Consider any DDP strategy σ that is dynamically persuasive (i.e., induces obedience) under γ', together with the receiver's obedient strategy ϕ_{ob}. The one-step obedience constraints compare current recommended action payoffs with any deviation payoffs *plus* the discounted continuation values. When γ increases from γ' to γ'', the continuation part of these constraints is (weakly) upweighted, making obedience *easier* to sustain. Hence any $(\sigma, \phi_{\text{ob}})$ feasible at γ' remains feasible at γ''.

Therefore the feasible set of DDP+obedience outcomes at γ'' contains that at γ', so the designer's optimal value cannot decrease:

$$\mathcal{U}_T^{**}(\mathcal{I}; \gamma') \;\le\; \mathcal{U}_T^{**}(\mathcal{I}; \gamma'').$$

Since γ', γ'' were arbitrary with $\gamma' < \gamma''$, $\mathcal{U}_T^{**}(\mathcal{I})$ is weakly increasing in γ. □

C.5 Proof of Proposition 3

Proof. Let π^* be an optimal signaling scheme that solves problem (22), with corresponding designer payoff $U(\pi^*, \rho_{\text{ob}}; \mu_D) = U_{IR}^*$.

Proof of upper bound $\frac{1}{T}\mathcal{U}_T^{**}(\mathcal{I}; \gamma) \le U_{IR}^*$. By Proposition 1, the dynamic benchmark $\mathcal{U}_T^{**}(\mathcal{I}; \gamma)$ can be achieved by some history-independent DDP strategies $\sigma = (\pi^{(t)})_{t=1}^T$ that is dynamically persuasiveness. Define $g_t = \sum_{t'=t}^T \gamma^{t'-t}\big(V(\pi^{(t')}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R)\big)$. By dynamic persuasiveness, we have $g_t \ge 0$. We also note that

$$V(\pi^{(t)}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R) + \gamma g_{t+1} = g_t. \tag{48}$$

Let $\bar{\pi} := \frac{1}{T}\sum_{t=1}^T \pi^{(t)}$ be the average signaling scheme of the DDP strategy σ. We will show $\sum_{t=1}^T \big[V(\pi^{(t)}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R)\big] \ge 0$. To see that, first, summing Eq. (48) over $t = 1, ..., T$,

$$\sum_{t=1}^T \big[V(\pi^{(t)}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R)\big] + \gamma \sum_{t=1}^T g_{t+1} = \sum_{t=1}^T g_t.$$

Rearranging,

$$\sum_{t=1}^T \big[V(\pi^{(t)}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R)\big] = \sum_{t=1}^T g_t - \gamma \sum_{t=1}^T g_{t+1} = g_1 + (1-\gamma)\sum_{t=2}^T g_t.$$

Because $g_1, g_t \ge 0$ and $\gamma \le 1$, we have

$$\sum_{t=1}^T \big[V(\pi^{(t)}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R)\big] \ge 0.$$

This means that the average signaling scheme $\overline{\pi}$ satisfies the receiver's individual rationality constraint in optimization problem (22), so it is a feasible solution, and the designer's payoff is at most

$$U(\overline{\pi}, \rho_{\text{ob}}; \mu_D) \leq U^*_{\text{IR}}.$$

By linearity, $U(\overline{\pi}, \rho_{\text{ob}}; \mu_D) = \frac{1}{T}\sum_{t=1}^{T} U(\pi^{(t)}, \rho_{\text{ob}}; \mu_D) = \frac{1}{T}\mathcal{U}_T^{**}(\mathcal{I}; \gamma)$. So $\frac{1}{T}\mathcal{U}_T^{**}(\mathcal{I}; \gamma) \leq U^*_{\text{IR}}$.

Proof of lower bound $\lim_{\gamma\to 1}\lim_{T\to\infty}\frac{1}{T}\mathcal{U}_T^{**}(\mathcal{I}) \geq U^*_{\text{IR}}$. Let π_{full} be the "full-information" signaling scheme that reveals the optimal action a^*_ω to the receiver at every state $\omega \in \Omega$, which gives the receiver expected payoff $V(\pi_{\text{full}}, \rho_{\text{ob}}; \mu_R)$. Under Assumptions 1 and 3, it is easy to show that $V(\pi_{\text{full}}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R) \geq p_0 G$. Let π' be a direct signaling scheme obtained by mixing π^* with weight $1-\delta'$ and π_{full} with weight δ'. Consider the DDP strategy $\sigma' = (\pi')_{t=1}^{T}$ that repeats π' unless the receiver has deviated. We claim that, under σ', the receiver is willing to obey the recommended actions in every period t from 1 to $T - \sqrt{T}$: to see this, given the recommendation at period t, the receiver's obedient payoff from period t to T is at least

$$\gamma^t \cdot 0 + \sum_{t'=t+1}^{T} \gamma^{t'} V(\pi', \rho_{\text{ob}}; \mu_R).$$

If the receiver deviates from the recommendation in period t, he will obtain payoff at most 1 in period t and be punished starting from period $t+1$, so his total payoff is at most

$$\gamma^t \cdot 1 + \sum_{t'=t+1}^{T} \gamma^{t'} V_{\text{uninformed}}(\mu_R).$$

The difference is

$$
\begin{aligned}
&\gamma^t \cdot 0 + \sum_{t'=t+1}^{T} \gamma^{t'} V(\pi', \rho_{\text{ob}}; \mu_R) - \gamma^t \cdot 1 - \sum_{t'=t+1}^{T} \gamma^{t'} V_{\text{uninformed}}(\mu_R) \\
&= -\gamma^t + \sum_{t'=t+1}^{T} \gamma^{t'} \Big[V(\pi', \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R) \Big] \\
&= -\gamma^t + \sum_{t'=t+1}^{T} \gamma^{t'} \Big[(1-\delta') \big(\underbrace{V(\pi^*, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R)}_{\geq 0} \big) + \delta' \big(\underbrace{V(\pi_{\text{full}}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R)}_{\geq p_0 G} \big) \Big] \\
&\geq -\gamma^t + \sum_{t'=t+1}^{T} \gamma^{t'} \delta' p_0 G \\
&= -\gamma^t + \delta' p_0 G \cdot \gamma^{t+1} \cdot \tfrac{1-\gamma^{T-t}}{1-\gamma} \\
&\geq 0
\end{aligned}
$$

by choosing $\delta' = \frac{1-\gamma}{p_0 G \gamma (1-\gamma^{T-t})}$. Note that $\delta' \to \frac{1-\gamma}{p_0 G \gamma} \to 0$ as $T \to \infty$ and $\gamma \to 1$. This means that the receiver will obey the recommendations from period 1 to period $T - \sqrt{T}$. So, the designer's average payoff is at least

$$
\begin{aligned}
\frac{1}{T} \mathcal{U}(\sigma', \phi^*(\sigma'); \mu_D) &\geq \frac{1}{T} \Bigg(\sum_{t=1}^{T-\sqrt{T}} U(\pi', \rho_{\text{ob}}; \mu_D) \;\; + \;\; 0 \Bigg) \\
&\geq \frac{1}{T} \Bigg(\sum_{t=1}^{T-\sqrt{T}} (1-\delta') U(\pi^*, \rho_{\text{ob}}; \mu_D) \;\; + \;\; 0 \Bigg) \\
&= \frac{T-\sqrt{T}}{T} (1-\delta') \cdot U(\pi^*, \rho_{\text{ob}}; \mu_D) \\
&\to U(\pi^*, \rho_{\text{ob}}; \mu_D)
\end{aligned}
$$

as $T \to \infty$ and $\gamma \to 1$, where the last step is because $\delta' \to 0$. □

D Omitted Proofs in Section 5

D.1 Proof of Lemma 9

Proof. Let π_{full} be the "full revelation" signaling scheme that reveals the optimal action $a^*_\omega = \max_{a \in A} v(a, \omega)$ for the receiver for every state $\omega \in \Omega$. By Assumption 3, π_{full} is unique. Let $\|\pi_{\hat{\sigma}} - \pi_{\text{full}}\| = \max_{\omega \in \Omega} \|\pi_{\hat{\sigma}}(\cdot|\omega) - \pi_{\text{full}}(\cdot|\omega)\|_1$ be the matrix ℓ_1-distance between direct signaling schemes π and π_{full}. Let $\theta = \sqrt{\frac{8\delta}{p_0^3 G}}$.

Case (1): $\|\pi_{\hat{\sigma}} - \pi_{\text{full}}\| \leq \theta$. In this case, we let $\pi_{\tilde{\sigma}}$ be π_{full}. Because π_{full} is persuasive regardless

of the receiver's prior, we immediately have

$$\mathbb{E}_{\omega\sim\mu(\cdot|a,\pi_{\tilde{\sigma}})}\big[v(a,\omega)-v(a',\omega)\big] \;\geq\; 0 \;\geq\; \min\big\{\mathbb{E}_{\omega\sim\hat{\mu}(\cdot|a,\pi_{\hat{\sigma}})}\big[v(a,\omega)-v(a',\omega)\big]+\delta D-\tfrac{2\varepsilon}{p_0},\;0\big\},$$

which satisfies the first claim of Lemma 9. The receiver's utility satisfies $V(\pi_{\tilde{\sigma}},\rho_{\text{ob}};\mu_R) \geq V(\pi_{\hat{\sigma}},\rho_{\text{ob}};\mu_R)$ because the full revelation signaling scheme maximizes the receiver's utility. For the designer's utility, by the Lipschitz continuity with respect to π (Lemma 13), we have

$$U(\pi_{\tilde{\sigma}},\rho_{\text{ob}};\mu_D) \;\geq\; U(\pi_{\hat{\sigma}},\rho_{\text{ob}};\mu_D)-\|\pi_{\text{full}}-\pi_{\hat{\sigma}}\| \;\geq\; U(\pi_{\hat{\sigma}},\rho_{\text{ob}};\mu_D)-\theta,$$

which satisfies the third claim of Lemma 9.

Case (2): $\|\pi_{\hat{\sigma}}-\pi_{\text{full}}\|>\theta$. In this case, we apply static robustification (Lemma 4) to π to obtain direct signaling scheme $\pi_{\tilde{\sigma}}$. According to the third claim of Lemma 4, $\|\pi_{\tilde{\sigma}}-\pi_{\hat{\sigma}}\|\leq\frac{4\delta}{p_0^2}$, so by the Lipschitz continuity of the receiver's utility (Lemma 13),

$$V(\pi_{\tilde{\sigma}},\rho_{\text{ob}};\mu_R) \;\geq\; V(\pi_{\hat{\sigma}},\rho_{\text{ob}};\mu_R)-\tfrac{4\delta}{p_0^2}.$$

Consider the convex combination of $\pi_{\tilde{\sigma}}$ and π_{full} with weights $1-\alpha$, α, denoted by $\pi_{\tilde{\sigma},\alpha}$. By the coalescing lemma (Lemma 15), $\tilde{\pi}_\alpha$ automatically satisfies the first claim of Lemma 9. By linearity, the receiver's utility satisfies

$$\begin{aligned}
&V(\pi_{\tilde{\sigma},\alpha},\rho_{\text{ob}};\mu_R)-V(\pi_{\hat{\sigma}},\rho_{\text{ob}};\mu_R)\\
&=\;(1-\alpha)\Big(V(\pi_{\tilde{\sigma}},\rho_{\text{ob}};\mu_R)-V(\pi_{\hat{\sigma}},\rho_{\text{ob}};\mu_R)\Big)\;+\;\alpha\Big(V(\pi_{\text{full}},\rho_{\text{ob}};\mu_R)-V(\pi_{\hat{\sigma}},\rho_{\text{ob}};\mu_R)\Big)\\
&\geq\;-(1-\alpha)\tfrac{4\delta}{p_0^2}\;+\;\alpha\cdot\tfrac{p_0G\theta}{2}\qquad\text{by Claim 6}\\
&\geq\;0
\end{aligned}$$

where we choose $\alpha=\frac{8\delta}{p_0^3G\theta}=\sqrt{\frac{8\delta}{p_0^3G}}$ ($\alpha\leq 1$ given $\delta\leq\frac{p_0^3G}{8}$). So, $\pi_{\tilde{\sigma},\alpha}$ weakly improves the receiver's utility compared to π, satisfying the second claim of Lemma 9.

The designer's utility under $\pi_{\tilde{\sigma},\alpha}$ is

$$
\begin{aligned}
&U(\pi_{\tilde{\sigma},\alpha}, \rho_{\text{ob}}; \mu_D) \\
&= (1-\alpha)U(\tilde{\pi}, \rho_{\text{ob}}; \mu_D) \;+\; \alpha U(\pi_{\text{full}}, \rho_{\text{ob}}; \mu_D) \\
&\geq (1-\alpha)\Big(U(\pi_{\hat{\sigma}}, \rho_{\text{ob}}; \mu_D) - \tfrac{3\delta}{p_0}\Big) \;+\; 0 && \text{by Lemma 4} \\
&\geq U(\pi_{\hat{\sigma}}, \rho_{\text{ob}}; \mu_D) - \tfrac{3\delta}{p_0} - \sqrt{\tfrac{8\delta}{p_0^3 G}} && \text{given } \alpha = \sqrt{\tfrac{8\delta}{p_0^3 G}} \\
&\geq U(\pi_{\hat{\sigma}}, \rho_{\text{ob}}; \mu_D) - 3\sqrt{\tfrac{\delta}{8p_0}} - \sqrt{\tfrac{8\delta}{p_0^3 G}} && \text{as } \delta \leq \tfrac{p_0^3 G}{8} \implies \tfrac{\delta}{p_0} \leq \tfrac{1}{8} \\
&\geq U(\pi_{\hat{\sigma}}, \rho_{\text{ob}}; \mu_D) - \sqrt{\tfrac{9\delta}{8p_0^3 G}} - \sqrt{\tfrac{8\delta}{p_0^3 G}} && p_0, G \leq 1 \\
&\geq U(\pi_{\hat{\sigma}}, \rho_{\text{ob}}; \mu_D) - 4\sqrt{\tfrac{\delta}{p_0^3 G}},
\end{aligned}
$$

which proves the lemma. □

Claim 6. $\|\pi_{\hat{\sigma}} - \pi_{\text{full}}\| > \theta \implies V(\pi_{\text{full}}, \rho_{\text{ob}}; \mu_R) - V(\pi_{\hat{\sigma}}, \rho_{\text{ob}}; \mu_R) > \frac{p_0 G\theta}{2}$,

Proof. For any state $\omega \in \Omega$, define the difference

$$
\begin{aligned}
\Delta(\omega) := v(a^*_\omega, \omega) - \sum_a \pi_{\hat{\sigma}}(a \mid \omega) v(a, \omega) &= \sum_{a \neq a^*_\omega} \pi_{\hat{\sigma}}(a \mid \omega)\big(v(a^*_\omega, \omega) - v(a, \omega)\big) \\
&\geq G\big(1 - \pi_{\hat{\sigma}}(a^*_\omega \mid \omega)\big) = \tfrac{G}{2}\big\|\pi_{\hat{\sigma}}(\cdot|\omega) - \pi_{\text{full}}(\cdot|\omega)\big\|_1
\end{aligned}
$$

The condition $\|\pi_{\hat{\sigma}} - \pi_{\text{full}}\| > \theta$ implies $\exists\, \bar{\omega} \in \Omega$ with $\|\pi_{\hat{\sigma}}(\cdot|\omega) - \pi_{\text{full}}(\cdot|\omega)\|_1 > \theta$, hence $\Delta(\bar{\omega}) \geq \frac{G\theta}{2}$. Therefore,

$$
V(\pi_{\text{full}}, \rho_{\text{ob}}; \mu_R) - V(\pi_{\hat{\sigma}}, \rho_{\text{ob}}; \mu_R) = \sum_\omega \mu_R(\omega)\Delta(\omega) \geq \mu_R(\bar{\omega})\Delta(\bar{\omega}) \geq \tfrac{p_0 G\theta}{2}.
$$

□

D.2 Proof of Theorem 2

Proof of Theorem 2. We convert $\hat{\sigma} = (\pi^{(1)}_{\hat{\sigma}}, \ldots, \pi^{(T)}_{\hat{\sigma}})$ to $\tilde{\sigma} = (\pi^{(1)}_{\tilde{\sigma}}, \ldots, \pi^{(T)}_{\tilde{\sigma}})$ by applying utility-improving robustification (Lemma 9) to the direct signaling schemes of σ in the backward order: namely, from $\pi^{(T)}$ to $\pi^{(1)}$.

Specifically, for each period $t \leq T$, having robustified $\pi^{(T)}_{\hat{\sigma}}, \ldots, \pi^{(t+1)}_{\hat{\sigma}}$, we robustify $\pi^{(t)}_{\hat{\sigma}}$ so that $(\pi^{(t)}_{\tilde{\sigma}}, \pi^{(t+1)}_{\tilde{\sigma}}, \ldots, \pi^{(T)}_{\tilde{\sigma}})$ is dynamically persuasive with respect to μ_R satisfying $\|\mu_R - \hat{\mu}\|_1 \leq \varepsilon$.

In particular, we want the following to hold: for any action a recommended by $\pi_{\tilde{\sigma}}^{(t)}$ and any $a' \in A$,

$$\gamma^t \mathbb{E}_{\omega\sim\mu_R(\cdot|a,\pi_{\tilde{\sigma}}^{(t)})}\Big[v(a,\omega) - v(a',\omega)\Big] + \sum_{t'=t+1}^{T} \gamma^{t'}\Big(V(\pi_{\tilde{\sigma}}^{(t')}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R)\Big) \geq 0. \quad (49)$$

Applying Lemma 9 to $\pi^{(t)}$ with parameter $\delta_t > 0$ (to be chosen later), we obtain another direct signaling scheme $\pi_{\tilde{\sigma}}^{(t)}$ that satisfies $\mathbb{E}_{\omega\sim\mu_R(\cdot|a,\pi_{\tilde{\sigma}}^{(t)})}\big[v(a,\omega) - v(a',\omega)\big] \geq \min\big\{\mathbb{E}_{\omega\sim\hat{\mu}(\cdot|a,\pi_{\hat{\sigma}}^{(t)})}\big[v(a,\omega) - v(a',\omega)\big] + \delta_t D - \frac{2\varepsilon}{p_0}, 0\big\}$. There are two cases:

Case (1): $\mathbb{E}_{\omega\sim\mu_R(\cdot|a,\pi_{\tilde{\sigma}}^{(t)})}\big[v(a,\omega) - v(a',\omega)\big] \geq 0$*:* In this case, by induction, we have

$$\sum_{t'=t+1}^{T} \gamma^{t'}\Big(V(\pi_{\tilde{\sigma}}^{(t')}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R)\Big) \geq 0$$

because the robustified DDP strategy $(\pi_{\tilde{\sigma}}^{(t')})_{t'=t+1}^{T}$ is dynamically persuasive. So, we immediately obtain (49).

Case (2): $\mathbb{E}_{\omega\sim\mu_R(\cdot|a,\pi_{\tilde{\sigma}}^{(t)})}\big[v(a,\omega) - v(a',\omega)\big] < 0$*:* in this case, by the first claim of Lemma 9,

$$0 \;>\; \mathbb{E}_{\omega\sim\mu_R(\cdot|a,\pi_{\tilde{\sigma}}^{(t)})}\big[v(a,\omega) - v(a',\omega)\big] \;\geq\; \mathbb{E}_{\omega\sim\hat{\mu}(\cdot|a,\pi_{\hat{\sigma}}^{(t)})}\big[v(a,\omega) - v(a',\omega)\big] + \delta_t D - \tfrac{2\varepsilon}{p_0}. \quad (50)$$

We then consider the receiver's continuation utility from period $t+1$ to T.

$$\begin{aligned}
&\sum_{t'=t+1}^{T} \gamma^{t'}\Big(V(\pi_{\tilde{\sigma}}^{(t')}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R)\Big) \\
&\geq \sum_{t'=t+1}^{T} \gamma^{t'}\Big(V(\pi_{\hat{\sigma}}^{(t')}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R)\Big) && \text{by the second claim of Lemma 9} \\
&\geq \sum_{t'=t+1}^{T} \gamma^{t'}\Big(V(\pi_{\hat{\sigma}}^{(t')}, \rho_{\text{ob}}; \hat{\mu}) - V_{\text{uninformed}}(\hat{\mu}) - 2\|\mu_R - \hat{\mu}\|_1\Big) && V \text{ is 1-Lipschitz (Lemma 13)} \\
&\geq \sum_{t'=t+1}^{T} \gamma^{t'}\Big(V(\pi_{\hat{\sigma}}^{(t')}, \rho_{\text{ob}}; \hat{\mu}) - V_{\text{uninformed}}(\hat{\mu})\Big) - \tfrac{\gamma^{t+1}}{1-\gamma} 2\varepsilon. && (51)
\end{aligned}$$

Adding (50) and (51),

$$\begin{aligned}
&\gamma^t \mathbb{E}_{\omega\sim\mu_R(\cdot|a,\pi^{(t)}_{\tilde\sigma})}\Big[v(a,\omega)-v(a',\omega)\Big] + \sum_{t'=t+1}^{T}\gamma^{t'}\Big(V(\pi^{(t')}_{\tilde\sigma},\rho_{\mathrm{ob}};\mu_R)-V_{\mathrm{uninformed}}(\mu_R)\Big)\\
&\geq \underbrace{\gamma^t \mathbb{E}_{\omega\sim\hat\mu(\cdot|a,\pi^{(t)}_{\hat\sigma})}\big[v(a,\omega)-v(a',\omega)\big] + \sum_{t'=t+1}^{T}\gamma^{t'}\Big(V(\pi^{(t')}_{\hat\sigma},\rho_{\mathrm{ob}};\hat\mu)-V_{\mathrm{uninformed}}(\hat\mu)\Big)}_{\geq 0 \text{ because } (\pi^{(t)}_{\hat\sigma},\dots,\pi^{(T)}_{\hat\sigma}) \text{ is dynamically persuasive for } \hat\mu}\\
&\quad + \gamma^t\delta_t D - \gamma^t\tfrac{2\varepsilon}{p_0} - \tfrac{\gamma^{t+1}}{1-\gamma}2\varepsilon\\
&\geq 0
\end{aligned}$$

where we choose $\delta_t = \left(\frac{1}{p_0}+\frac{\gamma}{1-\gamma}\right)\frac{2\varepsilon}{D}$. Thus, $(\pi^{(t)}_{\tilde\sigma},\pi^{(t+1)}_{\tilde\sigma},\dots,\pi^{(T)}_{\tilde\sigma})$ is dynamically persuasive for a receiver with prior μ_R. Performing the robustification from $t=T$ to $t=1$, we therefore obtain a DDP strategy $\tilde\sigma=(\pi^{(1)}_{\tilde\sigma},\dots,\pi^{(T)}_{\tilde\sigma})$ that is dynamically persuasive for a receiver with prior μ_R.

The designer's T-period total utility under the original DDP strategy $\sigma=(\pi^{(1)},\dots,\pi^{(T)})$ is $\mathcal{U}_T(\sigma,\phi_{\mathrm{ob}};\mu_D)=\sum_{t=1}^{T}U(\pi^{(t)},\rho_{\mathrm{ob}};\mu_D)$. Under the robustified DDP strategy $\tilde\sigma=(\pi^{(1)}_{\tilde\sigma},\dots,\pi^{(T)}_{\tilde\sigma})$, the designer's utility is

$$\begin{aligned}
\mathcal{U}_T(\tilde\sigma,\phi_{\mathrm{ob}};\mu_D) &= \sum_{t=1}^{T}U(\pi^{(t)}_{\tilde\sigma},\rho_{\mathrm{ob}};\mu_D)\\
&\geq \sum_{t=1}^{T}\left(U(\pi^{(t)},\rho_{\mathrm{ob}};\mu_D)-4\sqrt{\tfrac{\delta_t}{p_0^3 G}}\right) && \text{by the third claim of Lemma 9}\\
&= \mathcal{U}_T(\sigma,\phi_{\mathrm{ob}};\mu_D)-4T\sqrt{\left(\tfrac{1}{p_0}+\tfrac{\gamma}{1-\gamma}\right)\tfrac{2\varepsilon}{p_0^3 GD}},
\end{aligned}$$

which proves Theorem 2. □

D.3 Proof of Lemma 10

Proof. Let $\sigma^*_T=(\pi^{(1)}_{\sigma^*},\dots,\pi^{(T)}_{\sigma^*})$ be an optimal history-independent DDP strategy for the T-period game, which returns $\mathcal{U}^{**}_T(\mu_D,\hat\mu)$.

Let $\sigma^{\mathrm{tail}}=(\pi^{(T-T_0+1)}_{\sigma^*},\dots,\pi^{(T)}_{\sigma^*})$ be the tail strategy for σ^*_T. Because the DDP strategy ensures the receiver's obedience through the continuation incentives that depend only on the future, truncating the first T_0 periods does not affect incentives from period T_0+1 onward. Thus σ^{tail} is a feasible $(T-T_0)$-period DDP strategy for the same prior.

On the obedient path, the designer's per-period payoff is $U(\pi^{(t)},\rho_{\mathrm{ob}};\mu_D)\in[0,1]$. Therefore, the first T_0 periods contribute at most T_0 to the total:

$$\mathcal{U}_T^{**}(\mu_D,\hat{\mu}) = \sum_{t=1}^{T_0} U(\pi_{\sigma^*}^{(t)},\rho_{\text{ob}};\mu_D) + \sum_{t=T_0+1}^{T} U(\pi_{\sigma^*}^{(t)},\rho_{\text{ob}};\mu_D) \;\le\; T_0 \;+\; U(\sigma^{\text{tail}},\rho_{\text{ob}};\mu_D)$$

Maximizing over all feasible $(T-T_0)$-period strategies gives

$$\mathcal{U}_{T-T_0}^{**}(\mu_D,\hat{\mu}) \;\ge\; U(\sigma^{\text{tail}},\rho_{\text{ob}};\mu_D) \;\ge\; \mathcal{U}_T^{**}(\mu_D,\hat{\mu}) - T_0.$$

□

E Omitted Proofs in Section 6

E.1 Proof of Lemma 11

We summarize the information design instance here:

$$\mathcal{I} = \begin{cases} \text{designer's utility:} & u(a_1,\omega)=1, \quad u(a_0,\omega)=0, \quad \forall \omega\in\Omega=\{\omega_0,\omega_1\}; \\ \text{receiver's utility:} & v(a_0,\omega_0)=0, \quad v(a_0,\omega_1)=0, \quad v(a_1,\omega_0)=-1, \quad v(a_1,\omega_1)=1; \\ \text{designer's prior:} & \mu_D(\omega_0)=1-p_0, \quad \mu_D(\omega_1)=p_0, \quad 0<p_0<1/2; \\ \text{receiver's prior:} & \mu_R(\omega_0)=\frac{1}{1+\tilde{\mu}_R}, \quad \mu_R(\omega_1)=\frac{\tilde{\mu}_R}{1+\tilde{\mu}_R}, \quad 0\le\mu_R\le 1. \end{cases}$$

We present some useful facts before proving Lemma 11. Because $\mu_R(\omega_1)\le\mu_R(\omega_0)$, the receiver's ex-ante utility when receiving no information is

$$V_{\text{uninformed}}(\mu_R) \;=\; \max_{a\in A}\sum_{\omega\in\Omega}\mu_R(\omega)v(a,\omega) \;=\; \max\Big\{\underbrace{0}_{a_0\text{'s utility}}, \underbrace{\mu_R(\omega_1)-\mu_R(\omega_0)}_{a_1\text{'s utility}}\Big\} \;=\; 0. \quad (52)$$

When the designer uses a direct signaling scheme $\pi:\Omega\to\Delta(A)$, the ex-ante utility of an obedient receiver is

$$V(\pi,\rho_{\text{ob}};\mu_R) \;=\; \sum_{\omega\in\Omega}\mu_R(\omega)\sum_{a\in A}\pi(a|\omega)v(a,\omega) \;=\; -\mu_R(\omega_0)\pi(a_1|\omega_0)+\mu_R(\omega_1)\pi(a_1|\omega_1), \quad (53)$$

and the ex-ante utility of the designer is

$$U(\pi,\rho_{\text{ob}};\mu_D) \;=\; \sum_{\omega\in\Omega}\mu_R(\omega)\sum_{a\in A}\pi(a|\omega)u(a,\omega) \;=\; (1-p_0)\pi(a_1|\omega_0)+p_0\pi(a_1|\omega_1). \quad (54)$$

The persuasiveness constraints for a direct signaling scheme π are:

$$\text{(for } a_0) \quad \mathbb{E}_{\omega\sim\mu_R(\cdot|a_0,\pi)}\big[v(a_0,\omega)-v(a_1,\omega)\big] \;=\; \frac{\mu_R(\omega_0)\pi(a_0|\omega_0)-\mu_R(\omega_1)\pi(a_0|\omega_1)}{\mu_R(\omega_0)\pi(a_0|\omega_0)+\mu_R(\omega_1)\pi(a_0|\omega_1)} \;\geq\; 0; \tag{55}$$

$$\text{(for } a_1) \quad \mathbb{E}_{\omega\sim\mu_R(\cdot|a_1,\pi)}\big[v(a_1,\omega)-v(a_0,\omega)\big] \;=\; \frac{-\mu_R(\omega_0)\pi(a_1|\omega_0)+\mu_R(\omega_1)\pi(a_1|\omega_1)}{\mu_R(\omega_0)\pi(a_1|\omega_0)+\mu_R(\omega_1)\pi(a_1|\omega_1)} \;\geq\; 0; \tag{56}$$

The Bayesian persuasion benchmark $U^*_{\mathrm{BP}}(\mathcal{I})$ is the designer's optimal utility from a persuasive direct signaling scheme:

$$\begin{aligned} U^*_{\mathrm{BP}}(\mathcal{I}) \;=\; \max_{\pi:\Omega\to\Delta(A)} \; & U(\pi,\rho_{\mathrm{ob}};\mu_D) \\ \text{s.t.} \quad & \text{(for } a_0) \quad \mu_R(\omega_0)\pi(a_0|\omega_0)-\mu_R(\omega_1)\pi(a_0|\omega_1) \;\geq\; 0; \\ & \text{(for } a_1) \quad -\mu_R(\omega_0)\pi(a_1|\omega_0)+\mu_R(\omega_1)\pi(a_1|\omega_1) \;\geq\; 0. \end{aligned}$$

Claim 7. $U^*_{\mathrm{BP}}(\mathcal{I}) = (1-p_0)\tilde{\mu}_R + p_0$.

Proof. In the $U^*_{\mathrm{BP}}(\mathcal{I})$ optimization problem, the persuasiveness constraint for a_0

$$\begin{aligned} \mu_R(\omega_0)\pi(a_0|\omega_0)-\mu_R(\omega_1)\pi(a_0|\omega_1) &= \mu_R(\omega_0)\big(1-\pi(a_1|\omega_0)\big)-\mu_R(\omega_1)\big(1-\pi(a_1|\omega_1)\big) \\ &= \underbrace{\mu_R(\omega_0)-\mu_R(\omega_1)}_{\geq 0 \text{ by construction}} \; \underbrace{-\mu_R(\omega_0)\pi(a_1|\omega_0)+\mu_R(\omega_1)\pi(a_1|\omega_1)}_{\geq 0 \text{ when } a_1 \text{ is persuasive}} \\ &\geq 0 \end{aligned}$$

is automatically satisfied when a_1 is persuasive. So, we can drop the persuasiveness constraint for a_0 and write the optimization problem as

$$\begin{aligned} U^*_{\mathrm{BP}}(\mathcal{I}) \;=\; \max_{\pi:\Omega\to\Delta(A)} \; & (1-p_0)\pi(a_1|\omega_0)+p_0\pi(a_1|\omega_1) \\ \text{s.t.} \quad & -\mu_R(\omega_0)\pi(a_1|\omega_0)+\mu_R(\omega_1)\pi(a_1|\omega_1) \;\geq\; 0. \end{aligned}$$

The solution to the above optimization problem is clearly $\pi(a_1|\omega_1)=1$ and $\pi(a_1|\omega_0)=\frac{\mu_R(\omega_1)}{\mu_R(\omega_0)}=\tilde{\mu}_R$, with optimal objective value $U^*_{\mathrm{BP}}(\mathcal{I})=(1-p_0)\tilde{\mu}_R+p_0$. □

To prove Lemma 11, we analyze the designer's global benchmark $\mathcal{U}^{**}_T(\mathcal{I})$. According to Lemma **??**, the global benchmark $\mathcal{U}^{**}_T(\mathcal{I})$ can be achieved by some history-independent dynami-

cally persuasive DDP strategy $\sigma = (\pi^{(1)}, \ldots, \pi^{(T)})$:

$$\mathcal{U}_T^{**}(\mathcal{I}) \;=\; \max_{\sigma\,=\,(\pi^{(1)},\ldots,\pi^{(T)})} \;\sum_{t=1}^{T} U(\pi^{(t)}, \rho_{\text{ob}}; \mu_D)$$

subject to $\forall t \in \{1, \ldots, T\}$, $\forall a, a' \in A$,

$$\gamma_t \mathbb{E}_{\omega\sim\mu_R(\cdot|a,\pi^{(t)})}\Big[v(a,\omega) - v(a',\omega)\Big] + \sum_{t'=t+1}^{T} \gamma_{t'}\Big(V(\pi^{(t')}, \rho_{\text{ob}}; \mu_R) - V_{\text{uninformed}}(\mu_R)\Big) \geq 0.$$

Plugging in Equations (52)-(56), we obtain

$$\mathcal{U}_T^{**}(\mathcal{I}) \;=\; \max_{\sigma\,=\,(\pi^{(1)},\ldots,\pi^{(T)})} \;\sum_{t=1}^{T} \Big((1-p_0)\pi^{(t)}(a_1|\omega_0) + p_0\pi^{(t)}(a_1|\omega_1)\Big)$$

subject to $\forall t \in \{1, \ldots, T\}$,

$$\gamma_t \frac{\mu_R(\omega_0)\pi^{(t)}(a_0|\omega_0) - \mu_R(\omega_1)\pi^{(t)}(a_0|\omega_1)}{\mu_R(\omega_0)\pi^{(t)}(a_0|\omega_0) + \mu_R(\omega_1)\pi^{(t)}(a_0|\omega_1)} + \sum_{t'=t+1}^{T} \gamma_{t'}\Big(- \mu_R(\omega_0)\pi^{(t')}(a_1|\omega_0) + \mu_R(\omega_1)\pi^{(t')}(a_1|\omega_1)\Big) \geq 0,$$

$$\gamma_t \frac{-\mu_R(\omega_0)\pi^{(t)}(a_1|\omega_0) + \mu_R(\omega_1)\pi^{(t)}(a_1|\omega_1)}{\mu_R(\omega_0)\pi^{(t)}(a_1|\omega_0) + \mu_R(\omega_1)\pi^{(t)}(a_1|\omega_1)} + \sum_{t'=t+1}^{T} \gamma_{t'}\Big(- \mu_R(\omega_0)\pi^{(t')}(a_1|\omega_0) + \mu_R(\omega_1)\pi^{(t')}(a_1|\omega_1)\Big) \geq 0,$$

and $\displaystyle\sum_{t'=t+1}^{T} \gamma_{t'}\Big(- \mu_R(\omega_0)\pi^{(t')}(a_1|\omega_0) + \mu_R(\omega_1)\pi^{(t')}(a_1|\omega_1)\Big) \geq 0.$

We note that the optimal solution to the above problem must have $\pi^{(t)}(a_1|\omega_1) = 1$ for every t. This is because: (1) increasing $\pi^{(t)}(a_1|\omega_1)$ increases the designer's utility; (2) increasing $\pi^{(t)}(a_1|\omega_1)$ will not violate the three constraints because the right-hand-sides of the constraints are all increasing in $\pi^{(t)}(a_1|\omega_1)$. Note that in first constraint, the right-hand-side is decreasing in $\pi^{(t)}(a_0|\omega_1)$ while $\pi^{(t)}(a_0|\omega_1) = 1 - \pi^{(t)}(a_1|\omega_1)$ is decreasing in $\pi^{(t)}(a_1|\omega_1)$. So, we can increase

$\pi^{(t)}(a_1|\omega_1)$ to the maximal value of 1. With $\pi^{(t)}(a_1|\omega_1) = 1$, we can rewrite $\mathcal{U}_T^{**}(\mathcal{I})$ as

$$\begin{aligned}
\mathcal{U}_T^{**}(\mathcal{I}) \;=\; & \max_{\sigma\,=\,(\pi^{(1)},\ldots,\pi^{(T)})} \;\sum_{t=1}^{T}\Big((1-p_0)\pi^{(t)}(a_1|\omega_0)+p_0\Big) \\
\text{s.t. }\; & \gamma_t\frac{\mu_R(\omega_0)\pi^{(t)}(a_0|\omega_0)-0}{\mu_R(\omega_0)\pi^{(t)}(a_0|\omega_0)+0}+\sum_{t'=t+1}^{T}\gamma_{t'}\Big(-\mu_R(\omega_0)\pi^{(t')}(a_1|\omega_0)+\mu_R(\omega_1)\Big)\geq 0, \\
& \gamma_t\frac{-\mu_R(\omega_0)\pi^{(t)}(a_1|\omega_0)+\mu_R(\omega_1)}{\mu_R(\omega_0)\pi^{(t)}(a_1|\omega_0)+\mu_R(\omega_1)}+\sum_{t'=t+1}^{T}\gamma_{t'}\Big(-\mu_R(\omega_0)\pi^{(t')}(a_1|\omega_0)+\mu_R(\omega_1)\Big)\geq 0, \\
& \text{and }\;\sum_{t'=t+1}^{T}\gamma_{t'}\Big(-\mu_R(\omega_0)\pi^{(t')}(a_1|\omega_0)+\mu_R(\omega_1)\Big)\geq 0.
\end{aligned}$$

Because $\frac{\mu_R(\omega_0)\pi^{(t)}(a_0|\omega_0)-0}{\mu_R(\omega_0)\pi^{(t)}(a_0|\omega_0)+0}=1$, the first constraint is implied by the third constraint. So we can drop the first constraint. Letting $x^{(t)}=\pi^{(t)}(a_1|\omega_0)$, we have

$$\mathcal{U}_T^{**}(\mathcal{I}) \;=\; \max_{0\leq x^{(t)}\leq 1}\sum_{t=1}^{T}\Big((1-p_0)x^{(t)}+p_0\Big) \quad \text{or equivalently} \quad \max_{0\leq x^{(t)}\leq 1}\sum_{t=1}^{T}x^{(t)} \tag{57}$$

$$\text{s.t. }\; \gamma_t\frac{-\mu_R(\omega_0)x^{(t)}+\mu_R(\omega_1)}{\mu_R(\omega_0)x^{(t)}+\mu_R(\omega_1)}+\sum_{t'=t+1}^{T}\gamma_{t'}\Big(-\mu_R(\omega_0)x^{(t')}+\mu_R(\omega_1)\Big)\geq 0, \tag{58}$$

$$\text{and }\;\sum_{t'=t+1}^{T}\gamma_{t'}\Big(-\mu_R(\omega_0)x^{(t')}+\mu_R(\omega_1)\Big)\geq 0. \tag{59}$$

Claim 8. *A solution to the optimization problem* (57)-(59) *is* $x^{(t)}=\frac{\mu_R(\omega_1)}{\mu_R(\omega_0)}$ *for every* $t\in\{1,\ldots,T\}$.

Proof. We prove this claim by induction on T. When $T=1$, the optimization problem becomes

$$\begin{aligned}
&\max_{0\leq x^{(1)}\leq 1}\; x^{(1)} \\
&\quad\text{s.t. }\;\gamma_1\frac{-\mu_R(\omega_0)x^{(1)}+\mu_R(\omega_1)}{\mu_R(\omega_0)x^{(1)}+\mu_R(\omega_1)}\geq 0,
\end{aligned}$$

which is clearly maximized by $x^{(1)}=\frac{\mu_R(\omega_1)}{\mu_R(\omega_0)}$.

Consider $T\geq 2$. Let $x=(x^{(t)})_{t=1}^T$ be any feasible solution where $x^{(T)}<\frac{\mu_R(\omega_1)}{\mu_R(\omega_0)}$. We define $\tilde{x}=(\tilde{x}^{(t)})_{t=1}^T$ by

$$\tilde{x}^{(T)}=\tfrac{\mu_R(\omega_1)}{\mu_R(\omega_0)},\quad \tilde{x}^{(T-1)}=x^{(T-1)}-\tfrac{\gamma_T}{\gamma_{T-1}}\big(\tfrac{\mu_R(\omega_1)}{\mu_R(\omega_0)}-x^{(T)}\big),\quad \tilde{x}^{(t)}=x^{(t)}\;\;\forall t\leq T-2.$$

We show that $\tilde{x}$ is feasible and has a weakly larger objective value than x. To verify the feasibility

of $\tilde{x}$, we verify the first constraint (58) for every t:

- For $t \le T-2$, we have

$$\begin{aligned}
&\gamma_t \frac{-\mu_R(\omega_0)\tilde{x}^{(t)} + \mu_R(\omega_1)}{\mu_R(\omega_0)\tilde{x}^{(t)} + \mu_R(\omega_1)} + \sum_{t'=t+1}^{T} \gamma_{t'} \Big(-\mu_R(\omega_0)\tilde{x}^{(t')} + \mu_R(\omega_1) \Big) \\
&= \gamma_t \frac{-\mu_R(\omega_0)x^{(t)} + \mu_R(\omega_1)}{\mu_R(\omega_0)x^{(t)} + \mu_R(\omega_1)} + \sum_{t'=t+1}^{T-2} \gamma_{t'} \Big(-\mu_R(\omega_0)x^{(t')} + \mu_R(\omega_1) \Big) \\
&\quad + \gamma_{T-1} \Big(-\mu_R(\omega_0)\tilde{x}^{(T-1)} + \mu_R(\omega_1) \Big) + \gamma_T \cdot 0 \\
&= \gamma_t \frac{-\mu_R(\omega_0)x^{(t)} + \mu_R(\omega_1)}{\mu_R(\omega_0)x^{(t)} + \mu_R(\omega_1)} + \sum_{t'=t+1}^{T-2} \gamma_{t'} \Big(-\mu_R(\omega_0)x^{(t')} + \mu_R(\omega_1) \Big) \\
&\quad + \gamma_{T-1} \Big(-\mu_R(\omega_0) \Big(x^{(T-1)} - \tfrac{\gamma_T}{\gamma_{T-1}} \big(\tfrac{\mu_R(\omega_1)}{\mu_R(\omega_0)} - x^{(T)} \big) \Big) + \mu_R(\omega_1) \Big) \\
&= \gamma_t \frac{-\mu_R(\omega_0)x^{(t)} + \mu_R(\omega_1)}{\mu_R(\omega_0)x^{(t)} + \mu_R(\omega_1)} + \sum_{t'=t+1}^{T-2} \gamma_{t'} \Big(-\mu_R(\omega_0)x^{(t')} + \mu_R(\omega_1) \Big) \\
&\quad + \gamma_{T-1} \Big(-\mu_R(\omega_0)x^{(T-1)} + \mu_R(\omega_1) \Big) + \gamma_T \Big(-\mu_R(\omega_0)x^{(T)} + \mu_R(\omega_1) \Big) \\
&\ge 0
\end{aligned}$$

  where the "≥ 0" is because x satisfies (58).

- For $t = T-1$, we have

$$\begin{aligned}
&\gamma_t \frac{-\mu_R(\omega_0)\tilde{x}^{(t)} + \mu_R(\omega_1)}{\mu_R(\omega_0)\tilde{x}^{(t)} + \mu_R(\omega_1)} + \sum_{t'=t+1}^{T} \gamma_{t'} \Big(-\mu_R(\omega_0)\tilde{x}^{(t')} + \mu_R(\omega_1) \Big) \\
&= \gamma_{T-1} \frac{-\mu_R(\omega_0)\tilde{x}^{(T-1)} + \mu_R(\omega_1)}{\mu_R(\omega_0)\tilde{x}^{(T-1)} + \mu_R(\omega_1)} + \gamma_T \cdot 0 \\
&= \gamma_{T-1} \frac{-\mu_R(\omega_0) \Big(x^{(T-1)} - \tfrac{\gamma_T}{\gamma_{T-1}} \big(\tfrac{\mu_R(\omega_1)}{\mu_R(\omega_0)} - x^{(T)} \big) \Big) + \mu_R(\omega_1)}{\mu_R(\omega_0)\tilde{x}^{(T-1)} + \mu_R(\omega_1)} \\
&= \frac{\gamma_{T-1} \Big(-\mu_R(\omega_0)x^{(T-1)} + \mu_R(\omega_1) \Big) + \gamma_T \Big(-\mu_R(\omega_0)x^{(T)} + \mu_R(\omega_1) \Big)}{\mu_R(\omega_0)\tilde{x}^{(T-1)} + \mu_R(\omega_1)} \\
&\ge 0
\end{aligned}$$

  where the "≥ 0" is because x satisfies (59).

- For $t = T$, we have

$$\gamma_t \frac{-\mu_R(\omega_0)\tilde{x}^{(T)} + \mu_R(\omega_1)}{\mu_R(\omega_0)\tilde{x}^{(T)} + \mu_R(\omega_1)} = 0 \geq 0.$$

Thus, $\tilde{x}$ satisfies the first constraint (58). One can similarly verify that $\tilde{x}$ satisfies the second constraint (59). We then show that $\tilde{x}$ has a weakly larger objective value than x:

$$\begin{aligned}
\sum_{t=1}^{T} \tilde{x}^{(t)} &= \sum_{t=1}^{T-2} x^{(t)} + x^{(T-1)} - \tfrac{\gamma_T}{\gamma_{T-1}}\left(\tfrac{\mu_R(\omega_1)}{\mu_R(\omega_0)} - x^{(T)}\right) + \tfrac{\mu_R(\omega_1)}{\mu_R(\omega_0)} \\
&\geq \sum_{t=1}^{T-2} x^{(t)} + x^{(T-1)} - \left(\tfrac{\mu_R(\omega_1)}{\mu_R(\omega_0)} - x^{(T)}\right) + \tfrac{\mu_R(\omega_1)}{\mu_R(\omega_0)} \\
&= \sum_{t=1}^{T} x^{(t)}.
\end{aligned}$$

Since $\tilde{x}$ is feasible and weakly better than x, we can safely set $x^{(T)} = \frac{\mu_R(\omega_1)}{\mu_R(\omega_0)}$ and reduce the optimization problem to the $T-1$ period problem:

$$\begin{aligned}
&\max_{0 \leq x^{(t)} \leq 1} \sum_{t=1}^{T-1} x^{(t)} \\
&\quad \text{s.t. } \gamma_t \frac{-\mu_R(\omega_0)x^{(t)} + \mu_R(\omega_1)}{\mu_R(\omega_0)x^{(t)} + \mu_R(\omega_1)} + \sum_{t'=t+1}^{T-1} \gamma_{t'} \Big(-\mu_R(\omega_0)x^{(t')} + \mu_R(\omega_1) \Big) \geq 0, \\
&\qquad \text{and } \sum_{t'=t+1}^{T-1} \gamma_{t'} \Big(-\mu_R(\omega_0)x^{(t')} + \mu_R(\omega_1) \Big) \geq 0.
\end{aligned}$$

By induction, the $T-1$ period problem has solution $\left(x^{(t)} = \frac{\mu_R(\omega_1)}{\mu_R(\omega_0)}\right)_{t=1}^{T-1}$. Thus, the T period problem has solution $\left(x^{(t)} = \frac{\mu_R(\omega_1)}{\mu_R(\omega_0)}\right)_{t=1}^{T}$. □

By Claim 8, we have $x^{(t)} = \frac{\mu_R(\omega_1)}{\mu_R(\omega_0)}$ and thus

$$\mathcal{U}_T^{**}(\mathcal{I}) = \sum_{t=1}^{T} \left((1-p_0)\tfrac{\mu_R(\omega_1)}{\mu_R(\omega_0)} + p_0 \right) = T \cdot \left((1-p_0)\tilde{\mu}_R + p_0 \right) \overset{\text{by Claim 7}}{=} T \cdot U_{\mathrm{BP}}^*(\mathcal{I}),$$

which proves Lemma 11.

E.2 Proof of Lemma 12

Given an incentive compatible and individually rational menu $\mathcal{M} = \{\pi_\mu\}$, the receiver will report his prior μ_R truthfully. Consider a weighted welfare of the game, which is the sum of the information designer's utility and $(1+\tilde{\mu}_R)$ times the receiver's utility:

$$\begin{aligned}
&U^{\mathrm{sp}}(\mathcal{M}, \mu_R) + (1+\tilde{\mu}_R) V^{\mathrm{sp}}(\mu_R, \mu_R) \\
&= (1-p_0)\pi_{\mu_R}(a_1|\omega_0) + p_0 \pi_{\mu_R}(a_1|\omega_1) + \tilde{\mu}_R \pi_{\mu_R}(a_1|\omega_1) - \pi(a_1|\omega_0) \\
&= (\tilde{\mu}_R + p_0)\pi_{\mu_R}(a_1|\omega_1) - p_0 \pi_{\mu_R}(a_1|\omega_0). \qquad (60)
\end{aligned}$$

Replacing μ_R by μ (and replacing $\tilde{\mu}_R$ by $\tilde{\mu}$) in (60) and noting that $V^{\mathrm{sp}}(\mu, \mu) \geq V_{\text{uninformed}} \geq 0$ by individual rationality, we have

$$U^{\mathrm{sp}}(\mathcal{M}, \mu) \leq (\tilde{\mu} + p_0)\pi_\mu(a_1|\omega_1) - p_0 \pi_\mu(a_1|\omega_0). \qquad (61)$$

The information designer's single-period regret

$$\begin{aligned}
\mathrm{Reg}^{\mathrm{sp}}(\mathcal{M}, \mu_R) &= (1-p_0)\tilde{\mu}_R + p_0 - U^{\mathrm{sp}}(\mathcal{M}, \mu_R) \\
&\geq \pi_{\mu_R}(a_1|\omega_1) \cdot \Big((1-p_0)\tilde{\mu}_R + p_0 \Big) - U^{\mathrm{sp}}(\mathcal{M}, \mu_R) \\
&= (\tilde{\mu}_R + p_0)\pi_{\mu_R}(a_1|\omega_1) - U^{\mathrm{sp}}(\mathcal{M}, \mu_R) - p_0 \tilde{\mu}_R \pi_{\mu_R}(a_1|\omega_1) \\
\text{by (60)} &= (1+\tilde{\mu}_R) V^{\mathrm{sp}}(\mu_R, \mu_R) + p_0 \pi_{\mu_R}(a_1|\omega_0) - p_0 \tilde{\mu}_R \pi_{\mu_R}(a_1|\omega_1) \\
\text{incentive compatibility} &\geq (1+\tilde{\mu}_R) V^{\mathrm{sp}}(\mu_R, \mu) + p_0 \pi_{\mu_R}(a_1|\omega_0) - p_0 \tilde{\mu}_R \pi_{\mu_R}(a_1|\omega_1) \\
&= (1+\tilde{\mu}_R) \frac{1}{1+\tilde{\mu}_R} \Big(\tilde{\mu}_R \pi_\mu(a_1|\omega_1) - \pi_\mu(a_1|\omega_0) \Big) + p_0 \pi_{\mu_R}(a_1|\omega_0) - p_0 \tilde{\mu}_R \pi_{\mu_R}(a_1|\omega_1) \\
\text{by (61)} &\geq \left(\frac{\tilde{\mu}_R}{\tilde{\mu} + p_0} \Big(U^{\mathrm{sp}}(\mathcal{M}, \mu) + p_0 \pi_\mu(a_1|\omega_0) \Big) - \pi_\mu(a_1|\omega_0) \right) + p_0 \pi_{\mu_R}(a_1|\omega_0) - p_0 \tilde{\mu}_R \pi_{\mu_R}(a_1|\omega \\
\text{by (32)} &= \left(\frac{\tilde{\mu}_R}{\tilde{\mu} + p_0} \Big(U^{\mathrm{sp}}(\mathcal{M}, \mu) + p_0 \pi_\mu(a_1|\omega_0) \Big) - U^{\mathrm{sp}}(\mathcal{M}, \mu) - p_0 \pi_\mu(a_1|\omega_0) + p_0 \pi_\mu(a_1|\omega_1) \right) \\
&\quad + p_0 \pi_{\mu_R}(a_1|\omega_0) - p_0 \tilde{\mu}_R \pi_{\mu_R}(a_1|\omega_1) \\
&= \frac{\tilde{\mu}_R - \tilde{\mu} - p_0}{\tilde{\mu} + p_0} U^{\mathrm{sp}}(\mathcal{M}, \mu) \\
&\quad + p_0 \left(\frac{\tilde{\mu}_R - \tilde{\mu} - p_0}{\tilde{\mu} + p_0} \pi_\mu(a_1|\omega_0) + \pi_\mu(a_1|\omega_1) + \pi_{\mu_R}(a_1|\omega_0) - \tilde{\mu}_R \pi_{\mu_R}(a_1|\omega_1) \right) \\
&\geq \frac{\tilde{\mu}_R - \tilde{\mu} - p_0}{\tilde{\mu} + p_0} U^{\mathrm{sp}}(\mathcal{M}, \mu) + p_0 \left(\frac{\tilde{\mu}_R - \tilde{\mu} - p_0}{\tilde{\mu} + p_0} \pi_\mu(a_1|\omega_0) - \tilde{\mu}_R \right).
\end{aligned}$$

Let $\tilde{\mu}_R = 1$ and $\tilde{\mu} = \frac{1}{2} - p_0$. If $\mathrm{Reg}^{\mathrm{sp}}(\mathcal{M}, \mu_R) \geq \frac{1}{4}$, then the lemma is proved. If $\mathrm{Reg}^{\mathrm{sp}}(\mathcal{M}, \mu_R) < \frac{1}{4}$, then we have

$$\frac{1}{4} > \mathrm{Reg}^{\mathrm{sp}}(\mathcal{M}, \mu_R) \geq U^{\mathrm{sp}}(\mathcal{M}, \mu) + p_0\Big(\pi_\mu(a_1|\omega_0) - 1\Big) \geq U^{\mathrm{sp}}(\mathcal{M}, \mu) - p_0,$$

which implies that the information designer's regret under receiver prior μ is

$$\begin{aligned}\mathrm{Reg}^{\mathrm{sp}}(\mathcal{M}, \mu) &= (1 - p_0)\tilde{\mu} + p_0 - U^{\mathrm{sp}}(\mathcal{M}, \mu) \\ &\geq (1 - p_0)(\frac{1}{2} - p_0) + p_0 - \frac{1}{4} - p_0 \geq \frac{1}{4} - \frac{3}{2}p_0 \geq \frac{1}{16}\end{aligned}$$

given $p_0 = \frac{1}{8}$. Using $\mu(\omega_0) = \frac{1}{1+\tilde{\mu}}$ and $\mu(\omega_1) = \frac{\tilde{\mu}}{1+\tilde{\mu}}$, by direct calculation, one can verify that the condition $\mu(\omega) \geq p_0, \forall \omega \in \Omega$, is satisfied. So, the lemma holds with $\tilde{\mu}$.

E.3 Proof of Theorem 4

Let $\mathcal{G}$ be any learning algorithm of the information designer. Recall that a strategic receiver with prior μ_R best responds by using T-period strategy $\phi^*(\mathcal{G}, \mu_R) = (\phi^{(t)})_{t=1}^T$, where each $\phi^{(t)}$ maps the history $H^{(t-1)}$ and the current period signaling scheme $\pi^{(t)}$ to a current period strategy $\rho^{(t)} : S \to \Delta(A)$ which specifies a distribution over actions for each possible signal. The sequence $(\pi^{(t)}, \rho^{(t)})_{t=1}^T$ of signaling schemes and the receiver's single-period strategies is a stochastic process generated by $\mathcal{G}$ and $\phi^*(\mathcal{G}, \mu_R)$. Let $\sigma_{\mu_R}(a|\omega)$ be the "average probability" that the receiver takes action a conditioning on state ω when the receiver uses $\phi^*(\mathcal{G}, \mu_R)$ to interact with the designer's algorithm $\mathcal{G}$, weighted by the receiver's discount factor:

$$\sigma_{\mu_R}(a|\omega) = \mathbb{E}_{(\pi^{(t)}, \rho^{(t)}) \sim \mathcal{G}, \phi^*(\mathcal{G}, \mu_R)}\left[\frac{1}{T_\gamma}\sum_{t=1}^{T}\gamma_t \sum_{s \in S}\pi^{(t)}(s|\omega)\rho^{(t)}(a|s)\right], \tag{62}$$

where $T_\gamma := \sum_{t=1}^T \gamma_t$. Note that $\sigma_{\mu_R}(\cdot|\omega)$ is a distribution over A, so σ_{μ_R} can be regarded as a direct signaling scheme. Consider the menu $\mathcal{M} = \{\sigma_{\mu_R}\}_{\mu_R \in \Delta(\Omega)}$ of direct signaling schemes σ_{μ_R} associated with any prior μ_R. We claim that the $\mathcal{M}$ is incentive compatible and individually rational in the single-period game.

Claim 9. *The menu $\mathcal{M} = \{\sigma_{\mu_R}\}_{\mu_R \in \Delta(\Omega)}$ is incentive compatible and individually rational.*

Proof. Under menu $\mathcal{M}$, the receiver's expected utility with prior μ_R and a report μ in the single-

period is

$$\begin{aligned}
V^{\text{sp}}(\mathcal{M}, \mu_R, \mu) &= \sum_{\omega \in \Omega} \mu_R(\omega) \sum_{a \in A} \sigma_\mu(a|\omega) v(a, \omega) \\
&= \mathbb{E}_{(\pi^{(t)}, \rho^{(t)}) \sim \mathcal{G}, \phi^*(\mathcal{G}, \mu)} \left[\frac{1}{T_\gamma} \sum_{t=1}^{T} \gamma_t \sum_{\omega \in \Omega} \mu_R(\omega) \sum_{s \in S} \pi^{(t)}(s|\omega) \sum_{a \in A} \rho^{(t)}(a|s) v(a, \omega) \right] \\
&\leq \mathbb{E}_{(\pi^{(t)}, \rho^{(t)}) \sim \mathcal{G}, \phi^*(\mathcal{G}, \mu_R)} \left[\frac{1}{T_\gamma} \sum_{t=1}^{T} \gamma_t \sum_{\omega \in \Omega} \mu_R(\omega) \sum_{s \in S} \pi^{(t)}(s|\omega) \sum_{a \in A} \rho^{(t)}(a|s) v(a, \omega) \right] \\
&= V^{\text{sp}}(\mathcal{M}, \mu_R, \mu_R),
\end{aligned}$$

where the $\leq$ is because $\phi^*(\mathcal{G}, \mu_R)$ is a best response to $\mathcal{G}$ for a strategic receiver with prior μ_R in the T-period game. So, $\mathcal{M}$ is incentive compatible. By a similar argument, one can verify that $\mathcal{M}$ is individually rational. So the claim is proven. □

Then, we relate the designer's regret in the single-period game to her regret in the T-period game:

$$
\begin{aligned}
&T_\gamma \mathrm{Reg}^{\mathrm{sp}}(\mathcal{M}, \mu_R) \\
&= T_\gamma \Big[U^*(\mu_R) - U^{\mathrm{sp}}(\mathcal{M}, \mu_R) \Big] \\
&= T_\gamma \Bigg[U^*(\mu_R) - \sum_{\omega \in \Omega} \mu_D(\omega) \sum_{a \in A} \sigma_{\mu_R}(a|\omega) u(a, \omega) \Bigg] \\
&= T_\gamma U^*(\mu_R) - \mathbb{E}_{(\pi^{(t)}, \rho^{(t)}) \sim \mathcal{G}, \phi^*(\mathcal{G}, \mu_R)} \Bigg[\sum_{t=1}^{T} \gamma_t \sum_{\omega \in \Omega} \mu_D(\omega) \sum_{s \in S} \pi^{(t)}(s|\omega) \sum_{a \in A} \rho^{(t)}(a|s) u(a, \omega) \Bigg] \\
&= T U^*(\mu_R) - (T - T_\gamma) U^*(\mu_R) \\
&\quad - \mathbb{E}_{(\pi^{(t)}, \rho^{(t)}) \sim \mathcal{G}, \phi^*(\mathcal{G}, \mu_R)} \Bigg[\sum_{t=1}^{T} (1 - 1 + \gamma_t) \sum_{\omega \in \Omega} \mu_D(\omega) \sum_{s \in S} \pi^{(t)}(s|\omega) \sum_{a \in A} \rho^{(t)}(a|s) u(a, \omega) \Bigg] \\
&= T U^*(\mu_R) - \mathbb{E}_{(\pi^{(t)}, \rho^{(t)}) \sim \mathcal{G}, \phi^*(\mathcal{G}, \mu_R)} \Bigg[\sum_{t=1}^{T} \sum_{\omega \in \Omega} \mu_D(\omega) \sum_{s \in S} \pi^{(t)}(s|\omega) \sum_{a \in A} \rho^{(t)}(a|s) u(a, \omega) \Bigg] \\
&\quad - (T - T_\gamma) U^*(\mu_R) + \mathbb{E}_{(\pi^{(t)}, \rho^{(t)}) \sim \mathcal{G}, \phi^*(\mathcal{G}, \mu_R)} \Bigg[\sum_{t=1}^{T} (1 - \gamma_t) \sum_{\omega \in \Omega} \mu_D(\omega) \sum_{s \in S} \pi^{(t)}(s|\omega) \sum_{a \in A} \rho^{(t)}(a|s) u(a, \omega) \Bigg] \\
&= \mathrm{Reg}(\mathcal{G}; \mu_D, \mu_R) \\
&\quad \underbrace{- \mathbb{E}_{(\pi^{(t)}, \rho^{(t)}) \sim \mathcal{G}, \phi^*(\mathcal{G}, \mu_R)} \Bigg[\sum_{t=1}^{T} (1 - \gamma_t) \Big(U^*(\mu_R) - \sum_{\omega \in \Omega} \mu_D(\omega) \sum_{s \in S} \pi^{(t)}(s|\omega) \sum_{a \in A} \rho^{(t)}(a|s) u(a, \omega) \Big) \Bigg]}_{A}.
\end{aligned}
$$

Then, we want to show $A \geq 0$. According to the definition of utility (30) (32) in the two-state, two-action single-period game defined previously,

$$
\begin{aligned}
&U^*(\mu_R) - \sum_{\omega \in \Omega} \mu_D(\omega) \sum_{s \in S} \pi^{(t)}(s|\omega) \sum_{a \in A} \rho^{(t)}(a|s) u(a, \omega) \\
&= (1 - p_0) \tilde{\mu}_R + p_0 - (1 - p_0) \sum_{s \in S} \pi^{(t)}(s|\omega_0) \rho^{(t)}(a_1|s) - p_0 \sum_{s \in S} \pi^{(t)}(s|\omega_1) \rho^{(t)}(a_1|s) \\
&= (1 - p_0) \Big(\tilde{\mu}_R - \sum_{s \in S} \pi^{(t)}(s|\omega_0) \rho^{(t)}(a_1|s) \Big) + p_0 \Big(1 - \sum_{s \in S} \pi^{(t)}(s|\omega_1) \rho^{(t)}(a_1|s) \Big) \\
&\geq (1 - p_0)(1 + \tilde{\mu}_R) V^{\mathrm{sp}}(\pi^{(t)}, \rho^{(t)}) + 0,
\end{aligned}
$$

where $V^{\mathrm{sp}}(\pi^{(t)}, \rho^{(t)}) = \frac{1}{1 + \tilde{\mu}_R} \Big(\tilde{\mu}_R \sum_{s \in S} \pi^{(t)}(s|\omega_1) \rho^{(t)}(a_1|s) - \sum_{s \in S} \pi^{(t)}(s|\omega_0) \rho^{(t)}(a_1|s) \Big)$ is the receiver's single-period utility under signaling scheme $\pi^{(t)}$ and responding strategy $\rho^{(t)}$. The following lemma shows that $\mathbb{E}[\sum_{t=1}^{T} (1 - \gamma_t) V^{\mathrm{sp}}(\pi^{(t)}, \rho^{(t)})] \geq 0$. Note that this lemma is not immediate because the receiver's utility can be negative in some periods. See Appendix E.4 for the proof of

this lemma.

Lemma 19. *Assume that the receiver's discount factor sequence satisfies* $1 \geq \gamma_t \geq 0$ *and* $\gamma_t \geq \gamma_{t+1}, \forall t$. *For any learning algorithm* $\mathcal{G}$ *of the information designer, with the receiver best responding by* $\phi^*(\mathcal{G}, \mu_R)$, *we have* $\mathbb{E}[\sum_{t=1}^{T}(1-\gamma_t)V^{\mathrm{sp}}(\pi^{(t)}, \rho^{(t)})] \geq 0$.

So, we obtain

$$A \;\geq\; (1-p_0)(1+\tilde{\mu}_R)\mathbb{E}\bigg[\sum_{t=1}^{T}(1-\gamma_t)V^{\mathrm{sp}}(\pi^{(t)}, \rho^{(t)})\bigg] \;\geq\; 0,$$

which implies

$$\mathrm{Reg}(\mathcal{G}; \mu_D, \mu_R) \;\geq\; T_\gamma \mathrm{Reg}^{\mathrm{sp}}(\mathcal{M}, \mu_R).$$

By Lemma 12, there exists an instance where $\mathrm{Reg}^{\mathrm{sp}}(\mathcal{M}, \mu_R) \geq \frac{1}{16}$, so $\mathrm{Reg}(\mathcal{G}; \mu_D, \mu_R) \geq \frac{1}{16}T_\gamma$. □

E.4 Proof of Lemma 19

In this proof, the expectation is over the stochastic process $\{(\pi^{(t)}, \rho^{(t)})\}_{t=1}^{T}$ generated by $\mathcal{G}$ and $\phi^*(\mathcal{G}, \mu_R)$. Let

$$\mathbb{E}[V^{\mathrm{sp}}(\pi^{(t)}, \rho^{(t)})] = \mathbb{E}\bigg[\frac{1}{1+\tilde{\mu}_R}\Big(\tilde{\mu}_R \sum_{s\in S}\pi^{(t)}(s|\omega_1)\rho^{(t)}(a_1|\omega_1) - \sum_{s\in S}\pi^{(t)}(s|\omega_0)\rho^{(t)}(a_1|\omega_0)\Big)\bigg]$$

be the receiver's expected utility at period t, where the signaling scheme is $\pi^{(t)}$ and the receiver responds by strategy $\rho^{(t)}$. Denote the receiver's discounted and undiscounted total utility from period t_1 to t_2 by

$$DV(t_1, t_2) = \sum_{t=t_1}^{t_2}\gamma_t \mathbb{E}[V^{\mathrm{sp}}(\pi^{(t)}, \rho^{(t)})], \qquad UV(t_1, t_2) = \sum_{t=t_1}^{t_2}\mathbb{E}[V^{\mathrm{sp}}(\pi^{(t)}, \rho^{(t)})].$$

We will use an induction to prove $0 \leq DV(t, T) \leq \gamma_t UV(t, T)$ for every $t \in \{1, \ldots, T, T+1\}$. The base case where $t = T+1$ trivially holds. Consider any $t \leq T$. First, we have $DV(t, T) \geq \sum_{t'=1}^{T}\gamma_{t'}V_{\mathrm{uninformed}} \geq 0$ because the receiver is best-responding using $\phi^*(\mathcal{G}, \mu_R)$ and he can always guarantee the uninformed utility by ignoring the signals. We then note that

$$\begin{aligned} DV(t, T) \;&=\; DV(t, t) + DV(t+1, T) \\ &=\; \gamma_t UV(t, t) + DV(t+1, T) \\ &\leq\; \gamma_t UV(t, t) + \gamma_{t+1}UV(t+1, T) \quad \text{by induction.} \end{aligned}$$

Because $\gamma_t \geq \gamma_{t+1}$ and $UV(t+1,T) \geq \gamma_{t+1} DV(t+1,T) \geq 0$ by induction, we have

$$DV(t,T) \;\leq\; \gamma_t UV(t,t) + \gamma_t UV(t+1,T) \;=\; \gamma_t UV(t,T),$$

which proves the induction.

For $t = 1$, $\gamma_t \leq 1$, so $DV(1,T) \leq \gamma_1 UV(1,T) \leq UV(1,T)$, which implies

$$0 \;\leq\; UV(1,T) - DV(1,T) \;=\; \mathbb{E}\Big[\sum_{t=1}^{T} (1-\gamma_t) V^{\mathrm{sp}}(\pi^{(t)}, \rho^{(t)})\Big]$$

and proves the lemma.