Bridging Truthfulness and Corruption-Robustness in Multi-Armed Bandit Mechanisms

Jacob Abernethy [1] **Bhuvesh Kumar** [1] **Thodoris Lykouris** [2] **Yinglun Xu** [1]

Abstract

We study pay-per-click auctions where both the principal and the agents employ learning algorithms to learn the click-through-rates and intrinsic values respectively. In this setting, we illustrate a trade-off between a) effective learning in a truthful manner and b) robustness to the presence of adversarial corruptions. We design a mechanism that balances these two conflicting forces and achieves a graceful degradation in performance with the amount of corruption in the data without compromising the performance when there is no corruption. On the way we demonstrate that agent-learning introduces additional challenges in multi-armed bandit mechanisms even at the absence of corruptions, which may be of independent interest.

1. Introduction

The multi-armed bandit (MAB) problem is among the most basic and perhaps the most well-studied in sequential decision making. A learner (or *principal*) must select one among k possible actions, often referred to as "arms," and each arm $a \in [k]$ returns a reward $r^t(a)$ to the learner that is sampled according to some fixed distribution $\mathcal{F}(a)$. When the learner selects an arm a^t at round t, and observes and earns the reward $r^t(a^t)$, no new information about the other arms is made available. This information bottleneck presents the learner with the so-called *explore-exploit trade-off*: one can either try to maximize expected reward based on the currently-available information, or try to select from under-explored arms in order to enhance future decision-making.

In this paper we consider the application of MAB algorithms to the design of *ad auctions*, where advertising inventory is repeatedly allocated through advertiser bidding, as this presents a canonical example with an unavoidable explore-exploit trade-off. The learner in this case is effectively a sales platform which must sequentially select one among many ads to display in response to users' search queries. From the perspective of the advertising platform, an arm corresponds to one of the k advertisers competing to market their product. Let us give a rough overview of the challenges of this setting when the ad platform aims to maximize *social welfare*; more on this concept later. When the platform selects an ad to display to the user, the advertiser is only charged in the event that the user finds the advertisement of interest and then clicks on it. Clicking an ad doesn't automatically lead to a purchase, of course, and the ad only creates value to the advertiser when it leads to a buying the advertised product. We need to specify model parameters for these two stochastic events: assume that each ad a generates a click by a random user with probability $\rho(a)$ (the *click-through-rate*), and in the event the ad is clicked the user makes a purchasing decision of some value, which we assume to be an IID random variable with expectation $\mu(a)$. One of the central challenges of designing ad auctions is that *these two parameters are not known in advance*, the principal needs to carefully manage the explore-exploit trade-off as click statistics and purchase data slowly arrive.

One may notice, however, that the above model did not address the actions of the advertisers whose own incentives might lead to unexpected behavior. This poses additional hurdles to the principal, as the observed rewards may indeed not be IID but could instead be heavily influenced by strategic considerations. In online advertising, in particular, the arms are not passive actors but correspond rather to *agents* who wish to maximize their individual payoff over the long term.

There are two prominent forms of gaming that the agents can employ to increase their individual payoff, potentially compromising the social welfare. First, typically agents know their intrinsic value mean $\mu(a)$ and make a bid based on that; if misreporting this information can lead to higher individual payoff then agents will possibly do so. As a result, a line of work has focused on designing *truthful multi-armed bandit mechanisms*, i.e., making truth-telling a payoff-maximizing bidding strategy and thus rendering such attacks ineffective (Devanur & Kakade, 2009; Babaioff et al., 2014). Another attack aims to manipulate the feedback that the principal observes. For example, an agent may create a

[1]Georgia Tech [2]Microsoft Research NYC. Correspondence to: Bhuvesh Kumar <bhuvesh@gatech.edu>, Yinglun Xu <yxu647@gatech.edu>.

bot that either clicks her own ads or does not click competing ads to adversarially bias the click-through-rate estimates and make the ad seem less desirable to the principal. A separate line of work has designed algorithms that are *robust to the presence of such adversarial corruptions* in the data (Lykouris et al., 2018; Gupta et al., 2019).

The two attacks described above can, and unfortunately very frequently do, occur simultaneously. Indeed this brings us to the central challenge that we aim to address in the present work: existing techniques that have attempted to defend against the first attack actually increase vulnerability to the second, *and the same is true vice versa*; see the techniques section below for additional details. Our goal in this paper is to propose methods to mitigate both vulnerabilities at once, and this brings us to the following core question:

How can we design MAB mechanisms that are simultaneously truthful and corruption-robust?

Our results. Towards addressing this question, we consider the setting of a repeated ad auction where some adversary is allowed a certain number of corruptions to the feedback. Precisely, the adversary may adjust no more than C out of T many rounds, where C is unknown to the principal. There is no restriction on how the adversary distributes her corruption budget; it may well be the case that all C first rounds are corrupted. Mechanism performance is measured in terms of regret that compares the reward of the mechanism relative to the reward that would be achieved if the best arm were known in advance. Without the introduction of corruptions it is known that simple mechanisms (Devanur & Kakade, 2009; Babaioff et al., 2014) achieve regret on the order of $T^{2/3}$, a rate known to be unimprovable. In simple MAB, without the introduction of auctions and bids, there are several "robust" algorithms (Lykouris et al., 2018; Gupta et al., 2019) whose regret is optimal in T when $C = 0$ and gracefully degrades with C otherwise.

We provide a similar guarantee despite the additional requirement of truthfulness. In particular, our regret guarantee (Theorem 3) achieves the aforementioned desiderata and remains sublinear as long as the corruption level $C = o(T)$. Interestingly, when the arms are well separated (by a notion of *gap*), additional degradation in performance starts appearing only when $C = \Omega(T^{2/3})$. Subsequently, we move our attention to the revenue objective where the goal of the principal is to maximize the payment they collect from the advertisers. For this objective, we show that the same algorithm has essentially the same guarantee (Theorem 4) but the additional degradation starts when $C = \Omega(T^{1/3})$ as the revenue objective does not only require to identify the correct arm but also to ensure that the click-through-rates are accurate enough to induce the desired payments.

As a test of robustness of our findings, we study what occurs when the agents do not know their intrinsic values but instead need to also employ a learning algorithm. We show that, under a natural relaxation of truthfulness, our guarantees seamlessly extend to this setting where agents are also learning (Theorems 5 and 6). Notably, agent learning does add more complications to the setting; we show a simple setting where truthful bidding leads to logarithmic regret while agent learning renders guarantees $o(T^{2/3})$ unattainable (Remark 1).

Our techniques. Our technique builds on insights from prior work on truthful MAB mechanisms and corruption-robust MAB algorithms but needs to make careful adaptations to circumvent the shortcomings they exhibit at the presence of the additional attack.

The canonical approach for truthful MAB mechanisms (Babaioff et al., 2014) is the so-called *explore-then-commit*. In the *explore phase* that lasts $T^{2/3}$ rounds, the mechanism selects all arms to create accurate click-through-rate estimates while ignoring the agent bids. In the (*commit phase*), it applies a weighted second price auction with weights corresponding to the computed estimates. At the presence of corruptions, this algorithm can fail dramatically; in particular, if $C > T^{2/3}$), the adversary can corrupt the whole explore phase misleading the principal into believing that the best arm's click-through-rate is 0 and causing linear regret when there is a separation between the best arm and the second-best.

In contrast, corruption-robust algorithms (Lykouris et al., 2018; Gupta et al., 2019) employ adaptive exploration and continuously refine the estimates they have in order to achieve both gap-dependent regret and corruption-robustness. Since truthfulness requires *exploration separation*, i.e., disentangling exploration from exploitation, such techniques are not directly applicable as they continue exploring to enhance gap estimation and therefore do not provide a direct way for the principal to exploit and acquire revenue.

Our mechanism combines traits of both approaches by using a multi-layering approach to guarantee adaptive exploration and corruption-robustness (Lykouris et al., 2018) while also using exploration separation (Babaioff et al., 2014) within each layer. One nice technical contribution is a careful way to connect the layers via using the tightest confidence intervals among all layers and initializing the confidence intervals when they fail. This allows us to derive regret guarantees that are meaningful even for any $C = o(T)$ unlike previous multi-layering approaches.

Related work. Our work contributes to the literature of principal learning in auction design in repeated settings with strategic agents. The pay-per-click auction setting that we consider has a nice single-dimensional representation

for each agent which makes multi-armed bandit techniques amenable to it. At the absence of adversarial corruptions, truthful multi-armed bandit mechanisms for pay-per-click auctions were initially studied by Babaioff et al. (Babaioff et al., 2014), and Devanur and Kakade (Devanur & Kakade, 2009); their results were later extended to contextual multi-slot setting by Gatti et al (Gatti et al., 2012). Outside of pay-per-click auctions there are multiple auction settings where principal learning with strategic agents arises, including posted prices (Amin et al., 2013; 2014; Mohri & Munoz, 2014; 2015), reserve prices (Liu et al., 2018), or revenue maximizing auctions (Abernethy et al., 2019). These works assume either the bidders discount their future utilities or assume that the same bidders only shows up for a limited number of rounds and thus have limited impact over principal learning. We don't make any such assumptions about the agents' behavior. Finally, we note that repeated auction design with strategic agents also arises in many other works within dynamic mechanism design (Bergemann & Valimaki, 2006; Nazerzadeh et al., 2008; Kakade et al., 2013) for a non-exhaustive list. All the above works do not extend to the presence of adversarial corruptions and they also require the agents to know their exact values at the beginning of each round; in contrast, we allow learning to also arise at the agent level (Section 4).

Our work also adds to the growing literature that deviates from assuming that agents have ex-ante information about their values but rather studies settings where they need to learn how to act on the fly. Online learning for an agent that does not know their value was initially studied in the context of second-price auctions (Weed et al., 2016) and subsequently extended to more general settings (Feng et al., 2018). Another related research direction involves mechanism design questions when the principal knows that the agents are employing no-regret learning to decide their actions; this was initially suggested in the context of auctions by Braverman et al. (Braverman et al., 2017) and subsequently extended to more general learning strategies (Deng et al., 2019a) and other game settings (Deng et al., 2019b). The inverse problem of reacting as a buyer to a no-regret seller has also been studied in (Heidari et al., 2016).

Finally, the model of adversarial corruptions that we employ to capture applications such as click fraud was introduced in (Lykouris et al., 2018), the results were later refined (Gupta et al., 2019; Zimmert & Seldin, 2019) while another variant to capture adversarial corruptions was suggested in (Kapoor et al., 2018). This model provides a way to make stochastic algorithms robust to adversarial corruptions, retaining the stochastic guarantee when there is no corruption and achieving a degradation in performance that is graceful to the amount of corruption. As such, it has been widely applied in multiple settings such as linear optimization (Li et al., 2019), assortment optimization (Chen et al., 2019), reinforcement learning (Lykouris et al., 2019), Gaussian process bandit optimization (Bogunovic et al., 2020), contextual pricing (Krishnamurthy et al., 2020), and prediction with expert advice (Amir et al., 2020). A related line of work focuses on designing adversarial attacks against classical stochastic bandit algorithms allowing higher power to the adversary (Jun et al., 2018; Liu & Shroff, 2019). Closer to our work, Feng et al. (Feng et al., 2020) analyze different classical bandit algorithms with respect to their intrinsic robustness to particular classes of adversarial corruptions. Our work is the first to consider the design of algorithms robust to adversarial corruptions in the presence of strategic behavior by the participants in settings where classical algorithms are prone to simple attacks such as click fraud.

2. Model

We consider a single-slot pay-per-click (PPC) ad auction, consisting of repeated auctions with T rounds and K advertisers or *agents*. At each round, the advertisers compete for an ad impression and the auctioneer or *principal* selects one advertiser to display and a payment to charge.

Classical PPC setting. In the stochastic setting for PPC auctions, each advertiser or agent j is associated with a *click-through-rate* (CTR) ρ_j which determines the probability of getting clicked if she is selected by the principal; the click-through rates are unknown to the principal. More formally, at round t, each agent j makes a bid b_j^t and the principal displays the ad of agent a^t and charges a payment p^t which is not allowed to be above $b_{a^t}^t$. The click indicator c^t is a Bernoulli random variable with mean ρ_{a^t} that is 1 when the displayed ad gets clicked and 0 otherwise; if the click occurs, the agent pays p^t otherwise she does not pay anything (this is why the auction is called *pay-per-click*). Each agent j is also associated with an income $v_j^t \in [0, 1]$ that comes from a fixed agent-dependent distribution with mean μ_j; we refer to this mean as mean value (or mean income). Agents are assumed to bid in a way that maximizes their expected utility which is assumed to be quasilinear, i.e., value they obtain minus payment (their exact bidding strategy is formalized below). We initially assume that agent j knows μ_j (Section 3) and subsequently extend our results to a setting where the agent needs to learn it (Section 4). When players know their mean value, the above model is exactly the setting studied in previous work on truthful multi-armed-bandit mechanisms (Devanur & Kakade, 2009; Babaioff et al., 2014).

PPC with adversarial corruptions. We extend the classical PPC setting to allow for adversarial corruptions in observed rewards. Following the model of (Lykouris et al., 2018), we assume that an adversary can corrupt the results of the clicks c^t and the incomes v_j^t. The adversary can ob-

serve all the history of past outcomes until round t as well as the principal's distribution at round t but does not have access to the random selection of arm a^t; this is consistent to the adversarial bandit literature. The adversary has a corruption budget C which we term *corruption level* and captures the number of rounds that the adversary is allowed to corrupt. The corruption level is unknown to the principal.

Regret and performance of mechanism. The principal's performance is evaluated by *regret* which captures the loss in performance due to the principal not knowing the click-through-rates in advance. If the principal had access to the click-through-rates then the welfare-maximizing option is to select the agent j that has the highest utilization $\mu_j\rho_j$. For the revenue objective, in order to not encourage misreporting on the agent side (formalized below), the highest revenue that the principal could achieve is determined by the second highest utilization. More formally the two benchmarks are the following:

$$\textsc{WelOpt}(T) = T \cdot \max_j \rho_j\mu_j$$
$$\textsc{RevOpt}(T) = T \cdot \operatorname{smax}_j \rho_j\mu_j,$$

where smax refers to the second highest value. Without loss of generality, for the rest of the paper, we assume $\rho_1\mu_1 \geq \rho_2\mu_2 \geq \cdots \geq \rho_K\mu_K$. The welfare and revenue regret capture the extent to which the principal's performance is inferior to the above benchmarks:

$$\textsc{WelReg}(T) = T \cdot \max_j \rho_j\mu_j - \sum_{t=1}^{T} \rho_{a^t}\mu_{a^t}$$
$$\textsc{RevReg}(T) = T \cdot \operatorname{smax}_j \rho_j\mu_j - \sum_{t=1}^{T} p^t\rho_{a^t}.$$

Truthfulness requirement. In multi-armed bandit mechanisms such as pay-per-click auctions, it is important to induce the agents to not misreport their value; if the mechanism exploits the bidding pattern of an agent in order to charge them higher prices, then this creates incentive to the bidders to shade their bids which may cause the mechanism to experience unpredictable behavior. As a result, a desirable property in mechanism design is *truthfulness* which suggests that truth-telling is a dominant strategy for the agent, i.e., she cannot increase her utility by misreporting her value. The following definition quantifies this requirement for the case where the agents know their value means.

Definition 1 (Truthful Mechanism)**.** *A mechanism is* truthful *if for any sets of click-through-rates and value means, every agent j obtains higher expected utility by bidding her true value $b_j^t = \mu_j$ at every round; in other words, if the mechanism is truthful, misreporting does not help any agent.*

The above definition means that truthful bidding is a dominant strategy for the agents when they know their value. In Section 4, we suggest an extension of this bidding (c.f., Definition 2) that allows the agents to not know their value means in advance but learn it as time goes by.

Desiderata. At the absence of corruptions and when the agent knows her mean value, the principal can achieve a regret guarantee of $T^{2/3}$ via a truthful mechanism; this is unimprovable (Devanur & Kakade, 2009; Babaioff et al., 2014). Our goal is to recover this guarantee when $C = 0$, while gracefully degrading with the corruption level C. Interestingly our regret guarantees are sublinear for any sublinear corruption level $C = o(T)$.[1]

3. Truthful corruption-robust mechanism when agents know their value

In this section, we present a truthful mechanism for the basic setting where the agents know their mean value that satisfies the above desiderata with respect to both welfare and revenue.

Description of our mechanism. A typical way to achieve truthfulness when repeatedly interacting with the same set of agents is to not use the past and present bids of a particular agent to select the price charged. The classical mechanism based on this approach for the uncorrupted case (Babaioff et al., 2014) is a weighted second-price auction. The mechanism initially creates *weights* that are increasing with the bids (the weight functions are based on click-through-rate estimates). It subsequently selects the arm with the highest weight and charges a payment corresponding to the bid that would have incurred the second weight (this is the intuition behind the second highest value of the revenue benchmark). In the uncorrupted case, this can be achieved by simple explore-then-commit mechanisms.

To handle adversarial corruptions, inspired by (Lykouris et al., 2018), we employ a multi-layering version of exploration separation: some rounds help learn CTR estimates while others focus on maximizing welfare/revenue. We create $\log(T^{2/3})$ *layers*, each containing a confidence interval for the CTR of each arm. At round t, every layer ℓ has an upper confidence bound $\text{UCB}_{j,\ell}^t$ and a lower confidence bound for each arm j; the round is *exploration* with probability $\sim \log(T)KT^{-1/3}$ or *expolitation* otherwise. This enables exploration separation desired for truthfulness and allows us to subsample the corruption we encounter for the exploration rounds. In expectation, we only encounter $C \cdot \log(T)KT^{-1/3}$ corruption in exploration rounds as the adversary does not know the randomness in selecting arms at round t; if $C < T^{1/3}$, the expected effect of corruption is negligible. The multi-layering scheme addresses corruption

[1] When $C = \Theta(T)$, the regret needs to scale with C since just the corrupted rounds cause $\Theta(T)$ regret.

levels larger than $T^{1/3}$ as higher layers ℓ update with low probability their respective statistics: the corruption is thus subsampled by $2^{-\ell}$ which enables robustness to corruption up to $T^{1/3}2^{\ell}$. When layer ℓ is selected in an exploration round, it updates its respective statistics: number of trials $n^t_{j,\ell}$ for the selected arm j and empirical click-through rate $\hat{\rho}^t_{j,\ell}$

If the round is an exploitation round, we employ a weighted second-price auction similar to the uncorrupted setting but carefully design the weights to allow the more robust layers to supervise the performance of the less robust layers. In particular, each layer ℓ has a confidence interval for each arm j; we behave optimistically and treat all layers' confidence intervals of each arm as accurate (this is indeed the case with high probability when $C < T^{1/3}$). We then select the agent whose minimum upper confidence bound $\text{UCB}^t_{j,\ell}$ (across layers ℓ) is the highest. If at any point we have evidence that a layer is corrupted, i.e., the lowest confidence bound of an arm in a layer ℓ is higher than the upper confidence bound of the same arm in a less robust layer $\ell' < \ell$, we reset the estimates of layer ℓ'. This allows more robust layers to supervise less robust layers and correct mistakes due to corruption. The algorithm is formally provided in Algorithm 1. The following lemma shows that it is truthful.

Lemma 1. *Algorithm 1 is a truthful mechanism.*

Proof. Since bids affect only the current round, the arms do not have incentive to strategically bid to influence future outcomes. We now analyze the effect of the bids on the current round. If the round is an exploration round, then the bid is irrelevant as the allocation is random and the payment is 0. If the round is an exploitation round, we employ a weighted second price auction which is strictly truthful for any set of weights (bids do not affect these weights). As a result, Algorithm 1 is truthful. □

Welfare guarantee. For the welfare guarantee (Theorem 3), we first show that the weights w^t_j used by Algorithm 1 are always, with high probability, close the true CTRs ρ_j for all agents j.

Lemma 2. *If the corruption level for the mechanism is C, let* $\hat{C} = \max\{T^{1/3}, C\}$. *With probability at least* $1 - \frac{8K\log T}{3T}$, *it holds simultaneously for all arms* j *and all exploitation rounds* $t > 3\hat{C}\log^2(T)$ *of Algorithm 1, the weight* $w^t_j \in \left[\rho_j - 2\left(\sqrt{\frac{\log(T)}{t/2\hat{C}}} + \frac{7\log(T)}{t/2\hat{C}}\right), \rho_j + 2\left(\sqrt{\frac{\log(T)}{t/2\hat{C}}} + \frac{7\log(T)}{t/2\hat{C}}\right)\right]$.

Proof sketch. Similar to (Lykouris et al., 2018), due to the subsampling of corruption, smallest robust layer $\ell^\star = \log\hat{C}$ has its confidence interval inside the desired confidence intervals and does not contradict the confidence intervals of higher layers. By definition of the weight for j is less than $\text{UCB}^t_{j,\ell^\star}$. Since we reset any layer that contradicts

Algorithm 1 MuLES Mechanism (Multi-Layering Exploration Separation MAB Mechanism)

Parameters : Number of arms K, Number of rounds T
Initialize : $\log(T^{2/3})$ layers, set $\hat{\rho}^0_{j;\ell} \leftarrow 0$, $n^0_{j;\ell} \leftarrow 0$ for all $\ell \in [\log(T^{2/3})]$, all $j \in [K]$

for $t = 1, \dots, T$ **do**
  Receive bid b^t_j from arm j for all $j \in [K]$
  $\ell^t \leftarrow \begin{cases} \ell & \text{w.p. } K\log(T)T^{-1/3}2^{-\ell} \\ 0 & \text{otherwise} \end{cases}$
  if $\ell^t > 0$ **then**
    `/* ... Exploration Round ... */`
    $a^t \leftarrow j$ uniformly at random for $j \in [K]$
    Receive click result c^t
    Charge $p^t \leftarrow 0$ `/* Payment */`
    for $j \in [K]$ **do** `/* Update layer */`
      $n^t_{j;\ell} \leftarrow n^{t-1}_{j;\ell} + 1_{\{\ell=\ell^t, j=a^t\}}$
      $\hat{\rho}^t_{j;\ell} \leftarrow \frac{\hat{\rho}^{t-1}_{j;\ell} \cdot n^{t-1}_{j;\ell} + c_t \cdot 1_{\{\ell=\ell^t, j=a^t\}}}{n^{t-1}_{j;\ell}}$
    end
  else
    `/* ... Exploitation Round ... */`
    for $j \in [K]$ **do**
      $w^t_j \leftarrow \min_\ell \left\{\hat{\rho}^t_{j;\ell} + \sqrt{\frac{\log(T)}{n^t_{j;\ell}}} + \frac{7\log(T)}{n^t_{j;\ell}}\right\}$
    end
    $a^t \leftarrow \arg\max_j (w^t_j \cdot b^t_j)$
    Receive click result c^t
    Charge price $p^t \leftarrow \begin{cases} \frac{\text{smax}_j(w^t_j \cdot b^t_j)}{w^t_{a^t}} & \textbf{if } c^t = 1 \\ 0 & \text{otherwise} \end{cases}$
  end
  for $\ell = \log T^{2/3}, \log T^{2/3} - 1, \dots, 2$ **do**
    for $\ell' = \ell - 1, \ell - 2, \dots, 1$ **do**
      for $j = 1, \dots, K$ **do**
        $\text{LCB}^t_{j,\ell} \leftarrow \hat{\rho}^t_{j;\ell} - \sqrt{\frac{\log(T)}{n^t_{j;\ell}}} - \frac{7\log(T)}{n^t_{j;\ell}}$
        $\text{UCB}^t_{j,\ell} \leftarrow \hat{\rho}^t_{j;\ell} + \sqrt{\frac{\log(T)}{n^t_{j;\ell}}} + \frac{7\log(T)}{n^t_{j;\ell}}$
        $\text{LCB}^t_{j,\ell'} \leftarrow \hat{\rho}^t_{j;\ell'} - \sqrt{\frac{\log(T)}{n^t_{j;\ell'}}} - \frac{7\log(T)}{n^t_{j;\ell'}}$
        $\text{UCB}^t_{j,\ell'} \leftarrow \hat{\rho}^t_{j;\ell} + \sqrt{\frac{\log(T)}{n^t_{j;\ell'}}} + \frac{7\log(T)}{n^t_{j;\ell'}}$
        if $[\text{LCB}^t_{j,\ell}, \text{UCB}^t_{j,\ell}]$ and $[\text{LCB}^t_{j,\ell'}, \text{UCB}^t_{j,\ell'}]$ don't overlap **then**
          $\hat{\rho}^t_{j';\ell'} \leftarrow 0$ `/* Reset layer */`
          $n^t_{j';\ell'} \leftarrow 0$
        end
      end
    end
  end
end

more robust layers, it is also higher than $\text{LCB}^t_{j,\ell^\star}$. As a result, it also lies within the desired confidence interval. The complete proof is provided in Appendix A.1. □

Theorem 3. *If the corruption level for the mechanism is C, let* $\hat{C} = \max\{T^{1/3}, C\}$ *and* $\Delta = \min_{j>1}\{\frac{\rho_1\mu_1 - p_j\mu_j}{\mu_1+\mu_j}\}$ *then the expected welfare regret* $\text{WELREG}(T)$ *of Algorithm 1 satisfies*

$$\begin{aligned}\text{WELREG}(T) \leq & K\log(T)T^{2/3} + \\ & 16\cdot\min\left\{\frac{8\hat{C}\log(T)}{\Delta}, \sqrt{2\hat{C}T\log(T)}\right\} + \\ & 14\hat{C}\log^2(T) + \frac{8K\log(T)}{3}\end{aligned}$$

Note that the welfare regret degrades gracefully as the corruption level increases. It is $\widetilde{O}(T^{2/3})$ when $C = 0$ and the leading term in T remains $\tilde{O}(T^{2/3})$ as long as $C \leq O(T^{1/3})$. We also note that for $C > T^{2/3}$, if the utilization gap Δ is large enough, then we obtain gap-dependant bounds.

Proof sketch for Theorem 3. To bound the welfare regret $\text{WELREG}(T)$, we divide in two parts, the regret from exploration rounds $\text{WELREG}_{\text{explore}}(T)$ and the regret from exploitation rounds $\text{WELREG}_{\text{exploit}}(T)$. The expected number of exploration rounds is bounded by $K\log(T)T^{2/3}$ which bounds $\text{WELREG}_{\text{explore}}(T)$. For $\text{WELREG}_{\text{exploit}}(T)$, Lemma 2 implies that after $\frac{32\hat{C}\log(T)}{\Delta^2}$ rounds, the principal can converge to the best agent with high probability. In the cases when an agent other than the best agent is picked, we show that the welfare lost in those rounds is also bounded as the estimates for CTRs are tight. The complete proof is provided in Appendix A.2. □

Revenue guarantee. Revenue guarantees are harder than their welfare counterparts because the revenue is affected by the exact payments charged by the principal. For welfare, it suffices to show that once we converge to the best arm, then the principal incurs no further regret. For revenue, even when the principal converges to the best arm, if the estimates of the CTRs are not tight, then we may end up undercharging and lose more revenue. Our revenue guarantee (Theorem 4) is thus not gap-dependent even when the utilization gap Δ is large. Other than that, the guarantee achieves the same desiderata as Theorem 3. The proof follows similar steps and is deferred to Appendix A.3.

Theorem 4. *If the corruption level for the mechanism is C, let* $\hat{C} = \max\{T^{1/3}, C\}$ *then the expected revenue regret* $\text{REVREG}(T)$ *of Algorithm 1 satisfies*

$$\text{REVREG}(T) \leq \widetilde{O}\left(KT^{2/3} + \sqrt{\hat{C}T}\right)$$

where $\widetilde{O}(\cdot)$ *hides* $\text{polylog}(T)$ *terms.*

4. Extensions when agents do not know their value

When agents do not know their mean value, truthful bidding is not an option. Lemma 2 establishes that the principal can obtain a confidence interval of radius $W(t, C) = 4\sqrt{\frac{2\hat{C}\log(T)}{t}}$ where $\hat{C} = \max\{C, T^{1/3}\}$ on each agent's CTR via the exploration phase. When agent j is selected in the exploration phase, she receives information about her value with probability equal to ρ_j (if clicked). Our relaxation of truthful bidding is that the agents mostly bid based on these confidence intervals.

Definition 2 ((γ, δ)-Confidence Bidding)**.** *If the corruption level of the mechanism is C, let* $\hat{C} = \max\{T^{1/3}, C\}$ *and denote* $W(t, C) = 4\sqrt{\frac{2\hat{C}\log(T)}{t}}$. *Agents are* γ*-confidence bidding if, with probability at least* $1-\delta$*, their bids at all rounds are within* $\mu_j^t \pm \gamma W(\rho_j t, C)$.

Note that this is a reasonable assumption as the agents can create confidence intervals on the order of $W(\rho_j t, C)$ and have reason to do so. Regarding the ability, agents can, for example, follow the principal's estimation technique [2] to create a confidence interval around their empirical mean of size $W(\rho_j t, C)$ that with probability $1 - \frac{1}{T}$ includes the true mean μ_j^t; hence using $\gamma = 2$ and $\delta = \frac{1}{T}$ guarantees that the resulting confidence interval (centered around the empirical mean) includes the empirical mean with high probability. Regarding the purpose, when the true mean is indeed within the confidence interval, bidding anything outside the confidence interval is dominated by bidding either the upper or the lower confidence bound. Finally, a nice trait of the above definition is that the quality of learning is crisply described through the parameter γ. In particular, when agents know their value and bid truthfully, this corresponds to $\gamma = 0$.

We now show that the welfare guarantees seamlessly extend to the setting where agents do not know their mean value and therefore employ confidence bidding instead of truthful bidding.

Theorem 5. *If the corruption level for the mechanism is C, let* $\hat{C} = \max\{T^{1/3}, C\}$ *and* $\Delta = \min_{j>1}\frac{p_1\mu_1 - p_j\mu_j}{\mu_1+\mu_j+\gamma(4+\sqrt{\rho_1}+\sqrt{\rho_j})}$. *If all agents are* (γ, δ)*-confidence bidding, then the welfare regret* $\text{WELREG}(T)$ *of Algorithm 1 satisfies*

$$\text{WELREG}(T) \leq K(\delta T + \log(T)T^{2/3}) +$$

[2] They can also apply any other robust estimation technique and use information from the exploitation rounds which the principal is not allowed to do in order to maintain exploration separation.

$$16(1+3\gamma)\min\left\{\frac{8\log(T)\hat{C}}{\Delta}, \sqrt{2\log(T)\hat{C}T}\right\}+ 14\hat{C}\log^2(T)+\frac{8K\log(T)}{3}$$

We can see that if we have $\gamma = 0$ and $\delta = 0$, i.e. if the arms behave truthfully then we exactly recover the welfare regret from Theorem 3. For corruption level $C = o(T^{1/3})$, the regret from exploration rounds is dominant and independent of the bids reported. The regret decays smoothly with the quality of agent learning as expressed by γ.

The proof is similar to Theorem 3 but has the additional complication that we may need more rounds before we converge to the best agent as the allocation rule in exploitation rounds depends on the reported bids. If the latter are far from the true mean values, we are now more likely to select non-optimal agents and may lose more welfare in exploitation rounds. The complete proof for Theorem 5 is provided in Appendix B.1.

Regarding the revenue regret, we use similar analysis as Theorem 4, but this time the bids are not truthful. This leads to extra complications as the payments estimated and charged by the principal in exploit rounds might be even lower than what we had in Theorem 4.

Theorem 6. *If the corruption level for the mechanism is C, let* $\hat{C} = \max\{T^{1/3}, C\}$. *If all agents are* (γ, δ)*-confidence bidding, the revenue regret* $\textsc{RevReg}(T)$ *of Algorithm 1 satisfies*

$$\textsc{RevReg}(T) \leq \widetilde{O}\left(K(T^{2/3}+\delta T)+(1+\gamma)\sqrt{2\hat{C}T}\right)$$

where $\widetilde{O}(\cdot)$ *hides* $\text{polylog}(T)$ *terms.*

The exact bound and the proof of Theorem 6 can be found in the Appendix B.2.

If the agents do not know their mean values and are trying to learn it during the rounds of the mechanism, the agent learning adds significant challenges in itself even when there are no corruptions.

Remark 1. *Assume that there are no corruptions. When the agents know their true values, and the principal has a lower bound on the gap* $\min_{j>1}\frac{\rho_1\mu_1-\rho_j\mu_j}{\mu_1+\mu_2} \geq \Delta$, *then within* $\widetilde{O}(\frac{K}{\Delta^2})$ *rounds, the principal can converge on the best arm and incur only* $\widetilde{O}(\frac{K}{\Delta^2})$ *revenue regret.*

In contrast, when the agents don't know their values and even if the principal knows the exact CTRs, the principal has to spend $O(KT^{2/3})$ *rounds for exploration so that the agents get* $O(T^{2/3})$ *samples to estimate their values. If not, the agents' bids might be far off from the true value and the principal ends up picking non-optimal agents or undercharging. Thus the agent has to incur at least* $O(KT^{2/3})$ *revenue regret and cannot obtain a gap dependent regret even for large gaps.*

References

Abernethy, J. D., Cummings, R., Kumar, B., Taggart, S., and Morgenstern, J. H. Learning auctions with robust incentive guarantees. In *Advances in Neural Information Processing Systems*, pp. 11587–11597, 2019.

Amin, K., Rostamizadeh, A., and Syed, U. Learning prices for repeated auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, pp. 1169–1177, 2013.

Amin, K., Rostamizadeh, A., and Syed, U. Repeated contextual auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, pp. 622–630, 2014.

Amir, I., Attias, I., Koren, T., Livni, R., and Mansour, Y. Prediction with corrupted expert advice, 2020.

Babaioff, M., Sharma, Y., and Slivkins, A. Characterizing truthful multi-armed bandit mechanisms. *SIAM J. Comput.*, 43(1):194–230, 2014. doi: 10.1137/120878768. URL https://doi.org/10.1137/120878768.

Bergemann, D. and Valimaki, J. *Efficient dynamic auctions*. Cowles Foundation Discussion Paper, 2006.

Bogunovic, I., Krause, A., and Scarlett, J. Corruption-tolerant gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Braverman, M., Mao, J., Schneider, J., and Weinberg, S. M. Multi-armed bandit problems with strategic arms. *arXiv preprint arXiv:1706.09060*, 2017.

Chen, X., Krishnamurthy, A., and Wang, Y. Robust dynamic assortment optimization in the presence of outlier customers. *arXiv preprint arXiv:1910.04183*, 2019.

Deng, Y., Schneider, J., and Sivan, B. Prior-free dynamic auctions with low regret buyers. In *Advances in Neural Information Processing Systems*, pp. 4804–4814, 2019a.

Deng, Y., Schneider, J., and Sivan, B. Strategizing against no-regret learners. In *Advances in Neural Information Processing Systems*, pp. 1577–1585, 2019b.

Devanur, N. R. and Kakade, S. M. The price of truthfulness for pay-per-click auctions. In *Proceedings of the 10th ACM conference on Electronic commerce*, pp. 99–106, 2009.

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

Feng, Z., Parkes, D. C., and Xu, H. The intrinsic robustness of stochastic bandits to strategic manipulation. In *International Conference on Machine Learning (ICML)*, 2020.

Gatti, N., Lazaric, A., and Trovò, F. A truthful learning mechanism for contextual multi-slot sponsored search auctions with externalities. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pp. 605–622, 2012.

Gupta, A., Koren, T., and Talwar, K. Better algorithms for stochastic bandits with adversarial corruptions. In *Proceedings of the 32nd Conference on Learning Theory (COLT)*, 2019.

Heidari, H., Mahdian, M., Syed, U., Vassilvitskii, S., and Yazdanbod, S. Pricing a low-regret seller. In *International Conference on Machine Learning*, pp. 2559–2567, 2016.

Hoeffding, W. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pp. 409–426. Springer, 1994.

Jun, K.-S., Li, L., Ma, Y., and Zhu, X. Adversarial attacks on stochastic bandits. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS*, pp. 3644–3653, Red Hook, NY, USA, 2018. Curran Associates Inc.

Kakade, S. M., Lobel, I., and Nazerzadeh, H. Optimal dynamic mechanism design and the virtual-pivot mechanism. *Operations Research*, 61(4):837–854, 2013.

Kapoor, S., Patel, K., and Kar, P. Corruption-tolerant bandit learning. *Machine Learning*, 108, 08 2018. doi: 10.1007/s10994-018-5758-5.

Krishnamurthy, A., Lykouris, T., and Podimata, C. Corrupted multidimensional binary search: Learning in the presence of irrational agents, 2020.

Li, Y., Lou, E. Y., and Shan, L. Stochastic linear optimization with adversarial corruption. *arXiv preprint arXiv:1909.02109*, 2019.

Liu, F. and Shroff, N. B. Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning (ICML)*, volume abs/1905.06494, 2019.

Liu, J., Huang, Z., and Wang, X. Learning optimal reserve price against non-myopic bidders. In *Advances in Neural Information Processing Systems*, pp. 2038–2048, 2018.

Lykouris, T., Mirrokni, V., and Paes Leme, R. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th ACM Symposium on Theory of Computing (STOC)*, 2018.

Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. Corruption robust exploration in episodic reinforcement learning, 2019.

Mitzenmacher, M. and Upfal, E. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

Mohri, M. and Munoz, A. Optimal regret minimization in posted-price auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, pp. 1871–1879, 2014.

Mohri, M. and Munoz, A. Revenue optimization against strategic buyers. In *Advances in Neural Information Processing Systems*, pp. 2530–2538, 2015.

Nazerzadeh, H., Saberi, A., and Vohra, R. Dynamic cost-per-action mechanisms and applications to online advertising. In *Proceedings of the 17th international conference on World Wide Web*, pp. 179–188, 2008.

Weed, J., Perchet, V., and Rigollet, P. Online learning in repeated auctions. In *Conference on Learning Theory*, pp. 1562–1583, 2016.

Zimmert, J. and Seldin, Y. An optimal algorithm for stochastic and adversarial bandits. *CoRR*, abs/1807.07623, 2019. URL http://arxiv.org/abs/1807.07623.

Supplementary Material

A. Omitted proofs from Section 3

A.1. Proof of Lemma 2

Proof of Lemma 2. The idea in the proof is similar to (Lykouris et al., 2018). Due to the subsampling of corruption, the smallest robust layer $\ell^\star = \log \hat{C}$ has its confidence interval inside the desired confidence intervals and does not contradict the confidence intervals of higher layers. By definition of the weight for j, w_j^t is less than $\text{UCB}_{j,\ell^\star}^t$. Since we reset any layer that contradicts more robust layers, it is also higher than $\text{LCB}_{j,\ell^\star}^t$. As a result, it also lies within the desired confidence interval. Here are the details.

Let's call layer $\ell^\star = \max\{1, \lceil \log(C/T^{1/3}) \rceil\}$ the *dominant* layer. The probability that $\ell^\star$ is selected in an Explore round is $\log(T) T^{-1/3} 2^{-\ell^\star} \in [\log(T)/2\hat{C}, \log(T)/\hat{C}]$.

Let $\tilde{\rho}_{j;\ell^\star}^t$ be the empirical mean of the clicks received by arm j in layer $\ell^\star$ until round t and $n_{j;\ell^\star}^t$ be the number of times arm j is chosen in layer $\ell^\star$ by then. Using Hoeffding's inequality from Lemma 8, we have that $|\rho_j - \tilde{\rho}_{j;\ell^\star}^t| \leq \sqrt{\frac{\log(T)}{n_{j;\ell^\star}^t}}$ with probability at least $1 - 2/T^2$. Let $\hat{\rho}_{j;\ell^\star}^t$ be the empirical mean of clicks estimated by the layer in presence of corruption. Since the expected number of corruption that this layer experience is at most $\log(T)$, by Chernoff upper tail from Lemma 7, the corruption this layer experience is less than $7\log(T)$ with probability $1 - 1/T^2$ at least. So we have $|\hat{\rho}_{j;\ell^\star}^t \cdot n_{j;\ell^\star}^t - \tilde{\rho}_{j;\ell^\star}^t \cdot n_{j;\ell^\star}^t| \leq 7\log(T)$ with probability $1 - 1/T^2$ at least, which gives us that $|\hat{\rho}_{j;\ell^\star}^t - \tilde{\rho}_{j;\ell^\star}^t| \leq \frac{7\log(T)}{n_{j;\ell^\star}^t}$. Thus, by union bound, $|\rho_j - \hat{\rho}_{j;\ell^\star}^t| \leq \sqrt{\frac{\log(T)}{n_{j;\ell^\star}^t}} + \frac{7\log(T)}{n_{j;\ell^\star}^t}$ with probability at least $1 - 3/T^2$.

For any layer $\ell > \ell^\star$, it receives less corrupted data than $\ell^\star$ in expectation, so $\rho_j \in \left[\hat{\rho}_{j;\ell}^t - \left(\sqrt{\frac{\log(T)}{n_{j;l}^t}} + \frac{7\log(T)}{n_{j;\ell}^t}\right), \hat{\rho}_{j;\ell}^t + \left(\sqrt{\frac{\log(T)}{n_{j;\ell}^t}} + \frac{7\log(T)}{n_{j;\ell}^t}\right)\right]$ with probability at least $1 - \frac{3}{T^2}$. For any layer $\ell < \ell^\star$, it's intervals must have overlap with the dominant interval or they will be reset immediately in the Reset Phase of Algorithm 1.

If all layers $\ell \geq \ell^\star$ contain ρ_j in their interval at round t for all j, the dominant layer will never be reset. Let H be the event $\ell^\star$ is never reset and that for all arms j, ρ_j is inside the interval maintained by layer $\ell^\star$. Using union bound over all $\frac{2}{3}logT$ layers, all K arms and all T rounds, event H happens with probability at least $1 - \frac{6K\log(T)}{3T}$. When event H happens then for all t and arms j,

$$w_j^t = \min_\ell \hat{\rho}_{j;\ell}^t + \left(\sqrt{\frac{\log(T)}{n_{j;\ell}^t}} + \frac{7\log(T)}{n_{j;\ell}^t}\right) \leq \hat{\rho}_{j;\ell^\star}^t + \left(\sqrt{\frac{\log(T)}{n_{j;\ell^\star}^t}} + \frac{7\log(T)}{n_{j;\ell^\star}^t}\right)$$

and for any layer ℓ,

$$\hat{\rho}_{j;\ell}^t + \left(\sqrt{\frac{\log(T)}{n_{j;\ell}^t}} + \frac{7\log(T)}{n_{j;\ell}^t}\right) \geq \hat{\rho}_{j;\ell^\star}^t - \left(\sqrt{\frac{\log(T)}{n_{j;\ell^\star}^t}} + \frac{7\log(T)}{n_{j;\ell^\star}^t}\right)$$

Also, when H happens, for every round t and arm j, $\mathbb{E}[n_{j;\ell^\star}^t] = t\log(T)/T^{-1/3}2^{-\ell^\star}$ in expectation, which is at least $t\log(T)/2\hat{C}$. By Chernoff lower tail from Lemma 7, we have $n_{j;\ell^\star}^t > t/2\hat{C}$ with probability $1 - 1/T^2$ at least when $t > 3\hat{C}\log^2(T)$. Thus for all rounds $t > 3\hat{C}\log^2(T)$ and arms j,

$$w_j^t \in \left[\rho_j - 2\left(\sqrt{\frac{\log(T)}{t/2\hat{C}}} + \frac{7\log(T)}{t/2\hat{C}}\right), \rho_j + 2\left(\sqrt{\frac{\log(T)}{t/2\hat{C}}} + \frac{7\log(T)}{t/2\hat{C}}\right)\right]$$

with probability at least $1 - \frac{8K\log(T)}{3T}$.

□

Note that for $t > 14\hat{C}\log^2(T)$, with probability at least $1 - \frac{8K\log(T)}{3T}$,

$$w_j^t \in \Big[\rho_j - 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}, \rho_j + 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\Big] \tag{1}$$

A.2. Proof of Theorem 3

Proof of Theorem 3. Recall that $\textsc{WelReg}(T) = T\rho_1\mu_1 - \sum_t \rho_{a^t}\mu_{a^t}$ where a^t is the arm picked at round t. Let's divide the $\textsc{WelReg}(T)$ into the regret from exploration rounds $\textsc{WelReg}_{\text{explore}}(T)$ and the regret from exploitation rounds $\textsc{WelReg}_{\text{exploit}}(T)$.

The expected number of explore rounds is $\sum_{t=1}^{T}\sum_{\ell=1}^{\log T^{2/3}} K\log(T)\log T^{-1/3}2^{-\ell} \le K\log(T)T^{2/3}$. Thus $\textsc{WelReg}_{\text{explore}}(T) \le K\log(T)T^{2/3}$.

Define "good" event G such that at every round $t > 14\hat{C}\log^2(T)$, for every arm j, the weight $w_j^t \in [\rho_j - 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}, \rho_j + 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}]$. Using Lemma 2 (in particular Eq. (1)), G happens with probability at least $1 - \frac{8K\log(T)}{3T}$ since all rewards are in $[0,1]$. The regret from rounds where $t < 14\hat{C}\log^2(T)$ can be bound by $14\hat{C}\log^2(T)$, and henceforth we only focus on the rounds $t > 14\hat{C}\log^2(T)$ where event G is defined. Recall that $\Delta = \min_{j>1}\{\frac{\rho_1\mu_1 - p_j\mu_j}{\mu_1+\mu_j}\}$. Assuming G is true, using Eq. (1) and the fact that for $t > \frac{32\hat{C}\log(T)}{\Delta^2}$, it holds that:

$$\begin{aligned} w_1^t\mu_1 &\ge \Big(\rho_1 - 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\Big)\mu_1 \ge \Big(\rho_1 - 4\sqrt{\frac{\log(T)\Delta^2}{32\hat{C}\log T/2\hat{C}}}\Big)\mu_1 = (\rho_1 - \Delta)\mu_1 \\ &\ge \max_{j>1}(\rho_j + \Delta)\mu_j \ge \max_{j>1}\Big(\rho_j + 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\Big)\mu_j \ge \max_{j>1} w_j^t\mu_j \end{aligned}$$

Thus if G is true, then for $t > \frac{32\hat{C}\log(T)}{\Delta^2}$ the first arm will always be picked. For the earlier rounds, if an arm $j \neq 1$ is picked then

$$\begin{aligned} & w_j^t\mu_j \ge w_1^t\mu_1 \\ \implies & \Big(\rho_j + 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\Big)\mu_j \ge \Big(\rho_1 - 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\Big)\mu_1 \end{aligned}$$

Thus regret from this round is

$$\rho_1\mu_1 - \rho_j\mu_j \le 4(\mu_1+\mu_j)\sqrt{\frac{\log(T)}{t/2\hat{C}}} \le 8\sqrt{\frac{\log(T)}{t/2\hat{C}}}$$

The total expected regret from exploit rounds $\textsc{WelReg}_{\text{explore}}(T)$ when the "good" event G occurs can be bound by:

$$\begin{aligned} \textsc{WelReg}_{\text{exploit}|G}(T) &\le \sum_{t}^{\min\{\frac{32\hat{C}\log(T)}{\Delta^2}, T\}} 8\cdot\sqrt{\frac{\log(T)}{t/2\hat{C}}} \\ &\le 16\min\Big\{\frac{8\hat{C}\log(T)}{\Delta}, \sqrt{2CT\log(T)}\Big\} \end{aligned}$$

When good event G doesn't occur, $R_{\text{exploit}|\overline{G}}(T)$ is at most T. We get

$$\begin{aligned} \textsc{WelReg}_{\text{exploit}}(T) &\le Pr\{G\}\cdot 16\min\Big\{\frac{8\hat{C}\log(T)}{\Delta}, \sqrt{2CT\log(T)}\Big\} + Pr\{\overline{G}\}T \\ &\le 16\cdot min\Big\{\frac{8\hat{C}\log(T)}{\Delta}, \sqrt{2CT\log(T)}\Big\} + \frac{8K\log(T)}{3} \end{aligned} \tag{2}$$

Adding $\textsc{WelReg}_{\text{exploit}}(T)$ and $\textsc{WelReg}_{\text{exploit}}(T)$ and $14\hat{C}\log^2(T)$ completes the proof.

□

A.3. Proof of Theorem 4

Proof of Theorem 4. Recall that the revenue regret $\textsc{RevReg}(T) = T\rho_2\mu_2 - \sum_t P_t\rho_{a^t}$. Similar to the proof for Theorem 3, we divide $\textsc{RevReg}(T)$ into the exploration regret $\textsc{RevReg}_{\text{explore}}(T)$ and the exploitation regret $\textsc{RevReg}_{\text{exploit}}(T)$. Using similar arguments as Theorem 3, $\textsc{RevReg}_{\text{explore}}(T) \le K\log(T)T^{2/3}$

The arm chosen in an exploit round t is $a^t = \arg\max_j w_j^t\mu_j$. Recall that smax denotes the second maximum. The expected revenue from this round is then $p^t\rho_{a^t} = \frac{\text{smax}_j w_j^t\mu_j}{w_{a^t}^t}\cdot\rho_{a^t}$. As in Theorem 3, we define "good" event G that at every round $t > 14\hat{C}\log^2(T)$ and every arm j, the weight $w_j^t \in \left[\rho_j - 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}, \rho_j + 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\right]$. The regret from rounds $t < 14\hat{C}\log^2(T)$ can be bound by $14\hat{C}\log^2(T)$, and henceforth we only focus on rounds $t > 14\hat{C}\log^2(T)$ where event G is defined.

Denote $t^\star = \left(\frac{4}{\rho_1}\sqrt{2log(T)\hat{C}} + 1\right)^2$. If G happens and the best arm is not picked, i.e. $a^t \neq 1$, then the regret from exploit rounds can be bounded as

$$
\begin{aligned}
\textsc{RevReg}_{\text{exploit}|G,a^t>1}(T) &= \sum_{t:a^t>1}\rho_2\mu_2 - \frac{\text{smax}_j(w_j^t\mu_j)}{w_{a^t}^t}\cdot\rho_{a^t} \\
&= \sum_{t:a^t>1}\frac{\text{smax}_j(w_j^t\mu_j)}{w_{a^t}^t}\Big(\frac{\rho_2\mu_2}{\text{smax}_j(w_j^t\mu_j)}w_{a^t}^t - \rho_{a^t}\Big) \\
&\le \sum_{t:a^t>1}\mu_{a^t}\Big(\frac{\rho_2\mu_2}{\text{smax}_j(w_j^t\mu_j)}w_{a^t}^t - \rho_{a^t}\Big) \\
&\le \sum_{t:a^t>1}\mu_{a^t}\Big(\frac{\rho_2\mu_2}{\text{smax}_{j\in[2]}(w_j^t\mu_j)}w_{a^t}^t - \rho_{a^t}\Big) \\
&\le t^\star + \sum_{t>t^\star,t:a^t>1}\mu_{a^t}\Big(\frac{\rho_2\mu_2}{\rho_1\mu_1 - 4\mu_1\sqrt{\frac{\log(T)}{t/2\hat{C}}}}(\rho_{a^t} + 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}) - \rho_{a^t}\Big) \\
&\le t^\star + \sum_{t>t^\star,t:a^t>1}\mu_{a^t}\rho_{a^t}\Big(\frac{1 + \frac{4}{\rho_{a^t}}\sqrt{\frac{\log(T)}{t/2\hat{C}}}}{1 - \frac{4}{\rho_1}\sqrt{\frac{\log(T)}{t/2\hat{C}}}} - 1\Big) \\
&\le t^\star + \sum_{t>t^\star,t:a^t>1}4\mu_{a^t}\rho_{a^t}\frac{\left(\frac{1}{\rho_1} + \frac{1}{\rho_{a^t}}\right)\sqrt{\frac{\log(T)}{t/2\hat{C}}}}{1 - \frac{4}{\rho_1}\sqrt{\frac{\log(T)}{t/2\hat{C}}}} \\
&\le t^\star + 8\sum_{t>t^\star,t:a^t>1}\frac{\sqrt{\frac{\log(T)}{t/2\hat{C}}}}{1 - \frac{4}{\rho_1}\sqrt{\frac{\log(T)}{t/2\hat{C}}}} \\
&\le \Big(\frac{4}{\rho_1}\sqrt{2log(T)\hat{C}} + 1\Big)^2 + 16\sqrt{2\log(T)T\hat{C}} + \frac{64\log^2(T)\hat{C}}{\rho_1}
\end{aligned}
$$

If G happens and the best arm is indeed picked, i.e. $a^t = 1$, then regret from exploit rounds can be bounded as

$$
\begin{aligned}
\textsc{RevReg}_{\text{exploit}|G,a^t=1}(T) &= \sum_{t:a^t=1}\rho_2\mu_2 - \frac{\text{smax}_j(w_j^t\mu_j)}{w_{a^t}^t}\cdot\rho_{a^t} \\
&\le \sum_{t:a^t=1}\rho_2\mu_2 - \frac{w_2^t\mu_2}{w_1^t}\cdot\rho_1 \\
&\le \sum_{t:a^t=1}\rho_2\mu_2 - \frac{\left(\rho_2 - 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\right)\mu_2}{\rho_1 + 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}}\cdot\rho_1
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{t:a^t=1} \rho_2\mu_2\Big(1 - \frac{1-\frac{4}{\rho_2}\sqrt{\frac{\log(T)}{t/2\hat{C}}}}{1+\frac{4}{\rho_1}\sqrt{\frac{\log(T)}{t/2\hat{C}}}}\Big)\\
&\leq \sum_{t:a^t=1} 4\rho_2\mu_2 \frac{\Big(\frac{1}{\rho_1}+\frac{1}{\rho_2}\Big)\sqrt{\frac{\log(T)}{t/2\hat{C}}}}{1+\frac{4}{\rho_1}\sqrt{\frac{\log(T)}{t/2\hat{C}}}}\\
&\leq 8\sum_{t:a^t=1}\sqrt{\frac{\log(T)}{t/2\hat{C}}}\\
&\leq 16\sqrt{2\log(T)T\hat{C}}
\end{aligned}
$$

Hence $\textsc{RevReg}_{\text{exploit}}(T)$ can be bounded as

$$
\begin{aligned}
\textsc{RevReg}_{\text{exploit}}(T) &\leq Pr\{\overline{G}\}T + Pr\{G\}R_{\text{exploit}|G}(T)\\
&\leq \frac{8K\log(T)}{3} + 32\sqrt{2\log(T)T\hat{C}} + \Big(\frac{4}{\rho_1}\sqrt{2log(T)\hat{C}}+1\Big)^2 + \frac{64\log^2(T)\hat{C}}{\rho_1}
\end{aligned}
\tag{3}
$$

Adding all the terms, we get

$$
\begin{aligned}
\textsc{RevReg}(T) \leq& 14\hat{C}\log^2(T) + K\log(T)T^{2/3} + 32\sqrt{2\log(T)T\hat{C}} +\\
&\Big(\frac{4}{\rho_1}\sqrt{2\log(T)\hat{C}}+1\Big)^2 + \frac{64\log^2(T)\hat{C}}{\rho_1} + \frac{8K\log(T)}{3}
\end{aligned}
$$

□

B. Omitted Proofs From Section 4

B.1. Proof of Theorem 5

Proof of Theorem 5. The proof is similar to the proof for Theorem 3. The only difference is that arms cannot bid truthfully, but bid in a (γ,δ)-confidence truthful manner. Hence the arm j bids $b_j^t \in \Big[\mu_j - 4\gamma\sqrt{\frac{\log(T)}{t/2\hat{C}}}, \mu_j + 4\gamma\sqrt{\frac{\log(T)}{t/2\hat{C}}}\Big]$ for all but δT rounds. Recall that $\textsc{WelReg}(T) = T\rho_1\mu_1 - \sum_t \rho_{a^t}\mu_{a^t}$ where a^t is the arm picked at round t. First let's consider the "extreme" rounds when the arms bid outside their confidence interval. There can be at most $K\delta T$ and thus they contribute at most $K\delta T$ to the regret. For the rest of the proof, we will assume that each arm j bids $b_j^t \in \Big[\mu_j - 4\gamma\sqrt{\frac{\log(T)}{t/2\hat{C}}}, \mu_j + 4\gamma\sqrt{\frac{\log(T)}{t/\hat{2}C}}\Big]$ for every round t.

Similar to the proof for Theorem 3, we divide $\textsc{WelReg}(T)$ into $\textsc{WelReg}_{\text{explore}}(T)$ and $\textsc{WelReg}_{\text{exploit}}(T)$. We can show that $\textsc{WelReg}_{\text{explore}}(T) \leq K\log(T)T^{2/3}$. We again use the definition of the "good" event G that at every round $t > 14\hat{C}\log^2(T)$ and every arm j, the weight $w_j^t \in \Big[\rho_j - 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}, \rho_j + 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\Big]$. The regret from rounds where $t < 14\hat{C}\log^2(T)$ can be bound by $14\hat{C}\log^2(T)$, and henceforth we only focus on the rounds $t > 14\hat{C}\log^2(T)$ where event G is defined.

Recall that $\Delta = \min_{j>1}\frac{p_1\mu_1 - p_j\mu_j}{\mu_1+\mu_j+\gamma(4+\sqrt{\rho_1}+\sqrt{\rho_j})}$. When G is true, then for $t > \frac{32\log(T)\hat{C}}{\Delta^2}$ we have

$$
\begin{aligned}
w_1^t b_1^t &\geq \Big(\rho_1 - 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\Big)\Big(\mu_1 - 4\gamma\sqrt{\frac{\log(T)}{\rho_1 t/2\hat{C}}}\Big)\\
&\geq \max_{j>1}\Big(\rho_j + 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\Big)\Big(\mu_j + 4\gamma\sqrt{\frac{\log(T)}{\rho_j t/2\hat{C}}}\Big) \geq \max_{j>1} w_j^t b_j^t
\end{aligned}
$$

Thus when G is true, for $t > \frac{32\log(T)\hat{C}}{\Delta^2}$, the best arm is always picked and we incur no welfare regret for these rounds. For earlier rounds, if an arm $j \neq 1$ is picked then,

$$w_j^t b_j^t \geq w_1^t b_1^t$$
$$\implies \Big(\rho_j + 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\Big)\Big(\mu_j + 4\gamma\sqrt{\frac{\log(T)}{\rho_j t/2\hat{C}}}\Big) \geq \Big(\rho_1 - 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\Big)\Big(\mu_1 - 4\gamma\sqrt{\frac{\log(T)}{\rho_1 t/2\hat{C}}}\Big)$$

Thus regret from this round is

$$\rho_1\mu_1 - \rho_j\mu_j \leq 4\Big(\mu_j + \mu_1 + \gamma(4 + \sqrt{\rho_j} + \sqrt{\rho_1})\Big)\sqrt{2\log(T)\hat{C}/t} \leq 8(1+3\gamma)\sqrt{2\log(T)\hat{C}/t}$$

where the last inequality follows from the fact that μ_j, μ_1, ρ_1, and ρ_j are bounded between 0 and 1.

Combining these cases gives us that

$$\begin{aligned}\textsc{WelReg}_{\text{exploit}|G}(T) &\leq \sum_{t}^{\min\{\frac{32\log(T)\hat{C}}{\Delta^2}, T\}} 8\Big(1+3\gamma\Big)\sqrt{2\log(T)\hat{C}/t} \\ &\leq 16\Big(1+3\gamma\Big)\min\Big\{\frac{8\log(T)\hat{C}}{\Delta}, \sqrt{2\log(T)\hat{C}T}\Big\}\end{aligned}$$

When G is not true, the regret $\textsc{WelReg}_{\text{exploit}|\overline{G}}(T)$ is at most T Combining these regrets with the regret from the "extreme" rounds, we get

$$\begin{aligned}\textsc{WelReg}(T) &\leq \delta KT + \textsc{WelReg}_{\text{explore}}(T) + Pr\{G\}\textsc{WelReg}_{\text{exploit}|G}(T) + Pr\{\overline{G}\}\textsc{WelReg}_{\text{exploit}|\bar{G}}(T) \\ &\leq K\Big(\delta T + \log(T)T^{2/3}\Big) + 16\Big(1+3\gamma\Big)\min\Big\{\frac{8\log(T)\hat{C}}{\Delta}, \sqrt{2\log(T)\hat{C}T}\Big\} + \frac{8K\log(T)}{3}\end{aligned}$$

□

B.2. Proof of Theorem 6

Proof of Theorem 6. Recall that the welfare regret $\textsc{RevReg}(T) = T\rho_2\mu_2 - \sum_t P_t\rho_{a^t}$ where a^t is the arm picked at round t. Similar to the proof for Theorem 5, we divide $\textsc{RevReg}(T)$ into the exploration regret $\textsc{RevReg}_{\text{explore}}(T)$ and the exploitation regret $\textsc{RevReg}_{\text{exploit}}(T)$. Using similar arguments as Theorem 3, $\textsc{RevReg}_{\text{explore}}(T) \leq K(\log(T)T^{2/3} + \delta T)$

The proof for the exploit rounds follows similarly to Theorem 4. The arm chosen in round t, $a^t = \arg\max_j w_j^t b_j^t$ and revenue is $p^t \cdot \rho_{a^t} = \frac{\text{smax}_j\, w_j^t b_j^t}{w_{a^t}^t} \cdot \rho_{a^t}$. We again use the definition of the "good" event G that at every round $t > 14\hat{C}\log^2(T)$ and every arm j, the weight $w_j^t \in \Big[\rho_j - 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}, \rho_j + 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\Big]$. The regret from rounds where $t < 14\hat{C}\log^2(T)$ can be bound by $14\hat{C}\log^2(T)$, and henceforth we only focus on the rounds $t > 14\hat{C}\log^2(T)$ where event G is defined.

There are two possible cases, the best arm is chosen i.e. $a^t = 1$ or the non-optimal arm is chosen, i.e. $a^t > 1$. When G is true, let $t^\star = \Big(4\sqrt{2\hat{C}\log(T)}(\frac{1}{\rho_1} + \frac{1}{\sqrt{\rho_1\mu_1}}) + 1\Big)^2$.

$$\begin{aligned}\textsc{RevReg}_{Exploit|G,a^t>1}(T) &= \sum_{t:a^t>1} \rho_2\mu_2 - \frac{\text{smax}_j(w_j^t b_j^t)}{w_{a^t}^t} \cdot \rho_{a^t} \\ &= \sum_{t:a^t>1} \frac{\text{smax}_j(w_j^t b_j^t)}{w_{a^t}^t}\Big(\frac{\rho_2\mu_2}{\text{smax}_j(w_j^t b_j^t)} w_{a^t}^t - \rho_{a^t}\Big) \\ &\leq \sum_{t:a^t>1} b_{a^t}^t\Big(\frac{\rho_2\mu_2}{\text{smax}_j(w_j^t b_j^t)} w_{a^t}^t - \rho_{a^t}\Big) \\ &\leq \sum_{t:a^t>1} b_{a^t}^t\Big(\frac{\rho_2\mu_2}{\text{smax}_{j\in[2]}(w_j^t b_j^t)} w_{a^t}^t - \rho_{a^t}\Big)\end{aligned}$$

$$\leq t^\star + \sum_{t>t^\star, t:a^t>1} \Big(\mu_{a^t} + 4\gamma\sqrt{\frac{log(T)}{\rho_{a^t} t/2\hat{C}}}\Big) \cdot \Big(\frac{\rho_2\mu_2}{(\rho_1 - 4\sqrt{\frac{\log(T)}{t/2\hat{C}}})(\mu_1 - 4\gamma\sqrt{\frac{\log(T)}{\rho_1 t/2\hat{C}}})}\Big(\rho_{a^t} + 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\Big) - \rho_{a^t}\Big)$$

$$\leq t^\star + \sum_{t>t^\star, t:a^t>1} 4\Big(\mu_{a^t} + \frac{\mu_{a^t}\rho_{a^t}}{\rho_1} + \frac{\gamma\rho_{a^t}\mu_{a^t}}{\sqrt{\rho_1}\mu_1}\Big)\frac{\sqrt{\frac{\log(T)}{t/2\hat{C}}}}{1 - 4(\frac{1}{\rho_1} + \frac{\gamma}{\sqrt{\rho_1}\mu_1})\sqrt{\frac{\log(T)}{t/2\hat{C}}}} +$$

$$16\Big(1 + \frac{\rho_{a^t}}{\rho_1} + \frac{\gamma\rho_{a^t}}{\sqrt{\rho_1}\mu_1}\Big)\frac{log(T)2\hat{C}}{\sqrt{\rho_1}}\frac{1}{t - 4(\frac{1}{\rho_1} + \frac{\gamma}{\sqrt{\rho_1}\mu_1})\sqrt{log(T)2\hat{C}t}}$$

$$\leq \Big(8\frac{\sqrt{2\hat{C}\log(T)}}{\rho_1\mu_1} + 1\Big)^2 + 8(2+\gamma)\sqrt{2\log(T)\hat{C}T} + \frac{96(\gamma^2+2\gamma+2)}{(\rho_1\mu_1)^{3/2}}log^2(T)\hat{C}$$

For the case when $a^t = 1$, using that $\rho_2\mu_2 \leq \rho_1\mu_1$ and also $\mu_1, \mu_2 \leq 1$, similar to before it holds that:

$$\begin{aligned}
\textsc{RevReg}_{Exploit|G,a^t=1}(T) &= \sum_{t:a^t=1} \rho_2\mu_2 - \frac{\text{smax}_j(w_j^t b_j^t)}{w_{a^t}^t}\cdot\rho_{a^t} \\
&\leq \sum_{t:a^t=1} \rho_2\mu_2 - \frac{w_2^t b_2^t}{w_1^t}\cdot\rho_1 \\
&\leq \sum_{t:a^t=1} \rho_2\mu_2 - \frac{\Big(\rho_2 - 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}\Big)\Big(\mu_2 - 4\gamma\sqrt{\frac{\log(T)}{\rho_2 t/2\hat{C}}}\Big)}{\rho_1 + 4\sqrt{\frac{\log(T)}{t/2\hat{C}}}}\cdot\rho_1 \\
&\leq \sum_{t:a^t=1} \rho_2\mu_2\Big(1 - \frac{(1 - \frac{4}{\rho_2}\sqrt{\frac{\log(T)}{t/2\hat{C}}})(1 - \frac{4\gamma}{\sqrt{\rho_2}}\sqrt{\frac{\log(T)}{t/2\hat{C}}})}{1 + \frac{4}{\rho_1}\sqrt{\frac{\log(T)}{t/2\hat{C}}}}\Big) \\
&\leq \sum_{t:a^t=1} \rho_2\mu_2\frac{4\Big(\frac{1}{\rho_1} + \frac{1}{\rho_2} + \frac{\gamma}{\sqrt{\rho_2}}\Big)\sqrt{\frac{\log(T)}{t/2\hat{C}}}}{1 + \frac{4}{\rho_1}\sqrt{\frac{\log(T)}{t/2\hat{C}}}} \\
&\leq 8(2+\gamma)\sqrt{2\log(T)\hat{C}T}
\end{aligned}$$

So the total regret is bound by

$$\begin{aligned}
\textsc{RevReg}(T) &\leq 14\hat{C}\log^2(T) + K(\log(T)T^{2/3} + \delta T) + Pr\{\overline{G}\}T + Pr\{G\}\textsc{RevReg}_{Exploit|G}(T) \\
&\leq 14\hat{C}\log^2(T) + K\Big(\log(T)T^{2/3} + \delta T\Big) + \frac{8K\log(T)}{3} + \Big(8\frac{\sqrt{2\hat{C}\log(T)}}{\rho_1\mu_1} + 1\Big)^2 + \\
&16(2+\gamma)\sqrt{2\log(T)\hat{C}T} + \frac{96\Big(\gamma^2+2\gamma+2\Big)}{(\rho_1\mu_1)^{3/2}}log^2(T)\hat{C}
\end{aligned}$$

□

C. Auxiliary lemmas

Lemma 7 (Chernoff Bounds)**.** *Let $X_1, \ldots, X_n$ be independent random variables, and X_i lies in the interval $[0,1]$. Define $X = \sum_{i=1}^n X_i$ and denote $E[X] = \mu$. For any $\delta \in [0,1]$, we have **Chernoff lower tail:***

$$Pr\{X < (1-\delta)\mu\} \leq \exp(-\frac{\mu\delta^2}{3})$$

and we have ***Chernoff upper tail:***

$$Pr\{X > (1+\delta)\mu\} \leq \begin{cases} \exp(-\frac{\mu\delta}{3}) & \textbf{for } \delta > 1 \\ \exp(-\frac{\mu\delta^2}{3}) & \textbf{for } \delta \in [0,1] \end{cases}$$

The proofs for the inequalities in Lemma 7 can be found in Theorem 4.4 and Theorem 4.5 of (Mitzenmacher & Upfal, 2017)

Lemma 8 (Hoeffding's Inequality (Hoeffding, 1994))**.** *Let $X_1, \ldots, X_n$ be independent random variables, and X_i lies in the interval $[0,1]$. Define $X = \sum_{i=1}^{n} X_i$ and denote $E[X] = \mu$. For X and any $t > 0$, Hoeffding inequality gives*

$$Pr\{|X - \mu| > t\} \leq 2\exp(-2t^2/n)$$