Efficient Contextual Bandits with Uninformed Feedback Graphs

Mengxiao Zhang [1] **Yuheng Zhang** [2] **Haipeng Luo** [1] **Paul Mineiro** [3]

Abstract

Bandits with feedback graphs are powerful online learning models that interpolate between the full information and classic bandit problems, capturing many real-life applications. A recent work by Zhang et al. (2023) studies the contextual version of this problem and proposes an efficient and optimal algorithm via a reduction to online regression. However, their algorithm crucially relies on seeing the feedback graph before making each decision, while in many applications, the feedback graph is *uninformed*, meaning that it is either only revealed after the learner makes her decision or even never fully revealed at all. This work develops the first contextual algorithms for such uninformed settings, via an efficient reduction to online regression over both the losses and the graphs. Importantly, we show that it is critical to learn the graphs using *log loss* instead of squared loss to obtain favorable regret guarantees. We also demonstrate the empirical effectiveness of our algorithm on a bidding application using both synthetic and real-world data.

1. Introduction

In this paper, we consider efficient algorithm design for contextual bandits with directed feedback graphs, which generalizes classic contextual bandits (Auer et al., 2002; Langford & Zhang, 2007). The interaction between the learner and the environment lasts for T rounds. At each round, the learner first observes a context and then chooses one of K actions, while simultaneously an adversary decides the loss for each action and a *directed feedback graph* with the K actions as nodes. After that, the learner suffers the loss of the chosen action, and her observation is determined based on this feedback graph. Specifically, she observes the loss of every action to which the chosen action is connected.

[1]University of Southern California [2]University of Illinois Urbana-Champaign [3]Microsoft Research. Correspondence to: Mengxiao Zhang <mengxiao.zhang@usc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

This problem reduces to the classic contextual bandits when the feedback graph only contains self-loops, and generally captures more real-world applications, such as personalized web advertising (Mannor & Shamir, 2011):

The non-contextual version of this problem has been studied extensively in the literature (Mannor & Shamir, 2011; Alon et al., 2015a; 2017). Alon et al. (2015a) provided a full characterization on the minimax regret rate with respect to different graph theoretic quantities according to different types of the feedback graphs. However, the more practically useful contextual version is much less explored. A recent work by Zhang et al. (2023) first considered this problem with adversarial context and general feedback graphs, and proposed an efficient algorithm achieving minimax regret rates. However, their algorithm only works in the *informed* setting where the feedback graph is revealed to the learner before her decision. In many applications, such as online pricing (Cohen et al., 2016), the feedback graph is not available to the learner when (or even after) she makes the decision. Cohen et al. (2016) first considered this more challenging *uninformed* setting without context and derived algorithms achieving near-optimal regret guarantees. However, there is no prior work offering a solution for contextual bandits with uninformed feedback graphs. In this work, we take the first step in this direction and propose the first efficient algorithms achieving strong regret guarantees. Specifically, our contributions are as follows.

Contributions. Our algorithm, SquareCB.UG, is based on the SquareCB.G algorithm of Zhang et al. (2023). Assuming realizability on the loss function and an online square loss regression oracle, Zhang et al. (2023) extended the minimax framework for contextual bandits (Foster & Rakhlin, 2020; Foster & Krishnamurthy, 2021) to contextual bandits under informed feedback graphs. With uninformed graphs, we further assume that they are realizable by another function class $\mathcal{G}$ and propose to learn them simultaneously so that in each round we can plug in the predicted graph into SquareCB.G. While the idea is natural, our analysis is highly non-trivial and, perhaps more importantly, reveals that it is crucial to learn these graphs using *log loss* instead of squared loss.

More specifically, within the uninformed setting, we analyze two different types of feedback on the graph structure. In the

partially revealed graph setting, the learner only observes which actions are connected to the selected action, and our algorithm achieves $\widetilde{\mathcal{O}}(\sqrt{\alpha(\mathcal{G})T})$ regret (ignoring the regression overhead), where $\alpha(\mathcal{G})$ is the maximum expected independence number over all graphs in $\mathcal{G}$; in the easier *fully revealed graph setting*, the learner observes the entire graph after her decision, and our algorithm achieves an improved $\widetilde{\mathcal{O}}\left(\sqrt{\sum_{t=1}^{T}\alpha_t}\right)$ regret bound, where α_t denotes the expected independence number of the feedback graph at round t.[1] We note that this latter bound even matches the optimal regret for the easier informed setting (Zhang et al., 2023).

In addition to these strong theoretical guarantees, we also empirically test our algorithm on a bidding application with both synthetic and real-world data and show that it indeed outperforms the greedy algorithm or algorithms that ignore the additional feedback from the graphs.

Related works. (Non-contextual) multi-armed bandits with feedback graphs was first studied by Mannor & Shamir (2011). Alon et al. (2015a) characterized the minimax rates in terms of graph-theoretic quantities under deterministic feedback graphs and proposed algorithms achieving near-optimal guarantees. Since then, many different extensions have been studied, including stochastic feedback graphs (Kocák et al., 2016; Liu et al., 2018; Li et al., 2020; Esposito et al., 2022), uninformed feedback graphs (Cohen et al., 2016; Esposito et al., 2022), algorithms that adapt to both adversarial and stochastic losses (Erez & Koren, 2021; Ito et al., 2022; Dann et al., 2023), data-dependent regret bounds (Lykouris et al., 2018; Lee et al., 2020), and high-probability regret (Neu, 2015; Luo et al., 2023).

The contextual version of this problem has only been studied very recently. Wang et al. (2021) developed algorithms for adversarial linear bandits with uninformed graphs and stochastic contexts, but assumed several strong assumptions on both the policy class and the context space. Zhang et al. (2023) is the closest to our work. They also considered adversarial contexts and realizable losses, but as mentioned, their algorithm is only applicable to the informed setting. Moreover, their theoretical results are also restricted to deterministic feedback graphs only. Our work generalizes theirs from the informed setting to the uninformed setting and from deterministic graphs to stochastic graphs.

Our work is also closely related to the recent trend of designing efficient algorithms for contextual bandits. Since Langford & Zhang (2007) initiated the study of efficient learning in contextual bandits, many follow-ups develop efficient contextual bandits algorithms via reduction to either cost-sensitive classification (Dudik et al., 2011; Agarwal et al., 2014) or online/offline regression (Foster & Rakhlin, 2020; Foster & Krishnamurthy, 2021; Foster et al., 2021; Xu & Zeevi, 2020; Simchi-Levi & Xu, 2022). We follow the latter approach and reduce our problem to online regression on both the losses and the feedback graphs.

2. Preliminary

Throughout the paper, we denote the set $\{1, 2, \ldots, m\}$ for some positive integer m by $[m]$, the set of distributions over some set S by $\Delta(S)$, and the convex hull of some set S by $\mathrm{conv}(S)$. For a vector $v \in \mathbb{R}^m$ and a matrix $M \in \mathbb{R}^{m\times m}$, v_i denotes the i-th coordinate of v and $M_{i,j}$ denotes the (i, j)'s entry of M for $i, j \in [m]$.

The contextual bandits problem with uninformed feedback graphs proceeds in T rounds. At each round t, the environment (possibly randomly and adaptively) selects a context x_t from some arbitrary context space $\mathcal{X}$, a loss vector $\ell_t \in [0, 1]^K$ specifying the loss of each of the K possible actions, and finally a directed feedback graph $G_t = ([K], E_t)$ where $E_t \subseteq [K] \times [K]$ denotes the set of directed edges. The learner then observes the context x_t (but not ℓ_t or G_t) and has to select an action $i_t \in [K]$. At the end of this round, the learner suffers loss ℓ_{t,i_t} and observes the loss of every action connected to i_t (not necessarily including i_t itself): $\ell_{t,j}$ for all $j \in A_t$, where $A_t = \{j \in [K], (i_t, j) \in E_t\}$. In the *partially revealed graph* setting, the learner does not observe anything else about the graph (other than A_t), while in the *fully revealed graph* setting, the learner additionally observes the entire graph (that is, E_t).

Alon et al. (2015a) showed that in the non-contextual version of this problem, there are essentially only two types of nontrivial and learnable feedback graphs: strongly observable graphs and weakly observable graphs. For simplicity, our work focuses solely on the first type, that is, we assume that G_t is always strongly observable, meaning that for each node $i \in [K]$, either it can observe itself ($(i, i) \in E_t$) or it can at least be observed by any other nodes ($(j, i) \in E_t$ for any $j \in [K]\backslash\{i\}$). As mentioned in Footnote 1, our results can be directly generalized to weakly observable graphs as well; see Appendix D.

Bandits with uninformed feedback graphs naturally capture many applications such as online pricing, viral marketing, and recommendation in social networks (Kocák et al., 2014; Alon et al., 2015a; 2017; Rangi & Franceschetti, 2019; Cohen et al., 2016; Liu et al., 2018). By incorporating contexts, which are broadly available in practice, our model significantly increases its applicability in real world. For a concrete

[1] Definitions of all graph-theoretic numbers are formally introduced in Section 2. Also, for simplicity this paper focuses on strongly observable graphs where $\sqrt{T}$-regret is achievable, but our ideas can be directly generalized to weakly observable graphs as well (where $T^{2/3}$-regret is minimax optimal). See Appendix D for a detailed discussion.

example, see Section 5 for an application of bidding in a first-price auction.

Realizability and oracle assumptions. Following a line of recent works on developing efficient contextual bandit algorithms, we make the following standard realizability assumption on the loss function, stating that the expected loss of each action can be perfectly predicted by an unknown loss predictor from a known class:

Assumption 2.1 (Realizability of mean loss)**.** We assume that the learner has access to a function class $\mathcal{F} \subseteq (\mathcal{X} \times [K] \mapsto [0,1])$ in which there exists an unknown regression function $f^\star \in \mathcal{F}$ such that for any $i \in [K]$ and $t \in [T]$, we have $\mathbb{E}[\ell_{t,i} \mid x_t] = f^\star(x_t, i)$.

The goal of the learner is naturally to be comparable to an oracle strategy that knows $f^\star$ ahead of time, formally measured by the (expected) regret:

$$\mathbf{Reg}_{\mathsf{CB}} \triangleq \mathbb{E}\left[\sum_{t=1}^{T}(f^\star(x_t, i_t) - \min_{i\in[k]} f^\star(x_t, i))\right].$$

To efficiently minimize regret for a general class $\mathcal{F}$, it is important to assume some oracle access to this class. To this end, we follow prior works and assume that the learner is given an *online regression oracle* AlgSq for function class $\mathcal{F}$, which follows the following protocol: at each round $t \in [T]$, the oracle AlgSq produces an estimator $f_t \in \mathrm{conv}(\mathcal{F})$, then receives a context x_t and a set S_t of action-loss pairs in the form $(a, c) \in [K] \times [0, 1]$. The squared loss of the oracle for this round is $\sum_{(a,c)\in S_t}(f_t(x_t, a) - c)^2$, which is on average assumed to be close to that of the best predictor in $\mathcal{F}$:

Assumption 2.2 (Bounded squared loss regret)**.** The regression oracle AlgSq guarantees:

$$\sum_{t=1}^{T}\sum_{(a,c)\in S_t}(f_t(x_t, a) - c)^2 - \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\sum_{(a,c)\in S_t}(f(x_t, a) - c)^2 \le \mathbf{Reg}_{\mathsf{Sq}}.$$

Here, $\mathbf{Reg}_{\mathsf{Sq}}$ is a regret bound that is sublinear in T and depends on some complexity measure of $\mathcal{F}$; see e.g. Foster & Rakhlin (2020) for concrete examples of such oracles and the corresponding regret bounds. The point is that online regression is such a standard machine learning practice, so reducing our problem to online regression is both theoretically reasonable and practically desirable.

So far, we have made exactly the same assumptions as Zhang et al. (2023) which studies the informed setting. In our uninformed setting, however, since nothing is known about the feedback graph before deciding which action to take, we propose to additionally learn the feedback graphs, which requires the following extra realizability and oracle assumptions related to the graphs.

Assumption 2.3 (Realizability of mean graph)**.** We assume that the learner has access to a function class $\mathcal{G} \subseteq (\mathcal{X} \times [K] \times [K] \mapsto [0,1])$ in which there exists a regression function $g^\star \in \mathcal{G}$ such that for any $(i, j) \in [K] \times [K]$ and $t \in [T]$, we have $\mathbb{E}[\mathbb{1}\{(i, j) \in E_t\} \mid x_t] = g^\star(x_t, i, j)$.

Similarly, since we do not impose specific structures on $\mathcal{G}$, we assume that the learner can access $\mathcal{G}$ through the use of another online oracle AlgLog: at each round $t \in [T]$, AlgLog produces an estimator $g_t \in \mathrm{conv}(\mathcal{G})$, then receives a context x_t and a set S_t of tuples in the form $(i, j, b) \in [K] \times [K] \times \{0, 1\}$ where $b = 1$ ($b = 0$) means i is (is not) connected to j. Importantly, our analysis shows that it is critical for this oracle to learn the graphs using *log loss* instead of squared loss (hence the name AlgLog); see detailed explanations in Section 4.1. More specifically, we assume that the oracle satisfies the following regret bound measured by log loss:

Assumption 2.4 (Bounded log loss regret)**.** The regression oracle AlgLog guarantees:

$$\sum_{t=1}^{T}\sum_{(i,j,b)\in S_t}\ell_{\log}(g_t(x_t, i, j), b) - \inf_{g\in\mathcal{G}}\sum_{t=1}^{T}\sum_{(i,j,b)\in S_t}\ell_{\log}(g(x_t, i, j), b) \le \mathbf{Reg}_{\mathsf{Log}},$$

where for two scalars $u, v \in [0, 1]$, $\ell_{\log}(u, v)$ is defined as

$$\ell_{\log}(u, v) = v \log\frac{1}{u} + (1 - v)\log\frac{1}{1 - u}.$$

Once again, the bound $\mathbf{Reg}_{\mathsf{Log}}$ is sublinear in T and depends on some complexity measure of $\mathcal{G}$. We note that regression using log loss is also highly standard in practice. For concrete examples, we refer the readers to Foster & Krishnamurthy (2021) where the same log loss oracle was used (for a different purpose of obtaining first-order regret guarantees for contextual bandits). In our analysis, we also make use of the following important technical lemma that connects the log loss regret with something called the *triangular discrimination* under the realizability assumption.

Lemma 2.5 (Proposition 5 of Foster & Krishnamurthy (2021))**.** *Suppose for each t and $(i, j, b) \in S_t$, we have $\mathbb{E}[b|x_t] = g^\star(x_t, i, j)$. Then oracle* AlgLog *guarantees:*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{(i,j,b)\in S_t}\frac{(g_t(x_t, i, j) - g^\star(x_t, i, j))^2}{g_t(x_t, i, j) + g^\star(x_t, i, j)}\right] \le 2\mathbf{Reg}_{\mathsf{Log}}.$$

Algorithm 1 SquareCB.UG

Input: parameter $\gamma \geq 0$, a regression oracle AlgSq for loss prediction, and a regression oracle AlgLog for graph prediction
for $t = 1, 2, \ldots, T$ **do**
Receive context x_t.
Obtain a loss estimator f_t from the oracle AlgSq and a graph estimator g_t from AlgLog.
Compute $p_t = \operatorname{argmin}_{p \in \Delta([K])} \mathsf{dec}_\gamma(p; f_t, g_t, x_t)$ (a simple convex program; see Appendix C), where

$$\mathsf{dec}_\gamma(p; f, g, x) = \sup_{\substack{i^\star \in [K] \\ v^\star \in [0,1]^K}} \left[\sum_{i=1}^K p_i v_i^\star - v_{i^\star}^\star - \frac{\gamma}{4} \sum_{i=1}^K p_i \sum_{j=1}^K g(x,i,j) \left(f(x,j) - v_j^\star \right)^2 \right]. \tag{1}$$

The environment decides loss ℓ_t and feedback graph $G_t = ([K], E_t)$.
The learner samples i_t from p_t and observe $\{\ell_{t,j}\}_{j \in A_t}$ where $A_t = \{j \in [K] : (i_t, j) \in E_t\}$.
Feed x_t and $\{(j, \ell_{t,j})\}_{j \in A_t}$ to the oracle AlgSq.
Case 1 (Partially revealed graph): Construct $S_t = \{(i_t, j, 1)\}_{j \in A_t} \cup \{(i_t, j, 0)\}_{j \notin A_t}$.
Case 2 (Fully revealed graph): Observe E_t and construct $S_t = \{(i, j, 1)\}_{(i,j) \in E_t} \cup \{(i, j, 0)\}_{(i,j) \notin E_t}$.
Feed x_t and S_t to the oracle AlgLog.
end

Independence number. It is known that for strongly observable graphs, their *independence numbers* characterize the minimax regret (Alon et al., 2015a). Specifically, an independence set of a directed graph is a subset of nodes in which no two distinct nodes are connected. The size of the largest independence set in graph G is called its independence number, denoted by $\alpha(G)$. Since we consider stochastic graphs, we further define the independence number with respect to a $g \in \operatorname{conv}(\mathcal{G})$ and a context x as:

$$\alpha(g, x) \triangleq \inf_{q \in \mathcal{Q}(g,x)} \mathbb{E}_{G \sim q}[\alpha(G)],$$

where $\mathcal{Q}(g, x)$ denotes the set of all distributions of strongly observable graphs whose expected edge connections are specified by $g(x, \cdot, \cdot)$ (that is, for any $q \in \mathcal{Q}(g, x)$, we have $\mathbb{E}_{([K],E) \sim q}[\mathbb{1}\{(i,j) \in E\}] = g(x, i, j)$ for all (i, j)). With this notion, the difficulty of G_t is then characterized by the independence number $\alpha_t \triangleq \alpha(g^\star, x_t)$. In the more challenging partially revealed graph setting, however, our result depends on the worst-case independence number over the entire class $\mathcal{G}$: $\alpha(\mathcal{G}) \triangleq \sup_{g \in \mathcal{G}, x \in \mathcal{X}} \alpha(g, x)$.

3. Algorithms and Regret Guarantees

In this section, we introduce our algorithm SquareCB.UG and its regret guarantees. To describe our algorithm, we first briefly introduce the SquareCB.G algorithm of Zhang et al. (2023) for the informed setting: at each round t, given the loss estimator f_t obtained from the regression oracle AlgSq and the feedback graph G_t, SquareCB.G finds the action distribution $p_t \in \Delta([K])$ by solving $\operatorname{argmin}_p \mathsf{dec}_\gamma(p; f_t, G_t, x_t)$, where the *Decision-Estimation Coefficient* (DEC) is defined as:

$$\mathsf{dec}_\gamma(p; f_t, G_t, x_t) \triangleq \sup_{\substack{i^\star \in [K] \\ v^\star \in [0,1]^K}} \left[\sum_{i=1}^K p_i v_i^\star - v_{i^\star}^\star - \frac{\gamma}{4} \sum_{i=1}^K p_i \sum_{j=1}^K G_{t,i,j} \left(f_t(x_t, j) - v_j^\star \right)^2 \right] \tag{2}$$

for some parameter $\gamma > 0$, and we abuse the notation by letting G_t also represent its adjacent matrix. The idea of DEC originates from Foster & Rakhlin (2020) for contextual bandits and has become a general way to tackle interactive decision making problems since then (Foster et al., 2021). The first two terms within the supremum of Eq. (2) is the instantaneous regret of strategy p against the best action $i^\star$ with respect to a loss vector $v^\star$, and the third term corresponds to the expected squared loss between the loss predictor $f_t(x_t, \cdot)$ and the loss vector $v^\star$ on the observed actions (since each action i is selected with probability p_i and, conditioning on i being selected, each action j is observed with probability $G_{t,i,j}$). Because the true loss vector is unknown, a supremum over $v^\star$ is taken (that is, the worst case is considered). The goal of the learner is to pick p_t to minimize this DEC, since a small DEC means that the regret suffered by the learner is close to the regret of the regression oracle (which is assumed to be bounded). After selecting an action $i_t \sim p_t$ and seeing the loss of actions connected i_t, SquareCB.G naturally feeds these observations to AlgSq and proceeds to the next round.

The clear obstacle of running SquareCB.G in the uninformed setting is that G_t, required in Eq. (2), is unavailable at the beginning of round t. Therefore, we propose to learn the graphs simultaneously and simply use a predicted graph in place of the true graph G_t. More concretely, at the be-

ginning of round t, in addition to the loss estimator f_t, we also obtain a graph estimator g_t from the graph regression oracle AlgLog. Then, to find p_t, we solve the same problem but with G_t in Eq. (2) replaced by the estimator $g_t(x_t, \cdot, \cdot)$; see Eq. (1). After picking action $i_t \sim p_t$, the training of AlgSq remains the same, and additionally, we feed all the observations about the structure of G_t to AlgLog: for the partially revealed graph setting, the observations are the connections between i_t and all $j \in A_t$ and the disconnections between i_t and all $j \notin A_t$; while for the fully revealed graph setting, the observations are all the connections in E_t and the disconnections for all other action pairs. See Algorithm 1 for the complete pseudocode.

While the idea of our algorithm appears to be very natural, its analysis is in fact highly non-trivial and reveals that learning the graphs using log loss is critical; see Section 4 for more discussions. We also note that finding the minimizer of the DEC can be written as a simple convex program and solved efficiently; see more implementation details in Appendix C.

We prove the following regret guarantee of SquareCB.UG in the partially revealed graph setting.

Theorem 3.1. *Under Assumptions 2.1-2.4 and partially revealed feedback graphs,* SquareCB.UG *with* $\gamma = \max\left\{4, \sqrt{\frac{\alpha(\mathcal{G})T\log(KT)}{\max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\}}}\right\}$ *guarantees:*[2]

$$\mathbf{Reg}_{\mathsf{CB}} = \widetilde{\mathcal{O}}\left(\sqrt{\alpha(\mathcal{G})T\max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\}}\right).$$

Since $\mathbf{Reg}_{\mathsf{Sq}}$ and $\mathbf{Reg}_{\mathsf{Log}}$ are both sublinear in T, this regret bound is also sublinear in T. More importantly, it has no polynomial dependence on the total number of actions K, and instead only depends on the worst-case independence number $\alpha(\mathcal{G}) \le K$. There are indeed important applications where the independence number of every encountered feedback graph must be small or even independent of K (e.g., the inventory control problem discussed in Zhang et al. (2023) or the bidding application in Section 5), in which case it only makes sense to pick $\mathcal{G}$ such that $\alpha(\mathcal{G})$ is also small. While SquareCB.UG requires setting γ with the knowledge of $\alpha(\mathcal{G})$, in Appendix A.2, we show that applying certain doubling trick on the DEC value leads to the same regret bound even *without* the knowledge of $\alpha(\mathcal{G})$.

However, it would be even better if instead of paying $\alpha(\mathcal{G})$ every round, we only pay the independence number of the corresponding stochastic feedback graph at each round t, that is, α_t. While it is unclear to us whether this is achievable with partially revealed graphs, in the next theorem, we show that SquareCB.UG indeed achieves this in the easier fully revealed graph setting.

[2]The notation $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic dependence on K and T.

Theorem 3.2. *Under Assumptions 2.1-2.4 and fully revealed feedback graphs,* SquareCB.UG *with* $\gamma = \max\left\{12, \sqrt{\frac{\sum_{t=1}^T \alpha_t}{\max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\}}}\right\}$ *guarantees:*

$$\mathbf{Reg}_{\mathsf{CB}} = \widetilde{\mathcal{O}}\left(\sqrt{\sum_{t=1}^T \alpha_t \max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\}}\right).$$

In other words, we replace the $\alpha(\mathcal{G})T$ term in the regret with the smaller and more adaptive quantity $\sum_{t=1}^T \alpha_t$, indicating that the complexity of learning only depends on how difficult each encountered graph is, but not the worst case difficulty among all the possible graphs in $\mathcal{G}$.

4. Analysis

In this section, we provide some key steps of our analysis, highlighting 1) why it is enough to replace the true graph in Eq. (2) with the graph estimator g_t; 2) why using log loss in the graph regression oracle is important; and 3) why having fully revealed graphs helps improve the dependence from $\alpha(\mathcal{G})$ to α_t.

4.1. Analysis for Partially Revealed Graphs

While we present the DEC in Eq. (1) as a natural modification of Eq. (2) in the absence of the true graph, it in fact can be rigorously derived as an upper bound on another DEC more tailored to our original problem. Specifically, we define for two parameters $\gamma_1, \gamma_2 > 0$:

$$\begin{aligned}\mathsf{dec}^{\mathsf{P}}_{\gamma_1,\gamma_2}(p; f, g, x) \triangleq \sup_{\substack{i^\star \in [K], v^\star \in [0,1]^K \\ M^\star \in [0,1]^{K\times K}}} \Bigg[&\sum_{i=1}^K p_i v_i^\star - v_{i^\star}^\star \\ &- \gamma_1 \sum_{i=1}^K p_i \sum_{j=1}^K M^\star_{i,j}\left(f(x,j) - v_j^\star\right)^2 \\ &- \gamma_2 \sum_{i=1}^K p_i \sum_{j=1}^K \frac{(M^\star_{i,j} - g(x,i,j))^2}{M^\star_{i,j} + g(x,i,j)} \Bigg]. \end{aligned} \tag{3}$$

Similar to Eq. (2), the first two terms within the supermum represent the instantaneous regret of strategy p against the best action $i^\star$ with respect to a loss vector $v^\star$. The third term is also similar and represents the squared loss of AlgSq under strategy p, but since the true graph G_t is unknown, it is replaced with the worst-case adjacent matrix $M^\star$ (hence the supremum over $M^\star$). Finally, the last additional term is the expected triangular discrimination between $M^\star_{i,\cdot}$ and $g(x,i,\cdot)$ when i is sampled from p, which, according to Lemma 2.5, represents the log loss regret of AlgLog.

Once again, the idea is that if for every round t, we can find a strategy p_t with a small DEC value $\mathsf{dec}^{\mathsf{P}}_{\gamma_1,\gamma_2}(p_t; f_t, g_t, x_t)$,

then the learner's overall regret $\mathbf{Reg}_{\mathsf{CB}}$ will be close to the square loss regret $\mathbf{Reg}_{\mathsf{Sq}}$ of AlgSq plus the log loss regret $\mathbf{Reg}_{\mathsf{Log}}$ of AlgLog, both of which are assumed to be reasonably small. This is formally stated below.

Theorem 4.1. *Under Assumptions 2.1-2.4, for any $\gamma_1, \gamma_2 \geq 0$, the regret $\mathbf{Reg}_{\mathsf{CB}}$ of SquareCB.UG is at most*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathsf{dec}^{\mathsf{P}}_{\gamma_1,\gamma_2}(p_t; f_t, g_t, x_t)\right] + \gamma_1 \mathbf{Reg}_{\mathsf{Sq}} + 2\gamma_2 \mathbf{Reg}_{\mathsf{Log}}.$$

Now, instead of directly minimizing the DEC to find the strategy p_t (which is analytically complicated), the following lemma shows that the easier form of Eq. (1) serves as an upper bound of Eq. (3), further explaining our algorithm design.

Lemma 4.2. *For any $p \in \Delta([K])$, $g \in \mathrm{conv}(\mathcal{G})$, $f \in \mathrm{conv}(\mathcal{F})$, and $x \in \mathcal{X}$, we have*

$$\mathsf{dec}^{\mathsf{P}}_{\frac{3}{4}\gamma,\frac{1}{4}\gamma}(p; f, g, x) \leq \mathsf{dec}_{\gamma}(p; f, g, x).$$

To prove this lemma, we need to connect the squared loss with respect to $M^\star$ and that with respect to g. We achieve so using the following lemma, which also reveals why triangular discrimination naturally comes out.

Lemma 4.3. *For any $z, z' > 0$, the following holds:*

$$3z' \geq z - \frac{(z-z')^2}{z+z'}.$$

Proof. By AM-GM inequality, we have

$$2(z-z') \leq (z+z') + \frac{(z-z')^2}{z+z'}.$$

Rearranging then finishes the proof. □

Proof of Lemma 4.2. Using Lemma 4.3 with $z' = M^\star_{i,j}\left(f(x,j) - v^\star_j\right)^2$ and $z = g(x,i,j)\left(f(x,j) - v^\star_j\right)^2$, we have

$$\begin{aligned} 3M^\star_{i,j}\left(f(x,j) - v^\star_j\right)^2 &\geq g(x,i,j)\left(f(x,j) - v^\star_j\right)^2 \\ &\quad - \frac{(M^\star_{i,j} - g(x,i,j))^2}{M^\star_{i,j} + g(x,i,j)}\left(f(x,j) - v^\star_j\right)^2 \\ &\geq g(x,i,j)\left(f(x,j) - v^\star_j\right)^2 - \frac{(M^\star_{i,j} - g(x,i,j))^2}{M^\star_{i,j} + g(x,i,j)}, \end{aligned}$$

where the last step uses the fact $\left(f(x,j) - v^\star_j\right)^2 \leq 1$. Applying this inequality to the definition of Eq. (3) and plugging in the parameters $\gamma_1 = \frac{3}{4}\gamma$ and $\gamma_2 = \frac{1}{4}\gamma$, we see that the two triangular discrimination terms cancel, leading us to exactly Eq. (1). □

The importance of using log loss when learning the feedback graphs now becomes clear: the log loss regret turns out to be exactly the price one needs to pay by pretending that the graph estimator is the true graph.

The last step of the analysis is to show that the minimum DEC value, which our final strategy p_t achieves, is reasonably small and related to some independence number with no polynomial dependence on the total number of actions:

Lemma 4.4. *For any $g \in \mathrm{conv}(\mathcal{G})$, $f \in \mathrm{conv}(\mathcal{F})$, $x \in \mathcal{X}$, and $\gamma \geq 4$, we have*

$$\min_{p \in \Delta([K])} \mathsf{dec}_{\gamma}(p; f, g, x) = \mathcal{O}\left(\frac{\alpha(g,x)\log(K\gamma)}{\gamma}\right).$$

The proof of this lemma is deferred to Appendix A and is a refinement and generalization of Zhang et al. (2023, Theorem 3.2) which only concerns deterministic graphs. Combining everything, we are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. Setting $\gamma_1 = \frac{3}{4}\gamma$ and $\gamma_2 = \frac{1}{4}\gamma$ in Theorem 4.1 and combining it with Lemma 4.2, we know that $\mathbf{Reg}_{\mathsf{CB}}$ is at most

$$\sum_{t=1}^{T} \mathsf{dec}_{\gamma}(p_t; f_t, g_t, x_t) + \frac{3}{4}\gamma \mathbf{Reg}_{\mathsf{Sq}} + \frac{1}{2}\gamma \mathbf{Reg}_{\mathsf{Log}}. \quad (4)$$

Since p_t is chosen by minimizing $\mathsf{dec}_{\gamma}(p; f_t, g_t, x_t)$, we further apply Lemma 4.4 to bound the regret by

$$\mathcal{O}\left(\sum_{t=1}^{T} \frac{\alpha(g_t, x_t)\log(K\gamma)}{\gamma}\right) + \frac{3}{4}\gamma \mathbf{Reg}_{\mathsf{Sq}} + \frac{1}{2}\gamma \mathbf{Reg}_{\mathsf{Log}}.$$

Finally, realizing $\alpha(g_t, x_t) \leq \alpha(\mathcal{G})$ (see Lemma A.1) and plugging in the choice of γ finishes the proof. □

4.2. Analysis for Fully Revealed Graphs

From the analysis for the partially revealed graph setting, we see that the $\alpha(\mathcal{G})$ dependence in fact comes from $\alpha(g_t, x_t)$, the independence number with respect to the graph estimator g_t. To improve it to $\alpha_t = \alpha(g^\star, x_t)$, we again need to connect two different graphs in the DEC definition using Lemma 4.3, as shown below.

Lemma 4.5. *For $g, g' \in \mathrm{conv}(\mathcal{G})$, $f \in \mathrm{conv}(\mathcal{F})$, and $x \in \mathcal{X}$, we have*

$$\begin{aligned} \min_{p \in \Delta([K])} \mathsf{dec}_{\gamma}(p; f, g', x) &\leq \min_{p \in \Delta([K])} \mathsf{dec}_{\frac{\gamma}{3}}(p; f, g, x) \\ &\quad + \frac{\gamma}{12}\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{(g(x,i,j) - g'(x,i,j))^2}{g(x,i,j) + g'(x,i,j)}. \end{aligned}$$

Proof sketch. For a fix p, similarly to the proof of Lemma 4.2, we apply Lemma 4.3 with $z' =$

$g'(x,i,j)(f(x,j)-v_j^\star)^2$ and $z = g(x,i,j)(f(x,j)-v_j^\star)^2$ for each i and j and arrive at

$$\mathsf{dec}_\gamma(p; f, g', x) \le \mathsf{dec}_{\frac{\gamma}{3}}(p; f, g, x) + \frac{\gamma}{12}\sum_{i=1}^K p_i \sum_{j=1}^K \frac{(g(x,i,j)-g'(x,i,j))^2}{g(x,i,j)+g'(x,i,j)}.$$

Further upper bounding p_i by 1 in the last term and then taking min over p on both sides finishes the proof. □

This lemma allows us to connect the minimum DEC value with respect to g_t and that with respect to $g^\star$, but with the price of $\frac{\gamma}{12}\sum_{i=1}^K\sum_{j=1}^K \frac{(g_t(x_t,i,j)-g^\star(x_t,i,j))^2}{g_t(x_t,i,j)+g^\star(x_t,i,j)}$, which, under fully revealed graphs, is essentially the per-round log loss regret in light of Lemma 2.5 since the oracle AlgLog indeed receives observations for all (i,j) pairs (importantly, this does not hold for partially revealed graphs). With this insight, we are ready to prove the main theorem.

Proof of Theorem 3.2. First note that Theorem 4.1 in fact also holds for fully revealed graphs; this is intuitively true simply because the fully revealed case is easier than the partially revealed case, and is formally explained in the proof of Theorem 4.1. Therefore, combining it with Lemma 4.2, we still have $\mathbf{Reg}_{\mathsf{CB}}$ bounded by

$$\sum_{t=1}^T \min_p \mathsf{dec}_\gamma(p; f_t, g_t, x_t) + \frac{3}{4}\gamma\mathbf{Reg}_{\mathsf{Sq}} + \frac{1}{2}\gamma\mathbf{Reg}_{\mathsf{Log}}.$$

By Lemma 4.5, this is at most

$$\sum_{t=1}^T \min_p \mathsf{dec}_{\frac{\gamma}{3}}(p; f_t, g^\star, x_t) + \frac{3}{4}\gamma\mathbf{Reg}_{\mathsf{Sq}} + \frac{1}{2}\gamma\mathbf{Reg}_{\mathsf{Log}} + \frac{\gamma}{12}\sum_{t=1}^T\sum_{i=1}^K\sum_{j=1}^K \frac{(g_t(x_t,i,j)-g^\star(x_t,i,j))^2}{g_t(x_t,i,j)+g^\star(x_t,i,j)}.$$

Finally, using Lemma 4.4 and Lemma 2.5, the above is further bounded by

$$\mathcal{O}\left(\sum_{t=1}^T \frac{\alpha_t \log(K\gamma)}{\gamma}\right) + \frac{3}{4}\gamma\mathbf{Reg}_{\mathsf{Sq}} + \frac{2}{3}\gamma\mathbf{Reg}_{\mathsf{Log}},$$

which completes the proof with our choice of γ. □

5. Experiments

In this section, we show empirical results of our SquareCB.UG algorithm by testing it on a bidding application. We start by describing this application, followed by modelling it as an instance of our partially revealed graph setting. Specifically, consider a bidder (the learner) participating in a first-price auction. At each round t, the bidder observes some context x_t, while the environment decides a competing price $w_t \in [0,1]$ (the highest price of all other bidders) and the value of the learner $v_t \in [0,1]$ for the current item (unknown to the learner herself). Then, the learner decides her bid $c_t \in [0,1]$. If $c_t \ge w_t$, the learner wins the auction, pays c_t to the auctioneer (first-price), and observes her reward $v_t - c_t$; otherwise, the learner loses the auction without observing her value v_t, and her reward is 0. In either case, at the end of this round, the auctioneer announces the winning bid to all bidders. For the learner, this information is only meaningful when she loses the auction, in which case w_t (the winning bid) is revealed to her.

This problem is a natural instance of our model. Specifically, we let the learner choose her bid c_t from a discretized set $\mathcal{A}_\varepsilon = \{0, \varepsilon, \dots, 1-\varepsilon, 1\}$ of size $K = \frac{1}{\varepsilon}+1$ for some granularity ε. For ease of presentation, the i-th bid $(i-1)\varepsilon$ is denoted by a_i. The feedback graph G_t is completely determined by the competing price w_t in the following way:

$$G_{t,i,j} = \begin{cases} 1, & a_i < w_t \text{ and } a_j < w_t, \\ 1, & a_i \ge w_t \text{ and } j \ge i, \\ 0, & \text{otherwise}, \end{cases}$$

where we again overload the notation G_t to represent its adjacent matrix. This is because when bidding lower than the competing price w_t, the learner observes w_t and knows that bidding anything below w_t gives 0 reward; and when bidding higher than w_t, the bidder only knows that she would still win if she were to bid even higher, and the corresponding reward can be calculated since she knows her value v_t in this case. It is clear that this graph is strongly observable with independence number at most 2 and is only partially revealed at the end of each round if the learner wins (and fully revealed otherwise). On the other hand, the reward of action i is $\mathbb{1}[a_i \ge w_t]\cdot(v_t - a_i)$, which we translate to a loss in $[0,1]$ by shifting and scaling:

$$\ell_{t,i} = \tfrac{1}{2}\left(1 - \mathbb{1}[a_i \ge w_t]\cdot(v_t - a_i)\right) \in [0,1]. \quad (5)$$

Regression oracles. For the graph predictor, since the feedback graph is determined by the competing price w_t, we use a linear classification model to predict the distribution of w_t. Then at each round, we sample a competing price from this distribution, leading to the predicted feedback graph g_t. For the loss predictor f_t, since losses are determined by w_t and v_t, we use a two-layered fully connected neural network to predict the value v_t and construct the loss predictors according to Eq. (5) with w_t and v_t replaced by their predicted values. For more details of the oracles and their training, see Appendix C.

Implementation of SquareCB.UG. While Eq. (1) can be solved by a convex program, in order to implement SquareCB.UG even more efficiently, we use a closed-form

solution of p_t enabled by the specific structure of the predicted graph g_t in this application. See Appendix C for more details.

5.1. Empirical Results on Synthetic Data

Data. We first generate two synthetic datasets $\{(x_t, w_t, v_t)\}_{t=1}^T$ with $T = 5000$ and $x_t \in \mathbb{R}^{32}$ for all $t \in [T]$. The competing price w_t and the value v_t are generated by $w_t = \frac{1}{\sqrt{32}}\theta_1^\top x_t + \varepsilon_t, v_t = w_t + \max\{\frac{40}{\sqrt{32}}\theta_2^\top x_t, 0\}$, where $\theta_1, \theta_2 \in \mathbb{R}^{32}$ are sampled from standard Gaussian and $\varepsilon_t \sim \mathcal{N}(0, 0.05)$ is a small noise. All w_t and v_t's are then normalized to $[0, 1]$. The two datasets only differ in how $\{x_t\}_{t=1}^T$ are generated. Specifically, in the context of linear bandits, Bastani et al. (2021) showed that whether the simple greedy algorithm with no explicit exploration performs well depends largely on the context's diversity, roughly captured by the minimum eigenvalue of its covariance matrix. We thus follow their work and generate two datasets where the first one enjoys good diversity and second one does not; see Appendix C for details.

Results. We compare our SquareCB.UG with SquareCB (Foster & Rakhlin, 2020) (which ignores the additional feedback from graphs), the greedy algorithm (which simply picks the best action according to the loss predictors), and a trivial baseline that always bids 0. For the first three algorithms, we try three different granularity values $\varepsilon \in \{\frac{1}{25}, \frac{1}{50}, \frac{1}{75}\}$ leading to three increasing number of actions, and we run each of them 4 times and plot the averaged normalized regret ($\mathbf{Reg}_{\mathsf{CB}}/T$) together with the standard deviation in Figure 1. We observe that our algorithm performs the best and, unlike SquareCB, its regret almost does not increase when the number of action increases, matching our theoretical guarantee. In addition, consistent with Bastani et al. (2021), while greedy indeed performs quite well when the contexts are diverse (top figure), it performs almost the same as the trivial "not bidding" baseline and suffers linear regret in the absence of diverse contexts (bottom figure).

5.2. Empirical Results on Real Auction Data

Data. We also conduct experiments on a subset of 5000 samples of a real eBay auction dataset used in Mohri & Medina (2016); see Appendix C for details.

Results. We compare the four algorithms in the same way, with the only difference being the discretization granularity value $\varepsilon \in \{\frac{1}{50}, \frac{1}{100}, \frac{1}{150}\}$. The results are shown in Figure 2. Similar to what we observe in synthetic datasets, SquareCB.UG consistently outperforms other algorithms, demonstrating the advantage of exploration with graph information. The greedy algorithm's performance is unstable and has a relatively large variance due to the lack of exploration.

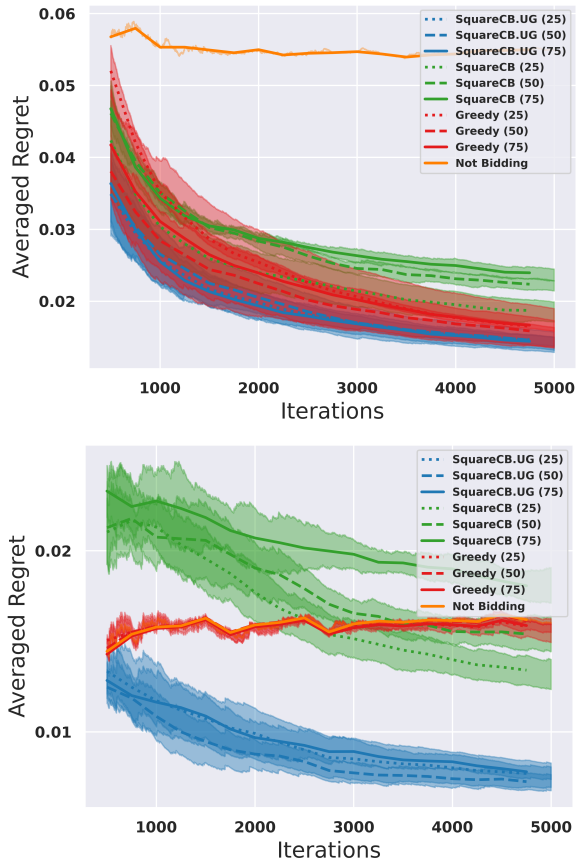

Figure 1. Comparison among SquareCB.UG, SquareCB, greedy, and a trivial baseline on one synthetic dataset with diverse contexts (top figure) and another one with poor diversity (bottom figure).

Figure 2. Comparison among SquareCB.UG, SquareCB, greedy, and a trivial baseline on a real auction dataset.

With all these results for both real and synthetic datasets, we show that our algorithm indeed effectively explores the environment utilizing the uninformed graph structure and is robust to different types of environments.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

HL and MZ are supported by NSF Award IIS-1943607.

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646. PMLR, 2014.

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

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. Online learning with feedback graphs: Beyond bandits. In *arXiv preprint arXiv:1502.07617*, 2015b.

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

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. In *SIAM Journal on computing*, volume 32, pp. 48–77. SIAM, 2002.

Bastani, H., Bayati, M., and Khosravi, K. Mostly exploration-free algorithms for contextual bandits. In *Management Science*, volume 67, pp. 1329–1349. INFORMS, 2021.

Cohen, A., Hazan, T., and Koren, T. Online learning with feedback graphs without the graphs. In *International Conference on Machine Learning*, pp. 811–819. PMLR, 2016.

Dann, C., Wei, C.-Y., and Zimmert, J. A blackbox approach to best of both worlds in bandits and beyond. In *Conference on Learning Theory*, pp. 5503–5570. PMLR, 2023.

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. Efficient optimal learning for contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, pp. 169–178, 2011.

Erez, L. and Koren, T. Towards best-of-all-worlds online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 34, pp. 28511–28521, 2021.

Esposito, E., Fusco, F., van der Hoeven, D., and Cesa-Bianchi, N. Learning on the edge: Online learning with stochastic feedback graphs. In *Advances in Neural Information Processing Systems*, 2022.

Foster, D. and Rakhlin, A. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR, 2020.

Foster, D. J. and Krishnamurthy, A. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18907–18919, 2021.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. In *arXiv preprint arXiv:2112.13487*, 2021.

Ito, S., Tsuchiya, T., and Honda, J. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, 2022.

Kocák, T., Neu, G., Valko, M., and Munos, R. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Kocák, T., Neu, G., and Valko, M. Online learning with erdos-renyi side-observation graphs. In *Conference on Uncertainty in Artificial Intelligence*, 2016.

Langford, J. and Zhang, T. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, volume 20, 2007.

Lee, C.-W., Luo, H., and Zhang, M. A closer look at small-loss bounds for bandits with graph feedback. In *Conference on Learning Theory*, pp. 2516–2564. PMLR, 2020.

Li, S., Chen, W., Wen, Z., and Leung, K.-S. Stochastic online learning with probabilistic graph feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4675–4682, 2020.

Liu, F., Buccapatnam, S., and Shroff, N. Information directed sampling for stochastic bandits with graph feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Luo, H., Tong, H., Zhang, M., and Zhang, Y. Improved high-probability regret for adversarial bandits with time-varying feedback graphs. In *International Conference on Algorithmic Learning Theory*, pp. 1074–1100. PMLR, 2023.

Lykouris, T., Sridharan, K., and Tardos, É. Small-loss bounds for online learning with partial information. In *Conference on Learning Theory*, pp. 979–986. PMLR, 2018.

Mannor, S. and Shamir, O. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Mohri, M. and Medina, A. M. Learning algorithms for second-price auctions with reserve. In *Journal of Machine Learning Research*, volume 17, pp. 2632–2656. JMLR. org, 2016.

Neu, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Rangi, A. and Franceschetti, M. Online learning with feedback graphs and switching costs. In *International Conference on Artificial Intelligence and Statistics*, pp. 2435–2444. PMLR, 2019.

Simchi-Levi, D. and Xu, Y. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. In *Mathematics of Operations Research*, volume 47, pp. 1904–1931. INFORMS, 2022.

Wang, L., Li, B., Zhou, H., Giannakis, G. B., Varshney, L. R., and Zhao, Z. Adversarial linear contextual bandits with graph-structured side observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10156–10164, 2021.

Xu, Y. and Zeevi, A. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. In *arXiv preprint arXiv:2007.07876*, 2020.

Zhang, M., Zhang, Y., Vrousgou, O., Luo, H., and Mineiro, P. Practical contextual bandits with feedback graphs. In *Conference on Neural Information Processing Systems*, 2023.

A. Omitted Details in Section 3

We start with restating Theorem 4.1 along with its proof.

Theorem 4.1. *Under Assumptions 2.1-2.4, for any $\gamma_1, \gamma_2 \geq 0$, the regret $\mathbf{Reg}_{\mathsf{CB}}$ of* SquareCB.UG *is at most*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathsf{dec}^{\mathsf{P}}_{\gamma_1,\gamma_2}(p_t; f_t, g_t, x_t)\right] + \gamma_1 \mathbf{Reg}_{\mathsf{Sq}} + 2\gamma_2 \mathbf{Reg}_{\mathsf{Log}}.$$

Proof. Define $\pi^\star(x) = \operatorname{argmin}_{i\in[K]} f^\star(x,i)$. We decompose $\mathbf{Reg}_{\mathsf{CB}}$ as follows:

$$\begin{aligned}
&\mathbf{Reg}_{\mathsf{CB}} \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} p_{t,i} f^\star(x_t,i) - \sum_{t=1}^{T} f^\star(x_t,\pi^\star(x_t))\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{i=1}^{K} p_{t,i} f^\star(x_t,i) - f^\star(x_t,\pi^\star(x_t)) - \gamma_1 \sum_{i=1}^{K}\sum_{j=1}^{K} p_{t,i} g^\star(x_t,i,j)\left(f^\star(x_t,j) - f_t(x_t,j)\right)^2\right)\right] \\
&\quad - \mathbb{E}\left[\sum_{t=1}^{T}\gamma_2 \sum_{i=1}^{K}\sum_{j=1}^{K} p_{t,i}\frac{(g^\star(x_t,i,j) - g_t(x_t,i,j))^2}{g^\star(x_t,i,j) + g_t(x_t,i,j)}\right] + \gamma_1 \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} p_{t,i} g^\star(x_t,i,j)\left(f^\star(x_t,j) - f_t(x_t,j)\right)^2\right] \\
&\quad + \gamma_2 \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} p_{t,i}\frac{(g^\star(x_t,i,j) - g_t(x_t,i,j))^2}{g^\star(x_t,i,j) + g_t(x_t,i,j)}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \max_{\substack{i^\star\in[K] \\ v^\star\in[0,1]^K, M^\star\in[0,1]^{K\times K}}}\left(\sum_{i=1}^{K} p_{t,i} v_i^\star - v_{i^\star}^\star - \gamma_1 \sum_{i=1}^{K}\sum_{j=1}^{K} p_{t,i} M_{i,j}^\star\left(v_j^\star - f_t(x_t,j)\right)^2\right.\right. \\
&\quad \left.\left. -\gamma_2 \sum_{i=1}^{K}\sum_{j=1}^{K} p_{t,i}\frac{(M_{i,j}^\star - g_t(x_t,i,j))^2}{M_{i,j}^\star + g_t(x_t,i,j)}\right)\right] + \gamma_1 \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} p_{t,i} g^\star(x_t,i,j)\left(f^\star(x_t,j) - f_t(x_t,j)\right)^2\right] \\
&\quad + \gamma_2 \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} p_{t,i}\frac{(g^\star(x_t,i,j) - g_t(x_t,i,j))^2}{g^\star(x_t,i,j) + g_t(x_t,i,j)}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathsf{dec}^{\mathsf{P}}_{\gamma_1,\gamma_2}(p_t, f_t, g_t, x_t)\right] + \gamma_1 \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} p_{t,i} g^\star(x_t,i,j)\left(f^\star(x_t,j) - f_t(x_t,j)\right)^2\right] \\
&\quad + \gamma_2 \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} p_{t,i}\frac{(g^\star(x_t,i,j) - g_t(x_t,i,j))^2}{g^\star(x_t,i,j) + g_t(x_t,i,j)}\right]. \qquad (6)
\end{aligned}$$

Next, since $\mathbb{E}[\ell_{t,a} \mid x_t] = f^\star(x_t,a)$ for all $t \in [T]$ and $a \in [K]$, we know that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} p_{t,i} g^\star(x_t,i,j)\left(f^\star(x_t,j) - f_t(x_t,j)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{A_t}\left[\sum_{i\in A_t}(f_t(x_t,i) - \ell_{t,i})^2 - \sum_{i\in A_t}(f^\star(x_t,i) - \ell_{t,i})^2\right]\right] \leq \mathbf{Reg}_{\mathsf{Sq}}, \qquad (7)
\end{aligned}$$

where the inequality is due to Assumption 2.2 and the way the algorithm feeds the oracle AlgSq. In addition, according to

Lemma 2.5 and the way the algorithm feeds the oracle AlgLog in the partially revealed graph setting, we know that

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} p_{t,i}\frac{(g^\star(x_t,i,j)-g_t(x_t,i,j))^2}{g^\star(x_t,i,j)+g_t(x_t,i,j)}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{i_t\sim p_t}\left[\sum_{i=1}^{K}\frac{(g^\star(x_t,i_t,i)-g_t(x_t,i_t,i))^2}{g^\star(x_t,i_t,i)+g_t(x_t,i_t,i)}\right]\right] \le 2\mathbf{Reg}_{\mathsf{Log}}. \tag{8}$$

Combining the above two inequalities, we obtain that

$$\mathbf{Reg}_{\mathsf{CB}} \le \mathbb{E}\left[\sum_{t=1}^{T}\mathsf{dec}^{\mathsf{P}}_{\gamma_1,\gamma_2}(p_t;f_t,g_t,x_t)\right] + \gamma_1\mathbf{Reg}_{\mathsf{Sq}} + 2\gamma_2\mathbf{Reg}_{\mathsf{Log}}. \tag{9}$$

Moreover, Eq. (9) also holds in the fully revealed feedback graph setting since in this case, according to Lemma 2.5 and the fact that the algorithm feeds all action-pairs to AlgLog, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} p_{t,i}\frac{(g^\star(x_t,i,j)-g_t(x_t,i,j))^2}{g^\star(x_t,i,j)+g_t(x_t,i,j)}\right] \le \mathbb{E}\left[\sum_{t=1}^{T}\left[\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{(g^\star(x_t,i,j)-g_t(x_t,i,j))^2}{g^\star(x_t,i,j)+g_t(x_t,i,j)}\right]\right] \le 2\mathbf{Reg}_{\mathsf{Log}}. \tag{10}$$

Plugging Eq. (10) and Eq. (7) into Eq. (6) finishes the proof in the fully revealed feedback graph setting.

□

A.1. Value of the Minimax Program

In this subsection, we show that for any context $x \in \mathcal{X}$, $g \in \mathrm{conv}(\mathcal{G})$ and $f \in \mathrm{conv}(\mathcal{F})$, the minimum DEC value is roughly of order $\frac{\alpha(g,x)}{\gamma}$.

Lemma 4.4. *For any $g \in \mathrm{conv}(\mathcal{G})$, $f \in \mathrm{conv}(\mathcal{F})$, $x \in \mathcal{X}$, and $\gamma \ge 4$, we have*

$$\min_{p\in\Delta([K])}\mathsf{dec}_\gamma(p;f,g,x) = \mathcal{O}\left(\frac{\alpha(g,x)\log(K\gamma)}{\gamma}\right).$$

Proof. To bound $\mathsf{dec}_\gamma(p;f,g,x)$, it suffices to bound $\overline{\mathsf{dec}}_\gamma(p;f,g,x)$ defined as follows, which relaxes the constraint from $v^\star \in [0,1]^K$ to $v^\star \in \mathbb{R}^K$:

$$\overline{\mathsf{dec}}_\gamma(p;f,g,x) = \max_{\substack{i^\star\in[K]\\ v^\star\in\mathbb{R}^K}}\left[\sum_{i=1}^{K}p_iv_i^\star - v_{i^\star}^\star - \frac{1}{4}\gamma\sum_{i=1}^{K}\sum_{j=1}^{K}p_jg(x,i,j)\left(f(x,i)-v_i^\star\right)^2\right].$$

For a positive definite matrix $M \in \mathbb{R}^{K\times K}$, we define norm $\|z\|_M = \sqrt{z^\top Mz}$. By taking the gradient with respect to $v^\star$ and setting it to zero, we know that for any $i^\star \in [K]$ and $p \in \Delta([K])$,

$$\begin{aligned}&\max_{v^\star\in\mathbb{R}^K}\left\{\sum_{i=1}^{K}p_iv_i^\star - v_{i^\star}^\star - \frac{1}{4}\gamma\sum_{i=1}^{K}\sum_{j=1}^{K}p_jg(x,i,j)(f(x,i)-v_i^\star)^2\right\}\\ &= \sum_{i=1}^{K}p_if(x,i) - f(x,i^\star) + \frac{1}{\gamma}\|p-e_{i^\star}\|^2_{W(p,g,x)},\end{aligned} \tag{11}$$

where $W(g,p,x)$ is a diagonal matrix with the i-th diagonal entry being $\sum_{j=1}^{K}p_jg(x,j,i)$ and $e_{i^\star} \in \mathbb{R}^K$ corresponds to the basic vector with the i-th coordinate being 1.

Then, direct calculation shows that

$$\min_{p\in\Delta([K])}\overline{\mathsf{dec}}_\gamma(p;f,g,x)$$

$$= \min_{p\in\Delta([K])} \max_{i^\star\in[K]} \max_{v^\star\in\mathbb{R}^K} \left\{ \sum_{i=1}^K p_i v_i^\star - v_{i^\star}^\star - \frac{1}{4}\gamma \sum_{i=1}^K \sum_{j=1}^K p_j g(x,i,j)(f(x,i)-v_i^\star)^2 \right\}$$

$$= \min_{p\in\Delta([K])} \max_{i^\star\in[K]} \left\{ \sum_{i=1}^K p_i f(x,i) - f(x,i^\star) + \frac{1}{\gamma}\|p-e_{i^\star}\|^2_{W(p,g,x)} \right\} \quad \text{(according to Eq. (11))}$$

$$= \min_{p\in\Delta([K])} \max_{i^\star\in[K]} \left\{ \sum_{i=1}^K p_i f(x,i) - f(x,i^\star) + \frac{1}{\gamma}\sum_{i\neq i^\star} \frac{p_i^2}{\sum_{i'=1}^K p_{i'} g(x,i',i)} + \frac{1}{\gamma}\frac{(1-p_{i^\star})^2}{\sum_{i'=1}^K p_{i'} g(x,i',i^\star)} \right\} \tag{12}$$

$$= \min_{p\in\Delta([K])} \max_{q\in\Delta([K])} \left\{ \sum_{i=1}^K p_i f(x,i) - \sum_{i=1}^K q_i f(x,i) + \frac{1}{\gamma}\sum_{i=1}^K \frac{(1-q_i)p_i^2}{\sum_{i'=1}^K p_{i'} g(x,i',i)} + \frac{1}{\gamma}\sum_{i=1}^K \frac{q_i(1-p_i)^2}{\sum_{i'=1}^K p_{i'} g(x,i',i)} \right\}$$

$$= \max_{q\in\Delta([K])} \min_{p\in\Delta([K])} \left\{ \sum_{i=1}^K p_i f_t(x,i) - \sum_{i=1}^K q_i f(x,i) + \frac{1}{\gamma}\sum_{i=1}^K \frac{(1-q_i)p_i^2}{\sum_{i'=1}^K p_{i'} g(x,i',i)} + \frac{1}{\gamma}\sum_{i=1}^K \frac{q_i(1-p_i)^2}{\sum_{i'=1}^K p_{i'} g(x,i',i)} \right\}, \tag{13}$$

where the last inequality is due to Sion's minimax theorem.

Picking $p_a = \left(1-\frac{1}{\gamma}\right) q_a + \frac{1}{\gamma K}$ for all $a\in[K]$, we obtain that for any distribution $\mu\in\mathcal{Q}(g,x)$,

$$\max_{q\in\Delta([K])} \min_{p\in\Delta([K])} \left\{ \sum_{i=1}^K p_i f(x,i) - \sum_{i=1}^K q_i f(x,i) + \frac{1}{\gamma}\sum_{i=1}^K \frac{(1-q_i)p_i^2}{\sum_{i'=1}^K p_{i'} g(x,i',i)} + \frac{1}{\gamma}\sum_{i=1}^K \frac{q_i(1-p_i)^2}{\sum_{i'\in\mathcal{A}} p_{i'} g(x,i',i)} \right\}$$

$$\overset{(i)}{\leq} \max_{q\in\Delta([K])} \left\{ -\frac{1}{\gamma}\sum_{i=1}^K q_i f(x,i) + \frac{1}{\gamma K}\sum_{i=1}^K f(x,i) + \frac{1}{\gamma}\sum_{i=1}^K \frac{(1-q_i)\left(q_i - \frac{1}{\gamma}q_i + \frac{1}{\gamma K}\right)^2 + q_i\left(1-(1-\frac{1}{\gamma})q_i - \frac{1}{\gamma K}\right)^2}{\sum_{i'=1}^K p_{i'} g(x,i',i)} \right\}$$

$$\overset{(ii)}{\leq} \max_{q\in\Delta([K])} \left\{ \frac{1}{\gamma} + \frac{1}{\gamma}\sum_{i=1}^K \frac{2\left((1-\frac{1}{\gamma})^2 q_i^2 + \frac{1}{\gamma^2K^2}\right)(1-q_i) + q_i\left(1-(1-\frac{1}{\gamma})q_i\right)^2}{\sum_{i'=1}^K p_{i'} g_t(x_t,i',i)} \right\}$$

$$\leq \max_{q\in\Delta([K])} \left\{ \frac{1}{\gamma} + \frac{1}{\gamma}\sum_{i=1}^K \frac{2\left(q_i^2 + \frac{1}{\gamma^2K^2}\right)(1-q_i) + q_i\left((1-q_i) + \frac{q_i}{\gamma}\right)^2}{\sum_{i'=1}^K p_{i'} g_t(x_t,i',i)} \right\}$$

$$\overset{(iii)}{\leq} \max_{q\in\Delta([K])} \left\{ \frac{1}{\gamma} + \frac{2}{\gamma^2} + \frac{1}{\gamma}\sum_{i=1}^K \frac{2q_i^2(1-q_i) + 2q_i(1-q_i)^2 + \frac{2q_i^3}{\gamma^2}}{\sum_{i'=1}^K p_{i'} g(x,i',i)} \right\}$$

$$= \max_{q\in\Delta([K])} \left\{ \frac{1}{\gamma} + \frac{2}{\gamma^2} + \frac{2}{\gamma}\sum_{i=1}^K \frac{q_i(1-q_i)}{\sum_{i'=1}^K p_{i'} g(x,i',i)} + \frac{2}{\gamma^3}\sum_{i=1}^K \frac{q_i^3}{\sum_{i'=1}^K p_{i'} g(x,i',i)} \right\}$$

$$\overset{(iv)}{=} \max_{q\in\Delta([K])} \left\{ \frac{1}{\gamma} + \frac{2}{\gamma^2} + \frac{2}{\gamma}\sum_{i=1}^K \frac{q_i(1-q_i)}{\mathbb{E}_{G\sim\mu}\left[\sum_{i'=1}^K p_{i'} G_{i',i}\right]} + \frac{2}{\gamma^3}\sum_{i=1}^K \frac{q_i^3}{\mathbb{E}_{G\sim\mu}\left[\sum_{i'=1}^K p_{i'} G_{i',i}\right]} \right\},$$

where (i) is by replacing p_i with $(1-\frac{1}{\gamma})q_i + \frac{1}{\gamma K}$ (except for the $p_{i'}$ in the denominators); (ii) holds since $f(x,i)\in[0,1]$ for all $i\in[K]$ and $(a+b)^2 \leq 2(a^2+b^2)$, and we drop the last $-\frac{1}{\gamma K}$ term; (iii) is because $\sum_{i'=1}^K p_{i'} g(x,i',i) \geq \frac{1}{\gamma K}$ for all $i\in[K]$ since $p_{i'} \geq \frac{1}{\gamma K}$ for all $i'\in[K]$ and $\sum_{i'=1}^K g(x,i',i) \geq 1$ as $g(x,\cdot,\cdot)$ is the mean graph of a distribution of strongly observable graphs; (iv) is by definition of $\mathcal{Q}(g,x)$ and with an abuse of notation, $G_{i,j}$ represents the (i,j)-th entry of the adjacent matrix of G.

For a feedback graph $G\in\{0,1\}^{K\times K}$ and a distribution $u\in\Delta([K])$, with an abuse of notation, define $W(G,u)\in[0,1]^K$ as the probability for each node $i\in[K]$ to be observed according to u and G. Specifically, for each $i\in[K]$, $W_i(G,u) = \sum_{j=1}^K u_j G_{j,i}$. In addition, let $S\subseteq[K]$ be the nodes in G that have self-loops, meaning that $G_{i,i}=1$ for all $i\in S$. Then,

using Jensen's inequality, we know that

$$
\begin{aligned}
&\min_{p\in\Delta([K])} \overline{\mathsf{dec}}_\gamma(p; f, g, x) \\
&\le \max_{q\in\Delta([K])} \mathbb{E}_{G\sim\mu}\left[\frac{1}{\gamma} + \frac{2}{\gamma^2} + \frac{2}{\gamma}\sum_{i=1}^K \frac{q_i(1-q_i)}{\sum_{i'=1}^K p_{i'}G_{i',i}} + \frac{2}{\gamma^3}\sum_{i=1}^K \frac{q_i^3}{\sum_{i'=1}^K p_{i'}G_{i',i}}\right] && \text{(Jensen's inequality)} \\
&\overset{(i)}{=} \max_{q\in\Delta([K])} \mathbb{E}_{G\sim\mu}\left[\frac{1}{\gamma} + \frac{2}{\gamma^2} + \frac{2}{\gamma}\sum_{i=1}^K \frac{q_i(1-q_i)}{W_i(G,p)} + \frac{2}{\gamma^3}\sum_{i\in S} \frac{q_i^3}{W_i(G,p)} + \frac{2}{\gamma^3}\sum_{i\notin S} \frac{q_i^3}{W_i(G,p)}\right] \\
&\overset{(ii)}{\le} \max_{q\in\Delta([K])} \mathbb{E}_{G\sim\mu}\left[\frac{1}{\gamma} + \frac{2}{\gamma^2} + \frac{2}{\gamma}\sum_{i=1}^K \frac{q_i(1-q_i)}{W_i(G,p)} + \frac{2}{\gamma^2(\gamma-1)}\sum_{i\in S} q_i^2 + \frac{2}{\gamma^3}\sum_{i\notin S} \frac{q_i^3}{\frac{K-1}{\gamma K}}\right] \\
&\overset{(iii)}{\le} \max_{q\in\Delta([K])} \mathbb{E}_{G\sim\mu}\left[\mathcal{O}\left(\frac{1}{\gamma} + \frac{1}{\gamma}\sum_{i=1}^K \frac{q_i(1-q_i)}{W_i(G,p)}\right)\right] \\
&\overset{(iv)}{\le} \mathcal{O}\left(\frac{1}{\gamma}\right) + \mathcal{O}\left(\frac{1}{\gamma}\right)\cdot \mathbb{E}_{G\sim\mu}\left[\max_{q\in\Delta([K])}\sum_{i=1}^K \frac{q_i(1-q_i)}{W_i(G,p)}\right],
\end{aligned} \tag{14}
$$

where (i) is by definition of $W(G,p)$; (ii) is because for all $i \in S$, $W_i(G,p) \ge p_i \ge (1-\frac{1}{\gamma})q_i$, and for all $i \notin S$, every other node in G can observe i and $W_i(G,p) = 1 - p_i \ge \frac{K-1}{\gamma K}$ since $p_i \ge \frac{1}{\gamma K}$; (iii) is because $K \ge 2$ and $\gamma \ge 4$; (iv) is again due to Jensen's inequality.

Next we bound $\sum_{i=1}^K \frac{q_i(1-q_i)}{W_i(G,p)}$ for any strongly observable graph G with independence number α. For notational convenience, we omit the index G and denote $W_i(G,p)$ by $W_i(p)$. If $i \in [K]\backslash S$, we know that $W_i(p) = 1 - p_i$ and

$$
\frac{q_i(1-q_i)}{W_i(p)} = \frac{q_i(1-q_i)}{1-\left(1-\frac{1}{\gamma}\right)q_i - \frac{1}{\gamma K}} = \frac{q_i(1-q_i)}{\left(1-\frac{1}{\gamma}\right)(1-q_i) + \frac{K-1}{\gamma K}} \le \frac{1}{1-\frac{1}{\gamma}}q_i \le 2q_i, \tag{15}
$$

where the first equality is because $p_i = (1-\frac{1}{\gamma})q_i + \frac{1}{\gamma K}$ for all $i \in [K]$ and the last inequality is because $\gamma \ge 4$. If $i \in S$, we know that

$$
\begin{aligned}
\sum_{i\in S}\frac{q_i(1-q_i)}{W_i(p)} &= \sum_{i\in S}\frac{q_i(1-q_i)}{\sum_{j:G_{j,i}=1}\left(\left(1-\frac{1}{\gamma}\right)q_j + \frac{1}{\gamma K}\right)} \\
&\le \frac{\gamma}{\gamma-1}\sum_{i\in S}\frac{\left((1-\frac{1}{\gamma})q_i + \frac{1}{\gamma K}\right)(1-q_i)}{\sum_{j:G_{j,i}=1}\left(\left(1-\frac{1}{\gamma}\right)q_j + \frac{1}{\gamma K}\right)} \\
&\le 2\sum_{i\in S}\frac{\left((1-\frac{1}{\gamma})q_i + \frac{1}{\gamma K}\right)(1-q_i)}{\sum_{j:G_{j,i}=1}\left(\left(1-\frac{1}{\gamma}\right)q_j + \frac{1}{\gamma K}\right)} \\
&\le \mathcal{O}(\alpha\log(K\gamma)),
\end{aligned} \tag{16}
$$

where the last inequality is due to Lemma A.2.

Combining Eq. (15) and Eq. (16), we know that for any $q \in \Delta([K])$ and strongly observable graph G with independence number α, $\sum_{i=1}^K \frac{q_i(1-q_i)}{W_i(G,p)} \le \mathcal{O}(\alpha\log(K\gamma))$. Plugging the above back to Eq. (14), we know that

$$
\min_{p\in\Delta([K])} \overline{\mathsf{dec}}_\gamma(p; f, g, x) \le \mathcal{O}\left(\frac{1}{\gamma}\right)\cdot \mathbb{E}_{G\sim\mu}\left[\mathcal{O}(\alpha(G)\log(K\gamma))\right] \le \mathcal{O}\left(\frac{\alpha(g,x)\log(K\gamma)}{\gamma}\right),
$$

where the last inequality is due to the definition of $\alpha(g,x)$.

□

The following two auxiliary lemmas have been used in our analysis.

Lemma A.1. *For all $\widehat{g} \in \mathrm{conv}(\mathcal{G})$ and context $x \in \mathcal{X}$, $\alpha(\widehat{g}, x) \le \alpha(\mathcal{G})$.*

Proof. Let $u \in \Delta(\mathcal{G})$ be such that $\mathbb{E}_{g\sim u}[g] = \widehat{g}$. For each $g \in \mathcal{G}$, consider any $q_g \in \mathcal{Q}(g, x)$. We have by definition $\widehat{q} \triangleq \mathbb{E}_{g\sim u}[q_g] \in \mathcal{Q}(\widehat{g}, x)$, leading to

$$\alpha(\widehat{g}, x) = \inf_{q\in\mathcal{Q}(\widehat{g},x)} \mathbb{E}_{G\sim q}[\alpha(G)] \le \mathbb{E}_{G\sim\widehat{q}}[\alpha(G)] = \mathbb{E}_{g\sim u}\left[\mathbb{E}_{G\sim q_g}[\alpha(G)]\right].$$

Since q_g can be any distribution in $\mathcal{Q}(g, x)$, the above implies

$$\alpha(\widehat{g}, x) \le \mathbb{E}_{g\sim u}\left[\inf_{\rho\in\mathcal{Q}(g,x)} \mathbb{E}_{G\sim\rho}[\alpha(G)]\right] = \mathbb{E}_{g\sim u}\left[\alpha(g, x)\right],$$

which is at most $\mathbb{E}_{g\sim u}[\sup_{x\in\mathcal{X}} \alpha(g, x)] \le \alpha(\mathcal{G})$, finishing the proof. □

Lemma A.2 (Lemma 5 in (Alon et al., 2015a))**.** *Let $G = (V, E)$ be a directed graph with $|V| = K$, in which $G_{i,i} = 1$ for all vertices $i \in [K]$. Assign each $i \in V$ with a positive weight w_i such that $\sum_{i=1}^n w_i \le 1$ and $w_i \ge \varepsilon$ for all $i \in V$ for some constant $0 < \varepsilon < \frac{1}{2}$. Then*

$$\sum_{i=1}^{K} \frac{w_i}{\sum_{j:G_{j,i}=1} w_j} \le 4\alpha(G) \ln \frac{4K}{\alpha(G)\varepsilon},$$

where $\alpha(G)$ is the independence number of G.

A.2. Parameter-Free Algorithm in the Partially Revealed Feedback Graphs Setting

In this section, we show that applying doubling trick to Algorithm 1 achieves the same regret without the knowledge of $\alpha(\mathcal{G})$ in the partially revealed feedback setting. The idea follows Zhang et al. (2023), which utilizes the value of the minimax problem Eq. (1) to guide the choice of γ.

Specifically, our algorithm goes in epochs with the parameter γ being γ_s in the s-th epoch and $\gamma_1 = \sqrt{\frac{T}{\max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\}}}$. Within each epoch s (with starting round b_s), at round t, we calculate the value

$$R_t = \sum_{\tau=b_s}^{t} \min_{p\in\Delta([K])} \mathsf{dec}_{\gamma_s}(p; f_\tau, g_\tau, x_\tau), \tag{17}$$

and decide whether to start a new epoch by checking whether $R_t \le \gamma_s \max\left\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\right\}$. Specifically, if $R_t \le \gamma_s \max\left\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\right\}$, we continue our algorithm using γ_s; otherwise, we set $\gamma_{s+1} = 2\gamma_s$ and restart the algorithm.

Now we analyze the performance of the above described algorithm. Denote the s-th epoch to be $I_s = \{b_s, b_s + 1, \ldots, e_s\} \triangleq [b_s, e_s]$ and let S be the total number of epochs. First, using Lemma A.1 and Lemma 4.4, we know that for any t within epoch s, we have

$$\gamma_s R_t \le \widetilde{\mathcal{O}}\left(\sum_{t\in I_s} \alpha(\mathcal{G})\right) \le \widetilde{\mathcal{O}}(\alpha(\mathcal{G})T).$$

Applying this to the last round of the $(S-1)$-th epoch, we obtain:

$$\gamma_{S-1}^2 \max\left\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\right\} \le \widetilde{\mathcal{O}}(\alpha(\mathcal{G})T),$$

which, together with $\gamma_{S-1} = 2^{S-2}\gamma_1$ and the definition of γ_1, implies $2^S = \widetilde{\mathcal{O}}(\sqrt{\alpha(G)})$.

Next, consider the regret in epoch s. According to Eq. (4), we know that the regret within epoch s is bounded as follows:

$$\mathbb{E}\left[\sum_{t\in I_s} f^\star(x_t, i_t) - \sum_{t\in I_s} \min_{i\in[K]} f^\star(x_t, i)\right]$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\sum_{t\in I_s} \mathsf{dec}_{\gamma_s}(p_t; f_t, g_t, x_t)\right] + \frac{3\gamma_s}{4}\mathbf{Reg}_{\mathsf{Sq}} + \frac{1}{2}\gamma_s \mathbf{Reg}_{\mathsf{Log}} \\
&\leq \mathbb{E}\left[\sum_{t\in[b_s, e_s-1]} \mathsf{dec}_{\gamma_s}(p_t; f_t, g_t, x_t)\right] + \widetilde{\mathcal{O}}\left(\frac{\alpha(\mathcal{G})}{\gamma_s}\right) + 2\gamma_s \max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\} \\
&\leq \widetilde{\mathcal{O}}(\gamma_s \max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\}), \qquad (18)
\end{aligned}
$$

where the second inequality uses Lemma 4.2 and Lemma 4.4 again, and the last inequality is because at round $t = e_s - 1$, $R_t \leq \gamma_s \max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\}$ is satisfied and $\frac{\alpha(\mathcal{G})}{\gamma_s} \leq \frac{T}{\gamma_s} \leq \gamma_s \max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\}$. Taking summation over all S epochs, we know that the overall regret is bounded as

$$
\begin{aligned}
\mathbb{E}[\mathbf{Reg}_{\mathsf{CB}}] &\leq \sum_{s=1}^{S} \widetilde{\mathcal{O}}(\gamma_s \max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\}) = \sum_{s=1}^{S} \widetilde{\mathcal{O}}\left(2^{s-1}\sqrt{T\max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\}}\right) \\
&\leq \widetilde{\mathcal{O}}\left(2^{S}\sqrt{T\max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\}}\right) = \widetilde{\mathcal{O}}\left(\sqrt{\alpha(\mathcal{G})T\max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\}}\right), \qquad (19)
\end{aligned}
$$

which is exactly the same as Theorem 3.1.

B. Omitted Details in Section 4.2

Lemma 4.5. *For $g, g' \in \mathrm{conv}(\mathcal{G})$, $f \in \mathrm{conv}(\mathcal{F})$, and $x \in \mathcal{X}$, we have*

$$
\begin{aligned}
\min_{p\in\Delta([K])} \mathsf{dec}_{\gamma}(p; f, g', x) &\leq \min_{p\in\Delta([K])} \mathsf{dec}_{\frac{\gamma}{3}}(p; f, g, x) \\
&+ \frac{\gamma}{12}\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{(g(x,i,j) - g'(x,i,j))^2}{g(x,i,j) + g'(x,i,j)}.
\end{aligned}
$$

Proof. For any $v \in [0,1]^K$, applying Lemma 4.3 with $z' = g'(x,i,j)(f(x,j) - v_j)^2$ and $z = g(x,i,j)(f(x,j) - v_j)^2$ for each i and j, we know that

$$
\begin{aligned}
&-3\gamma\sum_{i=1}^{K}\sum_{j=1}^{K} p_i g'(x,i,j)(f(x,j) - v_j)^2 - \gamma\sum_{i=1}^{K}\sum_{j=1}^{K} p_i \frac{(g(x,i,j) - g'(x,i,j))^2}{g(x,i,j) + g'(x,i,j)} \\
&\leq -\gamma\sum_{i=1}^{K}\sum_{j=1}^{K} p_i g(x,i,j)(f(x,j) - v_j)^2. \qquad (20)
\end{aligned}
$$

Plugging this back to the definition of $\mathsf{dec}_{\gamma}(p; f, g, x)$, we get

$$
\begin{aligned}
&\min_{p\in\Delta([K])} \mathsf{dec}_{\gamma}(p; f, g', x) \\
&= \min_{p\in\Delta([K])} \max_{\substack{i^\star\in[K] \\ v^\star\in[0,1]^K}} \left[\sum_{i=1}^{K} p_i v_i^\star - v_{i^\star}^\star - \frac{1}{4}\gamma\sum_{i=1}^{K}\sum_{j=1}^{K} p_i g'(x,i,j)\left(f(x,j) - v_j^\star\right)^2\right] \\
&\overset{(i)}{\leq} \min_{p\in\Delta([K])} \max_{\substack{i^\star\in[K] \\ v^\star\in[0,1]^K}} \left[\sum_{i=1}^{K} p_i v_i^\star - v_{i^\star}^\star - \frac{1}{12}\gamma\sum_{i=1}^{K}\sum_{j=1}^{K} p_i g(x,i,j)\left(f(x,j) - v_j^\star\right)^2\right. \\
&\qquad \left. + \frac{1}{12}\gamma\sum_{i=1}^{K}\sum_{j=1}^{K} p_i \frac{(g(x,i,j) - g'(x,i,j))^2}{g(x_t,i,j) + g'(x,i,j)}\right] \\
&\overset{(ii)}{\leq} \min_{p\in\Delta([K])} \max_{\substack{i^\star\in[K] \\ v^\star\in[0,1]^K}} \left[\sum_{i=1}^{K} p_i v_i^\star - v_{i^\star}^\star - \frac{1}{12}\gamma\sum_{i=1}^{K}\sum_{j=1}^{K} p_i g(x,i,j)\left(f(x,j) - v_j^\star\right)^2\right]
\end{aligned}
$$

$$+\frac{1}{12}\gamma\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{(g(x,i,j)-g'(x,i,j))^2}{g(x,i,j)+g'(x,i,j)}$$

$$=\min_{p\in\Delta([K])}\mathsf{dec}_{\frac{\gamma}{3}}(p;f,g,x)+\frac{1}{12}\gamma\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{(g(x,i,j)-g'(x,i,j))^2}{g(x,i,j)+g'(x,i,j)},$$

where (i) uses Eq. (20) and (ii) holds by trivially bounding p_i in the last term by 1. □

C. Implementation Details

We first point out that the DEC defined in Zhang et al. (2023) in fact relaxes the constraint $v^\star \in [0,1]^K$ to $v^\star \in \mathbb{R}^K$, which makes the problem of minimizing the DEC a simple convex program (see their Theorem 3.6). In the partially revealed graph setting, we in fact can do the exact same trick because it does not affect our analysis at all. Since our experiments are for an application with partially revealed graphs, we indeed implemented our algorithm in this way for simplicity.

However, this relaxation does not work for the fully revealed graph setting, since the analysis of Lemma 4.5 relies on $v^\star \in [0,1]^K$. Nevertheless, minimizing the DEC is still a relatively simple convex problem. To see this, we first fix an $i^\star \in [K]$ and work on the supremum over $v^\star \in [0,1]^K$. Specifically, define

$$v^\star(i^\star)=\operatorname*{argmax}_{v\in[0,1]^K}\left\{\sum_{i=1}^{K}p_iv_i-v_{i^\star}-\frac{\gamma}{4}\sum_{i=1}^{K}p_i\sum_{j=1}^{K}g(x,i,j)(f(x,j)-v_j)^2\right\}.$$

Direct calculation shows that $v^\star_j(i^\star)=\max\left\{f(x,j)-\frac{(p_j-\mathbb{1}\{j=i^\star\})^2}{\gamma\sum_{i=1}^{K}p_ig(x,i,j)},0\right\}$ for all $j\in[K]$. Let $h_{i^\star}(p)$ be the maximum value attained by $v^\star(i^\star)$, which is convex in p since it is a point-wise supremum over functions linear in p. It is then clear that solving $\operatorname{argmin}_{p\in\Delta([K])}\overline{\mathsf{dec}}_\gamma(p;f,g,x)$ is equivalent to solving the following constrained convex problem:

$$\min_{u\in\mathbb{R},p\in\Delta([K])} \quad u$$
$$s.t. \quad h_{i^\star}(p)\le u,\ \forall i^\star\in[K].$$

C.1. Omitted Details in Section 5

In this section, we include the omitted details for our experiments.

Dataset Details. For the synthetic datasets, as mentioned in Section 5.1, the two datasets differ in how $\{x_t\}_{t=1}^T$ are generated. In the first dataset, each coordinate of $x_t\in\mathbb{R}^{32}$ is independently drawn from $\mathcal{N}(0,1)$; in the second dataset, the first 8 coordinates of x_t is independently drawn from $\mathcal{N}(0,1)$ and the remaining coordinates are all 1. The real auction dataset we used in Section 5.2 is an eBay auction dataset (available at `https://cims.nyu.edu/~munoz/data/`) with the t-th datapoint consisting of a 78-dimensional feature vector x_t, a winning price of the auction v_t, and a competing price w_t. We treat the winning price as the value of the learner in our experiment. We randomly select a subset of 5000 data points whose winning price is in range $[100,300]$ and normalize the value and the competing price to range $[0,1]$.

Model Details. We implement the graph oracle as a linear classification model, aiming to predict the distribution of the competing price, denoted as $p_w(x_t)\in\Delta([K])$. With $p_w(x_t)$, we sample $\widehat{w}_t\sim p_w(x_t)$ and the predicted graph g_t is calculated as

$$g_t(x_t,i,j)=\mathbb{1}[i\ge\widehat{w}_t]\cdot\mathbb{1}[i\le j]+\mathbb{1}[i<\widehat{w}_t]\cdot\mathbb{1}[j<\widehat{w}_t]. \tag{21}$$

We implement the loss oracle as a two-layer fully connected neural network with hidden size 32. The neural network predicts the value of the data point x_t, denoted as $\widehat{v}_t$. The predicted loss of each arm i is then calculated as:

$$f_t(x_t,i)=\frac{1}{2}\left[1-\mathbb{1}\left[a_i\ge\widehat{w}_t\right]\cdot(\widehat{v}_t-a_i)\right], \tag{22}$$

where $a_i=(i-1)\varepsilon$.

Training Details. For the graph oracle, the loss function of each round t is calculated as: $\frac{1}{K}\sum_{(i,j,b)\in S_t}\ell_{\log}(\mathbb{E}_{\widehat{w}_t\sim p_w(x_t)}[g_t(x_t,i,j)],b)$, where S_t is the input dataset defined in SquareCB.UG. For the loss oracle, the loss function is calculated as $\frac{1}{|A_t|}\sum_{j\in A_t}(f_t(x_t,j)-\ell_{t,j})^2$.

We apply online gradient descent to train both models. Since the loss regression model aims to predict the value, we only update it when the learner wins the auction and observes the value. For experiments on the real auction dataset, learning rate is searched over $\{0.005, 0.01, 0.05\}$ for the loss oracle and over $\{0.01, 0.05\}$ for the graph regression oracle. For experiments on the synthetic datasets, they are searched over $\{0.005, 0.01, 0.02\}$ and $\{0.01, 0.05\}$ respectively. For SquareCB, we set the exploration parameter $\gamma = c\cdot\sqrt{KT}$ (based on what its theory suggests), where c is searched over $\{0.5, 1, 2\}$. For our SquareCB.UG, we set $\gamma = c\cdot\sqrt{T}$, where c is also searched over $\{0.5, 1, 2\}$. The experiment on the real auction dataset is repeated with 8 different random seeds and the experiment on the synthetic datasets is repeated with 4 different random seeds.

A Closed-Form Solution. In this part, we introduce a closed form of p_t which leads to a more efficient implementation of SquareCB.UG for the specific setting considered in the experiments. Specifically, given the predicted competing price $\widehat{w}_t$, let $b\in[\frac{1}{\varepsilon}+1]$ be the smallest action such that $(b-1)\varepsilon\geq\widehat{w}_t$. Given f_t and g_t defined in Eq. (22) and Eq. (21), define a closed-form p which concentrates on action 1 (bidding 0) and action b as follows,

$$p_1=\begin{cases}\frac{1}{2+\gamma(1/2-f_t(x_t,b))} & f_t(x_t,b)\leq\frac{1}{2}\\ 1-\frac{1}{2+\gamma(f_t(x_t,b)-1/2)} & f_t(x_t,b)>\frac{1}{2}\end{cases},\quad p_b=1-p_1, \tag{23}$$

In the following lemma, we prove that the closed-form probability distribution in Eq. (23) guarantees $\mathsf{dec}_\gamma(p;f_t,g_t,x_t)\leq\frac{4}{\gamma}$, which is enough for all our analysis to hold (despite the fact that it does not exactly minimize the DEC).

Lemma C.1. *For any $x_t\in\mathcal{X}$, $f_t\in\mathcal{F}$, and g_t in the form of Eq. (21), the probability distribution p defined in Eq. (23) guarantees $\mathsf{dec}_\gamma(p,f_t,g_t,x_t)\leq\frac{4}{\gamma}$.*

Proof. According to the analysis in Lemma 4.4, it suffices to bound $\overline{\mathsf{dec}}_\gamma(p,f_t,g_t,x_t)$. Based on Eq. (12), we have

$$\begin{aligned}\overline{\mathsf{dec}}_\gamma(p,f_t,g_t,x_t)&=\max_{i^\star\in[K]}\left\{\sum_{i=1}^K p_if_t(x_t,i)-f_t(x_t,i^\star)+\frac{1}{\gamma}\sum_{i\neq i^\star}\frac{p_i^2}{\sum_{i'=1}^K p_{i'}g_t(x_t,i',i)}+\frac{1}{\gamma}\frac{(1-p_{i^\star})^2}{\sum_{i'=1}^K p_{i'}g_t(x_t,i',i^\star)}\right\}\\&\leq\max_{i^\star\in[K]}\left\{\sum_{i=1}^K p_if_t(x_t,i)-f_t(x_t,i^\star)+\frac{1}{\gamma}\sum_{i=1}^K\frac{p_i^2}{\sum_{i'=1}^K p_{i'}g_t(x_t,i',i)}+\frac{1}{\gamma\sum_{i'=1}^K p_{i'}g_t(x_t,i',i^\star)}\right\}.\end{aligned}$$

Note that according to Eq. (22) and the definition of b, for $i<b$, $f_t(x,i)=\frac{1}{2}$. We now first consider the case $f_t(x_t,b)\leq\frac{1}{2}$ and $p_1=\frac{1}{2+\gamma(\frac{1}{2}-f_t(x_t,b))}\leq\frac{1}{2}$.

1. Suppose $1\leq i^\star<b$, we observe that $f_t(x_t,i^\star)=\frac{1}{2}$ and have

$$\begin{aligned}\overline{\mathsf{dec}}_\gamma(p,f_t,g_t,x_t)&\leq\frac{p_1}{2}+p_bf_t(x_t,b)-\frac{1}{2}+\frac{1}{\gamma}\left(\frac{1}{p_1}+1\right)\\&=\frac{p_1}{2}+(1-p_1)f_t(x_t,b)-\frac{1}{2}+\frac{1}{\gamma}\left(2+\gamma\left(\frac{1}{2}-f_t(x_t,b)\right)+1\right)\\&=\frac{p_1}{2}-p_1f_t(x_t,b)+\frac{3}{\gamma}\\&=p_1\left(\frac{1}{2}-f_t(x_t,b)\right)+\frac{3}{\gamma}\\&\leq\frac{4}{\gamma}.\end{aligned}$$

2. Suppose $b\leq i^\star\leq K$. According to Eq. (22), we know that $f_t(x_t,i^\star)\geq f_t(x_t,b)$ and obtain

$$\overline{\mathsf{dec}}_\gamma(p,f_t,g_t,x_t)\leq\frac{p_1}{2}+p_bf_t(x_t,b)-f_t(x_t,b)+\frac{1}{\gamma}\left(\frac{1}{p_b}+1\right)$$

$$
\begin{aligned}
&\leq \frac{p_1}{2} + (1-p_1) f_t(x_t, b) - f_t(x_t, b) + \frac{3}{\gamma} && \text{(since } p_b = 1 - p_1 \geq \tfrac{1}{2}\text{)} \\
&= p_1 \left(\frac{1}{2} - f_t(x_t, b) \right) + \frac{3}{\gamma} \\
&\leq \frac{4}{\gamma}.
\end{aligned}
$$

Then we consider the case when $f_t(x_t, b) > \frac{1}{2}$ and $p_1 = 1 - \frac{1}{2+\gamma(f_t(x_t,b)-\frac{1}{2})} > \frac{1}{2}$.

1. Suppose $1 \leq i^\star < b$, we have

$$
\begin{aligned}
\overline{\mathsf{dec}}_\gamma(p, f_t, g_t, x_t) &\leq \frac{p_1}{2} + p_b f_t(x_t, b) - \frac{1}{2} + \frac{1}{\gamma}\left(\frac{1}{p_1} + 1\right) \\
&= \frac{1}{2} + p_b \left(f_t(x_t, b) - \frac{1}{2} \right) - \frac{1}{2} + \frac{1}{\gamma}\left(\frac{1}{p_1} + 1\right) \\
&\leq p_b \left(f_t(x_t, b) - \frac{1}{2} \right) + \frac{3}{\gamma} && \text{(since } p_1 > \tfrac{1}{2}\text{)} \\
&\leq \frac{4}{\gamma}.
\end{aligned}
$$

2. Suppose $b \leq i^\star \leq K$, we have

$$
\begin{aligned}
\overline{\mathsf{dec}}_\gamma(p, f_t, g_t, x_t) &\leq \frac{p_1}{2} + p_b f_t(x_t, b) - f_t(x_t, b) + \frac{1}{\gamma}\left(\frac{1}{p_b} + 1\right) \\
&= \frac{1}{2} + p_b \left(f_t(x_t, b) - \frac{1}{2} \right) - f_t(x_t, b) + \frac{1}{\gamma}\left(2 + \gamma\left(f_t(x_t, b) - \frac{1}{2} \right) + 1\right) \\
&= p_b \left(f_t(x_t, b) - \frac{1}{2} \right) + \frac{3}{\gamma} \\
&\leq \frac{4}{\gamma}.
\end{aligned}
$$

Combining the two cases finishes the proof. □

D. Extension to Weakly Observable Graphs

In this section, we show that with a different choice of γ, Algorithm 1 also achieves favorable guarantees when all G_t's are weakly observable graphs, meaning that for each t, G_t is not strongly observable but each node $i \in [K]$ can be observed by certain node ($\exists j \in [K]$ such that $(j, i) \in E_t$). Before we introduce the results, we first introduce the concept of weakly domination number that is needed for the discussion.

Weakly domination number. For a weakly observable graph $G = ([K], E)$, a weakly dominating set is a subset D of nodes such that for any node i in G without a self-loop, there exists $j \in D$ such that $(j, i) \in E$. The size of the smallest weakly dominating set of G is called the weak domination number of G, denoted by $d(G)$.

In the non-contextual adversarial multi-armed bandits with feedback graphs, Alon et al. (2015a) showed that the minimax regret is $\widetilde{\Theta}(d(G)^{\frac{1}{3}} T^{\frac{2}{3}})$ when the feedback graphs $G_t = G$ are fixed and informed. However, $\widetilde{\mathcal{O}}(d(G)^{\frac{1}{3}} T^{\frac{2}{3}})$ is impossible when the feedback graphs are time varying and uninformed based on Theorem 9 of (Alon et al., 2015b), a sharp contrast to the informed case. Specifically, Alon et al. (2015b) showed that the minimax regret is $\widetilde{\Theta}(K^{\frac{1}{3}} T^{\frac{2}{3}})$ when the feedback graphs are all weakly observable and uninformed.

In the following, we introduce our results for the contextual setting when the feedback graphs are weakly observable and uninformed. The following theorem shows that Algorithm 1 achieves $\widetilde{\mathcal{O}}\left(K^{\frac{1}{3}} T^{\frac{2}{3}} (\max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\})^{\frac{1}{3}}\right)$ regret guarantee in weakly observable graphs in general. For notational convenience, similar to the strongly observable case, we

define $\mathcal{Q}'(g,x)$ to be the set of all distributions of weakly observable graphs whose expected edge connections are specified by $g(x,\cdot,\cdot)$.

Theorem D.1. *Under Assumptions 2.1-2.4, with either partially or fully revealed weakly observable feedback graphs,* SquareCB.UG *with* $\gamma = \max\{K, K^{\frac{1}{3}}T^{\frac{2}{3}}(\max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\})^{-\frac{2}{3}}\}$ *guarantees:*

$$\mathbf{Reg}_{\mathsf{CB}} = \mathcal{O}\left(K^{\frac{1}{3}}T^{\frac{2}{3}}(\max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\})^{\frac{1}{3}}\right).$$

Proof. Since Theorem 4.1 and Lemma 4.2 still hold for weakly observable graphs, we only need to show that $\mathsf{dec}_\gamma(p; f_t, G_t, x_t)$ defined in Eq. (2) is upper bounded by $\mathcal{O}\left(\sqrt{\frac{K}{\gamma}}\right)$. To prove this, following Eq. (13) in the proof of Lemma 4.4, we know that for any $x \in \mathcal{X}$, $g \in \mathrm{conv}(\mathcal{G})$, and $f \in \mathrm{conv}(\mathcal{F})$,

$$\begin{aligned}
&\mathsf{dec}_\gamma(p; f, g, x)\\
&\le \max_{q\in\Delta([K])}\min_{p\in\Delta([K])}\left\{\sum_{i=1}^K p_i f_t(x,i) - \sum_{i=1}^K q_i f(x,i) + \frac{1}{\gamma}\sum_{i=1}^K \frac{(1-q_i)p_i^2}{\sum_{i'=1}^K p_{i'} g(x,i',i)} + \frac{1}{\gamma}\sum_{i=1}^K \frac{q_i(1-p_i)^2}{\sum_{i'=1}^K p_{i'} g(x,i',i)}\right\}.
\end{aligned}$$

Picking $p_a = \left(1-\sqrt{\frac{K}{\gamma}}\right)q_a + \frac{1}{\sqrt{\gamma K}}$ for all $a \in [K]$ with $\gamma \ge K$, we know that for any distribution $\mu \in \mathcal{Q}'(g,x)$,

$$\begin{aligned}
&\mathsf{dec}_\gamma(p; f, g, x)\\
&\overset{(i)}{\le} \max_{q\in\Delta([K])}\left\{-\sqrt{\frac{K}{\gamma}}\sum_{i=1}^K q_i f(x,i) + \frac{1}{\sqrt{\gamma K}}\sum_{i=1}^K f(x,i)\right.\\
&\qquad\left. + \frac{1}{\gamma}\sum_{i=1}^K \frac{(1-q_i)\left(q_i - \sqrt{\frac{K}{\gamma}}q_i + \frac{1}{\sqrt{\gamma K}}\right)^2 + q_i\left(1-(1-\sqrt{\frac{K}{\gamma}})q_i - \frac{1}{\sqrt{\gamma K}}\right)^2}{\sum_{i'=1}^K p_{i'} g(x,i',i)}\right\}\\
&\overset{(ii)}{\le} \max_{q\in\Delta([K])}\left\{\sqrt{\frac{K}{\gamma}} + \frac{1}{\gamma}\sum_{i=1}^K \frac{2\left(\left(1-\sqrt{\frac{K}{\gamma}}\right)^2 q_i^2 + \frac{1}{\gamma K}\right)(1-q_i) + q_i\left(1-\left(1-\sqrt{\frac{K}{\gamma}}\right)q_i\right)^2}{\sum_{i'=1}^K p_{i'} g_t(x_t,i',i)}\right\}\\
&\le \max_{q\in\Delta([K])}\left\{\sqrt{\frac{K}{\gamma}} + \frac{1}{\gamma}\sum_{i=1}^K \frac{2\left(q_i^2 + \frac{1}{\gamma K}\right)(1-q_i) + q_i\left((1-q_i) + \sqrt{\frac{K}{\gamma}}q_i\right)^2}{\sum_{i'=1}^K p_{i'} g_t(x_t,i',i)}\right\}\\
&\overset{(iii)}{\le} \max_{q\in\Delta([K])}\left\{\sqrt{\frac{K}{\gamma}} + \frac{2\sqrt{K}}{\gamma^{1.5}} + \frac{1}{\gamma}\sum_{i=1}^K \frac{2q_i^2(1-q_i) + 2q_i(1-q_i)^2 + \frac{2Kq_i^3}{\gamma}}{\sum_{i'=1}^K p_{i'} g(x,i',i)}\right\}\\
&= \max_{q\in\Delta([K])}\left\{\sqrt{\frac{K}{\gamma}} + \frac{2\sqrt{K}}{\gamma^{1.5}} + \frac{2}{\gamma}\sum_{i=1}^K \frac{q_i(1-q_i)}{\sum_{i'=1}^K p_{i'} g(x,i',i)} + \frac{2K}{\gamma^2}\sum_{i=1}^K \frac{q_i^3}{\sum_{i'=1}^K p_{i'} g(x,i',i)}\right\}\\
&\overset{(iv)}{=} \max_{q\in\Delta([K])}\left\{\sqrt{\frac{K}{\gamma}} + \frac{2\sqrt{K}}{\gamma^{1.5}} + \frac{2}{\gamma}\sum_{i=1}^K \frac{q_i(1-q_i)}{\mathbb{E}_{G\sim\mu}\left[\sum_{i'=1}^K p_{i'} G_{i',i}\right]} + \frac{2K}{\gamma^2}\sum_{i=1}^K \frac{q_i^3}{\mathbb{E}_{G\sim\mu}\left[\sum_{i'=1}^K p_{i'} G_{i',i}\right]}\right\},
\end{aligned} \tag{24}$$

where (i) is by replacing p_i with $(1-\sqrt{\frac{K}{\gamma}})q_i + \sqrt{\frac{1}{\gamma K}}$ (except for the $p_{i'}$ in the denominators); (ii) holds since $f(x,i) \in [0,1]$ for all $i \in [K]$ and $(a+b)^2 \le 2(a^2+b^2)$, and we drop the last $-\frac{1}{\sqrt{\gamma K}}$ term; (iii) is because $\sum_{i'=1}^K p_{i'} g(x,i',i) \ge \frac{1}{\sqrt{\gamma K}}$ for all $i \in [K]$ since $p_{i'} \ge \frac{1}{\sqrt{\gamma K}}$ for all $i' \in [K]$ and $\sum_{i'=1}^K g(x,i',i) \ge 1$ as $g(x,\cdot,\cdot)$ is the mean graph of a distribution of weakly observable graphs; (iv) is by definition of $\mathcal{Q}'(g,x)$ and with an abuse of notation, $G_{i,j}$ represents the (i,j)-th entry of the adjacent matrix of G.

Since G is a weakly observable graph, we know that for each $i \in [K]$, $\sum_{i'=1}^{K} p_{i'} G_{i',i} \geq \min_{i' \in [K]} p_{i'} \geq \frac{1}{\sqrt{\gamma K}}$. Therefore, we know that

$$\begin{aligned}\mathsf{dec}_\gamma(p; f, g, x) &\leq \sqrt{\frac{K}{\gamma}} + \frac{2\sqrt{K}}{\gamma^{1.5}} + 2\sqrt{\frac{K}{\gamma}} \sum_{i=1}^{K} q_i(1-q_i) + \frac{2K^{1.5}}{\gamma^{1.5}} \\ &\leq 7\sqrt{\frac{K}{\gamma}}.\end{aligned}$$

Further picking $\gamma = \max\{K, K^{\frac{1}{3}} T^{\frac{2}{3}} (\max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\})^{-\frac{2}{3}}\}$ and applying Theorem 4.1 finish the proof. □

Next, we show that under certain cases, our algorithm is able to achieve even better regret. Specifically, for a mean graph $g \in \mathcal{G}$, we define the following condition:

Condition D.2. For each $x \in \mathcal{X}$, there exists a distribution $\mu(g, x) \in \mathcal{Q}'(g, x)$ with mean $g(x, \cdot, \cdot)$ such that every graph on the support of $\mu(g, x)$ has the same weakly dominating set $D(g, x)$ with size $d(g, x)$.

One example where Condition D.2 holds is when all the possible graphs have the same revealing action which can observe the loss of all the nodes. In this case, $d(g, x) = 1$ for all g and x. In the following, we prove two theorems showing that a better regret guarantee is achievable under partially revealed and fully revealed feedback graphs respectively.

Theorem D.3. *Suppose that Condition D.2 is satisfied for all $g \in \mathcal{G}$. Define $d(\mathcal{G}) = \max_{x \in \mathcal{X}, g \in \mathcal{G}} d(g, x)$. Under Assumptions 2.1-2.4 with partially revealed weakly observable feedback graphs,* SquareCB.UG *with $\gamma = \max\{16K^2, d(\mathcal{G})^{\frac{1}{3}} T^{\frac{2}{3}} (\max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\})^{-\frac{2}{3}}\}$ guarantees:*

$$\mathbf{Reg}_{\mathsf{CB}} = \mathcal{O}\left(d(\mathcal{G})^{\frac{1}{3}} T^{\frac{2}{3}} (\max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\})^{\frac{1}{3}}\right).$$

Proof. Since Theorem 4.1 and Lemma 4.2 still hold for weakly observable graphs, we only need to prove that $\mathsf{dec}_\gamma(p; f, g, x) \leq \mathcal{O}\left(\sqrt{\frac{d(g,x)}{\gamma}}\right)$ for all $x \in \mathcal{X}$, $g \in \mathrm{conv}(\mathcal{G})$, and $f \in \mathrm{conv}(\mathcal{F})$. Using Eq. (13), Eq. (24), and picking

$$p_a = \left(1 - \sqrt{\frac{d(g,x)}{\gamma}} - \frac{1}{\gamma}\right) q_a + \frac{1}{\sqrt{\gamma d(g,x)}} \cdot \mathbb{1}\{a \in D(g,x)\} + \frac{1}{\gamma K}, \tag{25}$$

for all $a \in [K]$ with $\gamma \geq 16K^2 \geq 16 d(g,x)$, we know that $p_a \geq \frac{1}{2} q_a$ for all $a \in [K]$ and

$$\begin{aligned}
&\mathsf{dec}_\gamma(p; f, g, x) \\
&\overset{(A)}{\leq} \max_{q \in \Delta([K])} \min_{p \in \Delta([K])} \left\{ \sum_{i=1}^{K} p_i f_t(x,i) - \sum_{i=1}^{K} q_i f(x,i) + \frac{1}{\gamma} \sum_{i=1}^{K} \frac{(1-q_i) p_i^2}{\sum_{i'=1}^{K} p_{i'} g(x,i',i)} + \frac{1}{\gamma} \sum_{i=1}^{K} \frac{q_i (1-p_i)^2}{\sum_{i'=1}^{K} p_{i'} g(x,i',i)} \right\} \\
&\overset{(B)}{\leq} \max_{q \in \Delta([K])} \left\{ \left(-\sqrt{\frac{d(g,x)}{\gamma}} - \frac{1}{\gamma}\right) \sum_{i=1}^{K} q_i f(x,i) + \frac{1}{\sqrt{\gamma d(g,x)}} \sum_{i \in D(g,x)} f(x,i) + \frac{1}{\gamma K} \sum_{i=1}^{K} f(x,i) \right. \\
&\quad + \frac{1}{\gamma} \sum_{i=1}^{K} \frac{(1-q_i) \left(\left(1 - \sqrt{\frac{d(g,x)}{\gamma}} - \frac{1}{\gamma}\right) q_i + \frac{1}{\sqrt{\gamma d(g,x)}} \mathbb{1}\{i \in D(g,x)\} + \frac{1}{\gamma K} \right)^2}{\sum_{i'=1}^{K} p_{i'} g(x,i',i)} \\
&\quad \left. + \frac{1}{\gamma} \frac{q_i \left(1 - \left(1 - \sqrt{\frac{d(g,x)}{\gamma}} - \frac{1}{\gamma}\right) q_i - \frac{1}{\sqrt{\gamma d(g,x)}} \mathbb{1}\{i \in D(g,x)\} - \frac{1}{\gamma K}\right)^2}{\sum_{i'=1}^{K} p_{i'} g(x,i',i)} \right\}
\end{aligned}$$

$$\overset{(C)}{\leq} \max_{q\in\Delta([K])} \left\{ \sqrt{\frac{d(g,x)}{\gamma}} + \frac{1}{\gamma} + \frac{1}{\gamma}\sum_{i=1}^{K} \frac{3\left(\left(1-\sqrt{\frac{d(g,x)}{\gamma}}-\frac{1}{\gamma}\right)^2 q_i^2 + \frac{\mathbb{1}\{i\in D(g,x)\}}{\gamma d(g,x)} + \frac{1}{\gamma^2K^2}\right)(1-q_i)}{\sum_{i'=1}^{K} p_{i'} g(x_t,i',i)} \right.$$
$$\left. + \frac{1}{\gamma}\frac{q_i\left(1-\left(1-\sqrt{\frac{d(g,x)}{\gamma}}-\frac{1}{\gamma}\right)q_i\right)^2}{\sum_{i'=1}^{K} p_{i'} g(x_t,i',i)} \right\}$$

$$\overset{(D)}{\leq} \max_{q\in\Delta([K])} \left\{ \sqrt{\frac{d(g,x)}{\gamma}} + \frac{1}{\gamma} + \frac{1}{\gamma}\sum_{i=1}^{K} \frac{3\left(q_i^2 + \frac{\mathbb{1}\{i\in D(g,x)\}}{\gamma d(g,x)} + \frac{1}{\gamma^2K^2}\right)(1-q_i) + 3q_i\left((1-q_i)^2 + \frac{d(g,x)}{\gamma}q_i^2 + \frac{q_i^2}{\gamma^2}\right)}{\sum_{i'=1}^{K} p_{i'} g(x_t,i',i)} \right\}$$

$$\overset{(E)}{\leq} \max_{q\in\Delta([K])} \left\{ \sqrt{\frac{d(g,x)}{\gamma}} + \frac{1}{\gamma} + \frac{3}{\gamma^2} + \frac{1}{\gamma}\sum_{i=1}^{K} \frac{3q_i(1-q_i)}{\sum_{i'=1}^{K} p_{i'} g(x_t,i',i)} + \frac{1}{\gamma^2}\sum_{i=1}^{K} \frac{3\left(d(g,x)q_i^3 + \frac{q_i^3}{\gamma}\right)}{\sum_{i'=1}^{K} p_{i'} g(x_t,i',i)} \right.$$
$$\left. + \frac{1}{\gamma}\sum_{i=1}^{K} \frac{3\cdot\mathbb{1}\{i\in D(g,x)\}}{\gamma d(g,x)\sum_{i'=1}^{K} p_{i'} g_t(x_t,i',i)} \right\}$$

$$\overset{(F)}{\leq} \max_{q\in\Delta([K])} \left\{ \sqrt{\frac{d(g,x)}{\gamma}} + \frac{4}{\gamma} + \frac{3}{\gamma}\sum_{i=1}^{K} \frac{q_i(1-q_i)}{\sum_{i'=1}^{K} p_{i'} g(x_t,i',i)} + \frac{6}{\gamma^2}\sum_{i=1}^{K} \frac{d(g,x)q_i^3}{\sum_{i'=1}^{K} p_{i'} g(x_t,i',i)} \right.$$
$$\left. + \frac{1}{\gamma}\sum_{i=1}^{K} \frac{3\cdot\mathbb{1}\{i\in D(g,x)\}}{\gamma d(g,x)\sum_{i'=1}^{K} p_{i'} g(x_t,i',i)} \right\}$$

$$\overset{(G)}{=} \max_{q\in\Delta([K])} \left\{ \sqrt{\frac{d(g,x)}{\gamma}} + \frac{4}{\gamma} + \frac{3}{\gamma}\sum_{i=1}^{K} \frac{q_i(1-q_i)}{\mathbb{E}_{G\sim\mu(g,x)}\left[\sum_{i'=1}^{K} p_{i'} G_{i',i}\right]} + \frac{6d(g,x)}{\gamma^2}\sum_{i=1}^{K} \frac{q_i^3}{\mathbb{E}_{G\sim\mu(g,x)}\left[\sum_{i'=1}^{K} p_{i'} G_{i',i}\right]} \right.$$
$$\left. + \frac{1}{\gamma}\sum_{i=1}^{K} \frac{3\cdot\mathbb{1}\{i\in D(g,x)\}}{\gamma d(g,x)\mathbb{E}_{G\sim\mu(g,x)}\left[\sum_{i'=1}^{K} p_{i'} G_{i',i}\right]} \right\}$$

$$\overset{(H)}{\leq} \max_{q\in\Delta([K])} \left\{ \mathbb{E}_{G\sim\mu(g,x)}\left[\sqrt{\frac{d(g,x)}{\gamma}} + \frac{4}{\gamma} + \frac{3}{\gamma}\sum_{i=1}^{K} \frac{q_i(1-q_i)}{\sum_{i'=1}^{K} p_{i'} G_{i',i}} + \frac{6d(g,x)}{\gamma^2}\sum_{i=1}^{K} \frac{q_i^3}{\sum_{i'=1}^{K} p_{i'} G_{i',i}} \right.\right.$$
$$\left.\left. + \frac{3}{\gamma^2 d(g,x)}\sum_{i=1}^{K} \frac{\mathbb{1}\{i\in D(g,x)\}}{\sum_{i'=1}^{K} p_{i'} G_{i',i}} \right] \right\}$$

$$\overset{(I)}{\leq} \mathbb{E}_{G\sim\mu(g,x)}\left[\max_{q\in\Delta([K])} \left\{ \sqrt{\frac{d(g,x)}{\gamma}} + \frac{4}{\gamma} + \frac{3}{\gamma}\sum_{i=1}^{K} \frac{q_i(1-q_i)}{\sum_{i'=1}^{K} p_{i'} G_{i',i}} + \frac{6d(g,x)}{\gamma^2}\sum_{i=1}^{K} \frac{q_i^3}{\sum_{i'=1}^{K} p_{i'} G_{i',i}} \right.\right.$$
$$\left.\left. + \frac{3}{\gamma^2 d(g,x)}\sum_{i=1}^{K} \frac{\mathbb{1}\{i\in D(g,x)\}}{\sum_{i'=1}^{K} p_{i'} G_{i',i}} \right\} \right],$$

$$\overset{(J)}{\leq} \mathbb{E}_{G\sim\mu(g,x)}\left[\max_{q\in\Delta([K])} \left\{ \sqrt{\frac{d(g,x)}{\gamma}} + \frac{4}{\gamma} + \frac{3\sqrt{d(g,x)}}{\gamma^{1.5}} + \frac{3}{\gamma}\sum_{i=1}^{K} \frac{q_i(1-q_i)}{\sum_{i'=1}^{K} p_{i'} G_{i',i}} + \frac{6d(g,x)}{\gamma^2}\sum_{i=1}^{K} \frac{q_i^3}{\sum_{i'=1}^{K} p_{i'} G_{i',i}} \right\} \right]$$
$$\leq \mathbb{E}_{G\sim\mu(g,x)}\left[\max_{q\in\Delta([K])} \left\{ 4\sqrt{\frac{d(g,x)}{\gamma}} + \frac{4}{\gamma} + \frac{3}{\gamma}\sum_{i=1}^{K} \frac{q_i(1-q_i)}{\sum_{i'=1}^{K} p_{i'} G_{i',i}} + \frac{6d(g,x)}{\gamma^2}\sum_{i=1}^{K} \frac{q_i^3}{\sum_{i'=1}^{K} p_{i'} G_{i',i}} \right\} \right],$$

where (A) uses Eq. (13); (B) replaces is by replacing p_a by the form in Eq. (25); (C) and (D) use $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$; (E) uses the fact that for each $i \in [K]$, $\sum_{i'=1}^{K} p_{i'} g(x_t,i',i) \geq \frac{1}{\gamma K}$; (F) uses $\gamma \geq 1$; (G) holds by the

definition of $\mu(g,x)$; (H) and (I) are by Jensen's inequality; (J) holds since $D(g,x)$ is a weakly dominating set for G according to Condition D.2. Let S be the set of nodes that have self-loops in G. Then, for any $i \in D(g,x)$, if $i \in S$, $\sum_{i'=1}^K p_{i'} G_{i',i} \geq p_i \geq \frac{1}{\sqrt{\gamma d(g,x)}}$; otherwise $\sum_{i'=1}^K p_{i'} G_{i',i} \geq \frac{1}{\sqrt{\gamma d(g,x)}}$ according to the definition of a weakly domination set and the construction of p.

To analyze the third and the fourth term, we consider two cases. For $i \in S$, we know that $\sum_{i'=1}^K p_{i'} G_{i',i} \geq p_i$ and $q_i \leq 2p_i$, meaning that

$$\sum_{i \in S} \frac{q_i(1-q_i)}{\sum_{i'=1}^K p_{i'} G_{i',i}} \leq 2K,$$
$$\sum_{i \in S} \frac{q_i^3}{\sum_{i'=1}^K p_{i'} G_{i',i}} \leq 1.$$

For $i \notin S$, we know that $\sum_{i'=1}^K p_{i'} G_{i',i} \geq \frac{1}{\sqrt{\gamma d(g,x)}}$, meaning that

$$\sum_{i \notin S} \frac{q_i(1-q_i)}{\sum_{i'=1}^K p_{i'} G_{i',i}} \leq \sqrt{\gamma d(g,x)},$$
$$\sum_{i \notin S} \frac{q_i^3}{\sum_{i'=1}^K p_{i'} G_{i',i}} \leq 1.$$

Combining the above two cases together, we know that

$$\mathsf{dec}_\gamma(p; f, g, x) \leq \mathcal{O}\left(\frac{K}{\gamma} + \sqrt{\frac{d(g,x)}{\gamma}}\right) \leq \mathcal{O}\left(\sqrt{\frac{d(g,x)}{\gamma}}\right), \tag{26}$$

where the second inequality is due to $\gamma \geq 16K^2$. This finishes the proof. □

Theorem D.4. *Suppose that Condition D.2 is satisfied for $g^\star \in \mathcal{G}$ and define $d_t = d(g^\star, x_t)$. Under Assumptions 2.1-2.4 with fully revealed weakly observable feedback graphs,* SquareCB.UG *with an appropriate choice of γ guarantees:*

$$\mathbf{Reg}_{\mathsf{CB}} = \mathcal{O}\left(\left(\sum_{t=1}^T \sqrt{d_t}\right)^{\frac{2}{3}} (\max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\})^{\frac{1}{3}}\right).$$

Proof. Since Lemma 2.5, Theorem 4.1, Lemma 4.2, and Lemma 4.5 still hold for weakly observable graphs, we know that $\mathbf{Reg}_{\mathsf{CB}}$ is upper bounded by

$$\mathbf{Reg}_{\mathsf{CB}} \leq \mathcal{O}\left(\sum_{t=1}^T \min_p \mathsf{dec}_{\frac{\gamma}{3}}(p; f_t, g^\star, x_t)\right) + \frac{3}{4}\gamma \mathbf{Reg}_{\mathsf{Sq}} + \frac{2}{3}\gamma \mathbf{Reg}_{\mathsf{Log}}.$$

Further applying Eq. (26) shows that

$$\mathbf{Reg}_{\mathsf{CB}} \leq \mathcal{O}\left(\sum_{t=1}^T \sqrt{\frac{d_t}{\gamma}} + \gamma \max\{\mathbf{Reg}_{\mathsf{Sq}}, \mathbf{Reg}_{\mathsf{Log}}\}\right).$$

Picking the optimal γ finishes the proof. □