Learning to Incentivize Information Acquisition: Proper Scoring Rules Meet Principal-Agent Model

Siyu Chen [1] **Jibang Wu** [2] **Yifan Wu** [3] **Zhuoran Yang** [1]

Abstract

We study the incentivized information acquisition problem, where a principal hires an agent to gather information on her behalf. Such a problem is modeled as a Stackelberg game between the principal and the agent, where the principal announces a scoring rule that specifies the payment, and then the agent then chooses an effort level that maximizes her own profit and reports the information. We study the online setting of such a problem from the principal's perspective, i.e., designing the optimal scoring rule by repeatedly interacting with the strategic agent. We design a provably sample efficient algorithm that tailors the UCB algorithm (Auer et al., 2002) to our model, which achieves a sublinear $T^{2/3}$-regret after T iterations. Our algorithm features a delicate estimation procedure for the optimal profit of the principal, and a conservative correction scheme that ensures the desired agent's actions are incentivized. Furthermore, a key feature of our regret bound is that it is independent of the number of states of the environment.

1. Introduction

Delegated information acquisition is a widely popular situation where one party (known as the principal) wants to acquire some information that assists decision-making, and thus hires another party (known as the agent) to gather information on its behalf. Consider a portfolio manager who aims to learn the potential of a company in order to decide whether to invest in its stock. The manager hires an analyst, who spends some effort to conduct the research, hands in a report to the manager, and gets payment according to the report she writes. Based on the report, the manager makes the investment decision and earns profit if the stock rises. The level of effort the analyst puts into the research affects the quality of the information gathered, i.e., the report, and her own cost incurred in conducting the research. As a result, a rational analyst would choose an effort level that maximizes her own profit – the difference between the payment and the cost. Whereas the manager also wants to maximize its own profit in expectation, which is given by the expected gain from the investment, subtracted by the payment. Knowing that the analyst is rational, the goal of the manager is to design a payment rule that incentivizes the analyst to spend a proper amount of effort, such that the acquired information leads to the investment decision that maximizes the portfolio manager's profit.

Mathematically, such a problem can be modeled under the principal-agent model (Laffont & Martimort, 2009), where the principal wants to know the state ω of a stochastic environment, and the information acquired by the agent is a distribution $\widehat{\sigma}$ over the state space Ω, also known as the reported belief. The game between the principal and the agent is as follows. At the beginning, the state is not realized and unknown to both parties. The principal chooses a scoring rule S (Savage, 1971; Gneiting & Raftery, 2007) as the payment rule and presents it to the agent. The agent chooses among K effort levels by selecting an action $b_k \in \{b_1, \ldots, b_K\}$ at a cost c_k. Then b_k determines the joint distribution $p(\omega, o \,|\, b_k)$ of the state ω and an observation o. Based on such a conditional distribution and the realized value of o, the agent reports $\widehat{\sigma}$ to the principal. Based on such acquired information, the principal chooses an action $a \in \mathcal{A}$. Finally, the state ω is revealed, the principal receives utility $u(a, \omega)$ and pays the agent $S(\widehat{\sigma}, \omega)$. Here, the payment to the agent is determined by the scoring rule S, which quantifies the value of the reported belief by comparing it with the realized state.

More generally, our model is a general Stackelberg game with information asymmetry (von Stackelberg, 1952; Mas-Colell et al., 1995), where the agent knows the distri-

[1]Department of Statistics and Data Science, Yale University [2]Department of Computer Science, University of Chicago [3]Department of Computer Science, Northwestern University. Correspondence to: Siyu Chen <siyu.chen.sc3226@yale.edu>, Jibang Wu <wujibang@uchicago.edu>, Yifan Wu <yifan.wu@u.northwestern.edu>, Zhuoran Yang <zhuoran.yang@yale.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

bution of the state while the principal does not. The principal first announces a scoring rule S. Then the principal chooses an effort level b_k which maximizes her own profit and reports a belief $\widehat{\sigma}$. That is, b_k is the best response of the agent to S. The expected profits of both the principal and the agents are functions of S, b_k, and $\widehat{\sigma}$. From such a perspective, designing the scoring rule that optimally elicits information is equivalent to finding the strong Stackelberg equilibrium of such a Stackelberg game.

In this work, we focus on the online setting of such a principal-agent model. In particular, we aim to answer:

How to learn the optimal scoring rule by interacting with a strategic agent from the principal's perspective?

The online setting comes with a few challenges. First, as the agent is strategic, the reported belief $\widehat{\sigma}$ might be untruthful, or even arbitrary. Second, similar to other online learning problems such as bandits (Lattimore & Szepesvári, 2020), we need to explore the stochastic environment. More importantly, in our problem, the distribution of the state is determined by the action b_k of the agent, which is beyond the control of the principal. Thus, any successful learning algorithm needs to execute scoring rules that incentivize the agent to explore her action space. Third, both the cost c_k and the distribution $p(\omega, o \mid b_k)$ are unknown to the principal. In other words, the profit functions of both the principal and the agent are unknown and needs to be estimated from the online data. In particular, this implies that the best response of the agent, as a function of the principal's scoring rule, is unknown. To find principal's optimal scoring rule, we need to know how to incentivize the agent to choose the most favorable b_k for principal, which requires learning the best response function.

We tackle these challenges by introducing a novel algorithm, OSRL-UCB, which leverages proper scoring rules (Gneiting & Raftery, 2007), the principle of optimism in the face of uncertainty (Auer et al., 2002; Lattimore & Szepesvári, 2020), and the particular constrained optimization formulation of our principal-agent model. In particular, to elicit truthful information, we prove a revelation principle (Myerson, 1979) that shows that it suffices to only focus on the class of proper scoring rules (Lemma 2.2). Then we show that the principal's profit maximization problem can be written as a K-armed bandit problem, where the reward of each arm $h^{k,*}$ is the optimal profit of the principal when the best response of the agent is fixed to b_k (Equation (3.1)). The value of $h^{k,*}$ is determined by the optimal objective of a constrained optimization problem — finding a proper scoring rule S^k that maximizes the principal's profit, subject to the constraint that the best response to S^k is b_k. Furthermore, we show that $h^{k,*}$ can be equivalently written as the optimal value of a constrained optimization linear program (LP) involves the unknown pairwise cost differences and the distribution of the truthful belief induced by b_k (Equation (3.3)). Following the optimism principle, we aim to construct upper confidence bounds (UCB) of each $h^{k,*}$ and incentive the agent to pick the action that maximizes the UCB. To this end, we construct confidence sets for the pairwise cost differences and belief distributions based on the online data, and obtain a UCB of $h^{k,*}$ by solving an optimistic variant of the LP by replacing the unknown quantities by elements in the corresponding confidence sets (Equation (3.7)). Furthermore, the optimal solution to such an optimistic LP, which is a scoring rule, might violate the condition that its best response is b_k. To remedy this, we devise a conservative modification scheme which simultaneously guarantees desired best response and optimism with high probability. Finally, we prove that the proposed algorithm achieves a $\widetilde{\mathcal{O}}(K^2\mathcal{C}_{\mathcal{O}} \cdot T^{2/3})$ sublinear regret upper bound after T rounds of interactions. Here K is the number of effort levels of the agent and $\mathcal{C}_{\mathcal{O}}$ is the number of all possible observations, and $\widetilde{\mathcal{O}}$ omits logarithmic terms. A key feature of our regret bound is that it is independent of the number of states of the environment.

1.1. Related Work

The information elicitation problem and its implications for decision making have been one of the most popular areas in economics. The seminal work by (Savage, 1971) characterizes the truthful elicitation with the design of proper scoring rules. More recently, several work have studied the optimization problem of scoring rules for the principal's different objectives (Chen & Yu, 2021; Li et al., 2022; Papireddygari & Waggoner, 2022; Neyman et al., 2021). Closer to our problem is the model studied by Papireddygari & Waggoner (2022); Oesterheld & Conitzer (2020), as we both explore the connection between information acquisition and contract design. However, our model is strictly more general in that we assume the agent may have multi-level efforts, and the information state may be affected by the agent's action. In addition, we primarily focus on the learning aspects, when the principal has limited knowledge of the game parameters.

More generally, our problem is related to a family of online learning problems under the principal-agent framework, which includes special cases such as security game, information design, auction design, contract design and etc. However, to our best knowledge, there is no previous work that considered any similar learning problem for our model of information acquisition. We also remark that, in many of these existing works, e.g., Balcan et al. (2015); Guo et al. (2022); Castiglioni et al. (2020); Wu et al. (2022), the learner is assumed to have sufficient knowledge about the other strategic player(s), and her uncertainty is regarding the environment or her own utility. Assumptions of such

kind can significantly simplify the problem into the standard online learning problems, once the learner can almost predict the best response of other player(s). Notably, our work does not make any of such assumption and the most challenging part of our learning algorithm design is indeed to ensure the desired agent response under uncertainty. We defer more discussions on the related work to Appendix A.

2. The Information Acquisition Model

2.1. Basics of Information Acquisition

To formulate the problem of optimally acquiring information under the principal-agent framework, we consider a stochastic environment with a principal and an almighty agent. At the t-th round, there is a hidden state $\omega_t \in \Omega$ that will affect the principal's utility, but is unknown to both the agent and the principal until the end of this round. To elicit refined *information*, i.e., the agent's belief of the hidden state, the principal moves first and offers a scoring rule to the agent, based on which the agent receives a payment according to the quality of her reported belief. The agent is allowed to choose an action from her finite action space $\mathcal{B}$ with some cost, obtain an observation related to the hidden state, and generate a report on her belief, which puts the principal in a better position to make a decision. In the end, the hidden state is revealed, and a utility is generated for the principal, who then pays the agent based on the scoring rule. For any $t \geq 0$, in the t-th round, the interactions between the principal and the agent are as follows.

1. The principal commits to a scoring rule $S_t : \Delta(\Omega) \times \Omega \to \mathbb{R}_+$, where $\Delta(\Omega)$ is the space of distributions over Ω.
2. Based on S_t, the agent chooses an action $b_{k_t} \in \mathcal{B}$ indexed by k_t and bears a cost $c_{k_t} \geq 0$. The action b_{k_t} can be observed by the principal.
3. The stochastic environment then selects a hidden state $\omega_t \in \Omega$ and emits an observation $o_t \in \mathcal{O}$ only for the agent according to $p(\omega_t, o_t \,|\, b_{k_t})$. The hidden state ω_t is unknown to both the agent and the principal at this moment.
4. The agent reports a belief $\widehat{\sigma}_t \in \Delta(\Omega)$ about the hidden state to the principal.
5. The principal makes a decision $a_t \in \mathcal{A}$ based on $(\widehat{\sigma}_t, k_t, S_t)$.
6. In the end, the hidden state ω_t is revealed. The principal obtains her utility $u(a_t, \omega_t)$ and pays the agent by $S_t(\widehat{\sigma}_t, \omega_t)$.

Here, the scoring rule $\mathcal{S}_t$ is a payment rule that depends on the agent's report $\widehat{\sigma}_t$ and the true state ω_t. Let $\mathscr{S}$ be the class of scoring rules with bounded norm $\|S\|_\infty \leq B_S$. In the sequel, we assume the principal picks $\mathcal{S}_t$ from $\mathscr{S}$. We assume that the reward function $u : \mathcal{A} \times \Omega \to \mathbb{R}$ also has a bounded norm $\|u\|_\infty \leq B_u$. We consider the agent's action set $\mathcal{B}$ and the observation set $\mathcal{O}$ to be finite. Specifically, the agent's action space $\mathcal{B} = \{b_1, \ldots, b_K\}$ has K actions. In the sequel, we also use the action index k_t to represent the agent's action. The agent's policy for choosing her action k_t, her report $\widehat{\sigma}_t$, and the principal's policy for choosing her action a_t will be introduced shortly after.

Notably, our modeling captures the endogenous effect that the agent's action choice may influence the environment state. This is more general than assuming the state is exogenous, and captures the real-world situations that the act of information acquisition, e.g., market investigation, affects the stochastic environment. Consider the example of a portfolio manager and financial analyst introduced in §1. The report written by the analyst about a particular stock, when released to the public, may generate considerable impact and affect the stock price (Lui et al., 2012).

Information Structure. In the remaining part of this subsection, we ignore the subscript t for a while. In this information acquisition process, we assume the agent is almighty that has full knowledge of the *information structure*, i.e., each action's cost c_k and the generating process $p(\omega, o \,|\, b_k)$ for the hidden state and the observation under action b_k. Therefore, after obtaining the observation o, the agent is able to refine her belief $\sigma \in \Delta(\Omega)$ of the hidden state via the Bayesian rule $\sigma(\omega) := p(\omega \,|\, o, b_k) = p(\omega, o \,|\, b_k)/p(o \,|\, b_k)$. Note that σ is a random measure mapping from the observation space to the probability space $\Delta(\Omega)$ and captures the randomness in o. Let $\Sigma \subset \Delta(\Omega)$ be the support of σ. Define M as the cardinality of Σ and it follows from the discreteness of $\mathcal{B}$ and $\mathcal{O}$ that $M \leq K \times C_{\mathcal{O}}$, where $C_{\mathcal{O}}$ is the cardinality of $\mathcal{O}$. Let $q_k(\sigma) \in \Delta(\Sigma)$ be the distribution of σ under the agent's action $k \in [K]$. Since σ already captures all the information about the hidden state from the observation, we ignore the observation o and refer to the costs $\{c_k\}_{k\in[K]}$ and the distributions of the belief $\{q_k\}_{k\in[K]}$ as the information structure, which are private to the agent.

Solving for the Optimal Scoring Rules. We start with the observation that the information acquisition process can be formulated as a general Stackelberg game. Let $\mu : \mathscr{S} \to [K]$ and $\nu : \mathscr{S} \times \Sigma \times [K] \to \Delta(\Omega)$ be the agent's policy for action selection and belief reporting, respectively. Here, the reporting policy ν depends on σ instead of o since σ captures all the information about the hidden state. Given any scoring rule S, the agent (as the follower of this game) finds her own action by $k = \mu(S)$, and reports $\widehat{\sigma} = \nu(S, \sigma, k)$ that maximizes her own expected profit, i.e., $g(\mu, \nu; S) := \mathbb{E}[S\left(\nu(S, \sigma, \mu(S)), \omega\right) - c_{\mu(S)}]$, where the expectation is taken over the randomness of ω, σ with respect to $q_{\mu(S)}$. Let $\iota : \mathscr{S} \times \Delta(\Omega) \times [K] \to \mathcal{A}$ be the principal's decision policy. Hence, the principal (as the

leader of this game) is to find the optimal scoring rule S and the best decision policy ι that maximize her own expected profit given the agent's best response (μ^*, ν^*), i.e., $h(\mu^*, \nu^*; S, \iota) := \mathbb{E}[u(\iota(S, \nu^*(S, \sigma, \mu^*(S)), \mu^*(S)), \omega) - S(\nu^*(S, \sigma, \mu^*(S)), \omega)]$, where the expectation is taken over the randomness of ω, σ with respect to $q_{\mu^*(S)}$. The optimal leader strategies are known as the strong Stackelberg equilibria that can be formulated as solutions of a bilevel optimization problem (Conitzer, 2016), i.e.,

$$\max_{\iota, S \in \mathscr{S}} h(\mu^*, \nu^*; S, \iota), \text{ s.t. } (\mu^*, \nu^*) \in \operatorname*{argmax}_{\mu, \nu} g(\mu, \nu; S). \tag{2.1}$$

However, the bilevel optimization in (2.1) is computationally intractable, since the agent's action space (particularly, the space of reporting scheme ν) is intractable.

To resolve such an issue, in the following, we establish a revelation principle result that guarantees that, without loss of generality, it suffices to only focus on the case where the agent is truthful. Specifically, there is a well-known class of scoring rules (in Definition 2.1) that characterizes all scoring rules, under which truthfully reporting is in the agent's best interest (Savage, 1971).

Definition 2.1 (Proper scoring rule)**.** A scoring rule $S : \Delta(\Omega) \times \Omega \to \mathbb{R}_+$ is proper if, for any belief $\sigma \in \Delta(\Omega)$ and any reported posterior $\widehat{\sigma} \in \Delta(\Omega)$, we have $\mathbb{E}_{\omega \sim \sigma} S(\widehat{\sigma}, \omega) \leq \mathbb{E}_{\omega \sim \sigma} S(\sigma, \omega)$. In addition, if the inequality holds strictly for any $\widehat{\sigma} \neq \sigma$, the scoring rule S is strictly proper.

By definition, reporting the true belief maximizes the payment for the agent. Following the revelation principle (Myerson, 1979), we can argue that any equilibrium with possibly untruthful report of belief can be implemented by an equilibrium with truthful report of belief. This means that the principal's optimal scoring rules can be assumed proper without loss of generality, and we can thereby restrict the reporting scheme ν to the truthful ones — we state the revelation principle for the information acquisition model in the following lemma and defer its proof to §C.1.

Lemma 2.2 (Revelation principle)**.** There exists a proper scoring rule S^* that is an optimal solution to (2.1).

In the sequel, we let $\mathcal{S}$ denote the class of proper scoring rules with bounded norm $\|S\|_\infty \leq B_S$. When $S \in \mathcal{S}$, the agent's best report scheme is $v^*(S, \sigma) = \sigma$ since being truth-telling maximizes the payment, and the principal's best decision policy ι^* can be simplified to $a^*(\sigma)$ since $\omega \perp\!\!\!\perp (S, k) \,|\, \sigma$. Using the notations $u(\sigma) = \mathbb{E}_{\omega \sim \sigma} u(a^*(\sigma), \omega)$, $S(\sigma) = \mathbb{E}_{\omega \sim \sigma} S(\sigma, \omega)$, and $k^*(S) = \mu^*(S)$, we can transform the optimization program in (2.1) into

$$\begin{aligned} \max_{S \in \mathcal{S}} \quad & \mathbb{E}_{\sigma \sim q_{k^*(S)}} \left[u(\sigma) - S(\sigma) \right], \\ \text{s.t.} \quad & k^*(S) \in \operatorname*{argmax}_{k \in [K]} \mathbb{E}_{\sigma \sim q_k} S(\sigma) - c_k, \end{aligned} \tag{2.2}$$

where we will denote the agent's utility function as $g(k, S) := \mathbb{E}_{\sigma \sim q_k} S(\sigma) - c_k$ and the principal's utility function under the agent's best response $k^*(S)$ as $h(S) := \mathbb{E}_{\sigma \sim q_{k^*(S)}} [u(\sigma) - S(\sigma)]$. This optimization program can be solved efficiently, e.g., by solving for the optimal proper scoring rule that induces each of the agent's actions as the best response. And we will revisit (2.2) in the online learning process where the information structure, i.e., q_k and c_k, is unknown to the principal.

We also remark that our model is fully capable of modeling the standard contract designing problem. If the information structure is full-revealing (e.g., σ is a point belief), our model with the endogenous states degenerates to a standard contract design problem, where the scoring rule S becomes a contract as a mapping from (truthfully reported) outcome to payment. We defer the details to §Appendix B.

2.2. Online Learning to Acquire Information

We now formalize the online learning problem of solving the optimal scoring rule for information acquisition. We consider the situation where the principal only has knowledge of her utility function u [1] and is able to observe the agent's action b_k [2]. In Table 1, we summarize all the information types discussed above together with their availability. Given $H_{t-1} = \{(S_\tau, k_\tau, \sigma_\tau, \omega_\tau)\}_{\tau \in [t-1]} \in \mathcal{H}_{t-1}$ as the history observed by the principal before round t, the principal is able to deploy a policy for the next round's scoring rule $\pi_t : \mathcal{H}_{t-1} \to \mathcal{S}$. Hence, the data generating process is described as the following,

$$\begin{aligned} p^\pi(S_t, k_t, \sigma_t, \omega_t \,|\, H_{t-1}) = \mathbb{1}\left(S_t = \pi_t(H_{t-1}), k_t = k^*(S_t)\right) \\ \cdot q_{k_t}(\sigma_t) \cdot \sigma_t(\omega_t). \end{aligned} \tag{2.3}$$

[1] Since the utility and the hidden state are both known to the principal at the end of each round, if the agent is truth-telling about her belief, the principal's decision problem reduces to a contextual bandit (with finite context set) and can be handled by the standard UCB algorithm with $\mathcal{O}(M\sqrt{T})$ regret. To simplify our discussion, we consider u to be known by the principal.

[2] In cases where the principal cannot observe the agent's action, there are still ways to distinguish different actions. For instance, when q_k has different support for different k and the agent is truth-telling about her belief under proper scoring rules (see §C.1), the principal is able to learn the support for a particular agent's action by repeating the same scoring rule multiple times. The next time the agent chooses the same action, the principal will be aware.

The regret for the online policy $\pi = \{\pi_t\}_{t\in[T]}$ is defined as,

$$\mathrm{Reg}^{\pi}(T) := T \cdot \max_{S\in\mathcal{S}} h(S) - \mathbb{E}_{\pi}\left[\sum_{t=1}^{T} u(a^*(\sigma_t),\omega_t) - S_t(\sigma_t,\omega_t)\right],$$

where the expectation is taken with respect to the data generating process. We aim to develop an online policy π_t that learns the optimal scoring rule with small regret.

3. The OSRL-UCB Algorithm

In this section, we introduce the algorithm for Online Scoring Rule Learning with Upper Confidence Bound (OSRL-UCB). We begin with an overview of the algorithm in §3.1 and introduce an action-informed oracle that is necessary for the algorithm in §3.2. In §3.3, we give a detailed description of the OSRL-UCB algorithm. To simplify notation, we let k to denote b_k in the sequel.

3.1. Algorithm Overview

Define $\mathcal{V}_k = \{S \in \mathcal{S} \,|\, g(k,S) \geq g(k',S), \forall k' \in [K]\}$ as the k-th section in which the agent takes action b_k as her best response. The principal's optimization problem (2.2) can be reformulated as

$$\max_{k\in[K]} h^{k,*} := \sup_{S\in\mathcal{V}_k} \mathbb{E}_{\sigma\sim q_k}\left[u(\sigma) - S(\sigma)\right], \tag{3.1}$$

where $h^{k,*}$ is the principal's optimal profit when the agent's best response is k. Let $S^{k,*}$ be the optimal solution to the inner problem of (3.1). If the principal knows the best scoring rule $S^{k,*}$ that achieves $h^{k,*}$, the problem immediately reduces to a multi-arm bandit where $k \in [K]$ is the arm and $h^{k,*}$ is the reward for arm k. Such a problem can thus be handled by the standard UCB algorithm (Auer et al., 2002). However, there are two obstacles: (i) the action region $\mathcal{V}_k$ is unknown; (ii) the belief distribution q_k is unknown. Recall the definition of $\mathcal{V}_k$:

$$\mathcal{V}_k = \{S \in \mathcal{S} \,|\, \langle q_k - q_{k'}, S\rangle_{\Sigma} \geq c_k - c_{k'}, \forall k' \in [K]\}.$$

To identify $\mathcal{V}_k$, we just need to estimate the belief distribution q_k and the pairwise cost difference defined as $C(i,j) = c_i - c_j$. Specifically, estimator $\widehat{q} = (\widehat{q}_k)_{k\in[K]}$ can be updated by the empirical distribution of σ_t such that $k_t = k$ while estimator $\widehat{C} = (\widehat{C}(i,j))_{i,j\in[K]}$ can be updated using the following identity

$$C(i,j) = \langle q_i - q_j, S\rangle_{\Sigma}, \quad \forall S \in \mathcal{V}_i \cap \mathcal{V}_j, \tag{3.2}$$

where we plug in estimator $\widehat{q}$. In addition, we must identify a scoring rule $S \in \mathcal{V}_i \cap \mathcal{V}_j$ to successfully employ (3.2). To this end, we employ a binary search method on the convex combination of S_1, S_2 such that $k^*(S_1) = i$ and $k^*(S_2) = j$. To inform the principal of the K actions and also to guarantee that the principal can find a scoring rule S such that $k^*(S) = i$ for the sake of the binary search, we introduce with examples an action-informed oracle in §3.2, which provides the principal with foreknown scoring rules $\widetilde{S}_1, \widetilde{S}_1, \ldots, \widetilde{S}_K$ such that $k^*(\widetilde{S}_i) = i$.

Now that the estimation problem of q_k and $C(i,j)$ is addressed, the inner problem of (3.1) can be solved by the following constrained linear program,

$$\begin{aligned} \mathrm{LP}_k: \quad & h^{k,*} = \max_{S\in\mathcal{S}} \quad \langle q_k, u - S\rangle_{\Sigma}, \\ & \text{s.t.} \quad \langle q_k - q_{k'}, S\rangle_{\Sigma} \geq C(k,k'), \quad \forall k' \in [K]. \end{aligned} \tag{3.3}$$

If (3.3) is solved, we can reduce the outer problem of (3.1) to a bandit by viewing $[K]$ as the set of arms and the $h^{k,*}$ as the reward for each arm $k \in [K]$. To illustrate the method for solving LP_k, let $\widetilde{\mathcal{Q}} = \{(\widetilde{q}_k)_{k\in[K]}\}$ and $\widetilde{\mathcal{C}} = \{(\widetilde{C}_{ij})_{i,j\in[K]}\}$ be the confidence sets for $\widehat{q}$ and $\widehat{C}$ that capture the real q and C with high probability. To encourage exploration, we follow the principle of optimism and obtain an upper confidence bound for $h^{k,*}$ by solving LP_k with plugged-in $(\widetilde{q}, \widetilde{C})$ that maximizes the optimization goal (principal's profit) over the confidence sets $\widetilde{\mathcal{Q}} \times \widetilde{\mathcal{C}}$. This optimistic version of LP_k is given by Opt-LP_k in (3.7) , in which we do not explicitly construct these confidence sets, but instead exploit the confidence intervals for estimating the utilities, the payments, and the cost differences using $\widehat{q}$ and $\widehat{C}$. With the optimistic reward for each action given by Opt-LP_k, the algorithm simply chooses action k_t^* that maximizes this reward, and obtain the scoring rule solution S_t^* corresponding to k_t^*.

However, there is one remaining issue if we want to deploy S_t^* to incentive the agent to choose action k_t^*. The fact that the constraints of (3.3) are relaxed in Opt-LP_k by using optimism may cause S_t^* to go beyond $\mathcal{V}_{k_t^*}$. To address such a problem, we propose to deploy a conservatively adjusted scoring rule $S_t = (1-\alpha_t)S_t^* + \alpha_t \widetilde{S}_{k_t^*}$, which guarantees the agent to choose k_t^* with high probability. By letting α_t to decay with t, the conservatively adjusted scoring rule also guarantees optimism. The algorithm is summarized in Algorithm 1 and more details are available in §3.3.

3.2. Oracle for Action Awareness

We assume that there is an oracle that provides the principal with a foreknown set of scoring rules that induce the agent to pick all her actions. This is a strictly weaker assumption than many existing online learning models in strategic environments, which assume that the principal can predict the best response of the agent, or know some parameters of the agent's utility function.

Assumption 3.1 (Action-informed oracle)**.** We assume

that there is an oracle which comes up with K scoring rules $(\widetilde{S}_1, \cdots, \widetilde{S}_K)$ such that under $\widetilde{S}_k$, the agent's best response is action k for $k \in [K]$. Moreover, for the agent's profit $g(k, S) = \mathbb{E}_{\sigma\sim q_k} S(\sigma) - c_k$, we assume that there exists $\varepsilon > 0$ such that for any $k' \neq k$, we have $g(k, \widetilde{S}_k) - g(k', \widetilde{S}_k) > \varepsilon$.

With the oracle in Assumption 3.1, the principal is initially aware of the K actions that the agent might respond, which can be easily done by trying $(\widetilde{S}_1, \cdots, \widetilde{S}_K)$ one by one. In addition, we assume that $\widetilde{S}_k$ lies within the interior of the k-th section $\mathcal{V}_k = \{S \in \mathcal{S} \,|\, g(k, S) \geq g(k', S), \forall k' \in [K]\}$ and keeps some distance from the boundary of $\mathcal{V}_k$. We remark that having $\widetilde{S}_k$ away from the boundary is essential to induce desired behavior in the agent with high probability when applying an approximating scoring rule based on $\widetilde{S}_k$ but with some errors. Notably, the assumption that the agent has marginal gain ε by choosing k under $\widetilde{S}_k$ also ensures a minimum radius of section $\mathcal{V}_k$, which corresponds to the non-degenerate assumption in strategic Stackelberg games (Letchford et al., 2009). We first present an impossible result for online learning the optimal proper scoring rule without the oracle.

Lemma 3.2 (Impossible result)**.** Suppose number of actions $K \geq 3$ and the number of possible beliefs $M \geq 3$. Without the action-informed oracle, for any online algorithm, there exists an instance such that $\text{Reg}(T) = \Omega(T)$.

See §E.1 for a construction of the hard instance. Without the action-informed oracle, any online algorithm inevitably suffers from a linear regret. The intuition behind the hard instance is that without the oracle, any algorithm can fail to locate a scoring rule that induces the desirable action from the agent. This happens because the feasible region, i.e., $\mathcal{V}_k$ in the space of bounded proper scoring rules for this desirable action k, can be arbitrarily small. We next give two examples that show the action-informed oracle is amenable to practical implementation. The first example is through random sampling, adapted from (Letchford et al., 2009). To do so, we first define the class of strongly proper scoring rules.

Definition 3.3 (Strongly proper scoring rules)**.** A scoring rule $S : \Delta(\Omega) \times \Omega \to \mathbb{R}$ is β-strongly proper if for all $p, q \in \Delta(\Omega)$, $\mathbb{E}_{\omega\sim q}[S(q, \omega)] - \mathbb{E}_{\omega\sim q}[S(p, \omega)] \geq \frac{\beta}{2} \|q - p\|_1^2$.

Let $\mathcal{S}_\beta$ be the class of β-strongly proper scoring rules. We sample from $\mathcal{S}_\beta$ in the hope that after a small number of samplings, e.g., $\log(T)$ the action-informed oracle can be constructed with high probability.

Example 3.4 (Action awareness via random sampling)**.** Let $d_1 = \min_{1\leq i<j\leq M} \|\sigma_i - \sigma_j\|_\infty$ and $d_2 = \min_{1\leq k<k'\leq K} \max_{i\in[M]} [q_k(i) - q_{k'}(i)]$. Since the proper scoring rule class $\mathcal{S}$ has bounded norm, the β-strongly proper scoring class also has bounded volume $\text{Vol}(\mathcal{S}_\beta) < \infty$. Set $\kappa = d_1^2\beta/2$ and let $\widetilde{\mathcal{V}}_k = \{S \in \mathcal{S}_\beta \,|\, g(k, S) \geq g(k', S) + \kappa, \forall k' \neq k, k' \in [K]\}$. We suppose $\text{Vol}(\widetilde{\mathcal{V}}_k) \geq \eta\, \text{Vol}(\mathcal{S}_\beta)$ for $k \in [K]$. Initiate $\mathcal{M} = \varnothing$ as the candidate set. Each time, we randomly sample a β-strongly proper scoring rule $S \in \mathcal{S}_\beta$ and obtain the agent's best response $k^*(S)$ with respect to S. Let $e_i(\sigma, \omega) = \mathbb{1}(\sigma = \sigma_i)$ for $i \in [M]$. By setting $\kappa = d_1^2 \cdot \beta/2$, we deploy $\{S - \kappa e_i\}_{i\in[M]}$ and see if the agent still responds $k^*(S)$. If so, Let $\mathscr{S} = \mathscr{S} \cup \{S\}$; if not, reject S. After $\mathcal{O}(M\eta^{-1}K \log K)$ rounds, with high probability, $\mathscr{S}$ serves as a valid action-informed oracle with parameter $\epsilon = d_2 \cdot \kappa$.

See §E.2 for more details. In Example 3.4, we ensure a set of scoring rules that successfully induces each action by randomly sampling in $\mathcal{S}$ for up to $\mathcal{O}(MK \log K)$ times. We next show another example via use of linear contract where the information structure satisfies some special properties.

Example 3.5 (Action awareness via linear scoring rule)**.** Suppose that these K actions are sorted in an increasing order with respect to the cost. Define $u_k = \mathbb{E}_{\omega\sim\sigma,\sigma\sim q_k}[u(a^*(\sigma), \omega)]$ and suppose $u_k > 0$. We assume the marginal information gain is strictly decaying, i.e., there exists a $\epsilon > 0$ such that

$$\frac{u_K - u_{K-1}}{c_K - c_{K-1}} > \epsilon, \quad \text{and}$$
$$\frac{u_k - u_{k-1}}{c_k - c_{k-1}} - \frac{u_K - u_k}{c_K - c_k} > \epsilon, \quad \forall k = 2, \cdots, K-1.$$

Moreover, we assume that $(u_2 - u_1)/(c_2 - c_1) \leq b$. The principal sets the scoring rule as $S(\sigma, \omega) = \lambda u(a^*(\sigma), \omega)$ and conducts binary search on $\lambda \in [0, 2/\epsilon]$. Specifically, the binary searches are iteratively conducted on all the segments (λ_1, λ_2) with $k^*(\lambda_1 u) \neq k^*(\lambda_2 u)$, where λ_1, λ_2 are neighboring points on $[0, 1/\epsilon]$ that are previously searched. With the maximal searching depth $m = \lceil \log_2(2b(b-\epsilon)/\epsilon^2) \rceil$, we can identify all the agent's actions. Suppose that $(\lambda_1^k, \lambda_2^k)$ is the largest segment with λ_1^k and λ_2^k searched before and $k^*(\lambda_1^k u) = k^*(\lambda_1^k u) = k$. By letting $\widetilde{S}_k = (\lambda_1^k + \lambda_2^k)u/2$, we obtain an oracle with $\varepsilon = \epsilon u_1/4b^2$. The procedure takes $\mathcal{O}(K \log_2(\varepsilon^{-1}))$ rounds.

See §E.3 for more details. In this example, we exploit the power of linear contract (Dütting et al., 2019) to identify all the agent's actions and produce a set of scoring rules based on the principal's utility. Specifically, the assumption of decaying marginal information gain is commonly seen in real world practice, and it further guarantees that all the actions are inducible using linear contracts. To obtain such an oracle, we just need at most $\mathcal{O}(K \log_2(\varepsilon^{-1}))$ trials, which is more efficient than random sampling. We remark that these foreknown scoring rules obtained by the oracle do not need to be optimal in each section $\mathcal{V}_k$. They can even be

obtained through random sampling from β-strongly proper scoring rules (See Example 3.4) for general setting, or discovered in a linear scoring rule class (See Example 3.5) if the marginal information gain is strictly decaying.

3.3. OSRL-UCB Algorithm

The OSRL-UCB algorithm contains the following four parts: (i) learning the belief distributions; (ii) learning the pairwise cost differences; (iii) solving for the optimal payment/scoring rule in each $\mathcal{V}_k$ via optimistic linear programming Opt-LP$_k$ and choosing the best "arm" via UCB; (iv) applying a conservative scoring rule and conduct binary search if $k_t \neq k_t^*$ to refine our estimation on $\mathcal{V}_k$. In this subsection, we describe the OSRL-UCB algorithm in detail.

Algorithm 1 Online Scoring Rule Learning with Upper Confidence Bound (OSRL-UCB)

Input: $\mathcal{S}$, $(\widetilde{S}_1, \cdots, \widetilde{S}_K)$.
while $t \leq T$ **do**
  Estimate the belief distributions $\widehat{q}_k^t$ by (3.4) and the corresponding confidence intervals $I_q^t(k)$ by (3.5);
  Estimate the pairwise cost differences $\widehat{C}^t(i,j)$ and the corresponding confidence intervals $I_c^t(i,j)$ by (3.6);
  Solve Opt-LP$_k$ in (3.7) and obtain the optimal value $h_{\text{LP}}^t(k)$ and the optimal solution $S_{\text{LP},k}^t$ for $k \in [K]$;
  Choose the best arm $k_t^* \leftarrow \text{argmax}_{k\in[K]} h_{\text{LP}}^t(k)$ and let $S_t^* \leftarrow S_{\text{LP},k_t^*}^t$;
  Pay with the conservative scoring rule $S_t \leftarrow \alpha_t \widetilde{S}_{k_t^*} + (1-\alpha_t) S_t^*$, and collect the agent's responding action k_t;
  if $k_t \neq k_t^*$ **then**
    Conduct binary search $\text{BS}(S_t, \widetilde{S}_{k_t^*}, k_t^*, t)$ as specified in Algorithm 3;
  end if
  Move on to a new round $t \leftarrow t+1$;
end while

Learning the Belief Distributions. Let n_k^t denote the total number of times the agent responds action k before round t. Then, q_k can be learned empirically as

$$\widehat{q}_k^t(\sigma) = \frac{1}{n_k^t} \sum_{\tau=1}^{t-1} \mathbb{1}(\sigma_\tau = \sigma, k_\tau = k), \forall \sigma \in \Sigma. \tag{3.4}$$

Following from a standard concentration result in Mardia et al. (2020), we define the confidence interval for $\widehat{q}_k^t$ as

$$I_q^t(k) = \sqrt{2\log((K)\cdot 2^M t)/n_k^t} \tag{3.5}$$

We state the concentration result in Lemma F.6. Under $\widehat{q}_k^t$, the estimated payment of scoring rule S if the agent responds action k is $\widehat{v}_S^t(k) = \langle S(\cdot), \widehat{q}_k^t(\cdot)\rangle_\Sigma$.

Learning the Pairwise Cost Differences. We next illustrate how to learn the pairwise cost difference. For each $\tau < t$ such that $k_\tau = i$ and $j \neq i$, define

$$C_+^t(i,j) = \min_{\tau<t:k_\tau=i} \widehat{v}_{S_\tau}(i) - \widehat{v}_{S_\tau}(j) + B_S\left(I_q^t(i) + I_q^t(j)\right),$$
$$C_-^t(i,j) = \max_{\tau<t:k_\tau=j} \widehat{v}_{S_\tau}(i) - \widehat{v}_{S_\tau}(j) - B_S\left(I_q^t(i) + I_q^t(j)\right).$$

For each pair (i,j), we also define

$$\theta^t(i,j) = \frac{C_+^t(i,j) + C_-^t(i,j)}{2},$$
$$\varphi^t(i,j) = \left[\frac{C_+^t(i,j) - C_-^t(i,j)}{2}\right]_+.$$

Directly using $\theta^t(i,j)$ as the estimation of pairwise cost difference is not the best option for two reasons: (i) For $\theta^t(i,j)$ to be accurate, we need to detect S_τ such that S_τ lies on the boundary $\mathcal{V}_i \cap \mathcal{V}_j$. However, it can happen that $\mathcal{V}_i \cap \mathcal{V}_j = \varnothing$, thus producing a constant error. (ii) Even if $\mathcal{V}_i \cap \mathcal{V}_j \neq \varnothing$, finding $S_\tau \in \mathcal{V}_i \cap \mathcal{V}_j$ for all (i,j) pairs can be costly and potentially increase the algorithm complexity. Instead, we observe that the number of unknown parameters in the cost is at most K, thus it suffices to search in a "tree" structure. Specifically, let $\mathcal{G} = (\mathcal{B}, \mathcal{E})$ denote the graph where the node set $\mathcal{B}$ corresponds to the K agent actions and the edge set $\mathcal{E}$ corresponds to the pairwise cost differences $C(i,j) = c_i - c_j$. In addition, we let $\varphi^t(i,j)$ be the "length" assigned to edge $C(i,j)$, which corresponds to the uncertainty of using θ^t to estimate the cost difference. Therefore, the problem of estimating $C(i,j)$ with minimal error becomes finding the *shortest path* between b_i and b_j on the graph $\mathcal{G}$, which can be efficiently handled by Dijkstra's algorithm or the Bellman-Ford algorithm (Ahuja et al., 1990). In summary, the cost difference is estimated by

$$l_{ij} = \text{shortest path}(\mathcal{G}, i, j), \quad \widehat{C}^t(i,j) = \sum_{(i',j')\in l_{ij}} \theta^t(i',j'),$$
$$I_c^t(i,j) = \sum_{(i',j')\in l_{ij}} \varphi^t(i',j'), \tag{3.6}$$

where $I_c^t(i,j)$ is the confidence interval for the pairwise cost estimator $\widehat{C}^t(i,j)$. Moreover, we can easily check that $\widehat{C}^t(i,j) = -\widehat{C}^t(j,i)$ since l_{ij} is the same as l_{ji} and $\theta^t(i,j) = -\theta^t(j,i)$ by definition. The validity of $I_c^t(i,j)$ serving as a confidence interval for $\widehat{C}^t$ is proved in Lemma F.7.

Solving for Opt-LP and Choosing the Best Arm. Note that we have the estimator $\widehat{q}_k^t$ for the belief distributions and

the estimator $\widehat{C}^t$ for the pairwise cost differences. We are now able to solve the linear program (LP) given in (3.3) for the best scoring rule corresponding to agent action b_k. To incorporate the optimism principle, we relax the constraint of (3.3) using the confidence interval I_q^t, I_c^t obtained before. Specifically, for agent action b_k, we consider the following optimistic linear program Opt-LP$_k$,

$$\begin{aligned} \text{Opt-LP}_k: \quad & \max_{S\in\mathcal{S}} \quad \langle u, \widehat{q}_k^t\rangle_\Sigma + B_u I_q^t(k) - v, \\ & \text{s.t.} \quad \left|v - \widehat{v}_S^t(k)\right| \le B_S \cdot I_q^t(k), \qquad (3.7) \\ v - \widehat{v}_S^t(i) \ge \widehat{C}^t(k,i) - & \left(I_c^t(k,i) + B_S \cdot I_q^t(i)\right), \forall i \neq k. \end{aligned}$$

Here, the optimization goal is to maximize the principal's profit under the agent's best response b_k, where we add $B_u I_q^t(k)$ to upper bound the true value with high probability. The first constraint actually constructs a confidence interval $B_S I_q^t(k)$ for the payment v, while the second constraint relax the initial boundary condition in (3.3) using I_c^t and I_q^t. The relaxations in the constraints guarantee that Opt-LP$_k$ is optimistic with high probability, as is verified in Lemma F.1. Suppose that the optimal value and solution for Opt-LP$_k$ are $h_{\text{LP}}^t(k)$ and $S_{\text{LP},k}^t$, respectively. If Opt-LP$_k$ is infeasible, we just let $h_{\text{LP}}^t(k) = \langle u, \widehat{q}_k^t\rangle_\Sigma - \mathbb{E}_{\sigma\sim\widehat{q}_k}\widetilde{S}_k(\sigma) + (B_S + B_u)I_q^t(k)$ and $S_k^t = \widetilde{S}_k$. Next, by viewing the problem as a K-arm bandit, we choose the best "arm" that maximizes the principal's optimistic expected profit, $k_t^* = \text{argmax}_{k\in[K]} h_{\text{LP}}^t(k)$. For simplicity, we let $S_t^* := S_{\text{LP},k_t^*}^t$.

Constructing Conservative Scoring Rules. However, it happens that we can sometimes be "overoptimistic" about the agent's best response. That is, as we relax the boundary constraint in Opt-LP$_k$, the agent might not respond action k_t^* under scoring rule S_t^*. This fact suggests that we ought to be *conservative* to a certain degree in the design of scoring rule. In particular, we consider the conservative scoring rule as, $S_t = (1-\alpha_t)S_t^* + \alpha_t \widetilde{S}_{k_t^*}$. The intuition is that since the agent strictly prefers the action k_t^* under the scoring rule $\widetilde{S}_{k_t^*}$, combining S_t^* with $\widetilde{S}_{k_t^*}$ increases the agent's relative preference of choosing action k_t^*. In Lemma F.2, we show that with a choice of $\alpha_t = \mathcal{O}(t^{-1/3})$, we can guarantee with high probability that the agent would response with action k_t^* under the conservative scoring rule S_t. This also means the optimism (reflected by the agent's choice of action k_t^*) is guaranteed.

Refining Parameter Estimations. Once S_t is deployed, we consider two types of outcomes. If the agent responds with action $k_t = k_t^*$, our estimates of agent's decision boundary $\mathcal{V}_{k_t^*}$ is good enough, and we can simply proceed to the next round. Otherwise, the agent responds with another action $k_t \neq k_t^*$, and we need to improve our estimations on $\mathcal{V}_{k_t^*}$ by updating $\widehat{q}^t, \widehat{C}^t$ in order to successfully induce the desired action k_t^* in the future. Specifically, due to the special conservative construction of S_t, it suffices to update the boundary of $\mathcal{V}_{k_t^*}$ that lies between S_t and $\widetilde{S}_{k_t^*}$. To achieve this, we conduct binary search on the segment connecting $\widetilde{S}_{k_t^*}$ and S_t and locate the first switch point where the agent's best response changes from action k_t^* to another action.

To be specific, in the process of searching the switching point, there are two useful information that we can utilize: Firstly, for the boundary of $\mathcal{V}_{k_t^*}$ that lies on the segment $(\widetilde{S}_{k_t^*}, S_t)$, we induce the two actions (i,j) that this boundary separates at least once, thus their belief distribution estimator $\widehat{q}_i^t$ and $\widehat{q}_j^t$ will be updated. Secondly, and more importantly, this switching point S lies near the boundary, we can refine the cost difference by $\widehat{C}(i,j) \approx \langle \widehat{q}_i^t - \widehat{q}_j^t, S\rangle_\Sigma$ in (3.2). More details are available in Algorithm 3. Thus, this binary search can refine both the belief estimator and the cost difference estimator, which leads to a better estimation of $\mathcal{V}_{k_t^*}$. Notably, such a binary search procedure achieves sufficient accuracy within logarithmic searching time.

4. Theoretical Results

In this subsection, we provide the learning result for the OSRL-UCB algorithm.

Theorem 4.1. Under Assumption 3.1 on the action-informed oracle, with $\alpha_t = \min\{Kt^{-1/3}, 1\}$, the OSRL-UCB algorithm 2 has regret

$$\text{Reg}(T) = \widetilde{\mathcal{O}}\big((B_S + B_u)B_S^2\varepsilon^{-2} \cdot K^2 C_{\mathcal{O}} \cdot T^{2/3}\big),$$

where B_S and B_u bound the magnitudes of the scoring rule and the utility function, respectively, ε is the marginal profit gain given by the action informed oracle, K is the number of the agent's actions, and $C_{\mathcal{O}}$ is the cardinality of the observation set.

Proof. We defer the detailed proof to §F. □

Here, we use $\widetilde{\mathcal{O}}$ to omit logarithmic factors. The regret depends quadratically on the agent's action number and linearly on the cardinality of the observation set. Notably, the regret is independent of the size of the hidden state $|\Omega|$. In addition, we achieve a $\mathcal{O}(T^{2/3})$ sublinear rate of regret in terms of the principal's accumulative profit for eliciting information under the scoring rule framework. Such a result is achieved with a mild assumption of an action-informed oracle that provides a set of scoring rules with marginal profit gain for the agent that induce all the agent's actions. We do not assume the learner to have sufficient knowledge about the other strategic player(s) in contrast to many existing works (Balcan et al., 2015; Guo et al., 2022; Wu et al., 2022). In addition, we only assume the principal to have knowledge of her utility and can observe the agent's action

choice. For discussion of the these two assumptions, we refer the readers to the footnote in §2.

For the action-informed oracle with a set of foreknown scoring rules, these foreknown scoring rules do not need to be optimal in each section $\mathcal{V}_k$. They can even be obtained through random sampling from β-strongly proper scoring rules (See Example 3.4) for general setting, or discovered in a linear scoring rule class (See Example 3.5) if the marginal information gain is strictly decaying. We also give the following corollaries that characterize the regret combined with the effort to find an action-informed oracle.

Corollary 4.2 (Regret with oracle from random sampling)**.** Let $\widetilde{\mathcal{V}}_k = \{S \in \mathcal{S}_\beta \mid g(k,S) \geq g(k',S) + \kappa, \forall k' \neq k, k' \in [K]\}$ where $\mathcal{S}_\beta \in \mathcal{S}$ is the class of β-strongly proper scoring rules, and suppose $\mathrm{Vol}(\widetilde{\mathcal{V}}_k) \geq \eta\, \mathrm{Vol}(\mathcal{S}_\beta)$ for $k \in [K]$. Running the oracle acquisition process in Example 3.4 for T^γ rounds before deploying the OSRL-UCB algorithm for $T - T^\gamma$ rounds, the online regret is bounded by

$$\begin{aligned}\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\big((d_2 d_1^2 \beta)^{-2} \cdot KM \cdot T^{2/3}\big) \\ + \mathcal{O}(KT\exp(-T^\gamma \eta/M) + T^\gamma),\end{aligned}$$

where $d_1 = \min_{i\neq j, \forall (i,j)\in[M]} \|\sigma_i - \sigma_j\|_\infty$, $d_2 = \min_{k'\neq k,(k,k')\in[K]} \max_{i\in[M]} [q_k(i) - q_{k'}(i)]$, and M is the cardinality of Σ.

And also we characterize the regret for the action-informed oracle obtained by linear scoring rule.

Corollary 4.3 (Regret with oracle from linear scoring rule)**.** Suppose the model assumption that the marginal information gain is strictly decaying in Example 3.5 holds. By running the oracle acquisition process in Example 3.5 for $\mathcal{O}(K\log_2(\varepsilon^{-1}))$ rounds and the OSRL-UCB algorithm for the remaining rounds, the online regret is bounded by

$$\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\big(\varepsilon^{-2} \cdot K^2 C_{\mathcal{O}} \cdot T^{2/3}\big) + \mathcal{O}(K\log_2(\varepsilon^{-1})),$$

where $\varepsilon = \epsilon u_1/4b^2$, and ϵ, u_1, b are constants defined in Example 3.5.

Corollaries Corollary 4.2 and Corollary 4.3 both provide regret bound without any using of oracle. Specifically, Corollary 4.2 considers a more general framework under the assumption of lower bounded action section volume while Corollary 4.3 assumes marginal information decay, which is commonly seen in real world practice. Specifically, Corollary 4.2 shows that by random sampling for T^γ rounds where $0 < \gamma < 2/3$, it suffices for the principal to have $\widetilde{\mathcal{O}}(T^{2/3})$ regret. In addition, γ can be significantly small since the second term diminishes exponentially on T^γ. In addition, Corollary 4.3 shows that running constant number of additional rounds in the oracle acquisition process does not deteriorate the regret bound.

Following the discussion in Jin et al. (2018), our algorithm also has PAC guarantee as the following.

Corollary 4.4 (PAC guarantee)**.** For every $\zeta > 0$, the OSRL-UCB algorithm with action informed oracle finds a ζ-optimal scoring rule using $\widetilde{\mathcal{O}}(\varepsilon^{-6}K^6 C_{\mathcal{O}}^3 \zeta^{-3})$ samples.

Zhu et al. (2022) provides an $\mathcal{O}(T^{2/3})$ regret lower bound for the online learning problem towards the optimal contract. Despite that standard contract design is a special case of our model, their setting is different from ours in that the agent may have possibly infinitely many actions. So the regret lower bound of our problem remains an open question. To close this gap, we believe the key question remains to be answered is whether the decision boundary of $\mathcal{V}_{k^*}$ can be determined efficiently. That is, even if the best action k^* is known, the learner is still unable to solve the optimal scoring rule from Opt-LP$_{k^*}$ without enough knowledge about $\mathcal{V}_{k^*}$. For now, we are able to construct a class of instances where such boundary of $\mathcal{V}_{k^*}$ can determine with binary search and thus avoid the costly learning of q_k for every $k \in [K]$. However, it still remains unclear if these efficient search techniques can possibly be generalized to arbitrary instances — a definitive answer should close up the regret lower and upper bound of this problem. We leave this intriguing direction to future work.

5. Conclusion

We study the problem of incentivizing information acquisition through proper scoring rules under the principal-agent framework with information asymmetry. We propose the OSRL-UCB algorithm and show that with a mild oracle assumption, it achieves a $\mathcal{O}(K^2 C_{\mathcal{O}} T^{2/3})$ sublinear regret. Future direction includes establishing regret lower bound and extensions to the contextual and dynamic settings.

References

Ahuja, R. K., Mehlhorn, K., Orlin, J., and Tarjan, R. E. Faster algorithms for the shortest path problem. *Journal of the ACM (JACM)*, 37(2):213–223, 1990.

Alon, T., Dütting, P., and Talgam-Cohen, I. Contracts with private cost per unit-of-effort. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 52–69, 2021.

Amin, K., Rostamizadeh, A., and Syed, U. Learning prices for repeated auctions with strategic buyers. *Advances in Neural Information Processing Systems*, 26, 2013.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

Balcan, M.-F., Blum, A., Haghtalab, N., and Procaccia, A. D. Commitment without regrets: Online learning in stackelberg security games. In *Proceedings of the sixteenth ACM conference on economics and computation*, pp. 61–78, 2015.

Camara, M. K., Hartline, J. D., and Johnsen, A. Mechanisms for a no-regret agent: Beyond the common prior. In *2020 ieee 61st annual symposium on foundations of computer science (focs)*, pp. 259–270. IEEE, 2020.

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., and Gatti, N. Bayesian agency: Linear versus tractable contracts. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 285–286, 2021.

Castiglioni, M., Marchesi, A., and Gatti, N. Designing menus of contracts efficiently: The power of randomization. *arXiv preprint arXiv:2202.10966*, 2022.

Chen, Y. and Yu, F.-Y. Optimal scoring rule design. *arXiv preprint arXiv:2107.07420*, 2021.

Conitzer, V. On stackelberg mixed strategies. *Synthese*, 193(3):689–703, 2016.

Dütting, P., Roughgarden, T., and Talgam-Cohen, I. Simple versus optimal contracts. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 369–387, 2019.

Dutting, P., Roughgarden, T., and Talgam-Cohen, I. The complexity of contracts. *SIAM Journal on Computing*, 50(1):211–254, 2021.

Feng, Z., Podimata, C., and Syrgkanis, V. Learning to bid without knowing your value. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 505–522, 2018.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Golrezaei, N., Javanmard, A., and Mirrokni, V. Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Advances in Neural Information Processing Systems*, 32, 2019.

Grossman, S. J. and Hart, O. D. An analysis of the principal-agent problem. In *Foundations of insurance economics*, pp. 302–340. Springer, 1992.

Guo, W., Jordan, M., and Vitercik, E. No-regret learning in partially-informed auctions. In *International Conference on Machine Learning*, pp. 8039–8055. PMLR, 2022.

Guruganesh, G., Schneider, J., and Wang, J. R. Contracts under moral hazard and adverse selection. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 563–582, 2021.

Haghtalab, N., Lykouris, T., Nietert, S., and Wei, A. Learning in stackelberg games with non-myopic agents. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 917–918, 2022.

Hartline, J., Shan, L., Li, Y., and Wu, Y. Optimal scoring rules for multi-dimensional effort. *working paper*, 2022.

Jagadeesan, M., Wei, A., Wang, Y., Jordan, M., and Steinhardt, J. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34:3323–3335, 2021.

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

Laffont, J.-J. and Martimort, D. The theory of incentives. In *The Theory of Incentives*. Princeton university press, 2009.

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

Letchford, J., Conitzer, V., and Munagala, K. Learning and approximating the optimal strategy to commit to. In *International symposium on algorithmic game theory*, pp. 250–262. Springer, 2009.

Li, Y., Hartline, J. D., Shan, L., and Wu, Y. Optimization of scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 988–989, 2022.

Lui, D., Markov, S., and Tamayo, A. Equity analysts and the market's assessment of risk. *Journal of Accounting Research*, 50(5):1287–1317, 2012.

Mardia, J., Jiao, J., Tánczos, E., Nowak, R. D., and Weissman, T. Concentration inequalities for the empirical distribution of discrete distributions: beyond the method of types. *Information and Inference: A Journal of the IMA*, 9(4):813–850, 2020.

Marecki, J., Tesauro, G., and Segal, R. Playing repeated stackelberg games with unknown opponents. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pp. 821–828, 2012.

Mas-Colell, A., Whinston, M., and Green, J. *Microeconomic Theory*. Oxford student edition. Oxford University Press, 1995. ISBN 9780195102680. URL https://books.google.com/books?id=sQGDQgAACAAJ.

Myerson, R. B. Incentive compatibility and the bargaining problem. *Econometrica: journal of the Econometric Society*, pp. 61–73, 1979.

Neyman, E., Noarov, G., and Weinberg, S. M. Binary scoring rules that incentivize precision. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 718–733, 2021.

Oesterheld, C. and Conitzer, V. Minimum-regret contracts for principal-expert problems. In *Web and Internet Economics: 16th International Conference, WINE 2020, Beijing, China, December 7–11, 2020, Proceedings 16*, pp. 430–443. Springer, 2020.

Papireddygari, M. and Waggoner, B. Contracts with information acquisition, via scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, pp. 703–704, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391504. doi: 10.1145/3490486.3538261. URL https://doi.org/10.1145/3490486.3538261.

Savage, L. J. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Smith, S. A. *Contract theory*. OUP Oxford, 2004.

von Stackelberg, H. *The Theory of the Market Economy*. William Hodge, 1952. URL https://books.google.com/books?id=fjIAtQEACAAJ.

Wu, J., Zhang, Z., Feng, Z., Wang, Z., Yang, Z., Jordan, M. I., and Xu, H. Sequential information design: Markov persuasion process and its efficient reinforcement learning. *arXiv preprint arXiv:2202.10678*, 2022.

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.

Zu, Y., Iyer, K., and Xu, H. Learning to persuade on the fly: Robustness against ignorance. EC '21, pp. 927–928, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385541. doi: 10.1145/3465456.3467593. URL https://doi.org/10.1145/3465456.3467593.

A. More Discussions on Related Work

Optimal Scoring Rules. This paper builds on a set of literature on optimizing scoring rules. Several papers consider the model with multiple levels of effort. Neyman et al. (2021) design outcome-optimal scoring rule under a binary-state model, and for integral levels of effort, where effort levels represent the number of samples drawn and are informationally ordered. Hartline et al. (2022) design effort-optimal scoring rule, under a multi-dimensional binary-state model, and each dimension of the effort corresponds to one of the independent multi-dimensional state. In contrast, our paper considers a general state space, with multiple levels of effort not necessarily ordered or independent. Li et al. (2022); Papireddygari & Waggoner (2022); Chen & Yu (2021) design optimal scoring rule for a binary effort model, which is different to our multiple-effort-level model. Oesterheld & Conitzer (2020) design regret-optimal scoring rule for multiple agents in a single round when the information structure is unknown to the principal, while our model only has one agent, and our learning algorithm achieves diminishing regret over multiple rounds. Also, all the papers mentioned above model the state as exogenously given, while in our paper, the prior of the state can potentially be affected by agent's endogenous action.

The Principal-Agent Problem. Our model of information acquisition can be viewed as a class of principal-agent problem, which has been established as a crucial branch of economics known as the contract theory (Grossman & Hart, 1992; Smith, 2004; Laffont & Martimort, 2009). Driven by an accelerating trend of contract-based markets deployed to Internet-based applications, the principal-agent problem recently started to receive a surging interest especially from the computer science community (Dütting et al., 2019; Dutting et al., 2021; Guruganesh et al., 2021; Alon et al., 2021; Castiglioni et al., 2021; 2022). As pointed out by Alon et al. (2021), this includes online markets for crowdsourcing, sponsored content creation, affiliate marketing, freelancing and etc. The economic value of these markets is substantial and the role of data and computation is pivotal. Different from the classic contract design problems, we focus on the design of contracts (i.e., scoring rules) that optimally elicit the information acquired by the agents at some cost.

Online Learning in Strategic Environment. More broadly, our work add to the literature on online learning in strategic environments, which has gained popularity in recent years. In particular, the online learner's utility at each round is determined by both her own action and the strategic response(s) of other player(s) in certain repeated game, and the typical goal of this learner is to find her optimal strategy under some equilibrium These repeated games are adopted from the influential economic models including, but not limited to, the Stackelberg (security) game (Marecki et al., 2012; Balcan et al., 2015; Haghtalab et al., 2022), auction design (Amin et al., 2013; Feng et al., 2018; Golrezaei et al., 2019; Guo et al., 2022), matching (Jagadeesan et al., 2021), contract design (Zhu et al., 2022), Bayesian persuasion (Castiglioni et al., 2020; Zu et al., 2021; Wu et al., 2022). Our model can be viewed as a generalized information elicitation problem in an online learning setup. To our best knowledge, there is no previous work that considered any similar learning problem. In addition, we remark that, in many of these existing works, e.g., Balcan et al. (2015); Guo et al. (2022); Wu et al. (2022), the learner is assumed to have sufficient knowledge about the other strategic player(s), and her uncertainty is regarding the environment or her own utility. Assumptions of such kind can significantly simplify the problem into the standard online learning problems, once the learner can almost predict the best response of other player(s). Our work, however, does not make any of such assumption and the most challenging part of our learning algorithm design is indeed to ensure the desired agent response under uncertainty. Camara et al. (2020) considers the online mechanism design problem, where both the principal and the agent may learn over time from the state history. However, our paper assumes the agent is myopic in each round.

B. Contract Design as a Special Case of Scoring Rule Design

In this section, we compare the contract design framework to the scoring rule design framework and reduces the standard contract designing problem to a special case of the scoring rule designing problem. In a contract designing problem, we refer to Ω as the outcome space. Consider the following contract designing problem:

Contract designing problem in the principal-agent framework

At the t-th round, the principal and the agent play as the following:

1. The principal announces a contract $C_t : \Omega \to \mathbb{R}_+$ to the agent.
2. Based on C_t, the agent chooses an action $b_{k_t} \in \mathcal{B}$ indexed by k_t and bears a cost $c_{k_t} \geq 0$. The action b_{k_t} can be observed by the principal, but the cost c_{k_t} is private to the agent.

3. The stochastic environment then selects an outcome $\omega_t \in \Omega$ according to $p(\omega_t \mid b_{k_t})$. The outcome ω_t is revealed as observation, but the generating process $p(\omega_t \mid b_{k_t})$ is private to the agent.
4. The principal makes a decision $a_t \in \mathcal{A}$ based on ω_t.
5. In the end, the principal obtains her utility $u(a_t, \omega_t)$ and pays the agent by $C_t(\omega_t)$.

The difference between this contract designing problem and the scoring rule designing problem is that ω_t is revealing, and the agent's action influences the principal's utility only through her action choice without giving any report. We remark that we can replace $u(a_t, \omega_t)$ by $u(\omega_t) = u(a^*(\omega_t), \omega_t)$ if the principal knows about the utility function and always takes the best action. In this contract design problem, the agent has an action policy $\pi : \mathcal{C} \to [K]$, where $\mathcal{C}$ is the contract space. The principal targets at designing the optimal contract that maximizes her profit, i.e., utility minus payment, subject to the agent's best response given by maximizing the agent's profit, i.e., payment minus cost. The Stackelberg game for this contract designing problem can be formulated as

$$\begin{aligned} \max_{C \in \mathcal{C}} \quad & \mathbb{E}_{\omega \sim p(\cdot \mid b_{\pi^*(C)})} \left[u(\omega) - C(\omega) \right], \\ \text{s.t.} \quad & \pi^*(C) \in \operatorname*{argmax}_{k \in [K]} \mathbb{E}_{\omega \sim p(\cdot \mid b_k)} C(\omega) - c_k, \end{aligned} \tag{B.1}$$

In the sequel, we aim to show in the scoring rule designing problem: (i) If the hidden state is perfectly revealing, i.e., $o_t = \omega_t$ as the agent's observation after taking her action, there exists a class of scoring rules such that the above contract designing problem is equivalent to the scoring rule designing problem. (ii) Using proper scoring rules, the principal's optimal profit under the scoring rule framework is no less than the optimal profit under the contract framework.

To show (i), consider the scoring rule class

$$\mathscr{S}^C = \{ S \in \mathscr{S} \mid S(\widehat{\sigma}, \omega) = \mathbb{1}(\widehat{\sigma} = e_\omega) \cdot C(\omega), \forall C \in \mathcal{C} \},$$

where $e_{\omega'}(\omega) = \mathbb{1}(\omega = \omega') \in \Delta(\Omega)$. Even though $S \in \mathscr{S}^C$ might not be a proper scoring rule, the agent will always be truth-telling, i.e., $\widehat{\sigma} = e_\omega$, since only by telling the truth can she gains nonzero payment. Therefore, this hidden state ω_t is also revealed to the principal through the agent's report. The Stackelberg game in (2.2) in the scoring rule problem can therefore be written as,

$$\begin{aligned} \max_{S \in \mathscr{S}^C} \quad & \mathbb{E}_{\omega \sim p(\omega \mid b_{k^*(S)})} \left[u(\omega) - S(e_\omega, \omega) \right], \\ \text{s.t.} \quad & k^*(S) \in \operatorname*{argmax}_{k \in [K]} \mathbb{E}_{\omega \sim p(\cdot \mid b_k)} S(e_\omega, \omega) - c_k, \end{aligned} \tag{B.2}$$

By noting that $S(e_\omega, \omega) = C(\omega)$, we have that (B.2) and (B.1) are actually the same problem. We thus conclude that the contract designing problem is perfectly reduced to this scoring rule designing problem.

We also remark that even if the hidden state is perfectly revealing, the principal need not be aware in advance. By sticking to a proper scoring rule, the agent always tells the truth. Moreover, using the revelation principle stated in Lemma 2.2, for any $S \in \mathscr{S}^C$, there always exists a proper scoring rule $\widetilde{S} \in \mathcal{S}$ that generates the same expected payment $\mathbb{E}_{\omega \sim p(\cdot \mid b_k)} \widetilde{S}(e_\omega, \omega) = \mathbb{E}_{\omega \sim p(\cdot \mid b_k)} C(\omega)$ for the agent, even though $\widetilde{S}(e_\omega, \omega)$ might not be equal to $C(\omega)$ pointwise. Therefore, statement (ii) is also justified and we conclude that the (proper) scoring rule framework has more power than the contract framework by asking one more question about the agent's belief.

C. More Details on Preliminaries

C.1. Eliciting Information via Proper Scoring Rules

In this section, we provide a formal argument on the revelation principal in our model. That is, it is without loss of generality to only design the proper scoring rules under which the agent is encouraged to be truth-telling.

Definition C.1 (Proper scoring rule)**.** A scoring rule $S : \Delta(\Omega) \times \Omega \to \mathbb{R}_+$ is proper if, for any belief $\sigma \in \Delta(\Omega)$ and any reported posterior $\widehat{\sigma} \in \Delta(\Omega)$, we have $\mathbb{E}_{\omega \sim \sigma} S(\widehat{\sigma}, \omega) \leq \mathbb{E}_{\omega \sim \sigma} S(\sigma, \omega)$. In addition, if the inequality holds strictly for any $\widehat{\sigma} \neq \sigma$, the scoring rule S is strictly proper.

Let S be a proper scoring rule and fix the agent's policy $\mu(\cdot)$ for action selection. For a reporting scheme ν and any true belief σ, it follows from definition 2.1 that

$$g^{\mu,\nu}(S) = \mathbb{E}_{\omega\sim\sigma} S(\nu(S,\sigma,\mu(S)),\omega) - c_{\mu(S)} \leq \mathbb{E}_{\omega\sim\sigma} S(\sigma,\omega) - c_{\mu(S)}.$$

Therefore, the agent's expected payment is maximized by always being truth-telling about her belief under the class of proper scoring rule. In the following, we let $S(\widehat{\sigma},\sigma) = \mathbb{E}_{\omega\sim\sigma} S(\widehat{\sigma},\omega)$. To give an example of proper scoring rules, let us consider the binary hidden state space $\Omega = \{0,1\}$ where the class of proper scoring rules admits the Schervish representation (Gneiting & Raftery, 2007), i.e., $S(p,1) = G(p) + (1-p)G'(p)$ and $S(p,0) = G(p) - pG'(p)$ where $p \in [0,1]$ and $G : [0,1] \to \mathbb{R}_+$ is a convex function. Intuitively, the expected payment of a proper scoring rule S given belief σ and report p is $S(p,\sigma) = G(p) + (\sigma - p)G'(p)$, which corresponds to the supporting line of G at p. In this example, the convexity of G guarantees that $S(p,\sigma) = G(p) + (\sigma - p)G'(p) \leq G(\sigma) = S(\sigma,\sigma)$. Moreover, the next observation in Lemma 2.2 suggests that restricting to the class of proper scoring rules does not incur any loss of generality for the principal's purpose.

Lemma C.2 (Restatement of the revelation principle)**.** There exists a proper scoring rule S^* that is an optimal solution to (2.1) if the agent is truth-telling under any proper scoring rule.

Proof. We first prove that for any scoring rule S such that $\|S\|_\infty \leq B_S$, there always exists a *proper* scoring rule $S'(\widehat{\sigma},\omega) = S(\nu^*(S,\widehat{\sigma},k),\omega)$ such that they make the same payment to the agent for any $\sigma \in \Delta(\Omega)$ and any agent's action choice. To prove that S' is a proper scoring rule with norm bounded by B_S, we have $B_S \geq S(\nu^*(S,\widehat{\sigma},k),\omega) = S'(\widehat{\sigma},\omega) \geq 0$ and

$$S'(\widehat{\sigma},\sigma) = S(\nu^*(S,\widehat{\sigma},k),\sigma) \leq S(\nu^*(S,\sigma,k),\sigma) = S'(\sigma,\sigma).$$

The fact that S' makes the same payment can be verified by plugging in $\widehat{\sigma} = \sigma$ in the definition of S' since the agent is truth-telling under proper scoring rules, and taking expectation with respect to $\omega \sim \sigma$, i.e., $S'(\sigma,\sigma) = S(\nu^*(S,\sigma,k),\sigma)$, which proves the first part.

Secondly, we prove that encouraging the agent to report the real belief σ makes the principal's revenue nondecreasing. Note that

$$\max_{\iota} \mathbb{E}_{\omega\sim\sigma,\sigma\sim q_k}\left[u(\iota(S,\nu^*(S,\sigma,k),k),\omega)\right] \leq \max_{\iota} \mathbb{E}_{\omega\sim\sigma,\sigma\sim q_k}\left[u(\iota(S,\sigma,k),\omega)\right] = \mathbb{E}_{\omega\sim\sigma,\sigma\sim q_k}\left[u(a^*(\sigma),\omega)\right],$$

where $a^*(\sigma) = \operatorname{argmax}_{a\in\mathcal{A}} \mathbb{E}_{\omega\sim\sigma} u(a,\omega)$. Here, the inequality holds by noting that $\nu^*(S,\sigma,k)$ is a function of σ, and the equality holds by noting that $\omega \perp\!\!\!\perp (S,k) \,|\, \sigma$. To conclude, by choosing S' instead of S, the payment is exactly the same while the principal's revenue is nondecreasing. Thus, the principal's profit is nondecreasing by choosing S' and there must exist a proper scoring rule that is also an optimal scoring rule. □

Following Lemma 2.2, the principal's optimal scoring rule lies within the class of proper scoring rules $\mathcal{S}$ with bounded norm $\|S\|_\infty \leq B_S$. One concern about the use of proper scoring rules is that being truth-telling might not be the unique maximizer to the agent's utility. However, we note that the class of proper scoring rules is a convex hull with strictly proper scoring rules as the interior. Thus, adding an infinitesimal portion of a *strictly proper* scoring rule to any *proper* scoring rule always yields a *strictly proper* scoring rule. In this sense, we can safely make the assumption that the agent always reports the true posterior under a proper scoring rule. The following table summarizes all the different information types in our model.

Public Info	Agent's Private Info	Principal's Private Info	Delayed Info
$S_t, b_{k_t}, \widehat{\sigma}_t$	$\{c_k\}_{k\in[K]}, \{q_k\}_{k\in[K]}, o_t, \sigma_t$	u	w_t

Table 1. Table for the information types. Here, "Public Info" refers to the information that is observable to both the principal and the agent at each round, "Agent's Private Info" refers to the information that is private to the agent, "Principal's Private Info" refers to the information that is private to the principal, and "Delayed Info" only contains the hidden state ω_t, which is unobservable when round t proceeds but revealed when round t terminates.

D. More Details on Algorithm

In this section, we give a summary of the OSRL-UCB algorithm in the following Algorithm 2. For details, we defer the readers to §3.3.

Algorithm 2 Online Scoring Rule Learning with Upper Confidence Bound (OSRL-UCB)

Input: $\mathcal{S}$, $(\widetilde{S}_1, \cdots, \widetilde{S}_K)$.
while $t \leq T$ **do**
  Estimate the belief distributions $\widehat{q}_k^t$ by (3.4) and the corresponding confidence intervals $I_q^t(k)$ by (3.5);
  Estimate the pairwise cost differences $\widehat{C}^t(i,j)$ and the corresponding confidence intervals $I_c^t(i,j)$ by (3.6);
  Solve Opt-LP$_k$ in (3.7) and obtain the optimal value $h_{\text{LP}}^t(k)$ and the optimal solution $S_{\text{LP},k}^t$ for $k \in [K]$;
  Choose the best arm $k_t^* \leftarrow \text{argmax}_{k\in[K]} h_{\text{LP}}^t(k)$ and let $S_t^* \leftarrow S_{\text{LP},k_t^*}^t$;
  Pay with the conservative scoring rule $S_t \leftarrow \alpha_t \widetilde{S}_{k_t^*} + (1-\alpha_t) S_t^*$, and collect the agent's responding action k_t;
  if $k_t \neq k_t^*$ **then**
    Conduct binary search BS$(S_t, \widetilde{S}_{k_t^*}, k_t^*, t)$ as specified in Algorithm 3;
  end if
  Move on to a new round $t \leftarrow t+1$;
end while

The details of Algorithm 3 are as follow:

Algorithm 3 Binary Search BS$(S_0, S_1, k^*(S_1), t)$

Input: $S_0, S_1, k^*(S_1), t, k_t, \{I_q^t(k)\}_{k\in[K]}$;
Output: t;
if $k_t = k^*(S_1)$ **then**
  Break the binary search algorithm;
end if
Initiate $\lambda_{\min} \leftarrow 0, \lambda_{\max} \leftarrow 1, t_0 \leftarrow t$;
while $\lambda_{\max} - \lambda_{\min} \geq I_q^{t_0}(k_t) \wedge I_q^{t_0}(k^*(S_1))$ **do**
  Start a new round $t \leftarrow t+1$;
  Pick $\lambda \leftarrow (\lambda_{\min} + \lambda_{\max})/2$ as the middle point;
  The principal announces scoring rule $S_t = (1-\lambda)S_0 + \lambda S_1$ and obtain the agent's response k_t;
  if $k_t = k^*(S_1)$ **then**
    Update $\lambda_{\max} \leftarrow \lambda$;
  else
    Update $\lambda_{\min} \leftarrow \lambda$;
  end if
end while

The binary searching algorithm at step t works on the segment connecting two scoring rules S_0 and S_1, where the agent's best response under S_1 is $k^*(S_1)$. The goal of this binary search is to find the first switching point of the agent's best response from k_1 to another action on this segment. The binary searching algorithm keeps updating on $\lambda_{\max}$ and $\lambda_{\min}$ as the candidate interval that contains the first switching point. Note that any $\lambda \in [0,1]$ corresponds to a scoring rule on the segment connecting S_0, S_1. Each time, the algorithm deploy a scoring rule corresponding to $(\lambda_{\min} + \lambda_{\max})/2$ and the candidate interval is thus cut by half. In addition, we consider binary searching to a finite depth m such that the searching error respects the minimal uncertainty of $\widehat{q}_{k'}^t, \widehat{q}_{k_t^*}^t$, i.e., $2^{-m} \leq I_q^t(k') \wedge I_q^t(k_t^*)$ in Algorithm 3. Thus, the binary search achieves sufficient accuracy with logarithmic searching time.

E. Proofs on Action-informed Oracle in Section 3.2

In this section, we prove the claims in Section 3.2.

E.1. Proof of Lemma 3.2

Without loss of generality, we just need to construct a hard instance by considering the case $K = 3$ and $M = 3$. To simplify the discussion, we consider two hidden states $\Omega = \{\omega_1, \omega_2\}$ and consider $B_S = 1$ as the boxing condition for the scoring rules. The idea of constructing the hard instance is as follows:

(i) Make sure that action a_2 is the optimal agent's response for the optimal scoring rule while any scoring rule not inducing the agent to chose a_2 yields a constant regret at least 1.

(ii) Let $\mathcal{V}_2$ be a "single point" on the boundary of $\mathcal{S}$ and show that with a great chance, any algorithm can fail to find the correct scoring rule located in $\mathcal{V}_2$ without the oracle.

Since $\Omega = \{\omega_1, \omega_2\}$, we can equally represent $\sigma_1, \sigma_2, \sigma_3$ by the their mass assigned to ω_1. With a little abuse of notation, we let σ_i denote $\sigma_i(\omega_1)$ and $1 - \sigma_i$ denote $\sigma_i(\omega_2)$. We consider the case $\Sigma = \{\sigma_1 = (0, 1), \sigma_2 = (.5, .5), \sigma_3 = (1, 0)\}$ for $M = 3$. For simplicity, we let $S_i = \mathbb{E}_\omega S(\sigma_i, \omega)$ for $i = 1, 2, 3$. Hence, the set of proper scoring rules is specified by conditions:

$$0 \le S_2 \le \frac{S_1 + S_3}{2}, \quad B_S \ge S_1 \ge 0, \quad B_S \ge S_3 \ge 0. \tag{E.1}$$

We remark that the first condition guarantees that the scoring rule is proper by requiring the curve of $G(\sigma) = \mathbb{E}_\omega[S(\sigma, \omega)]$ to be convex (Gneiting & Raftery, 2007), and the rest are just box conditions. In fact, one can easily construct a proper scoring rule $S(\sigma, \omega)$ for any $S = (S_1, S_2, S_3)$ satisfying the above constraints using the Schervish representation (Gneiting & Raftery, 2007) by fitting a convex function G defined on $[0, 1]$ passing through $(0, S_1), (.5, S_2), (1, S_3)$, and consider the supporting hyperplanes of G at these three points.

To ensure that $\mathcal{V}_2$ is a single point on the boundary in (ii), we let $q_2 - q_1$ and $q_2 - q_3$ be

$$q_2 - q_1 = \beta \cdot (0, -1, 1)^\top, \quad q_2 - q_3 = \beta \cdot (1, 1, -2)^\top,$$

where β should be considered a fixed constant and q_1, q_2, a_3 are what we want to design. These conditions imply the following expression for $\mathcal{V}_2$,

$$\mathcal{V}_2 = \{S \in \mathcal{S} \,|\, (0, -1, 1)S \ge \beta^{-1}(c_2 - c_1), \quad (1, 1, -2)S \ge \beta^{-1}(c_2 - c_3)\}.$$

We further let $e_1 = \beta^{-1}(c_2 - c_1)$ and $e_2 = \beta^{-1}(c_2 - c_3)$. One can easily verify that as long as $e_1 + e_2 = 1$ and $-1/2 \le e_1 \le 0$, $\mathcal{V}_2$ shrinks to a single point given by $S^* = (1, -e_1, 0)$ and the conditions in Equation (E.1) are always satisfied. Hence, we can safely restrict ourselves to the instances specified by parameter e_1 and c_1, c_2, c_3 can be adjusted accordingly. Hence, we are actually considering a single parameter linear system.

To ensure (i), i.e., the optimality of a_2, we first need to ensure that a_2 beats a_1 and a_2 at S^* by some constant $\gamma \ge 1$:

$$\begin{aligned} \langle q_2 - q_1, u - S^* \rangle &\ge \gamma, \\ \langle q_2 - q_3, u - S^* \rangle &\ge \gamma, \end{aligned} \tag{E.2}$$

where $S^* \in \mathcal{V}_2$. Note that since $q_2 - q_1$ does not lie on the same line with $q_2 - q_3$, we can always choose u such that (E.2) holds regardless of what S^* is. For any $S \in \mathcal{V}_1$, we can verify that for the principal's profit,

$$\langle q_2, u - S^* \rangle - \langle q_1, u - S \rangle = \langle q_2 - q_1, u - S^* \rangle + \langle q_1, S - S^* \rangle \ge \gamma - 1,$$

where the last inequality is a direct result of (E.2) and the fact that each entry of S should be within $[0, 1]$. The same holds for $S \in \mathcal{V}_3$. Hence, it suffices to find an instance satisfying (E.2) with $\gamma = 2$.

Construction of the hard instance: Following the above discussion, we construct the instance by letting $\gamma = 2$, $\beta = 1/16$, and

$$\begin{aligned} q_1 &= (3/4, 3/16, 1/16), \\ q_2 &= (3/4, 1/8, 1/8), \\ q_3 &= (11/16, 1/16, 1/4) \\ u &= (96, 0, 32) + (1, -e_1, 0). \end{aligned}$$

Here, the only thing unknown is the cost c_1, c_2, c_3, where the effect of the unknown cost is encoded in $e_1 \in [-1/2, 0]$, which only affects the position of S^*.

Since we are considering a one parameter linear system with the second coordinate of S^* undetermined, it is without loss of generality to consider finding the optimal scoring rule on the line $l = (1, \lambda, 0)$ where $0 \le \lambda \le 1/2$. Suppose we assign scoring rule $S^t = (1, b^t, 0)$ at step t and the agent responses $a^t \in \mathcal{A} = \{a_1, a_2, a_3\}$. In fact, for any $b^t < e_1$, the agent will response a_1 and for any $b^t > e_1$, the agent will response a_3. Then, we assert that any online strategy of the principal can result in linear regret since no algorithm can decide the exact position of e_1 under the feedback of the agent's actions only (Perhaps the best one can do is binary searching for e_1, which does not help either). In addition, any $b^t \neq e_1$ yields a suboptimality at least 1 at each round. Moreover, we remark that letting $\mathcal{V}_2$ shrink to a point still makes S^* the optimal scoring rule by the tie-breaking rule which is in favor of the principal. Nevertheless, one can always make $\mathcal{V}_2$ an interval around e_1 as small as possible. Hence, we conclude that for any algorithm, there always exists a hard instance that yield a linear regret Reg $\ge \Omega(T)$. We finish the proof of Lemma 3.2.

E.2. Proof of Example 3.4

To ensure the scoring rules we sample are ϵ better on the induced action, we sample from the strongly proper scoring rules.

Definition E.1 (Strongly proper scoring rules)**.** A scoring rule $S : \Delta(\Omega) \times \Omega \to \mathbb{R}$ is β-strongly proper if for all $p, q \in \Delta(\Omega)$, $\mathbb{E}_{\omega\sim q}[S(q,\omega)] - \mathbb{E}_{\omega\sim q}[S(p,\omega)] \ge \frac{\beta}{2}\|q-p\|_1^2$.

Example E.2 (Restatement of Example 3.4)**.** Let $d_1 = \min_{1\le i<j\le M}\|\sigma_i - \sigma_j\|_\infty$ and $d_2 = \min_{1\le k<k'\le K}\max_{i\in[M]}[q_k(i) - q_{k'}(i)]$. Since the proper scoring rule class $\mathcal{S}$ has bounded norm, the β-strongly proper scoring class also has bounded volume $\mathrm{Vol}(\mathcal{S}_\beta) < \infty$. Set $\kappa = d_1^2\beta/2$ and let $\widetilde{\mathcal{V}}_k = \{S \in \mathcal{S}_\beta \,|\, g(k,S) \ge g(k',S) + \kappa, \forall k' \neq k, k' \in [K]\}$. We suppose $\mathrm{Vol}(\widetilde{\mathcal{V}}_k) \ge \eta\,\mathrm{Vol}(\mathcal{S}_\beta)$ for $k \in [K]$. Initiate $\mathcal{M} = \varnothing$ as the candidate set. Each time, we randomly sample a β-strongly proper scoring rule $S \in \mathcal{S}_\beta$ and obtain the agent's best response $k^*(S)$ with respect to S. Let $e_i(\sigma,\omega) = \mathbb{1}(\sigma = \sigma_i)$ for $i \in [M]$. By setting $\kappa = d_1^2 \cdot \beta/2$, we deploy $\{S - \kappa e_i\}_{i\in[M]}$ and see if the agent still responds $k^*(S)$. If so, Let $\mathscr{S} = \mathscr{S} \cup \{S\}$; if not, reject S. After $\mathcal{O}(M\eta^{-1}K\log K)$ rounds, with high probability, $\mathscr{S}$ serves as a valid action-informed oracle with parameter $\epsilon = d_2 \cdot \kappa$.

Note that the volumn of $\widetilde{\mathcal{V}}_k$ is continuous, decreasing in β, and goes to $\mathrm{Vol}(\widetilde{\mathcal{V}}_k)$ as $\beta \to 0$. Hence, if $\mathcal{V}_k$ has a constant fraction of total volumn, we can find β such that $\widetilde{\mathcal{V}}_k$ has a constant fraction of total volumn.

Proof. First, if $\kappa = d_1^2 \cdot \beta/2$, $\{S - \epsilon e_i\}_{i\in[M]}$ is also proper since S is β-strongly proper.

Next, let $d_2 = \min_{k'}\max_i[q_k(i) - q_{k'}(i)]$. we show that any each time the probability of hitting a scoring rule in $\widetilde{\mathcal{V}}_k$ is at least η. Suppose S is not rejected if for all $i \in [M]$, the agent prefers action k to k': $g(k, S - \kappa e_i) - g(k', S - \kappa e_i) \ge 0$. By properness of $S - \kappa e_i$,

$$\begin{aligned} g(k, S - \kappa e_i) - g(k', S - \kappa e_i) &= \langle S - \kappa e_i, q_k - q_{k'}\rangle - (c_k - c_{k'}) \\ &= \langle S, q_k - q_{k'}\rangle - \kappa(q_k(i) - q_{k'}(i)) - (c_k - c_{k'}) \ge 0. \end{aligned}$$

If this holds for any $i \in [M]$ and $k' \in [K]$, it means $\langle S, q_k - q_{k'}\rangle - (c_k - c_{k'}) \ge \max_{i\in[M]}\kappa(q_k(i) - q_{k'}(i)) \ge \kappa d_2$. One the other hand, if S is in set $\widetilde{\mathcal{V}}_k$, it never gets rejected.

After $\frac{c}{\eta}K\log K$ rounds, the probability that one action is not induced is $(1-\eta)^{\frac{c}{\eta}K\log K} = (\frac{1}{K})^{cK}$. Taking a union bound, the probability that any action is not induced is at most $(\frac{1}{K})^{cK-1}$. Setting c large enough, this probability will be sufficiently small. To conclude, after $\mathcal{O}(M\eta^{-1}K\log K)$ rounds, we get an oracle with parameter $\epsilon = d_2 \cdot \kappa$.

□

E.3. Proof of Example 3.5

Example E.3 (Restatement of Example 3.5)**.** Suppose that these K signals are sorted in the increasing order of the cost. Define $u_k = \mathbb{E}_{\omega\sim\sigma,\sigma\sim q_k}[u(a^*(\sigma),\omega)]$ and suppose $u_k > 0$. We assume the marginal information gain is strictly decaying,

i.e., there exists a $\epsilon > 0$ such that

$$\frac{u_K - u_{K-1}}{c_K - c_{K-1}} > \epsilon, \quad \text{and}$$
$$\frac{u_k - u_{k-1}}{c_k - c_{k-1}} - \frac{u_{k+1} - u_k}{c_{k+1} - c_k} > \epsilon, \quad \forall k = 2, \cdots, K-1.$$

Moreover, we assume that $(u_2 - u_1)/(c_2 - c_1) \leq b$. The principal sets the scoring rule as $S(\sigma, \omega) = \lambda u(a^*(\sigma), \omega)$ and conducts binary search on $\lambda \in [0, 2/\epsilon]$. Specifically, the binary searches are iteratively conducted on all the segments (λ_1, λ_2) with $k^*(\lambda_1 u) \neq k^*(\lambda_2 u)$, where λ_1, λ_2 are neighboring points on $[0, 1/\epsilon]$ that are previously searched. With the maximal searching depth $m = \lceil \log_2(2b(b-\epsilon)/\epsilon^2) \rceil$, we can identify all the agent's actions. Suppose that $(\lambda_1^k, \lambda_2^k)$ is the largest segment with λ_1^k and λ_2^k searched before and $k^*(\lambda_1^k u) = k^*(\lambda_2^k u) = k$. By letting $\widetilde{S}_k = (\lambda_1^k + \lambda_2^k)u/2$, we obtain an oracle with $\varepsilon = \epsilon u_1/4b^2$. The procedure takes $\mathcal{O}(K \log_2(\varepsilon^{-1}))$ rounds.

Proof. First, we show that by setting $\lambda = \lambda_i = (c_i - c_{i-1}) / (u_i - u_{i-1})$ for any $i = 2, \ldots, K$, the agent is indifferent between taking actions i and $i-1$. In addition, we let $\lambda_1 = 0$ and $\lambda_{K+1} = 2/\epsilon$. For any $\lambda \in (\lambda_i, \lambda_{i+1})$, the agent best responds with action b_i. In addition, $\lambda_2 = (c_2 - c_1)/(u_2 - u_1) \geq 1/b$ and $\lambda_K = (c_K - c_{K-1})/(u_K - u_{K-1}) \leq 1/\epsilon$. Thus, by searching in $\lambda \in [0, 2/\epsilon]$, all the actions are guaranteed to be induced by this linear contract.

Since the sequence $(u_i - u_{i-1}) / (c_i - c_{i-1})$ is strictly decaying, λ_i is strictly increasing and $\lambda_i - \lambda_{i-1}$ is lower bounded by $1/(b-\epsilon) - 1/b = \epsilon/b(b-\epsilon)$, where the lower bound is reached by $\lambda_2 = 1/b$ and $\lambda_3 = 1/(b-\epsilon)$. Combined with the fact that search is conducted on $[0, 2/\epsilon]$, we get the maximal searching depth $m = \lceil \log_2(2b(b-\epsilon)/\epsilon^2) \rceil$ such that the neighboring searching points has a gap no more than $\epsilon/2b(b-\epsilon)$. Therefore, using the fact that $\lambda_i - \lambda_{i-1} \geq \epsilon/b(b-\epsilon)$, we can guarantee that all the actions are induced by this binary search.

For the chosen λ_1^k and λ_2^k such that $k^*(\lambda_1^k u) = k^*(\lambda_2^k u) = k$, we can guarantee that $\lambda_k \leq \lambda_1^k \leq \lambda_k + \epsilon/2b(b-\epsilon)$ and $\lambda_{k+1} - \epsilon/2b(b-\epsilon) \leq \lambda_2^k \leq \lambda_{k+1}$. Hence, $\lambda_k + \epsilon/4b(b-\epsilon) \leq (\lambda_1^k + \lambda_2^2)/2 \leq \lambda_{k+1} - \epsilon/4b(b-\epsilon)$. Therefore, by setting $\lambda = (\lambda_1^k + \lambda_2^k)/2$, the marginal profit for choosing b_k over other actions is lower bounded by $\epsilon/4b(b-\epsilon) \cdot u_k \geq \epsilon/4b(b-\epsilon) \cdot u_1 \geq \epsilon u_1/4b^2$. Hence, we obtain the oracle with $\varepsilon = \epsilon u_1/4b^2$.

□

F. Proof of Theorem 4.1

Proof of Theorem 4.1. Here we give a high-level analysis of the $\widetilde{\mathcal{O}}(K^2 T^{-2/3})$ regret for Algorithm 2. We consider two types of events in T rounds, $\mathbb{1}(k_t = k_t^*)$ or $\mathbb{1}(k_t \neq k_t^*)$. In the first type of events, by the construction of S_t, the suboptimality gap at each round can be decoupled into components contributed by $(1-\alpha_t)S_t^*$ or $\alpha_t \widetilde{S}_{k_t^*}$. The cumulative regret from $(1-\alpha_t)S_t^*$ follows the optimism principle of UCB and is on the order of $T^{1/2}$. The cumulative regret from $\alpha_t \widetilde{S}_{k_t^*}$ is on the order of $T^{2/3}$, since $\alpha_t = t^{-1/3}$ and the suboptimality gap of $S_{k_t^*}$ is constant. Meanwhile, for the second type of events, we can bound the total number of rounds such event occurs in the worst case as $\widetilde{\mathcal{O}}(K^2 T^{2/3})$. This requires some involved arguments, though it is intuitively the sample complexity to learn $\|\widehat{q}_k - q_k\| \leq t^{-1/3}$ for every $k \in [K]$. This corresponds to the worst case where these signals are competitive and $n_{k,t}$ grows at the same rate for any $k \in [K]$, which means that we have to precisely learn the belief distributions and costs for all the actions. Therefore, based on the dominating term, the total regret is $\widetilde{\mathcal{O}}(KT^{2/3})$.

We analyze the regret by the following two cases: (i) rounds t such that the agent responds with action $k_t = k_t^*$; (ii) rounds t such that the agent responds with action $k_t \neq k_t^*$. Let $\text{SubOpt}(k, S_t) = \max_{S \in \mathcal{S}} h(S) - \mathbb{E}_{\omega \sim \sigma, \sigma \sim q_k}[u(a^*(\sigma_t), \omega_t) - S_t(\sigma_t, \omega_t)]$ be the regret of implementing scoring rule S_t with the agent selecting action b_k in a single round. We have for the online regret that,

$$\begin{aligned}
\text{Reg}^\pi(T) &= \mathbb{E}_\pi \left[\sum_{t=1}^T \left(\max_{S \in \mathcal{S}} h(S) - \mathbb{E}_{\omega_t \sim \sigma_t, \sigma_t \sim q_{k_t}} \left[u(a^*(\sigma_t), \omega_t) - S_t(\sigma_t, \omega_t) \right] \right) \right] \\
&= \sum_{t=1}^T \underbrace{\mathbb{E}\left[\mathbb{1}(k_t = k_t^*) \text{SubOpt}(k_t, S_t) \right]}_{A_t} + \sum_{t=1}^T \underbrace{\mathbb{E}\left[\mathbb{1}(k_t \neq k_t^*) \text{SubOpt}(k_t, S_t) \right]}_{B_t}.
\end{aligned}$$

Here and in the sequel, we always use $\mathbb{E}$ to denote the expectation taken with respect to the data generating process described in (2.3).

Bounding regret A_t. Since $S_t = \alpha_t \widetilde{S}_{k_t^*} + (1-\alpha_t) S_t^*$, it follows from the linearity of $\text{SubOpt}(k_t, S_t)$ that

$$\text{SubOpt}(k_t, S_t) = \alpha_t \text{SubOpt}(k_t, \widetilde{S}_{k_t^*}) + (1-\alpha_t)\text{SubOpt}(k_t, S_t^*). \tag{F.1}$$

Here, the first term $\text{SubOpt}(k_t, \widetilde{S}_{k_t^*})$ is bounded by constant $C_1 = 2(B_S + B_u)$, and the second term $\text{SubOpt}(k_t, S_t^*)$ is bounded by $2\left(\|S_t^*\|_\infty + B_u\right) I_{q,k_t^*}^t$ with probability at least $1 - 1/t$ by the following lemma.

Lemma F.1. Define event $\mathcal{E}_t$ as $\mathcal{E}_t = \left\{\|q_k - \widehat{q}_k^t\|_1 \le I_q^t(k), \forall k \in [K]^+\right\}$. Then event $\mathcal{E}_t$ holds with probability at least $1 - 1/t$ and it holds on event $\mathcal{E}_t$ that

$$u_{k_t^*} - v_{S_t^*}(k_t^*) + 2\left(B_S + B_u\right) I_q^t(k_t^*) \ge \max_{S \in \mathcal{S}} h(S),$$

where $h(S) = \mathbb{E}_{\omega \sim \sigma, \sigma \sim q_{k^*(S)}}\left[u(a^*(\sigma), \omega) - S(\sigma, \omega)\right]$ and $k^*(S)$ is the agent's best response.

Proof. See §F.1 for a detailed proof. □

The first term on the right-hand side of (F.1) can be directly bounded by $\alpha_t C_1$. For the second term on event $\mathcal{E}_t$, it can be bounded by $C_1 I_q^t(k_t^*)$ using Lemma F.1. If event $\mathcal{E}_t$ does not hold, the second term is still bounded by C_1 with probability at most $1/t$. Thus, we have for $\alpha_t = K t^{-1/3} \wedge 1$ that

$$\begin{aligned}
\sum_{t=1}^T A_t &\le \sum_{t=1}^T C_1 \left(\alpha_t + \frac{1}{t} + I_q^t(k_t^*)\right) \\
&\le C_1 \left(K\left(\frac{3}{2}T^{2/3} + 1\right) + (\log T + 1) + \sum_{k \in [K]^+} \sum_{t: k_t^* = k} \sqrt{\frac{2\log(K \cdot 2^M t)}{n_k^t}}\right) \\
&\le C_1 \left(K\left(\frac{3}{2}T^{2/3} + 1\right) + \log T + 1\right) + C_1 \sqrt{2\log(K \cdot 2^M T)} \cdot \sum_{k \in [K]^+} \left(2\sqrt{n_k^T} + 1\right) \\
&\le \mathcal{O}(T^{2/3}) + \mathcal{O}(\sqrt{|M| + \log(KT)} \cdot \sqrt{KT}),
\end{aligned} \tag{F.2}$$

where the first equality follows from our previous discussions on (F.1), the second inequality holds from the definition of I_q^t, and the last inequality holds by the Jensen's inequality that $\sum_{k \in [K]^+} \sqrt{n_k^T}/K \le \sqrt{\sum_{k \in [K]^+} n_k^T / K}$.

Bounding regret B_t. We characterize regret $\sum_{t=1}^T B_t$ by bounding the expected number of rounds that $k_t \ne k_t^*$. To do so, we invoke the following lemma.

Lemma F.2. For any fixed $i \ne k_t^*$, under the condition that

$$\alpha_t \ge 2\varepsilon^{-1}\left(I_c^t(k_t^*, i) + B_S\left(I_q^t(i) + I_q^t(k_t^*)\right)\right) := \Delta_t(k_t^*, i),$$

the agent must respond $k_t \ne i$ under the scoring rule $S_t = \alpha_t \widetilde{S}_{k_t^*} + (1-\alpha_t) S_t^*$ on event $\mathcal{E}_t$.

Proof. See §F.2 for a detailed proof. □

Under the condition $k_t \ne k_t^*$, the algorithm must be doing binary search. Now, we introduce the notations that will be used in bounding B_t. Compare to the definition in Algorithm 3, we instead define $\text{BS}(S_0, S_1, \widetilde{\lambda}_{\min}, \widetilde{\lambda}_{\max}, \tau_0, \tau_1, k_0, k_1, t_0, t_1)$ for a binary search (BS), where (S_0, S_1) is the initial segment that this BS is conducted on, $(\widetilde{\lambda}_{\min}, \widetilde{\lambda}_{\max})$ corresponds to the value of $(\lambda_{\min}, \lambda_{\max})$ that this BS ends with, τ_0, τ_1 correspond to the rounds in which $(1 - \widetilde{\lambda}_{\min})S_0 + \widetilde{\lambda}_{\min} S_1$ and $(1 - \widetilde{\lambda}_{\max})S_0 + \widetilde{\lambda}_{\max} S_1$ are played, respectively, k_0, k_1 are the best response under $(1 - \widetilde{\lambda}_{\min})S_0 + \widetilde{\lambda}_{\min} S_1$ and

$(1-\widetilde{\lambda}_{\max})S_0+\widetilde{\lambda}_{\max}S_1$, and t_0, t_1 are the starting round and the ending round of this BS. Notably, k_0 is not necessarily the best response under S_0 but k_1 is guaranteed to be the best response under both S_1 and $(1-\widetilde{\lambda}_{\max})S_0+\widetilde{\lambda}_{\max}S_1$.

For the BS that lasts until round t, we let $t_0(t)$ be the first round of the BS (if round t is not doing BS, we just let $t_0(t)=t$), $(k_0(t), k_1(t))$ be the best response that this BS ends with, $(S_0(t), S_1(t))$ be the segment that this BS searches on. We consider the following two situations for the case $k_t \neq k_t^*$: (i) $\Delta_{t_0(t)}(k_0(t), k_1(t)) \leq \alpha_{t_0(t)}$; (ii) $\Delta_{t_0(t)}(k_0(t), k_1(t)) > \alpha_{t_0(t)}$. Following such an idea, we have

$$
\begin{aligned}
\sum_{t=1}^{T} B_t &= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(k_t \neq k_t^*)\text{SubOpt}(k_t, S_t)\right] \\
&\leq C_1 \left[\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\text{BS}=1, \Delta_{t_0(t)}(k_0(t), k_1(t)) \leq \alpha_{t_0(t)})\right] + \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\text{BS}=1, \Delta_{t_0(t)}(k_0(t), k_1(t)) > \alpha_{t_0(t)})\right]\right] \\
&\leq C_1 \Bigg(\underbrace{\sum_{t=1}^{T} \frac{1}{t_0(t)}}_{(B.1)} + \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\text{BS}=1, \Delta_{t_0(t)}(k_0(t), k_1(t)) > \alpha_{t_0(t)})\right]}_{(B.2)} \Bigg),
\end{aligned}
$$

where the first inequality holds by $\text{SubOpt}(k_t, S_t) \leq C_1$, the second inequality holds by Lemma F.2. Specifically, when doing a binary search, the deployed scoring rule S_t must lie on the segment $(S_0(t), S_1(t))$, where $S_0(t) = \alpha_{t_0(t)}\widetilde{S}_{k^*_{t_0(t)}} + (1-\alpha_{t_0(t)})S^*_{t_0(t)}$ and $S_1(t) = \widetilde{S}_{k^*_{t_0(t)}}$. Hence, S_t can be equivalently expressed as

$$
S_0(t) = \alpha_t' \widetilde{S}_{k^*_{t_0(t)}} + (1-\alpha_t')S^*_{t_0(t)},
$$

where $\alpha_t' \geq \alpha_{t_0(t)}$. Therefore, under the condition that $\Delta_{t_0(t)}(k_0(t), k_1(t)) \leq \alpha_{t_0(t)}$ and that the agent responds $k_0(t) \neq k_1(t)$ in this binary search, we can use Lemma F.2 to bound the probability by $1/t_0(t)$.

To bound $(B.2)$, we define a concept called *essential binary search*.

Definition F.3 (Essential binary search)**.** We call a binary search $\text{BS}(S_0, S_1, \widetilde{\lambda}_{\min}, \widetilde{\lambda}_{\max}, \tau_0, \tau_1, k_0, k_1, t_0, t_1)$ an *essential binary search* (essential BS) if $\alpha_{t_0} < \Delta_{t_0}(k_0, k_1)$ where $\alpha_t = Kt^{-1/3} \wedge 1$.

The following lemma bounds the total rounds of essential BSs.

Lemma F.4. Consider binary search $\text{BS}(S_0, S_1, \widetilde{\lambda}_{\min}, \widetilde{\lambda}_{\max}, \tau_0, \tau_1, k_0, k_1, t_0, t_1)$. Suppose that the total number of essential binary searches in T rounds is N. Then, we have

$$
N \leq K \cdot \left(12\varepsilon^{-1}B_S\right)^2 2\log\left(K2^M T\right) T^{2/3}. \tag{F.3}
$$

Proof. See §F.3 for a detailed proof. □

Let m be the maximal rounds in a binary search, i.e., $t_0(t) \geq t-m$. Using the terminal criteria for binary search, we have

$$
m \leq \max_{t,k} -\log_2\left(I_q^t(k)\right) = \mathcal{O}(\log(T)).
$$

The term $(B.1)$ is directly bounded by

$$
(B.1) \leq \log(T) + m = \mathcal{O}(\log(T)).
$$

The term $(B.2)$ is bounded using Lemma F.4,

$$
(B.2) \leq m \cdot K \cdot \left(12\varepsilon^{-1}B_S\right)^2 2\log\left(K2^M T\right) T^{2/3} = \mathcal{O}(\varepsilon^{-2}B_S^2(M+\log(KT))mKT^{2/3}).
$$

which finishes the proof.

□

F.1. Proof of Lemma F.1

Lemma F.5 (Restatement of Lemma F.1)**.** Define event $\mathcal{E}_t$ as $\mathcal{E}_t = \left\{\|q_k - \widehat{q}_k^t\|_1 \le I_q^t(k), \forall k \in [K]^+\right\}$. Then event $\mathcal{E}_t$ holds with probability at least $1 - 1/t$ and it holds on event $\mathcal{E}_t$ that

$$\langle q_{k_t^*}, u - S_t^* \rangle_\Sigma + 2\left(B_S + B_u\right) I_q^t(k_t^*) \ge \max_{S \in \mathcal{S}} h(S),$$

where $h(S) = \mathbb{E}_{\omega \sim \sigma, \sigma \sim q_{k^*(S)}}\left[u(a^*(\sigma), \omega) - S(\sigma, \omega)\right]$, $v_S(k) = \langle q_k, S \rangle_\Sigma$ and $k^*(S)$ is the agent's best response.

Proof of Lemma F.1. We first state the concentration result from Mardia et al. (2020).

Lemma F.6 (Mardia et al. (2020), Lemma 1, Concentration for empirical distribution)**.** Let p be a probability distribution supported in a finite set with cardinality at most M and p_n be the empirical distribution of n i.i.d. samples from p. Then, for all sample size $n \in \mathbb{N}_+$ and $0 < \delta < 1$,

$$\mathbb{P}\left(\|p - p_n\|_1 \ge \sqrt{\frac{2\log\left(2^M/\delta\right)}{n}}\right) \le \delta.$$

By Lemma F.6 and taking a union bound, with probability at least $1 - 1/t$, we have $\|\widehat{q}_k^t - q_k\|_1 \le I_q^t(k)$ for all $k \in [K]^+$, which gives event $\mathcal{E}_t$. In the sequel, we discuss under the condition that event $\mathcal{E}_t$ holds. We can verify for $v_S(k) = \langle q_k, S \rangle_\Sigma$ that

$$\left|\widehat{v}_S^t(k) - v_S(k)\right| = \left|\langle \widehat{q}_k^t - q_k, S \rangle_\Sigma\right| \le B_S I_q^t(k). \tag{F.4}$$

To study the confidence interval for $\widehat{C}^t(i,j)$, we invoke the following lemma.

Lemma F.7. Under event $\mathcal{E}_t = \left\{\|q_k - \widehat{q}_k^t\|_1 \le I_q^t(k), \forall k \in [K]^+\right\}$, we have

$$\left|\widehat{C}^t(i,j) - C(i,j)\right| \le I_c^t(i,j), \quad \forall (i,j) \in [K].$$

Proof. A direct corollary of (F.4) is

$$\begin{aligned} C_+^t(i,j) &\ge \min_{\tau < t : k_\tau = i} v_{S_\tau}(i) - v_{S_\tau}(j) \ge C(i,j), \\ C_-^t(i,j) &\le \max_{\tau < t : k_\tau = j} v_{S_\tau}(i) - v_{S_\tau}(j) \le C(i,j). \end{aligned} \tag{F.5}$$

Using the fact in (F.5), we have

$$\begin{aligned} \left|\widehat{C}^t(i,j) - C(i,j)\right| &= \left|\sum_{(i',j') \in l_{ij}} \frac{C_+^t(i',j') + C_-^t(i',j')}{2} - C(i',j')\right| \\ &\le \frac{1}{2} \sum_{(i',j') \in l_{ij}} \left(\left|C_+^t(i',j') - C(i',j')\right| + \left|C_-^t(i',j') - C(i',j')\right|\right) \\ &= \frac{1}{2} \sum_{(i',j') \in l_{ij}} \left|C_+^t(i',j') - C_-^t(i',j')\right| = I_c^t(i,j), \end{aligned} \tag{F.6}$$

where the inequality holds by triangular inequality, the second equality holds by (F.5), and the last equality holds by definition of I_C^t. □

let k^* and S^* be the best response and the optimal scoring rule that achieves the optimal objective $\sup_{S \in \mathcal{S}} h(S)$ in (2.2), respectively. Recall the linear programming Opt-LP$_k$,

$$\begin{aligned} \text{Opt-LP}_k: \quad & \max_{S \in \mathcal{S}} \quad \langle u, \widehat{q}_k^t \rangle_\Sigma + B_u I_q^t(k) - v, \\ & \text{s.t.} \quad \left|v - \widehat{v}_S^t(k)\right| \le B_S \cdot I_q^t(k), \\ & \qquad v - \widehat{v}_S^t(i) \ge \widehat{C}^t(k,i) - \left(I_c^t(k,i) + B_S \cdot I_q^t(i)\right), \quad \forall i \ne k. \end{aligned}$$

It is easy to verify that S^* satisfies all the constraint in Opt-LP$_{k^*}$ using (F.4) and (F.6) with $v = v_{S^*}(k^*)$ and $k = k^*$. Also, $\langle u, \widehat{q}_k^t \rangle_\Sigma + B_u I_q^t(k) \geq \langle u, q_k \rangle_\Sigma$ for $\forall k \in [K]^+$. Therefore, by noting that $k_t^* = \operatorname{argmax}_{k \in [K]^+} h_{\text{LP}}^t(k)$, we conclude that $h_{\text{LP}}^t(k_t^*) \geq h_{\text{LP}}^t(k^*) \geq h(S^*)$. Also, let v_{LP}^* be the solution to Opt-LP$_{k_t^*}$ and recall that $S_t^* = S_{\text{LP},k_t^*}^t$ we have

$$\begin{aligned} h_{\text{LP}}^t(k_t^*) &= \langle u, \widehat{q}_{k_t^*}^t \rangle_\Sigma + B_u I_q^t(k_t^*) - v_{\text{LP}}^* \\ &\leq (B_u + B_S) I_q^t(k^*) + \mathbb{E}_{\omega \sim \sigma, \sigma \sim \widehat{q}_{k_t^*}} \left[u(a^*(\sigma), \omega) - S_t^*(\sigma, \omega) \right] \\ &\leq 2(B_u + B_S) I_q^t(k^*) + \mathbb{E}_{\omega \sim \sigma, \sigma \sim q_{k_t^*}} \left[u(a^*(\sigma), \omega) - S_t^*(\sigma, \omega) \right], \end{aligned}$$

where the first inequality holds by noting that $\left| v_{\text{LP}}^* - \widehat{v}_{S_t^*}(k_t^*) \right| \leq B_S I_q^t(k_t^*)$, and the second inequality holds by definition of event $\mathcal{E}_t$. Thus, we finish the proof. □

F.2. Proof of Lemma F.2

Lemma F.8 (Restatement of Lemma F.2)**.** For $\forall i \neq k_t^*$, under the condition that

$$\alpha_t \geq 2\varepsilon^{-1} \left(I_c^t(k_t^*, i) + B_S \left(I_q^t(i) + I_q^t(k_t^*) \right) \right) := \Delta_t(k_t^*, i),$$

the agent will not respond $k_t = i$ under the scoring rule $S_t = \alpha_t \widetilde{S}_{k_t^*} + (1 - \alpha_t) S_t^*$ on event $\mathcal{E}_t$.

Proof of Lemma F.2. On event $\mathcal{E}_t$, for any $i \neq k_t^*$, if $\alpha_t \geq \Delta_t(k_t^*, i)$, we aim to show the agent prefers action $b_{k_t^*}$ to action b_i. Recall $S_t = \alpha_t \widetilde{S}_{k_t^*} + (1 - \alpha_t) S_t^*$. The expected profit for the agent generated by responding action b_i is

$$v_{S_t}(i) - c_i \leq \alpha_t v_{\widetilde{S}_{k_t^*}}(i) + (1 - \alpha_t) \widehat{v}_{S_t^*}^t(i) - c_i + (1 - \alpha_t) B_S I_q^t(i),$$

and the expected utility generated by responding action $b_{k_t^*}$ is

$$v_{S_t}(k_t^*) - c_{k_t^*} \geq \alpha_t v_{\widetilde{S}_{k_t^*}}(k_t^*) + (1 - \alpha_t) \widehat{v}_{S_t^*}^t(k_t^*) - c_{k_t^*} - (1 - \alpha_t) B_S I_q^t(k_t^*).$$

By Assumption 3.1, we already have $(v_{\widetilde{S}_{k_t^*}}(k_t^*) - c_{k_t^*}) - (v_{\widetilde{S}_{k_t^*}}(i) - c_i) > \varepsilon$. Since S_t^* is the solution to Opt-LP$_{k_t^*}$, following the constraints of Opt-LP$_{k_t^*}$ we have

$$\widehat{v}_{S_t^*}^t(k_t^*) - \widehat{v}_{S_t^*}^t(i) \geq \widehat{C}^t(k_t^*, i) - I_c^t(k_t^*, i) - B_S(I_q^t(i) + I_q^t(k_t^*)).$$

Combining the above inequalities, we get

$$\begin{aligned} \left[v_{S_t}(k_t^*) - c_{k_t^*} \right] - \left[v_{S_t}(i) - c_i \right] &\geq \alpha_t \cdot \varepsilon + (1 - \alpha_t) \left(\widehat{C}^t(k_t^*, i) - C(k_t^*, i) - I_c^t(k_t^*, i) - B_S(I_q^t(i) + I_q^t(k_t^*)) \right) \\ &\quad - (1 - \alpha_t) B_S \left(I_q^t(k_t^*) + I_q^t(i) \right) \\ &\geq \alpha_t \cdot \varepsilon + 2(1 - \alpha_t) \left[-I_c^t(k_t^*, i) - B_S(I_q^t(i) + I_q^t(k_t^*)) \right] \\ &> \alpha_t \cdot \varepsilon - 2 \left[I_c^t(k_t^*, i) + B_S(I_q^t(i) + I_q^t(k_t^*)) \right], \end{aligned}$$

where the second inequality holds by recalling that we have proved $C(k_t^*, i) \leq \widehat{C}^t(k_t^*, i) + I_c^t(k_t^*, i)$ in (F.6) under $\mathcal{E}_t$, and the last inequality holds by noting that $0 < \alpha_t \leq 1$. When $\alpha_t \geq 2\varepsilon^{-1} \left(I_c^t(k_t^*, i) + B_S \left(I_q^t(i) + I_q^t(k_t^*) \right) \right)$, the above $\left[v_{S_t}(k_t^*) - c_{k_t^*} \right] - \left[v_{S_t}(i) - c_i \right] > 0$, which means the agent prefers action k_t^* to i.

□

F.3. Proof of Lemma F.4

In the sequel, Proposition F.9 shows essential BS reduces confidence interval, while Proposition F.10 states the stopping criteria of essential BS. Combining Proposition F.9 and Proposition F.10, we can bound the total rounds in essential BS.

Proposition F.9. For a binary search BS$(S_0, S_1, \widetilde{\lambda}_{\min}, \widetilde{\lambda}_{\max}, \tau_0, \tau_1, k_0, k_1, t_0, t_1)$, we have

$$I_c^t(k_0, k_1) \leq 2\sqrt{2 \log(K \cdot 2^M T)} B_S \left(\frac{1}{\sqrt{n_{k_0}^{t_1}}} + \frac{1}{\sqrt{n_{k_1}^{t_1}}} \right). \tag{F.7}$$

Proof of Proposition F.9. By definition of $I_c^t(k_0, k_1)$, for any $t > t_1$,

$$I_c^t(k_0, k_1) = \sum_{(i',j') \in l_{k_0 k_1}} \varphi^t(i', j') \leq \varphi^t(k_0, k_1),$$

where the inequality holds by noting that $l_{k_0 k_1}$ is the "shortest path" with φ^t being the "length" of each edge. Using the definition of φ^t, we conclude that

$$\begin{aligned}
I_c^t(k_0, k_1) &\leq \frac{1}{2}\left[C_+^t(k_0, k_1) - C_-^t(k_0, k_1)\right]_+ \\
&= \frac{1}{2}\left[\min_{\tau' < t : k_{\tau'} = k_0} \left(\widehat{v}_{S_{\tau'}}^t(k_0) - \widehat{v}_{S_{\tau'}}^t(k_1)\right) - \max_{\tau' < t : k_{\tau'} = k_0} \left(\widehat{v}_{S_{\tau'}}^t(k_0) - \widehat{v}_{S_{\tau'}}^t(k_1)\right)\right]_+ \\
&\qquad + B_S\left(I_q^t(k_0) + I_q^t(k_1)\right) \\
&\leq \frac{1}{2}\left[\widehat{v}_{S_{\tau_0}}^t(k_0) - \widehat{v}_{S_{\tau_0}}^t(k_1) - \widehat{v}_{S_{\tau_1}}^t(k_0) + \widehat{v}_{S_{\tau_1}}^t(k_1)\right]_+ + B_S\left(I_q^t(k_0) + I_q^t(k_1)\right),
\end{aligned}$$

where the last inequality holds by noting that $t > t_1 \geq (\tau_0, \tau_1)$. Consequently, we have by the definition of $\widehat{v}_S^t$ that,

$$\begin{aligned}
I_c^t(k_0, k_1) &\leq \frac{1}{2}\left|\langle S_{\tau_0} - S_{\tau_1}, \widehat{q}_{k_0}^t - \widehat{q}_{k_1}^t \rangle_\Sigma\right| + B_S\left(I_q^t(k_0) + I_q^t(k_1)\right) \\
&= \frac{(\widetilde{\lambda}_{\max} - \widetilde{\lambda}_{\min})}{2}\left|\langle S_1 - S_0, \widehat{q}_{k_0}^t - \widehat{q}_{k_1}^t \rangle_\Sigma\right| + B_S\left(I_q^t(k_0) + I_q^t(k_1)\right) \\
&\leq \frac{\min\left(I_q^{t_1}(k_0), I_q^{t_1}(k_1)\right)}{2}\left(\left|\langle S_1 - S_0, \widehat{q}_{k_0}^t \rangle_\Sigma\right| + \left|\langle S_1 - S_0, \widehat{q}_{k_1}^t \rangle_\Sigma\right|\right) + B_S\left(I_q^t(k_0) + I_q^t(k_1)\right),
\end{aligned}$$

where $\widetilde{\lambda}_{\max}$ and $\widetilde{\lambda}_{\min}$ are the values of $\lambda_{\max}$ and $\lambda_{\min}$ when this BS terminates, respectively. Here, the last inequality holds by the terminal criteria of the BS. Furthermore, we can upper bound the right-hand side by

$$\begin{aligned}
I_c^t(k_0, k_1) &\leq \min\left(I_q^{t_1}(k_0), I_q^{t_1}(k_1)\right) B_S + B_S\left(I_q^t(k_0) + I_q^t(k_1)\right) \\
&\leq \left[\sqrt{2\log(K \cdot 2^M T)} \cdot \left(\frac{1}{\sqrt{n_{k_0}^{t_1}}} + \frac{1}{\sqrt{n_{k_1}^{t_1}}} + \frac{1}{\sqrt{\max\left\{n_{k_0}^{t_1}, n_{k_1}^{t_1}\right\}}}\right)\right] B_S \\
&\leq 2B_S\left[\sqrt{2\log(K \cdot 2^M T)} \cdot \left(\frac{1}{\sqrt{n_{k_0}^{t_1}}} + \frac{1}{\sqrt{n_{k_1}^{t_1}}}\right)\right],
\end{aligned}$$

where the first inequality holds by noting that $\widehat{q}_k^t$ is an empirical probability and the fact that $\|S_1 - S_0\|_\infty \leq B_S$ since S is nonnegative, and the second inequality holds by noting that n_k^t is nondecreasing and $t_1 < t$.

□

Proposition F.10. For a binary search BS$(S_0, S_1, \widetilde{\lambda}_{\min}, \widetilde{\lambda}_{\max}, \tau_0, \tau_1, k_0, k_1, t_0, t_1)$, if

$$\min\{n_{k_0}^{t_1}, n_{k_1}^{t_1}\} \geq \left(12\varepsilon^{-1}B_S\right)^2 2\log\left(K2^M T\right) T^{2/3} =: N^*, \tag{F.8}$$

then we have $t^{-1/3} \geq \Delta_t(k_0, k_1)$ for any $t_1 < t \leq T$ and no more essential BS ending with the same (k_0, k_1) is possible.

Proof of Proposition F.10. For any $t_1 < t \leq T$, we have

$$\begin{aligned}
t^{-1/3} \geq T^{-1/3} &= \left(12\varepsilon^{-1}B_S\right)\sqrt{\frac{2\log\left(K2^M T\right)}{N^*}} \\
&\geq 2\left(\varepsilon^{-1}B_S\right)\sqrt{2\log\left(K2^M T\right)}\left(2\frac{1}{\sqrt{n_{k_0}^{t_1}}} + 2\frac{1}{\sqrt{n_{k_1}^{t_1}}} + \frac{1}{\sqrt{n_{k_0}^{t}}} + \frac{1}{\sqrt{n_{k_1}^{t}}}\right) \\
&\geq 2\varepsilon^{-1}\left(B_S\left(I_q^t(k_0) + I_q^t(k_1)\right) + I_c^t(k_0, k_1)\right) = \Delta_t(k_0, k_1),
\end{aligned}$$

where the equality holds by definition of N^*, the second inequality holds by our condition $\min\{n_{k_0}^{t_1}, n_{k_1}^{t_1}\} \geq N^*$, and the last inequality holds by using Proposition F.9. Therefore, we conclude that by letting $\alpha_t = Kt^{-1/3} \wedge 1 \geq t^{-1/3}$, the condition for an essential BS, i.e., $\alpha_t < \Delta_t(k_0, k_1)$ no longer holds for any t such that $t_1 < t \leq T$. □

Now, we are ready to prove Lemma F.4.

Lemma F.11 (Restatement of Lemma F.4)**.** Consider binary search BS$(S_0, S_1, \widetilde{\lambda}_{\min}, \widetilde{\lambda}_{\max}, \tau_0, \tau_1, k_0, k_1, t_0, t_1)$. Suppose that the total number of essential binary searches in T rounds is N. Then, we have

$$N \leq \left(12\varepsilon^{-1}B_S\right)^2 2\log\left(K2^M T\right) KT^{2/3} = KN^*. \tag{F.9}$$

Proof of Lemma F.4. To prove Lemma F.4, we first introduce another concept called *critical binary search* (critical BS).

Definition F.12 (Critical binary search)**.** For an *essential* binary search BS$(S_0, S_1, \widetilde{\lambda}_{\min}, \widetilde{\lambda}_{\max}, \tau_0, \tau_1, k_0, k_1, t_0, t_1)$, we consider it to be critical if $\min\{n_{k_0}^{t_1}, n_{k_1}^{t_1}\} \geq N^*$.

Critical BSs. We claim that the number of critical BS is no more than K. Suppose the opposite holds, i.e., number of critical BS no less than $K+1$. Consider a graph $\mathcal{G} = (V, E)$ where $V = [K]^*$ and E is formed by the set the critical BSs ending with $(k_0, k_1) \in V$. We claim that one of the following two cases must hold

(i) There exists $(i, j) \in [K]^+$ such that there are at least two critical BSs ending with $(k_0, k_1) = (i, j)$;

(ii) There is a loop in $\mathcal{G}$ formed by the critical BSs.

We prove that both (i) and (ii) must not happen. For (i), following from the definition of critical BS that $\min\{n_{k_0}^{t_1}, n_{k_1}^{t_1}\} \geq N^*$ where t_1 is the ending round of a critical BS and using Proposition F.10 that there should be no more essential BSs ending with (k_0, k_1) after t_1, we directly have (i) impossible. For (ii), suppose that BS$(S_0, S_1, \widetilde{\lambda}_{\min}, \widetilde{\lambda}_{\max}, \tau_0, \tau_1, k_0, k_1, t_0, t_1)$ is the first that creates a loop. Then before t_0, there must be a path $\ell'_{k_0k_1}$ of critical BSs. For any edge (k'_0, k'_1) on $\ell'_{k_0k_1}$, there must be a critical Binary search BS$(S'_0, S'_1, \widetilde{\lambda}'_{\min}, \widetilde{\lambda}'_{\max}, \tau'_0, \tau'_1, k'_0, k'_1, t'_0, t'_1)$ with $t'_1 < t_0$. Using Proposition F.10 again, we have

$$t_0^{-1/3} \geq \Delta_{t_0}(k'_0, k'_1), \quad \forall (k'_0, k'_1) \in \ell'_{k_0,k_1}.$$

Since a path in $\mathcal{G}$ has length at most K, we conclude that

$$\begin{aligned}
\alpha_{t_0} = Kt_0^{-1/3} &\geq \sum_{(k'_0,k'_1)\in\ell'_{k_0k_1}} \Delta_{t_0}(k'_0, k'_1) \\
&= \sum_{(k'_0,k'_1)\in\ell'_{k_0k_1}} 2\varepsilon^{-1}\left(B_S\left(I_q^{t_0}(k'_0) + I_q^{t_0}(k'_1)\right) + I_c^{t_0}(k'_0, k'_1)\right) \\
&\geq 2\varepsilon^{-1}\left(B_S\left(I_q^{t_0}(k_0) + I_q^{t_0}(k_1)\right) + I_c^{t_0}(k_0, k_1)\right) = \Delta_{t_0}(k_0, k_1),
\end{aligned}$$

where the last inequality holds by noting that $I_c^{t_0}(k'_0, k'_1) \leq \varphi^{t_0}(k'_0, k'_1)$ and $I_c^{t_0}$ is obtained by the shortest path algorithm with edge length φ^{t_0}. Such a fact again suggests that there cannot be any essential BR ending with (k_0, k_1) that starts at t_0. Thus, a loop also cannot be formed, and we prove our claim that the total number of critical BSs is no more than K.

Essential but noncritical BSs. Next, we count the number of BSs that are essential but not critical. Recall the graph $\mathcal{G} = (V, E)$ where $V = [K]^*$ and E is formed by the set the critical BSs ending with $(k_0, k_1) \in V$. Let $w^t(i, j)$ be the total number of essential BSs ending with $(k_0, k_1) = (i, j)$ before round t and let $w(i, j) = w^T(i, j)$. Let $d_k^t = \sum_{k' \neq k} w(k, k')$ be the total weights for node $k \in V$. For any pair $(k_0, k_1) \in V$, if $d_{k_1}^t \wedge d_{k_2}^t \geq N^*$, there should be no more essential BSs after t by Proposition F.10. On the other hand, any BS ending with (k_0, k_1) guarantees that $d_{k_0}^t$ and $d_{k_1}^t$ increase at least by 1. Hence, the problem of counting the number of essential but noncritical BSs can be described as the following weight-placing game,

(i) initiate $d_k = 0$ for $\forall k \in V$;

(ii) at each round, select an edge $(k_0, k_1) \in V$ such that $d_{k_0} \wedge d_{k_1} < N^*$ and add 1 to both d_{k_0} and d_{k_1};

(iii) the game ends if no more edge can be selected.

The total rounds of this weight-placing game upper bound the total number of essential but noncritical BSs. We have the following proposition that gives the maximal number of rounds in this weight-placing game.

Proposition F.13 (Weight-placing game)**.** For the previously described weight-placing game, the total number of rounds should not exceed $(|V| - 1)(N^* - 1)$.

Proof. We prove this proposition by induction on $|V|$. The case $|V| = 1$ is trivial. For $|V| = 2$ the result is obvious since there is only one edge to choose. Suppose that the result holds for $|V| \leq v-1$. For $|V| = v$, suppose that $(v-1)(N^*-1)+1$ rounds are played. We study the quantity $d^- = \min_{k \in V} d_k$.

- If there exists $k \in V$ such that $d_k \leq N^* - 2$, then for the subgraph $\mathcal{G}^k = \mathcal{G} \backslash \{k\}$, we have at least $(v-2)(N^*-1)+2$ rounds played on $\mathcal{G}^k$. However, since $\mathcal{G}^k$ only has $v-1$ nodes, we have from the induction that there should be no more than $(v-2)(N^*-1)$ rounds played on $\mathcal{G}^k$, which causes a contradictory. Thus, we have $d^- \geq N^* - 1$ if $(v-1)(N^*-1)+1$ rounds are played.
- If $d^- \geq N^*$, then the last round in this play should be mistakenly played following rules (ii) and (iii).

Therefore, we conclude that $d^- = N^* - 1$ and we are able to choose a node $k \in V$ such that $d_k = N^* - 1$. On the subgraph $\mathcal{G}^k = \mathcal{G} \backslash \{k\}$, we have by induction that at most $(v-2)(N^*-1)$ rounds can be played. Therefore, the total rounds played on $\mathcal{G}$ is no more than

$$(v-2)(N^*-1) + d_k = (v-1)(N^*-1),$$

which contradicts our assumption that $(v-1)(N^*-1)+1$ rounds are played. Thus, for $|V| = v$, no more than $(v-1)(N^*-1)$ rounds can be played. By induction, we conclude that the proposition holds for any $|V|$. □

Hence, the total number of essential but noncritical BSs is bounded by $K(N^*-1)$. Given that the number of critical BSs bounded by K, the total number of essential BSs is thus bounded by KN^*. □