Learning to Bid in Discriminatory Auctions with Budget Constraints

Negin Golrezaei
MIT Sloan School of Management

Sourav Sahoo
MIT Operations Research Center

Abstract

We study repeated bidding in multi-unit discriminatory (pay-as-bid) auctions for a single bidder with per-round utility equal to value minus α times payment, where $\alpha \in [0,1]$ is a cost-of-capital parameter. The bidder aims to maximize cumulative utility over T rounds subject to a total budget B. The problem is challenging even without budgets: the action space is exponential in the bidder's maximum demand M, and the valuation vector (context) varies over time. Exploiting a decomposition of utility across units, we develop polynomial-time learning algorithms based on shortest paths in a directed acyclic graph, obtaining sublinear regret under both full-information and bandit feedback. In the bandit setting, the regret is independent of the number of contexts due to complete cross-learning: observing the utility of the chosen action under the realized context reveals the utility for the same action under all counterfactual contexts. With budget constraints, when the average normalized per-round budget $\rho = \frac{B}{MT} < 1$, we design a coupled primal-dual algorithm in which the DAG-based procedure uses dual-adjusted edge weights for primal updates, while online gradient descent updates the dual variable, yielding ρ-approximate sublinear regret. Finally, we give implementations whose per-round time and space are independent of the number of contexts, enabling scalability to large or even infinite context spaces.

1 INTRODUCTION

Multi-unit discriminatory auctions, also known as pay-as-bid (PAB) auctions, are commonly used in treasury auctions (Brenner et al., 2009) and electricity markets (Maurer and Barroso, 2011). In this format, K identical units of an item are sold to bidders who may demand multiple units, and each winning bidder pays exactly what they bid for each allocated unit. The well-known first-price auction (FPA) is a special case of PAB auction when $K = 1$.

Bidding in PAB auctions is notoriously challenging. Unlike truthful mechanisms such as VCG (Vickrey, 1961; Clarke, 1971; Groves, 1973), reporting one's true valuation is not optimal in PAB auctions (Ausubel et al., 2014). Even under restrictive assumptions like common priors and equilibria play, computing optimal strategies is intractable due to the multi-dimensional nature of valuations (Kasberger and Woodward, 2025). The complexity is further amplified in repeated settings with budget constraints. Here, bidders must adapt to strategic or even adversarial competitors, time-varying valuations, navigate an exponentially large bid space (even with discrete bids, as is common in practice), and resolve the "spend-or-save" dilemma, where budget allocation today affects outcomes in the future. These challenges underscore the need for efficient and computationally tractable bidding strategies for budget-constrained bidders in repeated PAB auctions.

1.1 Our Contributions

We design bidding algorithms for a budget-constrained bidder in repeated multi-unit PAB auctions.

Modeling. We consider the bidder's utility as value minus α times payment, where $\alpha \in [0,1]$ is a cost-of-capital parameter (Balseiro et al., 2019b). This unifies two existing common behavioral models: traditional *profit maximizers* ($\alpha = 1$) and *value maximizers* ($\alpha = 0$) (Balseiro et al., 2021b; Lv et al., 2023; Deng et al., 2024). We model the budget as a *hard constraint*, requiring it to be satisfied for every realized sequence of auctions, as is common in practice (Wang et al., 2023; Castiglioni et al., 2024). While prior works have focused on the setting where competing bids are drawn i.i.d. from a distribution (Feng et al., 2023; Han et al., 2024), we allow the competing bids to be generated adversarially. The individual bids are integral multiples of $\epsilon \in (0,1)$, reflecting the minimum bid increments in real-world auctions such as eBay (eBay, 2026) and U.S. Treasury auctions (Hortaçsu et al., 2018).

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

We focus on *no-overbidding (NOB)* strategies, which require that cumulative bids never exceed cumulative valuations (see Definition 1). This ensures that the bidder's utility is non-negative, regardless of the competing bids. For bidders with $\alpha = 1$, prior work often imposes a stricter *per-unit* NOB condition (Markakis and Telelis, 2015; Galgana and Golrezaei, 2024), which can be highly suboptimal when $\alpha < 1$. Our cumulative notion of NOB is therefore more general and necessitates new techniques for designing bidding algorithms.

No Budget Constraints. We first study the offline problem of computing the hindsight-optimal strategy without budget constraints, assuming the competing bids are known in advance. The main challenge is that the number of NOB strategies grows as $O(\epsilon^{-M})$, where M is the bidder's maximum demand. To address this, we exploit that the bidder's utility decomposes across units. Leveraging this structure, we construct an edge-weighted directed acyclic graph (DAG) of size $\text{poly}(M, 1/\epsilon)$ in which each s–d path corresponds to a NOB strategy and the associated path weight is an affine transformation of that strategy's total utility. So, computing the optimal offline NOB strategy reduces to finding a shortest path in the DAG (Theorem 3.1).

In the online setting, valuation vectors are drawn i.i.d. across rounds from a distribution $\mathcal{D}$ supported on $\mathcal{V}$. We first consider the case where $|\mathcal{V}|$ is finite (and small), and later extend the results to large or infinite context spaces. Treating each valuation vector $\mathbf{v} \in [0,1]^M$ as a context, the bidder maps each context to a NOB strategy (action) and, building on the offline reduction, maintains one DAG per context. The bidder then performs exponential-weights updates over s-d paths using a variant of the weight-pushing method of Takimoto and Warmuth (2003). Under full-information feedback, all the competing bids are revealed *ex post*,[1] allowing the bidder to update all DAGs. This algorithm does not require knowledge of $\mathcal{D}$ and achieves $\widetilde{O}(M^{3/2}\sqrt{T})$ regret, independent of $|\mathcal{V}|$ (Theorem 3.2).[2,3]

In the bandit setting, only the bidder's allocation is revealed after each auction. Under such limited feedback, a naive contextual bandit approach that updates only the DAG corresponding to the realized context typically incurs an unavoidable $O(\sqrt{|\mathcal{V}|})$ dependence in the regret bound. In contrast, in our setting the realized utility under the observed context helps infer counterfactual utilities that can be leveraged to update *all* contexts, yielding *complete cross learning* between contexts, a notion introduced by (Balseiro et al., 2019a). As a result, our regret bounds are independent of $|\mathcal{V}|$. Specifically, when the context distribution $\mathcal{D}$ is known, we exploit this knowledge to construct edge-weight estimators (cf. Eq. (11)), obtaining a regret bound of $\widetilde{O}(M^2\epsilon^{-3/2}\sqrt{T})$ (Theorem 3.3). When $\mathcal{D}$ is unknown, we obtain a slightly weaker bound of $\widetilde{O}(M^2\epsilon^{-1}T^{2/3})$ (Theorem 3.4). Additionally, we establish a regret lower bound of $\Omega(M\sqrt{T})$ (Theorem 3.5).

With Budget Constraints. Having addressed the unconstrained regime, we turn to the case where the bidder has a total budget $B = \rho MT$ over T rounds. This introduces a long-term constraint that couples decisions across time. Under NOB and the payment rule, the per-round payment is at most M, so when $\rho \geq 1$ the budget never binds and the problem reduces to the unconstrained setting. The interesting regime is $\rho < 1$, where the bidder must actively pace spending.

Prior work on budget-constrained bidding has primarily focused on single-item auctions and has developed primal-dual based algorithms where the dual variable acts as a *pacing multiplier* that directly scales values to set bids (Balseiro and Gur, 2019; Gaitonde et al., 2023; Lucier et al., 2024). In contrast, we develop a coupled primal–dual framework tailored to multi-unit PAB auctions. Building on Castiglioni et al. (2022a), we integrate the dual variable into our DAG-based primal regret minimizer via dual-adjusted edge weights (cf. Eq. (16)), while updating the dual variable using online gradient descent. The resulting algorithm jointly balances utility maximization and budget feasibility, yielding ρ-approximate regret of order $\widetilde{O}(R_T/\rho)$, where R_T is the regret of the underlying primal algorithm.

Efficient Implementation. We next give efficient implementations of our learning algorithms for the full information setting and bandit setting with unknown context distribution that preserve the regret guarantees stated above while making the per-round time and space complexity *independent* of the number of contexts $|\mathcal{V}|$. Thus, the algorithms remain tractable even for very large or infinite context spaces. The key observation is that edge utilities are affine in the context and decompose across edges, allowing us to maintain edge-specific coefficients shared across all contexts. Instead of a separate DAG for each $\mathbf{v} \in \mathcal{V}$, we use a *super DAG* whose edge set contains the edges of all context-dependent DAGs and has the same order of size. We then update the shared coefficients on the super DAG each round, while keeping the core exponential-weights update via weight-pushing unchanged. The resulting per-round time and space complexity is $O(|\overline{\mathsf{E}}|)$, where $|\overline{\mathsf{E}}| = O(M^2/\epsilon^3)$ is the number of edges in the super DAG (see Section 5 for details).

[1] Strictly speaking, it suffices to reveal the top K *competing* bids (equivalently, the top $K + r_t$ bids overall), where r_t is the number of units won by the bidder in round t.

[2] Here, $\widetilde{O}(\cdot)$ hides logarithmic factors in $1/\epsilon$.

[3] Since $\mathbf{v} \in [0,1]^M$, the per-round utility lies in $[0, M]$. Rescaling utilities by $1/M$ to lie in $[0,1]$ (as is standard in online learning) scales all stated regret bounds by $1/M$.

1.2 Related Works

Balseiro and Gur (2019) showed that, under adversarially varying competing bids, no-regret learning is impossible even in second price auctions (SPA): no algorithm can guarantee a competitive ratio better than the average per-round budget, ρ. Castiglioni et al. (2022a) studied FPA and obtained ρ-approximate sublinear regret. Castiglioni et al. (2022b) proposed a unified meta-algorithm that achieves *best-of-both-worlds* regret guarantees—simultaneously robust to adversarial sequences and near-optimal in stochastic environments—for both FPA and SPA under budget and return-on-investment constraints. Castiglioni et al. (2024) showed that when ρ is unknown, similar guarantees can be obtained by using weakly adaptive primal and dual regret minimizers, i.e., algorithms that ensure sublinear regret on every interval $[t_1, t_2] \subseteq [T]$. We extend this line of work by studying repeated PAB auctions under budget constraints in the adversarial setting.

Recently, several works have studied learning to bid in repeated multi-unit auctions (Brânzei et al., 2023; Golrezaei and Sahoo, 2025; Potfer et al., 2024). The closest to ours is Galgana and Golrezaei (2024), who study repeated PAB auctions and provide polynomial-time learning algorithms for profit maximizers ($\alpha = 1$). While our work builds on and generalizes theirs, it differs substantially in the problem formulation, assumptions, algorithmic techniques, and learning guarantees. We defer a detailed comparison to Table 1 in Appendix A, which also discusses further related work on multi-unit auctions, online learning under resource constraints, and cross-learning in multi-armed bandits.

2 MODEL

Notations. For $n \in \mathbb{N}$, let $[n] = \{1, 2, \ldots, n\}$. Let $\mathbb{I}[\cdot]$ denote the indicator function, i.e., $\mathbb{I}[X] = 1$ if the proposition X is true and 0 otherwise. We write $X \lesssim Y$ if $X \leq CY$ for some absolute constant $C > 0$, and define $X \gtrsim Y$ analogously. For any $\epsilon \in (0, 1]$, let $\mathbb{Z}_\epsilon = \{k\epsilon : k \in \mathbb{Z}_{\geq 0}\}$. For a given integer M, define $\mathcal{B} := \left\{\mathbf{b} \in \mathbb{Z}_\epsilon^M \cap [0,1]^M : b_1 \geq \cdots \geq b_M\right\}$, i.e., the set of all nonincreasing bid vectors on the ϵ-grid.

2.1 Pay-as-Bid Auctions

In a PAB auction, there are K identical units and bidders may demand multiple units. We study the bidding problem from the perspective of a single bidder whose maximum demand is $M \in [K]$. The bidder has a private valuation vector $\mathbf{v} \in [0,1]^M$ with diminishing marginal returns, i.e., $v_1 \geq \cdots \geq v_M$ (Goldner et al., 2020). The valuation vector $\mathbf{v}$ is drawn from a distribution $\mathcal{D}$ supported on $\mathcal{V}$. For simplicity, we assume $\mathcal{V}$ is finite; in Section 5 we extend the results to infinite context spaces. For any $\mathbf{v}$, define the cumulative valuation vector $\mathbf{W} \in \mathbb{R}_+^M$ where $W_j = \sum_{\ell=1}^j v_\ell, \forall j \in [M]$.

Allocation and Payment Rule. The bidder submits a vector of bids $\mathbf{b} = [b_1, \ldots, b_M] \in \mathcal{B}$ sorted in non-increasing order. Each bid is an integral multiple of ϵ, reflecting the discretization used in real-world auctions (e.g., increments of \$0.01). For $\mathbf{b} = [b_1, \ldots, b_M]$, define $\mathbf{B} \in \mathcal{B}$ as the cumulative bid vector where $B_j = \sum_{\ell=1}^j b_\ell, \forall j$. The competing bids are denoted by $\boldsymbol{\beta}_-$ and the bid profile is $\boldsymbol{\beta} := (\mathbf{b}; \boldsymbol{\beta}_-)$. The auctioneer elicits bids from all the bidders and the multiset of the top K bids are called *winning* bids. Ties are resolved according to a fixed, publicly known deterministic rule, such as favoring lower indexed bidders (Nisan et al., 2011; Chiesa et al., 2015). The bidder is allocated one unit for each of their bids in the multiset of winning bids. The total number of units allocated to the bidder under bid profile $\boldsymbol{\beta}$ is denoted by $x(\boldsymbol{\beta})$. If the valuation vector is $\mathbf{v} = [v_1, \ldots, v_M]$, the total value obtained by acquiring $x(\boldsymbol{\beta})$ units is $\mathsf{V}_\mathbf{v}(\boldsymbol{\beta}) := W_{x(\boldsymbol{\beta})} = \sum_{j \leq x(\boldsymbol{\beta})} v_j$ and total payment is the sum of the accepted bids of the bidder, i.e., $\mathsf{P}(\boldsymbol{\beta}) := B_{x(\boldsymbol{\beta})} = \sum_{j \leq x(\boldsymbol{\beta})} b_j$.

The bidder intends to maximize their quasilinear utility function $\mathsf{U}_\mathbf{v}(\cdot)$. Formally, for a bid profile $\boldsymbol{\beta} = (\mathbf{b}; \boldsymbol{\beta}_-)$ and cost of capital $\alpha \in [0, 1]$,

$$\mathsf{U}_\mathbf{v}(\boldsymbol{\beta}) := \mathsf{V}_\mathbf{v}(\boldsymbol{\beta}) - \alpha \mathsf{P}(\boldsymbol{\beta}) \,. \tag{1}$$

The utility function in Eq. (1) captures two well-studied bidder behavioral model. The traditional profit maximizer model is obtained when $\alpha = 1$ (Börgers, 2015), while $\alpha = 0$ corresponds to the value maximizer model, which has gained attention in the context of autobidders (Balseiro et al., 2021a,b; Deng et al., 2024).

Definition 1. For a given $\mathbf{v}$, $\mathbf{b} = [b_1, b_2, \ldots, b_M]$ is a *no-overbidding (NOB)* strategy if for all $\ell \in [M]$, $B_\ell \leq W_\ell$, where $B_\ell = \sum_{j=1}^\ell b_j$ and $W_\ell = \sum_{j=1}^\ell v_j$. The collection of all NOB strategies corresponding to any $\mathbf{v} \in \mathcal{V}$ is $\mathcal{B}_\mathbf{v} \subseteq \mathcal{B}$.

Assumption 1. The bidder follows NOB strategies.

Our notion of NOB in Assumption 1 is standard in multi-unit auctions (Christodoulou et al., 2016). In prior work on profit-maximizing bidders, a more restrictive *per-unit* NOB condition of $b_j \leq v_j$ for all j can be imposed without loss of generality (Markakis and Telelis, 2015; Galgana and Golrezaei, 2024). But this restriction can be highly suboptimal when $\alpha < 1$ (see Appendix B.1). Assumption 1 ensures that $\mathsf{U}_\mathbf{v}(\boldsymbol{\beta}) \geq 0$ for all $\boldsymbol{\beta}_-$ and $\alpha \in [0, 1]$.[4] To see this, fix any $\mathbf{v}$, $\alpha \in [0, 1]$,

[4] If we normalize the utility of the outside option (not participating in the auction) to 0, this is referred to as the *individual rationality* (IR) or *participation* constraint.

and competing bids $\boldsymbol{\beta}_{-}$. For any NOB strategy $\mathbf{b}$, let $\boldsymbol{\beta} = (\mathbf{b}; \boldsymbol{\beta}_{-})$. Then $\mathsf{U}_{\mathbf{v}}(\boldsymbol{\beta}) = \mathsf{V}_{\mathbf{v}}(\boldsymbol{\beta}) - \alpha \mathsf{P}(\boldsymbol{\beta}) = W_{x(\boldsymbol{\beta})} - \alpha B_{x(\boldsymbol{\beta})} \geq 0$. Moreover, for any overbidding strategy, i.e., $\exists \ell \in [M]$ such that $B_\ell > W_\ell$, there exists a competing bid profile for which $\mathsf{U}_{\mathbf{v}}(\cdot) < 0$ (see Appendix B.2). As competing bids can be adversarial, bidders follow NOB strategies to ensure that $\mathsf{U}_{\mathbf{v}}(\cdot) \geq 0$.

Example 1. *Consider a PAB auction with $K = 5$ identical units and $\epsilon = 0.1$. The bidder's maximum demand is $M = 5$, and $\mathbf{v} = [1, 0.9, 0.7, 0.6, 0.4]$. Suppose their submitted strategy is $\mathbf{b} = [0.9, 0.9, 0.7, 0.6, 0.5]$. It can be verified that $B_j \leq W_j, \forall j \in [M]$ implying $\mathbf{b}$ is a NOB strategy. Suppose $\boldsymbol{\beta}_{-} = [1, 0.8, 0.4, 0.3, 0.2]$. Thus, the winning bids are $[1, \underline{0.9}, \underline{0.9}, 0.8, \underline{0.7}]$ and the bidder is allocated 3 units (corresponding to the underlined bids). Hence, for any $\alpha \in [0, 1]$, $\mathsf{V}_{\mathbf{v}}(\boldsymbol{\beta}) = 1 + 0.9 + 0.7 = 2.6$ and $\mathsf{P}(\boldsymbol{\beta}) = 0.9 + 0.9 + 0.7 = 2.5$ which implies that $\mathsf{U}_{\mathbf{v}}(\boldsymbol{\beta}) = 2.6 - 2.5\alpha > 0$.*

2.2 Problem Statement

We study the bidding problem for a single budget-constrained bidder in a repeated setting over T rounds, where a PAB auction is conducted in each round. The bidder has a fixed budget B, and the total expenditure across the T rounds must not exceed B. Following prior work on budget-constrained bidders, we assume that $B = \rho MT$ for some constant $\rho \geq 0$ (Balseiro and Gur, 2019; Gaitonde et al., 2023; Castiglioni et al., 2024).

In round $t \in [T]$, the bidder first observes their valuation vector $\mathbf{v}^t \in \mathcal{V}$, sampled i.i.d. from the distribution $\mathcal{D}$, and then submits $\mathbf{b}^t \in \mathcal{B}_{\mathbf{v}^t}$. If the competing bids are $\boldsymbol{\beta}^t_{-}$, the resulting bid profile is $\boldsymbol{\beta}^t = (\mathbf{b}^t; \boldsymbol{\beta}^t_{-})$. Given cost of capital $\alpha \in [0, 1]$ and allocation $x(\boldsymbol{\beta}^t)$ units, the bidder's value $\mathsf{V}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)$, payment $\mathsf{P}(\boldsymbol{\beta}^t)$, and utility $\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)$ are defined as before. After each auction, the bidder receives feedback and updates their bidding policy for future rounds. We consider two standard feedback models: in the *full-information setting*, all competing bids $\boldsymbol{\beta}^t_{-}$ are revealed, whereas in the *bandit setting*, only the allocation $x(\boldsymbol{\beta}^t)$ is observed.

Baseline. We compare the bidder's performance against the optimal stationary policy subject to budget constraints. Let Π be the class of policies that maps valuation vectors to bidding strategies:

$$\Pi = \left\{ \pi : \mathcal{V} \to \bigcup_{\mathbf{v} \in \mathcal{V}} \mathcal{B}_{\mathbf{v}}, \text{ s.t. } \pi(\mathbf{v}) \in \mathcal{B}_{\mathbf{v}} \right\}.$$

The baseline in this setting is the expected utility subject to budget constraints over T rounds:

$$\mathsf{OPT} := \max_{\pi \in \Pi} \sum_{t=1}^{T} \mathbb{E}\left[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t); \boldsymbol{\beta}^t_{-})\right]$$
$$\text{s.t.} \quad \sum_{t=1}^{T} \mathsf{P}(\pi(\mathbf{v}^t); \boldsymbol{\beta}^t_{-}) \leq \rho MT\,. \qquad (\textsc{Opt})$$

Recall that the competing bids are adversarially chosen. The baseline assumes prior knowledge of the competing bids and finds the optimal policy $\pi^* \in \Pi$, where the expectation is taken with respect to the context distribution $\mathcal{D}$. The choice of baseline is consistent with prior works on repeated auctions (Balseiro et al., 2019a; Schneider and Zimmert, 2023; Kumar et al., 2024).

Performance Metric. We distinguish two regimes depending on ρ: (i) $\rho \geq 1$ and (ii) $\rho < 1$. When $\rho \geq 1$, the budget constraint is automatically satisfied. This is because, under the NOB assumption and the auction's payment rule, we have $\mathsf{P}(\boldsymbol{\beta}^t) \leq \mathsf{V}_{\mathbf{v}^t}(\boldsymbol{\beta}^t) \leq M$ for all $\boldsymbol{\beta}^t_{-}$. Hence, the total payment over T rounds is at most MT implying the budget constraint never binds. Equivalently, we can treat this regime as unconstrained. The benchmark is

$$\mathsf{OPT}_{nb} := \max_{\pi \in \Pi} \sum_{t=1}^{T} \mathbb{E}\left[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t); \boldsymbol{\beta}^t_{-})\right], \qquad (\textsc{Opt-nb})$$

and performance is measured via (standard) regret:

$$\mathsf{Reg}_{nb}(T) := \mathsf{OPT}_{nb} - \sum_{t=1}^{T} \mathbb{E}\left[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)\right].$$

When $\rho < 1$, the problem resembles *adversarial bandits with knapsacks* (BwK) (Immorlica et al., 2022; Kesselheim and Singla, 2020). In this regime, the standard choice of performance metric is *ρ-approximate regret*:

$$\rho \cdot \mathsf{Reg}(T) := \rho \cdot \mathsf{OPT} - \sum_{t=1}^{T} \mathbb{E}\left[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)\right].$$

Our goal in the remainder of the paper is to design algorithms that guarantee sublinear (ρ-approximate) regret, i.e., $\mathsf{Reg}_{nb}(T) = o(T)$ and $\rho \cdot \mathsf{Reg}(T) = o(T)$.

3 NO BUDGET CONSTRAINT

As stated earlier, we first consider the regime $\rho \geq 1$, which is equivalent to the setting without budget constraints. As will become clear in Section 4, the algorithms developed for this regime are a crucial building block for handling the case $\rho < 1$.

3.1 Offline Setting

To design our online algorithms, we begin by solving the *offline* optimization problem in (Opt-nb), which provides key structural insights.

For a broad class of auctions, bidders adopt NOB strategies precisely to satisfy the IR constraint (Bhawalkar and Roughgarden, 2011; De Keijzer et al., 2013).

Lemma 3.1. *For $\rho \geq 1$, an optimal stationary policy $\pi^* \in \Pi$ for (Opt-nb) satisfies*

$$\pi^*(\mathbf{v}) \in \arg\max_{\mathbf{b}\in\mathcal{B}_\mathbf{v}} \sum_{t=1}^{T} \mathsf{U}_\mathbf{v}(\mathbf{b}; \boldsymbol{\beta}_{-}^{t}), \qquad \forall \mathbf{v} \in \mathcal{V}. \tag{2}$$

Motivated by Lemma 3.1, we study the offline optimization problem above for a fixed valuation vector $\mathbf{v} = [v_1, \ldots, v_M] \in \mathcal{V}$, which will in turn guide the design of our online algorithms.

Let $\boldsymbol{\beta}_{-,t}^{-(k)}$ denote the k^{th} smallest bid among the top K competing bids in round t, i.e., among all bids excluding those of the bidder under consideration. If the bidder wins r units in round t, then for each $s \in [r]$ it must hold that $b_s \geq \boldsymbol{\beta}_{-,t}^{-(s)}$ (assume the tie-breaking rule is incorporated in the indicator function in Eq. (3)). Therefore, for any $\mathbf{b} = [b_1, \ldots, b_M] \in \mathcal{B}_\mathbf{v}$,

$$\mathsf{U}_\mathbf{v}(\mathbf{b}; \boldsymbol{\beta}_{-}^{t}) = \sum_{j=1}^{M} (v_j - \alpha b_j) \cdot \mathbb{I}\left[b_j \geq \boldsymbol{\beta}_{-,t}^{-(j)}\right]. \tag{3}$$

Summing Eq. (3) over all $t \in [T]$ yields the objective in Eq. (2). Moreover, Eq. (3) shows that this objective decomposes across units, a property we exploit to solve the offline problem efficiently.

3.1.1 Constructing the DAG

Since $|\mathcal{B}_\mathbf{v}| = O(\epsilon^{-M})$, naïvely enumerating all strategies is infeasible. Instead, for each context $\mathbf{v} \in \mathcal{V}$ we construct a context-dependent edge-weighted DAG of size $\text{poly}(M, 1/\epsilon)$, which will be crucial for computing the optimal policy $\pi^*(\mathbf{v})$ in Lemma 3.1. In particular, computing $\pi^*(\mathbf{v})$ is equivalent to finding a shortest (minimum-weight) s–d path in this DAG. Finally, we introduce the notion of a *super DAG*—a slight modification of the context-dependent DAGs—which will be useful for the analysis in Section 3.2 and is central to our efficient implementations in Section 5.

Context-Dependent DAG. Fix $\mathbf{v} \in \mathcal{V}$. The (context-dependent) DAG $\mathcal{G}(\mathsf{N}_\mathbf{v}, \mathsf{E}_\mathbf{v}, \omega_\mathbf{v})$ consists of a source node s, a destination node d, and M intermediate layers. Layer ℓ corresponds to choosing the ℓ^{th} bid.

Nodes. A node in layer ℓ is a triple (ℓ, b_ℓ, s_ℓ), where $b_\ell \in \mathbb{Z}_\epsilon \cap [0, 1]$ and

$$s_\ell \in \mathbb{Z}_\epsilon \quad \text{and} \quad s_\ell \leq W_\ell, \tag{4}$$

recalling that $W_\ell = \sum_{j=1}^{\ell} v_j$ and $\mathbb{Z}_\epsilon$ denotes the set of nonnegative integral multiples of ϵ. The constraint $s_\ell \leq W_\ell$ enforces the NOB assumption. For convenience, define the source as $s = (0, \infty, 0)$.

Edges and weights. Edges exist only between consecutive layers. A directed edge from $(\ell - 1, b_{\ell-1}, s_{\ell-1})$ to (ℓ, b_ℓ, s_ℓ) exists if

$$b_{\ell-1} \geq b_\ell \quad \text{and} \quad s_\ell = s_{\ell-1} + b_\ell. \tag{5}$$

The weight edge $e = (\ell - 1, b_{\ell-1}, s_{\ell-1}) \to (\ell, b_\ell, s_\ell)$ is

$$\omega_\mathbf{v}(e) := \sum_{t=1}^{T} \frac{1 - (v_\ell - \alpha b_\ell) \cdot \mathbb{I}\left[b_\ell \geq \boldsymbol{\beta}_{-,t}^{-(\ell)}\right]}{1 + \alpha}. \tag{6}$$

All edges from nodes in layer M to the destination node d have weight 0.

Size of the DAG. In layer ℓ, we have $b_\ell \in \mathbb{Z}_\epsilon \cap [0, 1]$ and $s_\ell \in \mathbb{Z}_\epsilon$ with $s_\ell \leq W_\ell \leq \ell$. Thus, the number of possible values for b_ℓ is $O(1/\epsilon)$ and for s_ℓ is $O(\ell/\epsilon)$, so layer ℓ contains at most $O\big(\frac{1}{\epsilon} \cdot \frac{\ell}{\epsilon}\big) = O\big(\frac{\ell}{\epsilon^2}\big)$ nodes. Summing over $\ell \in [M]$ yields $|\mathsf{N}_\mathbf{v}| = O\left(\frac{M^2}{\epsilon^2}\right)$.

For edges, fix a node $(\ell - 1, b_{\ell-1}, s_{\ell-1})$ in layer $\ell - 1$. Condition in Eq. (5) implies that the choice of b_ℓ uniquely determines s_ℓ, so the out-degree is $O(1/\epsilon)$. Therefore, $|\mathsf{E}_\mathbf{v}| = O\big(|\mathsf{N}_\mathbf{v}| \cdot \frac{1}{\epsilon}\big) = O\left(\frac{M^2}{\epsilon^3}\right)$, and hence the DAG has size $\text{poly}(M, 1/\epsilon)$.

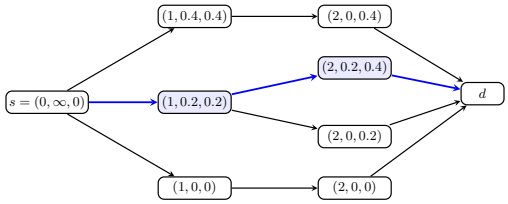

Figure 1: DAG for $M = 2$ with $\epsilon = 0.2$. Nodes are (ℓ, b_ℓ, s_ℓ) with $s_\ell = \sum_{j=1}^{\ell} b_j \leq W_\ell$, where $W_1 = 0.5$ and $W_2 = 0.55$. Each layer ℓ chooses one bid and enforces monotonicity ($b_{\ell-1} \geq b_\ell$) and NOB condition ($s_\ell \leq W_\ell$). The blue path corresponds to $\mathbf{b} = [0.2, 0.2]$.

We now state the main result of this section, which establishes a one-to-one correspondence between s–d paths in the context-dependent DAG $\mathcal{G}(\mathsf{N}_\mathbf{v}, \mathsf{E}_\mathbf{v}, \omega_\mathbf{v})$ and NOB strategies in $\mathcal{B}_\mathbf{v}$.

Theorem 3.1. *There is a bijection between s–d paths in $\mathcal{G}(\mathsf{N}_\mathbf{v}, \mathsf{E}_\mathbf{v}, \omega_\mathbf{v})$ and strategies in $\mathcal{B}_\mathbf{v}$.*

(Path $\leftrightarrow$ strategy). *A path $\mathfrak{p} = s \to (1, b_1, s_1) \to \cdots \to (M, b_M, s_M) \to d$ corresponds to the strategy $\mathbf{b} = [b_1, \ldots, b_M] \in \mathcal{B}_\mathbf{v}$. Conversely, any strategy $\mathbf{b} = [b_1, \ldots, b_M] \in \mathcal{B}_\mathbf{v}$ corresponds to the unique path $\mathfrak{p}$ stated earlier with $s_j = \sum_{\ell=1}^{j} b_\ell$ for all $j \in [M]$.*

(Path weight). *Let $\mathfrak{p}$ be the path corresponding to $\mathbf{b} = [b_1, \ldots, b_M]$. Then*

$$\omega_\mathbf{v}(\mathfrak{p}) := \sum_{e\in\mathfrak{p}} \omega_\mathbf{v}(e) = \frac{MT - \sum_{t=1}^{T} \mathsf{U}_\mathbf{v}(\mathbf{b}; \boldsymbol{\beta}_{-}^{t})}{1 + \alpha}.$$

In particular, maximizing $\sum_{t=1}^{T} \mathsf{U}_{\mathbf{v}}(\mathbf{b}; \boldsymbol{\beta}_{-}^{t})$ *over* $\mathbf{b} \in \mathcal{B}_{\mathbf{v}}$ *is equivalent to finding a shortest (minimum-weight)* s*-*d *path in* $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$*. Moreover, since* $|\mathsf{N}_{\mathbf{v}}| \lesssim |\mathsf{E}_{\mathbf{v}}|$*, a shortest path in* $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$ *can be computed in* $O(|\mathsf{E}_{\mathbf{v}}|) = O(M^2/\epsilon^3)$ *time and space.*

Super DAG. The *super DAG* $\mathcal{G}(\overline{\mathsf{N}}, \overline{\mathsf{E}})$ has the same layered structure as the context-dependent DAG $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$, but is independent of the context. In $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$, a node (ℓ, b_ℓ, s_ℓ) must satisfy $s_\ell \in \mathbb{Z}_\epsilon$ and $s_\ell \le W_\ell$, where $W_\ell = \sum_{j=1}^{\ell} v_j$. In the super DAG, we replace this with the context-free constraint

$$s_\ell \in \mathbb{Z}_\epsilon \quad \text{and} \quad s_\ell \le \ell. \tag{7}$$

Because $W_\ell \le \ell$ for all $\mathbf{v} \in \mathcal{V}$, every node and edge feasible in $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$ is also feasible in $\mathcal{G}(\overline{\mathsf{N}}, \overline{\mathsf{E}})$. So, for every $\mathbf{v} \in \mathcal{V}$, we have $\mathsf{N}_{\mathbf{v}} \subseteq \overline{\mathsf{N}}$ and $\mathsf{E}_{\mathbf{v}} \subseteq \overline{\mathsf{E}}$. Moreover, for any $\mathbf{v} \in \mathcal{V}$, $\mathcal{G}(\overline{\mathsf{N}}, \overline{\mathsf{E}})$ has the same order of size as $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$: $|\overline{\mathsf{N}}| = O(M^2/\epsilon^2)$ and $|\overline{\mathsf{E}}| = O(M^2/\epsilon^3)$.

3.2 Online Setting

We now build upon Theorem 3.1 to design a no-regret algorithm for the online setting, where a PAB auction is conducted in each round $t \in [T]$.

3.2.1 Full Information Feedback

In the full-information setting, after each round t the bidder observes the competing bid profile $\boldsymbol{\beta}_{-}^{t}$. In fact, it suffices to reveal only the top K competing bids, equivalently the top $K + r_t$ bids in $\boldsymbol{\beta}^t$, where r_t is the number of units won by the bidder in round t. We model the problem as contextual online learning: each valuation vector $\mathbf{v} \in \mathcal{V}$ is a context, and upon observing $\mathbf{v}$, the bidder selects a strategy in $\mathcal{B}_{\mathbf{v}}$. By Theorem 3.1, there is a bijection between strategies in $\mathcal{B}_{\mathbf{v}}$ and s–d paths in $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$. A natural idea is therefore to treat each path as an expert and run a no-regret algorithm such as Hedge (Freund and Schapire, 1997).[5] However, a naive implementation is computationally infeasible, since it requires tracking $O(\epsilon^{-M})$ experts. To circumvent this, we exploit the combinatorial structure of the action space and use a dynamic-programming-based variant of the weight-pushing method of Takimoto and Warmuth (2003), which maintains weights on edges rather than entire paths.

Overview of the Algorithm. The bidder maintains a context-dependent DAG $\mathcal{G}^t(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$ for each $\mathbf{v} \in \mathcal{V}$. At the beginning of round t, the bidder observes $\mathbf{v}^t \sim \mathcal{D}$ and updates the edge probabilities $\{\phi_{\mathbf{v}}^t(e)\}_{e \in \mathsf{E}_{\mathbf{v}}}$ for *all* $\mathbf{v} \in \mathcal{V}$ based on the previous-round values $\{\phi_{\mathbf{v}}^{t-1}(e)\}_{e \in \mathsf{E}_{\mathbf{v}}}$. Let $\omega_{\mathbf{v}}^t(e)$ denote the weight of edge $e \in \mathsf{E}_{\mathbf{v}}$ in round t (for any $\mathbf{v} \in \mathcal{V}$), and let η_t be the learning rate. Set $\eta_0 = 1$ and, for all $t \ge 1$, $\gamma_t := \frac{\eta_t}{\eta_{t-1}}$.

To compute $\phi_{\mathbf{v}}^t(\cdot)$, set $\Gamma_{\mathbf{v}}^{t-1}(d) = 1$ and compute $\Gamma_{\mathbf{v}}^{t-1}(\cdot)$ bottom-up as follows. For every $u \in \mathsf{N}_{\mathbf{v}}$,

$$\Gamma_{\mathbf{v}}^{t-1}(u) = \sum_{v : u \to v \in \mathsf{E}_{\mathbf{v}}} \Big(\Gamma_{\mathbf{v}}^{t-1}(v) \cdot [\phi_{\mathbf{v}}^{t-1}(u \to v)]^{\gamma_t} \cdot \exp(-\eta_t \, \omega_{\mathbf{v}}^{t-1}(u \to v)) \Big). \tag{8}$$

Then, for every edge $e = u \to v \in \mathsf{E}_{\mathbf{v}}$, update

$$\phi_{\mathbf{v}}^t(e) = [\phi_{\mathbf{v}}^{t-1}(e)]^{\gamma_t} \cdot \exp(-\eta_t \, \omega_{\mathbf{v}}^{t-1}(e)) \cdot \frac{\Gamma_{\mathbf{v}}^{t-1}(v)}{\Gamma_{\mathbf{v}}^{t-1}(u)}. \tag{9}$$

As we show in the proof of Theorem 3.2, the updates in Eqs. (8) and (9) recovers the standard Hedge algorithm by setting $\eta_t = \eta$ for all $t \in [T]$. When η_t time-decays, the update rule corresponds to the Decreasing Hedge algorithm (Mourtada and Gaïffas, 2019).

From Eq. (8) and Eq. (9), it is easy to verify that the edge probabilities $\phi_{\mathbf{v}}^t(u \to \cdot)$ form a valid distribution over the out-neighbors of u. This motivates sampling edges sequentially in a Markovian fashion over $M + 1$ steps. Starting at node s, select an outgoing edge $s \to u$ with probability $\phi_{\mathbf{v}^t}^t(s \to u)$ and transition to node u, repeating the process until the destination d is reached. The s-d path obtained in this manner maps to a strategy in $\mathcal{B}_{\mathbf{v}^t}$ as specified in Theorem 3.1, which the bidder then submits. After the auction concludes, the bidder observes $\boldsymbol{\beta}_{-}^{t}$ and sets the edge weights in all the DAGs. Specifically, for each $\mathbf{v} \in \mathcal{V}$, and edge $\mathsf{E}_{\mathbf{v}} \ni e = (\ell{-}1, b_{\ell-1}, s_{\ell-1}) \to (\ell, b_\ell, s_\ell)$ in layer $\ell \in [M]$, the edge weight is set as

$$\omega_{\mathbf{v}}^t(e) = \frac{1 - (v_\ell - \alpha b_\ell) \cdot \mathbb{I}\left[b_\ell \ge \boldsymbol{\beta}_{-,t}^{-(\ell)}\right]}{1 + \alpha}. \tag{10}$$

All edges from nodes in layer M to the destination node d have weight 0. Note that in this setting, the bidder does not require knowledge of the context distribution $\mathcal{D}$, and thus we may assume that $\mathcal{D}$ is unknown. The learning algorithm is formally presented in Algorithm 1.

Theorem 3.2. *In the full-information setting without budget constraint, Algorithm 1 runs in* $O(|\mathcal{V}|M^2/\epsilon^3)$ *space and time per round. For* $\epsilon \in (0, 1)$ *and any non-increasing sequence* $\{\eta_t\}_{t=1}^{T}$ *with* $\eta_t > 0$*,*

$$\mathsf{Reg}_{nb}(T) \lesssim \frac{M \log(1/\epsilon)}{\eta_T} + M^2 \sum_{t=1}^{T} \eta_t \,.$$

If $\eta_t = \sqrt{\frac{\log 1/\epsilon}{Mt}}, \forall t$*, then* $\mathsf{Reg}_{nb}(T) \lesssim M^{3/2}\sqrt{T \log 1/\epsilon}$*. Alternatively, if* $\eta_t = \eta = \sqrt{\frac{\log 1/\epsilon}{MT}}, \forall t$*, the same regret bound (up to constants) is also achieved.*[6]

[5] This is referred to as *Expanded Hedge* (Koolen et al., 2010) and as expanded exponential weights (EXP2) in Audibert et al. (2014).

[6] The variant with $\eta_t \propto t^{-1/2}$ is called an *anytime* algo-

Algorithm 1 No Budget Constraints (Full Information)

Require: Set of valuation vectors, $\mathcal{V}$, learning rates $\eta_t > 0, \forall t \geq 1$. Define $\eta_0 = 1, \phi^0_{\mathbf{v}}(e) = 1$ and $\omega^0_{\mathbf{v}}(e) = 0, \forall e \in \mathsf{E}_{\mathbf{v}}, \forall \mathbf{v} \in \mathcal{V}$.

1: **for** $t = 1, 2, \ldots$ **do**
2: Observe an i.i.d. valuation vector sample $\mathbf{v}^t \sim \mathcal{D}$.
3: Construct $\mathcal{G}^t(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$ without weights $\forall \mathbf{v} \in \mathcal{V}$.
4: **for** $\mathbf{v} \in \mathcal{V}$ **do**
5: Obtain edge probabilities $\phi^t_{\mathbf{v}}(\cdot), \forall \mathbf{v} \in \mathcal{V}$ following Eq. (8) and Eq. (9).
6: Define initial node $u = s$ and path $\mathfrak{p}^t = s$.
7: **while** $u \neq d$ **do**
8: Sample v with probability $\phi^t_{\mathbf{v}^t}(u \to v)$.
9: Append v to the path $\mathfrak{p}^t$; set $u \leftarrow v$.
10: Map $\mathfrak{p}^t = s \to (1, b_1, s_1) \to \cdots \to (M, b_M, s_M) \to d$, and submit $\mathbf{b}^t = [b_1, \ldots, b_M]$.
11: Set edge weights per Eq. (10) for all $\mathbf{v} \in \mathcal{V}$.

3.2.2 Bandit Feedback

Known Context Distribution. In the bandit setting, the bidder observes the context $\mathbf{v}^t$, selects a bid $\mathbf{b}^t \in \mathcal{B}_{\mathbf{v}^t}$, and after the auction observes only the allocation $x(\boldsymbol{\beta}^t)$. A natural approach is to treat this as a contextual bandit problem: maintain a context-dependent DAG for each $\mathbf{v} \in \mathcal{V}$ and update only the DAG corresponding to the realized context $\mathbf{v}^t$ using unbiased edge-weight estimators $\widehat{\omega}^t_{\mathbf{v}}(e)$.[7] However, this incurs an undesirable $O(\sqrt{|\mathcal{V}|})$ multiplicative dependence in the regret bound. We next exploit the structure of the utility function to obtain regret guarantees independent of $|\mathcal{V}|$.

We observe that the utility of a submitted strategy $\mathbf{b}$ reveals its counterfactual utility under every other context $\mathbf{v}' \in \mathcal{V}$. Specifically,

$$\mathsf{U}_{\mathbf{v}'}(\mathbf{b}; \boldsymbol{\beta}^t_-) = \mathsf{U}_{\mathbf{v}}(\mathbf{b}; \boldsymbol{\beta}^t_-) + \sum_{j=1}^{M} (v'_j - v_j) \cdot \mathbb{I}\left[b_j \geq \boldsymbol{\beta}^{-(j)}_{-,t}\right].$$

All quantities on the right-hand side are observable whenever $\mathbf{b}$ is submitted in round t, even under bandit feedback. Thus, submitting $\mathbf{b}$ reveals its utility under *all* contexts, yielding *complete cross-learning* (Balseiro et al., 2019a). This, in turn, allows us to construct edge-weight estimators for all context-dependent DAGs. Accordingly, the bidder runs Algorithm 1, replacing $\omega^t_{\mathbf{v}}(e)$ by its estimator $\widehat{\omega}^t_{\mathbf{v}}(e)$ as in Eq. (11).

In round t, suppose the bidder observes $\mathbf{v}^t$ and submits strategy $\mathbf{b}^t \in \mathcal{B}_{\mathbf{v}^t}$. If the corresponding path is $\mathfrak{p}^t$ (see Algorithm 1, Line 10), then for any $\mathbf{v} \in \mathcal{V}$ and $e \in \mathsf{E}_{\mathbf{v}}$,

$$\widehat{\omega}^t_{\mathbf{v}}(e) = \frac{\omega^t_{\mathbf{v}}(e)}{q^t(e)} \cdot \mathbb{I}\left[e \in \mathfrak{p}^t\right], \tag{11}$$

where $\omega^t_{\mathbf{v}}(e)$ is defined in Eq. (10), and

$$q^t(e) = \sum_{\mathbf{v} \in \mathcal{V}} \mathbb{P}[\mathbf{v}^t = \mathbf{v}] \sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}} : e \in \mathfrak{p}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}]\,. \tag{12}$$

Here, $\mathscr{P}_{\mathbf{v}}$ denotes the set of s–d paths in $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$, and $q^t(e)$ is the unconditional probability that edge e is selected in round t, averaged over $\mathbf{v} \sim \mathcal{D}$. Thus, computing Eq. (11) requires knowledge of the context distribution $\mathcal{D}$. The estimator resembles the standard importance-weighted bandit estimator, but with edge-marginal probability $q^t(e)$ in the denominator. The main computational bottleneck is therefore evaluating $q^t(e)$, which we show can be done in $O(|\mathcal{V}|M^2/\epsilon^3)$ time and space (see Appendix C.4).

Theorem 3.3. *In the bandit setting under known context distribution and no budget constraint, Algorithm 1 with the estimator in Eq. (11) runs in $O(|\mathcal{V}|M^2/\epsilon^3)$ space and time per round. For any $\epsilon \in (0,1)$ and non-increasing sequence $\eta_t > 0$,*

$$\mathsf{Reg}_{nb}(T) \lesssim \frac{M \log 1/\epsilon}{\eta_T} + \frac{M^3}{\epsilon^3} \sum_{t=1}^{T} \eta_t\,.$$

If $\eta_t = \frac{1}{M}\sqrt{\frac{\epsilon^3 \log 1/\epsilon}{t}}, \forall t$, $\mathsf{Reg}_{nb}(T) \lesssim \frac{M^2}{\epsilon^{3/2}}\sqrt{T \log 1/\epsilon}$. Alternatively, if $\eta_t = \eta = \frac{1}{M}\sqrt{\frac{\epsilon^3 \log 1/\epsilon}{T}}, \forall t$, the same regret bound (up to constants) is also achieved.

Unknown Context Distribution. Recall that knowledge of the distribution $\mathcal{D}$ was crucial for constructing the unbiased estimator $\widehat{\omega}^t_{\mathbf{v}}(e)$ in Eq. (11). For unknown $\mathcal{D}$, several changes are necessary.

Edge weights. For the purposes of regret analysis, we now maintain *gains* rather than losses. In any round t, for $\mathbf{v} \in \mathcal{V}$ and edge $e = (\ell - 1, b_{\ell-1}, s_{\ell-1}) \to (\ell, b_\ell, s_\ell)$ in layer $\ell \in [M]$, we define

$$\omega^t_{\mathbf{v}}(e) = \frac{\alpha + (v_\ell - \alpha b_\ell) \cdot \mathbb{I}\left[b_\ell \geq \boldsymbol{\beta}^{-(\ell)}_{-,t}\right]}{1 + \alpha}\,. \tag{13}$$

All edges from nodes in layer M to d have weight 0. As in Eq. (10), we have $\omega^t_{\mathbf{v}}(\cdot) \in [0,1]$ in Eq. (13) as well.

Estimator. If the observed context in round t is $\mathbf{v}^t = \mathbf{v}' \in \mathcal{V}$, then for any $\mathbf{v} \in \mathcal{V}$ and $e \in \mathsf{E}_{\mathbf{v}}$, we define

$$\widehat{\omega}^t_{\mathbf{v}}(e) = \frac{\omega^t_{\mathbf{v}}(e)}{p^t_{\mathbf{v}'}(e)} \cdot \mathbb{I}\left[e \in \mathfrak{p}^t\right], \tag{14}$$

where

$$p^t_{\mathbf{v}'}(e) = \sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}'} : e \in \mathfrak{p}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}']\,. \tag{15}$$

Here, $p^t_{\mathbf{v}'}(e)$ is the probability of selecting edge e in round t, conditional on the observed context $\mathbf{v}'$.

rithm (Lattimore and Szepesvári, 2020), since it attains the optimal regret bound with no prior knowledge of T.

[7] This is referred to as the S-EXP3 algorithm (Bubeck and Cesa-Bianchi, 2012).

Variance control. The estimator in Eq. (14) may suffer from high variance when $p^t_{\mathbf{v}'}(e)$ becomes too small. To mitigate this, we mix the exponential-weights distribution with a uniform distribution over a carefully chosen set of paths. For each $\mathbf{v} \in \mathcal{V}$, we construct a *edge path cover* $\mathcal{C}_{\mathbf{v}}$ (György et al., 2007; Golrezaei and Sahoo, 2025) which is a collection of s–d paths in $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$ such that every edge $e \in \mathsf{E}_{\mathbf{v}}$ lies on at least one path $\mathfrak{p} \in \mathcal{C}_{\mathbf{v}}$. In round t, after observing $\mathbf{v}^t = \mathbf{v}'$, the bidder selects a path according to a mixture strategy:

(i) with probability $\delta \in (0, 1]$, sample uniformly at random from $\mathcal{C}_{\mathbf{v}'}$;

(ii) with probability $1 - \delta$, sample using exponential weights $\propto \exp(\eta_t \sum_{s=1}^{t-1} \widehat{\omega}^s_{\mathbf{v}}(\mathfrak{p}))$.

Constructing an edge path cover (EPC). Since in any round t we only need the EPC for the realized context $\mathbf{v}^t$, we avoid constructing and storing one for every $\mathbf{v} \in \mathcal{V}$. Instead, when $\mathbf{v}^t = \mathbf{v}'$, we fix a deterministic *canonical* way to reach each node from the source and the destination from each node in $\mathcal{G}(\mathsf{N}_{\mathbf{v}'}, \mathsf{E}_{\mathbf{v}'}, \omega_{\mathbf{v}'})$ (e.g., via parent pointers from a topological pass). These pointers define, for each edge $e = (u, v)$, a canonical s–d path through e: take the canonical $s \to u$ prefix, traverse e, and then the canonical $v \to d$ suffix. Let $\mathcal{C}_{\mathbf{v}'}$ denote the resulting multiset of canonical s–d paths, one per edge, so that $|\mathcal{C}_{\mathbf{v}'}| = |\mathsf{E}_{\mathbf{v}'}|$. We then sample an edge uniformly at random using reservoir sampling in one pass over $\mathsf{E}_{\mathbf{v}'}$ and output its associated canonical path by following the pointers. This avoids explicitly materializing $\mathcal{C}_{\mathbf{v}'}$ and runs in $O(|\mathsf{E}_{\mathbf{v}'}|)$ time using $O(|\mathsf{N}_{\mathbf{v}'}|)$ space. The detailed sampling method appears in Algorithm 4, and the full procedure for the unknown context distribution setting is given in Algorithm 3.

Theorem 3.4. *In the bandit setting under unknown context distribution and no budget constraints, Algorithm 3 runs in $O(|\mathcal{V}|M^2/\epsilon^3)$ space and time per round. For any $\epsilon \in (0, 1)$, using $\delta \in (0, 1]$ and non-increasing sequences $\eta_t \le \frac{\delta}{M|\overline{\mathsf{E}}|}$, where $|\overline{\mathsf{E}}|$ is the number of edges in the super DAG, we get*

$$\mathsf{Reg}_{nb}(T) \lesssim \frac{M \log 1/\epsilon}{\eta_T} + \frac{M^2|\overline{\mathsf{E}}|}{\delta} \sum_{t=1}^{T} \eta_t + MT\delta\,.$$

If $\delta = \min\left(1, \left(\frac{M|\overline{\mathsf{E}}| \log 1/\epsilon}{T}\right)^{1/3}\right)$, and $\eta_t = \frac{\delta^2}{M|\overline{\mathsf{E}}|}$,

$$\mathsf{Reg}_{nb}(T) \lesssim \frac{M^2 T^{2/3} (\log 1/\epsilon)^{1/3}}{\epsilon} + \frac{M^4 \log 1/\epsilon}{\epsilon^3}\,.$$

Lower Bound. We conclude this section by establishing a regret lower bound showing that our learning algorithms in the full information setting and bandit setting with known context distribution are optimal in their dependence on T (up to logarithmic factors).

Theorem 3.5. *Fix any $\alpha \in (\frac{1}{2}, 1]$. There exists a sequence of competing bids $[\boldsymbol{\beta}^t_{-}]_{t\in[T]}$ such that, in the setting without budget constraints, every learning algorithm incurs $\mathbb{E}[\mathsf{Reg}_{nb}(T)] = \Omega(M\sqrt{T})$ in the full-information setting. This implies an equivalent lower bound in holds the bandit setting.*

4 WITH BUDGET CONSTRAINT

Having developed polynomial-time learning algorithms for the unconstrained regime, we now turn to the budget-constrained case, i.e., $\rho < 1$. Since the bids can be adversarial, the performance metric in this setting is *ρ-approximate regret*: $\rho \cdot \mathsf{Reg}(T) = \rho \cdot \mathsf{OPT} - \sum_{t=1}^{T} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)]$, where OPT is defined in (Opt).

Primal–dual algorithms are commonly used to *pace* bidders' spending under budget constraints: in each round, the primal step produces a bid by scaling the bidder's value using the current dual variable (a pacing multiplier), and the dual variable is updated via a gradient-based method (Balseiro and Gur, 2019; Gaitonde et al., 2023; Lucier et al., 2024). However, in multi-unit PAB auctions the optimal bid can be a nontrivial function of the entire valuation vector, so this simple "scale-the-value" primal update is no longer appropriate.

For ease of exposition, we focus on the full-information setting; the bandit case is deferred to Appendix C.7. The algorithm is inspired by the expression for the Lagrangian of the (Opt) which is of the form: $\mathcal{L}(\pi, \lambda) = \sum_{t=1}^{T} \{\mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t); \boldsymbol{\beta}^t_{-})] - \lambda \mathsf{P}(\pi(\mathbf{v}^t); \boldsymbol{\beta}^t_{-})\} + \lambda B$ where $\lambda \ge 0$. Following Castiglioni et al. (2022a, 2024); Fikioris and Tardos (2023), we can restrict $\lambda \in [0, \frac{1}{\rho}]$.

Overview of the Algorithm. We consider a primal-dual based algorithm consisting of two subroutines: a primal regret minimizer (primal RM) and a dual regret minimizer (dual RM).

Primal RM. We use Algorithm 1 as the primal RM with one crucial change: we replace Eq. (10) by dual-adjusted edge weights, i.e., for any $\mathbf{v} \in \mathcal{V}$, round t, and edge $e = (\ell - 1, b_{\ell-1}, s_{\ell-1}) \to (\ell, b_\ell, s_\ell)$ in layer $\ell \in [M]$, define

$$\omega^t_{\mathbf{v}}(e) := \frac{1 - (v_\ell - (\alpha + \lambda_t) b_\ell)\, \mathbb{I}\left[b_\ell \ge \boldsymbol{\beta}^{-(\ell)}_{-,t}\right]}{1 + \alpha + 1/\rho}, \quad (16)$$

where $\lambda_t \in [0, 1/\rho]$ is the dual variable. All edges from layer M to d have weight 0. Since $v_\ell, b_\ell \in [0, 1]$ and $\lambda_t \le 1/\rho$, $\omega^t_{\mathbf{v}}(e) \in [0, 1]$. To build intuition for Eq. (16), consider the unconstrained setting where the primal RM samples a path that maximizes utility. With dual-adjusted edge weights, the primal RM samples a path that maximizes utility minus λ_t times payment, thus penalizing strategies that result in higher spending.

Dual RM. We use the standard online gradient descent algorithm (Zinkevich, 2003) with affine cost functions $g_t \cdot \lambda$ as the dual RM where

$$g_t = -(\mathsf{P}(\mathbf{b}^t; \boldsymbol{\beta}_-^t) - \rho M) \in [-M, \rho M], \quad \forall t \geq 1\,.$$

The dual variables are updated following Eq. (17). Note that the dual RM always operates in the full information setting, independent of the primal RM's feedback model. The procedure is presented in Algorithm 2.

Algorithm 2 Budget Constraint (Full Information)

Require: Define $\lambda_1 = 0$, $B_1 = \rho MT$ and dual learning rates $\zeta_t = \frac{1}{\rho M\sqrt{t}}, \forall t \geq 1$.

for $t = 1, 2, \ldots$ **do**
  Observe an i.i.d. valuation vector sample $\mathbf{v}^t \sim \mathcal{D}$.
  if $B_t \geq M$ **then**
    Submit $\mathbf{b}^t \in \mathcal{B}_{\mathbf{v}^t}$ per Algorithm 1, Line 3 to 10.
    ▷ *Under bandit feedback, this line is altered. See details in Appendix C.7.*
  else
    Submit $\mathbf{b}^t = 0$.
  Observe $\mathsf{P}(\mathbf{b}^t; \boldsymbol{\beta}_-^t)$ and set: $B_{t+1} = B_t - \mathsf{P}(\mathbf{b}^t; \boldsymbol{\beta}_-^t)$.
  Set edge weights $\omega_{\mathbf{v}}^t(e)$ per Eq. (16) for all $\mathbf{v} \in \mathcal{V}$.
  Update the dual variable as follows:

$$\lambda_{t+1} = [\lambda_t + \zeta_t(\mathsf{P}(\mathbf{b}^t; \boldsymbol{\beta}_-^t) - \rho M)]_0^{1/\rho}\,. \qquad (17)$$

  Here, $[x]_a^b = \min(b, \max(x, a))$ is the Euclidean projection operator on the interval $[a, b]$.

Theorem 4.1. *Algorithm 2 runs in $O(|\mathcal{V}||\overline{\mathsf{E}}|) = O(|\mathcal{V}|M^2/\epsilon^3)$ space and time per round and for any $\epsilon \in (0, 1)$ achieves*

$$\rho \cdot \mathsf{Reg}(T) \lesssim \frac{M}{\rho} + \frac{M\sqrt{T}}{\rho} + \mathsf{R}_P^T,$$

where R_P^T is

- $\frac{M^{3/2}}{\rho}\sqrt{T \log 1/\epsilon}$ *in the full information setting.*
- $\frac{M^2}{\rho\epsilon^{3/2}}\sqrt{T \log 1/\epsilon}$ *in the bandit setting with known context distribution.*
- $\frac{M^2 T^{2/3}(\log 1/\epsilon)^{1/3}}{\rho\epsilon} + \frac{M^4 \log 1/\epsilon}{\rho\epsilon^3}$ *in the bandit setting with unknown context distribution.*

5 LARGE NUMBER OF CONTEXTS

In Section 3, we gave learning algorithms whose regret bounds are independent of the number of contexts $|\mathcal{V}|$, but whose per-round time and space complexity scale as $O(|\mathcal{V}|)$. This is acceptable when $|\mathcal{V}|$ is small, but in practice the context space may be large or even infinite. Now, we give implementations that preserve the same regret guarantees while making the per-round time and space complexity independent of $|\mathcal{V}|$. We focus on the unconstrained regime ($\rho \geq 1$), although the same implementation extends directly to the case $\rho < 1$.

We start with the super DAG $\mathcal{G}(\overline{\mathsf{N}}, \overline{\mathsf{E}})$ from Section 3, which has the same layered structure as the context-dependent DAGs $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$ but is independent of the context. Recall that, for every $\mathbf{v} \in \mathcal{V}$, $\mathsf{E}_{\mathbf{v}} \subseteq \overline{\mathsf{E}}$.

Key Idea. The efficient implementation relies on three properties: (i) utility is affine in the context, with *coefficients shared across all contexts*; (ii) utility decomposes over edges (Eq. (3)); and (iii) the weight-pushing algorithm efficiently maps cumulative edge weights to a sampling distribution over paths (Algorithm 1). As a result, after observing feedback *ex post*, each round only requires updating a common set of coefficients.

It remains to show that the edge weights are affine in the context. For ease of presentation, consider the full-information setting. For each $\mathbf{v} \in \mathcal{V}$ and each edge $e = (\ell - 1, b_{\ell-1}, s_{\ell-1}) \to (\ell, b_\ell, s_\ell)$ in layer $\ell \in [M]$,

$$\begin{aligned}\omega_{\mathbf{v}}^t(e) &= \frac{1 - (v_\ell - \alpha b_\ell) \cdot \mathbb{I}\left[b_\ell \geq \boldsymbol{\beta}_{-,t}^{-(\ell)}\right]}{1+\alpha} \\ &=: x^t(e)\, v_\ell + y^t(e). \qquad (\text{(10) restated})\end{aligned}$$

Edges from layer M to d have weight 0, where

$$\begin{aligned} x^t(e) &= -\frac{\mathbb{I}\left[b_\ell \geq \boldsymbol{\beta}_{-,t}^{-(\ell)}\right]}{1+\alpha}, \\ y^t(e) &= \frac{1 + \alpha b_\ell \mathbb{I}\left[b_\ell \geq \boldsymbol{\beta}_{-,t}^{-(\ell)}\right]}{1+\alpha}. \end{aligned} \qquad (18)$$

The coefficients $(x^t(e), y^t(e))$ are context-independent: they depend only on the edge e and the competing bids in round t. Therefore, once $\{x^t(e), y^t(e)\}_{e \in \overline{\mathsf{E}}}$ are known, the edge weights for any context-dependent DAG $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$ can be recovered for every $\mathbf{v} \in \mathcal{V}$. The details of the efficient implementation is presented in Appendix D.

6 CONCLUSION

We studied repeated bidding in multi-unit pay-as-bid auctions with budget constraints. In the unconstrained setting, we developed DAG-based algorithms with polynomial time and space complexity that obtain sublinear regret under multiple feedback models. For the budgeted setting, we built on these ideas to design a primal–dual framework achieving ρ-approximate sublinear regret. Important directions for future work include obtaining $O(\sqrt{T})$ regret in the bandit setting with unknown context distribution for combinatorially large action spaces, and understanding the equilibrium and market dynamics induced when all budget-constrained bidders use such no-regret learning algorithms.

ACKNOWLEDGEMENTS

N.G. and S.S. are partially supported by the MIT Junior Faculty Research Assistance Grant and the MIT Research Support Award. We thank the anonymous reviewers whose comments greatly improved the presentation of this manuscript.

References

Aggarwal, G., Fikioris, G., and Zhao, M. (2025). No-regret algorithms in non-truthful auctions with budget and roi constraints. In *Proceedings of the ACM on Web Conference 2025*, pages 1398–1415.

Agrawal, S. and Devanur, N. R. (2019). Bandits with global convex constraints and objective. *Operations Research*, 67(5):1486–1502.

Agrawal, S., Devanur, N. R., and Li, L. (2016). An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pages 4–18. PMLR.

Alvarez, F., Mazón, C., and André, F. J. (2019). Assigning pollution permits: are uniform auctions efficient? *Economic Theory*, 67(1):211–248.

Audibert, J.-Y., Bubeck, S., and Lugosi, G. (2014). Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45.

Ausubel, L. M., Cramton, P., Pycia, M., Rostek, M., and Weretka, M. (2014). Demand reduction and inefficiency in multi-unit auctions. *The Review of Economic Studies*, 81(4):1366–1400.

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

Badanidiyuru, A., Langford, J., and Slivkins, A. (2014). Resourceful contextual bandits. In *Conference on Learning Theory*, pages 1109–1134. PMLR.

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

Balseiro, S., Deng, Y., Mao, J., Mirrokni, V., and Zuo, S. (2021a). Robust auction design in the auto-bidding world. *Advances in Neural Information Processing Systems*, 34:17777–17788.

Balseiro, S., Deng, Y., Mao, J., Mirrokni, V. S., and Zuo, S. (2021b). The landscape of auto-bidding auctions: Value versus utility maximization. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 132–133.

Balseiro, S., Golrezaei, N., Mahdian, M., Mirrokni, V., and Schneider, J. (2019a). Contextual bandits with cross-learning. *Advances in Neural Information Processing Systems*, 32.

Balseiro, S., Golrezaei, N., Mirrokni, V., and Yazdanbod, S. (2019b). A black-box reduction in mechanism design with private cost of capital. *Available at SSRN 3341782*.

Balseiro, S. R., Besbes, O., and Weintraub, G. Y. (2015). Repeated auctions with budgets in ad exchanges: Approximations and design. *Management Science*, 61(4):864–884.

Balseiro, S. R. and Gur, Y. (2019). Learning in repeated auctions with budgets: Regret minimization and equilibrium. *Management Science*, 65(9):3952–3968.

Balseiro, S. R., Lu, H., and Mirrokni, V. (2023). The best of many worlds: Dual mirror descent for online allocation problems. *Operations Research*, 71(1):101–119.

Bhawalkar, K. and Roughgarden, T. (2011). Welfare guarantees for combinatorial auctions with item bidding. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 700–709. SIAM.

Börgers, T. (2015). *An introduction to the theory of mechanism design*. Oxford university press.

Brânzei, S., Derakhshan, M., Golrezaei, N., and Han, Y. (2023). Learning and collusion in multi-unit auctions. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Braverman, M., Liu, J., Mao, J., Schneider, J., and Xue, E. (2025). A new benchmark for online learning with budget-balancing constraints. *arXiv preprint arXiv:2503.14796*.

Brenner, M., Galai, D., and Sade, O. (2009). Sovereign debt auctions: Uniform or discriminatory? *Journal of Monetary Economics*, 56(2):267–274.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.

Castiglioni, M., Celli, A., and Kroer, C. (2022a). Online learning with knapsacks: the best of both worlds. In *International Conference on Machine Learning*, pages 2767–2783. PMLR.

Castiglioni, M., Celli, A., and Kroer, C. (2024). Online learning under budget and ROI constraints via weak adaptivity. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 5792–5816.

Castiglioni, M., Celli, A., Marchesi, A., Romano, G., and Gatti, N. (2022b). A unifying framework for online optimization with long-term constraints. *Advances in Neural Information Processing Systems*, 35:33589–33602.

Chen, Z., Wang, C., Wang, Q., Pan, Y., Shi, Z., Cai, Z., Ren, Y., Zhu, Z., and Deng, X. (2024). Dynamic budget throttling in repeated second-price auctions.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9598–9606.

Chiesa, A., Micali, S., and Zhu, Z. A. (2015). Knightian analysis of the vickrey mechanism. *Econometrica*, 83(5):1727–1754.

Christodoulou, G., Kovács, A., and Schapira, M. (2016). Bayesian combinatorial auctions. *Journal of the ACM (JACM)*, 63(2):1–19.

Clarke, E. H. (1971). Multipart pricing of public goods. *Public choice*, pages 17–33.

Cramton, P. and Ausubel, L. M. (2006). Dynamic auctions in procurement. *Handbook of Procurement*, pages 1–21.

Cramton, P. and Kerr, S. (2002). Tradeable carbon permit auctions: How and why to auction not grandfather. *Energy policy*, 30(4):333–345.

De Keijzer, B., Markakis, E., Schäfer, G., and Telelis, O. (2013). Inefficiency of standard multi-unit auctions. In *European Symposium on Algorithms*, pages 385–396. Springer.

Deng, Y., Golrezaei, N., Jaillet, P., Liang, J. C. N., and Mirrokni, V. (2024). Individual welfare guarantees in the autobidding world with machine-learned advice. In *Proceedings of the ACM on Web Conference 2024*, pages 267–275.

eBay (2026). Automatic bidding. Accessed: 2026-03-19.

Elsinger, H., Schmidt-Dengler, P., and Zulehner, C. (2019). Competition in treasury auctions. *American Economic Journal: Microeconomics*, 11(1):157–184.

Fabra, N., von der Fehr, N.-H., and Harbord, D. (2006). Designing electricity auctions. *The RAND Journal of Economics*, 37(1):23–46.

Feng, Y., Golrezaei, N., and Li, Q. (2026). Nash equilibria in uniform price auctions: Theory, computation, and market insights. In *Proceedings of the ACM SIGMETRICS 2026*. To appear.

Feng, Z., Padmanabhan, S., and Wang, D. (2023). Online bidding algorithms for return-on-spend constrained advertisers. In *Proceedings of the ACM Web Conference 2023*, pages 3550–3560.

Fikioris, G. and Tardos, É. (2023). Approximately stationary bandits with knapsacks. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3758–3782. PMLR.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Friedman, M. (1959). Testimony in employment, growth, and price levels. In *Hearings before the Joint Economic Committee, 86th Congress, 1st Session*, pages 3023–3026, Washington, D.C.

Gaitonde, J., Li, Y., Light, B., Lucier, B., and Slivkins, A. (2023). Budget pacing in repeated auctions: Regret and efficiency without convergence. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251, page 52.

Galgana, R. and Golrezaei, N. (2024). Learning in repeated multiunit pay-as-bid auctions. *Manufacturing & Service Operations Management*.

Garbade, K. and Ingber, J. (2005). The treasury auction process: Objectives, structure, and recent adaptations. *Structure, and Recent Adaptations*.

Garey, M. R. and Johnson, D. S. (1979). *Computers and intractability*, volume 174. freeman San Francisco.

Goldner, K., Immorlica, N., and Lucier, B. (2020). Reducing inefficiency in carbon auctions with imperfect competition. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Golrezaei, N. and Sahoo, S. (2025). Learning safe strategies for value maximizing buyers in uniform price auctions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 19841–19877. PMLR.

Groves, T. (1973). Incentives in teams. *Econometrica: Journal of the Econometric Society*, pages 617–631.

György, A., Linder, T., Lugosi, G., and Ottucsák, G. (2007). The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(10).

Han, Y., Weissman, T., and Zhou, Z. (2024). Optimal no-regret learning in repeated first-price auctions. *Operations Research*.

Hazan, E. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325.

Hortaçsu, A., Kastl, J., and Zhang, A. (2018). Bid shading and bidder surplus in the us treasury auction system. *American Economic Review*, 108(1):147–169.

Huang, R. and Huang, Z. (2025). High probability bound for cross-learning contextual bandits with unknown context distributions. In *Forty-second International Conference on Machine Learning*.

Immorlica, N., Sankararaman, K., Schapire, R., and Slivkins, A. (2022). Adversarial bandits with knapsacks. *Journal of the ACM*, 69(6):1–47.

Kasberger, B. and Woodward, K. (2025). Bidding in multi-unit auctions under limited information. *Journal of Economic Theory*, 226:106008.

Kesselheim, T. and Singla, S. (2020). Online learning with vector costs and bandits with knapsacks. In *Conference on Learning Theory*, pages 2286–2305. PMLR.

Koolen, W. M., Warmuth, M. K., Kivinen, J., et al.

(2010). Hedging structured concepts. In *COLT*, pages 93–105.

Kumar, R., Schneider, J., and Sivan, B. (2024). Strategically-robust learning algorithms for bidding in first-price auctions. In *Proceedings of the 25th ACM Conference on Economics and Computation*, page 893.

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

Lucier, B., Pattathil, S., Slivkins, A., and Zhang, M. (2024). Autobidders with budget and roi constraints: Efficiency, regret, and pacing dynamics. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247, pages 3642–3643.

Lv, H., Zhang, Z., Zheng, Z., Liu, J., Yu, C., Liu, L., Cui, L., and Wu, F. (2023). Utility maximizer or value maximizer: mechanism design for mixed bidders in online advertising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5789–5796.

Markakis, E. and Telelis, O. (2015). Uniform price auctions: Equilibria and efficiency. *Theory of Computing Systems*, 57:549–575.

Maurer, L. and Barroso, L. A. (2011). Electricity auctions: an overview of efficient practices.

Mourtada, J. and Gaïffas, S. (2019). On the optimality of the hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 20(83):1–28.

Nisan, N., Schapira, M., Valiant, G., and Zohar, A. (2011). Best-response auctions. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 351–360.

Nyborg, K. G., Rydqvist, K., and Sundaresan, S. M. (2002). Bidder behavior in multiunit auctions: Evidence from swedish treasury auctions. *Journal of Political Economy*, 110(2):394–424.

Potfer, M., Baudry, D., Richard, H., Perchet, V., and Wan, C. (2024). Improved learning rates in multi-unit uniform price auctions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Potfer, M. and Perchet, V. (2025). Comparing uniform price and discriminatory multi-unit auctions through regret minimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

Schneider, J. and Zimmert, J. (2023). Optimal cross-learning for contextual bandits with unknown context distributions. *Advances in Neural Information Processing Systems*, 36:51862–51880.

Takimoto, E. and Warmuth, M. K. (2003). Path kernels and multiplicative updates. *The Journal of Machine Learning Research*, 4:773–818.

Tierney, S. F., Schatzki, T., and Mukerji, R. (2008). Uniform-pricing versus pay-as-bid in wholesale electricity markets: does it make a difference? *New York ISO*.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 1st edition.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37.

Wang, Q., Yang, Z., Deng, X., and Kong, Y. (2023). Learning to bid in repeated first-price auctions with budgets. In *International Conference on Machine Learning*, pages 36494–36513.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Not Applicable**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes**
   (b) Complete proofs of all theoretical results. **Yes**
   (c) Clear explanations of any assumptions. **Yes**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Not Applicable**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Not Applicable**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Not Applicable**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Not Applicable**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Not Applicable**
   (b) The license information of the assets, if applicable. **Not Applicable**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable**
   (d) Information about consent from data providers/curators. **Not Applicable**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable**

Learning to Bid in Discriminatory Auctions with Budget Constraints: Supplementary Materials

Contents

A Additional Related Works

Multi-unit auctions are a special class of combinatorial auctions in which identical units of a single good are allocated. Examples of multi-unit auctions include emissions permit auctions (Cramton and Kerr, 2002; Alvarez et al., 2019), Treasury auctions (Nyborg et al., 2002; Garbade and Ingber, 2005; Elsinger et al., 2019), procurement auctions (Cramton and Ausubel, 2006), and wholesale electricity markets (Tierney et al., 2008; Fabra et al., 2006). The two most prominent payment rules in multi-unit auctions are the (a) *uniform price* rule, where each bidder pays the same per-unit price (usually the lowest accepted bid or the first rejected bid) and (b) *discriminatory price* rule, where each bidder pays their bid for each unit they receive. While the revenue ranking between these formats is often ambiguous (Baisa and Burkett, 2018; Ausubel et al., 2014), uniform price auctions are favored for their perceived fairness (since all bidders pay the same per-unit price) and simplified bidding (Friedman, 1959; Feng et al., 2026), whereas discriminatory price auctions are preferred for their transparency in payments.

For repeated uniform price auctions, Brânzei et al. (2023) study profit maximizers ($\alpha = 1$) and obtain sublinear regret bounds; these were later improved in the bandit setting by Potfer et al. (2024). Golrezaei and Sahoo (2025) study the same auctions for value maximizers ($\alpha = 0$) and give efficient learning algorithms with $O(\sqrt{T})$ regret under both full-information and bandit feedback. More recently, Potfer and Perchet (2025) compare the hardness of learning in uniform-price and PAB auctions for profit maximizers under stochastic competing bids.

Table 1: Comparison between Galgana and Golrezaei (2024) and this paper.

Facet	**Galgana and Golrezaei (2024)**	**This paper**
Bidder's objective	Profit-maximizing ($\alpha = 1$)	Cost-of-capital parameter $\alpha \in [0, 1]$
Notion of NOB	WLOG, Per-unit NOB ($b_j \le v_j$)	Must enforce *cumulative* NOB for general α; per-unit NOB can lose an $O(M)$ factor when $\alpha < 1$; see Appendix B.1
Offline optimization	Dynamic-programming graph based on per-unit NOB condition	Edge-weighted DAG whose nodes encode both bid values and the cumulative sum of bids along the path; edges enforce cumulative NOB. Naive extension of Galgana and Golrezaei (2024) becomes shortest path in a DAG with multiple constraints, which is NP-Hard in general (Garey and Johnson, 1979).
Budget constraints	–	Considered
Context distribution	Known	Both known and unknown
Horizon knowledge (T)	Assumes known T (or uses doubling trick for unknown T)	No prior horizon knowledge required when distribution is known. In the full information setting, this yields a *best-of-both-worlds* behavior: constant regret in the stochastic setting while remaining adversarially robust

Bidding in repeated second-price auctions (SPA) with budgets has been extensively studied in the stochastic setting (Balseiro et al., 2015; Balseiro and Gur, 2019; Balseiro et al., 2023; Feng et al., 2023), where $O(\sqrt{T})$ regret rate is achievable. These works leverage the truthfulness of SPA and adopt primal–dual frameworks in which the dual variables serve as *pacing multipliers*. Chen et al. (2024) studied an alternative budget management approach known as *throttling* and proposed a near-optimal throttling algorithm for SPA. For non-truthful auctions in the stochastic setting, Gaitonde et al. (2023) designed a learning algorithm with $O(T^{3/4})$ *pacing regret* relative to the optimal pacing multiplier, while Aggarwal et al. (2025) used the *best Lipschitz bidding function* as a benchmark and achieved $O(\sqrt{T})$ regret. Similarly, Wang et al. (2023) proposed a primal–dual algorithm for FPA that attains $O(\sqrt{T})$ regret against the best budget-feasible strategy.

Bidding in repeated environments naturally fits within the bandits with knapsack (BwK) framework, a multi-armed bandit problem with global resource constraints (Badanidiyuru et al., 2018; Immorlica et al., 2022; Kesselheim and Singla, 2020). The BwK model has been generalized to handle arbitrary reward and resource functions (Agrawal and Devanur, 2019), contextual information (Badanidiyuru et al., 2014; Agrawal et al., 2016), and combinatorial semi-bandit feedback (Sankararaman and Slivkins, 2018). In the stochastic setting, BwK admits vanishing $\widetilde{O}(\sqrt{T})$ regret, whereas in the adversarial case, no algorithm can achieve better than an $O(\log T)$ competitive ratio without additional assumptions. When the budget satisfies $B = \Omega(T)$, Castiglioni et al. (2022a) obtain ρ-approximate sublinear regret, where $\rho = \frac{B}{T}$. A related benchmark, conceptually similar to uniform spending, is studied by

Braverman et al. (2025), who also establish sublinear regret guarantees.

Balseiro et al. (2019a) introduced *cross learning* in contextual bandits, where the reward observed after playing an action in one context reveals (possibly partial) information about the reward that the same action would have obtained in other contexts. When such counterfactual rewards are revealed only for a subset of contexts, this is referred to as *partial* cross learning; when they are revealed for all contexts, it is termed *complete* cross learning. Balseiro et al. (2019a) model this structure via a directed graph over contexts and derive sublinear regret guarantees under both stochastic and adversarial reward and context models, with rates governed by graph-dependent parameters. Subsequently, Schneider and Zimmert (2023) studied the unknown context distribution setting with adversarial rewards and a polynomial number of arms, achieving an $O(\sqrt{T})$ regret bound (improving over the $O(T^{2/3})$ bound in Balseiro et al. (2019a) for this regime). More recently, Huang and Huang (2025) refined the analysis of the algorithm in Schneider and Zimmert (2023) and obtained high-probability guarantees.

B Omitted Details

B.1 Suboptimality of the Alternate No-Overbidding Rule

We give an example to show that the alternate notion of per-unit no-overbidding (NOB), $b_j \le v_j, \forall j \in [M]$, which is considered without loss of generality for profit maximizers ($\alpha = 1$), can be suboptimal in bidders with cost of capital $\alpha < 1$. To see this, fix $\alpha < 1$. For any $\epsilon \in (0, \frac{1}{20M}]$, let

$$\mathbf{v} = [1, 1 - 2M\epsilon, \dots, 1 - 2M\epsilon] \in \mathbb{R}_+^M .$$

Consider an auction with competing bid profile $\boldsymbol{\beta}_-$ in which each bid is $b = 1 - 2M\epsilon + \epsilon$. Any $\mathbf{b}$ which satisfies $b_\ell \le v_\ell, \forall \ell \in [M]$ can obtain at most 1 unit as $b_2 \le v_2 = 1 - 2M\epsilon < b$. To obtain the one possible unit, it is necessary that $b_1 \ge b$. Thus, the utility of the bidder is

$$\mathsf{U}_{\mathbf{v}}(\mathbf{b}; \boldsymbol{\beta}_-) = 1 - \alpha b_1 \le 1 - \alpha(1 - 2M\epsilon + \epsilon).$$

Now, define $\mathbf{b}' \in \mathcal{B}$ in which each entry is $1 - 2M\epsilon + 2\epsilon$. For any $j \in [M]$,

$$B_j = \sum_{\ell=1}^{j} b_\ell = (1 - 2M\epsilon + 2\epsilon)j \quad \text{and} \quad W_j = \sum_{\ell=1}^{j} v_\ell = 1 + (j-1)(1 - 2M\epsilon)$$

It can be verified that $\mathbf{b}'$ is a NOB strategy per Definition 1 as $B_j \le W_j, \forall j \in [M]$, and obtains M units since $b < 1 - 2M\epsilon + 2\epsilon$. Thus,

$$\mathsf{U}_{\mathbf{v}}(\mathbf{b}'; \boldsymbol{\beta}_-) = 1 + (M-1)(1 - 2M\epsilon) - \alpha M(1 - 2M\epsilon + 2\epsilon) .$$

Hence,

$$\begin{aligned} \frac{\mathsf{U}_{\mathbf{v}}(\mathbf{b}'; \boldsymbol{\beta}_-)}{\mathsf{U}_{\mathbf{v}}(\mathbf{b}; \boldsymbol{\beta}_-)} &\ge \frac{1 + (M-1)(1 - 2M\epsilon) - \alpha M(1 - 2M\epsilon + 2\epsilon)}{1 - \alpha(1 - 2M\epsilon + \epsilon)} = \frac{(1-\alpha)(M - 2M(M-1)\epsilon)}{1 - \alpha + \alpha(2M-1)\epsilon} \\ &\ge \frac{0.9M(1-\alpha)}{1 - 0.1\alpha} \gtrsim M . \end{aligned}$$

B.2 Overbidding May Result in Value < Payment

We show that for any overbidding strategy $\mathbf{b}$, there exists a competing-bid profile $\boldsymbol{\beta}_-$ such that $\mathsf{U}_{\mathbf{v}}(\boldsymbol{\beta}) < 0$. Consider a bidder with $\alpha = 1$. Recall that if $\mathbf{b}$ is an overbidding strategy, then there exists $\ell \in [M]$ such that $B_\ell > W_\ell$.

Fix such an ℓ and consider a competing-bid profile $\boldsymbol{\beta}_-$ in which the top $K - \ell$ competing bids are $2b_1$, and the remaining competing bids are all equal to some $\delta \in \left(0, \frac{b_M}{2}\right]$, where δ is an integer multiple of ϵ. Under this profile, the bidder wins exactly ℓ units. Hence, $\mathsf{V}_{\mathbf{v}}(\boldsymbol{\beta}) = W_\ell$ and $\mathsf{P}(\boldsymbol{\beta}) = \sum_{j=1}^{\ell} b_j = B_\ell$, which implies that $\mathsf{U}_{\mathbf{v}}(\boldsymbol{\beta}) = W_\ell - B_\ell < 0$.

C Omitted Proofs

C.1 Proof of Lemma 3.1

For any $\pi \in \Pi$, the objective function in Eq. (OPT) is

$$\sum_{t=1}^{T} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t); \boldsymbol{\beta}_-^t)] = \sum_{t=1}^{T} \sum_{\mathbf{v}\in\mathcal{V}} \mathbb{P}[\mathbf{v}^t = \mathbf{v}] \cdot \mathsf{U}_{\mathbf{v}}(\pi(\mathbf{v}); \boldsymbol{\beta}_-^t) = \sum_{\mathbf{v}\in\mathcal{V}} \mathbb{P}[\mathbf{v}^t = \mathbf{v}] \cdot \sum_{t=1}^{T} \mathsf{U}_{\mathbf{v}}(\pi(\mathbf{v}); \boldsymbol{\beta}_-^t),$$

where the last equality follows as $\mathbf{v}^t$ is sampled i.i.d. from the distribution $\mathcal{D}$ in each round. Since $\rho \geq 1$, the budget constraint is trivially satisfied. As the objective function is separable in $\mathbf{v}$, the optimal stationary policy π^* satisfies $\pi^*(\mathbf{v}) \in \operatorname{argmax}_{\pi(\mathbf{v})\in\mathcal{B}_{\mathbf{v}}} \sum_{t=1}^{T} \mathsf{U}_{\mathbf{v}}(\pi(\mathbf{v}); \boldsymbol{\beta}_-^t)$.

C.2 Proof of Theorem 3.1

(Path $\leftrightarrow$ strategy). In $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$, consider a s-d path,

$$\mathfrak{p} = s \to (1, b_1, s_1) \to \cdots \to (M, b_M, s_M) \to d, \tag{19}$$

mapped to the strategy $\mathbf{b} = [b_1, \ldots, b_M]$. We claim that $\mathbf{b} \in \mathcal{B}_{\mathbf{v}}$. Assume $b_0 = \infty$ and $s_0 = 0$.

By construction, $b_\ell \in \mathbb{Z}_\epsilon$ and $b_{\ell-1} \geq b_\ell, \forall \ell \in [M]$. Furthermore $s_\ell \overset{(5)}{=} s_{\ell-1} + b_\ell \implies s_\ell = \sum_{j=1}^{\ell} b_j$ as $s_0 = 0$. Hence, $B_\ell = \sum_{j=1}^{\ell} b_j = s_\ell \overset{(4)}{\leq} W_\ell, \forall \ell \in [M]$ which implies that $\mathbf{b}$ is a NOB strategy corresponding to the valuation vector $\mathbf{v}$ per Definition 1. Hence, $\mathbf{b} \in \mathcal{B}_{\mathbf{v}}$.

To show bijection, consider any $\mathbf{b} = [b_1, \ldots, b_M] \in \mathcal{B}_{\mathbf{v}}$. With this strategy, associate the s-d path stated in Eq. (19), where $S_j = \sum_{\ell=1}^{j} b_\ell, \forall j \in [M]$.

We show that $\mathfrak{p}$ is a valid path in the DAG, i.e., each of its nodes and edges exist in $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$. As $\mathbf{b} \in \mathcal{B}_{\mathbf{v}}$, $b_j \in \mathbb{Z}_\epsilon$ and $s_j = \sum_{\ell=1}^{j} b_\ell \leq W_j, \forall j \in [M]$. Hence, all nodes in $\mathfrak{p}$ exist in the DAG. Furthermore, $b_1 \geq \cdots \geq b_M$ and $s_\ell = s_{\ell-1} + b_\ell, \forall \ell \in [M]$ which implies that all the edges exist in $\mathfrak{p}$.

(Path weight). Let $\mathfrak{p} = s \to (1, b_1, s_1) \to \cdots \to (M, b_M, s_M) \to d$ be the path corresponding to $\mathbf{b} = [b_1, \ldots, b_M]$. The weight of path $\mathfrak{p}$ is

$$\begin{aligned}
\omega_{\mathbf{v}}(\mathfrak{p}) = \sum_{e\in\mathfrak{p}} \omega_{\mathbf{v}}(e) &= \sum_{j=1}^{M} \omega_{\mathbf{v}}((j-1, b_{j-1}, s_{j-1}) \to (j, b_j, s_j)) \\
&\overset{(6)}{=} \sum_{j=1}^{M} \sum_{t=1}^{T} \frac{1 - (v_j - \alpha b_j) \cdot \mathbb{I}\left[b_j \geq \boldsymbol{\beta}_{-,t}^{-(j)}\right]}{1+\alpha} \\
&\overset{(3)}{=} \frac{MT - \sum_{t=1}^{T} \mathsf{U}_{\mathbf{v}}(\mathbf{b}; \boldsymbol{\beta}_-^t)}{1+\alpha}.
\end{aligned}$$

Hence, maximizing $\sum_{t=1}^{T} \mathsf{U}_{\mathbf{v}}(\mathbf{b}; \boldsymbol{\beta}_-^t)$ over $\mathbf{b} \in \mathcal{B}_{\mathbf{v}}$ is equivalent to computing the shortest (minimum-weight) path in the DAG.

Space and Time Complexity. Since $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$ is a DAG, the shortest path in the DAG can be computed in $O(|\mathsf{N}_{\mathbf{v}}| + |\mathsf{E}_{\mathbf{v}}|) = O(|\mathsf{E}_{\mathbf{v}}|) = O(M^2/\epsilon^3)$ space and time complexity since $|\mathsf{N}_{\mathbf{v}}| \lesssim |\mathsf{E}_{\mathbf{v}}|$.

C.3 Proof of Theorem 3.2

Fix $\mathbf{v} \in \mathcal{V}$. Recall that the probability of selecting a path $\mathfrak{p}$ in round t given $\mathbf{v}^t = \mathbf{v}$ is

$$\mathbb{P}_{\mathbf{v}}^t(\mathfrak{p}) := \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}] = \prod_{e\in\mathfrak{p}} \phi_{\mathbf{v}}^t(e)\,. \tag{20}$$

For any edge $e = u \to v \in \mathsf{E}_{\mathbf{v}}$, the edge probabilities are

$$\phi^t_{\mathbf{v}}(e) = [\phi^{t-1}_{\mathbf{v}}(e)]^{\gamma_t} \cdot \exp(-\eta_t \omega^{t-1}_{\mathbf{v}}(e)) \cdot \frac{\Gamma^{t-1}_{\mathbf{v}}(v)}{\Gamma^{t-1}_{\mathbf{v}}(u)}, \qquad \text{((9) restated)}$$

where $\Gamma^{t-1}_{\mathbf{v}}(d) = 1$ and for $\Gamma^{t-1}_{\mathbf{v}}(\cdot)$ is computed recursively in a bottom-up fashion as follows:

$$\Gamma^{t-1}_{\mathbf{v}}(u) = \sum_{v : u \to v \in \mathsf{E}_{\mathbf{v}}} \Gamma^{t-1}_{\mathbf{v}}(v) \cdot [\phi^{t-1}_{\mathbf{v}}(u \to v)]^{\gamma_t} \cdot \exp(-\eta_t \omega^{t-1}_{\mathbf{v}}(u \to v)) \,. \qquad \text{((8) restated)}$$

Here, $\eta_0 = 1$ and

$$\gamma_t = \frac{\eta_t}{\eta_{t-1}} \quad \forall t \geq 1 \,. \tag{21}$$

Now, consider a naïve implementation of the Decreasing Hedge algorithm. In this case, the probability of selecting path $\mathfrak{p}$ in round t given $\mathbf{v}^t = \mathbf{v}$ is

$$\widehat{\mathbb{P}}^t_{\mathbf{v}}(\mathfrak{p}) = \frac{\exp(-\eta_t \sum_{s=1}^{t-1} \omega^s_{\mathbf{v}}(\mathfrak{p}))}{\sum_{\mathfrak{p}' \in \mathscr{P}_{\mathbf{v}}} \exp(-\eta_t \sum_{s=1}^{t-1} \omega^s_{\mathbf{v}}(\mathfrak{p}'))}, \tag{22}$$

where $\mathscr{P}_{\mathbf{v}}$ is the set of all s-d paths in $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$ and $\omega^s_{\mathbf{v}}(\mathfrak{p}) = \sum_{e \in \mathfrak{p}} \omega^s_{\mathbf{v}}(e)$. Note that if $\eta_t = \eta, \forall t$, then Eq. (22) describes the action selection probability under the classical Hedge algorithm.

Lemma C.1. *For any $t \in [T]$, $\mathbf{v} \in \mathcal{V}$, and $\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}$, $\mathbb{P}^t_{\mathbf{v}}(\mathfrak{p}) = \widehat{\mathbb{P}}^t_{\mathbf{v}}(\mathfrak{p})$. In words, in any round t, and $\mathbf{v} \in \mathcal{V}$, the probability of choosing path $\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}$ under Algorithm 1 is same as that under the naïve Decreasing Hedge algorithm.*

We first establish a few preliminaries necessary to prove Lemma C.1.

Claim 1. For any $t \in [T]$, $\mathbf{v} \in \mathcal{V}$, and $\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}$,

$$\widehat{\mathbb{P}}^t_{\mathbf{v}}(\mathfrak{p}) = \frac{[\widehat{\mathbb{P}}^{t-1}_{\mathbf{v}}(\mathfrak{p})]^{\gamma_t} \exp(-\eta_t \omega^{t-1}_{\mathbf{v}}(\mathfrak{p}))}{\sum_{\mathfrak{p}' \in \mathscr{P}_{\mathbf{v}}} [\widehat{\mathbb{P}}^{t-1}_{\mathbf{v}}(\mathfrak{p}')]^{\gamma_t} \exp(-\eta_t \omega^{t-1}_{\mathbf{v}}(\mathfrak{p}'))},$$

where γ_t is per Eq. (21), where we define $\widehat{\mathbb{P}}^0_{\mathbf{v}}(\mathfrak{p}) := 1$ and $\omega^0_{\mathbf{v}}(\mathfrak{p}) = 0$ for all $\mathbf{v} \in \mathcal{V}$ and $\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}$.

Proof. For any $t \in [T]$, $\mathbf{v} \in \mathcal{V}$, and $\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}$,

$$\begin{aligned}
\widehat{\mathbb{P}}^t_{\mathbf{v}}(\mathfrak{p}) &\overset{(22)}{=} \frac{\exp(-\eta_t \sum_{s=1}^{t-1} \omega^s_{\mathbf{v}}(\mathfrak{p}))}{\sum_{\mathfrak{p}' \in \mathscr{P}_{\mathbf{v}}} \exp(-\eta_t \sum_{s=1}^{t-1} \omega^s_{\mathbf{v}}(\mathfrak{p}'))} \\
&= \frac{\exp(-\eta_t \sum_{s=1}^{t-2} \omega^s_{\mathbf{v}}(\mathfrak{p})) \cdot \exp(-\eta_t \omega^{t-1}_{\mathbf{v}}(\mathfrak{p}))}{\sum_{\mathfrak{p}' \in \mathscr{P}_{\mathbf{v}}} \exp(-\eta_t \sum_{s=1}^{t-2} \omega^s_{\mathbf{v}}(\mathfrak{p}')) \cdot \exp(-\eta_t \omega^{t-1}_{\mathbf{v}}(\mathfrak{p}'))} \\
&= \frac{\exp(-\frac{\eta_t}{\eta_{t-1}} \cdot \eta_{t-1} \sum_{s=1}^{t-2} \omega^s_{\mathbf{v}}(\mathfrak{p})) \cdot \exp(-\eta_t \omega^{t-1}_{\mathbf{v}}(\mathfrak{p}))}{\sum_{\mathfrak{p}' \in \mathscr{P}_{\mathbf{v}}} \exp(-\frac{\eta_t}{\eta_{t-1}} \cdot \eta_{t-1} \sum_{s=1}^{t-2} \omega^s_{\mathbf{v}}(\mathfrak{p}')) \cdot \exp(-\eta_t \omega^{t-1}_{\mathbf{v}}(\mathfrak{p}'))} \\
&= \frac{[\widehat{\mathbb{P}}^{t-1}_{\mathbf{v}}(\mathfrak{p})]^{\gamma_t} \exp(-\eta_t \omega^{t-1}_{\mathbf{v}}(\mathfrak{p}))}{\sum_{\mathfrak{p}' \in \mathscr{P}_{\mathbf{v}}} [\widehat{\mathbb{P}}^{t-1}_{\mathbf{v}}(\mathfrak{p}')]^{\gamma_t} \exp(-\eta_t \omega^{t-1}_{\mathbf{v}}(\mathfrak{p}'))},
\end{aligned}$$

where the last line follows from Eq. (22) corresponding round $t-1$. □

Claim 2. For any node $u \in \mathsf{N}_{\mathbf{v}}$, let $\mathscr{P}_{\mathbf{v}}(u)$ be the set of paths starting at u and terminating in d. Then,

$$\Gamma^{t-1}_{\mathbf{v}}(u) = \sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}(u)} \prod_{e \in \mathfrak{p}} [\phi^{t-1}_{\mathbf{v}}(e)]^{\gamma_t} \cdot \exp(-\eta_t \omega^{t-1}_{\mathbf{v}}(e)) \,.$$

Proof. We prove the result by backward induction. For the base case, $\Gamma_{\mathbf{v}}^{t-1}(d) = 1$. Suppose the result holds true for all the nodes in layer $\ell+1$ for some $0 \le \ell \le M$. Then, for any node u in layer ℓ,

$$\begin{aligned}\Gamma_{\mathbf{v}}^{t-1}(u) &\overset{(8)}{=} \sum_{v:u\to v\in \mathsf{E}_{\mathbf{v}}} \Gamma_{\mathbf{v}}^{t-1}(v)\cdot[\phi_{\mathbf{v}}^{t-1}(u\to v)]^{\gamma_t}\cdot\exp(-\eta_t\omega_{\mathbf{v}}^{t-1}(u\to v))\\ &= \sum_{v:u\to v\in \mathsf{E}_{\mathbf{v}}}\Big(\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}(v)}\prod_{e\in\mathfrak{p}}[\phi_{\mathbf{v}}^{t-1}(e)]^{\gamma_t}\cdot\exp(-\eta_t\omega_{\mathbf{v}}^{t-1}(e))\Big)\cdot[\phi_{\mathbf{v}}^{t-1}(u\to v)]^{\gamma_t}\cdot\exp(-\eta_t\omega_{\mathbf{v}}^{t-1}(u\to v))\\ &= \sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}(u)}\prod_{e\in\mathfrak{p}}[\phi_{\mathbf{v}}^{t-1}(e)]^{\gamma_t}\cdot\exp(-\eta_t\omega_{\mathbf{v}}^{t-1}(e))\,.\end{aligned}$$

Here, the second equality follows from the induction hypothesis. □

Having all the necessary pieces, we now prove Lemma C.1 by induction on round t.

Proof of Lemma C.1. For $t=1$, $\widehat{\mathbb{P}}_{\mathbf{v}}^1(\mathfrak{p}) \overset{(22)}{=} \frac{1}{|\mathscr{P}_{\mathbf{v}}|}$. For $t=1$,

$$\mathbb{P}_{\mathbf{v}}^1(\mathfrak{p}) \overset{(20)}{=} \prod_{e\in\mathfrak{p}}\phi_{\mathbf{v}}^1(e) \overset{(9)}{=} \frac{\Gamma_{\mathbf{v}}^0(d)}{\Gamma_{\mathbf{v}}^0(s)},$$

where the last equality holds as $\omega_{\mathbf{v}}^0(\cdot)=0$ and $\phi_{\mathbf{v}}^0(\cdot)=1$. Recall that $\Gamma_{\mathbf{v}}^0(d)=1$ and by Claim 2,

$$\Gamma_{\mathbf{v}}^0(s) = \sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}(s)} 1 = |\mathscr{P}_{\mathbf{v}}(s)| = |\mathscr{P}_{\mathbf{v}}| \implies \mathbb{P}_{\mathbf{v}}^1(\mathfrak{p}) = \frac{1}{|\mathscr{P}_{\mathbf{v}}|}\,.$$

Hence, we have that $\mathbb{P}_{\mathbf{v}}^1(\mathfrak{p}) = \widehat{\mathbb{P}}_{\mathbf{v}}^1(\mathfrak{p})$. Suppose the result holds for any round $t-1$, i.e., $\mathbb{P}_{\mathbf{v}}^{t-1}(\mathfrak{p}) = \widehat{\mathbb{P}}_{\mathbf{v}}^{t-1}(\mathfrak{p})$. Then,

$$\begin{aligned}\mathbb{P}_{\mathbf{v}}^t(\mathfrak{p}) &\overset{(20)}{=} \prod_{e\in\mathfrak{p}}\phi_{\mathbf{v}}^t(e)\\ &\overset{(9)}{=} \prod_{e=u\to v\in\mathfrak{p}}[\phi_{\mathbf{v}}^{t-1}(e)]^{\gamma_t}\cdot\exp(-\eta_t\omega_{\mathbf{v}}^{t-1}(e))\cdot\frac{\Gamma_{\mathbf{v}}^{t-1}(v)}{\Gamma_{\mathbf{v}}^{t-1}(u)}\\ &= [\widehat{\mathbb{P}}_{\mathbf{v}}^{t-1}(\mathfrak{p})]^{\gamma_t}\cdot\exp(-\eta_t\omega_{\mathbf{v}}^{t-1}(\mathfrak{p}))\cdot\frac{\Gamma_{\mathbf{v}}^{t-1}(d)}{\Gamma_{\mathbf{v}}^{t-1}(s)},\end{aligned}$$

where the last line follows as $\widehat{\mathbb{P}}_{\mathbf{v}}^{t-1}(\mathfrak{p}) = \mathbb{P}_{\mathbf{v}}^{t-1}(\mathfrak{p}) = \prod_{e\in\mathfrak{p}}\phi_{\mathbf{v}}^{t-1}(e)$. Now, recall that $\Gamma_{\mathbf{v}}^{t-1}(d)=1$ and by Claim 2,

$$\Gamma_{\mathbf{v}}^{t-1}(s) = \sum_{\mathfrak{p}'\in\mathscr{P}_{\mathbf{v}}}\prod_{e\in\mathfrak{p}'}[\phi_{\mathbf{v}}^{t-1}(e)]^{\gamma_t}\cdot\exp(-\eta_t\omega_{\mathbf{v}}^{t-1}(e)) = \sum_{\mathfrak{p}'\in\mathscr{P}_{\mathbf{v}}}[\widehat{\mathbb{P}}_{\mathbf{v}}^{t-1}(\mathfrak{p}')]^{\gamma_t}\cdot\exp(-\eta_t\omega_{\mathbf{v}}^{t-1}(\mathfrak{p}'))\,.$$

Combining everything together and using Claim 1, we get that

$$\mathbb{P}_{\mathbf{v}}^t(\mathfrak{p}) = \frac{[\widehat{\mathbb{P}}_{\mathbf{v}}^{t-1}(\mathfrak{p})]^{\gamma_t}\cdot\exp(-\eta_t\omega_{\mathbf{v}}^{t-1}(\mathfrak{p}))}{\sum_{\mathfrak{p}'\in\mathscr{P}_{\mathbf{v}}}[\widehat{\mathbb{P}}_{\mathbf{v}}^{t-1}(\mathfrak{p}')]^{\gamma_t}\cdot\exp(-\eta_t\omega_{\mathbf{v}}^{t-1}(\mathfrak{p}'))} = \widehat{\mathbb{P}}_{\mathbf{v}}^t(\mathfrak{p}),$$

which is the desired result. □

Regret Analysis. For any $\mathbf{v}\in\mathcal{V}$ and $\eta>0$, define

$$\Phi_{\mathbf{v}}^t(\eta) := \frac{1}{\eta}\log\Big(\frac{1}{|\mathscr{P}_{\mathbf{v}}|}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\exp\Big(-\eta\sum_{s=1}^t\omega_{\mathbf{v}}^s(\mathfrak{p})\Big)\Big)$$

Then,

$$\Phi_{\mathbf{v}}^t(\eta_t) - \Phi_{\mathbf{v}}^{t-1}(\eta_t) = \frac{1}{\eta_t}\log\left(\frac{\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\exp(-\eta_t\sum_{s=1}^t\omega_{\mathbf{v}}^s(\mathfrak{p}))}{\sum_{\mathfrak{p}'\in\mathscr{P}_{\mathbf{v}}}\exp(-\eta_t\sum_{s=1}^{t-1}\omega_{\mathbf{v}}^s(\mathfrak{p}'))}\right)$$

$$= \frac{1}{\eta_t} \log \left(\frac{\sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}} \exp(-\eta_t \sum_{s=1}^{t-1} \omega_{\mathbf{v}}^s(\mathfrak{p})) \cdot \exp(-\eta_t \omega_{\mathbf{v}}^t(\mathfrak{p}))}{\sum_{\mathfrak{p}' \in \mathscr{P}_{\mathbf{v}}} \exp(-\eta_t \sum_{s=1}^{t-1} \omega_{\mathbf{v}}^s(\mathfrak{p}'))} \right)$$

$$\overset{(22)}{=} \frac{1}{\eta_t} \log \left(\sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}} \widehat{\mathbb{P}}_{\mathbf{v}}^t(\mathfrak{p}) \cdot \exp(-\eta_t \omega_{\mathbf{v}}^t(\mathfrak{p})) \right)$$

$$= \frac{1}{\eta_t} \log \left(\sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}} \mathbb{P}_{\mathbf{v}}^t(\mathfrak{p}) \cdot \exp(-\eta_t \omega_{\mathbf{v}}^t(\mathfrak{p})) \right) .$$

By Eq. (20), $\mathbb{P}_{\mathbf{v}}^t(\mathfrak{p}) = \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}]$. For $x \geq 0$, $e^{-x} \leq 1 - x + x^2$. Thus,

$$\Phi_{\mathbf{v}}^t(\eta_t) - \Phi_{\mathbf{v}}^{t-1}(\eta_t) \leq \frac{1}{\eta_t} \log \left(\sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}](1 - \eta_t \omega_{\mathbf{v}}^t(\mathfrak{p}) + \eta_t^2 \omega_{\mathbf{v}}^t(\mathfrak{p})^2) \right)$$

$$= \frac{1}{\eta_t} \log \left(1 + \sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}](-\eta_t \omega_{\mathbf{v}}^t(\mathfrak{p}) + \eta_t^2 \omega_{\mathbf{v}}^t(\mathfrak{p})^2) \right) .$$

Since $-\eta_t \omega_{\mathbf{v}}^t(\mathfrak{p}) + \eta_t^2 \omega_{\mathbf{v}}^t(\mathfrak{p})^2 \geq -\frac{1}{4}$, $\sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}](-\eta_t \omega_{\mathbf{v}}^t(\mathfrak{p}) + \eta_t^2 \omega_{\mathbf{v}}^t(\mathfrak{p})^2) \geq -\frac{1}{4}$. Moreover, $\log(1+x) \leq x$ for all $x > -1$. Thus,

$$\Phi_{\mathbf{v}}^t(\eta_t) - \Phi_{\mathbf{v}}^{t-1}(\eta_t) \leq - \sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}] \omega_{\mathbf{v}}^t(\mathfrak{p}) + \eta_t \sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}] \omega_{\mathbf{v}}^t(\mathfrak{p})^2$$

Summing from $t = 1$ to $t = T$ and using $\Phi_{\mathbf{v}}^0(\eta_1) = 0$, the left hand side becomes

$$\sum_{t=1}^{T} \left(\Phi_{\mathbf{v}}^t(\eta_t) - \Phi_{\mathbf{v}}^{t-1}(\eta_t) \right) = \Phi_{\mathbf{v}}^T(\eta_T) - \Phi_{\mathbf{v}}^0(\eta_1) + \sum_{t=1}^{T-1} \left(\Phi_{\mathbf{v}}^t(\eta_t) - \Phi_{\mathbf{v}}^t(\eta_{t+1}) \right) .$$

By Theorem 3.1, there is a bijection between paths $\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}$ and strategies $\mathbf{b} \in \mathcal{B}_{\mathbf{v}}$. Recall that

$$\Pi = \left\{ \pi : \mathcal{V} \to \bigcup_{\mathbf{v} \in \mathcal{V}} \mathcal{B}_{\mathbf{v}} \text{ s.t. } \pi(\mathbf{v}) \in \mathcal{B}_{\mathbf{v}} \right\}.$$

Composing this bijection with the policy mapping, we define a path-based policy class as follows:

$$\widetilde{\Pi} := \left\{ \widetilde{\pi} : \mathcal{V} \to \bigcup_{\mathbf{v} \in \mathcal{V}} \mathscr{P}_{\mathbf{v}}, \text{ s.t. } \widetilde{\pi}(\mathbf{v}) \in \mathscr{P}_{\mathbf{v}} \right\}. \tag{23}$$

By construction, there is a one-to-one correspondence between $\pi \in \Pi$ and $\widetilde{\pi} \in \widetilde{\Pi}$. So, for any $\widetilde{\pi} \in \widetilde{\Pi}$,

$$\Phi_{\mathbf{v}}^T(\eta_T) \geq \frac{1}{\eta_T} \log \left(\frac{1}{|\mathscr{P}_{\mathbf{v}}|} \cdot \exp \left(-\eta_T \sum_{t=1}^{T} \omega_{\mathbf{v}}^t(\widetilde{\pi}(\mathbf{v})) \right) \right) = -\frac{\log |\mathscr{P}_{\mathbf{v}}|}{\eta_T} - \sum_{t=1}^{T} \omega_{\mathbf{v}}^t(\widetilde{\pi}(\mathbf{v}))$$

$$\implies \sum_{t=1}^{T} \left(\Phi_{\mathbf{v}}^t(\eta_t) - \Phi_{\mathbf{v}}^{t-1}(\eta_t) \right) \geq -\frac{\log |\mathscr{P}_{\mathbf{v}}|}{\eta_T} - \sum_{t=1}^{T} \omega_{\mathbf{v}}^t(\widetilde{\pi}(\mathbf{v})) + \sum_{t=1}^{T-1} \left(\Phi_{\mathbf{v}}^t(\eta_t) - \Phi_{\mathbf{v}}^t(\eta_{t+1}) \right) .$$

This implies

$$-\frac{\log |\mathscr{P}_{\mathbf{v}}|}{\eta_T} - \sum_{t=1}^{T} \omega_{\mathbf{v}}^t(\widetilde{\pi}(\mathbf{v})) + \sum_{t=1}^{T-1} \left(\Phi_{\mathbf{v}}^t(\eta_t) - \Phi_{\mathbf{v}}^t(\eta_{t+1}) \right)$$

$$\leq - \sum_{t=1}^{T} \sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}] \omega_{\mathbf{v}}^t(\mathfrak{p}) + \sum_{t=1}^{T} \sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}] \eta_t \omega_{\mathbf{v}}^t(\mathfrak{p})^2 . \tag{24}$$

Claim 3. For any $\mathbf{v} \in \mathcal{V}$ and $t \in [T]$, $\Phi_{\mathbf{v}}^t(\eta)$ is non-decreasing for $\eta > 0$.

Proof. Recall that $\Phi^t_{\mathbf{v}}(\eta) = \frac{1}{\eta}\log\left(\frac{1}{|\mathscr{P}_{\mathbf{v}}|}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\exp\left(-\eta\sum_{s=1}^t \omega^s_{\mathbf{v}}(\mathfrak{p})\right)\right)$. Define

$$g(\eta) = \log\left(\frac{1}{|\mathscr{P}_{\mathbf{v}}|}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\exp\left(-\eta\sum_{s=1}^t \omega^s_{\mathbf{v}}(\mathfrak{p})\right)\right),$$

such that $\Phi^t_{\mathbf{v}}(\eta) = g(\eta)/\eta$. Note that $g(\eta)$ is convex because it is the composition of a convex function (LogSumExp) with an affine map of η,$-\eta\sum_{s=1}^t \omega^s_{\mathbf{v}}(\mathfrak{p})$, plus a constant, $-\log|\mathscr{P}_{\mathbf{v}}|$. As $\Phi^t_{\mathbf{v}}(\eta)$ is differentiable, we have

$$[\Phi^t_{\mathbf{v}}(\eta)]' = \frac{\eta g'(\eta) - g(\eta)}{\eta^2}.$$

Define $h(\eta) = \eta g'(\eta) - g(\eta)$ which implies $h'(\eta) = \eta g''(\eta)$. As $g(\eta)$ is convex and twice differentiable, $g''(\eta) \geq 0$ which implies for $\eta > 0$, $h'(\eta) \geq 0$, i.e., $h(\eta)$ is a non-decreasing function. Furthermore, $h(0) = -g(0) = 0$. Hence, $h(\eta) \geq 0$, for all $\eta > 0$ which implies $[\Phi^t_{\mathbf{v}}(\eta)]' = h(\eta)/\eta \geq 0, \forall \eta > 0$. Thus, $\Phi^t_{\mathbf{v}}(\eta)$ is non-decreasing in $\eta > 0$. □

Since $\eta_t \geq \eta_{t+1}$, by Claim 3, we have $\sum_{t=1}^{T-1}(\Phi^t_{\mathbf{v}}(\eta_t) - \Phi^t_{\mathbf{v}}(\eta_{t+1})) \geq 0$. Thus, after rearranging and using the fact that $|\mathscr{P}_{\mathbf{v}}| \lesssim \left(\frac{1}{\epsilon}\right)^M$, Eq. (24) becomes

$$\sum_{t=1}^T\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\mathbb{P}[\mathfrak{p}^t = \mathfrak{p}|\mathbf{v}^t = \mathbf{v}]\omega^t_{\mathbf{v}}(\mathfrak{p}) - \sum_{t=1}^T \omega^t_{\mathbf{v}}(\widetilde{\pi}(\mathbf{v})) \lesssim \frac{M\log 1/\epsilon}{\eta_T} + \sum_{t=1}^T\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\mathbb{P}[\mathfrak{p}^t = \mathfrak{p}|\mathbf{v}^t = \mathbf{v}]\eta_t\omega^t_{\mathbf{v}}(\mathfrak{p})^2\,. \tag{25}$$

Recall that for edge $e = (\ell-1, b_{\ell-1}, s_{\ell-1}) \to (\ell, b_\ell, s_\ell)$ in layer $\ell \in [M]$, the edge weight is:

$$\omega^t_{\mathbf{v}}(e) = \frac{1 - (v_\ell - \alpha b_\ell)\cdot\mathbb{I}\left[b_\ell \geq \boldsymbol{\beta}^{-(\ell)}_{-,t}\right]}{1+\alpha}\,. \qquad \text{((10) restated)}$$

$$\implies \omega^t_{\mathbf{v}}(\mathfrak{p}) = \sum_{e\in\mathfrak{p}}\omega^t_{\mathbf{v}}(e) \overset{(3)}{=} \frac{M - \mathsf{U}_{\mathbf{v}}(\mathbf{b};\boldsymbol{\beta}^t_-)}{1+\alpha}, \tag{26}$$

where $\mathbf{b} \in \mathcal{B}_{\mathbf{v}}$ is bidding strategy corresponding to the path $\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}$. Using the facts that $\omega^t_{\mathbf{v}}(\mathfrak{p}) \leq M$ and existence of a bijective mapping between policies in $\widetilde{\Pi}$ and Π, Eq. (25) becomes

$$\begin{aligned}
&\sum_{t=1}^T \mathsf{U}_{\mathbf{v}}(\pi(\mathbf{v});\boldsymbol{\beta}^t_-) - \sum_{t=1}^T\sum_{\mathbf{b}\in\mathcal{B}_{\mathbf{v}}}\mathbb{P}[\mathbf{b}^t = \mathbf{b}|\mathbf{v}^t = \mathbf{v}]\mathsf{U}_{\mathbf{v}}(\mathbf{b};\boldsymbol{\beta}^t_-)\\
&\lesssim (1+\alpha)\left[\frac{M\log 1/\epsilon}{\eta_T} + M^2\sum_{t=1}^T\eta_t\sum_{\mathbf{b}\in\mathcal{B}_{\mathbf{v}}}\mathbb{P}[\mathbf{b}^t = \mathbf{b}|\mathbf{v}^t = \mathbf{v}]\right]\\
&= (1+\alpha)\left[\frac{M\log 1/\epsilon}{\eta_T} + M^2\sum_{t=1}^T\eta_t\right]
\end{aligned}$$

Taking expectations with respect to the randomness of the valuation vectors, and using the facts that $\alpha \leq 1$, and that the choice of π was arbitrary, we obtain

$$\begin{aligned}
\mathsf{Reg}_{nb}(T) &= \mathsf{OPT}_{nb} - \sum_{t=1}^T\mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)]\\
&\overset{(\text{OPT-NB})}{=} \max_{\pi\in\Pi}\sum_{t=1}^T\mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t);\boldsymbol{\beta}^t_-)] - \sum_{t=1}^T\mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)] \lesssim \frac{M\log 1/\epsilon}{\eta_T} + M^2\sum_{t=1}^T\eta_t,
\end{aligned}$$

where $\boldsymbol{\beta}^t = (\mathbf{b};\boldsymbol{\beta}^t_-), \forall t \geq 1$. Setting $\eta_t = \sqrt{\frac{\log 1/\epsilon}{Mt}}$, and using the fact that $\sum_{t=1}^T\frac{1}{\sqrt{t}} \leq 2\sqrt{T}$, we get

$$\mathsf{Reg}_{nb}(T) \lesssim M^{3/2}\sqrt{T\log 1/\epsilon}\,.$$

Space and Time Complexity. The main bottleneck of implementing Algorithm 1 in the full information setting is updating the edge probabilities, $\phi^t_{\mathbf{v}}(\cdot)$ for all the contexts. For a fixed $\mathbf{v} \in \mathcal{V}$, updating $\phi^t_{\mathbf{v}}(\cdot)$ requires a forward pass and a backward pass over the edges $\mathsf{E}_{\mathbf{v}}$. Thus, Algorithm 1 requires $O(|\mathcal{V}| \cdot \max_{\mathbf{v}\in\mathcal{V}} |\mathsf{E}_{\mathbf{v}}|) = O(|\mathcal{V}|M^2/\epsilon^3)$ space and time per round.

C.4 Proof of Theorem 3.3

Recall that for any $\mathbf{v} \in \mathcal{V}$, $e \in \mathsf{E}_{\mathbf{v}}$ and $t \in [T]$,

$$\widehat{\omega}^t_{\mathbf{v}}(e) = \frac{\omega^t_{\mathbf{v}}(e)}{q^t(e)} \cdot \mathbb{I}\big[e \in \mathfrak{p}^t\big], \tag{(11) restated}$$

where $q^t(e) = \sum_{\mathbf{v}\in\mathcal{V}} \mathbb{P}[\mathbf{v}^t = \mathbf{v}] \sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}: e\in\mathfrak{p}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p}|\mathbf{v}^t = \mathbf{v}]$. Then,

Lemma C.2. *For any $\mathbf{v} \in \mathcal{V}$, $e \in \mathsf{E}_{\mathbf{v}}$ and $t \in [T]$,*

1. *$\widehat{\omega}^t_{\mathbf{v}}(e)$ is well defined,*
2. *$\widehat{\omega}^t_{\mathbf{v}}(e) \in [0, \infty)$,*
3. *$\mathbb{E}[\widehat{\omega}^t_{\mathbf{v}}(e)] = \omega^t_{\mathbf{v}}(e)$, and*
4. *$\mathbb{E}[\widehat{\omega}^t_{\mathbf{v}}(e)^2] \le \frac{1}{q^t(e)}$.*

Here, the expectation is taken with respect to the randomness in contexts as well as the learning algorithm conditioned on the history up to round t.

Proof. Fix any $\mathbf{v} \in \mathcal{V}$, $e \in \mathsf{E}_{\mathbf{v}}$ and $t \in [T]$.

1. Without loss of generality, assume $\mathbb{P}[\mathbf{v}^t = \mathbf{v}] > 0$.[8] This implies that $q^t(e) > 0$, since every $e \in \mathsf{E}_{\mathbf{v}}$ lies on some path $\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}$ and every $\mathfrak{p}$ satisfies $\mathbb{P}[\mathfrak{p}^1 = \mathfrak{p}|\mathbf{v}^1 = \mathbf{v}] > 0$. The edge probabilities updates in Algorithm 1 ensure that $\mathbb{P}[\mathfrak{p}^t = \mathfrak{p}|\mathbf{v}^t = \mathbf{v}] > 0$ for all subsequent rounds t. Hence, $\widehat{\omega}^t_{\mathbf{v}}(e)$ is well defined.
2. The result follows directly from the definition in Eq. (11).
3. Since
$$\mathbb{E}\left[\mathbb{I}\big[e \in \mathfrak{p}^t\big]\right] = \sum_{\mathbf{v}\in\mathcal{V}} \mathbb{P}[\mathbf{v}^t = \mathbf{v}] \sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}: e\in\mathfrak{p}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p}|\mathbf{v}^t = \mathbf{v}] = q^t(e),$$
we get that $\mathbb{E}[\widehat{\omega}^t_{\mathbf{v}}(e)] = \omega^t_{\mathbf{v}}(e)$.
4. By definition,
$$\mathbb{E}[\widehat{\omega}^t_{\mathbf{v}}(e)^2] \overset{(11)}{=} \frac{\omega^t_{\mathbf{v}}(e)^2}{q^t(e)^2} \cdot \mathbb{E}\left[\mathbb{I}\big[e \in \mathfrak{p}^t\big]\right] \le \frac{1}{q^t(e)}, \tag{27}$$
where the first inequality follows as $0 \le \omega^t_{\mathbf{v}}(e) \le 1$ (see Eq. (10)).

□

Following the proof of Theorem 3.2 up to Eq. (25),[9] for any stationary policy $\widetilde{\pi} \in \widetilde{\Pi}$, we have

$$\sum_{t=1}^{T} \sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p}|\mathbf{v}^t = \mathbf{v}]\widehat{\omega}^t_{\mathbf{v}}(\mathfrak{p}) - \sum_{t=1}^{T} \widehat{\omega}^t_{\mathbf{v}}(\widetilde{\pi}(\mathbf{v}))$$

[8] If there exists context $\mathbf{v} \in \mathcal{V}$ such that $\mathbb{P}[\mathbf{v}^t = \mathbf{v}] = 0$, consider $\mathcal{V}' = \mathcal{V} \setminus \{\mathbf{v}\}$.

[9] The analysis of Theorem 3.2 up to Eq. (25) only requires $\widehat{\omega}^t_{\mathbf{v}}(\cdot) \ge 0$ which is ensured by Lemma C.2 (2).

$$\lesssim \frac{M\log 1/\epsilon}{\eta_T} + \sum_{t=1}^{T}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}} \mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\eta_t\widehat{\omega}^t_{\mathbf{v}}(\mathfrak{p})^2$$

$$\leq \frac{M\log 1/\epsilon}{\eta_T} + M\sum_{t=1}^{T}\eta_t \sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}} \mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\sum_{e\in\mathfrak{p}}\widehat{\omega}^t_{\mathbf{v}}(e)^2, \tag{28}$$

where the last line follows due to Cauchy-Schwarz inequality.

Taking expectations with respect to the randomness of contexts and Algorithm 1 conditioned on the history and using Eq. (27), we have

$$\sum_{t=1}^{T}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}} \mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\omega^t_{\mathbf{v}}(\mathfrak{p}) - \sum_{t=1}^{T}\omega^t_{\mathbf{v}}(\widetilde{\pi}(\mathbf{v})) \lesssim \frac{M\log 1/\epsilon}{\eta_T} + M\sum_{t=1}^{T}\eta_t\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}} \mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\sum_{e\in\mathfrak{p}}\frac{1}{q^t(e)}\,.$$

Relating the path weights in the DAG to the utility using Eq. (26) and leveraging the bijective mapping between policies in $\widetilde{\Pi}$ and Π, the left hand side becomes

$$\sum_{t=1}^{T}\mathsf{U}_{\mathbf{v}}(\pi(\mathbf{v});\boldsymbol{\beta}^t_-) - \sum_{t=1}^{T}\sum_{\mathbf{b}\in\mathcal{B}_{\mathbf{v}}}\mathbb{P}[\mathbf{b}^t=\mathbf{b}|\mathbf{v}^t=\mathbf{v}]\mathsf{U}_{\mathbf{v}}(\mathbf{b};\boldsymbol{\beta}^t_-)\,.$$

Taking expectations with respect to the contexts and the history, we get

$$\begin{aligned}\mathsf{Reg}_{nb}(T) &= \sum_{t=1}^{T}\mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t);\boldsymbol{\beta}^t_-)] - \sum_{t=1}^{T}\mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)]\\ &\lesssim \frac{M\log 1/\epsilon}{\eta_T} + M\mathbb{E}\left[\sum_{t=1}^{T}\eta_t\sum_{\mathbf{v}\in\mathcal{V}}\mathbb{P}[\mathbf{v}^t=\mathbf{v}]\sum_{e\in\mathsf{E}_{\mathbf{v}}}\frac{1}{q^t(e)}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}:e\in\mathfrak{p}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\right]\end{aligned}$$

Define $\mathsf{E}:=\bigcup_{\mathbf{v}\in\mathcal{V}}\mathsf{E}_{\mathbf{v}}$. This implies

$$\begin{aligned}\mathsf{Reg}_{nb}(T) &\lesssim \frac{M\log 1/\epsilon}{\eta_T} + M\mathbb{E}\left[\sum_{t=1}^{T}\eta_t\sum_{e\in\mathsf{E}}\frac{1}{q^t(e)}\sum_{\mathbf{v}\in\mathcal{V}}\mathbb{P}[\mathbf{v}^t=\mathbf{v}]\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}:e\in\mathfrak{p}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\right]\\ &\overset{(12)}{=} \frac{M\log 1/\epsilon}{\eta_T} + M\mathbb{E}\left[\sum_{t=1}^{T}\eta_t\sum_{e\in\mathsf{E}}\frac{1}{q^t(e)}\cdot q^t(e)\right]\\ &= \frac{M\log 1/\epsilon}{\eta_T} + M|\mathsf{E}|\sum_{t=1}^{T}\eta_t\,.\end{aligned}$$

To bound $|\mathsf{E}|$, recall the super DAG $\mathcal{G}(\overline{\mathsf{N}},\overline{\mathsf{E}})$ constructed in Section 5 and that $\mathsf{E}_{\mathbf{v}}\subseteq\overline{\mathsf{E}}$ for all $\mathbf{v}\in\mathcal{V}$. Hence, $\mathsf{E}=\bigcup_{\mathbf{v}\in\mathcal{V}}\mathsf{E}_{\mathbf{v}}\subseteq\overline{\mathsf{E}}$. Moreover, $|\overline{\mathsf{E}}|=O\left(\frac{M^2}{\epsilon^3}\right)$. Since, $|\mathsf{E}|\leq|\overline{\mathsf{E}}|$, we have

$$\mathsf{Reg}_{nb}(T)\lesssim\frac{M\log 1/\epsilon}{\eta_T}+\frac{M^3}{\epsilon^3}\sum_{t=1}^{T}\eta_t\,.$$

Setting $\eta_t=\frac{1}{M}\sqrt{\frac{\epsilon^3\log 1/\epsilon}{t}}$, and using the fact that $\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\leq 2\sqrt{T}-1$, we get

$$\mathsf{Reg}_{nb}(T)\lesssim\frac{M^2}{\epsilon^{3/2}}\sqrt{T\log 1/\epsilon}\,.$$

Space and Time Complexity. In this setting, the main bottleneck of implementing Algorithm 1 is computing $q^t(e)$ (equivalently $\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}:e\in\mathfrak{p}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]$ for each $\mathbf{v}\in\mathcal{V}$) per Eq. (12). For a fixed $\mathbf{v}\in\mathcal{V}$ and $e\in\mathsf{E}_{\mathbf{v}}$,

$$\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}:e\in\mathfrak{p}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\overset{(20)}{=}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}:e\in\mathfrak{p}}\prod_{e'\in\mathfrak{p}}\phi^t_{\mathbf{v}}(e') \tag{29}$$

Leveraging Golrezaei and Sahoo (2025, Theorem 4.4), for any $\mathbf{v}\in\mathcal{V}$, Eq. (29) can be computed in $O(|\mathsf{E}_{\mathbf{v}}|)$ time and space per round. Hence, Algorithm 1 runs in $O(|\mathcal{V}|\cdot\max_{\mathbf{v}\in\mathcal{V}}|\mathsf{E}_{\mathbf{v}}|)=O\Big(\frac{|\mathcal{V}|M^2}{\epsilon^3}\Big)$ time and space per round.

C.5 Proof of Theorem 3.4

We first present the learning algorithm in the bandit setting with unknown context distribution.

Algorithm 3 No Budget Constraint, Bandit Setting, Unknown Distribution

Require: Set of valuation vectors $\mathcal{V}$, learning rates $\eta_t>0$, $\delta\in(0,1]$. Define $\eta_0=1,\phi^0_{\mathbf{v}}(e)=1$ and $\omega^0_{\mathbf{v}}(e)=0,\forall e\in\mathsf{E}_{\mathbf{v}},\forall\mathbf{v}\in\mathcal{V}$.

1: **for** $t=1,2,\dots,T$ **do**
2: Observe an i.i.d. valuation vector sample $\mathbf{v}^t\sim\mathcal{D}$.
3: Construct $\mathcal{G}^t(\mathsf{N}_{\mathbf{v}},\mathsf{E}_{\mathbf{v}},\omega_{\mathbf{v}})$ without weights $\forall\mathbf{v}\in\mathcal{V}$.
4: Sample $Z_t\sim\text{Unif}[0,1]$.
5: **if** $Z_t\le\delta$ **then**
6: Obtain path $\mathfrak{p}^t$ using Algorithm 4.
7: **else**
8: **for** $\mathbf{v}\in\mathcal{V}$ **do**
9: Obtain edge probabilities $\phi^t_{\mathbf{v}}(\cdot),\forall\mathbf{v}\in\mathcal{V}$ using Eq. (8) and Eq. (9), with $\omega^t_{\mathbf{v}}(\cdot)$ replaced by $-\widehat{\omega}^t_{\mathbf{v}}(\cdot)$, where $\widehat{\omega}^t_{\mathbf{v}}(\cdot)$ is defined in Eq. (14).
10: Define initial node $u=s$ and path $\mathfrak{p}^t=s$.
11: **while** $u\neq d$ **do**
12: Sample v with probability $\phi^t_{\mathbf{v}^t}(u\to v)$.
13: Append v to the path $\mathfrak{p}^t$; set $u\leftarrow v$.
14: Map $\mathfrak{p}^t=s\to(1,b_1,s)\to\cdots\to(M,b_M,s_M)\to d$, and submit $\mathbf{b}^t=[b_1,\dots,b_M]$.
15: Set edge weights per Eq. (14) for all $\mathbf{v}\in\mathcal{V}$.

Recall that for any $\mathbf{v}\in\mathcal{V}$, $e\in\mathsf{E}_{\mathbf{v}}$ and $t\in[T]$, if $\mathbf{v}^t=\mathbf{v}'$, the estimator is given as

$$\widehat{\omega}^t_{\mathbf{v}}(e)=\frac{\omega^t_{\mathbf{v}}(e)}{p^t_{\mathbf{v}'}(e)}\cdot\mathbb{I}\big[e\in\mathfrak{p}^t\big], \tag{(14) restated}$$

where $\omega^t_{\mathbf{v}}(e)$ is per Eq. (13), and $p^t_{\mathbf{v}'}(e)=\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}'}:e\in\mathfrak{p}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}']$.

Lemma C.3. *For any $\mathbf{v}\in\mathcal{V}$, $e\in\mathsf{E}_{\mathbf{v}}$ and $t\in[T]$,*

1. *$\widehat{\omega}^t_{\mathbf{v}}(e)$ is well defined,*
2. *$\widehat{\omega}^t_{\mathbf{v}}(e)\in[0,\frac{|\overline{\mathsf{E}}|}{\delta}]$,*
3. *$\mathbb{E}[\widehat{\omega}^t_{\mathbf{v}}(e)]=\omega^t_{\mathbf{v}}(e)$, and*
4. *$\mathbb{E}[\widehat{\omega}^t_{\mathbf{v}}(e)^2]\le\frac{|\overline{\mathsf{E}}|}{\delta}$.*

where $\overline{\mathsf{E}}$ is the set of edges in the super DAG $\mathcal{G}(\overline{\mathsf{N}},\overline{\mathsf{E}})$ constructed in Section 3. Here, the expectation is taken with respect to the randomness in contexts as well as the learning algorithm conditioned on the history up to round t.

Proof. Following the proof of Theorem 3.2, it is easy to verify that Algorithm 3, Line 7 to Algorithm 3, Line 13 implements the Decreasing Hedge algorithm. Concretely, the probability of selecting path $\mathfrak{p}$, given the context $\mathbf{v}^t=\mathbf{v}'$ is

$$\widetilde{\mathbb{P}}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}']=\frac{\exp(\eta_t\sum_{s=1}^{t-1}\widehat{\omega}^s_{\mathbf{v}'}(\mathfrak{p}))}{\sum_{\mathfrak{p}'\in\mathscr{P}_{\mathbf{v}'}}\exp(\eta_t\sum_{s=1}^{t-1}\widehat{\omega}^s_{\mathbf{v}'}(\mathfrak{p}'))}. \tag{30}$$

Algorithm 4 Sample a Random s-d Path

Require: Context-dependent DAG $\mathcal{G}(\mathsf{N}_{\mathbf{v}'}, \mathsf{E}_{\mathbf{v}'}, \omega_{\mathbf{v}'})$ with source s and sink d
Compute a topological order of $\mathcal{G}(\mathsf{N}_{\mathbf{v}'}, \mathsf{E}_{\mathbf{v}'}, \omega_{\mathbf{v}'})$
Compute canonical pointers $\texttt{parS}[\cdot]$ (one $s \to u$ path for each u)
Compute canonical pointers $\texttt{parD}[\cdot]$ (one $u \to d$ path for each u)
$e^\star \leftarrow \perp$; $\quad k \leftarrow 0$
for all $e = (u, v) \in \mathsf{E}_{\mathbf{v}'}$ in a fixed order **do** ▷ *reservoir sampling*
  $k \leftarrow k + 1$
  With prob. $1/k$, set $e^\star \leftarrow (u, v)$
return $\textsc{Path}(s \to u; \texttt{parS}) \parallel e^\star \parallel \textsc{Path}(v \to d; \texttt{parD})$

Hence, from the description of Algorithm 3, we get

$$\mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}'] := (1-\delta) \cdot \widetilde{\mathbb{P}}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}'] + \delta \cdot \frac{\textsc{Count}[\mathfrak{p} \in \mathcal{C}_{\mathbf{v}'}]}{|\mathcal{C}_{\mathbf{v}'}|}, \tag{31}$$

where $\textsc{Count}[\mathfrak{p} \in \mathcal{C}_{\mathbf{v}'}]$ is the number of times $\mathfrak{p}$ appears in the EPC $\mathcal{C}_{\mathbf{v}'}$

1. If $\mathbf{v}^t = \mathbf{v}'$, then

$$p^t_{\mathbf{v}'}(e) = \sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}'} : e \in \mathfrak{p}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}'] \geq \delta \sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}'} : e \in \mathfrak{p}} \frac{\textsc{Count}[\mathfrak{p} \in \mathcal{C}_{\mathbf{v}'}]}{|\mathcal{C}_{\mathbf{v}'}|} \geq \frac{\delta}{|\mathcal{C}_{\mathbf{v}'}|} = \frac{\delta}{|\mathsf{E}_{\mathbf{v}'}|}, \tag{32}$$

   where the second inequality follows as every edge e is contained in at least one path $\mathfrak{p} \in \mathcal{C}_{\mathbf{v}'}$ by definition of EPC and the third equality holds as $|\mathcal{C}_{\mathbf{v}'}| = |\mathsf{E}_{\mathbf{v}'}|$. Since $p^t_{\mathbf{v}'}(e) > 0$, the estimator in Eq. (14) is well-defined.

2. Since, $\omega^t_{\mathbf{v}}(\cdot) \in [0, 1]$ per Eq. (13), by the definition of the estimator in Eq. (14), it follows that $\widehat{\omega}^t(e) \geq 0$. By using the result from previous part and the fact that $|\mathsf{E}_{\mathbf{v}'}| \leq |\overline{\mathsf{E}}|$, we have

$$\frac{\omega^t_{\mathbf{v}}(e)}{p^t_{\mathbf{v}'}(e)} \cdot \mathbb{I}\big[e \in \mathfrak{p}^t\big] \leq \frac{1}{p^t_{\mathbf{v}'}(e)} \leq \frac{|\mathsf{E}_{\mathbf{v}'}|}{\delta} \leq \frac{|\overline{\mathsf{E}}|}{\delta}\,.$$

3. Taking expectations with respect to all randomness in contexts and the learning algorithm conditioned on the history up to round t,

$$\mathbb{E}[\widehat{\omega}^t_{\mathbf{v}}(e)] = \omega^t_{\mathbf{v}}(e) \sum_{\mathbf{v}' \in \mathcal{V}} \mathbb{P}[\mathbf{v}^t = \mathbf{v}'] \cdot \frac{p^t_{\mathbf{v}'}(e)}{p^t_{\mathbf{v}'}(e)} = \omega^t_{\mathbf{v}}(e)\,.$$

4. By definition of the estimator in Eq. (14),

$$\begin{aligned}\mathbb{E}[\widehat{\omega}^t_{\mathbf{v}}(e)^2] &= \omega^t_{\mathbf{v}}(e)^2 \sum_{\mathbf{v}' \in \mathcal{V}} \mathbb{P}[\mathbf{v}^t = \mathbf{v}'] \cdot \frac{p^t_{\mathbf{v}'}(e)}{p^t_{\mathbf{v}'}(e)^2} \leq \sum_{\mathbf{v}' \in \mathcal{V}} \mathbb{P}[\mathbf{v}^t = \mathbf{v}'] \cdot \frac{1}{p^t_{\mathbf{v}'}(e)} \\ &\overset{(32)}{\leq} \sum_{\mathbf{v}' \in \mathcal{V}} \mathbb{P}[\mathbf{v}^t = \mathbf{v}'] \cdot \frac{|\mathsf{E}_{\mathbf{v}'}|}{\delta} \leq \max_{\mathbf{v}' \in \mathcal{V}} \frac{|\mathsf{E}_{\mathbf{v}'}|}{\delta} \leq \frac{|\mathsf{E}|}{\delta}\end{aligned}$$

   where the first inequality holds true as $\omega^t_{\mathbf{v}}(\cdot) \in [0, 1]$.

□

Regret Analysis. For any $\mathbf{v} \in \mathcal{V}$ and $\eta > 0$, $\delta \in (0, 1]$ and $t \in [T]$, let

$$\Phi^t_{\mathbf{v}}(\eta) := \frac{1}{\eta} \log \left(\frac{1}{|\mathscr{P}_{\mathbf{v}}|} \sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}}} \exp\left(\eta \sum_{s=1}^{t} \widehat{\omega}^s_{\mathbf{v}}(\mathfrak{p}) \right) \right)\,.$$

So, for some $\eta_t \le \frac{\delta}{M|\mathsf{E}|}$ and $\delta \in (0,1]$,

$$\begin{aligned}
&\Phi^t_{\mathbf{v}}(\eta_t) - \Phi^{t-1}_{\mathbf{v}}(\eta_t) \\
&= \frac{1}{\eta_t}\log\left(\frac{\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\exp(\eta_t\sum_{s=1}^{t-1}\widehat{\omega}^s_{\mathbf{v}}(\mathfrak{p}))\cdot\exp(\eta_t\widehat{\omega}^t_{\mathbf{v}}(\mathfrak{p}))}{\sum_{\mathfrak{p}'\in\mathscr{P}_{\mathbf{v}}}\exp(\eta_t\sum_{s=1}^{t-1}\widehat{\omega}^s_{\mathbf{v}}(\mathfrak{p}'))}\right) \\
&\overset{(30)}{=} \frac{1}{\eta_t}\log\left(\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\widetilde{\mathbb{P}}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\cdot\exp(\eta_t\widehat{\omega}^t_{\mathbf{v}}(\mathfrak{p}))\right) \\
&\overset{(i)}{\le} \frac{1}{\eta_t}\log\left(\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\widetilde{\mathbb{P}}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\cdot(1+\eta_t\widehat{\omega}^t_{\mathbf{v}}(\mathfrak{p})+\eta_t^2\widehat{\omega}^t_{\mathbf{v}}(\mathfrak{p})^2)\right) \\
&= \frac{1}{\eta_t}\log\left(1+\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\widetilde{\mathbb{P}}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\cdot(\eta_t\widehat{\omega}^t_{\mathbf{v}}(\mathfrak{p})+\eta_t^2\widehat{\omega}^t_{\mathbf{v}}(\mathfrak{p})^2)\right) \\
&\overset{(ii)}{\le} \frac{1}{\eta_t}\log\left(1+\frac{\eta_t}{1-\delta}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\widehat{\omega}^t_{\mathbf{v}}(\mathfrak{p})+\frac{\eta_t^2}{1-\delta}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\widehat{\omega}^t_{\mathbf{v}}(\mathfrak{p})^2\right) \\
&\overset{(iii)}{\le} \frac{1}{1-\delta}\left(\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\widehat{\omega}^t_{\mathbf{v}}(\mathfrak{p})+\eta_t\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\widehat{\omega}^t_{\mathbf{v}}(\mathfrak{p})^2\right).
\end{aligned}$$

Here, (i) uses $e^x \le 1+x+x^2$ for $x\le 1$ because for $\eta_t \le \frac{\delta}{M|\mathsf{E}|}$, any $\mathbf{v}\in\mathcal{V}$ and $\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}$,

$$\eta_t\widehat{\omega}^t_{\mathbf{v}}(\mathfrak{p}) = \eta_t\sum_{e\in\mathfrak{p}}\widehat{\omega}^t_{\mathbf{v}}(e) \overset{(a)}{\le} \eta_t\sum_{e\in\mathfrak{p}}\frac{|\mathsf{E}|}{\delta} \le \frac{\eta_t M|\mathsf{E}|}{\delta} \le 1\,.$$

where the first inequality follows from Lemma C.3 (2). The inequality (ii) holds

$$\widetilde{\mathbb{P}}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}] \overset{(31)}{=} \frac{\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}] - \frac{\delta}{|\mathcal{C}_{\mathbf{v}}|}\cdot\textsc{Count}[\mathfrak{p}\in\mathcal{C}_{\mathbf{v}}]}{1-\delta} \le \frac{\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]}{1-\delta}$$

and (iii) is true as $\log(1+x)\le x, \forall x>-1$. Summing over $t=1,\dots,T$, the left hand side becomes

$$\sum_{t=1}^{T}\left(\Phi^t_{\mathbf{v}}(\eta_t)-\Phi^{t-1}_{\mathbf{v}}(\eta_t)\right) = \Phi^T_{\mathbf{v}}(\eta_T) - \Phi^0_{\mathbf{v}}(\eta_1) + \sum_{t=1}^{T-1}\left(\Phi^t_{\mathbf{v}}(\eta_t)-\Phi^t_{\mathbf{v}}(\eta_{t+1})\right)$$

By definition, $\Phi^0_{\mathbf{v}}(\eta_1)=0$. Recall the definition of the policy class $\widetilde{\Pi}$ from Eq. (23). For any $\widetilde{\pi}\in\widetilde{\Pi}$,

$$\begin{aligned}
\Phi^T_{\mathbf{v}}(\eta_T) &= \frac{1}{\eta_T}\log\left(\frac{1}{|\mathscr{P}_{\mathbf{v}}|}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\exp\left(\eta_T\sum_{s=1}^{T}\widehat{\omega}^s_{\mathbf{v}}(\mathfrak{p})\right)\right) \\
&\ge \frac{1}{\eta_T}\log\left(\frac{1}{|\mathscr{P}_{\mathbf{v}}|}\cdot\exp\left(\eta_T\sum_{s=1}^{T}\widehat{\omega}^s_{\mathbf{v}}(\widetilde{\pi}(\mathbf{v}))\right)\right) \\
&= -\frac{\log|\mathscr{P}_{\mathbf{v}}|}{\eta_T} + \sum_{t=1}^{T}\widehat{\omega}^t_{\mathbf{v}}(\widetilde{\pi}(\mathbf{v}))\,. \qquad (33)
\end{aligned}$$

By Claim 3, $\Phi^t_{\mathbf{v}}(\cdot)$ is an non-decreasing function which implies that $\Phi^t_{\mathbf{v}}(\eta_t)\ge\Phi^t_{\mathbf{v}}(\eta_{t+1})$ for $\eta_t\ge\eta_{t+1}$.[10] Hence,

$$\sum_{t=1}^{T}\left(\Phi^t_{\mathbf{v}}(\eta_t)-\Phi^{t-1}_{\mathbf{v}}(\eta_t)\right) \overset{(33)}{\ge} -\frac{\log|\mathscr{P}_{\mathbf{v}}|}{\eta_T} + \sum_{t=1}^{T}\widehat{\omega}^t_{\mathbf{v}}(\widetilde{\pi}(\mathbf{v}))\,.$$

[10] Observe that the proof of Claim 3 does not rely on the sign of the coefficient of η in the exponential.

Thus, we have

$$
\begin{aligned}
&-\frac{\log|\mathscr{P}_{\mathbf{v}}|}{\eta_T}+\sum_{t=1}^{T}\widehat{\omega}_{\mathbf{v}}^{t}(\widetilde{\pi}(\mathbf{v}))\\
&\leq\frac{1}{1-\delta}\left(\sum_{t=1}^{T}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\widehat{\omega}_{\mathbf{v}}^{t}(\mathfrak{p})+\sum_{t=1}^{T}\eta_t\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\widehat{\omega}_{\mathbf{v}}^{t}(\mathfrak{p})^2\right)\\
&\leq\frac{1}{1-\delta}\left(\sum_{t=1}^{T}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\widehat{\omega}_{\mathbf{v}}^{t}(\mathfrak{p})+M\sum_{t=1}^{T}\eta_t\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\sum_{e\in\mathfrak{p}}\widehat{\omega}_{\mathbf{v}}^{t}(e)^2\right)
\end{aligned}
$$

where, the last inequality holds due to Cauchy-Schwarz inequality. Taking expectations over the randomness of the contexts and of Algorithm 3, conditioned on the history, and using Lemma C.3 (4), we have

$$
\begin{aligned}
&-\frac{\log|\mathscr{P}_{\mathbf{v}}|}{\eta_T}+\sum_{t=1}^{T}\omega_{\mathbf{v}}^{t}(\widetilde{\pi}(\mathbf{v}))\\
&\leq\frac{1}{1-\delta}\left(\sum_{t=1}^{T}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\omega_{\mathbf{v}}^{t}(\mathfrak{p})+M\sum_{t=1}^{T}\eta_t\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\sum_{e\in\mathfrak{p}}\frac{|\overline{\mathsf{E}}|}{\delta}\right)\\
&\leq\frac{1}{1-\delta}\left(\sum_{t=1}^{T}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\omega_{\mathbf{v}}^{t}(\mathfrak{p})+\frac{M^2|\overline{\mathsf{E}}|}{\delta}\sum_{t=1}^{T}\eta_t\right),
\end{aligned}
$$

where $\overline{\mathsf{E}}$ is the set of edges in the super DAG. Multiplying both sides by $(1-\delta)$, rearranging the terms and using the fact that $\omega_{\mathbf{v}}^{t}(\widetilde{\pi}(\mathbf{v}))\leq M$, we get

$$
\sum_{t=1}^{T}\omega_{\mathbf{v}}^{t}(\widetilde{\pi}(\mathbf{v}))-\sum_{t=1}^{T}\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}}\mathbb{P}[\mathfrak{p}^t=\mathfrak{p}|\mathbf{v}^t=\mathbf{v}]\omega_{\mathbf{v}}^{t}(\mathfrak{p})\leq\frac{\log|\mathscr{P}_{\mathbf{v}}|}{\eta_T}+\frac{M^2|\overline{\mathsf{E}}|}{\delta}\sum_{t=1}^{T}\eta_t+MT\delta\,.
$$

Recall that in this setting, the weight of edge $e=(\ell-1,b_{\ell-1},s_{\ell-1})\to(\ell,b_\ell,s_\ell)$ in layer $\ell\in[M]$ is:

$$
\begin{aligned}
\omega_{\mathbf{v}}^{t}(e)&=\frac{\alpha+(v_\ell-\alpha b_\ell)\cdot\mathbb{I}\left[b_\ell\geq\boldsymbol{\beta}_{-,t}^{-(\ell)}\right]}{1+\alpha}\,. && \text{((13) restated)}\\
\implies\omega_{\mathbf{v}}^{t}(\mathfrak{p})&=\sum_{e\in\mathfrak{p}}\omega_{\mathbf{v}}^{t}(e)\overset{(3)}{=}\frac{M\alpha+\mathsf{U}_{\mathbf{v}}(\mathbf{b};\boldsymbol{\beta}_{-}^{t})}{1+\alpha},
\end{aligned}
$$

where $\mathbf{b}\in\mathcal{B}_{\mathbf{v}}$ is strategy corresponding to the path $\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}$. Due to the bijective mapping between policies in $\widetilde{\Pi}$ (defined in Eq. (23)) and Π, we have

$$
\begin{aligned}
\sum_{t=1}^{T}\mathsf{U}_{\mathbf{v}}(\pi(\mathbf{v});\boldsymbol{\beta}_{-}^{t})-\sum_{t=1}^{T}\sum_{\mathbf{b}\in\mathcal{B}_{\mathbf{v}}}\mathbb{P}[\mathbf{b}^t=\mathbf{b}|\mathbf{v}^t=\mathbf{v}]\mathsf{U}_{\mathbf{v}}(\mathbf{b};\boldsymbol{\beta}_{-}^{t})&\leq\frac{\log|\mathscr{P}_{\mathbf{v}}|}{\eta_T}+\frac{M^2|\overline{\mathsf{E}}|}{\delta}\sum_{t=1}^{T}\eta_t+MT\delta\\
&\lesssim\frac{M\log 1/\epsilon}{\eta_T}+\frac{M^2|\overline{\mathsf{E}}|}{\delta}\sum_{t=1}^{T}\eta_t+MT\delta,
\end{aligned}
$$

where the last line follows since $|\mathscr{P}_{\mathbf{v}}|\lesssim(\frac{1}{\epsilon})^M$. Taking expectations with respect to the contexts and the history, and maximizing over the policies in Π, we have

$$
\mathsf{Reg}_{nb}(T)=\max_{\pi\in\Pi}\sum_{t=1}^{T}\mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t);\boldsymbol{\beta}_{-}^{t})]-\sum_{t=1}^{T}\mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)]\lesssim\frac{M\log 1/\epsilon}{\eta_T}+\frac{M^2|\overline{\mathsf{E}}|}{\delta}\sum_{t=1}^{T}\eta_t+MT\delta\,.
$$

Suppose for all $t \geq 1$, $\eta_t = \eta = \frac{\delta^2}{M|\overline{\mathsf{E}}|}$. Then, for $T \geq T_0 =: M|\overline{\mathsf{E}}| \log 1/\epsilon$, set

$$\delta = \left(\frac{M|\overline{\mathsf{E}}| \log 1/\epsilon}{T}\right)^{1/3}.$$

Note that $\eta_t \leq \frac{\delta}{M|\overline{\mathsf{E}}|}$ since $\delta \leq 1$. Recall from Section 3 that $|\overline{\mathsf{E}}| = O(M^2/\epsilon^3)$. Hence, for $T \geq T_0$,

$$\mathsf{Reg}_{nb}(T) \lesssim \frac{M^2 T^{2/3} (\log 1/\epsilon)^{1/3}}{\epsilon}.$$

Since regret in any round is bounded by M, the regret in the first T_0 rounds is at most MT_0. Hence,

$$\mathsf{Reg}_{nb}(T) \lesssim \frac{M^2 T^{2/3} (\log 1/\epsilon)^{1/3}}{\epsilon} + \frac{M^4 \log 1/\epsilon}{\epsilon^3}.$$

Space and Time Complexity. In this setting, the bidder alternates between two modes depending on the mixing parameter δ. If the bidder samples a path from an EPC, the per-round time and space complexity is $O(M^2/\epsilon^3)$ (see Algorithm 4). If instead the bidder follows the exponential-weights updates, the complexity matches the bandit setting with known context distribution, namely $O(|\mathcal{V}|M^2/\epsilon^3)$. Therefore, the overall per-round time and space complexity is $O(|\mathcal{V}|M^2/\epsilon^3)$.

C.6 Proof of Theorem 3.5

We build upon and generalise the proof of regret lower bound for FPA in Han et al. (2024) intended for profit maximizers ($\alpha = 1$). We construct a stochastic adversary whose distribution makes it harder for the bidder to determine their optimal bid.

Concretely, let $\mathcal{V} = \{\mathbf{v}\}$, $K = M$ and $\mathbf{v} = [1, \ldots, 1] \in \mathbb{R}^M$. Assume ties are always resolved in favor of the bidder in consideration and the cost of capital $\alpha \in (\frac{1}{2}, 1]$. Define

$$\boldsymbol{\beta}_-^{\clubsuit} = [c, \ldots, c] \quad \text{and} \quad \boldsymbol{\beta}_-^{\diamondsuit} = [\gamma c, \ldots, \gamma c], \tag{34}$$

where

$$\gamma > \frac{\alpha}{2\alpha - 1}, \qquad c = \frac{1}{\alpha(2\gamma - 1)}.$$

It is easy to verify that $\gamma > 1$ and

$$\gamma c = \frac{\gamma}{\alpha(2\gamma - 1)} < 1,$$

since $f(x) = \frac{x}{2x-1}$ is decreasing over $\gamma \in (1, \infty)$.

Consider the following two scenarios:

Scenario 1. In this scenario, for every $t \in [T]$, the competing bids $\boldsymbol{\beta}_-^t$ are:

$$\boldsymbol{\beta}_-^t = \begin{cases} \boldsymbol{\beta}_-^{\clubsuit}, & \text{w.p. } \frac{1}{2} + \delta, \\ \boldsymbol{\beta}_-^{\diamondsuit}, & \text{w.p. } \frac{1}{2} - \delta, \end{cases}$$

Scenario 2. In this scenario, for every $t \in [T]$, the competing bids $\boldsymbol{\beta}_-^t$ are:

$$\boldsymbol{\beta}_-^t = \begin{cases} \boldsymbol{\beta}_-^{\clubsuit}, & \text{w.p. } \frac{1}{2} - \delta, \\ \boldsymbol{\beta}_-^{\diamondsuit}, & \text{w.p. } \frac{1}{2} + \delta, \end{cases}$$

for some $\delta \in (0, \frac{1}{4})$ to be determined shortly. Assume the randomness used in different rounds are independent. Then, for $\delta \in (0, \frac{1}{4})$,

$$\mathrm{KL}(P||Q) = T \cdot \mathrm{KL}(\textsc{Bern}(0.5 + \delta)||\textsc{Bern}(0.5 - \delta)) = 2T\delta \log\left(\frac{1 + 2\delta}{1 - 2\delta}\right) \leq \frac{8T\delta^2}{1 - 2\delta} \leq 16T\delta^2$$

where the first inequality follows from $\log(\frac{1+x}{1-x}) \leq \frac{2x}{1-x}$. By Tsybakov (2009, Lemma 2.6),

$$1 - \mathrm{TV}(P, Q) \geq \frac{1}{2}\exp\left(-\mathrm{KL}(P||Q)\right) \geq \frac{1}{2}\exp\left(-16T\delta^2\right).$$

Now, we inspect the separation between the two scenarios. Observe that

$$\max_{\mathbf{b}\in\mathcal{B}_\mathbf{v}} \mathbb{E}_P[\mathsf{U}_\mathbf{v}(\mathbf{b};\boldsymbol{\beta}_-^t)] \geq \mathbb{E}_P[\mathsf{U}_\mathbf{v}([c,\ldots,c];\boldsymbol{\beta}_-^t)] = \Big(\frac{1}{2}+\delta\Big)M(1-\alpha c),$$
$$\max_{\mathbf{b}\in\mathcal{B}_\mathbf{v}} \mathbb{E}_Q[\mathsf{U}_\mathbf{v}(\mathbf{b};\boldsymbol{\beta}_-^t)] \geq \mathbb{E}_Q[\mathsf{U}_\mathbf{v}([\gamma c,\ldots,\gamma c];\boldsymbol{\beta}_-^t)] = M(1-\gamma\alpha c).$$

Now, we aim to compute $\max_{\mathbf{b}\in\mathcal{B}_\mathbf{v}} \mathbb{E}_{(P+Q)/2}[\mathsf{U}_\mathbf{v}(\mathbf{b};\boldsymbol{\beta}_-^t)]$. Since $\mathbf{v} = [1,\ldots,1]$,

$$\begin{aligned}\max_{\mathbf{b}\in\mathcal{B}_\mathbf{v}} \mathbb{E}_{(P+Q)/2}[\mathsf{U}_\mathbf{v}(\mathbf{b};\boldsymbol{\beta}_-^t)] &\leq \max_{\mathbf{b}:1\geq b_1\geq\cdots\geq b_M\geq 0} \mathbb{E}_{(P+Q)/2}[\mathsf{U}_\mathbf{v}(\mathbf{b};\boldsymbol{\beta}_-^t)]\\ &= \frac{1}{2}\Big(\sum_{i=1}^{M}(1-\alpha b_i)\cdot(\mathbb{I}[b_i\geq\gamma c]+\mathbb{I}[b_i\geq c])\Big).\end{aligned}$$

Thus, we can only consider threshold strategies of the form $\mathbf{b}_j = [\gamma c,\ldots,\gamma c, c,\ldots,c]$ that contain j entries of γc and $M-j$ entries of c for $j\in\{0,\ldots,M\}$. For any $j\in\{0,\ldots,M\}$,

$$\mathbb{E}_{(P+Q)/2}[\mathsf{U}_\mathbf{v}(\mathbf{b}_j;\boldsymbol{\beta}_-^t)] = \frac{1}{2}\left(M(1-\alpha c)+j(1-\alpha c(2\gamma-1))\right) = \frac{1}{2}M(1-\alpha c),$$

since $\alpha c(2\gamma-1)=1$. Hence, $\max_{\mathbf{b}\in\mathcal{B}_\mathbf{v}} \mathbb{E}_{(P+Q)/2}[\mathsf{U}_\mathbf{v}(\mathbf{b};\boldsymbol{\beta}_-^t)] \leq \frac{1}{2}M(1-\alpha c)$. For any $\mathbf{b}\in\mathcal{B}_\mathbf{v}$,

$$\begin{aligned}&\max_{\mathbf{b}^*\in\mathcal{B}_\mathbf{v}} \mathbb{E}_P[\mathsf{U}_\mathbf{v}(\mathbf{b}^*;\boldsymbol{\beta}_-^t)-\mathsf{U}_\mathbf{v}(\mathbf{b};\boldsymbol{\beta}_-^t)] + \max_{\mathbf{b}^*\in\mathcal{B}_\mathbf{v}} \mathbb{E}_Q[\mathsf{U}_\mathbf{v}(\mathbf{b}^*;\boldsymbol{\beta}_-^t)-\mathsf{U}_\mathbf{v}(\mathbf{b};\boldsymbol{\beta}_-^t)]\\ &\geq \max_{\mathbf{b}^*\in\mathcal{B}_\mathbf{v}} \mathbb{E}_P[\mathsf{U}_\mathbf{v}(\mathbf{b}^*;\boldsymbol{\beta}_-^t)] + \max_{\mathbf{b}^*\in\mathcal{B}_\mathbf{v}} \mathbb{E}_Q[\mathsf{U}_\mathbf{v}(\mathbf{b}^*;\boldsymbol{\beta}_-^t)] - 2\max_{\mathbf{b}^*\in\mathcal{B}_\mathbf{v}} \mathbb{E}_{(P+Q)/2}[\mathsf{U}_\mathbf{v}(\mathbf{b}^*;\boldsymbol{\beta}_-^t)]\\ &= \Big(\frac{1}{2}+\delta\Big)M(1-\alpha c) + M(1-\gamma\alpha c) - M(1-\alpha c) = M(1-\alpha c)\delta = \frac{2M\delta(\gamma-1)}{2\gamma-1},\end{aligned}$$

where the last line uses $\alpha c(2\gamma-1)=1$.

Thus, any $\mathbf{b}\in\mathcal{B}_\mathbf{v}$ incurs a total regret of $\frac{MT\delta(\gamma-1)}{2\gamma-1}$ under P (Scenario 1), or a total regret of $\frac{MT\delta(\gamma-1)}{2\gamma-1}$ under Q (Scenario 2). By two-point method from Tsybakov (2009, Theorem 2.2),

$$\mathbb{E}_{(P+Q)/2}[\mathsf{Reg}_{nb}(T)] \geq \frac{MT\delta(\gamma-1)}{2\gamma-1}\cdot(1-\mathrm{TV}(P,Q)) \geq \frac{MT\delta(\gamma-1)}{2(2\gamma-1)}\exp\left(-16T\delta^2\right)$$

Setting $\delta = \frac{1}{4\sqrt{2T}}$, we get $\mathbb{E}_{(P+Q)/2}[\mathsf{Reg}_{nb}(T)] = \Omega(M\sqrt{T})$.

C.7 Proof of Theorem 4.1

Before analyzing regret, we first state the necessary changes in Algorithm 2 in the bandit setting.

Known context distribution. The algorithm is identical to Algorithm 2, except that edge weights $\omega_\mathbf{v}^t(e)$ are replaced with their unbiased estimators from Eq. (11), where $\omega_\mathbf{v}^t(e)$ is defined in Eq. (16).

Unknown context distribution. For any $\mathbf{v}\in\mathcal{V}$, round t, and edge $e=(\ell-1,b_{\ell-1},s_{\ell-1})\to(\ell,b_\ell,s_\ell)$ in layer $\ell\in[M]$, the edge weight is

$$\omega_\mathbf{v}^t(e) = \frac{\alpha+1/\rho+(v_\ell-(\alpha+\lambda_t)b_\ell)\cdot\mathbb{I}\left[b_\ell\geq\boldsymbol{\beta}_{-,t}^{-(\ell)}\right]}{1+\alpha+1/\rho}, \tag{35}$$

with the edge weights of all edges between later M and d as 0. Here $\lambda_t\in[0,1/\rho]$ is the dual variable, ensuring $\omega_\mathbf{v}^t(e)\in[0,1]$. In this case, the bidder computes the strategy in Algorithm 2, Line 4 via Algorithm 3, using the estimator in Eq. (14) with $\omega_\mathbf{v}^t(e)$ per Eq. (35).

Regret Analysis. We begin with the following result about the primal regret minimizer:

Lemma C.4. *For any sequence of dual variables $\{\lambda_t\}_{t\geq 1}$ where $\lambda_t \in [0, \frac{1}{\rho}]$, we have:*

$$\max_{\pi\in\Pi} \sum_{t=1}^{T} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t); \boldsymbol{\beta}_-^t) - \lambda_t \mathsf{P}(\pi(\mathbf{v}^t); \boldsymbol{\beta}_-^t)] - \sum_{t=1}^{T} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t) - \lambda_t \mathsf{P}(\boldsymbol{\beta}^t)] \lesssim \mathsf{R}_P^T \tag{36}$$

where $\boldsymbol{\beta}^t = (\mathbf{b}^t; \boldsymbol{\beta}_-^t), \forall t \geq 1$ and R_P^T is

- $\frac{M^{3/2}}{\rho}\sqrt{T \log 1/\epsilon}$ *in the full information setting.*
- $\frac{M^2}{\rho\epsilon^{3/2}}\sqrt{T \log 1/\epsilon}$ *in the bandit setting with known context distribution.*
- $\frac{M^2 T^{2/3} (\log 1/\epsilon)^{1/3}}{\rho\epsilon} + \frac{M^4 \log 1/\epsilon}{\rho\epsilon^3}$ *in the bandit setting with unknown context distribution.*

Proof. We prove the result for the full information setting. The results in the other settings follows accordingly. Following the analysis of Theorem 3.2 up to Eq. (25), we get

$$\sum_{t=1}^{T} \sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}] \omega_{\mathbf{v}}^t(\mathfrak{p}) - \sum_{t=1}^{T} \omega_{\mathbf{v}}^t(\widetilde{\pi}(\mathbf{v})) \lesssim \frac{M \log 1/\epsilon}{\eta_T} + \sum_{t=1}^{T} \sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}] \eta_t \omega_{\mathbf{v}}^t(\mathfrak{p})^2 .$$

Using the definition of edge weights in Eq. (16), we get

$$\omega_{\mathbf{v}}^t(\mathfrak{p}) = \sum_{e\in\mathfrak{p}} \omega_{\mathbf{v}}^t(e) = \frac{M - (\mathsf{U}_{\mathbf{v}}(\mathbf{b}; \boldsymbol{\beta}_-^t) - \lambda_t \mathsf{P}(\mathbf{b}; \boldsymbol{\beta}_-^t))}{1 + \alpha + 1/\rho} .$$

Then, continuing with analysis of Theorem 3.2 and using the fact that $1 + \alpha + 1/\rho \leq 3/\rho$ since $\alpha \leq 1$ and $\rho < 1$, we get the desired regret bound. □

Furthermore, the OGD algorithm (dual regret minimizer) satisfies the following regret bound:

Proposition C.1 (Hazan (2016))**.** *Let $\zeta_t > 0, \forall t \geq 1$ be the learning rate in round t. For any fixed $\lambda \in [0, \frac{1}{\rho}]$,*

$$\mathsf{Reg}_D(T) := \sum_{t=1}^{T} (\mathsf{P}(\boldsymbol{\beta}^t) - \rho M)(\lambda - \lambda_t) \lesssim \frac{1}{\zeta_T \rho^2} + M^2 \sum_{t=1}^{T} \zeta_t . \tag{37}$$

where $\boldsymbol{\beta}^t = (\mathbf{b}^t; \boldsymbol{\beta}_-^t), \forall t \geq 1$. Setting $\zeta_t = \frac{1}{\rho M\sqrt{t}}$ yields $\mathsf{Reg}_D(T) \lesssim \frac{M\sqrt{T}}{\rho}$.

Let $\tau = \min\{t \in [T] : B_t < M\}$, the stopping time of Algorithm 2. If no such round exists, define $\tau = T$. Let the primal regret upper bound (the right hand side of Eq. (36)) up to τ be denoted as R_P^τ. Rearranging Eq. (36) we get,

$$\sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)] \gtrsim \max_{\pi\in\Pi} \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t); \boldsymbol{\beta}_-^t) - \lambda_t \mathsf{P}(\pi(\mathbf{v}^t); \boldsymbol{\beta}_-^t)] + \sum_{t=1}^{\tau} \mathbb{E}[\lambda_t \mathsf{P}(\boldsymbol{\beta}^t)] - \mathsf{R}_P^\tau . \tag{38}$$

Similarly, define the OGD regret bound up to round τ as $\mathsf{R}_D^\tau = \frac{M\sqrt{\tau}}{\rho}$. Rearranging Eq. (37), we get

$$\sum_{t=1}^{\tau} \lambda_t \mathsf{P}(\boldsymbol{\beta}^t) \gtrsim \sum_{t=1}^{\tau} \rho M(\lambda_t - \lambda) + \sum_{t=1}^{\tau} \lambda \mathsf{P}(\boldsymbol{\beta}^t) - \mathsf{R}_D^\tau . \tag{39}$$

Substituting in Eq. (38), we get

$$\begin{aligned}
&\sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)] \gtrsim \max_{\pi\in\Pi} \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t); \boldsymbol{\beta}_-^t) - \lambda_t \mathsf{P}(\pi(\mathbf{v}^t); \boldsymbol{\beta}_-^t)] + \sum_{t=1}^{\tau} \mathbb{E}[\rho M(\lambda_t - \lambda)] + \sum_{t=1}^{\tau} \lambda \mathbb{E}[\mathsf{P}(\boldsymbol{\beta}^t)] - \mathsf{R}_D^\tau - \mathsf{R}_P^\tau \\
&= \max_{\pi\in\Pi} \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t); \boldsymbol{\beta}_-^t) + \lambda_t(\rho M - \mathsf{P}(\pi(\mathbf{v}^t); \boldsymbol{\beta}_-^t))] + \sum_{t=1}^{\tau} \lambda \mathbb{E}[\mathsf{P}(\boldsymbol{\beta}^t)] - \mathsf{R}_D^\tau - \mathsf{R}_P^\tau - \lambda\rho M\tau .
\end{aligned} \tag{40}$$

Define

$$\mathsf{OPT}^{\tau}_{nb} = \max_{\pi\in\Pi} \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t);\boldsymbol{\beta}^t_-)] \tag{41}$$

$$\pi^* = \operatorname*{argmax}_{\pi\in\Pi} \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t);\boldsymbol{\beta}^t_-)] \tag{42}$$

In words, OPT^{τ}_{nb} is the maximum utility obtained by a stationary policy in Π over first τ rounds without the budget constraints and π^* is the stationary policy achieving it. Then,

Claim 4. For any sequence of competing bids $\{\boldsymbol{\beta}^t_-\}_{t\geq 1}$, dual variables $\{\lambda_t\}_{t\geq 1}$ and $\rho < 1$, we have

$$\max_{\pi\in\Pi} \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t);\boldsymbol{\beta}^t_-) + \lambda_t(\rho M - \mathsf{P}(\pi(\mathbf{v}^t);\boldsymbol{\beta}^t_-))] \geq \rho \cdot \mathsf{OPT}^{\tau}_{nb}\,.$$

where OPT^{τ}_{nb} is defined per Eq. (41).

Proof. By definition,

$$\max_{\pi\in\Pi} \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t);\boldsymbol{\beta}^t_-) + \lambda_t(\rho M - \mathsf{P}(\pi(\mathbf{v}^t);\boldsymbol{\beta}^t_-))] \geq \max_{\pi^*,\perp} \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t);\boldsymbol{\beta}^t_-) + \lambda_t(\rho M - \mathsf{P}(\pi(\mathbf{v}^t);\boldsymbol{\beta}^t_-))], \tag{43}$$

where π^* is per Eq. (42) and $\perp \in \Pi$ is the policy that maps all valuation vectors to the bidding strategy $\mathbf{0}$. Now, consider two cases:

1. Suppose $\sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-) - \lambda_t \mathsf{P}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-))] \geq 0$. Then,

$$\begin{aligned}
\max_{\pi\in\Pi} \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t);\boldsymbol{\beta}^t_-) + \lambda_t(\rho M - \mathsf{P}(\pi(\mathbf{v}^t);\boldsymbol{\beta}^t_-))] &\geq \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-) + \lambda_t(\rho M - \mathsf{P}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-))] \\
&\geq \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-) - (1-\rho)\lambda_t \mathsf{P}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-)] \\
&\geq \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-) - (1-\rho)\mathsf{U}_{\mathbf{v}^t}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-)] \\
&= \rho \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-)] \\
&= \rho \cdot \mathsf{OPT}^{\tau}_{nb}\,.
\end{aligned}$$

   Here, the second inequality holds as $\mathsf{P}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-) \leq M$ and third inequality holds because by assumption, $\sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-)] \geq \sum_{t=1}^{\tau} \mathbb{E}[\lambda_t \mathsf{P}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-))]$.

2. Suppose $\sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-) - \lambda_t \mathsf{P}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-))] < 0$. Setting $\pi = \perp$ in the right hand side of Eq. (43),

$$\begin{aligned}
\max_{\pi\in\Pi} \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t);\boldsymbol{\beta}^t_-) + \lambda_t(\rho M - \mathsf{P}(\pi(\mathbf{v}^t);\boldsymbol{\beta}^t_-))] &\geq \rho M \sum_{t=1}^{\tau} \mathbb{E}[\lambda_t] \\
&\geq \rho \sum_{t=1}^{\tau} \mathbb{E}[\lambda_t \mathsf{P}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-)] \\
&\geq \rho \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi^*(\mathbf{v}^t);\boldsymbol{\beta}^t_-)] \\
&= \rho \cdot \mathsf{OPT}^{\tau}_{nb}\,.
\end{aligned}$$

□

Substituting the result of Claim 4 in Eq. (40) and rearranging, we get

$$\sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t) - \lambda \mathsf{P}(\boldsymbol{\beta}^t)] \gtrsim \rho \cdot \mathsf{OPT}^{\tau}_{nb} - \mathsf{R}^{\tau}_D - \mathsf{R}^{\tau}_P - \lambda \rho M \tau$$

Claim 5. Recall that OPT is the optimal value of (Opt). Then, $\mathsf{OPT}^{\tau}_{nb} \geq \mathsf{OPT} - M(T-\tau)$.

Proof. With slight abuse of notation, define

$$\mathsf{OPT}^{s} := \max_{\pi \in \Pi} \sum_{t=1}^{s} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t); \boldsymbol{\beta}^t_{-})] \quad \text{such that} \quad \sum_{t=1}^{s} \mathsf{P}(\pi(\mathbf{v}^t); \boldsymbol{\beta}^t_{-}) \leq \rho M T\,.$$

Observe that $\mathsf{OPT}^{\tau}_{nb} \geq \mathsf{OPT}^{\tau}$ because OPT^{τ} is the optimal objective value of the constrained problem whereas OPT^{τ}_{nb} is the optimal objective value of the unconstrained problem.

Let $\pi^{\dagger} \in \Pi$ be the policy that maximizes (Opt). Then,

$$\begin{aligned}
\mathsf{OPT} &= \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi^{\dagger}(\mathbf{v}^t); \boldsymbol{\beta}^t_{-})] + \sum_{t=\tau+1}^{T} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi^{\dagger}(\mathbf{v}^t); \boldsymbol{\beta}^t_{-})] \quad \text{such that} \quad \sum_{t=1}^{T} \mathsf{P}(\pi^{\dagger}(\mathbf{v}^t); \boldsymbol{\beta}^t_{-}) \leq \rho M T \\
&\leq \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi^{\dagger}(\mathbf{v}^t); \boldsymbol{\beta}^t_{-})] + M(T-\tau) \quad \text{such that} \quad \sum_{t=1}^{T} \mathsf{P}(\pi^{\dagger}(\mathbf{v}^t); \boldsymbol{\beta}^t_{-}) \leq \rho M T \\
&\leq \max_{\pi \in \Pi} \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\pi(\mathbf{v}^t); \boldsymbol{\beta}^t_{-})] + M(T-\tau) \quad \text{such that} \quad \sum_{t=1}^{\tau} \mathsf{P}(\pi(\mathbf{v}^t); \boldsymbol{\beta}^t_{-}) \leq \rho M T \\
&= \mathsf{OPT}^{\tau} + M(T-\tau)\,.
\end{aligned}$$

Hence, $\mathsf{OPT}^{\tau} \geq \mathsf{OPT} - M(T-\tau) \implies \mathsf{OPT}^{\tau}_{nb} \geq \mathsf{OPT} - M(T-\tau)$. □

Hence, by Claim 5, we have

$$\sum_{t=1}^{T} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)] \geq \sum_{t=1}^{\tau} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)] \gtrsim \rho \cdot (\mathsf{OPT} - M(T-\tau)) + \sum_{t=1}^{\tau} \lambda \mathbb{E}[\mathsf{P}(\boldsymbol{\beta}^t)] - \mathsf{R}^{\tau}_D - \mathsf{R}^{\tau}_P - \lambda \rho M \tau\,.$$

If $\tau = T$, set $\lambda = 0$. Rearranging the terms and substituting the value of R^{T}_D yields

$$\rho \cdot \mathsf{Reg}(T) = \rho \cdot \mathsf{OPT} - \sum_{t=1}^{T} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)] \lesssim \frac{M\sqrt{T}}{\rho} + \mathsf{R}^{T}_P. \tag{44}$$

On the other hand, if $\tau < T$, $\sum_{t=1}^{\tau} \mathsf{P}(\boldsymbol{\beta}^t) + M \geq \rho M T$. So,

$$\sum_{t=1}^{T} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)] \gtrsim \rho \cdot (\mathsf{OPT} - M(T-\tau)) + \lambda(\rho M T - M) - \mathsf{R}^{\tau}_D - \mathsf{R}^{\tau}_P - \lambda \rho M \tau.$$

Setting $\lambda = \frac{1}{\rho}$,

$$\sum_{t=1}^{T} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)] \gtrsim \rho \cdot (\mathsf{OPT} - M(T-\tau)) + M(T-\tau) - \frac{M}{\rho} - \mathsf{R}^{\tau}_D - \mathsf{R}^{\tau}_P.$$

Rearranging,

$$\begin{aligned}
\rho \cdot \mathsf{Reg}(T) = \rho \cdot \mathsf{OPT} - \sum_{t=1}^{T} \mathbb{E}[\mathsf{U}_{\mathbf{v}^t}(\boldsymbol{\beta}^t)] &\lesssim \rho M(T-\tau) - M(T-\tau) + \frac{M}{\rho} + \mathsf{R}^{\tau}_D + \mathsf{R}^{\tau}_P \\
\implies \rho \cdot \mathsf{Reg}(T) &\lesssim \frac{M}{\rho} + \frac{M\sqrt{T}}{\rho} + \mathsf{R}^{T}_P,
\end{aligned}$$

where the last inequality follows as $\rho \leq 1$ and $\tau \leq T$. Combining with Eq. (44), we get the desired regret bound.

D Efficient Implementation

D.1 Full Information Setting

In the full-information setting, for each $\mathbf{v} \in \mathcal{V}$ and each edge $e = (\ell-1, b_{\ell-1}, s_{\ell-1}) \to (\ell, b_\ell, s_\ell)$ in layer $\ell \in [M]$,

$$\omega^t_{\mathbf{v}}(e) = \frac{1 - (v_\ell - \alpha b_\ell) \cdot \mathbb{I}\left[b_\ell \geq \boldsymbol{\beta}^{-(\ell)}_{-,t}\right]}{1+\alpha} =: x^t(e)\, v_\ell + y^t(e). \qquad \text{((10) restated)}$$

All edges from nodes in layer M to the destination node d have weight 0, and

$$x^t(e) = -\frac{\mathbb{I}\left[b_\ell \geq \boldsymbol{\beta}^{-(\ell)}_{-,t}\right]}{1+\alpha}, \qquad y^t(e) = \frac{1 + \alpha b_\ell \cdot \mathbb{I}\left[b_\ell \geq \boldsymbol{\beta}^{-(\ell)}_{-,t}\right]}{1+\alpha}. \qquad \text{((18) restated)}$$

Having defined the coefficients $x^t(e)$ and $y^t(e)$, we now describe how to update them across rounds. To this end, for each edge e define the cumulative (scaled) coefficients

$$\widehat{x}^t(e) := -\eta_{t+1} \sum_{s=1}^{t} x^s(e), \qquad \widehat{y}^t(e) := -\eta_{t+1} \sum_{s=1}^{t} y^s(e),$$

for $t \geq 1$, and let $\widehat{x}^0(e) = \widehat{y}^0(e) = 0$. It follows that

$$\widehat{x}^t(e) = \frac{\eta_{t+1}}{\eta_t}\, \widehat{x}^{t-1}(e) - \eta_{t+1} x^t(e), \tag{45}$$

$$\widehat{y}^t(e) = \frac{\eta_{t+1}}{\eta_t}\, \widehat{y}^{t-1}(e) - \eta_{t+1} y^t(e). \tag{46}$$

Overview of the Algorithm. The algorithm is similar to Algorithm 1 but differs in three key ways: (i) how we instantiate context-dependent DAGs (cf. Algorithm 1, Line 3); (ii) how we compute the sampling probabilities $\phi^t_{\mathbf{v}}(\cdot)$ (cf. Eqs. (8) and (9)); and (iii) how we set the edge weights across contexts (cf. Algorithm 1, Line 11). Concretely:

(i) On-the-fly instantiation. We instantiate the context-dependent DAG only for the current context $\mathbf{v}^t = \mathbf{v}$, rather than for all $\mathbf{v} \in \mathcal{V}$.

(ii) Computing $\phi^t_{\mathbf{v}}(\cdot)$. For any $\mathbf{v}$, set $\Gamma^{t-1}_{\mathbf{v}}(d) = 1$ and compute $\Gamma^{t-1}_{\mathbf{v}}(\cdot)$ bottom-up. For each node $u = (\ell-1, b_{\ell-1}, s_{\ell-1}) \in \mathsf{N}_{\mathbf{v}}$, define

$$\Gamma^{t-1}_{\mathbf{v}}(u) = \sum_{v=(\ell, b_\ell, s_\ell):\, u \to v =: e \in \mathsf{E}_{\mathbf{v}}} \Gamma^{t-1}_{\mathbf{v}}(v) \cdot \exp\big(\widehat{x}^{t-1}(e)\, v_\ell + \widehat{y}^{t-1}(e)\big). \tag{47}$$

Then, for each edge $e = (\ell-1, b_{\ell-1}, s_{\ell-1}) \to (\ell, b_\ell, s_\ell) \in \mathsf{E}_{\mathbf{v}}$, set

$$\phi^t_{\mathbf{v}}(e) = \exp\big(\widehat{x}^{t-1}(e)\, v_\ell + \widehat{y}^{t-1}(e)\big) \cdot \frac{\Gamma^{t-1}_{\mathbf{v}}(\ell, b_\ell, s_\ell)}{\Gamma^{t-1}_{\mathbf{v}}(\ell-1, b_{\ell-1}, s_{\ell-1})}. \tag{48}$$

Computing $\phi^t_{\mathbf{v}}(\cdot)$ for any $\mathbf{v}$ only requires maintaining the shared coefficients $\{\widehat{x}^{t-1}(e), \widehat{y}^{t-1}(e)\}_{e \in \overline{\mathsf{E}}}$. Once $\phi^t_{\mathbf{v}}(\cdot)$ is available, we sample a path $\mathfrak{p}$ in a Markovian fashion as in Section 3.

(iii) Updating shared coefficients. Finally, instead of explicitly instantiating per-round edge weights as in Eq. (10), we update the shared coefficients using Eqs. (45) and (46).

The complete algorithm is given in Algorithm 5. Our main result is the following.

Theorem D.1. *In the full-information setting without budget constraints, Algorithm 5 implements Algorithm 1 using $O(|\overline{\mathsf{E}}|) = O(M^2/\epsilon^3)$ time and space per round, while achieving the same regret bound.*

Algorithm 5 No Budget Constraints (Full Information) – Efficient Implementation

Require: Learning rates $\eta_t > 0, \forall t \geq 1$. Define $\eta_0 = 1$. For all $e \in \overline{\mathsf{E}}$, define $\widehat{x}^0(e) = \widehat{y}^0(e) = 0$.
for $t = 1, 2, \ldots$ **do**
  Observe an i.i.d. valuation vector sample $\mathbf{v}^t \sim \mathcal{D}$. Suppose $\mathbf{v}^t = \mathbf{v}$.
  Construct $\mathcal{G}^t(\mathsf{N}_\mathbf{v}, \mathsf{E}_\mathbf{v}, \omega_\mathbf{v})$ and obtain edge probabilities $\phi^t_\mathbf{v}(\cdot)$ following Eq. (47) and Eq. (48).
  Define initial node $u = s$ and path $\mathfrak{p}^t = s$.
  while $u \neq d$ **do**
    Sample v with probability $\phi^t_\mathbf{v}(u \to v)$.
    Append v to the path $\mathfrak{p}^t$; set $u \leftarrow v$.
  Map $\mathfrak{p}^t = s \to (1, b_1, s_1) \to \cdots \to (M, b_M, s_M) \to d$, and submit $\mathbf{b}^t = [b_1, \ldots, b_M]$.
  Update $\widehat{x}^t(e)$ and $\widehat{y}^t(e)$ for all $e \in \overline{\mathsf{E}}$ per Eq. (45) and Eq. (46) respectively.

D.2 Bandit Setting

Known Context Distribution. In this setting, for any edge $e = (\ell - 1, b_{\ell-1}, s_{\ell-1}) \to (\ell, b_\ell, s_\ell)$, the edge-weight estimator $\widehat{\omega}^t_\mathbf{v}(e)$ in Eq. (11) can be written as $\widehat{\omega}^t_\mathbf{v}(e) = x^t(e) \cdot v_\ell + y^t(e)$, where

$$x^t(e) = \frac{-\mathbb{I}\left[b_\ell \geq \boldsymbol{\beta}^{-(\ell)}_{-,t}\right] \cdot \mathbb{I}[e \in \mathfrak{p}^t]}{q^t(e)\,(1+\alpha)}, \qquad y^t(e) = \frac{\left(1 + \alpha b_\ell \cdot \mathbb{I}\left[b_\ell \geq \boldsymbol{\beta}^{-(\ell)}_{-,t}\right]\right) \cdot \mathbb{I}[e \in \mathfrak{p}^t]}{q^t(e)\,(1+\alpha)}.$$

The main computational bottleneck is evaluating $q^t(e)$, the unconditional probability of selecting edge e in round t, which requires taking an expectation with respect to the context distribution $\mathcal{D}$ (cf. Eq. (12)). In Appendix C.4, we give a procedure to compute $q^t(e)$ in $O(|\mathcal{V}| \cdot \max_{\mathbf{v} \in \mathcal{V}} |\mathsf{E}_\mathbf{v}|) = O\big(|\mathcal{V}|\,|\overline{\mathsf{E}}|\big)$ time. Whenever $q^t(e)$ can be computed more efficiently, the same implementation ideas used in the full-information setting apply directly here as well.

Unknown Context Distribution. Recall that in this setting, depending on the mixing parameter $\delta \in (0, 1]$, the algorithm either (i) samples a path from the EPC associated with the realized context, or (ii) samples according to exponential-weight updates; see the overview in Section 3.2.2 and the full procedure in Algorithm 3 in Appendix C.5.

In case (i), sampling from the EPC corresponding to the realized context can be implemented on the fly and requires $O(\max_{\mathbf{v} \in \mathcal{V}} |\mathsf{E}_\mathbf{v}|) = O(|\overline{\mathsf{E}}|)$ time and space per round (as discussed in Section 3.2.2). In case (ii), the shared-coefficient idea from the full-information setting extends to the bandit setting with minor modifications.

Recall that for edge $e = (\ell - 1, b_{\ell-1}, s_{\ell-1}) \to (\ell, b_\ell, s_\ell)$ in layer $\ell \in [M]$, and valuation vector $\mathbf{v}$, we defined

$$\omega^t_\mathbf{v}(e) = \frac{\alpha + (v_\ell - \alpha b_\ell) \cdot \mathbb{I}\left[b_\ell \geq \boldsymbol{\beta}^{-(\ell)}_{-,t}\right]}{1 + \alpha}. \qquad \text{((13) restated)}$$

All edges from nodes in layer M to the destination node d have weight 0. If $\mathbf{v}^t = \mathbf{v}'$, the edge-weight estimator in this setting is

$$\widehat{\omega}^t_\mathbf{v}(e) = \frac{\omega^t_\mathbf{v}(e)}{p^t_{\mathbf{v}'}(e)} \cdot \mathbb{I}\big[e \in \mathfrak{p}^t\big], \qquad \text{((14) restated)}$$

where $p^t_{\mathbf{v}'}(e) = \sum_{\mathfrak{p} \in \mathscr{P}_{\mathbf{v}'}: e \in \mathfrak{p}} \mathbb{P}[\mathfrak{p}^t = \mathfrak{p} | \mathbf{v}^t = \mathbf{v}']$. Following ideas similar to the full information setting, for $\mathbf{v} \in \mathcal{V}$, and edge $e = (\ell - 1, b_{\ell-1}, s_{\ell-1}) \to (\ell, b_\ell, s_\ell)$ in layer $\ell \in [M]$, we can express

$$\widehat{\omega}^t_\mathbf{v}(e) = x^t(e) \cdot v_\ell + y^t(e),$$

where

$$x^t(e) = \frac{\mathbb{I}\left[b_\ell \geq \boldsymbol{\beta}^{-(\ell)}_{-,t}\right] \cdot \mathbb{I}[e \in \mathfrak{p}^t]}{p^t_{\mathbf{v}'}(e)(1+\alpha)}, \qquad y^t(e) = \frac{(\alpha - \alpha b_\ell \cdot \mathbb{I}\left[b_\ell \geq \boldsymbol{\beta}^{-(\ell)}_{-,t}\right]) \cdot \mathbb{I}[e \in \mathfrak{p}^t]}{p^t_{\mathbf{v}'}(e)(1+\alpha)}. \qquad (49)$$

Theorem D.2. *In the bandit setting under unknown context distribution without budget constraints, Algorithm 6 implements Algorithm 3 using $O(|\overline{\mathsf{E}}|) = O(M^2/\epsilon^3)$ time and space per round, while achieving the same regret bound.*

Algorithm 6 No Budget Constraint, Bandit Setting, Unknown Distribution – Efficient Implementation

Require: Learning rates $\eta_t > 0$, $\delta \in (0, 1]$. Define $\eta_0 = 1$. For all $e \in \overline{\mathsf{E}}$, define $\widehat{x}^0(e) = \widehat{y}^0(e) = 0$.
for $t = 1, 2, \ldots, T$ **do**
  Observe an i.i.d. valuation vector sample $\mathbf{v}^t \sim \mathcal{D}$. Let $\mathbf{v}^t = \mathbf{v}'$.
  Sample $Z_t \sim \text{Unif}[0, 1]$.
  if $Z_t \le \delta$ **then**
    Construct $\mathcal{G}^t(\mathsf{N}_{\mathbf{v}'}, \mathsf{E}_{\mathbf{v}'}, \omega_{\mathbf{v}'})$ and select path $\mathfrak{p}^t$ using Algorithm 4.
  else
    Construct $\mathcal{G}^t(\mathsf{N}_{\mathbf{v}'}, \mathsf{E}_{\mathbf{v}'}, \omega_{\mathbf{v}'})$ and obtain edge probabilities $\phi^t_{\mathbf{v}'}(\cdot)$ following Eq. (47) and Eq. (48).
    Define initial node $u = s$ and path $\mathfrak{p}^t = s$.
    while $u \neq d$ **do**
      Sample v with probability $\phi^t_{\mathbf{v}'}(u \to v)$.
      Append v to the path $\mathfrak{p}^t$; set $u \leftarrow v$.
  Map $\mathfrak{p}^t = s \to (1, b_1, s) \to \cdots \to (M, b_M, s_M) \to d$, and submit $\mathbf{b}^t = [b_1, \ldots, b_M]$.
  Update $\widehat{x}^t(e)$ and $\widehat{y}^t(e)$ for all $e \in \overline{\mathsf{E}}$ per Eq. (45) and Eq. (46) respectively, where $x^t(e)$ and $y^t(e)$ are defined in Eq. (49).

Proof. Observe that once the coefficients $\{x^t(e), y^t(e)\}_{e \in \overline{\mathsf{E}}}$ are available, the Hedge-style updates in Algorithm 6 follow the same structure as in the full-information implementation (cf. Algorithm 5). Together with the sampling subroutine described in Algorithm 4, this shows that Algorithm 6 is equivalent to Algorithm 3. Thus the regret bound remains the same in this setting.

To obtain the space and time complexity, note that if the algorithm samples a path uniformly at random, this can be done in $O(|\mathsf{E}_{\mathbf{v}'}|)$ time via Algorithm 4. If instead it performs exponential-weight updates, then for the realized context $\mathbf{v}'$, the edge marginals $\{p^t_{\mathbf{v}'}(e)\}_{e \in \mathsf{E}_{\mathbf{v}'}}$ can be computed in $O(|\mathsf{E}_{\mathbf{v}'}|)$ time and space (see Appendix C.5). The remaining steps mirror the full-information implementation and likewise take $O(|\mathsf{E}_{\mathbf{v}'}|)$ time and space. Therefore, the overall per-round complexity of Algorithm 6 is $O(|\overline{\mathsf{E}}|)$ in both time and space. □

D.3 Proof of Theorem D.1

The key idea of the proof is to show that Algorithm 5 is an efficient implementation of Decreasing Hedge (equivalently, Algorithm 1) and, crucially, that its time and space complexity are independent of the number of contexts.

Recall that in the full-information setting, for each $\mathbf{v} \in \mathcal{V}$ and each edge $e = (\ell - 1, b_{\ell-1}, s_{\ell-1}) \to (\ell, b_\ell, s_\ell)$ in layer $\ell \in [M]$, the edge weight is of the form

$$\omega^t_{\mathbf{v}}(e) = x^t(e)\, v_\ell + y^t(e)$$

and all edges from nodes in layer M to the destination node d have weight 0. Here,

$$x^t(e) = -\frac{\mathbb{I}\left[b_\ell \ge \boldsymbol{\beta}^{-(\ell)}_{-,t}\right]}{1+\alpha}, \qquad y^t(e) = \frac{1 + \alpha b_\ell \cdot \mathbb{I}\left[b_\ell \ge \boldsymbol{\beta}^{-(\ell)}_{-,t}\right]}{1+\alpha}.$$

We further defined

$$\widehat{x}^t(e) := -\eta_{t+1} \sum_{s=1}^{t} x^s(e), \qquad \widehat{y}^t(e) := -\eta_{t+1} \sum_{s=1}^{t} y^s(e),$$

where $\widehat{x}^0(e) = \widehat{y}^0(e) = 0$. This gives

$$\widehat{x}^t(e) = \frac{\eta_{t+1}}{\eta_t}\, \widehat{x}^{t-1}(e) - \eta_{t+1} x^t(e), \qquad \text{((45) restated)}$$

$$\widehat{y}^t(e) = \frac{\eta_{t+1}}{\eta_t}\, \widehat{y}^{t-1}(e) - \eta_{t+1} y^t(e). \qquad \text{((46) restated)}$$

Recall that $\omega^s_{\mathbf{v}}(\mathfrak{p}) = \sum_{e \in \mathfrak{p}} \omega^s_{\mathbf{v}}(e)$. Then, for any $\mathfrak{p} = s \to (1, b_1, s_1) \to \cdots \to (M, b_M, s_M) \to d$,

$$-\eta_t \sum_{s=1}^{t-1} \omega^s_{\mathbf{v}}(\mathfrak{p}) = \langle \widehat{\mathbf{x}}^{t-1}(\mathfrak{p}), \mathbf{v} \rangle + \widehat{y}^{t-1}(\mathfrak{p}), \tag{50}$$

where $\widehat{\mathbf{x}}^{t-1}(\mathfrak{p}) \in \mathbb{R}^M$ and the ℓ^{th} coordinate corresponding to $e = (\ell-1, b_{\ell-1}, s_{\ell-1}) \to (\ell, b_\ell, s_\ell)$ is

$$[\widehat{\mathbf{x}}^{t-1}(\mathfrak{p})]_\ell = \widehat{x}^{t-1}(e), \quad \text{and} \quad \widehat{y}^{t-1}(\mathfrak{p}) = \sum_{e \in \mathfrak{p}} \widehat{y}^{t-1}(e)\,.$$

Recall that in round t, the probability of selecting a path $\mathfrak{p}$ is $\mathbb{P}^t_{\mathbf{v}}(\mathfrak{p}) = \prod_{e\in\mathfrak{p}} \phi^t_{\mathbf{v}}(e)$, and for any edge $e = (\ell-1, b_{\ell-1}, s_{\ell-1}) \to (\ell, b_\ell, s_\ell) \in \mathsf{E}_{\mathbf{v}}$, the edge probabilities are

$$\phi^t_{\mathbf{v}}(e) = \exp(\widehat{x}^{t-1}(e)\cdot v_\ell + \widehat{y}^{t-1}(e)) \cdot \frac{\Gamma^{t-1}_{\mathbf{v}}(\ell, b_\ell, s_\ell)}{\Gamma^{t-1}_{\mathbf{v}}(\ell-1, b_{\ell-1}, s_{\ell-1})}, \qquad \text{((48) restated)}$$

where $\Gamma^{t-1}_{\mathbf{v}}(d) = 1$ and for $\Gamma^{t-1}_{\mathbf{v}}(\cdot)$ is computed recursively in a bottom-to-top fashion for every $u = (\ell-1, b_{\ell-1}, s_{\ell-1}) \in \mathsf{N}_{\mathbf{v}}$:

$$\Gamma^{t-1}_{\mathbf{v}}(u) = \sum_{v=(\ell,b_\ell,s_\ell): u\to v =: e \in \mathsf{E}_{\mathbf{v}}} \Gamma^{t-1}_{\mathbf{v}}(v) \cdot \exp(\widehat{x}^{t-1}(e)\cdot v_\ell + \widehat{y}^{t-1}(e))\,. \qquad \text{((47) restated)}$$

Here, $\eta_0 = 1$ and $\gamma_t = \frac{\eta_t}{\eta_{t-1}}, \forall t \geq 1$. Recall that in a naïve implementation of the Decreasing Hedge algorithm, the probability of selecting path $\mathfrak{p}$ in round t is

$$\widehat{\mathbb{P}}^t_{\mathbf{v}}(\mathfrak{p}) = \frac{\exp(-\eta_t \sum_{s=1}^{t-1} \omega^s_{\mathbf{v}}(\mathfrak{p}))}{\sum_{\mathfrak{p}'\in\mathscr{P}_{\mathbf{v}}} \exp(-\eta_t \sum_{s=1}^{t-1} \omega^s_{\mathbf{v}}(\mathfrak{p}'))}, \qquad \text{((22) restated)}$$

where $\mathscr{P}_{\mathbf{v}}$ is the set of all s-d paths in $\mathcal{G}(\mathsf{N}_{\mathbf{v}}, \mathsf{E}_{\mathbf{v}}, \omega_{\mathbf{v}})$. To show that Algorithm 5 is an efficient implementation of Decreasing Hedge, we prove that $\mathbb{P}^t_{\mathbf{v}}(\mathfrak{p}) = \widehat{\mathbb{P}}^t_{\mathbf{v}}(\mathfrak{p})$. Since $\mathbb{P}^t_{\mathbf{v}}(\mathfrak{p}) = \prod_{e\in\mathfrak{p}} \phi^t_{\mathbf{v}}(e)$ and

$$\widehat{\mathbb{P}}^t_{\mathbf{v}}(\mathfrak{p}) \overset{(22)}{=} \frac{\exp(-\eta_t \sum_{s=1}^{t-1} \omega^s_{\mathbf{v}}(\mathfrak{p}))}{\sum_{\mathfrak{p}'\in\mathscr{P}_{\mathbf{v}}} \exp(-\eta_t \sum_{s=1}^{t-1} \omega^s_{\mathbf{v}}(\mathfrak{p}'))} \overset{(50)}{=} \frac{\exp(\langle \widehat{\mathbf{x}}^{t-1}(\mathfrak{p}), \mathbf{v}\rangle + \widehat{y}^{t-1}(\mathfrak{p}))}{\sum_{\mathfrak{p}'\in\mathscr{P}_{\mathbf{v}}} \exp(\langle \widehat{\mathbf{x}}^{t-1}(\mathfrak{p}'), \mathbf{v}\rangle + \widehat{y}^{t-1}(\mathfrak{p}'))},$$

it suffices to show that

$$\prod_{e\in\mathfrak{p}} \phi^t_{\mathbf{v}}(e) = \frac{\exp(\langle \widehat{\mathbf{x}}^{t-1}(\mathfrak{p}), \mathbf{v}\rangle + \widehat{y}^{t-1}(\mathfrak{p}))}{\sum_{\mathfrak{p}'\in\mathscr{P}_{\mathbf{v}}} \exp(\langle \widehat{\mathbf{x}}^{t-1}(\mathfrak{p}'), \mathbf{v}\rangle + \widehat{y}^{t-1}(\mathfrak{p}'))}\,.$$

To this end, we first show the following result (similar to Claim 2):

Claim 6. For any node $u = (\ell-1, b_{\ell-1}, s_{\ell-1}) \in \mathsf{N}_{\mathbf{v}}$, let $\mathscr{P}_{\mathbf{v}}(u)$ be the set of paths starting at u and terminating in d. Then,

$$\Gamma^{t-1}_{\mathbf{v}}(u) = \sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}(u)} \prod_{e\in\mathfrak{p}} \exp(\widehat{x}^{t-1}(e)\cdot v_k + \widehat{y}^{t-1}(e))\,.$$

where $e = (k-1, b_{k-1}, s_{k-1}) \to (k, b_k, s_k)$ is an edge between layer $k-1$ and layer k for $\ell \leq k \leq M$.

Proof. We prove the result by backward induction. For the base case, $\Gamma^{t-1}_{\mathbf{v}}(d) = 1$. Suppose the result holds true for all the nodes in layer ℓ for some $\ell \in [M+1]$. Then, for any node $v = (\ell, b_\ell, s_\ell)$ in layer ℓ, we have

$$\Gamma^{t-1}_{\mathbf{v}}(v) = \sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}(v)} \prod_{e'\in\mathfrak{p}} \exp(\widehat{x}^{t-1}(e')\cdot v_j + \widehat{y}^{t-1}(e'))\,. \tag{51}$$

where $e = (j-1, b_{j-1}, s_{j-1}) \to (j, b_j, s_j)$ is an edge between layer $j-1$ and layer j for $\ell+1 \leq j \leq M$. Thus, for $u = (\ell-1, b_{\ell-1}, s_{\ell-1})$ in layer $\ell-1$,

$$\begin{aligned}
\Gamma^{t-1}_{\mathbf{v}}(u) &\overset{(47)}{=} \sum_{v=(\ell,b_\ell,s_\ell): u\to v =: e\in\mathsf{E}_{\mathbf{v}}} \Gamma^{t-1}_{\mathbf{v}}(v) \cdot \exp(\widehat{x}^{t-1}(e)\cdot v_\ell + \widehat{y}^{t-1}(e)) \\
&\overset{(51)}{=} \sum_{v=(\ell,b_\ell,s_\ell): u\to v =: e\in\mathsf{E}_{\mathbf{v}}} \Big(\sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}(v)} \prod_{e'\in\mathfrak{p}} \exp(\widehat{x}^{t-1}(e')\cdot v_j + \widehat{y}^{t-1}(e')) \Big) \cdot \exp(\widehat{x}^{t-1}(e)\cdot v_\ell + \widehat{y}^{t-1}(e)) \\
&= \sum_{\mathfrak{p}\in\mathscr{P}_{\mathbf{v}}(u)} \prod_{e\in\mathfrak{p}} \exp(\widehat{x}^{t-1}(e)\cdot v_k + \widehat{y}^{t-1}(e))\,.
\end{aligned}$$

□

Hence, the probability of sampling path $\mathfrak{p}$ in round t when the context is $\mathbf{v}$ is

$$\begin{aligned}\prod_{e\in\mathfrak{p}}\phi^t_{\mathbf{v}}(e) &\overset{(48)}{=} \prod_{e=(\ell-1,b_{\ell-1},s_{\ell-1})\to(\ell,b_\ell,s_\ell)\in\mathfrak{p}} \exp(\widehat{x}^{t-1}(e)\cdot v_\ell+\widehat{y}^{t-1}(e))\cdot\frac{\Gamma^{t-1}_{\mathbf{v}}(\ell,b_\ell,s_\ell)}{\Gamma^{t-1}_{\mathbf{v}}(\ell-1,b_{\ell-1},s_{\ell-1})}\\ &= \frac{\Gamma^{t-1}_{\mathbf{v}}(d)}{\Gamma^{t-1}_{\mathbf{v}}(s)}\prod_{e=(\ell-1,b_{\ell-1},s_{\ell-1})\to(\ell,b_\ell,s_\ell)\in\mathfrak{p}} \exp(\widehat{x}^{t-1}(e)\cdot v_\ell+\widehat{y}^{t-1}(e))\\ &= \frac{\exp(\langle\widehat{\mathbf{x}}^{t-1}(\mathfrak{p}),\mathbf{v}\rangle+\widehat{y}^{t-1}(\mathfrak{p}))}{\Gamma^{t-1}_{\mathbf{v}}(s)},\end{aligned}$$

where the last inequality follows from the definition of $\widehat{\mathbf{x}}^{t-1}(\mathfrak{p}),\widehat{y}^{t-1}(\mathfrak{p})$ and the fact that $\Gamma^{t-1}_{\mathbf{v}}(d)=1$. Finally, by Claim 2, we get that

$$\Gamma^{t-1}_{\mathbf{v}}(s)=\sum_{\mathfrak{p}'\in\mathscr{P}_{\mathbf{v}}}\prod_{e\in\mathfrak{p}}\exp(\widehat{x}^{t-1}(e)\cdot v_\ell+\widehat{y}^{t-1}(e))=\sum_{\mathfrak{p}'\in\mathscr{P}_{\mathbf{v}}}\exp(\langle\widehat{\mathbf{x}}^{t-1}(\mathfrak{p}'),\mathbf{v}\rangle+\widehat{y}^{t-1}(\mathfrak{p}'))\,.$$

Thus,

$$\prod_{e\in\mathfrak{p}}\phi^t_{\mathbf{v}}(e)=\frac{\exp(\langle\widehat{\mathbf{x}}^{t-1}(\mathfrak{p}),\mathbf{v}\rangle+\widehat{y}^{t-1}(\mathfrak{p}))}{\sum_{\mathfrak{p}'\in\mathscr{P}_{\mathbf{v}}}\exp(\langle\widehat{\mathbf{x}}^{t-1}(\mathfrak{p}'),\mathbf{v}\rangle+\widehat{y}^{t-1}(\mathfrak{p}'))},$$

which was the desired result.

Space and Time Complexity. Constructing the context-dependent DAG $\mathcal{G}(\mathsf{N}_{\mathbf{v}},\mathsf{E}_{\mathbf{v}},\omega_{\mathbf{v}})$ for the realized context and performing the updates in Eqs. (47) and (48) require $O(|\mathsf{E}_{\mathbf{v}}|)$ time and space. Sampling a path takes $O(M/\epsilon)$ time. Updating the shared coefficients via Eqs. (45) and (46) requires $O(|\overline{\mathsf{E}}|)$ time and space, where $|\overline{\mathsf{E}}|$ denotes the number of edges in the super DAG. Since $|\mathsf{E}_{\mathbf{v}}|\le|\overline{\mathsf{E}}|$ for all $\mathbf{v}\in\mathcal{V}$ (and in fact the two are of the same order), the overall per-round complexity of Algorithm 5 is $O(|\overline{\mathsf{E}}|)$ in both time and space. *In particular, the implementation is independent of $|\mathcal{V}|$, and thus applies even when the context space is infinite.*