Journal of Machine Learning Research 24 (2023) 1-45 Submitted 10/20; Revised 7/22; Published 1/23

VCG Mechanism Design with Unknown Agent Values under Stochastic Bandit Feedback

Kirthevasan Kandasamy KANDASAMY@CS.WISC.EDU
University of Wisconsin—Madison, WI 53706, USA

Joseph E Gonzalez JEGONZAL@EECS.BERKELEY.EDU
Michael I Jordan JORDAN@CS.BERKELEY.EDU
Ion Stoica STOICA@CS.BERKELEY.EDU
University of California, Berkeley, CA 94723, USA

Editor: Vahab Mirrokni

Abstract

We study a multi-round welfare-maximising mechanism design problem in instances where agents do not know their values. On each round, a mechanism first assigns an allocation to a set of agents and charges them a price; at the end of the round, the agents provide (stochastic) feedback to the mechanism for the allocation they received. This setting is motivated by applications in cloud markets and online advertising where an agent may know her value for an allocation only after experiencing it. Therefore, the mechanism needs to explore different allocations for each agent so that it can learn their values, while simultaneously attempting to find the socially optimal set of allocations. Our focus is on truthful and individually rational mechanisms which imitate the classical VCG mechanism in the long run. To that end, we first define three notions of regret for the welfare, the individual utilities of each agent and that of the mechanism. We show that these three terms are interdependent via an $\Omega(T^{2/3})$ lower bound for the maximum of these three terms after T rounds of allocations, and describe an algorithm which essentially achieves this rate. Our framework also provides flexibility to control the pricing scheme so as to trade-off between the agent and seller regrets. Next, we define asymptotic variants for the truthfulness and individual rationality requirements and provide asymptotic rates to quantify the degree to which both properties are satisfied by the proposed algorithm.

Keywords: Mechanism design, VCG Mechanism, Truthfulness, Game Theory, Bandits

1. Introduction

Mechanism design is one of the most important problems in economics and computer science (Nisan and Ronen, 2001). A mechanism chooses *allocations* for multiple rational agents with possibly conflicting goals and charges them a price. It is necessary to find an *outcome* (an allocation to each agent) that is as beneficial as possible to all agents and the mechanism designer. Agents who act in their own self interest might choose to misrepresent their values in order to obtain an advantageous allocation. Mechanism design aims to elicit values from agents, such that the agents are incentivised to report truthfully (truthfulness), while ensuring that they are not worse off than if they had not participated in the mechanism (individual rationality).

As a motivating example, consider a Platform-as-a-Service (PaaS) provider who serves multiple customers using the same compute cluster. The service provider (seller) chooses a service level (allocation) for each customer (agent) and charges them accordingly. The service level determines the resources allocated to the customer, and consequently her value for that service, which could be tied to her own revenue. A customer's experience of a service level at a given instant is affected by exogenous stochastic factors such as traffic, machine failures, etc., which are beyond the control of the customer. The celebrated VCG mechanism (Vickrey, 1961; Groves, 1979; Clarke, 1971) provides a means to find outcomes which maximise the social welfare (sum of agent and seller utilities) in such situations while satisfying truthfulness and individual rationality. For instance, if one customer's application is memory intensive and another's is compute intensive, they can be co-located on the same set of machines instead of using separate machines. This might be a better outcome for the service provider as she can serve both customers at a cheaper cost, and for the customers, since the service provider can now charge them less and they achieve the same end result. The VCG mechanism requires customers to submit bids for each service level and encourages truthful behaviour; i.e., the dominant strategy for each customer is to submit their true value as the bid.

Despite the many success stories of mechanism design, deploying it in some real world use cases has remained challenging since most mechanism design work assumes that agents know their own values for each allocation. For instance, the VCG mechanism requires that customers submit bids representing these values. This may not be true in many real world situations, especially when there are many unsophisticated agents and/or when the number of allocations is very large. However, having experienced an allocation, it is often the case that a customer can provide feedback based on their experience. She can either measure this directly via the impact on her own revenue, such as in online advertising where an ad impression might lead to a click and then a purchase, or gauge it from performance metrics, such as in the PaaS example where the service level affects the fraction of queries completed on time, which in turn affects her revenue.

Setting: In a departure from prior work, we study mechanisms where agents do not know their values a priori. However, the mechanism can learn them over multiple rounds of allocations and feedback, while simultaneously finding the socially optimal outcome. At the beginning of each round, the mechanism chooses an outcome, i.e. an allocation for each agent, and charges each agent a price. At the end of the round, the agents report stochastic feedback on their experience in using the given allocation, which we will call *reward*. When choosing an outcome for a given round, the mechanism may use the rewards reported by the agents in previous rounds.

This problem ushers in the classical explore-exploit dilemma encountered in bandit settings. Provided that all agents report their rewards truthfully, choosing the outcome that appears to be the best according to feedback provided by agents up to the current round will likely have large welfare. However, exploring other outcomes might improve the estimate of the best outcome for future rounds.

As is the case in prior mechanism design work, we assume that agents are strategic and rational, which necessitates the truthfulness and individual rationality requirements. A strategic agent wishes to maximise her total utility after T rounds, which is simply the sum of her instantaneous utilities (value of the allocation received minus price). An individually rational agent wishes to ensure that her utility after T rounds is non-negative, so that she stands to gain by participating in the mechanism. Both these requirements are more challenging in our setting. A mechanism cannot learn agent values if she does not report back truthfully. Since she reports a reward at the end of each round,

	Truthfulness $\mathbb{E}[U_{iT}^{\pi} - U_{iT}] \in \tilde{\mathcal{O}}(?)$ (Theorem 2)	Individual rationality $-\mathbb{E}[U_{iT}] \in \tilde{\mathcal{O}}(?)$ (Theorem 3)	VCG regret $\mathbb{E}[R_T^{\mathrm{VCG}}] \in \tilde{\mathcal{O}}(?)$ (Theorem 4)
$\zeta = \mathsf{ETC}$	$K^{1/3}T^{2/3}$	$n^\star(\lambda)K^{1/3}T^{2/3}$, DSIC	$n^2K^{1/3}T^{2/3}$
$\zeta = \mathsf{OPT}$	$nK^{1/3}T^{2/3}$, NIC	$n^\star(\lambda)\lvert\mathcal{S}\rvert^{1/2}T^{1/2}$	$n^2K^{1/3}T^{2/3}$

Table 1: Summary of asymptotic rates for Algorithm 1 for truthfulness, individual rationality, and the VCG regret R_T^{VCG} (3). Above, n denotes the number of agents, T the number of rounds, $|\mathcal{S}|$ the number of different allocations, K is a problem-specific parameter defined in Section 3, U_{iT} is the total utility of agent i after T rounds when being truthful, and U_{iT}^{π} is the total utility of agent i after T rounds when following any other *adaptive* nontruthful strategy π. DSIC (NIC) indicates that the mechanism is asymptotically dominant–strategy (Nash) incentive-compatible. Algorithm 1 has two binary hyperparameters $\zeta \in \{\mathsf{ETC}, \mathsf{OPT}\}$, and $\lambda \in \{\mathsf{AGE}, \mathsf{SEL}\}$ (Section 5). First, when $\zeta = \mathsf{ETC}$, the algorithm follows an explore-then-commit strategy while when $\zeta = \mathsf{OPT}$, the algorithm follows an optimistic strategy with interleaved exploration rounds. When $\zeta = \mathsf{ETC}$, we have better truthfulness guarantees, but weak rates for individual rationality, and vice versa when $\zeta = \mathsf{OPT}$. Second, when $\lambda = \mathsf{AGE}$ ($\lambda = \mathsf{SEL}$), the pricing strategy is favourable to the agents (seller), via the $n^\star$ term; here, we have $n^\star(\mathsf{AGE}) = 1$ and $n^\star(\mathsf{SEL}) = n$.

a strategic agent has significantly more opportunity to manipulate outcomes in her favour, than in typical mechanism design settings where she submits a single bid once. In particular, she may be strategic over multiple rounds, say, by incurring losses in early rounds in order to gain in the long run. Additionally, since an agent's true values cannot be exactly known, the mechanism runs the risk of overcharging them, which might cause her to withdraw from the mechanism.

We design an algorithm that accounts for the above considerations. Applications such as PaaS or online advertising, where there are repeated agent-mechanism interactions and where values can be reported back in an automated way, are suitable for such methods.

Our Contributions: Our goal in this work is to study the fundamental limitations and trade-offs when designing a repeated VCG mechanism where agents do not know their valuations, but can provide bandit feedback based on their experience. To that end, our contribution is threefold:

1. First, we formalise mechanism design with bandit feedback for settings where agents do not know their values, but the mechanism is repeated for several rounds. In order to quantify how close the mechanism is to the VCG mechanism, our formalism defines the *VCG regret*; this is derived via three regret terms for the welfare, the seller, and the agents relative to the VCG mechanism. Additionally, given the above challenges in achieving truthfulness and individual rationality exactly, we define asymptotic variants to make the problem tractable.

2. Second, we establish a hardness result via an $\Omega(T^{2/3})$ lower bound after T rounds for the VCG regret even under truthful reporting from agents. This result captures the interaction between the three regret terms mentioned in 1. For instance, a seller can achieve small regret by demanding large payments from the agents, but this will result in large agent regret. Moreover,

the value shared by the agents and the seller is limited by the total value generated, and hence the agent and seller regrets are inherently tied to the welfare regret.

3. Third, we describe VCG-Learn, an algorithm whose behaviour is determined by two binary hyperparameters. For all values of these hyperparameters, the algorithm is asymptotically individually rational and truthful, and moreover matches the above lower bound on the VCG regret up to factors that are polylogarithmic in T and polynomial in other problem-dependent terms. However, the asymptotic rates and the regrets of the agent and the seller are affected by the choice of these hyperparameters. Table 1 summarises the results for Algorithm 1.

This manuscript is organised as follows. First, in Section 2, we discuss related work. In Section 3, we briefly review the VCG mechanism and describe our formalism. Section 4 presents the lower bound on the VCG regret. Section 5 present our algorithm, VCG-Learn, and Section 6 presents the theoretical results for VCG-Learn. Section 7 presents some simulation results. The proof of the lower bound is given in Section 8 and the proofs of results in Section 6 are given in Section 9.

2. Related Work

Bandit problems were first studied by Thompson (1933) and have since become an attractive framework to study exploration-exploitation trade-offs that arise in online decision-making. Optimistic methods, which usually choose an arm on a given round by maximising an upper confidence bound on the mean rewards, are known to be minimax optimal in a variety of stochastic optimisation settings (Lai and Robbins, 1985; Auer, 2002; Bubeck et al., 2011). Explore-then-commit strategies use separate rounds for exploration and exploitation. While they are provably sub-optimal (Garivier et al., 2016), they separate exploration from exploitation facilitating a cleaner analysis when we need to combine optimisation with other side objectives, such as in our problem, where we need to provide truthfulness guarantees and compute the prices.

Mechanism design has historically been one of the core areas of research in the economics and game theory literature with applications in kidney exchange (Roth et al., 2004), matching markets (Roth, 1986), and fair division (Procaccia, 2013). Our work is on auction-like settings for mechanism design. In addition to a rich history of research on this topic, there has also been a recent flurry of work due to the rise in popularity of sponsored search markets (Lahaie et al., 2007; Mehta et al., 2007; Aggarwal et al., 2006), wireless spectrum auctions (Cramton, 2013; Milgrom, 2017), and cloud spot markets (Toosi et al., 2016).

There is a long history of work in the intersection of machine learning and mechanism design. Some examples include online learning formulations (Dudik et al., 2017; Amin et al., 2013; Kakade et al., 2010), learning bidder values from past observations (Balcan et al., 2016; Blum et al., 2015; Balcan et al., 2008), and learning in other settings with truthfulness constraints (Mansour et al., 2015). Some work in this space study settings where individual agents may learn to bid in a repeated auction. Here, an agent may not know her value at the beginning, but needs to submit bids at the *beginning* of each round. The agent may calibrate her bid based on past rewards. In this line of work, Weed et al. (2016) and Feng et al. (2018) study a setting where the behaviour of the mechanism is fixed over multiple rounds, while Nedelec et al. (2019) study a setting where the mechanism may adapt its behaviour so as to maximize revenue. In a similar vein, Liu et al. (2019) develop bandit methods where agents on one side of a matching market learn to bid for arms on the other side. In the above work, the

regret is defined for the agent in question, defined relative to an oracle which knows the true values. In contrast to these works, in our setting, learning happens entirely on the mechanism side and the role of each agent is very simple: submit the reward at the *end* of the round. This imposes minimal burden on agents who, while being strategic and rational, may not be very sophisticated.

A body of work studies multi-armed bandit formalisms for auctions with canonical use cases in online advertising (Babaioff et al., 2015, 2014; Devanur and Kakade, 2009). In the above works, there is a single item (ad slot) with different and unknown click-through rates for each agent $c_i \in \mathbb{R}$. The agent has a *known* private value $v_i \in \mathbb{R}$ for each click and she submits a bid $b_i \in \mathbb{R}$ once ahead of time representing this value. On each round, the mechanism chooses one of the agents for the slot and observes the number of clicks c; if agent i was chosen, then $\mathbb{E}[c] = c_i$ The agent's reward for this round is $c \cdot v_i$. They formalise this problem where the agents are viewed as the arms and define regret with respect to the optimal arm, i.e. the agent with the highest expected reward $v_i \times c_i$. Importantly, the stochastic component c of the reward is observed by *both* the agent and the mechanism. In both works, truthfulness means that the agent is incentivised to submit a bid $b_i = v_i$ at the beginning of all rounds. There are a number of differences between these works and our setting. First, while they formalise each agent as a different arm competing for the item, in our setting, the allocations are viewed as arms with multiple agents being able to experience different arms simultaneously. We do not believe their results, and their lower bound in particular, can be straightforwardly extended to settings where multiple agents might receive an allocation. Second, in these works the agent can only submit a single bid and the stochastic component of the reward (number of clicks) is observed by the mechanism on each round. In contrast, in our setting, the reward on each round is only revealed to the agent, and she may misreport this reward to the mechanism on each round; therefore, she has significantly more opportunity to manipulate outcomes in her favour. Due to these differences, their results are not comparable to ours. We will elaborate in other differences between our results and theirs in further detail at the end of Sections 4 and 6.

In other related work, Braverman et al. (2019) consider a setting where a seller chooses one of n agents to receive an item on each round of a repeated auction. The agents submit a payment at the end of the round to the seller based on the reward they observed. They study mechanisms that allow the seller to extract as much payment as possible from the agents who themselves are trying to maximise their long term utility. Nazerzadeh et al. (2016) study a multi-round setting where the seller chooses an agent and a price on each round; the agent may choose to purchase the item at the price in which case the seller receives some revenue. Their goal is to maximise the revenue for the seller over a finite horizon of T rounds. Gatti et al. (2012) study an online advertising setting when there are multiple ad slots with different click-through rates which are the same for all agents, and design a mechanism which charges the agents only when an ad is clicked. Finally, some works on dynamic auctions (Bergemann and Valimaki, 2006; Athey and Segal, 2013; Kakade et al., 2013) study settings where agent values are unknown at the beginning but there is a known prior on the agent value. Over time, she receives side information and the mechanism needs to incentivise truth telling so as to update the posterior.

Perhaps the closest work to ours is Nazerzadeh et al. (2008), who study a single item auction with a feedback method similar to ours: agents report rewards at the end of the round and the learning happens at the mechanism. While they consider asymptotic efficiency, truthfulness, and individual

rationality (with definitions that differ from ours), they do not provide rates, establish lower bounds, or study the regrets of the agents and seller.

3. Problem Description

3.1 A brief review of the VCG mechanism

We begin with a brief review of mechanism design adapted to our setting. There are n agents (customers) $\{1, \ldots, n\}$, a mechanism (seller), and a set of possible outcomes Ω. The mechanism chooses an *outcome* $\omega \in \Omega$ and charges a price p_i to agent i. For agent i, there exists a function $s_i : \Omega \to \mathcal{S}$ which maps outcomes to *allocations* relevant to the agent; i.e., different outcomes ω, ω' might yield the same allocation to the agent, $s_i(\omega) = s_i(\omega')$. In this work, $|\mathcal{S}| < \infty$. $\mathcal{S}$ could be as large as Ω, but could be much smaller in some applications. This distinction between $\mathcal{S}$ and Ω will be important when we consider the learning problem in Section 3.2; as we will see shortly, our regret bounds will scale with $|\mathcal{S}|$ and not $|\Omega|$.

Agent i has a *value function*, $v_i : \mathcal{S} \to [0, 1]$, where $v_i(s)$ represents her private independent value for the allocation s. For an outcome $\omega \in \Omega$, we will overload notation and write $v_i(\omega) = v_i(s_i(\omega))$. After the agent experiences an allocation, she realises a *reward* X_i drawn from a σ sub-Gaussian distribution with mean $v_i(s)$. We let $v_0 : \Omega \to \mathbb{R}$ denote the value function of the mechanism designer. In the PaaS example, $v_0(\omega)$ may denote the cost to the service provider for providing the service where the allocations are as specified in ω. For an outcome ω and prices $\{p_i\}_{i=1}^n$, the *utility* of agent i is $u_i = \mathbb{E}[X_i] - p_i = v_i(\omega) - p_i$. The utility of the seller (which may represent profit) is $u_0 = v_0(\omega) + \sum_{i=1}^n p_i$. The *welfare* $V(\omega)$ is the sum of the agent and seller values $V(\omega) = \sum_{i=1}^n v_i(\omega) + v_0(\omega)$, which is also the sum of all utilities regardless of the prices $\{p_i\}_{i=1}^n$.

The expectations above are taken with respect to the rewards, i.e. the exogenous stochasticity arising when agents experience their allocation. In applications of interest, the agent does not have control over nor is able to predict this stochasticity.

The VCG Mechanism: Assume that the agents know their value functions v_i and submit them truthfully as bids to the seller. The VCG mechanism stipulates that we choose the outcome $\omega_\star$ which maximises the welfare. We then charge agent i an amount $p_{i\star}$, which is the loss her presence causes to the others. Precisely, denoting $V^{-i}(\omega) = v_0(\omega) + \sum_{j \neq i} v_j(\omega)$, we have

$$\omega_\star = \operatorname*{argmax}_{\omega \in \Omega} \; V(\omega), \qquad p_{i\star} = \max_{\omega \in \Omega} V^{-i}(\omega) - V^{-i}(\omega_\star). \tag{1}$$

In general, an agent may submit a bid $b_i : \mathcal{S} \to [0, 1]$ (not necessarily truthfully), and the mechanism computes the outcomes and prices by replacing v_i with b_i above. The VCG mechanism satisfies the following three fundamental desiderata in mechanism design (Karlin and Peres, 2017):

1. *Truthfulness*: A mechanism is truthful or dominant strategy incentive-compatible if, regardless of the bids submitted by other agents, the utility u_i of agent i is maximised when bidding truthfully, i.e. $b_i = v_i$.

2. *Individual rationality:* A mechanism is individually rational if it does not charge an agent more than her bid for an allocation. Thus, if she bids truthfully, her utility is nonnegative.

3. *Efficiency:* If all agents bid truthfully, a mechanism is efficient if it maximises welfare.

Since the agents cannot control the exogenous stochasticity, it is meaningful for agents to submit bids based on their expected rewards, i.e. their value. This is different from Bayesian formalisms for mechanism design where agent values are drawn from a *known* prior and she may submit bids based on this value. (A Bayesian formulation of this setting would assume priors over the values, i.e. the expected rewards, themselves.) The following examples illustrate our motivations.

Example 1 (PaaS) In the PaaS example from Section 1, $\mathcal{S}$ are the service levels (allocations) available to a customer. $\Omega = \mathcal{S}^n$ are the possible outcomes. $-v_0(\omega)$ is the cost for providing the service as specified in ω. An agent's reward X_i for a service level s could denote her instantaneous revenue, which is affected by exogenous stochastic factors such as traffic, machine failures, etc., but it concentrates around her expected revenue, i.e. her value, $v_i(s)$. A strategic agent who cannot control such stochastic effects would hence submit bids so as to maximise her utility (expected reward minus price). Such PaaS services can take place in a competitive market or internally within an organisation where the provider is one team providing a service to other teams.

Example 2 (Online Advertising) A publisher (mechanism) has a set of advertising slots $\mathcal{S}$ and must assign them to n advertisers (agents). Typically, $|\mathcal{S}| \ll n$ and there exists $\varnothing \in \mathcal{S}$ indicating no assignment. When a slot is assigned to an advertiser, her reward is her instantaneous revenue which is simply the number of people who clicked the ad and then purchased the product. Consequently, it is a random quantity. Different agents could have different values for different slots. Ω is the set of possible ways in which the mechanism can assign slots to advertisers.

Henceforth, when we say that an agent is truthful, we mean that she reports her values truthfully, whereas when we say that a mechanism is truthful, we mean that it incentivises truthful behaviour from the agents. We are now ready to describe the learning problem when agents do not know their values, but when the mechanism is repeated for multiple rounds.

3.2 Learning a VCG mechanism under bandit feedback from agents

In the multi-round setting, agent and seller values $\{v_i\}_{i=0}^n$ remain fixed throughout all the rounds. On round t, the mechanism chooses an outcome $\omega_t \in \Omega$ and sets prices $\{p_{it}\}_{i=1}^n$ for the agents. Then, agent i realises a stochastic reward X_{it} which has expectation $v_i(\omega_t)$. At the end of the round, she reports a reward Y_{it}; if she is being truthful, she would report $Y_{it} = X_{it}$, but she may also choose to misreport the reward. While the agent does not know her values v_i, by reporting the reward at the end of each round, the mechanism could learn these values over multiple rounds.

While the primary focus of this work is on agents who do not know their values, our mechanism can also accommodate agents who know their values up front. Hence, we will also permit agents to submit bids $b_i : \mathcal{S} \to [0, 1]$ (not necessarily $b_i = v_i$) which represent her values for all rounds; she may do so once before the first round. We will refer to agents who submit rewards at the end of each round as those participating *by rewards*, and those who submit bids once at the beginning as those participating *by bids*. As we will see shortly, stronger results are possible for agents who participate by bids as their values need not be learned.

When choosing outcome ω_t on round t, the mechanism may use the information gathered from previous rewards $\{X_{i\ell}\}_{\ell=1}^{t-1}$ for agents participating by rewards and the bids b_i for agents participating by bids. The utility of agent i on round t is $u_{it} = \mathbb{E}[X_{it}] - p_{it} = v_i(\omega_t) - p_{it}$, where the expectation is only with respect to the rewards (exogenous stochasticity) at round t. The utility of the seller is $u_{0t} = v_0(\omega_t) + \sum_{i=1}^n p_{it}$. Let U_{iT}, U_{0T}, defined below, denote the sum of utilities of agent i and the mechanism respectively over T rounds. We have:

$$U_{iT} = \sum_{t=1}^{T} u_{it}, \qquad\qquad U_{0T} = \sum_{t=1}^{T} u_{0t}. \tag{2}$$

Our goal is to design an anytime algorithm which imitates the VCG mechanism over time. To that end, we quantify the performance of an algorithm via the following regret terms, defined relative to the VCG mechanism (1), after T rounds of interactions:

$$R_T^{\mathrm{w}} = TV(\omega_\star) - \sum_{t=1}^{T} V(\omega_t), \qquad R_{iT} = Tu_{i\star} - U_{iT}, \qquad R_T^{\mathrm{a}} = \sum_{i=1}^{n} R_{iT},$$
$$R_{0T} = Tu_{0\star} - U_{0T}, \qquad R_T^{\mathrm{VCG}} = \max(nR_T^{\mathrm{w}}, R_T^{\mathrm{a}}, R_{0T}). \tag{3}$$

Here $\omega_\star$ is the optimal outcome (1), which we will assume is unique. Moreover, $u_{0\star} = v_0(\omega_\star) + \sum_i p_{i\star}$ and $u_{i\star} = v_i(\omega_\star) - p_{i\star}$ are the utilities of the seller and agent i respectively in the VCG mechanism. R_T^{w} is the welfare regret over T rounds; it measures the welfare of the chosen outcomes ω_t relative to $\omega_\star$. R_{iT} is the regret of agent i and R_{0T} is the regret of the seller, both defined relative to the VCG mechanism. R_T^{a} is the sum of all agents' regrets. Finally, we also define the *VCG regret* R_T^{VCG}. In (2) and (3), we have followed pseudo-regret convention, which takes an expectation with respect to the rewards at the current round.

Our goal is to imitate the VCG mechanism over time, and R_T^{VCG} captures how well the welfare, and all agent/seller utilities converge uniformly to their VCG values. As we will see shortly, R_T^{VCG} will be a fundamental quantity in this problem, and we will use it to establish a hardness result. We focus on the VCG mechanism because it is one of the well-studied paradigms in multi-parameter mechanism design and is therefore a natural starting point. Moreover, even in competitive markets, sellers may be motivated to maximise welfare for long-term customer retention. This is similar in spirit to Devanur and Kakade (2009) who study a seller's regret, and Weed et al. (2016) who study an agent's regret when the agent bids in a repeated single item auction—in both cases, the regret is defined relative to the Vickrey auction.

A truthful agent simply reports her rewards at the end of each round. In general though, a strategic agent follows some strategy π so as to maximise her sum of utilities over several rounds. If she is participating by rewards, π is a map from her past information $\{(s_i(\omega_\ell), p_{i\ell}, X_{i\ell}, Y_{i\ell})\}_{\ell=1}^{t-1}$ and current allocation, price, and reward $(s_i(\omega_t), p_{it}, X_{it})$ to a (possibly random) scalar to report as Y_{it}. In particular, the agent may adopt a non-truthful strategy π, where she misreports her reward at the end of the current round so as to manipulate the allocations she may receive in future rounds, with the intent of maximising her long-term utility U_{iT}^{π} for large T. We also mention that if an agent is participating by bids, π is simply the bid that she submits ahead of time.

In addition to obtaining sublinear VCG regret (3), we would like to achieve the three desiderata for mechanism design given in Section 3.1. Here we define variants of those desiderata in order to precisely delineate the extent to which they can be achieved in our setting.

1. *Truthfulness*: Let U_{iT} and U_{iT}^{π} respectively denote the sum of utilities of agent i when she is being truthful and when she is following any other (non-truthful) strategy π. A mechanism is *truthful*, if, for all π, T, $U_{iT}^{\pi} \leq U_{iT}$ almost surely (a.s), regardless of the behaviour of other agents. It is *asymptotically truthful* if, for all π, T, $\mathbb{E}[U_{iT}^{\pi} - U_{iT}] \in o(T)$, regardless of the behaviour of other agents. A mechanism is *asymptotically Nash incentive-compatible* (NIC) if, for all π, T, $\mathbb{E}[U_{iT}^{\pi} - U_{iT}] \in o(T)$, when the other agents are behaving truthfully.

2. *Individual rationality:* Assume that agent i is truthful. A mechanism is *individually rational* if, for all T, $U_{iT} \geq 0$ a.s, regardless of the behaviour of other agents. It is *asymptotically individually rational* if $\lim_{T\to\infty} \frac{1}{T}\mathbb{E}[U_{iT}] \geq 0$, regardless of the behaviour of other agents.

3. *Efficiency:* A mechanism is *asymptotically efficient* if $\mathbb{E}[R_T^{\mathrm{w}}] \in o(T)$ when all agents are reporting truthfully.

To undestand the difference between the almost-sure and in-expectation definitions above, recall that $u_{it} = v_i(\omega_t) - p_{it}$ contains an expectation with respect to the reward at round t, but is a random quantity as the outcome ω_t and price p_{it} depend on the rewards realised/reported by all agents in previous rounds. In our almost sure definitions above, the statements should hold regardless of this randomness, whereas in our in-expectation definitions, they need to hold in expectation over the past exogenous randomness.

While achieving dominant-strategy incentive-compatibility is a desirable goal, it can be difficult, especially in multi-round mechanisms (Babaioff et al., 2014, 2013). A common approach to sidestep this difficulty is to adopt a Bayesian formalism which assumes that agent values are drawn from *known* prior beliefs and consider *ex ante* or *ex interim* versions of incentive-compatibility. However, Bayesian assumptions can be strong (Schummer, 2004) as it may not be possible to know the prior distributions ahead of time. In contrast, we do not make such distributional assumptions, but rely on asymptotic notions of truthfulness to make the problem tractable. If a mechanism is asymptotically truthful, the maximum value an agent may gain by not being truthful vanishes over time. In many applications it is reasonable to assume that agents would be truthful if the benefit of deviating is negligible, especially in settings where they may not know their value. It is worth pointing out that prior work has explored similar ideas of approximate incentive-compatibility in various contexts (Nazerzadeh et al., 2008; Lipton et al., 2003; Kojima and Manea, 2010; Roberts and Postlewaite, 1976; Feder et al., 2007; Daskalakis et al., 2006).

Finally, we will define two problem-dependent terms for what follows. First, let K be the minimum number of rounds necessary to assign all allocations to all agents. In Example 1, we can do this in $K = |\mathcal{S}|$ rounds, provided that there are no constraints on assigning different service levels to different agents. In Example 2, this can be done in $K = n$ rounds if $|\mathcal{S}| < n$. Second, let $V_{\max}$ be an upper bound on the expected welfare. Since $v_i(s) \in [0, 1]$, $V_{\max}$ could be as large as $\mathcal{O}(n)$. However, it can be small in settings such as Example 2 where it is $\mathcal{O}(1)$ if there is only one ad slot.

We make two observations before we proceed. First, we consider a fairly unadorned version of this problem as it provides the simplest platform to study how truthfulness and individual rationality

constraints affect learning in this setting. One could study richer models which assume structure between the allocations in $\mathcal{S}$ or that the values v_i change on each round. For instance, we may assume that $\mathcal{S} \subset \mathbb{R}^d$ and that v_i is either linear or smooth in these d attributes. We may also consider variations which incorporate changing values and/or contextual information. While these settings are beyond the scope of this work, we believe the analysis techniques and intuitions developed in this work would be useful in analysing such settings. Second, while our feedback model requires agents to share their observed reward at the end of each round, this is not too dissimilar from agents sharing their values in mechanism design. For instance, in Example 2, in usual truthful mehanisms, the agents would share their expected revenue from an ad slot when its known, whereas in our setting they would submit their instantaneous revenue on each round so that its expectation can be learned.

4. A Hardness Result

We first establish a lower bound on the VCG regret, defined in (3), even when all agents are truthful. To formalise this, let Θ be the class of problems with n agents, and $\mathcal{A}$ be the class of algorithms for this setting. Note that the regret terms in (3) depend on the specific problem in Θ and algorithm in $\mathcal{A}$.

Theorem 1 *Let $n \geq 2$ and assume all agents are truthful. Let the VCG regret R_T^{VCG} be as defined in* (3)*. Then, for $T \geq 128n$,*

$$\inf_{\mathcal{A}} \sup_{\Theta} \mathbb{E}\left[R_T^{\mathrm{VCG}}\right] \geq \frac{1}{50}(n-1)^{4/3}T^{2/3}.$$

The above result on the VCG regret captures the interaction between the welfare, agent, and sellller regret terms in (3). For instance, regardless of the chosen outcome, the seller can achieve small regret by demanding large payments from the agents; however, this will result in large agent regret. Hence, there is a natural trade-off between agent and seller regrets, which is determined by how the prices $\{p_{it}\}_{i,t}$ are set when the VCG prices $\{p_{i\star}\}_i$ are unknown[1]. Achieving small VCG regret requires that we estimate the prices accurately, which in turn requires that we estimate the welfare of the optimal outcomes without each given agent (1). Unlike in typical bandit settings, where $\mathcal{O}(\sqrt{T})$ regret is possible since it is sufficient to simply choose the optimal outcome, simultaneously estimating optimal outcomes without each agent necessitates incurring $\Omega(T^{2/3})$ regret.

It is necessary to study $\max(nR_T^{\mathrm{w}}, R_T^{\mathrm{a}}, R_{0T})$ instead of $\max(R_T^{\mathrm{a}}, R_{0T})$ as we need to account for the fact that the value being shared by the agents and the mechanism is constrained by the total welfare generated, which is factored into the welfare regret R_T^{w}. Precisely, the total welfare generated over T rounds is $TV(\omega_\star) - R_T^{\mathrm{w}}$ and not the maximum achievable $TV(\omega_\star)$ when the values are known.

It is instructive to compare this result with prior lower bounds in similar settings where learning happens on the mechanism side (Babaioff et al., 2014, 2015; Devanur and Kakade, 2009). In the online advertising setting described in Section 2, Babaioff et al. (2014) show that $\Omega(T^{2/3})$ welfare regret is unavoidable for deterministic a.s. truthful mechanisms. Devanur and Kakade (2009) establish a similar lower bound for the seller regret, defined relative to the seller's revenue in a

1. In fact, we will also see this phenomenon manifest in our algorithm, where, while there is flexibility to handle this trade-off in a way that is favourable to either the agent or the seller, the maximum of R_T^{a} and R_{0T} is always large (Proposition 6).

Vickrey auction for online advertising. Both hardness results rely on a necessary and sufficient condition for truthfulness in single parameter auctions (Archer and Tardos, 2001; Myerson, 1981). In contrast, our result for the VCG regret is obtained by studying the estimation error of the prices and applies to the maximum of the welfare, agent, and seller regrets, even when agents are reporting truthfully. Moreover, while our result applies to the VCG regret for general mechanisms, their results apply to the welfare/seller regrets only in the online advertising use case described in Section 2. Babaioff et al. (2015) design a randomised multi-round mechanism for this online advertising use case which is truthful in expectation and achieves $\mathcal{O}(\sqrt{T})$ welfare regret. This result does not contradict our result above which, as described in Section 2, considers a different feedback model and additionally accounts for the agent and seller regrets along with the welfare regret.

Proof sketch of Theorem 1: Minimising all regret terms requires that we estimate the VCG prices (1) correctly, which is the main bottleneck as the best outcome omitting any given agent might be very different from the optimal outcome. We first use a series of manipulations to lower bound the VCG regret via $R_T^{\mathrm{VCG}} \geq n\mathbb{E}R_T^{\mathrm{w}} + \mathbb{E}W_T$ where R_T^{w} is the welfare regret and W_T captures how well we have estimated the prices. These two terms are conflicting: keeping $\mathbb{E}R_T^{\mathrm{w}}$ small requires that we repeatedly choose the optimal outcome $\omega_\star$, but if we do so then we may not be able to estimate the optimal outcomes omitting each agent, leading to large $\mathbb{E}W_T$. We reduce the task of minimising the supremum of this sum over Θ to a binary hypothesis testing problem between two carefully chosen problems in Θ. We then apply a high probability version of Fano's inequality to obtain the result. The complete proof is given in Section 8.

5. Algorithm

We now describe our algorithm for this setting, called VCG-Learn, which is outlined in Algorithm 1. The algorithm has two binary hyperparameters λ, ζ, which control the trade-offs between the agent and seller regrets and properties such as truthfulness and individual rationality. We will first describe the algorithm then explain how these hyperparameters may be used to control the above properties.

VCG-Learn proceeds over a sequence of brackets. Brackets are indexed by q and rounds by t. Each bracket begins with an explore-phase of K rounds where the mechanism assigns all allocations in $\mathcal{S}$ to all agents at least once. It does not charge the agents during this phase but collects their realised rewards. This is then followed by $\lfloor \frac{5}{6} K q^{1/2} \rfloor$ rounds, during which the mechanism sets the outcome and prices based on the rewards collected thus far. The outcomes in the latter phase are chosen dependent on hyperparameter ζ; if $\zeta = \textsf{ETC}$ (explore-then-commit), we only use rewards from the explore phase to determine the outcomes, whereas when $\zeta = \textsf{OPT}$ (optimistic), we use rewards from all rounds thus far. By proceeding in brackets in the above manner, we are able to optimally control the time spent in the different phases. As the length of the latter phase increases with each bracket, we spend more rounds in this phase than in the explore phase as the mechanism is repeated for longer.

To describe how we compute the outcome, we first define the following three quantities, $\overline{v}_{it}, \widehat{v}_{it}, \breve{v}_{it} : \mathcal{S} \to [0, 1]$. For an agent participating by rewards, $\overline{v}_{it}(s)$ is the sample mean of the rewards when agent i was assigned outcome $s \in \mathcal{S}$, which serves as an estimate for $v_i(s)$. Next, $\widehat{v}_{it}(s)$ and $\breve{v}_{it}(s)$ are upper and lower confidence bounds respectively for $v_i(s)$. They are computed as shown in (4). Below, D_t denotes the round indices belonging to explore phases up to round $t-1$ when $\zeta = \textsf{ETC}$ and $D_t = \{1, \ldots, t-1\}$ when $\zeta = \textsf{OPT}$. $N_{it}(s)$ denotes the number of observations from agent i

for allocation s in the first $t-1$ rounds that are used in the computation for $\overline{v}_{it}, \widehat{v}_{it}, \breve{v}_{it}$. σ is the sub-Gaussian constant for the reward distributions (see Section 3.1), The $\overline{v}_{it}, \widehat{v}_{it}, \breve{v}_{it}$ quantities are first computed when $t > K$ and $N_{it}(s) \geq 1$ so they are well defined. We have:

$$N_{it}(s) = \sum_{\ell \in D_t} \mathbb{1}\big(s_i(\omega_\ell) = s\big), \quad \overline{v}_{it}(s) = \text{clip}\left(\frac{1}{N_{it}(s)} \sum_{\ell \in D_t} X_{i\ell}\, \mathbb{1}\big(s_i(\omega_\ell) = s\big),\ 0,\ 1\right),$$
$$\widehat{v}_{it}(s) = \overline{v}_{it}(s) + \sigma\sqrt{\frac{5\log(t - qK + 1) + 2\log(|\mathcal{S}|)}{N_{it}(s)}},$$
$$\breve{v}_{it}(s) = \overline{v}_{it}(s) - \sigma\sqrt{\frac{5\log(t - qK + 1) + 2\log(|\mathcal{S}|)}{N_{it}(s)}}. \tag{4}$$

Since $v_i(s) \in [0, 1]$, we clip the initial estimate between 0 and 1 to obtain $\overline{v}_{it}$. We will assume that each agent experiences each allocation in $\mathcal{S}$ *exactly once* during the exploration phase at the beginning of each bracket. If an agent was assigned the same allocation multiple times, we will use the reported value of only one of them, picked arbitrarily. For an agent i who participates by bidding b_i, we simply set, for all $s \in \mathcal{S}$,

$$\overline{v}_{it}(s) = \widehat{v}_{it}(s) = \breve{v}_{it}(s) = b_i(s). \tag{5}$$

We now define $\widehat{V}_t$, an upper confidence bound on the welfare at time t. In line 8, the algorithm chooses the outcome which maximises $\widehat{V}_t$ in round t:

$$\widehat{V}_t(\omega) = v_0(\omega) + \sum\nolimits_{i=1}^{n} \widehat{v}_{it}(\omega). \tag{6}$$

Finally, we describe how the prices are computed in line 9, which depend on the hyperparameter $\lambda \in \{\textsf{AGE}, \textsf{SEL}\}$. First define the functions $f_{it}, g_{it} : \mathcal{S} \to [0, 1]$ for all i, t as follows: if $\lambda = \textsf{AGE}$ (agent favourable pricing), set $f_{it} = \breve{v}_{it}$ and $g_{it} = \widehat{v}_{it}$; if $\lambda = \textsf{SEL}$ (seller favourable pricing), set $f_{it} = \widehat{v}_{it}$ and $g_{it} = \breve{v}_{it}$. Then define F_t^{-i}, G_t^{-i} as follows:

$$F_t^{-i}(\omega) = v_0(\omega) + \sum_{j \neq i} f_{jt}(s_j(\omega)), \qquad G_t^{-i}(\omega) = v_0(\omega) + \sum_{j \neq i} g_{jt}(s_j(\omega)). \tag{7}$$

As described in line 9, we charge price $p_{it} = \max_{\omega \in \Omega} F_t^{-i}(\omega) - G_t^{-i}(\omega)$ from agent i on rounds t that are not in the exploration phase. This completes the description of the algorithm, outlined in Algorithm 1.

To warm us up for the theoretical analysis in the next section, we discuss the implications of the hyperparameter choices ζ, λ. First, when $\zeta = \textsf{ETC}$, Algorithm 1 behaves similarly to explore-then-commit-style bandit algorithms (Perchet et al., 2013). It first explores all options at the beginning of each bracket. It then switches to an exploit phase for the remainder of the bracket during which it commits to the best outcome found during previous explore phases[2]. The main advantage of this

2. As all agents experience all allocations exactly once during each explore phase, when $\zeta = \textsf{ETC}$, the maximiser of the upper confidence bound $\widehat{V}_t$ and the mean $\overline{V}_t = v_0 + \sum_i \overline{v}_{it}$ coincide.

Algorithm 1 VCG-Learn

Require: $\zeta \in \{\mathsf{ETC}, \mathsf{OPT}\}$, $\lambda \in \{\mathsf{AGE}, \mathsf{SEL}\}$. # ζ, λ are hyperparameters. See lines 9 and 10.

1: Collect bids $\{b_i\}_i$ from agents participating by bids.
2: $t \leftarrow 0$. # t indexes rounds
3: **for** $q = 1, 2, \ldots,$ **do** # q indexes brackets
4: *Explore phase:* Assign each allocation $s \in \mathcal{S}$ to each agents at least once and charge them price 0 on each round. Collect reported rewards $\{Y_{it}\}_{i,t}$ from agents participating by rewards.
5: $t \leftarrow t + K$. # K is the number of rounds required for the explore phase. See Sec. 3
6: **for** $r = 1, \ldots, \lfloor \frac{5}{6} K q^{1/2} \rfloor$, **do**
7: $t \leftarrow t + 1$.
8: Choose outcome $\omega_t \leftarrow \operatorname{argmax}_\omega \widehat{V}_t(\omega)$. # $\widehat{V}_t$ is a UCB on the welfare. See (4),(5),(6)
9: Charge each agent i, $p_{it} \leftarrow \max_\omega F_t^{-i}(\omega) - G_t^{-i}(\omega_t)$. # λ determines computation of F_t^{-i} and G_t^{-i} respectively. See (7)
10: **if** $\zeta = \mathsf{OPT}$ **then**
11: Collect reported reward $\{Y_{it}\}_i$ from each agent i participating by rewards.
12: **end if**
13: **end for**
14: **end for**

two-phase strategy is a clean separation between preference learning and outcome/pricing selection which gives rise to strong truthfulness guarantees. When $\zeta = \mathsf{OPT}$, the procedure is reminiscent of optimistic strategies (Lai and Robbins, 1985) which maximise an upper confidence bound using rewards from all rounds. Not only is this empirically sample-efficient as it uses rewards from all rounds, but it also enjoys better welfare regret and individual rationality properties over $\zeta = \mathsf{ETC}$ as we shall demonstrate shortly. Unfortunately, this comes at the cost of weaker guarantees on truthfulness. We will elucidate this in Section 6. While optimistic strategies do not usually require an explore phase, this is necessary in our problem to accurately estimate the prices and to guarantee asymptotic NIC. Consequently, our bounds on the welfare, agent, seller regrets are worse than the typical $\sqrt{T}$ rates one comes to expect of optimistic strategies in stochastic bandit problems.

Next, consider λ, which is used in computing the F_t^{-i}, G_t^{-i} quantities (7), and consequently determines the pricing calculation in line (9) of Algorithm 1. While λ does not affect the outcome and the welfare generated on each round, it determines how this welfare is shared between the agents and the seller; therefore, it affects the agent and seller regrets R_{iT}, R_{0T}. For instance, suppose we choose $\lambda = \mathsf{SEL}$. In line (9), this uses the most optimistic estimate of the maximum welfare omitting agent i in F_t^{-i}, and the most pessimistic estimate of the values of the current outcome for the other agents in G_t^{-i}. This results in large payments and consequently is the most favourable pricing scheme to the seller, while still ensuring asymptotic truthfulness, individual rationality, and sublinear agent regret. Similarly, when $\lambda = \mathsf{AGE}$, the pricing is favourable to the agents. We will illustrate these trade-offs, along with their effects on individual rationality and truthfulness, in the next section. These options give a practitioner a fair amount of flexibility when applying Algorithm 1 for their specific use case.

6. Main Results for Algorithm 1

We now present our main theoretical results for VCG-Learn, providing rates for asymptotic truthfulness, individual rationality, and VCG regret in Sections 6.1, 6.2, and 6.3 respectively. In Section 6.4 we also provide bounds on the agent, seller, and welfare regrets defined in (3). We wish to remind the reader that Table 1 summarises the main results (Theorems 2, 3, and 4) of this section. The proofs of all results are given in Section 9.

6.1 Asymptotic truthfulness

We first state the truthfulness/NIC properties of the proposed algorithm. Theorem 2 establishes that Algorithm 1 is asymptotically truthful when $\zeta = \texttt{ETC}$ and is asymptotically NIC when $\zeta = \texttt{OPT}$. In fact, we will state a slightly stronger result for the $\zeta = \texttt{OPT}$ case. For this, we say that a strategy π by agent j is *stationary* if she either participates by bids, or if participating by rewards, when assigned an allocation $s \in \mathcal{S}$, she reports a sample from some fixed distribution dependent on s. Any other strategy is *non-stationary*. Intuitively, when an agent participates by rewards, if we view the rewards reported for any allocation as a time series, the strategy is stationary if this time series is stationary.

While truthfulness implies stationarity, a non-truthful player can be either stationary or non-stationary. For example, when participating by rewards, an agent may choose to report $Y_{it} = \phi_s(X_{it})$ when assigned an allocation s, where the functions $\{\phi_s\}_s \in \mathcal{S}$ may be designed to squash or amplify rewards for certain allocations, say, so as to discourage or encourage the mechanism from assigning said allocation to the agent in the future. Such reports, while non-truthful, come from a stationary distribution. An agent is also stationarily non-truthful if, when participating by bids, she submits false values. We have the following theorem.

Theorem 2 *Let π be any non-truthful strategy for agent i. Fix the strategies adopted by the other agents. Let U^{π}_{iT}, U_{iT} be the sum of agent i's utilities when she follows π and when being truthful respectively. The following statements hold for any $\lambda \in (\texttt{AGE}, \texttt{SEL})$ for all $T > 2K$.*

1. *First let $\zeta = \texttt{ETC}$. If an agent participates by bids, then regardless of the behaviour of others, then $U^{\pi}_{iT} - U_{iT} \leq 0$ a.s; i.e. Algorithm 1 is truthful. If the agent participates by rewards, then, regardless of the behaviour of the others, we have,*

$$\mathbb{E}[U^{\pi}_{iT} - U_{iT}] \leq 10\sigma\sqrt{\log(|\mathcal{S}|T)}K^{1/3}T^{2/3} + 4 \ \in \ \tilde{\mathcal{O}}\left(K^{1/3}T^{2/3}\right).$$

2. *Next, let $\zeta = \texttt{OPT}$ and assume that all agents other than i adopt stationary policies. Then, for any (stationary or non-stationary) strategy π for agent i,*

$$\mathbb{E}[U^{\pi}_{iT} - U_{iT}] \leq 10\sigma(6n+2)\sqrt{\log(|\mathcal{S}|T)}K^{1/3}T^{2/3} + 12n \ \in \ \tilde{\mathcal{O}}\left(nK^{1/3}T^{2/3}\right).$$

The above imply that Algorithm 1 is asymptotically truthful when $\zeta = \texttt{ETC}$ and asymptotically Nash incentive-compatible when $\zeta = \texttt{OPT}$ for an agent participating by rewards.

The guarantees when $\zeta = \texttt{OPT}$ is weak when compared to $\zeta = \texttt{ETC}$ in two regards. Not only does the asymptotic bound scale with n, but it also holds only when the other agents are adopting stationary policies. However, since truthfulness implies stationarity, it does imply an asymptotic Nash equilibrium; that is, when $\zeta = \texttt{OPT}$, if all other agents are truthful, then the amount by which an agent stands to gain by misreporting her rewards vanishes over multiple rounds.

However, as we will see shortly, when $\zeta = \textsf{OPT}$, we have better empirical results and theoretical bounds on the welfare regret, individual rationality and the agent and seller regrets since we use data from all rounds. Using only a small fraction of the data can be wasteful if we do not expect agents to be very strategic. The $\zeta = \textsf{OPT}$ option is primarily motivated by this practical consideration. It allows us to efficiently learn in such environments, while providing some protection against a nontruthful agent, not just in settings where the other agents are being truthful, but also when they may try to manipulate the mechanism with "simple" methods, such as squashing/amplifying their rewards for certain allocations.

Proof sketch: We write the instantaneous difference in the utilities as $u_{it}^{\pi} - u_{it} = (u_{it}^{\pi} - \widetilde{u}_{it}) + (\widetilde{u}_{it} - u_{it})$. Here, u_{it}^{π} is the utility on round t when the agent reports according to strategy π up to round $t-1$, $\widetilde{u}_{it}$ is the utility on round t when she follows π up to round $t-2$ and switches to truth-telling on round $t-1$, and u_{it} is the utility on round t when she is truthful on all rounds. The first term captures the benefit of misreporting in the current round; this can be bound using proof techniques for truthfulness of the VCG mechanism. The latter term captures the benefit of misreporting in previous rounds; this can be large, since, an agent's false reports will have affected the outcomes and prices chosen by the mechanism not just in the current round but in previous rounds as well. To control this term, we use properties of our algorithm to show that the agent's past actions cannot have changed the outcomes by too much; for instance, for the harder $\zeta = \textsf{OPT}$ case, this term is dominated by values reported by the other agents; since they are adopting stationary policies, the reported values concentrate around the respective means which cannot be influenced by the agent.

6.2 Asymptotic individual rationality

Our next theorem establishes the asymptotic individual rationality properties of Algorithm 1.

Theorem 3 *Consider any agent i. Let U_{iT} be the sum of her utilities after T rounds when she participates truthfully (while others may not). The following statements are true for all $T > 2K$.*

***1.** First let $\zeta = \textsf{ETC}$. When $\lambda = \textsf{AGE}$, $U_{iT} \geq 0$ a.s for all T for an agent participating by bids; i.e, Algorithm 1 is (almost surely) individually rational. For an agent participating by rewards,*

$$\mathbb{E}[U_{iT}] \geq -10\sigma\sqrt{\log(|\mathcal{S}|T)}K^{1/3}T^{2/3} - 4, \qquad \textit{i.e. } \mathbb{E}[-U_{iT}] \in \tilde{\mathcal{O}}\left(K^{1/3}T^{2/3}\right)$$

That is, Algorithm 1 is asymptotically individually rational. Similarly, when $\lambda = \textsf{SEL}$,

$$\mathbb{E}[U_{iT}] \geq -10\sigma n\sqrt{\log(|\mathcal{S}|T)}K^{1/3}T^{2/3} - 4, \qquad \textit{i.e. } \mathbb{E}[-U_{iT}] \in \tilde{\mathcal{O}}\left(nK^{1/3}T^{2/3}\right)$$

***2.** Next, let $\zeta = \textsf{OPT}$. When $\lambda = \textsf{AGE}$, for all agents, $U_{iT} \geq 0$ a.s for all T for an agent participating by bids; i.e, Algorithm 1 is individually rational. Moreover, for an agent participating by rewards,*

$$\mathbb{E}[U_{iT}] \geq -9\sigma\sqrt{|\mathcal{S}|T\log(|\mathcal{S}|T)} - 6, \qquad \textit{i.e. } \mathbb{E}[-U_{iT}] \in \tilde{\mathcal{O}}\left(|\mathcal{S}|^{1/2}T^{1/2}\right)$$

That is, Algorithm 1 is asymptotically individually rational. Similarly, when $\lambda = \textsf{SEL}$,

$$\mathbb{E}[U_{iT}] \geq -9\sigma n\sqrt{|\mathcal{S}|T\log(|\mathcal{S}|T)} - 6, \qquad \textit{i.e. } \mathbb{E}[-U_{iT}] \in \tilde{\mathcal{O}}\left(n|\mathcal{S}|^{1/2}T^{1/2}\right).$$

While the above theorem implies asymptotic individual rationality for all ζ, λ values, let us consider how the different hyperparameter choices affect the rates in the theorem. We see that when $\zeta = \texttt{OPT}$, the $\mathcal{O}(T^{1/2})$ rates are better than when $\zeta = \texttt{ETC}$, where the rate is $\mathcal{O}(T^{2/3})$. This demonstrates the first trade-off determined by the ζ hyperparameter: when $\zeta = \texttt{ETC}$, we have stronger truthfulness guarantees but weaker individual rationality guarantees than when $\zeta = \texttt{OPT}$. The stronger rates are possible in the latter case because we use all data to learn an agent's preferences. Next, when $\lambda = \texttt{SEL}$, the asymptotic rates for individual rationality have an additional n dependence than when $\lambda = \texttt{AGE}$. In the former case, the agents bear the brunt of the uncertainty in the price estimation leading to worse rates; we will see this manifest in the agent and seller regret bounds as well in Section 6.4. Finally, we also see that when $\lambda = \texttt{AGE}$, the individual rationality holds exactly and almost surely for agents participating by bids while it only does so asymptotically for agents participating by rewards. Hence, if an agent knows her values, she is better off submitting them as bids up front.

It is also worth highlighting that the above bounds above have dependence on the size of the allocation set $|\mathcal{S}|$ and not the size of the outcomes $|\Omega|$ (recall from Section 3 that K also may depend on $|\mathcal{S}|$, but not $|\Omega|$). While $|\Omega|$ can be quite large, possibly as large as $|\mathcal{S}|^n$, the rates scale with $|\mathcal{S}|$ since the updates to the means and confidence intervals for one agent occur independent of the rewards observed by the others (4).

Proof sketch: All agents have non-negative utility in the exploration phase so we can restrict our attention to rounds not in the exploration phase. We first show that we can decompose the utility of agent i on round t as $u_{it} = c_t + d_t$ where $c_t = v_i(s_i(\omega_t)) - g_{it}(s_i(\omega_t))$ and $d_t = G_t(\omega_t) - \max_\omega F_t^{-i}(\omega)$. Intuitively, c_t, if negative, can be viewed as negative utility that an agent may accrue due to the mechanism mis-estimating her values, and d_t, if negative, can be viewed as negative utility that an agent may accrue due to the mechanism mis-estimating the values of the other agents, and consequently the prices. When $\lambda = \texttt{SEL}$, we show that c_t is small (its sum can be bound by a constant) but d_t is large; in this case, the agents bear most of the effects of uncertainty in values leading to large asymptotic rates which scale with n. When $\lambda = \texttt{AGE}$, d_t is small and c_t is large; however, as the seller bears the effects of the uncertainty, it does not scale with n. In the remainder of the analysis we show that when $\zeta = \texttt{ETC}$, the information obtained in the explore phase rounds lead to a $T^{2/3}$ rate, whereas when $\zeta = \texttt{OPT}$, the information obtained in all rounds lead to a $T^{1/2}$ rate.

6.3 Bounding the VCG regret

Finally, we will upper bound the VCG regret R_T^{VCG} for Algorithm 1. Recall that R_T^{VCG} captures how well the welfare, all agent utilities and the seller utility converge uniformly to the VCG values.

Theorem 4 *Assume all agents are truthful. Let R_T^{VCG} be as defined in* (3)*. The following statements hold for any $\lambda \in (\texttt{AGE}, \texttt{SEL})$. Whe $\zeta = \texttt{ETC}$, for all $T > 2K$,*

$$\begin{aligned}\mathbb{E}[R_T^{\text{VCG}}] &\leq \left(3V_{\max}(n+3) + 10(5n^2+n)\sqrt{\log(|\mathcal{S}|T)}\right)K^{1/3}T^{2/3} + 4V_{\max}(n^2+3n) \\ &\in \tilde{\mathcal{O}}\left(n^2K^{1/3}T^{2/3}\right).\end{aligned}$$

Next, when $\zeta = \texttt{OPT}$, we have that for all $T > 2K$,

$$\mathbb{E}[R_T^{\text{VCG}}] \leq 9\sigma(3n^2+n)\sqrt{|\mathcal{S}|T\log(|\mathcal{S}|T)} + \left(3V_{\max}(n+3) + 20\sigma n^2\sqrt{\log(|\mathcal{S}|T)}\right)K^{1/3}T^{2/3}$$

$$+\ 6V_{\max}(n^2+3n) \qquad \in\ \tilde{\mathcal{O}}\left(n^2K^{1/3}T^{2/3}\right).$$

We find that for both choices of ζ, we have an $\tilde{\mathcal{O}}(n^2K^{1/3}T^{2/3})$ upper bound on the VCG regret R_T^{VCG}. It is worth noting that since we use data from all rounds, the constants in the higher order $n^2K^{1/3}T^{2/3}$ terms are smaller when $\zeta = \mathsf{OPT}$ than when $\zeta = \mathsf{ETC}$. While both upper bounds differ by a $\mathrm{poly}(n)$ factor from the lower bound in Theorem 1, it achieves the $T^{2/3}$ rate. This establishes minimax optimality for VCG-Learn.

Proof sketch: We first decompose the VCG regret as follows:

$$R_T^{\mathrm{VCG}} \lesssim nR_T^{\mathrm{w}} + \sum_{i=1}^{n}\sum_{t}\mathbb{E}[|A_t^{-i}|\,|\mathcal{E}_t] + n\sum_{t}\mathbb{E}[|B_t|\,|\mathcal{E}_t],$$

where, $A_t^{-i} = V^{-i}(\omega_\star^{-i}) - F_t^{-i}(\omega_\star^{-i})$, $B_t = G_t(\omega_t) - V(\omega_\star)$, and the $\lesssim$ notation ignores lower order terms. Here, B_t is due to the error in estimating the optimum outcome and A_t^{-i} is due to the error in estimating the optimum without agent i. This decomposition bounds the VCG regret in terms of the difference between the true values of the agents and their upper or lower confidence bounds. To bound the $\sum_t A_t^{-i}$ terms, we use the fact that the information obtained in the explore phase rounds lead to a $n^2T^{2/3}$ rate for both ζ choices. For the R_T^{w} and $\sum_t B_t$ terms, we similarly obtain a $nT^{2/3}$ rate when $\zeta = \mathsf{ETC}$ and a $nT^{1/2}$ rate when $\zeta = \mathsf{OPT}$.

6.4 Bounding the welfare, agent, and seller regrets

Finally, in this section, to better understand the behaviour of the algorithm under various hyperparameter choices, we individually bound the welfare, agent, and seller regrets defined in (3). While the VCG regret provides a bound on the welfare and seller regrets, we find that tighter bounds are possible based on the different hyperparameters. First, in Proposition 5, we bound the welfare regret.

Proposition 5 *Assume all agents are truthful. Let R_T^{w} be as defined in* (3)*. The following statements hold for any $\lambda \in (\mathsf{AGE}, \mathsf{SEL})$. When $\zeta = \mathsf{ETC}$, for all $T > 2K$,*

$$\mathbb{E}[R_T^{\mathrm{w}}] \le \left(3V_{\max} + 10n\sqrt{\log(|\mathcal{S}|T)}\right)K^{1/3}T^{2/3} + 4V_{\max}n\ \in\ \tilde{\mathcal{O}}\left(nK^{1/3}T^{2/3}\right).$$

Moreover, when $\zeta = \mathsf{OPT}$, for all $T > 2K$, the welfare regret satisfies,

$$\mathbb{E}[R_T^{\mathrm{w}}] \le 9n\sqrt{|\mathcal{S}|T\log(|\mathcal{S}|T)} + 3V_{\max}K^{1/3}T^{2/3} + 6V_{\max}n\ \in\ \tilde{\mathcal{O}}\left(n|\mathcal{S}|^{1/2}T^{1/2} + V_{\max}K^{1/3}T^{2/3}\right).$$

The above results imply that in both cases Algorithm 1 is asymptotically efficient.

When $\zeta = \mathsf{OPT}$, the welfare regret is $\tilde{\mathcal{O}}(n|\mathcal{S}|^{1/2}T^{1/2} + V_{\max}K^{1/3}T^{2/3})$ whereas, when $\zeta = \mathsf{ETC}$, it is $\tilde{\mathcal{O}}(nK^{1/3}T^{2/3})$. When $V_{\max} \in o(n)$, the former is better (recall from Section 3, that the maximum welfare $V_{\max} \in \mathcal{O}(n)$, but could be much smaller). More precisely, there are two factors contributing to the welfare regret: first, the rounds spent in the exploration phase during which the instantaneous regret may be arbitrarily bad; second, the effects of the estimation errors of the values. For both choices of ζ, the former can be bound by $V_{\max}K^{1/3}T^{2/3}$. In contrast, when $\zeta = \mathsf{OPT}$, the latter can be bound by $n|\mathcal{S}|^{1/2}T^{1/2}$ as we use data from all the rounds, whereas when $\zeta = \mathsf{ETC}$, it can only

be bound by $nK^{1/3}T^{2/3}$. We will see this effect empirically as well, with $\zeta = \mathsf{OPT}$ performing significantly better than $\zeta = \mathsf{ETC}$.

Next we will consider the agent and seller regrets. For this, we define $v_i^\dagger$ below, which can be used to bound the instantaneous regret of agent i during the exploration phase, i.e. $u_{i\star} - u_{it} \leq v_i^\dagger$. If the agent prefers any allocation $s \in \mathcal{S}$ for free than paying the VCG price (1) for the socially optimal outcome, she will incur no regret during the exploration rounds, and correspondingly, $v_i^\dagger = 0$.

$$v_i^\dagger = \max\big(v_i(\omega_\star) - p_{i\star} - \min_s v_i(s),\ 0\big). \tag{8}$$

Proposition 6 bounds the agent and seller regrets for the different ζ, λ choices.

Proposition 6 *Assume all agents are truthful. Let R_{iT} and R_T^{a} be as defined in* (3)*. Let $\kappa_i = 1$ if agent i participates by rewards and 0 if she participates by bids. The following statements hold after $T > 2K$ rounds for the ζ, λ choices specified.*

1. *Let $\zeta = \mathsf{ETC}$. Then, when $\lambda = \mathsf{AGE}$, we have*

$$\mathbb{E}[R_{iT}] \leq \big(3v_i^\dagger + 10\sigma\kappa_i\sqrt{\log(|\mathcal{S}|T)}\big)K^{1/3}T^{2/3} + 4n \ \in\ \tilde{\mathcal{O}}\left(K^{1/3}T^{2/3}\right)$$
$$\mathbb{E}[R_{0T}] \leq \big(3V_{\max} + 20\sigma n^2\sqrt{\log(|\mathcal{S}|T)}\big)K^{1/3}T^{2/3} + 4V_{\max}n \ \in\ \tilde{\mathcal{O}}\left(n^2K^{1/3}T^{2/3}\right)$$

If $\lambda = \mathsf{SEL}$, we have

$$\mathbb{E}[R_{iT}] \leq \big(3v_i^\dagger + 20\sigma n\sqrt{\log(|\mathcal{S}|T)}\big)K^{1/3}T^{2/3} + 4n \ \in\ \tilde{\mathcal{O}}\left(nK^{1/3}T^{2/3}\right),$$
$$\mathbb{E}[R_{0T}] \leq 3V_{\max}K^{1/3}T^{2/3} + 4V_{\max}n \ \in\ \tilde{\mathcal{O}}\left(V_{\max}K^{1/3}T^{2/3}\right)$$

2. *Let $\zeta = \mathsf{OPT}$. Then, when $\lambda = \mathsf{AGE}$, we have*

$$\mathbb{E}[R_{iT}] \leq 9\sigma\kappa_i\sqrt{|\mathcal{S}|T\log(|\mathcal{S}|T)} + 3v_i^\dagger K^{1/3}T^{2/3} + 6n \ \in\ \tilde{\mathcal{O}}\left(|\mathcal{S}|^{1/2}T^{1/2} + v_i^\dagger K^{1/3}T^{2/3}\right)$$
$$\mathbb{E}[R_{0T}] \leq 9\sigma n^2\sqrt{|\mathcal{S}|T\log(|\mathcal{S}|T)} + \big(3V_{\max} + 10\sigma n^2\sqrt{\log(|\mathcal{S}|T)}\big)K^{1/3}T^{2/3} + 6V_{\max}n$$
$$\in\ \tilde{\mathcal{O}}\left(n^2|\mathcal{S}|^{1/2}T^{1/2} + n^2K^{1/3}T^{2/3}\right).$$

If $\lambda = \mathsf{SEL}$, we have

$$\mathbb{E}[R_{iT}] \leq 9\sigma n\sqrt{|\mathcal{S}|T\log(|\mathcal{S}|T)} + \big(3v_i^\dagger + 20\sigma n\sqrt{\log(|\mathcal{S}|T)}\big)K^{1/3}T^{2/3} + 6n$$
$$\in\ \tilde{\mathcal{O}}\left(n|\mathcal{S}|^{1/2}T^{1/2} + nK^{1/3}T^{2/3}\right),$$
$$\mathbb{E}[R_{0T}] \leq 3V_{\max}K^{1/3}T^{2/3} + 6V_{\max}n \ \in\ \tilde{\mathcal{O}}\left(V_{\max}K^{1/3}T^{2/3}\right).$$

While, generally speaking, the agent and seller regrets scale at rate $T^{2/3}$, the dependence on other problem parameters are determined by the choices for ζ and λ. First consider the case $\zeta = \mathsf{ETC}$. If we choose $\lambda = \mathsf{SEL}$, which, as we explained before, is favourable to the seller, the seller's regret

scales at rate $V_{\max} K^{1/3} T^{2/3}$, with at most linear dependence on n. However, this is disadvantageous for an agent—her regret and asymptotic individual rationality bounds (Theorem 3) scale linearly with n. On the other hand, if we choose $\lambda = \texttt{AGE}$, then the agent regret is the smallest, but the seller suffers some disadvantageous consequences. Since $\breve{v}_{jt} \leq \widehat{v}_{jt}$, in line 9 of Algorithm 1, p_{it} could be negative, i.e., the seller makes a payment to the customer. This violates the no-positive-transfers property which is considered desirable in mechanism design. The seller's regret is also poor, with n^2 scaling. We may draw similar conclusions when $\zeta = \texttt{OPT}$, with the main difference being that some terms can be bounded by $T^{1/2}$ rates. It is also worth noting that when $\zeta = \texttt{OPT}$, for agents for whom $v_i^\dagger = 0$, we achieve $\sqrt{T}$ regret.

It is worth observing that while the welfare regret is simply the sum of the agent and seller regrets $R_T^{\mathrm{w}} = R_{0T} + \sum_{i=1}^n R_{iT}$ (see (3)), the bounds for R_T^{w} given in Proposition 5 is smaller than the sum of the agent and seller bounds in Proposition 5 for all ζ choices. For instance, when $\zeta = \texttt{ETC}$, we have $R_T^{\mathrm{w}} \in \tilde{\mathcal{O}}(nK^{1/3}T^{2/3})$, but naively summing the bounds on the agent and seller regrets would yield a bound $\tilde{\mathcal{O}}(n^2 K^{1/3} T^{2/3})$. This discrepancy can be explained by the fact that the prices do not affect the welfare and therefore the error in estimating the prices need not be accounted for in the welfare regret. However, the agent and seller utilities depend on the price, and consequently their regret bounds should account for this error. As we explained in Section 4, estimating the prices is one of the main bottlenecks in this set up. We are able to bound the welfare regret separately and obtain a better bound than the sum of individual regrets. For example, in our simulation in Figure 1, the regret of the mechanism and the first agent are fairly large while the regret of many of the other agents is negative. This highlights the fact that the regret of any one agent or the seller might be large due to the error in estimating the prices, even though the sum of these regret terms, which is the welfare regret, is small.

6.5 Discussion

It is worth contextualising the above results with prior work on mechanism design with bandit feedback in the online advertising setting. As explained in Section 2, these settings, where an agent submits a single bid ahead of time and the stochasticity is observed by the mechanism on each round, is different from ours, where the mechanism needs to rely on the agents to report their values on each round. In a fixed-horizon version of this problem, Babaioff et al. (2014) describe an almost surely truthful mechanism with $T^{2/3}$ welfare regret and Devanur and Kakade (2009) describe an almost surely truthful mechanism with $T^{2/3}$ seller regret. While they focus on a simpler problem and provide stronger truthfulness guarantees, it is worth noting that both works use an explore-then-commit style algorithm to guarantee truthfulness. Babaioff et al. (2015) describe a truthful-in-expectation mechanism with $\sqrt{T}$ welfare regret. However, they do not bound the agent and seller regrets.

Finally, we note that our algorithm and analysis assumes that seller values are known. If this is unknown, one can define lower and upper confidence bounds for the seller similar to (4) and use them in Algorithm 1 in place of v_0, similar to those of the agents. While $T^{2/3}$ rates are still possible, there are additional considerations. First, in many applications, it may not be reasonable to assume that this distribution has the same sub-Gaussian constant σ (e.g. PaaS); the variance of the seller might scale with n and this will invariably be reflected in the regret bounds, including that of the agents. Second, since Ω may be much larger than $\mathcal{S}$, this results in long exploration phases and worse regret bounds reflected via the parameter K.

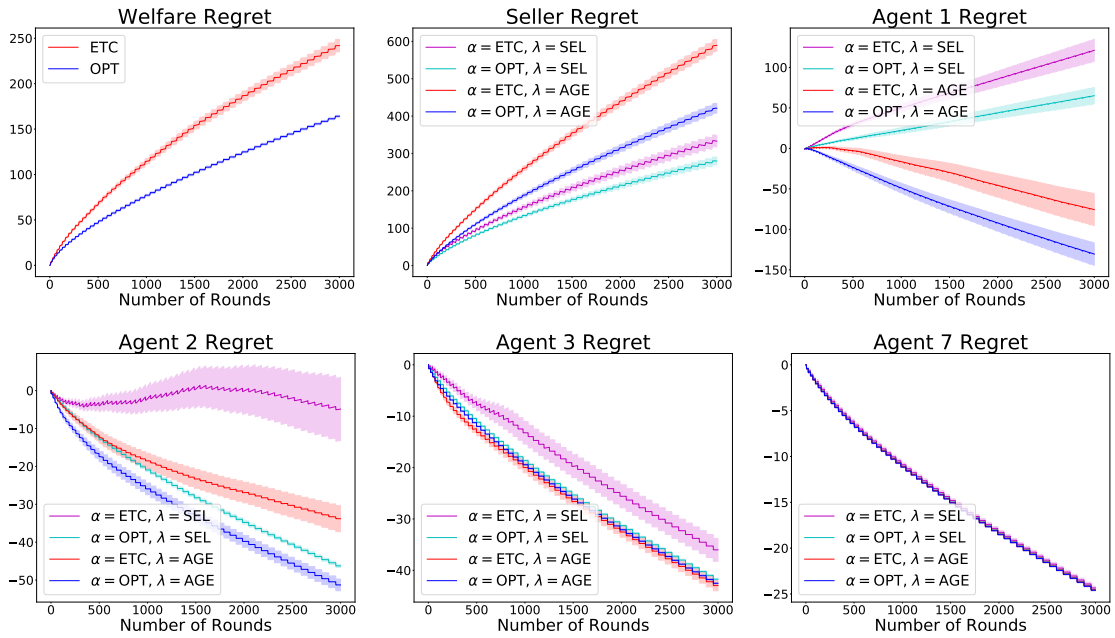

Figure 1: Results for a single-item simulation study. We show the welfare regret, the seller regret, and the regret of four agents for the four difference choices for (ζ, λ) over 3000 rounds. Lower is better in all plots. Agent 1 has the highest value, and if the values were known, the item would be assigned to agent 1. The figures were obtained by averaging over 50 independent runs and the shaded regions represent two standard errors. The jagged shape of the curves is due to the periodic exploration phase in the algorithm. Some of the curves overlap in the last two plots for agents 3 and 7.

7. A Simulation

We present some simulation results in a single-parameter single-item environment. Here, ten agents are competing for a single item and all of them are participating by rewards. When an agent receives the item, her value is drawn stochastically from a $\mathcal{N}(\mu, 0.5)$ distribution where μ is chosen uniformly on a grid in the interval $(0.2, 0.9)$. Agent 1 has a value of 0.9 for receiving the item (and will be the agent who receives the item if values are known) and agent 10 has a value of 0.2. If an agent does not receive the item, their value is non-stochastically zero. Observe that this is environment is rather noisy—the variance of the reward distribution is large when compared to the range of the values of the agents. The game is repeated for 3000 rounds.

We have shown the pseudo-regrets for the welfare, the seller, and some of the agents in Figure 1 for all possible choices of the ζ and λ hyperparameters. As we see, $\zeta = \textsf{OPT}$ performs better than $\zeta = \textsf{ETC}$ on all plots as it uses all the data. Moreover, we see that the agents have lower regret when $\lambda = \textsf{AGE}$ than when $\lambda = \textsf{SEL}$, and vice versa for the seller. The regrets of agents 2 to 10 decrease indefinitely leading to negative regret since their utility at the socially optimal outcome is zero, but they occasionally get the item assigned to them during the exploration phase.

8. Proof of Theorem 1

In this section, we present our proof of the lower bound in Section 4. We will first describe notation and definitions that will be used throughout our proofs in Sections 8 and 9.

Notation: $\mathbb{E}, \mathbb{P}$ will denote expectations and probabilities. $\mathbb{E}_t, \mathbb{P}_t$ will denote expectation and probability when conditioned on observations up to time $t-1$; for example, $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot|D_t)$, where $D_t = \{s_{i\ell}, Y_{i\ell}\}_{i\le n, \ell \le t-1}$.

Recall that $\omega_\star = \operatorname{argmax}_\omega V(\omega)$ is the socially optimal outcome. Let $s_{i\star} = s_i(\omega_\star)$ be the allocation for agent i at the optimum. Similarly, V^{-i} and $\omega_\star^{-i}$, defined below, will denote the welfare without agent i and its optimiser respectively.

$$V^{-i}(\omega) = v_0(\omega) + \sum_{j \neq i} v_i(\omega_t), \qquad \omega_\star^{-i} = \operatorname*{argmax}_{\omega \in \Omega} V^{-i}(\omega). \tag{9}$$

We will first state the following fact, which is straightforward to verify, regarding agent and seller utilities in the VCG mechanism.

Fact 7 *When the outcome and the prices are chosen according to the VCG mechanism,*

$$u_{i\star} = v_i(s_{i\star}) - p_{i\star} = V(\omega_\star) - V^{-i}(\omega_\star^{-i}),$$
$$u_{0\star} = v_0(\omega_\star) + \sum_{i=1}^n p_{i\star} = \sum_{i=1}^n V^{-i}(\omega_\star^{-i}) - (n-1)V(\omega_\star).$$

Our second result expresses the regret terms in (3) in a way that is convenient for analysis. For this, we define quantities H_T, W_T below.

$$H_T = \frac{1}{T}\sum_{i=1}^n \sum_{t=1}^T \left(p_{it} + V^{-i}(\omega_t)\right), \qquad W_T = H_T - \sum_{i=1}^n V^{-i}(\omega_\star^{-i}). \tag{10}$$

H_T is computed using observations from rounds 1 to T, and can be thought of as the algorithm's estimate of $\sum_i V^{-i}(\omega_\star)$ at the end of T rounds. The following lemma expresses R_T^{a} and R_{0T} in terms of R_T^{w} and W_T.

Lemma 8 *Let $R_T^{\mathrm{a}}, R_{0T}, R_T^{\mathrm{w}}$ be as defined in* (3)*. Then,*

$$R_T^{\mathrm{a}} = nR_T^{\mathrm{w}} + TW_T, \qquad R_{0T} = -(n-1)R_T^{\mathrm{w}} - TW_T.$$

Proof Let $h_{it} = p_{it} + V^{-i}(\omega_t)$ so that $H_T = \frac{1}{T}\sum_{i=1}^n \sum_{t=1}^T h_{it}$. For agent i, we can use Fact 7 and the fact that $u_{it} = v_i(\omega_t) - p_{it} = V(\omega_t) - h_{it}$ to obtain,

$$u_{i\star} - u_{it} = (V(\omega_\star) - V(\omega_\star^{-i})) - (V(\omega_t) - h_{it}) = (V(\omega_\star) - V(\omega_t)) + (h_{it} - V(\omega_\star^{-i})).$$

Then, since $R_T^{\mathrm{a}} = \sum_i \sum_t (u_{i\star} - u_{it})$, we have

$$R_T^{\mathrm{a}} = \sum_{t=1}^T \sum_{i=1}^n \left((V(\omega_\star) - V(\omega_t)) + (h_{it} - V(\omega_\star^{-i}))\right)$$

$$= n\sum_{t=1}^{T}(V(\omega_\star) - V(\omega_t)) \;+\; T\left(H_T - \sum_{i=1}^{n} V^{-i}(\omega_\star^{-i})\right).$$

This proves the first claim. For the seller, at time t, we observe

$$u_{0t} = v_0(\omega_t) + \sum_{i=1}^{n} p_{it} = v_0(\omega_t) + \sum_{i=1}^{n} h_{it} - \sum_{i=1}^{n} V^{-i}(\omega_t) = \sum_{i=1}^{n} h_{it} - (n-1)V(\omega_t).$$

As before, we can now use Fact 7 to write,

$$R_{0T} = \sum_{t=1}^{T}(u_{0\star} - u_{0t}) = \sum_{t=1}^{T}\sum_{i=1}^{n}\left(V^{-i}(\omega_\star^{-i}) - h_{it}\right) + (n-1)\sum_{t=1}^{T}(V(\omega_t) - V(\omega_\star)).$$

The claim follows by observing that the first term in the RHS is $-TW_T$ and that the second term is $-R_T^{\mathrm{w}}$. ■

Our proof of Theorem 1 uses techniques from binary hypothesis testing to establish a lower bound on the VCG regret. For this, we begin by reviewing some facts about the KL divergence $\mathrm{KL}(\cdot\|\cdot)$. Recall that for two probabilities P, Q with Q absolutely continuous with respect to P, the KL divergence is $\mathrm{KL}(P\|Q) = \mathbb{E}_P\big[(\frac{\mathrm{d}P}{\mathrm{d}Q}(X))\big]$. For distributions P, P', Q, Q' with $\mathbf{supp}\,(P) = \mathbf{supp}\,(Q)$ and $\mathbf{supp}\,(P') = \mathbf{supp}\,(Q')$, the KL divergence between the product distributions satisfies $\mathrm{KL}(P \times P'\|Q \times Q') = \mathrm{KL}(P\|Q) + \mathrm{KL}(P'\|Q')$. Additionally, for two univariate Gaussians $\mathcal{N}(\mu_1, 1), \mathcal{N}(\mu_2, 1)$, we know $\mathrm{KL}(\mathcal{N}(\mu_1, 1)\|\mathcal{N}(\mu_2, 1)) = (\mu_1 - \mu_2)^2/2$. The following result from Tsybakov (2008) will be useful in our proof.

Lemma 9 (Tsybakov (2008), Lemmas 2.1 and 2.6) *Let P, Q be probabilities such that Q is absolutely continuous with respect to P. Let A be any event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2}\exp\big(-\mathrm{KL}(P\|Q)\big).$$

We are now ready to prove the theorem.

Proof of Theorem 1. Let $n > 1$. Since the maximum is larger than an average, for any set of real numbers $\{a_i\}_i$, we have $\max(\{a_i\}_i) \geq \sum_i \alpha_i a_i$ for any $\{\alpha_i\}_i$ such that $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$. Using Lemma 8 and the fact that R_T^{w} is positive, we obtain the following two upper bounds on $\max(nR_T^{\mathrm{w}}, R_T^{\mathrm{a}}, R_{0T})$:

$$\max(nR_T^{\mathrm{w}}, R_T^{\mathrm{a}}, R_{0T}) \geq \frac{4}{5}nR_T^{\mathrm{w}} + \frac{1}{5}R_{0T} \geq \frac{2}{5}nR_T^{\mathrm{w}} - \frac{1}{5}TW_T,$$
$$\max(nR_T^{\mathrm{w}}, R_T^{\mathrm{a}}, R_{0T}) \geq \frac{4}{5}nR_T^{\mathrm{w}} + \frac{1}{5}TR_T^{\mathrm{a}} = nR_T^{\mathrm{w}} + \frac{1}{5}TW_T \geq \frac{2}{5}nR_T^{\mathrm{w}} + \frac{1}{5}TW_T.$$

The LHS should be larger than both of the above lower bounds. Since $\max(a + b, a + c) = a + \max(b, c)$, we have,

$$R_T^{\mathrm{VCG}} = \max(nR_T^{\mathrm{w}}, R_T^{\mathrm{a}}, R_{0T}) \geq \frac{2}{5}nR_T^{\mathrm{w}} + \frac{1}{5}T|W_T| \;\triangleq Q_T.$$

We will obtain a lower bound on $\inf_{\mathcal{A}} \sup_{\Theta} \mathbb{E}Q_T$ which translates to a lower bound on the desired quantity. Our strategy for doing so is to consider two problems in Θ and show that any algorithm will not be able to distinguish between them. Both problems will have the same set of outcomes $\Omega = \{0, 1, \ldots, \ldots, |\Omega| - 1\}$ with $\Omega = \mathcal{S}$ and $|\Omega| \geq n + 1$. In the first problem, henceforth called θ_0, the optimal outcome is 0 with $v_i(0) = 1/2$ for all agents i. For outcome $j \in \{1, \ldots, n\}$, $v_j(j) = 0$ and $v_i(j) = 1/2$ for every other agent $i \neq j$. For $j > n$, $v_i(j) < 1/4$ for all i. When an outcome ω is chosen, agent i realises a value drawn from $\mathcal{N}(v_i(\omega), 1)$. Finally, the seller has 0 value for all outcomes, $v_0(j) = 0$ for all $j \in \Omega$. The following statements are true about problem θ_0:

$$V(\omega_\star) = V(0) = \frac{n}{2}, \qquad V^{-i}(\omega_\star^{-i}) = V(i) = \frac{n}{2} - \frac{1}{2}, \qquad \sum_{i=1}^{n} V^{-i}(\omega_\star^{-i}) = \frac{n^2}{2} - \frac{n}{2}.$$

The second problem, henceforth called θ_1, is the same as θ_0 but differs in outcomes $j \in \{1, \ldots, n\}$, as shown below. Here, the value of $\delta \in (0, 1/(2(n-1)))$ will be specified shortly. We have:

$$v_i(j) = \begin{cases} 0 & \text{if } i = j, \\ \frac{1}{2} + \delta & \text{if } i \neq j. \end{cases}$$

The following statements are true about problem θ_1:

$$V(\omega_\star) = V(0) = \frac{n}{2}, \qquad V^{-i}(\omega_\star^{-i}) = V(i) = \frac{n}{2} - \frac{1}{2} + (n-1)\delta,$$
$$\sum_{i=1}^{n} V^{-i}(\omega_\star^{-i}) = \frac{n^2}{2} - \frac{n}{2} + n(n-1)\delta.$$

In the above problems, if δ is set to be too large, then it becomes easier to distinguish between the values of different outcomes using stochastic observations thus making the problem easy. If δ is set to be too small, then the regret terms become small since all outcomes have similar values. The largest lower bound is obtained by careful choice of δ (dependent on both n and T) so as to balance between these two cases.

We will make the dependence of Q_T on the problem explicit and write $Q_T(\theta_0), Q_T(\theta_1)$ respectively. Consider any algorithm in $\mathcal{A}$. Expectations and probabilities when we execute this algorithm problem in θ_0 will be denoted $\mathbb{E}_{\theta_0}, \mathbb{P}_{\theta_0}$, and in problem θ_1, they will be denoted $\mathbb{E}_{\theta_1}, \mathbb{P}_{\theta_1}$. Let $N_t(\omega) = \sum_{i=1}^{t-1} \mathbb{1}(\omega_t = \omega)$ denote the number of times outcome $\omega \in \Omega$ was chosen in the first $t-1$ time steps. With this notation, we can upper bound the welfare regret in problem $\theta \in \{\theta_0, \theta_1\}$ as,

$$\mathbb{E}_\theta[R_T^{\mathrm{w}}] = \sum_{j \geq 1} (V(0) - V(j)) \mathbb{E}_\theta[N_t(j)] \geq \sum_{j=1}^{n} (V(0) - V(j)) \mathbb{E}_\theta[N_t(j)].$$

Using the observation that the gap between the optimal and any other outcome in problem θ_0 is at least $1/2$, and that when $H_T > n^2/2 - n/2 + n(n-1)\delta/2$, $|W_T|$ is at least $n(n-1)\delta/2$, we obtain the following lower bound on $\mathbb{E}_{\theta_0}[Q_T(\theta_0)]$:

$$\mathbb{E}_{\theta_0}[Q_T(\theta_0)] \geq \frac{2n}{5} \sum_{k=1}^{n} \frac{1}{2} \mathbb{E}_{\theta_0}\left[N_{t+1}(k)\right] + \frac{T}{5} \frac{n(n-1)\delta}{2} \mathbb{P}_{\theta_0}\Bigg(\underbrace{H_T > \frac{n^2}{2} - \frac{n}{2} + \frac{1}{2} n(n-1)\delta}_{\text{event } A} \Bigg),$$

$$\geq \frac{n}{10}\left(\sum_{k=1}^{n} 2\mathbb{E}_{\theta_0}\left[N_{t+1}(k)\right] \; + \; Tn(n-1)\delta \; \mathbb{P}_{\theta_0}(A) \right). \tag{11}$$

By a similar argument regarding H_T under the event A^c in problem θ_1, we obtain the following. Here, we have dropped the $\mathbb{E}_{\theta_1}\left[N_{t+1}(k)\right]$ terms which are positive.

$$\mathbb{E}_{\theta_1}[Q_T(\theta_1)] \geq \frac{n}{10} Tn(n-1)\delta \; \mathbb{P}_{\theta_1}(A^c).$$

To combine these results we will apply Lemma 9 on $\mathbb{P}_{\theta_0}(A) + \mathbb{P}'_{\theta_0}(A^c)$ in a manner similar to Bubeck et al. (2013). Letting θ_0^T, θ_1^T denote the probability laws of the observed rewards up to round T in problems θ_0, θ_1 respectively, we obtain

$$\mathbb{P}_{\theta_0}(A) + \mathbb{P}'_{\theta_0}(A^c) \geq \frac{1}{2}\exp\left(-\mathrm{KL}(\theta_0^T \| \theta_1^T)\right) = \frac{1}{2}\exp\left(-\frac{(n-1)\delta^2}{2} \sum_{j=1}^{n} \mathbb{E}_{\theta_0}[N_{T+1}(j)] \right).$$

For the first step we have used the fact that A is measurable with respect to the σ-field generated by observations up to round T. For the second step, observe that the outcomes $0, n+1, n+2, |\Omega|-1$ have the same distributions under both θ_0 and θ_1. For any outcome $i \in \{1, \ldots, n\}$, the distribution of agent i is also the same in both problems. For all other agents $j \neq i$, the KL divergence between the corresponding distributions in the two problems is $\delta^2/2$. By combining the three previous bounds, we obtain an upper bound on $\mathbb{E}_{\theta_0}[Q_T(\theta_0)] + \mathbb{E}_{\theta_1}[Q_T(\theta_1)]$:

$$\begin{aligned}
&\frac{10}{n}\Big(\mathbb{E}_{\theta_0}[Q_T(\theta_0)] + \mathbb{E}_{\theta_1}[Q_T(\theta_1)]\Big) \\
&\qquad \geq \sum_{k=1}^{n} 2\mathbb{E}_{\theta_0}\left[N_{t+1}(k)\right] \; + \; T(n-1)\delta \left(\mathbb{P}_{\theta_0}(A) + \mathbb{P}'_{\theta_0}(A^c)\right), \\
&\qquad \geq 2\sum_{k=1}^{n} \mathbb{E}_{\theta_0}\left[N_{t+1}(k)\right] \; + \; \frac{1}{2}T(n-1)\delta \; \exp\left(-\frac{(n-1)\delta^2}{2} \sum_{k=1}^{n} \mathbb{E}_{\theta_0}[N_{t+1}(k)] \right) \\
&\qquad \geq \min_x \left\{ 2x \; + \; \frac{1}{2}T(n-1)\delta \; \exp\left(-\frac{(n-1)\delta^2}{2} x \right) \right\} \\
&\qquad \geq \frac{4}{(n-1)\delta^2} \log\left(\frac{T(n-1)^2\delta^3}{8} \right).
\end{aligned}$$

Finally, we choose $\delta = \left(\frac{16}{T(n-1)^2}\right)^{1/3}$ so that the $\log$ term above can be upper bounded by a constant. This results in the following bound:

$$\frac{1}{2}\Big(\mathbb{E}_{\theta_0}[Q_T(\theta_0)] + \mathbb{E}_{\theta_1}[Q_T(\theta_1)]\Big) \geq \frac{\log(2)}{5 \cdot 16^{2/3}} \cdot T^{2/3}(n-1)^{4/3},$$

where $\delta < \frac{1}{2(n-1)}$ is satisfied if $T > 128n$. The claim follows by observing $\sup_{\theta \in \Theta} \mathbb{E}[Q_T(\theta)] \geq \max\left(\mathbb{E}_{\theta_0}[Q_T(\theta_0)], \mathbb{E}_{\theta_1}[Q_T(\theta_1)]\right) \geq \frac{1}{2}\mathbb{E}_{\theta_0}[Q_T(\theta_0)] + \frac{1}{2}\mathbb{E}_{\theta_1}[Q_T(\theta_1)]$. ■

9. Proofs of Results in Section 6

In this section, we analyse Algorithm 1. Section 9.1 controls the probability that the confidence intervals given in (4) capture the true values. The proofs of Theorems 2, 3, and 4 are given in Sections 9.2, 9.3, and 9.6 respectively. Sections 9.4 and 9.5 prove Propositions 5 and 6 respectively. The bounds on R_{0T} and R_T^{w} will be useful in bounding R_T^{VCG}. In Section 9.7, we state some technical results that are used in our proofs. We begin with some notation and definitions.

Notation & Definitions: Recall that Algorithm 1 proceeds in a sequence of brackets. In our proofs, q_t will denote the bracket index round t belongs to and T_q will be the number of rounds completed by q brackets. Then,

$$T_{q_t-1} < t \leq T_{q_t}. \tag{12}$$

$\mathcal{E}_{it}$, defined below, will denote the event that agent i's values are trapped by the lower and upper confidence bounds at round t when she participates truthfully. $\mathcal{E}_t$ denotes the same for all agents. Here $\widehat{v}_{it}, \breve{v}_{it}$ are as defined in (4). We have:

$$\mathcal{E}_{it} = \Big(\forall\, s \in \mathcal{S},\ \ v_i(s) \in [\breve{v}_{it}(s), \widehat{v}_{it}(s)]\Big), \qquad \mathcal{E}_t = \bigcap_{i=1}^{n} \mathcal{E}_{it}. \tag{13}$$

For the outcome ω_t at time t, let $s_{it} = s_i(\omega_t)$ be the allocation for agent i. Hence, for instance, we can write $V(\omega_t) = v_0(\omega_t) + \sum_{i=1}^n v_i(\omega_t) = v_0(\omega_t) + \sum_{i=1}^n v_i(s_{it})$. We will similarly use the following definitions for the upper and lower bound on the welfare at time t, the functions F_t^{-i}, G_t^{-i} used in the pricing calclulation, and their optimisers. Some of these terms have been defined before.

$$\begin{aligned}
&\widehat{V}_t(\omega) = v_0(\omega) + \sum_{i=1}^{n} \widehat{v}_{it}(\omega_t), \quad \omega_t = \operatorname*{argmax}_{\omega\in\Omega} \widehat{V}_t(\omega), \qquad \breve{V}_t(\omega) = v_0(\omega) + \sum_{i=1}^{n} \breve{v}_{it}(\omega_t),\\
&F_t^{-i}(\omega) = v_0(\omega) + \sum_{j\neq i} f_{it}(\omega_t), \quad \omega_t^{-i} = \operatorname*{argmax}_{\omega\in\Omega} F_t^{-i}(\omega),\\
&G_t^{-i}(\omega) = v_0(\omega) + \sum_{j\neq i} g_{it}(\omega_t), \qquad G_t(\omega) = v_0(\omega) + \sum_{i=1}^{n} g_{it}(\omega_t).
\end{aligned} \tag{14}$$

Next, we define some quantities related to the mean and confidence intervals defined in (4). For brevity, we will denote the unclipped empirical mean in (4) by $\overline{v}'_{it}(s)$. Next, we define β_t, σ_{it} as shown below. With this, we can rewrite the upper and lower confidence bounds in (4) as follows:

$$\begin{aligned}
&\beta_t = \sqrt{5\log(t-qK+1) + 2\log(|\mathcal{S}|)}, \qquad \sigma_{it}(s) = \begin{cases} 0 & \text{if } i \text{ plays by bids} \\ \frac{\sigma}{\sqrt{N_{it}(s)}} & \text{otherwise} \end{cases},\\
&\widehat{v}_{it}(s) = \overline{v}_{it}(s) + \beta_t\sigma_{it}(s), \qquad \breve{v}_{it}(s) = \overline{v}_{it}(s) - \beta_t\sigma_{it}(s).
\end{aligned} \tag{15}$$

9.1 Bounding $\mathbb{P}_t(\mathcal{E}_t^c)$

In this section, we control the probability that the upper and lower confidence bounds do not trap the true values $\{v_i(s)\}_{i,s}$. Recall that sub-Gaussian random variables satisfy the following concentration property. Let $\{X_i\}_{i=1}^n$ be n i.i.d samples from a σ sub-Gaussian distribution and $\overline{X} = \frac{1}{n}\sum_i X_i$ be its sample mean. Then,

$$\mathbb{P}(\overline{X} > \epsilon) \le e^{\frac{-n\epsilon^2}{2\sigma^2}}, \qquad \mathbb{P}(\overline{X} < \epsilon) \le e^{\frac{-n\epsilon^2}{2\sigma^2}}.$$

Lemma 10 *Assume that agent i participates truthfully and let $\mathcal{E}_{it}$ be as defined in* (13)*. When $\zeta = \mathsf{ETC}$, for $t \notin \mathbf{E}$ in bracket q, $\mathbb{P}_t(\mathcal{E}_{it}^c) \le 2(t - qK)^{-5/2}$. Moreover, for all T, $\sum_{t=1,t\notin\mathbf{E}}^T \mathbb{P}_t(\mathcal{E}_t^c) \le 4$. When $\zeta = \mathsf{OPT}$, for $t \notin \mathbf{E}$ in bracket q, $\mathbb{P}_t(\mathcal{E}_{it}^c) \le 2(t - qK)^{-3/2}$. Moreover, for all T, $\sum_{t=1,t\notin\mathbf{E}}^T \mathbb{P}_t(\mathcal{E}_t^c) \le 6$.*

Proof If the agent participates by bids truthfully, then $\widehat{v}_{it} = \breve{v}_{it} = v_i$ and the claim is trivially true. For agents participating by rewards, we will first prove this for $\zeta = \mathsf{OPT}$. Consider the event $\{v_i(s) > \widehat{v}_{it}(s)\}$ and recall the definitions in (4). Let $\overline{v}'_{it}(s)$ be the unclipped empirical mean in (4). Let $\overline{v}_{it}(s) = \max(0, \overline{v}'_{it}(s))$ and $\widehat{\overline{v}}_{it}(s) = \overline{v}_{it}(s) + \beta_t\sigma_{it}(s)$. Since $\overline{v}_{it}(s) = \min(1, \overline{v}_{it}(s))$, we have $\widehat{\overline{v}}_{it}(s) \ge \widehat{v}_{it}(s)$. However, the following calculations show that $\mathbb{P}\left(v_i(s) > \widehat{v}_{it}(s)\right) = \mathbb{P}(v_i(s) > \widehat{\overline{v}}_{it}(s))$.

$$\begin{aligned}
&\mathbb{P}_t\left(v_i(s) > \widehat{v}_{it}(s)\right) \\
&= \mathbb{P}_t\left(v_i(s) > \widehat{v}_{it}(s) | \overline{v}'_{it}(s) \ge 1\right)\mathbb{P}_t(\overline{v}'_{it}(s) \ge 1) + \mathbb{P}_t\left(v_i(s) > \widehat{v}_{it}(s) | \overline{v}'_{it}(s) < 1\right)\mathbb{P}_t(\overline{v}'_{it}(s) < 1) \\
&= \mathbb{P}_t\left(v_i(s) > \widehat{\overline{v}}_{it}(s) | \overline{v}_{it}(s) \ge 1\right)\mathbb{P}_t(\overline{v}_{it}(s) \ge 1) + \mathbb{P}_t\left(v_i(s) > \widehat{\overline{v}}_{it}(s) | \overline{v}_{it}(s) < 1\right)\mathbb{P}_t(\overline{v}_{it}(s) < 1) \\
&= \mathbb{P}_t\left(v_i(s) > \widehat{\overline{v}}_{it}(s)\right).
\end{aligned}$$

Here, the second step uses two arguments. First, when $\overline{v}'_{it}(s) < 1$, then $\widehat{v}_{it}(s) = \widehat{\overline{v}}_{it}(s)$. Second, when $\overline{v}'_{it}(s) \ge 1$, then $\mathbb{P}_t(v_i(s) > \widehat{v}_{it}(s)) = \mathbb{P}_t(v_i(s) > \widehat{\overline{v}}_{it}(s)) = 1$ since $v_i(s) \le 1 < \widehat{v}_{it}(s) \le \widehat{\overline{v}}_{it}(s)$. We can now bound,

$$\begin{aligned}
\mathbb{P}_t(v_i(s) > \widehat{v}_{it}(s)) &= \mathbb{P}_t\big(v_i(s) > \max(0, \overline{v}'_{it}(s)) + \beta_t\sigma_{it}(s)\big) \\
&\le \mathbb{P}_t\left(v_i(s) > \frac{1}{N_{it}(s)}\sum_{\ell=1}^{t-1} X_{i\ell}\mathbb{1}(s_{it} = s) + \beta_t\frac{\sigma}{\sqrt{N_{it}(s)}}\right) \\
&\le \mathbb{P}_t\left(\exists\tau \in \{q, \ldots, t - (K-1)q\},\ v_i(s) > \frac{1}{\tau}\sum_{\ell=1}^{\tau} X'_{i\ell} + \beta_t\frac{\sigma}{\sqrt{\tau}}\right) \\
&\le \sum_{\tau=q}^{t-qK+q} \mathbb{P}_t\left(v_i(s) > \frac{1}{\tau}\sum_{\ell=1}^{\tau} X'_{i\ell} + \beta_t\frac{\sigma}{\sqrt{\tau}}\right) \le (t - qK + 1)e^{-\beta_t^2/2} \\
&\le \frac{1}{|\mathcal{S}|(t - qK + 1)^{3/2}}
\end{aligned}$$

In the second step, if $\overline{v}'_{it}(s)$ was clipped below at 0, then we can replace it with a smaller quantity. In the third step, we have used the fact that $N_{it}(s)$ would take a value in $\{q, \ldots, t - (K-1)q\}$ since there have been qK exploration rounds thus far, during which we have collected rewards from agent

i for allocation s exactly q times. $\{X'_{i\ell}\}_{\ell=1}^{\tau}$ denotes the rewards $X_{i\ell}$ collected when $N_{it}(s) = \tau$. The fourth step uses a union bound and the fourth step applies the sub-Gaussian condition. A similar bound can be shown for the event $\{v_i(s) < \breve{v}_{it}(s)\}$. The first claim follows by applying a union bound over these two events and over all $s \in \mathcal{S}$. The second claim follows from the observation $\sum_{t=1}^{\infty} t^{-3/2} \leq 1 + \int_1^{\infty} t^{-3/2} \leq 3$.

Now consider $\zeta = \mathsf{ETC}$. The calculations above can be repeated, except $N_{it}(s) = q_t$ (12) deterministically for all i, t. (When $\zeta = \mathsf{OPT}$, $N_{it}(s)$ is random and depends on the reward realised.) Therefore, we will not need the sum over $\tau \in \{q, \ldots, t - (K-1)q\}$, resulting in the bound $e^{-\beta_t^2/2} \leq \frac{1}{|\mathcal{S}|(t-qK)^{5/2}}$. The second claim follows from $\sum_t t^{-5/2} \leq 2$. ■

Lemma 11 *Assume that all agents participate truthfully and let $\mathcal{E}_t$ be as defined in* (13)*. When $\zeta = \mathsf{ETC}$, for $t \notin \mathbf{E}$ in bracket q, $\mathbb{P}_t(\mathcal{E}_t^c) \leq 2n(t-qK)^{-5/2}$. Moreover, for all T, $\sum_{t=1,t\notin\mathbf{E}}^{T} \mathbb{P}_t(\mathcal{E}_t^c) \leq 4n$. When $\zeta = \mathsf{OPT}$, for $t \notin \mathbf{E}$ in bracket q, $\mathbb{P}_t(\mathcal{E}_t^c) \leq 2n(t-qK)^{-3/2}$. Moreover, for all T, $\sum_{t=1,t\notin\mathbf{E}}^{T} \mathbb{P}_t(\mathcal{E}_t^c) \leq 6n$.*

Proof This follows by an application of the union bound over the agents $i \in \{1, \ldots, n\}$ on the results of Lemma 10. ■

9.2 Proof of Theorem 2

We will first prove Theorem 2. We begin with the following Lemma. To state it, consider any strategy π that agent i may follow when reporting her rewards. Let u_{it} be the utility of the agent on round t when she reports truthfully on rounds 1 to $t-1$ (recall that the allocation the agent receives on round t depends on the rewards $\{Y_{i\ell}\}_{\ell=1}^{t-2}$ she reported on rounds in the first $t-1$ rounds), let u_{it}^{π} be the utility of the agent when she follows strategy π from rounds 1 through $t-1$, and let u_{it}^{t-1} be the utility of agent on round t when she follows π on rounds 1 thorough $t-2$ and then switches to truth-telling at the end of round $t-1$. If participating by bids, this means it will change the bid function, and if participating by rewards, it means it will replace the reported rewards $Y_{i\ell}$ for rounds $1, \ldots, t-2$ with the true rewards $X_{i\ell}$ and then report truthfully at round $t-1$. Agent i's allocation at round t when the agent replaces her rewards this way will be different to the allocation chosen when simply reporting truthfully since her past untruthful behaviour will have affected the outcomes chosen by the mechanism in the previous rounds (this is particularly the case when $\zeta = \mathsf{OPT}$). We should also emphasise that this behaviour of replacing the rewards is only for the purposes of our proof below. We have the following result.

Lemma 12 *Let $u_{it}, u_{it}^{\pi}, u_{it}^{t-1}$ be as defined above. Then,*

$$U_{iT}^{\pi} - U_{iT} = \sum_{t=1}^{T} u_{it}^{\pi} - u_{it}^{t-1} + \sum_{t=2}^{T} u_{it}^{t-1} - u_{it}. \tag{16}$$

Proof *The claim follows by adding and subtracting $\sum_{t=1}^{T} u_{it}^{t-1}$, rearranging the terms, and noting that $u_{i1}^{0} = u_{i1}$.* ■

When applying the above Lemma, we will denote the strategy which follows π up to round $t-2$ and switches to truth-telling at the end of round $t-1$ as π^{t-1}. We will denote the outcomes at round t when following π, π^{t-1} and truth-telling by $\omega_t^\pi, \omega_t^{t-1}$ and ω_t respectively, and the allocations for agent i by $s_{it}^\pi, s_{it}^r, s_{it}$ for respectively; therefore, $s_{it}^\pi = s_i(\omega_t^\pi)$, $s_{it}^r = s_i(\omega_t^{t-1}$, and $s_{it} = s_i(\omega_t)$.

9.2.1 PROOF OF THEOREM 2.1

We begin with Lemma 12. First, consider the second summation in its RHS, where we claim that each term inside the summation is 0. To see this, note that π^{t-1} is also participating truthfully at round t. It has replaced its reported rewards with its true realised rewards in the previous rounds. The mechanism only uses rewards reported in the exploration rounds to decide outcomes on the exploitation rounds, and the outcomes in the exploration rounds are chosen independent of the bids/rewards reported by the agent. As the outcome and prices in round t will be the same for both policies, we have $u_{it}^{t-1} = u_{it}$. (As we will see shortly in Section 9.2.2, this will not be the case when $\zeta = \mathsf{OPT}$, and the second sum will be non-zero.)

Now turn to the first summation in the RHS of Lemma 12, the bound of which will leverage intuitions from the proof of truthfulness of the VCG mechanism. In the remainder of the proof, g_{it} will denote the appropriate quantity, either $\breve{v}_{it}$ or $\widehat{v}_{it}$ depending on the value of hyperparameter λ, for agent i when following π^{t-1}. Since, at time t, she has switched to being truthful and only rewards from the exploration phase are used in computing outcomes, this will be the same as had she been truthful throughout. Similarly, let G_t denote either $\breve{V}_t$ or $\widehat{V}_t$ when agent i follows π^{t-1}. Using these, we can write for $t \notin \mathbf{E}$,

$$
\begin{aligned}
u_{it}^\pi - u_{it}^{t-1} \;=\; & \left(v_i(s_{it}^\pi) + \left(v_0(\omega_t^\pi) + \sum_{j \neq i} g_{jt}(\omega_t^\pi) \right) - \max_\omega F_t^{-i}(\omega) \right) \qquad (17) \\
& \qquad - \left(v_i(s_{it}^{t-1}) + \left(v_0(\omega_t^{t-1}) + \sum_{j \neq i} g_{jt}(\omega_t^{t-1}) \right) - \max_\omega F_t^{-i}(\omega) \right), \\
=\; & v_i(s_{it}^\pi) - v_i(s_{it}^{t-1}) + \left(v_0(\omega_t^\pi) + \sum_{j \neq i} g_{jt}(\omega_t^\pi) \right) - \left(v_0(\omega_t^{t-1}) + \sum_{j \neq i} g_{jt}(\omega_t^{t-1}) \right), \\
=\; & (v_i(s_{it}^\pi) - g_{it}(s_{it}^\pi)) + (g_{it}(s_{it}^{t-1}) - v_i(s_{it}^{t-1})) + \\
& \qquad \underbrace{\left(v_0(\omega_t^\pi) + \sum_{i=1}^n g_{jt}(\omega_t^\pi) \right)}_{G_t(\omega_t^\pi)} - \underbrace{\left(v_0(\omega_t^{t-1}) + \sum_{i=1}^n g_{jt}(\omega_t^{t-1}) \right)}_{G_t(\omega_t^{t-1})}, \\
\leq\; & (v_i(s_{it}^\pi) - g_{it}(s_{it}^\pi)) + (g_{it}(s_{it}^{t-1}) - v_i(s_{it}^{t-1})).
\end{aligned}
$$

Here, the first step substitutes expressions for u_{it}^π, u_{it}^{t-1} from Fact 7. The $\max_\omega F_t^{-i}(\omega)$ terms are cancelled out in the second step; they will be the same for both policies π, π^{t-1} since it is computed using the rewards reported by other agents in rounds $1, \ldots, t-1$ and hence does not depend on the fact that agent i has switched policies in the current round. The third step adds and subtracts $g_{it}(s_{it}^\pi) + g_{it}(s_{it}^{t-1})$ and observes that the last two terms are $G_t(\omega_t^\pi), G_t(\omega_t^{t-1})$, where, recall G_t is the appropriate quantity computed *after* agent i switches to truthful reporting.

To obtain the last step, recall that $\omega_t = \omega_t^{t-1} = \operatorname{argmax}_\omega \widehat{V}_t(\omega)$ by line 8 of Algorithm 1. Moreover, when $\zeta = \textsf{ETC}$, $\breve{v}_{it}, \overline{v}_{it}, \widehat{v}_{it}$ are vertically shifted functions; for agents participating by bids, they are identical while for agents participating by rewards, we use only one observation per allocation per agent in each exploration phase. Therefore, $\breve{V}_t, \overline{V}_t, \widehat{V}_t$ are also vertically shifted functions and hence $\omega_t = \omega_t^{t-1} = \operatorname{argmax} \widehat{V}_t = \operatorname{argmax} \breve{V}_t = \operatorname{argmax} \overline{V}_t$. Therefore, regardless of the value of λ, we have $G_t(\omega_t^{t-1}) \geq G_t(\omega_t^\pi)$. We emphasise that the above calculations do not use the fact that $\widehat{v}_{jt}$ is an upper confidence bound on v_j for agents $j \neq i$; this may not be true since agent j may not be truthful. Instead, it is simply treated as a function of rewards reported by agent j in previous rounds.

To complete the proof, we can use the fact that that the g_{it} terms are computed under truthful reporting from agent i. If the agent participates by bids, then $g_{it} = v_i$ and hence $u_{it}^\pi - u_{it}^{t-1} \leq 0$ a.s. Combining this with the fact that the utilities for all policies are the same during $t \in \mathbf{E}$, we have $U_{iT}^\pi - U_{iT} \leq 0$ a.s.. For an agent participating by rewards, under $\mathcal{E}_{it}$,

$$\begin{aligned} v_i(s_{it}^\pi) - g_{it}(s_{it}^\pi) + g_{it}(s_{it}^{t-1}) - v_i(s_{it}^{t-1}) &= v_i(s_{it}^\pi) - \widehat{v}_{it}(s_{it}^\pi) + \widehat{v}_{it}(s_{it}^{t-1}) - v_i(s_{it}^{t-1}) \\ &\leq 2\beta_t \sigma_{it}(s_i(\omega_t^{-i})) \;\leq\; 2\sqrt{2}\beta_t \sigma K^{1/3} t^{-1/3}. \end{aligned} \tag{18}$$

Above, we have used the fact the widths of the confidence intervals are all equal. For the last step, we use the following argument to bound $\sigma_{it}(s)$ for any $s \in \mathcal{S}$. It uses Lemma 18 and the fact that at time t, agent i will have experienced all allocations $s \in \mathcal{S}$ at least q_t times.

$$\forall\, i \in \{1, \ldots, n\},\ t \geq 1,\ s \in \mathcal{S}, \quad \sigma_{it}(s) = \sigma / \sqrt{N_{it}(s)} \leq \sigma / \sqrt{q_t} \leq \sqrt{2} K^{1/3} t^{-1/3}. \tag{19}$$

We will use the bound in (19) repeatedly in our proofs.

Therefore, (18) leads us to $\mathbb{E}[u_{it}^\pi - u_{it}^{t-1} | \mathcal{E}_{it}] \leq 2\sqrt{2}\beta_t \sigma K^{1/3} t^{-1/3}$ and consequently,

$$\mathbb{E}[U_{iT}^\pi - U_{iT}] = \sum_t \mathbb{E}[u_{it}^\pi - u_{it}^{t-1} | \mathcal{E}_{it}] + \sum_t \mathbb{P}(\mathcal{E}_{it}) \leq 3\sqrt{2}\beta_T K^{1/3} T^{2/3} + 4.$$

The last step uses Lemma 10 to bound $\sum_t \mathbb{P}(\mathcal{E}_{it})$. The claim follows by substituting for β_T (15). □

9.2.2 PROOF OF THEOREM 2.2

The main difference in applying Lemma 12 in the $\zeta = \textsf{OPT}$ case is that now the mechanism uses all of the rewards reported by the agents, and this needs to be accounted for when bounding the two summations. Unlike in Section 9.2.1, we cannot take values such as $\widehat{v}_{it}, \breve{v}_{it}$ to be the same for π^{t-1} and truth-telling because now the mechanism is using reported rewards from all rounds to determine the outcome at round t; while we have swapped all false reports with the true rewards in π^{t-1}, the outcomes in the rounds outside the exploration phase will have been different, and therefore so are the rewards realised and the quantities computed based on the rewards. Therefore, in this proof, we will annotate quantities related to strategy π_t^{t-1} at time t with a prime. For example, $\widehat{v}'_{it} : \mathcal{S} \to \mathbb{R}$ (see (4)) will be the upper confidence bounds at time t for agent i when following π^{t-1}. On the same note, $\mathcal{E}'_{it}$ denotes the event that agent i's true values fall within the confidence interval at time t when she follows π^{t-1}.

For the terms in the first summation in the RHS of Lemma 12, by repeating the calculations in (17), we obtain (using our above notation),

$$u_{it}^\pi - u_{it}^{t-1} = (v_i(s_{it}^\pi) - g'_{it}(s_{it}^\pi)) + (g'_{it}(s_{it}^{t-1}) - v_i(s_{it}^{t-1})) + G'_t(\omega_t^\pi) - G'_t(\omega_t^{t-1}).$$

Recall that (g'_{it}, G'_t) denote either $(\check{v}'_{it}, \check{V}_t)$ or $(\widehat{v}'_{it}, \widehat{V}'_t)$ as per the value of λ being SEL or AGE. They are computed in round t under truthful reporting. If agent i participates by bids, then $g'_{it} = v_i$ and hence $v_i(s) - g'_{it}(s) = 0$ for all s. If she participates by rewards, then for all choices of λ,

$$\mathbb{E}[(v_i(s^\pi_{it}) - g'_{it}(s^\pi_{it})) + (g'_{it}(s^{t-1}_{it}) - v_i(s^{t-1}_{it})|\mathcal{E}'_{it}] \leq 2\sqrt{2}\sigma\beta_t K^{1/3}t^{-1/3}. \tag{20}$$

This follows from the observation that when $\lambda = \text{AGE}$, the first term is less than 0 while the second is less than $2\beta_t\sigma_{it}(s^{t-1}_{it}) \leq 2\sqrt{2}\sigma\beta_t K^{1/3}t^{-1/3}$ by Lemma 18 and the fact that there have been q_t exploration phases (see (19)); a similar argument holds for $\lambda = \text{SEL}$, but with the terms reversed.

Next, we use $\omega^{t-1}_t = \text{argmax}_\omega \widehat{V}'_t(\omega)$ to bound the difference $G'_t(\omega^\pi_t) - G'_t(\omega^{t-1}_t)$. When $\lambda = \text{AGE}$,

$$G'_t(\omega^\pi_t) - G'_t(\omega^{t-1}_t) = \widehat{V}'_t(\omega^\pi_t) - \widehat{V}'_t(\omega^{t-1}_t) \leq 0.$$

When $\lambda = \text{SEL}$, we can use the fact that at round $t \notin \mathbf{E}$ all allocations will have been experienced by each agent at least q_t times (12) to obtain,

$$\begin{aligned} G'_t(\omega^\pi_t) - G'_t(\omega^{t-1}_t) &= \check{V}_t(\omega^\pi_t) - \check{V}_t(\omega^{t-1}_t) \\ &\leq \widehat{V}'_t(\omega^\pi_t) - \widehat{V}'_t(\omega^{t-1}_t) + 2\beta_t \sum_i (\sigma'_{it}(s^{t-1}_{it}) - \sigma'_{it}(\omega^\pi_t)) \\ &\leq 2\beta_t \sum_i \sigma'_{it}(\omega^{t-1}_t) \leq 2\sigma\beta_t \sum_i \frac{1}{\sqrt{q_t}} \leq 2\sqrt{2}\sigma\beta_t n K^{1/3}t^{-1/3}. \end{aligned}$$

The last step uses Lemma 18. Now summing over all t and using Lemma 20, we obtain

$$\sum_{t\notin\mathbf{E}} \mathbb{E}[u^\pi_{it} - u^{t-1}_{it}|\mathcal{E}'_{it}] \leq \begin{cases} 3\sqrt{2}\sigma\kappa_i\beta_T K^{1/3}T^{2/3} & \text{if } \lambda = \text{AGE}, \\ 3\sqrt{2}\sigma(n+\kappa_i)\beta_T K^{1/3}T^{2/3} & \text{if } \lambda = \text{SEL}. \end{cases} \tag{21}$$

We now move to the second summation in the RHS of Lemma 12. To bound this term, we will use the fact that all agents except i are adopting stationary policies. Therefore, the rewards reported by any agent $j \neq i$ for any $s \in \mathcal{S}$ concentrates around some mean, and we can apply Lemma 10 for that agent. For the remainder of this proof, $v_j(s)$ will denote the mean of this distribution. (This may not be equal to the true value of agent j for allocation s since she may not be truthful.) $\mathcal{E}_t, \mathcal{E}'_t$ denote the events that $v_j(s)$ falls within the confidence intervals $(\check{v}_{jt}(s), \widehat{v}_{jt}(s))$, $(\check{v}'_{jt}(s), \widehat{v}'_{jt}(s))$ respectively for all agents j at round t. Here, recall, the former interval is obtained for agent j when agent i is being truthful from the beginning and the latter when i is following π^{t-1}. We now expand each term in the second summation as follows,

$$\begin{aligned} u^{t-1}_{it} - u_{it} \;=\; & \left(v_i(s^{t-1}_{it}) + \left(v_0(\omega^{t-1}_t) + \sum_{j\neq i} g'_{jt}(\omega^{t-1}_t)\right) - \max_\omega F'^{-i}_t(\omega)\right) \\ & \qquad - \left(v_i(s_{it}) + \left(v_0(\omega_t) + \sum_{j\neq i} g_{jt}(\omega_t)\right) - \max_\omega F^{-i}_t(\omega)\right), \\ =\; & \left(v_i(s^{t-1}_{it}) - g'_{it}(s^{t-1}_{it})\right) + \left(g_{it}(s_{it}) - v_i(s_{it})\right) \\ & \qquad + \left(G'_t(\omega^{t-1}_t) - G_t(\omega_t)\right) + \left(\max_\omega F^{-i}_t(\omega) - \max_\omega F'^{-i}_t(\omega)\right)\Big). \end{aligned} \tag{22}$$

The first step uses the expressions in Fact 17, while the second step adds and subtracts $g'_{it}(s_{it}^{t-1}) + g_{it}(s_{it})$ and rearranges the terms. To bound all four terms in (22), we will use that g_{it}, g'_{it}, G_t, G'_t, F_t^{-i}, $F_t'^{-i}$ are all computed under truthful reporting from agent i, that all other agents are adopting stationary policies, and that each agent has experienced each allocation at least q_t times (12) in round t. The first two terms are 0 for an agent participating by bids. If participating by rewards, via a similar reasoning to that used in (20),

$$\mathbb{E}[(v_i(s_{it}^{t-1}) - g'_{it}(s_{it}^{t-1})) + (g_{it}(s_{it}) - v_i(s_{it}))|\mathcal{E}'_{it}, \mathcal{E}_{it}] \leq 2\sqrt{2}\sigma\beta_t K^{1/3}t^{-1/3}.$$

To bound the third term in (22), observe that $\widehat{V}'_t - \widehat{V}_t$ is uniformly bounded under $\mathcal{E}_t \cap \mathcal{E}'_t$.

$$\begin{aligned}\widehat{V}'_t(\omega) - \widehat{V}_t(\omega) &= \sum_i (\widehat{v}'_{it}(\omega) - \widehat{v}_{it}(\omega)) = \sum_i (\widehat{v}'_{it}(\omega) - v_i(\omega)) + \sum_i (v_i(\omega) - \widehat{v}_{it}(\omega)) \\ &\leq 2\sqrt{2}\sigma\beta_t n K^{1/3}t^{-1/3}.\end{aligned}$$

Observing that $\omega_t^{t-1} = \operatorname{argmax}_\omega \widehat{V}'_t(\omega_t)$ and $\omega_t = \operatorname{argmax}_\omega \widehat{V}_t(\omega_t)$, we use Lemma 19 to obtain,

$$\begin{aligned}G'_t(\omega_t^{t-1}) - G_t(\omega_t) &= \widehat{V}'_t(\omega_t^{t-1}) - \widehat{V}_t(\omega_t) + \beta_t\kappa_\lambda \sum_j (\sigma_{jt}(\omega_t) - \sigma'_{jt}(\omega_t^{t-1})) \\ &\leq (2 + \kappa_\lambda)\sqrt{2}\sigma\beta_t n K^{1/3}t^{-1/3}.\end{aligned}$$

Here $\kappa_\lambda = 0$ if $\lambda = \textsf{AGE}$ and $\kappa_\lambda = 2$ if $\lambda = \textsf{SEL}$. Above, the first step rewrites the expression for G_t, G'_t in terms of $\widehat{V}_t$, $\widehat{V}'_t$, and κ_λ. The second step drops the $\sigma'_{it}(\omega_t^{t-1})$ terms and bounds the $\sigma_{it}(\omega_t^{t-1})$ terms using Lemma 18. To bound the last term, we observe that $F_t'^{-i} - F_t^{-i}$ is uniformly bounded under $\mathcal{E}_t \cap \mathcal{E}'_t$. Using a similar reasoning to (20),

$$\begin{aligned}F'_t(\omega) - F_t(\omega) &= \sum_i (f'_{it}(\omega) - f_{it}(\omega)) = \sum_i (f'_{it}(\omega) - v_i(\omega)) + \sum_i (v_i(\omega) - f_{it}(\omega)) \\ &\leq 2\sqrt{2}\sigma\beta_t n K^{1/3}t^{-1/3}.\end{aligned}$$

By Lemma 19, we therefore have, $\max_\omega F_t^{-i}(\omega) - \max_\omega F_t'^{-i}(\omega) \leq 2\sqrt{2}\sigma\beta_t n K^{1/3}t^{-1/3}$. Summing over all t and using Lemma 20, we can now bound the second summation in Lemma 12.

$$\sum_{t \notin \mathbf{E}} \mathbb{E}[u_{it}^{t-1} - u_{it}|\mathcal{E}'_t, \mathcal{E}_t] \leq 3\sqrt{2}\sigma\big(\kappa_i + n(4 + \kappa_\lambda/2)\big)\beta_T K^{1/3}T^{2/3}. \tag{23}$$

Finally, we can combine the results in (21), (23) to obtain

$$\begin{aligned}\mathbb{E}[U_{iT}^\pi - U_{iT}] &= \sum_{t \notin \mathbf{E}} \mathbb{E}[u_{it}^\pi - u_{it}] = \sum_{t \notin \mathbf{E}} \mathbb{E}[u_{it}^\pi - u_{it}|\mathcal{E}_t, \mathcal{E}'_t] + \sum_{t \notin \mathbf{E}} \mathbb{P}(\mathcal{E}_t^c \cup \mathcal{E}_t'^c) \\ &= \sum_{t \notin \mathbf{E}} \mathbb{E}[u_{it}^\pi - u_{it}^{t-1}|\mathcal{E}'_{it}] + \sum_{t \notin \mathbf{E}} \mathbb{E}[u_{it}^{t-1} - u_{it}|\mathcal{E}_t, \mathcal{E}'_t] + 12n \\ &\leq 3\sqrt{2}\sigma\big(2\kappa_i + n(4 + \kappa_\lambda)\big)\beta_T K^{1/3}T^{2/3} + 12n.\end{aligned}$$

The first step observes that the allocations and prices are the same during the exploration phase rounds $\mathbf{E}$. The third step uses Lemma 11, although $\mathcal{E}'_t$ now refers to an event when the strategy changes at each step. The claim follows by substituting for β_T (15) and then observing $\kappa_\lambda \leq 2$ and $\kappa_i \leq 1$. □

9.3 Proof of Theorem 3

In this Section, we prove the individual rationality properties of Algorithm 1. In our proofs, we will only assume that agent i is participating truthfully. While the computed upper/lower confidence bounds of all agents will appear in the analysis, we will not use the fact that $\check{v}_{jt} \leq v_j \leq \hat{v}_{jt}$ for $j \neq i$. We will however use Lemma 10 to control the probability of the event $\check{v}_{it} \leq v_i \leq \hat{v}_{it}$ for agent i.

9.3.1 PROOF OF THEOREM 3.1

We will first consider the $\zeta = \mathsf{ETC}$ case. For all agents, $u_{it} \geq 0$ when $t \in \mathbf{E}$, so let us consider $t \notin \mathbf{E}$. By Fact 17, we have for $t \notin \mathbf{E}$,

$$u_{it} = \underbrace{v_i(s_{it}) - g_{it}(s_{it})}_{c_t} + \underbrace{G_t(\omega_t) - F_t^{-i}(\omega_t^{-i})}_{d_t}. \tag{24}$$

We will first bound c_t. If agent i participates by bids truthfully, $g_{it} = v_i$ and hence $c_t = 0$ a.s. To bound c_t when she participates by rewards truthfully, let $\tilde{c}_t = \max(0, g_{it}(s_{it}) - v_i(s_{it}))$. Clearly, $\tilde{c}_t \geq 0$ and $c_t \geq -\tilde{c}_t$. Observing that $\check{v}_{it} \leq v_i \leq \hat{v}_{it}$ under $\mathcal{E}_{it}$ (13), and that $g_{it} = \hat{v}_{it}$ when $\lambda = \mathsf{AGE}$ and $g_{it} = \check{v}_{it}$ when $\lambda = \mathsf{SEL}$, we have,

$$\mathbb{E}[\tilde{c}_t|\mathcal{E}_{it}] \leq 0 \;\text{ if } \lambda = \mathsf{SEL}, \qquad \mathbb{E}[\tilde{c}_t|\mathcal{E}_{it}] \leq 2\beta_t\sigma_{it}(s_{it}) \;\text{ if } \lambda = \mathsf{AGE}.$$

To bound d_t, we first observe that $\omega_t = \operatorname{argmax}_\omega G_t(\omega)$ since $\check{V}_t, \overline{V}_t, \hat{V}_t$ are vertically shifted functions (using the same argument used in Section 9.2.1). Now, consider the case $\lambda = \mathsf{AGE}$. Since, $\hat{V}_t = \hat{V}_t^{-i} + \hat{v}_{it}$ and $\hat{v}_{it} \geq \overline{v}_{it} \geq 0$ (recall from (4) that we clip $\overline{v}_{it}$ between 0 and 1), we have that $\hat{V}_t^{-i} \geq \hat{V}_t^{-i}$. By observing $\check{V}_t^{-i} \leq \hat{V}_t^{-i}$, we have

$$d_t = \max_\omega G_t - \max_\omega F_t^{-i} = \max \hat{V}_t - \max \check{V}_t^{-i} \geq \max \hat{V}_t - \max \hat{V}_t^{-i} \geq 0.$$

When $\lambda = \mathsf{SEL}$, and therefore $G_t = \check{V}_t$ and $F_t^{-i} = \hat{V}_t^{-i}$, one no longer has $d_t \geq 0$ since $\hat{V}_t^{-i}$ can be larger than $\check{V}_t$. However, we can obtain a weaker bound of the form,

$$d_t = \check{V}_t(\omega_t) - \hat{V}_t^{-i}(\omega_t^{-i}) = \max \hat{V}_t - \max \hat{V}_t^{-i} - 2\sum_j \sigma_{jt}(s_{jt}) \geq -2\sqrt{2}\sigma\beta_t n K^{1/3} t^{-1/3}.$$

Above, the last step uses that $\hat{V}_t \geq \hat{V}_t^{-i}$ as before and (19) to bound the $\sigma_{jt}(s_{jt})$ terms.

We can now bound the utilities for the various cases in the theorem for agent i. First, when $\lambda = \mathsf{SEL}$ and agent i is participating by bids, we have, $u_{it} = c_t + d_t \geq 0$ a.s. for all $t \notin \mathbf{E}$. Therefore, $U_{iT} \geq 0$ for all T and the mechanism is individually rational for this agent. That is, the algorithm is (almost surely) individually rational. If agent i participates by rewards truthfully, we have

$$\mathbb{E}[-U_{iT}] \leq \sum_{t \notin \mathbf{E}} \mathbb{E}_t[-c_t|\mathcal{E}_{it}] + \sum_{t \notin \mathbf{E}} \mathbb{E}_t[-d_t|\mathcal{E}_{it}] + \sum_{t \notin \mathbf{E}} \mathbb{P}_t(\mathcal{E}_{it}^c).$$

By combining the results above and applying Lemma 10, we obtain,

$$\mathbb{E}[U_{iT}] \geq \sum_{t \notin \mathbf{E}} \mathbb{E}[u_{it}] \geq \begin{cases} -3\sqrt{2}\sigma\beta_t K^{1/3} T^{2/3} - 4, & \text{if } \lambda = \mathsf{AGE}, \\ -3\sqrt{2}n\sigma\beta_t K^{1/3} T^{2/3} - 4, & \text{if } \lambda = \mathsf{SEL}. \end{cases}$$

The claim follows by substituting for β_T (15). □

9.3.2 PROOF OF THEOREM 3.2

Now we will consider the $\zeta = \mathsf{OPT}$ case. As in Section 9.3.1, we will write $u_{it} = c_t + d_t$ where c_t, d_t are as defined in (24), and consider rounds $t \notin \mathbf{E}$. First consider c_t. If agent i participates by bids truthfully, $g_{it} = v_i$ and hence $c_t = 0$. To bound c_t when she participates by rewards truthfully, let $\tilde{c}_t = \max(0, g_{it}(s_{it}) - v_i(s_{it})) \geq 0$. Using a similar argument as above, we have

$$\mathbb{E}[\tilde{c}_t|\mathcal{E}_{it}] \leq 0 \;\; \text{if } \lambda = \mathsf{SEL}, \qquad \mathbb{E}[\tilde{c}_t|\mathcal{E}_{it}] \leq 2\beta_t \sigma_{it}(s_{it}) \;\; \text{if } \lambda = \mathsf{AGE}.$$

To bound d_t, first note that $\widehat{V}_t(\omega_t) - \max \widehat{V}_t^{-i} \geq 0$ since $\widehat{V}_t = \widehat{V}_t^{-i} + \widehat{v}_{it}$ and $\widehat{v}_{it} \geq 0$. Therefore, when $\lambda = \mathsf{AGE}$,

$$d_t = \widehat{V}_t(\omega_t) - F_t^{-i}(\omega_t^{-i}) \geq \widehat{V}_t(\omega_t) - \max \widehat{V}_t^{-i} \geq 0,$$

and when $\lambda = \mathsf{SEL}$,

$$d_t = \breve{V}_t(\omega_t) - F_t^{-i}(\omega_t^{-i}) \geq \widehat{V}_t(\omega_t) - \max \widehat{V}_t^{-i} - 2\beta_t \sum_j \sigma_{jt}(s_{jt}) \geq -2\beta_t \sum_j \sigma_{jt}(s_{jt}).$$

To bound the $\sum_t \sigma_{it}(s_{it})$ terms in c_t and d_t when $\zeta = \mathsf{OPT}$, we use the following argument which leads to a tighter upper bound bound.

$$\begin{aligned} \sum_{t \notin \mathbf{E}} \frac{\sigma_{it}(s_{it})}{\sigma} &= \sum_{t \notin \mathbf{E}} \frac{1}{\sqrt{N_{it}(s_{it})}} \leq \sum_{t > K} \frac{1}{\sqrt{N_{it}(s_{it})}} \leq \sum_{s \in \mathcal{S}} \sum_{j=1}^{N_{iT}(s)} \frac{1}{\sqrt{j}} \\ &\leq 2 \sum_{s \in \mathcal{S}} \sqrt{N_{iT}(s)} \leq 2\sqrt{|\mathcal{S}|(T - qK)} \leq 2\sqrt{|\mathcal{S}|T}. \end{aligned} \tag{25}$$

The first step simply adds more terms to the summation. The second step observes that the summation can be written as $|\mathcal{S}|$ different summations, one for each $s \in \mathcal{S}$. The third step uses Lemma 20. We will use the above bound in (25) elsewhere in our proofs for the $\zeta = \mathsf{OPT}$ case.

By the same argument for other agents j and using the fact that $\beta_t \leq \beta_T$ for all $t < T$, we have $\sum_t \beta_t \sigma_{jt}(s_{jt}) \leq 2\sigma\beta_T |\mathcal{S}|^{1/2} T^{1/2}$ for all agents j. Therefore, for an agent participating by bids, $U_{iT} \geq 0$ if $\lambda = \mathsf{AGE}$ and $U_{iT} \geq -4\beta_T \sigma n |\mathcal{S}|^{1/2} T^{1/2}$ if $\lambda = \mathsf{SEL}$. For an agent participating by rewards, we have:

$$\mathbb{E}[U_{iT}] \geq -\sum_{t \notin \mathbf{E}} \mathbb{E}[\tilde{c}_t|\mathcal{E}_{it}] + \sum_{t \notin \mathbf{E}} \mathbb{E}[d_t] - \sum_{t \notin \mathbf{E}} \mathbb{P}(\mathcal{E}_{it}^c) \geq \begin{cases} -4\beta_T \sigma |\mathcal{S}|^{1/2} T^{1/2} - 6 & \text{if } \lambda = \mathsf{AGE}, \\ -4\beta_T \sigma n |\mathcal{S}|^{1/2} T^{1/2} - 6 & \text{if } \lambda = \mathsf{SEL}. \end{cases}$$

□

9.4 Proof of Proposition 5

In this section, we bound the welfare regret R_T^{w}. The bounds we establish for the welfare regret here and for the seller regret in Section 9.5 will be useful when we bound the VCG regret in Section 9.6. The following lemma provides a bound that will be useful in the proof of Proposition 5.

Lemma 13 *The welfare regret R_T^{w} (3) satisfies the following bound.*

$$\mathbb{E}[R_T^{\text{w}}] \;\leq\; 3V_{\max}K^{1/3}T^{2/3} + 2\beta_T \sum_{t\notin\mathbf{E}}\sum_{i=1}^{n}\mathbb{E}_t[\sigma_{it}(s_{it})] + V_{\max}\sum_{t\notin\mathbf{E}}\mathbb{P}_t(\mathcal{E}_t^c).$$

Proof Write $R_T^{\text{w}} = \sum_{t=1}^{T} r_t$ where $r_t = V(\omega_\star) - V(\omega_t)$. Recall that $\mathbf{E}$ denotes time indices belonging to the explore phase. We split the instantaneous regret terms to obtain,

$$R_T^{\text{w}} = \sum_{t=1,t\in\mathbf{E}}^{T} r_t + \sum_{t=1,t\notin\mathbf{E}}^{T} r_t.$$

First consider the second summation. Using the notation in (15), we obtain,

$$\begin{aligned}
\mathbb{E}_t[r_t] \;&\leq\; \mathbb{E}[r_t|\mathcal{E}_t] + \mathbb{E}[r_t|\mathcal{E}_t^c]\mathbb{P}(\mathcal{E}_t^c) \;=\; \mathbb{E}_t[V(\omega_\star) - \widehat{V}_t(\omega_t) + \widehat{V}_t(\omega_t) - V(\omega_t)|\mathcal{E}_t] + V_{\max}\mathbb{P}_t(\mathcal{E}_t^c)\\
&\leq\; \mathbb{E}_t[V(\omega_\star) - \widehat{V}_t(\omega_\star) + \widehat{V}_t(\omega_t) - V(\omega_t)|\mathcal{E}_t] + V_{\max}\mathbb{P}_t(\mathcal{E}_t^c)\\
&\leq\; \mathbb{E}_t[\widehat{V}_t(\omega_t) - \check{V}_t(\omega_t)|\mathcal{E}_t^c] + V_{\max}\mathbb{P}_t(\mathcal{E}_t^c) \;\leq 2\beta_t\sum_{i=1}^{n}\mathbb{E}_t[\sigma_{it}(s_{it})] + V_{\max}\mathbb{P}_t(\mathcal{E}_t^c). \qquad (26)
\end{aligned}$$

Here, the third step uses the fact that $\widehat{V}_t$ is maximised at ω_t. The fourth step uses that $\widehat{V}_t \geq V$ and $\check{V}_t \leq V$ under $\mathcal{E}_t$. Now summing over all t, we obtain

$$\mathbb{E}[R_T^{\text{w}}] \;\leq\; \sum_{t\in\mathbf{E}}\mathbb{E}_t[r_t] \;+\; \sum_{t\notin\mathbf{E}}\mathbb{E}_t[r_t] \;\leq\; \sum_{t\in\mathbf{E}}V_{\max} \;+\; \sum_{t\notin\mathbf{E}}\Big(2\sum_{i=1}^{n}\beta_t\mathbb{E}_t[\sigma_{it}(s_{it})] + V_{\max}\mathbb{P}_t(\mathcal{E}_t^c)\Big).$$

Now, the number of terms in the first summation can be bound by $q_T K \leq 3K^{1/3}T^{2/3}$ using Lemma 18. The claim follows by observing $\beta_t \leq \beta_T$ for all $t \leq T$. ■

Proof of Proposition 5. We will first consider the case $\zeta = \texttt{ETC}$, and apply Lemma 13. By Lemma 11, we have $\sum_t \mathbb{P}_t(\mathcal{E}_t^c) \leq 4n$. By following a similar argument to (19), we obtain $\sigma_{it}(s_{it}) \leq \sqrt{2}K^{1/3}t^{-1/3}$. Then, using Lemma 20 to bound $\sum t^{-1/3}$, we have

$$\mathbb{E}[R_T^{\text{w}}] \;\leq\; 4nV_{\max} + 3V_{\max}K^{1/3}T^{2/3} + 3\sqrt{2}\beta_T n K^{1/3}T^{2/3}. \qquad (27)$$

Next, consider $\zeta = \texttt{OPT}$. In order to use Lemma 13, we will use a similar argument as in (25) to obtain $\sum_{t\notin\mathbf{E}}\sigma_{it}(s_{it}) \leq 2\sigma|\mathcal{S}|^{1/2}T^{1/2}$. Next, by Lemma 11, we have $\sum_t \mathbb{P}_t(\mathcal{E}_t^c) \leq 6n$. These results when applied with Lemma 13 yield:

$$\mathbb{E}[R_T^{\text{w}}] \leq 6V_{\max}n + 4\sigma n\beta_T|\mathcal{S}|^{1/2}T^{1/2} + 3K^{1/3}T^{2/3}. \qquad (28)$$

The claims follow by substituting for β_T (15) in (27) and (28). □

9.5 Proof of Proposition 6

In this section, we bound the agent and seller regrets. First, in Lemma 14 we provide an upper bound on the agent regret. Recall that $v_i^\dagger = \max(u_{i\star} - \min_s v_i(s), 0)$ from (8). If the agent prefers receiving any item in $\mathcal{S}$ for free instead of the socially optimal outcome at the VCG price, then this term will be 0 and the agent does not incur any regret during the exploration phase rounds.

Lemma 14 *Consider any agent i and define a_t, b_t as follows for $t \geq 0$.*

$$a_t = F_t^{-i}(\omega_t^{-i}) - \breve{V}_t(\omega_\star^{-i}), \qquad b_t = g_{it}(s_{it}) - \breve{v}_{it}(s_{it}) + \widehat{V}_t(\omega_t) - G_t(\omega_t).$$

Then, the following bound holds on the regret of agent i.

$$\mathbb{E}[R_{iT}] \;\leq\; 3v_i^\dagger K^{1/3}T^{2/3} \;+\; \sum_{t\notin\mathbf{E}} \mathbb{E}_t[a_t|\mathcal{E}_t] \;+\; \sum_{t\notin\mathbf{E}} \mathbb{E}_t[b_t|\mathcal{E}_t] \;+\; \sum_{t\notin\mathbf{E}} \mathbb{P}_t(\mathcal{E}_t^c).$$

Proof As above, we will write $R_{iT} = \sum_{t\in\mathbf{E}} r_{it} + \sum_{t\notin\mathbf{E}} r_{it}$, where $r_{it} = u_{i\star} - u_{it}$. We will first bound the second summation in expectation. For $t \notin \mathbf{E}$, we use Facts 7 and 17 to obtain,

$$r_{it} = g_{it}(s_{it}) - v_i(s_{it}) + V(\omega_\star) - G_t(\omega_t) + F_t^{-i}(\omega_t^{-i}) - V^{-i}(\omega_\star^{-i}).$$

Under $\mathcal{E}_t$, the following are true; $v_i(s_{it}) \geq \breve{v}_{it}(s_{it})$; $V(\omega_\star) \leq \widehat{V}_t(\omega_\star) \leq \widehat{V}_t(\omega_t)$ since ω_t maximises $\widehat{V}_t$, and $V^{-i}(\omega_\star^{-i}) \geq V^{-i}(\omega_t^{-i}) \geq \breve{V}_t(\omega_t^{-i})$ since $\omega_\star^{-i}$ maximises V^{-i}. This leads us to,

$$\mathbb{E}_t[r_{it}] \;\leq\; \mathbb{E}_t[\underbrace{F_t^{-i}(\omega_t^{-i}) - \breve{V}_t(\omega_t^{-i})}_{a_t}|\mathcal{E}_t] \;+\; \mathbb{E}_t[\underbrace{g_{it}(s_{it}) - \breve{v}_{it}(s_{it}) + \widehat{V}_t(\omega_t) - G_t(\omega_t)}_{b_t}|\mathcal{E}_t] \;+\; \mathbb{P}_t(\mathcal{E}_t^c).$$

Summing over all t yields the following bound on the agent regret:

$$\begin{aligned}\mathbb{E}[R_{iT}] \;&\leq\; \sum_{t\in\mathbf{E}} v_i^\dagger \;+\; \sum_{t\notin\mathbf{E}} \mathbb{E}_t[r_{it}|\mathcal{E}_t] \;+\; \sum_{t\notin\mathbf{E}} \mathbb{P}_t(\mathcal{E}_t^c) \\ &\leq\; 3v_i^\dagger K^{1/3}T^{2/3} \;+\; \sum_{t\notin\mathbf{E}} \mathbb{E}_t[a_t|\mathcal{E}_t] \;+\; \sum_{t\notin\mathbf{E}} \mathbb{E}_t[b_t|\mathcal{E}_t] \;+\; \sum_{t\notin\mathbf{E}} \mathbb{P}_t(\mathcal{E}_t^c).\end{aligned} \tag{29}$$

Here, for the first summation, we applied Lemma 18 to obtain $v_i^\dagger q_T K \leq 3v_i^\dagger K^{1/3}T^{2/3}$. ■

When applying the above lemma, the value of hyperparameter λ in Algorithm 1 will decide the bounds for a_t, b_t respectively. Additionally, note that a_t, b_t are measurable with respect to the sigma field generated by observations up to time $t-1$. Hence, $\mathbb{E}_t[a_t], \mathbb{E}_t[b_t]$ are deterministic quantities. Our next lemma bounds the seller regret. For this, we first define A_t^{-i}, for $i \in \{1, \ldots, n\}$ and B_t as follows for $t \geq 0$:

$$A_t^{-i} = V^{-i}(\omega_\star^{-i}) - F_t^{-i}(\omega_\star^{-i}), \qquad B_t = G_t(\omega_t) - V(\omega_\star). \tag{30}$$

Lemma 15 *Let A_t^{-i} and B_t be as defined in* (30)*. Then, the following bound holds on the regret of the seller* (3)*:*

$$\mathbb{E}[R_{0T}] \;\le\; 3V_{\max}K^{1/3}T^{2/3} \;+\; \sum_{t\notin\mathbf{E}}\sum_{i=1}^{n}\mathbb{E}_t[A_t^{-i}|\mathcal{E}_t] \;+\; (n-1)\sum_{t\notin\mathbf{E}}\mathbb{E}_t[B_t|\mathcal{E}_t] \;+\; V_{\max}\sum_{t\notin\mathbf{E}}\mathbb{P}_t(\mathcal{E}_t^c).$$

Proof Write $R_{0T} = \sum_{t\in\mathbf{E}}^{T} r_{0t} + \sum_{t\notin\mathbf{E}}^{T} r_{0t}$, where $r_{0t} = u_{0\star} - u_{0t}$. To bound the second summation, we use Facts 7 and 17 to obtain the following expression for r_{0t} when $t \notin \mathbf{E}$:

$$r_{0t} = u_{0\star} - u_{0t} \;=\; \sum_{i=1}^{n}\left(V^{-i}(\omega_\star^{-i}) - F_t^{-i}(\omega_t^{-i})\right) \;+\; (n-1)\left(G_t(\omega_t) - V(\omega_\star)\right). \tag{31}$$

Hence, for $t \notin \mathbf{E}$, we have $\mathbb{E}_t[r_{0t}] \le \mathbb{E}_t[r_{0t}|\mathcal{E}_t] + \mathbb{E}_t[r_{0t}|\mathcal{E}_t^c]\mathbb{P}(\mathcal{E}_t^c) \le \sum_{i=1}^{n}\mathbb{E}[A_t^{-i}|\mathcal{E}_t] + (n-1)\mathbb{E}[B_t|\mathcal{E}_t] + V_{\max}\mathbb{P}_t(\mathcal{E}_t^c)$. Summing over all t yields the following bound on the seller regret:

$$\begin{aligned}\mathbb{E}[R_{0T}] \;&\le\; \sum_{t\in\mathbf{E}} V_{\max} \;+\; \sum_{t\notin\mathbf{E}}\mathbb{E}_t[r_{0t}] \\ &\le\; 3V_{\max}K^{1/3}T^{2/3} \;+\; \sum_{t\notin\mathbf{E}}\sum_{i=1}^{n}\mathbb{E}_t[A_t^{-i}|\mathcal{E}_t] \;+\; (n-1)\sum_{t\notin\mathbf{E}}\mathbb{E}_t[B_t|\mathcal{E}_t] \;+\; V_{\max}\sum_{t\notin\mathbf{E}}\mathbb{P}_t(\mathcal{E}_t^c).\end{aligned}$$

Here, for the first summation, we applied Lemma 18 to obtain $q_T K \le 3K^{1/3}T^{2/3}$. ■

9.5.1 PROOF OF PROPOSITION 6.1, AGENT REGRET

Let us first consider R_{iT}, the regret for agent i, when $\zeta = \mathsf{ETC}$. We will apply Lemma 14 and proceed to control the a_t, b_t terms for the two different choices for λ when $\mathcal{E}_t$ holds. First consider a_t. When $\lambda = \mathsf{AGE}$, we have $F_t^{-i} = \breve{V}_t^{-i}$ and therefore $a_t = 0$ a.s. When $\lambda = \mathsf{SEL}$, we have $F_t^{-i} = \widehat{V}_t^{-i}$ and therefore under $\mathcal{E}_t$,

$$a_t \;=\; \widehat{V}_t^{-i}(\omega_t^{-i}) - \breve{V}_t^{-i}(\omega_t^{-i}) \;=\; \sum_{i=1}^{n} 2\beta_t\sigma_{it}(s_i(\omega_t^{-i})) \;\le\; 2\sqrt{2}\beta_t n\sigma K^{1/3}t^{-1/3}.$$

The last step uses an argument similar to (19) followed by Lemma 18. Along with Lemma 20, we have the following bounds on the sum of a_t's:

$$\sum_{t\notin\mathbf{E}}\mathbb{E}[a_t|\mathcal{E}_t] \;\le\; \begin{cases} 0 & \text{if } \lambda = \mathsf{AGE}, \\ 3\sqrt{2}\beta_T n\sigma K^{1/3}T^{2/3} & \text{if } f = \mathsf{SEL}. \end{cases} \tag{32}$$

Now consider b_t and assume the agent participates by rewards. When $\lambda = \mathsf{AGE}$, $g_{it} = \widehat{v}_{it}$ and $G_t^{-i} = \widehat{V}_t^{-i}$. We therefore have, $b_t = \widehat{v}_{it}(s_{it}) - \breve{v}_{it}(s_{it}) = 2\beta_t\sigma_{it}(s_{it}) \le 2\sqrt{2}\sigma\beta_t K^{1/3}t^{-1/3}$ under $\mathcal{E}_t$. Similarly, when $\lambda = \mathsf{SEL}$, $g_{it} = \breve{v}_{it}$ and $G_t^{-i} = \breve{V}_t^{-i}$, which results in $b_t = \widehat{V}_t(\omega_t) - \breve{V}_t(\omega_t) = 2\beta_t\sum_i \sigma_{it}(\omega_t) \le 2\sqrt{2}\sigma\beta_t n K^{1/3}t^{-1/3}$. For an agent participating by bids $\breve{v}_{it} = g_{it} = v_i$. The only change in the analysis is that now $b_t = \widehat{V}_t(\omega_t) - G_t(\omega_t)$ which can be bound in a similar fashion to

above depending on the value of λ. Accounting for these considerations, and using Lemma 20, we have the following bounds on the sum of b_t's:

$$\sum_{t\notin\mathbf{E}}\mathbb{E}[b_t|\mathcal{E}_t] \;\le\; \begin{cases} 3\sqrt{2}\beta_T\sigma\kappa_i K^{1/3}T^{2/3} & \text{if } \lambda = \mathsf{AGE}, \\ 3\sqrt{2}\beta_T\sigma n K^{1/3}T^{2/3} & \text{if } \lambda = \mathsf{SEL}. \end{cases} \tag{33}$$

Recall that $\kappa_i = 1$ if the agent participates by rewards and 0 if she participates by bids. Finally, an application of Lemma 11 leads to the following bounds for the agent regret:

$$\mathbb{E}[R_{iT}] \le \begin{cases} 4n + \big(3v_i^\dagger + 3\sqrt{2}\kappa_i\sigma\beta_T\big)K^{1/3}T^{2/3}, & \text{if } \lambda = \mathsf{AGE}, \\ 4n + \big(3v_i^\dagger + 6\sqrt{2}\sigma\beta_T n\,\big)K^{1/3}T^{2/3}, & \text{if } \lambda = \mathsf{SEL}. \end{cases}$$

9.5.2 PROOF OF PROPOSITION 6.1, SELLER REGRET

Now, we will consider R_{0T}, the seller regret, when $\zeta = \mathsf{ETC}$. We will apply Lemma 15 and proceed to control the A_t^{-i}, B_t terms for the two different choices for λ under $\mathcal{E}_t$. First consider the A_t^{-i} terms. When $\lambda = \mathsf{SEL}$, then $F_t^{-i} = \widehat{V}_t^{-i}$ is an upper bound for V^{-i} under $\mathcal{E}_t$. Hence, $A_t^{-i} = \max V^{-i} - \max \widehat{V}_t^{-i} \le 0$. When $\lambda = \mathsf{AGE}$, under $\mathcal{E}_t$, we obtain the following uniform bound on $V^{-i}(\omega) - F_t^{-i}(\omega)$:

$$\begin{aligned} \forall\omega\in\Omega,\; V^{-i}(\omega) - F_t^{-i}(\omega) &\le \widehat{V}_t^{-i}(\omega) - \breve{V}_t^{-i}(\omega) = \sum_{j\neq i} 2\beta_t\sigma_{jt}(s_j(\omega)) \\ &\le 2\sqrt{2}\sigma\beta_t(n-1)K^{1/3}t^{-1/3}. \end{aligned} \tag{34}$$

Here, the last step uses an argument similar to (19). Hence, by Lemma 19, we have $A_t^{-i} = \max V^{-i} - \max F_t^{-i} \le 2\sqrt{2}\sigma\beta_t(n-1)K^{1/3}t^{-1/3}$ under $\mathcal{E}_t$. Along with Lemma 20, we obtain the following.

$$\sum_{i=1}^{n}\sum_{t\notin\mathbf{E}}\mathbb{E}[A_t^{-i}|\mathcal{E}_t] \;\le\; \begin{cases} 3\sqrt{2}\sigma\beta_T n(n-1)K^{1/3}T^{2/3} & \text{if } \lambda = \mathsf{AGE}, \\ 0 & \text{if } \lambda = \mathsf{SEL}. \end{cases} \tag{35}$$

Now, we turn to B_t. For this, note that under $\mathcal{E}_t$, $V(\omega_\star) \ge V(\omega_t) \ge \breve{V}_t(\omega_t)$. When $\lambda = \mathsf{SEL}$, we have $G_t = \breve{V}_t$ and therefore $B_t \le 0$. When $\lambda = \mathsf{AGE}$, we have $G_t = \widehat{V}_t$ and therefore,

$$B_t \;\le\; \widehat{V}_t(\omega_t) - \breve{V}_t(\omega_t) \;=\; 2\beta_t\sum_{i=1}^{n}\sigma_{it}(s_{it}) \;\le\; 2\sqrt{2}\sigma n\beta_t K^{1/3}t^{-1/3}.$$

This yields the following bounds for the sum of B_t's.

$$(n-1)\sum_{t\notin\mathbf{E}}\mathbb{E}[B_t|\mathcal{E}_t] \;\le\; \begin{cases} 3\sqrt{2}\sigma\beta_T n(n-1)K^{1/3}T^{2/3} & \text{if } \lambda = \mathsf{AGE}, \\ 0 & \text{if } \lambda = \mathsf{SEL}. \end{cases} \tag{36}$$

Combining the above results with Lemma 15 and Lemma 11 leads to the following bounds for the seller regret:

$$\mathbb{E}[R_{0T}] \le \begin{cases} 4nV_{\max} + \Big(3V_{\max} + 6\sqrt{2}\sigma\beta_T n(n-1)\Big)K^{1/3}T^{2/3}. & \text{if } \lambda = \mathsf{AGE}, \\ 4nV_{\max} + 3V_{\max}K^{1/3}T^{2/3}. & \text{if } \lambda = \mathsf{SEL}. \end{cases}$$

9.5.3 PROOF OF PROPOSITION 6.2, AGENT REGRET

Next, we will consider agent i's regret R_{iT}, when $\zeta = \mathsf{OPT}$. As in Section 9.5.1, we will use Lemma 14 and control the a_t, b_t terms for the two different choices for λ under $\mathcal{E}_t$. First, a_t is bounded identically to obtain the upper bound in (32); this uses the fact that even when $\zeta = \mathsf{OPT}$, there will have been q_T exploration phases by round T.

Next, consider b_t. When $\lambda = \mathsf{AGE}$, $g_{it} = \widehat{v}_{it}$ and $G_t^{-i} = \widehat{V}_t^{-i}$. We therefore have, $b_t = \widehat{v}_{it}(s_{it}) - \breve{v}_{it}(s_{it}) = 2\beta_t \sigma_{it}(s_{it})$ if the agent is participating by rewards and $b_t = 0$ if she is participating by bids. Similarly, when $\lambda = \mathsf{SEL}$, $b_t = \widehat{V}_t(\omega_t) - \breve{V}_t(\omega_t) \le 2\beta_t \sum_i \sigma_{it}(\omega_t)$ under $\mathcal{E}_t$. Using a similar argument to (25), we obtain the following bounds on the sum of b_t's when $t \notin \mathbf{E}$:

$$\sum_{t \notin \mathbf{E}} \mathbb{E}[b_t | \mathcal{E}_t] \;\le\; \begin{cases} 4\kappa_i \beta_T \sigma |\mathcal{S}|^{1/2} T^{1/2} & \text{if } \lambda = \mathsf{AGE}, \\ 4\beta_T \sigma n |\mathcal{S}|^{1/2} T^{1/2} & \text{if } \lambda = \mathsf{SEL}. \end{cases} \tag{37}$$

Combining the above results with Lemma 14 and Lemma 11 leads to the following bounds for the agent regret:

$$\mathbb{E}[R_{iT}] \le \begin{cases} 6n + 4\sigma\beta_T \kappa_i |\mathcal{S}|^{1/2} T^{1/2} + 3v_i^\dagger K^{1/3} T^{2/3}, & \text{if } \lambda = \mathsf{AGE}, \\ 6n + 4\sigma\beta_T n |\mathcal{S}|^{1/2} T^{1/2} + (3v_i^\dagger + 3\sqrt{2}\sigma\beta_T n) K^{1/3} T^{2/3}, & \text{if } \lambda = \mathsf{SEL}. \end{cases}$$

9.5.4 PROOF OF PROPOSITION 6.2, SELLER REGRET

Finally, we will consider the seller regret R_{0T}, when $\zeta = \mathsf{OPT}$. Following along the same lines as Section 9.5.2, we will use Lemma 15 to control the seller regret, and moreover, use the expression in (35) to bound the $\mathbb{E}[A_t^{-i} | \mathcal{E}_t]$ terms. The same bounding technique can be used since, even when $\zeta = \mathsf{OPT}$, there will have been q_T exploration phases by round T.

To bound the B_t terms, we first observe that under $\mathcal{E}_t$, $V(\omega_\star) \ge V(\omega_t) \ge \breve{V}_t(\omega_t)$. When $\lambda = \mathsf{SEL}$, we have $G_t = \breve{V}_t$, and therefore $B_t \le 0$. When $\lambda = \mathsf{AGE}$, we have $G_t = \widehat{V}_t$, and therefore $B_t \le \widehat{V}_t(\omega_t) - \breve{V}_t(\omega_t) = 2\beta_t \sum_i \sigma_{it}(s_{it})$. Putting these results together and using a similar argument to (25), we obtain the following bounds on the sum of B_t's:

$$(n-1) \sum_{t \notin \mathbf{E}} \mathbb{E}[B_t | \mathcal{E}_t] \;\le\; \begin{cases} 4\sigma\beta_T n(n-1) |\mathcal{S}|^{1/2} T^{1/2} & \text{if } \lambda = \mathsf{AGE}, \\ 0 & \text{if } \lambda = \mathsf{SEL}. \end{cases} \tag{38}$$

Combining the above results with Lemma 15 and Lemma 11 leads to the following bounds for the seller regret:

$$\mathbb{E}[R_{0T}] \le \begin{cases} 6nV_{\max} + 4\sigma\beta_T n(n-1)|\mathcal{S}|^{1/2} T^{1/2} + \Big(3V_{\max} + 3\sqrt{2}\sigma\beta_T n(n-1)\Big) K^{1/3} T^{2/3}, & \text{if } \lambda = \mathsf{AGE}, \\ 6nV_{\max} + 3V_{\max} K^{1/3} T^{2/3}, & \text{if } \lambda = \mathsf{SEL}. \end{cases}$$

9.6 Proof of Theorem 4

We now bound R_T^{VCG}. First, we provide a bound on R_T^{VCG} in terms of A_t^{-i} and B_t defined in (30).

Lemma 16 *Let A_t^{-i}, B_t be as defined in* (30)*. Then, the following bound holds on the VCG regret R_T^{VCG} defined in* (3)*.*

$$\mathbb{E}[R_T^{\mathrm{VCG}}] \leq 2\sum_{i=1}^{n}\sum_{t\notin\mathbf{E}}\mathbb{E}[|A_t^{-i}|\,|\mathcal{E}_t] + 2(n-1)\sum_{t\notin\mathbf{E}}\mathbb{E}[|B_t|\,|\mathcal{E}_t] + 6V_{\max}K^{1/3}T^{2/3} + (n+1)R_T^{\mathrm{w}} + 2V_{\max}\sum_{t\notin\mathbf{E}}\mathbb{P}_t(\mathcal{E}_t^c).$$

Proof Recall that R_T^{w} is always non-negative while R_T^{a} and R_{0T} may be positive or negative. From Lemma 8, we have $R_T^{\mathrm{a}} + R_{0T} = R_T^{\mathrm{w}}$. Since the maximum is smaller than the sum, we have

$$R_T^{\mathrm{VCG}} = \max(nR_T^{\mathrm{w}}, R_T^{\mathrm{a}}, R_{0T}) \leq nR_T^{\mathrm{w}} + |R_T^{\mathrm{a}}| + |R_{0T}| \leq (n+1)R_T^{\mathrm{w}} + 2|R_{0T}|. \tag{39}$$

By the triangle inequality, we obtain the following bound on $|R_{0T}|$, similar to Lemma 15:

$$\begin{aligned}\mathbb{E}[|R_{0T}|] \;&\leq\; \sum_{t\in\mathbf{E}}\mathbb{E}[|r_{0t}|] \;+\; \sum_{t\notin\mathbf{E}}\mathbb{E}[|r_{0t}|\,|\mathcal{E}_t] \;+\; \sum_{t\notin\mathbf{E}}\mathbb{E}[|r_{0t}|\,|\mathcal{E}_t] \\ &\leq\; V_{\max}Kq_T \;+\; \sum_{i=1}^{n}\sum_{t\notin\mathbf{E}}\mathbb{E}[|A_t^{-i}|\,|\mathcal{E}_t] \;+\; (n-1)\sum_{t\notin\mathbf{E}}\mathbb{E}[|B_t|\,|\mathcal{E}_t] + V_{\max}\sum_{t\notin\mathbf{E}}\mathbb{P}_t(\mathcal{E}_t^c).\end{aligned}$$

The claim follows by combining the above bound with (39) and then applying Lemma 18 for a bound on Kq_T. ■

We are now ready to prove Theorem 4.

Proof of Theorem 4. First consider the $\zeta = \mathsf{ETC}$ case. We will use Lemma 16 to control R_T^{VCG}. We already have an upper bound on R_T^{w} from Section 9.4, and upper bounds on A_t^{-i} and B_t from Sections 9.5.2 and 9.5.4. The lower bounds for A_t^{-i}, B_t are obtained by simply reversing the argument.

First consider, $|A_t^{-i}|$. If $\lambda = \mathsf{SEL}$, we already saw $A_t^{-i} \leq 0$. By using an argument similar to (34), we obtain $-A_t^{-i} \leq 2\sqrt{2}\sigma\beta_t(n-1)K^{1/3}t^{-1/3}$. Similarly, if $\lambda = \mathsf{AGE}$, we already saw $A_t^{-i} \leq 2\sqrt{2}\sigma\beta_t(n-1)K^{1/3}t^{-1/3}$. Moreover,

$$-A_t^{-i} = \breve{V}_t^{-i}(\omega_t^{-i}) - V^{-i}(\omega_\star) \leq V^{-i}(\omega_t^{-i}) - V^{-i}(\omega_\star) \leq 0.$$

Therefore, for both λ values we have $|A_t^{-i}| \leq 2\sqrt{2}\sigma\beta_t(n-1)K^{1/3}t^{-1/3}$. After an application of Lemma 20, we obtain,

$$\sum_i\sum_{t\notin\mathbf{E}}\mathbb{E}[|A_t^{-i}|\,|\mathcal{E}_t] \leq 3\sqrt{2}\sigma\beta_t n(n-1)K^{1/3}T^{2/3}. \tag{40}$$

By following a similar argument, we can obtain $(n-1)\sum_{t\notin\mathbf{E}}\mathbb{E}[|B_t|\,|\mathcal{E}_t] \leq 3\sqrt{2}\sigma\beta_t n(n-1)K^{1/3}T^{2/3}$. The claim follows by combining the above with Lemma 16, the bound on the welfare regret in (27), and Lemma 11 to control $\sum_{t\notin\mathbf{E}}\mathbb{P}_t(\mathcal{E}_t^c)$.

Now consider the $\zeta = \mathsf{OPT}$ case. Since there will have been q_T exploration phases in T rounds, we use the expression in (40) to bound the sum of $\mathbb{E}[|A_t^{-i}||\mathcal{E}_t]$ terms. Next, let us turn to the $|B_t|$ terms

in the RHS of Lemma 16. When $\lambda = \textsf{AGE}$, we already saw in Section 9.5.4 that $B_t \leq 0$. Moreover, $-B_t = V(\omega_\star) - \check{V}_t(\omega_t) \leq \widehat{V}_t(\omega_t) - \check{V}_t(\omega_t) \leq 2\beta_t \sum_i \sigma_{it}(s_{it})$ under $\mathcal{E}_t$. Similarly, when $\lambda = \textsf{SEL}$, we already saw $B_t \leq 2\beta_t \sum_i \sigma_{it}(s_{it})$. Moreover, $-B_t = V(\omega_\star) - \widehat{V}_t(\omega_t) \leq 0$. In all cases, we have $|B_t| \leq 2\beta_t \sum_i \sigma_{it}(s_{it})$ and therefore, by following the same calculations in (25), we have,

$$(n-1)\sum_{t \notin \mathbf{E}} \mathbb{E}[|B_t| \,|\mathcal{E}_t] \leq 4\sigma\beta_T n(n-1)|\mathcal{S}|^{1/2}T^{1/2}.$$

The claim follows by combining the above with Lemma 16, the bound on the welfare regret in (28), and Lemma 11 to control $\sum_{t \notin \mathbf{E}} \mathbb{P}_t(\mathcal{E}_t^c)$. □

9.7 Some Technical Lemmas

This section states some technical results that were used throughout our proofs. The following fact, akin to Fact 7, is straightforward to verify.

Fact 17 *In round t of Algorithm 1, the agent and seller utilities satisfy the following for the given $\{f_{it}, g_{it}\}_{i,t}$ choices.*

$$\text{if } t \in \mathbf{E},\ \ u_{it} = v_i(s_{it}), \qquad \text{if } t \notin \mathbf{E},\ \ u_{it} = v_i(s_{it}) - g_{it}(s_{it}) + G_t(\omega_t) - F_t^{-i}(\omega_t^{-i}),$$
$$\text{if } t \in \mathbf{E},\ \ u_{0t} = v_0(\omega_t), \qquad \text{if } t \notin \mathbf{E},\ \ u_{0t} = \sum_{i=1}^n F_t^{-i}(\omega_t^{-i}) - (n-1)G_t(\omega_t).$$

The following result bounds the number of brackets q_T (12) after a given number of rounds T.

Lemma 18 *Consider Algorithm 1 on the T^{th} round, where $T > 2K$, and let q_T denote the current bracket index. If T is an exploration round, i.e. $T \in \mathbf{E}$, then $q_T \leq 3K^{-2/3}T^{2/3}$. If $T \notin \mathbf{E}$, then, $\frac{1}{2}K^{-2/3}T^{2/3} \leq q_T \leq 3K^{-2/3}T^{2/3}$.*

Proof For brevity, write $q = q_T$, $c = 5/6$, $d = 1/2$. First let $T \notin \mathbf{E}$. Using the notation in (12), we have

$$T_{q-1} + K < T \leq T_q, \quad \text{where, } T_m = Km + \sum_{t=1}^m \lfloor cKt^d \rfloor.$$

To bound T_m, letting $S_m = \sum_{t=1}^m t^d$ and bounding the sum of an increasing function by an integral we have,

$$\int_0^m t^d \mathrm{d}t < S_m < \int_1^{m+1} t^d \mathrm{d}t \quad \implies \quad \frac{m^{d+1}}{d+1} < S_m < \frac{(m+1)^{d+1}-1}{d+1}. \tag{41}$$

This leads to the following bounds on T,

$$\begin{aligned} T \leq T_q &\leq qK + \sum_{t=1}^q cKt^d \leq qK + \frac{cK}{d+1}\big((q+1)^{d+1} - 1\big) \\ &\leq q^{d+1}K + \frac{cK}{d+1}(2q)^{d+1} \leq c_1 q^{3/2}K, \end{aligned} \tag{42}$$

$$T \geq T_{q-1} + K \geq qK + \sum_{t=1}^{q-1}(cKt^d - 1) \geq qK - (q-1) + \frac{cK}{d+1}(q-1)^{d+1} \tag{43}$$
$$\geq \frac{cK}{d+1}\left(\frac{q}{2}\right)^{d+1} = c_2 q^{3/2} K.$$

In (42), we have used the upper bound in (41) with $m = q$, and the facts $q \leq q^{3/2}$, $q+1 \leq 2q$. In (43), we have used the lower bound in (41) with $m = q-1$, and the facts $qK > q-1$, $q-1 \geq q/2$; the last inequality holds when $q \geq 2$ which is true when $T \geq 2K$. Now, by substituting the values for c and d, we have $c_1 = 1 + 10\sqrt{2}/9$ and $c_2 = 5\sqrt{2}/36$. Thus,

$$q \leq \left(\frac{T}{c_2 K}\right)^{2/3} \leq 3\frac{T^{2/3}}{K^{2/3}}, \qquad q \geq \left(\frac{T}{c_1 K}\right)^{2/3} \geq \frac{1}{2}\frac{T^{2/3}}{K^{2/3}}.$$

This proves the result for $T \notin \mathbf{E}$. If $T \in \mathbf{E}$, by noting that $T \geq T_{q-1}$, we can repeat the calculations in (43) to obtain the same bound. ■

The following two results were used repeatedly throughout our proofs.

Lemma 19 *Let $f_1, f_2 : \mathcal{X} \to \mathbb{R}$ for some finite set $\mathcal{X}$ such that $f_1(x) - f_2(x) \leq \epsilon$ for all $x \in \mathcal{X}$ and a given $\epsilon \geq 0$. Then* $\max f_1 - \max f_2 \leq \epsilon$.

Proof *Let $x_i = \operatorname{argmax} f_i$. Then, $f_1(x_1) - f_2(x_2) \leq f_2(x_1) - f_2(x_2) + \epsilon \leq \epsilon$.* ■

Lemma 20 $\sum_{t=1}^{n} t^{-1/2} \leq 2n^{1/2}, \quad \sum_{t=1}^{n} t^{-1/3} \leq \frac{3}{2}n^{2/3}$.

Proof *By bounding the summation of a decreasing function by an integral we have for $r \in [0,1]$, $\sum_{t=1}^{n} t^{-r} \leq 1 + \int_1^n t^{-r}\mathrm{d}t \leq 1 + \frac{n^{1-r}}{1-r} - \frac{1}{1-r} \leq \frac{n^{1-r}}{1-r}$. Setting $r = 1/2, 1/3$ yields the results.* ■

10. Conclusion

We have studied mechanism design in settings where agents may not know their values, but can experience an allocation and report back a realised reward. The goal of the mechanism is to learn the values of the users while simultaneously finding the optimal outcome and satisfying game-theoretic desiderata such as individual rationality and truthfulness. We established a lower bound on the VCG regret for this problem, and presented an algorithm that essentially achieves this rate. The proposed framework allows a practitioner to control trade-offs between various properties that they might be interested in, such as agent and seller regrets, individual rationality, and truthfulness. We conclude with two avenues for future work.

First, we have assumed that we can maximise the upper confidence bound (6) exactly. In many settings, this might be computationally prohibitive, and we might only be able to obtain an approximate solution. It will be instructive to study which of the desiderata carry through in this case. If we have

an α-approximate solver ($\alpha < 1$), it is straightforward to show that sublinear welfare regret (3) is possible under truthful reporting, if it is defined as $R_T^{\mathrm{w}} = \alpha T V(\omega_\star) - \sum_t V(\omega_t)$. However, bounding the agent and seller regrets requires more careful analysis as their utility depends on the near-optimal outcome chosen by the solver. Implications on truthfulness are even less clear, especially as an agent can be strategic over multiple rounds.

Second, the lower bound in Theorem 1 only captures one of the two key difficulties in this problem, namely pricing calculation for agent/seller trade-offs; the other being truthfulness. It is worth studying the implications of even asymptotic truthfulness on learning. While our algorithm is optimal with respect to the lower bound for the VCG regret, in some applications, it is not necessary to minimise all three regret terms in R_T^{VCG}. For instance, the PaaS setting in Example 1 could occur within an organisation, where the service provider is one team providing a service to other (agent) teams. In such cases, the seller regret is not a meaningful quantity. In this setting, it is possible to obtain $\sqrt{T}$ regret for both the welfare and the agents *if* the agents report truthfully: at *all* time steps, select the outcome which maximises the upper confidence bound on the welfare and choose a favourable pricing scheme to the agents, such as the one obtained by setting $\lambda = \textsc{age}$. However, this is not a truthful mechanism. In situations like this, we believe that truthfulness will prevent obtaining $\sqrt{T}$ regret. For instance, Babaioff et al. (2013) and Devanur and Kakade (2009) show that $T^{2/3}$ regret is unavoidable for deterministic truthful algorithms in their online advertising problem. Their proof relies heavily on a necessary and sufficient condition for truthfulness in single-parameter auctions where agents submit bids (Myerson, 1981; Archer and Tardos, 2001). Extensions of this condition to multi-parameter auctions exist (Rochet, 1987), but only in instances where agents submit bids ahead of time. This characterisation does not apply in our problem where the agent does not know her value and reports a reward at the end of the round.

Acknowledgments

We would like to thank the reviewers for the valuable feedback in improving the presentation of our results and Matthew Wright for providing feedback on an initial draft of this manuscript. Michael I Jordan is partially supported by NSF Grant IIS-1901252.

References

Gagan Aggarwal, Ashish Goel, and Rajeev Motwani. Truthful Auctions for Pricing Search Keywords. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 1–7, 2006.

Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Learning Prices for Repeated Auctions with Strategic Buyers. In *Advances in Neural Information Processing Systems*, pages 1169–1177, 2013.

Aaron Archer and Éva Tardos. Truthful Mechanisms for One-parameter Agents. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*, pages 482–491. IEEE, 2001.

Susan Athey and Ilya Segal. An Efficient Dynamic Mechanism. *Econometrica*, 81(6):2463–2485, 2013.

Peter Auer. Using Confidence Bounds for Exploitation-exploration Trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Moshe Babaioff, Robert Kleinberg, and Aleksandrs Slivkins. Multi-parameter mechanisms with implicit payment computation. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, pages 35–52, 2013.

Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. *SIAM Journal on Computing*, 43(1):194–230, 2014.

Moshe Babaioff, Robert D Kleinberg, and Aleksandrs Slivkins. Truthful Mechanisms with Implicit Payment Computation. *Journal of the ACM*, 62(2):1–37, 2015.

Maria-Florina Balcan, Avrim Blum, Jason D Hartline, and Yishay Mansour. Reducing mechanism design to algorithm design via machine learning. *Journal of Computer and System Sciences*, 74 (8):1245–1270, 2008.

Maria-Florina F Balcan, Tuomas Sandholm, and Ellen Vitercik. Sample complexity of automated mechanism design. In *Advances in Neural Information Processing Systems*, pages 2083–2091, 2016.

Dirk Bergemann and Juuso Valimaki. Efficient Dynamic Auctions. 2006.

Avrim Blum, Yishay Mansour, and Jame Morgenstern. Learning Valuation Distributions from Partial Observation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Mark Braverman, Jieming Mao, Jon Schneider, and S Matthew Weinberg. Multi-armed Bandit Problems with Strategic Arms. In *Conference on Learning Theory*, pages 383–416. PMLR, 2019.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695, 2011.

Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded Regret in Stochastic Multi-armed Bandits. In *Conference on Learning Theory*, pages 122–134, 2013.

Edward H Clarke. Multipart Pricing of Public Goods. *Public Choice*, 1971.

Peter Cramton. Spectrum Auction Design. *Review of Industrial Organization*, 42(2):161–190, 2013.

Constantinos Daskalakis, Aranyak Mehta, and Christos Papadimitriou. A Note on Approximate Nash Equilibria. In *International Workshop on Internet and Network Economics*, pages 297–306. Springer, 2006.

Nikhil R Devanur and Sham M Kakade. The Price of Truthfulness for Pay-per-click Auctions. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 99–106, 2009.

Miroslav Dudik, Nika Haghtalab, Haipeng Luo, Robert E Schapire, Vasilis Syrgkanis, and Jennifer Wortman Vaughan. Oracle-efficient Online Learning and Auction Design. In *IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 528–539. IEEE, 2017.

Tomas Feder, Hamid Nazerzadeh, and Amin Saberi. Approximating Nash Equilibria using Small-support Strategies. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 352–354, 2007.

Zhe Feng, Chara Podimata, and Vasilis Syrgkanis. Learning to Bid without Knowing Your Value. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 505–522, 2018.

Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On Explore-then-commit Strategies. In *Advances in Neural Information Processing Systems*, 2016.

Nicola Gatti, Alessandro Lazaric, and Francesco Trovò. A Truthful Learning Mechanism for Contextual Multi-slot Sponsored Search Auctions with Externalities. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 605–622, 2012.

Theodore Groves. Efficient Collective Choice when Compensation is Possible. *The Review of Economic Studies*, 1979.

Sham M Kakade, Ilan Lobel, and Hamid Nazerzadeh. An Optimal Dynamic Mechanism for Multi-armed Bandit Processes. *arXiv preprint arXiv:1001.4598*, 2010.

Sham M Kakade, Ilan Lobel, and Hamid Nazerzadeh. Optimal Dynamic Mechanism Design and the Virtual-pivot Mechanism. *Operations Research*, 61(4):837–854, 2013.

Anna R Karlin and Yuval Peres. *Game Theory, Alive*. American Mathematical Society, 2017.

Fuhito Kojima and Mihai Manea. Incentives in the Probabilistic Serial Mechanism. *Journal of Economic Theory*, 145(1):106–123, 2010.

Sébastien Lahaie, David M Pennock, Amin Saberi, and Rakesh V Vohra. Sponsored Search Auctions. *Algorithmic Game Theory*, 1:699–716, 2007.

Tze Leung Lai and Herbert Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 1985.

Richard J Lipton, Evangelos Markakis, and Aranyak Mehta. Playing Large Games using Simple Strategies. In *Proceedings of the 4th ACM conference on Electronic commerce*, pages 36–41, 2003.

Lydia T Liu, Horia Mania, and Michael I Jordan. Competing Bandits in Matching Markets. In *Proceedings of the Twenty-Third Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian Incentive-compatible Bandit Exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582, 2015.

Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. Adwords and Generalized Online Matching. *Journal of the ACM*, 54(5):22–es, 2007.

Paul Milgrom. *Discovering Prices*. Columbia University Press, 2017.

Roger B Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

Hamid Nazerzadeh, Amin Saberi, and Rakesh Vohra. Dynamic Cost-per-action Mechanisms and Applications to Online Advertising. In *Proceedings of the 17th International Conference on World Wide Web*, pages 179–188, 2008.

Hamid Nazerzadeh, Renato Paes Leme, Afshin Rostamizadeh, and Umar Syed. Where to Sell: Simulating Auctions from Learning Algorithms. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 597–598, 2016.

Thomas Nedelec, Noureddine El Karoui, and Vianney Perchet. Learning to Bid in Revenue-maximizing Auctions. In *International Conference on Machine Learning*, pages 4781–4789. PMLR, 2019.

Noam Nisan and Amir Ronen. Algorithmic Mechanism Design. *Games and Economic Behavior*, 35 (1-2):166–196, 2001.

Vianney Perchet, Philippe Rigollet, et al. The Multi-armed Bandit Problem with Covariates. *The Annals of Statistics*, 41(2):693–721, 2013.

Ariel D Procaccia. Cake Cutting: Not just Child's Play. *Communications of the ACM*, 56(7):78–87, 2013.

Donald John Roberts and Andrew Postlewaite. The Incentives for Price-taking Behavior in Large Exchange Economies. *Econometrica: Journal of the Econometric Society*, pages 115–127, 1976.

Jean-Charles Rochet. A necessary and sufficient condition for rationalizability in a quasi-linear context. *Journal of Mathematical Economics*, 16(2):191–200, 1987.

Alvin E Roth. On the Allocation of Residents to Rural Hospitals: A General Property of Two-sided Matching Markets. *Econometrica: Journal of the Econometric Society*, pages 425–427, 1986.

Alvin E Roth, Tayfun Sönmez, and M Utku Ünver. Kidney Exchange. *The Quarterly journal of economics*, 119(2):457–488, 2004.

James Schummer. Almost-dominant Strategy Implementation: Exchange Economies. *Games and Economic Behavior*, 48(1):154–170, 2004.

W. R. Thompson. On the Likelihood that one Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 1933.

Adel Nadjaran Toosi, Kurt Vanmechelen, Farzad Khodadadi, and Rajkumar Buyya. An Auction Mechanism for Cloud Spot Markets. *ACM Transactions on Autonomous and Adaptive Systems*, 11 (1):1–33, 2016.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008.

William Vickrey. Counterspeculation, Auctions, and Competitive Sealed Tenders. *The Journal of Finance*, 1961.

Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions. In *Conference on Learning Theory*, pages 1562–1583, 2016.