Non-stochastic Budgeted Online Pricing with Semi-Bandit Feedback

Xiang Liu[1, 2*], **Hau Chan**[3], **Minming Li**[4], **Weiwei Wu**[1], **Long Tran-Thanh**[5]
[1]Southeast University
[2]The Chinese University of Hong Kong
[3]University of Nebraska-Lincoln
[4]City University of Hong Kong
[5]The University of Warwick
xiangliu@seu.edu.cn, hchan3@unl.edu, minming.li@cityu.edu.hk, weiweiwu@seu.edu.cn, long.tran-thanh@warwick.ac.uk

Abstract

We consider a general non-stochastic online pricing bandit setting in a procurement scenario where a buyer with a budget wants to procure items from a fixed set of sellers to maximize the buyer's reward by dynamically offering purchasing prices to the sellers, where the sellers' costs and values at each time period can change arbitrarily and the sellers determine whether to accept the offered prices to sell the items. This setting models online pricing scenarios of procuring resources or services in multi-agent systems. We first consider the offline setting when sellers' costs and values are known in advance and investigate the best fixed-price policy in hindsight. We show that it has a tight approximation guarantee with respect to the offline optimal solutions. In the general online setting, we propose an online pricing policy, Granularity-based Pricing (GAP), which exploits underlying side-information from the feedback graph when the budget is given as the input. We show that GAP achieves an upper bound of $O(n\frac{v_{max}}{c_{min}}\sqrt{B/c_{min}}\ln B)$ on the α-regret where n, v_{max}, c_{min}, and B are the number, the maximum value, the minimum cost of sellers, and the budget, respectively. We then extend it to the unknown budget case by developing a variant of GAP, namely Doubling-GAP, and show its α-regret is at most $O(n\frac{v_{max}}{c_{min}}\sqrt{B/c_{min}}\ln^2 B)$. We also provide an α-regret lower bound $\Omega(v_{max}\sqrt{Bn/c_{min}})$ of any online policy that is tight up to sub-linear terms. We conduct simulation experiments to show that the proposed policy outperforms the baseline algorithms.

1 Introduction

In many online multi-agent procurement situations (Badanidiyuru, Kleinberg, and Singer 2012; Singer and Mittal 2013; Balkanski and Hartline 2016; Jain et al. 2016), a buyer is often interested in procuring some subset of resources or services from sellers over a period of time by offering sellers certain prices (or payments) to maximize the buyer's cumulative reward without exceeding their budget on the total payment. For instance, in online crowdsourcing platforms (Ho and Vaughan 2012), a group of workers (*i.e.*, sellers) arrive over time to provide services (*e.g.* working on tasks such as labeling images or collecting sensing data). The requester (*i.e.*, buyer) with a budget can submit bids (*e.g.*, prices for workers) to recruit workers to perform tasks for the purpose of maximizing the total value of recruited workers (*e.g.*, the quality of labels) of finished tasks (Singer and Mittal 2013). In addition, in online advertising markets (Ha 2008), website publishers (*i.e.*, sellers) register advertising placements for sale over time. The advertiser (*i.e.*, the buyer, who naturally has a limited operational budget within a fixed period (Lee, Jalali, and Dasdan 2013)) needs to bid (*e.g.*, through the generalized second price auction) to procure these placements to display their advertisements (Yuan, Wang, and Zhao 2013). If the advertiser wins the bidding auction, they can obtain impressions (*i.e.*, the value of the placement) such as advertising exposure, and the reward of the advertiser can be the cumulative impression in display advertising.

Our Challenges. These problems are often known as the *online procurement pricing problem* (as highlighted in the above-mentioned situations) where a buyer with a limited budget aims to determine what prices to offer for sellers' resources or services at each period to optimize the buyer's reward. While these problems have been studied (Badanidiyuru, Kleinberg, and Singer 2012; Singla and Krause 2013; Balkanski and Hartline 2016), existing studies have two main limitations that make previous results not applicable to many real-world situations.

First, most of the previous work in online procurement pricing focuses on the stochastic setting that assumes that sellers' costs (or accepted prices) and rewards for the buyer typically follow some (unknown) underlying stationary distribution (Badanidiyuru, Kleinberg, and Singer 2012; Singla and Krause 2013; Balkanski and Hartline 2016). However, this assumption does not hold in many real-world situations because costs and rewards can change drastically due to external shocks (*e.g.*, market crashes, inflation) and may not always obey any fixed distribution. For example, in crowdsourcing platforms, workers' accepted prices and performance qualities can vary largely depending on factors such as task preferences, tastes, experiences, and expertise. All of these can affect the prices and quality unpredictably in a non-stochastic manner (Jagabathula, Subramanian, and Venkataraman 2017). In online advertising, the winning price of each advertising placement can be arbitrary (*e.g.*, depending on bidding strategies of other advertisers), and the corresponding gained impression is uncertain, *e.g.*,

*Corresponding author.

clicks of advertisements rely on variable preferences and individual behaviors of website users (Wilbur and Zhu 2009).

Second, existing studies often consider offering a price to a single seller at every time period, which is unrealistic for many applications. For instance, in crowdsourcing platforms, multiple workers are available to provide services during every time slot. Similarly, in online advertising, the advertiser can also bid for multiple advertising placements on every webpage simultaneously.

In this paper, to address these two challenges, we consider the non-stochastic setting of budgeted online procurement pricing problems, where the costs (or accepted prices) and values of the sellers vary arbitrarily over time. Specially, this non-stochastic setting encompasses various general scenarios, including adversarial and non-stationary settings.

Main Contributions

We propose and consider the non-stochastic budgeted online procurement pricing problem, in which a budget-limited buyer chooses an appropriate price from a continuous price space for each seller without information about sellers' costs and values at each time step. Given the offered prices, each seller makes the decision to either accept or reject the offer.

Facing sellers' unknown information, we naturally apply the widely known framework of (combinatorial) multi-armed bandit method (Auer, Cesa-Bianchi, and Fischer 2002) where we are able to choose a price vector for multiple sellers (set a price for each seller) at each time period, *i.e.*, pull a set of arms in the bandit, and we can observe the outcome from sellers due to the selected prices. Our main contributions can be summarized as follows:

1. **Approximation ratio of the offline oracle**: We first consider the offline setting where sellers' values and costs in the full horizon are known in advance, and investigate the best fixed-price policy under the discrete price set with granularity ϵ which decides a fixed-price in hindsight. We show that it achieves $(2 - c_{min})(1 + \frac{\epsilon}{c_{min}})\frac{v_{max}}{v_{min}}$-approximation to the offline optimal solutions, where v_{max} (v_{min}) is the maximum (minimum) value of sellers and c_{min} is the minimum cost of sellers. Moreover, any fixed-price policy has a matching lower bound $\Omega(\frac{v_{max}}{v_{min}})$.
2. **Tight regret bound for the online setting**: For the general online setting, we design a Granularity-based Pricing (GAP) policy when the budget B is given as the input. In particular, we show that GAP can provably achieve an upper bound of $O(n\frac{v_{max}}{c_{min}}\sqrt{B/c_{min}}\ln B)$ on the α-regret where n is the number of sellers and $\alpha = \frac{v_{min}}{(2-c_{min})v_{max}}$. We then extend it to the case of the unknown budget by developing a variant of GAP, namely Doubling-GAP, which runs GAP over multiple phases by applying the doubling method. We prove that Doubling-GAP achieves an upper bound of $O(n\frac{v_{max}}{c_{min}}\sqrt{B/c_{min}}\ln^2 B)$ on the α-regret. Furthermore, we provide a tight lower bound up to sub-linear terms, *i.e.*, the α-regret for any online policy for the non-stochastic budgeted pricing setting should be at least $\Omega(v_{max}\sqrt{Bn/c_{min}})$.
3. **Numerical validation**: We conduct numerical simulations in the non-stochastic environment. Our results show that our proposed policy significantly outperforms baseline pricing algorithms in terms of the cumulative reward.

Related Work

Procurement Pricing. In the pricing literature, many works focus on seller-centric markets where the seller learns the optimal price to sell items to buyers to maximize the revenue (the achieved total payment from buyers), *e.g.*, (Misra, Schwartz, and Abernethy 2019; Romano et al. 2021; Feldman et al. 2016). In this paper, we focus on the procurement pricing problem under the budget constraint in the reverse markets, where the buyer needs to procure items/services from sellers by offering prices. This body of work can be categorized as follows:

(1) *Procurement Pricing with Known Information:* Relevant works typically focus on designing pricing mechanisms under known stochastic costs and values, and achieves constant approximation ratios to the optimal solution, under some restrictive conditions, *e.g.*, subadditive functions (Bei et al. 2012), independent identical distributed sellers (Anari, Goel, and Nikzad 2014), and independent identical distributed sellers with submodular functions (Balkanski and Hartline 2016).

Apart from the above pricing mechanism design problem, many works consider the online version where agents may arrive online, and the decision makers should make decisions immediately under the strong assumption that the decision-makers have access to the perfect information of agents (Liu et al. 2024). For example, (Dütting and Kesselheim 2019) further take into account the inaccurate prior beliefs in the posted-pricing design, *i.e.*, the prices are chosen with respect to similar but different probability distributions. For a more thorough coverage of these works, please see (Singer and Mittal 2011; Badanidiyuru, Kleinberg, and Singer 2012; Singer and Mittal 2013; Zhao, Li, and Ma 2014; Amanatidis, Kleer, and Schäfer 2022). Since there is no prior knowledge available in our setting, the above-mentioned body of work cannot directly address our problem.

(2) *Procurement Pricing with Unknown Information.* Facing the uncertainty caused by the unknown information of sellers, *e.g.*, unknown costs or values, many works study the learning-based algorithms for stochastic environments with unknown distributions. For example, (Singla and Krause 2013) propose a UCB-based posted pricing mechanism for the budgeted procurement in stochastic online settings. (Hu and Zhang 2017) introduce an optimal algorithm to recruit workers by exploiting the unique features of micro-task crowdsourcing. However, in our model, we consider the non-stochastic online procurement pricing problem involving a diverse set of sellers, each with varying arbitrary costs and values over time. Although (Avadhanula et al. 2021) also consider the budgeted procurement problems, their payment of procuring the item is determined by the seller and derived from the unknown stationary distribution. In contrast, our model focuses on the pricing problem, *i.e.*, the payment to the seller is the price offered by the buyer, and the

non-stochastic setting. Consequently, their techniques cannot be used to address our problem.

Comparison with Existing Works. Closest to our work are settings investigated by (Kleinberg and Leighton 2003; Weed, Perchet, and Rigollet 2016), which also deal with non-stochastic pricing problems. (Kleinberg and Leighton 2003) focus on seller-centric markets (selling items to buyers) and assume that the seller's bid, if successful, will be incorporated into the payoff as a revenue loss. Under this assumption, they show that $O(T^{\frac{2}{3}})$ is the best possible regret bound. Our model, on the other hand, separates the loss induced by bids from rewards (as the former is controlled by a bidding budget B in our setting, and our reward is binary). The rationale behind our setting is that we consider reverse scenarios, *i.e.*, buyer-centric markets (or procurement markets) where the buyer sets prices to procure sellers' items or services. Different from the pricing problem in seller-centric markets, the buyer in procurement markets wants to learn the optimal price to improve the total obtained value of items rather than the revenue. In such procurement scenarios, it is natural that the buyer faces a budget constraint while maximizing the procured value of items (Singer and Mittal 2011). Such difference in our model allows us to achieve an $O(\sqrt{B})$ regret bound (so the exponent of the main term in the bound is $1/2$ instead of $2/3$). It is worth noting that while (Weed, Perchet, and Rigollet 2016) also addresses the online procurement problem from the buyer's perspective with a squared-root regret bound, they do not consider the budget constraints (and they also incorporate the bid into the payoff as a revenue loss). In addition, they assume that the algorithm always possesses knowledge of the accepted price at each round. However, in our problem, this price becomes unknown when the seller rejects the posted price. Thus, their techniques cannot be applied to our setting.

As we tackle this online pricing problem as a bandit model, many bandit algorithms can be considered to be related to our approach, ranging from non-stochastic (or adversarial) bandits (Auer et al. 2002) and combinatorial (semi) bandits (Combes et al. 2015; Neu and Bartók 2016) to budgeted bandits (Tran-Thanh et al. 2012; Badanidiyuru, Kleinberg, and Slivkins 2018) and bandits with feedback graph (Alon et al. 2015, 2017) (for more comprehensive coverage of bandit models, we refer the reader to (Lattimore and Szepesvári 2020)). Among these, the bandit with feedback graph model is the most relevant one. However, they are not designed for cases with budgets and, thus, will fail to perform in our setting. On the other hand, while budgeted bandits can deal with budget limits, they cannot deal with the combinatorial and non-stochastic nature of our setting. In particular, they either fail to yield meaningful regrets due to the combinatorial nature of our problem (Rangi, Franceschetti, and Tran-Thanh 2019; Immorlica et al. 2019), or they are designed for stochastic setting only (Sankararaman and Slivkins 2018).

Therefore, we propose a new bandit model that tackles the combinatorial and non-stochastic setting of the budgeted online pricing, utilizing a special feedback graph structure based on the monotonicity of sellers' acceptance functions.

2 Preliminaries

Non-stochastic Pricing Bandits. In this paper, we consider a procurement scenario where a buyer, with a budget B, wants to procure items from sellers. There are n fixed sellers $S = \{s_1, s_2, ..., s_n\}$ where s_i is the i-th seller. At every time slot t, any seller s_i holds an item with a private cost $c_t^i \in [c_{min}, 1]$ (only known by themselves) and a value $v_t^i \in [v_{min}, v_{max}]$ with $v_{min} \leq v_{max}$ for the buyer[1]. Given the continuous price space $A_{con} = [c_{min}, 1]$, we focus on a pricing problem that at every time slot t, the buyer chooses a price vector for n sellers $\boldsymbol{e}_t = (e_t^1, e_t^2, ..., e_t^n)$ where $e_t^i \in A_{con}$ is the price offered to seller s_i. If the cost of seller s_i at time slot t is not higher than the given price e_t^i (*i.e.*, $c_t^i \leq e_t^i$), then s_i accepts to sell their item with the price e_t^i and the buyer obtains value v_t^i after paying s_i price e_t^i. In our model, we consider the *non-stochastic* setting where the cost c_t^i and the value v_t^i for any $s_i \in S$ at each time slot t can be arbitrary.

Semi-bandit Feedback. The above model can be interpreted as a bandit setting, where the buyer chooses candidate prices from a continuous range for sellers, which are considered as combinatorial arms to be pulled at each time slot, and the reward is the semi-bandit feedback observed from sellers' decisions. Note that, as sellers' costs are private information, at time slot t, the buyer can only observe the binary decision (accept or reject) of seller s_i under the given price e_t^i, and the corresponding value v_t^i can be observed only if s_i accepts the price e_t^i. Let $x_t^i(e_t^i) = 1$ indicate that s_i accepts the price e_t^i at time slot t, otherwise $x_t^i(e_t^i) = 0$.

The Objective. Note that $\tau(\mathcal{A})$ depends on the budget B of the buyer and $\boldsymbol{e}_t$ is the chosen price vector at time t. Let $G(\mathcal{A})$ denote the expected cumulative reward of policy $\mathcal{A}$, *i.e.*, $G(\mathcal{A}) = \mathbb{E}\left[\sum_{t \leq \tau(\mathcal{A})} \sum_{i \leq n} v_t^i \cdot x_t^i(e_t^i)\right]$, where the expectation is over all possible randomization coming from policy $\mathcal{A}$. The objective is to design an online pricing policy $\mathcal{A}$ such that $G(\mathcal{A})$ is maximized as follows,

$$\max_{(\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_{\tau(\mathcal{A})})} G(\mathcal{A}),\ s.t.,\ \sum_{t=1}^{\tau(\mathcal{A})} \sum_{i=1}^{n} e_t^i \cdot x_t^i(e_t^i) \leq B.$$

Dynamic Regrets. Let $\mathcal{A}^*$ denote the optimal policy for the above optimization problem. Then, the corresponding expected regret under budget B is $\mathcal{R}(B) = G(\mathcal{A}^*) - G(\mathcal{A})$. For this pricing problem, it is not difficult to show that its offline version, where we know all sellers' costs and values in advance, is computationally hard. This can be done by, *e.g.*, reducing the well known Knapsack Problem to ours[2]. For such a problem, it is a common approach to consider the α-*regret* (Chen, Wang, and Yuan 2013) which is the worst case difference between an α-optimal sequence of actions and the performance under a particular policy $\mathcal{A}$ for $\alpha \in (0, 1]$, *i.e.*,

$$\mathcal{R}_\alpha(B) = \alpha G(\mathcal{A}^*) - G(\mathcal{A}). \quad (1)$$

[1]Normalize sellers' costs and suppose that any seller has a basic cost $c_{min} \in (0, 1]$, *e.g.*, the basic production cost (Li et al. 2017).

[2]Prices are viewed as items' weights and the objective is to maximize the total value of selected items under knapsack budget.

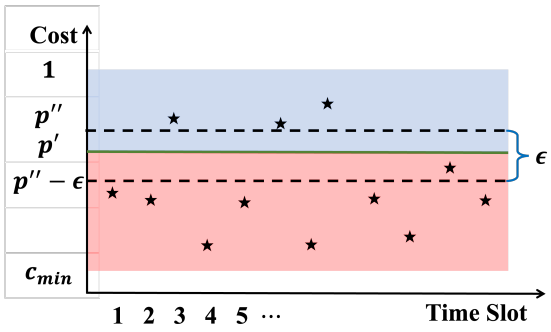

Figure 1: An example for the proof of Theorem 1.

3 Best Fixed-Price Policy in Hindsight

In this section, based on the definition of α-regret, we first turn to look for learning a computationally efficient benchmark oracle, with prior knowledge of sellers' costs and values, ensuring an approximation to the general optimum $G(\mathcal{A}^*)$. To this end, we consider the *best fixed-price policy in hindsight* (without ambiguity, we only say the best fixed-price policy in the remaining paper), which decides for each seller their best fixed price over time (corresponding to the best fixed arm in the bandit setting). We show that this policy: (i) enjoys a *provable approximation ratio* (this section); and (ii) is *learnable* (Section 4) in the semi-bandit feedback setting.

We first divide the continuous price space $A_{con} = [c_{min}, 1]$ into a discrete price set A with granularity ϵ, *i.e.*, $A = \{c_{min}, c_{min}+\epsilon, .., c_{min}+(K-2)\epsilon, 1\}$ where $K = |A| = \lceil \frac{1-c_{min}}{\epsilon} \rceil + 1$. Specially, we use $\mathcal{F}_{dis}$ to denote the fixed-price policy that chooses prices from the discrete price space A to maximize the cumulative reward. Let $(e_1, e_2, ..., e_n)$ denote the selected fixed-price vector for n sellers where e_i is the fixed-price for seller s_i among all time slots. Then, the objective of $\mathcal{F}_{dis}$ is

$$\begin{aligned} &\max_{(e_1,e_2,...,e_n)} \sum_{t\le\tau(\mathcal{F}_{dis})} \sum_{i\le n} x_t^i(e_i)\cdot v_t^i \\ s.t., &\sum_{t=1}^{\tau(\mathcal{F}_{dis})} \sum_{i=1}^{n} e_i \cdot x_t^i(e_i) \le B; \forall i \le n, e_i \in A \end{aligned} \tag{2}$$

Let $\mathcal{F}_{dis}^*$ be the best fixed-price policy for the optimization in (2). Then, we show the approximation performance of $\mathcal{F}_{dis}^*$.

Theorem 1. *The best fixed-price policy $\mathcal{F}_{dis}^*$ under the discrete price set A with granularity ϵ achieves $(2-c_{min})(1+\frac{\epsilon}{c_{min}})\frac{v_{max}}{v_{min}}$-approximation to the optimal policy $\mathcal{A}^*$.*

Proof. We first consider the case of a single seller. As shown in Fig. 1, the stars represent sellers' items and their costs at different time slots. Suppose that the best fixed-price policy $\mathcal{F}_{dis}^*$ for the single seller case chooses the fixed-price $p^* \in A$. Let $ALG_F(p)$ be the cumulative reward of the fixed-price policy with price $p \in A_{con}$. Denote by $N(p)$ the number of items with costs no higher than p across all time slots.

We define the price $p' \in [c_{min}, 1]$ as the maximum price that satisfies $N(p') \cdot p' \le B$. Let p'' denote the minimum price in A which is no lower than p'. Thus, we have $ALG_F(p'') \le ALG_F(p^*)$ as p^* is the best fixed-price. Then, we divide all items into two parts: the items in the blue and red area have costs in range $(p', 1]$ and $[c_{min}, p']$, respectively. Next, we consider the following two cases: 1) In the red area: the optimal solution $\mathcal{A}^*$ in the best case can obtain at most $N(p') \le \frac{B}{p'}$ items[3] with total payment at least $\frac{B}{p'} \cdot c_{min}$ according to the definition of p'. 2) In the blue area: given the budget B, $\mathcal{A}^*$ can procure at most $\frac{B}{p'}$ items. By combining the above two cases, we have

$$G(\mathcal{A}^*) \le (\frac{B}{p'} + \frac{B - \frac{B}{p'}c_{min}}{p'}) v_{max} \le (2-c_{min}) v_{max} \frac{B}{p'}.$$

Due to the definition of p', there are at least $\frac{B}{p''}$ items with costs no higher than p'' as $p'' \ge p'$. Thus, given the fixed price p'', $ALG_F(p^*) \ge ALG_F(p'') \ge v_{min}\frac{B}{p''}$. Then,

$$\begin{aligned} \frac{G(\mathcal{A}^*)}{G(\mathcal{F}_{dis}^*)} &\le \frac{(2-c_{min}) v_{max}\frac{B}{p'}}{v_{min}\frac{B}{p''}} \\ &\le (2-c_{min})\frac{v_{max}(c_{min}+\epsilon)}{v_{min}c_{min}}. \end{aligned} \tag{3}$$

Next, we extend the above approximation to the case with n sellers. Suppose that $\mathcal{A}^*$ pays overall $z_i \cdot B, z_i \in [0,1]$ to seller s_i and $\sum_{i\le n} z_i \le 1$. Let $V_F(z_i \cdot B)$ and $V_O(z_i \cdot B)$ denote the maximum value achieved by the fixed-price policy $\mathcal{F}_{dis}^*$ and $\mathcal{A}^*$ from seller s_i with budget $z_i \cdot B$. Due to Eq. (3), we have $\frac{V_O(z_i \cdot B)}{V_F(z_i \cdot B)} \le (2-c_{min})\frac{v_{max}(c_{min}+\epsilon)}{v_{min}c_{min}}$. As $\mathcal{F}_{dis}^*$ is the best fixed-price policy, then $G(\mathcal{F}_{dis}^*) \ge \sum_{i\le n} V_F(z_i \cdot B)$. We therefore have $\frac{G(\mathcal{A}^*)}{G(\mathcal{F}_{dis}^*)} \le \frac{\sum_{i\le n} V_O(z_i \cdot B)}{\sum_{i\le n} V_F(z_i \cdot B)} \le (2-c_{min})\frac{v_{max}(c_{min}+\epsilon)}{v_{min}c_{min}}$. □

Next, we provide an asymptotically tight lower bound for all fixed-price policies which can choose any price in A_{con}.

Theorem 2. *No fixed-price policy $\mathcal{F}$ under A_{con} obtains an approximation ratio better than $\Omega(\frac{v_{max}}{v_{min}})$.*

Proof. We consider a scenario involving a single seller with all costs equal to 1. During the first $\lfloor B \rfloor$ time slots, the values are v_{min}, while in the remaining time slots, the values are v_{max}. The best fixed-price policy sets the price at 1, yielding a total value of $\lfloor B \rfloor v_{min}$. In contrast, the optimal solution can achieve a total value of $\lfloor B \rfloor v_{max}$ by setting the price to c_{min} during the first $\lfloor B \rfloor$ time slots and 1 in the remaining time slots. Thus, the performance ratio between the optimal solution and the fixed-price policy is at least $\frac{v_{max}}{v_{min}}$. □

4 The Non-stochastic Pricing Bandit

Given the above best fixed-price policy $\mathcal{F}_{dis}^*$, we now show that it is learnable in the semi-bandit feedback setting by introducing an online pricing policy. The main idea is as follows. Firstly, because the continuous price space can result in infinite action space when the buyer chooses a price vector for sellers at each time slot, we first divide the continuous

[3] We do not round the number of items here and its effect is negligible since the budget is typically much larger than the cost.

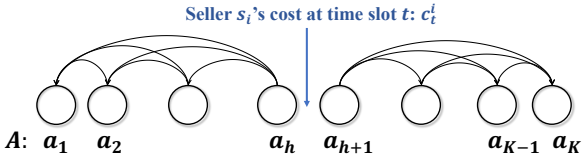

Figure 2: Example of the feedback graph G_t^i of seller s_i.

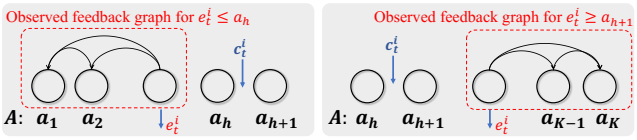

Figure 3: The observed feedback graph G_t^i of seller s_i.

price space into discrete candidate prices with granularity ϵ by which the best fixed-price policy on such discrete price set can achieve the granularity-dependent approximation to the best fixed-price policy on the continuous price (see Theorem 1). By choosing an appropriate granularity, we can obtain the sub-linear regret with respect to the budget. Secondly, by leveraging the monotonicity property of the sellers' acceptance function in the pricing problem, we design a feedback graph based online learning policy which can efficiently utilize sellers' binary decisions and thus improve the cumulative reward. In the following, we first introduce the feedback graph structure in our non-stochastic pricing problem. We then propose an online learning policy with a given budget as the input and analyze its regret. Finally, we extend the policy to deal with the unknown budget case.

Feedback

Recall that we divide the continuous price space $A_{con} = [c_{min}, 1]$ into a discrete price set A with granularity ϵ, *i.e.*, $A = \{c_{min}, .., c_{min} + (K-2)\epsilon, 1\}$ where $K = \lceil \frac{1-c_{min}}{\epsilon} \rceil + 1$. Let a_j denote the j-th price in A. Unlike the classical non-stochastic bandit problem in which only the selected action revealed the reward, additional decisions at prices apart from the given price can also be revealed in the pricing problem, *e.g.*, if seller s_i accepts a price e_t^i, we can further have the information that s_i also accepts any price higher than e_t^i.

Graph-based Feedback. Following the description in (Alon et al. 2017), we adopt a graph-theoretic interpretation to describe the additional information revealed by the chosen price. At each time slot t, there exists a directed graph $G_t^i = \left(V, D_t^i\right)$ for seller s_i, called *feedback graph*, where the node set V represents the price set A and D_t^i is the set of arcs, *i.e.*, ordered pairs of nodes. For any $e', e'' \in A$, the arc (e', e'') is included in D_t^i if and only if the decision of s_i regarding price e' also reveals the acceptance/rejection of price e'' (there is also an arc (e, e) pointing to themselves). Thus, at time slot t, there are n feedback graphs represented as $\mathcal{G}_t = \{G_t^1, G_t^2, ..., G_t^n\}$. It is important to note that we do not possess any information about the feedback graph prior to the buyer's actions. To provide a visual representation, we use Fig. 2 as an example to illustrate feedback graph G_t^i.

Example 1. *Suppose that the cost of seller s_i at time slot t is $c_t^i \in (a_h, a_{h+1}), h \le K-1$. Specifically, when choosing any price $e_t^i \ge a_{h+1}$, s_i obviously accepts e_t^i, and we have the information that s_i accepts any price higher than e_t^i. Thus, as shown in Fig. 2, there exists an arc (e', e'') for any price pair e', e'' such that $e' \le e''$ and $e' \ge a_{h+1}, e'' \ge a_{h+1}$ (the arc (e, e) is excluded in the figure for simplification). On the contrary, when choosing any price $e_t^i \le a_h$, seller s_i rejects the price e_t^i and will also reject any price no higher than e_t^i. Then, there exists an arc (e', e'') for any price pair e', e'' such that $e' \ge e''$ and $e' \le a_h, e'' \le a_h$ in Fig. 2.*

Observed Feedback System. Since the cost of each seller at each time slot is unknown, the buyer can only observe part of feedback graph G_t^i after selecting a price e_t^i for seller s_i. Recall that $x_t^i(e_t^i) \in \{0, 1\}$ represents whether s_i accepts the price $e_t^i \in A$ or not. Let $G_t^{i'}$ be the observed feedback graph, which is a sub-graph of G_t^i containing the prices whose decisions are revealed by e_t^i and the directed edges connecting these prices. Specially, the observed feedback graph is because if $x_t^i(e_t^i) = 1$ (or $x_t^i(e_t^i) = 0$), when setting any price $e \ge e_t^i$ (or $e \le e_t^i$) for s_i, we obtain the information that s_i accepts (or rejects) any price $e' \ge e$ (or $e' \le e$). For example, if $x_t^i(e_t^i) = 1$, the observed feedback graph $G_t^{i'}$ for s_i is displayed on the right side of Fig. 3. The buyer can also obtain v_t^i value by choosing any price $e \ge e_t^i$. Otherwise, if $x_t^i(e_t^i) = 0$, the observed feedback graph $G_t^{i'}$ is displayed on the left side of Fig. 3, and the buyer obtains 0 value by setting any price $e \le e_t^i$. Let $\mathcal{G}_t' = \{G_t^{i'}\}_{i \le n}$ be the observed feedback system at time slot t. We write $e \overset{(i,t)}{\to} e'$ if there exists an arc (e, e') in $G_t^{i'}$, and A_t^i is the set of prices in $G_t^{i'}$.

The GAP Policy

Given the defined feedback graph, as shown in Algorithm 1, we develop an online learning policy, *Granularity-based Pricing (GAP)*, which contains three main components: price selection, feedback observation and estimation update.

Price Selection. At each time slot, GAP first chooses a price vector for n sellers. Denote by $w_t^i(e)$ the weight of price $e \in A$ for seller s_i at time slot t. We maintain a set of time-varying weight vectors $W = \{W_t^i\}_{t \le \mathcal{T}, i \le n}$ where $W_t^i = \sum_{e \in A} w_t^i(e)$. At each time t, we compute the *importance sampling probability*, denoted by $p_t^i(e)$, for each seller s_i and each possible price $e \in A$, which is dependent of the time-varying weights and the exploration constant γ/K as

$$p_t^i(e) = (1-\gamma)\frac{w_t^i(e)}{\sum_{e' \in A} w_t^i(e')} + \frac{\gamma}{K}, \forall e \in A, i \le n. \quad (4)$$

The exploration constant γ/K guarantees that, for each seller, the buyer always has an underlying positive probability to explore prices in set A. Then, we choose price e_t^i drawn according to the distribution $p_t^i = (p_t^i(a_1), ..., p_t^i(a_j), ..., p_t^i(a_K))$ for each seller s_i. Let $B(t)$ denote the remaining budget at the beginning of time slot t. If the sum of the chosen prices for sellers at time slot t is higher than $B(t)$, *i.e.*, $\sum_{i \le n} e_t^i > B(t)$, GAP terminates. Otherwise, GAP sets price vector $\boldsymbol{e}_t = (e_t^1, e_t^2, ..., e_t^n)$.

Algorithm 1: The Granularity-based Pricing (GAP) Policy

Input: $B, S, c_{min}, v_{min}, v_{max}$

Initiation: ϵ, γ, η;
Generate price set $A = \{c_{min}, c_{min}+\epsilon, .., c_{min}+(K-2)\epsilon, 1\}$ where $K = \lceil \frac{1-c_{min}}{\epsilon} \rceil + 1$;
Weight $w_1^i(e) := 1, \forall e \in A, \forall i \leq n$;
$t \leftarrow 1, B(t) \leftarrow B$;
while $B(t) > 0$ and $t \leq \mathcal{T}$ **do**
  Compute $p_t^i(e), \forall e \in A, \forall i \leq n$ according to Eq. (4);
  Draw price $e_t^i \sim (p_t^i(a_1), ..., p_t^i(a_K)), \forall i \leq n$;
  If $\sum_{i \leq n} e_t^i > B(t)$, then GAP terminates;
  Observe $\forall i \leq n, x_t^i(e_t^i), r_t^i(e_t^i)$ and $\mathcal{G}_t'$;
  $B(t+1) \leftarrow B(t) - \sum_{i \leq n} e_t^i \cdot x_t^i(e_t^i)$;
  $\forall e \in A_t^i, i \leq n$, compute $\boldsymbol{\ell}_t^i(e) = \frac{r_t^i(e)}{e}$
  $\forall e \in A, i \leq n$, compute $\hat{\boldsymbol{\ell}}_t^i(e) = \frac{\boldsymbol{\ell}_t^i(e)}{q_t^i(e)} \mathbb{I}\{e \in A_t^i\}$;
  $\forall e \in A, i \leq n, w_{t+1}^i(e) = w_t^i(e) \cdot \exp\left(\eta \hat{\boldsymbol{\ell}}_t^i(e)\right)$;
  $t \leftarrow t+1$
end while

Feedback Observation. Following the given price e_t^i, the buyer observes s_i's decision $x_t^i(e_t^i)$, the obtained value denoted by $r_t^i(e_t^i)$, and the feedback system $\mathcal{G}_t'$. Specially, v_t^i can be observed when $x_t^i(e_t^i) = 1$, *i.e.*, $r_t^i(e_t^i) = v_t^i \cdot x_t^i(e_t^i)$. Let $m_t^i(e_t^i) = e_t^i \cdot x_t^i(e_t^i)$ be the payment paid to s_i under the price e_t^i. Recall that A_t^i contains prices whose decisions are revealed after the selection of e_t^i for s_i. Thus, when choosing price $e \in A_t^i$ for s_i, the obtained value $r_t^i(e)$ equals $r_t^i(e_t^i)$.

Estimation Update. Next, we use $\boldsymbol{\ell}_t^i(e)$ to measure the price-efficiency of $\forall e \in A_t^i$ for s_i, *i.e.*, $\boldsymbol{\ell}_t^i(e) = \frac{r_t^i(e)}{e}$. We further define the *side-observation yielded unbiased estimate* as $\hat{\boldsymbol{\ell}}_t^i(e) = \frac{\boldsymbol{\ell}_t^i(e)}{q_t^i(e)} \mathbb{I}\{e \in A_t^i\}, \forall e \in A, i \leq n$, that divides $\boldsymbol{\ell}_t^i(e)$ by the probability of $q_t^i(e) = \sum_{e': e' \overset{(i,t)}{\rightarrow} e} p_t^i(e')$, where $q_t^i(e)$ is the sum of the probabilities of choosing prices who also reveal s_i's decision for price e based on the feedback system $\mathcal{G}_t'$. Obviously, this estimate leverages the side-observations of other prices based on the particular characterization of the pricing problem. Finally, we use the estimate $\hat{\boldsymbol{\ell}}_t^i(e)$ to update weights as $w_{t+1}^i(e) = w_t^i(e) \cdot \exp\left(\eta \hat{\boldsymbol{\ell}}_t^i(e)\right), \forall e \in A, i \leq n$. The main difference between weights $w_t^i(e)$ and $w_{t+1}^i(e)$ is controlled by scaling estimated efficiency $\hat{\boldsymbol{\ell}}_t^i(e)$.

Regret Analysis

Upper Bound of Regret. We compare GAP against an α-approximation solution with $\alpha = \frac{v_{min}}{(2-c_{min})v_{max}}$. According to Theorem 1, the α-regret in Eq. (1) can be bounded by $\mathcal{R}_\alpha(B) = \alpha G(\mathcal{A}^*) - G(\mathcal{A}) \leq (1 + \frac{\epsilon}{c_{min}}) G(\mathcal{F}_{dis}^*) - G(\mathcal{A})$. Then, by choosing appropriate ϵ, we have the following regret upper bound.

Theorem 3. *By setting $\epsilon = c_{min}^{3/2}/\sqrt{B}$, $\eta = \epsilon / v_{max}$ and $\gamma = \epsilon / c_{min}$, the expected α-regret of GAP is at most $O\left(n \frac{v_{max}}{c_{min}} \sqrt{B/c_{min}} \ln B\right)$ where $\alpha = \frac{v_{min}}{(2-c_{min})v_{max}}$.*

Lower Bound of Regret. We now show that our α-regret upper bounds are tight up to sub-linear terms. In particular, we provide the following regret lower bound.

Theorem 4. *For any buyer's policy $\mathcal{M}$, there exists a sequence of sellers' costs and values where the expected α-regret of $\mathcal{M}$ is at least $\Omega\left(v_{max}\sqrt{Bn/c_{min}}\right)$.*

Extension to Unknown Budget

Note that GAP depends on the given budget to determine the price granularity ϵ that determines the candidate price set A, and tuning parameters η and γ, respectively. We now further extend it to a scenario where the budget is not given in advance, *e.g.*, the online algorithm may not know the advertisers' budgets in advance (Udwani 2024). To address it, we develop a variant of GAP, referred to as Doubling-GAP, by adopting the doubling method. Specifically, we can define a doubling sequence $(T_l)_{l \in N}$, where $T_l = e^l$ and $T_0 = 1$, and divide the learning process into multiple phases. At the beginning of each phase, *e.g.*, the l-th phase, we fully restart the underlying policy GAP by initializing the price granularity ϵ and the corresponding parameters γ, η based on the current budget $B_0 T_l$ and setting the running time slot for this phase to $\frac{B_0 T_l}{n c_{min}}$. Let L_B denote the last phase running GAP, *i.e.*, $L_B = \min_{l \in N}\{\sum_{h \leq l} B_0 \cdot e^h \geq B\}$, which implies $L_B = \lceil \ln(B+1) \rceil - 1$. Then, we have the regret bound:

Corollary 1. *The α-regret of Doubling-GAP is at most $O(n \frac{v_{max}}{c_{min}} \sqrt{B/c_{min}} \ln^2 B)$ where $\alpha = \frac{v_{min}}{(2-c_{min})v_{max}}$.*

5 Simulation Experiments

We empirically evaluate the proposed main policy GAP by comparing it with four state-of-the-art baseline algorithms from the literature, which represent the typical variations of online learning algorithms one may use in the budgeted non-stochastic pricing bandit problem. In particular, they are: the classic non-stochastic Exp3 algorithm (Exp3) (Auer et al. 2002); the stochastic pricing mechanism BP-UCB (Singla and Krause 2013); KUBE (Tran-Thanh et al. 2012) which estimates sellers' costs and values by using traditional confidence bound; and the Thompson Sampling (TSampling) algorithm (Gopalan, Mannor, and Mansour 2014) which chooses prices for sellers at each time slot according to the generated value sampled from the dynamic beta distribution. These algorithms are representative standard methods designed for non-stochastic (adversarial) bandits, budgeted bandits, and online pricing, respectively. Additionally, they often assume that the budget information is known.

We follow the standard setup for evaluating non-stochastic bandits (Zimmert, Luo, and Wei 2019; Alipour-Fanid, Dabaghchian, and Zeng 2021). In particular, in our experiments the degree of the non-stochasticity can be quantified by the stochastically constrained adversaries. That is, the mean cost/value of each seller switches while staying unchanged for phases that are increasing exponentially in

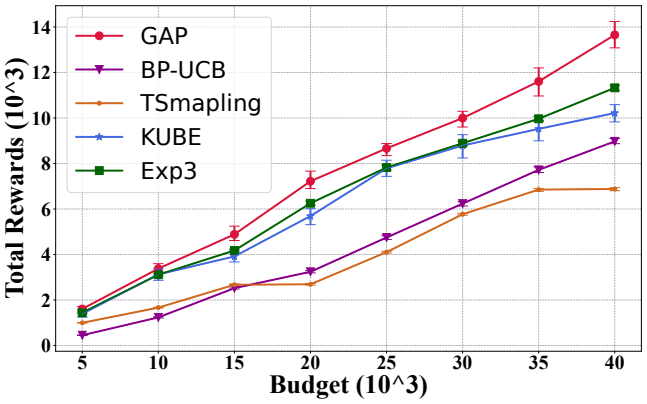

Figure 4: Reward vs. Budget

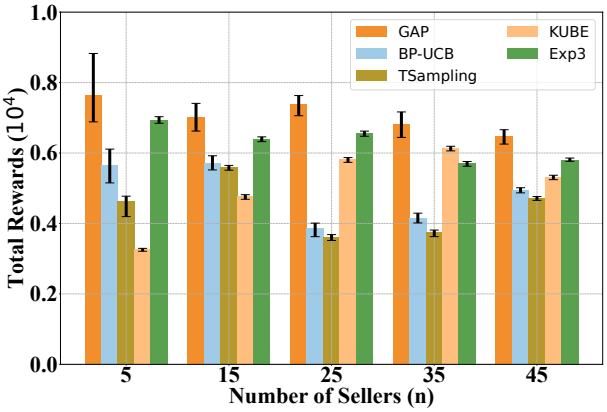

Figure 5: Reward vs. Number of Sellers

length (which is a common metric in non-stochastic bandit literature). This approach has proven effective in testing non-stochastic algorithms via extensive experiments (Zimmert and Seldin 2021). We divide the horizon into phases

$$\underbrace{1, ..., t_1}_{T_1}, \underbrace{t_1 + 1, ..., t_2}_{T_2}, ..., \underbrace{t_{n-1} + 1, ..., \mathcal{T}}_{T_n},$$

in which the length of the s-th phase is set as $T_s = 1.6^s$. The cost of seller s_i is uniformly set as follows,

$$c_t^i \in \begin{cases} [0.1, 0.2] & \text{if } t \text{ belongs to an odd phase,} \\ [0.2, 1] & \text{otherwise} \end{cases}$$

and the value of each seller similarly switches between range $[0.1, 0.2]$ and $[0.2, 1]$. Such a model can be justified by real-world applications, *e.g.*, in a network routing problem, an adversary might periodically attack the network, making the delay and the throughput of every edge change dynamically (Pongle and Chavan 2015).

Results. We compare the performance of the algorithms by the total reward over at least 100 rounds. To evaluate the impact of the budget, we vary the budget in the range $[5000, 40000]$ with the increment of 5000 and let $B = 20000$ by default. We also let the number of sellers, *i.e.*, n, be selected from $\{5, 15, 25, 35, 45\}$, and let $n = 25$ by default. Besides, we keep the ratio between the budget B and the total horizon fixed so as to measure the performance of all algorithms in a fair comparison with the increase of budget.

When the budget varies, the achieved total reward of algorithms are shown in Fig. 4. We conclude that our proposed GAP policy outperforms all baseline algorithms. **Firstly**, the total reward achieved by GAP is significantly higher than that of BP-UCB, TSampling, and KUBE. Precisely, the total reward of GAP is 55.59%, 59.89% and 26.32% larger than those of BP-UCB, TSampling, and KUBE on average, respectively. This is because these algorithms make efforts learning distributions of sellers' costs and values, which falls into inaccurate estimation due to the "non-stochastic" environment. **Secondly**, for the non-stochastic algorithms, GAP also achieves a better total reward than Exp3. In detail, the total reward of GAP is 14.16% larger than that of Exp3 on average. This is because GAP can pay more precise prices than that of Exp3 by utilizing the side-information which leverages the characterization of the feedback graph in the pricing problem. **Lastly**, we observe that the gap between GAP and the compared algorithms is enlarged with the increase of budget. This is because GAP can choose prices more accurately with the increase of running time slots, which results in more efficient utilization of the budget than the baseline algorithms, and thus obtains greater achieved reward. Moreover, we analyze the impact of the number of sellers, and the corresponding results are shown in Fig. 5. The total reward achieved by GAP are much higher than those of baseline algorithms. Precisely, the total reward of GAP is 48.69%, 62.22%, 47.45% and 12.31% larger than those of the BP-UCB, TSampling, KUBE and Exp3 algorithms on average, respectively. In summary, these results validate that the proposed method significantly outperforms the baseline algorithms.

6 Conclusion

In this paper, we investigate the problem of non-stochastic online pricing with semi-bandit feedback in the procurement scenario where sellers' costs and values at each time slot are arbitrary. Firstly, for the offline setting where sellers' costs and values over time are all known in advance, we investigate the learnable benchmark oracle, that is, the best fixed-price policy. We show its tight approximation ratio to the global optimum. In the general online setting, we first investigate the known budget case and propose feedback graph based policy GAP which achieves an upper bound of $O(n\frac{v_{max}}{c_{min}}\sqrt{B/c_{min}}\ln B)$ on the α-regret where $\alpha = \frac{v_{min}}{(2-c_{min})v_{max}}$. We then extend it to the case where the budget of the buyer is unknown and introduce Doubling-GAP which divides time slots into multiple phases which run GAP. We show that the α-regret of Doubling-GAP is at most $O(n\frac{v_{max}}{c_{min}}\sqrt{B/c_{min}}\ln^2 B)$. We also prove a tight lower bound up to a logarithmic factor, that is, the regret is $\Omega(v_{max}\sqrt{Bn/c_{min}})$. The conducted simulation experiments show that the proposed policy outperforms the compared baseline algorithms.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grant 62402102, 61972086, the Natural Science Foundation of Jiangsu Province under Grant No. BK20241275, BK20230024. Hau Chan is supported by the National Institute of General Medical Sciences of the National Institutes of Health [P20GM130461], the Rural Drug Addiction Research Center at the University of Nebraska-Lincoln, and the National Science Foundation under grants IIS:RI #2302999 and IIS:RI #2414554. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

Alipour-Fanid, A.; Dabaghchian, M.; and Zeng, K. 2021. Self-unaware adversarial multi-armed bandits with switching costs. *IEEE Transactions on Neural Networks and Learning Systems*.

Alon, N.; Cesa-Bianchi, N.; Dekel, O.; and Koren, T. 2015. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, 23–35. PMLR.

Alon, N.; Cesa-Bianchi, N.; Gentile, C.; Mannor, S.; Mansour, Y.; and Shamir, O. 2017. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6): 1785–1826.

Amanatidis, G.; Kleer, P.; and Schäfer, G. 2022. Budget-feasible mechanism design for non-monotone submodular objectives: Offline and online. *Mathematics of Operations Research*.

Anari, N.; Goel, G.; and Nikzad, A. 2014. Mechanism design for crowdsourcing: An optimal 1-1/e competitive budget-feasible mechanism for large markets. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, 266–275. IEEE.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2): 235–256.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1): 48–77.

Avadhanula, V.; Colini Baldeschi, R.; Leonardi, S.; Sankararaman, K. A.; and Schrijvers, O. 2021. Stochastic bandits for multi-platform budget optimization in online advertising. In *Proceedings of the Web Conference 2021*, 2805–2817.

Badanidiyuru, A.; Kleinberg, R.; and Singer, Y. 2012. Learning on a budget: posted price mechanisms for online procurement. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, 128–145.

Badanidiyuru, A.; Kleinberg, R.; and Slivkins, A. 2018. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3): 1–55.

Balkanski, E.; and Hartline, J. D. 2016. Bayesian budget feasibility with posted pricing. In *Proceedings of the 25th International Conference on World Wide Web*, 189–203.

Bei, X.; Chen, N.; Gravin, N.; and Lu, P. 2012. Budget feasible mechanism design: from prior-free to bayesian. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, 449–458.

Chen, W.; Wang, Y.; and Yuan, Y. 2013. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, 151–159. PMLR.

Combes, R.; Talebi Mazraeh Shahi, M. S.; Proutiere, A.; et al. 2015. Combinatorial bandits revisited. *Advances in Neural Information Processing Systems*, 28.

Dütting, P.; and Kesselheim, T. 2019. Posted pricing and prophet inequalities with inaccurate priors. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, 111–129.

Feldman, M.; Koren, T.; Livni, R.; Mansour, Y.; and Zohar, A. 2016. Online pricing with strategic and patient buyers. *Advances in Neural Information Processing Systems*, 29.

Gopalan, A.; Mannor, S.; and Mansour, Y. 2014. Thompson sampling for complex online problems. In *International Conference on Machine Learning*, 100–108. PMLR.

Ha, L. 2008. Online advertising research in advertising journals: A review. *Journal of Current Issues & Research in Advertising*, 30(1): 31–48.

Ho, C.-J.; and Vaughan, J. 2012. Online task assignment in crowdsourcing markets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, 45–51.

Hu, Z.; and Zhang, J. 2017. Optimal Posted-Price Mechanism in Microtask Crowdsourcing. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 228–234.

Immorlica, N.; Sankararaman, K. A.; Schapire, R.; and Slivkins, A. 2019. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, 202–219. IEEE.

Jagabathula, S.; Subramanian, L.; and Venkataraman, A. 2017. Identifying unreliable and adversarial workers in crowdsourced labeling tasks. *The Journal of Machine Learning Research*, 18(1): 3233–3299.

Jain, S.; Ghalme, G.; Bhat, S.; Gujar, S.; and Narahari, Y. 2016. A deterministic mab mechanism for crowdsourcing with logarithmic regret and immediate payments. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 86–94.

Kleinberg, R.; and Leighton, T. 2003. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, 594–605. IEEE.

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

Lee, K.-C.; Jalali, A.; and Dasdan, A. 2013. Real time bid optimization with smooth budget delivery in online advertising. In *Proceedings of the seventh international workshop on data mining for online advertising*, 1–9.

Li, C.; Chen, X.; Tang, Y.; and Li, L. 2017. Selection of optimum parameters in multi-pass face milling for maximum

energy efficiency and minimum production cost. *Journal of Cleaner Production*, 140: 1805–1818.

Liu, X.; Chan, H.; Li, M.; and Wu, W. 2024. Budget Feasible Mechanisms: A Survey. In *33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*, 8132–8141. International Joint Conferences on Artificial Intelligence.

Misra, K.; Schwartz, E. M.; and Abernethy, J. 2019. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38(2): 226–252.

Neu, G.; and Bartók, G. 2016. Importance Weighting Without Importance Weights: An Efficient Algorithm for Combinatorial Semi-Bandits. *Journal of Machine Learning Research*, 17: 1–21.

Pongle, P.; and Chavan, G. 2015. A survey: Attacks on RPL and 6LoWPAN in IoT. In *2015 International Conference on Pervasive Computing (ICPC)*, 1–6. IEEE.

Rangi, A.; Franceschetti, M.; and Tran-Thanh, L. 2019. Unifying the stochastic and the adversarial bandits with knapsack. *Proc. of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*, 3311–3317.

Romano, G.; Tartaglia, G.; Marchesi, A.; and Gatti, N. 2021. Online posted pricing with unknown time-discounted valuations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 5682–5689.

Sankararaman, K. A.; and Slivkins, A. 2018. Combinatorial semi-bandits with knapsacks. In *International Conference on Artificial Intelligence and Statistics*, 1760–1770. PMLR.

Singer, Y.; and Mittal, M. 2011. Pricing tasks in online labor markets. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Singer, Y.; and Mittal, M. 2013. Pricing mechanisms for crowdsourcing markets. In *Proceedings of the 22nd international conference on World Wide Web*, 1157–1166.

Singla, A.; and Krause, A. 2013. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *Proceedings of the 22nd International Conference on World Wide Web*, 1167–1178.

Tran-Thanh, L.; Chapman, A.; Rogers, A.; and Jennings, N. 2012. Knapsack based optimal policies for budget–limited multi–armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, 1134–1140.

Udwani, R. 2024. Adwords with unknown budgets and beyond. *Management Science*.

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

Wilbur, K. C.; and Zhu, Y. 2009. Click fraud. *Marketing Science*, 28(2): 293–308.

Yuan, S.; Wang, J.; and Zhao, X. 2013. Real-time bidding for online advertising: measurement and analysis. In *Proceedings of the seventh international workshop on data mining for online advertising*, 1–8.

Zhao, D.; Li, X.-Y.; and Ma, H. 2014. Budget-feasible online incentive mechanisms for crowdsourcing tasks truthfully. *IEEE/ACM Transactions on Networking*, 24(2): 647–661.

Zimmert, J.; Luo, H.; and Wei, C.-Y. 2019. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, 7683–7692. PMLR.

Zimmert, J.; and Seldin, Y. 2021. Tsallis-INF: An Optimal Algorithm for Stochastic and Adversarial Bandits. *J. Mach. Learn. Res.*, 22(28): 1–49.