Robust Learning with Private Information

Kyohei Okumura *

February 10, 2026

Abstract

Firms increasingly delegate decisions to learning algorithms in platform markets. Standard algorithms perform well when platform policies are stationary, but firms often face ambiguity about whether policies are stationary or adapt strategically to their behavior. When policies adapt, efficient learning under stationarity may backfire: it may reveal a firm's persistent private information, allowing the platform to personalize terms and extract information rents. We study a repeated screening problem in which an agent with a fixed private type commits ex ante to a learning algorithm, facing ambiguity about the principal's policy. We show that a broad class of standard algorithms, including all no-external-regret algorithms, can be manipulated by adaptive principals and permit asymptotic full surplus extraction. We then construct a misspecification-robust learning algorithm that treats stationarity as a testable hypothesis. It achieves the optimal payoff under stationarity at the minimax-optimal rate, while preventing dynamic rent extraction: against any adaptive principal, each type's long-run utility is at least its utility under the menu that maximizes revenue under the principal's prior.

* kyohei.okumura@gmail.com.
I thank Hersh Chopra, Annie Liang, Harry Pei, Jon Schneider, Marciano Siniscalchi, and Bruno Strulovici for insightful comments.

1 Introduction

Algorithmic decision-making is now routine in platform markets, and firms increasingly delegate their decisions to learning algorithms. Advertisers rely on automated bidding tools, while sellers and hosts use algorithmic pricing. They do so, in part, because the mapping from actions to outcomes is imperfectly known: market conditions such as demand and product quality are uncertain, and platform policies—including reserve prices, ranking scores, eligibility rules, and fee schedules—are often only partially understood ex ante and are learned through interaction. In such settings, a learning algorithm serves as a practical commitment device, specifying how the firm will experiment and adapt given its limited knowledge of the environment.

Many firms deploy off-the-shelf learning algorithms that are guaranteed to perform well in stationary environments. Stationarity can be a reasonable approximation when policy updates are infrequent, and rapid personalization is limited by operational or institutional frictions. At the same time, modern platforms may implement adaptive, history-dependent policies firm-by-firm through automated experimentation and revenue optimization, so the terms a firm faces may respond to its past behavior. Do standard algorithms perform well when the environment is not stationary? If not, can we design algorithms that are robust to this form of model misspecification?

This paper studies the design of learning algorithms for a firm facing ambiguity about whether a platform's policy is stationary or adaptive. The ambiguity matters because efficient learning under stationarity requires payoff-responsive experimentation. With a persistent private type, payoff-responsive experimentation induces type-revealing behavior. When a platform conditions future terms of trade on observed behavior, it may be possible to use early play to infer private type and then extract information rents by personalizing future terms. We show that this tension is real for many standard learning objectives, but it can be resolved by an appropriately designed algorithm.

The contribution of this paper is threefold. First, we formalize a robustness concern that arises when a privately informed agent commits ex ante to a learning algorithm in a screening environment where the principal's policy may be stationary or adaptive. Second, we show that standard learning objectives do not address this concern: a broad class of off-the-shelf algorithms, including all no-external-regret algorithms, is vulnerable to adaptive principals and allows asymptotic full surplus extraction. Third, we construct a learning algorithm that is minimax-rate-optimal under stationarity and guarantees that, against any adaptive principal, each type's long-run utility is at least its benchmark utility under the static revenue-maximizing menu.

We model a two-player repeated screening problem between a privately informed agent and a principal. The agent privately observes his type before the interaction begins, and commits ex ante to an online learning algorithm. The agent's type is persistent and fixed throughout the game. Each period, the agent chooses an action and observes the realized allocation and payment for the chosen action. The principal observes the committed algorithm and the agent's actions, but not the agent's type. The agent faces am-

biguity about the principal's behavior. Under *stationarity*, the principal's per-period mechanism is drawn i.i.d. from an unknown distribution. Under *adaptation*, the principal chooses mechanisms strategically to maximize expected revenue given a prior over types, taking the committed algorithm as given.

We evaluate learning algorithms along two dimensions. The goal is to achieve the best possible performance when the principal's policy lies in a reference class, while guaranteeing a meaningful payoff when it does not. Specifically, under *stationarity*, the algorithm should learn efficiently: the agent's time-average payoff should approach that of the best action. Under *adaptation*, the algorithm should limit dynamic rent extraction. To formalize this second requirement, we first introduce a conservative notion, *weak extraction robustness*, which rules out principal strategies that extract the agent's full surplus for all types asymptotically. We then introduce a sharper benchmark, *no dynamic rent extraction*, which compares the agent's long-run utility to its utility under the static revenue-maximizing screening menu.

Our first results show that standard learning algorithms can be highly vulnerable when the principal adapts. The common force is that efficient learning induces type-dependent behavior. In particular, any algorithm that satisfies no external regret must respond approximately optimally to mechanisms that are held fixed for long periods. An adaptive principal can exploit this by running a short probing phase to elicit the agent's type, then switching to personalized terms of trade that extract the agent's information rent. We show that standard learning algorithms, including all no-external-regret algorithms, fail to achieve weak extraction robustness, thereby permitting asymptotic full surplus extraction.

We then construct an algorithm that achieves rate-optimal learning under stationarity while satisfying no dynamic rent extraction. The key idea is to combine a statistical test with a credible opt-out threat, calibrated so that deviations from behavior consistent with stationarity are detected at asymptotically negligible cost. The algorithm begins with a short, type-independent exploration phase that establishes a baseline, and then runs a rate-optimal learning procedure while monitoring outcomes against that baseline. If the principal's behavior becomes statistically inconsistent with stationarity in a manner indicative of adaptation, the algorithm triggers and permanently switches to an opt-out action that eliminates profitable trade. This opt-out threat disciplines the adaptive principal: to sustain long-run revenue, the principal must behave as if committed to a stationary policy before it can infer the agent's private information.

The remainder of the paper proceeds as follows. Section 2 presents a motivating example. Section 3 introduces the model. Section 4 defines the design goals. Section 5 establishes impossibility results for standard objectives and shows how canonical algorithms can be exploited. Section 6 presents the proposed algorithms and their guarantees. Section 7 discusses the conceptual limitations of regret-based objectives in strategic environments. Section 8 concludes. All proofs are in the Appendix.

1.1 Related literature

This paper contributes to three strands of literature.

■Exploiting no-regret learners. A growing literature studies how a principal can exploit an agent who relies on a no-regret learning algorithm in repeated interaction, beginning with Braverman et al. (2018). Most existing analyses focus on two benchmark environments. In repeated normal-form games (e.g., Deng et al., 2019b), the principal knows the agent's payoff function. In repeated Bayesian games (e.g., Mansour et al., 2022; Arunachaleswaran et al., 2025), the agent's type is privately observed but is independently re-drawn each period from a fixed distribution; the principal does not observe the realized type.[*1]

In these benchmark environments, *manipulability* (or *strategic robustness*) is typically defined by comparing the principal's payoff under adaptive play against a static benchmark achievable against no-external-regret behavior using a fixed strategy based on the principal's available information. In normal-form games, this benchmark corresponds to the Stackelberg value (Deng et al., 2019b); in Bayesian games, it corresponds to the Myerson revenue (Arunachaleswaran et al., 2025; Kumar et al., 2024). Accordingly, it is shown that some no-external-regret algorithms are manipulable, while no-swap-regret algorithms, a specific subclass of no-external-regret algorithms, and their variants are non-manipulable.

We study a different, economically relevant information structure: repeated buyer-seller interactions with *one-sided fixed private information* (Shalev, 1994; Baron and Besanko, 1984). The agent learns his type before interaction begins, and this type is fixed over time. Relative to normal-form games, the principal initially lacks information about the agent's payoff; relative to Bayesian games, the agent does not receive a fresh private signal each period. In this setting, protecting private information over time is especially important because once information leaks during play, it can be exploited by the other player. Our main result shows that this change in information structure substantially strengthens the scope for exploitation: all no-external-regret algorithms, including algorithms regarded as "non-manipulable" in the previous benchmark settings, are subject to asymptotic full surplus extraction by the principal. This impossibility highlights that when private information is persistent and economically valuable, preventing rent extraction requires going beyond standard no-regret criteria.

■Designing learning rules under ambiguity about opponent's play. No-regret learning has long been studied to microfound static equilibrium notions when players are strategically unsophisticated (e.g., Hart and Mas-Colell, 2000, 2001; Fudenberg and Levine, 1999; Foster and Vohra, 1997).[*2] More recently, motivated by the growing use of machine learning in economic settings, learning dynamics have been studied for their practical relevance, including applications of reinforcement learning in market and industrial-organization environments (Calvano et al., 2020; Asker et al., 2024; Banchio and Mantegazza, 2023), and broader discussions of no-regret learning in economics (Hartline, 2026). A common approach is to treat the learning rule as exogenous and abstract from the agent's incentives for choosing or designing an algorithm.

Our paper instead treats the learning algorithm as a practical commitment device for an agent seeking a better outcome, facing ambiguity about the principal's strategy. A similar perspective is taken in a recent

[*1] A prior-free variant has also been studied; see, for example, Deng et al. (2019a).

[*2] See Fudenberg and Levine (1998) and Roughgarden (2016) for textbook treatments. For Bayesian approaches, see also Kalai and Lehrer (1993); Nachbar (1997).

paper by Arunachaleswaran et al. (2024), who study a setting in which the learner commits ex ante to a learning algorithm while facing ambiguity about the principal's payoff function. In our buyer-seller setting, the principal's objective is clear; the ambiguity instead concerns the class of policies the principal may follow. We assume the agent has limited knowledge and data about the environment and do not endow the agent with a prior over the principal's policies. There is no single standard concept that captures the agent's rationality in such settings. We propose a design goal that targets best performance under a reference class (stationary policies) while guaranteeing a meaningful payoff lower bound when the policy lies outside the reference class (adaptive policies).

■Repeated games with incomplete information. Finally, our paper relates to repeated games with incomplete information, including repeated games with known payoffs (Hart, 1985; Shalev, 1994); see Renault (2020) for a survey. These models typically assume strategically sophisticated players who understand the game structure and best respond given beliefs.

We depart from this approach by combining incomplete information with behavioral and informational constraints motivated by practical applications. The agent does not know about the principal's policy and commits to an online learning algorithm as a practical decision rule. The principal observes this commitment and, if adaptive, best responds to the behavior induced by the algorithm. This asymmetry in knowledge and sophistication both motivates our impossibility result for standard no-regret algorithms and guides our construction of learning rules that perform well under ambiguity about the principal's policy.

2 A motivating example: a vulnerable learning rule

This section illustrates a typical failure mode that motivates the paper. Standard learning algorithms can perform well when the principal's policy is stationary, yet become vulnerable when the principal adapts to the algorithm's behavior. The key force is that efficient learning may induce *type-dependent* action paths. An adaptive principal can use early outcomes to infer private information and then tailor future terms of trade.

Consider a T-period interaction between an agent (an advertiser) and a principal (an auctioneer). The agent has a fixed private valuation

$$\theta \in \{\theta_1, \theta_2\}, \quad 0 < \theta_1 < \theta_2,$$

and chooses in each period whether to submit a low bid ($a_t = 0$) or a high bid ($a_t = 1$). The principal sets a reserve price $p_t \geq 0$. If the agent bids low, he loses and pays 0. If he bids high, he wins and pays p_t. The agent's period payoff is

$$u_t = \begin{cases} 0 & \text{if } a_t = 0, \\ \theta - p_t & \text{if } a_t = 1, \end{cases}$$

and the principal's period revenue is p_t whenever $a_t = 1$.

Suppose that the agent commits ex ante to a standard bandit algorithm such as EXP3 (Auer et al., 2002). At a high level, EXP3 randomizes over actions using weights and shifts probability toward actions that have yielded higher payoffs. In this two-action problem, EXP3 increases the probability of bidding high when bidding high has performed well, and decreases it when bidding high has performed poorly.

■Optimal learning under stationarity. First consider a stationary principal: $(p_t)_t$ is i.i.d.-drawn from a fixed distribution with mean $\overline{p}$. Then, the best fixed action is immediate:

- if $\overline{p} \leq \theta$, bidding high yields positive expected payoff and is optimal;
- if $\overline{p} > \theta$, bidding high yields negative expected payoff, and bidding low is optimal.

Standard bandit algorithms are designed so that, in such stationary environments, the agent's long-run average payoff approaches that of the best fixed action.

■Surplus extraction by an adaptive principal. Now consider an adaptive principal who observes the agent's committed algorithm and past bids, and chooses $(p_t)_t$ to maximize long-run revenue without observing θ. The crucial observation is that EXP3's behavior is *type-dependent*: facing the same price sequence, a higher-valuation agent is more willing to bid high, so the algorithm tends to shift probability toward $a_t = 1$ more quickly for θ_2 than for θ_1.

An adaptive principal can exploit this with a two-phase strategy.

Phase 1 (probing). For an initial block of periods, the principal posts the intermediate reserve price

$$p_t = \frac{\theta_1 + \theta_2}{2}.$$

At this price, bidding high yields negative payoff for type θ_1 and positive payoff for type θ_2. Because EXP3 rapidly shifts weight toward the better-performing action, the frequency of $a_t = 1$ quickly separates across types. By monitoring whether the agent increasingly bids high or increasingly avoids bidding high during this probing phase, the principal can infer the agent's type with high accuracy after relatively few periods.

Phase 2 (extraction). After forming an estimate $\hat{\theta} \in \{\theta_1, \theta_2\}$, the principal posts a personalized reserve price just below it,

$$p_t = \hat{\theta} - \varepsilon,$$

for the remaining periods. If the type is correctly estimated (i.e., $\hat{\theta} = \theta$), it is optimal for the agent to bid high under the new price (it yields payoff $\varepsilon > 0$), so the algorithm will learn to place high probability on $a_t = 1$. The principal then earns revenue approximately $\theta - \varepsilon$ in almost every remaining period in the long run. Since the probing phase can be short relative to T, the principal's average revenue over the entire horizon approaches $\theta - \varepsilon$, while the agent's average surplus approaches ε.

The example isolates a general failure mode: algorithms that learn quickly under stationarity can inad-

vertently act as signals of private information. When the principal can adapt terms of trade to observed behavior, early experimentation can be used to infer the agent's type, and subsequent personalization can extract the agent's information rent. The rest of the paper formalizes this tension and develops learning algorithms that retain good performance under stationarity while limiting dynamic rent extraction under adaptation.

3 Model

We study repeated interaction between a privately informed *agent* and a *principal*. The agent has a fixed private type and commits ex ante to an online learning algorithm. The agent faces ambiguity about whether the principal's behavior is stationary over the learning horizon or instead adapts to the agent's behavior.

3.1 Agent's problem and online learning algorithms

Two players interact for $T \in \mathbb{Z}_{>0}$ periods. The agent's stage-game action set is denoted by a finite set $\mathcal{A}$. In each period, the principal chooses a *mechanism*. A mechanism is a pair (x, p) where $x : \mathcal{A} \to \{0, 1\}$ is an allocation rule and $p : \mathcal{A} \to [0, p_{\max}]$ is a payment rule, with $p_{\max} = 1/2$.[*3] Let

$$\mathcal{M} := \left\{ \{0,1\}^{\mathcal{A}} \times [0, p_{\max}]^{\mathcal{A}} \colon x(a_0) = 0,\ p(a_0) = 0 \right\}$$

denote the set of mechanisms, where $a_0 \in \mathcal{A}$ denotes an opt-out action of the agent such that for all $(x, p) \in \mathcal{M}$, $x(a_0) = 0$ and $p(a_0) = 0$. We assume the existence of the opt-out action.

The agent's type θ is drawn once before the interaction begins from a commonly known finite set $\Theta \subseteq [0, p_{\max})$ and is privately observed by the agent. In each period $t \in \{1, \ldots, T\}$, the agent chooses $a_t \in \mathcal{A}$ and the principal chooses $(x_t, p_t) \in \mathcal{M}$ simultaneously. The agent's stage payoff is

$$u(a_t, (x_t, p_t), \theta) := \theta\, x_t(a_t) - p_t(a_t),$$

and the principal's stage payoff is revenue,

$$v(a_t, (x_t, p_t)) := p_t(a_t).$$

Both players evaluate outcomes by undiscounted time-average payoffs. We assume that Θ contains at least two strictly positive elements.

After period t, the agent observes $(a_t, x_t(a_t), p_t(a_t))$ but not $(x_t(a), p_t(a))$ for actions $a \neq a_t$.[*4] The princi-

[*3] Allowing (x, p) to specify outcomes for each $a \in \mathcal{A}$ is a reduced-form way to represent any per-period policy that maps the agent's current action into an allocation and a payment (e.g., posted terms, scoring rules, reserve prices, or eligibility rules).

[*4] In the online learning literature, this type of feedback is called a partial or bandit feedback. It is closely related to repeated games with imperfect monitoring (Lehrer and Solan, 2016) and repeated games with incomplete information and private learning (Wiseman, 2012).

pal observes the agent's action a_t and the mechanism (x_t, p_t) it chose, but does not observe θ.

A (behavioral) strategy of the agent, typically written as σ_A^T, is a mapping from the agent's private history to a distribution over stage game actions. The agent's private history up to period t consists of his past actions and corresponding outcomes:

$$\Big(a_s, x_s(a_s), p_s(a_s)\Big)_{s=1}^{t-1}.$$

We consider the case in which the agent has limited knowledge of the principal's strategy and therefore relies on an online learning algorithm to guide his decision-making over time. Formally, before the interaction begins, the agent provides (θ, T) to an online learning algorithm and commits to the strategy generated by the algorithm.

Definition 1 (Online learning algorithm)**.** An *online learning algorithm* $\mathcal{L}$ maps (θ, T) into an agent strategy for the T-period game. For each $(\theta, T) \in \Theta \times \mathbb{Z}_{>0}$, we denote by $\mathcal{L}(\theta, T)$ the agent's behavioral strategy generated by the algorithm.

3.2 Stationary and adaptive principals

The principal's behavior may belong to one of two classes.

A *stationary principal* is described by a distribution $\mu \in \Delta(\mathcal{M})$ over mechanisms. In each period t, the principal draws $(x_t, p_t) \sim \mu$ independently across periods. The agent does not know μ ex ante.

An *adaptive principal* observes the agent's committed algorithm before the interaction begins and, during play, observes the history of both players' actions. Formally, an adaptive principal chooses a behavioral strategy σ_P^T that maps its observed history to a distribution over mechanisms. The adaptive principal has a full-support prior $\pi \in \Delta(\Theta)$ over the agent's type (not necessarily known to the agent) and chooses σ_P^T to maximize expected time-average revenue.

The timing is as follows: the horizon T is exogenously given.

1. The agent privately observes his type θ.
2. The agent chooses an online learning algorithm $\mathcal{L}$, inputs (θ, T) to the algorithm, and commits to the generated strategy $\sigma_A^T = \mathcal{L}(\theta, T)$.
3. The principal observes $\mathcal{L}$, but does not observe θ and σ_A^T.
4. The interaction then begins and lasts for T periods.

The agent faces ambiguity about which class the principal belongs to and therefore seeks learning algorithms that perform well against stationary behavior while providing protection against adaptive behavior. We formalize these performance criteria in Section 4.

■Examples The model applies to a range of contemporary platform markets. In many such settings, an agent delegates repeated decisions to an algorithm, observes only realized outcomes, and interacts with a platform that controls exposure and/or fees through policies that are opaque to the agent. These policies

may remain approximately stable over the agent's learning horizon, yet may adapt to observed behavior. The following examples illustrate how the model's primitives map to ad-auction and short-term rental pricing environments. We expect additional applications to arise as advances in machine learning and artificial intelligence further expand the scope of algorithmic decision-making across economic domains.

Example 1 (Ad auctions on search and retail media)**.** An advertiser delegates bidding to an algorithm. The advertiser inputs its value per conversion (or a target CPA/ROAS) into the algorithm. For each query or impression opportunity, the algorithm chooses a bid, and the platform runs an auction to determine whether the advertiser's ad is shown, in which position (allocation), and how much the advertiser pays per click (payment). The advertiser observes only realized outcomes (impressions, clicks, conversions, and spend), not the platform's full auction state. In particular, allocation and payment depend on additional factors such as Ad Rank thresholds (which act like position-specific reserve prices) and auction-time ad-quality signals (predicted relevance) that are computed dynamically for each auction and are not fully observable to advertisers. [*5] This creates ambiguity about the platform's policy.

Example 2 (Pricing on short-term rental platforms)**.** A host delegates pricing to an algorithm. The host inputs constraints and objectives into the algorithm, such as a minimum acceptable nightly rate and a maximum rate, reflecting the host's opportunity cost of occupancy. For each booking opportunity, the algorithm sets a nightly price. The platform then uses a search-ranking algorithm to determine whether the listing is shown and where it appears in search results (allocation). The host observes only realized outcomes (views, bookings, and payouts), not the platform's full search state. In particular, the allocation rule depends on additional factors such as listing quality and popularity, and on search-specific context, which are combined by the platform's ranking algorithm and are not fully observable for hosts. [*6]

Remark 1 (Normalizations and auxiliary assumptions)**.** The bound $p_{\max} = 1/2$ is a normalization; only boundedness matters. We assume $\Theta \subseteq [0, p_{\max})$ so that the principal can post prices above all types. When needed, we impose additional size conditions on $\mathcal{A}$ relative to Θ; these are stated alongside the results that use them.

[*5] Google Ads documents that Ad Rank depends on multiple factors in addition to the bid, and that Ad Rank thresholds are determined dynamically at the time of each auction (https://support.google.com/google-ads/answer/142918 and https://support.google.com/google-ads/answer/7634668). For Amazon sponsored ads, Amazon Advertising states that the final CPC is usually determined by an auction and is based on an adjusted bid plus additional factors (https://advertising.amazon.com/library/guides/cost-per-click).

[*6] The private type in this example is value per booking (revenue minus opportunity cost); the payment is the platform's commission or fee charged to the host. Airbnb documents that search results are ordered by an algorithm that uses many factors, and highlights price, quality, popularity, and (for homes) location as important determinants of how a listing appears in search results (https://www.airbnb.com/help/article/39. Airbnb also documents Smart Pricing as a built-in pricing tool (https://www.airbnb.com/help/article/1168.)

4 Goal setting: optimal learning under stationarity and extraction robustness

As the agent faces ambiguity about the principal's behavior, we evaluate an online learning algorithm along two dimensions. First, the algorithm should achieve nearly optimal payoff when the principal's behavior falls within the reference class, namely, the class of stationary principals. Second, when the principal's behavior falls outside the reference class, that is, when the principal strategically adapts to the agent's behavior, the algorithm should guarantee a meaningful lower bound on the agent's payoff. In particular, it should limit dynamic rent extraction arising from the use of the algorithm's behavior to infer the agent's private type.

We introduce the following notation. Fix a horizon T, agent's type $\theta \in \Theta$, and a strategy profile (σ_A^T, σ_P^T). Let $(a_t, (x_t, p_t))_{t=1}^T$ denote a realized play path. The ex-ante expected time-average payoffs of the agent and the principal are defined as follows:

$$\overline{u}(\sigma_A^T, \sigma_P^T; \theta) := \mathop{\mathbb{E}}_{\sigma_A^T, \sigma_P^T}\left[\frac{1}{T}\sum_{t=1}^{T} u(a_t, (x_t, p_t), \theta)\right],$$
$$\overline{v}(\sigma_A^T, \sigma_P^T) := \mathop{\mathbb{E}}_{\sigma_A^T, \sigma_P^T}\left[\frac{1}{T}\sum_{t=1}^{T} v(a_t, (x_t, p_t))\right],$$

where the expectation is taken over the distribution of $(a_t, (x_t, p_t))_{t=1}^T$ induced by (σ_A^T, σ_P^T).

4.1 Optimal learning under stationarity (OLS)

When the principal is stationary with distribution $\mu \in \Delta(\mathcal{M})$, the per-period mechanism (x_t, p_t) is drawn i.i.d. from μ. In this case, the relevant benchmark is the best *fixed* action in hindsight, i.e., the action that maximizes the agent's expected stage payoff under μ.

Definition 2 (Optimal learning under stationarity (OLS))**.** Fix $(T, \theta, \mu) \in \mathbb{Z}_{>0} \times \Theta \times \Delta(\mathcal{M})$. For an agent's strategy σ_A^T, define its *weak external regret* against μ as

$$\mathrm{WER}(\sigma_A^T; T, \theta, \mu) := T \cdot \max_{a \in \mathcal{A}} \mathop{\mathbb{E}}_{(x,p)\sim\mu}[u(a, (x, p), \theta)] - \mathop{\mathbb{E}}_{\sigma_A^T, \mu}\left[\sum_{t=1}^{T} u(a_t, (x_t, p_t), \theta)\right].$$

An online learning algorithm $\mathcal{L}$ achieves *OLS* if there exists a function $R : \mathbb{Z}_{>0} \to \mathbb{R}_{\geq 0}$ with $R(T) = o(T)$ such that for all (T, θ, μ),

$$\mathrm{WER}(\mathcal{L}(\theta, T); T, \theta, \mu) \leq R(T).$$

When $\mathcal{L}$ achieves OLS, the agent's time-average payoff converges to the stationary benchmark payoff as $T \to \infty$. Many standard bandit algorithms achieve OLS (see Section 5). Some achieve $R(T) = \widetilde{O}(\sqrt{T})$,

and this $\sqrt{T}$ rate is known to be minimax-optimal over stationary principals.[*7] We call an OLS algorithm *rate-optimal* if it achieves $R(T) = \widetilde{O}(\sqrt{T})$.

4.2 Weak extraction robustness and individual rationality

OLS evaluates learning performance when the principal's strategy lies in the reference class. It provides no guarantees when the principal's strategy falls outside the reference class, in particular when the principal is adaptive and uses the agent's early behavior to infer its type and personalize future terms of trade. This subsection and the next formalize robustness requirements for such cases.

We start with a deliberately conservative benchmark that rules out an extreme pathology: no adaptive principal should be able to extract the agent's full surplus uniformly across types as the horizon grows. Formally, we say an online learning algorithm $\mathcal{L}$ *admits uniform asymptotic extraction* if for every $\varepsilon > 0$ and for any sufficiently large T, there exists a principal's strategy σ_P^T such that all $\theta \in \Theta$,

$$\overline{v}\Big(\mathcal{L}(\theta, T), \sigma_P^T\Big) > \theta - \varepsilon.$$

We define weak extraction robustness as the absence of uniform asymptotic extraction.

Definition 3 (Weak extraction robustness (weak-XR))**.** An online learning algorithm $\mathcal{L}$ is *weakly extraction-robust* (weak-XR) if it does not admit uniform asymptotic extraction.

Another conservative requirement for learning algorithms is ex-ante individual rationality.

Definition 4 (Ex-ante asymptotic individual rationality (AIR))**.** A learning algorithm $\mathcal{L}$ satisfies *ex-ante asymptotic individual rationality (AIR)* if there exists a sequence $\varepsilon_T \searrow 0$ such that for any type $\theta \in \Theta$, any horizon T, and any principal's strategy σ_P^T,

$$\overline{u}\left(\mathcal{L}(\theta, T), \sigma_P^T; \theta\right) \geq -\varepsilon_T.$$

Combined with AIR, failure of weak-XR implies that under an adaptive principal who best responds, the agent's long-run payoff converges to zero for every type. To state this formally, fix a learning algorithm $\mathcal{L}$, a prior $\pi \in \Delta(\Theta)$, and a horizon T, and let

$$\mathrm{BR}_T(\pi, \mathcal{L}) \coloneqq \arg\max_{\sigma_P^T} \mathop{\mathbb{E}}_{\theta \sim \pi}\Big[\overline{v}\big(\mathcal{L}(\theta, T), \sigma_P^T\big)\Big]$$

denote the set of the principal's best responses.

Proposition 1. *Fix any full-support prior $\pi \in \Delta(\Theta)$, and let $\mathcal{L}$ be a learning algorithm. Given T, let $\sigma_P^T \in \mathrm{BR}_T(\pi, \mathcal{L})$ be a best response of the adaptive principal.*

[*7] See Lattimore and Szepesvári (2020) for a textbook reference. We use the standard $\widetilde{O}(\cdot)$ notation to suppress polylogarithmic factors: for functions $f, g\colon \mathbb{Z}_{>0} \to \mathbb{R}_{>0}$, we write $f(T) = \widetilde{O}(g(T))$ if $f(T) = O\big(g(T)\,(\log T)^c\big)$ for some constant $c \geq 0$.

If $\mathcal{L}$ satisfies AIR and fails weak-XR, then for every $\theta \in \Theta$,

$$\lim_{T\to\infty} \overline{u}\Big(\mathcal{L}(\theta, T), \sigma_P^T; \theta\Big) = 0.$$

4.3 No dynamic rent extraction

Weak-XR is conservative and its implication for the agent's surplus can be limited: it is stated in terms of the principal's revenue and does not directly guarantee that each type retains a meaningful surplus. A sharper benchmark compares the dynamic interaction to a canonical *static* adverse-selection problem.

■Static benchmark. Consider a one-shot principal-agent problem in which the principal chooses a menu of (stochastic) allocations and payments indexed by the agent's action set,

$$(\overline{x}, \overline{p}) \in [0,1]^{\mathcal{A}} \times [0, p_{\max}]^{\mathcal{A}},$$

and an agent with type θ chooses an action $a \in \mathcal{A}$ to maximize $\theta\overline{x}(a) - \overline{p}(a)$. Let $a^*(\theta, (\overline{x}, \overline{p}))$ denote a principal-favorable selection from $\arg\max_{a\in\mathcal{A}}\{\theta\overline{x}(a) - \overline{p}(a)\}$. Given a prior $\pi \in \Delta(\Theta)$, define the set of revenue-optimal menus

$$\mathcal{M}^*(\pi) := \arg\max_{(\overline{x},\overline{p})} \mathop{\mathbb{E}}_{\theta\sim\pi}\left[\overline{p}\big(a^*(\theta, (\overline{x}, \overline{p}))\big)\right],$$

and define the agent's benchmark utility (robust to multiplicity of optimal menus) by

$$u^*(\theta, \pi) := \min_{(\overline{x},\overline{p})\in\mathcal{M}^*(\pi)} \left\{\theta\overline{x}\big(a^*(\theta, (\overline{x}, \overline{p}))\big) - \overline{p}\big(a^*(\theta, (\overline{x}, \overline{p}))\big)\right\}.$$

Definition 5 (No dynamic rent extraction (no-DRE))**.** We say that $\mathcal{L}$ satisfies *no dynamic rent extraction* if for every $\pi \in \Delta(\Theta)$, every type $\theta \in \Theta$, and every selection $\sigma_P^T \in \mathrm{BR}_T(\pi, \mathcal{L})$ (for each T),

$$\liminf_{T\to\infty} \overline{u}\left(\mathcal{L}(\theta, T), \sigma_P^T; \theta\right) \geq u^*(\theta, \pi).$$

No-DRE gives a direct interpretation: against an adaptive principal, the agent's long-run utility is at least the utility guaranteed by the static revenue-maximizing benchmark, type by type. In particular, dynamic interaction does not allow the principal to extract more rent from the agent than in the static benchmark.

Under AIR, the relationship between weak-XR and no-DRE is simple: no-DRE is the stronger requirement.

Proposition 2. *If the learning algorithm $\mathcal{L}$ satisfies AIR and no-DRE, then $\mathcal{L}$ is weak extraction-robust.*

Remark 2. Without AIR, no-DRE does not imply weak-XR in general, and failure of weak-XR does not imply that *all* types have asymptotically nonpositive payoffs, since it may be possible that there exists a

type θ from which the principal extracts strictly more than θ. Our proposed algorithms (see Section 6) satisfy AIR and no-DRE, and thus they also satisfy weak-XR.

5 Limits of standard learning algorithms

This section shows that standard learning algorithms can be highly vulnerable when the principal can adapt to the algorithm's behavior. The key economic force is simple: algorithms that learn quickly typically generate type-dependent action paths, which allow an adaptive principal to infer the agent's private type and then adjust terms to extract the agent's information rent.

We proceed in three steps. First, we show that the widely used design criterion of *no external regret* (no-ER) is incompatible with weak extraction robustness (weak-XR), thereby permitting asymptotic full surplus extraction. Because many standard online learning algorithms satisfy no-ER, this result implies that a broad class of off-the-shelf algorithms is not weakly extraction-robust, even though some of them are regarded as "non-manipulable" in other settings. Second, we show that the same conclusion holds for several canonical bandit algorithms that achieve optimal learning under stationarity (OLS) but do not satisfy the no-ER condition.

5.1 No-external-regret algorithms

Satisfying *no external regret* (no-ER) is a popular design goal in economics and computer science. It is also known as *Hannan consistency* (Hannan, 1957) or *universal consistency* (Fudenberg and Levine, 1995). In the current setup, external regret compares the agent's cumulative payoff to that of the best fixed action in hindsight, holding the principal's realized mechanism sequence fixed. The no-ER requirement demands that, for any realized sequence of mechanisms chosen by the principal, the agent's expected payoff is asymptotically as large as the best fixed action in hindsight.

Definition 6 (No external regret (no-ER))**.** Given horizon T, type θ, and a principal's mechanism sequence $(x_t, p_t)_{t=1}^T$, the *external regret* of an agent strategy σ_A^T is

$$\mathrm{ER}\left(\sigma_A^T; T, \theta, (x_t, p_t)_{t=1}^T\right) := \max_{a \in \mathcal{A}} \sum_{t=1}^T u(a, (x_t, p_t), \theta) - \mathop{\mathbb{E}}_{\sigma_A^T}\left[\sum_{t=1}^T u(a_t, (x_t, p_t), \theta)\right], \tag{1}$$

where the expectation is taken over the distribution of the action sequence $(a_t)_{t=1}^T$ induced by σ_A^T. An online learning algorithm $\mathcal{L}$ has *no external regret (no-ER)* if there exists $R\colon \mathbb{Z}_{>0} \to \mathbb{R}_{\geq 0}$ such that for any T, θ, and $(x_t, p_t)_{t=1}^T$, we have

$$\mathrm{ER}\left(\mathcal{L}(\theta, T); T, \theta, (x_t, p_t)_{t=1}^T\right) \leq R(T), \tag{2}$$

with $R(T) = o(T)$.[*8]

A canonical example is EXP3, discussed in Section 2.

Example 3 (EXP3)**.** EXP3 (Algorithm 1 in Appendix A.2) has no-ER and is rate-optimal.[*9] The algorithm puts more weight on actions that historically performed well. Because it only observes the payoff of the chosen action, it updates weights using an importance-weighted estimate of each action's payoff, whose conditional expectation equals the true per-round payoff.

A key reason no-ER is widely used is that it implies efficient learning under stationarity.

Proposition 3. *Any no-ER algorithm achieves OLS.*

Despite its popularity, no-ER algorithms can be exploited by an adaptive principal. The no-ER criterion benchmarks performance against the best fixed action for the *realized* sequence of mechanisms. Consequently, if the principal holds a particular mechanism fixed for sufficiently many periods, a no-ER algorithm must play approximately a best response to that mechanism. An adaptive principal can use this implication to run a short "probing" phase that makes the agent's play informative about θ, infer the agent's type from the resulting action frequencies, and then adjust future mechanisms to extract the agent's information rent.

Theorem 1. *Any no-ER algorithm fails weak extraction robustness.*

This impossibility extends to more stringent regret notions that retain no-ER as a baseline requirement. For example, no-swap-regret algorithms (Deng et al., 2019b; Arunachaleswaran et al., 2024) and online gradient descent (Kumar et al., 2024) imply no-ER and therefore inherit the same vulnerability, even though previous work establishes "non-manipulability" of these algorithms in alternative settings.[*10]

■Agent's payoff under no-ER algorithms The opt-out action a_0 yields payoff 0 in every period under any principal behavior. Combined with no-ER, this implies AIR: the agent cannot do asymptotically worse than always opting out.

Lemma 1. *Any no-ER algorithm satisfies AIR.*

As no-ER algorithms fail weak-XR, by Proposition 1, this implies that all types of agents experience a payoff of zero asymptotically when using a no-ER algorithm against an adaptive principal.

Corollary 1. *Suppose the principal is adaptive with prior π. If the agent uses a no-ER algorithm $\mathcal{L}$ and the principal plays a best response $\sigma_P^T \in \mathrm{BR}_T(\pi, \mathcal{L})$ for each T, then for every $\theta \in \Theta$,*

$$\lim_{T\to\infty} \overline{u}\Big(\mathcal{L}(\theta, T), \sigma_P^T; \theta\Big) = 0.$$

[*8] For a function $R\colon \mathbb{Z}_{>0} \to \mathbb{R}_{\geq 0}$, $R(T) = o(T)$ means $\limsup_{T\to\infty} \frac{R(T)}{T} = 0$.

[*9] See Theorem 11.2 of Lattimore and Szepesvári (2020) for a textbook reference.

[*10] Other examples of learning algorithms with more stringent regret notions include learning rules that satisfy conditional universal consistency (Fudenberg and Levine, 1999), as well as algorithms with low dynamic regret (Hazan and Seshadhri, 2007, 2009) or low adaptive regret (Herbster and Warmuth, 1998; Zinkevich, 2003).

Remark 3. Theorem 6 of Brown et al. (2023) implies the same conclusion as Theorem 1 under an additional assumption that the no-ER condition holds *anytime* (see Brown et al. (2023) for the definition).[*11]

5.2 Other standard learning algorithms

Theorem 1 does not imply that OLS is incompatible with weak extraction robustness, because no-ER is only a sufficient condition for OLS. Indeed, the stochastic bandit literature provides canonical algorithms that achieve OLS without satisfying no-ER. Below, we describe three standard examples.[*12]

Example 4 (Uniform Exploration)**.** Uniform Exploration (UE) tries each action T_1 times, and then chooses the action with the highest average realized payoff in all remaining periods. It is well known that choosing $T_1 := T^{2/3}(\log T)^{1/3}$ yields regret of order $\widetilde{O}(T^{2/3})$, and hence UE achieves OLS. UE is not rate-optimal because it explores for too long, trying suboptimal actions too many times.

Example 5 (Successive Elimination)**.** Successive Elimination (SE) achieves OLS with regret of order $\widetilde{O}(T^{1/2})$, which is rate-optimal. It constructs confidence intervals for the expected payoff of each action and eliminates actions that are confidently suboptimal: these occur when the upper bound of an action's confidence interval falls strictly below the lower bound of another action's confidence interval. (see Algorithm 2 in Appendix A.2). Under stationarity, the confidence intervals cover the true means $\mathbb{E}_\mu[u(a,(x,p),\theta)]$ with high probability in all periods (see Appendix A.1).

Example 6 (UCB)**.** UCB is another rate-optimal OLS algorithm. It balances exploration and exploitation by choosing the action with the highest upper confidence bound in each period (see Algorithm 3 in Appendix A.2).

These algorithms achieve OLS, but they also fail weak extraction robustness. The proofs highlight two failure modes. First, an algorithm can fail weak-XR if it commits to a specific action after a finite learning phase; then an adaptive principal can raise the payment for that action once the commitment occurs. This applies to UE and SE. Second, an algorithm can fail weak-XR if its early exploration is sufficiently type-revealing; then the principal can infer the agent's type and tailor future mechanisms. This applies to UCB and, as above, to no-ER algorithms.

Theorem 2. *1. Uniform Exploration and Successive Elimination achieve OLS but fail weak-XR.*

2. UCB achieves OLS. Moreover, if $|\mathcal{A}| \geq |\Theta| + 1$, then UCB fails weak-XR.

Remark 4. The condition $|\mathcal{A}| \geq |\Theta| + 1$ is natural in applications where the action space is sufficiently rich relative to the finite type space. We expect the conclusion to extend beyond this condition, but do not pursue it here.

[*11] In our setting, the opt-out action allows a proof without the anytime requirement. Brown et al. (2023) impose anytime no-ER because they analyze a broader class of games. I thank an anonymous referee for pointing this out.

[*12] See Chapter 1 of Slivkins (2019) for textbook references. UE, SE, and UCB are not no-ER algorithms as their behavior is deterministic: it is known that any deterministic algorithm has a linear external regret (see, for example, Roughgarden (2016).)

6 Robust algorithms

This section presents two learning algorithms that achieve optimal learning under stationarity (OLS) while achieving no dynamic rent extraction (no-DRE). Both algorithms implement the same design principle: The agent first collects data to form a baseline in a type-independent manner. This pins down the allocation and payment outcomes that are consistent with stationarity. The algorithm then learns efficiently as long as realized outcomes remain consistent with this baseline. If outcomes drift in a statistically detectable way, the algorithm permanently opts out. This "test-and-opt-out" structure makes history-dependent price discrimination unprofitable before the principal can infer the agent's persistent private information.

The first algorithm, *UE-Trigger* (UE-T), is deliberately simple and highlights the core idea. The second, *UE-SE-Trigger* (UE-SE-T), applies the idea while incorporating a rate-optimal learning procedure. It achieves rate-optimal learning under stationarity while still satisfying no-DRE.

Throughout this section, we assume $|\mathcal{A}| \geq |\Theta| + 1$.[*13]

6.1 A simple robust algorithm

We first establish the existence of a simple algorithm that achieves OLS and no dynamic rent extraction. We term it *UE-Trigger* (UE-T; see Algorithm 5 in the Appendix for pseudocode.) The algorithm has three phases:

Phase 1 (type-independent exploration). For the first $|\mathcal{A}|T_1$ periods, the agent cycles through actions in round-robin fashion so that each action is played exactly T_1 times. At the end of Phase 1, the algorithm selects an action a^* that attains the highest empirical payoff in Phase 1. Using only the Phase-1 samples of a^*, it constructs baseline confidence intervals for the mean allocation and mean payment induced by a^*.

Phase 2 (exploitation with protection). The agent repeatedly plays a^*. Using the Phase-2 samples, it continuously recomputes confidence intervals for the mean allocation and mean payment induced by a^*. The algorithm *triggers* and exits Phase 2 as soon as either the Phase-2 allocation interval or the Phase-2 payment interval becomes disjoint from its Phase-1 counterpart.

Phase 3 (opt out). After triggering, the agent plays the opt-out action a_0 forever.

Under stationarity, the distribution of outcomes induced by a fixed action is stable. As a result, the Phase-1 and Phase-2 estimates of $(x(a^*), p(a^*))$ remain consistent with high probability, so the algorithm rarely opts out. Conditional on no trigger, the agent plays a^* for all but a vanishing fraction of periods, which yields OLS.

Under adaptation, any attempt to worsen the continuation terms for a^* after it is selected must change

[*13] In our model, a principal's per-period policy is a menu indexed by actions in $\mathcal{A}$. A larger $|\mathcal{A}|$ enlarges the set of menus the principal can implement and thus weakly increases its screening power in the adaptive case. For this reason, we view the condition as conservative for our robustness guarantees.

either the expected allocation or the expected payment induced by a^*. Because the algorithm monitors both allocation and payment, such changes are detected quickly (within a sublinear number of periods), which triggers opt-out. Anticipating this, a revenue-maximizing adaptive principal keeps the continuation outcomes following a^* essentially fixed after Phase 1 ends. Moreover, because Phase 1 is type-independent, the principal's behavior during Phase 1 cannot condition on the realized type and must be chosen based on the prior. Consequently, dynamic interaction does not allow the principal to extract more rent than in the static benchmark, which delivers the no-DRE guarantee.

Theorem 3 (Properties of UE-T)**.** *For sufficiently large T, UE-T (Algorithm 5) with*

$$T_1 := T^{2/3}(\log T)^{1/3}$$

achieves OLS and no dynamic rent extraction.

6.2 Rate-optimal improvement

Algorithm UE-T achieves OLS and no-DRE, but it does not achieve rate-optimal learning under stationarity: its weak external regret is $\widetilde{O}(T^{2/3})$, which is larger than $\widetilde{O}(\sqrt{T})$, because Phase 1 lasts $\widetilde{O}(T^{2/3})$ periods. We now introduce UE-SE-T, which preserves the same discipline on an adaptive principal while attaining the minimax-optimal $\widetilde{O}(\sqrt{T})$ regret rate under stationarity.

A naive modification would replace the exploitation phase of UE-T with a rate-optimal bandit algorithm such as Successive Elimination or UCB. This approach fails for the reason emphasized in Section 5: rate-optimal exploration under these algorithms is payoff-responsive, and in our model, payoffs depend on the private type θ. Hence, the induced action path can become type-revealing, allowing an adaptive principal to infer θ and then personalize terms. The design of our proposed algorithm (UE-SE-T; see Algorithm 6 in the Appendix for pseudocode) resolves this by separating baseline formation phase from efficient learning phase: Phase 1 generates a baseline for the allocation and payment induced by each action by uniform type-independent exploration, while Phase 2 runs a rate-optimal learning rule à la Successive Elimination but is constrained to remain consistent with the baseline.

Phase 1 (type-independent exploration). Let $m := \lceil \sqrt{T} \rceil$. For $|\mathcal{A}|m$ periods, the agent cycles through actions in round-robin fashion. For each $a \in \mathcal{A}$, it records the empirical mean allocation $\overline{x}_a$ and empirical mean payment $\overline{p}_a$ from the m plays of a, and forms baseline confidence intervals $I^{(1)}_{x,a}$ and $I^{(1)}_{p,a}$ for the corresponding mean allocation and mean payment.

Phase 2 (successive elimination with protection). Starting from the full active set $\mathcal{A}_{\text{act}} = \mathcal{A}$, the agent runs Successive Elimination using period payoff $u_t = \theta x_t(a_t) - p_t(a_t)$. Each time an action a_t is played, the algorithm updates the empirical mean allocation and payment for that action and forms Phase-2 confidence intervals $I^{(2)}_{x,a_t}(t)$ and $I^{(2)}_{p,a_t}(t)$. The protection rule is action-by-action: the algorithm *triggers* and exits Phase 2

as soon as, for the currently played action a_t, either

$$I^{(2)}_{x,a_t}(t) \cap I^{(1)}_{x,a_t} = \emptyset \qquad \text{or} \qquad I^{(2)}_{p,a_t}(t) \cap I^{(1)}_{p,a_t} = \emptyset.$$

If no trigger occurs, the algorithm eliminates actions according to the standard SE rule based on payoff estimates and confidence radii.

Phase 3 (opt out). After triggering, the agent plays a_0 forever.

Under stationarity, for each fixed action a the induced allocation and payment outcomes are drawn i.i.d. from time-invariant distributions with stable means. With high probability, the Phase-2 confidence intervals for $(x(a), p(a))$ remain consistent with the Phase-1 baseline intervals whenever a is played, so the trigger rarely fires. Conditional on not triggering, Phase 2 is exactly a standard run of Successive Elimination with m warm-start samples per action, and thus achieves the minimax-optimal $\widetilde{O}(\sqrt{T})$ regret rate. Since Phase 1 lasts $|\mathcal{A}|m = O(\sqrt{T})$ rounds, it contributes only $O(\sqrt{T})$ regret, so the overall regret remains $\widetilde{O}(\sqrt{T})$.

Under adaptation, the action-by-action trigger rule disciplines the principal's behavior after Phase 1. Any statistically detectable deviation in the allocation or payment induced by a played action causes permanent opt-out after a sublinear delay. Thus, along any history in which the algorithm remains in Phase 2, the principal must keep the induced outcomes for each played action a consistent with the baseline set in Phase 1. In this sense, the principal is forced to behave as if it were committed to an approximately fixed menu $(\overline{x}_a, \overline{p}_a)_a$ pinned down by the Phase-1 baselines before it can profitably infer the agent's type.

The new technical issue relative to UE-T is that the played action can vary over time in Phase 2. A key lemma (proved in the Appendix) shows that, conditional on not triggering, the actions chosen by Successive Elimination are approximately optimal for the menu $(\overline{x}_a, \overline{p}_a)_a$ pinned down at the end of Phase 1. This approximate best-response behavior implies that an adaptive principal's best-response revenue is asymptotically bounded by the static revenue benchmark. Applying an approximate mechanism-design argument to the Phase-2 averages then yields no-DRE.

Theorem 4 (Properties of UE-SE-T)**.** *For sufficiently large T, UE–SE–T (Algorithm 6) with $m := \lceil \sqrt{T} \rceil$ achieves OLS and no dynamic rent extraction. Moreover, it is rate-optimal under stationarity.*

7 Discussion: a conceptual problem of no external regret

This paper highlights a mismatch between standard regret-based objectives and economic environments in which the opponent's behavior is endogenous to the learner's strategy. The issue is not that no-external-regret (no-ER) guarantees are weak. Rather, the counterfactual underlying external regret—holding the realized sequence of opponent actions fixed while changing only the learner's actions in the benchmark—need not correspond to a meaningful comparison when the opponent adapts strategically.

No-ER algorithms (Definition 6) are widely studied and used (e.g., Nekipelov et al., 2015; Lomys and

Magnolfi, 2024) because they provide a distribution-free guarantee: near-optimal performance relative to the best fixed action in hindsight for any realized sequence of payoff-relevant objects (such as states and other players' actions). This guarantee is built around external regret, which evaluates the learner against a benchmark that replaces the realized learner's action sequence with a single fixed action while keeping the rest of the realized history unchanged. This benchmark is meaningful when the relevant sequence can be treated as exogenous to the learner's deviations, as in stationary bandit problems; accordingly, no-ER implies optimal learning under stationarity (Proposition 3). It is also well known that in two-player zero-sum games, no-ER suffices to secure the value (i.e., the Nash equilibrium payoff) asymptotically (Freund and Schapire, 1999). Outside these settings, however, the opponent's action sequence may respond to the learner's behavior, and the external-regret benchmark need not correspond to a payoff-relevant counterfactual.

Our fixed-type adverse-selection environment provides a sharp illustration. With persistent private information, payoff-responsive experimentation that achieves efficient learning under stationarity is inherently type-dependent. Theorem 1 shows that this exposes no-ER algorithms to probing and subsequent personalization by an adaptive principal, permitting asymptotic full surplus extraction. In this sense, interpreting no-ER as a form of robustness can be misleading in screening environments: the requirement that the learner achieve low external regret against any realized opponent behavior is precisely what forces informative, type-revealing experimentation.

Related concerns about external regret in nonstationary environments have been noted in the learning literature. For example, Arora et al. (2012, 2018) propose *policy regret*, which evaluates a learner against counterfactual histories induced by alternative learner policies, and show that meaningful guarantees generally require restricting the class of adversaries (e.g., to m-memory-bounded adversaries). Our approach follows the same general logic—obtaining meaningful guarantees by imposing structure on the adversary—but we build in structure by assuming that the principal is an optimizing agent rather than an arbitrary adversary: we assume that, when the principal is adaptive, it best responds to the committed algorithm to maximize revenue, which restricts the class of policies the agent needs to guard against. More generally, when the data-generating process of payoff-relevant objects responds to the learner's behavior, treating agents as no-regret learners is a substantive behavioral assumption rather than a neutral robustness benchmark, and such agents may deliberately avoid no-regret learning in screening environments.[*14]

8 Conclusion

This paper studies learning in economic environments where a learning agent with fixed private information repeatedly interacts with a principal, facing ambiguity about the principal's policy. We model learning algorithms as practical commitment devices and evaluate them not only by their performance in the ref-

[*14] Camara et al. (2020) and Collina et al. (2024) introduce a regret notion to capture learners' rationality in prior-free settings, where both a principal and an agent simultaneously learn the state distribution. Unlike our model, the agent in their framework does not have private types.

erence class (stationary principal), but also by the payoff guarantee outside the reference class (adaptive principal).

Our main negative result shows that these objectives can conflict. In a fixed-type adverse-selection setting, a broad class of standard algorithms, including all no-external-regret algorithms, can be exploited by an adaptive principal: typical payoff-responsive experimentation strategies reveal persistent private information, enabling history-dependent personalization and asymptotic full surplus extraction. Compared with existing work in which agents' types are redrawn each period, the persistence of private information substantially increases the scope for exploitation.

Our main positive result shows that this vulnerability is not inherent to efficient learning under stationarity. We construct a learning algorithm that achieves the minimax-optimal $\widetilde{O}(\sqrt{T})$ regret rate under stationary policies while guaranteeing no dynamic rent extraction against an adaptive principal. The algorithm implements a test-and-opt-out architecture: it forms a type-independent baseline, learns efficiently as long as outcomes remain statistically consistent with that baseline, and otherwise switches permanently to an opt-out action. This credible exit threat makes revenue-relevant deviations unprofitable before the principal can infer the agent's type.

More broadly, our results suggest that when learning agents with private information operate in environments where some components of the game can adapt to their behavior, standard regret-based criteria are incomplete. Designing learning rules for such environments requires performance criteria that explicitly incorporate economically motivated ambiguity—particularly the possibility of strategic policy adaptation—and that are aligned with payoff-relevant benchmarks. Developing such frameworks for richer market settings, including interactions among multiple learning agents and alternative payoff and information structures, remains an important direction for future research.

References

Arora, Raman, Michael Dinitz, Teodor Vanislavov Marinov, and Mehryar Mohri, "Policy Regret in Repeated Games," in "Advances in Neural Information Processing Systems," Vol. 31 2018, pp. 6732–6741.

_ , Ofer Dekel, and Ambuj Tewari, "Online Bandit Learning against an Adaptive Adversary: from Regret to Policy Regret," in "Proceedings of the 29th International Conference on Machine Learning" 2012, pp. 1503–1510.

Arunachaleswaran, Eshwar Ram, Natalie Collina, and Jon Schneider, "Pareto-Optimal Algorithms for Learning in Games," in "Proceedings of the 25th ACM Conference on Economics and Computation" 2024, pp. 490–510.

_ , _ , Yishay Mansour, Mehryar Mohri, Jon Schneider, and Balasubramanian Sivan, "Swap Regret and Correlated Equilibria Beyond Normal-Form Games," in "Proceedings of the 26th ACM Conference on Economics and Computation" Association for Computing Machinery 2025, pp. 130–157.

Asker, John, Chaim Fershtman, and Ariel Pakes, "The impact of artificial intelligence design on pricing," *Journal of Economics & Management Strategy*, 2024, *33* (2), 276–304.

Auer, Peter, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire, "The Nonstochastic Multiarmed Bandit Problem," *SIAM Journal on Computing*, 2002, *32* (1), 48–77.

Banchio, Martino and Giacomo Mantegazza, "Adaptive Algorithms and Collusion via Coupling," in "Proceedings of the 24th ACM Conference on Economics and Computation" 2023.

Baron, David P. and David A. Besanko, "Regulation, Asymmetric Information, and Auditing," *The RAND Journal of Economics*, 1984, *15* (4), 447–470.

Braverman, Mark, Jieming Mao, Jon Schneider, and S. Matthew Weinberg, "Selling to a No-Regret Buyer," in "Proceedings of the 2018 ACM Conference on Economics and Computation" 2018, pp. 523–538.

Brown, William, Jon Schneider, and Kiran Vodrahalli, "Is Learning in Games Good for the Learners?," in "Advances in Neural Information Processing Systems," Vol. 36 2023, pp. 1–22.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello, "Artificial Intelligence, Algorithmic Pricing, and Collusion," *American Economic Review*, 2020, *110* (10), 3267–3297.

Camara, Modibo Khane, Jason D. Hartline, and Aleck C. Johnsen, "Mechanisms for a No-Regret Agent: Beyond the Common Prior," in "Proceedings of the 61st Annual IEEE Symposium on Foundations of Computer Science" 2020, pp. 259–270.

Collina, Natalie, Aaron Roth, and Han Shao, "Efficient Prior-Free Mechanisms for No-Regret Agents," in "Proceedings of the 25th ACM Conference on Economics and Computation" Association for Computing Machinery 2024, pp. 511–541.

Deng, Yuan, Jon Schneider, and Balasubramanian Sivan, "Prior-Free Dynamic Auctions with Low Regret Buyers," in "Advances in Neural Information Processing Systems," Vol. 32 2019, pp. 4803–4813.

_ , _ , and _ , "Strategizing against No-regret Learners," in "Advances in Neural Information Processing

Systems," Vol. 32 2019, pp. 1579–1587.

Foster, Dean P. and Rakesh V. Vohra, "Calibrated Learning and Correlated Equilibrium," *Games and Economic Behavior*, 1997, *21* (1–2), 40–55.

Freund, Yoav and Robert E. Schapire, "Adaptive Game Playing Using Multiplicative Weights," *Games and Economic Behavior*, 1999, *29* (1-2), 79–103.

Fudenberg, Drew and David K. Levine, "Consistency and cautious fictitious play," *Journal of Economic Dynamics and Control*, 1995, *19* (5–7), 1065–1089.

_ **and David K Levine**, *The Theory of Learning in Games*, Vol. 2 of *Economic Learning and Social Evolution*, MIT Press, 1998.

_ **and David K. Levine**, "Conditional Universal Consistency," *Games and Economic Behavior*, 1999, *29* (1-2), 104–130.

Hannan, James F., "Approximation to Bayes risk in repeated play," in "Contributions to the Theory of Games, Volume 3, Annals of Mathematics Studies 39," Vol. 3 1957, pp. 97–139.

Hart, Sergiu, "Non-zero-sum 2-person Repeated Games with Incomplete Information," *Mathematics of Operations Research*, 1985, *10* (1), 117–153.

_ **and Andreu Mas-Colell**, "A Simple Adaptive Procedure Leading to Correlated Equilibrium," *Econometrica*, 2000, *68* (5), 1127–1150.

_ **and** _ , "A General Class of Adaptive Strategies," *Journal of Economic Theory*, 2001, *98* (1), 26–54.

Hartline, Jason, "The Economics of No-regret Learning Algorithms," *arXiv preprint arXiv:2601.22079*, 2026.

Hazan, Elad and Comandur Seshadhri, "Adaptive Algorithms for Online Decision Problems," *Electronic Colloquium on Computational Complexity*, 2007, *14* (88), 1–19.

_ **and** _ , "Efficient learning algorithms for changing environments," in "Proceedings of the 26th International Conference on Machine Learning" 2009, pp. 393–400.

Herbster, Mark and Manfred K. Warmuth, "Tracking the Best Expert," *Machine Learning*, 1998, *32* (2), 151–178.

Kalai, Ehud and Ehud Lehrer, "Rational Learning Leads to Nash Equilibrium," *Econometrica*, 1993, *61* (5), 1019–1045.

Kumar, Rachitesh, Jon Schneider, and Balasubramanian Sivan, "Strategically-Robust Learning Algorithms for Bidding in First-Price Auctions," in "Proceedings of the 25th ACM Conference on Economics and Computation" 2024.

Lattimore, Tor and Csaba Szepesvári, *Bandit Algorithms*, Cambridge University Press, 2020.

Lehrer, Ehud and Eilon Solan, "A General Internal Regret-Free Strategy," *Dynamic Games and Applications*, 2016, *6* (1), 112–138.

Lomys, Niccolò and Lorenzo Magnolfi, "Estimation of Games under No Regret: Structural Econometrics for AI," *SSRN Electronic Journal*, 2024, pp. 1–76.

Mansour, Yishay, Mehryar Mohri, Jon Schneider, and Balasubramanian Sivan, "Strategizing against Learners in Bayesian Games," in "Proceedings of Thirty Fifth Conference on Learning Theory," Vol. 178

2022, pp. 5221–5252.

Nachbar, John H., "Prediction, Optimization, and Learning in Repeated Games," *Econometrica*, 1997, *65* (2), 275–309.

Nekipelov, Denis, Vasilis Syrgkanis, and ÃL'va Tardos, "Econometrics for Learning Agents," in "Proceedings of the Sixteenth ACM Conference on Economics and Computation" 2015, pp. 1–18.

Renault, Jérôme, "Repeated Games with Incomplete Information," in Marilda Sotomayor, David Pérez-Castrillo, and Filippo Castiglione, eds., *Complex Social and Behavioral Systems : Game Theory and Agent-Based Models*, New York, NY: Springer US, 2020, pp. 157–184.

Roughgarden, Timothy Avelin, *Twenty Lectures on Algorithmic Game Theory*, Cambridge University Press, 2016.

Shalev, Jonathan, "Nonzero-Sum Two-Person Repeated Games with Incomplete Information and Known-Own Payoffs," *Games and Economic Behavior*, 1994, *7* (2), 246–259.

Slivkins, Aleksandrs, "Introduction to Multi-Armed Bandits," *Foundations and Trends in Machine Learning*, 2019, *12* (1–2), 1–286.

Wiseman, Thomas E., "A partial folk theorem for games with private learning," *Theoretical Economics*, 2012, *7* (2), 217–239.

Zinkevich, Martin, "Online Convex Programming and Generalized Infinitesimal Gradient Ascent," in "Proceedings of the Twentieth International Conference on Machine Learning" 2003, pp. 928–936.

Appendix

A Preliminaries

A.1 Confidence intervals

Lemma A.1 (Hoeffding's inequality)**.** *For any independent random variables $\mu_1, \mu_2, \ldots, \mu_T$ such that $a_t \leq \mu_t \leq b_t$ almost surely, and for their empirical mean $\hat{\mu}_T = \frac{1}{T}\sum_{t=1}^{T} \mu_t$, the following inequality holds for all $\varepsilon > 0$:*

$$\Pr\left(\hat{\mu}_T - \mathbb{E}[\hat{\mu}_T] \geq \varepsilon\right) \leq \exp\left(-\frac{2T^2\varepsilon^2}{\sum_{t=1}^{T}(b_t - a_t)^2}\right).$$

Similarly,

$$\Pr\left(\hat{\mu}_T - \mathbb{E}[\hat{\mu}_T] \leq -\varepsilon\right) \leq \exp\left(-\frac{2T^2\varepsilon^2}{\sum_{t=1}^{T}(b_t - a_t)^2}\right).$$

Combining these, we get:

$$\Pr\left(|\hat{\mu}_T - \mathbb{E}[\hat{\mu}_T]| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2T^2\varepsilon^2}{\sum_{t=1}^{T}(b_t - a_t)^2}\right).$$

Corollary A.1. *Fix any $T \in \mathbb{Z}_{>0}$. Suppose that $(\mu_t(a))_{t=1}^{T}$ is iid-drawn from some $P(a)$ with mean $\mu(a)$, and $\mu_t(a) \in [0,1]$ for any $t \in [T]$ and $a \in \mathcal{A}$ almost surely. Let*

$$\rho_t(a) \coloneqq \sqrt{\frac{2\log T}{t}}.$$

Then, for any $t \leq T$, we have

$$\Pr_P\left(|\hat{\mu}_t(a) - \mu(a)| \geq \rho_t(a)\right) < \frac{2}{T^4}.$$

Moreover, for any strategy σ_A^T, t, and a, we have

$$\Pr_{P,\sigma_A^T}\left(|\hat{\mu}_t(a) - \mu(a)| \geq \rho_t(a)\right) < \frac{2}{T^4},$$

where

$$n_t(a) \coloneqq \sum_{s=1}^{t} \mathbb{1}\{a_s = a\}, \quad \rho_t(a) \coloneqq \sqrt{\frac{2\log T}{n_t(a)}}.$$

Proof. Fix any $t \leq T$. Let $a_t \equiv 0$, $b_t \equiv 1$, and $\varepsilon := \sqrt{(2\log T)/t}$. Applying Hoeffding's inequality, we obtain

$$2\exp\left(-4\frac{T^2}{t^2}\log T\right) \leq 2\exp\left(-4\log T\right) = \frac{2}{T^4}.$$

□

The confidence intervals are defined as follows (see also Example 5):

$$\begin{aligned} n_t(a) &:= \textstyle\sum_{s=1}^{t} \mathbb{1}\{a_s = a\}, & \hat{\mu}_t(a) &:= \tfrac{1}{n_t(a)} \textstyle\sum_{s=1}^{t} \mathbb{1}\{a_s = a\}\mu_s(a), \\ \rho_t(a) &:= \sqrt{\tfrac{2\log T}{n_t(a)}}, & & \\ \ell_t(a) &:= \hat{\mu}_t(a) - \rho_t(a), & u_t(a) &:= \hat{\mu}_t(a) + \rho_t(a). \end{aligned} \tag{3}$$

Corollary A.2. *Fix any $T \in \mathbb{Z}_{>0}$ such that $T \geq |\mathcal{A}|$. Suppose that $(\mu_t(a))_{t=1}^{T}$ is iid-drawn from some $P(a)$ with mean $\mu(a)$, and $\mu_t(a) \in [0,1]$ for any $t \in T$ and $a \in \mathcal{A}$ almost surely. Then, for any strategy σ_A^T, we have*

$$\Pr_{P,\sigma_A^T}\left(\forall a \in \mathcal{A}\ \forall t \leq T, \quad |\hat{\mu}_t(a) - \mu(a)| \leq \rho_t(a)\right) \geq 1 - \frac{2}{T^2}.$$

Proof.

$$\begin{aligned} \Pr_{P,\sigma_A^T}\left(\forall a \in \mathcal{A}\ \forall t \leq T, \quad |\hat{\mu}_t(a) - \mu(a)| \leq \rho_t(a)\right) &= 1 - \Pr_{P,\sigma_A^T}\left(\exists a \in \mathcal{A}\ \exists t \leq T, \quad |\hat{\mu}_t(a) - \mu(a)| > \rho_t(a)\right) \\ &= 1 - \Pr_{P,\sigma_A^T}\left(\bigcup_{a\in\mathcal{A}}\bigcup_{t\leq T}\{|\hat{\mu}_t(a) - \mu(a)| > \rho_t(a)\}\right) \\ &\geq 1 - \sum_{a\in\mathcal{A}}\sum_{t\leq T}\Pr\left(|\hat{\mu}_t(a) - \mu(a)| > \rho_t(a)\right) \\ &\geq 1 - |\mathcal{A}|T\frac{2}{T^4} \\ &\geq 1 - \frac{2}{T^2} \quad (\because\ |\mathcal{A}| \leq T) \end{aligned}$$

□

A.2 Examples of No-ER and No-WER algorithms

We first introduce the following definition:

Definition A.1 (γ-no-ER, γ-no-WER)**.** We say an algorithm has *γ-no-ER* if, in equation (1) of Definition 6, the regret upper bound satisfies $R(T) = C\left(\log T\right)^{\delta} T^{\gamma}$ for some constants $C > 0$, $\delta \in [0,1)$ and $\gamma \in (0,1)$ that are independent of T. Note that if $\gamma < 1$, regardless of C and γ, we have $R(T) = o(T)$.

The class of *γ-no-WER* algorithms is defined in the same manner for weak-external-regret.

Note that an online learning algorithm achieves OLS (Definition 2) if and only if it has γ-no-WER with

$\gamma < 1$. It is known that EXP3 (Algorithm 1) with $\eta := \sqrt{\frac{\log|\mathcal{A}|}{T|\mathcal{A}|}}$ is $\frac{1}{2}$-no-ER, and SE (Algorithm 2) and UCB (Algorithm 3) are $\frac{1}{2}$-no-WER (Slivkins, 2019; Lattimore and Szepesvári, 2020).

Algorithm 1: EXP3

Input: Horizon T, learner's type θ, action set $\mathcal{A}$, learning rate $\eta_t > 0$

Rescale u so that its range is $[0,1]$;
Initialize weights $w_1(a) := 1/|\mathcal{A}|$ for each $a \in \mathcal{A}$
for $t \in \{1,\dots,T\}$ **do**
  Let $q_t(a) = \dfrac{w_t(a)}{\sum_a w_t(a)}$ for $a \in \mathcal{A}$;
  Draw action a_t from the multinomial distribution $(q_t(a))_{a\in\mathcal{A}}$;
  Observe $(x_t(a_t), p_t(a_t))$. Let $u_t := u(a_t,(x_t,p_t),\theta)$;
  for $a \in \mathcal{A}$ **do**
    if $a = a_t$ **then**
      $w_{t+1}(a) = w_t(a)\exp\left(\eta_t \cdot \left(1 - \frac{1}{q_t(a)}(1-u_t)\right)\right)$
    if $a \neq a_t$ **then**
      $w_{t+1}(a) = w_t(a)$

Algorithm 2: Successive Elimination

Input: Horizon T, learner's type θ, action set $\mathcal{A}$

Rescale u so that its range is $[0,1]$;
Initialize the active action set $\mathcal{A}_{\text{active}} := \mathcal{A}$ and time counter $t := 1$;
while $t \le T$ **do**
  for $a \in \mathcal{A}_{\text{active}}$ **do**
    Choose arm $a_t = a$;
    Observe $u_t := u(a_t,(x_t(a_t),p_t(a_t)),\theta)$; Compute $n_t(a)$, $\hat{\mu}_t(a)$, $\text{LCB}_t(a)$, and $\text{UCB}_t(a)$;
    $t \leftarrow t+1$;
  for $a \in \mathcal{A}_{\text{active}}$ **do**
    if $\exists a' \in \mathcal{A}_{\text{active}}$, $\text{LCB}_t(a') > \text{UCB}_t(a)$ **then**
      $\mathcal{A}_{\text{active}} \leftarrow \mathcal{A}_{\text{active}} \setminus \{a\}$;

B Omitted proofs

B.1 Proof of Proposition 1

Proof of Proposition 1. Fix a full-support prior $\pi \in \Delta(\Theta)$ and an algorithm $\mathcal{L}$. For each horizon T, let $\sigma_P^T \in \text{BR}_T(\pi,\mathcal{L})$ be a Bayesian-adaptive best response.

Fix $\theta \in \Theta$ and T. Given (σ_A^T, σ_P^T), we have

$$\overline{u}(\sigma_A^T,\sigma_P^T;\theta) \le \theta - \overline{v}(\sigma_A^T,\sigma_P^T). \tag{4}$$

Algorithm 3: UCB

Input: Horizon T, learner's type θ, action set $\mathcal{A}$

Rescale u so that its range is $[0,1]$;
Initialize $t := 1$;
// Cold start
for $t \in \{1, 2, \ldots, |\mathcal{A}|\}$ **do**
    Choose arm a_t;
    Observe $u_t := u(a_t, (x_t, p_t), \theta)$; Compute $n_t(a)$, $\hat{\mu}_t(a)$, $\mathrm{LCB}_t(a)$, and $\mathrm{UCB}_t(a)$;
    $t \leftarrow t + 1$;
// Main loop
while $t \leq T$ **do**
    Choose arm $a_t = \arg\max_{a \in \mathcal{A}} \mathrm{UCB}_{t-1}(a)$;
    Observe $u(a_t, (x_t, p_t), \theta)$; Compute $n_t(a)$, $\hat{\mu}_t(a)$, $\mathrm{LCB}_t(a)$, and $\mathrm{UCB}_t(a)$;
    $t \leftarrow t + 1$;

Because $\mathcal{L}$ fails weak-XR, it implies that for every $\varepsilon > 0$ there exist a sequence of principal strategies $(\widetilde{\sigma}_P^T)_T$ and $\overline{T} \in \mathbb{Z}_{>0}$ such that for all $T \geq \overline{T}$ and all $\theta \in \Theta$,

$$\overline{v}\Big(\mathcal{L}(\theta, T), \widetilde{\sigma}_P^T\Big) > \theta - \varepsilon. \tag{5}$$

Taking expectation over $\theta \sim \pi$ in (5) gives, for all $T \geq \overline{T}$,

$$\mathbb{E}_{\theta \sim \pi}\left[\overline{v}\Big(\mathcal{L}(\theta, T), \widetilde{\sigma}_P^T\Big)\right] > \mathbb{E}_{\theta \sim \pi}[\theta] - \varepsilon. \tag{6}$$

Since $\sigma_P^T \in \mathrm{BR}_T(\pi, \mathcal{L})$ maximizes the π-expected time-average revenue at horizon T, (6) implies that for all $T \geq \overline{T}$,

$$\mathbb{E}_{\theta \sim \pi}\left[\overline{v}\Big(\mathcal{L}(\theta, T), \sigma_P^T\Big)\right] \geq \mathbb{E}_{\theta \sim \pi}[\theta] - \varepsilon. \tag{7}$$

Because $\varepsilon > 0$ is arbitrary, (7) yields

$$\liminf_{T \to \infty} \mathbb{E}_{\theta \sim \pi}\left[\overline{v}\Big(\mathcal{L}(\theta, T), \sigma_P^T\Big)\right] \geq \mathbb{E}_{\theta \sim \pi}[\theta]. \tag{8}$$

By asymptotic ex-ante IR, there exists $\varepsilon_T \searrow 0$ such that for every $\theta \in \Theta$, every T, and every principal's strategy (in particular, σ_P^T),

$$\overline{u}\Big(\mathcal{L}(\theta, T), \sigma_P^T; \theta\Big) \geq -\varepsilon_T. \tag{9}$$

By (4) and (9), for each θ, we have

$$\overline{v}\Big(\mathcal{L}(\theta, T), \sigma_P^T\Big) \leq \theta - \overline{u}\Big(\mathcal{L}(\theta, T), \sigma_P^T; \theta\Big) \leq \theta + \varepsilon_T.$$

Taking expectation over $\theta \sim \pi$ yields

$$\mathbb{E}_{\theta\sim\pi}\left[\overline{v}\left(\mathcal{L}(\theta,T),\sigma_P^T\right)\right] \le \mathbb{E}_{\theta\sim\pi}[\theta] + \varepsilon_T. \tag{10}$$

Combining (8) and (10) gives

$$\lim_{T\to\infty} \mathbb{E}_{\theta\sim\pi}\left[\overline{v}\left(\mathcal{L}(\theta,T),\sigma_P^T\right)\right] = \mathbb{E}_{\theta\sim\pi}[\theta]. \tag{11}$$

For each T and θ, define the revenue gap

$$g_T(\theta) := \theta - \overline{v}\left(\mathcal{L}(\theta,T),\sigma_P^T\right).$$

From the previous step we have $g_T(\theta) \ge -\varepsilon_T$ for all θ, and (11) implies

$$\lim_{T\to\infty} \mathbb{E}_{\theta\sim\pi}[g_T(\theta)] = 0. \tag{12}$$

We claim that for each fixed $\theta \in \Theta$,

$$\lim_{T\to\infty} g_T(\theta) = 0, \quad \text{equivalently} \quad \lim_{T\to\infty} \overline{v}\left(\mathcal{L}(\theta,T),\sigma_P^T\right) = \theta. \tag{13}$$

Suppose not. Then there exist $\theta^* \in \Theta$, $\delta > 0$, and a subsequence $(T_k)_k$ such that $g_{T_k}(\theta^*) \ge \delta$ for all k. Since π has full support, $\pi(\theta^*) > 0$. Therefore, for each k,

$$\mathbb{E}_{\theta\sim\pi}[g_{T_k}(\theta)] \ge \pi(\theta^*)\, g_{T_k}(\theta^*) + \sum_{\theta\neq\theta^*} \pi(\theta)\,(-\varepsilon_{T_k}) \ge \pi(\theta^*)\delta - \varepsilon_{T_k}.$$

Letting $k \to \infty$ and using $\varepsilon_{T_k} \to 0$ yields $\liminf_{k\to\infty} \mathbb{E}_{\theta\sim\pi}[g_{T_k}(\theta)] \ge \pi(\theta^*)\delta > 0$, contradicting (12). This proves (13).

Fix $\theta \in \Theta$. By asymptotic ex-ante IR we already have the lower bound

$$\overline{u}\left(\mathcal{L}(\theta,T),\sigma_P^T;\theta\right) \ge -\varepsilon_T.$$

For the upper bound, apply (4) with $\sigma_A^T = \mathcal{L}(\theta,T)$:

$$\overline{u}\left(\mathcal{L}(\theta,T),\sigma_P^T;\theta\right) \le \theta - \overline{v}\left(\mathcal{L}(\theta,T),\sigma_P^T\right) = g_T(\theta).$$

By (13), $g_T(\theta) \to 0$, and by assumption $\varepsilon_T \to 0$. Hence

$$-\varepsilon_T \le \overline{u}\left(\mathcal{L}(\theta,T),\sigma_P^T;\theta\right) \le g_T(\theta) \quad\Longrightarrow\quad \lim_{T\to\infty} \overline{u}\left(\mathcal{L}(\theta,T),\sigma_P^T;\theta\right) = 0,$$

as desired. □

B.2 Proof of Proposition 2

Suppose toward contradiction that $\mathcal{L}$ fails weak-XR.

Fix any (θ, T, σ_P^T). For each t and realized history,

$$v(a_t, (x_t, p_t)) \;=\; p_t(a_t) \;=\; \theta x_t(a_t) - u(a_t, (x_t, p_t), \theta) \;\leq\; \theta - u(a_t, (x_t, p_t), \theta).$$

Taking time averages and expectations yields

$$\overline{v}\Big(\mathcal{L}(\theta, T), \sigma_P^T\Big) \;\leq\; \theta - \overline{u}\Big(\mathcal{L}(\theta, T), \sigma_P^T; \theta\Big). \tag{14}$$

By asymptotic ex-ante IR, there exists $\varepsilon_T \searrow 0$ such that $\overline{u}(\mathcal{L}(\theta, T), \sigma_P^T; \theta) \geq -\varepsilon_T$ for all (θ, T, σ_P^T). Substituting into (14) gives the uniform bound

$$\overline{v}\Big(\mathcal{L}(\theta, T), \sigma_P^T\Big) \leq \theta + \varepsilon_T \qquad \forall \theta, T, \sigma_P^T. \tag{15}$$

Because $|\Theta \setminus \{0\}| \geq 2$, there exist two distinct positive types $0 < \theta_L < \theta_H$ in Θ. Fix any full-support prior $\pi \in \Delta(\Theta)$ that assigns sufficiently large probability to θ_L and positive probability to θ_H (and positive probability to every other type). Then the static benchmark satisfies

$$u^*(\theta_H, \pi) =: \delta > 0.$$

Let $\sigma_P^T \in \mathrm{BR}_T(\pi, \mathcal{L})$. By no-DRE,

$$\liminf_{T\to\infty} \overline{u}\Big(\mathcal{L}(\theta_H, T), \sigma_P^T; \theta_H\Big) \geq u^*(\theta_H, \pi) = \delta.$$

Hence there exists T_1 such that for all $T \geq T_1$,

$$\overline{u}\Big(\mathcal{L}(\theta_H, T), \sigma_P^T; \theta_H\Big) \geq \frac{\delta}{2}. \tag{16}$$

Combining (14) with (16) yields, for all $T \geq T_1$,

$$\overline{v}\Big(\mathcal{L}(\theta_H, T), \sigma_P^T\Big) \leq \theta_H - \frac{\delta}{2}.$$

For all other types $\theta \in \Theta \setminus \{\theta_H\}$, apply the uniform bound (15). Taking expectations over $\theta \sim \pi$, for all

$T \geq T_1$,

$$\begin{aligned}\mathbb{E}_{\theta\sim\pi}\left[\overline{v}\left(\mathcal{L}(\theta,T),\sigma_P^T\right)\right] &\leq \sum_{\theta\neq\theta_H}\pi(\theta)\,(\theta+\varepsilon_T) \;+\; \pi(\theta_H)\,\left(\theta_H-\frac{\delta}{2}\right)\\ &\leq \mathbb{E}_{\theta\sim\pi}[\theta] \;+\; \varepsilon_T \;-\; \pi(\theta_H)\,\frac{\delta}{2}. \end{aligned} \tag{17}$$

Let $\eta := \pi(\theta_H)\delta/4 > 0$. Since $\varepsilon_T \searrow 0$, choose T_2 such that $\varepsilon_T < \eta$ for all $T \geq T_2$. Because $\mathcal{L}$ fails weak-XR, apply Definition 3 with $\varepsilon = \eta$ and $T_0 = \max\{T_1, T_2\}$. Then there exist $T \geq T_0$ and $\widetilde{\sigma}_P^T$ such that for all $\theta \in \Theta$,

$$\overline{v}\left(\mathcal{L}(\theta,T),\widetilde{\sigma}_P^T\right) \geq \theta - \eta.$$

Taking expectation over $\theta \sim \pi$ gives

$$\mathbb{E}_{\theta\sim\pi}\left[\overline{v}\left(\mathcal{L}(\theta,T),\widetilde{\sigma}_P^T\right)\right] \geq \mathbb{E}_{\theta\sim\pi}[\theta] - \eta.$$

Since $\sigma_P^T \in \mathrm{BR}_T(\pi,\mathcal{L})$ maximizes expected revenue under π,

$$\mathbb{E}_{\theta\sim\pi}\left[\overline{v}\left(\mathcal{L}(\theta,T),\sigma_P^T\right)\right] \;\geq\; \mathbb{E}_{\theta\sim\pi}[\theta] - \eta.$$

Combining this with (17) and $\varepsilon_T \leq \eta$ (for $T \geq T_2$) yields

$$\mathbb{E}_{\theta\sim\pi}[\theta] - \eta \leq \mathbb{E}_{\theta\sim\pi}\left[\overline{v}\left(\mathcal{L}(\theta,T),\sigma_P^T\right)\right] < \mathbb{E}_{\theta\sim\pi}[\theta] + \eta - \pi(\theta_H)\frac{\delta}{2} = \mathbb{E}_{\theta\sim\pi}[\theta] - \eta,$$

a contradiction. □

B.3 Proof of Proposition 3

Assume that an online learning algorithm $\mathcal{L}$ has no-ER with regret upper bound R. Fix any a, T, θ, and $\mu \in \Delta(\mathcal{M})$. Let $\sigma_A^T := \mathcal{L}(\theta,T)$. We have

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mu}\left[\sum_{t=1}^T u(a,(x_t,p_t),\theta)\right] - \mathop{\mathbb{E}}_{\sigma_A^T,\mu}\left[\sum_{t=1}^T u(a_t,(x_t,p_t),\theta)\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=1}^T u(a,(x_t,p_t),\theta) - \mathop{\mathbb{E}}_{\sigma_A^T}\left[\sum_{t=1}^T u(a_t,(x_t,p_t),\theta)\right] \mid (x_t,p_t)_{t=1}^T\right]\right]. \end{aligned} \tag{18}$$

Since $\mathcal{L}$ has no-ER, for any $(x_t,p_t)_{t=1}^T$, we have

$$\sum_{t=1}^T u(a,(x_t,p_t),\theta) - \mathop{\mathbb{E}}_{\sigma_A^T}\left[\sum_{t=1}^T u(a_t,(x_t,p_t),\theta)\right] \leq R(T).$$

Therefore, the RHS of (18) is bounded from above by $R(T)$, with $R(T) = o(T)$ by assumption. □

B.4 Proof of Theorem 1

Let $\Theta := \{\theta_1, \dots, \theta_K\}$, where $0 \leq \theta_1 < \theta_2 < \cdots < \theta_K < 1/2$. In this section, we show the following theorem, which implies Theorem 1.

Theorem A.1. *Suppose that the agent uses a no-ER algorithm and agent's type is $\theta_k \in \Theta$. Then, for any $\varepsilon > 0$ and any sufficiently large T, there exists a principal's strategy such that the principal obtains the average ex-ante expected payoff at least $(1-\varepsilon)\theta_k + o(1)$.*

■Notation For $A \subseteq \mathcal{A}$ and $T' \leq T$, let $n_{T'}(A) := \sum_{t=1}^{T'} \mathbb{1}\{a_t \in A\}$. Define $n_{T'}(a) := n_{T'}(\{a\})$ and $n_{T'}(\neg a) := n_{T'}(\mathcal{A} \setminus \{a\})$. Let $\delta_k := (\theta_k - \theta_{k-1})/2 > 0$ for $k \geq 2$. For ease of exposition, we omit the ceiling operator $\lceil \cdot \rceil$ throughout the subsequent analysis.

B.4.1 Proof of Theorem A.1

Suppose that $\mathcal{L}$ is a no-ER algorithm.

Lemma A.2. *Fix any $\theta \in \Theta$. Suppose that the agent's private type is θ and he uses a no-ER algorithm, and T is sufficiently large.* [*15]

For any $\varepsilon' > 0$, there exists principal's strategy such that the principal can correctly learn whether $\theta \geq \theta_K$ or not with probability at least $1 - \varepsilon'$ at the end of period T_K, where $T_K := \frac{2R(T)}{\varepsilon' \delta_K} = o(T)$.

Proof. Choose $\bar{a} \in \mathcal{A} \setminus \{a_0\}$. Fix any $T_K \leq T$. First, consider a principal's strategy such that

$$x_t(a) := \begin{cases} 0 & (a \neq \bar{a}, \text{ or } t \geq T_K + 1) \\ 1 & (a = \bar{a}, \text{ and } t \leq T_K) \end{cases}, \quad p_t(a) := \begin{cases} 0 & (a \neq \bar{a}, \text{ or } t \geq T_K + 1) \\ \theta_K - \delta_K & (a = \bar{a}, \text{ and } t \leq T_K) \end{cases}.$$

Throughout the proof, we fix this $(x_t, p_t)_{t=1}^T$. If $n_{T_K}(\bar{a}) \geq T_K/2$, then the principal concludes $\theta = \theta_K$; otherwise, she concludes $\theta \leq \theta_{K-1}$.

■Case (i): $\theta = \theta_K$. By the no-ER condition, we have

$$\sum_{t=1}^{T_K} u\left(\bar{a}, (x_t, p_t), \theta\right) - \mathbb{E}_{\sigma_A^T}\left[\sum_{t=1}^{T_K} u\left(a_t, (x_t, p_t), \theta\right)\right] \leq R(T) = o(T).$$

[*15] T should be large enough so that $T_K \leq T$ holds below, which is necessary for the proposed strategy well-defined.

As for the LHS, we have

$$\begin{aligned}
&\sum_{t=1}^{T} u\left(\overline{a}, (x_t, p_t), \theta\right) - \mathbb{E}_{\sigma_A^T}\left[\sum_{t=1}^{T} u\left(a_t, (x_t, p_t), \theta\right)\right] \\
&= \sum_{t=1}^{T_K} u\left(\overline{a}, (x_t, p_t), \theta\right) - \mathbb{E}_{\sigma_A^T}\left[\sum_{t=1}^{T_K} u\left(a_t, (x_t, p_t), \theta\right)\right] \\
&= \sum_{t=1}^{T_K} \left(\theta - (\theta_K - \delta_K)\right) - \mathbb{E}_{\sigma_A^T}\left[\sum_{t=1}^{T_K} \mathbb{1}\{a_t = \overline{a}\}(\theta - (\theta_K - \delta_K))\right] \\
&= \mathbb{E}_{\sigma_A^T}\left[\sum_{t=1}^{T_K} \mathbb{1}\{a_t \neq \overline{a}\}(\theta - (\theta_K - \delta_K))\right] \\
&= \sum_{t=1}^{T_K} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{a_t \neq \overline{a}\}(\theta - (\theta_K - \delta_K)) \mid a_{1:t-1}\right]\right] \\
&= \delta_K \mathbb{E}\left[\mathbb{E}\left[n_{T_K}(\neg\overline{a}) \mid a_{1:t-1}\right]\right] \quad (\because\ \theta = \theta_K) \\
&= \delta_K \mathop{\mathbb{E}}_{\sigma_A^T}[n_{T_K}(\neg\overline{a})].
\end{aligned}$$

Thus, we have

$$\mathop{\mathbb{E}}_{\sigma_A^T}[n_{T_K}(\neg\overline{a})] \leq \frac{R(T)}{\delta_K}.$$

By Markov's inequality, we have

$$\Pr\left(n_{T_K}(\neg\overline{a}) < \frac{T_K}{2}\right) \geq 1 - \frac{2}{T_K}\frac{R(T)}{\delta_K}.$$

Let $T_K \coloneqq \frac{2R(T)}{\varepsilon'\delta_K} = o(T)$. Then, we have

$$\Pr\left(n_{T_K}(\neg\overline{a}) < \frac{T_K}{2}\right) \geq 1 - \varepsilon'.$$

■Case (ii): $\theta \leq \theta_{K-1}$ By the no-ER condition, considering action a_0, we have

$$-\mathbb{E}_{\sigma_A^T}\left[\sum_{t=1}^{T_K} u\left(a_t, (x_t, p_t), \theta\right)\right] \leq R(T) = o(T).$$

By the similar argument to Case (i), we have

$$\delta_K \mathbb{E}_{\sigma_A^T}\left[n_{T_K}(\overline{a})\right] \leq R(T),$$

and thus

$$\Pr\left(n_{T_K}(\overline{a}) < \frac{T_K}{2}\right) \geq 1 - \varepsilon',$$

with $T_K := \frac{2R(T)}{\varepsilon'\delta_K} = o(T)$. □

Lemma A.3. *Fix any $\varepsilon' > 0$. Suppose that $\theta \leq \theta_{K-1}$ and T is sufficiently large. There exists principal's strategy such that at the end of period $T_K + T_{K-1}$, with probability at least $(1-\varepsilon')^2$, the principal can correctly learn whether $\theta = \theta_{K-1}$ or $\theta \leq \theta_{K-2}$, where*

$$T_K := \frac{2}{\varepsilon'\delta_K} R(T) = o(T), \quad T_{K-1} := \frac{2}{\varepsilon'(1-\varepsilon')\delta_{K-1}} \left(R(T) + \delta_K T_K\right) = o(T).$$

Proof. Suppose that $\theta \leq \theta_{K-1}$. Define the *K-clean event* as

$$\mathrm{CE}_K := \{\text{Principal correctly learn } \theta \leq \theta_{K-1} \text{ at the end of period } T_K\},$$

which happens with probability at least $1-\varepsilon'$ by Lemma A.2. Let $I_K := \{1, \ldots, T_K\}$, $I_{K-1} := \{T_K + 1, \ldots, T_K + T_{K-1}\}$, and $n_I(A) := \sum_{t \in I} \mathbb{1}\{a_t \in A\}$ for any $I \subseteq [T]$.

Consider the following strategy: for $t \in \{1, \ldots, T_K + T_{K-1}\}$,

$$x_t(a) := \begin{cases} 0 & (a \neq \bar{a}, \text{ or } t \geq T_K + T_{K-1} + 1) \\ 1 & o.w. \end{cases}, \quad p_t(a) := \begin{cases} 0 & (a \neq \bar{a}, \text{ or } t \geq T_K + T_{K-1} + 1) \\ \theta_K - \delta_K & (a = \bar{a}, t \leq T_K) \\ \theta_{K-1} - \delta_{K-1} & (a = \bar{a}, T_K + 1 \leq t \leq T_K + T_{K-1}) \end{cases}.$$

■Case (i): $\theta = \theta_{K-1}$ By a similar argument to the one in the proof of Lemma A.2, the no-ER condition for action $\bar{a}$ implies

$$-\delta_K \mathop{\mathbb{E}}_{\sigma_A^T}\left[n_{I_K}(\neg\bar{a})\right] + \delta_{K-1} \mathop{\mathbb{E}}_{\sigma_A^T}\left[n_{I_{K-1}}(\neg\bar{a})\right] \leq R(T),$$

and thus

$$\begin{aligned} \mathop{\mathbb{E}}_{\sigma_A^T}\left[n_{I_{K-1}}(\neg\bar{a})\right] &\leq \frac{1}{\delta_{K-1}} \left(R(T) + \delta_K \mathop{\mathbb{E}}_{\sigma_A^T}\left[n_{I_K}(\neg\bar{a})\right]\right) \\ &\leq \frac{1}{\delta_{K-1}} \left(R(T) + \delta_K T_K\right). \end{aligned}$$

Regarding the LHS, we have

$$\begin{aligned} \mathop{\mathbb{E}}_{\sigma_A^T}\left[n_{I_{K-1}}(\neg\bar{a})\right] &= \mathop{\mathbb{E}}_{\sigma_A^T}\left[n_{I_{K-1}}(\neg\bar{a}) \mid \mathrm{CE}_K\right] \Pr(\mathrm{CE}_K) + \mathop{\mathbb{E}}_{\sigma_A^T}\left[n_{I_{K-1}}(\neg\bar{a}) \mid \neg\mathrm{CE}_K\right] \Pr(\neg\mathrm{CE}_K) \\ &\geq \mathop{\mathbb{E}}_{\sigma_A^T}\left[n_{I_{K-1}}(\neg\bar{a}) \mid \mathrm{CE}_K\right] (1-\varepsilon'). \end{aligned}$$

We then have

$$\mathop{\mathbb{E}}_{\sigma_A^T}\left[n_{I_{K-1}}(\neg\overline{a}) \mid \mathrm{CE}_K\right] \leq \frac{1}{1-\varepsilon'}\frac{1}{\delta_{K-1}}\left(R(T)+\delta_K T_K\right).$$

By Markov's inequality, we have

$$\Pr\left(n_{I_{K-1}}(\neg\overline{a}) \leq \frac{T_{K-1}}{2} \mid \mathrm{CE}_K\right) \geq 1-\frac{2}{T_{K-1}}\frac{1}{1-\varepsilon'}\frac{1}{\delta_{K-1}}\left(R(T)+\delta_K T_K\right) \geq 1-\varepsilon',$$

with $T_{K-1} \coloneqq \frac{2}{\varepsilon'(1-\varepsilon')\delta_{K-1}}\left(R(T)+\delta_K T_K\right)$.

■Case (ii): $\theta \leq \theta_{K-2}$ By the no-ER condition for action a_0, we have

$$(\theta_K-(\theta+\delta_K))\,\mathbb{E}\left[n_{I_K}(\overline{a})\right]+(\theta_{K-1}-(\theta+\delta_{K-1}))\,\mathbb{E}\left[n_{I_{K-1}}(\overline{a})\right] \leq R(T).$$

This implies

$$\mathbb{E}\left[n_{I_{K-1}}(\overline{a})\right] \leq \frac{1}{\delta_{K-1}}R(T) \leq \frac{1}{\delta_{K-1}}(R(T)+\delta_K T_K).$$

By the same argument as in Case (i), we have

$$\Pr\left(n_{I_{K-1}}(\overline{a}) \leq \frac{T_{K-1}}{2} \mid \mathrm{CE}_K\right) \geq 1-\varepsilon',$$

with $T_{K-1} \coloneqq \frac{2}{\varepsilon'(1-\varepsilon')\delta_{K-1}}\left(R(T)+\delta_K T_K\right)$.

□

Repeating this argument inductively yields the following result.

Corollary A.3. *Fix any $\varepsilon' > 0$. Suppose that T is sufficiently large. Let*

$$T_{K-\ell} \coloneqq \frac{2}{\varepsilon'}\frac{1}{(1-\varepsilon')^\ell}\frac{1}{\delta_{K-\ell}}\left(R(T)+\sum_{s=K-\ell+1}^{K} w_s^{(\ell)} T_s\right) \quad (\ell \in \{0,\dots,K-1\}),$$

where

$$w_{K-m}^{(\ell)} \coloneqq \begin{cases} \delta_{K-m}+2\delta_{K-m-1}+\cdots+2\delta_{K-\ell+1} & (m \in \{0,1,\dots,\ell-2\}) \\ \delta_{K-\ell+1} & (m=\ell-1) \end{cases}.$$

Suppose that agent's private type is $\theta_k \in \Theta$. There exists a principal's strategy such that, with probability at least $(1-\varepsilon')^{K-k+1}$, the principal correctly learns $\theta=\theta_k$ up to period $S_k \coloneqq \sum_{s=k}^{K} T_s = o(T)$.

Lemma A.4. *Fix any T, ε', and $(\Delta_k)_k \in \mathbb{R}_{>0}^K$. Suppose that the true mean is $\theta_k \in \Theta$. If the principal uses Algorithm 4, then with probability at least $(1-\varepsilon')^K$, her undiscounted payoff is $\theta_k - \Delta_k - o(1)$.*

Theorem A.2. *Suppose that the agent uses a no-ER algorithm, agent's type is $\theta_k \in \Theta$, and T is sufficiently large. For any $\varepsilon \in (0,1)$, Algorithm 4 with*

$$\varepsilon' := 1 - \left(1 - \frac{\varepsilon}{2}\right)^{1/K}$$

learns that $\theta = \theta_k$ at the end of period $T_K + \cdots + T_k = o(T)$ with probability at least $(1 - \varepsilon/2)$.[*16] *Moreover, with*

$$\Delta_k := \frac{\varepsilon}{2 - \varepsilon}\theta_k,$$

the principal can get the average payoff of $(1-\varepsilon)\theta_k + o(1)$.

Proof. Note that Δ_k is chosen so that

$$\left(1 - \frac{\varepsilon}{2}\right)(\theta_k - \Delta_k) \geq (1-\varepsilon)\theta_k.$$

Let $S_k := T_K + \cdots T_k$ and $J_k := \{S_k + 1, \ldots, T\}$. Given $\theta_k \in \Theta$, we define the clean event CE as the event on which the principal learns $\theta = \theta_k$ correctly at the end of period S_k. To claim the second part of the statement, it suffices to show that

$$\mathbb{E}\left[n_{J_k}(\overline{a}) \mid \mathrm{CE}\right] \geq T - o(T).$$

Note that there exists $\overline{\theta} \in [0, 1/2]$ such that, for any k, $\theta_k \leq \overline{\theta}$. By the same argument as before, the no-ER condition for action $\overline{a}$ implies

$$\mathbb{E}\left[\sum_{t \in J_k} \mathbb{1}\{a_t \neq \overline{a}\}(\theta - p_t(\overline{a}))\right] \leq R(T) + \overline{\theta} S_k = o(T).$$

Regarding the LHS, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t \in J_k} \mathbb{1}\{a_t \neq \overline{a}\}(\theta - p_t(\overline{a}))\right] \\
&= \mathbb{E}\left[\sum_{t \in J_k} \mathbb{1}\{a_t \neq \overline{a}\}(\theta - p_t(\overline{a})) \mid \mathrm{CE}\right] \Pr(\mathrm{CE}) + \mathbb{E}\left[\sum_{t \in J_k} \mathbb{1}\{a_t \neq \overline{a}\}(\theta - p_t(\overline{a})) \mid \neg\mathrm{CE}\right] \Pr(\neg\mathrm{CE}) \\
&\geq (1-\varepsilon')^K \Delta_k \mathbb{E}\left[n_{J_k}(\neg\overline{a}) \mid \mathrm{CE}\right],
\end{aligned}$$

where the inequality follows since the second term of the second line is positive by construction. We then have

$$\mathbb{E}\left[n_{J_k}(\neg\overline{a}) \mid \mathrm{CE}\right] \leq \frac{1}{(1-\varepsilon')^K \Delta_k}\left(R(T) + \overline{\theta} S_k\right) =: B_k = o(T).$$

[*16] ε' is defined to satisfy $1 - \varepsilon/2 = (1-\varepsilon')^K$.

Therefore, we have

$$\begin{aligned}\mathbb{E}\left[n_{J_k}(\overline{a}) \mid \mathrm{CE}\right] &= T - S_k - \mathbb{E}\left[n_{J_k}(\neg\overline{a}) \mid \mathrm{CE}\right] \\ &\geq T - (S_k + B_k) \\ &= T - o(T).\end{aligned}$$

□

Algorithm 4: Principal's strategy against a no-ER learner

Input: $T \in \mathbb{Z}_{>0}, \varepsilon' > 0, R(T), \overline{a} \in \mathcal{A} \setminus \{a_0\}, (\Delta_k)_k, \Theta = \{\theta_1, \dots, \theta_K\}$

Let

$$T_{K-\ell} \coloneqq \frac{2}{\varepsilon'} \frac{1}{(1-\varepsilon')^\ell} \frac{1}{\delta_{K-\ell}} \left(R(T) + \sum_{s=K-\ell+1}^{K} w_s^{(\ell)} T_s \right) \quad (\ell \in \{0, \dots, K-1\}),$$

where

$$w_{K-m}^{(\ell)} \coloneqq \begin{cases} \delta_{K-m} + 2\delta_{K-m-1} + \cdots + 2\delta_{K-\ell+1} & (m \in \{0, 1, \dots, \ell-2\}) \\ \delta_{K-\ell+1} & (m = \ell - 1) \end{cases}.$$

Initialize $t \leftarrow 0, k \leftarrow K$;

while $k \geq 2$ **do**

  `// Exploration: Phase` k `consists of` T_k `periods`

  for $s = 1, \dots, T_k$ **do**

    $t \leftarrow t + 1$;

    $x_t(a) \leftarrow \mathbb{1}\{a = \overline{a}\}$;

    $p_t(a) \leftarrow (\theta_k - \delta_k)\mathbb{1}\{a = \overline{a}\}$;

    observe a_t;

  if $n_{I_k}(\overline{a}) \coloneqq \sum_{t \in I_k} \mathbb{1}\{a_t = \overline{a}\} \geq T_k/2$ **then**

    break; `// Conclude` $\theta = \theta_k$

  else

    $k \leftarrow k - 1$;

while $t \leq T$ **do**

  `// Exploitation: price` $\theta_k - \Delta_k \approx \theta_k$ `is charged for` $\overline{a}$ `for the remaining periods`

  $t \leftarrow t + 1$;

  $x_t(a) \leftarrow \mathbb{1}\{a = \overline{a}\}$;

  $p_t(a) \leftarrow (\theta_k - \Delta_k)\mathbb{1}\{a = \overline{a}\}$;

B.5 Proof of Lemma 1

Proof. Fix $\theta \in \Theta$, horizon T, and an arbitrary principal's strategy $\sigma_P^T \in \Sigma_P^T$. Condition on any realized sequence of principal's actions $((x_t, p_t))_{t=1}^T$ induced by σ_P^T. In Definition 6, the benchmark maximizes over fixed actions $a \in \mathcal{A}$, and the opt-out action a_0 is available. Since $u(a_0, (x_t, p_t), \theta) = 0$ for all t, the benchmark payoff satisfies

$$\max_{a \in \mathcal{A}} \sum_{t=1}^{T} u(a, (x_t, p_t), \theta) \geq 0.$$

If $\mathcal{L}$ has no-ER with bound $R(T) = o(T)$, then by (2),

$$\mathbb{E}\left[\sum_{t=1}^{T} u(a_t, (x_t, p_t), \theta)\right] \geq -R(T),$$

where the expectation is over the learner's randomization under $\sigma_A^T = \mathcal{L}(\theta, T)$. Dividing by T yields

$$\overline{u}\Big(\mathcal{L}(\theta, T), \sigma_P^T; \theta\Big) \geq -\frac{R(T)}{T}.$$

Setting $\varepsilon_T := R(T)/T$ proves asymptotic ex-ante IR. □

B.6 Proof of Theorem 2

It is known that these three algorithms are no-WER.

For UE and SE, consider the following strategy of the adaptive principal: for some $a_1 \neq a_0$, for any period $t \leq T_1$ before the learner drops all actions except a_1 at the end of period T_1,

$$(x_t(a), p_t(a)) = \begin{cases} (1, 0) & (a = a_1) \\ (0, 0) & (a = a_0) \\ (0, 1/2) & (o.w.) \end{cases}.$$

Both UE and SE remove all actions but a_1 with sublinear T_1. For $t \geq T_1 + 1$, the principal sets $p_t(a_1) = 1/2$, so she achieves $1/2 + o(1)$ average payoff. This proves that UE and SE are unsafe.

For UCB, suppose that there are K possible learner types $\{\theta_1, \ldots, \theta_K\}$ with $0 \leq \theta_1 < \theta_2 < \cdots < \theta_K < 1/2$. By assumption, we have $|\mathcal{A}| \geq K + 1$. Let $\varepsilon := \min_{k \in [K]} |\theta_k - \theta_{k-1}| > 0$, where $\theta_0 := 0$.

For sufficiently large T, for the early periods, UCB chooses the least explored actions (i.e., chooses action $a \in \arg\min_{a \in \mathcal{A}} n_t(a)$ in period t) as the effect of the confidence intervals dominates. If the number of explorations is the same for all actions, then UCB chooses the one with the best past performance.

Let $M := |\mathcal{A}|$. Note that $M \geq K + 1$ by assumption. Consider the following principal's strategy: for

$t \leq M$,

$$(x_t(a_k), p_t(a_k)) = \begin{cases} (1, \theta_k - \varepsilon) & (k \in \{1, \dots, K\}) \\ (0, 0) & (k \in \{0, K+1, \dots, M\}), \end{cases}$$

and for $t \in \{M+1, \dots, 2M\}$, $x_t(a) = p_t(a) = 0$ for all $a \in \mathcal{A}$.

For $t \in \{1, \dots, M\}$, UCB chooses each of the M actions once. Then, for $t \in \{M+1, \dots, 2M\}$, UCB chooses actions in descending order of past performance observed during $t \leq M$. Define $a_0, a_{K+1}, \dots a_M$ as *null actions*. If the learner has type $\theta = 0$, then UCB chooses a null action at $t = M+1$. Suppose that the learner's type is $\theta_k > 0$. Then, UCB selects actions in the order $a_1, a_2, \dots, a_k$, followed by the null actions, and then $a_{k+1}, \dots, a_M$. Therefore, by observing the behavior during periods $t \in \{M+1, \dots, 2M\}$, the principal can perfectly learn the learner's type θ. For the remainder of the game, the adaptive principal sets $x_t(a_k) = 1, p_t(a_k) = \theta - \varepsilon$ with ε for all $k \geq 1$, obtaining the payoff $\theta - \varepsilon$ asymptotically if T is sufficiently long.

□

B.7 Proof of Theorem 3

Assume $T \geq T(\underline{\delta})$, where

$$\begin{aligned} \delta_k &:= \frac{\theta_k - \theta_{k-1}}{2} \quad (k \geq 2), \\ \underline{\delta} &:= \min_{k \geq 2} \delta_k, \\ T(\underline{\delta}) &:= \min \left\{ T \in \mathbb{N} \colon \frac{\log T}{T} < \left(\frac{\underline{\delta}}{\sqrt{18}} \right)^3 \right\}. \end{aligned}$$

The condition guarantees that the radius of Phase-1 confidence interval $\overline{\rho}_{T_1} < \underline{\delta}/3$.

Lemma A.5 (Preventive use of confidence intervals)**.** *Fix any horizon $T_0 \in \mathbb{Z}_{>0}$, margin $\Delta > 0$, and reference point $B \in \mathbb{R}$. For $s \in \mathbb{Z}_{>0}$, let $\overline{\rho}_s := \sqrt{(2 \log T)/s}$. Given a real sequence $(c_t)_{t \geq 1}$, let $\overline{c}_s := s^{-1} \sum_{t=1}^{s} c_t$.*

Suppose that the upper confidence interval never goes below B up to period T_0, i.e. $\overline{c}_s + \overline{\rho}_s \geq B$ for all $s \leq T_0$.

1. *If the empirical average of the sequence is strictly below than the reference point by the size of the margin at the end of period T_0, i.e., $\overline{c}_{T_0} \leq B - \Delta$, then we have*

$$T_0 \leq \frac{2 \log T}{\Delta^2}. \tag{19}$$

2. *Similarly, if $\overline{c}_s - \overline{\rho}_s \leq B$ for all $s \leq T_0$ and $B + \Delta \leq \overline{c}_{T_0}$, then we have (19).*

Algorithm 5: UE-Trigger (UE-T)

Input: Action set $\mathcal{A}$, time horizon T, agent's type θ, default action a_0, exploration length T_1

Phase 1: Exploration

for $t \in [|\mathcal{A}|T_1]$ **do**

  Choose $a_t := a_k$, where $k := t \pmod{|\mathcal{A}|}$;

  Observe allocation $x_t(a_t)$, and payment $p_t(a_t)$;

  $t \leftarrow t+1$;

end

Choose action $a^* \in \mathcal{A}$ with the highest empirical reward, i.e.,

$$a^* \in \arg\max_{a\in\mathcal{A}} \frac{1}{T_1} \sum_{t=1}^{|\mathcal{A}|T_1} \mathbb{1}\{a_t = a\} \left(\theta x_t(a_t) - p_t(a_t)\right).$$

Record the upper and lower confidence bounds for allocation and payment for a^*:

$$\mathrm{UCB}_x^1 := \frac{1}{T_1} \sum_{t=1}^{|\mathcal{A}|T_1} \mathbb{1}\{a_t = a^*\} x_t(a_t) + \overline{\rho}_{T_1}, \quad \mathrm{LCB}_x^1 := \frac{1}{T_1} \sum_{t=1}^{|\mathcal{A}|T_1} \mathbb{1}\{a_t = a^*\} x_t(a_t) - \overline{\rho}_{T_1},$$

$$\mathrm{UCB}_p^1 := \frac{1}{T_1} \sum_{t=1}^{|\mathcal{A}|T_1} \mathbb{1}\{a_t = a^*\} p_t(a_t) + \overline{\rho}_{T_1}, \quad \mathrm{LCB}_p^1 := \frac{1}{T_1} \sum_{t=1}^{|\mathcal{A}|T_1} \mathbb{1}\{a_t = a^*\} p_t(a_t) - \overline{\rho}_{T_1},$$

where

$$\overline{\rho}_s := \sqrt{\frac{2\log T}{s}} \quad (s \in \mathbb{Z}_{>0}).$$

Phase 2: Exploitation with Protection

Initialize $s := 1$

while *True* **do**

  Play action a^*

  Observe $x_t(a^*)$ and $p_t(a^*)$

  Compute Phase-2 confidence bounds:

$$\mathrm{UCB}_x^2(s) := \frac{1}{s} \sum_{t=|\mathcal{A}|T_1}^{|\mathcal{A}|T_1+s} x_t(a^*) + \overline{\rho}_s, \quad \mathrm{LCB}_x^2(s) := \frac{1}{s} \sum_{t=|\mathcal{A}|T_1}^{|\mathcal{A}|T_1+s} x_t(a^*) - \overline{\rho}_s,$$

$$\mathrm{UCB}_p^2(s) := \frac{1}{s} \sum_{t=|\mathcal{A}|T_1}^{|\mathcal{A}|T_1+s} p_t(a^*) + \overline{\rho}_s, \quad \mathrm{LCB}_p^2(s) := \frac{1}{s} \sum_{t=|\mathcal{A}|T_1}^{|\mathcal{A}|T_1+s} p_t(a^*) - \overline{\rho}_s,$$

  if $(\mathrm{LCB}_x^2(s), \mathrm{UCB}_x^2(s)) \cap (\mathrm{LCB}_x^1, \mathrm{UCB}_x^1) = \emptyset$ *or* $(\mathrm{LCB}_p^2(s), \mathrm{UCB}_p^2(s)) \cap (\mathrm{LCB}_p^1, \mathrm{UCB}_p^1) = \emptyset$ **then**

    break

  end

  $s \leftarrow s+1$

end

Phase 3: Opt-out

while *True* **do**

  Play the default action a_0.

end

Proof. Since $\bar{c}_s + \bar{\rho}_s \geq B$ with $s := T_0$, we have

$$\bar{c}_{T_0} \geq B - \sqrt{\frac{2 \log T}{T_0}}.$$

Since $\bar{c}_{T_0} \leq B - \Delta$, we have

$$B - \sqrt{\frac{2 \log T}{T_0}} \leq \bar{c}_{T_0} \leq B - \Delta,$$

and thus

$$\Delta \leq \sqrt{\frac{2 \log T}{T_0}},$$

which is equivalent to (19).

□

Lemma A.6. *UE-T is 2/3-no-WER.*

Proof. Suppose that the principal is stationary.

Given $T_1 = T^{2/3}(\log T)^{1/3}$, the length of Phase 1 is $O(T^{2/3}(\log T)^{1/3})$, and the contribution of the outcomes during Phase 1 to the regret is $O(T^{2/3}(\log T)^{1/3}) = \widetilde{O}(T^{2/3})$.[*17]

By Corollary A.2, the confidence intervals formed during Phase 1 include the true mean allocations and payments for all actions with probability at least $1 - \widetilde{O}(T^{-4/3})$. Let's call the event on which the Phase-1 confidence intervals include true means for all actions *the clean event*, and its complement *the Phase-1 rare event*. The Phase-1 rare event happens with probability at most $\widetilde{O}(T^{-4/3})$. Since the period payoff is bounded, the contribution of the outcomes on the Phase-1 rare event to the regret is at most $\widetilde{O}(T^{-1/3})$, and thus negligible. Below, we focus on the Phase-1 clean event.

For Phase 2, again by Corollary A.2, conditional on the Phase-1 clean event, with probability at least $1 - O(T^{-2})$, the new confidence interval contains the true mean allocation and payment until the game ends (*Phase-2 clean event*). If the Phase-2 clean event happens, the algorithm never enters Phase 3, and the behavior of UE-T during Phase 2 is exactly the same as UE, which is $\frac{2}{3}$-no-WER. Therefore, the regret of UE-T is bounded above by $\widetilde{O}(T^{2/3})$. □

Lemma A.7. *UE-T satisfies no-DRE.*

Proof. Suppose that the principal is adaptive.

Since the length of Phase 1 is sublinear, for sufficiently large T, the outcomes during Phase 1 are negligible. We can focus on the outcomes after Phase 2 starts when discussing the asymptotic payoff of the principal.

[*17] $\widetilde{O}$-notation ignores log-related terms in O-notation.

Let

$$\overline{x}_{T_1}(a) = \frac{1}{T_1}\sum_{t=1}^{|\mathcal{A}|T_1} \mathbb{1}\{a_t = a\}x_t(a), \quad \overline{p}_{T_1}(a) = \frac{1}{T_1}\sum_{t=1}^{|\mathcal{A}|T_1} \mathbb{1}\{a_t = a\}p_t(a).$$

To avoid obtaining asymptotically zero average payoff, the principal must stay in Phase 2 for $T - o(T)$ periods. Then, by Lemma A.5 and the construction of UE-T, the principal's revenue during Phase 2, denoted by $\overline{p}$, is in an interval $\overline{p}_{T_1} \pm (\overline{\rho}_{T_1} + o(1)) = \overline{p}_{T_1} + o(1)$. Since the length of Phase 1 is $o(T)$ and the contribution to the average payoff is $o(1)$, we have $\overline{v} = \overline{p} + o(1)$. Thus, we have

$$\overline{v} = \overline{p}_{T_1} + o(1).$$

Note that the principal can choose any $\overline{p}_{T_1} \in [0, 1/2]$ by adjusting $(p_t)_{t \le |\mathcal{A}|T_1}$, but she needs to choose it before acquiring any additional information about learner's type. Thus, the principal essentially posts one single menu $(\overline{x}, \overline{p})$ against the learner with his private type. Thus, the algorithm satisfies no-DRE. □

B.8 Proof of Theorem 4

Note that proofs below use notation and results from Appendix C.

Proposition A.1. *UE-SE-T is $1/2$-no-WER.*

Proof. By Corollary A.2, the clean event happens with probability $1 - O(T^{-2})$ even for UE-SE-T. The proof for the no-WER property of Successive Elimination (Proposition A.7) applies without modification. □

Proposition A.2. *UE-SE-T satisfies no-DRE.*

We prove this proposition by proving a series of lemmas.

Lemma A.8. *Let $T_1 := |A|m$ for a positive integer m, and τ be the (possibly random) first time UE-SE-T enters Phase 3. Suppose that Phase 1 empirical means are $(\overline{x}_a, \overline{p}_a)_a$. Let*

$$\overline{u}_a(\theta) := \theta\overline{x}_a - \overline{p}_a, \quad \overline{u}^*(\theta) := \max_a \overline{u}_a(\theta).$$

Fix any θ and $t \in [T_1 + 1, \tau)$ (i.e., the algorithm is in Phase 2). If action a_t is chosen by UE-SE-T in period t, then we have

$$\overline{u}_{a_t}(\theta) \ge \overline{u}^*(\theta) - 8\rho_m.$$

Proof. Let $T_1 := m|\mathcal{A}|$ be the length of Phase 1. Denote by τ the (random) first time the algorithm enters Phase 3. If the algorithm does not enter Phase 3, let $\tau := T + 1$.

Algorithm 6: UE-SE-Trigger (UE–SE–T)

Input: Action set $\mathcal{A}$ (with default action a_0), horizon T, agent's type θ

$m := \lceil \sqrt{T} \rceil, \quad \overline{\rho}_s := \sqrt{\frac{2 \log T}{s}}$

Phase 1: Uniform Exploration

for $t \in \{1, \ldots, |\mathcal{A}|m\}$ **do**

  Choose $a_t := a_k$, where $k \equiv t \pmod{|\mathcal{A}|}$;

  Observe allocation $x_t(a_t)$ and payment $p_t(a_t)$;

  $t \leftarrow t + 1$;

end

for *each* $a \in \mathcal{A}$ **do**

  Compute empirical means

$$\overline{x}_a := \frac{1}{m} \sum_{t=1}^{|\mathcal{A}|m} \mathbb{1}\{a_t = a\} x_t(a_t), \quad \overline{p}_a := \frac{1}{m} \sum_{t=1}^{|\mathcal{A}|m} \mathbb{1}\{a_t = a\} p_t(a_t)$$

  Record Phase-1 confidence intervals

$$I^{(1)}_{x,a} := [\overline{x}_a \pm \overline{\rho}_m], \quad I^{(1)}_{p,a} := [\overline{p}_a \pm \overline{\rho}_m]$$

end

Phase 2: Successive Elimination with Protection

Initialize active set $\mathcal{A}_{\text{act}} := \mathcal{A}$;

Initialize counters $n(a) := m$ for all $a \in \mathcal{A}$;

while $t \leq T$ **do**

  Select action $a_t \in \mathcal{A}_{\text{act}}$ according to Successive Elimination (using reward $u_t = \theta x_t(a_t) - p_t(a_t)$);

  Observe $x_t(a_t)$ and $p_t(a_t)$;

  Update $n(a_t) \leftarrow n(a_t) + 1$;

  Update empirical means `// We use all observations up to the current period, including the ones from Phase 1, when computing empirical means and confidence intervals.`

$$\hat{x}_t(a_t) := \frac{1}{n(a_t)} \sum_{s \leq t} \mathbb{1}\{a_s = a_t\} x_s(a_t), \quad \hat{p}_t(a_t) := \frac{1}{n(a_t)} \sum_{s \leq t} \mathbb{1}\{a_s = a_t\} p_s(a_t)$$

  Compute Phase-2 confidence intervals

$$I^{(2)}_{x,a_t}(t) := [\hat{x}_t(a_t) \pm \overline{\rho}_{n(a_t)}], \quad I^{(2)}_{p,a_t}(t) := [\hat{p}_t(a_t) \pm \overline{\rho}_{n(a_t)}]$$

  if $I^{(2)}_{x,a_t}(t) \cap I^{(1)}_{x,a_t} = \emptyset$ ***or*** $I^{(2)}_{p,a_t}(t) \cap I^{(1)}_{p,a_t} = \emptyset$ **then**

    break

  end

  Eliminate actions from $\mathcal{A}_{\text{act}}$ according to Successive Elimination;

  $t \leftarrow t + 1$;

end

Phase 3: Opt-out

while $t \leq T$ **do**

  Play default action a_0;

  $t \leftarrow t + 1$;

end

Fix any $t \in [T_1 + 1, \tau)$. By construction, we have

$$|\hat{x}_t(a) - \overline{x}_a| \le \rho_m + \rho_{n_t(a)} \le 2\rho_m,$$
$$|\hat{p}_t(a) - \overline{p}_a| \le \rho_m + \rho_{n_t(a)} \le 2\rho_m.$$

Let $\hat{u}_t(\theta, a) \coloneqq \theta \hat{x}_t(a) - \hat{p}_t(a)$. Then, we have

$$|\hat{u}_t(\theta, a) - \overline{u}_a(\theta)| \le \theta |\hat{x}_t(a) - \overline{x}_a| + |\hat{p}_t(a) - \overline{p}_a| \le 3\rho_m.$$

Denote by $\mathcal{A}_{\text{act}}(t)$ the set of active actions at the beginning of period t. Suppose that $b \in \mathcal{A}_{\text{act}}(t)$. This implies that for any $s \in [T_1 + 1, t-1]$ and $a \in \mathcal{A}_{\text{act}}(s)$, we have $\text{UCB}_s(b) \ge \text{LCB}_s(a)$. Thus,

$$\begin{aligned} 0 &\ge (\hat{u}_s(a, \theta) - \rho_{n_s(a)}) - (\hat{u}_s(b, \theta) + \rho_{n_s(b)}) \\ &\ge \overline{u}_a(\theta) - \overline{u}_b(\theta) - 8\rho_m. \end{aligned}$$

Since all actions are active at the beginning of period $T_1 + 1$, this implies that, if action a_t is chosen in period t, we have

$$\overline{u}_{a_t}(\theta) \ge \overline{u}^*(\theta) - 8\rho_m.$$

□

Lemma A.9. *Fix any agent's tie-breaking rule. For any $\delta > 0$ menu (x, p) with revenue $R(x, p)$, there exists a menu (x, p') with revenue $R(x, p')$ such that*

1. *$R(x, p') \ge R(x, p) - \delta$;*
2. *For each θ, the best response correspondence under (x, p') is a singleton;*

Proof. See Proposition A.5. □

Lemma A.10. *If the agent uses UE-SE-T, the principal's equilibrium revenue is $\text{Rev}^*(\pi) + o(1)$.*

Proof. First, we prove that $\text{Rev}^*(\pi) + o(1)$ forms an upper bound for any principal's strategy. Fix any Phase-1-proposed menu $(\overline{x}, \overline{p})$. The upper bound clearly holds if the length of Phase 2 is sublinear. Suppose that the length of Phase 2 is $\gamma T - o(1)$ for some $\gamma \in (0, 1]$. Let $\varepsilon_T \coloneqq 8\rho_m = o(1)$. By Lemma A.8, during Phase 2, the agent chooses ε_T-optimal actions given $(\overline{x}, \overline{p})$. Since $\overline{x}(a)$ should be a multiple of $1/m$, the principal's revenue is upper bounded by $\gamma \text{Rev}^{\varepsilon_T}_{(m)}(\pi) + o(1) \le \text{Rev}^{\varepsilon_T}_{(m)}(\pi) + o(1)$. Since $m \to \infty$ and $\varepsilon_T \to 0$ as $T \to \infty$, by Corollary A.5, we have $\text{Rev}^{\varepsilon_T}_{(m)}(\pi) + o(1) = \text{Rev}^*(\pi) + o(1)$.

Next, we prove that the principal's revenue under her optimal strategy is bounded below by $\text{Rev}^*(\pi) + o(1)$. It suffices to show that, for any $\delta > 0$, there exists an principal's strategy that achieves revenue $\text{Rev}^*(\pi) - \delta + o(1)$.

Fix any $\delta > 0$. Let (x^*, p^*) be the menu be the revenue optimal menu that achieves the Myerson-revenue

when the agent uses the principal-favorable tie-breaking rule. The menu exists due to the taxation principle. Note that we do not need to assume that the agent breaks ties in this way. Then, by Lemma A.9, there exists another menu $M' = (x^*, p')$ such that each type θ has a unique best action, and the principal's expected revenue under M' in the static setup is at least $\text{Rev}^*(\pi) - \delta$. Consider the principal's strategy such that x_t is iid-drawn from $\text{Ber}(x^*)$ and $p_t = p'$ in each period t. Since UE-SE-T is no-WER, for $T - o(T)$ periods, type-θ agent chooses the unique best action under M'. Therefore, the principal's expected revenue under this strategy is $\text{Rev}^*(\pi) - \delta + o(1)$. This proves the lower bound. □

Lemma A.11. *Suppose that the agent uses $\mathcal{L} :=$ UE-SE-T, his type is θ, and σ_T^P denotes the optimal strategy of the Bayesian adaptive principal with belief π given UE-SE-T. Then, we have*

$$\liminf_{T\to\infty} \overline{u}\left(\mathcal{L}(\theta, T), \sigma_P^T, \theta\right) \geq u^*(\theta, \pi).$$

Proof. By Lemma A.10, under the principal's optimal strategy, the expected length of Phase 2 is $T - o(T)$. Denote the (possibly random) length of Phase 2 by T_2, and the last period of Phase 1 by τ_1. Let $\widetilde{x}(\theta)$ and $\widetilde{p}(\theta)$ be a direct mechanism, which is defined as the ex ante expected empirical average of allocations and payments during Phase 2, i.e.,

$$\widetilde{x}(\theta) := \mathop{\mathbb{E}}_{(a_t)_t}\left[\frac{1}{T_2}\sum_{t=\tau_1+1}^{\tau_1+T_2} x_t(a_t)\right], \quad \widetilde{p}(\theta) := \mathop{\mathbb{E}}_{(a_t)_t}\left[\frac{1}{T_2}\sum_{t=\tau_1+1}^{\tau_1+T_2} p_t(a_t)\right],$$

where the distribution of $(a_t)_t$ is determined by $\mathcal{L}$, θ, and σ_P^T. The type-θ agent's payoff is $U_T(\theta) := \theta\widetilde{x}(\theta) - \widetilde{p}(\theta) + o(1)$ since $T_2 = T - o(T)$. By Lemma A.8, during Phase 2, the agent ε_T-best responds to $(\widetilde{x}, \widetilde{p})$, where $\varepsilon_T = o(1)$. By Lemma A.10, the principal's revenue tends to $\text{Rev}^*(\pi)$ as $T \to \infty$. Since $\varepsilon_T \to 0$, by Proposition A.6, we have

$$\liminf_{T\to\infty} U_T(\theta) \geq u^*(\theta, \pi).$$

□

C Results for Mechanism Design with Approximation

C.1 Setup

C.1.1 Baseline model

Consider the following canonical static principal-agent problem: Let $\mathcal{A}$ and $\Theta := \{0 \leq \theta_1 < \theta_2 < \cdots < \theta_n \leq 1/2\} \subseteq [0, 1/2]$ be the action and type sets of the agent. We assume both sets are finite and $|\mathcal{A}| \geq |\Theta|$. The game proceeds as follows.

1. The agent privately observes his type $\theta \in \Theta$.

2. The principal with prior belief $\pi \in \Delta(\Theta)$ proposes a menu $(\overline{x}(a), \overline{p}(a))_{a\in\mathcal{A}} \in [0,1]^{\mathcal{A}} \times [0,1/2]^{\mathcal{A}}$.
3. Observing the menu, the agent chooses the menu (equivalently, action) that maximizes his payoff $u(\theta, a, (\overline{x}, \overline{p})) := \theta\overline{x}(a) - \overline{p}(a)$ given his private information θ.

We assume that there is a special action $a_0 \in \mathcal{A}$ such that for any menu (x, p), $x(a_0) = p(a_0) = 0$. We assume a fixed deterministic tie-breaking rule of the agent so that for each θ and $(\overline{x}, \overline{p})$, the best-response of a type-θ agent is uniquely determined.[*18] Since $|\mathcal{A}| \geq |\Theta|$, by the taxation principle, it is without loss of generality that the principal designs a direct revelation mechanism under IC and IR constraints.[*19] Recall that $n = |\Theta|$. Denote by $(x_i, p_i)_{i=1}^n$ a direct revelation mechanism, where $(x_i, p_i) \in [0,1] \times [0,1/2]$ is the allocation and payment for type θ_i under the IR and IC constraints. The set of direct mechanisms is denoted by $\mathcal{M} := ([0,1] \times [0,1/2])^n$. Let $\pi_i := \pi(\theta_i)$.

Let

$$\begin{aligned}
\mathrm{Rev}^*(\pi) &:= \max_{(x,p)\in\mathcal{M}} \left\{ \sum_{i=1}^n \pi_i p_i \colon (x,p) \text{ satisfies IC and IR} \right\}, \\
\mathcal{M}^*(\pi) &:= \arg\max_{(x,p)\in\mathcal{M}} \left\{ \sum_{i=1}^n \pi_i p_i \colon (x,p) \text{ satisfies IC and IR} \right\}, \\
u^*(\theta_i, \pi) &:= \min_{(x,p)\in\mathcal{M}^*(\pi)} \{\theta_i x_i - p_i\}.
\end{aligned}$$

Remark A.1. $\mathcal{M}^*(\pi)$ is nonempty since the principal's problem can be written as a linear programming problem and the feasible set is nonempty. Moreover, $\mathcal{M}^*(\pi)$ is compact since the feasible set of the revenue-maximization problem is compact and the objective function is continuous. Thus, the minimizer in the definition of u^* is well-defined. Since the type space is finite, there are several optimal mechanisms, i.e., $\mathcal{M}^*(\pi)$ is not a singleton. Therefore, without the minimization in the definition of u^*, the payoff for each type under the optimal mechanism is not uniquely pinned down. For a fixed allocation rule, the revenue-maximizing payment rule is uniquely pinned down.

C.1.2 Model with agents with ε-BR

We next consider a variant of the model in which the agent ε-best-responds to the proposed menu, i.e., the agent chooses a principal-favorable action from the following set:

$$A^\varepsilon(\overline{x}, \overline{p}) := \left\{ a \in \mathcal{A} \colon \forall a' \neq a,\ u(\theta, a, (\overline{x}, \overline{p})) \geq u(\theta, a', (\overline{x}, \overline{p})) - \varepsilon \right\}.$$

[*18] For example, we can assume the principal-favorable tie-breaking.

[*19] "Without loss of generality"means: fixing a (deterministic) selection rule for the agent's best responses, any outcome induced by some menu can be replicated by a truthful direct mechanism (and conversely). Hence, the set of implementable allocation–payment rules and the equilibrium payoffs of the players are the same under menus and direct mechanisms.

Given a tie-breaking rule, we can still apply the taxation principle; we focus on the class of direct mechanisms. A direct mechanism satisfies ε-IC and ε-IR conditions:

$$\begin{aligned}&\theta_i x_i - p_i \geq \theta_i x_j - p_j - \varepsilon \quad (\forall i, j),\\ &\theta_i x_i - p_i \geq -\varepsilon \quad (\forall i).\end{aligned}$$

Let

$$\mathrm{Rev}^{\varepsilon}(\pi) \coloneqq \max_{(x,p)\in\mathcal{M}} \left\{ \sum_{i=1}^{n} \pi_i p_i \colon (x,p) \text{ satisfies } \varepsilon\text{-IC and } \varepsilon\text{-IR} \right\}.$$

Note that when we set $\varepsilon \coloneqq 0$, these objects coincide with the ones for the baseline model.

C.1.3 Model with integral constraints on allocations

Lastly, we consider another version of the baseline model in which the principal's allocation $\overline{x}(a)$ should be a multiple of $1/m$ given $m \in \mathbb{Z}_{>0}$. Again, by the taxation principle, we focus on the direct mechanisms. Let $\mathcal{M}_{(m)} \coloneqq \{(x,p) \in \mathcal{M} \colon \forall i,\ x_i = k/m, k \in \mathbb{N}\}$, and

$$\mathrm{Rev}^{\varepsilon}_{(m)}(\pi) \coloneqq \max_{(x,p)\in\mathcal{M}_{(m)}} \left\{ \sum_{i=1}^{n} \pi_i p_i \colon (x,p) \text{ satisfies } \varepsilon\text{-IC and } \varepsilon\text{-IR} \right\}.$$

C.2 Results

C.2.1 Approximation of Optimal Revenues

Proposition A.3. *Fix any $m \in \mathbb{Z}_{>0}$. We have*[*20]

$$0 \leq \mathrm{Rev}^{0}(\pi) - \mathrm{Rev}^{0}_{(m)}(\pi) \leq \frac{1}{m}.$$

Proof. By the standard "integration by part" argument, we can rewrite $\mathrm{Rev}^{0}(\pi)$ as follows:

$$\mathrm{Rev}^{0}(\pi) = \left\{ \sum_{k=1}^{n} w_k x_k \colon 0 \leq x_1 \leq \cdots \leq x_n \leq 1 \right\},$$

[*20] More generally, we can obtain the upper bound $\sup \Theta / m$.

where

$$
\begin{aligned}
w_k &\coloneqq \pi_k \theta_k - \Delta_k^\theta Q_k \quad (k \le n-1) \\
w_n &\coloneqq \pi_n \theta_n \\
\Delta_k^\theta &\coloneqq \theta_{k+1} - \theta_k \\
Q_k &\coloneqq \sum_{i=k+1}^{n} \pi_i
\end{aligned}
$$

Denote by $R(x) \coloneqq \sum_{k=1}^{n} w_k x_k$ the optimal revenue given allocation x, and let $p^0(x)$ be the corresponding payment rule that makes (x, p) incentive compatible.

Fix any allocation x. Denote the optimal revenue given x by $R(x)$. Let

$$
x_k^{(m)} \coloneqq \frac{1}{m} \lfloor m x_k \rfloor.
$$

Since $(x_k^{(m)})_k$ is monotone and feasible, $(x^{(m)}, p^0(x^{(m)})) \in \mathcal{M}^0_{(m)}$. Note that for each k, we have

$$
0 \le x_k - x_k^m < \frac{1}{m}.
$$

Let x be an optimal allocation among $\mathcal{M}$. We have

$$
R(x^{(m)}) \le \mathrm{Rev}^0_{(m)}(\pi),
$$

and thus

$$
\begin{aligned}
\mathrm{Rev}^0(\pi) - \mathrm{Rev}^0_{(m)}(\pi) &\le \mathrm{Rev}^0(\pi) - R(x^{(m)}) \\
&= \sum_k w_k (x_k - x_k^{(m)}) \\
&\le \frac{1}{m} \sum_k |w_k|.
\end{aligned}
$$

We have

$$
\begin{aligned}
\sum_k |w_k| &\le \sum_k \pi_k \theta_k + \sum_{k=1}^{n-1} \Delta_k^\theta Q_k \\
&\le \max \Theta + \sum_{k=1}^{n-1} \Delta_k^\theta \\
&\le 2 \max \Theta = 1.
\end{aligned}
$$

□

Proposition A.4. *For any π,*

$$\lim_{\varepsilon \downarrow 0} \mathrm{Rev}^{\varepsilon}(\pi) = \mathrm{Rev}^{0}(\pi).$$

Proof. Let

$$f(x,p) \coloneqq \sum_i \pi_i p_i, \quad g_{ij}(x,p) \coloneqq \theta_i(x_j - x_i) - (p_j - p_i), \quad g_{i0} \coloneqq p_i - \theta_i x_i.$$

Define

$$v(\varepsilon) = \max\left\{ f(x,p) \colon \forall i,j, g_{ij}(x,p) \leq \varepsilon, g_{i0} \leq \varepsilon,\ (x,p) \in [0,1]^n \times [0,1/2]^n \right\}.$$

Note that $v(0) = \mathrm{Rev}^0(\pi)$ and $v(\varepsilon) = \mathrm{Rev}^{\varepsilon}(\pi)$. For $\varepsilon \geq 0$, $(x,p) \coloneqq (0.0)$ is feasible, hence $v(\varepsilon)$ is finite. Thus, v is concave and continuous on $(0,\infty)$. To show $\lim_{\varepsilon \to 0} v(\varepsilon) = v(0)$, it remains to show that v is right-continuous at 0.

Take any sequence $\varepsilon_k \searrow 0$. For each k, let

$$(x^k, p^k) \in \arg\max \left\{ f(x,p) \colon \forall i,j, g_{ij}(x,p) \leq \varepsilon_k, g_{i0} \leq \varepsilon_k,\ (x,p) \in [0,1]^n \times [0,1/2]^n \right\}.$$

Since v is increasing, $v^* \coloneqq \lim_k v(\varepsilon_k)$ exists. Since $[0,1]^n \times [0,1/2]^n$ is compact, there exists a convergent subsequence of (x^k, p^k) (we do not relabel.) Let $(x^*, p^*) \coloneqq \lim_k (x^k, p^k)$. Note that $v^* = \lim_k f(x^k, p^k)$. This mechanism is feasible with $\varepsilon = 0$ given the structure of the constraints, and thus $v(0) \geq v^*$. Since v is increasing in ε, it is also clear that $v(0) \leq v^*$. Thus, $v^* = v(0)$. □

Corollary A.4. *For any π and $m \in \mathbb{Z}_{>0}$,*

$$\lim_{\varepsilon \downarrow 0} \mathrm{Rev}^{\varepsilon}_{(m)}(\pi) = \mathrm{Rev}^{0}_{(m)}(\pi).$$

Proof. The same proof applies even with the integral constraints on x, since with the additional constraints, the feasible set is still compact and $v(\varepsilon)$ is finite for all $\varepsilon \in [0,\infty)$. v is still concave as it is the maximum of linear functions. □

Corollary A.5. *For any finite $\Theta \subseteq \mathbb{R}_+$, $\pi \in \Delta(\Theta)$, we have*

$$\lim_{(\varepsilon,m) \to (0,\infty)} \left| \mathrm{Rev}^{\varepsilon}_{(m)}(\pi) - \mathrm{Rev}^{0}(\pi) \right| = 0$$

Proof. By Propositions A.3 and A.4, we have

$$\left| \mathrm{Rev}^{\varepsilon}_{(m)}(\pi) - \mathrm{Rev}^{0}(\pi) \right| \leq \left| \mathrm{Rev}^{\varepsilon}_{(m)}(\pi) - \mathrm{Rev}^{0}_{(m)}(\pi) \right| + \left| \mathrm{Rev}^{0}_{(m)}(\pi) - \mathrm{Rev}^{0}(\pi) \right| \to 0,$$

as $\varepsilon \to 0$ and $m \to \infty$. □

■Breaking ties.

Proposition A.5 (Breaking ties with arbitrarily small revenue loss)**.** *Fix a belief $\pi \in \Delta(\Theta)$. Fix any deterministic tie-breaking rule τ that selects, for each $\theta \in \Theta$ and each menu M, an action*

$$\tau(\theta, M) \in \arg\max_{a \in \mathcal{A}} \{\theta x(a) - p(a)\}.$$

Let $M = (x(a), p(a))_{a \in \mathcal{A}}$ satisfy $x(a) \in [0,1]$ and $p(a) \in [0, 1/2]$ for all a, and assume there is an outside option $a_0 \in \mathcal{A}$ with $x(a_0) = p(a_0) = 0$. Define expected revenue under τ by

$$R(M) := \sum_{i=1}^{n} \pi(\theta_i)\, p(\tau(\theta_i, M)).$$

Then for every $\delta > 0$, there exists a menu $M' = (x'(a), p'(a))_{a \in \mathcal{A}}$ with $x'(a) \in [0,1]$, $p'(a) \in [0, 1/2]$ such that:

1. *For every $\theta_i \in \Theta$, the set $\arg\max_{a \in \mathcal{A}} \{\theta_i x'(a) - p'(a)\}$ is a singleton.*
2. *$R(M') \geq R(M) - \delta$, where $R(M')$ is the expected revenue induced by the (now unique) best responses.*

Moreover, one can take $x' = x$ and make $\sup_{a \in \mathcal{A}} |p'(a) - p(a)|$ arbitrarily small.

Proof. Write $u_i(a) := \theta_i x(a) - p(a)$ for the utility of type θ_i from action a under M. Let

$$U_i := \max_{a \in \mathcal{A}} u_i(a), \qquad B_i := \{a \in \mathcal{A} : u_i(a) = U_i\}$$

be the maximal utility and the best-response set for type i.

Define the maximal payment among best responses:

$$P_i := \max_{a \in B_i} p(a),$$

and choose $a_i^\star \in B_i$ with $p(a_i^\star) = P_i$. Since $\tau(\theta_i, M) \in B_i$, we have $p(\tau(\theta_i, M)) \leq P_i$ for each i.

Since $\mathcal{A}$ is finite, there exist distinct numbers $(\kappa(a))_{a \in \mathcal{A}} \subset [0,1]$. Fix one such assignment and impose $\kappa(a_0) = 0$.

Let

$$\Delta_p := \begin{cases} \min\{|p(a) - p(b)| : a, b \in \mathcal{A},\ p(a) \neq p(b)\}, & \text{if payments are not all equal,} \\ +\infty, & \text{if } p(a) \text{ is constant on } \mathcal{A}, \end{cases}$$

so $\Delta_p \in (0, +\infty]$.

Next define the minimal positive strict gap below the maximum:

$$g := \begin{cases} \min\{U_i - u_i(a) : i \in \{1, \ldots, n\},\ a \in \mathcal{A} \setminus B_i\}, & \text{if some } \mathcal{A} \setminus B_i \neq \emptyset, \\ +\infty, & \text{if } B_i = \mathcal{A} \text{ for all } i. \end{cases}$$

Since the set in the minimum is finite, $g \in (0, +\infty]$.

Fix $\delta > 0$. Choose $\eta > 0$ such that

$$\eta \leq \min\left\{\frac{1}{4},\ 2\delta,\ \frac{g}{4},\ \frac{\Delta_p}{2}\right\}. \tag{20}$$

Define $M' = (x', p')$ by keeping allocations unchanged and perturbing payments:

$$x'(a) := x(a), \qquad p'(a) := (1-\eta)p(a) + \eta^2\kappa(a), \qquad \forall a \in \mathcal{A}.$$

Feasibility of bounds. For each a, $p'(a) \geq 0$ since $p(a) \geq 0$ and $\kappa(a) \geq 0$. Moreover, using $p(a) \leq 1/2$, $\kappa(a) \leq 1$, and $\eta \leq 1/4$,

$$p'(a) \leq (1-\eta)\cdot\frac{1}{2} + \eta^2 = \frac{1}{2} - \frac{\eta}{2} + \eta^2 \leq \frac{1}{2},$$

where the last inequality uses $\eta^2 \leq \eta/2$ for $\eta \leq 1/2$. Also $p'(a_0) = 0$ and $x'(a_0) = 0$ because $p(a_0) = x(a_0) = 0$ and $\kappa(a_0) = 0$.

Under M', type i's utility from a is

$$u_i'(a) := \theta_i x'(a) - p'(a) = \theta_i x(a) - \big((1-\eta)p(a) + \eta^2\kappa(a)\big) = u_i(a) + \eta p(a) - \eta^2\kappa(a).$$

Lemma A.12. *Fix i. If $b \in \mathcal{A} \setminus B_i$, then b is not optimal under M'.*

Proof. Let $b \in \mathcal{A} \setminus B_i$. Then $U_i - u_i(b) \geq g$. Since $p(b) \leq 1/2$ and $\kappa(b) \geq 0$,

$$u_i'(b) = u_i(b) + \eta p(b) - \eta^2\kappa(b) \leq u_i(b) + \frac{\eta}{2}.$$

On the other hand, for $a_i^\star \in B_i$,

$$u_i'(a_i^\star) = U_i + \eta P_i - \eta^2\kappa(a_i^\star) \geq U_i - \eta^2.$$

Hence

$$u_i'(a_i^\star) - u_i'(b) \geq (U_i - \eta^2) - \left(u_i(b) + \frac{\eta}{2}\right) = (U_i - u_i(b)) - \frac{\eta}{2} - \eta^2 \geq g - \frac{\eta}{2} - \eta^2.$$

Using (20), we have $\eta \leq g/4$ and $\eta \leq 1/4$, so $\eta/2 + \eta^2 \leq \eta/2 + \eta/4 = 3\eta/4 \leq 3g/16 < g$, hence $g - \eta/2 - \eta^2 > 0$. Therefore $u_i'(a_i^\star) > u_i'(b)$, so b cannot be optimal under M'. □

By Lemma A.12, any best response under M' lies in B_i.

Lemma A.13. *Fix i. The maximizer of $u_i'(\cdot)$ over B_i is unique.*

Proof. Take distinct $a, b \in B_i$. Then

$$u_i'(a) - u_i'(b) = \eta(p(a) - p(b)) - \eta^2(\kappa(a) - \kappa(b)).$$

If $p(a) \neq p(b)$, then $|p(a) - p(b)| \geq \Delta_p$ and $|\kappa(a) - \kappa(b)| \leq 1$. If $p(a) > p(b)$, then

$$u_i'(a) - u_i'(b) \geq \eta\Delta_p - \eta^2 = \eta(\Delta_p - \eta) > 0,$$

where the last inequality uses $\eta \leq \Delta_p/2$ from (20). Thus, within B_i, actions with larger $p(\cdot)$ give strictly larger $u_i'(\cdot)$.

If instead $p(a) = p(b)$, then

$$u_i'(a) - u_i'(b) = -\eta^2(\kappa(a) - \kappa(b)) \neq 0,$$

since κ is injective. Hence, no tie remains. Therefor,e the maximizer over B_i is unique. □

Combining Lemmas A.12 and A.13, type θ_i has a unique best response in $\mathcal{A}$ under M'. Since i was arbitrary, every type has a unique optimal action under M'.

Let a_i' denote the unique best response of type θ_i under M'. By Lemma A.12 and Lemma A.13, we have $a_i' \in B_i$ and it maximizes $p(a)$ over B_i. Hence $p(a_i') = P_i$, and therefore

$$R(M') = \sum_{i=1}^{n} \pi(\theta_i)p'(a_i') = \sum_{i=1}^{n} \pi(\theta_i)\big((1-\eta)P_i + \eta^2\kappa(a_i')\big) \geq (1-\eta)\sum_{i=1}^{n} \pi(\theta_i)P_i.$$

Since $P_i \geq p(\tau(\theta_i, M))$ for each i,

$$R(M') \geq (1-\eta)\sum_{i=1}^{n} \pi(\theta_i)p(\tau(\theta_i, M)) = (1-\eta)R(M).$$

Because $p(\cdot) \in [0, 1/2]$, we have $R(M) \leq 1/2$, and thus

$$(1-\eta)R(M) \geq R(M) - \frac{\eta}{2}.$$

By (20), $\eta \leq 2\delta$, hence $\eta/2 \leq \delta$ and therefore

$$R(M') \geq R(M) - \delta.$$

This completes the proof. □

C.2.2 Approximation of Agent's Payoffs

Let

$$\begin{aligned}\mathcal{M}^{\varepsilon}_{(m)} &:= \left\{(x,p) \in \mathcal{M}_{(m)} \colon (x,p) \text{ satisfies } \varepsilon\text{-IC and } \varepsilon\text{-IR}\right\},\\ R(x,p;\pi) &:= \sum_i \pi_i p_i,\\ u(\theta_i,(x,p)) &:= \theta_i x_i - p_i.\end{aligned}$$

Proposition A.6. *Fix any principal's belief π and a sequence $(\varepsilon^T)_T$ such that $\varepsilon_T \searrow 0$ as $T \to \infty$. Suppose that a sequence of direct mechanisms $(x^T, p^T)_T$ satisfies*

$$\begin{aligned}&\forall T \in \mathbb{N}, (x^T,p^T) \in \mathcal{M}^{\varepsilon^T},\\ &R(x^T,p^T;\pi) \to \mathrm{Rev}^*(\pi)\end{aligned}$$

For any agent's type θ_i, we have

$$\liminf_{T\to\infty} u\left(\theta_i,(x^T,p^T)\right) \geq u^*(\theta_i,\pi).$$

Proof. Note that the set $\mathcal{M}^{\varepsilon}$ is nonempty and compact for any $\varepsilon > 0$. Fix any $\theta_i \in \Theta$. Below, we subsume π and θ_i: let $R(x,p) := \sum_i \pi_i p_i$ and $u_i(x,p) := \theta_i x_i - p_i$.

Let $v^T := u_i(x^T,p^T)$. We can take a subsequence of $(\varepsilon^T)_T$, denoted by $(\varepsilon^s)_s$, such that $\lim_s v^s = \liminf_T v^T$. We have $v^s = u_i(x^s,p^s)$. $(x^s,p^s)_s$ is a sequence in a compact set $[0,1]^n \times [0,1/2]^n$. Take a converging subsequence $(x^\tau,p^\tau)_\tau$, which converges to $(x^*,p^*) \in [0,1]^n \times [0,1/2]^n$. For each τ, it satisfies ε^τ-IC and ε^τ-IR. Since $\varepsilon^\tau \to 0$, (x^*,p^*) satisfies 0-IC and 0-IR.

By construction, $R(x^\tau,p^\tau) = \sum_i \pi_i p_i^\tau$. By assumption, we have

$$\lim_\tau R(x^\tau,p^\tau) = \mathrm{Rev}^*(\pi).$$

By continuity of R, the LHS coincides with $R(x^*,p^*)$. Thus, we have $R(x^*,p^*) = \mathrm{Rev}^*(\pi)$, and thus $(x^*,p^*) \in \mathcal{M}^*(\pi)$. Therefore, we have $u_i(x^*,p^*) \geq u^*(\theta_i,\pi)$.

By continuity of u_i, we have

$$\liminf_T u\left(\theta_i,(x^T,p^T)\right) = \liminf_T v^T = \lim_\tau v^\tau = \lim_\tau u_i(x^\tau,p^\tau) = u_i(x^*,p^*) \geq u^*(\theta_i,\pi).$$

□

D No-WER algorithms

For completeness, we present the relevant results for no-WER algorithms. All the results and proofs are from Slivkins (2019). Throughout this section, we fix the agent's type $\theta \in \Theta$ and subsume it. Assume that the principal's actions $(x_t, p_t)_{t=1}^T$ is iid-drawn from $\sigma_P \in \Delta(\mathcal{M})$. Let $r_t(a) := u(a, (x_t, p_t)) + 1/2 \in [0,1]$ (Here, r stands for "reward".) Then, $(r_t(a))_t$ is iid-drawn from $P_a \in \Delta([0,1])$ with mean $\mu(a) \in [0,1]$ for each $a \in \mathcal{A}$. Let

$$a^* \in \arg\max_{a \in \mathcal{A}} \mu(a), \quad \Delta(a) := \mu(a^*) - \mu(a) \geq 0.$$

Note that for any strategy σ_A^T, we have

$$\mathrm{WER}(\sigma_A^T; T, \sigma_P) = \mathbb{E}\left[\sum_{t=1}^T \Delta(a_t)\right] = \sum_{a \in \mathcal{A}} \Delta(a)\mathbb{E}[n_T(a)],$$

where $n_T(a)$ is defined in (3). Note that $n_T(a)$ is a random variable depending on σ_A^T.

Definition A.2 (Clean Events)**.** Suppose that the confidence intervals are defined as (3). For a fixed agent's strategy σ_A^T and principal's stationary strategy σ_P, the following event is called the *clean event (CE)*:

$$\{\forall a \in \mathcal{A}\ \forall t \leq T, \quad |\mu_t(a) - \mu(a)| \leq \rho_t(a)\}.$$

Conditional on CE, for each action, the confidence interval includes the true mean throughout the game play.

Remark A.2. Corollary A.2 states that the clean event happens with probability at least $1 - 2/T^2$.

Lemma A.14. *Fix any agent's strategy σ_A^T, horizon $T \in \mathbb{Z}_{>0}$, and stationary strategy σ_P. Suppose that, conditional on the clean event, we have*

$$\forall a \in \mathcal{A}, \quad \Delta(a) \leq O\left(\sqrt{\frac{\log T}{n_T(a)}}\right). \tag{21}$$

Then, we have $\mathrm{WER}(\sigma_A^T; T, \sigma_P) \leq O(\sqrt{|\mathcal{A}| T \log T})$.

Proof. We will subsume σ_A^T and σ_P throughout the proof. By (21), on the clean event, we have

$$\sum_a \Delta(a) n_T(a) \leq O(\sqrt{\log T}) \sum_a \sqrt{n_T(a)}.$$

By Jensen's inequality, we have

$$\sum_a \sqrt{n_T(a)} = |\mathcal{A}| \sum_a \frac{1}{|\mathcal{A}|} \sqrt{n_T(a)} \leq |\mathcal{A}| \sqrt{\sum_a \frac{n_T(a)}{|\mathcal{A}|}} = \sqrt{|\mathcal{A}|T}.$$

Therefore, we have

$$\begin{aligned}
\mathrm{WER}(T) &= \mathbb{E}\left[\sum_a \Delta(a) n_T(a) \mid \mathrm{CE}\right] \Pr(\mathrm{CE}) + \mathbb{E}\left[\sum_a \Delta(a) n_T(a) \mid \neg\mathrm{CE}\right] \Pr(\neg\mathrm{CE}) \\
&\leq O\left(\sqrt{|\mathcal{A}| T \log T}\right) + \frac{1}{T^2} T \\
&= O\left(\sqrt{|\mathcal{A}| T \log T}\right).
\end{aligned}$$

□

Proposition A.7. *Successive Elimination (Algorithm 2) has* $\left(5\sqrt{|\mathcal{A}|}, \frac{1}{2}, \frac{1}{2}\right)$*-no-WER.*

Proof. First, suppose that the clean event happens. Suppose that action a is eliminated at the end of period t. Conditional on the clean event (CE), we have

$$\Delta(a) \leq 2(\rho_t(a) + \rho_t(a^*)) = 4\rho_t(a) = 4\sqrt{\frac{\log T}{n_t(a)}} = 4\sqrt{\frac{\log T}{n_T(a)}},$$

where the last equality follows since $n_t(a) = n_T(a)$ if a is removed at the end of period t. Then, we have

$$\begin{aligned}
\mathrm{WER}(T) &= \mathbb{E}\left[\sum_a \Delta(a) n_T(a)\right] \\
&= \mathbb{E}\left[\sum_a \Delta(a) n_T(a) \mid \mathrm{CE}\right] \Pr(\mathrm{CE}) + \mathbb{E}\left[\sum_a \Delta(a) n_T(a) \mid \neg\mathrm{CE}\right] \Pr(\neg\mathrm{CE}) \\
&\leq 4\sqrt{\log T} \sum_a \sqrt{n_T(a)} + \frac{1}{T^2} T \\
&\leq 4\sqrt{\log T} \sqrt{|\mathcal{A}| T} + \frac{1}{T} \\
&= 5\sqrt{|\mathcal{A}| \log T} \sqrt{T}.
\end{aligned}$$

□

Proposition A.8. *UCB has* $\frac{1}{2}$*-no-WER.*

Proof. See Slivkins (2019). □

There is another more primitive no-WER algorithm called *Uniform Exploration*: try each arm N times, and then choose the arm with the highest average reward in all remaining rounds. If we choose N suitably to balance the exploration-exploitation trade-off, it achieves no-weak-ER.

Proposition A.9. *If we set* $N \coloneqq T^{2/3}(\log T)^{1/3}$*, then Uniform Exploration has* $\frac{2}{3}$*-no-WER.*

Proof. Let $\overline{\rho}_N \coloneqq \sqrt{(2\log T)/N}$ and $K \coloneqq |\mathcal{A}|$.

For each action, true mean $\mu(a)$ is included in the confidence interval $\overline{\mu}_N(a) \pm \overline{\rho}_N$ with probability at least $1 - 2/T^4$. Thus, the clean event (CE), on which the true means are in the confidence intervals for all actions, happens with probability $1 - O(T^{-4})$. Conditional on the CE, if action a is chosen at the end of the exploration phase, we have $\mu(a^*) - \mu(a) \leq 2\overline{\rho}_N$. Therefore, conditional on CE, we have the regret bound

$$KN + (T - KN)2\overline{\rho}_N.$$

Therefore, by setting $N \coloneqq T^{2/3}(\log T)^{1/3}$, we have the following regret bound

$$\begin{aligned} R(T) &\leq \left(1 - O(T^{-4})\right)\left[KN + (T - KN)2\overline{\rho}_N\right] + O(T^{-4})T \\ &= O\left(T^{2/3}\left(\log T\right)^{1/3}\right). \end{aligned}$$

□