MIT CEEPR

MIT Center for Energy and Environmental Policy Research

Working Paper Series

Learning in Repeated Multi-Unit Pay-As-Bid Auctions

Rigel Galgana and Negin Golrezaei

OCTOBER 2023

CEEPR WP 2023-18

Working Paper Series.

Since 1977, the Center for Energy and Environmental Policy Research (CEEPR) has been a focal point for research on energy and environmental policy at MIT. CEEPR promotes rigorous, objective research for improved decision making in government and the private sector, and secures the relevance of its work through close cooperation with industry partners from around the globe. Drawing on the unparalleled resources available at MIT, affiliated faculty and research staff as well as international research associates contribute to the empirical study of a wide range of policy issues related to energy supply, energy demand, and the environment.

An important dissemination channel for these research efforts is the MIT CEEPR Working Paper series. CEEPR releases Working Papers written by researchers from MIT and other academic institutions in order to enable timely consideration and reaction to energy and environmental policy research, but does not conduct a selection process or peer review prior to posting. CEEPR's posting of a Working Paper, therefore, does not constitute an endorsement of the accuracy or merit of the Working Paper. If you have questions about a particular Working Paper, please contact the authors or their home institutions.

Learning in Repeated Multi-Unit Pay-As-Bid Auctions

Rigel Galgana
Operations Research Center, Massachusetts Institute of Technology, rgalgana@mit.edu,

Negin Golrezaei
Sloan School of Management, Massachusetts Institute of Technology, golrezaei@mit.edu,

Motivated by Carbon Emissions Trading Schemes, Treasury Auctions, Procurement Auctions, and Wholesale Electricity Markets, which all involve the auctioning of homogeneous multiple units, we consider the problem of learning how to bid in repeated multi-unit pay-as-bid auctions. In each of these auctions, a large number of (identical) items are to be allocated to the largest submitted bids, where the price of each of the winning bids is equal to the bid itself. In this work, we study the problem of optimizing bidding strategies from the perspective of a single bidder.

This problem is challenging due to the combinatorial nature of the action space. We overcome this challenge by focusing on the offline setting, where the bidder optimizes their vector of bids while only having access to the past submitted bids by other bidders. We show that the optimal solution to the offline problem can be obtained using a polynomial time dynamic programming (DP) scheme under which the bidder's utility is decoupled across units. We leverage the structure of the DP scheme to design online learning algorithms with polynomial time and space complexity under full information and bandit feedback settings. Under these two feedback structures, we achieve an upper bound on regret of $O(M\sqrt{T\log|\mathcal{B}|})$ and $O(M\sqrt{|\mathcal{B}|T\log|\mathcal{B}|})$ respectively, where M is the number of units demanded by the bidder, T is the total number of auctions, and $|\mathcal{B}|$ is the size of the discretized bid space. We accompany these results with a regret lower bound, which match the linear dependency in M.

Our numerical results suggest that when all agents behave according to our proposed no regret learning algorithms, the resulting market dynamics mainly converge to a welfare maximizing equilibrium where bidders submit uniform bids. We further show that added competition reduces the impact of strategization and bidders converge more rapidly to a higher revenue and welfare steady state. Lastly, our experiments demonstrate that the pay-as-bid auction consistently generates significantly higher revenue compared to its popular alternative, the uniform price auction. This advantage positions the pay-as-bid auction as an appealing auction format in settings where earning high revenue holds significant social value, such as the Carbon Emissions Trading Scheme.

Keywords. Multi-unit pay-as-bid auctions, Bidding strategies, Regret analysis, Market dynamics.

---

1. Introduction

Homogeneous multi-unit auctions, a special case of combinatorial auctions, are commonly used to auction off large quantities of identical items, for example in Carbon Emissions Trading Schemes Goulder and Schein (2013), Schmalensee and Stavins (2017), Goldner et al. (2019), US Treasury Auctions Garbade and Ingber (2005), Binmore and Swierzbinski (2000), Procurement Auctions Ausubel and Cramton (2005), and Wholesale Electricity Markets Tierney et al. (2008), Federico and Rahman (2003), Fabra et al. (2006). In these multi-unit auctions, bidders submit a bid vector and then are allocated their goods and charged payments according to the auction format.

Of widespread use are the uniform price and pay-as-bid (PAB) mechanisms. As natural multi-unit generalizations of the second and first price sealed bid auctions, bidders are allocated units in decreasing order of bids and, for each unit won, are charged as payment the lowest winning bid (uniform price) or their own bid (PAB). In this work, we focus on the PAB auction in light of the recent industry and research community-wide push towards first price auctions, which is mainly due to demand for price transparency and ease of revenue management (Deospotakis et al. 2021, Bigler 2019).

Nevertheless, it has been observed that participants find it challenging to determine how to bid effectively in PAB auctions, as highlighted by Porter et al. (2003). Bidders face a fundamental dilemma: larger bids increase their chances of winning units but also lead to higher payments. This predicament is further complicated by the necessity to submit monotone bid vectors. Bidding too conservatively reduces the probability of winning units in subsequent slots, while bidding excessively may inflate the payment.

In this paper, we address the issue of learning optimal bidding strategies in repeated multi-unit PAB auctions. As we will elaborate later, we develop efficient no-regret algorithms that simplify the bidding complexity associated with PAB auctions. Through simulating the market dynamics derived from these learning algorithms, we empirically analyze the equilibria of PAB auctions, which have been poorly understood prior to our research. Our empirical findings demonstrate that in the equilibria resulting from these market dynamics, bidders' winning bids converge to the same value, thus addressing concerns regarding price fairness in PAB auctions (Binmore and Swierzbinski 2000, Akbarpour and Li 2020).

We also consistently observe high revenue from these equilibria, especially when compared to its uniform price counterpart. In the context of carbon markets, this additional revenue can be invested into clean-up efforts and green technology; see. e.g., phase 4 of the European Union Emissions Trading System (EU-ETS) (Gregor 2023).

1.1. Technical Contributions

New Framework to Study Learning How to Bid in PAB Auctions (Section 2). Let there be N bidders/agents with M-unit demand in a PAB auction with $\overline{M}$ supply. Each bidder n is endowed with valuation vector $\boldsymbol{v}_n = (v_{n,1}, \dots, v_{n,M}) \in [0,1]^M$ and submits a bid vector $\boldsymbol{b}_n = (b_{n,1}, \dots, b_{n,M}) \in \mathcal{B}^M$, where $\mathcal{B}$ is some discretization of $[0,1]$ that represents the set of all possible bids. Agents then receive allocation $x_n \in [M]$ and utility $\sum_{m=1}^{x_n}(v_{n,m} - b_{n,m})$ according to the PAB auction rule; see Section 2. Repeating this auction across T rounds, each agent's goal is to minimize their regret with respect to their hindsight, utility maximizing bid vector. Here, we refer to discretized regret as the regret incurred as a function of $M, \mathcal{B}, T$ when restricting our bid space to $\mathcal{B}$. Conversely, we refer to continuous regret as the regret incurred as a function of only M and T when optimizing for the discretization error.

Dynamic Programming Scheme for Hindsight Optimal Offline Solution (Section 3). To design low-regret bidding algorithms, we crucially leverage the structure of the hindsight optimal offline solution. In the offline/hindsight problem, the bidder has access to the (historical) dataset of submitted bids by competitors and seeks to find the utility maximizing bid vector on that dataset. (See Roughgarden and Wang (2016), Derakhshan et al. (2022), Golrezaei et al. (2021b), Derakhshan et al. (2021) for works that study similar problems from an auctioneer's perspective.)

We show that the optimal solution to the offline problem—which is our benchmark in computing the regret of our online learning algorithms—can be solved using a polynomial time Dynamic Programming (DP) scheme. To do so, we make the following key observation: to win m units (or equivalently, slots), an agent n must have at least m bids larger than the smallest m among the largest $\overline{M}$ bids of all other bidders. This observation allows us to devise a DP where in each step of the DP, we decide about the bid for one unit, while considering the externality that this bid will impose on the bids and utilities for other units. This externality is precisely the fundamental tradeoff of PAB auctions aforementioned: bidding too small decreases the probability of winning the current or any subsequent units, however, bidding too large increases the payment of the current and previous units.

Decoupled Exponential Weights Algorithm (Section 4). We present our first set of algorithms to learn in the online setting, in both the full information and bandit feedback regimes. We leverage our DP scheme to obtain decoupled rewards, or reward estimates in the bandit setting, for each unit-bid value pair. In particular, we can obtain an exact expression for the utility estimate for bidding b_m for unit m that is independent of $b_1, \dots, b_{m-1}$ or $b_{m+1}, \dots, b_M$, subject to bid vector monotonicity. This allows us to mimic the exponential weights algorithm on the exponentially large bid space in polynomial time and space complexity. We show that our decoupled exponential weights algorithm (Algorithm 2) achieves time and space complexities of $O(M|\mathcal{B}|T)$ and $O(M|\mathcal{B}|)$ respectively, with discretized regret $O(M^{\frac{3}{2}}\sqrt{T\log|\mathcal{B}|})$; see Theorem 4. Optimizing for the discretization error from using finite bid space $\mathcal{B}$, we show that in the full information setting, our algorithm achieves $O(M^{\frac{3}{2}}\sqrt{T\log T})$ regret.

THEOREM 1 **(Informal)**. *Under full information feedback, there exists an algorithm that achieves discretized and continuous regrets $O(M^{\frac{3}{2}}\sqrt{T\log|\mathcal{B}|})$ and $O(M^{\frac{3}{2}}\sqrt{T\log T})$ respectively in the repeated multi-unit PAB auction.*

We also analyze the bandit version of our algorithm and show that it achieves discretized regret $O(M^{\frac{3}{2}}\sqrt{|\mathcal{B}|T\log|\mathcal{B}|})$; see Theorem 5. Balancing with the discretization error, this algorithm achieves continuous regret of $O(M^{\frac{4}{3}}T^{\frac{2}{3}}\sqrt{\log T})$ (We defer the proofs of these and our other main results to the appendix unless otherwise stated).

Online Mirror Descent Algorithm (Section 5). In this section, we present an alternative learning algorithm based on Online Mirror Descent (OMD) that improves the regret upper bounds of our decoupled exponential weight algorithm by a factor of $\sqrt{M}$ at the cost of additional computation. We once again leverage our DP scheme and in particular the graph induced by the DP (see Section 3 for a formal definition). Given this DP graph, one idea is to maintain probability measures over the edges of the DP graph to learn how to bid. Such an idea, which has been studied in other contexts (see Takimoto and Warmuth (2003), Zimin and Neu (2013), Chen et al. (2013)), leads to a sub-optimal $O(M|\mathcal{B}|\sqrt{T\log|\mathcal{B}|})$ discretized regret and $O(MT^{\frac{3}{4}}\sqrt{\log T})$ continuous regret. We obtain an improved regret bound by making an important observation that the utility of a bid vector depends only on the *nodes* of the DP graph and not the *edges*. Leveraging this fact enables us to maintain probability measures over the nodes in the DP graph, rather than its edges, resulting in an algorithm that achieves time and space complexities polynomial in $M, |\mathcal{B}|, T$, with discretized regrets $O(M\sqrt{T\log|\mathcal{B}|})$ and $O(M\sqrt{|\mathcal{B}|T\log|\mathcal{B}|})$ in the full information and bandit settings, respectively. See Theorem 6. This algorithm achieves a factor of $\sqrt{M}$ better than the decoupled exponential weights algorithm. Optimizing for the discretization error from $\mathcal{B}$, we show that in the bandit feedback setting, we derive an algorithm that achieves $O(MT^{\frac{2}{3}})$ regret.

THEOREM 2 **(Informal)**. *Under bandit feedback, there exists an algorithm that achieves discretized and continuous regrets $O(M\sqrt{T\log|\mathcal{B}|})$ and $O(MT^{\frac{2}{3}}\sqrt{\log T})$ respectively in the repeated multi-unit PAB auction.*

Regret Lower Bound (Section 6). To complement our discretized regret upper bound, we construct a regret lower bound for the full information setting which matches our upper bound up to a factor of $\sqrt{\log|\mathcal{B}|}$. We do this by constructing two distributions over adversary bid vectors for which any learning agent is guaranteed to incur regret linear in M when trying to learn the optimal bid under these distributions.

1.2. Experimental Results and Managerial Implications

Our experiments yield valuable practical insights for both auction designers and participants. These insights are primarily derived from simulations of PAB market dynamics using the no-regret learning algorithms outlined in our paper. Additionally, we compare these results with the market dynamics of uniform price auctions using the algorithms described in Brânzei et al. (2023). It is important to emphasize that conducting such systematic comparisons was previously challenging due to the inherent difficulty of characterizing equilibria in these auctions prior to our research.

1. **Uniform Bidding in PAB Auctions is Optimal.** As shown in Figures 2 and 8, the market dynamics consistently yield convergence of the winning bids and largest losing bids to a common price across all bidders. This partially addresses one of the main concerns over the fairness of the PAB auction. That is, while the payment for each unit *can* be different across units and across bidders, under a reasonable learning and bidding strategy, these payments across units and bidders converge to the same value in the long run.
2. **Simplified Bidding Interface is Sufficient for PAB but not for Uniform Price.** In a recent trend of bidding simplification and automation (e.g., Aggarwal et al. (2019), Deng et al. (2023), Susan et al. (2023), Lucier et al. (2023)), auctioneers may find it easier to restrict bidders' demand expressiveness by requiring only a single price and quantity, rather than a vector of bids. As per our previous insight, the bid value convergence of the market dynamics suggest appropriateness of this simplified bidding interface. In contrast, we show that the market dynamics of the uniform price auction converge to a staggered bid vector (Figure 8), suggesting that the simplified bidding interface may significantly damage the uniform price auction's welfare, revenue, or bidders' utility.
3. **PAB Obtains High Revenue but Slightly Lower Welfare than Uniform Price.** From the insights provided by Figure 7, it is evident that the PAB auction surpasses the uniform price auction in terms of revenue generation. However, it slightly lags behind in welfare, though to a lesser extent. Consequently, auctioneers who prioritize revenue (resp. welfare) should favor the PAB (resp. uniform price) auction over uniform price (resp. PAB) auctions.

1.3. Other Related Works

Learning in Auctions. Most of the recent learning-theory-flavored auction design research has either focused on the single unit setting (Han et al. 2020b,a, Balseiro et al. 2022), the perspective of the auctioneer setting reserve prices (Morgenstern and Roughgarden 2016, Mohri and Medina 2016, Cai and Daskalakis 2017, Dudík et al. 2017, Kanoria and Nazerzadeh 2014, Golrezaei et al. 2023a, 2021a, 2023b), or uniform price auctions (Mohri and Medina 2013, 2016, Huang et al. 2018, Nedelec et al. 2019, Haoyu and Wei 2020). However, in PAB auctions, as the space of possible bid vectors is exponentially large, the task of learning how to bid optimally in these multi-unit auctions is more challenging, compared with the single unit setting. This is especially true when the number of units demanded is large which necessitates not only low-regret but also tractable algorithms to learn how to bid optimally. In this paper, we contribute to this line of works by proposing a novel framework under which to analyze and derive efficient, low-regret learning algorithms for these inherently combinatorial multi-unit PAB auctions.

PAB Mechanism. There are several multi-unit auction formats that are commonly used in practice; e.g., uniform price (Binmore and Swierzbinski 2000, Burkett and Woodward 2020, Beyhaghi et al. 2021, Goldner et al. 2019), PAB (Pycia and Woodward 2020, Pesendorfer 2000, Baisa and Burkett 2018), Vickrey-Clarke-Groves (VCG) (Dobzinski and Nisan 2007, Ausubel and Milgrom 2006), ascending price (Cramton 1998, Ausubel 2004). The literature is divided as to which auction is appropriate for various settings. For example, while the PAB mechanism has desirable revenue and welfare guarantees compared to the uniform price auction (Pycia and Woodward 2020), the empirical revenue of the two auctions is often comparable (Hortaçsu and McAdams 2010), and some argue that guess-the-clearing-price and other strategic behavior (Heim and Götz 2021) along with collusion (Tierney et al. 2008) can further damage its performance. Furthermore, there are ethical and fairness concerns regarding PAB auctions, as their discriminatory nature implies that agents pay unequally for the same unit. Despite this criticism, and other arguments for (and against) other auction formats (Porter et al. 2003, Cramton et al. 2004, de Keijzer et al. 2013a, Ausubel et al. 2014, Akbarpour and Li 2020), we focus on the PAB mechanism due to the simplicity and transparency of its payment rule, as well as, its widespread use.

The economics literature has only recently addressed several of the questions regarding the equilibria, bidding dynamics, efficiency, and other key properties of multi-unit PAB mechanisms in the static or Bayesian setting. For example, the Bayesian optimal bidding strategy is known for the case of 2-unit demand and supply multi-unit auctions (Nautz 1995), for when valuations follow a class of parametric distributions (Pycia and Woodward 2020), or when the bidders have symmetric valuations (Ausubel et al. 2014). The PAB mechanism is also known to be smooth (Syrgkanis and Tardos 2012), which yields a number of desirable guarantees on the price of anarchy of the auction (Roughgarden 2015), even with the presence of an aftermarket (Babaioff et al. 2022). An additional attractive property of the PAB mechanism is complete transparency of payments, as given one's allocation, an agent knows precisely how much they will pay. This is in stark contrast to the uniform price auction, where shill bids can inflate payments by artificially increasing demand (Akbarpour and Li 2020).

Relationship to Uniform Price Auctions. The uniform price auction is an alternative mechanism of allocating multiple homogeneous goods. Closely related to the PAB auction, bidders are allocated units in decreasing order of bids, but instead of charging each bidder their corresponding winning bid, each bidder instead pays the smallest winning bid. The EU-ETS carbon license auctions use the uniform price auction, rather than the PAB auction, largely due to fairness considerations, as each agent pays the same amount per unit allocated (EEX 2023). However, our work shows that price fairness is not a concern in the long term when bidders learn how to bid and converge to a common price.

In a study closely aligned with our research, Brânzei et al. (2023) investigated the problem of learning optimal bidding strategies in multi-unit uniform pricing auctions. Since the uniform price auction employs a distinct payment rule, the bid optimization problem necessitates different approaches compared to the PAB auction. While Brânzei et al. (2023) reformulated the offline bid optimization problem as a path weight maximization problem over a directed acyclic graph (DAG), the construction of these DAGs differs fundamentally from those we constructed in our manuscript for PAB auctions. We also emphasize that the decoupling of utility across units is unique to the PAB setting, which enables improved regret guarantees. Consequently, the path kernel-based algorithms described in Brânzei et al. (2023) for both the full and bandit settings yield sub-optimal regret in terms of dependence on the number of units in the PAB mechanism. As previously mentioned, we introduce an alternative method based on OMD, which achieves regret matching the lower bound.

Structured Bandits. The crux of our paper is constructing time and space efficient no-regret bandit algorithms for bid optimization under a combinatorially large bid space for identical multi-unit PAB auctions. As such, naive implementations of bandit algorithms, such as Exp3, that consider each bid vector as its own arm in isolation incur exponential regret, computational and memory costs. It is similarly difficult to generalize existing efficient cross-learning based algorithms in the single-unit case (Han et al. 2020a,b, Badanidiyuru et al. 2021). To combat this, we take from the expansive structured bandit literature, which includes linear bandits (Dani et al. 2008, Abbasi-yadkori et al. 2011, Chu et al. 2011, Lattimore and Szepesvári 2020) and combinatorial bandits (Chen et al. 2013, Audibert et al. 2011, Niazadeh et al. 2022), and convex uncertainty set bandit (Van Parys and Golrezaei 2023). In particular, our algorithm most closely resembles existing algorithms exploiting both these combinatorial and linear aspects in episodic Markov Decision Processes (Zimin and Neu 2013) or cost minimization on graphs (Takimoto and Warmuth 2003). The primary difference is that our algorithm seeks to minimize the sum of the costs of nodes in a path, rather than the edges. We describe how to efficiently perform negentropy regularized OMD updates in our setting.

Multi-Agent Learning. While we seek to derive efficient, low-regret algorithms for a single bidder, as we also explored, it is equally important to understand the implications of the bidding dynamics induced by such adaptive behavior. For example, it may be possible for these adaptive agents to learn to collude (Hendricks and Porter 1989, Pesendorfer 2000, Aoyagi 2003, Calvano et al. 2021, Heim and Götz 2021) which can significantly reduce revenue and welfare. Precisely how much they do so is characterized in the Bayesian setting as the Price of Anarchy (PoA) (Lucier and Borodin 2009, de Keijzer et al. 2013a, Hartline et al. 2014, Roughgarden et al. 2017). The auction efficiency in the dynamic setting is less well understood, though there have been several results for specific auctions and learning algorithms (Blum et al. 2008, Lykouris et al. 2016, Hartline et al. 2015, Golrezaei et al. 2020, Kolumbus and Nisan 2022). However, most of the standard learning theory literature takes the perspective of the auctioneer who optimizes revenue through reserves or supply. More standard in literature is assuming that the auctioneer is an adaptive agent themselves, optimizing over the set of reserves or supply (Mohri and Medina 2013, Roughgarden and Wang 2016, Mohri and Medina 2015). This line of work is closely related to the vast and rapidly growing literature on multi-agent learning (Phelps et al. 2008, Buşoniu et al. 2010, Yang and Wang 2020, Golrezaei et al. 2020, 2023b, Zhang et al. 2021).

The multi-agent learning dynamics in the PAB auction (and also uniform price auction) is studied in our experiments section. To our knowledge, these experiments are the first systematic comparison between the equilibria of PAB and uniform price auctions, showing a noticeable revenue improvement of the PAB over the uniform price auction.

2. Preliminaries

Notation. We let $|\mathcal{S}|$ denote the size of set $\mathcal{S}$, and define $[k] = \{1, \ldots, k\}$ to be the set of the first k positive integers. We define $\mathcal{S}^{+k}$ to be the set of non-increasing k-vectors of elements from set $\mathcal{S}$. Similarly, $\mathcal{S}^{-k}$ denotes the set of non-decreasing k-vectors of elements from set $\mathcal{S}$. We let $\Delta(\mathcal{S})$ denote the set of valid probability measures over set $\mathcal{S}$. We also say that the quantity x is $\lesssim$, $\propto$, or $\gtrsim$ than $f(a_1, \ldots, a_k)$ some function of k algorithm parameters if $x \in \mathcal{O}(f(a_1, \ldots, a_k))$, $x \in \Theta(f(a_1, \ldots, a_k))$, or $x \in \Omega(f(a_1, \ldots, a_k))$, respectively.

Auction format: Pay-as-bid. Consider a scenario in which there are N bidders and $\overline{M}$ identical units available for auction in a PAB format. In this context, we assume that each agent n within the set $[N]$ desires a maximum of M units. (It should be noted that our results can be extended to a situation where each agent n may demand a different maximum number of units, denoted by M_n, which are not necessarily identical.)

Let $\boldsymbol{v}_n = (v_{n,m})_{m \in [M]} \in [0,1]^{+M}$ represent agent n's non-increasing marginal valuation profile. This implies that for any given n in the set $[N]$, the following conditions hold: (i) valuation monotonicity $v_{n,1} \geq v_{n,2} \geq \ldots \geq v_{n,M}$ and (ii) the total valuation of agent n after receiving m units is given by $\sum_{k=1}^{m} v_{n,k}$.

Let $\boldsymbol{b}_n = \{b_{n,m}\}_{m \in [M]} \in [0,1]^{+M}$ represent the non-increasing bids submitted by bidder n. Here, $b_{n,m}$ refers to the bid made by bidder n for the m-th slot or, equivalently, the m-th unit. Similar to $\boldsymbol{v}_n$, we have the following conditions for any $n \in [N]$: (i) bid monotonicity $b_{n,1} \geq b_{n,2} \geq \ldots \geq b_{n,M}$ and (ii) individual rationality (IR) $b_{n,m} \leq v_{n,m}$ for all $m \in [M]$. It is important to note that the bid monotonicity condition is not an assumptions; it is implied by the auction rule that will be stated shortly. Consequently, the total payment made by bidder n after receiving m units is given by $\sum_{k=1}^{m} b_{n,k}$. We define $\boldsymbol{b}_{-n} = (b_{-n,m})_{m \in [\overline{M}]} \in [0,1]^{-\overline{M}}$ as the set of the $\overline{M}$ largest bids not belonging to agent n, arranged in increasing order (i.e., $b_{-n,1} \leq \ldots \leq b_{-n,\overline{M}}$).

The auction operates according to the following rules: In a PAB auction, all bids submitted across the N bidders are arranged in descending order. The m-th unit is assigned to the bidder with the m-th highest bid, and they are charged the amount of their bid. We denote the allocation to agent n as $x_n(\boldsymbol{b}_n) = x(\boldsymbol{b}_n, \boldsymbol{b}_{-n})$, and the (quasi-linear) utility as $\mu_n(\boldsymbol{b}_n) = \mu(\boldsymbol{b}_n, \boldsymbol{b}_{-n})$, where

$$x(\boldsymbol{b}_n, \boldsymbol{b}_{-n}) = \sum_{m=1}^{M} \mathbf{1}_{b_{n,m} \geq b_{-n,m}} \quad \text{and} \quad \mu(\boldsymbol{b}_n, \boldsymbol{b}_{-n}) = \sum_{m=1}^{x_n(\boldsymbol{b}_n)} (v_{n,m} - b_{n,m})\,. \tag{1}$$

respectively. Here, $b_{-n,m}$ represents the m-th smallest bid among the $\overline{M}$ largest bids of all other bidders except bidder n. It should be noted that $x(\boldsymbol{b}_n, \boldsymbol{b}_{-n})$ denotes the number of units that agent n receives in the auction. In the case of tied bids, we assume the use of an arbitrary, publicly known deterministic tie-breaking rule, denoted as $\textsc{TieBreak}_n : \mathcal{R}^{+M} \times \mathcal{R}^{-\overline{M}} \to [M]$, to determine the allocation for agent n. This tie-breaking rule is incorporated into the allocation as $x(\boldsymbol{b}_n, \boldsymbol{b}_{-n}) = \sum_{m=1}^{M} \mathbf{1}_{b_{n,m} > b_{-n,m}} + \textsc{TieBreak}_n(\boldsymbol{b}_n, \boldsymbol{b}_{-n})$, and we use this shorthand notation in Equation (1).

Online/Repeated Setting. Consider a repeated setting where the PAB auction is conducted over T rounds. In this repeated setting, we will focus on the perspective of agent $n \in [N]$ and remove additional indexing when it is evident from the context. For each auction round, agent n has a fixed valuation profile represented by $\boldsymbol{v} = (v)_{m\in[M]} \in [0,1]^{+M}$, and their bid vector in the t-th auction is denoted by $\boldsymbol{b}^t = (b^t_m)_{m\in[M]} \in [0,1]^{+M}$.[1] Similarly, $\boldsymbol{b}^t_- = (b^t_{-m})_{m\in[\overline{M}]} \in [0,1]^{-\overline{M}}$ represents the competing bids in round t. In each round, agent n receives $x^t_n(\boldsymbol{b}^t)$ units and earns a utility of $\mu^t_n(\boldsymbol{b}^t)$, where:

$$x^t_n(\boldsymbol{b}^t) = x(\boldsymbol{b}^t, \boldsymbol{b}^t_-) \quad \text{and} \quad \mu^t_n(\boldsymbol{b}^t) = \mu(\boldsymbol{b}^t, \boldsymbol{b}^t_-)\,. \tag{2}$$

Recall that functions x and μ are defined in Equation (1). The goal of agent n is to choose a sequence of bid vectors $(\boldsymbol{b}^t)_{t\in[T]}$ that maximizes their total utility, given by $\sum_{t=1}^{T} \mu^t_n(\boldsymbol{b}^t)$. However, the main challenge is that the vectors $\boldsymbol{b}^t_-$, representing the competing bids, are not known in advance. Instead, they are revealed in an online manner. Consequently, agents must learn how to bid optimally throughout the sequence of auctions, taking into account their previous allocations $H^{t-1} = (x^\tau(\boldsymbol{b}^\tau))_{\tau\in[t-1]}$ and their valuation profile $\boldsymbol{v}$. The performance of an agent's learning strategy is evaluated in terms of regret, which quantifies the difference between their expected utility using their learning strategy and the optimal utility achievable with perfect knowledge of the competing bidding vectors in hindsight:

$$\textsc{Regret} = \max_{\boldsymbol{b}\in[0,1]^{+M}} \sum_{t=1}^{T} \mu^t_n(\boldsymbol{b}) - \mathbb{E}\left[\sum_{t=1}^{T} \mu^t_n(\boldsymbol{b}^t)\right]. \tag{Continuous Regret}$$

Here, $\boldsymbol{b}^t_-$ can be selected by an adaptive adversary; i.e. an adversary who can select $\boldsymbol{b}^t_-$ as a function of the entire auction history, which includes $\boldsymbol{b}^1, \ldots, \boldsymbol{b}^{t-1}$, but does not have access to the possible

[1] In Section 9.9, we extend our results to the case of time-varying valuations.

randomness when selecting $\boldsymbol{b}^t$. One example is when the other competitors are also behaving according to no-regret learning algorithms. We note that it is known that the time averaged iterates in the game dynamics induced by agents running no-regret learning algorithms converges to a coarse correlated equilibrium (CCE), which have desirable revenue and welfare guarantees via smooth-auction PoA analysis (de Keijzer et al. 2013b, Feldman et al. 2017). As few theoretical results are known regarding the structure of CCE's in PAB auctions, we hope our proposed algorithms and experiments will provide useful insights in this direction. We will discuss this further in Section 7.

The benchmark, $\max_{\boldsymbol{b}\in[0,1]^M} \sum_{t=1}^{T} \mu_n^t(\boldsymbol{b})$ used in the definition of continuous regret, is constructed considering all possible $\boldsymbol{b}_-^t$ for every round $t \in [T]$, where in the hindsight optimal solution, the bid vector can be chosen from any vector $\boldsymbol{b} \in [0,1]^{+M}$. However, in practice, bid vectors are often restricted to a discretization of $[0,1]$ denoted by $\mathcal{B} = (B_1, \ldots, B_{|\mathcal{B}|})$, where $0 = B_1 < \ldots < B_{|\mathcal{B}|} = 1$. For such cases, we define an analogous version of regret:

$$\textsc{Regret}_{\mathcal{B}} = \max_{\boldsymbol{b}\in\mathcal{B}^{+M}} \sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}) - \mathbb{E}\left[\sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}^t)\right]. \tag{Discretized Regret}$$

In both the definitions of continuous and discretized regret, we henceforth implicitly assume that $b_m \le v_m$ for any $m \in [M]$; that is, we have the bid vector $\boldsymbol{b}$ is subject to individual rationality and overbidding is not allowed. Observe that the benchmark in defining $\textsc{Regret}_{\mathcal{B}}$ is weaker than that in $\textsc{Regret}$. Nevertheless, they are not too far from each other as we discuss in the following sections. We wish to derive a learning algorithm that achieves an upper bound on $\textsc{Regret}$ that is polynomial in M and sub-linear in T. To do so, we consider the discretized setting, and bound $\textsc{Regret}_{\mathcal{B}}$; an upper bound on $\textsc{Regret}$ will be obtained by accounting for the discretization errors.

We consider two feedback structures: (i) full information and (ii) bandit. In the full information setting, the agent's allocation and the values of $\boldsymbol{b}_-^t$ are revealed after the end of each round, whereas in the bandit setting, only the agent's allocation is revealed.

3. Hindsight Optimal Offline Solution

In the offline setting, our goal is to determine agent n's optimal fixed bidding strategy for the T rounds of PAB auctions. Recall the following optimization problem with bid space $\mathcal{B}$:

$$\max_{\boldsymbol{b}\in\mathcal{B}^{+M}} \sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}), \tag{Offline}$$

where $\mu_n^t(\boldsymbol{b})$ (defined in Equation (2)) represents the utility of agent n in round t given bid vector $\boldsymbol{b}$ and competing bids $\boldsymbol{b}_-^t$. As mentioned earlier, the solution to this optimization problem serves as a benchmark for evaluating the performance of online learning algorithms in the repeated setting. Furthermore, it provides valuable insights for designing algorithms with polynomial time and space complexity for the repeated setting.

ALGORITHM 1: OFFLINE($\boldsymbol{v}, \{\boldsymbol{b}_{-}^t\}_{t\in[T]}$)

Input: Valuation $\boldsymbol{v}$ for $\boldsymbol{v} \in [0,1]^{+M}$, Other bids $\{\boldsymbol{b}_{-}^t\}_{t\in[T]}$ for $\boldsymbol{b}_{-}^t \in \mathcal{B}^{-\overline{M}}$.

Output: Optimal bid vector $\boldsymbol{b}^* = \text{argmax}_{\boldsymbol{b}\in\mathcal{B}^M} \mu_n^T(\boldsymbol{b})$ and its corresponding utility.

Let $W_m^{T+1}(b) \leftarrow \sum_{t=1}^T \mathbf{1}_{b\geq b_{-m}^t}(v_m - b)$, $b \in \mathcal{B}, m \in [M]$, define $U_{M+1}(b) \leftarrow 0$, $b \in \mathcal{B}$, and set $b_0^* = \max(\mathcal{B})$;

for $m \in [M,\ldots,1], b \in \mathcal{B} : U_m(b) \leftarrow \max_{b'\in\mathcal{B};b'\leq b} W_m^{T+1}(b') + U_{m+1}(b')$;

for $m \in [1,\ldots,M] : b_m^* \leftarrow \arg\max_{b\leq b_{m-1}^*} U_m(b)$;

Return $U_1(\max(\mathcal{B}))$ and $\boldsymbol{b}^* = (b_m^*)_{m\in[M]}$.

To solve Problem (Offline), we take advantage of the following decomposition:

$$\sum_{t=1}^T \mu_n^t(\boldsymbol{b}) = \sum_{t=1}^T \sum_{m=1}^M (v_m - b_m)\mathbf{1}_{b_m\geq b_{-m}^t} = \sum_{m=1}^M \sum_{t=1}^T (v_m - b_m)\mathbf{1}_{b_m\geq b_{-m}^t} \tag{3}$$

$$:= \sum_{m=1}^M \sum_{t=1}^T w_m^t(b_m) := \sum_{m=1}^M W_m^{T+1}(b_m)\,, \tag{4}$$

where $w_m^t(b) = \mathbf{1}_{b\geq b_{-m}^t}(v_m - b)$ represents the utility in the t-th auction for winning the m-th item with bid b, and $W_m^{T+1}(b) = \sum_{t=1}^T w_m^t(b)$ represents the cumulative utility gained from winning the m-th item with bid b across the T auctions. (Here, in $w_m^t(b)$, the same tie-breaking rule in Equation (1) is applied). To solve Problem (Offline), we develop a polynomial-time DP scheme utilizing these $w_m^t(b)$ and $W_m^{T+1}(b)$. In particular, for any $m \in [M]$ and any bid $b \in \mathcal{B}$, let $U_m(b)$ be the optimal cumulative utility of the agents from units $m, m+1, \ldots, M$ over T auctions assuming that bids for unit m is less than or equal to b.

We then have

$$U_m(b) = \max_{b'\leq b, b'\in\mathcal{B}} \left\{W_m^{T+1}(b') + U_{m+1}(b')\right\} \quad b \in \mathcal{B}, m \in [M] \quad \text{and} \quad U_{M+1}(b) = 0 \quad b \in \mathcal{B}\,. \tag{5}$$

Algorithm 1 uses the aforementioned DP scheme to devise an optimal solution to Problem (Offline). The following theorem, proven in Section 9.1, shows the optimality of Algorithm 1.

THEOREM 3. *Algorithm 1 returns the optimal solution to problem* (Offline) *with time and space complexity of $O(M|\mathcal{B}|^2)$ and $O(M|\mathcal{B}|)$ respectively.*

This algorithm to solve the offline bid optimization problem enables us to compute the hindsight optimal utility, which serves as a benchmark for evaluating the effectiveness of our online learning algorithms. It is worth mentioning that we can represent our DP algorithm as an equivalent graph with M layers, with $|\mathcal{B}|$ nodes in each. More precisely, we define the (offline) DP graph as follows:

1. **DP nodes/states.** There are M layers, each with $|\mathcal{B}|$ nodes in each, denoted by $\{(m,b)\}_{m\in[M],b\in\mathcal{B}}$.

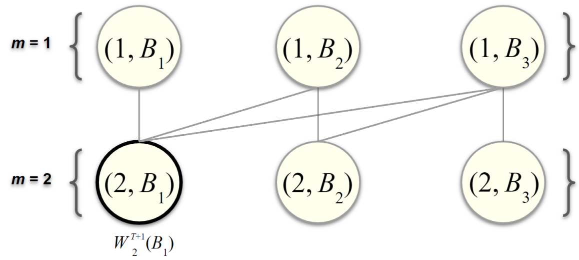

Figure 1 **DP graph for problem (Offline): Bid optimization problem cast as a graph problem, for $m = 2$ and $|\mathcal{B}| = 3$ with $B_1 < B_2 < B_3$. Node (m, b)—the node in the m'th layer with bid value b—has weight $W_m^{T+1}(b)$.**

2. **DP edges.** In this graph, there are only (directed) edges between two consecutive layers, i.e., from layer m to layer $m+1$ for any $m \in [M-1]$. In particular, node (m, b) only has an edge to node $(m+1, b')$ for $b' \leq b$ and if $b' \leq v_{m+1}, b \leq v_m$.

3. **DP weights.** We define the weight of node/state (m, b) to be $W_m^{T+1}(b) = \sum_{\tau=1}^{T} \mathbf{1}_{b \geq b_{-m}^{\tau}}(v_m - b)$.

For the online setting, we also note that we can define the DP graph at time t, as opposed to T, by setting $W_m^t(b) = \sum_{\tau=1}^{t-1} \mathbf{1}_{b \geq b_{-m}^{\tau}}(v_m - b)$. This allows us to construct algorithms for the full information and bandit settings by taking advantage of the structure of the DP graph to enhance efficiency and optimize storage of necessary computations.

4. Decoupled Exponential Weights Algorithms

In this section, we present our first algorithm for learning in the online setting. In particular, we construct a decoupled version of the Exponential Weights algorithm which circumvent the large space and time complexity of maintaining and updating the sampling distributions of all possible bid vectors. Our algorithms instead sequentially sample a singular bid value from each layer of our DP graph such that the probability of sampling a particular vector of bids $\boldsymbol{b}^t$ is precisely equal to the probability of the exponential weights algorithm selecting $\boldsymbol{b}^t$.[2]

In the following sections, we provide a description of our algorithm in both the full information and the bandit settings. It is important to note that our decoupled exponential weights algorithm achieves regret that is sub-optimal by a factor of $O(\sqrt{M})$. Nonetheless, we present an alternative regret optimal algorithm based on OMD in a subsequent section. Despite this, our decoupled exponential weights algorithm remains practical as it does not necessitate solving a convex optimization problem at each time step $t \in [T]$.

[2] In Section 9.9, we present a generalization of our decoupled exponential weights algorithm to the setting with time varying valuations, where the valuations are drawn from some known, finite support distribution $F_{\boldsymbol{v}}$.

4.1. Full Information Setting

Now let us focus on learning optimal bidding in an online fashion with full information feedback. One straightforward approach in this context is to apply the exponential weights algorithm (Littlestone and Warmuth 1994) to the entire set of bid vectors. This algorithm guarantees per-round rewards within the range of $[-M, M]$. However, the challenge lies in the exponentially large bid space $\mathcal{B}^{+M}$. Tracking and updating weights for all possible bid vectors naively would lead to a non-polynomial time and space complexity. Although this approach achieves a small regret of $O(M^{\frac{3}{2}}\sqrt{T\log|\mathcal{B}|})$, we need a more efficient solution with a polynomial time and space complexity.

To do so, we leverage the DP scheme developed in Section 3. By utilizing the DP graph and the information it provides about bid vector utilities, we can effectively mimic the exponential weights algorithm without explicitly tracking weights for every bid vector. In Algorithm 2, instead of associating weights to each possible bid vector, we associate weights with each (m, b) pair for any $m \in [M]$ and $b \in \mathcal{B}$. These weights are then updated via variables $S_m^t(b)$ for $b \in \mathcal{B}, m \in [M]$, which are inspired by the DP scheme. For any round t, we define

$$\begin{aligned} S_m^t(b) &= \exp(\eta W_m^t(b)) \sum_{\boldsymbol{b}'_{m+1:M} \in \mathcal{B}^{+(M-m)}, b'_{m+1} \leq b'_m = b} \exp(\eta \sum_{m'=m+1}^{M} W_{m'}^t(b'_{m'})) \\ &= \exp(\eta W_m^t(b)) \sum_{b' \in \mathcal{B}; b' \leq b} S_{m+1}(b') \,. \end{aligned}$$

Here, $\eta > 0$ is the learning rate of the algorithm and we recall that $W_m^t(b) = \sum_{\tau=1}^{t-1} w_m^\tau(b)$ is cumulative utility gained across the first $t-1$ auctions from the winning the m'th item with bid b, respectively. Computing $S_m^t(\cdot)$ is done in step $\textsc{Compute}-S_m$ of the algorithm. In step $\textsc{Sample}-\boldsymbol{b}$, the bid vector is then sampled according to S_m^t's subject to bid monotonicity. To disallow overbidding, we initialize weights $W_m^0(b) = -\infty$ for all $m \in [M], b \in \mathcal{B}$ such that $b > v_m$.

The concept of utility decoupling shares similarities with solutions used in combinatorial bandits, tabular reinforcement learning (Chen et al. 2013, Zimin and Neu 2013), and problems such as shortest path algorithms involving weight pushing or path kernels (Takimoto and Warmuth 2003, Koolen et al. 2010). These methods are employed to solve variants of the shortest path or maximum weight path problems, where costs or weights are associated with edges rather than nodes. By exploiting the graph structure and the linearity of utilities with respect to the weights of each edge, these algorithms efficiently compute path weights based on edge weights, similar to how our algorithm computes path weights based on node weights. In addition to investigating a fundamentally different problem, the key distinction is that our approach considers weights associated with nodes instead of edges. In our setting, the reward associated with selecting bid b' in slot $m+1$ is independent of selecting bid $b \geq b'$ in slot m. This allows us to get an improved regret bound and

ALGORITHM 2: Decoupled Exponential Weights - Full Information

Input: Learning rate $0 < \eta < \frac{1}{M}$, $\boldsymbol{v} \in [0,1]^{+M}$.

Output: The aggregate utility $\sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}^t)$

$W_m^0(b) \leftarrow 0$ for all $m \in [M], b \in \mathcal{B}$ such that $b \leq v_m$; else $W_m^0(b) \leftarrow -\infty$;

$b_0^t \leftarrow \max \mathcal{B}$, and $S_{M+1}^t(\min \mathcal{B}) = 1$ for any $t \in [T]$;

for $t \in [1, \ldots, T]$: **do**

  Recursively Computing Exponentially Weighted Partial Utilities $\boldsymbol{S}^t$;

  for $m \in [M, \ldots, 1], b \in \mathcal{B} : S_m^t(b) \leftarrow \exp(\eta W_m^t(b)) \sum_{b' \leq b} S_{m+1}^t(b')$. \\ Compute $- S_m$;

  Determining the Bid Vector $\boldsymbol{b}^t$ Recursively;

  for $m \in [1, \ldots, M], b \leq b_{m-1}^t : b_m^t \leftarrow b$ with probability $\frac{S_m^t(b)}{\sum_{b' \leq b_{m-1}^t} S_m^t(b')}$; \\ Sample $- \boldsymbol{b}$;

  Observe $\boldsymbol{b}_-^t$ and receive reward $\mu_n^t(\boldsymbol{b}^t)$;

  Update Weight Estimates ;

  for $m \in [M], b \in \mathcal{B} : W_m^{t+1}(b) \leftarrow W_m^t(b) + (v_m - b)\mathbf{1}_{b \geq b_{-m}^t}$ if $b \leq v_m$; else $W_m^{t+1}(b) \leftarrow -\infty$;

end

Return $\sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}^t)$

save a factor of $|\mathcal{B}|$ in terms of time and space complexity. Instead of storing and updating weights for $O(M|\mathcal{B}|^2)$ possible (m, b, b') slot-value-next value triplets, we only need to handle $O(M|\mathcal{B}|)$ possible (m, b) unit-bid pairs.

The following statement is the main result of this section.

Theorem 4 **(Decoupled Exponential Weights: Full Information)**. *With* $\eta = \Theta(\sqrt{\frac{\log |\mathcal{B}|}{MT}})$, *Algorithm 2 achieves (discretized) regret* $O(M^{\frac{3}{2}}\sqrt{T \log |\mathcal{B}|})$, *with total time and space complexity polynomial in* M, $|\mathcal{B}|$, *and* T. *Optimizing for discretization error from restricting the bid space to* $\mathcal{B}$, *we obtain a continuous regret of* $O(M^{\frac{3}{2}}\sqrt{T \log T})$.

It is worth noting that both the time and space complexity exhibit polynomial scaling with M and $|\mathcal{B}|$. Given that M can be large in practical scenarios, such as carbon emissions license auctions or electricity markets, it becomes crucial to minimize the dependence on M.

4.2. Bandit Feedback Setting

We extend Algorithm 2 for the bandit feedback setting. In the bandit feedback setting, the bidder's allocation and utility are not available for all possible bid vectors, unlike in the full information setting. Instead, the agent only observes their utility for the submitted bid vector. To handle this, we use inverse probability weighted (IPW) node weight estimates $\widehat{w}_m^t(b)$ instead of the node weights $w_m^t(b)$ in Algorithm 2. This adaptation results in a regret of $O(M^{\frac{3}{2}}\sqrt{|\mathcal{B}| T \log |\mathcal{B}|})$, as shown in Theorem 5. This regret includes an additional factor of $\sqrt{|\mathcal{B}|}$ compared to the full information setting.

ALGORITHM 3: Decoupled Exponential Weights - Bandit Feedback

Input: Learning rate $0 < \eta < \frac{1}{M}$, $\boldsymbol{v} \in [0,1]^{+M}$

Output: The aggregate utility $\sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}^t)$

$\widehat{W}_m^0(b) \leftarrow 0$ for all $m \in [M], b \in \mathcal{B}$ such that $b \leq v_m$; else $\widehat{W}_m^0(b) \leftarrow -\infty$.;

$b_0^t \leftarrow \max \mathcal{B}$, and $\widehat{S}_{M+1}^t(\min \mathcal{B}) = 1$ for any $t \in [T]$;

for $t \in [1, \ldots, T]$: **do**

  Recursively Computing Exponentially Weighted Partial Utilities $\boldsymbol{S}^t$;

  for $m \in [M, \ldots, 1], b \in \mathcal{B}: \widehat{S}_m^t(b) \leftarrow \exp(\eta \widehat{W}_m^t(b)) \sum_{b' \leq b} \widehat{S}_{m+1}^t(b')$ \\ Compute $- \widehat{S}_m$;

  Determining the Bid Vector $\boldsymbol{b}^t$ Recursively;

  for $m \in [1, \ldots, M], b \leq b_{m-1}^t : b_m^t \leftarrow b$ with probability $\frac{\widehat{S}_m^t(b)}{\sum_{b' \leq b_{m-1}^t} \widehat{S}_m^t(b')}$; \\ Sample $- \boldsymbol{b}$;

  Observe $\boldsymbol{b}_{-}^t$ and receive reward $\mu_n^t(\boldsymbol{b}^t)$;

  Recursively Computing Probability Measure $\boldsymbol{q}$;

  $q_1^t(b) \leftarrow \frac{\widehat{S}_m^t(b)}{\sum_{b' \in \mathcal{B}} \widehat{S}_m^t(b')}$ for all $b \in \mathcal{B}$;

  for $m \in [2, \ldots, M], b \in \mathcal{B}: q_m^t(b) \leftarrow \sum_{b' \geq b} \frac{q_{m-1}^t(b') \widehat{S}_m^t(b)}{\sum_{b" \geq b'} \widehat{S}_m^t(b")}$ for all $b \in \mathcal{B}$;

  Update Weight Estimates;

  for $m \in [M], b \in \mathcal{B}: \widehat{W}_m^{t+1}(b) \leftarrow \widehat{W}_m^t(b) + (1 - \frac{1-(v_m - b)\mathbf{1}_{b \geq b_m^t}}{q_m^t(b)} \mathbf{1}_{b_m^t = b})$ if $b \leq v_m$; else $\widehat{W}_m^{t+1}(b) \leftarrow -\infty$;

end

Return $\sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}^t)$

The structure of Algorithm 3 is similar to that of Algorithm 2. Both algorithms maintain node weight estimates, compute the sum of exponentiated partial bid vector estimated utilities recursively, and sample bids for each unit recursively proportional to these summed exponentiated utilities. Specifically, Algorithm 3 samples bid vectors with probabilities proportional to the sum of the cumulative *estimated* utility $\widehat{W}_m^t(b) = \sum_{\tau=1}^{t-1} \widehat{w}_m^\tau(b)$ over each unit-bid value pair, where $\widehat{w}_m^t(b) = 1 - \frac{1-(v_m - b)\mathbf{1}_{b \geq b_{-m}^t}}{q_m^t(b)} \mathbf{1}_{b_m^t = b}$ and $q_m^t(b)$ is the (unconditional) probability that bid b is chosen for unit m at time t.

Note that this bid vector utility estimator is a slightly different estimator than the one used in the standard Exp3 algorithm (See Chapter 11 of Lattimore and Szepesvári (2020)). In particular, the standard IPW estimator $\widehat{w}_m^t(b) = \frac{v_m - b}{q_m^t(b)} \mathbf{1}_{b = b_m^t > b_{-m}^t}$, while unbiased, can be unboundedly large when $q_m^t(b)$ approaches 0, whereas our proposed estimator is bounded above by 1. As a consequence, we have that $\widehat{\mu}_n^t(\boldsymbol{b})$ is upper bounded by M, and therefore $= \eta \widehat{\mu}^t(\boldsymbol{b}) = \eta \sum_{m=1}^{M} \widehat{w}_m^t(b)$ is upper bounded by 1 for $\eta < \frac{1}{M}$, which we crucially use in the proof.

The primary difference in the implementation of Algorithm 3 as compared to Algorithm 2 is that we require additional steps in order to obtain unbiased node weight estimates $\widehat{W}_m^{t+1}(b) = \sum_{\tau=1}^{t} \widehat{w}_m^\tau(b)$ which we compute using an IPW estimator. In order to do this, we must compute $q_m^t(b)$—the probabilities of selecting bid b at slot m.

Theorem 5 **(Decoupled Exponential Weights: Bandit Feedback)**. *With* $\eta = \Theta(\sqrt{\frac{\log |\mathcal{B}|}{M|\mathcal{B}|T}})$ *such that* $\eta < \frac{1}{M}$, *Algorithm 3 achieves (discretized) regret* $O(M^{\frac{3}{2}} \sqrt{|\mathcal{B}| T \log |\mathcal{B}|})$, *with* *total time and space complexity polynomial in M, $|\mathcal{B}|$, and T. Optimizing for discretization error from restricting the bid space to $\mathcal{B}$, we obtain a continuous regret of $O(M^{\frac{4}{3}}T^{\frac{2}{3}}\sqrt{\log T})$.*

5. Online Learning Algorithms: Mirror Descent

In this section, we propose our second online learning algorithm. Instead of mimicking the exponential weights algorithm, we reformulate the problem as online linear optimization over node probabilities in our DP graph. We solve this using OMD and construct a policy that sequentially samples bids based on these probabilities. We first present the regret analysis for the bandit setting, followed by the full information setting. We provide a single algorithm for both settings, with only one line changing based on the feedback structure. Our OMD algorithm is regret optimal in both feedback structures (up to a factor of $\sqrt{|\mathcal{B}|\log|\mathcal{B}|}$). However, it requires solving a convex optimization problem at each iteration, which may slow it down in practice. Nonetheless, this algorithm is preferred when prioritizing regret optimality over computational complexity.

5.1. Algorithm Statement

Recall that in the DP graph, we have M layers, where in each layer there are $|\mathcal{B}|$ nodes and $|\mathcal{B}|^2$ edges. Given the structure of the DP graph, one idea is to maintain some policy $\boldsymbol{\pi}:[M]\times\mathcal{B}\times\mathcal{B}\to[0,1]$ which induces a family of probability measures $\boldsymbol{\rho}:[M]\times\mathcal{B}\times\mathcal{B}\to[0,1]$ over the edges in the DP graph. In particular, let $\pi((m,b),b')=\mathbb{P}(b_{m+1}=b'\mid b_m=b)$ be the probability that the agent selects bid b' for slot $m+1$ conditional on having already selected bid b for slot m. Further, define $\rho((m,b),b')=\mathbb{P}(b_m=b,b_{m+1}=b')$ as the unconditional probability that agent selects bids b and b' for slots m and $m+1$, respectively. Following this idea, one can transform the bid optimization problem as an equivalent online linear optimization (OLO) problem over the space of possible $\boldsymbol{\rho}$, which we will show in the following section. However, this approach would lead to an algorithm with sub-optimal regret of $O(M|\mathcal{B}|\sqrt{T\log|\mathcal{B}|})$ as it fails to capture the additional structure within the DP graph; cf. (Chen et al. 2013, Takimoto and Warmuth 2003, Zimin and Neu 2013). We show later that it is possible to improve this regret by a factor of $\sqrt{|\mathcal{B}|}$.

In this section, as our main contribution, instead of maintaining probability measures $\boldsymbol{\rho}$ over the edges in the DP graph, we maintain probability measures $\boldsymbol{q}$ over nodes. This idea is based on an important observation that, in the DP formulation, the weight of a path depends only on the nodes traversed and not the edges. In other words, regardless of the value of b_m, selecting the edge from (m,b_m) to $(m+1,b_{m+1})$ at round t always yields the same utility $w^t_{m+1}(b_{m+1})$. We then construct some policy $\boldsymbol{\pi}$ that generates the desired node probability measures $\boldsymbol{q}$, where there may be many such choices of $\boldsymbol{\pi}$, though we argue the specific choice will not affect the regret. Consequently, we can reduce the higher dimensional problem of regret minimization over policies to the simpler one of regret minimization over node measures.

ALGORITHM 4: OMD - Bid Optimization in Multi-Unit Pay as Bid Auctions

Input: Learning rate $\eta > 0$, Valuation $\boldsymbol{v} \in [0,1]^{+M}$

Output: The aggregate utility $\sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}^t)$.

$\pi_0((m,b),b') \leftarrow \frac{1}{|\{b" \in \mathcal{B}: b" \leq b, b" \leq v_{m+1}\}|}$ for all $m \in [M], b \geq b' \in \mathcal{B}, b' \leq v_{m+1}$. Let $\boldsymbol{q}^0 \in [M] \times \mathcal{B} \rightarrow [0,1]$ be the corresponding unit-bid value pair occupancy measure;

for $t \in [T]$: **do**

Determining the Bid Vector $\boldsymbol{b}^t$ recursively. Set b_1 to $b \in \mathcal{B}$ with probability $q_1^t(b)$;

for $m \in [1, \ldots, M-1], b \in \mathcal{B}: b_{m+1} \leftarrow b$ with probability $\pi^t((m, b_m), b)$;

Receive reward $\mu_n^t(\boldsymbol{b}^t) = \sum_{m=1}^{M} w_m^t(b_m^t)$ and observe $w_m^t(b_m^t)$ where $w_m^t(b) = (v_m - b)\mathbf{1}_{b \geq b_{-m}^t}$;

Update Reward Estimates;

for $m \in [M], b \in \mathcal{B}: \widehat{w}_m^t(b) \leftarrow \frac{w_m^t(b)}{q_m^{t-1}(b)} \mathbf{1}_{b=b_m^t}$ if *Bandit Feedback*, $\widehat{w}_m^t(b) \leftarrow w_m^t(b)$ if *Full Information*;

Determining Probability Measure $\boldsymbol{q}^t$ over any unit-bid value pair (m,b) Set

$$\boldsymbol{q}^t \leftarrow \text{argmin}_{\boldsymbol{q} \in \mathcal{Q}} \eta \langle \boldsymbol{q}, -\widehat{\boldsymbol{w}}^t \rangle + D(\boldsymbol{q} || \boldsymbol{q}^{t-1}),$$

where $\mathcal{Q}$ is as in Equation (6) and $D(\boldsymbol{q}||\boldsymbol{q}') = \sum_{m \in [M], b \in \mathcal{B}} q_m(b) \log \frac{q_m(b)}{q'_m(b)} - (q_m(b) - q'_m(b))$.

Convert $\boldsymbol{q}^t$ to Policy $\boldsymbol{\pi}^t$;

Compute any feasible solution $\boldsymbol{\pi}^t$ to constraints $q_m^t(b) = \sum_{b' \geq b} q_m^t(b') \pi^t((m-1, b'), b)$ and $\sum_{b" \leq b} \pi^t((m,b), b") = 1$ for all $m \in [M], b \in \mathcal{B}$.

end

Return $\sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}^t)$**.**

Algorithm Summary. Our algorithm (Algorithm 4) consists of four steps. First, we recursively sample $\boldsymbol{b}^t$ according to the policy $\boldsymbol{\pi}^t$. Second, we compute node utility estimates $\widehat{w}_m^t(b)$, either as the true loss $w_m^t(b)$ in the full information setting or the inverse probability weighted version $\frac{w_m^t(b)}{q_m^{t-1}(b)} \mathbf{1}_{b=b_m^t}$ in the bandit setting. Third, we optimize the negentropy-regularized expected estimated utility with respect to the probability measure over states $\boldsymbol{q}^t$, using OMD or Follow-the-Negentropy-Regularized-Leader updates. We show how this update can be efficiently computed by projecting the unconstrained optimizer of the regularized utility to the feasible space of $\boldsymbol{q}$, denoted by $\mathcal{Q}$, where the set $\mathcal{Q}$

$$\mathcal{Q} = \left\{ \boldsymbol{q} \in [0,1]^{M \times |\mathcal{B}|} : \sum_{b \in \mathcal{B}} q_m(b) = 1, \sum_{b \leq b'} q_{m+1}(b) \geq \sum_{b \leq b'} q_m(b) \forall b, b' \in \mathcal{B}, m \in [M] \right\}. \tag{6}$$

consists of probability distributions over bids satisfying certain stochastic dominance constraints which reflect the bid monotonicity constraints. That is, under $\mathcal{Q}$, $\boldsymbol{q}_{m+1}$ stochastically dominates $\boldsymbol{q}_m$ for all $m \in [M-1]$. Fourth, we convert $\boldsymbol{q}^t$ to a corresponding policy representation $\boldsymbol{\pi}^t$, ensuring that a feasible solution $\boldsymbol{\pi}^t$ exists as long as $\boldsymbol{q}^t \in \mathcal{Q}$.

We next discuss the main ideas and provide insights regarding our algorithm design.

Main Idea: Using the DP formulation to reduce our problem to online linear optimization. To design our algorithm, we observe that our DP formulation allows us to reduce the bidding problem to OLO over the space of possible node probability measures $\boldsymbol{q}$. Of course, we must justify why it is reasonable for our optimization procedure to only consider node probability measures instead of the larger space of possible policies $\boldsymbol{\pi}$. Recall that the reward at round t for bidding $\boldsymbol{b}$ is given by the sum of utilities of unit-bid values $\mu_n^t(\boldsymbol{b}) = \sum_{m=1}^{M} w_m^t(b_m)$. We then take expectations over the bid vector $\boldsymbol{b}$, sampled from the following policy $\boldsymbol{\pi}$ which induces probability measures $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_M$ over bid values; i.e., $q_m(b) = \mathbb{P}_{\boldsymbol{b}\sim\boldsymbol{\pi}}(b_m = b)$:

$$\mathbb{E}_{\boldsymbol{b}\sim\boldsymbol{\pi}}\left[\mu_n^t(\boldsymbol{b})\right] = \sum_{m=1}^{M} \mathbb{E}_{\boldsymbol{b}\sim\boldsymbol{\pi}}\left[w_m^t(b_m)\right] = \sum_{m=1}^{M} \mathbb{E}_{b_m\sim\boldsymbol{q}_m}\left[w_m^t(b_m)\right] = \sum_{m=1}^{M}\sum_{b\in\mathcal{B}} q_m(b) w_m^t(b) = \langle \boldsymbol{q}, \boldsymbol{w}^t \rangle\,. \tag{7}$$

Here, the last term is an inner product over the space $[M]\times\mathcal{B}$, and in the first equation, we invoke the linearity of bid vector utilities on its unit-bid value utilities. The second equality is justified because we are taking an expectation over possible bid vectors $\boldsymbol{b}$, as the $\boldsymbol{q}_m$'s are by definition the probabilities of selecting bid b and unit m. This addresses the question of why we concern ourselves only with the node probability measures $\boldsymbol{q}$ when optimizing, as the regret depends only on $\boldsymbol{q}$, rather than the associated policy. In other words, for a fixed $\boldsymbol{q}$, any policy $\boldsymbol{\pi}$ that induces node probability measures $\boldsymbol{q}$ will yield the same expected utility. Intuitively, this reflects the fact that the utilities are associated with nodes and not edges in our DP graph.

Letting $\boldsymbol{q}^t$ denote the probability measures induced by the (condensed) policy at round t, the instantaneous utility at round t is given by $\langle \boldsymbol{q}, \boldsymbol{w}^t \rangle$. Seeing this inner product begs use of OLO algorithms. However, most OLO algorithms require convexity of the action space which is, in our setting, the space of possible $\boldsymbol{q}$. To show that this space is convex, we invoke the following lemma.

Lemma 1 (**$\mathcal{Q}$-Space Equivalence**). *Let*

$$\Pi = \left\{\pi \in [0,1]^{M\times|\mathcal{B}|\times|\mathcal{B}|} : \pi((m,b),b') = 0 \quad \forall b' > b, m\in[M], \sum_{b'\leq b} \pi((m,b),b') = 1, m\in[M]\right\}$$

denote the space of policies on our DP graph. With a slight abuse of notation, for any $\pi\in\Pi$, define

$$q(\pi) = \{\mathbf{q}\in[0,1]^{M\times|\mathcal{B}|} : \forall b\in\mathcal{B}, q_1(b) = \pi((0,b_0),b), q_{m+1}(b) = \sum_{b'\in\mathcal{B}} q_m(b')\pi((m,b'),b), m\in[M-1]\}$$

as the node probabilities induced by π. Here, $b_0 = \max\mathcal{B}$. Let $\mathcal{Q}_\Pi = \cup_{\pi\in\Pi} q(\pi)$. Then, $\mathcal{Q}_\Pi$ is equivalent to the set $\mathcal{Q}$ where $\mathcal{Q}$ is defined in Equation (6).

At a high level, the proof requires constructing a bijection between elements of $\mathcal{Q}_\Pi$ and $\mathcal{Q}$. Showing that $\boldsymbol{q}\in\mathcal{Q}_\Pi$ implies $\boldsymbol{q}\in\mathcal{Q}$ follows straightforwardly by applying the linear transform q to the $\boldsymbol{\pi}$ associated with $\boldsymbol{q}$. The reverse direction requires a careful construction of a sequence of nested, non-empty subsets of Π that satisfy the $q_{m+1}(b) = \sum_{b'\in\mathcal{B}} q_m(b')\pi((m,b'),b)$ constraints.

Lemma 1 establishes that during the execution of Algorithm 4, we can focus on the node probabilities in set $\mathcal{Q}$ without loss of generality. We recall that within $\mathcal{Q}$, the stochastic dominance conditions are enforced solely over node probabilities across layers. In other words, when determining $\mathbf{q}^t$ in Algorithm 4, it is sufficient to consider the feasible set restricted to $\mathcal{Q}$, which is a convex set as $\mathcal{Q}$ is a polyhedron.

Now, we argue that we only need to consider optimizing over $\mathcal{Q}$ as opposed to Π, as the regret can be rewritten strictly in terms of $\boldsymbol{q}$, independently of the corresponding $\boldsymbol{\pi}$.

Lemma 2. *Any sequence of policies $\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^T$ over our DP graph with associated node probability measures $\boldsymbol{q}^1, \ldots, \boldsymbol{q}^T$ has discretized regret $\text{Regret}_{\mathcal{B}} = \max_{\boldsymbol{q} \in \mathcal{Q}} \sum_{t=1}^T \langle \boldsymbol{q} - \boldsymbol{q}^t, \boldsymbol{w}^t \rangle$. Here, $\boldsymbol{w}^t = \{w_m^t(b)\}_{m \in [M], b \in \mathcal{B}}$ represents vector of the round t rewards for all possible (m, b) unit-bid value pairs.*

Having shown convexity of our action space and the mapping to an equivalent OLO problem, we are ready to state the key result for Algorithm 4, for the bandit setting.

Theorem 6 **(Online Mirror Descent: Bandit Feedback)**. *With $\eta = \Theta(\sqrt{\frac{\log|\mathcal{B}|}{|\mathcal{B}|T}})$, Algorithm 4 achieves (discretized) regret $O(M\sqrt{|\mathcal{B}|T\log|\mathcal{B}|})$, with total time and space complexity polynomial in M, $|\mathcal{B}|$, and T. Optimizing for discretization error from restricting the bid space to $\mathcal{B}$, we obtain a continuous regret of $O(MT^{\frac{2}{3}})$.*

Under full information, we recover the regret bound of Algorithm 2 by replacing the node weight estimates with the true weights.

Corollary 1 **(Online Mirror Descent: Full Information)**. *With $\eta = \Theta(\sqrt{\frac{\log|\mathcal{B}|}{T}})$, Algorithm 4 achieves (discretized) regret $O(M\sqrt{T\log|\mathcal{B}|})$, with total time and space complexity polynomial in M, $|\mathcal{B}|$, and T. Optimizing for discretization error from restricting the bid space to $\mathcal{B}$, we obtain a continuous regret of $O(M\sqrt{T\log T})$.*

6. Regret Lower Bound

We remark that our OMD algorithms, under both full information and bandit settings, were designed to be robust to adversarial environments and incur discretized regret linear in M. In this section, we show that this is the best one can do, even in the stochastic setting. More specifically, we construct a corresponding (discretized) regret lower bound for our online bid optimization problem. At a high level, we will construct two bid vectors with nearly optimal expected utility under stochastic highest other bids. We derive the precise distribution of highest other bids and, using Le Cam's method, show that no algorithm in the full information or bandit feedback setting can learn the optimal bid vector quickly enough to avoid incurring $O(M\sqrt{T})$ regret.

Theorem 7. *Under the full information setting, the discretized regret is lower bounded with $\text{Regret}_{\mathcal{B}} \in \Omega(M\sqrt{T})$. This implies an equivalent regret lower bound in the bandit feedback setting.*

We remark that our regret lower bound matches our upper bound for the OMD algorithm in the full information setting (up to a $\sqrt{\log|\mathcal{B}|}$ factor), as well as in the bandit setting up to a factor of $\sqrt{|\mathcal{B}|\log|\mathcal{B}|}$ factor.

7. Experiments

In our experiments, we simulate the market dynamics induced by our learning algorithms; see the performance of our algorithms under a stochastic setting in Section 9.7.3. To better give context into the meaning of these experiments, we briefly discuss the notions of coarse correlated equilibria. We also provide specifics as to the slight modifications made in our algorithms to improve its empirical performance.

Coarse Correlated Equilibrium (CCE). In our experiments, we simulate the market dynamics in which every agent behaves according to our algorithms. This will allow us to obtain some insight into the structure, welfare, and revenues of the PAB CCEs recovered by our algorithms. CCEs are solution concepts that generalize Nash equilibria by allowing for dependence between bidder strategies. It is well known that the time-averaged behavior of agents running no-regret learning algorithms converges to a CCE and that these CCEs possess strong welfare and revenue guarantees in smooth games, such as PAB auctions Syrgkanis and Tardos (2012), de Keijzer et al. (2013a), Roughgarden (2015), Roughgarden et al. (2017), Feldman et al. (2017). While there are some limiting results describing the efficiency, revenue, and structure of Bayes-Nash equilibria of PAB auctions (Nautz 1995, de Keijzer et al. 2013a, Pycia and Woodward 2020), the CCEs have eluded an analytic characterization. We conduct several simulations of market dynamics under these no-regret learning algorithms to better understand the properties of these CCEs.

Algorithm Implementation. In our experiments, we run a slightly modified version of our algorithms in the bandit feedback setting. We do this as the variance of the regret of our algorithms is high, as the node weight estimators normalize over small probabilities $q^t_m(b)$. To mitigate the effect of such normalization, we use the implicit EXP3-IX estimator as described in (Neu 2015, Lattimore and Szepesvári 2020). Under this estimator, rather than reward estimate $\widehat{w}^t_m(b^t_m)$ of selecting bid b^t_m for unit m at time t by $q^t_m(b^t_m)$, we instead normalize it by $q^t_m(b^t_m)+\gamma$. That is, we use node reward estimator $\widehat{w}^t_m(b) = \frac{w^t_m(b)}{q^t_m(b)+\gamma}\mathbf{1}_{b=b^t_m}$ for specially chosen $\gamma > 0$ (see Section (9.7.1)) in our OMD algorithm. (Note that in the standard K-armed bandit setting, despite being a biased estimator, this algorithm still achieves the same sublinear expected regret guarantee with a smaller variance.) Aside from this modified estimator, the remainder of the Algorithm 4 remains the same.

Experiments. To that end, we analyze the bidding behavior of multiple learning agents and the induced market dynamics under full information with Algorithm 2 and under bandit feedback with Algorithm 4 (see Section 7.1). Note that we omit the bandit feedback decoupled exponential

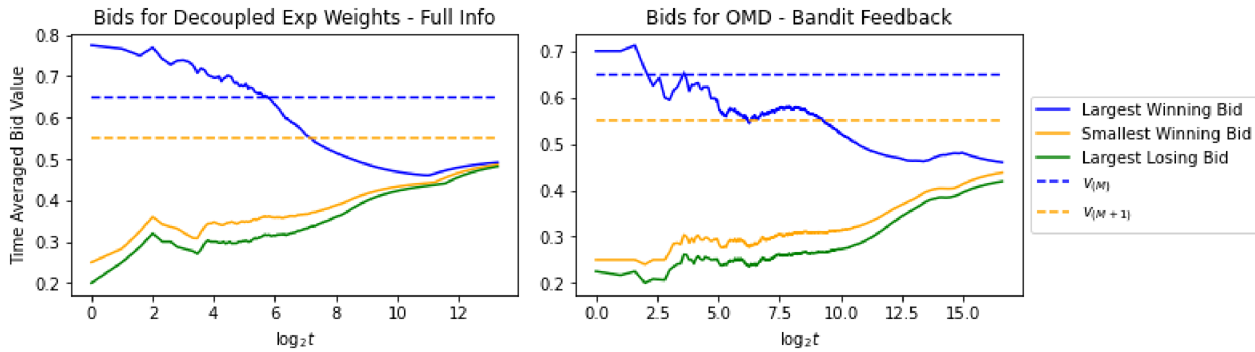

Figure 2 **Time averaged bids under market dynamics for the setting described in Section 7.1. The left (resp. right) figure corresponds with the full information setting (resp. bandit setting). In this specific instance, valuations are given by** $v_1 = [0.89, 0.7, 0.55, 0.51, 0.29], v_2 = [0.89, 0.44, 0.2, 0.12, 0.05], v_3 = [0.67, 0.64, 0.45, 0.27, 0.02]$**.**

weights (Algorithm 3) as its regret guarantees are dominated by the OMD variant both theoretically and empirically (See Section 9.7.3). Similarly, we omit the full information OMD algorithm as the convex optimization step is prohibitively computationally expensive, considering the marginal improvement in regret. We additionally compare the PAB market dynamics to the uniform pricing dynamics recovered under the no-regret learning algorithms described in Brânzei et al. (2023).

7.1. PAB Market Dynamics

In this experiment, we let there be $N = 3$ bidders, $\overline{M} = M = 5$ items, all valuations drawn from $\text{Unif}(0,1)$ which are then sorted, and the bid space is $\mathcal{B} = \{\frac{i}{20}\}_{i \in [20]}$, with higher indexed bidders receiving tie-break priority. We run the market dynamics where each agent is using the decoupled exponential weights algorithm under full information (Algorithm 2, with $T = 10^4, \eta = \sqrt{\frac{\log|\mathcal{B}|}{T}} = 0.036$) and the OMD algorithm under bandit feedback (Algorithm 4 with $T = 10^5$ and $\eta = \sqrt{\frac{\log|\mathcal{B}|}{|\mathcal{B}|T}} = 0.0008$).

Bids Dynamics. In Figure 2, we analyze the bids over time of the market dynamics induced when all agents bid according to our Algorithms 2 and 4. We observe that the winning bids (and largest losing bid) converge to approximately the same value. Informally, while the converged prices are slightly different, our learning algorithms induce market dynamics under which prices are almost the same for all the bidders. (See also the left plot of Figure 8 for a similar observation.)

Utilities Dynamics. In Figure 3, we analyze the time averaged utility of the agent in the market dynamics induced when all agents bid according to Algorithm 2 under the full info setting and Algorithm 4 under the bandit setting. In both the full information and bandit feedback settings, the utilities converge to the optimal utilities overtime; albeit significantly faster in the full information setting. As a consequence, the algorithms converge to a welfare optimal CCE. We also note that

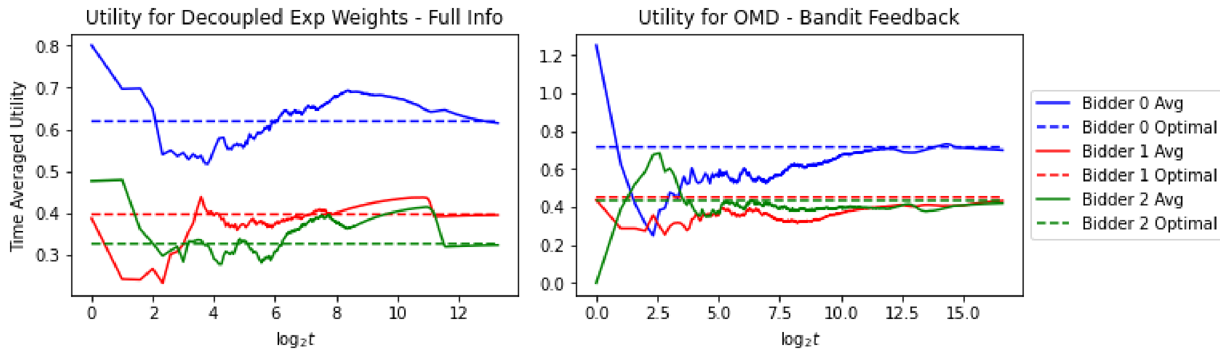

Figure 3 **Time averaged utilities under market dynamics for the setting described in Section 7.1. The left (resp. right) figure corresponds with the full information setting (resp. bandit setting). In this specific instance, valuations are given by** $\boldsymbol{v}_1 = [0.89, 0.7, 0.55, 0.51, 0.29], v_2 = [0.89, 0.44, 0.2, 0.12, 0.05], v_3 = [0.67, 0.64, 0.45, 0.27, 0.02]$**.**

the hindsight optimal utility for each bidder differs across the two settings, potentially due to the different prices converged to. One interesting observation is that the shape of the regret curves are similar for both settings, indicating that the agents learned similar sets of strategies under, albeit more slowly and with higher variance in the bandit feedback setting.

Distribution of Revenues. In Figure 4, we compare the distribution of per-item time averaged revenue under Algorithm 2 for the full information setting and Algorithm 4 for the bandit setting. Here, we normalize by either the per-item average welfare or by the $\overline{M}$'th and $\overline{M}+1$ largest valuations (i.e., $v_{(\overline{M})}$ and $v_{(\overline{M}+1)}$). We also plot both the time averaged and last iterate versions to see that the bids did indeed converge last-iterate wise. We note that while the revenues are generally lower in the bandit setting, the distribution of revenue of both algorithms generally maintain the same shape and it is not too heavy tailed towards low revenue. In fact, in most cases, the revenue returned is at least half of the maximum welfare. This is hopeful considering that the impact of strategizing and bid shading is large as there are a small number $N = 3$ agents. We further note that revenue recovered is actually *smaller* than both $v_{(\overline{M})}$ and $v_{(\overline{M}+1)}$. This suggests that in some cases, there exists an agent whose valuation exceeds the clearing price, but does not actually learn to win this item; as they have foregone exploring this possibility in order to lower their payments for units they are already winning. Indeed, in several instances, some agents fail to learn their hindsight optimal bid and incur large regret. [3]

Welfare and Revenue Over Time. In Figure 5, we further compare the distribution of welfare and revenue over time showing the 10th, 25th, 50th, 75th, and 90th percentiles in different shades under Algorithm 2 and Algorithm 4. We normalize both welfare and revenue by the maximum

[3] We note that the occasional non-convergence to 0 regret does not contradict our results as our algorithms only guarantee convergence of the *expected* regret. See Section 11.5 in Lattimore and Szepesvári (2020).

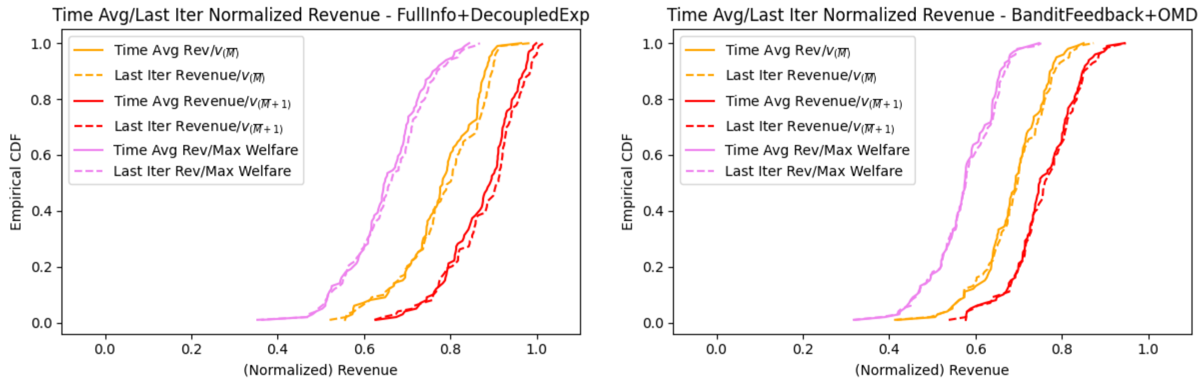

Figure 4 **Distribution of revenue for the setting described in Section 7.1. The left (resp. right) figure corresponds with the full information setting (resp. bandit setting).**

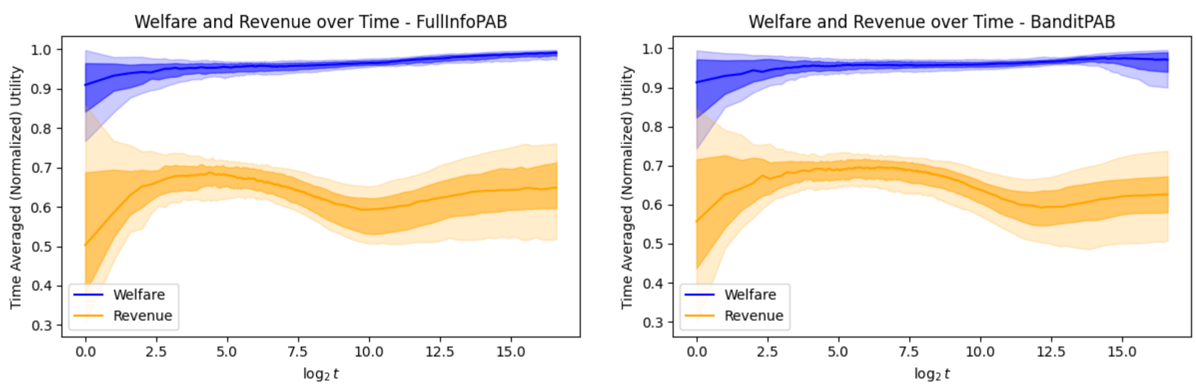

Figure 5 **Welfare and revenue over time under the market dynamics for the setting described in Section 7.1. The left (resp. right) figure corresponds with the full information setting (resp. bandit setting).**

.

possible welfare (sum of the largest M valuations) in each instance. In comparison to the full information version, it takes longer for the bidders to settle to an approximately welfare maximizing steady state in the bandit setting. Furthermore, the revenue at the recovered steady state under bandit feedback is lower than that of the full information setting, albeit with lower variance.

Impact of Competition. In Figure 6, we compare the distribution of welfare and revenue over time showing the 10th, 25th, 50th, 75th, and 90th percentiles in different shades under Algorithm 4 for the bandit setting for a varying number of market participants $N \in \{12, 48\}$ with $T = 10^4$. Compared with the previous experiment with $N = 3$ bidders, we see that as N grows larger, there is less incentive for agents to bid strategically as there is more competition, and thus, the agents' bids, welfare, and revenue converge more quickly. Additionally, the increased competition also improves the revenue, as expected.

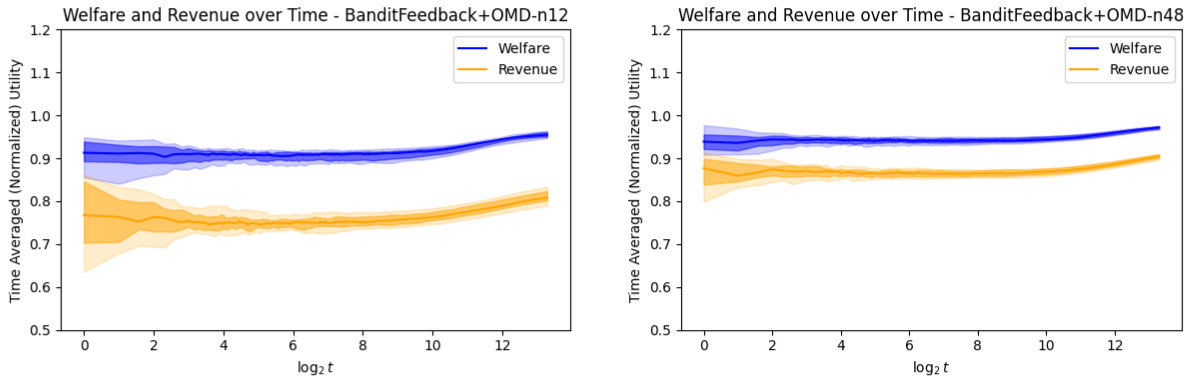

Figure 6 **Effect of competition under the market dynamics for the setting described in Section 7.1. In the left figure, the number of agents is 12 while in the right one, it is 48. In both figures, we run Algorithm 4.**

.

7.2. PAB vs. Uniform Price Auction Market Dynamics

In the following set of experiments, we compare the bids, welfare, and revenue of the no-regret market dynamics recovered in the uniform price auction (using the same parameterizations $N = 3, M = 5, |\mathcal{B}| = 20, T = 10^5$ and valuations as for the PAB learning algorithms). More specifically, we implement the no-regret learning algorithm for the uniform price auction as described in Brânzei et al. (2023), with the payment equal to the smallest winning bid. (An alternative approach would be to set the payment equal to the largest losing bid. However, as argued in Burkett and Woodward (2020), Brânzei et al. (2023), this strategy yields low revenue under Nash Equilibria. This insight has inspired us to set the payment equal to the smallest winning bid in our implementation of uniform price auctions.)

In our implementation of these algorithms, we impose no-overbidding constraint just as we had in the PAB setting, which prevents the learning dynamics from converging to a non-IR equilibrium. Similarly, we also use the EXP3-IX based estimator in the bandit uniform price algorithm to be consistent with the PAB implementation. In Sections 9.7.3 and 9.8, we run additional experiments that provide more detail regarding the regret rates, bid dynamics, and the evolution of revenue and welfare of the uniform price auction.

Time Averaged Welfare and Revenue. In Figure 7, we present a direct comparison of the time-averaged welfare and revenue (normalized by maximum welfare) between the PAB and uniform price auctions, considering both full information and bandit feedback scenarios. The plot includes the median (represented by a horizontal line), mean (indicated by an ‘x’), 25th and 75th percentiles, as well as the minimum and maximum values for welfare and revenue. Our observations reveal that in both full information and bandit feedback settings, the uniform price dynamics yield slightly higher welfare but significantly lower revenue compared to the PAB dynamics. This

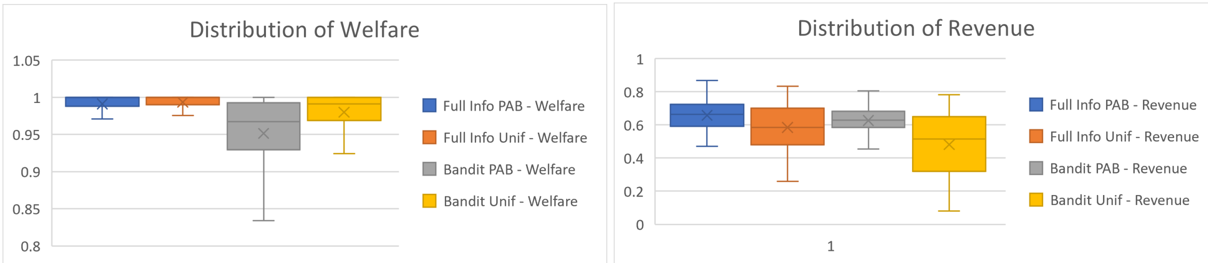

Figure 7 **Welfare (left) and revenue (right) comparisons between the PAB and Uniform Price auctions, under both full information and bandit feedback. We use the setting described in Section 7.2.**

disparity is further amplified in the bandit feedback scenario. Specifically, in bandit feedback, the mean welfare for the uniform price auction is 0.980 (with a standard deviation of 0.028), while the PAB mechanism achieves a mean welfare of 0.952 (with a standard deviation of 0.049). Similarly, the mean revenue for the uniform price auction is 0.481 (with a standard deviation of 0.194), whereas the PAB mechanism achieves a mean revenue of 0.626 (with a standard deviation of 0.091). It is worth noting that the variance in uniform price revenue under bandit feedback is particularly concerning, as the revenue can be as low as 8% of the maximum welfare, compared to 39% under the PAB mechanism. Conversely, the minimum welfare achieved under bandit feedback for the PAB mechanism is 77%, while the uniform price auction attains a minimum welfare of 86%.

The slight decrease in welfare observed in PAB can be attributed to the increased incentive for bidders to understate their bids in order to reduce their payment. In PAB, the payment and utility for each individual unit directly depend on the bid placed on that specific unit. On the other hand, in the uniform price auction, the payment is determined by the lowest winning bid, and there is consequently less motivation to manipulate higher bids since they are less correlated with the payment. This observation aligns with findings from other variants of discriminatory versus uniform pricing, such as generalized first and second price auctions. Conversely, the improved stability and higher revenue observed in PAB can be attributed to the ability of bidders to strategically shade their bids on a per-unit basis. In PAB, bidders have more direct control over their payments compared to the uniform price auction, where the payment per unit can fluctuate significantly depending on a single clearing price. This increased control allows bidders to optimize their bids strategically and leads to enhanced stability and higher revenue in the PAB mechanism.

Convergence of Bids. In Figure 8, we compare the ratios of (i) the largest to the smallest winning bid and (ii) the smallest winning bid to the largest losing bid. We plot these ratios over time for both the PAB and uniform price auctions, under bandit feedback only (we exclude full information feedback as the differences in welfare and revenue are negligible). We make the crucial observation that while both ratios converge to 1 ($\log_2 1 = 0$ in the plot) in the PAB auction, the

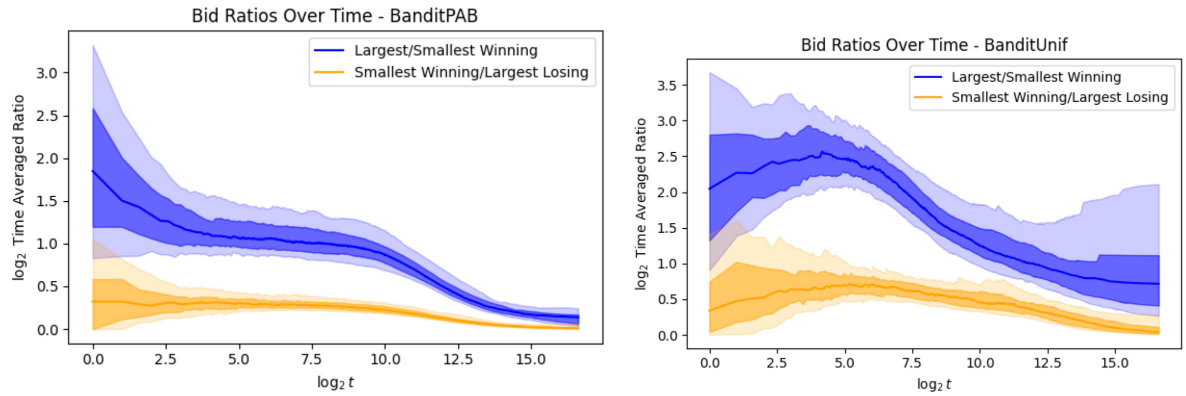

Figure 8 **We compare the ratios of the largest winning bid to the smallest winning bid and also the smallest winning bid to the largest losing bid for the setting described in Section 7.1. The left (resp. right) figure corresponds with the PAB dynamics (Uniform pricing dynamics).**

.

ratio of the largest to smallest winning bid in the uniform price auction does not converge to 1. The convergence of these ratios for the PAB setting indicate that the optimal bidding strategy for the PAB auction is uniform bidding. In contrast, the non-convergence for the uniform price setting suggests that the optimal bidding in the uniform price auction requires discriminatory bidding (different prices for each unit). See our discussions in Section 1.2 regarding the practical implications of this observation.

Furthermore, we observe that the ratio of the smallest winning to largest losing bids converges noticeably more slowly under the uniform price auction. This slower convergence can be attributed to the increased number of edge weights $O(M|\mathcal{B}|^2)$ required to learn in the uniform price auction as compared to the $O(M|\mathcal{B}|)$ node weights in the PAB auction.

8. Concluding Remarks

We have provided low-regret learning algorithms for PAB auctions in the full information and bandit settings with corresponding polynomial time and space complexities. In particular, we utilize our DP formulation and its equivalent graph representation to decouple the utility associated with bidding $b_m = b$ for all $m \in [M], b \in \mathcal{B}$. We derived two algorithms, one that mimics the exponential weights algorithm and another based on OMD, both of which allowed us to achieve polynomial (in M, $|\mathcal{B}|$, and T) regret upper bounds, as well as time and space complexities, despite the combinatorially large bid space.

There are several intriguing avenues for future research that can be explored based on the current work. A promising direction is to leverage the structure induced by bid monotonicity in PAB auctions. Recent advancements in a simpler single-unit setting have demonstrated the efficacy of cross-learning between bids under certain feedback structures (Han et al. 2020a,b). It would be intriguing to investigate the potential benefits of applying cross-learning techniques in our multi-unit setting. By incorporating such methods, we can explore whether they can enhance our regret bounds. Furthermore, inspired by our numerical results—where we show that the winning bids in PAB market dynamics converge to the same value—we can explore the design of online learning algorithms for the setting where bidders are restricted to a simplified bidding interface, wherein they are only allowed to submit a single price and quantity for the units demanded rather than an entire vector of bids.

References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (Eds.), Vol. 24. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf

Gagan Aggarwal, Ashwinkumar Badanidiyuru, and Aranyak Mehta. 2019. Autobidding with constraints. In *Web and Internet Economics: 15th International Conference, WINE 2019, New York, NY, USA, December 10–12, 2019, Proceedings 15*. Springer, 17–30.

Mohammad Akbarpour and Shengwu Li. 2020. Credible Auctions: A Trilemma. *Econometrica* 88, 2 (2020), 425–467. https://doi.org/10.3982/ECTA15925 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA15925

Masaki Aoyagi. 2003. Bid rotation and collusion in repeated auctions. *Journal of Economic Theory* 112, 1 (2003), 79–105. https://doi.org/10.1016/S0022-0531(03)00071-1

Jean-Yves Audibert, Sebastien Bubeck, and Gabor Lugosi. 2011. Minimax Policies for Combinatorial Prediction Games. (2011). https://doi.org/10.48550/ARXIV.1105.4871

Lawrence Ausubel and Paul Milgrom. 2006. The Lovely but Lonely Vickrey Auction. *Comb. Auct.* 17 (01 2006). https://doi.org/10.7551/mitpress/9780262033428.003.0002

Lawrence M. Ausubel. 2004. An Efficient Ascending-Bid Auction for Multiple Objects. *American Economic Review* 94, 5 (December 2004), 1452–1475. https://doi.org/10.1257/0002828043052330

Lawrence M. Ausubel and Peter Cramton. 2005. Dynamic Auctions in Procurement. *Handbook of Procurement* (2005). https://www.cramton.umd.edu/papers2005-2009/ausubel-cramton-dynamic-auctions-in-procurement.pdf

Lawrence M. Ausubel, Peter Cramton, Marek Pycia, Marzena Rostek, and Marek Weretka. 2014. Demand Reduction and Inefficiency in Multi-Unit Auctions. *The Review of Economic Studies* 81, 4 (289) (2014), 1366–1400. http://www.jstor.org/stable/43551738

Moshe Babaioff, Nicole Immorlica, Yingkai Li, and Brendan Lucier. 2022. Making Auctions Robust to Aftermarkets. arXiv:2107.05853 [econ.TH]

Ashwinkumar Badanidiyuru, Zhe Feng, and Guru Guruganesh. 2021. Learning to Bid in Contextual First Price Auctions. https://doi.org/10.48550/ARXIV.2109.03173

Brian Baisa and Justin Burkett. 2018. Large multi-unit auctions with a large bidder. *Journal of Economic Theory* 174 (2018), 1–15.

Santiago Balseiro, Negin Golrezaei, Mohammad Mahdian, Vahab Mirrokni, and Jon Schneider. 2022. Contextual Bandits with Cross-Learning. *Mathematics of Operations Research* (2022).

G. Bartok, D. P´al, C. Szepesv´ari, and I. Szita. 2011. Online learning. Lecture notes, University of Alberta. (2011). http://david.palenica.com/papers/online-learning-lecture-notes/online-learning-lecture-notes-2011-Oct-20.pdf#page44

Hedyeh Beyhaghi, Negin Golrezaei, Renato Paes Leme, Martin Pál, and Balasubramanian Sivan. 2021. Improved Revenue Bounds for Posted-Price and Second-Price Mechanisms. *Operations research.* 69, 6 (2021).

Jason Bigler. 2019. An update on first price auctions for Google Ad Manager. (2019). https://blog.google/products/ads-commerce/update-first-price-auctions-google-ad-manager/

Ken Binmore and Joe Swierzbinski. 2000. Treasury auctions: Uniform or discriminatory? *Review of Economic Design* 5 (2000), 387–410.

Avrim Blum, MohammadTaghi Hajiaghayi, Katrina Ligett, and Aaron Roth. 2008. Regret Minimization and the Price of Total Anarchy. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing* (Victoria, British Columbia, Canada) *(STOC '08)*. Association for Computing Machinery, New York, NY, USA, 373–382. https://doi.org/10.1145/1374376.1374430

Simina Brânzei, Mahsa Derakhshan, Negin Golrezaei, and Yanjun Han. 2023. Online Learning in Multi-unit Auctions. arXiv:2305.17402 [cs.GT]

Justin Burkett and Kyle Woodward. 2020. Uniform price auctions with a last accepted bid pricing rule. *Journal of Economic Theory* 185 (2020), 104954. https://doi.org/10.1016/j.jet.2019.104954

Lucian Buşoniu, Robert Babuška, and Bart De Schutter. 2010. *Multi-agent Reinforcement Learning: An Overview*. Springer Berlin Heidelberg, Berlin, Heidelberg, 183–221. https://doi.org/10.1007/978-3-642-14435-6_7

Yang Cai and Constantinos Daskalakis. 2017. Learning Multi-item Auctions with (or without) Samples. In *FOCS*.

Emilio Calvano, Giacomo Calzolari, Vincenzo Denicoló, and Sergio Pastorello. 2021. Algorithmic collusion with imperfect monitoring. *International Journal of Industrial Organization* 79 (2021), 102712. https://doi.org/10.1016/j.ijindorg.2021.102712

Wei Chen, Yajun Wang, and Yang Yuan. 2013. Combinatorial Multi-Armed Bandit: General Framework and Applications. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, Atlanta, Georgia, USA, 151–159. https://proceedings.mlr.press/v28/chen13a.html

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. 2011. Contextual Bandits with Linear Payoff Functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 15)*, Geoffrey Gordon, David Dunson, and Miroslav Dudík (Eds.). PMLR, Fort Lauderdale, FL, USA, 208–214. https://proceedings.mlr.press/v15/chu11a.html

Peter Cramton. 1998. Ascending auctions. *European Economic Review* 42, 3 (1998), 745–756. https://doi.org/10.1016/S0014-2921(97)00122-0

Peter Cramton, Yoav Shoham, and Richard Steinberg. 2004. *Combinatorial Auctions.* Papers of Peter Cramton 04mit. University of Maryland, Department of Economics - Peter Cramton. https://ideas.repec.org/p/pcc/pccumd/04mit.html

Varsha Dani, Thomas Hayes, and Sham Kakade. 2008. Stochastic Linear Optimization under Bandit Feedback. 355–366.

Bart de Keijzer, Evangelos Markakis, Guido Schäfer, and Orestis Telelis. 2013a. Inefficiency of Standard Multi-unit Auctions. In *Algorithms – ESA 2013*, Hans L. Bodlaender and Giuseppe F. Italiano (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 385–396.

Bart de Keijzer, Evangelos Markakis, Guido Schäfer, and Orestis Telelis. 2013b. On the Inefficiency of Standard Multi-Unit Auctions. *CoRR* abs/1303.1646 (2013). arXiv:1303.1646 http://arxiv.org/abs/1303.1646

Yuan Deng, Negin Golrezaei, Patrick Jaillet, Jason Cheuk Nam Liang, and Vahab Mirrokni. 2023. Multi-channel Autobidding with Budget and ROI Constraints. *arXiv e-prints* (2023), arXiv–2302.

S. Deospotakis, R. Ravi, and A. Sayedi. 2021. First-Price Auctions in Online Display Advertising. Journal of Marketing Research. *Journal of Marketing Research* 58(5) (2021), 888–907. https://doi.org/10.1177/00222437211030201

Mahsa Derakhshan, Negin Golrezaei, and Renato Paes Leme. 2022. Linear program-based approximation for personalized reserve prices. *Management Science* 68, 3 (2022), 1849–1864.

Mahsa Derakhshan, David M Pennock, and Aleksandrs Slivkins. 2021. Beating greedy for approximating reserve prices in multi-unit VCG auctions. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 1099–1118.

Shahar Dobzinski and Noam Nisan. 2007. Mechanisms for Multi-Unit Auctions. In *Proceedings of the 8th ACM Conference on Electronic Commerce* (San Diego, California, USA) *(EC '07)*. Association for Computing Machinery, New York, NY, USA, 346–351. https://doi.org/10.1145/1250910.1250960

Miroslav Dudík, Nika Haghtalab, Haipeng Luo, Robert E. Schapire, Vasilis Syrgkanis, and Jennifer Wortman Vaughan. 2017. Oracle-Efficient Learning and Auction Design. In *FOCS*.

EEX. 2023. EU ETS Auctions. (2023). https://www.eex.com/en/markets/environmental-markets/eu-ets-auctions#:~:text=Bids%20are%20sorted%20in%20descending,selection%20according%20to%20an%20algorithm

Natalia Fabra, Nils-Henrik von der Fehr, and David Harbord. 2006. Designing Electricity Auctions. *The RAND Journal of Economics* 37, 1 (2006), 23–46. http://www.jstor.org/stable/25046225

Giulio Federico and David Rahman. 2003. Bidding in an Electricity Pay-as-Bid Auction. *Journal of Regulatory Economics* (2003). https://link.springer.com/article/10.1023/A:1024738128115

Michal Feldman, Brendan Lucier, and Noam Nisan. 2017. Correlated and Coarse equilibria of Single-item auctions. arXiv:1601.07702 [cs.GT]

Kenneth Garbade and Jeffrey Ingber. 2005. The Treasury Auction Process: Objectives, Structure, and Recent Adaptations. (2005). https://ssrn.com/abstract=678301

Kira Goldner, Nicole Immorlica, and Brendan Lucier. 2019. Reducing Inefficiency in Carbon Auctions with Imperfect Competition. https://doi.org/10.48550/ARXIV.1912.06428

Negin Golrezaei, Patrick Jaillet, and Jason Cheuk Nam Liang. 2020. No-regret learning in price competitions under consumer reference effects. *Advances in Neural Information Processing Systems* 33 (2020), 21416–21427.

Negin Golrezaei, Patrick Jaillet, and Jason Cheuk Nam Liang. 2023a. Incentive-aware contextual pricing with non-parametric market noise. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 9331–9361.

Negin Golrezaei, Patrick Jaillet, Jason Cheuk Nam Liang, and Vahab Mirrokni. 2023b. Pricing against a Budget and ROI Constrained Buyer. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 9282–9307.

Negin Golrezaei, Adel Javanmard, and Vahab Mirrokni. 2021a. Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Operations Research* 69, 1 (2021), 297–314.

Negin Golrezaei, Max Lin, Vahab Mirrokni, and Hamid Nazerzadeh. 2021b. Boosted second price auctions: Revenue optimization for heterogeneous bidders. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 447–457.

Larence H. Goulder and Andrew R. Schein. 2013. Carbon Taxes versus Cap and Trade: A Critical Review. *Climate Change Economics* 4 (2013). https://www.worldscientific.com/doi/10.1142/S2010007813500103

ERBACH Gregor. 2023. Review of the EU ETS: Fit for 55. (2023).

Yanjun Han, Zhengyuan Zhou, Aaron Flores, Erik Ordentlich, and Tsachy Weissman. 2020b. Learning to Bid Optimally and Efficiently in Adversarial First-price Auctions. https://doi.org/10.48550/ARXIV.2007.04568

Yanjun Han, Zhengyuan Zhou, and Tsachy Weissman. 2020a. Optimal No-regret Learning in Repeated First-price Auctions. https://doi.org/10.48550/ARXIV.2003.09795

Zhao Haoyu and Chen Wei. 2020. Online Second Price Auction with Semi-Bandit Feedback under the Non-Stationary Setting. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 04 (Apr. 2020), 6893–6900. https://doi.org/10.1609/aaai.v34i04.6171

Jason Hartline, Darrell Hoy, and Sam Taggart. 2014. Price of Anarchy for Revenue. (04 2014).

Jason Hartline, Vasilis Syrgkanis, and Eva Tardos. 2015. No-Regret Learning in Bayesian Games. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2015/file/3e7e0224018ab3cf51abb96464d518cd-Paper.pdf

Sven Heim and Georg Götz. 2021. Do Pay-As-Bid Auctions Favor Collusion? Evidence from Germany's market for reserve power. *Energy Policy* 155 (2021), 112308. https://doi.org/10.1016/j.enpol.2021.112308

Kenneth Hendricks and Robert H. Porter. 1989. Collusion in Auctions. *Annales d'Économie et de Statistique* 15/16 (1989), 217–230. http://www.jstor.org/stable/20075758

Ali Hortaçsu and David McAdams. 2010. Mechanism Choice and Strategic Bidding in Divisible Good Auctions: An Empirical Analysis of the Turkish Treasury Auction Market. *Journal of Political Economy* 118, 5 (2010), 833–865. http://www.jstor.org/stable/10.1086/657948

Zhiyi Huang, Jinyan Liu, and Xiangning Wang. 2018. Learning Optimal Reserve Price against Non-myopic Bidders. https://doi.org/10.48550/ARXIV.1804.11060

Yash Kanoria and Hamid Nazerzadeh. 2014. Dynamic Reserve Prices for Repeated Auctions: Learning from Bids. In *Web and Internet Economics: 10th International Conference, WINE 2014, Beijing, China, December 14-17, 2014, Proceedings*, Vol. 8877. Springer, 232.

Yoav Kolumbus and Noam Nisan. 2022. Auctions between Regret-Minimizing Agents. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 100–111. https://doi.org/10.1145/3485447.3512055

Wouter M Koolen, Manfred K Warmuth, and Jyrki Kivinen. 2010. Hedging Structured Concepts. *Conference on Learning Theory* 23 (2010), 93–105.

Tor Lattimore and Csaba Szepesvári. 2020. *Bandit Algorithms*. Cambridge University Press. https://doi.org/10.1017/9781108571401

Nick Littlestone and Manfred K. Warmuth. 1994. The Weighted Majority Algorithm. *Inf. Comput.* 108, 2 (1994), 212–261. https://doi.org/10.1006/inco.1994.1009

Brendan Lucier and Allan Borodin. 2009. Price of anarchy for greedy auctions. In *ACM-SIAM Symposium on Discrete Algorithms*.

Brendan Lucier, Sarath Pattathil, Aleksandrs Slivkins, and Mengxiao Zhang. 2023. Autobidders with budget and roi constraints: Efficiency, regret, and pacing dynamics. *arXiv preprint arXiv:2301.13306* (2023).

Thodoris Lykouris, Vasilis Syrgkanis, and Éva Tardos. 2016. Learning and Efficiency in Games with Dynamic Population. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms* (Arlington, Virginia) *(SODA '16)*. Society for Industrial and Applied Mathematics, USA, 120–129.

Mehryar Mohri and Andres Muñoz Medina. 2013. Learning Theory and Algorithms for Revenue Optimization in Second-Price Auctions with Reserve. `https://doi.org/10.48550/ARXIV.1310.5665`

Mehryar Mohri and Andrés Muñoz Medina. 2015. Revenue Optimization against Strategic Buyers. In *NIPS*.

Mehryar Mohri and Andrés Munoz Medina. 2016. Learning algorithms for second-price auctions with reserve. *The Journal of Machine Learning Research* 17, 1 (2016), 2632–2656.

Jamie Morgenstern and Tim Roughgarden. 2016. Learning Simple Auctions. In *29th Annual Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 49)*, Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (Eds.). PMLR, Columbia University, New York, New York, USA, 1298–1318.

D. Nautz. 1995. Optimal bidding in multi-unit auctions with many bidders. *Economics Letters* 48, 3 (1995), 301–306. `https://doi.org/10.1016/0165-1765(94)00641-E`

Thomas Nedelec, Noureddine El Karoui, and Vianney Perchet. 2019. Learning to bid in revenue-maximizing auctions. *CoRR* abs/1902.10427 (2019). arXiv:1902.10427 `http://arxiv.org/abs/1902.10427`

Gergely Neu. 2015. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. arXiv:1506.03271 [cs.LG]

Rad Niazadeh, Negin Golrezaei, Joshua Wang, Fransisca Susan, and Ashwinkumar Badanidiyuru. 2022. Online Learning via Offline Greedy Algorithms: Applications in Market Design and Optimization. *Management Science* (2022).

Martin Pesendorfer. 2000. A Study of Collusion in First-Price Auctions. *The Review of Economic Studies* 67, 3 (2000), 381–411. `http://www.jstor.org/stable/2566959`

Steve Phelps, Kai Cai, Peter McBurney, Jinzhong Niu, Simon Parsons, and Elizabeth Sklar. 2008. Auctions, Evolution, and Multi-agent Learning. In *Adaptive Agents and Multi-Agent Systems III. Adaptation and Multi-Agent Learning*, Karl Tuyls, Ann Nowe, Zahia Guessoum, and Daniel Kudenko (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 188–210.

David Porter, Stephen Rassenti, Anil Roopnarine, and Vernon Smith. 2003. Combinatorial auction design. *Proceedings of the National Academy of Sciences* 100, 19 (2003), 11153–11157. `https://doi.org/10.1073/pnas.1633736100` arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1633736100

Market Pycia and Kyle Woodward. 2020. Auctions of Homogeneous Goods: A Case for Pay-as-Bid. (2020). `https://ssrn.com/abstract=3865760orhttp://dx.doi.org/10.2139/ssrn.3865760`

Alexander Rakhlin. 2009. Lecture Notes on Online Learning. (2009). `https://www.mit.edu/~rakhlin/papers/online_learning.pdf`

Tim Roughgarden. 2015. The Price of Anarchy in Games of Incomplete Information. *ACM Trans. Econ. Comput.* 3, 1, Article 6 (mar 2015), 20 pages. https://doi.org/10.1145/2737816

Tim Roughgarden, Vasilis Syrgkanis, and Éva Tardos. 2017. The Price of Anarchy in Auctions. *J. Artif. Int. Res.* 59, 1 (may 2017), 59–101.

Tim Roughgarden and Joshua R. Wang. 2016. Minimizing Regret with Multiple Reserves. In *Proceedings of the 2016 ACM Conference on Economics and Computation, EC '16, Maastricht, The Netherlands, July 24-28, 2016*. 601–616.

Richard Schmalensee and Robert N. Stavins. 2017. Lessons Learned from Three Decades of Experience with Cap and Trade. *Review of Environmental Economics and Policy* 11 (2017). https://www.journals.uchicago.edu/doi/full/10.1093/reep/rew017

Fransisca Susan, Negin Golrezaei, and Okke Schrijvers. 2023. Multi-Platform Budget Management in Ad Markets with Non-IC Auctions. *arXiv preprint arXiv:2306.07352* (2023).

Vasilis Syrgkanis and Eva Tardos. 2012. Composable and Efficient Mechanisms. arXiv:1211.1325 [cs.GT]

Eiji Takimoto and Manfred K. Warmuth. 2003. Path Kernels and Multiplicative Updates. *J. Mach. Learn. Res.* 4, null (dec 2003), 773–818.

Susan Tierney, Todd Schatzki, and Rana Mukerji. 2008. Uniform-Pricing versus Pay-as-Bid in Wholesale Electricity Markets: Does it Make a Difference? (2008). https://kylewoodward.com/blog-data/pdfs/references/tierney+schatzki+mukerji-new-york-iso-2008A.pdf

Alexandre B. Tsybakov. 2008. *Introduction to Nonparametric Estimation* (1st ed.). Springer Publishing Company, Incorporated.

Bart PG Van Parys and Negin Golrezaei. 2023. Optimal learning for structured bandits. *Management Science* (2023).

Yaodong Yang and Jun Wang. 2020. An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective. *CoRR* abs/2011.00583 (2020). arXiv:2011.00583 https://arxiv.org/abs/2011.00583

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2021. *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms.* Springer International Publishing, Cham, 321–384. https://doi.org/10.1007/978-3-030-60990-0_12

Alexander Zimin and Gergely Neu. 2013. Online learning in episodic Markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf

9. Appendix

9.1. Proof of Theorem 3: Offline Bid Optimization Algorithm

We give a proof of correctness of the offline bid optimization algorithm used to compute the hindsight optimal bid vector across T rounds of PAB auctions. Our proof shows that the variables $U_m(b)$ are path weights of the optimal partial bid vector with weights $W_m^{T+1}(b)$. Thus, $U_1(b)$ is the optimal bid vector and b_m^*'s can be used to back out the optimal bid vector recursively in polynomial time.

Proof of Theorem 3 By definition,

$$
\begin{aligned}
U_m(b) &= \max_{b \geq b_m \geq \cdots \geq b_M} \sum_{m'=m}^{M} W_{m'}^{T+1}(b_{m'}) \\
&= \max_{b \geq b_m \geq \cdots \geq b_M} W_m^{T+1}(b_m) + \sum_{m'=m+1}^{M} W_{m'}^{T+1}(b_{m'}) \\
&= \max_{b' \leq b} W_m^{T+1}(b') + U_{m+1}(b') \,.
\end{aligned}
$$

Since we have that $U_M(b) = W_M^{T+1}(b)$ trivially correct from the base case, and the optimality of $U_m(b)$ follows from induction. Consequently, optimality of b_m^* follows from induction. The base case trivially holds as $b_1^* = \mathrm{argmax}_{b \in \mathcal{B}} U_1(b)$. The recursive case also follows straightforwardly by definition of b_m^*:

$$
b_m^* = \mathrm{argmax}_{b \leq b_{m-1}^*} U_m(b) \,.
$$

As b_{m-1}^* was optimal by the induction hypothesis, then b_m^* must also be optimal.

We finish this proof by discussing the time and space complexity of Algorithm 1. Table U is of size $O(M|\mathcal{B}|)$, with each entry requiring taking a maximum over $O(|\mathcal{B}|)$ terms, yielding time and space complexities of $O(M|\mathcal{B}|^2)$ and $O(M|\mathcal{B}|)$ respectively.

9.2. Proof of Theorems 4 and 5: Decoupled Exponential Weights Algorithm

We give the proofs of correctness, complexity analysis, and regret analysis for the decoupled exponential weights algorithms for both the full information (Algorithm 2) and the bandit setting (Algorithm 3). Our proof comes in 5 parts. We first prove correctness of the bandit version of our algorithm. In particular, we show that defining the node and bid vector utility estimates to be $\widehat{w}_m^t(b) = 1 - \frac{1-(v_m-b)1_{b \geq b^t_{-m}}}{q_m^t(b)} 1_{b=b_m^t}$ and $\widehat{\mu}^t(\boldsymbol{b}) = \sum_{m=1}^M \widehat{w}_m^t(b_m)$, our algorithm samples bid vector $\boldsymbol{b}^t$ with probability proportional to $\exp(\eta \sum_{\tau=1}^{t-1} \widehat{\mu}^\tau(\boldsymbol{b}))$ via the same recursive sampling procedure as in Algorithm 2. In the second part and third parts, we derive a corresponding regret upper bound and obtain the time and space complexities of our algorithm with bandit feedback. In the fourth part, we optimize the continuous regret w.r.t. the selection of $\mathcal{B}$. In the fifth part, we show how to extend our algorithm and results to the full information setting.

Part 1: Algorithm Correctness. In this part of the proof, we argue that our choice of estimator is unbiased and that Algorithm 3 samples bid vectors with the same probability that the exponential weights algorithm would have, given the same node utility estimates $\widehat{w}_m^t(b)$. To show unbiasedness of $\widehat{w}_m^t(b)$, we have:

$$
\mathbb{E}\left[\widehat{w}_m^t(b)\right] = \mathbb{E}\left[1 - \frac{1-w_m^t(b)}{q_m^t(b)} \mathbf{1}_{b_m^t=b}\right] = \mathbb{E}\left[1 - \frac{\mathbf{1}_{b_m^t=b}}{q_m^t(b)} + \frac{\mathbf{1}_{b_m^t=b} \cdot w_m^t(b)}{q_m^t(b)}\right] = w_m^t(b) \,.
$$

Now, it remains to show that our sampling procedure $\textsc{Sample}-\boldsymbol{b}$ w.r.t. $\widehat{S}_m^t$ indeed samples bid vectors $\boldsymbol{b}$ with the same probability as the exponential weights algorithm under weights $\widehat{\mu}_n^t(\boldsymbol{b})$. In particular, we want to show that our algorithm samples bid vectors $\boldsymbol{b}^t$ with probability proportional to $\exp(\eta \sum_{\tau=1}^{t-1} \widehat{\mu}_m^\tau(b_m))$ for any $m \in [M]$. This follows from analyzing the dynamic programming variables that represent the sum of exponentiated (estimated) partial bid vector utilities, $\widehat{S}$.

In exponential weights, the bidder selects at round $t+1$ some action $\boldsymbol{b}$ with probability $P^t(\boldsymbol{b})$ proportional to $\sum_{\tau=1}^{t} \widehat{\mu}_n^t(b)$. Using our representation of $\widehat{\mu}_n^t(\boldsymbol{b})$ as a function $\widehat{w}_m^\tau(b)$, we have:

$$P^t(\boldsymbol{b}) = \frac{\exp(\eta \sum_{\tau=1}^{t-1} \widehat{\mu}_n^\tau(\boldsymbol{b}))}{\sum_{\boldsymbol{b}' \in \mathcal{B}^{+M}} \exp(\eta \sum_{\tau=1}^{t-1} \widehat{\mu}_n^\tau(\boldsymbol{b}'))} = \frac{\exp(\eta \sum_{m=1}^{M} \widehat{W}_m^t(b_m))}{\sum_{\boldsymbol{b}' \in \mathcal{B}^{+M}} \exp(\eta \sum_{m=1}^{M} \widehat{W}_m^t(b'_m))}\,.$$

Hence, we wish to construct a sampler that samples $\boldsymbol{b}$ with the above probability. Defining $b_0 = \max_{b\in\mathcal{B}} b$, we begin by decomposing the denominator as follows:

$$\begin{aligned}
&\sum_{\boldsymbol{b}\in\mathcal{B}^{+M}} \exp(\eta \sum_{m=1}^{M} \widehat{W}_m^t(b_m)) = \sum_{b_1\in\mathcal{B}, b_1\le b_0} \; \sum_{b_2\in\mathcal{B}, b_2\le b_1} \cdots \sum_{b_M\in\mathcal{B}, b_M\le b_{M-1}} \exp(\eta \sum_{m=1}^{M} \widehat{W}_m^t(b_m)) \\
&= \sum_{b_1\in\mathcal{B}, b_1\le b_0} \exp(\eta \widehat{W}_1^t(b_1)) \sum_{b_2\in\mathcal{B}, b_2\le b_1} \exp(\eta \widehat{W}_2^t(b_2)) \cdots \sum_{b_M\in\mathcal{B}, b_M\le b_{M-1}} \exp(\eta \widehat{W}_M^t(b_M))\,.
\end{aligned}$$

Recall a key object $\widehat{S}_m^t(b)$, which is the sum of exponentially weighted utilities of partial bid vectors $\boldsymbol{b}'_{m:M} \in \mathcal{B}^{+(M-m+1)}$ over slots $m, \ldots, M$ subject to $b_m = b$.

$$\widehat{S}_m^t(b) = \exp(\eta \widehat{W}_m^t(b)) \sum_{\boldsymbol{b}'_{m+1:M} \in \mathcal{B}^{+(M-m)}, b'_{m+1} \le b'_m = b} \exp(\eta \sum_{m'=m+1}^{M} \widehat{W}_{m'}^t(b'_{m'})) \tag{8}$$

$$= \exp(\eta \widehat{W}_m^t(b)) \sum_{b'\in\mathcal{B}; b'\le b} \widehat{S}_{m+1}^t(b')\,. \tag{9}$$

With the trivial base case $\widehat{S}_M^t(b) = \exp(\eta \widehat{W}_M^t(b))$, we can recover all of the exponentially weighted partial utilities $\{\widehat{S}_m^t(b)\}_{m\in[M], b\in\mathcal{B}}$ given $\boldsymbol{W}^t$. Once we have computed $\{\widehat{S}_m^t(b)\}_{m\in[M], b\in\mathcal{B}}$, we can sample $\boldsymbol{b}$ according to its exponentially weighted utility $\exp(\eta \widehat{\mu}_m^t(\boldsymbol{b}))$ by sequentially sampling each $b_1, \ldots, b_M$.

Let $P_{\mathrm{D}}^t(\boldsymbol{b})$ be the probability that our Algorithm 3 returns bid vector $\boldsymbol{b} \in \mathcal{B}^{+M}$ in round t. Recall that we sample $\boldsymbol{b}$ by setting b_m^t to $b \in \mathcal{B}, b \le b_{m-1}^t$ with probability $\frac{\widehat{S}_m^t(b)}{\sum_{b'\le b_{m-1}^t} \widehat{S}_m^t(b')}$. Hence, the probability of selecting $\boldsymbol{b}$ is the product of m conditional probability mass functions (pmf's) and we have

$$P_{\mathrm{D}}^t(\boldsymbol{b}) = \prod_{m=1}^{M} \frac{\widehat{S}_m^t(b_m)}{\sum_{b'\le b_{m-1}} \widehat{S}_m^t(b')} = \left(\prod_{m=1}^{M-1} \frac{\exp(\eta \widehat{W}_m^t(b_m)) \sum_{b\le b_m} \widehat{S}_{m+1}^t(b)}{\sum_{b'\le b_{m-1}} \widehat{S}_m^t(b')}\right) \left(\frac{\exp(\eta \widehat{W}_M^t(b_M))}{\sum_{b'\le b_{M-1}} \widehat{S}_M^t(b')}\right).$$

Moving the $\exp(\eta \widehat{W}_m^t(b_m))$ outside of the product, we obtain:

$$\begin{aligned}
P_{\mathrm{D}}^t(\boldsymbol{b}) &= \left(\prod_{m=1}^{M-1} \exp(\eta \widehat{W}_m^t(b_m))\right) \left(\prod_{m=1}^{M-1} \frac{\sum_{b\le b_m} \widehat{S}_{m+1}^t(b)}{\sum_{b'\le b_{m-1}} \widehat{S}_m^t(b')}\right) \left(\frac{\exp(\eta \widehat{W}_M^t(b_M))}{\sum_{b'\le b_{M-1}} \widehat{S}_M^t(b')}\right) \\
&= \left(\prod_{m=1}^{M} \exp(\eta \widehat{W}_m^t(b_m))\right) \left(\frac{\sum_{b\le b_{M-1}} \widehat{S}_M^t(b)}{\sum_{b'\le b_0} \widehat{S}_1^t(b')}\right) \left(\frac{1}{\sum_{b'\le b_{M-1}} \widehat{S}_M^t(b')}\right) \\
&= \frac{\prod_{m=1}^{M} \exp(\eta \widehat{W}_m^t(b_m))}{\sum_{b'\le b_0} \widehat{S}_1^t(b')}\,.
\end{aligned}$$

We now rearrange the last expression to see that our algorithm samples $\boldsymbol{b}$ with the same probability as the exponential weights algorithm:

$$P_{\mathrm{D}}^t(\boldsymbol{b}) = \frac{\prod_{m=1}^{M} \exp(\eta \widehat{W}_m^t(b_m))}{\sum_{b\le b_0} S_1^t(b)} = \frac{\exp(\eta \sum_{m=1}^{M} \widehat{W}_m^t(b_m))}{\sum_{\boldsymbol{b}'\in\mathcal{B}^{+M}} \exp(\eta \sum_{m=1}^{M} \widehat{W}_m^t(b'_m))} = P^t(\boldsymbol{b})\,.$$

Part 2: Regret Analysis. We are now ready to derive the regret upper bound on Algorithm 3. First, we show that the bid vector utility estimators $\widehat{\mu}^t(\boldsymbol{b})$ are both unbiased and have a finite upper bound. To show the upper bound, we take expectation with respect to the bid vectors selected by our algorithm and observe that

$$\mathbb{E}[\widehat{\mu}^t(\boldsymbol{b})] = \sum_{m=1}^{M} \mathbb{E}[\widehat{w}_m^t(b_m)] = \sum_{m=1}^{M} \mathbb{E}\left[1 - \frac{1 - (v_m - b_m) 1_{b_m > b_{-m}^t}}{q_m^t(b_m)} 1_{b_m = b_m^t}\right]$$

As we are considering the expectation ex-post, keeping the b_{-m}^t's fixed, we have independence between the two indicator functions and we obtain:

$$= \sum_{m=1}^{M} \mathbb{E}\left[1 - \frac{1-w_m^t(b_m)}{q_m^t(b_m)} 1_{b_m=b_m^t}\right] = M - \sum_{m=1}^{M} \frac{1-w_m^t(b_m)}{q_m^t(b_m)} \mathbb{E}\left[1_{b_m=b_m^t}\right] = \mu^t(\boldsymbol{b}).$$

As for the finite upper bound, we have that $\widehat{\mu}^t(\boldsymbol{b}) = \sum_{m=1}^{M} \widehat{w}_m^t(b_m)$ is the sum over M node utility estimators, each of which is upper bounded by 1. Hence, $\widehat{\mu}^t(\boldsymbol{b}) \le M$ for all $\boldsymbol{b} \in \mathcal{B}^{+M}$. Now, we make the following claim:

Lemma 3. *Let $\widehat{\mu}^t(\boldsymbol{b}) = \sum_{m=1}^{M}(1 - \frac{1-(v_m-b_m)1_{b_m > b_{-m}^t}}{q_m^t(b_m)} 1_{b_m=b_m^t})$ be our bid vector utility estimate as discussed. Then, any algorithm which samples bid vectors $\boldsymbol{b}$ with probability proportional to $\exp(\eta \sum_{\tau=1}^{t-1} \widehat{\mu}^t(\boldsymbol{b}))$ at round t for $\eta \le \frac{1}{M}$ has regret upper bound*

$$\textsc{Regret}_{\mathcal{B}} \lesssim \eta^{-1} M \log|\mathcal{B}| + \eta \sum_{t=1}^{T} \sum_{\boldsymbol{b}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}) \mathbb{E}[(\sum_{m=1}^{M} \widehat{w}_m^t(b_m))^2]. \tag{10}$$

Proof of Lemma 3 We will closely follow the analysis of the Exp3 algorithm as presented in Chapter 11.4 of Lattimore and Szepesvári (2020). In particular, we follow their regret analysis until Equation 11.13. Define $\Phi^t = \sum_{\boldsymbol{b} \in \mathcal{B}^{+M}} \exp(\eta \sum_{\tau=1}^{t} \widehat{\mu}^\tau(\boldsymbol{b}))$ to be the *potential* at round t. As per our initial conditions in Algorithm 3, we have $\widehat{\mu}^0(\boldsymbol{b}) = 0$, and consequently, $\Phi^0 = |\mathcal{B}^{+M}|$. While it is not immediately apparent how the potentials Φ^t relate to the regret, we begin by upper bounding $\exp(\eta \sum_{t=1}^{T} \widehat{\mu}^t(\boldsymbol{b}))$ for a fixed $\boldsymbol{b}'$:

$$\exp(\eta \sum_{t=1}^{T} \widehat{\mu}^t(\boldsymbol{b}')) \le \sum_{\boldsymbol{b} \in \mathcal{B}^{+M}} \exp(\eta \sum_{t=1}^{T} \widehat{\mu}^t(\boldsymbol{b})) = \Phi^T = \Phi^0 \prod_{t=1}^{T} \frac{\Phi^t}{\Phi^{t-1}}. \tag{11}$$

Now, we upper bound each $\frac{\Phi^t}{\Phi^{t-1}}$:

$$\frac{\Phi^t}{\Phi^{t-1}} = \sum_{\boldsymbol{b} \in \mathcal{B}^{+M}} \frac{\exp(\eta \sum_{\tau=1}^{t} \widehat{\mu}^\tau(\boldsymbol{b}))}{\Phi^{t-1}} = \sum_{\boldsymbol{b} \in \mathcal{B}^{+M}} \frac{\exp(\eta \sum_{\tau=1}^{t-1} \widehat{\mu}^\tau(\boldsymbol{b}))}{\Phi^{t-1}} \exp(\eta \widehat{\mu}^t(\boldsymbol{b})) = \sum_{\boldsymbol{b} \in \mathcal{B}^{+M}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}) \exp(\eta \widehat{\mu}^t(\boldsymbol{b})),$$

where in the last equality, we used the condition that our algorithm samples bid vector $\boldsymbol{b}$ with probability proportional to $\exp(\eta \sum_{\tau=1}^{t-1} \widehat{\mu}^t(\boldsymbol{b}))$ at round t. In order to continue the chain of inequalities, we note that for $\eta \le \frac{1}{M}$, we have that the quantity $\eta \widehat{\mu}^t(\boldsymbol{b})$ is upper bounded by 1 as $\eta \widehat{\mu}^t(\boldsymbol{b}) \le \eta M \le 1$. In the first inequality, we used the fact that $\widehat{\mu}^t(\boldsymbol{b}) \le M$. Now, we can apply the inequalities $\exp(x) \le 1 + x + x^2$ and $1 + x \le \exp(x)$ for all $x \le 1$, with $x = \eta \widehat{\mu}^t(\boldsymbol{b})$ and $x = \eta \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}) \widehat{\mu}^t(\boldsymbol{b})$, respectively, to obtain:

$$\begin{aligned}
\frac{\Phi^t}{\Phi^{t-1}} &\le \sum_{\boldsymbol{b} \in \mathcal{B}^{+M}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}) \exp(\eta \widehat{\mu}^t(\boldsymbol{b})) \\
&\le \sum_{\boldsymbol{b} \in \mathcal{B}^{+M}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}) \left[1 + \eta \widehat{\mu}^t(\boldsymbol{b}) + \eta^2 \widehat{\mu}^t(\boldsymbol{b})^2\right] \\
&= 1 + \sum_{\boldsymbol{b} \in \mathcal{B}^{+M}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}) \left[\eta \widehat{\mu}^t(\boldsymbol{b}) + \eta^2 \widehat{\mu}^t(\boldsymbol{b})^2\right] \\
&\le \exp(\sum_{\boldsymbol{b} \in \mathcal{B}^{+M}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}) \left[\eta \widehat{\mu}^t(\boldsymbol{b}) + \eta^2 \widehat{\mu}^t(\boldsymbol{b})^2\right]).
\end{aligned}$$

Combining this with Equation (11) and then taking logarithms, we obtain:

$$\eta \sum_{t=1}^{T} \widehat{\mu}^t(\boldsymbol{b}') \le \log \Phi^0 + \eta \sum_{t=1}^{T} \sum_{\boldsymbol{b} \in \mathcal{B}^{+M}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}) \widehat{\mu}^t(\boldsymbol{b}) + \eta^2 \sum_{t=1}^{T} \sum_{\boldsymbol{b} \in \mathcal{B}^{+M}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}) \widehat{\mu}^t(\boldsymbol{b})^2.$$

Dividing both sides by η, applying the upper bound on Φ^0, and rearranging, we obtain that for any $\boldsymbol{b}' \in \mathcal{B}^{+M}$:

$$\sum_{t=1}^{T} \widehat{\mu}^t(\boldsymbol{b}') - \sum_{t=1}^{T} \sum_{\boldsymbol{b} \in \mathcal{B}^{+M}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}) \widehat{\mu}^t(\boldsymbol{b}) \lesssim \eta^{-1} M \log|\mathcal{B}| + \eta \sum_{t=1}^{T} \sum_{\boldsymbol{b} \in \mathcal{B}^{+M}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}) \widehat{\mu}^t(\boldsymbol{b})^2.$$

Replacing $\widehat{\mu}^t(\boldsymbol{b})$ with its definition in terms of $\widehat{w}_m^t(b_m)$ and taking expectations, we obtain the right hand side of the lemma:

$$\mathbb{E}\left[\sum_{t=1}^{T}\widehat{\mu}^t(\boldsymbol{b}') - \sum_{t=1}^{T}\sum_{\boldsymbol{b}\in\mathcal{B}^{+M}}\mathbb{P}(\boldsymbol{b}^t=\boldsymbol{b})\widehat{\mu}^t(\boldsymbol{b})\right] \lesssim \eta^{-1}M\log|\mathcal{B}| + \eta\sum_{t=1}^{T}\sum_{\boldsymbol{b}}\mathbb{P}(\boldsymbol{b}^t=\boldsymbol{b})\mathbb{E}[(\sum_{m=1}^{M}\widehat{w}_m^t(b_m))^2]\,.$$

Replacing $\sum_{\boldsymbol{b}\in\mathcal{B}^{+M}}\mathbb{P}(\boldsymbol{b}^t=\boldsymbol{b})\widehat{\mu}^t(\boldsymbol{b})$ with $\mathbb{E}[\widehat{\mu}^t(\boldsymbol{b}^t)]$ and recalling that the bid vector utility estimates $\widehat{\mu}^t$ were unbiased, we have:

$$\sum_{t=1}^{T}\mu^t(\boldsymbol{b}') - \sum_{t=1}^{T}\mathbb{E}[\mu^t(\boldsymbol{b}^t)] \lesssim \eta^{-1}M\log|\mathcal{B}| + \eta\sum_{t=1}^{T}\sum_{\boldsymbol{b}}\mathbb{P}(\boldsymbol{b}^t=\boldsymbol{b})\mathbb{E}[(\sum_{m=1}^{M}\widehat{w}_m^t(b_m))^2]\,.$$

Notice that as this is true for any $\boldsymbol{b}'$, we can replace it with the bid vector that maximizes the true cumulative utility $\sum_{t=1}^{T}\mu^t(\boldsymbol{b}')$ and see that the left hand side becomes precisely $\textsc{Regret}_{\mathcal{B}}$, which completes the proof. □

Now, it remains to show an upper bound on the second moment of the bid vector utility estimate $\mathbb{E}[(\sum_{m=1}^{M}\widehat{w}_m^t(b_m))^2]$. A crude attempt would be to say that $\mathbb{E}[(\sum_{m=1}^{M}\widehat{w}_m^t(b_m))^2] \le M\sum_{m=1}^{M}\mathbb{E}[\widehat{w}_m^t(b_m)^2]$:

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{\boldsymbol{b}}\mathbb{P}(\boldsymbol{b}^t=\boldsymbol{b})\mathbb{E}[(\sum_{m=1}^{M}\widehat{w}_m^t(b_m))^2] \le M\sum_{t=1}^{T}\sum_{\boldsymbol{b}}\mathbb{P}(\boldsymbol{b}^t=\boldsymbol{b})\sum_{m=1}^{M}\mathbb{E}[\widehat{w}_m^t(b_m)^2] \\
&= M\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{b\in\mathcal{B}}\mathbb{E}[\widehat{w}_m^t(b)^2]\sum_{\boldsymbol{b}:b_m=b}\mathbb{P}(\boldsymbol{b}^t=\boldsymbol{b}) = M\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{b\in\mathcal{B}}\mathbb{E}[\widehat{w}_m^t(b)^2]q_m^t(b) \\
&\le M\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{b\in\mathcal{B}}\frac{2}{q_m^t(b)}q_m^t(b) = O(M^2|\mathcal{B}|T)\,.
\end{aligned} \tag{12}$$

Where the last inequality follows from:

$$\mathbb{E}[\widehat{w}_m^t(b)^2] = \mathbb{E}\left[\left(1-\frac{1-w_m^t(b)}{q_m^t(b)}\mathbf{1}_{b_m^t=b}\right)^2\right] = 1-2\mathbb{E}\left[\frac{1-w_m^t(b)}{q_m^t(b)}\mathbf{1}_{b_m^t=b}\right] + \mathbb{E}\left[\left(\frac{1-w_m^t(b;\boldsymbol{v})}{q_m^t(b)}\right)^2\mathbf{1}_{b_m^t=b}\right].$$

Evaluating the expectations with $\mathbb{E}\left[\mathbf{1}_{b_m^t=b}\right] = q_m^t(b)$, we have:

$$\mathbb{E}[\widehat{w}_m^t(b)^2] = 1-\left[2-2w_m^t(b)\right] + \left[\frac{(1-w_m^t(b))^2}{q_m^t(b)}\right] = 2w_m^t(b) - 1 + \frac{1}{q_m^t(b)} \le 1 + \frac{1}{q_m^t(b)} \le \frac{2}{q_m^t(b)}\,.$$

Plugging this back into our upper bound yields stated regret bound for $\eta = \Theta(\sqrt{\frac{\log|\mathcal{B}|}{M|\mathcal{B}|T}})$ such that $\eta < \frac{1}{M}$:

$$\textsc{Regret}_{\mathcal{B}} \lesssim \eta^{-1}M\log|\mathcal{B}| + \eta M^2|\mathcal{B}|T = O(M^{\frac{3}{2}}\sqrt{|\mathcal{B}|T\log|\mathcal{B}|})\,.$$

Part 3: Complexity Analysis. We note that the time and space complexity analysis is identical to that of Algorithm 2, as the only additional computational work being done is computing the normalization terms $q_m^t(b)$, which requires $O(M|\mathcal{B}|)$ space and $O(MT|\mathcal{B}|)$ time respectively. Hence, discarding old tables, the total time and space complexities of Algorithm 3 are $O(M|\mathcal{B}|T)$ and $O(M|\mathcal{B}|)$ respectively. As this is polynomial in $M, |\mathcal{B}|, T$, we have proven the claim of polynomial space and time complexities.

Part 4: Selecting $\mathcal{B}$. We claim that the sub-optimality due to the discretization is on the order of $\frac{MT}{|\mathcal{B}|}$. Assume that $\mathcal{B} \equiv \{\frac{i}{|\mathcal{B}|}\}_{i\in[|\mathcal{B}|]}$ is an even discretization of $[0,1]$., and recall the continuous regret benchmark:

$$\textsc{Regret} = \max_{\boldsymbol{b}\in[0,1]^{+M}}\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b}) - \mathbb{E}\left[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b}^t)\right],$$

where the maximum is taken over the entire space $[0,1]^{+M}$ rather than $\mathcal{B}^{+M}$. Let $\boldsymbol{b}^*$ denote the maximizer of the continuous regret. Then, bidder n could have obtained at least the same allocation by rounding up each bid in $\boldsymbol{b}^*$ to the next largest multiple of $\frac{1}{|\mathcal{B}|}$. Let this rounded bid vector be denoted by $\boldsymbol{b}^+$. As their allocation, thus value for the set of items received, does not decrease, and their total payment increases by a maximum of $\frac{M}{|\mathcal{B}|}$ at each round, then we have that $\mu_n^t(\boldsymbol{b}^+) \geq \mu_n^t(\boldsymbol{b}^*) - \frac{M}{|\mathcal{B}|}$. Let $\boldsymbol{b}_{\mathcal{B}}^* \in \mathcal{B}^{+M}$ denote the hindsight optimal utility vector returned by our offline dynamic programming (Algorithm 1), which serves as the regret benchmark in the definition of discretized regret. Noting that $\boldsymbol{b}^+ \in \mathcal{B}^{+M}$, we have that the total utility of bidding $\boldsymbol{b}_{\mathcal{B}}^*$ must be at least that of $\boldsymbol{b}^+$. Thus,

$$\sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}_{\mathcal{B}}^*) \geq \sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}^+) \geq \sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}^*) - \frac{MT}{|\mathcal{B}|}$$

We balance this with the discretized regret $O(M^{\frac{3}{2}}\sqrt{|\mathcal{B}|T\log|\mathcal{B}|})$ with $|\mathcal{B}| = M^{-\frac{1}{3}}T^{\frac{1}{3}}$. This yields continuous regret $\textsc{Regret} = O(M^{\frac{4}{3}}T^{\frac{2}{3}}\sqrt{\log T})$.

Part 5: Extending to the Full Information Setting. Thus far, we have only discussed the bandit feedback algorithm. Fortunately, the full information setting algorithm is exactly the same except for two differences: 1) we do not need to compute $\boldsymbol{q}$ and 2) we can replace the reward estimates $\widehat{\mu}^t(\boldsymbol{b})$ with the true rewards $\mu^t(\boldsymbol{b})$ in Equation 12. The first difference can only serve to improve the time and space complexity of our algorithm. The second difference allows us to improve the bound on $\sum_{t=1}^{T}\sum_{\boldsymbol{b}}\mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b})\mathbb{E}[(\sum_{m=1}^{M}\widehat{w}_m^t(b_m))^2]$ in the left hand sight of Equation 12 by replacing $\widehat{w}_m^t(b_m))$ with $w_m^t(b_m)$:

$$\sum_{t=1}^{T}\sum_{\boldsymbol{b}}\mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b})\mathbb{E}[(\sum_{m=1}^{M}\widehat{w}_m^t(b_m))^2] = \sum_{t=1}^{T}\sum_{\boldsymbol{b}}\mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b})\mathbb{E}[(\sum_{m=1}^{M}\widehat{w}_m^t(b_m))^2] \leq M^2\sum_{t=1}^{T}\sum_{\boldsymbol{b}}\mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}) = M^2T$$

Notice that this bound is a factor of $|\mathcal{B}|$ improvement over that in the bandit setting. Consequently, we obtain our stated regret bound of $O(M^{\frac{3}{2}}\sqrt{T\log|\mathcal{B}|})$ with the choice of $\eta = \Theta(\sqrt{\frac{\log|\mathcal{B}|}{MT}})$. Balancing this regret with the error term, which is of order $O\left(\frac{M|\mathcal{B}|}{T}\right)$, the optimal choice of $|\mathcal{B}|$ is given by $\Theta(\sqrt{\frac{T}{M}})$. This yields corresponding continuous regret of $O(M^{\frac{3}{2}}\sqrt{T\log T})$.

9.3. Proof of Lemma 1

Proof of Lemma 1 In order to show equivalence, we show that (1) for any $\pi \in \Pi$, that $q(\pi) \in \mathcal{Q}$ and (2) for any $\boldsymbol{q} \in \mathcal{Q}$, there exists a $\pi \in \Pi$ such that $q(\pi) = \boldsymbol{q}$. We first prove (1). To do this, we simply need to check that for a given $\pi \in \Pi$, that $q^\pi = q(\pi)$ satisfies the constraints prescribed by $\mathcal{Q}$.

The non-negativity constraint holds trivially as each $\pi((m,b),b')$ is non-negative. Since all $q_1^\pi(b) = \pi((0,\max\mathcal{B}),b) \geq 0$ for all $b \in \mathcal{B}$, by induction, $q_{m+1}^\pi(b) = \sum_{b" \geq b} q_m^\pi(b")\pi((m,b"),b)$ is also non-negative.

Now we prove that each layer m sums to 1, i.e., $\sum_{b\in\mathcal{B}} q_m^\pi(b) = 1$. Since $\sum_{b\in\mathcal{B}} q_1^\pi(b)$, the policy has total node probability 1 in the first layer, we can prove $\sum_{b\in\mathcal{B}} q_m^\pi(b) = 1$, that the policy has total node probability 1 in the m'th layer, via induction. This follows immediately from the fact that the DP graph is layered, i.e., edges exist only from nodes in layer m to nodes in layer $m+1$, thus the only edges leading to layer $m+1$ are from layer m, in which there are no other edges. Hence, the total node probability in layer $m+1$ must be exactly that of layer m. More formally, we have:

$$\sum_{b\in\mathcal{B}} q_{m+1}^\pi(b) = \sum_{b\in\mathcal{B}}\sum_{b"\geq b} q_m^\pi(b")\pi((m,b"),b) = \sum_{b"\in\mathcal{B}} q_m^\pi(b")\sum_{b\leq b"}\pi((m,b"),b) = \sum_{b"\in\mathcal{B}} q_m^\pi(b")\,.$$

To show the stochastic domination constraint $\sum_{b\leq b'} q_{m+1}^\pi(b) \geq \sum_{b\leq b'} q_m^\pi(b)$, we use the bid monotonicity constraint; i.e., the fact that the edges between layers are only from larger bids to (weakly) smaller bids. Recall that $\pi((m,b'),b")$ is the probability of transitioning from unit-bid value pair (m,b') to $(m+1,b")$ and that the only edges leading to $(m+1,b")$ come from nodes (m,b') for $b' \geq b"$. Then, we have:

$$\sum_{b\leq b'} q_{m+1}^\pi(b) = \sum_{b\leq b'}\sum_{b"\geq b} q_m^\pi(b")\pi((m,b"),b)$$

$$
\begin{aligned}
&= \sum_{b">b'} q_m^\pi(b") \sum_{b\le b'} \pi((m,b"),b) + \sum_{b"\le b'} q_m^\pi(b") \sum_{b\le b"} \pi((m,b"),b) \\
&= \sum_{b">b'} q_m^\pi(b") \sum_{b\le b'} \pi((m,b"),b) + \sum_{b\le b'} q_m^\pi(b) \\
&\ge \sum_{b\le b'} q_m^\pi(b)\,.
\end{aligned}
$$

Hence, we have shown that for any $\pi \in \Pi$, that the corresponding $q(\pi) \in \mathcal{Q}$.

Now we show the other direction (2), that for any $\boldsymbol{q} \in \mathcal{Q}$, there exists a $\pi \in \Pi$ such that $q(\pi) = \boldsymbol{q}$. We proceed by showing that for all m, b^*, there exists $\{\pi((m,b),b')\}_{b,b'\in\mathcal{B}}$ such that the following conditions hold:

1. $\pi((m,b),b') \ge 0$ for all $b, b' \ge b^*$.
2. $\pi((m,b),b') = 0$ for all $b' > b \ge b^*$.
3. $\sum_{b'\le b, b'\ge b^*} \pi((m,b),b') \le 1$ for all $b^* \in \mathcal{B}$, with equality if and only if $b^* = b_{\min}$ where $b_{\min} = \min \mathcal{B}$.
4. $\sum_{b'\ge b^*} \sum_{b\ge b'} q_m(b)\pi((m,b),b') = \sum_{b'\ge b^*} q_{m+1}(b')$.

Let $\Pi(b^*;\mathbf{q})$, $b^* \in \mathcal{B}$, be the set of all policies under which the four conditions hold at b^* and $\mathbf{q} \in \mathcal{Q}$.

These conditions trivially hold for $m = 0$, as we can set $\pi((0, \max\mathcal{B}), b) = q_1(b)$ and $\pi((0,b),b') = \mathbf{1}_{b=b'}$. To solve for general m, we must show that there exists $\{\pi((m,b),b')\}_{b,b'\in\mathcal{B}}$ that satisfies the constraints prescribed by Π and that $\sum_{b\ge b'} q_m(b)\pi((m,b),b') = q_{m+1}(b')$ for all $b' \in \mathcal{B}$. In order to do this, we show that conditions (1), (2), (3), and (4) for each $b^* \in \mathcal{B}$. In particular, if we show conditions (1) and (2) for $b^* = b_{\min}$, then we have already satisfied the first two conditions of Π. If we show that (3) holds for $b^* = b_{\min}$, then by condition (1), then (3) holds for all $b^* \in \mathcal{B}$ as well, as the summation only includes fewer terms as b^* increases. Similarly, if we show condition (4) holds for two adjacent values of $b^*_- < b^*$, then we have that $\sum_{b\ge b'\ge b^*} q_m(b)\pi((m,b),b') = q_{m+1}(b^*)$. Thus, if condition (4) holds for all possible pairs of adjacent bid values, then we have that $\sum_{b\ge b'} q_m(b)\pi((m,b),b') = q_{m+1}(b')$ for all b'. These observations suggest use of induction over b^*, and indeed, we begin by showing that these conditions hold for $b^* = b_{\min}$. We then show that this implies that the conditions hold for the next smallest value of b^*, which would complete the induction proof.

Base Case: Recall $b^* = b_{\min}$. We now show that there exists $\{\pi((m,b),b')\}_{b,b'\in\mathcal{B}}$ satisfying all four conditions. For any $m \in [M]$, let we set $\pi((m,b),b') = \mathbf{1}_{b=b'}$. Then, condition (4) is clearly satisfied:

$$
\sum_{b'\ge b^*} \sum_{b\ge b'} q_m(b)\pi((m,b),b') = \sum_{b'\ge b^*} q_{m+1}(b') \leftrightarrow \sum_{b\in\mathcal{B}} q_m(b) \sum_{b'\le b} \pi((m,b),b') = \sum_{b\in\mathcal{B}} q_{m+1}(b) = 1\,.
$$

It is also easy to check that conditions (1)-(3) are also satisfied when we set $\pi((m,b),b') = \mathbf{1}_{b=b'}$ for any m. This shows that $\Pi(b^*;\mathbf{q})$ is non-empty, as desired.

Recursive Case: For any $b \in \mathcal{B}$, let b_- be the largest $b' \in \mathcal{B}$, which is strictly smaller than b. Here, we assume that $\Pi(b^*_-;\mathbf{q})$ is not empty, and under this assumption, we show that set $\Pi(b^*;\mathbf{q})$ is not empty, where $\Pi(b^*;\mathbf{q}) \subseteq \Pi(b^*_-;\mathbf{q})$. Let us start with condition (4). We would like to show that there exists a $\boldsymbol{\pi}$ that satisfies condition (4) at b^* along with the other three conditions. By the induction assumption, we have

$$
\begin{aligned}
&\sum_{b'\ge b^*_-} \sum_{b\ge b'} q_m(b)\pi((m,b),b') = \sum_{b'\ge b^*_-} q_{m+1}(b') \rightarrow \\
&\sum_{b'\ge b^*} \sum_{b\ge b'} q_m(b)\pi((m,b),b') + \sum_{b\ge b^*_-} q_m(b)\pi((m,b),b^*_-) = \sum_{b'\ge b^*} q_{m+1}(b') + q_{m+1}(b^*_-) \rightarrow \\
&\sum_{b'\ge b^*} \sum_{b\ge b'} q_m(b)\pi((m,b),b') = \sum_{b'\ge b^*} q_{m+1}(b') + \left[q_{m+1}(b^*_-) - \sum_{b\ge b^*_-} q_m(b)\pi((m,b),b^*_-) \right] \rightarrow \\
&\sum_{b\ge b^*} q_m(b) \sum_{b'\le b; b'\ge b^*} \pi((m,b),b') = \sum_{b'\ge b^*} q_{m+1}(b') + \left[q_{m+1}(b^*_-) - \sum_{b\ge b^*_-} q_m(b)\pi((m,b),b^*_-) \right]
\end{aligned}
$$

Thus, we can satisfy condition (4) if $q_{m+1}(b^*_-) = \sum_{b \geq b^*_-} q_m(b)\pi((m,b),b^*_-)$. We now observe that the latter summation depends linearly (and hence, continuously) in the values of $\pi((m,b),b^*_-)$. If we can show that there exists an assignment of these variables that satisfy $q_{m+1}(b^*_-) \geq \sum_{b \geq b^*_-} q_m(b)\pi((m,b),b^*_-)$ and also $q_{m+1}(b^*_-) \leq \sum_{b \geq b^*_-} q_m(b)\pi((m,b),b^*_-)$, then by the intermediate value theorem, there must be some assignment that achieves exact equality.

In order to show the first inequality, notice that if we set $\pi((m,b),b^*_-) = 1 - \sum_{b' < b^*_-} \pi((m,b),b')$ for all $b \geq b^*_-$ (this is required in order to guarantee conditions (1) and (3) are satisfied), then:

$$\begin{aligned}
\sum_{b \geq b^*_-} q_m(b)\pi((m,b),b^*_-) &= \sum_{b \geq b^*_-} q_m(b) - \sum_{b \geq b^*_-} q_m(b) \sum_{b' < b^*_-} \pi((m,b),b') \\
&= \sum_{b \geq b^*_-} q_m(b) - \sum_{b \in \mathcal{B}} q_m(b) \sum_{b' < b^*_-} \pi((m,b),b') + \sum_{b < b^*_-} q_m(b) \sum_{b' < b^*_-} \pi((m,b),b') \\
&= \sum_{b \geq b^*_-} q_m(b) - \sum_{b' < b^*_-} q_{m+1}(b) + \sum_{b < b^*_-} q_m(b) \sum_{b' < b^*_-} \pi((m,b),b') \\
&= \sum_{b \geq b^*_-} q_m(b) - \sum_{b' < b^*_-} q_{m+1}(b) + \sum_{b < b^*_-} q_m(b) \\
&\geq \sum_{b \geq b^*_-} q_{m+1}(b) - \sum_{b' < b^*_-} q_{m+1}(b) + \sum_{b < b^*_-} q_{m+1}(b) \\
&= \sum_{b \geq b^*_-} q_{m+1}(b) \\
&\geq q_{m+1}(b^*_-)\,.
\end{aligned}$$

Here, the third equality follows from the (strong) inductive hypothesis, and the first inequality is a result of the stochastic domination constraint in $\mathcal{Q}$. We also note that the values $\sum_{b' < b^*_-} \pi((m,b),b')$ have already been fixed as these were required to satisfy condition (4) in the previous iterates, and as condition (3) holds for b^*_- by the inductive hypothesis, then $1 - \sum_{b' < b^*_-} \pi((m,b),b') \geq 0$. Conversely, if we set $\pi((m,b),b^*_-) = 0$ for all $b \geq b^*_-$, then:

$$\sum_{b \geq b^*_-} q_m(b)\pi((m,b),b^*_-) = 0 \leq q_{m+1}(b^*_-)\,.$$

As the sum $\sum_{b \geq b^*_-} q_m(b)\pi((m,b),b^*_-)$ linearly (thus, continuously) depends on the values of $\pi((m,b),b^*_-)$, by the intermediate value theorem, there exists an assignment of $\{\pi((m,b),b^*_-)\}_{b \geq b^*_-}$ with each $\pi((m,b),b^*_-) \in [0, 1 - \sum_{b' < b^*_-} \pi((m,b),b')]$ such that the sum is precisely equal to $q_{m+1}(b^*_-) \in [0,1]$. Now we observe that these values of $\pi((m,b),b^*_-) \in [0, 1 - \sum_{b' < b^*_-} \pi((m,b),b')]$ do not violate conditions (1), (2), or (3). Furthermore, note that any $\boldsymbol{\pi} \in \Pi$ also satisfied conditions (1), (2), and (3) under b^*_- for $\{\pi((m,b),b')\}_{b \geq b^*_-, b' \leq b^*_-}$, then the assignment to $\{\pi((m,b),b')\}_{b \geq b^*_-, b' < b^*_-}$ will not violate these conditions as our new constraint on the variables $\{\pi((m,b),b^*_-)\}_{b \geq b^*_-}$ is independent of the values of $\{\pi((m,b),b')\}_{b \geq b^*_-, b' < b^*_-}$. Thus, the set $\Pi(b^*)$ is non-empty:

$$\Pi(b^*) = \{\{\pi((m,b),b')\}_{b,b' \in \mathcal{B}} \in \Pi(b^*_-) : \sum_{b \geq b^*_-} q_m(b)\pi((m,b),b^*_-) = q_{m+1}(b^*_-)\} \neq \emptyset$$

With this, we have proven via induction that our four conditions hold for all $b^* \in \mathcal{B}$, implying that for a fixed m, every constraint in Π pertaining to variables $\pi((m,b),b')$ is satisfied, as well as the node-measure constraints $\sum_{b \geq b'} q_m(b)\pi((m,b),b') = q_{m+1}(b')$ for all b'. By induction, this works for all $m \in [M]$, which concludes the proof.

9.4. Proof of Lemma 2

We have by the definition of discretized regret:

$$\textsc{Regret}_{\mathcal{B}} = \max_{\boldsymbol{b} \in \mathcal{B}} \sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}) - \mathbb{E}\left[\sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}^t)\right] = \max_{\boldsymbol{q} \in \mathcal{Q}} \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{q}, \boldsymbol{w}^t \rangle - \sum_{t=1}^{T} \langle \boldsymbol{q}^t, \boldsymbol{w}^t \rangle\right],$$

where in the first equality, we applied Equation (7) which equated the dot product of utilities $\boldsymbol{w}^t$ and node probability weights $\boldsymbol{q}$ to the expected utility of bid vector $\boldsymbol{b} \sim \boldsymbol{\pi}$ with utilities $\{w_m^t(b)\}_{m \in [M], b \in \mathcal{B}} = \boldsymbol{w}^t$. Combining the two summations yields the desired result.

9.5. Proof of Theorem 6: Online Mirror Descent Algorithm

Proof of Theorem 6: Online Mirror Descent Algorithm The proof is divided into four parts, similar to the analysis of Algorithm 2. In the first part, we rigorously show how our algorithm achieves the stated regret. In the second, we verify correctness of our procedure that recovers a policy $\boldsymbol{\pi}^t$ from $\boldsymbol{q}^t$. Then, we show the corresponding time and space complexity of our algorithm. Afterwards, we optimize over discretization error to obtain the continuous regret.

Part 1: Regret of Online Linear Optimization. Recall that from Lemma 2, we have

$$\textsc{Regret}_{\mathcal{B}} = \max_{\boldsymbol{q}\in\mathcal{Q}} \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{q} - \boldsymbol{q}^t, \boldsymbol{w}^t \rangle\right] = \max_{\boldsymbol{q}\in\mathcal{Q}} \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{q}^t - \boldsymbol{q}, -\boldsymbol{w}^t \rangle\right], \tag{13}$$

where we negate the utility function into a loss function to be consistent with the OLO convention. We follow a standard analysis of OMD, which shows that the optimization step can be solved efficiently and the resulting iterates have bounded regret. For the former, we show that solution to the $\boldsymbol{q}$ optimization step in our algorithm $\boldsymbol{q}^t = \operatorname{argmin}_{\boldsymbol{q}\in\mathcal{Q}} \eta\langle \boldsymbol{q}, -\boldsymbol{w}^t\rangle + D(\boldsymbol{q}||\boldsymbol{q}^{t-1})$ can be obtained as the projection of the unconstrained minimizer of

$$\tilde{q}^t = \operatorname{argmin}_{\boldsymbol{q}\in[0,1]^{M\times|\mathcal{B}|}} \eta\langle \boldsymbol{q}, -\boldsymbol{w}^t\rangle + D(\boldsymbol{q}||\boldsymbol{q}^{t-1})$$

to the space $\mathcal{Q}$ (See Projection Lemma, Lemma 8.6 of Bartok et al. (2011)). Having characterized the exact form of the OMD iterates, all that remains is to upper bound the regret of OMD with the regret of Be-the-regularized-leader.

Lemma 4 **(Lemma 9.2 of Bartok et al. (2011))**. *Letting $D(\boldsymbol{q}||\boldsymbol{q}')$ denote the unnormalized KL divergence between $\boldsymbol{q}$ and $\boldsymbol{q}'$, we have:*

$$\textsc{Regret}_{\mathcal{B}} \le \max_{\boldsymbol{q}\in\mathcal{Q}} \mathbb{E}\Big[\eta^{-1} D(\boldsymbol{q}||\boldsymbol{q}^1) + \sum_{t=1}^{T} \langle \boldsymbol{q}^t - \tilde{\boldsymbol{q}}^{t+1}, \boldsymbol{w}^t\rangle\Big].$$

The remainder of the regret analysis closely follows that of Theorem 1 in Zimin and Neu (2013). At a high level, we want to bound the regret of Online Mirror Descent by the regret of the unconstrained Be the (Negentropy) Regularized leader, via Lemma 4 (see Lemma 13 of Rakhlin (2009) for the more general statement and proof of this lemma). We then upper the contribution of the summation term by using the specific definition of the node weight estimators. Similarly, we upper bound the divergence term as a function of the dimension of the space $\mathcal{Q}$.

To begin, note that our node utility estimators $\widehat{w}_m^t(b)$ are unbiased:

$$\mathbb{E}_{\boldsymbol{b}\sim\boldsymbol{\pi}^t}[\widehat{w}_m^t(b)] = \mathbb{E}_{\boldsymbol{b}\sim\boldsymbol{\pi}^t}[\frac{w_m^t(b)}{q_m^t(b)} \mathbf{1}_{b=b_m^t}] = \frac{w_m^t(b)}{q_m^t(b)} \mathbb{P}_{\boldsymbol{b}\sim\boldsymbol{\pi}^t}(b=b_m^t) = \frac{w_m^t(b)}{q_m^t(b)} q_m^{t-1}(b) = w_m^t(b). \tag{14}$$

Now, consider the right hand side of the inequality in Lemma 4. As the node utility estimators are unbiased, so we can replace $\boldsymbol{w}^t$ with $\widehat{\boldsymbol{w}}^t$. Now, as per Lemma 4, we can upper bound the expected estimated regret as a function of the unconstrained optimizer $\tilde{\boldsymbol{q}}^{t+1}$ and the unregularized relative entropy with respect to the initial state-edge occupancy measure $\boldsymbol{q}^1$. Applying the aforementioned lemma to Equation (14), we obtain:

$$\textsc{Regret}_{\mathcal{B}} = \max_{\boldsymbol{q}\in\mathcal{Q}} \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{q}^t - \boldsymbol{q}, -\widehat{\boldsymbol{w}}^t\rangle\right] \le \max_{\boldsymbol{q}\in\mathcal{Q}} \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{q}^t - \tilde{\boldsymbol{q}}^{t+1}, -\widehat{\boldsymbol{w}}^t\rangle + \eta^{-1} D(\boldsymbol{q}||\boldsymbol{q}^1)\right] \tag{15}$$

Applying $\exp(x) \ge 1+x$ for $x = \exp(\eta\widehat{\boldsymbol{w}}^t)$, we obtain $\tilde{\boldsymbol{q}}^{t+1} = \boldsymbol{q}^t \exp(\eta\widehat{\boldsymbol{w}}^t) \ge \boldsymbol{q}^t + \eta\boldsymbol{q}^t\widehat{\boldsymbol{w}}^t$, which yields $\boldsymbol{q}^t - \boldsymbol{q}^t\exp(\eta\widehat{\boldsymbol{w}}^t) \ge -\eta\boldsymbol{q}^t\widehat{\boldsymbol{w}}^t$. Plugging this back in:

$$\textsc{Regret}_{\mathcal{B}} \le \max_{\boldsymbol{q}\in\mathcal{Q}} \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{q}^t - \boldsymbol{q}^t\exp(\eta\widehat{\boldsymbol{w}}^t), -\widehat{\boldsymbol{w}}^t\rangle + \eta^{-1} D(\boldsymbol{q}||\boldsymbol{q}^1)\right] \tag{16}$$

$$\le \max_{\boldsymbol{q}\in\mathcal{Q}} \mathbb{E}\left[\eta \sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{b\in\mathcal{B}} q_m^t(b)\widehat{w}_m^t(b)^2 + \eta^{-1} D(\boldsymbol{q}||\boldsymbol{q}^1)\right]. \tag{17}$$

Note that $\widehat{w}_m^t(b) = \frac{w_m^t(b)}{q_m^{t-1}(b)} \mathbf{1}_{b=b_m^t}$ for all $m \in [M]$ and $b \in \mathcal{B}$ by definition. Since $w_m^t(b) \le 1$ and $\mathbf{1}_{b=b_m^t} \le 1$ we have $\widehat{w}_m^t(b) \le \frac{1}{q_m^t(b)}$ and we continue the above chain of inequalities with:

$$\begin{aligned}
\text{REGRET}_{\mathcal{B}} &\le \max_{\boldsymbol{q} \in \mathcal{Q}} \mathbb{E}\left[\eta \sum_{t=1}^T \sum_{m=1}^M \sum_{b \in \mathcal{B}} q_m^t(b) \widehat{w}_m^t(b) \frac{1}{q_m^t(b)} + \eta^{-1} D(\boldsymbol{q} || q^1)\right] && (18)\\
&= \max_{\boldsymbol{q} \in \mathcal{Q}} \mathbb{E}\left[\eta \sum_{t=1}^T \sum_{m=1}^M \sum_{b \in \mathcal{B}} \widehat{w}_m^t(b) + \eta^{-1} D(\boldsymbol{q} || \boldsymbol{q}^1)\right]. && (19)
\end{aligned}$$

Recalling that $D(\boldsymbol{q}||\boldsymbol{q}^1) = \sum_{m \in [M], b \in \mathcal{B}} q_m(b) \log \frac{q_m(b)}{q_m^1(b)} - (q_m(b) - q_m^1(b))$, we note that:

$$\begin{aligned}
D(\boldsymbol{q}||\boldsymbol{q}^1) &= \sum_{m=1}^M \sum_{b \in \mathcal{B}} q_m(b) \frac{\log q_m(b)}{\log q_m^1(b)} - q_m(b) + q_m^1(b)\\
&= \sum_{m=1}^M \sum_{b \in \mathcal{B}} q_m(b) \log q_m(b) - q_m(b) \log q_m^1(b),
\end{aligned}$$

where in the second equality, we used the fact that the elements both $\boldsymbol{q}$ and $\boldsymbol{q}^1$ all sum to M. Selecting $\boldsymbol{q}_m^1(\cdot)$ to be the uniform distribution over all $b \in \mathcal{B}$ and using the fact that the entropy of a discrete distribution over $|\mathcal{B}|$ items is $\log |\mathcal{B}|$, we obtain:

$$\begin{aligned}
D(\boldsymbol{q}||\boldsymbol{q}^1) &= -\sum_{m=1}^M H(\boldsymbol{q}_m) + \log |\mathcal{B}| \sum_{m=1}^M \sum_{b \in \mathcal{B}} q_m(b)\\
&\le \sum_{m=1}^M \log |\mathcal{B}| + \log |\mathcal{B}| \sum_{m=1}^M \sum_{b \in \mathcal{B}} q_m(b) = \Theta(M \log |\mathcal{B}|),
\end{aligned}$$

where $H(\boldsymbol{x}) = -\sum_{x \in \boldsymbol{x}} x \log x$ denotes the discrete entropy function. Plugging this back in:

$$\begin{aligned}
\text{REGRET}_{\mathcal{B}} &\lesssim \mathbb{E}\left[\eta \sum_{t=1}^T \sum_{m=1}^M \sum_{b \in \mathcal{B}} \widehat{w}_m^t(b) + \eta^{-1} M \log |\mathcal{B}|\right] && (20)\\
&\le \mathbb{E}\left[\eta \sum_{t=1}^T \sum_{m=1} \sum_{b \in \mathcal{B}} \widehat{w}_m^t(b) + \eta^{-1} M \log |\mathcal{B}|\right] && (21)\\
&\le \eta \sum_{t=1}^T \sum_{m=1} \sum_{b \in \mathcal{B}} w_m^t(b) + \eta^{-1} M \log |\mathcal{B}| && (22)\\
&= \eta T M |\mathcal{B}| + \eta^{-1} M \log |\mathcal{B}|, && (23)
\end{aligned}$$

where in the last equality, we used the unbiasedness property of $\widehat{\boldsymbol{w}}^t$. Setting $\eta = \sqrt{\frac{\log |\mathcal{B}|}{|\mathcal{B}| T}}$, we obtain $\text{REGRET}_{\mathcal{B}}(T) \le M \sqrt{|\mathcal{B}| T \log |\mathcal{B}|}$.

Part 2: Determining Policy $\boldsymbol{\pi}$ from Node Probability Measures $\boldsymbol{q}$. Notice that in our regret analysis for both the bandit and full information setting, we do not require explicit knowledge of the policy $\boldsymbol{\pi}^t$, so long as it generates the desired node occupancy measure $\boldsymbol{q}^t$. In particular, we require a method of converting $\boldsymbol{q}^t$ to policy $\boldsymbol{\pi}^t$ which, in turn, is required in order to sample $\boldsymbol{b}^t$. Recall from Lemma 1 that the mapping from the space of policies Π to the space of node weight measures $\mathcal{Q}_\Pi = \mathcal{Q}$ is injective. Thus, for any $\boldsymbol{q} \in \mathcal{Q}$, there must exists a $\boldsymbol{\pi} \in \Pi$ such that $q(\boldsymbol{\pi}) = \boldsymbol{q}$. Moreover, the set $\Pi(\boldsymbol{q})$ of such $\boldsymbol{\pi}$ can be written as the intersection of two polyhedrons, and hence a polyhedron, from which a feasible solution can be computed efficiently (e.g., ellipsoid method), where $\Pi(\boldsymbol{q})$ is the set of policies $\pi \in [0,1]^{M \times |\mathcal{B}| \times |\mathcal{B}|}$ such that

- $\pi((0, \max \mathcal{B}), b) = q_1(b)$, for any $b \in \mathcal{B}$;
- $\pi((0, b), b') = \mathbf{1}_{b=b'}$ for any $b, b' < \max \mathcal{B}$;
- $q_{m+1}(b') = \sum_{b \in \mathcal{B}} q_m(b) \pi((m, b), b')\}$ for any $b' \in \mathcal{B}$ and $m \in [M-1]$.

Part 3: Complexity analysis. One may wonder how to efficiently update the state occupancy measures by computing the minimizer of $\eta\langle \boldsymbol{q}, -\widehat{\boldsymbol{w}}^t\rangle + D(\boldsymbol{q}||\boldsymbol{q}^{t-1})$. The idea is to first solve the unconstrained entropy regularized minimizer with $\tilde{\boldsymbol{q}}^{t+1} = \boldsymbol{q}^t \exp(\eta \widehat{\boldsymbol{w}}^t)$. We then project this unconstrained minimizer to $\mathcal{Q}$ with:

$$\boldsymbol{q}^{t+1} = \operatorname{argmin}_{\boldsymbol{q}\in\mathcal{Q}} D(\boldsymbol{q}||\tilde{\boldsymbol{q}}^{t+1}) \tag{24}$$

Relegating the details to Zimin and Neu (2013), the above constrained optimization problem can be solved as the minimizer of an equivalent unconstrained convex optimization problem with a polynomial (in M and $|\mathcal{B}|$) number of variables, and therefore, can be computed efficiently. Combining with finding an initial feasible solution to $\Pi(\boldsymbol{q})$ as well as the optimization step, we achieve polynomial in $M, |\mathcal{B}|, T$ total time complexity. For the space complexity, we only need store the values of $\boldsymbol{\pi}^t$, $\boldsymbol{q}^t$, and $\widehat{\boldsymbol{w}}^t$, for a total space complexity of $O(M|\mathcal{B}|^2)$.

Part 4: Continuous Regret. To obtain the continuous regret, recall that the discretization error is $O(\frac{MT}{|\mathcal{B}|})$. As the discretized regret is $O\left(M\sqrt{|\mathcal{B}|T\log|\mathcal{B}|}\right)$ in the bandit feedback setting, the optimal choice of $|\mathcal{B}|$ is $\Theta(T^{\frac{1}{3}})$, which achieves continuous regret $\textsc{Regret} = O(MT^{\frac{2}{3}}\sqrt{\log T})$.

9.5.1. Proof of Corollary 1

We can straightforwardly extend Algorithm 4 to the full information setting. To do this, we note that we can improve Equation (18) by instead replacing $\widehat{\boldsymbol{w}}^t$ with $\boldsymbol{w}^t$ in Equation (17) to obtain:

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{b\in\mathcal{B}} q_m^t(b)\widehat{w}_m^t(b)^2 = \sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{b\in\mathcal{B}} q_m^t(b) w_m^t(b)^2 \le \sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{b\in\mathcal{B}} q_m^t(b) = \sum_{t=1}^{T}\sum_{m=1}^{M} 1 = TM\,.$$

Setting $\eta = \sqrt{\frac{\log|\mathcal{B}|}{T}}$, we obtain in the full information setting $\textsc{Regret}_{\mathcal{B}} = O(M\sqrt{T\log|\mathcal{B}|})$. We can also compute the optimal choice of $|\mathcal{B}|$ to obtain optimal continuous regret. Using the optimal choice of $|\mathcal{B}|$ being $\Theta(\sqrt{T})$, we achieve continuous regret of $\textsc{Regret} = O(M\sqrt{T\log T})$. Note that due to the complexity of the optimization sub-routine in the projection step of OMD, for the full information setting, it is preferable to use Algorithm 2 instead.

9.6. Proof of Theorem 7: Regret Lower Bound

To construct our lower bounds, we construct a stochastic adversary whose distribution across their bids makes it difficult for the bidder to determine their optimal bid, and thus, occurs $\Omega(M\sqrt{T})$ regret while doing so. We define $\boldsymbol{b}'_- = (0,\dots,0,c,\dots,c)$, where there are k and $M-k$ values of 0 and c each. We additionally define $\boldsymbol{b}"_- = (c,\dots,c)$ as the M-vector of bids at c. Restricting the adversary's bid vectors to be in $\{\boldsymbol{b}'_-, \boldsymbol{b}"_-\}$, we construct two adversary bid vector distributions F and G over $\{\boldsymbol{b}'_-, \boldsymbol{b}"_-\}^T$ such that under F, we have $\mathbb{P}(\boldsymbol{b}^t_- = \boldsymbol{b}'_-) = \frac{1}{2}+\delta$ and $\mathbb{P}(\boldsymbol{b}^t_- = \boldsymbol{b}"_-) = \frac{1}{2}-\delta$ and under G, we have $\mathbb{P}(\boldsymbol{b}^t_- = \boldsymbol{b}'_-) = \frac{1}{2}-\delta$ and $\mathbb{P}(\boldsymbol{b}^t_- = \boldsymbol{b}"_-) = \frac{1}{2}+\delta$ for some $\delta\in[0,\frac{1}{2}]$ to be optimized over later.

Assume that $\boldsymbol{v} = (1,\dots,1)$, all tiebreaks are won for simplicity, and the competitors' bids over time are independent. Then, for certain choices of c and k (which we show below), the expected utility maximizing bid vector under $\{\boldsymbol{b}^t_-\}_{t\in[T]} \sim F$ is $(0,\dots,0)$ and under $\{\boldsymbol{b}^t_-\}_{t\in[T]} \sim G$ is $(c,\dots,c)$. In particular, we can compute precisely the expected value of bidding $\boldsymbol{b}^t = \boldsymbol{b}$ for all $t\in[T]$ under both F and G. Note that as adversary bid values only take values in $\{0,c\}$ and bidder n wins all tiebreaks, then the bidder only need consider bid vectors consisting only of all 0 or c. Letting m denote the number of bids in $\boldsymbol{b}$ equal to c, we have:

$$\mathbb{E}_F\left[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b})\right] = T\left[(\frac{1}{2}+\delta)\left((1-c)m + \max(0, M-k-m)\right) + (\frac{1}{2}-\delta)(1-c)m\right].$$

Where $\mathbb{E}_F$ denotes the expectation with respect to the adversary bids drawn from F, namely $\{\boldsymbol{b}^t_-\}_{t\in[T]} \sim F$ (and similarly for $\mathbb{E}_G$ below). In particular, we have that with probability $\frac{1}{2}+\delta$, the adversary will select bid $\boldsymbol{b}'_-$. We are then guaranteed to win m units at a price of c, for a utility of $1-c$ per unit. If $m<k$, then $M-k-m$ of the items

were won with price 0, for a utility of 1 per unit. With probability $\frac{1}{2}-\delta$, all of the adversary bids are c, and we obtain m units at a cost of c each, which corresponds to utility $1-c$. Similarly, we have:

$$\mathbb{E}_G\left[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b})\right]=T\left[(\frac{1}{2}-\delta)\left((1-c)m+\max(0,M-k-m)\right)+(\frac{1}{2}+\delta)(1-c)m\right]$$

For the case $m+k\leq M$, we have that $(1-c)m+\max(0,M-k-m)=M-k-mc$, and the above two equations simplify to:

$$\mathbb{E}_F[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b})]=T\left[(\frac{1}{2}+\delta)(M-k)+m(\frac{1}{2}-\delta-c)\right];$$
$$\mathbb{E}_G[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b})]=T\left[(\frac{1}{2}-\delta)(M-k)+m(\frac{1}{2}+\delta-c)\right].$$

In the case where $m+k\geq M$, we have that $(1-c)m+\max(0,M-k-m)=m-mc$ and we obtain:

$$\mathbb{E}_F\left[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b})\right]=\mathbb{E}_G\left[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b})\right]=T(1-c)m.$$

Note that in either case, in the case where we sample $\{\boldsymbol{b}_-^t\}_{t\in[T]}$ according to the mixture $\frac{F+G}{2}$, this corresponds to the case where $\delta=0$, i.e., the probability of observing either $\boldsymbol{b}'_-$ or $\boldsymbol{b}"_-$ is equal. We have for all $\boldsymbol{b}$:

$$\mathbb{E}_{(F+G)/2}[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b})]=\frac{1}{2}((1-c)m+\max(0,M-k-m))+\frac{1}{2}(1-c)m\geq(1-c)m$$

Note that under F, the optimal occurs at the all 0's vector for $c>\frac{1}{2}-\delta$ and $(\frac{1}{2}+\delta)(M-k)>(1-c)m=0$. Similarly, the optimal occurs at the all c's vector for $c>\frac{1}{2}-\delta$ and $(\frac{1}{2}-\delta)(M-k)>(1-c)M$. These obtain utilities of $(\frac{1}{2}+\delta)(M-k)$ and $M-Mc$ respectively. One choice of c and k is $\frac{2}{3}$ and $\frac{M}{3}$, with $0<\delta<\frac{1}{6}$. Looking at the regret incurred each step of the algorithm by selecting any action $\boldsymbol{b}$, we have:

$$\begin{aligned}
&\max_{\boldsymbol{b}'}\left(\mathbb{E}_F[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b}')-\mu_n^t(\boldsymbol{b})]\right)+\max_{\boldsymbol{b}'}\left(\mathbb{E}_G[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b}')-\mu_n^t(\boldsymbol{b})]\right)\\
&\geq\max_{\boldsymbol{b}'}\left(\mathbb{E}_F[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b}')]\right)+\max_{\boldsymbol{b}'}\left(\mathbb{E}_G[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b}')]\right)-2\max_{\boldsymbol{b}'}\mathbb{E}_{(F+G)/2}\left(\mathbb{E}_F[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b}')]\right)\\
&\geq\mathbb{E}_F[\sum_{t=1}^{T}\mu_n^t((0,\ldots,0))]+\mathbb{E}_G[\sum_{t=1}^{T}\mu_n^t((c,\ldots,c))]-2\max_{\boldsymbol{b}'}\mathbb{E}_{(F+G)/2}\left(\mathbb{E}_F[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b}')]\right)\\
&=(\frac{1}{2}+\delta)(M-k)+(M-Mc)-2\max_{\boldsymbol{b}'}\mathbb{E}_{(F+G)/2}\left(\mathbb{E}_F[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b}')]\right)\\
&\geq(\frac{1}{2}+\delta)(M-k)+(M-Mc)-2(1-c)M.
\end{aligned}$$

Now, for example, we can set $k=\frac{M}{3}$ and $c=\frac{2}{3}$ to obtain a per step incurred regret of $\Theta(M\delta)$. We invoke the useful lemma relating the regret under $(F+G)/2$ to the Kullback-Leilber divergence:

Lemma 5 **(Tsybakov (2008) Theorem 2.2.)**. *We have for any two discrete distributions F and G:*

$$\mathbb{E}_{(F+G)/2}\left[Regret_{\mathcal{B}}(T)\right]=\Omega\left(\frac{\Delta}{2}\exp(-D_{\mathrm{KL}(F||G)})\right)\tag{25}$$

where Δ denotes the sum of the total regret incurred under F or G.

When F and G are independent Bernoulli processes with parameters $\frac{1}{2}+\delta$ and $\frac{1}{2}-\delta$ respectively, then $D_{\mathrm{KL}}(F||G)\leq cT\delta^2$ for some constant c. Using $\Delta\in\Theta(MT\delta)$, we have that the previous lemma implies:

$$\text{Regret}_{\mathcal{B}}\in\Omega\left(M\sqrt{T}\right)\tag{26}$$

where δ is chosen to be $\Omega(\frac{1}{\sqrt{T}})$.

9.7. Appendix to Section 7: Additional Experiments

In this section, we run additional experiments to (1) show the impact of the modified EXP3-IX estimator, and (2) empirically verify the regret guarantees of Algorithms 3 and 4. First, we explain the modified, EXP3-IX based versions of the algorithms as used in the experiments section, as well as why we chose to use these modified versions instead of the original ones. We show that the change in the algorithms is marginal, as the only step that is different between Algorithm 4 and its EXP3-IX variant is in updating the node weights. We then run experiments in the $M = 1$ unit setting to illustrate the impact of our modified algorithm. Second, in order to empirically gauge the regret guarantees of our proposed algorithms, we compare their performance against an adversary that bids stochastically. We analytically derive the optimal bidding strategy and compare how quickly our algorithms converge to this optimal solution. We repeat similar experiments for the uniform price auctions.

9.7.1. EXP3-IX vs. Unbiased Reward Estimator In the experiments section, we ran a slightly modified version of our existing algorithms in the bandit feedback setting. We do this as the variance of the accumulated regret of our algorithms are high, as the node weight estimators normalize over vanishingly small probabilities $q_m^t(b)$. To mitigate the effect of such normalization, we use the EXP3-IX estimator as described in Neu (2015), Lattimore and Szepesvári (2020). Under this estimator, rather than normalizing the probability of selecting bid b_m^t for unit m at time t by $q_m^t(b_m^t)$, we instead normalize it by $q_m^t(b_m^t) + \gamma$ for some constant $\gamma > 0$. In the standard K-armed bandit setting, despite being a biased estimator, still achieves the same sublinear expected regret guarantee with a smaller variance. This smaller variance indeed allows for stronger high probability guarantees on the magnitude of our regret; i.e., for $\delta > 0$ and $\gamma = \sqrt{\frac{\log(K)+\log(\frac{K+1}{\delta})}{4KT}}$, the EXP3-IX algorithm guarantees with probability at least $1 - \delta$ that the regret is upper bounded by $C\sqrt{KT \log K}$ for some absolute constant $c > 0$. We extend this algorithm to the multi-unit PAB setting algorithms, where for each node (m, b), we set $\gamma = \sqrt{\frac{\log(K)+\log(\frac{K+1}{\delta})}{4KT}}$ and $K = |\{b \in \mathcal{B} : b \leq v_m\}|$, for $\delta = 0.05$. Aside from the change in node weight estimators, the EXP3-IX versions of Algorithms 3 and 4 are exactly the same.

9.7.2. Empirical Performance of Original Algorithms vs. EXP3-IX Variants In this section, we empirically analyze the modified variants of our algorithms which use the biased, but lower variance EXP3-IX node-weight estimators (see Appendix 9.7.1). We compare the distribution of the regret recovered by these modified algorithms versus the non-modified versions when the number of units is one. The bidder, endowed with valuation vector $\boldsymbol{v} = [1]$, will compete against a single adversary over the course of T rounds for $\overline{M} = M = 1$ item. This is the standard first price auction (FPA). Here, we compare performance when the adversary is stochastic (bids drawn uniformly random from $[0, 1]$) versus adaptive adversary (running the same algorithm, with a valuation drawn uniformly random from $[0, 1]$).

We plot the regret of the bidder against the stochastic and adversarial competitors for moderate $T \in \{100, 500, 2000, 10000\}$. The stochastic adversary setting is shown in Figure 9 (a) and the adversarial setting is shown in Figure 9 (b). We observe that while the EXP3-IX variants marginally worsens regret for small values of $T \in \{100, 500\}$ for both the stochastic and adaptive settings, it significantly mitigates the heavy tailed distribution of regret for large $T \in \{2000, 10000\}$, especially in the adversarial setting.

9.7.3. Stochastic Setting To verify our algorithms' theoretical regret guarantees, we consider the setting where the bidder competes in a stochastic setting with multi-unit. Here, the bidder, endowed with valuation vector $\boldsymbol{v} = [1, 1, 1]$, will compete over the course of $T = 10^4$ rounds for $\overline{M} = M = 3$ items. The competing bids are $\boldsymbol{b}^{-1} = [0.1, 0.1, 0.1]$, $[0.3, 0.3, 1.0]$, or $[0.4, 1.0, 1.0]$ with probabilities $\frac{1}{4}, \frac{1}{4}$, and $\frac{1}{2}$, respectively. Assuming that the bidder

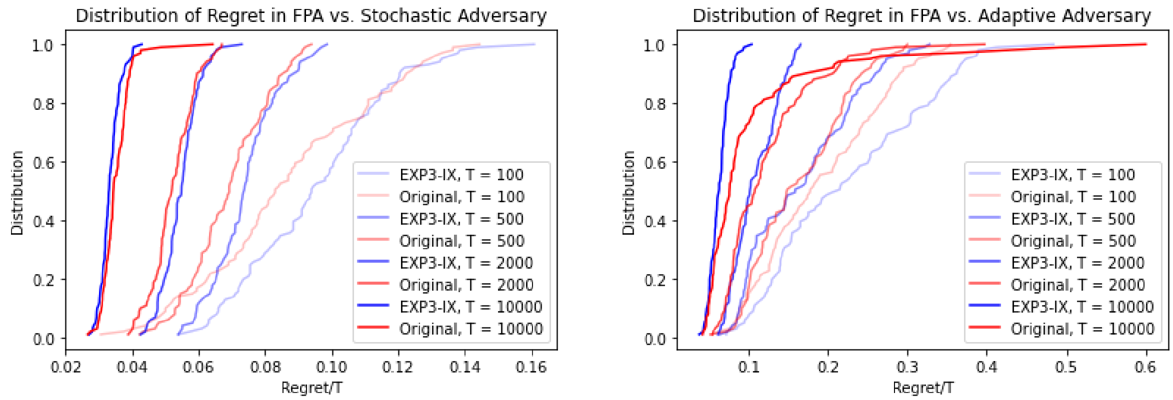

Figure 9 **Distribution of regret when using OMD vs its EXP3-IX variant against stochastic and adaptive adversaries for varying T.**

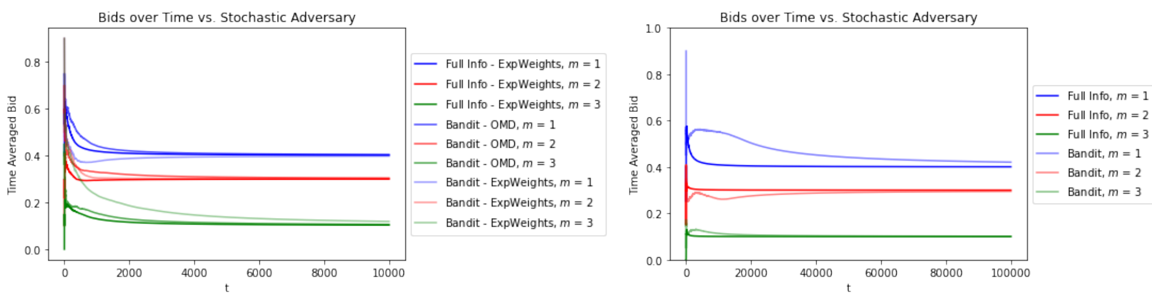

Figure 10 **Bid convergence over time under the stochastic setting in Section 9.7.3 for the PAB auction (left) and the uniform price auction (right).**

receives priority in tiebreaks, with $\mathcal{B} = \{\frac{i}{10}\}_{i\in[10]}$, the expected utility $\sum_{m=1}^{3} \mathbb{P}(b_m \geq b_m^{-1})(1-b_m)$ maximizing bid vector is given by $\boldsymbol{b} = [0.4, 0.3, 0.1]$, which yields utility $(1)(1-0.4)+(0.75)(1-0.3)+(0.5)(1-0.1) = 0.6+0.525+0.45 = 1.575$. We select learning rates $\eta = \sqrt{\frac{\log(|\mathcal{B}|)}{|\mathcal{B}|T}} \approx 0.005$ and $\eta = \sqrt{\frac{\log(|\mathcal{B}|)}{T}} = 0.002$ for the full information and bandit settings respectively (and for the EXP3-IX estimator, we choose an exploration rate of $\sqrt{\frac{2\log(|\mathcal{B}|/\delta)}{4|\mathcal{B}|T}} \approx 0.003$, for high probability bound parameter $\delta = 0.05$).

In Figure 10, we plot the average value of each bid over time. Here, the bidder's objective is to learn the optimal bid vector under each of our three algorithms: decoupled exponential weights algorithm (Algorithm 2) for the full information, modified (i.e., EXP3-IX) version of decoupled exponential weights algorithm (Algorithm 3) for the bandit setting, and modified version of the OMD algorithm (Algorithm 4) for the bandit setting. In this figure, we further compare the rate of convergence to the optimal bid vector of $[0.4, 0.3, 0.1]$ with our three algorithms. We observe that the full information decoupled exponential weights algorithm converges the fastest to the optimal bid, and the bandit feedback decoupled exponential weights algorithm converges the slowest, and the bandit feedback mirror descent algorithm is in between. This behavior is consistent with our theoretical findings.

We repeat this experiment for the algorithms to learn in uniform price auctions described in Brânzei et al. (2023). Though we do not perform the calculations, the optimal bid vector in the uniform price setting is still $[0.4, 0.3, 0.1]$. We note that it takes noticeably longer for the bandit algorithm to converge as compared to either its full information variant or our Algorithms 3 or 4, as predicted by the looser regret upper bounds:

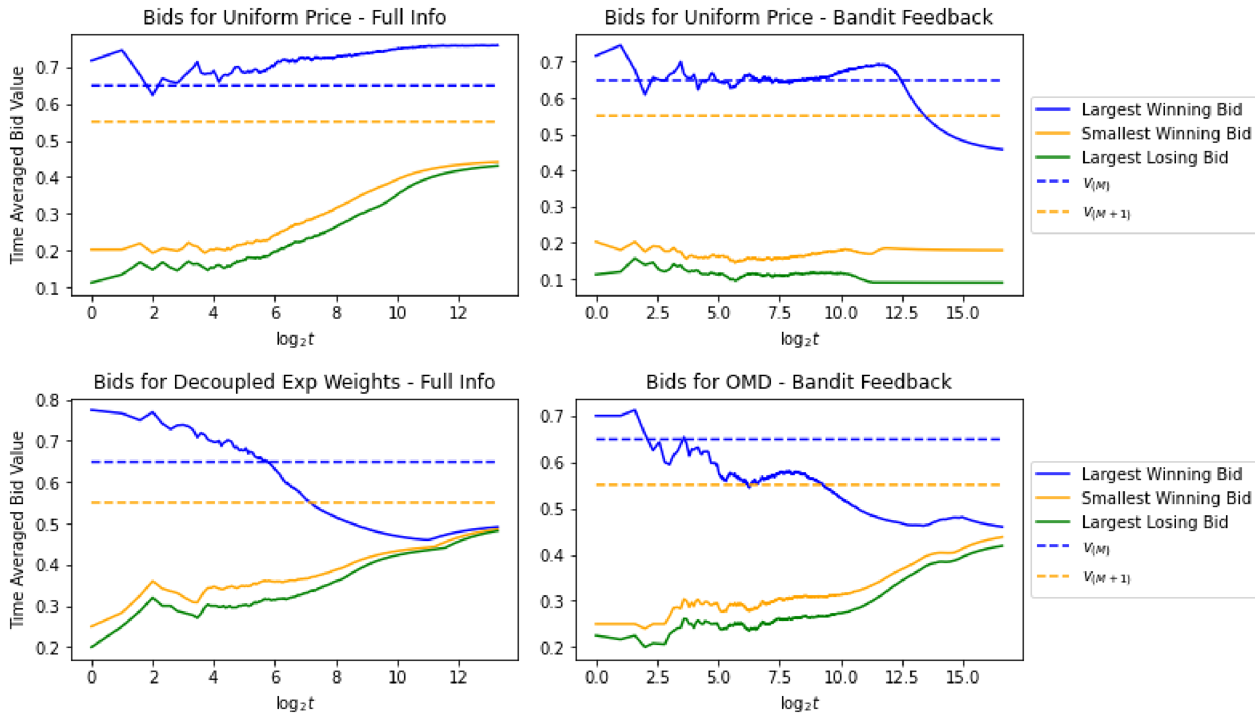

Figure 11 **Time averaged bids under market dynamics for the setting described in Section 7.1. The left (resp. right) figures correspond with the full information setting (resp. bandit setting) [and the top (resp. bottom) figures correspond with the uniform price (PAB) auctions. In this specific instance, valuations are given by** $\boldsymbol{v}_1 = [0.89, 0.7, 0.55, 0.51, 0.29], v_2 = [0.89, 0.44, 0.2, 0.12, 0.05], v_3 = [0.67, 0.64, 0.45, 0.27, 0.02]$**].**

THEOREM 8 **((Discrete) Regret in Uniform Price Auctions, Brânzei et al. (2023))**. *Under full information feedback (resp. bandit feedback), there exists an algorithm which achieves* $O(M^{\frac{3}{2}}\sqrt{T\log|\mathcal{B}|})$ *(resp.* $O(M^{\frac{5}{2}}|\mathcal{B}|T^{\frac{1}{2}}\log|\sqrt{\log|\mathcal{B}|}+M^2\log|\mathcal{B}|)$*) discrete regret.*

9.8. Uniform Price Market Dynamic Analysis

In this section, we provide some additional experiments or analyses in the market dynamics for the uniform price as prescribed in Section 7.1 in order to directly compare to the dynamics of the PAB auction. In particular, we run a more complete analysis of the uniform price auction bidding dynamics that parallels Section 7.1.

Uniform Price Learning Dynamics: Here, we repeat the setup and analyses of Section 7.1 except using the uniform price auction with bidders bidding according to the learning algorithms as prescribed in Brânzei et al. (2023).

Bid Dynamics. In Figure 11, we observe that the winning bids and largest losing bids noticeably diverge, indicating that the regret minimizing bid strategies are non-uniform (and so are the true hindsight optimal bid vectors which we verified using an offline optimization protocol described in Brânzei et al. (2023)). In particular, as in Section 7.1, there are $N = 3$ bidders, $\overline{M} = M = 5$ items, the bid space is $\mathcal{B} = \{\frac{i}{20}\}_{i\in[20]}$. The valuations (which are drawn i.i.d. Unif$(0,1)$ which are then sorted) for this specific instance are given by $\boldsymbol{v}_1 = [0.89, 0.7, 0.55, 0.51, 0.29], v_2 = [0.89, 0.44, 0.2, 0.12, 0.05], v_3 = [0.67, 0.64, 0.45, 0.27, 0.02]$. For convenience and ease of comparison, we include the PAB bidding dynamics counterpart (Figure 2) as well.

Welfare and Revenue Over Time. In Figure 12, we compare the distribution of welfare and revenue (normalized by maximum welfare) of the uniform price auction over time showing the 10th, 25th, 50th, 75th, and 90th percentiles

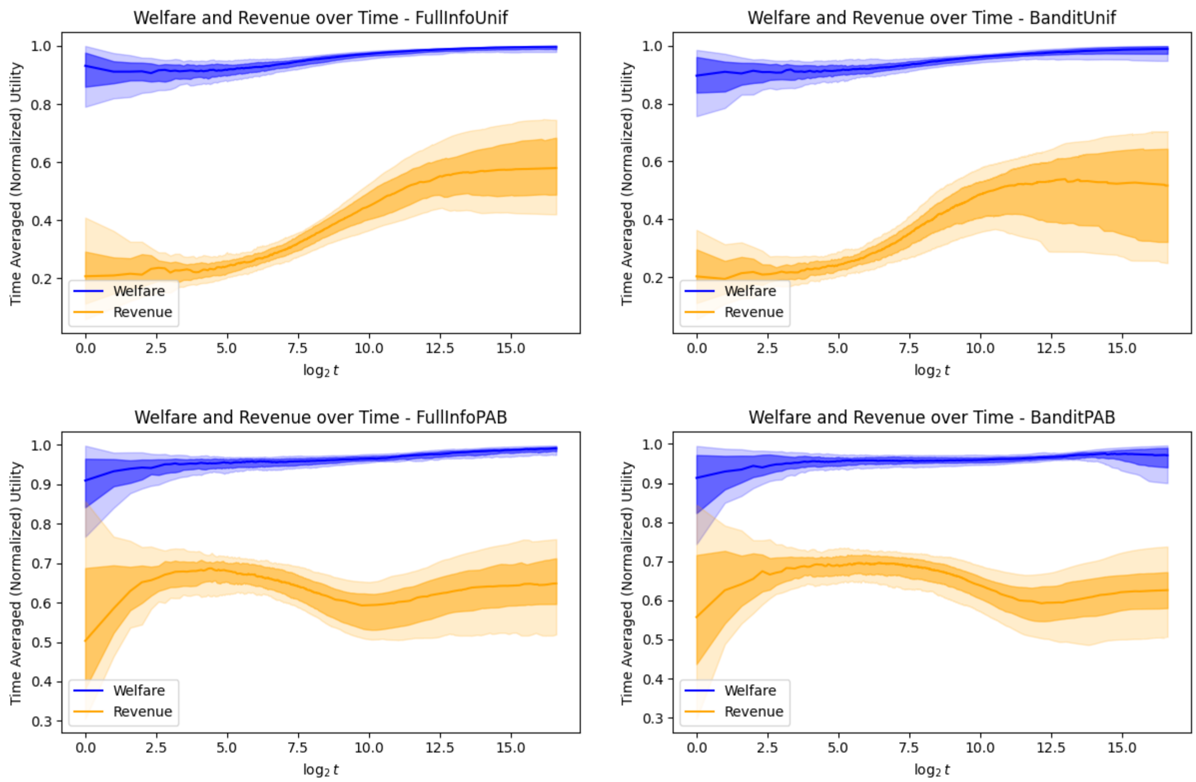

Figure 12 **Welfare and revenue over time under the market dynamics for the uniform price auction, under the setting described in Section 7.1. The left (resp. right) figures correspond with the full information setting (resp. bandit setting) and the top (resp. bottom) figures correspond with the uniform price (PAB) auctions.**

in different shades. In particular, we run the full information and bandit feedback learning algorithms for the uniform price auction Brânzei et al. (2023). Note that this figure parallels Figure 5 under PAB. We observe that the welfare rapidly converges to 1 in both the full information and bandit feedback settings. However, the revenue under bandit feedback has noticeably larger variance compared to the full information revenue. Once again, to compare the welfare and revenue evolution over time with the PAB auction, we include the PAB counterpart (Figure 5) for convenience.

9.9. Time Varying Valuations

We extend Algorithms 2 and 3 to the time varying valuations setting. In particular, we assume that the valuations $\boldsymbol{v}$ are no longer fixed, and instead, in every round t, $\boldsymbol{v}^t$ is independently drawn i.i.d. from some known distribution $F_{\boldsymbol{v}}$ with discrete, finite support $\mathcal{V}$. This contextual setting requires a stronger benchmark oracle in comparison to our original setup with a fixed valuation. The new benchmark oracle, which we will formalize shortly, possesses knowledge of the hindsight optimal bid vector for each context. That is, under this benchmark, we have the optimal mapping from any context (valuation vector) to an action (bid vector). Consequently, our current definitions of $\textsc{Regret}$ and $\textsc{Regret}_{\mathcal{B}}$ need to be updated to accommodate these contextual factors:

$$\textsc{Regret}(F_{\boldsymbol{v}}) = \max_{\boldsymbol{b}:\mathcal{V}\to[0,1]^{+M}} \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{v}\sim F_{\boldsymbol{v}}}[\mu_n^t(\boldsymbol{b}(\boldsymbol{v});\boldsymbol{v})] - \mathbb{E}\left[\sum_{t=1}^{T} \mu_n^t(\boldsymbol{b}^t;\boldsymbol{v}^t)\right]. \qquad \text{(Continuous Contextual Regret)}$$

Here, $\mu_n^t(\boldsymbol{b};\boldsymbol{v})$ denotes the utility of bidder n by submitting bid vector $\boldsymbol{b}$ with valuations $\boldsymbol{v}$ at round t where the competing bids are $\boldsymbol{b}_-^t$. Observe that in the benchmark of $\textsc{Regret}(F_{\boldsymbol{v}})$, i.e., $\max_{\boldsymbol{b}:\mathcal{V}\to[0,1]^{+M}} \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{v}\sim F_{\boldsymbol{v}}}[\mu_n^t(\boldsymbol{b}(\boldsymbol{v});\boldsymbol{v})]$, we abuse notation and define valuation-to-bid vector mapping $\boldsymbol{b}:\mathcal{V}\to[0,1]^{+M}$. We have an equivalent definition of discretized contextual regret:

$$\textsc{Regret}_{\mathcal{B}}(F_{\boldsymbol{v}}) = \max_{\boldsymbol{b}:\mathcal{V}\to\mathcal{B}^{+M}} \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{v}\sim F_{\boldsymbol{v}}}[\mu_n^t(\boldsymbol{b}(\boldsymbol{v});\boldsymbol{v})] - \mathbb{E}\left[\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b}^t;\boldsymbol{v}^t)\right]. \qquad \text{(Discretized Contextual Regret)}$$

An agent's goal is to minimize their contextual regret with respect to their valuation distribution $F_{\boldsymbol{v}}$. Using naive contextual bandit algorithms would lead to a large regret, as the regret of these algorithms scales with the square root of the number of contexts. However, we make an observation that we have complete cross-learning over these contexts as in Balseiro et al. (2022). As such, we borrow from the results described in Balseiro et al. (2022); specifically those explaining the cross-learning-across-contexts generalizations of the EXP3 algorithm in the stochastic contexts (valuations) and adversarial rewards setting (adversarial competing bids).

We assume that the agent has access to their valuation distribution. Moreover, as stated earlier, we assume that the support of this valuation distribution is finite; i.e., $|\mathcal{V}|<\infty$. This scenario occurs often in practice where bidders' valuations depend naturally on some natural events. For example, investors may prescribe a 'low' or 'high' value to certain assets depending on various market indices.

We generalize the EXP3-CL algorithm described in Balseiro et al. (2022) to our PAB setting, specifically Algorithm 3, and achieve exactly the same regret rates as our non-contextual variants, albeit requiring an additional $O(|\mathcal{V}|)$ factor of memory and computation.

In order to make the generalization more clear, at a high level, the EXP3-CL algorithm on a set of K arms and C contexts with full cross learning constructs a reward estimator $\hat{r}(k;c) = \frac{r(k;c)}{\sum_c \mathbb{P}(c)\mathbb{P}(k^t=k|c^t=c)}\mathbf{1}_{k^t=k}$ for each arm k and context c pair. Here, the term $\sum_c \mathbb{P}(c)\mathbb{P}(k^t=k|c^t=c)$ is the expected probability that arm $k^t=k$ was selected under context $c^t=c$, where in the summation we take expectation with respect to the stochasticity over contexts c. This estimator mirrors that of standard EXP3 using the IPW estimator, except that the IPW is averaged over the context distribution.

To apply this to our setting, we wish to mimic the behavior of the EXP3-CL algorithm with our decoupled exponential weights algorithm. This can be done by running the EXP3-CL estimator on all of the nodes $b\in\mathcal{B}$ within each layer $m\in[M]$. In particular, we use the following estimator $\widehat{w}_m^t(b;\boldsymbol{v}) = 1-\frac{1-w_m^t(b;\boldsymbol{v})}{Q_m^t(b)}\mathbf{1}_{b_m^t=b}$, where the normalizer $Q_m^t(b)=\sum_{\boldsymbol{v}\in\mathcal{V}}\mathbb{P}(\boldsymbol{v}^t=\boldsymbol{v})q_m^t(b;\boldsymbol{v})$ in this estimator is the expected probability of selecting bid b with corresponding valuation $\boldsymbol{v}=\boldsymbol{v}^t$, where the expectation is taken with respect to all valuation vectors $\boldsymbol{v}\in\mathcal{V}$. This procedure, formally described in Algorithm 5 yields the following regret upper bound:

Theorem 9 **(Time Varying Valuations - Decoupled Exponential Weights)**. *Under bandit feedback (resp. full information feedback), Algorithm 5, with appropriately chosen* η, *achieves contextual continuous regret* $\textsc{Regret}(F_{\boldsymbol{v}})$ *of order* $O(M^{\frac{3}{2}}\sqrt{T\log T})$ *(resp.* $O(M\sqrt{T\log T})$*) with total time time and space complexity polynomial in* M, $|\mathcal{B}|$, $|\mathcal{V}|$, *and* T.

Proof of Theorem 9 To begin, we can once again 'decouple' the utility per unit-bid pair, but this time conditional on the valuation vector context. In particular, we have:

$$\mu_n^t(\boldsymbol{b};\boldsymbol{v}) = \sum_{m=1}^{M} w_m^t(b_m;\boldsymbol{v}) = \sum_{m=1}^{M}(v_m-b_m)1_{b_m\geq b_{-m}^t} \quad \text{and} \quad \widehat{\mu}_n^t(\boldsymbol{b};\boldsymbol{v}) = \sum_{m=1}^{M}\widehat{w}_m^t(b_m;\boldsymbol{v}).$$

As stated earlier, we define reward-weight estimates based on Equation (6) of Balseiro et al. (2022) and our Algorithm 3:

$$\widehat{w}_m^t(b;\boldsymbol{v}) = 1-\frac{1-w_m^t(b;\boldsymbol{v})}{\sum_{\boldsymbol{v}\in\mathcal{V}}\mathbb{P}(\boldsymbol{v}^t=\boldsymbol{v})q_m^t(b;\boldsymbol{v})}\mathbf{1}_{b_m^t=b} = 1-\frac{1-w_m^t(b;\boldsymbol{v})}{Q_m^t(b)}\mathbf{1}_{b_m^t=b}.$$

Here, $q_m^t(b;\boldsymbol{v}) = \mathbb{P}(b_m^t = b|\boldsymbol{v}^t = \boldsymbol{v}) = \sum_{\boldsymbol{b}:b_m^t=b} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}|\boldsymbol{v}^t = \boldsymbol{v})$ is the probability of selecting bid b in slot m with valuation $\boldsymbol{v}$. Similarly, $Q_m^t(b)$ is the probability of selecting bid b for unit m, averaged across all possible valuations. One can verify unbiasedness of this estimator $\mathbb{E}[\hat{w}_m^t(b;\boldsymbol{v})] = w_m^t(b;\boldsymbol{v})$ for all $m \in [M], b \in \mathcal{B}, \boldsymbol{v} \in \mathcal{V}$. The second moment can similarly be computed as:

$$\mathbb{E}[\widehat{w}_m^t(b;\boldsymbol{v})^2] = \mathbb{E}\left[\left(1 - \frac{1 - w_m^t(b;\boldsymbol{v})}{Q_m^t(b)}\mathbf{1}_{b_m^t=b}\right)^2\right] = 1 - 2\mathbb{E}\left[\frac{1 - w_m^t(b;\boldsymbol{v})}{Q_m^t(b)}\mathbf{1}_{b_m^t=b}\right] + \mathbb{E}\left[\left(\frac{1 - w_m^t(b;\boldsymbol{v})}{Q_m^t(b)}\right)^2\mathbf{1}_{b_m^t=b}\right].$$

Evaluating the expectations and recalling that $\mathbb{E}[\mathbf{1}_{b_m^t=b}] = Q_m^t(b)$, we have:

$$\mathbb{E}[\widehat{w}_m^t(b;\boldsymbol{v})^2] = 1 - \left[2 - 2w_m^t(b;\boldsymbol{v})\right] + \left[\frac{(1 - w_m^t(b;\boldsymbol{v}))^2}{Q_m^t(b)}\right] = 2w_m^t(b;\boldsymbol{v}) - 1 + \frac{1}{Q_m^t(b)} \leq 1 + \frac{1}{Q_m^t(b)} \leq \frac{2}{Q_m^t(b)}.$$

Using this, the proof largely follows that of Algorithm 3 up until Equation (12). In particular, we have that the contextual regret can be written as:

$$\begin{aligned}
\textsc{Regret}_{\mathcal{B}}(F_{\boldsymbol{v}}) &= \mathbb{E}_{F_{\boldsymbol{v}}}\left[\sum_{t=1}^T \mu_n^t(\boldsymbol{b}';\boldsymbol{v}^t) - \sum_{t=1}^T \mathbb{E}[\mu^t(\boldsymbol{b}^t;\boldsymbol{v}^t)]\right] \\
&\lesssim \eta^{-1} M\log|\mathcal{B}| + \eta\mathbb{E}_{F_{\boldsymbol{v}}}\left[\sum_{t=1}^T \sum_{\boldsymbol{b}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}|\boldsymbol{v}^t = \boldsymbol{v})\mathbb{E}[(\sum_{m=1}^M \widehat{w}_m^t(b_m;\boldsymbol{v}))^2]\right] \\
&= \eta^{-1} M\log|\mathcal{B}| + \eta\left[\sum_{t=1}^T \sum_{\boldsymbol{v}\in\mathcal{V}} \mathbb{P}(\boldsymbol{v}^t = \boldsymbol{v})\sum_{\boldsymbol{b}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}|\boldsymbol{v}^t = \boldsymbol{v})\mathbb{E}[(\sum_{m=1}^M \widehat{w}_m^t(b_m;\boldsymbol{v}))^2]\right] \\
&= \eta^{-1} M\log|\mathcal{B}| + \eta M\left[\sum_{t=1}^T \sum_{m=1}^M \sum_{\boldsymbol{v}\in\mathcal{V}} \mathbb{P}(\boldsymbol{v}^t = \boldsymbol{v})\sum_{b\in\mathcal{B}} \mathbb{E}[\widehat{w}_m^t(b;\boldsymbol{v})^2]\sum_{\boldsymbol{b}:b_m=b} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}|\boldsymbol{v}^t = \boldsymbol{v})\right] \\
&= \eta^{-1} M\log|\mathcal{B}| + \eta M\left[\sum_{t=1}^T \sum_{m=1}^M \sum_{\boldsymbol{v}\in\mathcal{V}} \mathbb{P}(\boldsymbol{v}^t = \boldsymbol{v})\sum_{b\in\mathcal{B}} \mathbb{E}[\widehat{w}_m^t(b;\boldsymbol{v})^2]q_m^t(b;\boldsymbol{v})\right] \\
&= \eta^{-1} M\log|\mathcal{B}| + 2\eta M\left[\sum_{t=1}^T \sum_{m=1}^M \sum_{\boldsymbol{v}\in\mathcal{V}} \mathbb{P}(\boldsymbol{v}^t = \boldsymbol{v})\sum_{b\in\mathcal{B}} \frac{1}{Q_m^t(b)}q_m^t(b;\boldsymbol{v})\right] \\
&= \eta^{-1} M\log|\mathcal{B}| + 2\eta M\left[\sum_{t=1}^T \sum_{m=1}^M \sum_{b\in\mathcal{B}} \frac{1}{Q_m^t(b)}\sum_{\boldsymbol{v}\in\mathcal{V}} \mathbb{P}(\boldsymbol{v}^t = \boldsymbol{v})q_m^t(b;\boldsymbol{v})\right] \\
&= \eta^{-1} M\log|\mathcal{B}| + 2\eta M\left[\sum_{t=1}^T \sum_{m=1}^M \sum_{b\in\mathcal{B}} \frac{1}{Q_m^t(b)}Q_m^t(b)\right] \\
&\leq \eta^{-1} M\log|\mathcal{B}| + \eta M^2|\mathcal{B}|T.
\end{aligned}$$

(We will show the first inequality shortly.) With $\eta = \Theta(\sqrt{\frac{\log|\mathcal{B}|}{M|\mathcal{B}|T}})$, this yields the discretized contextual regret upper bounds of $O(M^{\frac{3}{2}}\sqrt{|\mathcal{B}|T\log|\mathcal{B}|})$ under the bandit setting. Accounting for the rounding error of order $O(\frac{MT}{|\mathcal{B}|})$, we obtain the stated continuous contextual regret upper bounds. To obtain the full information results, we simply replace $\widehat{w}_m^t(b_m;\boldsymbol{v}^t)$ with $w_m^t(b_m;\boldsymbol{v}^t)$ in the second line of the above equations, which leads to the discretized contextual regret upper bounds of $O(M^{\frac{3}{2}}\sqrt{T\log|\mathcal{B}|})$, as desired.

Next, following the proof of Algorithm 3, we show the first inequality. We define the potentials with respect to a fixed valuation vector $\boldsymbol{v}$: $\Phi^t(\boldsymbol{v}) = \sum_{\boldsymbol{b}\in\mathcal{B}^{+M}} \exp(\eta\sum_{\tau=1}^t \widehat{\mu}^\tau(\boldsymbol{b};\boldsymbol{v}^\tau))$. Taking the ratio of adjacent terms, we obtain:

$$\frac{\Phi^t(\boldsymbol{v})}{\Phi^{t-1}(\boldsymbol{v})} = \sum_{\boldsymbol{b}\in\mathcal{B}^{+M}} \frac{\exp(\eta\sum_{\tau=1}^{t-1} \widehat{\mu}^\tau(\boldsymbol{b};\boldsymbol{v}^\tau))}{\Phi^{t-1}(\boldsymbol{v})}\exp(\eta\widehat{\mu}^t(\boldsymbol{b};\boldsymbol{v}^t)) = \sum_{\boldsymbol{b}\in\mathcal{B}^{+M}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}|\boldsymbol{v}^t = \boldsymbol{v})\exp(\eta\widehat{\mu}^t(\boldsymbol{b};\boldsymbol{v}^t)),$$

Where in the last equality, we used the condition that our algorithm samples bid vector $\boldsymbol{b}$ with probability proportional to $\exp(\eta\sum_{\tau=1}^{t-1} \widehat{\mu}^t(\boldsymbol{b};\boldsymbol{v}))$ at round t with valuations $\boldsymbol{v}^t = \boldsymbol{v}$. Combining this with inequalities $\exp(x) \leq 1 + x + x^2$ and $1 + x \leq \exp(x)$ for all $x \leq 1$, we obtain:

$$\frac{\Phi^t(\boldsymbol{v})}{\Phi^{t-1}(\boldsymbol{v})} \leq \sum_{\boldsymbol{b}\in\mathcal{B}^{+M}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}|\boldsymbol{v}^t = \boldsymbol{v})\exp(\eta\widehat{\mu}^t(\boldsymbol{b};\boldsymbol{v})) \leq \exp(\sum_{\boldsymbol{b}\in\mathcal{B}^{+M}} \mathbb{P}(\boldsymbol{b}^t = \boldsymbol{b}|\boldsymbol{v}^t = \boldsymbol{v})\left[\eta\widehat{\mu}^t(\boldsymbol{b};\boldsymbol{v}) + \eta^2\widehat{\mu}^t(\boldsymbol{b};\boldsymbol{v})^2\right]).$$

ALGORITHM 5: DECOUPLED EXP3-CL - TIME VARYING VALUATIONS

Input: Learning rate $0<\eta<\frac{1}{M}$, Valuation Distribution $F_{\boldsymbol{v}}$

Output: The aggregate utility $\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b}^t;\boldsymbol{v}^t)$

$\widehat{W}_m^0(b;\boldsymbol{v})\leftarrow 0$ for all $m\in[M], b\in\mathcal{B}, \boldsymbol{v}\in\mathcal{V}$ such that $b\le v_m$; else $\widehat{W}_m^0(b;\boldsymbol{v})\leftarrow -\infty$.;

for $t\in[1,\dots,T]$: **do**

  Observe Valuation Vector $\boldsymbol{v}^t\sim F_{\boldsymbol{v}}$;

  $b_0^t\leftarrow\max\mathcal{B}$, and $\widehat{S}_{M+1}^t(\min\mathcal{B};\boldsymbol{v}^t)=1$ for any $t\in[T]$;

  Recursively Computing Exponentially Weighted Partial Utilities $\boldsymbol{S}^t$;

  for $m\in[M,\dots,1], b\in\mathcal{B}:\widehat{S}_m^t(b;\boldsymbol{v}^t)\leftarrow\exp(\eta\widehat{W}_m^t(b;\boldsymbol{v}^t))\sum_{b'\le b}\widehat{S}_{m+1}^t(b';\boldsymbol{v}^t)$ \\ COMPUTE $-\widehat{S}_m$;

  Determining the Bid Vector $\boldsymbol{b}^t$ **Recursively**;

  for $m\in[1,\dots,M], b\le b_{m-1}^t: b_m^t\leftarrow b$ with probability $\frac{\widehat{S}_m^t(b;\boldsymbol{v}^t)}{\sum_{b'\le b_{m-1}^t}\widehat{S}_m^t(b';\boldsymbol{v}^t)}$; \\ SAMPLE $-\boldsymbol{b}$;

  Observe $\boldsymbol{b}_-^t$ and receive reward $\mu_n^t(\boldsymbol{b}^t;\boldsymbol{v}^t)$;

  $Q_m^t(b)\leftarrow 0$ for all $m\in[M], b\in\mathcal{B}$;

  for $\boldsymbol{v}\in\mathcal{V}$: **do**

    Recursively Computing Probability Measure $\boldsymbol{q}$ **Under** $\boldsymbol{v}\in\mathcal{V}$;

    $\widehat{S}_{M+1}^t(b;\boldsymbol{v})\leftarrow 1$ for all $m\in[M], b\in\mathcal{B}$;

    for $m\in[M,\dots,1], b\in\mathcal{B}:\widehat{S}_m^t(b;\boldsymbol{v})\leftarrow\exp(\eta\widehat{W}_m^t(b;\boldsymbol{v}))\sum_{b'\le b}\widehat{S}_{m+1}^t(b';\boldsymbol{v})$;

    $q_1^t(b;\boldsymbol{v})\leftarrow\frac{\widehat{S}_m^t(b;\boldsymbol{v})}{\sum_{b'\in\mathcal{B}}\widehat{S}_m^t(b';\boldsymbol{v})}$ for all $b\in\mathcal{B}$;

    for $m\in[2,\dots,M], b\in\mathcal{B}: q_m^t(b;\boldsymbol{v})\leftarrow\sum_{b'\ge b}\frac{q_{m-1}^t(b';\boldsymbol{v})\widehat{S}_m^t(b;\boldsymbol{v})}{\sum_{b"\ge b'}\widehat{S}_m^t(b";\boldsymbol{v})}$ for all $b\in\mathcal{B}$;

    $Q_m^t(b)\leftarrow Q_m^t(b)+\mathbb{P}(\boldsymbol{v}^t=\boldsymbol{v})q_m^t(b;\boldsymbol{v})$

  end

  Update Weight Estimates;

  if BANDIT FEEDBACK, **for** $m\in[M], b\in\mathcal{B}, \boldsymbol{v}\in\mathcal{V}$;

  $\widehat{W}_m^{t+1}(b;\boldsymbol{v})\leftarrow\widehat{W}_m^t(b;\boldsymbol{v})+(1-\frac{1-(v-b)\mathbf{1}_{b\ge b_m^t}}{Q_m^t(b)}\mathbf{1}_{b_m^t=b})$ if $b\le v$; else $\widehat{W}_m^{t+1}(b;\boldsymbol{v})\leftarrow-\infty$;

  if FULL INFORMATION, **for** $m\in[M], b\in\mathcal{B}, \boldsymbol{v}\in\mathcal{V}$;

  $\widehat{W}_m^{t+1}(b;\boldsymbol{v})\leftarrow\widehat{W}_m^t(b;\boldsymbol{v})+(v-b)\mathbf{1}_{b\ge b_m^t}$ if $b\le v$; else $\widehat{W}_m^{t+1}(b;\boldsymbol{v})\leftarrow-\infty$;

end

Return $\sum_{t=1}^{T}\mu_n^t(\boldsymbol{b}^t;\boldsymbol{v}^t)$

Combining this with Equations 11 and the fact that $\Phi^0(\boldsymbol{v})=M\log|\mathcal{B}|$, for any fixed bid vector $\boldsymbol{b}'$, we have:

$$\begin{aligned}\sum_{t=1}^{T}\widehat{\mu}^t(\boldsymbol{b}';\boldsymbol{v})-\sum_{t=1}^{T}\sum_{\boldsymbol{b}}\mathbb{P}(\boldsymbol{b}^t=\boldsymbol{b}|\boldsymbol{v}^t=\boldsymbol{v})\widehat{\mu}^t(\boldsymbol{b};\boldsymbol{v})&\lesssim\eta^{-1}M\log|\mathcal{B}|+\eta\sum_{t=1}^{T}\sum_{\boldsymbol{b}}\mathbb{P}(\boldsymbol{b}^t=\boldsymbol{b}|\boldsymbol{v}^t=\boldsymbol{v})\widehat{\mu}^t(\boldsymbol{b};\boldsymbol{v})^2\\&=\eta^{-1}M\log|\mathcal{B}|+\eta\sum_{t=1}^{T}\sum_{\boldsymbol{b}}\mathbb{P}(\boldsymbol{b}^t=\boldsymbol{b}|\boldsymbol{v}^t=\boldsymbol{v})(\sum_{m=1}^{M}\widehat{w}_m^t(b_m;\boldsymbol{v}))^2\,.\end{aligned}$$

Taking expectations over $\boldsymbol{b}$ and the supremum over all $\boldsymbol{b}'$ yields the desired first crucial regret inequality.

As for the time and space complexity, notice that the only algorithmic difference between Algorithm 5 and Algorithm 3 is precisely in computing the estimator, which in the former, requires having to compute the weights $Q_m^t(b)$ by iterating over all $\boldsymbol{v}\in\mathcal{V}$. As we also have to store reward estimates for each possible valuations, both the time complexity and space complexity of Algorithm 5 are a factor $|\mathcal{V}|$ larger than in Algorithm 3, which are $O(M|\mathcal{B}||\mathcal{V}|T)$ and $O(M|\mathcal{B}||\mathcal{V}|)$ respectively.

Contact.

MIT CEEPR Working Paper Series
is published by the MIT Center for Energy and Environmental Policy Research from submissions by affiliated researchers.
For inquiries and/or for permission to reproduce material in this working paper, please contact:

General inquiries: ceepr@mit.edu
Media inquiries: ceepr-media@mit.edu

MIT Center for Energy and
Environmental Policy Research

MIT Center for Energy and
Environmental Policy Research
Massachusetts Institute of Technology
77 Massachusetts Avenue, E19-411
Cambridge, MA 02139-4307
USA

ceepr.mit.edu

MASSACHUSETTS INSTITUTE OF TECHNOLOGY