Sequential Information Design: Markov Persuasion Process and Its Efficient Reinforcement Learning

Jibang Wu[*] Zixuan Zhang[†] Zhe Feng[‡] Zhaoran Wang[§]
Zhuoran Yang[¶] Michael I. Jordan[‖] Haifeng Xu[**]

Abstract

In today's economy, it becomes important for Internet platforms to consider the sequential information design problem to align its long term interest with incentives of the gig service providers (e.g., drivers, hosts). This paper proposes a novel model of sequential information design, namely the Markov persuasion processes (MPPs). Specifically, in an MPP, a sender, with informational advantage, seeks to persuade a stream of myopic receivers to take actions that maximizes the sender's cumulative utilities in a finite horizon Markovian environment with varying prior and utility functions. Planning in MPPs thus faces the unique challenge in finding a signaling policy that is simultaneously persuasive to the myopic receivers and inducing the optimal long-term cumulative utilities of the sender. Nevertheless, in the population level where the model is known, it turns out that we can efficiently determine the optimal (resp. ϵ-optimal) policy with finite (resp. infinite) states and outcomes, through a modified formulation of the Bellman equation that additionally takes persuasiveness into consideration.

Our main technical contribution is to study the MPP under the online reinforcement learning (RL) setting, where the goal is to learn the optimal signaling policy by interacting with with the underlying MPP, without the knowledge of the sender's utility functions, prior distributions, and the Markov transition kernels. For such a problem, we design a provably efficient no-regret learning algorithm, the Optimism-Pessimism Principle for Persuasion Process (`OP4`), which features a novel combination of both optimism and pessimism principles. In particular, we obtain optimistic estimates of the value functions to encourage exploration under the unknown environment. Meanwhile, we additionally robustify the signaling policy with respect to the uncertainty of prior estimation to prevent receiver's detrimental equilibrium behavior. Our algorithm enjoys

[*]University of Virginia. Email: jw7jb@virginia.edu.
[†]University of Science and Technology of China. Email: zhangzixuan@mail.ustc.edu.cn.
[‡]Google. Email: zhef@google.com.
[§]Northwestern University. Email: zhaoranwang@gmail.com.
[¶]Yale University. Email: zhuoran.yang@yale.edu.
[‖]UC Berkeley. Email: jordan@cs.berkeley.edu.
[**]University of Virginia. Email: hx4ad@virginia.edu.

sample efficiency by achieving a sublinear $\sqrt{T}$-regret upper bound. Furthermore, both our algorithm and theory can be applied to MPPs with large space of outcomes and states via function approximation, and we showcase such a success under the linear setting.

1 Introduction

Most sequential decision models assume that there is a sole agent who possesses and processes all relevant (online or offline) information and takes an action accordingly. However, the economic literature on information design [26, 8] highlights the importance of considering information asymmetry in decision making, where the decision maker and information possessor may be two parties having different interests and goals. For example, a ride-sharing platform holds historical and real-time data on active riders and driver types in different locations, based on which they have developed centralized combinatorial optimization algorithms and reinforcement learning algorithms for vehicle repositioning, routing and order matching to optimize their operational efficiency and profit [33, 42, 34, 43]. But the de facto decision makers are the drivers. Moreover, as increasingly many drivers are freelancers instead of employees, the platform cannot expect to give mandatory orders to them. On the other hand, if the platform shares no information on rider demand, most drivers will not be able to efficiently find profitable trips. Therefore, it is not only realistic but also necessary to consider an information design problem that aligns the interests of the two parties in sequential decision making processes of this kind.

Given the large data sets being collected by corporations and governments, with avowed goals that relate data analysis to social welfare, it is timely to pursue formal treatments of sequential information design, to understand how to strategically inform the (sequential) decision makers (e.g., users, clients or citizens) impacted by centralized data analysis. In particular, we wish to understand the resulting equilibrium outcomes of both parties. As a concrete example, consider an online shopping platform which may make use of learning tools such as reinforcement learning or online convex optimization to manage inventory and ensure profitability [20, 36]. The platform cannot single-handedly manage its inventory, instead it requires information design (a.k.a., Bayesian persuasion) in its interactions with its suppliers and consumers. On the supply side, it could strategically reveal aspects of consumer sentiment (e.g., rough number of visits, search) to the suppliers in order to guide their sales expectation and negotiate for lower unit prices. On the demand side, it could tactically control displayed product information (e.g., last five remaining, editor's choice) so as to influence consumers' perception of products and consequently their purchase decisions. Similar situations can be anticipated for a recommendation platform. On the one hand, it should recommend most relevant items to its users for click-through and engagement. On the other hand, its recommendations are subject to misalignments with long-term objectives such as profits (e.g., from paid promotion), social impact (e.g., to prevent misinformation and filter bubbles) or development of a creator ecosystem [49, 53, 37].

1.1 Our Results and Contributions

To provide a formal foundation for the study of sequential information design, we introduce the *Markov persuasion process* (MPP), where a sender, with informational advantage, seeks to persuade a stream of myopic receivers to take actions that maximize the sender's cumulative utility in a finite-horizon Markovian environment with varying prior and utility functions. We need to address a key challenge regarding the planning problem in MPPs; specifically, how to find persuasive signaling policies that are also optimized for the sender's long-term objective. Moreover, in face of the uncertainty for both the environment and receivers, there is a dilemma that the optimal policy based on estimated prior is not necessarily persuasive and thus cannot induce the desired trajectory, whereas a full information revelation policy is always persuasive but usually leads to suboptimal cumulative utility. So the reinforcement learning algorithm in MPPs has to ensure optimality under the premise of robust persuasiveness. This makes our algorithm design non-trivial and regret analysis highly challenging.

We show how to surmount these analysis and design challenges, and present a no-regret learning algorithm, which we refer to as Optimism-Pessimism Principle for Persuasion Process (`OP4`), that provably achieves a $\widetilde{O}\big(\sqrt{d_\phi^2 d_\psi^3 H^4 T}\big)$ regret with high probability, where d_ϕ, d_ψ are dimensions of the feature spaces, H is the horizon length in each episode, T is the number of episodes, and $\widetilde{O}(\cdot)$ hides logarithmic factors as well as problem-dependent parameters. To establish this result, in Section 3.3 we start by constructing a modified formulation of the Bellman equation that can efficiently determine the optimal (resp. ϵ-optimal) policy with finite (resp. infinite) states and outcomes. Section 4.2 then considers the learning problem, in particular the design of the `OP4` that adopts both the optimistic principle in utility estimation to incentivize exploration and the pessimism principle in prior estimation to prevent a detrimental equilibrium for the receiver. In Sections 4.3 and 4.4, we showcase `OP4` in the tabular MPPs and contextual Bayesian persuasion problem, respectively, both of which are practical special cases of MPPs. In Section 5, we then generalize these positive results to MPPs with large outcome and state spaces via linear function approximation and generalized linear models.

In summary, our contributions are threefold. At the conceptual level, we identify the need for sequential information design in real-world problems and accordingly formulate a novel model, the MPP, to capture the misaligned incentives between the (sequential) decision makers and information possessors. At the methodological level, our key insight is a new algorithmic principle—optimism to encourage exploration and pessimism to induce robust equilibrium behavior. Finally, at the technical level, we develop a novel regret decomposition tailored to this combination of optimism and pessimism in the design of online learning algorithms. The fact that the combined optimism-pessimism concept can still lead to $O(\sqrt{T})$ regret for strategic setups was not clear before our new regret decomposition lemma. We expect this design principle and our proof techniques can be useful for other strategic learning problems.

1.2 Related Work

Our work is built on the foundation of information design and reinforcement learning. We refer the readers to Section 2.1 and 2.2 for background and formal introductions. Here we focus on the technical and modeling comparisons with related work from dynamic Bayesian persuasion and efficient reinforcement learning.

Dynamic Bayesian persuasion. Starting from seminal work by Kamenica and Gentzkow [26], the study of Bayesian persuasion looks at the design problem to influence an uninformed decision maker through strategic information revelation. Many variants of this model have been studied, with applications in security, advertising, finance, etc. [44, 54, 21, 6]. More recently, several dynamic Bayesian persuasion frameworks have been proposed to model the long-term interest of the sender. Many papers [16, 45, 17, 29] consider the setting where the sender observes the evolving states of a Markov chain, seeks to influence the receiver's belief of the state through signaling and thereby persuade him to take certain actions. In contrast to our setting, at each round, the receiver's action has no influence on the evolution of the Markov process and thus can only maximizes his utility on his belief of current state, given all the historical signals received from the sender. In Ely [16], Farhadi and Teneketzis [17], the Markov chain has two states (one is absorbing): the receiver is interested in detecting the jump to the absorbing state, whereas the sender seeks to prolong the time to detection of such a jump. Renault et al. [45] shows a greedy disclosure policy that ignores its influence to the future utility can be optimal in Markov chain with special utility functions. Lehrer and Shaiderman [29] characterize optimal strategies under different discount factors as well as the optimal values the sender could achieve. Closer to our model is that of Gan et al. [19]—we both assume the Markov environment with state transition influenced by receiver's action, as well as a separate persuasion state drawn from a prior independent of receiver's action. However, Gan et al. [19] focus on the planning problem for the infinite-horizon MDP, solving sender's optimal signaling policy when the environment is known in cases when the receiver is myoptic or far-sighted. In particular, it is shown as NP-hard to approximate an optimal policy against a far-sighted receiver, which also justifies our interest on the myoptic receiver. Another closely related work [61] studies the learning problem in repeated persuasion setting (without Markov state transition) between a stream of myopic receivers and a sender without initial knowledge of the prior. It introduces the notion of regret as well as the robustness principle to this learning problem that we adopt and generalize to our model.

Bayesian Incentive-Compatible Bandit Exploration. Our work is also loosely related to a seminal result by Mansour et al. [35], who model the misaligned incentives between a system (i.e., sender) and a stream of myopic agents (i.e., receivers). Mansour et al. [35] shows that using information asymmetry, the system can create intrinsic incentives for agents to follow its recommendations. In this problem, the sender's objective is limited to the social welfare, i.e, the cumulative utility of all agents, whereas we make no assumption on the sender's utility function and thus her long-term objective. Besides our model is designed to capture more general situations where each receiver could have different priors and utility functions, and the environment might be Markovian with dynamics under the influence of the receivers' actions.

Efficient Reinforcement Learning. Reinforcement learning has seen its successful applications in various domains, such as robotics, finance and dialogue systems [27, 58, 30]. Along with the empirical success, we have seen a growing quest to establish provably efficient RL methods. Classical sample efficiency results focus on tabular environments with small, finite state spaces [2, 39, 4, 12, 47, 24, 46]. Notably, through the design principle, known as optimism in the face of uncertainty [28], an RL algorithm would provably incur a $\Omega(\sqrt{|\mathcal{S}||\mathcal{A}|T})$ regret under the tabular setting, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces respectively [24, 4]. More recently, there have been advances in RL with function approximation, especially the linear case. Jin et al. [25] proposed an efficient algorithm for a setting where the transition kernel and the utility function are both linear functions with respect to a feature mapping: $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. A similar assumption has been studied for different settings and has led to sample efficiency results [55, 14, 38, 57, 22]. Moreover, other general function approximations have been studied in parallel, including generalized linear function approximation [52], linear mixture MDPs based on a ternary feature mapping [3, 60, 9, 59], kernel approximation [56] as well as models based on the low Bellman rank assumption [23, 11]. We make use of these function approximation techniques to model our conditional prior, and we show how to integrate the persuasion structure into these efficient reinforcement learning frameworks, thereby obtaining sample efficient result for large-scale MPPs.

2 Preliminaries

This section provides some necessary background in information design and Markov decision processes, as preparation for our model of Markov persuasion processes presented in the next section.

2.1 Basics of Information Design

Classic information design [26] considers the persuasion problem between a single sender (she) and receiver (he). The receiver is the only actor, and looks to take an action $a \in \mathcal{A}$ which results in receiver utility $u(\omega, a)$ and sender utility $v(\omega, a)$. Here $\omega \in \Omega$ is the realized *outcome* of certain environment uncertainty, which is drawn from a prior distribution $\mu \in \Delta(\Omega)$, and $\mathcal{A}$ is a finite set of available actions for the receiver. While $u, v : \Omega \times \mathcal{A} \to [0, 1]$ and the prior distribution μ are all common knowledge, the sender possesses an informational advantage and can privately observe the realized outcome ω. The persuasion problem studies how the sender can selectively reveal her private information about ω to influence the receiver's decisions and ultimately maximize her own expected utility v.

To model the sender's strategic revelation of information, it is standard to use a *signaling scheme*, which essentially specifies the conditional distribution of a random variable (namely the *signal*), given the outcome ω. Before the realization of the outcome, the sender commits to such a signaling scheme. Given the realized outcome, the sender samples a *signal* from the conditional distribution according to the *signaling scheme* and reveals it to the receiver. Upon receiving this *signal*, the receiver infers a posterior belief about the

outcome via Bayes' theorem (based on the correlation between the signal and outcome ω as promised by the signaling scheme) and then chooses an action a that maximizes the expected utility.

A standard revelation-principle-style argument shows that it is without loss of generality to focus on *direct* and *persuasive* signaling schemes [26]. A scheme is direct if each signal corresponds to an action recommendation to the receiver, and is persuasive if the recommended action indeed maximizes the receiver's a posteriori expected utility. More formally, in a direct signaling scheme , $\pi = (\pi(a|\omega) : \omega \in \Omega, a \in \mathcal{A})$, $\pi(a|\omega)$ denotes the probability of recommending action a given realized outcome ω. Upon receiving an action recommendation a, the receiver computes a posterior belief for ω: $\Pr(\omega|a) = \frac{\mu(\omega)\pi(a|\omega)}{\sum_{\omega'}\mu(\omega')\pi(a|\omega')}$. Thus, the action recommendation a is persuasive if and only if a maximizes the expected utility w.r.t. the posterior belief about ω; i.e., $\sum_\omega \Pr(\omega|a) \cdot u(\omega, a) \geq \sum_\omega \Pr(\omega|a) \cdot u(\omega, a')$ for any $a' \in \mathcal{A}$. Equivalently, we define *persuasiveness* as

$$\text{Persuasiveness:} \qquad \sum_{\omega\in\Omega} \mu(\omega)\pi(a|\omega) \cdot [u(\omega, a) - u(\omega, a')] \geq 0, \forall a, a' \in \mathcal{A}.$$

Let $\mathcal{P} = \{\pi : \pi(\cdot|\omega) \in \Delta(\mathcal{A}) \text{ for each } \omega \in \Omega\}$ denote the set of all signaling schemes. To emphasize that the definition of persuasiveness depends on the prior μ, we denote the set of persuasive schemes on prior μ by

$$\text{Pers}(\mu) \coloneqq \left\{ \pi \in \mathcal{P} : \sum_{\omega\in\Omega} \mu(\omega)\pi(a|\omega) \left[u(\omega, a) - u\left(\omega, a'\right)\right] \geq 0, \quad \forall a, a' \in \mathcal{A} \right\}.$$

Given a persuasive signaling scheme $\pi \in \text{Pers}(\mu)$, it is in the receiver's best interest to take the recommended action and thus the sender's expected utility becomes $V(\mu, \pi) \coloneqq \sum_{\omega\in\Omega}\sum_{a\in\mathcal{A}} \mu(\omega)\pi(a|\omega)v(\omega, a)$.

Therefore, given full knowledge of the persuasion instance, the sender can solve for an optimal persuasive signaling scheme that maximizes her expected utility through the following linear program (LP) which searches for a persuasive signaling scheme that maximizes $V(\mu, \pi)$ (see, e.g., [15] for details):

$$\text{Persuasion as an LP:} \qquad \text{OPT}\,(\mu) \coloneqq \max_{\pi\in\text{Pers}(\mu)} \quad V(\mu, \pi).$$

2.2 Basics of Reinforcement Learning and Markov Decision Processes

The Markov decision process (MDP) [41, 48] is a classic mathematical framework for the sequential decision making problem. In this work, we focus on the model of episodic MDP. Specifically, at the beginning of the episode, the environment has an initial state s_1 (possibly picked by an adversary). Then, at each step $h \geq 1$, the agent takes some action $a_h \in \mathcal{A}$ to interact with environment at state $s_h \in \mathcal{S}$. The state s_h obeys a Markov property and thus captures all relevant information in the history $\{s_i\}_{i<h}$. Accordingly, the agent receives the utility $v_h(s_h, a_h) \in [0, 1]$ and the system evolves to the state of the next step $s_{h+1} \sim P_h(\cdot|s_h, a_h)$. Such a process terminates after $h = H$, where H is also known as the horizon length. Here, $\mathcal{A}$ is a finite set of available actions for the agent, $\mathcal{S}$ is the (possibly infinite) set of MDP states. The utility

function $v_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ and transition kernel $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ may vary at each step. A policy of the agent $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ characterizes her decision making process at step h—after observing the state s, the agent takes action a with probability $\pi_h(a|s)$.

In an episodic MDP with H steps, under policy $\boldsymbol{\pi} = \{\pi_h\}_{h\in[H]}$, we define the value function as the expected value of cumulative utilities starting from an arbitrary state,

$$V_h^{\pi}(s) \coloneqq \mathbb{E}_{P,\boldsymbol{\pi}}\left[\sum_{h'=h}^{H} v_h(s_{h'}, a_{h'}) \middle| s_{h'} = s\right], \quad \forall s \in \mathcal{S}, h \in [H].$$

Here $\mathbb{E}_{P,\boldsymbol{\pi}}$ means that the expectation is taken with respect to the trajectory $\{s_h, a_h\}_{h\in[H]}$, which is generated by policy $\boldsymbol{\pi}$ on the transition model $P = \{P_h\}_{h\in[H]}$. Similarly, we define the action-value function as the expected value of cumulative utilities starting from an arbitrary state-action pair,

$$Q_h^{\pi}(s,a) \coloneqq v_h(s_h, a_h) + \mathbb{E}_{P,\boldsymbol{\pi}}\left[\sum_{h'=h+1}^{H} v_h(s_{h'}, a_{h'}) \middle| s_{h'} = s, a_{h'} = a\right], \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, h \in [H].$$

The optimal policy is defined as $\boldsymbol{\pi}^* \coloneqq \arg\max_{\boldsymbol{\pi}} V_h^{\boldsymbol{\pi}}(s_1)$, which maximizes the (expected) cumulative utility. Since the agent's action affects both immediate utility and next states that influences its future utility, it thus demands careful planning to maximize total utility. Notably, $\boldsymbol{\pi}^*$ can solved by dynamic programming based on the Bellman equation [7]. Specifically, with $V_{H+1}^*(s) = 0$ and for each h from H to 1, iteratively update $Q_h^*(s,a) = v_h(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)} V_{h+1}^*(s', a)$, $V_h^*(s) = \max_{a\in\mathcal{A}} Q_h^*(s,a)$, and determine the optimal policy $\boldsymbol{\pi}^*$ as the greedy policy with respect to $\{Q_h^*\}_{h\in[H]}$.

In online reinforcement learning, the agent has no prior knowledge of the environment, namely, $\{v_h, P_h\}_{h\in[H]}$, but aims to learn the optimal policy by interacting with the environment for T episodes. For each $t \in [T]$, at the beginning of the t-th episode, after observing the initial state s_1^t, the agent chooses a policy $\boldsymbol{\pi}^t$ based on the observations before t-th episode. The discrepancy between $V_1^{\boldsymbol{\pi}^t}(s_1^t)$ and $V_1^*(s_1^t)$ serves as the suboptimality of the agent at the t-th episode. The performance of the online learning algorithm is measured by the expected regret, $\mathrm{Reg}(T) \coloneqq \sum_{t=1}^{T} [V_1^*(s_1^t) - V_1^{\boldsymbol{\pi}^t}(s_1^t)]$.

3 Markov Persuasion Processes

This section introduces the Markov Persuasion Process (MPP), a novel model for sequential information design in *Markovian environments*. It notably captures the motivating yet intricate real-world problems in Section 1. Furthermore, our MPP model is readily applicable to generalized settings with large state spaces by incorporating function approximation techniques.

3.1 A Model of Markov Persuasion Processes (MPPs)

We start by abstracting the sequential information design problem instances in Section 1 into MPPs. Taking as an example recommendation platform for ad keywords, we view the platform as the *sender*, the advertisers as the *receivers*. The advertisers decide the *actions* $a \in \mathcal{A}$ such as whether to accept the recommended keyword. To better reflect the nature of reality, we model two types of information for MPPs, *outcome* and *state*. We use the notion of *outcome* $\omega \in \Omega$ to characterize the sender's private information in face of each receiver, such as the features of searchers for some keyword. The *outcome* follows a prior distribution such as the general demographics of Internet users on the platform. The platform can thus leverage such fine-grained knowledge on keyword features, matching with the specific ad features of each advertiser, to persuade the advertisers to take a recommendation of keywords. Meanwhile, we use the notion of *state* $s \in \mathcal{S}$ to characterize the Markovian state of the environment, e.g., the availability of ad keyword slots. The state is affected by the receiver's action, as the availability changes after some keywords get brought.[1] Naturally, both sender's and receiver's utility are determined by the receiver's action a jointly with the state of environment s and realized outcome ω, i.e., $u, v : \mathcal{S} \times \Omega \times \mathcal{A} \to [0, 1]$. Meanwhile, as these applications could serve thousands or millions of receivers every day, to reduce the complexity of our model we assume each receiver appears only once and thus will *myopically* maximizes his utility at that particular step, whereas the sender is a system planner who aim to maximizes her long-term accumulated expected utility.

More specifically, an MPP is built on top of a standard episodic MDP with *state* space $\mathcal{S}$, action space $\mathcal{A}$, and transition kernel P. In this paper, we restrict our attention to *finite-horizon* (i.e., episodic) MPPs with H steps denoted by $[H] = \{1, \cdots, H\}$, and leave the study of infinite-horizon MPPs as an interesting future direction. At a high level, there are two major differences between MPPs and MDPs. First, in a MPP, the planner *cannot* directly take an action but instead can leverage its information advantage and "persuade" a receiver to take a desired action a_h at each step $h \in [H]$. Second, in an MPP, the state transition is affected not only by the current action a_h and state s_h, but also by the realized outcome ω_h of Nature's probability distribution. Specifically, the state transition kernel at step h is denoted as $P_h(s_{h+1}|s_h, \omega_h, a_h)$. To capture the sender's persuasion of a receiver to take actions at step h, we assume that a fresh receiver arrives at time h with a prior μ_h over the outcome ω_h. The planner, who is the *sender* here, can observe the realized outcome ω_h and would like to strategically reveal information about ω_h in order to persuade the receiver to take a certain action a_h.

Differing from classical single-shot information design, the immediate utility functions u_h, v_h for the

[1] Similarly, we can view the online shopping platform as the sender who persuades a stream of receivers (supplier, consumer) to take certain action, whether to take an offer or make a purchase. In this case, sender can privately observe the outcomes such as the consumer sentiments on some random products based on the search and click logs, whereas the states are product reviews, sales or shipping time commonly known to the public and affected by the actions of both supply and demand sides. In case of rider-sharing, outcome represents the fine-grained knowledge of currently active rider types that are privately known to the platform and generally stochastic in accordance to some user demographics, whereas the state captures the general driver supply or rider demand at locations that is affected by the drivers' decisions.

receiver and sender vary not only at each step h but also additionally depend on the commonly observed state s_h of the environment. We assume the receiver to have full knowledge of his utility u_h and prior μ_h at each step h, and would take the recommended action a_h if and only if a_h maximizes his expected utility under the posterior for ω_h.[2]

Formally, an MPP with a horizon length H proceeds as follows at each step $h \in [H]$:

1. A fresh receiver with prior distribution $\mu_h \in \Delta(\Omega)$ and utility $u_h : \mathcal{S} \times \Omega \times \mathcal{A} \to [0, 1]$ arrives.
2. The sender commits to a *persuasive* signaling policy $\pi_h : \mathcal{S} \to \mathcal{P}$, which is public knowledge.
3. After observing the realized state s_h and outcome ω_h, the sender accordingly recommends the receiver to take an action $a_h \sim \pi_h(\cdot|s_h, \omega_h)$.
4. Given the recommended action a_h, the receiver takes an action a'_h, receives utility $u_h(s_h, \omega_h, a'_h)$ and then leaves the system. Meanwhile, the sender receives utility $v_h(s_h, \omega_h, a'_h)$.
5. The next state $s_{h+1} \sim P_h(\cdot|s_h, \omega_h, a'_h)$ is generated according to $P_h : \mathcal{S} \times \Omega \times \mathcal{A} \to \Delta(\mathcal{S})$, the state transition kernel at the h-th step.

Here we coin the notion of a *signaling policy* π_h as a mapping from state to a signaling scheme at the h-th step. It captures a possibly multi-step procedure in which the sender commits to a signaling scheme after observing the realized state and then samples a signal after observing the realized outcome. For notational convenience, we denote $\pi(a|s, \omega)$ as the probability of recommending action a given state s and realized outcome ω. We can also generalize the notion of persuasiveness from classic information design to MPPs. Specifically, we define $\mathrm{Pers}(\mu, u)$ as the persuasive set that contains all signaling policies that are persuasive to the receiver with utility u and prior μ for all possible state $s \in \mathcal{S}$:

$$\mathrm{Pers}(\mu, u) := \Bigg\{ \pi : \mathcal{S} \to \mathcal{P} : \int_{\omega \in \Omega} \mu(\omega)\pi(a|s, \omega)\big[u(s, \omega, a) - u(s, \omega, a')\big]\mathrm{d}\omega \geq 0, \quad \forall a, a' \in \mathcal{A}, s \in \mathcal{S} \Bigg\}.$$

Recall that $\mathcal{P}$ consists of all mappings from Ω to $\Delta(\mathcal{A})$. As such, the sender's persuasive signaling scheme $\pi_h \in \mathrm{Pers}(\mu_h, u_h)$ is essentially a stochastic policy as defined in standard MDPs—π_h maps a state s_h to a stochastic action a_h—except that here the probability of suggesting action a_h by policy π_h depends additionally on the realized outcome ω_h that is only known to the sender.

[2]This assumption is not essential but just for technical rigor. Because even if receivers have limited knowledge or computational power to accurately determine the utility-maximizing actions, the sender should have sufficient ethical or legal reasons to comply with the persuasive constraints in practice. And the receivers would only take the recommendation if the platform has good reputation (i.e., persuasive with high probability).

We say $\boldsymbol{\pi} \coloneqq \{\pi_h\}_{h\in[H]}$ is a *feasible* policy of the MPP if $\pi_h \in \mathrm{Pers}(\mu_h, u_h), \forall h \in [H]$, because the state transition trajectory would otherwise be infeasible if the receiver is not guaranteed to take the recommended action, i.e., $a'_h \neq a_h$. We denote the set of all *feasible* policies as $\mathcal{P}^H \coloneqq \prod_{h\in[H]} \mathrm{Pers}(\mu_h, u_h)$.

3.2 MPPs: the Generalized Version with Contexts and Linear Parameterization

To provide a broadly useful modeling concept, we also study a generalized setting of the Markov Persuasion Process with contextual prior and a possibly large space of states, outcomes and contexts.

Contextual Prior. At the beginning of each episode, a sequence of contexts $C = \{c_h \in \mathcal{C}\}_{h\in[H]}$ is realized by Nature and becomes public knowledge. And we allow the prior μ_h to be influenced by the context c_h at each step h, and thus denote it by $\mu_h(\cdot|c_h)$. Specifically, the contextual information is able to model the uncertainty such as the varying demographics of active user group affected by events (e.g., scheduled concerts or sport games in ride-sharing) at different time of the day.[3] Here we allow the sequence of contexts to be adversarially generated.

Linear Parameterization. We also relax the state, context and outcome space $\mathcal{S}, \mathcal{C}, \Omega$ to be continuous and additionally assume that the transition kernels and utility functions are linear, and the conditional priors of outcomes are generalized linear models (GLM) of the context at each steps. More formally, for each step $h \in [H]$, our linearity condition assumes:

- The sender's utility is $v_h \coloneqq v_h^*(s_h, \omega_h, a_h) = \psi(s_h, \omega_h, a_h)^\top \gamma_h^*$, where (1) $\psi(\cdot,\cdot,\cdot) \in \mathbb{R}^{d_\psi}$ is a known feature vector; (2) $\gamma_h^* \in \mathbb{R}^{d_\psi}$ is the unknown linear parameter at step h.
- The next state s_{h+1} is drawn from the distribution $P_{M,h}(\cdot|s_h, \omega_h, a_h) = \psi(s_h, \omega_h, a_h)^\top M_h(\cdot)$, where $M_h = (M_h^{(1)}, M_h^{(2)}, \ldots, M_h^{(d_\psi)})$ is a vector of d_ψ unknown measures over $\mathcal{S}$ at step h.
- The outcome $\omega_h \in \mathbb{R}$ subjects to a generalized linear model (GLM), which models a wider range of hypothesis function classes.[4] Given the context c_h, there exists a link function $f : \mathbb{R} \to \mathbb{R}$ such that $\omega_h = f(\phi(c_h)^\top \theta_h^*) + z_h$, where $\phi(\cdot) \in \mathbb{R}^{d_\phi}$ is a feature vector and $\theta_h^* \in \mathbb{R}^{d_\phi}$ is an unknown parameter. The noises $\{z_h\}_{h\in[H]}$ are independent σ-sub-Gaussian variables with zero mean. We denote the prior of ω_h with parameter θ at context c as $\mu_\theta(\cdot|c)$.

Without loss of generality, we assume that there exist Φ, Ψ such that $\|\phi(s)\| \leq \Phi$,[5] $\|\psi(s, \omega, a)\| \leq \Psi$ for all $s \in \mathcal{S}, \omega \in \Omega$ and $a \in \mathcal{A}$. We also assume that $\|\theta_h^*\| \leq L_\theta$, $\|\gamma_h^*\| \leq L_\gamma$, $\|M_h^*\| \leq L_M$, $|\mathcal{A}| \geq 2$, $|\Omega| \geq 2$. Such a regularity condition is common in the RL literature.

[3] In the case of the online shopping platform, the prior of consumer interests may be affected by the different holidays or seasons at different time of year.

[4] We note that GLM is a strictly generalization of the linear model assumption that we have for the distribution of transition kernel P. While we could use similar technique to extend the distribution of P to GLM using techniques similar to that in Wang et al. [52], but we save such an extension for simplicity, since it is not the primary focus of our work.

[5] For the simplicity of notation, we will omit the subscript of the norm whenever it is an L_2 norm in this paper.

3.3 Optimal Signaling Policy in MPPs

In order to maximize the sender's utility, we study the optimal policy in MPPs, in analogy to that of standard MDPs. We start by considering the value of any feasible policy $\boldsymbol{\pi}$. For each step $h \in [H]$, we define the value function for the sender $V_h^{\pi} : \mathcal{S} \to \mathbb{R}$ as the expected value of cumulative utilities under $\boldsymbol{\pi}$ when starting from an arbitrary state at the h-th step. That is, for any $s \in \mathcal{S}, h \in [H]$, we define

$$V_h^{\pi}(s) \coloneqq \mathbb{E}_{P,\mu,\pi}\left[\sum_{h'=h}^{H} v_{h'}\left(s_{h'}, \omega_{h'}, a_{h'}\right) \middle| s_h = s\right],$$

where the expectation $\mathbb{E}_{P,\mu,\pi}$ is taken with respect to the randomness of the trajectory (i.e., randomness of state transition), realized outcome and the stochasticity of $\boldsymbol{\pi}$. Accordingly, we define the Q-function (action-value function) $Q_h^{\pi} : \mathcal{S} \times \Omega \times \mathcal{A} \to \mathbb{R}$ which gives the expected value of cumulative utilities when starting from an arbitrary state-action pair at the h-step following the signaling policy $\boldsymbol{\pi}$, that is,

$$Q_h^{\pi}(s, \omega, a) \coloneqq v_h(s, \omega, a) + \mathbb{E}_{P,\mu,\pi}\left[\sum_{h'=h+1}^{H} v_{h'}\left(s_{h'}, \omega_{h'}, a_{h'}\right) \middle| s_h = s, \omega_h = \omega, a_h = a\right].$$

By definition, $Q_h(\cdot,\cdot,\cdot), V_h(\cdot) \in [0, h]$, since $v_h(\cdot,\cdot,\cdot) \in [0, 1]$. To simplify notation, for any Q-function Q_h and any distributions μ_h and π_h over Ω and $\mathcal{A}$, we additionally denote

$$\left\langle Q_h, \mu_h \otimes \pi_h \right\rangle_{\Omega \times \mathcal{A}}(s) \coloneqq \mathbb{E}_{\omega \sim \mu_h, a \sim \pi_h(\cdot|s,\omega)}\left[Q_h(s, \omega, a)\right].$$

Using this notation, the Bellman equation associated with signaling policy $\boldsymbol{\pi}$ becomes

$$Q_h^{\pi}(s, \omega, a) = (v_h + P_h V_{h+1}^{\pi})(s, \omega, a), \;\; V_h^{\pi}(s) = \left\langle Q_h^{\pi}, \mu_h \otimes \pi_h \right\rangle_{\Omega \times \mathcal{A}}(s), \;\; V_{H+1}^{\pi}(s) = 0, \tag{3.1}$$

which holds for all $s \in \mathcal{S}, \omega \in \Omega, a \in \mathcal{A}$. Similarly, the Bellman optimality equation is

$$Q_h^{*}(s, \omega, a) = (v_h + P_h V_{h+1}^{*})(s, \omega, a), \; V_h^{*}(s) = \max_{\pi_h' \in \mathrm{Pers}(\mu_h, u_h)} \left\langle Q_h^{*}, \mu_h \otimes \pi_h' \right\rangle_{\Omega \times \mathcal{A}}(s), \; V_{H+1}^{*}(s) = 0. \tag{3.2}$$

We remark that the above equations implicitly assume the context $C = \{c_h\}_{h \in [H]}$ (and thus the priors) are determined in advance. To emphasize the values' dependence on context which will be useful for the analysis of later learning algorithms, we extend the notation to $V_h^{\pi}(s; C), Q_h^{\pi}(s, \omega, a; C)$ to specify that the value (resp. Q) function is estimated based on which prior μ conditioned on which sequence of context C.

A Note on Computational Efficiency. We note that the above Bellman Optimality Equation in (3.2) also implies an efficient dynamic program to compute the optimal policy $\boldsymbol{\pi}^*$ in the basic tabular model of MPP in Subsection 3.1, i.e., when $s \in \mathcal{S}, \omega \in \Omega, a \in \mathcal{A}$ are all discrete. This is because the maximization problem in

equation (3.2) can be solved efficiently be a linear program. The generalized MPP of subsection 3.2 imposes some computational challenge due to infinitely many outcomes and states. Fortunately, it is already known that planning in the infinite state MDP with linear function approximation can also be solved efficiently [25]. Following a similar analysis, we can determine $Q_h^*(\cdot,\cdot,\cdot)$ through a linear function of $q_h^* \in \mathbb{R}^{d_\psi}$ with the observed feature $\psi(\cdot,\cdot,\cdot)$. Hence, the dominating operation is to compute $\max_{\pi \in \mathrm{Pers}(\mu_h, u_h)} \langle Q_h^*, \mu_h \otimes \pi_h \rangle_{\Omega \times \mathcal{A}}(s)$ at each step. Let the sender utility function be Q_h^*; such an optimization is exactly the problem of optimal information design with infinitely many outcomes but finitely many actions, which has been studied in previous work [15]. It turns out that there is an efficient algorithm that can signal on the fly for any given outcome ω and obtains an ϵ-optimal persuasive signaling scheme in $\mathrm{poly}(1/\epsilon)$ time. Therefore, in our later studies of learning, we will take these algorithms as given and simply assume that we can compute the optimal signaling scheme efficiently at any given state s. One caveat is that our regret guarantee will additionally lose an additive ϵ factor at each step due to the availability of only an ϵ-optimal algorithm, but this loss can be negligible when we set $\epsilon = O(1/(TH))$ by using a $\mathrm{poly}(TH)$ time algorithm.

4 Reinforcement Learning in MPPs and the Optimism-Pessimism Principle

In this section, we study online reinforcement learning (RL) for learning the optimal signaling policy on an MPP. Here the learner only knows the utility functions of the receivers[6] and has no prior knowledge about the prior distribution, the sender's utility function, and the transition kernel. While the computation of optimal policy in MPPs in Section 3.3 may appear analogous to that of a standard MDP, as we will see that the corresponding RL problem turns out to be significantly different, partially due to the presence of the stream of receivers, whose decisions are *self-interested* and not under the learner's control. This makes the learning challenging because if the receivers' incentives are not carefully addressed, they may take actions that are extremely undesirable to the learner. Such concern leads to the integration of the *pessimism* principle into our learning algorithm design. Specifically, our learner will be optimistic to the estimation of the Q-function, similar to many other RL algorithms, in order to encourage exploration. But more interestingly, it will be pessimistic to the uncertainty in the estimation of the prior distributions in order to prepare for detrimental equilibrium behavior. Such dual considerations lead to an interesting *optimism-pessimism principle* (OPP) for learning MPPs under the online setting. From a technical point of view, our main contribution is to prove how the mixture of optimism and pessimism principle can still lead to no regret algorithms, and this proof crucially hinges on a robust property of the MPP model which we develop and carefully apply to the regret analysis. To the best of our knowledge, this is the first time that OPP is employed to learn the optimal information design in an online fashion. We prove that it can not only satisfy incentive constraints but also guarantees efficiency in terms of both sample complexity and computational complexity.

In order to convey our key design ideas before diving into the intricate technicalities, this section singles

[6] The receiver's utility is known to the sender because the pricing rules are usually transparent, some are even set by the platform. For example, a rider-sharing platform usually sets per hour or mile payment rules for the drivers.

out two representative special cases of the online sequential information design problem. In a nutshell, we present a learning algorithm `OP4` that combines the principle of optimism and pessimism such that the sender can learn to persuade without initially knowing her own utility or the prior distribution of outcomes. In the *tabular MPP*, we illustrate the unique challenges of learning to persuade arising from the dynamically evolving environment state according to a Markov process. Through the *contextual Bayesian persuasion*, we showcase the techniques necessary for learning to persuade with infinitely many states (i.e., contexts) and outcomes. We shall omit most proofs in this section to focus on the high-level ideas, because the proof for the general setting presented in Section 5 suffices to imply all results for the two special cases here.

4.1 Learning Optimal Policies in MPPs: Setups and Benchmarks

We consider the episodic reinforcement learning problem in finite-horizon MPPs. Different from the full knowledge setting in Section 3.3, the transition kernel, the sender's utility function and the outcome prior at each step of the episode, $\{P_h, v_h, \mu_h\}_{h\in[H]}$, are all unknown. The sender has to learn the optimal signaling policy by interacting with the environment as well as a stream of receivers in T number of episodes. For each $t \in [T] = \{1, \cdots, T\}$, at the beginning of t-th episode, given the data $\{(c_h^\tau, s_h^\tau, \omega_h^\tau, a_h^\tau, v_h^\tau)\}_{h\in[H],\tau\in[t-1]}$, the adversary picks the context sequence $\{c_h^t\}_{h\in[H]}$ as well as the initial state s_1^t, and the agent accordingly chooses a signaling policy $\boldsymbol{\pi}^t = \{\pi_h^t\}_{h\in[H]}$. Here v_h^τ is the utility collected by the sender at step h of episode τ.

Regret To evaluate the online learning performance, given the ground-truth outcome prior $\boldsymbol{\mu}^* = \{\mu_h^*\}_{h\in[H]}$, we define the sender's total (expected) regret over the all T episodes as

$$\mathrm{Reg}(T, \boldsymbol{\mu}^*) \coloneqq \sum_{t=1}^{T} \left[V_1^*(s_1^t; C^t) - V_1^{\boldsymbol{\pi}^t}(s_1^t; C^t) \right]. \tag{4.1}$$

Note that if $\boldsymbol{\pi}^t$ is not always feasible under $\boldsymbol{\mu}^*$, but is only persuasive with high probability, so the corresponding regret under $\boldsymbol{\pi}^t$ should be also in high probability sense.

It turns out that in certain degenerate cases it is impossible to achieve a sublinear regret. For example, if the set of possible posterior outcome distributions that induce some $a \in \mathcal{A}$ as the optimal receiver action has zero measure, then such posterior within a zero-measure set can never be exactly induced by a signaling scheme without a precise knowledge of the prior. Thus, the regret could be $\Omega(T)$ if receiver cannot be persuaded to play such action a. Therefore, to guarantee no regret, it is necessary to introduce certain regularity assumption on the MPP instance. Towards that end, we shall assume that the receivers' utility u and prior μ at any step of the MPP instance always satisfies a minor assumption of (p_0, D)-regularity as defined below.

Regularity Conditions An instance satisfies (p_0, D)-*regularity*, if for any feasible state $s \in \mathcal{S}$ and context $c \in \mathcal{C}$, we have

$$\mathbb{P}_{\omega \sim \mu(\cdot|c)} \left[\omega \in \mathcal{W}_{s,a}(D)\right] \geq p_0, \quad \forall a \in \mathcal{A},$$

where μ is the ground-truth prior of outcomes and $\mathcal{W}_{s,a}(D) \triangleq \{\omega : u(s,\omega,a) - u(s,\omega,a') \geq D, \forall a' \in \mathcal{A}/\{a\}\}$ is the set of outcomes ω for which the action a is optimal for the receiver by at least D at state s. . In other words, an instance is (p_0, D)-regular if every action a has at least probability p_0, under randomness of the outcome, to be strictly better than other actions by at least D. This regularity condition is analogous to a regularity condition of Zu et al. [61] but is generalizable to infinite outcomes as we consider here.

4.2 Algorithm: Optimism-Pessimism Principle for Persuasion Process (OP4)

The learning task in MPPs involves two intertwined challenges: (1) How to persuade the receiver to take desired actions under unknown μ_h? (2) Which action to persuade the receiver to take in order to explore the underlying environment? For the first challenge, due to having finite data, it is impossible to perfectly recover μ_h. We can only hope to construct an approximately accurate estimator of μ_h. To guard against potentially detrimental equilibrium behavior of the receivers due to the prior estimation error, we propose to adopt the pessimism principle. Specifically, before each episode, we conduct uncertainty quantification for the estimator of the prior distributions, which enables us to construct a confidence region containing the true prior with high probability. Then we propose to find the signaling policy within a pessimistic candidate set—signaling policies that are simultaneously persuasive with respect to all prior distributions in the confidence region. When the confidence region is valid, such a pessimism principle ensures that the executed signaling policy is always persuasive with respect to the true prior. Furthermore, to address the second challenge, we adopt the celebrated principle of optimism in the face of uncertainty [28], which has played a key role in the online RL literature. The main idea of this principle is that, the uncertainty of the Q-function estimates essentially reflects our uncertainty about the underlying model. By adding the uncertainty as a bonus function, we encourage actions with high uncertainty to be recommended and thus taken by the receiver when persuasiveness is satisfied. We then fuse the two principles into the `OP4` algorithm in Algorithm 1.

Algorithm 1 OP4 Overview

for episode $t = 1 \dots T$ **do**
  Receive the initial state $\{s_1^t\}$ and context $C^t = \{c_h^t\}_{h=1}^H$.
  For each step $h \in [H]$, estimate prior μ_h^t along with the confidence region $\mu_{\mathcal{B}_h^t}$, and construct an optimistic Q-function Q_h^t iteratively with the value function V_h^t.
  for step $h = 1, \dots, H$ **do**
    Choose robust signaling scheme $\pi_h^t \in \arg\max_{\pi_h \in \mathrm{Pers}(\mu_{\mathcal{B}_h^t}, u_h)} \left\langle Q_h^t, \mu_h^t \otimes \pi_h \right\rangle_{\Omega \times \mathcal{A}} (s_h^t; C^t)$.
    Observe state s_h, outcome ω_h and accordingly recommend action $a \sim \pi_h^t(\omega_h, \cdot)$ to the receiver.
  end for
end for

Pessimism to Induce Robust Equilibrium Behavior From the data in the past episode, the sender can estimate the mean of the prior as well as obtain a confidence region through concentration inequalities. Given this partial knowledge of the prior distribution, the sender needs to design a signaling scheme that works in the face of any possible priors in the confidence region in order to ensure the receiver will take its recommended action with high probability. Specifically, we let $\mathrm{B}_\Sigma(\theta, \beta) := \{\theta' : \|\theta' - \theta\|_\Sigma \leq \beta\}$ denote the closed ball in $\|\cdot\|_\Sigma$ norm of radius $\beta > 0$ centered at $\theta \in \mathbb{R}^{d_\theta}$. For any set $\mathcal{B} \subseteq \mathbb{R}^{d_\theta}$, we let $\mathrm{Pers}(\mu_\mathcal{B}, u)$ denote the set of signaling policies that are simultaneously persuasive under all weigh vectors $\theta \in \mathcal{B}$: $\mathrm{Pers}(\mu_\mathcal{B}, u) := \bigcap_{\theta \in \mathcal{B}} \mathrm{Pers}(\mu_\theta, u)$. For any non-empty set $\mathcal{B}$, the set $\mathrm{Pers}(\mu_\mathcal{B}, u)$ is convex since it is an intersection of convex sets $\mathrm{Pers}(\mu_\theta, u)$, and is non-empty since it must contain the full-information signaling scheme. We note that since $\mathrm{Pers}(\mu_\mathcal{B}, u)$ is a convex set, we can solve the linear optimization among the policies in $\mathrm{Pers}(\mu_\mathcal{B}, u)$ in polynomial time (see e.g., [61]).

Optimism to Encourage Exploration In order to balance exploration and exploitation, we adopt the principle of optimism in face of uncertainty to the value iteration algorithm based on Bellman equation, following in a line of work in online RL such as Q-learning with UCB exploration [24], UCBVI [4], LSVI-UCB [25] (also see [50, 56, 51] and the references therein). The additional UCB bonus on the Q-value encourages exploration and has been shown to be a provably efficient online method to improve policies in MDPs. Moreover, this method not only works for the simple tabular setting, but also generalizes to settings with infinite state spaces by exploiting linearity of the Q-function and a regularized least-squares program to determine the optimal estimation of Q-value. In fact, within our framework, we could obtain efficient learning result in the infinite state space setting through other optimism-based online RL methods and general function approximators, such as linear mixture MDPs [3, 60, 9, 59], or kernel approximation [56] or bilinear classes [13].

To provide a concrete picture of the learning process, we instantiate the OP4 algorithm in two special cases and showcase our key ideas and techniques before delving into the more involved analysis of the generalized MPP setting. Nevertheless, we remark that whether the problem instance is tabular or in the form of linear or generalized linear approximations is not essential and not the focus of our study. OP4 itself

only relies on two things, i.e., the uncertainty quantification for Q-function and prior estimation. So even the model-free RL framework can be replaced by model-based RL, as we can just construct confidence region for the transition models.

4.3 Warm-up I: Reinforcement Learning in the Tabular MPP

We first consider MPPs in tabular setting with finite states and outcomes, as described in Section 3.1. In this case, the prior on outcomes at each step degenerates to an unknown but fixed discrete distribution independent of context. As linear parameterization is not required for discrete probability distribution, the algorithm can simply update the empirical estimation of μ_h^t through counting. Similarly, the transition kernel P_h^* is estimated through the occurrence of observed samples, and we uses this estimated transition to compute the Q-function $\widehat{Q}_h^t$ from the observed utility and estimated value function in the next step, according to the Bellman equation. To be specific, for each $s \in \mathcal{S}, \omega \in \Omega, a \in \mathcal{A}$, μ_h^t and $\widehat{Q}_h^t$ are estimated through the following equations:

$$\mu_h^t(\omega) \leftarrow \frac{\lambda/|\Omega| + N_{t,h}(\omega)}{\lambda + t - 1},$$
$$\widehat{Q}_h^t(s,\omega,a) \leftarrow \frac{1}{\lambda + N_{t,h}(s,\omega,a)} \sum_{\tau \in [t-1]} \left\{ \mathbb{I}(s_h^\tau = s, \omega_h^\tau = \omega, a_h^\tau = a) \left[v_h^\tau + V_{h+1}^t(s_{h+1}^\tau) \right] \right\},$$

where $N_{t,h}(\omega) = \sum_{\tau \in [t-1]} \mathbb{I}(\omega_h^\tau = \omega)$ and $N_{t,h}(s,\omega,a) = \sum_{\tau \in [t-1]} \mathbb{I}(s_h^\tau = s, \omega_h^t = \omega, a_h^t = a)$ respectively count the effective number of samples that the sender has observed arriving at ω, or the combination $\{s,\omega,a\}$), and $\lambda > 0$ is a constant for regularization.

In our learning algorithm, we determine the radius of confidence region $\mathcal{B}_h^t$ for μ_h^t according to confidence bound $\epsilon_h^t = O(\sqrt{\log(HT)/t})$. Moreover, we add a UCB bonus term of form $\rho/\sqrt{N_{t,h}(s,\omega,a)}$ to $\widehat{Q}_h^t$ to obtain the *optimistic Q-function* Q_h^t. Then, it selects a robustly persuasive signaling scheme that maximizes an optimistic estimation of Q-function with respect to the current prior estimation μ_h^t. Finally, it makes an action recommendation a_h^t using this signaling scheme, given the state and outcome realization $\{s_h^t, \omega_h^t\}$.

Theorem 4.1. *Let $\epsilon_h^t = \widetilde{O}(\sqrt{1/t})$, and $\rho = \widetilde{O}(|S| \cdot |\Omega| \cdot |A|H)$. Then under (p_0, D)-regularity, with probability at least $1 - 3H^{-1}T^{-1}$, OP4 has regret of order $\widetilde{O}\big(|C|(|S| \cdot |\Omega| \cdot |A|)^{3/2} \cdot H^2\sqrt{T}/(p_0 D)\big)$ in tabular MPPs.*

To obtain the regret of OP4, we have to consider the regret arising from different procedures. Formal decomposition of the regret is described in Lemma 6.1. Separately, we upper bound errors incurred from estimating Q-function (Lemma 6.2), the randomness of of choosing the outcome, action and next state (Lemma A.5) as well as estimating the prior of outcome and choosing a persuasive signaling scheme that is robustly persuasive for a subset of priors (Lemmas 6.3 and 6.4). As the two warm-up models are special

cases of the general MPP, the proof of the above properties follows from that of the general MPP setting in Section 6, and thus is omitted here.

4.4 Warm-up II: Reinforcement Learning in Contextual Bayesian Persuasion

We now move to another special case with $H = 1$, such that the MPP problem reduces to a contextual-bandit-like problem, where transitions no longer exist. Given a context c and a persuasive signaling policy π, the value function is simply the sender's expected utility for any $s \in \mathcal{S}$,

$$V^{\pi}(s; c) \coloneqq \int_{\omega} \sum_{a \in A} \mu(\omega|c)\pi(a|s, \omega)v(s, \omega, a)\mathrm{d}\omega.$$

The sender's optimal expected utility is defined as $V^*(s; c) \coloneqq \max_{\pi \in \mathrm{Pers}(\mu(\cdot|c),u)} V^{\pi}(s; c)$.

Meanwhile, we consider the general setting where outcome ω is a continuous random variable that subjects to a generalized linear model. To be specific, the prior μ is conditioned on the context c with the mean value $f(\phi(c)^{\top}\theta)$. For the prior μ and link function f, we assume the smoothness of the prior and the bounded derivatives of the link function:

Assumption 4.2. *There exists a constant $L_{\mu} > 0$ such that for any parameter θ_1, θ_2, we have $\left\|\mu_{\theta_1}(\cdot|c) - \mu_{\theta_2}(\cdot|c)\right\|_1 \leq L_{\mu}\left\|f\left(\phi(c)^{\top}\theta_1\right) - f\left(\phi(c)^{\top}\theta_2\right)\right\|$ for any given context c.*

Assumption 4.3. *The link function f is either monotonically increasing or decreasing. Moreover, there exists absolute constants $0 < \kappa < K < \infty$ and $0 < M < \infty$ such that $\kappa \leq |f'(z)| \leq K$ and $|f''(z)| \leq M$ for all $|z| \leq \Phi L_{\theta}$.*

It is natural to assume a Lipschitz property of the distribution in Assumption 4.2. For instance, Gaussian distributions and uniform distributions satisfy this property. Assumption 4.3 is standard in the literature [18, 52, 32]. Two example link functions are the identity map $f(z) = z$ and the logistic map $f(z) = 1/(1 + e^{-z})$ with bounded z. It is easy to verify that both maps satisfy this assumption.

Different from the tabular setting, we are now unable to use the counting-based estimator to keep track of the distribution of the possibly infinite states and outcomes. Instead, we resort to function approximation techniques and estimate the linear parameters θ^* and γ^*. In each episode, `OP 4` respectively updates the estimation and confidence region of θ^t and γ^t, with which it can determine the outcome prior under pessimism and sender's utility under optimism. To be specific, the update of θ^t solves a constrained least-squares problem and the update of q^t solves precisely a regularized one:

$$\theta^t \leftarrow \arg\min_{\|\theta\| \leq L_{\theta}} \sum_{\tau \in [t-1]} \left[\omega^{\tau} - f(\phi(c^{\tau})^{\top}\theta_h)\right]^2,$$

$$\gamma^t \leftarrow \arg\min_{\gamma \in \mathbb{R}^{\psi}} \sum_{\tau \in [t-1]} \left\|v^{\tau} - \psi(s^{\tau}, \omega^{\tau}, a^{\tau})^{\top}\gamma\right\|^2 + \lambda \left\|\gamma\right\|^2.$$

We then estimate the prior by setting $\mu^t(\cdot|c)$ to the distribution of $f\left(\phi(c)^\top \theta^t\right) + z$ and estimate the sender's utility by setting $v^t(\cdot,\cdot,\cdot) = \psi(\cdot,\cdot,\cdot)^\top \gamma^t$. On one hand, to encourage exploration, OP4 adds the UCB bonus term of form $\rho \left\|\psi(\cdot,\cdot,\cdot)\right\|_{(\Gamma^t)^{-1}}$ to the Q-function, where $\Gamma^t = \lambda I_{d_\psi} + \sum_{\tau\in[t]} \psi(s^\tau,\omega^\tau,a^\tau)\psi(s^\tau,\omega^\tau,a^\tau)^\top$ is the Gram matrix of the regularized least-squares problem and ρ is equivalent to a scalar. This is a common technique for linear bandits. On the other hand, OP4 determines the confidence region of θ^t with radius β, and ensures that signaling scheme is robustly persuasive for any possible (worst case) prior induced by linear parameters θ in this region. Combining optimism and pessimism, OP4 picks the signaling scheme among the robust persuasive set that maximizes the sender's optimistic utility.

Theorem 4.4. *Under (p_0, D)-regularity and Assumption 4.2 and 4.3, there exists an absolute constant $C_1, C_2 > 0$ such that, if we set $\lambda = \max\{1, \Psi^2\}$, $\beta = C_1(1+\kappa^{-1}\sqrt{K+M+d_\phi\sigma^2\log(T)})$, and $\rho = C_2 d_\psi\sqrt{\log(4d_\psi\Psi^2T^3)}$, then with probability at least $1-3T^{-1}$, OP4 has regret of order $\widetilde{O}\big(d_\phi\sqrt{d_\psi^3}\sqrt{T}/(p_0D)\big)$ in contextual Bayesian persuasion problems.*

Since we estimate the prior by computing an estimator θ^t, we evaluate the persuasiveness of OP4 through the probability that θ^* lies in the confidence region centered at θ^t with the radius $\beta = O(\sqrt{d_\phi\log(T)})$ in weighted norm. Due to the smoothness of the prior and the assumption of link function, the error of the estimated prior is bounded by the product of β and the weighted norm of feature vector $\|\phi(c^t)\|_{\Sigma^t} = O(1/\sqrt{t})$, which yields the same conclusion for ϵ^t in the tabular MPP case. Also compared to Li et al. [31], we do not require any regularity for Σ^t, since we add a constant matrix $\Phi^2 I$ to the Gram matrix Σ^t. This ensures that Σ^t is always lower bounded by the constant $\Phi^2 > 0$. The proof of the persuasiveness and sublinear regret of contextual bandit can be viewed as a direct reduction of the MPP case when the total step $H = 1$. We decompose the regret in the same way as that in Lemma 6.1 for MPPs and then estimate the upper bound for each item to measure the regret loss.

5 No-Regret Learning in the General Markov Persuasion Process

In this section, we present the full version of the OP4 algorithm for MPPs and show that it is persuasive with high probability and meanwhile achieves average regret $\widetilde{O}\big(d_\phi \cdot d_\psi^{3/2}H^2\sqrt{T}/(p_0D)\big)$.

In the general MPP setting with the linear utility and transition, a crucial property is that the Q-functions under any signaling policy is always linear in the feature map ψ (akin to linear MDPs [25]). Therefore, when designing learning algorithms, it suffices to focus on linear Q-functions. In our OP4 algorithm, we iteratively fit the optimal Q-function, which is parameterized by q_h^* as $\psi(\cdot,\cdot,\cdot)^\top q_h^*$ at each step $h \in [H]$. OP4 learns the Q-functions of MPPs and the prior of persuasion states simultaneously. It operates similarly as that in tabular MPPs and contextual Bayesian persuasion. At the t-th episode, given the historical data $\{(c_h^\tau, s_h^\tau, \omega_h^\tau, a_h^\tau, v_h^\tau)\}_{h\in[H],\tau\in[t-1]}$, we can estimate the unknown vectors $\theta_h^*, q_h^*, \forall h \in [H]$ by solving the following constrained or regularized least-squares problems:

$$\theta_h^t \leftarrow \operatorname*{argmin}_{\|\theta_h\| \le L_\theta} \sum_{\tau \in [t-1]} \left[\omega_h^\tau - f(\phi(c_h^\tau)^\top \theta_h)\right]^2,$$
$$q_h^t \leftarrow \operatorname*{argmin}_{q \in \mathbb{R}^{d_\psi}} \sum_{\tau \in [t-1]} \left[v_h^\tau + V_{h+1}^t(s_{h+1}^\tau; C^t) - \psi(s_h^\tau, \omega_h^\tau, a_h^\tau)^\top q\right]^2 + \lambda \|q\|^2.$$

Additionally, V_{h+1}^t is the estimated value function with the observed context C^t at the episode t, which we describe formally later. This estimator is used to replace the unknown transition P_h and distribution ν_h in equation (3.2). Moreover, we can update the estimate of outcome prior μ_h^t and Q-function Q_h^t respectively. Here `OP4` adds UCB bonus to Q_h^t to encourage exploration. The formal description is given in Algorithm 2.

Likewise, the MPP setting inherits the regularity conditions and Assumption 4.2 and 4.3 in the last section. Combining the insights from both the tabular MPPs and contextual Bayesian persuasion, we can show that the `OP4` is persuasive and guarantees sublinear regret with high probability for general MPPs.

Theorem 5.1. *Under (p_0, D)-regularity and Assumption 4.2 and 4.3, there exists absolute constants $C_1, C_2 > 0$ such that, if we set $\lambda = \max\{1, \Psi^2\}$, $\beta = C_1(1+\kappa^{-1}\sqrt{K + M + d_\phi \sigma^2 \log(HT)})$,and $\rho = C_2 d_\psi H \sqrt{\log(4 d_\psi \Psi^2 H^2 T^3)}$, then with probability at least $1 - 3H^{-1}T^{-1}$, `OP4` has regret of order $\widetilde{O}\big(d_\phi d_\psi^{3/2} H^2 \sqrt{T}/(p_0 D)\big)$.*

Recall that the novelty of `OP4` is that we adopt pessimism and optimism to induce robust equilibrium behavior and encourage exploration simultaneously. Specifically, pessimism tackles the uncertainty in the prior estimation by selecting a signaling policy that is persuasive w.r.t. all the priors in the confidence region, while optimism in Q-function estimation encourages exploration. To evaluate the regret of `OP4`, we provide a novel regret decomposition, which is tailored to this pessimism and optimism combination. Each term represents different aspects of regret loss incurred by either estimation or randomness.

6 Proof Sketch and Technical Highlights

In this section, we present the proof sketch for Theorem 5.1. We first decompose the regret into several terms tailored to MPPs and briefly introduce how to bound each term. Then we highlight our technical contribution about regularity when measuring the loss in the sender's utility for choosing a signaling scheme that is persuasive for a subset of priors close to each other.

6.1 Proof of Theorem 5.1

In order to prove the sublinear regret for `OP4`, we construct a novel regret decomposition tailored to MPPs. Our proof starts from decomposing the regret into several terms, each of which indicates the regret loss either from estimation or from the randomness of trajectories. Next, we evaluate each term and then add

them together to conclude the upper bound of the regret of OP4. For simplicity of presentation, denote $\widetilde{V}_h^t(\cdot; C) = \left\langle Q_h^t, \mu_h^* \otimes \pi_h^t \right\rangle_{\Omega \times \mathcal{A}}(\cdot; C)$ as the expectation of Q_h^t with respect to the ground-truth prior μ_h^* and signaling scheme π_h^t at the h-th step. Then we can define the temporal-difference (TD) error as

$$\delta_h^t(s, \omega, a) = (v_h^t + P_h V_{h+1}^t - Q_h^t)(s, \omega, a; C^t). \tag{6.1}$$

Here δ_h^t is a function on $\mathcal{S} \times \Omega \times \mathcal{A}$ for all $h \in [H]$ and $t \in [T]$. Intuitively, $\{\delta_h^t\}_{h \in [H]}$ quantifies how far the Q-functions $\{Q_h^t\}_{h \in [H]}$ are from satisfying the Bellman optimality equation in equation (3.2). Moreover, define $\zeta_{t,h}^1$ and $\zeta_{t,h}^2$ for the trajectory $\{c_h^t, s_h^t, \omega_h^t, a_h^t\}_{h \in [H]}$ generated by Algorithm 2 at the t-th episode as follows

$$\begin{aligned} \zeta_{t,h}^1 &= (\widetilde{V}_h^t - V_h^{\pi^t})(s_h^t; C^t) - (Q_h^t - Q_h^{\pi^t})(s_h^t, \omega_h^t, a_h^t; C^t), \\ \zeta_{t,h}^2 &= P_h(V_{h+1}^t - V_{h+1}^*)(s_h^t, \omega_h^t, a_h^t; C^t) - (V_{h+1}^t - V_{h+1}^*)(s_{h+1}^t; C^t). \end{aligned} \tag{6.2}$$

By definition, $\zeta_{t,h}^1$ capture the randomness of realizing the outcome $\omega_h^t \sim \mu_h^*(\cdot | c_h)$ and signaling the action $a_h^t \sim \pi_h^t(s_h^t, \omega_h^t, \cdot)$, while $\zeta_{t,h}^2$ captures the randomness of drawing the next state s_{h+1}^t from $P_h(\cdot | s_h^t, \omega_h^t, \cdot)$. With the notations above, we can decompose the regret into six parts to facilitate the establishment of the upper bound of the regret.

Lemma 6.1 (Regret Decomposition)**.** *With the notations defines in equation* (6.1) *and* (6.2)*, we can write the regret as:*

$$\begin{aligned} \mathrm{Reg}(T, \mu^*) = & \underbrace{\sum_{t \in [T]} \sum_{h \in [H]} \left\{ \mathbb{E}_{\mu_h^*, \pi_h^*}[\delta_h^t(s_h, \omega_h, a_h^t) | s_1 = s_1^t] - \delta_h^t(s_h^t, \omega_h^t, a_h^t) \right\}}_{\text{(i)}} + \underbrace{\sum_{t \in [T]} \sum_{h \in [H]} (\zeta_{t,h}^1 + \zeta_{t,h}^2)}_{\text{(ii)}} \\ & + \underbrace{\sum_{t \in [T]} \sum_{h \in [H]} \mathbb{E}_{\mu_h^*, \pi_h^*} \left[\left\langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^t \otimes \pi_h^t \right\rangle_{\Omega \times \mathcal{A}}(s_h; C^t) | s_1 = s_1^t \right]}_{\text{(iii)}} \\ & + \underbrace{\sum_{t \in [T]} \sum_{h \in [H]} \left\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t \right\rangle_{\Omega \times \mathcal{A}}(s_h^t, C^t)}_{\text{(iv)}}. \end{aligned} \tag{6.3}$$

In this novel regret decomposition, term (i) indicates the optimism in OP4. Provably, δ_h^t in term (i) is always non-positive due to the optimistic Q-value estimation, which could simplify this term. Term (iii) corresponds to the pessimism in OP4 for inducing a robust equilibria. It evaluates the regret loss incurred by choosing a robustly persuasive signaling policy. Since the signaling policy has to be persuasive to ensure that receivers will always take recommended actions, we cannot simply choose a greedy policy for a fixed prior

estimation. Instead, we first apply optimism to construct the optimistic Q-value estimation and then apply pessimism to select a signaling policy that is robustly persuasive for all the priors in the confidence region. Therefore, we design the regret decomposition, especially term (iii) in this form to reflect the optimism and pessimism principle in `OP4`. Notice that this decomposition does not depend on specific function approximation forms in the algorithm, since not only the estimation of prior and Q-function but also the chosen signaling policy has no influence on this formula. Therefore, it generally suits all the algorithms for MPPs.

Unlike the regret decomposition in [56], Lemma 6.1 also captures the randomness of realizing the outcome. Since we have to estimate the prior of the outcome and choose a robustly persuasive policy in MPPs, we add term (iii) and (iv) to evaluate the further regret loss.

The rigorous arguments turn out to be technical, and thus we shall defer the proof of most lemmas to the appendix while aiming to present all the key ideas and conclusions in the following. For term (i) in equation (6.3), although we do not observe the trajectories under prior μ^* and signaling policy π^*, we can upper bound both δ_h^t and $-\delta_h^t$. The following lemma states this result.

Lemma 6.2 (Optimism)**.** *There exists an absolute constant $c > 0$ such that, for any fixed $\delta \in (0,1)$, if we set $\lambda = \max\{1, \Psi^2\}$ and $\rho = cd_\psi H\sqrt{\iota}$ in Algorithm 2 with $\iota = \log(2d_\psi\Psi^2 T/\delta)$, then with probability at least $1-\delta/2$, we have*

$$-2\rho\|\psi(s,\omega,a)\|_{(\Gamma_h^t)^{-1}} \le \delta_h^t(s,\omega,a) \le 0.$$

for all $s \in \mathcal{S}, \omega \in \Omega, a \in \mathcal{A}, h \in [H]$ and $t \in [T]$.

Term (ii) in equation (6.3) can be bounded by Lemma 5.3 from [56] using martingale techniques and the Azuma-Hoeffding inequality [5]. We state the upper bound for term (ii) in Lemma A.5. Moreover, term (iii) in equation (6.3) evaluates the regret loss caused by estimating the prior and choosing a robustly persuasive signaling policy. Here, we apply the robustness gap Gap defined later to bound this term.

Lemma 6.3 (Bounding Term (iii))**.** *On the event of $\{\theta_h^* \in \mathcal{B}_h^t\}$, under Assumption 4.2 and 4.3, we have*

$$\sum_{t\in[T]}\sum_{h\in[H]} \mathbb{E}_{\mu_h^*,\pi_h^*}\left[\left\langle Q_h^t, \mu_h^*\otimes\pi_h^* - \mu_h^t\otimes\pi_h^t\right\rangle_{\Omega\times\mathcal{A}}(s_h;C^t)|s_1 = s_1^t\right] \le \left(\frac{3HL_\mu K}{p_0 D} + \frac{HL_\mu K}{2}\right)\beta \sum_{h\in[H]}\sum_{t\in[T]} \|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}.$$

It remains to bound term (iv) in equation (6.3). This bound can be derived from Holder inequality and the property of the prior.

Lemma 6.4 (Bounding Term (iv))**.** *On the event of $\{\theta_h^* \in \mathcal{B}_h^t\}$, under Assumption 4.2 and 4.3, we have*

$$\sum_{t\in[T]}\sum_{h\in[H]} \left\langle Q_h^t, (\mu_h^t - \mu_h^*)\otimes\pi_h^t\right\rangle_{\Omega\times\mathcal{A}}(s_h^t;C^t) \le HL_\mu K\beta \sum_{h\in[H]}\sum_{t\in[T]} \|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}.$$

Now we are ready to prove our main result, Theorem 5.1. By the decomposition in Lemma 6.1 and all previous lemmas, let $\beta = C(1+\kappa^{-1}\sqrt{K+M+d_\phi\sigma^2\log(HT)})$, and then we obtain the following upper bound for regret:

$$\begin{aligned}\mathrm{Reg}(T,\mu^*) \leq & 4\sqrt{2TH^3\log(2HT)} \\ & + \sum_{t\in[T]}\sum_{h\in[H]}\left[2\rho\big\|\psi(s_h^t,\omega_h^t,a_h^t)\big\|_{(\Gamma_h^t)^{-1}} + \left(\frac{3HL_\mu K}{p_0 D}+\frac{3HL_\mu K}{2}\right)\beta\big\|\phi(c_h^t)\big\|_{(\Sigma_h^t)^{-1}}\right],\end{aligned}$$

With the probability of the given event by Lemma 6.8 and appropriately chosen δ in previous lemmas, the above inequality holds for the probability at least $1-3H^{-1}T^{-1}$.

By Lemma A.6, we have

$$\begin{aligned}\mathrm{Reg}(T,\mu^*) \leq & \, 2\sqrt{2TH^3\log(2HT)} + 2\rho H\sqrt{2d_\psi T\log\left(1+T\Psi^2/(\lambda d_\psi)\right)} \\ & + \beta H^2 L_\mu K\left(\frac{3}{p_0 D}+\frac{3}{2}\right)\sqrt{2d_\phi T\log\left(1+T/(d_\phi)\right)}.\end{aligned}$$

Since β is in $\widetilde{O}(\sqrt{d_\phi})$ and ρ is in $\widetilde{O}(d_\psi H)$, we can conclude that the regret of Algorithm 2 is $\widetilde{O}\big(d_\phi d_\psi^{3/2}H^2\sqrt{T}/(p_0 D)\big)$.

6.2 Inducing Robust Equilibria via Pessimism

One necessary prerequisite is that the signaling policy given by `OP 4` has to be persuasive to ensure receivers to take recommended actions. However, the optimal signaling policy that is persuasive for the estimated prior can hardly be also persuasive for the true prior, even if the estimation is quite close to it. To ensure persuasiveness under the prior estimation error, we adopt pessimism principle to select a signaling policy that is robustly persuasive for all the priors in the confidence region. And we shall quantify the extra utility loss suffered by the pessimism principle. In this subsection, we start by showing that there exists a robust signaling scheme that suffers only $O(\epsilon)$ utility loss compared to the optimal expected utility of persuasion algorithm designed with precise knowledge of the prior. Formally, in basic MPP, given any fixed Q-function $Q(\cdot,\cdot,\cdot)$, we define the *robustness gap* for some state $s\in\mathcal{S}$ and any prior $\mu\in\mathcal{B}\subseteq\Delta(\Omega)$ as

$$\mathrm{Gap}\left(s,\mu,\mathcal{B};Q\right) \triangleq \max_{\pi\in\mathrm{Pers}(\mu,u)}\left\langle Q,\mu\otimes\pi\right\rangle_{\Omega\times\mathcal{A}}(s) - \max_{\pi\in\mathrm{Pers}(\mathcal{B},u)}\left\langle Q,\mu\otimes\pi\right\rangle_{\Omega\times\mathcal{A}}(s). \tag{6.4}$$

We let $\mathrm{B}(\mu,\epsilon)=\{\mu'\in\Delta(\Omega):\|\mu-\mu'\|_1\leq\epsilon\}$ be the ℓ_1-norm ball centered the prior distribution μ with radius ϵ.

Lemma 6.5 (Pessimism)**.** *Under (p_0,D)-regularity, for all $\epsilon>0$, given a Q-function Q, for any state $s\in\mathcal{S}$, we have*

$$\mathrm{Gap}\left(s,\mu,\mathrm{B}(\mu,\epsilon);Q\right)\leq\frac{H\epsilon}{p_0 D}.$$

The proof is given in Appendix A.2. This result extends Proposition 1 in [61]. Notice that the upper bound of $\mathrm{Gap}(\cdot;\cdot)$ does not depend on the value of Q, which is important for our analysis. Once given a signaling algorithm, at each episode $t \in [T]$ and each step $h \in [H]$, we are able to obtain an estimation of Q-function with an explicit form. It is equivalent to the "known" Q-function mentioned in equation (6.4). Using $\mathrm{Gap}(\cdot;\cdot)$, we can estimate the expected sender's utility loss for choosing a signaling mechanism that is persuasive for all priors in a subset. Moreover, if we consider the dependence on context for priors and add the linear assumption of priors to the proceeding lemma, we can bound $\mathrm{Gap}(\cdot;\cdot)$ by the difference of linearity parameter θ.

Corollary 6.6. *Under (p_0, D)-regularity and Assumption 4.2 and 4.3, given a Q-function Q and context c, for any state $s \in \mathcal{S}$, prior $\mu_\theta(\cdot|c)$ and confidence region $\mathcal{B} = \{\mu_{\theta'}(\cdot|c) : \theta' \in \mathrm{B}_\Sigma(\theta, \epsilon)\}$, we have $\mathrm{Gap}(s, \mu_\theta(\cdot|c), \mathcal{B}; Q) \leq HL_\mu K \|\phi(c)\|_{\Sigma^{-1}} \epsilon/(p_0 D)$.*

In MPPs, we have to estimate the prior of the outcome since we cannot observe the ground-truth prior. However, the estimation may not satisfy the regularity conditions, which conflicts with the requirements for the prior when proving Lemma 6.5. To address this problem, we give another upper bound of the robustness gap for the prior estimation in Lemma A.1. In addition, to handle the regret loss incurred by estimating the prior, we compute the difference in Q-functions when choosing respectively persuasive scheme for different priors in Lemma A.2.

We now prove that the above pessimism design guarantees persuasiveness w.r.t. the true prior with high probability. And it suffices to show that the estimation θ_h^t is close enough to the real parameter θ_h^* such that the confidence region $\mathcal{B}_h^t$ centered at θ_h^t given in Algorithm 2 contains θ_h^*. If so, the signaling scheme chosen to be persuasive for the whole set $\mu_{\mathcal{B}_h^t}$ is also persuasive for μ_h^*, where $\mu_{\mathcal{B}} \coloneqq \{\mu_{\theta'} : \theta' \in \mathcal{B}\}$ denotes the set of priors that are determined by the parameters $\theta' \in \mathcal{B}$.

Lemma 6.7. *There exists a constant $C > 0$, such that for $\beta = C(1 + \kappa^{-1}\sqrt{K + M + d_\phi \sigma^2 \log(HT)})$, `OP4` Algorithm is persuasive with probability at least $1 - H^{-1}T^{-1}$, i.e.,*

$$\mathbb{P}_{\theta^*}\Big(\bigcup_{h\in[H]} \{\theta_h^* \notin \cap_{t\in[T]} \mathcal{B}_h^t\} \Big) \leq H^{-1}T^{-1}.$$

Proof. We first analyze the probability for being non-persuasive. For any $\|\theta_h^*\| \leq L_\theta$, using the union bound, we have

$$\begin{aligned} P_{\theta^*}\Big(\bigcup_{t\in[T], h\in[H]} \{\theta_h^* \notin \cap_{t\in[T]} \mathcal{B}_h^t\} \Big) &\leq \sum_{t\in[T]} \sum_{h\in[H]} P_{\theta_h^*}\big(\theta_h^* \notin \cap_{t\in[T]} \mathcal{B}_h^t\big) \\ &\leq \sum_{t\in[T]} \sum_{h\in[H]} P_{\theta_h^*}\big(\|\theta_h^t - \theta_h^*\|_{\Sigma_h^t} > \beta\big). \end{aligned}$$

The following lemma gives the belief of confidence region for the linear parameter θ_h^*. The proof can be

directly derived from Lemma 6 in Wang et al. [52].

Lemma 6.8 (Belief of Confidence Region)**.** *For any $t \in [T]$ and $h \in [H]$, there exists a constant $C > 0$, such that for $\beta = C(1 + \kappa^{-1}\sqrt{K + M + d_\phi \sigma^2 \log(1/\delta)})$, given $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have $\|\theta_h^t - \theta_h^*\|_{\Sigma_h^t} \leq \beta$.*

By Lemma 6.8, taking $\delta = H^{-2}T^{-2}$, then we have $\mathbb{P}_{\theta^*}(\|\theta_h^t - \theta_h^*\|_{\Sigma_h^t} > \beta) \leq H^{-2}T^{-2}$. Summing up the failure probabilities over $t \in [T]$, we have $\mathbb{P}_{\theta^*}(\theta^* \notin \cap_{t\in[T]}\mathcal{B}^t) \leq H^{-1}T^{-1}$. □

7 Conclusion

We have presented a novel model, the MPP, which captures the misaligned incentives of uninformed decision makers and the long-term objective of an information possessor for the first time. We then provide a reinforcement learning algorithm, `OP4`, that is provably efficient in terms of both computational complexity and sample complexity, under mild assumptions. We remark that while we showcase this algorithm in particular problem instances with linear approximation or GLMs, the framework of `OP4` does not rely on the function approximation form, as long as we can quantify the uncertainty of the prior estimation and Q-function (or transition model). In addition, we expect this optimism-pessimism design principle and its corresponding proof techniques to be generally useful for some other strategic learning problems with misaligned incentives involved.

Besides extending our techniques to other design problems, we point out that several other open problems arises from our work. First, while it is natural that the sender have knowledge of receiver's utility functions in many cases (see Footnote 6), we hope to also study the problem even without initially knowing receiver's utility. Similar problem has been studied in Stackelberg games [40, 10] yet without measuring the performance in terms of the cumulative utility of sender (leader). Second, another interesting direction is to study the setting of Markov Bayesian persuasion with one sender and one receiver, both aiming at maximizing their own long-term cumulative utilities when the environment involves Markovian transitions.

References

[1] Agarwal, A., Jiang, N., Kakade, S. M. and Sun, W. (2019). Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*.

[2] Auer, P., Jaksch, T. and Ortner, R. (2008). Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, **21**.

[3] Ayoub, A., Jia, Z., Szepesvari, C., Wang, M. and Yang, L. (2020). Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*. PMLR.

[4] Azar, M. G., Osband, I. and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*. PMLR.

[5] Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, **19** 357–367.

[6] Badanidiyuru, A., Bhawalkar, K. and Xu, H. (2018). Targeting and signaling in ad auctions. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM.

[7] Bellman, R. (1957). A markovian decision process. *Indiana Univ. Math. J.*, **6** 679–684.

[8] Bergemann, D. and Morris, S. (2019). Information design: A unified perspective. *Journal of Economic Literature*, **57** 44–95.

[9] Cai, Q., Yang, Z., Jin, C. and Wang, Z. (2020). Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*. PMLR.

[10] Conitzer, V. and Sandholm, T. (2006). Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM conference on Electronic commerce*.

[11] Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J. and Schapire, R. E. (2018). On oracle-efficient pac rl with rich observations. *Advances in neural information processing systems*, **31**.

[12] Dann, C., Lattimore, T. and Brunskill, E. (2017). Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, **30**.

[13] Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W. and Wang, R. (2021). Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*. PMLR.

[14] Du, S. S., Kakade, S. M., Wang, R. and Yang, L. F. (2019). Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*.

[15] Dughmi, S. and Xu, H. (2019). Algorithmic bayesian persuasion. *SIAM Journal on Computing* STOC16–68.

[16] Ely, J. C. (2017). Beeps. *American Economic Review*, **107** 31–53.

[17] Farhadi, F. and Teneketzis, D. (2021). Dynamic information design: a simple problem on optimal sequential information disclosure. *Dynamic Games and Applications* 1–42.

[18] Filippi, S., Cappe, O., Garivier, A. and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, **23**.

[19] Gan, J., Majumdar, R., Radanovic, G. and Singla, A. (2021). Bayesian persuasion in sequential decision-making. *arXiv preprint arXiv:2106.05137*.

[20] Giannoccaro, I. and Pontrandolfo, P. (2002). Inventory management in supply chains: a reinforcement learning approach. *International Journal of Production Economics*, **78** 153–161.

[21] Goldstein, I. and Leitner, Y. (2018). Stress tests and information disclosure. *Journal of Economic Theory*, **177** 34–69.

[22] He, J., Zhou, D. and Gu, Q. (2021). Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*. PMLR.

[23] Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J. and Schapire, R. E. (2017). Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*. PMLR.

[24] Jin, C., Allen-Zhu, Z., Bubeck, S. and Jordan, M. I. (2018). Is q-learning provably efficient? *Advances in neural information processing systems*, **31**.

[25] Jin, C., Yang, Z., Wang, Z. and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*. PMLR.

[26] Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, **101** 2590–2615.

[27] Kober, J., Bagnell, J. A. and Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, **32** 1238–1274.

[28] Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

[29] Lehrer, E. and Shaiderman, D. (2021). Markovian persuasion. *arXiv preprint arXiv:2111.14365*.

[30] Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J. and Jurafsky, D. (2016). Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

[31] Li, L., Lu, Y. and Zhou, D. (2017). Provable optimal algorithms for generalized linear contextual bandits. *CoRR*, **abs/1703.00048**.
http://arxiv.org/abs/1703.00048

[32] Li, L., Lu, Y. and Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*. PMLR.

[33] Li, M., Qin, Z., Jiao, Y., Yang, Y., Wang, J., Wang, C., Wu, G. and Ye, J. (2019). Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning. In *The world wide web conference*.

[34] Liang, E., Wen, K., Lam, W. H., Sumalee, A. and Zhong, R. (2021). An integrated reinforcement learning and centralized programming approach for online taxi dispatching. *IEEE Transactions on Neural Networks and Learning Systems*.

[35] Mansour, Y., Slivkins, A., Syrgkanis, V. and Wu, Z. S. (2021). Bayesian exploration: Incentivizing exploration in bayesian games. *Operations Research*.

[36] Meisheri, H., Baniwal, V., Sultana, N. N., Khadilkar, H. and Ravindran, B. (2020). Using reinforcement learning for a large variable-dimensional inventory management problem. In *Adaptive Learning Agents Workshop at AAMAS*.

[37] Milano, S., Taddeo, M. and Floridi, L. (2020). Recommender systems and their ethical challenges. *Ai & Society*, **35** 957–967.

[38] Neu, G. and Pike-Burke, C. (2020). A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, **33** 1392–1403.

[39] Osband, I., Van Roy, B. and Wen, Z. (2016). Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*. PMLR.

[40] Peng, B., Shen, W., Tang, P. and Zuo, S. (2019). Learning optimal strategies to commit to. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33.

[41] Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

[42] Qin, Z., Tang, X., Jiao, Y., Zhang, F., Xu, Z., Zhu, H. and Ye, J. (2020). Ride-hailing order dispatching at didi via reinforcement learning. *INFORMS Journal on Applied Analytics*, **50** 272–286.

[43] Qin, Z. T., Zhu, H. and Ye, J. (2021). Reinforcement learning for ridesharing: A survey. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE.

[44] Rabinovich, Z., Jiang, A. X., Jain, M. and Xu, H. (2015). Information disclosure as a means to security. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. Citeseer.

[45] Renault, J., Solan, E. and Vieille, N. (2017). Optimal dynamic information provision. *Games and Economic Behavior*, **104** 329–349.

[46] Russo, D. (2019). Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, **32**.

[47] Strehl, A. L., Li, L., Wiewiora, E., Langford, J. and Littman, M. L. (2006). Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*.

[48] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

[49] Tang, T. Y. and Winoto, P. (2016). I should not recommend it to you even if you will like it: the ethics of recommender systems. *New Review of Hypermedia and Multimedia*, **22** 111–138.

[50] Wang, R., Salakhutdinov, R. and Yang, L. F. (2020). Provably efficient reinforcement learning with general value function approximation.

[51] Wang, T., Zhou, D. and Gu, Q. (2021). Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *Advances in Neural Information Processing Systems*, **34**.

[52] Wang, Y., Wang, R., Du, S. S. and Krishnamurthy, A. (2019). Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*.

[53] Xiao, W., Zhao, H., Pan, H., Song, Y., Zheng, V. W. and Yang, Q. (2019). Beyond personalization: Social content recommendation for creator equality and consumer satisfaction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

[54] Xu, H., Rabinovich, Z., Dughmi, S. and Tambe, M. (2015). Exploring information asymmetry in two-stage security games. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[55] Yang, L. and Wang, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*. PMLR.

[56] Yang, Z., Jin, C., Wang, Z., Wang, M. and Jordan, M. I. (2020). Bridging exploration and general function approximation in reinforcement learning: Provably efficient kernel and neural value iterations. *CoRR*, **abs/2011.04622**.
https://arxiv.org/abs/2011.04622

[57] Zanette, A., Lazaric, A., Kochenderfer, M. and Brunskill, E. (2020). Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*. PMLR.

[58] Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D. C. and Socher, R. (2020). The ai economist: Improving equality and productivity with ai-driven tax policies. *arXiv preprint arXiv:2004.13332*.

[59] Zhou, D., Gu, Q. and Szepesvari, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*. PMLR.

[60] Zhou, D., He, J. and Gu, Q. (2021). Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*. PMLR.

[61] Zu, Y., Iyer, K. and Xu, H. (2021). Learning to persuade on the fly: Robustness against ignorance. EC '21, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3465456.3467593

A Omitted Proofs and Descriptions

A.1 Formal Description of the OP4

The formal description of the OP4 for MPPs is stated as follows:

Algorithm 2 The OP4 for MPPs

Input: Number of Episodes T, Number of Step H
Parameters: $\beta > 0$, $\rho > 0$, $\lambda \in \mathbb{R}^+$.
Output: $a_h^t \in \mathcal{A}$ for each $h \in [H], t \in [T]$
for episode $t = 1 \dots T$ **do**
  Receive the initial state s_1^t and context $C^t = (c_1^t, \dots, c_H^t)$.
  for step $h = H, \dots, 1$ **do**
    Compute the constrained least square problem

$$\theta_h^t \leftarrow \operatorname*{argmin}_{\|\theta_h\| \le L_\theta} \sum_{\tau \in [t-1]} \left[\omega_h^\tau - f(\phi(c_h^\tau)^\top \theta_h)\right]^2.$$

    Calculate $\Sigma_h^t = \Phi^2 I_{d_\phi} + \sum_{\tau \in [t-1]} \phi(c_h^\tau)\phi(c_h^\tau)^\top$. Update $\mathcal{B}_h^t \leftarrow \mathrm{B}_{\Sigma_h^t}(\theta_h^t, \beta)$.
    Set $\mu_h^t(\cdot|c)$ to the distribution of $f\left(\phi(c)^\top \theta_h^t\right) + z_h$.
    Calculate

$$\Gamma_h^t = \lambda I_{d_\psi} + \sum_{\tau \in [t-1]} \psi(s_h^\tau, \omega_h^\tau, a_h^\tau)\psi(s_h^\tau, \omega_h^\tau, a_h^\tau)^\top,$$
$$\iota_h^t = \sum_{\tau \in [t-1]} \psi(s_h^\tau, \omega_h^\tau, a_h^\tau)[v_h^\tau + V_{h+1}^t(s_{h+1}^\tau; C^t)]$$

    Update $q_h^t \leftarrow (\Gamma_h^t)^{-1}\iota_h^t$.
    Set $\begin{cases} Q_h^t(\cdot,\cdot,\cdot; C^t) \leftarrow \min\{\psi(\cdot,\cdot,\cdot)^\top q_h^t + \rho\|\psi(\cdot,\cdot,\cdot)\|_{(\Gamma_h^t)^{-1}}, H\}, \\ V_h^t(\cdot; C^t) \leftarrow \max_{\pi_h \in \mathrm{Pers}(\mu_{\mathcal{B}_h^t}, u_h)} \left\langle Q_h^t, \mu_h^t \otimes \pi_h \right\rangle_{\Omega \times \mathcal{A}}(\cdot; C^t). \end{cases}$
  end for
  for step $h = 1, \dots, H$ **do**
    Choose $\pi_h^t \in \arg\max_{\pi_h \in \mathrm{Pers}(\mu_{\mathcal{B}_h^t}, u_h)} \left\langle Q_h^t, \mu_h^t \otimes \pi_h \right\rangle_{\Omega \times \mathcal{A}}(s_h^t; C^t)$.
  end for
  Execute π^t to sample a trajectory $\{(s_h^t, \omega_h^t, a_h^t, v_h^t)\}_{h \in [H]}$.
end for

A.2 Proof of Lemma 6.5

Proof. We prove with an explicit construction of a signaling scheme that is robustly persuasive for any prior in $\mathrm{B}(\mu, \epsilon)$ and achieve the expected utility at least $\max_{\pi \in \mathrm{Pers}(\mu,u)} \left\langle Q, \mu \otimes \pi \right\rangle_{\Omega \times \mathcal{A}}(s) - H\epsilon/(p_0 D)$. To simplify the notation, we omit the s in u, Q and $\mathcal{W}$.

Let $\pi^* = \arg\max_{\pi \in \mathrm{Pers}(\mu,u)} \left\langle Q, \mu \otimes \pi \right\rangle_{\Omega \times \mathcal{A}}$ be a direct scheme without loss of generality [26]. For

each $a \in \mathcal{A}$, let $\mu_a(\cdot) \coloneqq \mu(\cdot) \odot \pi^*(a|\cdot)$ denote the posterior of outcome (i.e., kernel [7]) that action a is recommended by π, so the prior can be composed as $\mu(\cdot) = \sum_{a\in\mathcal{A}} \mu_a(\cdot)$. Since π is persuasive, we know $\int_{\omega\in\Omega} \mu_a(\omega)\left[u(\omega,a) - u(\omega,a')\right] \geq 0, \forall a' \in \mathcal{A}$.

Let π^0 be the fully revealing signaling scheme that always recommends (signals) the action that maximizes the receivers' utility at the realized outcome. For each $a \in \mathcal{A}$, let $\eta_a(\cdot) \coloneqq \mu(\cdot) \odot \pi^0(a|\cdot)$ denote the posterior of outcome that action a is recommended by π^0, so the prior can be composed as $\mu(\cdot) = \sum_{a\in\mathcal{A}} \eta_a(\cdot)$. By regularity condition, we have

$$\int_{\omega\in\Omega} \eta_a(\omega)\left[u(\omega,a) - u(\omega,a')\right] \geq \int_{\omega\in\mathcal{W}_a(D)} \eta_a(\omega)\left[u(\omega,a) - u(\omega,a')\right] \geq p_0 D, \quad \forall a' \in \mathcal{A}.$$

We now show that the signaling scheme $\pi' = (1-\delta)\pi^* + \delta\pi^0$ is persuasive for any prior $\widetilde{\mu} \in \mathrm{B}(\mu,\epsilon)$ with $\delta = \frac{\epsilon}{p_0 D}$. One simple way to interpret this "compound" signaling scheme is to follow π^* with probability $(1-\delta)$ and follow π^0 with probability δ. Hence, given a recommended action a, the receiver would compute the posterior as $\mu'_a = (1-\delta)\mu_a(\omega) + \delta\eta_a(\omega)$. Let $\mu'_a, \widetilde{\mu}_a$ be the outcome posterior of π' recommending action a under the true prior μ (resp. the perturbed prior $\widetilde{\mu}$). So $\mu'_a(\cdot) = \mu(\cdot) \odot \pi'(a|\cdot)$ and $\widetilde{\mu}_a(\cdot) = \widetilde{\mu}(\cdot) \odot \pi'(a|\cdot)$. By definition of persuasiveness, we need to show that for any recommended action (signal from π') $a \in \mathcal{A}$, the action a maximizes the receiver's utility under μ'_a. This follows from the decomposition below,

$$\begin{aligned}
&\int_{\omega\in\Omega} \widetilde{\mu}_a \cdot \left[u(\omega,a) - u(\omega,a')\right] \\
\geq &\int_{\omega\in\Omega} \mu'_a \cdot \left[u(\omega,a) - u(\omega,a')\right] - \left\|\widetilde{\mu}_a - \mu'_a\right\|_1 \\
\geq &\int_{\omega\in\Omega} \left[(1-\delta)\mu_a(\omega) + \delta\eta_a(\omega)\right] \cdot \left[u(\omega,a) - u(\omega,a')\right] - \left\|\widetilde{\mu}_a - \mu_a\right\|_1 \\
= &\int_{\omega\in\Omega} (1-\delta)\mu_a(\omega)\left[u(\omega,a) - u(\omega,a')\right] + \int_{\omega\in\Omega} \delta\eta_a(\omega)\left[u(\omega,a) - u(\omega,a')\right] - \left\|\widetilde{\mu}_a - \mu_a\right\|_1 \\
\geq &\, \delta p_0 D - \left\|\widetilde{\mu}_a - \mu_a\right\|_1 \\
= &\, \epsilon - \left\|\widetilde{\mu}_a - \mu_a\right\|_1 \geq 0.
\end{aligned}$$

The first inequality is by the fact that $u(\omega,a) \in [0,1]$ for any ω, a and thus $\sum_a (\widetilde{\mu}_a - \mu'_a)\cdot[u(\omega,a) - u(\omega,a')] \leq \|\widetilde{\mu}_a - \mu'_a\|_1$. The second inequality is from $\mu'_a = (1-\delta)\mu_a(\omega) + \delta\eta_a(\omega)$. The third inequality is by construction of μ_a and η_a induced by signaling scheme π and π^0. The last inequality is by the fact that $\|\widetilde{\mu}_a - \mu'_a\|_1 = \|(\widetilde{\mu} - \mu') \odot \pi'(a|\cdot)\|_1 \leq \|\widetilde{\mu} - \mu'\|_1 = \epsilon$, since $\|\pi'(a|\cdot)\|_\infty \leq 1$

It remains to show the expected utility under signaling scheme π' is at least $\left\langle Q, \mu\otimes\pi^*\right\rangle_{\Omega\times\mathcal{A}} - H\epsilon/(p_0 D)$.

[7] In this proof, we will directly work with the posterior without normalization (kernel) to simplify our notations and derivations, because $\int_{\omega\in\Omega} \mu_a(\omega)\left[u(\omega,a) - u(\omega,a')\right] \geq 0 \iff \int_{\omega\in\Omega} \frac{\mu_a(\omega)}{\int_{\omega\in\Omega}\mu_a(\omega)}\left[u(\omega,a) - u(\omega,a')\right] \geq 0$. We use $\odot$ to denote the Hadamard product.

This is due to the following inequalities,

$$\begin{aligned}\left\langle Q, \mu \otimes \pi' \right\rangle_{\Omega\times\mathcal{A}} - \left\langle Q, \mu \otimes \pi^* \right\rangle_{\Omega\times\mathcal{A}} &= \int_{\omega\in\Omega, a\in\mathcal{A}} \mu(\omega)\left[\pi'(a|\omega) - \pi^*(a|\omega)\right] Q(\omega, a) \\ &= \int_{\omega\in\Omega, a\in\mathcal{A}} \mu(\omega)\left[\delta\pi^0(a|\omega) - \delta\pi^*(a|\omega)\right] Q(\omega, a) \\ &\geq -\delta \int_{\omega\in\Omega, a\in\mathcal{A}} \mu(\omega)\pi(a|\omega)Q(\omega, a) \\ &\geq -H\delta = -\frac{H\epsilon}{p_0 D}.\end{aligned}$$

The first and second equalities use the definition and linearity. The third and last inequalities use the fact that $\mathbb{E}[Q(\omega, a)] \in [0, H]$ and remove the positive term. □

A.3 Properties for the Robustness Gap

We present the robustness gap Gap for the ground-truth prior in Lemma 6.5. For the estimation of prior μ_h^t given in Algorithm 2 which may not satisfy the regularity condition, we also have corresponding robustness gap.

Lemma A.1. *For any $h \in [H], t \in [T]$ and $s \in \mathcal{S}$, on the event of $\{\theta_h^* \in \mathcal{B}_h^t\}$, we have*

$$\mathrm{Gap}(s, \mu_h^t, \mathrm{B}(\mu_h^t, \epsilon_h^t); Q_h^t) \leq \frac{2H\epsilon}{p_0 D}.$$

Proof. For any fixed action $a \in \mathcal{A}$, on the given event, we have

$$\begin{aligned}\mathbb{P}_{\omega\sim\mu_h^t(\cdot)}[\omega \in \mathcal{W}_{s,a}(D)] &= \int_{\omega\in\Omega} \mu_h^t(\omega)\mathbb{I}(\omega \in \mathcal{W}_{s,a}(D))\mathrm{d}\omega \\ &= \int_{\omega\in\Omega} \mu_h^*(\omega)\mathbb{I}(\omega \in \mathcal{W}_{s,a}(D))\mathrm{d}\omega + \int_{\omega\in\Omega} [\mu_h^t(\omega) - \mu_h^*(\omega)]\mathbb{I}(\omega \in \mathcal{W}_{s,a}(D))\mathrm{d}\omega \\ &\geq \int_{\omega\in\Omega} \mu_h^*(\omega)\mathbb{I}(\omega \in \mathcal{W}_{s,a}(D))\mathrm{d}\omega + \|\mu_h^t - \mu_h^*\|_1 \\ &\geq p_0 - \epsilon_h^t,\end{aligned}$$

where $\mathbb{I}$ is the indicating function. The last inequality uses the regularity condition for the real prior μ_h^*. For $\epsilon_h^t \leq p_0/2$, we have $\mathbb{P}_{\omega\sim\mu_h^t(\cdot)}[\omega \in \mathcal{W}_{s,a}] \leq p_0/2$. Then by Lemma 6.5, we can arrive at

$$\mathrm{Gap}(s, \mu_h^t, \mathrm{B}(\mu_h^t, \epsilon_h^t); Q_h^t) \leq \frac{2H\epsilon_h^t}{p_0 D}.$$

For $\epsilon_h^t > p_0/2$, the bound holds trivially since $2H\epsilon_h^t/(p_0 D) > H$. □

The robustness gap Gap defined in equation (6.4) measures the loss in value functions for being robustly

persuasive for a subset of priors. In the following lemma, we show that we can also use Gap to bound the difference in expected optimal Q-functions between different priors.

Lemma A.2. *Denote $\mathcal{B}_{1,2} \coloneqq \mathrm{B}\big(\mu_1, \|\mu_1 - \mu_2\|_1\big)$ for any fixed state $s \in \mathcal{S}$ and $\mu_1, \mu_2 \in \Delta(\Omega)$. Then given a known Q-function $Q(\cdot,\cdot,\cdot)$, we have*

$$\max_{\pi_1 \in \mathrm{Pers}(\mu_1,u)} \big\langle Q, \mu_1 \otimes \pi_1 \big\rangle_{\Omega\times\mathcal{A}}(s) - \max_{\pi_2 \in \mathrm{Pers}(\mu_2,u)} \big\langle Q, \mu_2 \otimes \pi_2 \big\rangle_{\Omega\times\mathcal{A}}(s) \le \mathrm{Gap}(s, \mu_1, \mathcal{B}_{1,2}; Q) + \frac{H}{2}\|\mu_1 - \mu_2\|_1.$$

Proof. Fix $\mu_1, \mu_2 \in \Delta(\Omega)$, we respectively choose the optimal signaling scheme

$$\pi_i = \operatorname*{argmax}_{\pi_i \in \mathrm{Pers}(\mu_i,u)} \big\langle Q, \mu_i \otimes \pi_i \big\rangle_{\Omega\times\mathcal{A}}(s),\ i = 1, 2.$$

Then among all the signaling schemes persuasive for all $\mathcal{B}_{1,2}$, let π_3 maximize $\big\langle Q, \mu_1 \otimes \pi \big\rangle_{\Omega\times\mathcal{A}}(s)$. Since π_3 is persuasive for μ_2, we know $\big\langle Q, \mu_2 \otimes \pi_2 \big\rangle_{\Omega\times\mathcal{A}}(s) \ge \big\langle Q, \mu_2 \otimes \pi_3 \big\rangle_{\Omega\times\mathcal{A}}(s)$ by definition. Therefore, we have

$$\begin{aligned}
\big\langle Q, \mu_1 \otimes \pi_1 - \mu_2 \otimes \pi_2 \big\rangle_{\Omega\times\mathcal{A}}(s) &\le \big\langle Q, \mu_1 \otimes \pi_1 - \mu_2 \otimes \pi_3 \big\rangle_{\Omega\times\mathcal{A}}(s) \\
&\le \big\langle Q, \mu_1 \otimes \pi_1 - \mu_1 \otimes \pi_3 \big\rangle_{\Omega\times\mathcal{A}}(s) + \big\langle Q, \mu_1 \otimes \pi_3 - \mu_2 \otimes \pi_3 \big\rangle_{\Omega\times\mathcal{A}}(s) \\
&= \mathrm{Gap}(s, \mu_1, \mathcal{B}_{1,2}; Q) + \frac{H}{2}\|\mu_1 - \mu_2\|_1.
\end{aligned}$$

The last equality uses the definition of Gap and Lemma A.3. □

Lemma A.3. *Given a Q-function $Q(\cdot,\cdot,\cdot) \in [0, H]$, for any fixed state $s \in \mathcal{S}$, $\mu_1, \mu_2 \in \Delta(\Omega)$ and any signaling scheme π, we have*

$$\big|\big\langle Q, \mu_1 \otimes \pi \big\rangle_{\Omega\times\mathcal{A}}(s) - \big\langle Q, \mu_2 \otimes \pi \big\rangle_{\Omega\times\mathcal{A}}(s)\big| \le \frac{H}{2}\|\mu_1 - \mu_2\|_1.$$

Proof. Fix $\mu_1(\cdot), \mu_2(\cdot) \in \Delta(\Omega)$. For any $x \in \mathbb{R}$, we have

$$\begin{aligned}
\big|\big\langle Q, \mu_1 \otimes \pi - \mu_2 \otimes \pi \big\rangle_{\Omega\times\mathcal{A}}(s) &= \left| \int_{\omega\in\Omega} [\mu_1(\omega) - \mu_2(\omega)] \left[\int_{a\in\mathcal{A}} \pi(a|s,\omega) Q(s,\omega,a) \mathrm{d}a - x \right] \mathrm{d}\omega \right| \\
&\le \|\mu_1 - \mu_2\|_1 \cdot \sup_{\omega\in\Omega} \left| \int_{a\in\mathcal{A}} \pi(a|s,\omega) Q(s,\omega,a) \mathrm{d}a - x \right|,
\end{aligned}$$

where the last inequality is derived from Holder's inequality. With Q-function taking values in $[0, H]$, we can set $x = H/2$ and achieve the optimality.

□

A.4 Proof of Lemma 6.1

Proof. Before presenting the proof, we first define two operators $\mathbb{J}_h^*$ and $\mathbb{J}_h^t$:

$$(\mathbb{J}_h^* f)(s;C) = \langle f, \mu_h^* \otimes \pi_h^* \rangle_{\Omega\times\mathcal{A}}(s;C), \quad (\mathbb{J}_h^t f)(s;C) = \langle f, \mu_h^t \otimes \pi_h^t \rangle_{\Omega\times\mathcal{A}}(s;C), \tag{A.1}$$

for any $h \in [H], t \in [T]$ and any function $f(\cdot,\cdot,\cdot;C) : \mathcal{S}\times\Omega\times\mathcal{A} \to \mathbb{R}$ under the context C. Moreover, for any $h \in [H], t \in [T]$ and any state $s \in \mathcal{S}$, we define

$$\xi_h^t(s;C) = (\mathbb{J}_h^* Q_h^t)(s;C) - (\mathbb{J}_h^t Q_h^t)(s;C) = \langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^t \otimes \pi_h^t \rangle_{\Omega\times\mathcal{A}}(s;C). \tag{A.2}$$

After introducing these notations, we decompose the instantaneous regret at the t-th episode into two terms,

$$V_1^*(s_1^t;C^t) - V_1^{\pi^t}(s_1^t;C^t) = \underbrace{V_1^*(s_1^t;C^t) - V_1^t(s_1^t;C^t)}_{\mathbf{p}_1} + \underbrace{V_1^t(s_1^t;C^t) - V_1^{\pi^t}(s_1^t;C^t)}_{\mathbf{p}_2}. \tag{A.3}$$

Then we consider these two terms separately. By the definition of value functions in (3.1) and the operator $\mathbb{J}_h^*$ in (A.1), we have $V_h^* = \mathbb{J}_h^* Q_h^*$. By the construction of Algorithm 2, we have $V_h^t = \mathbb{J}_h^t Q_h^t$ similarly. Thus, for the first term $\mathbf{p}_1$ defined in equation (A.3), using ξ_h^t defined in (A.2), for any $h \in [H], t \in [T]$, we have

$$\begin{aligned} V_h^* - V_h^t &= \mathbb{J}_h^* Q_h^* - \mathbb{J}_h^t Q_h^t = (\mathbb{J}_h^* Q_h^* - \mathbb{J}_h^* Q_h^t) + (\mathbb{J}_h^* Q_h^t - \mathbb{J}_h^t Q_h^t) \\ &= \mathbb{J}_h^*(Q_h^* - Q_h^t) + \xi_h^t. \end{aligned}$$

Next, by the definition of the temporal-difference error δ_h^t in (6.1) and the Bellman optimality equation in equation (3.2), we have

$$Q_h^* - Q_h^t = (v_h + P_h V_{h+1}^*) - (v_h + P_h V_{h+1}^t - \delta_h^t) = P_h(V_{h+1}^* - V_{h+1}^t) + \delta_h^t.$$

Hence we get

$$V_h^* - V_h^t = \mathbb{J}_h^* P_h (V_{h+1}^* - V_{h+1}^t) + + \mathbb{J}_h^* \delta_h^t + \xi_h^t.$$

Then, by recursively applying the above formula, we have

$$V_1^* - V_1^t = \Big(\prod_{h\in[H]} \mathbb{J}_h^* P_h\Big)(V_{H+1}^* - V_{H+1}^t) + \sum_{h\in[H]} \Big(\prod_{i\in[h]} \mathbb{J}_i^* P_i\Big)\mathbb{J}_h^* \delta_h^t + \sum_{h\in[H]} \Big(\prod_{i\in[h]} \mathbb{J}_i^* P_i\Big)\xi_h^t.$$

By the definition of ξ_h^t in equation (A.2) and $\zeta_{t,h}^3$ in equation (6.2), we get

$$\sum_{h\in[H]} \Big(\prod_{i\in[h]} \mathbb{J}_i^* P_i\Big)\xi_h^t(s_h;C^t) = \sum_{h\in[H]} \mathbb{E}_{\mu^*,\pi^*}\big\{\big[\langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^t \otimes \pi_h^t \rangle_{\Omega\times\mathcal{A}}(s_h;C^t) | s_1 = s_1^t\big]\big\}.$$

Notice that $V_{H+1}^* = V_{H+1}^t = 0$. Therefore, for any episode $t \in [T]$, we have

$$V_1^*(s_1^t; C^t) - V_1^t(s_1^t; C^t) = \sum_{h\in[H]} \mathbb{E}_{\mu^*,\pi^*}\left\{\left[\langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^t \otimes \pi_h^t\rangle_{\Omega\times\mathcal{A}}(s_h; C^t)|s_1 = s_1^t\right]\right\} + \sum_{h\in[H]} \mathbb{E}_{\mu^*,\pi^*}\left[\delta_h^t(s_h, \omega_h, a_h)|s_1 = s_1^t\right].$$

Now we come to bound the second term $\mathbf{p}_2$ in equation (A.3). By the definition of the temporal-difference error δ_h^t in (6.1), for any $h \in [H], t \in [T]$, we note that

$$\begin{aligned}\delta_h^t(s_h^t, \omega_h^t, a_h^t) &= (v_h^t + P_h V_{h+1}^t - Q_h^t)(s_h^t, \omega_h^t, a_h^t; C^t)\\ &= (v_h^t + P_h V_{h+1}^t - Q_h^{\pi^t})(s_h^t, \omega_h^t, a_h^t; C^t) + (Q_h^{\pi^t} - Q_h^t)(s_h^t, \omega_h^t, a_h^t; C^t)\\ &= (P_h V_{h+1}^t - P_h V_{h+1}^{\pi^t})(s_h^t, \omega_h^t, a_h^t) + (Q_h^{\pi^t} - Q_h^t)(s_h^t, \omega_h^t, a_h^t).\end{aligned}$$

where the last equality follows the Bellman equation (3.1). Furthermore, using $\zeta_{t,h}^1$ and $\zeta_{t,h}^2$ defined in (6.2), we have

$$\begin{aligned}&V_h^t(s_h^t;C^t) - V_h^{\pi^t}(s_h^t;C^t)\\ =&(V_h^t - V_h^{\pi^t})(s_h^t;C^t) - \delta_h^t(s_h^t, \omega_h^t, a_h^t) + (Q_h^{\pi^t} - Q_h^t)(s_h^t, \omega_h^t, a_h^t; C^t)\\ &+ (P_h V_{h+1}^t - P_h V_{h+1}^{\pi^t})(s_h^t, \omega_h^t, a_h^t; C^t)\\ =&(V_h^t - \widetilde{V}_h^t)(s_h^t;C^t) - \delta_h^t(s_h^t, \omega_h^t, a_h^t) + (\widetilde{V}_h^t - V_h^{\pi^t})(s_h^t;C^t) + (Q_h^{\pi^t} - Q_h^t)(s_h^t, \omega_h^t, a_h^t; C^t)\\ &+ \big(P_h(V_{h+1}^t - V_{h+1}^{\pi^t})\big)(s_h^t, \omega_h^t, a_h^t; C^t) - (V_{h+1}^t - V_{h+1}^{\pi^t})(s_{h+1}^t; C^t) + (V_{h+1}^t - V_{h+1}^{\pi^t})(s_{h+1}^t; C^t)\\ =&\big[V_{h+1}^t(s_{h+1}^t; C^t) - V_{h+1}^{\pi^t}(s_{h+1}^t; C^t)\big] + \big[V_h^t(s_h^t; C^t) - \widetilde{V}_h^t(s_h^t; C^t)\big] - \delta_h^t(s_h^t, \omega_h^t, a_h^t) + \zeta_{t,h}^1 + \zeta_{t,h}^2.\end{aligned}$$

Applying the above equation recursively, we get that

$$\begin{aligned}V_1^t(s_1^t; C^t) - V_1^{\pi^t}(s_1^t; C^t) =& V_{H+1}^t(s_H^t; C^t) - V_{H+1}^{\pi^t}(s_H^t; C^t) + \sum_{h\in[H]}\big[V_h^t(s_h^t; C^t) - \widetilde{V}_h^t(s_h^t; C^t)\big]\\ &- \sum_{h\in[H]} \delta_h^t(s_h^t, \omega_h^t, a_h^t) + \sum_{h\in[H]}(\zeta_{t,h}^1 + \zeta_{t,h}^2).\end{aligned}$$

Again by Bellman equation (3.1), we have,

$$V_h^t(s_h^t; C^t) - \widetilde{V}_h^t(s_h^t; C^t) = \big\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t\big\rangle_{\Omega\times\mathcal{A}}(s_h^t; C^t).$$

Then we use $V_{H+1}^t = V_{H+1}^{\pi^t} = 0$ to simplify the decomposition to the following form:

$$V_1^t(s_1^t; C^t) - V_1^{\pi^t}(s_1^t; C^t) = \sum_{h\in[H]} \left\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t \right\rangle_{\Omega\times\mathcal{A}}(s_h^t; C^t) - \sum_{h\in[H]} \delta_h^t(s_h^t, \omega_h^t, a_h^t) + \sum_{h\in[H]} (\zeta_{t,h}^1 + \zeta_{t,h}^2).$$

Therefore, combining $\mathbf{p}_1$ and $\mathbf{p}_2$, we can conclude the proof of this lemma.

$$\begin{aligned}
\mathrm{Reg}(T, \mu^*) &= \sum_{t\in[T]} \left[V_1^*(s_1^t; C^t) - V_1^{\pi^t}(s_1^t; C^t) \right] \\
&= \sum_{t\in[T]} \sum_{h\in[H]} \left\{ \mathbb{E}_{\mu_h^*, \pi_h^*}[\delta_h^t(s_h, \omega_h, a_h) | s_1 = s_1^t] - \delta_h^t(s_h^t, \omega_h^t, a_h^t) \right\} + \sum_{t\in[T]} \sum_{h\in[H]} (\zeta_{t,h}^1, \zeta_{t,h}^2) \\
&\quad + \sum_{t\in[T]} \sum_{h\in[H]} \mathbb{E}_{\mu_h^*, \pi_h^*} \left[\left\langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^t \otimes \pi_h^t \right\rangle_{\Omega\times\mathcal{A}}(s_h; C^t) | s_1 = s_1^t \right] \\
&\quad + \sum_{t\in[T]} \sum_{h\in[H]} \left\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t \right\rangle_{\Omega\times\mathcal{A}}(s_h^t; C^t).
\end{aligned}$$

Therefore, we conclude the proof of the lemma. □

A.5 Proof of Lemma 6.2

Proof. In the following lemma, we firstly bound the difference between the Q-function maintained in Algorithm 2 (without bonus) and the real Q-function of any policy π by their expected difference at next step, plus an error term. This error term can be upper bounded by our bonus with high probability. This lemma can be derived from Lemma B.4 in [25] with slight revisions.

Lemma A.4. *Set $\lambda = \max\{1, \Psi^2\}$. There exists an absolute constant c_ρ such that for $\rho = c_\rho d_\psi H\sqrt{\iota}$ where $\iota = \log(2d_\psi\Psi^2 T/\delta)$, and for any fixed policy π, with probability at least $1-\delta/2$, we have for all $s \in \mathcal{S}$, $\omega \in \Omega$, $a \in \mathcal{A}$, $h \in [H]$, $t \in [T]$,*

$$\psi(s, \omega, a)^\top q_h^t - Q_h^\pi(s, \omega, a) = P_h(V_{h+1}^t - V_{h+1}^\pi)(s, \omega, a) + \triangle_h^t(s, \omega, a),$$

for some $\triangle_h^t(s, \omega, a)$ that satisfies $|\triangle_h^t(s, \omega, a)| \le \rho\|\psi(s, \omega, a)\|_{(\Gamma_h^t)^{-1}}$.

Now we are ready to prove Lemma 6.2. By the definition of δ_h^t in (6.1), we have $\delta_h^t = (v_h + P_h V_{h+1}^t - Q_h^t) = (P_h V_{h+1}^t - P_h V_{h+1}^{\pi^t}) + (Q_h^{\pi^t} - Q_h^t)$. Therefore, by the construction of Q_h^t in Algorithm 2, we obtain

that

$$\begin{aligned}\delta_h^t(s,\omega,a) &\geq (P_h V_{h+1}^t - P_h V_{h+1}^{\pi^t})(s,\omega,a) + Q_h^{\pi^t}(s,\omega,a) - \left(\psi(s,\omega,a)^\top q_h^t + \rho\|\psi(s,\omega,a)\|_{(\Gamma_h^t)^{-1}}\right) \\ &= -\triangle_h^t(s,\omega,a) - \rho\|\psi(s,\omega,a)\|_{(\Gamma_h^t)^{-1}} \geq -2\rho\|\psi(s,\omega,a)\|_{(\Gamma_h^t)^{-1}},\end{aligned}$$

which concludes the proof. □

A.6 Proof of Lemma 6.3

Proof. Denote the optimal signaling schemes corresponding to the real prior μ_h^* and the estimated prior μ_h^t respectively as

$$\pi_h' = \operatorname*{argmax}_{\pi_h \in \mathrm{Pers}(\mu_h^*)} \left\langle Q_h^t, \mu_h^* \otimes \pi_h \right\rangle_{\Omega\times\mathcal{A}}(\cdot; C^t) \quad \text{and} \quad \pi_h'' = \operatorname*{argmax}_{\pi_h \in \mathrm{Pers}(\mu_h^t)} \left\langle Q_h^t, \mu_h^t \otimes \pi_h \right\rangle_{\Omega\times\mathcal{A}}(\cdot; C^t),$$

where the Q-function Q_h^t is given by Algorithm 2. Notably, π_h' is different from the truly optimal policy μ_h^*, since π_h' is computed based on the approximate Q-function Q_h^t. By definition, we can decompose the difference as follows:

$$\begin{aligned}\left\langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^t \otimes \pi_h^t \right\rangle_{\Omega\times\mathcal{A}}(s_h; C^t) =& \left\langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^* \otimes \pi_h' \right\rangle_{\Omega\times\mathcal{A}}(s_h; C^t) && \text{(A.4)} \\ &+ \left\langle Q_h^t, \mu_h^* \otimes \pi_h' - \mu_h^t \otimes \pi_h'' \right\rangle_{\Omega\times\mathcal{A}}(s_h; C^t) && \text{(A.5)} \\ &+ \left\langle Q_h^t, \mu_h^t \otimes \pi_h'' - \mu_h^t \otimes \pi_h^t \right\rangle_{\Omega\times\mathcal{A}}(s_h; C^t). && \text{(A.6)}\end{aligned}$$

By definition, equation (A.4) is always non-positive. Apply Lemma A.2 to equation (A.5) and we can get

$$\begin{aligned}\left\langle Q_h^t, \mu_h^* \otimes \pi_h' - \mu_h^t \otimes \pi_h'' \right\rangle_{\Omega\times\mathcal{A}}(s_h; C^t) \leq& \mathrm{Gap}\left(s_h, \mu_h^*(\cdot|c_h^t), \mathrm{B}_1\left(\mu_h^*(\cdot|c_h^t), \left\|\mu_h^*(\cdot|c_h^t) - \mu_h^t(\cdot|c_h^t)\right\|_1\right); Q_h^t\right) \\ &+ \frac{H}{2}\left\|\mu_h^*(\cdot|c_h^t) - \mu_h^t(\cdot|c_h^t)\right\|_1.\end{aligned}$$

According to Corollary 6.6, we can bound the above equation with the norm of feature vector and the radius of confidence region for θ_h.

$$\left\langle Q_h^t, \mu_h^* \otimes \pi_h' - \mu_h^t \otimes \pi_h'' \right\rangle_{\Omega\times\mathcal{A}}(s_h; C^t) \leq \left(\frac{HL_\mu K}{p_0 D} + \frac{HL_\mu K}{2}\right)\beta\|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}.$$

We also note that equation (A.6) is equal to $\mathrm{Gap}\left(s_h, \mu_h^t(\cdot|c_h^t), \mu_{\mathcal{B}_h^t}(\cdot|c_h^t); Q_h^t\right)$. By Lemma A.1, on the event $\{\theta_h^* \in \mathcal{B}_h^t\}$, we have

$$\mathrm{Gap}\left(s_h, \mu_h^t(\cdot|c_h^t), \mu_{\mathcal{B}_h^t}(\cdot|c_h^t); Q_h^t\right) \leq \frac{2HL_\mu K}{p_0 D}\beta\|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}.$$

Therefore, on the given event, we have

$$\mathbb{E}_{\mu_h^*,\pi_h^*}\left[\left\langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^t \otimes \pi_h^t \right\rangle_{\Omega\times\mathcal{A}}(s_h; C^t) | s_1 = s_1^t\right] \leq \left(\frac{3HL_\mu K}{p_0 D} + \frac{HL_\mu K}{2}\right)\beta \|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}.$$

Summing up together, we get

$$\begin{aligned}\sum_{t\in[T]}\sum_{h\in[H]} \mathbb{E}_{\mu_h^*,\pi_h^*}\left[\left\langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^t \otimes \pi_h^t \right\rangle_{\Omega\times\mathcal{A}}(s_h; C^t) | s_1 = s_1^t\right] \\ \leq \left(\frac{3HL_\mu K}{p_0 D} + \frac{HL_\mu K}{2}\right)\beta \sum_{t\in[T]}\sum_{h\in[H]} \|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}.\end{aligned}$$

Therefore, we conclude the proof of Lemma 6.3. □

A.7 Proof of Lemma 6.4

Proof. By definition, we can rewrite the difference in Lemma 6.4 as

$$\begin{aligned}\left\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t \right\rangle_{\Omega\times\mathcal{A}}(s_h^t; C^t) &= \int_{\Omega\times\mathcal{A}} \left[\mu_h^t(\omega|c_h^t) - \mu_h^*(\omega|c_h^t)\right] \pi_h^t(s,\omega,a) Q_h^t(s,\omega,a) \mathrm{d}a\mathrm{d}\omega \\ &= \int_{\Omega} \left[\mu_h^t(\omega|c_h^t) - \mu_h^*(\omega|c_h^t)\right] \int_{\mathcal{A}} \pi_h^t(s,\omega,a) Q_h^t(s,\omega,a) \mathrm{d}a\mathrm{d}\omega.\end{aligned}$$

By Holder's inequality, we have

$$\left|\left\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t \right\rangle_{\Omega\times\mathcal{A}}(s_h^t; C^t)\right| \leq \left\|\mu_h^t(\cdot|c_h^t) - \mu_h^*(\cdot|c_h^t)\right\|_1 \sup_{\omega\in\Omega}\left|\int_{\mathcal{A}} \pi_h^t(s,\omega,a) Q_h^t(s,\omega,a) \mathrm{d}a\right|.$$

Since $Q_h^t \leq H$ for any $h \in [H]$ and $t \in [T]$, the inequality can be simplified to

$$\left|\left\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t \right\rangle_{\Omega\times\mathcal{A}}(s_h^t; C^t)\right| \leq H\left\|\mu_h^t(\cdot|c_h^t) - \mu_h^*(\cdot|c_h^t)\right\|_1.$$

With the assumption of the prior and link function, on the given event, we obtain that

$$\sum_{t\in[T]}\sum_{h\in[H]} \left\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t \right\rangle_{\Omega\times\mathcal{A}}(s_h^t; C^t) \leq HL_\mu K\beta \sum_{h\in[H]}\sum_{t\in[T]} \|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}.$$

Therefore, we conclude the proof of Lemma 6.4. □

A.8 Proof of Corollary 6.6

Proof. According to Assumption 4.2 for the prior, we can show that for any $\mu_{\theta'}(\cdot|c) \in \mathcal{B}$,

$$\|\mu_\theta(\cdot|c) - \mu_{\theta'}(\cdot|c)\|_1 \leq L_\mu \big\| f(\phi(c)^\top \theta) - f(\phi(c)^\top \theta') \big\|.$$

Moreover, by Assumption 4.3 for the link function $f(\cdot)$, we have

$$\|\mu_\theta(\cdot|c) - \mu_{\theta'}(\cdot|c)\|_1 \leq L_\mu K \big\| \phi(c)^\top (\theta - \theta') \big\| \leq L_\mu K \, \|\phi(c)\|_{\Sigma^{-1}} \, \epsilon.$$

Therefore, $\mathcal{B} \subseteq \mathrm{B}(\mu_\theta(\cdot|c), L_\mu K \|\phi(c)\|_{\Sigma^{-1}} \epsilon)$, and by Lemma 6.5, we can conclude the result. □

A.9 Auxiliary Lemmas

This section presents several auxiliary lemmas and their proofs.

Lemma A.5 (Martingale Bound; [9])**.** *For $\zeta^1_{t,h}$ and $\zeta^2_{t,h}$ defined in (6.2) and for any fixed $\delta \in (0,1)$, with probability at least $1 - \delta/2$, we have*

$$\sum_{t\in[T]} \sum_{h\in[H]} (\zeta^1_{t,h} + \zeta^2_{t,h}) \leq \sqrt{16 T H^3 \log(4/\delta)}.$$

Proof. See [9] for a detailed proof. □

Lemma A.6. *Suppose that $\phi_1, \phi_2, \ldots, \phi_T \in R^{d_\phi \times d}$ and for any $1 \leq i \leq T$, there exists a constant $\Phi > 0$ such that $\|\phi_i\| \leq \Phi$. Let $\Sigma_t = \lambda I_{d_\phi} + \sum_{i\in[t-1]} \phi_i \phi_i'$ for some $\lambda \geq \Phi^2$. Then,*

$$\sum_{t\in[T]} \|\phi_t\|_{(\Sigma_t)^{-1}} \leq \sqrt{2 d_\phi T \log(1 + T\Phi^2/(\lambda d_\phi))}.$$

Proof. Firstly, we apply Cauchy-Schwartz inequality,

$$\sum_{t\in[T]} \|\phi_t\|_{(\Sigma_t)^{-1}} \leq \sqrt{T \sum_{t\in[T]} \|\phi_t\|^2_{(\Sigma_t)^{-1}}}.$$

Since $\|\phi_t\|_{(\Sigma_t)^{-1}} = \sqrt{\phi_t^\top (\Sigma_t)^{-1} \phi_t} \leq \sqrt{\lambda^{-1} \phi_t^\top \phi_t} \leq \Phi/\sqrt{\lambda} \leq 1$, we can use Lemma A.7 to bound the sum of squares:

$$\begin{aligned}\sum_{t\in[T]} \|\phi_t\|_{(\Sigma_t)^{-1}} &\leq \sqrt{2T \log(\det(\Sigma_T) \det(\Sigma_1)^{-1})} \\ &\leq \sqrt{2 d_\phi T \log(1 + T\Phi^2/(\lambda d_\phi))}.\end{aligned}$$

The last inequality is derived from Lemma A.8. □

Lemma A.7 (Sum of Potential Function; [1]). *For any sequence of* $\{\phi_t\}_{t\in[T]}$, *let* $\Sigma_t = \lambda I_h + \sum_{t\in[t-1]} \phi_i \phi_i'$ *for some* $\lambda \geq 0$. *Then we have*

$$\sum_{t\in[T]} \min\{\|\phi_t\|^2_{(\Sigma_t)^{-1}}, 1\} \leq 2\log(\det(\Sigma_T)\det(\Sigma_1)^{-1}).$$

Proof. See [1] for a detailed proof. □

Lemma A.8 (Determinant-Trace Inequality). *Suppose that* $\phi_1, \phi_2, \ldots, \phi_T \in R^{d_\phi \times d}$ *and for any* $1 \leq i \leq T$, *there exists a constant* $\Phi > 0$ *such that* $\|\phi_i\| \leq \Phi$. *Let* $\Sigma_t = \lambda I_{d_\phi} + \sum_{i\in[t-1]} \phi_i \phi_i'$ *for some* $\lambda \geq 0$. *Then,*

$$\det(\Sigma_t) \leq \left(\lambda + t\Phi^2/d_\phi\right)^{d_\phi}.$$

Proof. Let $\lambda_1, \lambda_2, \ldots, \lambda_h$ be the eigenvalues of Σ_t. Since Σ_t is positive definite, its eigenvalues are positive. Also, note that $\det(\Sigma_t) = \prod_{s=1}^{d_\phi} \lambda_s$ and $\mathrm{tr}(\Sigma_t) = \sum_{s=1}^{h} \lambda_s$. By inequality of arithmetic and geometric means

$$\det(\Sigma_t) \leq (\mathrm{tr}(\Sigma_t)/d_\phi)^{d_\phi}.$$

It remains to upper bound the trace:

$$\mathrm{tr}(\Sigma_t) = \mathrm{tr}(\lambda I_{d_\phi}) + \sum_{i=1}^{t-1} \mathrm{tr}(\phi_i \phi_i') = d_\phi \lambda + \sum_{i=1}^{t-1} \|\phi_i\|^2 \leq d_\phi \lambda + t\Phi^2$$

and the lemma follows. □