Incentivized Learning in Principal-Agent Bandit Games

Antoine Scheid [1] **Daniil Tiapkin** [1 2] **Etienne Boursier** [3] **Aymeric Capitaine** [1] **Eric Moulines** [1]
Michael I. Jordan [4 5] **El-Mahdi El-Mhamdi** [1] **Alain Durmus** [1]

Abstract

This work considers a repeated principal-agent bandit game, where the principal can only interact with her environment through the agent. The principal and the agent have misaligned objectives and the choice of action is only left to the agent. However, the principal can influence the agent's decisions by offering incentives which add up to his rewards. The principal aims to iteratively learn an incentive policy to maximize her own total utility. This framework extends usual bandit problems and is motivated by several practical applications, such as healthcare or ecological taxation, where traditionally used mechanism design theories often overlook the learning aspect of the problem. We present nearly optimal (with respect to a horizon T) learning algorithms for the principal's regret in both multi-armed and linear contextual settings. Finally, we support our theoretical guarantees through numerical experiments.

1 Introduction

Decision-making under uncertainty is a ubiquitous feature of real-world applications of machine learning, arising in domains as diverse as recommendation systems (Li et al., 2010), healthcare (Yu et al., 2021), and agriculture (Evans et al., 2017). Multi-armed bandits provide a classical point of departure for decision-making under uncertainty in these settings (Thompson, 1933; Woodroofe, 1979; Lattimore & Szepesvári, 2020; Slivkins et al., 2019). The basic bandit solution involves an agent who learns which decisions yield high rewards via repeated experimentation. Real-world decision-making problems, however, often present challenges that are not addressed in this simple optimization framework. These include the challenge of scarcity when there are multiple decision-makers, issues of misaligned objectives, and problems arising from information asymmetries and signaling. The economics literature addresses these issues through the design of game-theoretic mechanisms, including auctions and contracts (see, e.g., Myerson, 1989; Laffont & Martimort, 2009), aiming to achieve favorable outcomes despite agents' self-interest and limited information set. Unfortunately, the economics literature tends to neglect the learning aspect of the problem, often assuming that preferences, or distributions on preferences, are known a priori. Our work focuses on the blend of mechanism design and learning. We study a principal-agent model with information asymmetry and we develop a learning framework in which the principal aims to uncover the true preferences of the agent while optimizing her own gains.

Building on the work of Dogan et al. (2023a;b), we consider a repeated game between a principal and an agent, where, at each round, the principal proposes an incentive transfer associated with any action. The agent greedily chooses the action that maximizes the sum of his expected reward and the incentive. The goal of the principal is to learn an incentive policy which maximizes her own utility over time, taking into account both the rewards that she reaps and the costly incentives that she offers.

Our contributions are as follows:

- We present the `Incentivized Principal-Agent Algorithm` (`IPA`) framework, which comprises two steps. First, `IPA` estimates the minimal level of incentive needed to make the agent select any desired action. Subsequently, forming an upper estimate of these incentives, `IPA` uses a regret-minimization algorithm in a black-box fashion. The overall algorithm achieves both nearly optimal distribution-free and instance-dependent regret bounds.
- We extend `IPA` to the linear contextual bandit setting (see, e.g., Abe & Long, 1999; Auer, 2002; Dani et al., 2008), significantly broadening its applicability in various applications. Here `Contextual IPA` achieves

[1]Centre de Mathématiques Appliquées – CNRS – École polytechnique – Institut Polytechnique de Paris – route de Saclay 91128 Palaiseau cedex [2]Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France [3]INRIA, Universite Paris Saclay, LMO, Orsay, France [4]University of California, Berkeley [5]Inria, Ecole Normale Supérieure, PSL Research University. Correspondence to: Antoine Scheid <antoine.scheid@polytechnique.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

a $\mathcal{O}(d\sqrt{T}\log(T))$ regret bound. We emphasize that `Contextual IPA` is the first known algorithm for incentivized learning in a contextual setting. Moreover it matches, up to logarithmic factors, the minimax lower bound $\Omega(d\sqrt{T})$ for the easier problem of stochastic linear bandits (Rusmevichientong & Tsitsiklis, 2010).

2 Related Work

While classical work on bandit problems and reinforcement learning has predominantly focused on single-agent scenarios, many emerging applications require considering multiple agents. Recent literature has accordingly begun to study frameworks for learning in multi-agent multi-armed bandit settings (see, e.g., Boursier & Perchet, 2022).

Mansour et al. (2020) discuss how a social planner can simultaneously learn and influence self-interested agents' decisions through Bayesian-Incentive Compatible (BIC) recommendations. The rationale behind this notion is that a BIC recommendation guarantees to each agent a maximal reward given the past, at any step. The social planner objective is to design BIC recommendations that maximize the global welfare. Mansour et al. (2020) propose an algorithm for solving this problem in both multi-armed and contextual bandit settings. Notably, their work turns any black-box bandit algorithm into a BIC algorithm. For this problem, Sellke & Slivkins (2021) show that Thompson sampling can be made BIC, with a sufficient number of initial observations. Hu et al. (2022) extend this work to the combinatorial bandits problem.

Another line of work due to Banihashem et al. (2023) and Simchowitz & Slivkins (2023) studies how a principal can provide recommendations to agents so that they explore all reachable states in a Markov Decision Process (MDP). To this end, the principal supplies the agents with a modified history, with the modifications carefully chosen to retain the agents' trust. This line of work is closely related to the online Bayesian persuasion literature (see, e.g., Castiglioni et al., 2020; 2021), which dates to the seminal work of Kamenica & Gentzkow (2011). Online Bayesian persuasion consists of the principal sequentially influencing agents' decision with signals in her own interest and has been extended to the online Information Design setting (Doval & Ely, 2020; Bernasconi et al., 2022).

In these works, the information asymmetry favors the principal, so that the principal can influence the agent's decision at little or no cost. In our problem, the agent instead has perfect knowledge of the problem parameters and his action can only be influenced through utility transfers.

Ben-Porat et al. (2024) study a principal and an agent sharing a common Markov Decision Process (MDP) with different reward functions. Similarly to our setting, at each step the action is chosen by an agent with full knowledge of the game. The objective of the principal is to minimize her cumulative regret under a constraint on the incentive budget. Despite extending our setup to MDPs, Ben-Porat et al. (2024) do not consider uncertainty in the principal's side, turning the game into an optimization problem. Other leader/follower dynamics with Reinforcement Learning have for instance been studied in the works of Chen et al. (2023); Zhong et al. (2021) to find a Stackelberg equilibrium or an optimal policy for the leader.

Related issues arise in the study of dynamic pricing (Den Boer, 2015; Javanmard & Nazerzadeh, 2019; Mao et al., 2018; Golrezaei et al., 2023). Our work diverges from dynamic pricing in that in our case the principal not only faces uncertainty with respect to the agent's utility, but also with respect to her own utility.

Some works study similar principal-agent games but with a specific focus on the achievable optimality of the contract Cohen et al. (2022) or a specific stochastic model for the agent's behavior Conitzer & Garera (2006). The work of Zhu et al. (2022) interestingly studies online contract design but with a specific focus on the Stackelberg regret of the task, which provides results concerning the sample complexity.

Finally, our study is inspired by the work of Dogan et al. (2023b) to explore the principal's learning mechanism within a principal-agent setting. They propose an ε-Greedy algorithm with suboptimal regret guarantees. In particular, it suffers an exponential dependence in the number of actions. Dogan et al. (2023a) extend the work of (Dogan et al., 2023b), taking into account the presence of uncertainty on the agent's side but with the same limitation. We consider the same setup and problem as introduced in (Dogan et al., 2023b;a). However, it corresponds to the only common ground between our work and theirs. First, their algorithm as well as their regret bound depends on a hyperparameter m that controls the level of exploration, as mentioned in Appendix B. If m is too small, there is not enough exploration and the regret scales poorly with respect to T whereas if m is too large, the convergence takes longer to occur. m needs to be carefully chosen and depends on the specific bandit instance considered, which makes things more complicated. Furthermore, their algorithm is harder to analyze and to implement in practical settings: a lot of difficulties come from a simultaneous exploration/exploitation of both the principal's rewards and the agent's ones. In contrast, the algorithmic solution that we provide for the multi-armed case is completely orthogonal, relying on a two phases approach and avoids fine-tuning of any kind. We provide both distribution free and instance-dependent regret bounds that nearly match the known lower bounds. Also note that we extend our approach to the non-trivial contextual case.

3 Multi-Armed Principal-Agent Learning

Setup. We consider a repeated principal-agent game. A contextual version of the game is introduced in Section 4. The action set for the agent (or set of arms) is fixed to be $\mathcal{A} \coloneqq [K] = \{1, \dots, K\}, K \in \mathbb{N}^\star$. We assume that the agent's rewards, $\mathsf{s} = (\mathsf{s}_1, \dots, \mathsf{s}_K) \in \mathbb{R}_+^K$, are deterministic and that they are known to the agent and unknown to the principal.

For each action $a \in [K]$, the rewards of the principal are given by a random, i.i.d. sequence $(X_a(t))_{t\in[T]}$, where $X_a(t) \sim \nu_a$ and ν_a is the arm distribution. The distributions $\{\nu_a\}_{a=1}^K$ are unknown to the principal and are learned as a consequence of the following principal-agent interaction.

At each round $t \in [T]$, where T is the game horizon, the principal proposes an incentive $\pi(t) \in \mathbb{R}_+$ associated with an action $a_t \in [K]$. The agent then greedily chooses action A_t maximizing his utility:

$$A_t \in \operatorname{argmax}_{a\in[K]}\{\mathsf{s}_a + \mathbb{1}_{a_t}(a)\pi(t)\}\,, \tag{1}$$

breaking ties arbitrarily.[1] The principal then observes the arm A_t selected by the agent, as well as her reward given by $X_{A_t}(t)$. The utility of the principal on the round is given by $X_{A_t}(t) - \mathbb{1}_{a_t}(A_t)\pi(t)$. For any $a \in [K]$, the principal's mean reward is $\theta_a \coloneqq \mathbb{E}[X_a(t)]$. See Table 1 in Appendix A for a summary of the main definitions used in this section.

The sequence of incentives $(a_t, \pi(t))_{t\in[T]}$ defines a sequence of actions $(A_t)_{t\in[T]}$ chosen by the agent. The goal of the principal is to maximize her total utility. On a single round, she thus aims at proposing an optimal incentive π^{opt} on an arm $a^{\mathrm{opt}} \in [K]$, which solves

$$\begin{aligned} &\text{maximize } \textstyle\int x\nu_a(\mathrm{d}x) - \pi \text{ over } \pi \in \mathbb{R}_+, a \in [K] \\ &\text{such that } a \in \operatorname{argmax}_{a'\in[K]}\{\mathsf{s}_{a'} + \mathbb{1}_a(a')\pi\}\,. \end{aligned} \tag{2}$$

This is consistent with the conventional framework for utility in bandit problems, where we subtract the cost of incentives to the principal. Here, the principal's influence is exerted solely through the strategic use of incentives, carefully designed to guide the agent's behavior. We define $\mu^\star \coloneqq \int x\nu_{a^{\mathrm{opt}}}(\mathrm{d}x) - \pi^{\mathrm{opt}}$. Maximizing the total utility of the principal over T rounds is equivalent to minimizing the expected regret, defined as

$$\mathfrak{R}(T) \coloneqq T\,\mu^\star - \sum_{t=1}^{T} \mathbb{E}[X_{A_t}(t) - \mathbb{1}_{a_t}(A_t)\pi(t)]\,. \tag{3}$$

Remark. In the prior work of Dogan et al. (2023b), incentives were defined as a vector of size K, where the incentivize associated with an action $a \in [K]$ was denoted π_a. In our setting, since the goal of the principal is to make sure that the agent picks one prescribed action, it is enough to consider a restricted family of the form $\pi_a = \mathbb{1}_{a_t}(a)\pi(t)$, where $(a_t, \pi(t))$ are incentives in the sense defined above.

We make the following assumption.

H1. *For any $a \in [K]$, $\mathsf{s}_a \in [0, 1]$.*

Neither the distributions ν_a nor the preferences of the agent are known to the principal. Another difficulty arises from designing the magnitude of the incentive $\pi(t)$: if it is too small, the agent might not choose the arm a_t proposed by the principal whereas using an overly large amount leads the principal to overpay, decreasing her utility.

This trade-off also arises in dynamic pricing, where sellers must strike a balance between attractive pricing and profitability. For discussion of the results in that literature, see the comprehensive overview by Den Boer (2015). In addition, there are links between dynamic pricing and bandit problems (see, e.g., Javanmard & Nazerzadeh, 2019; Cai et al., 2023).

Optimal incentives. Before introducing IPA, we highlight a pivotal observation. For any given round $t \geqslant 1$, action $a \in [K]$ and $\varepsilon > 0$, the principal can entice the agent to choose a by offering an incentive, $\pi_a^{\star,\varepsilon} \in \mathbb{R}_+$, defined as:

$$\pi_a^{\star,\varepsilon} = \max_{a'\in[K]} \mathsf{s}_{a'} - \mathsf{s}_a + \varepsilon\,. \tag{4}$$

With this incentive, it holds that for any $a' \in [K], a' \neq a$:

$$\mathsf{s}_{a'} < \mathsf{s}_a + \pi_a^{\star,\varepsilon}\,,$$

which ensures that the agent chooses $A_t = a$, given that action a yields a superior reward. Consequently, $\pi_a^\star \coloneqq \lim_{\varepsilon\to 0} \pi_a^{\star,\varepsilon} = \max_{a'\in[K]} \mathsf{s}_{a'} - \mathsf{s}_a$ represents the infimal incentive necessary to make arm a the agent's selection. Assuming s is known to the principal, then using $\pi_a^{\star,\varepsilon}$ for any $\varepsilon > 0$ across all arms $a \in [K]$, will provide an expected reward of $\theta_a - \pi_a^\star$ per arm, which can be found using a standard bandit algorithm. Lemma 1 allows us to define the regret in a more convenient way.

Lemma 1. *For any $T \in \mathbb{N}$, the regret of any algorithm on our problem instance can be written as*

$$\begin{aligned} \mathfrak{R}(T) = T \max_{a\in[K]}\{\theta_a + \mathsf{s}_a - \max_{a'\in[K]} \mathsf{s}_{a'}\} \\ - \mathbb{E}\left[\sum_{t=1}^{T}\{\theta_{A_t} - \mathbb{1}_{a_t}(A_t)\pi(t)\}\right]\,. \end{aligned}$$

Warm up: fixed horizon solution and regret analysis. IPA separates the problem of learning optimal incentives $\pi_a^\star$ for each action a—a problem that can be solved efficiently

[1] Note that the related works (Ben-Porat et al., 2024; Simchowitz & Slivkins, 2023) assume a tie-breaking in favor of the principal, an assumption that we do not need here.

via binary search (see Algorithm 3)—from estimation of the principal's expected reward $(\theta_a)_{a\in[K]}$, which is achieved using a standard multi-armed bandit algorithm. With a known horizon T, the algorithm unfolds in two stages. First, for each action $a \in [K]$, the principal devotes $N_T := \lceil \log_2 T \rceil$ rounds of binary search per arm to estimate $\pi_a^\star$, maintaining lower $\underline{\pi}_a(N_T)$ and upper $\overline{\pi}_a(N_T)$ bounds with $\underline{\pi}_a(N_T) \leqslant \pi_a^\star \leqslant \overline{\pi}_a(N_T)$. Denoting $\overline{\pi}_a := \overline{\pi}_a(\lceil \log_2 T \rceil)$ for simplicity, we compute the estimate

$$\hat{\pi}_a := \overline{\pi}_a(\lceil \log_2 T \rceil) + 1/T \;, \tag{5}$$

where $1/T$ is added to avoid any tie-breaking situation. We show formally in Lemma 8 that

$$\hat{\pi}_a - 2/T \leqslant \pi_a^\star < \hat{\pi}_a \;. \tag{6}$$

In the second phase, IPA then employs an arbitrary multi-armed bandit subroutine Alg in a black-box manner to learn θ.

Bandit instance. For any distributions $(\tilde{\nu}_a)_{a\in[K]}$ and sequence of i.i.d. random variables $(Y_a(t))_{t\in\mathbb{N}^\star}$, $a \in [K]$, with $Y_a(t) \sim \tilde{\nu}_a$, we define the history $\mathcal{G}_t := (A_s, U_s, Y_{A_s}(s))_{s\leqslant t}$ where for any $s \in \mathbb{N}^\star$, $(U_s)_{s\in\mathbb{N}^\star}$ is a family of independent uniform random variables on $[0,1]$ allowing for randomization in the subroutine. Let Alg be a bandit algorithm, i.e., $\texttt{Alg}\colon (U_t, \mathcal{G}_{t-1}) \mapsto a_t^{\mathrm{Rec}}$. We define the expected regret of Alg as

$$R_{\texttt{Alg}}(T,\tilde{\nu}) := T \max_{a\in[K]} \mathbb{E}_{Y\sim\tilde{\nu}}[Y_a(1)] - \mathbb{E}_{Y\sim\tilde{\nu}}\left[\sum_{t=1}^{T} Y_{\texttt{Alg}(U_t,\mathcal{G}_{t-1})}(t)\right] .$$

After the binary search phase, for $t > K\lceil \log_2 T \rceil$, the principal plays Alg on her bandit instance driven by her own mean rewards $(\theta_a)_{a\in[K]}$ and the approximated incentives $(\hat{\pi}_a)_{a\in[K]}$. Alg will be fed with a shifted history, defined for any $t > K\lceil \log_2 T \rceil$ as

$$\tilde{\mathcal{H}}_t := (a_s^{\mathrm{Rec}}, U_s, X_{A_s}(s) - \hat{\pi}_{a_s^{\mathrm{Rec}}})_{s\in[K\lceil \log_2 T\rceil+1,t]} \;, \tag{7}$$

with a_t^{Rec} the action recommended by Alg at time t and A_t the action pulled by the agent. At time t, IPA offers the incentive $\hat{\pi}_{a_t^{\mathrm{Rec}}}$ to the agent if he chooses action a_t^{Rec}. Equation (6) ensures that this incentive makes a_t^{Rec} strictly preferable to any other action for the agent and so a_t^{Rec} is eventually played. As can be seen in (7), the shift of each arm's mean by $\hat{\pi}_a$ is taken into account while Alg is learning. We also define the shifted distribution ρ_a^T for any $a \in [K]$ as the distribution of $X_a(1) - \hat{\pi}_a$.

Theorem 1. IPA *run with any multi-armed bandit subroutine* Alg *has an overall regret* $\mathfrak{R}(T)$ *such that*

$$\mathfrak{R}(T) \leqslant 2 + (1 + \max_{a\in[K]}\{\theta_a\} - \min_{a\in[K]}\{\theta_a\})(1 + K\log_2 T) + R_{\texttt{Alg}}(T - K\lceil \log_2 T\rceil, \rho^T) \;,$$

where $R_{\texttt{Alg}}$ *stands for the regret induced by* Alg *on the shifted vanilla multi-armed bandit problem* ρ^T.

Algorithm 1 IPA

Input: Set of actions $\mathcal{A} := [K]$, time horizon T, subroutine Alg
Compute $\tilde{\mathcal{H}}_s := \varnothing$ for any $s \leqslant K\lceil \log_2 T\rceil$
for $a \in [K]$ **do**
  # See Algorithm 3
  $\underline{\pi}_a, \overline{\pi}_a = \texttt{Binary Search}(a, \lceil \log_2 T\rceil, 0, 1)$
end for
For any action $a \in [K]$, $\hat{\pi}_a = \overline{\pi}_a + 1/T$
for $t = K\lceil \log_2 T\rceil + 1, \dots, T$ **do**
  Sample $U_t \sim \mathrm{U}(0,1)$ and get a recommended action by Alg, $a_t^{\mathrm{Rec}} = \texttt{Alg}(U_t, \tilde{\mathcal{H}}_{t-1})$
  Offer an incentive $\hat{\pi}_{a_t^{\mathrm{Rec}}}$ on action a_t^{Rec} and nothing for any other action $a' \in [K], a' \neq a$
  Observe $A_t, X_{A_t}(t)$ and compute history $\tilde{\mathcal{H}}_t = (a_s^{\mathrm{Rec}}, U_s, X_{A_s}(s) - \hat{\pi}_{a_s^{\mathrm{Rec}}})_{s\in[K\lceil \log_2 T\rceil+1,t]}$
end for

The proof is postponed to Appendix C.

Corollary 1. *Assume the principal's reward distribution* ν_a *for any action* $a \in [K]$ *is 1-subgaussian. Then,* IPA *run with the bandit subroutine* Alg = *UCB has a regret bounded for any* $T \in \mathbb{N}$ *as follows*

$$\mathfrak{R}(T) \leqslant 3 + 3\sum_{a\in[K],\Delta_a^\star>0} \Delta_a^\star + (1 + \max_{a\in[K]}\{\theta_a\} - \min_{a\in[K]}\{\theta_a\})(1 + 9K\log_2 T) + 8\min\left\{\sqrt{TK\log T}\,; \sum_{a\in[K],\Delta_a^\star>0} \frac{4\log T}{\Delta_a^\star}\right\},$$

where $\Delta_a^\star := \max_{a'\in[K]}\{\theta_{a'} + \mathtt{s}_{a'}\} - (\theta_a + \mathtt{s}_a)$ *are the reward gaps.*

Note that any black-box algorithm, not necessarily UCB, can be employed, yielding other concrete bounds in the corollary. We recover the usual multi-armed UCB bounds (both distribution-free and instance-dependent): this is why IPA achieves the bound provided in Corollary 1. For completeness, the UCB subroutine is given in Appendix E.

4 Contextual Principal-Agent Learning

In this section, we study the same interaction between a principal and an agent, but in a contextual setting (see, e.g., Abe & Long, 1999; Auer, 2002; Dani et al., 2008). We use the simplified model of stochastic linear bandits for both the

agent and the principal. Consider a set of possible actions in $\mathrm{B}(0,1)$, where $\mathrm{B}(0,1)$ stands for the unit closed ball in $\mathbb{R}^d$, and a family of zero-mean distributions indexed by $\mathrm{B}(0,1)$, $(\tilde{\nu}_a)_{a\in\mathrm{B}(0,1)}$ such that for any $a \in \mathrm{B}(0,1), t \in [T], \eta_a(t) \sim \tilde{\nu}_a$. The principal's reward is given by the sequence $\{(X_a(t))_{a\in\mathrm{B}(0,1)} \colon t \in [T]\}$ of independent random variables such that for any $t \in [T], a \in \mathrm{B}(0,1)$,

$$X_a(t) \coloneqq \langle \theta^\star, a\rangle + \eta_a(t) \;,$$

where $\theta^\star$ is unknown to the principal. The agent's reward function is defined as $a \mapsto \langle \mathsf{s}^\star, a\rangle$, where $\mathsf{s}^\star \in \mathbb{R}^d$ is known to the agent and unknown to the principal. With this notation, the agent and the principal observe an action set $\mathcal{A}_t \subseteq \mathrm{B}(0,1)$, at each round $t \geqslant 1$. Note that this set is no longer stationary. The precise timeline is as follows. At each round, the principal proposes an incentive function, $\kappa(t,\cdot)\colon \mathcal{A}_t \to \mathbb{R}_+$, associating any action $a \in \mathcal{A}_t \subseteq \mathrm{B}(0,1)$ with a transfer of incentives $\kappa(t,a)$ from the principal to the agent. The principal chooses $\kappa(t,\cdot)$ as a function with a finite support, which makes it upper semi-continuous. The agent then greedily chooses the action A_t as follows

$$A_t \in \mathrm{argmax}_{a\in\mathcal{A}_t}\{\langle \mathsf{s}^\star, a\rangle + \kappa(t,a)\} \;, \tag{8}$$

which is well-defined since $\kappa(t,\cdot)$ is upper semi-continuous and $\mathcal{A}_t$ satisfies the following assumption.

H 2. *For any $t \geqslant 1$, $\mathcal{A}_t$ is closed, therefore compact. Moreover, $\mathsf{s}^\star \in \mathrm{B}(0,1)$ and $\theta^\star \in \mathrm{B}(0,1)$.*

The principal then observes the arm A_t selected by the agent, as well as her incurred reward given by $X_{A_t}(t)$. The utility of the principal on the round is given by $X_{A_t}(t) - \kappa(t,A_t)$. This defines, for a sequence of principal's incentive functions $\{\kappa(t,\cdot), t \in [T]\}$, the sequence of actions $\{A_t \colon t \in [T]\}$ chosen by the agent. The goal of the principal is to maximize her total utility. On a single round t, she thus aims at proposing an optimal incentive function $\kappa(t,\cdot)$ which solves

$$\begin{aligned} &\text{maximize } \langle \theta^\star, a\rangle - \kappa(t,a) \text{ over } \kappa(t,\cdot)\colon \mathcal{A}_t \to \mathbb{R}_+, \\ &\text{such that } a \in \mathrm{argmax}_{a'\in\mathcal{A}_t}\{\langle \mathsf{s}^\star, a'\rangle + \kappa(t,a')\} \;. \end{aligned} \tag{9}$$

In addition, we define the optimal average reward at t as

$$\begin{aligned} \mu_t^\star \coloneqq \sup_{\kappa(t,\cdot)\colon \mathcal{A}_t\to\mathbb{R}_+} &\{\langle \theta^\star, a\rangle - \kappa(t,a)\} \\ \text{such that } a \in \mathrm{argmax}_{a'\in\mathcal{A}_t}&\{\langle \mathsf{s}^\star, a'\rangle + \kappa(t,a')\} \;. \end{aligned} \tag{10}$$

Maximizing the total utility of the principal over T rounds is equivalent to minimizing the expected regret, defined as

$$\mathfrak{R}(T) = \sum_{t=1}^T \mu_t^\star - \mathbb{E}\left[\sum_{t=1}^T (X_{A_t}(t) - \kappa(t,A_t))\right] \;. \tag{11}$$

Similarly to Lemma 1, the following result provides an alternative definition for the regret.

Lemma 2. *For any $T \in \mathbb{N}$, the regret of any algorithm on our contextual problem instance can be written as*

$$\begin{aligned} \mathfrak{R}(T) = \sum_{t=1}^T \max_{a\in\mathcal{A}_t}\{\langle \theta^\star + \mathsf{s}^\star, a\rangle - \max_{a'\in\mathcal{A}_t}\langle \mathsf{s}^\star, a'\rangle\} \\ - \mathbb{E}\left[\sum_{t=1}^T (\langle \theta^\star, A_t\rangle - \kappa(t,A_t))\right] \;. \end{aligned}$$

The proof is deferred to Appendix D.

Design of the optimal incentives. At any round $t \geqslant 1$, for the agent to necessarily choose action $a \in \mathcal{A}_t$, the principal can provide the agent with the incentive function $\kappa_a^{\star,\varepsilon}(t,a') \coloneqq \mathbb{1}_a(a')\pi^{\star,\varepsilon}(t,a)$, where for any $\varepsilon > 0$,

$$\pi^{\star,\varepsilon}(t,a) \coloneqq \max_{a'_t\in\mathcal{A}_t}\{\langle \mathsf{s}^\star, a'_t - a\rangle + \varepsilon\} \;.$$

Lemma 10 in Appendix D guarantees that this choice of $\kappa_a^{\star,\varepsilon}$ gives $A_t = a$, where A_t is defined in (8). Define

$$\begin{aligned} a_t^{\mathrm{ag}} &\coloneqq \mathrm{argmax}_{a'\in\mathcal{A}_t}\langle \mathsf{s}^\star, a'\rangle \;, \\ \pi^\star(t,a) &\coloneqq \langle \mathsf{s}^\star, a_t^{\mathrm{ag}}\rangle - \langle \mathsf{s}^\star, a\rangle \end{aligned} \tag{12}$$

and $\kappa_a^\star(t,a') \coloneqq \mathbb{1}_a(a')\pi^\star(t,a)$. As in the non-contextual case, taking $\varepsilon \to 0$ makes the incentive function $\kappa_a^\star$ the infimal function that makes the choice of a strictly preferable to any other arm $a' \in \mathcal{A}_t$ at time t.

Similarly to the multi-armed setting, we decompose the problem into two distinct components. First, we aim to estimate the agent's reward $a \mapsto \langle \mathsf{s}^\star, a\rangle$ based on the observation of agent's selected actions given an appropriate choice of incentives. As discussed below, this can be achieved with a binary-search-like procedure. Second, once this function is accurately estimated, the principal can use a contextual bandit algorithm `CtxAlg` in a black-box manner to minimize her own regret with the estimated incentive function to determine the agent's behavior.

Estimation of the agent's reward. The approach that we propose is based on a sequence of confidence sets $\{\mathcal{S}_t\}_{t\in[T]}$ that satisfy $\mathsf{s}^\star \in \mathcal{S}_t$ for any $t \in [T]$. We construct the sequence $(\mathcal{S}_t)_{t\in[T]}$ recursively such that their diameters decrease along the iterations. This is motivated by Lemma 3 which allows us to control the estimation error of $\pi^\star$ and relates it to the diameter of these sets. The proof is postponed to Appendix D.

Lemma 3. *For any $t \in [T]$ and closed subset $\mathcal{S} \subset \mathrm{B}(0,1)$ with $\mathsf{s}^\star \in \mathcal{S}$, it holds, for any $a \in \mathcal{A}_t$, $|\max_{\mathsf{s}\in\mathcal{S},a'\in\mathcal{A}_t}\langle \mathsf{s}, a' - a\rangle - \pi^\star(t,a)| \leqslant 2\,\mathrm{diam}(\mathcal{S},\mathcal{A}_t)$ where $\mathrm{diam}(\mathcal{S},\mathcal{A}_t) \coloneqq \max_{a'\in\mathcal{A}_t}\max_{\mathsf{s}_1,\mathsf{s}_2\in\mathcal{S}}|\langle \mathsf{s}_1 - \mathsf{s}_2, a'\rangle|$.*

In the light of Lemma 3, we thus aim to build confidence sets $\mathcal{S}_t$ with decreasing diameters such that $\mathsf{s}^\star \in \mathcal{S}_t$ for any t. To

this end, the principal can offer an incentive function $\kappa(t,\cdot)$ concentrated on a *single point* $a \in \mathcal{A}_t$ as in the multi-armed case: $\kappa(t,a') = \pi(t)\cdot \mathbb{1}_a(a')$ for $\pi(t)\in\mathbb{R}_+$. In this case, the principal receives the agent's choice as a feedback; by (8), either $A_t = a_t$ or $A_t = \operatorname{argmax}_{a'\in\mathcal{A}_t}\langle \mathsf{s}^\star, a'\rangle = a_t^{\mathrm{ag}}$. In addition, $A_t = a_t$ is equivalent to the fact that

$$\langle \mathsf{s}^\star, a_t^{\mathrm{ag}} - a_t\rangle \leqslant \pi(t)\ .$$

The information $A_t = a_t$ or $A_t = a_t^{\mathrm{ag}}$ can be used as *binary search feedback* in the direction $a_t^{\mathrm{ag}} - a_t$, as follows. Given a current confidence set $\mathcal{S}_t$ at time t it can be updated either as $\mathcal{S}_{t+1} = \mathcal{S}_t \cap \{\mathsf{s}\colon \langle \mathsf{s}, a_t^{\mathrm{ag}} - a_t\rangle \leqslant \pi(t)\}$ if $A_t = a_t$ or $\mathcal{S}_{t+1} = \mathcal{S}_t \cap \{\mathsf{s}\colon \langle \mathsf{s}, a_t^{\mathrm{ag}} - a_t\rangle \geqslant \pi(t)\}$ otherwise.

However, since the action set $\mathcal{A}_t$ is non-stationary, we cannot determine the action a_t^{ag} by just observing the first round. Consequently, the new set $\mathcal{S}_{t+1}$ cannot be computed as previously in the non-contextual setting. This makes the *single-point* incentive functions not suited for an efficient learning of $\mathsf{s}^\star$ over the iterations. Instead, at any time t, we seek for a form of binary-search feedback in the direction $a_t^1 - a_2^t$ for any two arms $a_t^1 \neq a_t^2 \in \mathcal{A}_t$. As we will see, this can be achieved by considering an incentive function κ with support $\{a_t^1, a_t^2\} \subseteq \mathcal{A}_t$.

Indeed, an important remark is that the amount of incentive needed to make the agent play any particular action is bounded under **H**2 since

$$\max_{a\in\mathcal{A}_t} \pi^\star(t,a) = \max_{a\in\mathcal{A}_t} \langle \mathsf{s}^*, a_t^{\mathrm{ag}} - a\rangle \leqslant 2\ , \qquad (13)$$

this bound being known by the principal. For any $a \in \mathrm{B}(0,1)$, this makes the incentive function $a' \mapsto 3\cdot\mathbb{1}_a(a')$ sufficient to ensure $A_t = a$ from (8). The value 3 in the definition of the incentive function is chosen instead of 2 to avoid an arbitrary tie-breaking.

Consequently, under the choice $\kappa(t,a_t^1) = 3$, $\kappa(t,a_t^2) = 3+\pi(t)$ for $\pi(t)\geqslant 0$ and $\kappa(t,a') = 0$ for any other arm a', (13) guarantees that only a_t^1 and a_t^2 may be chosen by the agent, helping the principal to update her confidence set $\mathcal{S}_t$ in a known direction. Specifically for such an incentive κ, the choice $A_t = a_t^1$ reveals the following information on $\mathsf{s}^\star$:

$$\langle \mathsf{s}^\star, a_t^1 - a_t^2\rangle \geqslant \pi(t),$$

that permits the definition of a *binary search-like feedback* in the direction $a_t^1 - a_t^2$ and thus allows us to update the confidence set $\mathcal{S}_t$ following $\mathcal{S}_{t+1} = \mathcal{S}_t \cap \{\mathsf{s}\colon \langle \mathsf{s}, a_t^1 - a_t^2\rangle \geqslant \pi(t)\}$ if $A_t = a_t^1$ or $\mathcal{S}_{t+1} = \mathcal{S}_t \cap \{\mathsf{s}\colon \langle \mathsf{s}, a_t^1 - a_t^2\rangle \leqslant \pi(t)\}$ otherwise.

Binary search. This update turns our estimation of the optimal incentives $\pi^\star$ into a multidimensional binary search where the unknown quantity is the vector $\mathsf{s}^\star$. At each iteration t, a vector w_t from the unit sphere is given. Then, the algorithm has to guess the value of $\langle \mathsf{s}^\star, w_t\rangle$ using its previous observations. Finally, an oracle reveals as feedback whether the guess is above or below the true value $\langle \mathsf{s}^\star, w_t\rangle$, and the algorithm updates its observation history. In our case, $w_t \coloneqq (a_t^1 - a_t^2)/\|a_t^1 - a_t^2\|$ for $a_t^1, a_t^2 \in \mathcal{A}_t$ and the resulting feedback is given through the agent picking either a_t^1 or a_t^2. However, extending the binary search to the multidimensional case is non-trivial for two reasons.

Direction of the multidimensional binary search. In the contextual bandit setting, we cannot divide the horizon into two successive phases. Indeed, the principal cannot choose any binary search direction in $\mathbb{R}^d$, since w_t depends on the action set $\mathcal{A}_t$ available at each iteration. For instance, action sets $\mathcal{A}_t$ could be restricted to a small dimensional subspace of $\mathbb{R}^d$ during the whole binary search procedure, so that the principal can only get a good estimate of $\mathsf{s}^\star$ in this subspace. After this phase, received action sets could be totally different (e.g., in the orthogonal subspace or the whole of $\mathbb{R}^d$) during the remainder of the game.

We solve the issue of constraint directions for the binary search by running it in an adaptive way, depending on the available action set at each time step and on the current level of estimation on this set. More precisely, at iteration t, the principal's estimate of the true value $\langle \mathsf{s}^\star, w_t\rangle$ is $\langle \hat{\mathsf{s}}_t, w_t\rangle$, where $\hat{\mathsf{s}}_t$ is defined as the centroid of $\mathcal{S}_t$:

$$\hat{\mathsf{s}}_t \coloneqq \frac{1}{\mathrm{Vol}(\mathcal{S}_t)}\int_{\mathcal{S}_t} x\mathrm{d}x \quad \text{with} \quad \mathrm{Vol}(\mathcal{S}_t) = \int_{\mathcal{S}_t}\mathrm{d}x\ .$$

Whenever $|\langle \hat{\mathsf{s}}_t, w_t\rangle - \langle \mathsf{s}^\star, w_t\rangle| < 1/T$, the principal incurs a negligible cost to incentivize the agent to choose her desired action. Then, in this context, for any action $a\in\mathcal{A}_t$ that the principal wants to play, she designs the incentive

$$\hat{\pi}(t,a) \coloneqq \max_{a'\in\mathcal{A}_t}\langle \hat{\mathsf{s}}_t, a'\rangle - \langle \hat{\mathsf{s}}_t, a\rangle + 2/T\ ,$$
$$\hat{\kappa}_a(t,a') = \mathbb{1}_a(a')\hat{\pi}(t,a) \quad \text{for any } a'\in\mathcal{A}_t\ .$$

To control the precision of the estimation $\hat{\pi}(t,a)$ of $\pi^\star(t,a)$ for any $a\in\mathcal{A}_t$, Lemma 4 shows that it is sufficient to consider the event $\mathcal{E}_t$, defined as

$$\mathcal{E}_t \coloneqq \left\{\max_{a_t^1\neq a_t^2\in\mathcal{A}_t} \mathrm{diam}\left(\mathcal{S}_t, \frac{a_t^1 - a_t^2}{\|a_t^1 - a_t^2\|}\right) < \frac{1}{T}\right\}\ , \qquad (14)$$

where we recall the definition of the projected diameter: $\mathrm{diam}(\mathcal{S}_t, x) \coloneqq \max_{\mathsf{s}_1,\mathsf{s}_2\in\mathcal{S}_t}|\langle \mathsf{s}_1 - \mathsf{s}_2, x\rangle|$ for any $x\in\mathbb{R}^d$. When $\mathcal{E}_t$ is false, the principal does not have a good characterization of the incentive function that she needs to provide and thus `Contextual IPA` runs a multidimensional binary search step, which is explained in the paragraph below. Otherwise, `Contextual IPA` runs a contextual bandit subroutine `CtxAlg` in a black-box manner on her bandit instance driven by the principal's own mean rewards $\langle \theta^\star, a\rangle$ and the approximated incentives $\hat{\pi}(t,a)$ for any

$a \in \mathcal{A}_t$. Lemma 4 guarantees that these approximations are upper estimates of $\pi^\star(t,a)$. The principal proposes an incentive function $\hat{\kappa}_{a_t^{\mathrm{Rec}}}$ depending on the estimate to make the agent select the action a_t^{Rec} recommended by the bandit subroutine. Again, we do not impose any assumption on the tie-breaking, which can be arbitrary.

Lemma 4. *Consider $t \in [T]$, $\mathcal{A}_t \subseteq \mathrm{B}(0,1)$, $\mathcal{S}_t \subseteq \mathrm{B}(0,1)$ such that $\mathcal{E}_t$ defined in (14) is true. Then for any action $a \in \mathcal{A}_t$, we have: $\pi^\star(t,a) < \hat{\kappa}_a(t,a) \leqslant \pi^\star(t,a) + 4/T$.*

A corollary of Lemma 4 is that running `Contextual IPA`, under $\mathcal{E}_t$, $A_t = a_t^{\mathrm{Rec}}$.

Issue of the diameter reduction. We illustrate the challenge of the multidimensional constrained binary search on a very simple problem. At time t, we can only run a binary search step in one of the directions $w_t = a - a'$ for $a, a' \in \mathcal{A}_t$. Suppose that we have two directions of interest, v_1, v_2 in $\mathbb{R}^d$, such that we aim to decrease the diameter of $\mathcal{S}_t$ in the direction of v_1 or v_2. Even if we divide the diameter of $\mathcal{S}_t$ in two in a direction w_t, which is always possible, this does not necessarily imply that the diameter of $\mathcal{S}_t$ would reduce along any direction v_i, as illustrated on Figure 1.

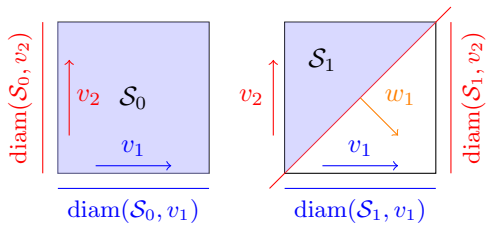

Figure 1: Illustration of a case where the volume $\mathcal{S}_0$ is cut along a direction w_1 to give a new confidence set $\mathcal{S}_1$; while the diameter is not reduced along the directions v_1 nor v_2.

An early attempt to tackle this multidimensional binary search problem with adversarial directions was presented by Cohen et al. (2020), who used ellipsoid methods. Here, we use the recent strategy proposed by Lobel et al. (2018) and their `Projected Volume` subroutine, which is described further in Appendix E.

Non-stationarity of the reward shift. For any rewards $\{(Y_a(t))_{a\in\mathcal{A}_t} : t \in [T]\}$, we define the history as $\mathcal{G}_t := (\mathcal{A}_s, A_s, U_s, Y_{A_s}(s))_{s\leqslant t}$ where $(U_s)_{s\in\mathbb{N}}$ is a family of independent uniform random variables on $[0,1]$ to allow randomization in the decision making. Let `CtxAlg` be a linear contextual bandit algorithm, i.e., $\texttt{CtxAlg}\colon (U_t, \mathcal{G}_{t-1}, \mathcal{A}_t) \mapsto a_t^{\mathrm{Rec}} \in \mathcal{A}_t$. When the principal is not running a binary search step, i.e., when $\mathcal{E}_t$ is true, she plays the `CtxAlg` subroutine on her bandit instance. We define a subset I_t of all the iterations during which `CtxAlg` is run and a shifted history $\tilde{\mathcal{H}}_t$ available at time t as

$$I_t := \{s \in [t] \text{ such that } \mathcal{E}_s \text{ is true}\}\,, \tag{15}$$

and $\tilde{\mathcal{H}}_t := (\mathcal{A}_s, A_s, U_s, X_{a_s^{\mathrm{Rec}}}(s) - \hat{\kappa}_{a_t^{\mathrm{Rec}}}(s, A_s))_{s\in I_t}$.

In our setup, `CtxAlg` will be fed in a black-box manner with this shifted history to issue recommendations a_t^{Rec}, $\texttt{CtxAlg} : (U_t, \tilde{\mathcal{H}}_{t-1}, \mathcal{A}_t) \mapsto a_t^{\mathrm{Rec}}$. At time t, for an action a_t^{Rec} recommended by `CtxAlg`, our meta-algorithm `Contextual IPA` proposes an incentive designed so that the agent eventually picks a_t^{Rec} (Lemma 4): $A_t = a_t^{\mathrm{Rec}}$.

However, the last difference between the non-contextual and contextual cases is that the shift between $\hat{\pi}(t, a_t)$ and $\pi^\star(t, a_t)$ is not constant anymore on bandit steps $t \in I_T$. This shift of rewards is interpreted as adversarial corruption (Bubeck & Slivkins, 2012; Lykouris et al., 2018).

At each round, taking into account this shift, the optimal average utility associated with action $a \in \mathcal{A}_t$ for the principal is $r^\star(t,a) := \langle \theta^\star, a\rangle - \pi^\star(t,a)$, while the principal can only estimate a *non-stationary* expected reward[2] $r^\star(t,a) + \varepsilon_t^{\text{corrupt}}$ with the *corruption level* $\varepsilon_t^{\text{corrupt}}$ defined as

$$\varepsilon_t^{\text{corrupt}} := \pi^\star(t, A_t) - \hat{\pi}(t, A_t) \tag{16}$$

and $\varepsilon_{I_t}^{\text{corrupt}} := (\varepsilon_s^{\text{corrupt}})_{s\in I_t}$. In this setup, we can define a corrupted regret as follows

$$R^{\text{corrupt}}_{\texttt{CtxAlg}}(I_T, \varepsilon^{\text{corrupt}}_{I_T}) = \mathbb{E}\left[\sum_{t\in I_T} \max_{a\in\mathcal{A}_t} r^\star(t,a) - r^\star(t, a_t^{\mathrm{Rec}})\right], \tag{17}$$

where $a_t^{\mathrm{Rec}} = \texttt{CtxAlg}(U_t, \tilde{\mathcal{H}}_{t-1}(\varepsilon^{\text{corrupt}}_{I_{t-1}}), \mathcal{A}_t)$ and $\tilde{\mathcal{H}}_t(\varepsilon^{\text{corrupt}}_{I_t}) = (\mathcal{A}_s, A_s, U_s, r^\star(s, a_s^{\mathrm{Rec}}) + \eta_{a_s^{\mathrm{Rec}}}(s) + \varepsilon_s^{\text{corrupt}})_{s\in I_t}$. Then, we aim to minimize the corrupted regret with `CtxAlg`, which is not possible using a naive linear contextual bandit algorithm.

Regret analysis. We split the regret into three components, each of them being bounded separately. One of these components comes from the bias in the estimation of the optimal incentives. Secondly, the principal incurs a cost due to the iterations of `CtxAlg` on the corrupted bandit instance. We use the results from He et al. (2022) with a known corruption level to bound this term. Finally, the last term follows from the multidimensional binary search steps used to estimate $\mathtt{s}^\star$. Lemma 5 allows us to bound the number of such steps; see Appendix D for a proof which builds on the work of Lobel et al. (2018).

Lemma 5. *Consider $\mathcal{E}_t$ defined by (14) with $(\mathcal{S}_t)_{t\in[T]}$ defined by* `Contextual IPA`. *Then it holds almost surely*

[2] Even if we were to feed the stochastic observations $(X_{A_s}(s) - \hat{\pi}(t, A_s))_{s\leqslant t}$ at time t, past algorithmic decisions would depend on different observation distributions, making the direct use of classical regret bounds of the bandit subroutine impossible.

Algorithm 2 `Contextual IPA`

Input: horizon T, subroutine `CtxAlg`, $\bar{\delta} = 1/16T^2d(d+1)^2$
Initialize: $\tilde{\mathcal{H}}_0 = V_0 = \varnothing$, $\mathcal{S}_0 = \{s \in \mathbb{R}^d : \|s\| \leqslant 1\}$
for $t = 1, \dots, T$ **do**
  Observe available action set $\mathcal{A}_t$
  if $\mathcal{E}_t$ is `FALSE` **then**
    # Where $\mathcal{E}_t$ is defined in (14)
    $\mathcal{S}_{t+1}, \mathsf{V}_{t+1}$
    $=$ `Projected Volume`$(T, \bar{\delta}, \mathcal{S}_t, \mathsf{V}_t, a_t^1, a_t^2)$
  else
    Compute $\hat{\mathsf{s}}_t$ as the centroid of $\mathcal{S}_t$
    Sample $U_t \sim \mathrm{U}(0,1)$
    Get $a_t^{\mathrm{Rec}} = \texttt{CtxAlg}(U_t, \tilde{\mathcal{H}}_t, \mathcal{A}_t)$
    Let $\hat{\pi}(t, a_t^{\mathrm{Rec}}) = \max_{a' \in \mathcal{A}_t} \langle \hat{\mathsf{s}}_t, a' \rangle - \langle \hat{\mathsf{s}}_t, a_t^{\mathrm{Rec}} \rangle + \frac{2}{T}$
    Propose incentive function $\hat{\kappa}_{a_t^{\mathrm{Rec}}}(t, a) = \mathbb{1}_{a_t^{\mathrm{Rec}}}(a) \cdot \hat{\pi}(t, a_t^{\mathrm{Rec}})$
    Observe A_t as defined in (8), X_{A_t}
    Update $\tilde{\mathcal{H}}_t$ with $(\mathcal{A}_t, A_t, U_t, X_{A_t} - \kappa(t, A_t))$
  end if
end for

that

$$\sum_{t \geqslant 1} \mathbb{1}_{\mathcal{E}_t^c} \leqslant 192 \cdot d \log(dT) \ ,$$

where $\mathcal{E}_t$ is defined by (14).

Theorem 2. *If* `Contextual IPA` *is run with any linear contextual subroutine* `CtxAlg`*, then with the same constant* 192 *as in Lemma 5, the regret of* `Contextual IPA` *is bounded as*

$$\mathfrak{R}(T) \leqslant 2 + R^{\mathit{corrupt}}_{\texttt{CtxAlg}}(I_T, \varepsilon^{\mathit{corrupt}}_{I_T}) + 1344\, d \log(dT) \ .$$

We emphasize that our results still hold with any contextual linear bandit algorithm and that the overall regret is mostly driven by the term $R^{\text{corrupt}}_{\texttt{CtxAlg}}$: although the principal has to solve simultaneously a *pricing-like* problem and a *stochastic bandit* problem, she almost achieves linear bandit state-of-the art regret.

The main difference between traditional and corrupted rewards in the bandit setting lies in the fact that, in the former case, rewards are typically assumed to be i.i.d., whereas in the latter case, they may be chosen adversarially. An algorithm is robust to corruptions if it yields regret guarantees for any possible reward corruption within a specific budget. This kind of problem was first considered by Bubeck & Slivkins (2012) and has been extensively studied since (see, e.g., Kapoor et al., 2019). In our setting, a bound for the corruption budget is available, thanks to Lemma 6 below.

Lemma 6. *Consider $(\mathcal{E}_t)_{t\in[T]}$ defined by (14) with $(\mathcal{S}_t)_{t\in[T]}$ defined by* `Contextual IPA`*. Let I_T and $(\varepsilon_t^{\mathit{corrupt}})_{t\in[T]}$ as defined in (15) and (16) and $t \in [T]$ such that $\mathcal{E}_t$ is true. Then $|\varepsilon_t^{\mathit{corrupt}}| \leqslant 4/T$ and $\sum_{t\in I_T} |\varepsilon_t^{\mathit{corrupt}}| := C_{\mathit{corrupt}} \leqslant 4$.*

With standard bandit assumptions, we can then consider a corruption robust algorithms, such as $\texttt{CW} - \texttt{OFUL}$ from He et al. (2022).

H 3. *At each round $t \geqslant 1$, for any action $a \in \mathcal{A}_t$, the principal's reward $X_a(t)$ is $\tilde{\mathcal{H}}_t$-conditionally 1-subgaussian, i.e., for any $\lambda \in \mathbb{R}$, we have $\mathbb{E}\big[\mathrm{e}^{\lambda(X_a(t) - \mathbb{E}[X_a(t)])} \big| \mathcal{H}_{t-1}\big] \leqslant \mathrm{e}^{\lambda^2/2}$.*

Corollary 2. *Suppose that **H**3 is true. If Algorithm 2 is run with the subroutine* $\texttt{CtxAlg} := \texttt{CW} - \texttt{OFUL}$ *proposed by He et al. (2022), the regularization parameter $\lambda = 1$ and a confidence level $\delta = 1/T$, the following bound holds*

$$\mathfrak{R}(T) \leqslant 11 + 1344\, d \log(dT) + C_{\texttt{CtxAlg}} d\sqrt{T} \log(T),$$

with $C_{\texttt{CtxAlg}}$ being an universal constant.

As in the multi-armed setting, the obtained regret bounds are comparable to the achievable best performance in the standard bandit settings, where the principal does not need to estimate the agent's parameters $s^\star$.

5 Experiments

We illustrate our theoretical findings with experiments on a toy example and compare `IPA` with the `Principal's` ε`-Greedy algorithm` of Dogan et al. (2023b). We also compare with a `UCB Oracle` baseline that runs `UCB` on the shifted bandit instance with arm means $\mu_a := \theta_a - \pi_a^\star$ and no principal-agent consideration. This baseline corresponds to the case where the principal knows the agent's reward vector and therefore only has to consider a bandit algorithm. Experimental details can be found in Appendix B. We observe in Figure 2 that the Principal ε-Greedy Algorithm from Dogan et al. (2023b) exhibits suboptimal performance. Additionally, another issue arises from its computational complexity, requiring an optimization step at every round. In comparison, `IPA` yields a regret nearly equal to the one of `Oracle UCB`, illustrating that the cost of estimating the agent's preferences, obtained from binary search, is negligible for `IPA`.

6 Lower Bounds

For the sake of clarity, we stick to the multi-armed case of Section 3 in this section. A simple observation yields that

$$\mathfrak{R}(T) \geqslant \mathbb{E}\left[\sum_{t=1}^{T} \mu^\star - \mu_{A_t}\right] \ ,$$

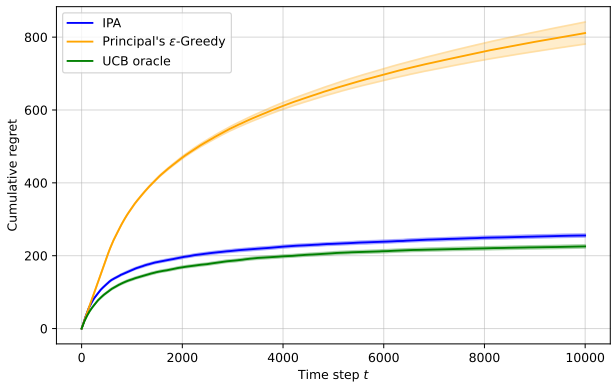

Figure 2: Cumulative regret for different algorithms on a 5 arms instance.

where $\mu_a = \theta_a - \pi^\star_a$ and $\mu^\star = \max_{a\in[K]} \mu_a$. Even if the principal was to know the optimal incentives $(\pi^\star_a)_{a\in[K]}$, she would still face a bandit instance with arm means μ_a. From there, we can directly extend standard lower bounds from the bandit literature to our setting (Lai & Robbins, 1985; Burnetas & Katehakis, 1996).

Proposition 1. *Let $\mathcal{D}$ be a class of distributions. Consider the multi-armed case of Section 3 and a policy satisfying for any instance $\nu \in \mathcal{D}^K$ and $\alpha > 0$, $\mathfrak{R}(T) = o(T^\alpha)$. Then, for any $\nu \in \mathcal{D}^K$,*

$$\liminf_{T\to\infty} \frac{\mathfrak{R}(T)}{\log T} \geqslant \sum_{a,\mu_a<\mu^\star} \frac{\mu^\star - \mu_a}{\mathrm{KL}_{\inf}(\nu_a - \pi^\star_a, \mu^\star, \mathcal{D})} \,,$$

where denoting by KL *the Kullback-Leibler divergence,*

$$\mathrm{KL}_{\inf}(\rho, \mu^\star, \mathcal{D}) \coloneqq \inf\{\mathrm{KL}(\rho,\rho') : \rho' \in \mathcal{D} \,, \int x\rho'(\mathrm{d}x) > \mu^\star\} \,,$$

The complete proof is postponed to Appendix F. Proposition 1 states that `IPA` yields a nearly optimal regret. Similar arguments can be made in the contextual setting.

7 Conclusion and Possible Extensions

This paper presents two novel algorithms called `IPA` and `Contextual IPA`, tackling generalizations of both multi-armed and contextual bandits that account for principal-agent interactions. By decoupling the learning of the agent and the estimation of the principal's parameters, we are able to obtain a nearly optimal algorithm, improving over the previous work of Dogan et al. (2023b). Overall, we obtain an efficient principal-agent bandit framework that allows us to take into account an interaction between a principal and an agent with misaligned interests in a bandit environment. There are various possible extensions of our work, among which considering strategic behaviors for repeated interactions with a single agent or uncertainty on the agent's side.

Information rent. Again, we consider the multi-armed case for the sake of clarity. We assumed that the agent is always greedy and therefore chooses at time t an action following

$$A_t = \mathrm{argmax}_{a\in\mathcal{A}_t}\{s_a + \mathbb{1}_{a_t}(a)\pi(t)\} \,. \tag{18}$$

However, nothing prevents the agent from lying and choosing another action instead of (18). The maximal total welfare that can be extracted at each round is $\max_{a\in\mathcal{A}}\{s_a + \theta_a\}$. In our setting, with a trustful agent, this reward was shared between the two actors with an average reward $\max_{a'\in\mathcal{A}} s_{a'}$ for the agent and $\max_{a\in\mathcal{A}}\{s_a + \theta_a\} - \max_{a'\in\mathcal{A}} s_{a'}$ for the principal. However, as it is exposed by Dogan et al. (2023b, Section 4), the agent could play with a malicious policy and choose A_t as if he had different $(s_a)_{a\in[K]}$. In that case, he can extract an individual reward $\max_{a\in\mathcal{A}}\{s_a + \theta_a\} - \min_{a'\in\mathcal{A}} \theta_{a'}$, while letting a reward $\min_{a'\in[K]} \theta_{a'}$ to the principal. In that case, the agent exploits his *information rent* to increase his profits. Against such adversarial and powerful agents, the principal cannot do more than play and learn with the $(s_a)_{a\in[K]}$ *announced* by the agent. However, this situation is not an issue with myopic agents who act greedily since each of them tries to maximize his own instantaneous reward and eventually select A_t from (18). This situation is encountered in many applications, where each agent has a single round interaction with the principal for the whole game.

Learning agents. A possible extension would be to incorporate uncertainty on the agent's side and consider learning agents (see, e.g., Dogan et al., 2023a). However again, when considering single round interactions with the agents, each agent myopically maximizes his reward *a priori*. Consequently, the agent policy is stationary and would be driven by s_a, the expected *beliefs* on the action rewards. The single interaction model is already well suited for numerous real world applications. In the case of repeated interactions between the principal and a single learning agent, it becomes much more complex, as this agent can both learn his true rewards on the run while trying to influence future actions of the principal with his own choices. Restricting the agent's policy to a specific set might then be necessary, as done by Dogan et al. (2023a) with Agent's ε-Greedy strategy. The major learning difficulty from the principal side would then come from the non-stationarity of the agents decisions and could be handled using non-stationary bandits algorithms (see, e.g., Gittins, 1979; Lattimore & Szepesvári, 2020).

Acknowledgements

Funded by the European Union (ERC, Ocean, 101071601). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. The work of DT has been supported by the Paris Île-de-France Région in the framework of DIM AI4IDF.

Impact Statement

This paper aims to advance the field of principal-agent problems, particularly in the online setting. We believe that this study could lead to several positive real-world outcomes, though we do not feel it is necessary to highlight any specific consequences here. However, we did mention a few relevant domains of interest in the introduction.

References

Abe, N. and Long, P. M. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pp. 3–11, 1999.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Banihashem, K., Hajiaghayi, M., Shin, S., and Slivkins, A. Bandit social learning: Exploration under myopic behavior. *arXiv preprint arXiv:2302.07425*, 2023.

Ben-Porat, O., Mansour, Y., Moshkovitz, M., and Taitler, B. Principal-agent reward shaping in mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 9502–9510, 2024.

Bernasconi, M., Castiglioni, M., Marchesi, A., Gatti, N., and Trovò, F. Sequential information design: Learning to persuade in the dark. *Advances in Neural Information Processing Systems*, 35:15917–15928, 2022.

Bertsimas, D. and Vempala, S. Solving convex programs by random walks. *Journal of the ACM (JACM)*, 51(4): 540–556, 2004.

Boursier, E. and Perchet, V. A survey on multi-player bandits. *arXiv preprint arXiv:2211.16275*, 2022.

Bubeck, S. and Slivkins, A. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 42–1. JMLR Workshop and Conference Proceedings, 2012.

Burnetas, A. N. and Katehakis, M. N. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Cai, J., Chen, R., Wainwright, M. J., and Zhao, L. Doubly high-dimensional contextual bandits: An interpretable model for joint assortment-pricing. *arXiv preprint arXiv:2309.08634*, 2023.

Castiglioni, M., Celli, A., Marchesi, A., and Gatti, N. Online bayesian persuasion. *Advances in Neural Information Processing Systems*, 33:16188–16198, 2020.

Castiglioni, M., Marchesi, A., Celli, A., and Gatti, N. Multi-receiver online bayesian persuasion. In *International Conference on Machine Learning*, pp. 1314–1323. PMLR, 2021.

Chen, S., Wang, M., and Yang, Z. Actions speak what you want: Provably sample-efficient reinforcement learning of the quantal stackelberg equilibrium from strategic feedbacks. *arXiv preprint arXiv:2307.14085*, 2023.

Cohen, A., Deligkas, A., and Koren, M. Learning approximately optimal contracts. In *International Symposium on Algorithmic Game Theory*, pp. 331–346. Springer, 2022.

Cohen, M. C., Lobel, I., and Paes Leme, R. Feature-based dynamic pricing. *Management Science*, 66(11):4921–4943, 2020.

Conitzer, V. and Garera, N. Learning algorithms for online principal-agent problems (and selling goods online). In *Proceedings of the 23rd international conference on Machine learning*, pp. 209–216, 2006.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. 2008.

Den Boer, A. V. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science*, 20(1): 1–18, 2015.

Dogan, I., Shen, Z.-J. M., and Aswani, A. Estimating and incentivizing imperfect-knowledge agents with hidden rewards. *arXiv preprint arXiv:2308.06717*, 2023a.

Dogan, I., Shen, Z.-J. M., and Aswani, A. Repeated principal-agent games with unobserved agent rewards and perfect-knowledge agents. *arXiv preprint arXiv:2304.07407*, 2023b.

Doval, L. and Ely, J. C. Sequential information design. *Econometrica*, 88(6):2575–2608, 2020.

Evans, K. J., Terhorst, A., and Kang, B. H. From data to decisions: helping crop producers build their actionable knowledge. *Critical reviews in plant sciences*, 36(2): 71–88, 2017.

Gittins, J. C. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(2):148–164, 1979.

Golrezaei, N., Jaillet, P., and Liang, J. C. N. Incentive-aware contextual pricing with non-parametric market noise. In *International Conference on Artificial Intelligence and Statistics*, pp. 9331–9361. PMLR, 2023.

Grötschel, M., Lovász, L., and Schrijver, A. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.

He, J., Zhou, D., Zhang, T., and Gu, Q. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *Advances in Neural Information Processing Systems*, 35:34614–34625, 2022.

Hu, X., Ngo, D., Slivkins, A., and Wu, S. Z. Incentivizing combinatorial bandit exploration. *Advances in Neural Information Processing Systems*, 35:37173–37183, 2022.

Javanmard, A. and Nazerzadeh, H. Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research*, 20(1):315–363, 2019.

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

Kapoor, S., Patel, K. K., and Kar, P. Corruption-tolerant bandit learning. *Machine Learning*, 108(4):687–715, 2019.

Laffont, J.-J. and Martimort, D. The theory of incentives: the principal-agent model. In *The Theory of Incentives*. Princeton University Press, 2009.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1): 4–22, 1985.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on the World Wide Web*, pp. 661–670, 2010.

Lobel, I., Leme, R. P., and Vladu, A. Multidimensional binary search for contextual decision-making. *Operations Research*, 66(5):1346–1361, 2018.

Lykouris, T., Mirrokni, V., and Paes Leme, R. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, 2018.

Mansour, Y., Slivkins, A., and Syrgkanis, V. Bayesian incentive-compatible bandit exploration. *Operations Research*, 68(4):1132–1161, 2020.

Mao, J., Leme, R., and Schneider, J. Contextual pricing for lipschitz buyers. *Advances in Neural Information Processing Systems*, 31, 2018.

Myerson, R. B. *Mechanism design*. Springer, 1989.

Rademacher, L. A. Approximating the centroid is hard. In *Proceedings of the Twenty-Third Annual Symposium on Computational Geometry*, pp. 302–305, 2007.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Sellke, M. and Slivkins, A. The price of incentivizing exploration: A characterization via thompson sampling and sample complexity. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 795–796, 2021.

Simchowitz, M. and Slivkins, A. Exploration and incentives in reinforcement learning. *Operations Research*, 2023.

Slivkins, A. et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Smith, D. J. and Vamanamurthy, M. K. How small is a unit ball? *Mathematics Magazine*, 62(2):101–107, 1989.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Woodroofe, M. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

Yu, C., Liu, J., Nemati, S., and Yin, G. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Zhong, H., Yang, Z., Wang, Z., and Jordan, M. I. Can reinforcement learning find stackelberg-nash equilibria in general-sum markov games with myopic followers? *arXiv preprint arXiv:2112.13521*, 2021.

Zhu, B., Bates, S., Yang, Z., Wang, Y., Jiao, J., and Jordan, M. I. The sample complexity of online contract design. *arXiv preprint arXiv:2211.05732*, 2022.

A Notation

$\mathcal{A} \coloneqq [K]$	Set of possible arms.
T	Horizon.
$N_T \coloneqq \log_2 T$	Number of steps dedicated to the binary search on each arm in `IPA`.
a_t	Arm on which the principal offers an incentive.
$\pi(t)$	Amount of incentive offered by the principal on action a_t.
A_t	Arm chosen by the agent, maximizing his utility, known by everyone.
$\mathsf{s}_a + \mathbb{1}_{a_t}(a)\pi(t)$	Agent's utility for action a.
ν_a	Principal's reward distribution for action a.
$X_a(t) - \mathbb{1}_{a_t}(a)\pi(t)$	Principal's utility for action a.
μ_a	Principal's expected utility for action a, using the optimal incentive $\pi_a^\star$.
$\mu^\star$	Maximal expected utility for the principal.
$\theta_a \coloneqq \mathbb{E}[X_a(1)]$	Principal's expected reward.
$\pi_a^\star$	Infimum amount of incentives to be offered on action $a_t = a$ to make the agent choose it.
$\tilde{\mathcal{H}}$	Shifted history used to feed `Alg`.
$R_{\texttt{Alg}}(T)$	Regret of the subroutine `Alg` on a horizon T.
$\mathfrak{R}(T)$	Overall regret of `IPA` on a horizon T.

Table 1: Notations used in Section 3.

T	Horizon.
$\mathrm{B}(0,1)$	Unit ball in $\mathbb{R}^d$, $d \geqslant 1$.
$\mathcal{A}_t \subseteq \mathrm{B}(0,1)$	Action set at time t among which the agent selects A_t.
$\pi(t)$	Amount of incentive offered by the principal on some action.
$\eta_a(t)$	Noise distribution of the principal's reward associated with action a at time t.
$r^\star(t,a) = \langle \theta^\star, a\rangle + \eta_a(t) - \pi^\star(t,a)$	Utility collected by the principal on action a at time t if the optimal amount of incentive is used.
$\mu_t^\star$	Principal's maximal expected utility at time t.
$\pi^\star(t,a)$	Infimal amount of incentives to be offered on action a with $\kappa(t,a') = \mathbb{1}_a(a')\pi_a^\star$ so that the agent eventually chooses action a.
$\hat{\pi}(t,a)$	Principal's estimation of $\pi^\star(,a)$
$\kappa(t,\cdot)\colon \mathrm{B}(0,1) \to \mathbb{R}_+$	Incentive function, associating each action with some amount of incentives.
$\mathsf{s}^\star \in \mathrm{B}(0,1)$	Agent's true reward vector.
$\mathcal{S}_t \ni \mathsf{s}^\star$	Principal's confidence set for $\mathsf{s}^\star$ at time t.
$\langle \mathsf{s}^\star, a\rangle + \kappa(t,a)$	Agent's utility for action a.
$a_t^{\mathrm{ag}} = \mathrm{argmax}_{a\in\mathcal{A}_t}\langle \mathsf{s}^\star, a\rangle$	Agent's optimal action with null incentives at time t.
a_t^{Rec}	Action recommended by `CtxAlg` at time t.
$\langle \theta^\star, a\rangle + \eta_a(t) - \kappa(t,a)$	Principal's utility for action a.
$\mu_t^\star$	Maximal expected utility for the principal at time t.
$\varepsilon_t^{\text{corrupt}} = \pi^\star(t,a) - \hat{\pi}(t,a)$	Shift between the optimal incentives and the estimated ones.
C_{corrupt}	Total corruption budget due to the shift between $\hat{\pi}(t,a)$ and $\pi^\star(t,a)$ over the rounds.
$\mathcal{E}_t$	Event being true if the diameter of $\mathcal{S}_t$ projected $\mathcal{A}_t$ is small than $1/T$.
I_t	Rounds up to time t during which $\mathcal{E}_s, s \leqslant t$ is true.
$\tilde{\mathcal{H}}$	Shifted history used to feed `Alg`.
$R_{\texttt{CtxAlg}}(T)$	Regret of the subroutine `CtxAlg` on a horizon T.
$\mathfrak{R}(T)$	Overall regret of `Contextual IPA` on a horizon T.

Table 2: Notation used in Section 4.

B Experimental details

We ran the experiments in Figure 2 for a horizon $T = 10\,000$ on an average of 100 runs on a five arms bandit. We plotted the standard error across the different runs. The expected rewards for the principal (θ) and the agent (s) are given in Table 3. The principal's rewards $X_a(t)$ are drawn from an i.i.d. distribution $X_a(t) \sim \mathcal{N}(\theta_a, 1)$ for any $a \in [K], t \in [T]$. We also run an oracle UCB instance with rewards following a Gaussian distribution $\mathcal{N}(\mu_a, 1)$ where for any $a \in [K]$, $\mu_a := \theta_a + \mathtt{s}_a - \max_{a' \in [K]} \mathtt{s}_{a'}$, as if a UCB algorithm was run with the full knowledge of the optimal incentives and was learning his own mean rewards (μ), taking into account these incentives. The mean rewards μ are also given in Table 3. We observe that the additional exploration steps needed to learn the optimal incentives in IPA are not very costly compared to the regret achieved by the UCB oracle.

For the Principal's ε-Greedy algorithm, we use the hyperparameters $\alpha = 1$ and $m = 500$. The hyperparameter m controls the number of exploration steps. We ran the Principal's ε-Greedy algorihtm on the same bandit setting for different values $m = 30, m = 100, m = 200, m = 300, m = 400, m = 500, m = 600, m = 800, m = 1000, m = 2000, m = 5000, m = 10\,000$. Below $m = 500$, the algorithm does not explore enough and incurs a linear regret on some runs, consequenly yielding a poor mean regret, whereas above $m = 500$, the algorithm explores excessively, leading to a higher regret due to overexploration. We ran the same experiments on longer horizons $T = 100\,000$ and $T = 1\,000\,000$ and the algorithms exhibited the same behavior. In practice, the tuning of the ε-Greedy algorithm depends on the reward gaps and is not common to use. This is why another advantage of IPA compared to the Principal's ε-Greedy algorithm of Dogan et al. (2023b) lies in the fact that it does not need any tuning of hyperparameters, leading to a better use in practice, on potentially broader bandit instances.

s	0.64	0.99	0.73	0.61	0.59
θ	0.30	0.24	0.88	0.07	0.65
μ	-0.05	0.24	0.62	-0.31	0.25

Table 3: Experimental parameters for Figure 2.

We did not run Contextual IPA in a contextual bandit setting because it is quite tedious to implement, due to the use of the Projected Volume subroutine from the work of Lobel et al. (2018). Even though they obtain an excellent regret bound, the computations raise specific challenges. The first issue is the computation of the centroid which is known to be a #P-hard problem (Rademacher, 2007). However, it can be solved through an approximation of the centroid, which is computable in polynomial time (see, Bertsimas & Vempala, 2004, Lemma 5 and Theorem 12). A second issue is finding directions along which the set $\mathcal{S}_t$ has a small diameter, which is needed to compute the set V_t. It is solved by Lobel et al. (2018) with an ellipsoidal approximation E of $\mathcal{S}_t$ such that $E \subseteq \mathcal{S}_t \subseteq \alpha E$ with $\alpha > 1$, since such an ellipsoid can be computed in polynomial time, (see, Grötschel et al., 2012, Corollary 4.6.9). Such a variation of the Projected Volume subroutine is presented in the work of Lobel et al. (2018, Section 9.3). It is shown that one can achieve polynomial time computations with still the same regret bound for the multidimensional binary search steps (Lobel et al., 2018, Theorem 9.4). This line of work needs to be explored for implementing Contextual IPA in practice, which is feasible but still requires a significant amount of work.

C Regret Bound for Non-Contextual Setting

Notations. We define $\overline{\pi}_a(t) \in \mathbb{R}_+$ as the upper estimate and $\underline{\pi}_a(t) \in \mathbb{R}_+$ as the lower estimate of $\pi^\star_a$ after t rounds of binary search on arm a. For any $t \in [T]$ and $a \in [K]$, we define $\pi^{\text{mid}}_a(t) := (\overline{\pi}_a(t) + \underline{\pi}_a(t))/2$. We define $N_T := \lceil \log_2 T \rceil$ as the number of binary search steps per arm and $\hat{\pi}_a$ is the estimated incentive to make the agent choose action a after N_T steps of binary search: $\hat{\pi}_a = \overline{\pi}_a(N_T) + 1/T$. Since our problem is stationary, we write $a^{\text{pr}} := \text{argmax}_{a \in [K]} \{\theta_a - \pi^\star_a\}$ for the optimal action that the principal could aim to play at each round.

Lemma 1. *For any $T \in \mathbb{N}$, the regret of any algorithm on our problem instance can be written as*

$$\mathfrak{R}(T) = T \max_{a \in [K]} \{\theta_a + \mathtt{s}_a - \max_{a' \in [K]} \mathtt{s}_{a'}\} - \mathbb{E}\left[\sum_{t=1}^{T} \{\theta_{A_t} - \mathbb{1}_{a_t}(A_t)\pi(t)\}\right].$$

Proof of Lemma 1. Recall that the regret is defined in (3) as $\mathfrak{R}(T) \coloneqq T\,\mu^\star - \sum_{t=1}^T \mathbb{E}[X_{A_t}(t) - \mathbb{1}_{a_t}(A_t)\pi(t)]$, where $\mu^\star = \sup_{a\in[K],\pi\in\mathbb{R}_+} \mathbb{E}_\nu[X_a(1)] - \pi$, such that $a \in \operatorname{argmax}_{a'\in[K]}\{\mathsf{s}_{a'} + \pi\}$. Note that we can write $\mu^\star = \sup_{a\in[K],\pi\in\mathbb{R}_+}\{\theta_a - \mathbb{1}_{\tilde{A}}(a,\pi)\pi\}$, where $\tilde{A} \coloneqq \{(a,\pi)\colon \mathsf{s}_a + \pi \geqslant \max_{a'} \mathsf{s}_{a'} + \mathbb{1}_a(a')\pi\}$. First note that if $(a,\pi) \in \tilde{A}$, then $\mathsf{s}_a - \mathsf{s}_{a'} \geqslant -\pi$ for any $a' \in [K]$, which implies by definition of the optimal incentives (4) that $\pi_a^\star \leqslant \pi$. Consequently,

$$\mu^\star = \max_{a\in[K]}\{\mathbb{E}_\nu[X_{a^{\mathrm{opt}}}(1)] - \pi_a^\star\} = \max_{a\in[K]}\{\theta_a - \max_{a'\in[K]}\{\mathsf{s}_{a'}\} + \mathsf{s}_a\} \;,$$

hence our result about the regret. □

Lemma 7. *Assume **H**1 and that we run Algorithm 3 for an action $a \in [K]$ and a number of binary searches $N_T \in \mathbb{N}$. Then, for any $t \in [N_T]$, $0 \leqslant \underline{\pi}_a(t) \leqslant \pi_a^{\mathrm{mid}}(t) \leqslant \overline{\pi}_a(t) \leqslant 1$.*

Proof. The proof is by induction on $t \in [N_T]$. For $t = 0$, it is defined by definition. Then suppose that it holds true for $t \geqslant 0$. Note that line 6 in Algorithm 3 can be written as

$$\begin{aligned} \overline{\pi}_a(t+1) &= \mathbb{1}_a(A_t)\pi_a^{\mathrm{mid}}(t) + (1 - \mathbb{1}_a(A_t))\overline{\pi}_a(t) \\ \underline{\pi}_a(t+1) &= (1 - \mathbb{1}_a(A_t))\pi_a^{\mathrm{mid}}(t) + \mathbb{1}_a(A_t)\underline{\pi}_a(t) \;, \end{aligned} \tag{19}$$

which completes the proof by applying the induction hypothesis. □

Lemma 8. *Assume **H**1 and that we run Algorithm 3 for an action $a \in [K]$ and a number of binary searches $N_T \in \mathbb{N}$. Then, for any $t \in [N_T]$,*

$$\pi_a^\star \in [\underline{\pi}_a(t), \overline{\pi}_a(t)] \;\; \textit{and} \;\; |\overline{\pi}_a(t) - \underline{\pi}_a(t)| \leqslant 1/2^t \;.$$

Proof. The proof is by induction on t. The case $t = 0$ is trivial by the initialization of Algorithm 3 and **H**1.

Suppose that the statement holds for t. Note that $\mathbb{1}(\{A_t = a\}) = \mathbb{1}(\{\pi_a^{\mathrm{mid}}(t) \geqslant \pi_a^\star\})$, therefore using (19), we obtain by using the induction hypothesis that

$$\pi_a^\star \in [\underline{\pi}_a(t+1), \overline{\pi}_a(t+1)] \;, \quad \overline{\pi}_a(t+1) - \underline{\pi}_a(t+1) = \frac{\overline{\pi}_a(t) - \underline{\pi}_a(t)}{2} \;,$$

which completes the proof. □

Proof of Theorem 1. Recall that Lemma 1 implies that

$$\mathfrak{R}(T) = \mathbb{E}\left[\sum_{t=1}^T \max_{a\in[K]}\{\theta_a - \pi_a^\star\} - (X_{A_t}(t) - \mathbb{1}_{a_t}(A_t)\pi(t))\right] \tag{20}$$

We decompose the regret between the $K\lceil\log_2 T\rceil = KN_T$ first steps during which we run the `Binary Search Subroutine` and all the subsequent ones

$$\begin{aligned} \mathfrak{R}(T) &= (\mathbf{A}) + (\mathbf{B}) \\ (\mathbf{A}) &= \mathbb{E}\left[\sum_{t=1}^{K\lceil\log_2 T\rceil} \max_{a\in[K]}\{\theta_a - \pi_a^\star\} - \{X_{A_t}(t) - \mathbb{1}_{a_t}(A_t)\pi(t)\}\right] \\ (\mathbf{B}) &= \mathbb{E}\left[\sum_{t=K\lceil\log_2 T\rceil+1}^{T} \max_{a\in[K]}\{\theta_a - \pi_a^\star\} - \{X_{A_t}(t) - \mathbb{1}_{a_t}(A_t)\pi(t)\}\right] . \end{aligned}$$

We separate the analysis of the regret, bounding independently two terms in the right-hand side of the previous decomposition. Since $\pi(t)$ is always equal to $\pi_a^{\mathrm{mid}}(t)$ for some $t \in [N_T], a \in [K]$, during the binary search phase, we use Lemma 7 to bound $\pi(t)$ by 1 for any $t \leqslant K\lceil\log_2 T\rceil$ in $(\mathbf{A})$, giving

$$(\mathbf{A}) = \mathbb{E}\left[\sum_{t=1}^{K\lceil\log_2 T\rceil} \max_{a\in[K]}\{\theta_a - X_{A_t}(t) + \mathbb{1}_{a_t}(A_t)(t)\pi(t) - \pi_{a^{\mathrm{pr}}}^\star\}\right] \tag{21}$$

$$\leqslant \sum_{t=1}^{K\lceil \log_2 T\rceil} (1+\max_{a\in[K]}\{\theta_a\}-\min_{a\in[K]}\{\theta_a\}) \leqslant (1+\max_{a\in[K]}\{\theta_a\}-\min_{a\in[K]}\{\theta_a\})K(1+\log_2 T)\,. \tag{22}$$

At the end of the binary search phase and for all the subsequent rounds $t > K\lceil \log_2 T\rceil$, $\texttt{Alg}$ recommends an action a_t^{Rec} and the principal proposes the incentive $\pi(t) = \hat{\pi}_{a_t^{\mathrm{Rec}}} = \overline{\pi}_{a_t^{\mathrm{Rec}}}(\lceil \log_2 T\rceil) + 1/T$ on action a_t^{Rec} to make the agent choose it. Lemma 8 ensures that after $\lceil \log_2 T\rceil$ rounds of binary search on action $a \in [K]$, we have

$$\underline{\pi}_a(\lceil \log_2 T\rceil) \leqslant \pi_a^\star \leqslant \overline{\pi}_a(\lceil \log_2 T\rceil) \quad \text{and} \quad \overline{\pi}_a(\lceil \log_2 T\rceil) - \underline{\pi}_a(\lceil \log_2 T\rceil) \leqslant 1/2^{\lceil \log_2 T\rceil} \leqslant 1/T\,.$$

Therefore,

$$\pi_a^\star < \hat{\pi}_a \text{ and } \hat{\pi}_a - \pi_a^\star \leqslant 2/T\,.$$

For the agent, the utility associated with action a_t^{Rec} is $\mathsf{s}_{a_t^{\mathrm{Rec}}} + \hat{\pi}_{a_t^{\mathrm{Rec}}} > \mathsf{s}_{a_t^{\mathrm{Rec}}} + \pi^\star_{a_t^{\mathrm{Rec}}}$, which guarantees that he eventually selects a_t^{Rec} at time t because of (1) and (4). It ensures that for any $t > K\lceil \log_2 T\rceil$, $A_t = a_t^{\mathrm{Rec}}$.

To compute these recommendations, $\texttt{Alg}$ is fed at any time $t \in \mathbb{N}$ with the shifted history defined in (7): $\tilde{\mathcal{H}}_t = (a_s^{\mathrm{Rec}}, U_s, X_{a_s^{\mathrm{Rec}}}(s) - \hat{\pi}_{a_s^{\mathrm{Rec}}})_{s\in[K\lceil \log_2 T\rceil+1,t]}$. Recall that we defined the shifted distribution ρ_a^T for any $a \in [K]$ as the distribution of $X_a(1) - \hat{\pi}_a$. For any $t > K\lceil \log_2 T\rceil$, $a_t^{\mathrm{Rec}} = \texttt{Alg}(U_t, \tilde{\mathcal{H}}_{t-1})$ and we define $Y_a(t) \sim \rho_a^T$ for any $t \in [K\lceil \log_2 T\rceil + 1, T]$, $a \in [K]$. In this setup, the regret of $\texttt{Alg}$ after τ subsequent steps is defined as

$$R_{\texttt{Alg}}(\tau, \rho^T) = \tau \max_{a\in[K]} \mathbb{E}_{\rho^T}[Y_a(K\lceil \log_2 T\rceil + 1)] - \mathbb{E}\left[\sum_{s=K\lceil \log_2 T\rceil+1}^{K\lceil \log_2 T\rceil+\tau} Y_{\texttt{Alg}(U_s,\tilde{\mathcal{H}}_{s-1})}(s)\right]\,.$$

Consequently, since $a_t = a_t^{\mathrm{Rec}} = A_t$

$$\begin{aligned}
(\mathbf{B}) &= \mathbb{E}\left[\sum_{t=K\lceil \log_2 T\rceil+1}^{T} \max_{a\in[K]}\{\theta_a - \pi_a^\star\} - \left(X_{a_t^{\mathrm{Rec}}}(t) - \hat{\pi}_{a_t^{\mathrm{Rec}}}\right)\right] \\
&\leqslant \mathbb{E}\left[\sum_{t=K\lceil \log_2 T\rceil+1}^{T} \max_{a\in[K]}\left\{\theta_a - \hat{\pi}_a - \left(X_{a_t^{\mathrm{Rec}}}(t) - \hat{\pi}_{a_t^{\mathrm{Rec}}}\right)\right\} + \max_{a'\in[K]}\{\hat{\pi}_{a'} - \pi^\star_{a'}\}\right] \\
&= \mathbb{E}\left[\sum_{t=K\lceil \log_2 T\rceil+1}^{T} \max_{a\in[K]}\{\theta_a - \hat{\pi}_a\} - \left(X_{a_t^{\mathrm{Rec}}}(t) - \hat{\pi}_{a_t^{\mathrm{Rec}}}\right)\right] + \mathbb{E}\left[\sum_{t=K\lceil \log_2 T\rceil+1}^{T} \max_{a'\in[K]}\{\hat{\pi}_{a'} - \pi^\star_{a'}\}\right] \\
&= (T - K\lceil \log_2 T\rceil) \max_{a\in[K]} \mathbb{E}[Y_a(K\lceil \log_2 T\rceil + 1)] - \mathbb{E}\left[\sum_{t=K\lceil \log_2 T\rceil+1}^{T} \left(X_{a_t^{\mathrm{Rec}}}(t) - \hat{\pi}_{a_t^{\mathrm{Rec}}}\right)\right] \\
&\quad + (T - K\lceil \log_2 T\rceil) \max_{a'\in[K]}\{\hat{\pi}_{a'} - \pi^\star_{a'}\} \\
&\leqslant R_{\texttt{Alg}}\big(T - K\lceil \log_2 T\rceil, (\rho_a^T)_{a\in[K]}\big) + 2.
\end{aligned}$$

Plugging $(\mathbf{A})$ and $(\mathbf{B})$ together finally gives a bound for the regret

$$\mathfrak{R}(T) \leqslant 2 + (1+\max_{a\in[K]}\{\theta_a\} - \min_{a\in[K]}\{\theta_a\})(1 + K\log_2 T) + R_{\texttt{Alg}}(T - K\lceil \log_2 T\rceil, \rho^T)\,.$$

□

Proof of Corollary 1. The case $T \leqslant 9K$ is trivial. Assume then that $T \geqslant 9K$. Note that after $\lceil \log_2 T\rceil$ rounds of binary search, $\pi_a^\star \leqslant \overline{\pi}_a \leqslant \pi_a^\star + 1/T$. We define $\Delta_a^\star := \max_{a'\in[K]}\{\theta_{a'} - \pi^\star_{a'}\} - (\theta_a - \pi_a^\star) = \max_{a'\in[K]}\{\theta_{a'} + \mathsf{s}_{a'}\} - (\theta_a + \mathsf{s}_a)$ and $\tilde{\Delta}_a := \max_{a'\in[K]}\{\theta_{a'} - \hat{\pi}_{a'}\} - (\theta_a - \hat{\pi}_a) = \max_{a'\in[K]}\{\theta_{a'} - \overline{\pi}_{a'}\} - (\theta_a - \overline{\pi}_a)$ using $\hat{\pi}_a = \overline{\pi}_a + 1/T$. Since $\pi_a^\star \leqslant \overline{\pi}_a \leqslant \pi_a^\star + 1/T$, we have $|\Delta_a^\star - \tilde{\Delta}_a| \leqslant 2/T$.

Using the results about UCB algorithm that can be found in (Lattimore & Szepesvári, 2020, Theorems 7.1 and 7.2), since Alg is run in a black-box manner on a shifted bandit instance ρ^T for $T - K\lceil \log_2 T \rceil$ rounds with reward gaps $\tilde{\Delta}_a$, we have

$$\begin{aligned} R_{\texttt{Alg}}(T - K\lceil \log_2 T\rceil, \rho^T) &\leqslant 3 \sum_{\tilde{\Delta}_a > 0} \tilde{\Delta}_a \\ &\quad + 8 \min\left\{ \sqrt{(T - K\lceil \log_2 T\rceil) K \log(T - K\lceil \log_2 T\rceil)}\,; \sum_{\tilde{\Delta}_a > 0} \frac{2\log(T - K\lceil \log_2 T\rceil)}{\tilde{\Delta}_a} \right\} \\ &\leqslant 3 \sum_{a\in[K], \Delta_a^\star > 0} \left(\Delta_a^\star + \frac{2}{T}\right) + 8 \min\left\{ \sqrt{TK\log T}\,; \sum_{a\in[K], \tilde{\Delta} > 0} \frac{2\log T}{\tilde{\Delta}_a} \right\} \\ &\leqslant 1 + 3 \sum_{a\in[K], \Delta_a^\star > 0} \Delta_a^\star + 8 \min\left\{ \sqrt{TK\log T}\,; \sum_{a\in[K], \tilde{\Delta}_a > 0} \frac{2\log T}{\tilde{\Delta}_a} \right\}, \end{aligned} \tag{23}$$

where the last line holds because of $T \geqslant 9K > 6K$.

We now analyse the sum $\sum_{a\in[K], \tilde{\Delta} > 0} 2\log T/\tilde{\Delta}_a$ and consider two cases: either there exists $\tilde{a} \in [K]$ such that $\Delta_{\tilde{a}}^\star \leqslant 4/T$ or not.

First case: if there exists $\tilde{a} \in [K]$ such that $\Delta_{\tilde{a}}^\star \leqslant 4/T$, since $T > 9K$, we have for such an action $\tilde{a}$: $2\log T/\Delta_{\tilde{a}}^\star \geqslant T\log T/2 > \sqrt{TK\log T}$ as well as $\tilde{\Delta}_{\tilde{a}} \leqslant \Delta_{\tilde{a}}^\star + 2/T \leqslant 6/T$ which is equivalent to $2\log T/\tilde{\Delta}_{\tilde{a}} \geqslant T\log T/3 \geqslant \sqrt{TK\log T}$. Consequently,

$$\min\left\{ \sqrt{TK\log T}\,; \sum_{a\in[K], \tilde{\Delta}_a > 0} \frac{2\log T}{\tilde{\Delta}_a} \right\} = \sqrt{TK\log T} = \min\left\{ \sqrt{TK\log T}; \sum_{a\in[K], \Delta_a^\star > 0} \frac{2\log T}{\Delta_a^\star} \right\}. \tag{24}$$

Second case: for any $a \in [K], \Delta_a^\star > 4/T$. Therefore $\tilde{\Delta}_a \geqslant \Delta_a^\star - 2/T > \Delta_a^\star - \Delta_a^\star/2 = \Delta_a^\star/2$. Consequently

$$\begin{aligned} R_{\texttt{Alg}}(T - K\lceil \log_2 T\rceil, \rho^T) &\leqslant 1 + 3 \sum_{a\in[K], \Delta_a^\star > 0} \Delta_a^\star + 8 \min\left\{ \sqrt{TK\log T}\,; \sum_{a\in[K], \Delta_a^\star > 0} \frac{2\log T}{\tilde{\Delta}_a} \right\} \\ &\leqslant 1 + 3 \sum_{a\in[K], \Delta_a^\star > 0} \Delta_a^\star + 8 \min\left\{ \sqrt{TK\log T}\,; \sum_{a\in[K], \Delta_a^\star > 0} \frac{4\log T}{\Delta_a^\star} \right\}. \end{aligned} \tag{25}$$

Finally, combining (24) and (25) in (23) completes the proof. □

D Regret Bound for the Contextual Setting

For the whole section, we define $\hat{a}_t^{\text{ag}} \coloneqq \operatorname{argmax}_{a\in\mathcal{A}_t} \langle \hat{\mathsf{s}}_t, a\rangle$ and recall that $a_t^{\text{ag}} = \operatorname{argmax}_{a\in\mathcal{A}_t} \langle \mathsf{s}^\star, a\rangle$.

D.1 Technical lemmas

Lemma 2. *For any $T \in \mathbb{N}$, the regret of any algorithm on our contextual problem instance can be written as*

$$\begin{aligned} \mathfrak{R}(T) = \sum_{t=1}^T \max_{a\in\mathcal{A}_t} \{\langle \theta^\star + \mathsf{s}^\star, a\rangle - \max_{a'\in\mathcal{A}_t} \langle \mathsf{s}^\star, a'\rangle\} \\ - \mathbb{E}\left[\sum_{t=1}^T (\langle \theta^\star, A_t\rangle - \kappa(t, A_t)) \right]. \end{aligned}$$

Proof of Lemma 2. Recall that the regret is defined in (11) as $\mathfrak{R}(T) = \sum_{t=1}^T \mu_t^\star - \mathbb{E}\left[\sum_{t=1}^T (X_{A_t}(t) - \kappa(t, A_t))\right]$, where $\mu_t^\star \coloneqq \sup_{\kappa(t,\cdot)\colon \mathbb{R}^d \to \mathbb{R}_+} \{\langle \theta^\star, a\rangle - \kappa(t, a)\}$ such that $a \in \operatorname{argmax}_{a'\in\mathcal{A}_t} \{\langle \mathsf{s}^\star, a'\rangle + \kappa(t, a')\}$. Note that we can write

$\mu_t^\star = \sup_{a\in\mathcal{A}_t,\pi\in\mathbb{R}_+}\{\langle\theta^\star,a\rangle - \mathbb{1}_{\tilde{A}_t}(a,\pi)\pi\}$ where $\tilde{A}_t \coloneqq \{(a,\pi)\colon \langle \mathsf{s}^\star,a\rangle + \pi \geqslant \max_{a'\in\mathcal{A}_t}\langle \mathsf{s}^\star,a'\rangle + \mathbb{1}_a(a')\pi\}$. First note that if $(a,\pi)\in\tilde{A}_t$, then $\langle \mathsf{s}^\star, a-a'\rangle \geqslant -\pi$ for any $a'\in\mathcal{A}_t$, which implies by definition of the optimal incentives (12) that $\pi^\star(t,a)\leqslant\pi$. Consequently

$$\mu_t^\star = \max_{a\in\mathcal{A}_t}\{\langle\theta^\star,a\rangle - \pi^\star(t,a)\} = \max_{a\in\mathcal{A}_t}\{\langle\theta^\star,a\rangle - \max_{a'\in\mathcal{A}_t}\{\langle \mathsf{s}^\star,a'\rangle\} + \langle \mathsf{s}^\star,a\rangle\}\ ,$$

hence our result about the regret. □

Lemma 3. *For any $t\in[T]$ and closed subset $\mathcal{S}\subset \mathrm{B}(0,1)$ with $\mathsf{s}^\star\in\mathcal{S}$, it holds, for any $a\in\mathcal{A}_t$, $|\max_{\mathsf{s}\in\mathcal{S},a'\in\mathcal{A}_t}\langle \mathsf{s},a'-a\rangle - \pi^\star(t,a)| \leqslant 2\operatorname{diam}(\mathcal{S},\mathcal{A}_t)$ where $\operatorname{diam}(\mathcal{S},\mathcal{A}_t) \coloneqq \max_{a'\in\mathcal{A}_t}\max_{\mathsf{s}_1,\mathsf{s}_2\in\mathcal{S}}|\langle \mathsf{s}_1-\mathsf{s}_2,a'\rangle|$.*

Proof of Lemma 3. For any $t\in[T], \mathcal{S}\in\mathrm{B}(0,1)$ with $s^\star\in\mathcal{S}$, $\mathcal{A}_t\in\mathrm{B}(0,1)$ and $a\in\mathcal{A}_t$, recall that we defined $a_t^{\mathrm{ag}} = \operatorname{argmax}_{a'\in\mathcal{A}_t}\langle \mathsf{s}^\star,a'\rangle$ and $\pi^\star(t,a) = \langle \mathsf{s}^\star,a_t^{\mathrm{ag}}\rangle - \langle \mathsf{s}^\star,a\rangle$. Consequently, defining for any $\mathsf{s}\in\mathcal{S}$, $a_t^{\mathsf{s}} \coloneqq \operatorname{argmax}_{a'\in\mathcal{A}_t}\langle \mathsf{s},a'\rangle$ (the compactness of both $\mathcal{S}$ and $\mathcal{A}_t$ as well as the continuity of the applications that we consider guarantee the existence of such an argmax), we have, since $\langle \mathsf{s}^\star,a_t^{\mathrm{ag}}\rangle \geqslant \langle \mathsf{s}^\star,a_t^{\mathsf{s}}\rangle$ for any $\mathsf{s}\in\mathcal{S}$ and associated a_t^{s},

$$\begin{aligned}
\max_{\mathsf{s}\in\mathcal{S},a'\in\mathcal{A}_t}\langle \mathsf{s},a'-a\rangle - \pi^\star(t,a) &= \max_{\mathsf{s}\in\mathcal{S}}\max_{a'\in\mathcal{A}_t}\{\langle \mathsf{s},a'-a\rangle - \langle \mathsf{s}^\star,a_t^{\mathrm{ag}}-a\rangle\}\\
&= \max_{\mathsf{s}\in\mathcal{S}}\{\langle \mathsf{s},a_t^{\mathsf{s}}-a\rangle - \langle \mathsf{s}^\star,a_t^{\mathrm{ag}}-a\rangle\}\\
&\leqslant \max_{s\in\mathcal{S}}\{\langle \mathsf{s},a_t^{\mathsf{s}}-a\rangle - \langle \mathsf{s}^\star,a_t^{\mathsf{s}}-a\rangle\}\\
&\leqslant \max_{s\in\mathcal{S}}|\langle \mathsf{s}-\mathsf{s}^\star,a_t^{\mathsf{s}}\rangle| + \max_{s\in\mathcal{S}}|\langle \mathsf{s}-\mathsf{s}^\star,a\rangle|\\
&\leqslant 2\operatorname{diam}(\mathcal{S},\mathcal{A}_t)\ .
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\pi^\star(t,a) - \max_{\mathsf{s}\in\mathcal{S},a'\in\mathcal{A}_t}\langle \mathsf{s},a'-a\rangle &\leqslant \langle \mathsf{s}^\star,a_t^{\mathrm{ag}}-a\rangle - \max_{\mathsf{s}\in\mathcal{S}}\langle \mathsf{s},a_t^{\mathrm{ag}}-a\rangle\\
&\leqslant \max_{s\in\mathcal{S}}|\langle \mathsf{s}^\star-\mathsf{s},a_t^{\mathrm{ag}}\rangle| + \max_{s\in\mathcal{S}}|\langle \mathsf{s}^\star-\mathsf{s},a\rangle|\\
&\leqslant 2\operatorname{diam}(\mathcal{S},\mathcal{A}_t)\ ,
\end{aligned}$$

and the proof follows. □

Lemma 4. *Consider $t\in[T]$, $\mathcal{A}_t\subseteq\mathrm{B}(0,1)$, $\mathcal{S}_t\subseteq\mathrm{B}(0,1)$ such that $\mathcal{E}_t$ defined in (14) is true. Then for any action $a\in\mathcal{A}_t$, we have: $\pi^\star(t,a) < \hat{\kappa}_a(t,a) \leqslant \pi^\star(t,a) + 4/T$.*

Proof. The proof is similar to the proof of Lemma 3. Consider $t\geqslant 1$, $\mathcal{A}_t\subseteq\mathrm{B}(0,1)$, $\mathcal{S}_t\subseteq\mathrm{B}(0,1)$ such that $\mathcal{E}_t$ holds, $a_t\in\mathcal{A}_t$. Then $1/T > \max_{a_t^1\neq a_t^2\in\mathcal{A}_t}\operatorname{diam}\big(\mathcal{S}_t,(a_t^1-a_t^2)/\|a_t^1-a_t^2\|\big)$. Recall that we defined $a_t^{\mathrm{ag}} = \operatorname{argmax}_{a\in\mathcal{A}_t}\langle \mathsf{s}^\star,a\rangle$ and $\hat{a}_t^{\mathrm{ag}} = \operatorname{argmax}_{a\in\mathcal{A}_t}\langle \hat{\mathsf{s}}_t,a\rangle$, $\pi^\star(t,a_t) = \langle \mathsf{s}^\star,a_t^{\mathrm{ag}}\rangle - \langle \mathsf{s}^\star,a_t\rangle$ and $\hat{\pi}(t,a_t) = \langle \hat{\mathsf{s}}_t,\hat{a}_t^{\mathrm{ag}}\rangle - \langle \hat{\mathsf{s}}_t,a_t\rangle + 2/T$, giving

$$\hat{\pi}(t,a_t) \geqslant \langle \hat{\mathsf{s}}_t,a_t^{\mathrm{ag}}\rangle - \langle \hat{\mathsf{s}}_t,a_t\rangle + 2/T\ .$$

Therefore, under $\mathcal{E}_t$, it holds

$$\begin{aligned}
\hat{\pi}(t,a_t) - \pi^\star(t,a_t) &\geqslant \langle \hat{\mathsf{s}}_t,a_t^{\mathrm{ag}}\rangle - \langle \hat{\mathsf{s}}_t,a_t\rangle + 2/T - \langle \mathsf{s}^\star,a_t^{\mathrm{ag}}\rangle + \langle \mathsf{s}^\star,a_t\rangle\\
&= \langle \hat{\mathsf{s}}_t-\mathsf{s}^\star,a_t^{\mathrm{ag}}-a_t\rangle + 2/T\\
&> -\operatorname{diam}\left(\mathcal{S}_t,\frac{a_t^{\mathrm{ag}}-a_t}{\|a_t^{\mathrm{ag}}-a_t\|}\right)\underbrace{\|a_t^{\mathrm{ag}}-a_t\|}_{\leqslant 2} + 2\max_{a_t^1\neq a_t^2\in\mathcal{A}_t}\operatorname{diam}\left(\mathcal{S}_t,\frac{a_t^1-a_t^2}{\|a_t^1-a_t^2\|}\right) \geqslant 0\ . \qquad (26)
\end{aligned}$$

Similarly, since $\langle \mathsf{s}^\star,a_t^{\mathrm{ag}}\rangle \geqslant \langle \mathsf{s}^\star,\hat{a}_t^{\mathrm{ag}}\rangle$, under $\mathcal{E}_t$, we have

$$\begin{aligned}
\hat{\pi}(t,a_t) - \pi^\star(t,a_t) &= \langle \hat{\mathsf{s}}_t,\hat{a}_t^{\mathrm{ag}}-a_t\rangle + 2/T - \langle \mathsf{s}^\star,a_t^{\mathrm{ag}}-a_t\rangle\\
&\leqslant \langle \hat{\mathsf{s}}_t,\hat{a}_t^{\mathrm{ag}}-a_t\rangle - \langle \mathsf{s}^\star,\hat{a}_t^{\mathrm{ag}}-a_t\rangle + 2/T = \langle \hat{\mathsf{s}}_t-\mathsf{s}^\star,\hat{a}_t^{\mathrm{ag}}-a_t\rangle + 2/T\\
&\leqslant \|\hat{a}_t^{\mathrm{ag}}-a\|\cdot\operatorname{diam}\left(\mathcal{S}_t,\frac{\hat{a}_t^{\mathrm{ag}}-a_t}{\|\hat{a}_t^{\mathrm{ag}}-a_t\|}\right) + 2/T \leqslant 4/T\ . \qquad (27)
\end{aligned}$$

Combining (26) and (27) with the definition of $\hat{\kappa}$, we obtain the result. □

Lemma 9. *Let $t \in [T]$ such that $\mathcal{E}_t$ does not hold. Then* $\mathrm{Vol}(\Pi_{\mathsf{V}_t^\perp}\mathcal{S}_t) \geqslant \bar{\delta}^{2d}/d^{2d}$, *where* $\bar{\delta} = 1/16T^2d(d+1)^2$ *and* V_t *is defined in* (31).

Proof. Since $\mathcal{E}_t$ does not hold, there exists a direction $u \in \mathcal{A}_t \subseteq \mathrm{B}(0,1)$ such that $\mathrm{diam}(\mathcal{S}_t, u) \geqslant 1/T \geqslant \bar{\delta}$. For any $u \in \mathsf{V}_t^\perp, \mathrm{diam}(\mathcal{S}_t, u) \geqslant \bar{\delta}$. Lemma 6.3 of Lobel et al. (2018, Section 6) guarantees that $\Pi_{\mathsf{V}_t^\perp}(\mathcal{S}_t)$ contains a k-dimensional ball of radius $\bar{\delta}/k$, where $k \coloneqq \dim(\mathsf{V}_t^\perp)$. Therefore, $\mathrm{Vol}(\Pi_{\mathsf{V}_t^\perp}(\mathcal{S}_t)) \geqslant \bar{\delta}^k\pi^{k/2}/k^k\Gamma(k/2+1)$, where Γ stands for Euler's Gamma function (Smith & Vamanamurthy, 1989). Therefore, we have by definition of Euler's Gamma function: $\mathrm{Vol}(\Pi_{\mathsf{V}_t^\perp}(\mathcal{S}_t)) \geqslant \bar{\delta}^k\pi^{k/2}/k^kk^k \geqslant \bar{\delta}^{2d}/d^{2d}$. □

Lemma 5. *Consider $\mathcal{E}_t$ defined by* (14) *with $(\mathcal{S}_t)_{t\in[T]}$ defined by* `Contextual IPA`. *Then it holds almost surely that*

$$\sum_{t\geqslant 1} \mathbb{1}_{\mathcal{E}_t^c} \leqslant 192 \cdot d\log(dT)\ ,$$

where $\mathcal{E}_t$ is defined by (14).

Proof of Lemma 5. This proof follows the same line as the proof of the main theorem of Lobel et al. (2018, Section 7). Let $t \in [T]$ such that $\mathcal{E}_t$ does not hold. We define w_t as

$$w_t \coloneqq \mathrm{argmax}\left\{\mathrm{diam}\left(\mathcal{S}_t, \frac{a_t^1 - a_t^2}{\|a_t^1 - a_t^2\|}\right) \ :\ \frac{a_t^1 - a_t^2}{\|a_t^1 - a_t^2\|} \text{ such that } a_t^1 \neq a_t^2 \in \mathcal{A}_t\right\}\ , \tag{28}$$

where $\mathcal{S}_t$ is defined in (30). Our goal is to bound the number of steps for which the diameter of $\mathcal{S}_t$ in the direction w_t is strictly superior to $1/T$. Define

$$b_t \coloneqq \mathbb{1}\{\mathrm{diam}(\mathrm{Cyl}(\mathcal{S}_t, \mathsf{V}_t^\perp), w_t) \geqslant 1/T\}\ ,$$

where V_t is defined in (31). Let Π_{E} denote the orthogonal projection onto the subspace $\mathsf{E} \subset \mathbb{R}^d$ and $\mathrm{Vol}(\Pi_{\mathsf{E}}\mathcal{S}) = \mu_{\mathsf{E}}(\Pi_E\mathcal{S})$, where μ_{E} is the Lebesgue measure of $\Pi_E\mathcal{S}$, well-defined for $\Pi_E\mathcal{S}$ being a convex body. Setting $\bar{\delta} = T^{-2}/16d(d+1)^2$, we can apply the projected Grünbaum lemma of Lobel et al. (2018, Lemma 7.1) to obtain that

$$\mathrm{Vol}(\Pi_{\mathsf{V}_t^\perp}\mathcal{S}_{t+1}) \leqslant \left(1 - \mathrm{e}^{-2}\right)^{b_t}\mathrm{Vol}(\Pi_{\mathsf{V}_t^\perp}\mathcal{S}_t)\ .$$

By definition of V_t in (31), we have for any $u \in \mathsf{V}_t$, $\mathrm{diam}(\mathcal{S}_t, u) \geqslant \bar{\delta}$. Therefore, by definition of $\mathcal{S}_{t+1}$ in (30), the directional Grünbaum Theorem (Lobel et al., 2018, Theorem 5.3) guarantees that we have: $\mathrm{diam}(\mathcal{S}_{t+1}, u) \geqslant \bar{\delta}/(d+1)$. Note that for $\mathcal{S}_t$ being a convex body, Lemma 11 ensures that $\mathcal{S}_{t+1}$ remains a convex body.

If $J_{t+1} = 0$, where J_t is defined in (32), then $\mathsf{V}_{t+1}^\perp = \mathsf{V}_t^\perp$, and we have $\mathrm{Vol}(\Pi_{\mathsf{V}_{t+1}^\perp}\mathcal{S}_{t+1}) = \mathrm{Vol}(\Pi_{\mathsf{V}_t^\perp}\mathcal{S}_{t+1})$.

Otherwise, let $i \in [J_{t+1} - 1]$ and $v \in \mathsf{V}_t^{(i),\perp}$ such that $\mathsf{V}_t^{(i+1)} = \mathsf{V}_t^{(i)} \cup \{v\}$. We have $\mathsf{V}_t^{(i),\perp} \cap \mathrm{span}(v)^\perp = \{x \in \mathsf{V}_t^{(i),\perp} \ :\ v^\intercal x = 0\} \subseteq \mathsf{V}_t^{(i+1),\perp} \subseteq \mathsf{V}_t^{(i),\perp}$ and $\dim(\mathsf{V}_t^{(i+1),\perp}) = \dim(\mathsf{V}_t^{(i),\perp}) - 1$. Then, applying the Cylindrification Lemma from Lobel et al. (2018, Lemma 6.1) we obtain

$$\mathrm{Vol}(\Pi_{\mathsf{V}_t^{(i+1),\perp}}\mathcal{S}_{t+1}) \leqslant \frac{d(d+1)^2}{\bar{\delta}}\mathrm{Vol}(\Pi_{\mathsf{V}_t^{(i),\perp}}\mathcal{S}_{t+1})\ .$$

If $J_{t+1} = r$, the volume can blow up by at most $\left(d(d+1)^2/\bar{\delta}\right)^r$. In particular, since the initial volume is bounded by $\mathrm{Vol}(\mathrm{B}(0,1)) \leqslant 8\pi^2/15 \leqslant 6$ (Smith & Vamanamurthy, 1989), then by Lemma 9, we obtain: $\bar{\delta}^{2d}/d^{2d} \leqslant \mathrm{Vol}(\Pi_{L_t}\mathcal{S}_t) \leqslant 6 \cdot \left(d(d+1)^2/\bar{\delta}\right)^d \cdot \left(1 - 1/\mathrm{e}^2\right)^{\sum_{t=1}^T b_t}$. Therefore, applying the logarithm function, we obtain

$$\sum_{t=1}^T b_t \leqslant -1/\log(1-\mathrm{e}^{-2})(\log 6 + 2d\log(16) + 5d\log(d) + 6d\log(d+1) + 4d\log(T))\ ,$$

giving, since $-1/\log(1-\mathrm{e}^{-2}) < 16$: $\sum_{t=1}^T \mathbb{1}\{\mathrm{diam}(\mathrm{Cyl}(\mathcal{S}_t, L_t), w_t) \geqslant 1/T\} \leqslant 192d\log(dT)$. Therefore, lemma 12 ensures that $\mathrm{diam}(\mathcal{S}_t, w_t) \leqslant \mathrm{diam}(\mathrm{Cyl}(\mathcal{S}_t, \mathsf{V}_t^\perp), w_t)$ and we get

$$\sum_{t=1}^T \mathbb{1}\left\{\max_{s\in\mathcal{S}_t} |\langle s, w_t\rangle| \geqslant 1/T\right\} \leqslant 192\, d\log(dT)\ ,$$

which concludes the proof by definition of w_t in (28) and $\mathcal{E}_t$ in (14). □

Lemma 6. *Consider $(\mathcal{E}_t)_{t\in[T]}$ defined by (14) with $(\mathcal{S}_t)_{t\in[T]}$ defined by* `Contextual IPA`*. Let I_T and $(\varepsilon_t^{corrupt})_{t\in[T]}$ as defined in (15) and (16) and $t \in [T]$ such that $\mathcal{E}_t$ is true. Then $|\varepsilon_t^{corrupt}| \leqslant 4/T$ and $\sum_{t\in I_T} |\varepsilon_t^{corrupt}| := C_{corrupt} \leqslant 4$.*

Proof of Lemma 6. Consider $t \in I_T$, $a \in \mathcal{A}_t$: by (15), $\mathcal{E}_t$ is true. Consider $\varepsilon_t^{\text{corrupt}}$ as defined in (16)

Using Lemma 4 gives $|\varepsilon_t^{\text{corrupt}}| \leqslant 4/T$, and summing over all the iterations $t \in I_T$, since $|I_T| \leqslant T$, we have $\sum_{t\in I_t} |\varepsilon_t^{\text{corrupt}}| \leqslant 4 \cdot 1/T \cdot |I_T| \leqslant 4$. □

Lemma 10. *For any $t \in [T]$, $\varepsilon > 0$ and action $a \in \mathcal{A}_t$, set $\pi^{\star,\varepsilon}(t, a) := \max_{a_t^{\text{ag}}\in\mathcal{A}_t} \langle \mathsf{s}^\star, a_t^{\text{ag}}\rangle - \langle \mathsf{s}^\star, a\rangle + \varepsilon$, and define the incentive function $\kappa_a^{\star,\varepsilon}(t, a') = \mathbb{1}_a(a')\pi^{\star,\varepsilon}(t, a)$ for any $a' \in \mathcal{A}_t$. Then $A_t = a$, where A_t is defined by (8).*

Proof of Lemma 10. Note that for any $a' \in \mathcal{A}_t, a' \neq a$, $\langle \mathsf{s}^\star, a'\rangle + \kappa_a^{\star,\varepsilon}(t, a') < \langle \mathsf{s}^\star, a\rangle + \kappa_a^{\star,\varepsilon}(t, a)$ and, as a result, $A_t = a$. □

D.2 Proof of Theorem 2

Proof of Theorem 2. Denote for any $t \in [T]$, $a_t^{\text{pr}} := \text{argmax}_{a\in\mathcal{A}_t}\{\langle\theta^\star, a\rangle - \pi^\star(t, a)\} = \text{argmax}_{a\in\mathcal{A}_t}\langle\theta^\star + \mathsf{s}^\star, a\rangle$, $a_t^{\text{ag}} := \text{argmax}_{a\in\mathcal{A}_t}\langle \mathsf{s}^\star, a\rangle$ and $\hat{a}_t^{\text{ag}} = \text{argmax}_{a\in\mathcal{A}_t}\langle \hat{\mathsf{s}}_t, a\rangle$. Note that if $\mathcal{E}_t$ holds, for any $a_t^1, a_t^2 \in \mathcal{A}_t, a_t^1 \neq a_t^2$, **H**2 gives that if

$$\text{diam}\left(\mathcal{S}_t, \frac{a_t^1 - a_t^2}{\|a_t^1 - a_t^2\|}\right) < \frac{1}{T} \text{ , then } \max_{s\in\mathcal{S}_t}\langle s, a_t^1 - a_t^2\rangle = \max_{s\in\mathcal{S}_t}\left\langle s, \frac{a_t^1 - a_t^2}{\|a_t^1 - a_t^2\|}\right\rangle \underbrace{\|a_t^1 - a_t^2\|}_{\leqslant 2} < 2 \cdot \frac{1}{T} ,$$

and therefore $\text{diam}\big(\mathcal{S}_t, a_t^1 - a_t^2\big) < 2/T$.

If $\mathcal{E}_t$ does not hold, the incentive function proposed in Algorithm 5 is given by $\kappa(t, a) = 3 \cdot \mathbb{1}_{a_t^1}(a) + (3 + \langle \hat{\mathsf{s}}_t, a_t^1 - a_t^2\rangle) \cdot \mathbb{1}_{a_t^2}(a)$. Therefore: $\kappa(t, a_t^1) = 3, \kappa(t, a_t^2) = 3 + \langle \hat{\mathsf{s}}_t, a_t^1 - a_t^2\rangle \leqslant 5$. When $\mathcal{E}_t$ holds, the incentive function proposed in `Contextual IPA` is defined as $\kappa(t, a) = \mathbb{1}_{a_t^{\text{Rec}}}(a)\hat{\pi}(t, a_t^{\text{Rec}}) \leqslant 2$.

For any $t \in [T]$, we define the instantaneous regret reg_t at t as

$$\text{reg}_t := \mu_t^\star - (X_{A_t}(t) - \kappa(t, A_t)) ,$$

where $\mu_t^\star$ is defined in (10) and we decompose the regret into two terms, making use of the Cauchy-Schwarz inequality as well as **H**2 to obtain

$$\begin{aligned}
\mathfrak{R}(T) &= \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}_{\{\mathcal{E}_t\}}\text{reg}_t + \sum_{t=1}^T \mathbb{1}_{\{\mathcal{E}_t^c\}}\text{reg}_t\right] \\
&\leqslant \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}_{\{\mathcal{E}_t\}}\text{reg}_t\right] + \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}_{\{\mathcal{E}_t^c\}}\big(\max_{a\in\mathcal{A}_t}\langle\theta^\star + \mathsf{s}^\star, a\rangle - \max_{a'\in\mathcal{A}_t}\langle \mathsf{s}^\star, a'\rangle - \langle\theta^\star, A_t\rangle + \kappa(t, A_t)\big)\right] \\
&\leqslant \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}_{\{\mathcal{E}_t\}}\text{reg}_t\right] + \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}_{\{\mathcal{E}_t^c\}}\big(\max_{a\in\mathcal{A}_t}\{\underbrace{\langle\theta^\star, a\rangle}_{\leqslant 1} + \underbrace{\langle \mathsf{s}^\star, a - a_t^{\text{ag}}\rangle}_{\leqslant 0}\} \underbrace{-\langle\theta^\star, A_t\rangle}_{\leqslant 1} + \underbrace{\kappa(t, A_t))}_{\leqslant 5}\big)\right] \\
&\leqslant \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}_{\{\mathcal{E}_t\}}\text{reg}_t\right] + 7\,\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}_{\{\mathcal{E}_t^c\}}\right] .
\end{aligned}$$

Using Lemma 5, we can bound the second term

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}_{\{\mathcal{E}_t^c\}}\right] \leqslant 192\, d \log(dT) .$$

Now we bound the first term. Working on steps t such that $\mathcal{E}_t$ is true with incentive function $\hat{\kappa}_{a_t^{\text{Rec}}}(t, \cdot) = \mathbb{1}_{a_t^{\text{Rec}}}(\cdot)\hat{\pi}(t, a_t^{\text{Rec}})$, Lemma 4 guarantees that $A_t = a_t^{\text{Rec}}$, and we have

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}_{\{\mathcal{E}_t\}}\text{reg}_t\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{E}_t\}} \left(\max_{a\in\mathcal{A}_t} \{\langle \theta^\star + \mathsf{s}^\star, a\rangle - \max_{a'\in\mathcal{A}_t} \langle \mathsf{s}^\star, a'\rangle\} - \left\{ \langle \theta^\star, a_t^{\mathrm{Rec}}\rangle - \left(\langle \hat{\mathsf{s}}_t, \hat{a}_t^{\mathrm{ag}}\rangle - \langle \hat{\mathsf{s}}_t, a_t^{\mathrm{Rec}}\rangle + \frac{2}{T} \right) \right\} \right) \right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{E}_t\}} \big(\langle \theta^\star, a_t^{\mathrm{pr}}\rangle + \langle \mathsf{s}^\star, a_t^{\mathrm{pr}}\rangle - \langle \mathsf{s}^\star, a_t^{\mathrm{ag}}\rangle - \langle \theta^\star, a_t^{\mathrm{Rec}}\rangle + \langle \hat{\mathsf{s}}_t, \hat{a}_t^{\mathrm{ag}}\rangle - \langle \hat{\mathsf{s}}_t, a_t^{\mathrm{Rec}}\rangle \big) + \sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{E}_t\}} \frac{2}{T} \right] \\
&\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{E}_t\}} \big(\langle \theta^\star, a_t^{\mathrm{pr}}\rangle + \langle \mathsf{s}^\star, a_t^{\mathrm{pr}}\rangle - \langle \mathsf{s}^\star, a_t^{\mathrm{ag}}\rangle - \langle \theta^\star, a_t^{\mathrm{Rec}}\rangle \big) \right] + 2T\frac{1}{T} \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{E}_t\}} \big(-\langle \mathsf{s}^\star, a_t^{\mathrm{Rec}}\rangle + \langle \mathsf{s}^\star, a_t^{\mathrm{ag}}\rangle + \langle \hat{\mathsf{s}}_t, \hat{a}_t^{\mathrm{ag}}\rangle - \langle \hat{\mathsf{s}}_t, a_t^{\mathrm{Rec}}\rangle + \langle \mathsf{s}^\star, a_t^{\mathrm{Rec}}\rangle - \langle \mathsf{s}^\star, a_t^{\mathrm{ag}}\rangle \big) \right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{E}_t\}} \big(\langle \theta^\star, a_t^{\mathrm{pr}}\rangle + \langle \mathsf{s}^\star, a_t^{\mathrm{pr}}\rangle - \langle \mathsf{s}^\star, a_t^{\mathrm{ag}}\rangle - \langle \theta^\star, a_t^{\mathrm{Rec}}\rangle - \langle \mathsf{s}^\star, a_t^{\mathrm{Rec}}\rangle + \langle \mathsf{s}^\star, a_t^{\mathrm{ag}}\rangle \big) \right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{E}_t\}} \big(\langle \hat{\mathsf{s}}_t, \hat{a}_t^{\mathrm{ag}}\rangle + \langle \mathsf{s}^\star - \hat{\mathsf{s}}_t, a_t^{\mathrm{Rec}}\rangle - \langle \mathsf{s}^\star, a_t^{\mathrm{ag}}\rangle \big) \right] + 2 \,.
\end{aligned}
$$

Since $\langle \mathsf{s}^\star, a_t^{\mathrm{ag}}\rangle \geqslant \langle \mathsf{s}^\star, \hat{a}_t^{\mathrm{ag}}\rangle$, we have $-\langle \mathsf{s}^\star, a_t^{\mathrm{ag}}\rangle \leqslant -\langle \mathsf{s}^\star, \hat{a}_t^{\mathrm{ag}}\rangle$. Therefore

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{E}_t\}} \big(\langle \hat{\mathsf{s}}_t, \hat{a}_t^{\mathrm{ag}}\rangle + \langle \mathsf{s}^\star - \hat{\mathsf{s}}_t, a_t^{\mathrm{Rec}}\rangle - \langle \mathsf{s}^\star, a_t^{\mathrm{ag}}\rangle \big) \right] \\
&\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{E}_t\}} \big(\langle \hat{\mathsf{s}}_t, \hat{a}_t^{\mathrm{ag}}\rangle + \langle \mathsf{s}^\star - \hat{\mathsf{s}}_t, a_t^{\mathrm{Rec}}\rangle - \langle \mathsf{s}^\star, \hat{a}_t^{\mathrm{ag}}\rangle \big) \right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{E}_t\}} \big(\langle \mathsf{s}^\star - \hat{\mathsf{s}}_t, a_t^{\mathrm{Rec}}\rangle - \langle \mathsf{s}^\star - \hat{\mathsf{s}}_t, \hat{a}_t^{\mathrm{ag}}\rangle \big) \right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{E}_t\}} \big(\langle \mathsf{s}^\star - \hat{\mathsf{s}}_t, a_t^{\mathrm{Rec}} - \hat{a}_t^{\mathrm{ag}}\rangle \big) \right] \\
&\leqslant \sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{E}_t\}} \mathbb{1}\{a_t^{\mathrm{Rec}} \neq \hat{a}_t^{\mathrm{ag}}\} \underbrace{\|a_t^{\mathrm{Rec}} - \hat{a}_t^{\mathrm{ag}}\|}_{\leqslant 2} \operatorname{diam}\left(\mathcal{S}_t, \frac{a_t^{\mathrm{Rec}} - \hat{a}_t^{\mathrm{ag}}}{\|a_t^{\mathrm{Rec}} - \hat{a}_t^{\mathrm{ag}}\|} \right) < 2T \cdot \frac{1}{T} = 2 \,,
\end{aligned}
$$

and plugging this inequality gives

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{E}_t\}} \mathrm{reg}_t\right] &\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{\mathcal{E}_t\}} \big(\langle \theta^\star + \mathsf{s}^\star, a_t^{\mathrm{pr}}\rangle - \langle \mathsf{s}^\star, a_t^{\mathrm{ag}}\rangle - \langle \theta^\star + \mathsf{s}^\star, a_t^{\mathrm{Rec}}\rangle + \langle \mathsf{s}^\star, a_t^{\mathrm{ag}}\rangle \big) \right] + 4 \\
&= \mathbb{E}\left[\sum_{t\in I_T} \max_{a\in\mathcal{A}_t} \{\langle \theta^\star + \mathsf{s}^\star, a\rangle - \max_{a'\in\mathcal{A}_t} \langle \mathsf{s}^\star, a'\rangle\}\right] - \mathbb{E}\left[\sum_{t\in I_T} \langle \theta^\star + \mathsf{s}^\star, a_t^{\mathrm{Rec}}\rangle - \max_{a\in\mathcal{A}_t} \langle \mathsf{s}^\star, a\rangle\right] + 4 \\
&= R^{\mathrm{corrupt}}_{\texttt{CtxAlg}}(I_T, \varepsilon^{\mathrm{corrupt}}_{I_T}) + 4 \,,
\end{aligned}
$$

where $R^{\mathrm{corrupt}}_{\texttt{Contextual IPA}}$ is defined in (17) with $\pi^\star(t,a) = \max_{a'\in\mathcal{A}_t} \langle \mathsf{s}^\star, a'\rangle - \langle \mathsf{s}, a\rangle$. Plugging all the terms together gives the following upper-bound for the regret:

$$
\mathfrak{R}(T) \leqslant 1344\, d \log(dT) + 4 + R^{\mathrm{corrupt}}_{\texttt{CtxAlg}}(I_T, \varepsilon^{\mathrm{corrupt}}_{I_T}) \,.
$$

□

Proof of Corollary 2. Here, Lemma 6 guarantees that $C_{\mathrm{corrupt}} = 4$. With the setup of He et al. (2022), we take $L = 1, S = 1, R = 1$, and $\alpha = \sqrt{d}/4$.

Choose $\lambda = 1$ as a regularization parameter and $\delta = 1/T$ as a confidence level. Using $\texttt{CW} - \texttt{OFUL}$ proposed in He et al. (2022) as a subroutine robust to corruption in the stochastic linear case, since our subroutine is fed with the same reward r' as in their model while $R^{\text{corrupt}}_{\texttt{CtxAlg}}(I_T, \varepsilon^{\text{corrupt}}_{I_T})$ is defined compared to the true reward r, we can use the result provided in He et al. (2022, Theorem 4.2) to bound $R^{\text{corrupt}}_{\texttt{CtxAlg}}(I_T, \varepsilon^{\text{corrupt}}_{I_T})$ with probability $1 - 1/T$ and use the fact that our instantaneous regret is always bounded by 7 as it is shown in the proof of Theorem 2 to get in expectation for some universal constant $B > 0$

$$
\begin{aligned}
R^{\text{corrupt}}_{\texttt{CtxAlg}}(I_T, \varepsilon^{\text{corrupt}}_{I_T}) &\leqslant (1-\delta) B \left(2d\sqrt{T} \log\left(\frac{1+T}{\delta}\right) + \sqrt{d\lambda T}\sqrt{\log(1+T)} + 4d\sqrt{\log\left(\frac{1+T}{\delta}\right)^3} \right) + 7T\delta \\
&\leqslant \left(1 - \frac{1}{T}\right) B \left(2d\sqrt{T} \log\left(\frac{1+T}{\frac{1}{T}}\right) + \sqrt{dT}\sqrt{\log\left(\frac{1+T}{\frac{1}{T}}\right)} + 4d\sqrt{\log\left(\frac{1+T}{\frac{1}{T}}\right)^3} \right) + 7T\frac{1}{T} \\
&\leqslant \left(1 - \frac{1}{T}\right) B \left(2d\sqrt{T} \log\left(T + T^2\right) + \sqrt{dT}\sqrt{\log(T+T^2)} + 4d\left(\log\left(T+T^2\right)\right)^{\frac{3}{2}} \right) + 7 \,,
\end{aligned}
$$

therefore, since $C_{\text{corrupt}} \leqslant 4$, there exists a constant $C_{\texttt{CtxAlg}}$ such that

$$
R^{\text{corrupt}}_{\texttt{CtxAlg}}(I_T, \varepsilon^{\text{corrupt}}_{I_T}) \leqslant 7 + C_{\texttt{CtxAlg}} d\sqrt{T} \log T \,.
$$

Finally plugging this term in the bound from Theorem 2 and integrating the 3 factor in constant B gives the result

$$
\mathfrak{R}(T) \leqslant 11 + 1344\, d \log(dT) + C_{\texttt{CtxAlg}} d\sqrt{T} \log T \,.
$$

□

E Algorithms

E.1 UCB subroutine

We present the `Binary search` subroutine and the `UCB` algorithm that we use as a subroutine in `IPA` as formulated in Lattimore & Szepesvári (2020, Algorithm 3).

Algorithm 3 `Binary Search Subroutine`

Input: action a, N_T
Initialize: $\underline{\pi}_a(0), \overline{\pi}_a(0) = 0, 1$
for $t = 1, \dots, N_T$ **do**
  $\pi_a^{\text{mid}}(t-1) = \frac{\overline{\pi}_a(t-1) + \underline{\pi}_a(t-1)}{2}$
  Propose incentive $\pi_a^{\text{mid}}(t-1)$ on arm a
  If $A_{t-1} = a$ **then** $\overline{\pi}_a(t) = \pi_a^{\text{mid}}(t-1)$ and $\underline{\pi}_a(t) = \underline{\pi}_a(t-1)$ **else** $\underline{\pi}_a(t) = \pi_a^{\text{mid}}(t)$ and $\overline{\pi}_a(t) = \overline{\pi}_a(t-1)$
end for
Return $\underline{\pi}_a(N_T), \overline{\pi}_a(N_T)$

E.2 Projected volume algorithm

We present the `Projected volume algorithm` from Lobel et al. (2018) that we use as a subroutine in `Contextual IPA`. For any horizon T and $0 < \bar{\delta} < T^{-2}/(16d(d+1)^2)$, this algorithm defines recursively a sequence $(\mathcal{S}_t, \mathsf{V}_t)_{t\in[T]}$ such that $(\mathcal{S}_t)_{t\in[T]}$ is a sequence of decreasing subsets (for the inclusion) of $\mathrm{B}(0,1)$ including $\mathsf{s}^\star$ and $(\mathsf{V}_t)_{t\in[T]}$ is a sequence of increasing sets of $\mathbb{R}^d$ containing orthogonal directions $\{v_i\}_{i\in[n]}$ along which the principal has a good knowledge of $\langle \mathsf{s}^\star, v_i \rangle$.

The main ingredient in Lobel et al. (2018) allowing low regret is *cylindrification*. Given a compact convex set $\mathcal{S} \subseteq \mathbb{R}^d$, $\mathsf{V} = \{v_1, \dots, v_n\}$, and $\Pi_{\mathsf{V}^\perp}(\mathcal{S})$ being the orthogonal projection of $\mathcal{S}$ onto $\mathsf{V}^\perp$, we define the cylindrification of S on V as

$$
\mathrm{Cyl}(S, \mathsf{V}) \coloneqq \Pi_L(\mathcal{S}) + \Pi_{\text{span}(v_1)}(\mathcal{S}) + \ldots + \Pi_{\text{span}(v_n)}(\mathcal{S})
$$

Algorithm 4 `UCB Subroutine`

Input: Set of arms K, horizon T
Initialize: For any arm $a \in [K]$, set $\hat{\mu}_a \coloneqq 0$, $T_a \coloneqq 0$
for $1 \leqslant t \leqslant K$: **do**
  Pull arm $a = t$
  Update $\hat{\mu}_a = X_a(t)$, $T_a(t) = 1$
end for
for $t \geqslant K+1$ **do**
  Pull arm $a_{\max} \in \operatorname{argmax}_{a\in[K]}\left\{\hat{\mu}_a(t-1) + 2\sqrt{\frac{\log T}{T_a(t-1)}}\right\}$
  Update $T_a(t) = T_a(t-1)+1$, $\hat{\mu}_a(t) = \frac{1}{T_a(t)}(T_a(t-1)\hat{\mu}_a(t-1) + X_a(t))$
end for

$$= \left\{ x + \sum_{i=1}^{n} y_i v_i \colon x \in \Pi_L(\mathcal{S}), \min_{s\in\mathcal{S}} \langle s, v_i\rangle \leqslant y_i \leqslant \max_{s\in\mathcal{S}} \langle s, v_i\rangle \right\}.$$

At iteration t, given $(\mathcal{S}_t, \mathsf{V}_t)$, we define the estimate $\hat{\mathsf{s}}_t$ of $\mathsf{s}^\star$ as the centroid of $\mathrm{Cyl}(\mathcal{S}_t, \mathsf{V}_t)$:

$$\hat{\mathsf{s}}_t \coloneqq \frac{1}{\mathrm{Vol}(\mathrm{Cyl}(\mathcal{S}_t, \mathsf{V}_t))} \int_{\mathrm{Cyl}(\mathcal{S}_t, \mathsf{V}_t)} x \mathrm{d}x \,, \tag{29}$$

Note that the iterative construction of $\mathcal{S}_t$ described below together with Lemma 11 guarantees that $\mathcal{S}_t$ is always a convex body, making $\mathrm{Vol}(\mathcal{S}_t)$ well-defined as the Lebesgue measure of $\mathcal{S}_t$ in dimension d and $\mathrm{Vol}(\mathcal{S}_t) > 0$. Combined with Lemma 12, it guarantees that $\mathrm{Vol}(\mathrm{Cyl}(\mathcal{S}_t, \mathsf{V}_t))$ is well-defined, and $\mathrm{Vol}(\mathrm{Cyl}(\mathcal{S}_t, \mathsf{V}_t)) > 0$. Therefore, (29) is well-defined.

At iteration t, given $(\mathcal{S}_t, \mathsf{V}_t)$ and two actions $a_t^1, a_t^2 \in \mathcal{A}_t$, recall that we defined w_t as

$$w_t \coloneqq \operatorname{argmax}\left\{ \mathrm{diam}\left(\mathcal{S}_t, \frac{a_t^1 - a_t^2}{\|a_t^1 - a_t^2\|}\right) \,:\, \frac{a_t^1 - a_t^2}{\|a_t^1 - a_t^2\|} \text{ such that } a_t^1 \neq a_t^2 \in \mathcal{A}_t \right\}.$$

Then, the principal offers the incentive function $\kappa(t,a) = 5 \cdot \mathbb{1}_{a_t^1}(a) + (5 + \langle \hat{\mathsf{s}}, a_t^1 - a_t^2\rangle) \cdot \mathbb{1}_{a_t^2}(a)$. Recall that $\kappa(t,a)$ is always bounded by 5 and that a_t^1, a_t^2 are chosen such that $\langle \hat{\mathsf{s}}_t, a_t^1 - a_t^2\rangle \geqslant 0$, ensuring that we either have $A_t = a_t^1$ or $A_t = a_t^2$.

If $A_t = a_t^1$, it means that $\langle \mathsf{s}^\star, w_t\rangle \geqslant 0$ and we update $\mathcal{S}_{t+1} = \mathcal{S}_t \cap \{\mathsf{s} | \langle \mathsf{s}, w_t\rangle \geqslant x_t\}$. Otherwise, if $A_t = a_t^2$, it means that $\langle \mathsf{s}^\star, w_t\rangle \leqslant 0$ and we update $\mathcal{S}_{t+1} = \mathcal{S}_t \cap \{\mathsf{s} | \langle \mathsf{s}, w_t\rangle \leqslant x_t\}$. This defines the subset $\mathcal{S}_{t+1}$ such that $\mathsf{s}^\star \in \mathcal{S}_{t+1}$:

$$\mathcal{S}_{t+1} = \mathcal{S}_t \cap (\mathbb{1}_{a_t^1}(A_t)\{\mathsf{s} | \langle \mathsf{s}, w_t\rangle \geqslant x_t\} + \mathbb{1}_{a_t^2}(A_t)\{\mathsf{s} | \langle \mathsf{s}, w_t\rangle \leqslant x_t\}) \tag{30}$$

Regarding the subspace V_{t+1}, we consider $\mathsf{V}_{t+1} = \mathsf{V}_{t+1}^{J_{t+1}}$ where $(\mathsf{V}_{t+1}^i)_{i\in\mathbb{N}}$ is defined as

$$\begin{aligned}
\mathsf{V}_{t+1}^{(0)} &\coloneqq \mathsf{V}_t \\
V_{t+1}^{(i+1)} &\coloneqq \begin{cases} V_{t+1}^{(i)} \cup \{v\} \text{ if } \exists\, v \in (\mathsf{V}_{t+1}^{(i)})^\perp \,:\, \mathrm{diam}(\mathcal{S}_{t+1}, v) \leqslant \bar{\delta} \\ \mathsf{V}_{t+1}^{(i)} \text{ otherwise} \end{cases} \qquad (31) \\
J_{t+1} &\coloneqq \min\{i \colon \text{ it does not exist } v \in (\mathsf{V}_{t+1}^{(i)})^\perp \text{ such that } \mathrm{diam}(\mathcal{S}_{t+1}, v) \leqslant \bar{\delta}\}\,, \qquad (32)
\end{aligned}$$

which exists since if it does not exist i such that it does not exist $v \in (\mathsf{V}_{t+1}^{(i)})^\perp$ such that $\mathrm{diam}(\mathcal{S}_{t+1}, v) \geqslant \bar{\delta}$, we would have $\dim(\mathrm{span}(\mathsf{V}_t^{(i+1)})) = \dim(\mathrm{span}(\mathsf{V}_t^{(i)})) + 1$ for any $i \in \mathbb{N}$, which would imply a contradiction.

In what follows, we provide technical results ensuring that $\mathcal{S}_t$ is a convex body for any $t \in [T]$ and $\mathcal{S}_t \subseteq \mathrm{Cyl}(\mathcal{S}_t, \mathsf{V}_t)$, which implies that $\mathrm{Cyl}(\mathcal{S}_t, \mathsf{V}_t)$ has non-empty interior.

Lemma 11. *Let $\mathcal{S}$ be a convex body in $\mathbb{R}^d$ and s be a point in the interior of $\mathcal{S}$: $\mathsf{s} \in \mathring{\mathcal{S}}$. Let G be the half-space defined by $\mathsf{G} \coloneqq \{x \in \mathbb{R}^d \colon \langle h^\star, x\rangle \geqslant 0\}$, for some $h^\star \in \mathbb{R}^d$. Suppose that $\mathsf{s} \in \mathsf{G}$. Then $\mathcal{S} \cap \mathsf{G}$ is a convex body.*

Proof. Note that we only need to show that $\mathcal{S} \cap \mathsf{G}$ has non-empty interior since the intersection of two compact convex sets is compact and convex.

The only case that we consider is $\mathsf{s} \in \mathsf{H}$ where H is the hyperplane defined by $\mathsf{H} \coloneqq \{x \in \mathbb{R}^d \colon \langle h^\star, x\rangle = 0\}$. The other case simply follows from the fact that the intersection of two open sets is also open.

Since $\mathcal{S}$ is a convex body and $\mathsf{s} \in \mathring{\mathcal{S}}$, there exists a ball $\mathrm{B}(\mathsf{s}, r)$ centered in s with $r > 0$, such that $\mathrm{B}(\mathsf{s}, r) \subseteq \mathring{\mathcal{S}}$. For any $x \in \mathrm{B}(\mathsf{s}, r) \cap \mathsf{G}$ is equivalent to $\langle h^\star, x - \mathsf{s}\rangle \geqslant 0$ since $\langle h^\star, \mathsf{s}\rangle = 0$ and $\|x - \mathsf{s}\| \leqslant r$.

Now we define $y_0 = \mathsf{s} + rh^\star/(2\|h^\star\|)$ and consider the ball $\mathrm{B}(y_0, r/2)$. For any $y \in \mathrm{B}(y_0, r/2)$, we can write $y = y_0 + \tilde{y}$ with $\|\tilde{y}\| \leqslant r/2$. Using $\langle h^\star, \mathsf{s}\rangle = 0$, we have $\|y - \mathsf{s}\| \leqslant \|y_0 - \mathsf{s}\| + \|\tilde{y}\| \leqslant r$ and $\langle h^\star, y\rangle = \langle h^\star, y_0\rangle + \langle h^\star, \tilde{y}\rangle$ with $\langle h^\star, y_0\rangle = r/2$ and $\langle h^\star, \tilde{y}\rangle \geqslant -\|h^\star\| \cdot \|\tilde{y}\| \geqslant -1 \cdot r/2 = -r/2$. Therefore $\langle h^\star, y\rangle \geqslant 0$ and we obtain $y \in \mathrm{B}(\mathsf{s}, r) \cap \mathsf{G}$, which gives $\mathrm{B}(y_0, r/2) \subseteq \mathrm{B}(\mathsf{s}, r) \cap \mathsf{G}$.

□

Lemma 12. *Given a convex body $\mathcal{S} \subseteq \mathbb{R}^d$ and a set of orthonormal vectors $\mathsf{V} = \{v_1, \ldots, v_n\} \subseteq \mathbb{R}^d$, let $\Pi_{\mathsf{V}^\perp}(\mathcal{S})$ be the projection of $\mathcal{S}$ on the subspace $\mathsf{V}^\perp = \{x \in \mathbb{R}^d \mid \langle x, v_i\rangle = 0\}$. Define the cylindrification of S onto V as*

$$\begin{aligned}\mathrm{Cyl}(S, \mathsf{V}) &\coloneqq \Pi_{\mathsf{V}^\perp}(\mathcal{S}) + \Pi_{\mathit{span}(v_1)}(\mathcal{S}) + \ldots + \Pi_{\mathit{span}(v_n)}(\mathcal{S}) \\ &= \left\{ x + \sum_{i=1}^{n} y_i v_i | x \in \Pi_L(\mathcal{S}), \min_{s \in \mathcal{S}} \langle s, v_i\rangle \leqslant y_i \leqslant \max_{s \in \mathcal{S}} \langle s, v_i\rangle \right\}.\end{aligned}$$

Then it holds that $\mathcal{S} \subseteq \mathrm{Cyl}(\mathcal{S}, \mathsf{V})$.

Proof. Define $\Pi_{\mathsf{V}^\perp}$ as the orthogonal projector on $\mathsf{V}^\perp$. Then we have for any $s \in \mathcal{S}$

$$s = \Pi_{\mathsf{V}^\perp} s + (I - \Pi_{\mathsf{V}^\perp}) s \,,$$

where $(I - \Pi_{\mathsf{V}^\perp})$ is an orthogonal projector on the space span$(\{v_1, \ldots, v_n\})$, thus $(I - \Pi_{\mathsf{V}^\perp})s = \sum_{i=1}^n y_i v_i$ for $y_i = \langle s, v_i\rangle$. This decomposition allows us to conclude. □

Algorithm 5 `Projected Volume`

Input: $T, \bar{\delta}, \mathcal{S}_t$ such that $\mathrm{diam}(\mathcal{S}_t) \geqslant \bar{\delta}, \mathsf{V}_t, a_t^1, a_t^2$
Compute $\mathrm{Cyl}(\mathcal{S}_t, \mathsf{V}_t)$ and its centroid $\hat{\mathsf{s}}_t$, $w_t = (a_t^1 - a_t^2)/\|a_t^1 - a_t^2\|$, $x_t = \langle \hat{\mathsf{s}}_t, w_t\rangle$, $\bar{\delta} \in \left(0, 1/16d(d+1)^2T^2\right)$
Propose an incentive function $\kappa(t, a) = 5 \cdot \mathbb{1}_{a_t^1}(a) + (5 + \langle \hat{\mathsf{s}}_t, a_t^1 - a_t^2\rangle) \cdot \mathbb{1}_{a_t^2}(a)$
if $A_t = a_t^1$ **then**
    $S_{t+1} = \mathcal{S}_t \cap \{s | \langle s, w_t\rangle \geqslant x_t\}$
else
    $S_{t+1} = \mathcal{S}_t \cap \{s | \langle s, w_t\rangle \leqslant x_t\}$
end if
Let $\mathsf{V}_{t+1} = \mathsf{V}_t$
if $\exists v \perp \mathsf{V}_{t+1}$ such that $\mathrm{diam}(\mathcal{S}_{t+1}, v) \leqslant \bar{\delta}$ **then**
    add v to V_t
    Repeat this step as many times as necessary
end if
Output: $S_{t+1}, \mathsf{V}_{t+1}$.

F Lower Bound

Proof of Proposition 1. Suppose that the principal was to know $(s_a)_{a \in [K]}$. For any incentive $\pi(t)$ offered on action a_t at round t, the agent selects his action following (1): $A_t = \mathrm{argmax}_{a \in [K]} \mathsf{s}_a + \mathbb{1}_{a_t}(a)\pi(t)$. The principal's expected reward is

$$\theta_{A_t} - \mathbb{1}_{a_t}(A_t)\pi(t) \leqslant \theta_{A_t} - \pi^\star_{A_t} = \mu_a \,,$$

by definition of $\pi_a^\star \coloneqq \max_{a' \in [K]} \mathsf{s}_{a'} - \mathsf{s}_a$ as the infimal amount of incentive to be offered on action a to make the agent choose it. Consequently, we have

$$\begin{aligned}\mathfrak{R}(T) &= T\,\mu^\star - \sum_{t=1}^{T} \mathbb{E}[\theta_{A_t} - \mathbb{1}_{a_t}(A_t)\pi(t)] \\ &\geqslant \mathbb{E}\left[\sum_{t=1}^{T} \mu^\star - (\theta_{A_t} - \pi^\star_{A_t})\right] \\ &\geqslant \mathbb{E}\left[\sum_{t=1}^{T} \mu^\star - \mu_{A_t}\right] .\end{aligned}$$

Assuming the principal knows $(\mathsf{s}_a)_{a\in[K]}$, observing $X_a(t)$ is equivalent to observing $X_a(t) - \pi_a^\star$. Using the result of Burnetas & Katehakis (1996) (see, e.g., Lattimore & Szepesvári, 2020, Theorem 16.2.), it then comes

$$\liminf_{T\to\infty} \frac{\mathfrak{R}(T)}{\log T} \geqslant \sum_{a,\mu_a<\mu^\star} \frac{\mu^\star - \mu_a}{\mathrm{KL}_{\inf}(\nu_a - \pi_a^\star, \mu^\star, \mathcal{D})} .$$

□