Proceedings of Machine Learning Research vol 201:1–32, 2023 34th International Conference on Algorithmic Learning Theory

Best-of-Both-Worlds Algorithms for Partial Monitoring

Taira Tsuchiya TSUCHIYA@SYS.I.KYOTO-U.AC.JP
Kyoto University and RIKEN AIP

Shinji Ito I-SHINJI@NEC.COM
NEC Corporation

Junya Honda HONDA@I.KYOTO-U.AC.JP
Kyoto University and RIKEN AIP

Editors: Shipra Agrawal and Francesco Orabona

Abstract

This study considers the partial monitoring problem with k-actions and d-outcomes and provides the first best-of-both-worlds algorithms, whose regrets are favorably bounded both in the stochastic and adversarial regimes. In particular, we show that for non-degenerate locally observable games, the regret is $O(m^2 k^4 \log(T) \log(k_\Pi T)/\Delta_{\min})$ in the stochastic regime and $O(mk^{3/2}\sqrt{T \log(T) \log k_\Pi})$ in the adversarial regime, where T is the number of rounds, m is the maximum number of distinct observations per action, $\Delta_{\min}$ is the minimum suboptimality gap, and k_Π is the number of Pareto optimal actions. Moreover, we show that for globally observable games, the regret is $O(c_{\mathcal{G}}^2 \log(T) \log(k_\Pi T)/\Delta_{\min}^2)$ in the stochastic regime and $O((c_{\mathcal{G}}^2 \log(T) \log(k_\Pi T))^{1/3} T^{2/3})$ in the adversarial regime, where $c_{\mathcal{G}}$ is a game-dependent constant. We also provide regret bounds for a stochastic regime with adversarial corruptions. Our algorithms are based on the follow-the-regularized-leader framework and are inspired by the approach of exploration by optimization and the adaptive learning rate in the field of online learning with feedback graphs.

Keywords: partial monitoring, best-of-both-worlds, follow-the-regularized-leader, stochastic regime with adversarial corruptions

1. Introduction

Partial monitoring (PM) is a general sequential decision-making problem with limited feedback, which can be seen as a generalization of the bandit problem. A PM game $\mathcal{G} = (\mathcal{L}, \Phi)$ is defined by the pair of a loss matrix $\mathcal{L} \in [0,1]^{k \times d}$ and feedback matrix $\Phi \in \Sigma^{k \times d}$, where k is the number of actions, d is the number of outcomes, and Σ is a set of feedback symbols. The game is sequentially played by a learner and opponent for $T \geq 3$ rounds. At the beginning of the game, the learner observes $\mathcal{L}$ and Φ. At every round $t \in [T]$, the opponent chooses an outcome $x_t \in [d]$, and then the learner chooses an action $A_t \in [k]$, suffers an unobserved loss $\mathcal{L}_{A_t x_t}$, and receives a feedback symbol $\sigma_t = \Phi_{A_t x_t}$, where $\mathcal{L}_{ax}$ is the (a,x)-th element of $\mathcal{L}$. In general, the learner cannot directly observe the outcome and loss, and can only observe the feedback symbol. The learner's goal is to minimize their cumulative loss over all rounds. The performance of the learner is evaluated by the regret R_T, which is defined as the difference between the cumulative loss of the learner and the single optimal action a^* fixed in hindsight, that is, $a^* = \arg\min_{a \in [k]} \mathbb{E}\bigl[\sum_{t=1}^T \mathcal{L}_{a x_t}\bigr]$ and $R_T = \mathbb{E}\bigl[\sum_{t=1}^T \bigl(\mathcal{L}_{A_t x_t} - \mathcal{L}_{a^* x_t}\bigr)\bigr] = \mathbb{E}\bigl[\sum_{t=1}^T \langle \ell_{A_t} - \ell_{a^*}, e_{x_t} \rangle\bigr]$, where $\ell_a \in \mathbb{R}^d$ is the a-th row of $\mathcal{L}$, and $e_x \in \{0,1\}^d$ is the x-th orthonormal basis of $\mathbb{R}^d$.

PM has been investigated in two regimes: the *stochastic* and *adversarial* regimes. In the stochastic regime, outcomes $(x_t)_{t=1}^T$ are sampled from a fixed distribution ν^* in an i.i.d. manner, whereas in the adversarial regime, the outcomes are arbitrarily decided from the set of outcomes $[d]$ possibly depending on the history of the actions $(A_s)_{s=1}^{t-1}$.

Some of the first investigations on PM originate from work by Rustichini (1999); Piccolboni and Schindelhauer (2001). The seminal work was conducted by Cesa-Bianchi et al. (2006); Bartók et al. (2011), the latter of which showed that all PM games can be classified into four classes based on their minimax regrets. They classified PM games into trivial, easy, hard, and hopeless games, for which their minimax regrets are 0, $\tilde{\Theta}(\sqrt{T})$, $\Theta(T^{2/3})$, and $\Theta(T)$, respectively. The easy and hard games are also called *locally observable* and *globally observable* games, respectively.

PM algorithms have been established for both the stochastic and adversarial regimes. In the adversarial regime, the most common form of algorithms is an *Exp3-type* one (Freund and Schapire, 1997; Auer et al., 2002). Recently, Lattimore and Szepesvári (2020b) showed that an Exp3-type algorithm with the approach of *exploration by optimization* obtains the aforementioned minimax bounds. Notably, they proved the regret bounds of $O(mk^{3/2}\sqrt{T \log k})$ for non-degenerate locally observable games, and $O((c_{\mathcal{G}}T)^{2/3}(\log k)^{1/3})$ for globally observable games, where $m \leq \min\{|\Sigma|, d\}$ is the maximum number of distinct observations per action and $c_{\mathcal{G}}$ is a game-dependent constant defined in Section 5. PM has also been investigated in the stochastic regime and some algorithms exploiting the stochastic structure of the problem can achieve $O(\log T)$ regret bounds (Vanchinathan et al., 2014; Komiyama et al., 2015; Tsuchiya et al., 2020).

Algorithms assuming the stochastic model for losses can suffer linear regret in the adversarial regime, whereas algorithms for the adversarial regime tend to perform poorly in the stochastic regime. Since knowing the underlying regime is difficult in practice, obtaining favorable performance for both the stochastic and adversarial regimes *without* knowing the underlying regime is desirable.

To achieve this goal, particularly in the classical multi-armed bandits, the Best-of-Both-Worlds (BOBW) algorithms that perform well in both stochastic and adversarial regimes have been developed. The first BOBW algorithm was developed in a seminal paper by Bubeck and Slivkins (2012), and the celebrated Tsallis-INF algorithm was recently proposed by Zimmert and Seldin (2021). BOBW algorithms have also been developed beyond the multi-armed bandits (*e.g.,* Gaillard et al. 2014; Luo and Schapire 2015; Erez and Koren 2021; Zimmert et al. 2019; Lee et al. 2021; Jin and Luo 2020; Huang et al. 2022; Saha and Gaillard 2022), whereas they have never been investigated in PM.

Some BOBW algorithms are known to perform well also in the *stochastic regime with adversarial corruptions* (Lykouris et al., 2018), which is an intermediate regime between the stochastic and adversarial regimes. This regime is advantageous in practice, since the stochastic assumption on outcomes is too strong whereas the adversarial assumption is too pessimistic. Therefore it is also practically important to develop BOBW algorithms that cover this intermediate regime.

1.1. Contribution of This Study

This study establishes new BOBW algorithms for PM based on the Follow-the-Regularized-Leader (FTRL) framework (McMahan, 2011). We rely on two recent theoretical advances: (i) the Exp3-type algorithm for PM developed with the approach of exploration by optimization (Lattimore and Szepesvári, 2020b) and (ii) the adaptive learning rate for online learning with feedback graphs (Ito et al., 2022b), for which BOBW algorithms have been developed (Erez and Koren, 2021; Ito et al.,

Table 1: Regret upper bounds for PM. The constant $C \geq 0$ is the corruption level, and $\mathcal{R}^{\mathrm{loc}}$ and $\mathcal{R}^{\mathrm{glo}}$ are the regret upper bounds of the proposed algorithm in the stochastic regime for locally and globally games, respectively. "observ." means observability. TSPM is the bound by Tsuchiya et al. (2020); refer to the paper for the definition of Λ'. ExpPM is by Lattimore and Szepesvári (2020b). PM-DEMD is by Komiyama et al. (2015), and $D(\nu^*)$ is a distribution-dependent constant.

observ.	algorithm	stochastic (stoc.)	adversarial	stoc. w/ corruptions
locally	TSPM	$O\left(\frac{mk^2 d\log(T)}{\Lambda'^2}\right)$	–	–
obs.	ExpPM	–	$O(mk^{3/2}\sqrt{T\log k})$	–
	Proposed	$O\left(\frac{m^2k^4\log(T)\log(k_\Pi T)}{\Delta_{\min}}\right)$	$O(mk^{3/2}\sqrt{T\log(T)\log k_\Pi})$	$\mathcal{R}^{\mathrm{loc}} + \sqrt{C\mathcal{R}^{\mathrm{loc}}}$
globally	PM-DMED	$O(D(\nu^*)\log T)$	–	–
obs.	ExpPM	–	$O((c_{\mathcal{G}}T)^{2/3}(\log k)^{1/3})$	–
	Proposed	$O\left(\frac{c_{\mathcal{G}}^2\log(T)\log(k_\Pi T)}{\Delta_{\min}^2}\right)$	$O((c_{\mathcal{G}}T)^{2/3}(\log(T)\log(k_\Pi T))^{1/3})$	$\mathcal{R}^{\mathrm{glo}} + (C^2\mathcal{R}^{\mathrm{glo}})^{1/3}$

2022a; Rouyer et al., 2022; Kong et al., 2022). Note that it is known that the FTRL with the (negative) Shannon entropy regularizer corresponds to the Exp3 algorithm.

The regret bounds of the proposed algorithms are as follows. We define the number of Pareto optimal actions by $k_\Pi \leq k$, and the minimum suboptimality gap by $\Delta_{\min} = \min_{a\in[k]\setminus\{a^*\}} \Delta_a$, where $\Delta_a = (\ell_a - \ell_{a^*})^\top \nu^* \geq 0$ for $a \in [k]$ is the loss gap between action a and optimal action a^*. We show that for non-degenerate locally observable games, the regret is $O(m^2k^4\log(T)\log(k_\Pi T)/\Delta_{\min})$ in the stochastic regime and $O(mk^{3/2}\sqrt{T\log(T)\log k_\Pi})$ in the adversarial regime. We also show that for globally observable games, the regret is $O(c_{\mathcal{G}}^2\log(T)\log(k_\Pi T)/\Delta_{\min}^2)$ in the stochastic regime and $O((c_{\mathcal{G}}T)^{2/3}(\log(T)\log(k_\Pi T))^{1/3})$ in the adversarial regime. In addition, we also consider some intermediate regimes, such as the stochastic regime with adversarial corruptions (Lykouris et al., 2018), which we define in PM based on the corruptions on outcomes. To our knowledge, the proposed algorithms are the first BOBW algorithms for PM. Table 1 lists the regret bounds provided in this study and summarizes comparisons with existing work. Our algorithm is not the best in the strict sense. For example in the stochastic regime, compared to Komiyama et al. (2015), the dependence on T of their bound is $\log T$, whereas that of ours is $(\log T)^2$. Nevertheless, this kind of looseness often appears in the BOBW literature (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Seldin and Lugosi, 2017; Ito et al., 2022a) and it is an important future work to close this gap as was done by Zimmert and Seldin (2021) in the case of multi-armed bandits.

1.2. Technical Summary

For locally observable games, we develop the algorithm based on the approach of *exploration by optimization* (Lattimore and Szepesvári, 2020b) with the Shannon entropy regularizer. This approach is promising especially in locally observable games for bounding a component of regret, in which we consider a certain optimization problem with respect to the action selection probability. To obtain BOBW guarantees, we consider using a self-bounding technique (Zimmert and Seldin, 2021). In the self-bounding technique, we first derive upper and lower bounds of regret using a random variable depending on the action selection probability, and then derive a regret bound by combining the upper and lower bounds. However, using the exploration by optimization may make some action selection probabilities extremely small, preventing deriving a meaningful lower bound. To handle this problem, we consider an optimization over a *restricted* feasible set. This restriction enables us to lower bound the regret such that the self-bounding technique is applicable, and we show that even with the optimization over the restricted feasible set, the component of regret is favorably bounded. In addition, we consider the upper truncation of the learning rate developed by Ito et al. (2022a) to collaborate with the theory of exploration by optimization.

For globally observable games, we develop the algorithm using the Shannon entropy regularizer as for locally observable games. To derive BOBW guarantees, we use the technique of adaptive learning rate developed in online learning with feedback graphs by Ito et al. (2022a), but in a modified way. Their work uses a regularization called hybrid regularizers, which combines a Shannon entropy of the compensation of the action selection probability with typical regularizers (Zimmert et al., 2019; Ito et al., 2022b,a). We think that naively applying this regularization also yields BOBW guarantees, but it loses the closed form of the action selection probability in FTRL updates and requires solving an optimization problem each round. This study shows that we can obtain the BOBW guarantee even only with the standard Shannon entropy regularization, and consequently, the proposed algorithm does not need to solve the optimization problem every round and can be implemented efficiently.

1.3. Related Work

In the adversarial regime, FeedExp3 is a first Exp3-type algorithm, which has a first non-asymptotic regret bound (Piccolboni and Schindelhauer, 2001) and is known to achieve a minimax regret of $O(T^{2/3})$ (Cesa-Bianchi et al., 2006). Since then, Exp3-type algorithms have been used in many contexts. Bartók (2013) relied on an Exp3-type algorithm as a subroutine of their algorithm. Lattimore and Szepesvári (2019a) showed that for a variant of the locally observable game (point-locally observable games), an Exp3-type algorithm achieves an $O(\sqrt{T})$ regret. Recently, Lattimore and Szepesvári (2020b) showed that an Exp3-type algorithm using exploration by optimization can obtain bounds with good leading constants for both easy and hard games. There are also a few algorithms that are not Exp3-type (Bartók et al., 2011; Foster and Rakhlin, 2012).

PM has also been investigated in the stochastic regime, although less extensively than the adversarial regime (Bartók et al., 2012). One study (Komiyama et al., 2015) is based on DMED (Honda and Takemura, 2011), in which the algorithm heavily exploits the stochastic structure, and the algorithm was shown to achieve an $O(\log T)$ regret with a distribution-optimal constant factor for globally observable games. Two other approaches (Vanchinathan et al., 2014; Tsuchiya et al., 2020) are based on Thompson sampling (Thompson, 1933). They focus on another variant of locally observable games (strongly locally observable games), and the algorithms presented a strong empirical performance in the stochastic regime with an $O(\log T)$ regret bound (Tsuchiya et al., 2020).

It is worth noting that PM has been studied in a variety of contexts with somewhat different settings, *e.g.,* with feedback graphs (Alon et al., 2015) or with linear feedback (Lin et al., 2014). While our focus in this paper is the locally and globally observable games, there has been some literature for hopeless games; we basically cannot do anything with the current definition of the regret, but some research has been done by modifying the definition of the regret (Rustichini, 1999; Mannor and Shimkin, 2003; Perchet, 2011; Mannor et al., 2014).

2. Background

Notation Let $\|x\|$, $\|x\|_1$, and $\|x\|_\infty$ be the Euclidian, ℓ_1-, and ℓ_∞-norms for a vector x respectively, and $\|A\|_\infty = \max_{i,j}|A_{ij}|$ be the maximum norm for a matrix A. Let $\mathcal{P}_k = \{p \in [0,1]^k : \|p\|_1 = 1\}$ be the $(k-1)$-dimensional probability simplex. A vector $e_a \in \{0,1\}^k$ is the a-th orthonormal basis of $\mathbb{R}^k$, and $\mathbf{1}$ is the all-one vector.

Partial Monitoring Consider any PM game $\mathcal{G} = (\mathcal{L}, \Phi)$. Let $m \le |\Sigma|$ be the maximum number of distinct symbols in a single row of $\Phi \in \Sigma^{k\times d}$ over all rows. In the following, we introduce several concepts in PM. Different actions a and b are *duplicate* if $\ell_a = \ell_b$. We can decompose possible distributions of d outcomes in $\mathcal{P}_d$ based on the loss matrix: for every action $a \in [k]$, *cell* $\mathcal{C}_a = \{u \in \mathcal{P}_d : \max_{b\in[k]}(\ell_a - \ell_b)^\top u \le 0\}$ is the set of probability vectors in $\mathcal{P}_d$ for which action a is optimal. Each cell is a convex closed polytope. Let $\dim(\mathcal{C}_a)$ be the dimension of the affine hull of $\mathcal{C}_a$. If $\mathcal{C}_a = \emptyset$, action a is *dominated*. For non-dominated actions, if $\dim(\mathcal{C}_a) = d-1$ then action a is *Pareto optimal*, and if $\dim(\mathcal{C}_a) < d-1$ then action a is *degenerate*. We denote the set of Pareto optimal actions by Π, and the number of Pareto optimal actions by $k_\Pi = |\Pi|$. Two Pareto optimal actions $a, b \in \Pi$ are *neighbors* if $\dim(\mathcal{C}_a \cap \mathcal{C}_b) = d-2$, and this notion is used to define the difficulty of PM games. It is known that the undirected graph induced by the above neighborhood relations is connected (see *e.g.,* Bartók et al. 2012, Lattimore and Szepesvári 2020a, Lemma 37.7), and this is useful for loss difference estimations between distinct Pareto optimal actions. A PM game is called non-degenerate if it has no degenerate actions. An example of cell decomposition is given in Figure 1. From hereon, we assume that PM game $\mathcal{G}$ is non-degenerate and contains no duplicate actions. The following *observability* conditions characterize the difficulty of PM games.

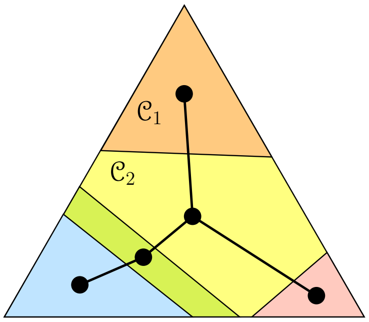

Figure 1: An example of cell decomposition. The points • correspond to Pareto-optimal actions. Cells $\mathcal{C}_1$ and $\mathcal{C}_2$ are neighbors.

Definition 1 *Neighbouring actions a and b are globally observable if there exists function $w_e : [k] \times \Sigma \to \mathbb{R}$ such that*

$$\sum_{c=1}^{k} w_e(c, \Phi_{cx}) = \mathcal{L}_{ax} - \mathcal{L}_{bx} \textit{ for all } x \in [d]\,. \tag{1}$$

Neighbouring actions a and b are locally observable if there exists $w_e = w_{ab}$ satisfying (1) and $w_e(c,\sigma) = 0$ for $c \notin \{a,b\}$. A PM game is called globally (resp. locally) observable if all neighboring actions are globally (resp. locally) observable.

It is easy to see from the above definition that any locally observable games are globally observable, and this paper assumes that $\mathcal{G}$ is globally observable.

Loss Difference Estimation Next, we introduce a method of loss difference estimations used in PM. Let $\mathcal{H}$ be the set of all functions from $[k] \times \Sigma$ to $\mathbb{R}^d$. In the following, we show that for globally observable games we can estimate loss differences between *any* Pareto optimal actions using some $G \in \mathcal{H}$ based on (1).

Lemma 2 (Lemma 4 of Lattimore and Szepesvári 2020b) *Consider any globally observable game. Then there exists a function $G \in \mathcal{H}$ such that for all $b, c \in \Pi$, we have*

$$\sum_{a=1}^{k} (G(a, \Phi_{ax})_b - G(a, \Phi_{ax})_c) = \mathcal{L}_{bx} - \mathcal{L}_{cx} \text{ for all } x \in [d] \,. \tag{2}$$

This result straightforwardly follows from the fact that the graph induced by the set of Pareto optimal actions is connected. Let $\mathscr{T}$ be a tree over Π induced by the neighborhood relations. Lattimore and Szepesvári (2020b) provides the following example of G:

$$G(a, \sigma)_b = \sum_{e \in \mathrm{path}_{\mathscr{T}}(b)} w_e(a, \sigma) \text{ for } a \in \Pi \,, \tag{3}$$

where $\mathrm{path}_{\mathscr{T}}(b)$ is the set of edges from $b \in \Pi$ to an arbitrarily chosen root $c \in \Pi$ on $\mathscr{T}$.

Intermediate Regimes between Stochastic and Adversarial Regimes Here, we discuss intermediate regimes between the stochastic and adversarial regimes: the stochastic regime with adversarial corruptions and an adversarial regime with a self-bounding constraint.

The stochastic regime with adversarial corruptions was originally considered by Lykouris et al. (2018) in the classical multi-armed bandits. We define this regime in PM by considering the corruptions on the sequence of outcomes $(x_t)_{t=1}^T$. In this regime, a temporary outcome $x'_t \in [d]$ is sampled from an unknown distribution ν^*, and the adversary then corrupts x'_t to x_t without knowing A_t. We define the corruption level by $C = \mathbb{E}\big[\sum_{t=1}^T \|\mathcal{L} e_{x_t} - \mathcal{L} e_{x'_t}\|_\infty\big] \geq 0$. If $C = 0$, this regime corresponds to the stochastic regime, and if $C \geq T$, this regime corresponds to the adversarial regime. As we will see, the proposed algorithms work without knowing the corruption level C. We also define another intermediate regime, a *stochastically constrained adversarial regime*, in Appendix A.

In this work, we consider an *adversarial regime with a self-bounding constraint*, developed in the multi-armed bandits (Zimmert and Seldin, 2021) and includes the regimes that appeared so far.

Definition 3 *Let $\Delta \in [0, 1]^k$ and $C \geq 0$. The environment is in an adversarial regime with a (Δ, C, T) self-bounding constraint if it holds for any algorithm that $R_T \geq \mathbb{E}\big[\sum_{t=1}^T \Delta_{A_t} - C\big]$.*

We can show that the regimes that have appeared so far are included in the adversarial regime with a self-bounding constraint; the details are discussed in Appendix A.

In this study, we assume that there exists a unique optimal action. This assumption has been employed by many studies aiming to develop BOBW algorithms (Gaillard et al., 2014; Luo and Schapire, 2015; Wei and Luo, 2018; Ito, 2021; Zimmert and Seldin, 2021).

3. Follow-the-Regularized-Leader

This section introduces the FTRL framework and provides some fundamental bounds used in the analysis. We recall that Π is the set of Pareto optimal actions. In the FTRL framework, a probability vector $p_t \in \mathcal{P}_k$ over the action set $[k]$ is given as

$$q_t \in \operatorname*{arg\,min}_{q \in \mathcal{P}(\Pi)} \left[\left\langle \sum_{s=1}^{t-1} \widehat{y}_s, \, q \right\rangle + \psi_t(q) \right], \quad p_t = \mathcal{T}_t(q_t) \,, \tag{4}$$

where the set $\mathcal{P}(\mathcal{B}) \coloneqq \{p \in \mathcal{P}_k : p_a = 0 \text{ for } a \notin \mathcal{B}\}$ for $\mathcal{B} \subset [k]$ is a convex closed polytope on the probability simplex with nonzero elements at indices in $\mathcal{B}$, $\widehat{y}_s \in \mathbb{R}^k$ is an estimator of the loss at round t, $\psi_t : \mathcal{P}_k \to \mathbb{R}$ is a convex regularizer, and $\mathcal{T}_t : \mathcal{P}(\Pi) \to \mathcal{P}_k$ is a map from q_t to an action selection probability vector p_t. We use the Shannon entropy for ψ_t, which is defined as

$$\psi_t(p) = \frac{1}{\eta_t} \sum_{a=1}^{k} p_a \log(p_a) = -\frac{1}{\eta_t} H(p) \,. \tag{5}$$

We can easily check that if we use the Shannon entropy with learning rate η_t, $q_t \in \mathcal{P}(\Pi)$ is expressed as

$$q_{t,a} = \frac{\mathbb{1}[a \in \Pi] \exp\left(-\eta_t \sum_{s=1}^{t-1} \widehat{y}_{sa}\right)}{\sum_{b \in \Pi} \exp\left(-\eta_t \sum_{s=1}^{t-1} \widehat{y}_{sb}\right)} \quad \text{for } a \in [k] \,. \tag{6}$$

We set an estimator to $\widehat{y}_t = G_t(A_t, \sigma_t)/p_{t,A_t}$ (Lattimore and Szepesvári, 2020b), where for locally observable games, G_t is obtained by minimizing a certain optimization problem, whereas for globally observable games G_t is set to (3). The regret analysis of FTRL boils down to the evaluation of $\sum_{t=1}^{T} \sum_{a=1}^{k} p_{t,a}(\widehat{y}_{ta} - \widehat{y}_{ta^*})$. We can decompose this quantity into

$$\begin{aligned}\sum_{t=1}^{T} \sum_{a=1}^{k} p_{t,a}(\widehat{y}_{ta} - \widehat{y}_{ta^*}) &\le \underline{\sum_{t=1}^{T} \Big(\psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1})\Big) + \psi_{T+1}(e_{a^*}) - \psi_1(q_1)} \\ &\quad + \underline{\sum_{t=1}^{T} \Big(\langle q_t - q_{t+1}, \widehat{y}_t \rangle - D_t(q_{t+1}, q_t)\Big)} + \underline{\sum_{t=1}^{T} \sum_{a=1}^{k} (q_{t,a} - p_{t,a})(\widehat{y}_{ta} - \widehat{y}_{ta^*})} \,,\end{aligned} \tag{7}$$

where the inequality follows from the standard analysis of the FTRL framework (see *e.g.,* Lattimore and Szepesvári, 2020a, Exercise 28.12), and $D_t : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_+$ is the *Bregman divergence* induced by ψ_t, *i.e.,* $D_t(p, q) = \psi_t(p) - \psi_t(q) - \langle \nabla \psi_t(q), p - q \rangle$. We refer to the terms with dashed, wavy, and straight underlines in (7) as the *penalty*, *stability*, and *transformation* terms, respectively.

We use a self-bounding technique to bound the regret in the stochastic regime, which requires a lower bound of the regret. To this end, we introduce parameters $Q(a^*)$ and $\bar{Q}(a^*)$ given by

$$Q(a^*) = \sum_{t=1}^{T} (1 - q_{t,a^*}) \quad \text{and} \quad \bar{Q}(a^*) = \mathbb{E}\left[Q(a^*)\right] \,. \tag{8}$$

Note that $0 \le \bar{Q}(a^*) \le T$ for any $a^* \in [k]$. Based on quantity $\bar{Q}(a^*)$, the regret in the adversarial regime with a self-bounding constraint can be bounded from below as follows.

Lemma 4 *In the adversarial regime with a self-bounding constraint, if there exists $c \in (0, 1]$ such that $p_{t,a} \ge c\, q_{t,a}$ for $t \in [T]$ and $a \in [k]$, the regret is bounded as $R_T \ge c\, \Delta_{\min} \bar{Q}(a^*) - C$.*

All omitted proofs are given in Appendix B. This lemma is used to derive poly-logarithmic regret bounds in the adversarial regime with a self-bounding constraint.

4. Locally Observable Case

This section provides a BOBW algorithm for locally observable games and derives its regret bounds.

4.1. Exploration by Optimization in PM

We first briefly explain the approach of exploration by optimization by Lattimore and Szepesvári (2020b), based on which our algorithm for locally observable games is developed. In locally observable games, the achievable regret is generally smaller than in globally observable games. Hence, we need to exploit this easiness to achieve small regret, for which we rely on exploration-by-optimization. Intuitively, in locally observable games, a loss estimator may suffer a large variance because an informative action might not be selected due to its large losses. To overcome this issue, Lattimore and Szepesvári (2020b) proposed exploration-by-optimization, which improves regret bound by optimizing the stability that corresponds to the variance.

The key idea behind the approach is to minimize a part of a regret upper bound of an Exp3-type algorithm (equivalently, FTRL with the Shannon entropy). In particular, they consider the optimization on variables $G : [k] \times \Sigma \to \mathbb{R}^k$ and $p \in \mathcal{P}_k$. Their algorithm computes every round the function G and the action selection probability vector p by optimizing a part of the regret upper bound of FTRL, expressed as

$$\underset{G \in \mathcal{H},\, p \in \mathcal{P}_k}{\text{minimize}} \quad \max_{x \in [d]} \left[\frac{(p-q)^\top \mathcal{L} e_x}{\eta} + \frac{\text{bias}_q(G; x)}{\eta} + \frac{1}{\eta^2} \sum_{a=1}^{k} p_a \left\langle q, \xi\left(\frac{\eta G(a, \Phi_{ax})}{p_a} \right) \right\rangle \right], \tag{9}$$

where $\xi(x) = \mathrm{e}^{-x} + x - 1$ (we abuse the notation by applying ξ in an element-wise manner), and

$$\text{bias}_q(G; x) = \left\langle q,\, \mathcal{L} e_x - \sum_{a=1}^{k} G(a, \Phi_{ax}) \right\rangle + \max_{c \in \Pi} \left(\sum_{a=1}^{k} G(a, \Phi_{ax})_c - \mathcal{L}_{cx} \right), \tag{10}$$

is the bias function. In the optimization problem (9), the first term corresponds to the transformation term, the second term corresponds to the regret for using a biased estimator, and the third term comes from a part of the stability term. Note that the bias term does not appear when G satisfies (2). Note also that the optimization problem in (9) is convex and can be solved numerically by using standard solvers as discussed in Lattimore and Szepesvári (2020b).

4.2. Proposed Algorithm

This section describes the proposed algorithm for locally observable games. Although exploration-by-optimization significantly improves the regret bound for locally observable games, they only consider the adversarial regimes, and some modification is required for making it valid also for the stochastic regime. To obtain BOBW guarantees, we often rely on a self-bounding technique, which requires a certain lower bound on the action selection probability p (Gaillard et al., 2014; Wei and Luo, 2018; Zimmert and Seldin, 2021). However, solving the optimization problem (9) may result in $p_a = 0$ for a certain $a \in [k]$, which precludes the use of the technique. The proposed algorithm considers the minimization problem over a restricted feasible set for p instead of over $\mathcal{P}_k$. Let $\mathcal{P}'_k(q)$ for $q \in \mathcal{P}(\Pi)$ be $\mathcal{P}'_k(q) = \{p \in \mathcal{P}_k : p_a \geq q_a/(2k) \text{ for all } a \in [k]\} \subset \mathcal{P}_k$. We then consider the

Algorithm 1: BOBW algorithm for locally observable games

```
input: $B$
for $t = 1, 2, \dots$ do
    Compute $\eta_t$ using (12) and $q_t$ using (6)
    Solve (11) with $\eta \leftarrow \eta_t$ and $q \leftarrow q_t$ to determine $V'_t = \max\{0, \mathrm{opt}'_{q_t}(\eta_t)\}$ and the
      corresponding solution $p_t$ and $G_t$
    Sample $A_t \sim p_t$, observe $\sigma_t \in \Sigma$, compute $\widehat{y}_t = G_t(A_t, \sigma_t)/p_{t,A_t}$, update $\beta'_t$ using (12)
```

following optimization problem:

$$\underset{G \in \mathcal{H},\, p \in \mathcal{P}'_k(q)}{\text{minimize}} \quad \max_{x \in [d]} \left[\frac{(p-q)^\top \mathcal{L} e_x}{\eta} + \frac{\mathrm{bias}_q(G; x)}{\eta} + \frac{1}{\eta^2} \sum_{a=1}^{k} p_a \left\langle q, \xi\left(\frac{\eta G(a, \Phi_{ax})}{p_a} \right) \right\rangle \right], \tag{11}$$

which implies that the solution p of the optimization problem (11) satisfies $p \geq q/(2k)$. This property is useful when applying the self-bounding technique to bound the regret in the stochastic regime (possibly with adversarial corruptions). We define the optimal value of the optimization problem (11) by $\mathrm{opt}'_q(\eta)$ and its truncation at round t by $V'_t = \max\{0, \mathrm{opt}'_{q_t}(\eta_t)\}$.

Regularizer and Learning Rate We use the Shannon entropy with learning rate η_t in (5) as a regularizer. The learning rate η_t is defined as follows. Let $\beta'_1 = c_1 \geq 1$ and

$$\beta'_{t+1} = \beta'_t + \frac{c_1}{\sqrt{1 + (\log k_\Pi)^{-1} \sum_{s=1}^{t} H(q_s)}}, \quad \beta_t = \max\left\{B, \beta'_t\right\}, \quad \text{and} \quad \eta_t = \frac{1}{\beta_t} \tag{12}$$

for $c_1 > 0$ (determined in Theorem 6). The fundamental idea of this learning rate was developed by Ito et al. (2022a), and we use its variant by the upper truncation of β'_t. The truncation is required when applying the following lemma to bound $\mathrm{opt}'_q(\eta)$.

Lemma 5 *For non-degenerate locally observable games and $\eta \leq 1/(2mk^2)$, we have*

$$\mathrm{opt}'_*(\eta) \coloneqq \sup_{q \in \mathcal{P}_k} \mathrm{opt}'_q(\eta) \leq 3m^2k^3 .$$

This lemma is a slightly stronger version of Proposition 8 of Lattimore and Szepesvári (2020b), in which the same upper bound is derived for the minimum value over larger feasible set $\mathcal{P}_k \supset \mathcal{P}'_k(q)$ in (9) instead of (11). Since the objective function of (9) and (11) originally comes from a component of the regret, this lemma means that the restriction of the feasible set does not harm the regret bound. Algorithm 1 provides the proposed algorithm for locally observable games.

4.3. Regret Analysis for Locally Observable Games

With the above algorithm, we can prove the following regret bound for locally observable games.

Theorem 6 *Consider any locally observable non-degenerate partial monitoring game. If we run Algorithm 1 with $B \geq 2mk^2$ and $c_1 = \Theta\big(mk^{3/2}\sqrt{(\log T)/(\log k_\Pi)}\big)$, we have the following* *bounds. For the adversarial regime with a (Δ, C, T) self-bounding constraint, we have*

$$R_T = O\left(\frac{m^2 k^4 \log(T) \log(k_\Pi T)}{\Delta_{\min}} + \sqrt{\frac{C m^2 k^4 \log(T) \log(k_\Pi T)}{\Delta_{\min}}} \right), \tag{13}$$

and for the adversarial regime, we have

$$R_T = O\big(m k^{3/2} \sqrt{T \log(T) \log k_\Pi}\big) + B \log k_\Pi \,.$$

Note that (13) with $C = 0$ yields the bound in the stochastic regime. The bound for the adversarial regime is a factor of $\sqrt{\log(T)\log(k_\Pi)/\log k}$ worse for large enough T than the algorithm by Lattimore and Szepesvári (2020b). This comes from the difficulty of obtaining the BOBW guarantee, where we need to aggressively change the learning rate when the environment looks not so much adversarial. Note that exactly solving the optimization problem (11) is not necessary, and we discuss regret bounds for this case in Appendix C. In the rest of this section, we provide a sketch of the analysis.

We start by decomposing the regret as follows.

Lemma 7 $R_T \le \mathbb{E}\Big[\sum_{t=1}^T \big(\eta_{t+1}^{-1} - \eta_t^{-1}\big) H(q_{t+1}) + H(q_1)/\eta_1 + \sum_{t=1}^T \eta_t V_t'\Big]$.

This can be proven by refining the analysis of the penalty term of Theorem 6 in Lattimore and Szepesvári (2020b), in which we rely on the standard analysis in (7), and the first and remaining terms correspond to the penalty term and the sum of the transformation and stability terms, respectively. As will be shown in the proof of Theorem 6, the RHS of Lemma 7 can be bounded in terms of $\sum_{t=1}^T H(q_t)$, for which we have the following bound.

Lemma 8 *For any $a^* \in [k]$, we have $\sum_{t=1}^T H(q_t) \le Q(a^*) \log(\mathrm{e} k_\Pi T / Q(a^*))$.*

We can show this lemma similarly to Lemma 4 of Ito et al. (2022a) by noting that $q_{t,a} = 0$ for $a \notin \Pi$. Finally, we are ready to prove Theorem 6. Here, we only sketch the proof and provide the complete proof can be found in Appendix B.5.

Proof sketch of Theorem 6. We prove this theorem by bounding the RHS of Lemma 7.

(Bounding the penalty term) Since β_{t+1}' is non-decreasing and $\beta_t' \le \beta_t$ from the definition of learning rate in (12), it holds that

$$\begin{aligned} \sum_{t=1}^T ({\eta_{t+1}}^{-1} - {\eta_t}^{-1}) H(q_{t+1}) &\le \sum_{t=1}^T (\beta_{t+1}' - \beta_t') H(q_{t+1}) = \sum_{t=1}^T \frac{c_1 \sqrt{\log k_\Pi}\, H(q_{t+1})}{\sqrt{\log k_\Pi + \sum_{s=1}^t H(q_s)}} \\ &\le c_1 \sqrt{\log k_\Pi} \sum_{t=1}^T \frac{2 H(q_{t+1})}{\sqrt{\sum_{s=1}^{t+1} H(q_s)} + \sqrt{\sum_{s=1}^t H(q_s)}} \le 2 c_1 \sqrt{\log k_\Pi} \sqrt{\sum_{t=1}^T H(q_t)} \,, \end{aligned} \tag{14}$$

where the second inequality follows from $0 \le H(q_{t+1}) \le \log k_\Pi$, and the last inequality follows by sequentially applying $b/(\sqrt{a+b} + \sqrt{a}) = \sqrt{a+b} - \sqrt{a}$ for $a, b > 0$, the telescoping argument, $\sqrt{a+b} - \sqrt{b} \le \sqrt{a}$ for $a, b \ge 0$, and $H(q_{T+1}) \le H(q_1)$.

(Bounding the sum of the transformation and part of stability terms) It holds that

$$\sum_{t=1}^{T} \eta_t V_t' \le \max_{s \in [T]} V_s' \sum_{t=1}^{T} \eta_t \le 3m^2 k^3 \sum_{t=1}^{T} \eta_t \le \frac{3m^2 k^3 (1 + \log T)}{c_1} \sqrt{1 + \frac{1}{\log k_\Pi} \sum_{t=1}^{T} H(q_t)}\,, \quad (15)$$

where the second inequality follows from Lemma 5 and the last inequality follows since the lower bound $\beta_t' = c_1 + \sum_{u=1}^{t-1} \frac{c_1}{\sqrt{1+(\log k_\Pi)^{-1} \sum_{s=1}^{u} H(q_s)}} \ge \frac{c_1 t}{\sqrt{1+(\log k_\Pi)^{-1} \sum_{s=1}^{t} H(q_s)}}$ implies that

$$\sum_{t=1}^{T} \eta_t \le \sum_{t=1}^{T} \frac{1}{\beta_t'} \le \sum_{t=1}^{T} \frac{1}{c_1 t} \sqrt{1 + \frac{1}{\log k_\Pi} \sum_{s=1}^{t} H(q_s)} \le \frac{1 + \log T}{c_1} \sqrt{1 + \frac{1}{\log k_\Pi} \sum_{t=1}^{T} H(q_t)}\,.$$

(Summing up arguments and applying a self-bounding technique) By bounding the RHS of Lemma 7 by (14) and (15) with $c_1 = \Theta\big(mk^{3/2}\sqrt{\log(T)/\log k_\Pi}\big)$, we have $R_T = O\Big(mk^{3/2} \sqrt{\log(T) \sum_{t=1}^{T} H(q_t)} + mk^{3/2}\sqrt{\log(T) \log k_\Pi}\Big) + 2mk^2 \log k_\Pi$. Since $\sum_{t=1}^{T} H(q_t) \le T \log k_\Pi$, the desired bound for the adversarial regime is obtained. We consider the adversarial regime with a self-bounding constraint in the following. Here, we only consider the case of $Q(a^*) \ge \mathrm{e}$, since otherwise we easily obtain the desired bound. Note that Lemma 8 with $Q(a^*) \ge \mathrm{e}$ implies $\sum_{t=1}^{T} H(q_t) \le Q(a^*) \log(k_\Pi T)$. Hence, for any $\lambda > 0$

$$\begin{aligned} R_T &= (1+\lambda) R_T - \lambda R_T \le \mathbb{E}\Big[(1+\lambda) O\left(mk^{3/2}\sqrt{\log(T)\log(k_\Pi T) Q(a^*)}\right) - \frac{\lambda \Delta_{\min} Q(a^*)}{2k}\Big] + \lambda C \\ &\le O\big(\mathcal{R}^{\mathrm{loc}} + \lambda(\mathcal{R}^{\mathrm{loc}} + C) + \mathcal{R}^{\mathrm{loc}}/\lambda\big)\,, \end{aligned}$$

where the first inequality follows by Lemma 4 with $c = 1/(2k)$, and the second inequality follows from $a\sqrt{x} - bx/2 \le a^2/(2b)$ for $a, b, x \ge 0$ and $\mathcal{R}^{\mathrm{loc}} = m^2 k^4 \log(T) \log(k_\Pi T)/\Delta_{\min}$. Appropriately choosing λ gives the desired bound. ■

5. Globally Observable Case

This section proposes an algorithm for globally observable games and derives its BOBW regret bound. We use G defined in (3) and let $c_{\mathcal{G}} = \max\{1, k\|G\|_\infty\}$ be the game-dependent constant.

5.1. Proposed Algorithm

In the proposed algorithm for globally observable games, we use the regularizer ψ_t in (5) as used in the locally observable case, but with different parameters. We define $\beta_t, \gamma_t \in \mathbb{R}$ by $\beta_1 = \max\{c_2, 2c_{\mathcal{G}}\}$ and

$$\gamma_t' = \frac{1}{4} \frac{c_1 b_t}{c_1 + \left(\sum_{s=1}^{t} b_s\right)^{1/3}}\,, \quad \beta_{t+1} = \beta_t + \frac{c_2 b_t}{\gamma_t' \left(c_1 + \sum_{s=1}^{t-1} \frac{b_s a_{s+1}}{\gamma_s'}\right)^{1/2}}\,, \quad \gamma_t = \gamma_t' + \frac{c_{\mathcal{G}}}{2\beta_t}\,, \quad (16)$$

where c_1 and c_2 are parameters satisfying $c_1 \ge \max\{1, \log k_\Pi\}$, and a_t and b_t are defined by

$$a_t = H(q_t) = -\sum_{a \in \Pi} q_{t,a} \log(q_{t,a}) \quad \text{and} \quad b_t = 1 - \max_{a \in \Pi} q_{t,a}\,. \quad (17)$$

Algorithm 2: BOBW algorithm for globally observable games

```
for $t = 1, 2, \ldots$ do
    Compute $q_t$ using (6)
    Compute $a_t, b_t$ in (17), $\gamma'_t, \gamma_t$ in (16), and $p_t$ from $q_t$ by (18)
    Sample $A_t \sim p_t$, observe $\sigma_t \in \Sigma$, compute $\widehat{y}_t = G(A_t, \sigma_t)/p_{t,A_t}$, and update $\beta_t$ using (16)
```

Note that we have $\psi_t(0) = 0$, and using $\beta_t \geq \beta_1 \geq 2c_{\mathcal{G}}$ and $b_t \leq \sum_{a=1}^k q_{t,a} \leq 1$ we have $\gamma_t \leq c_1 b_t/(4c_1) + c_{\mathcal{G}}/(2c_{\mathcal{G}}) \leq 1/2$. We use the following transform from q_t to p_t:

$$p_t = \mathcal{T}_t(q_t) = (1-\gamma_t)q_t + \frac{\gamma_t}{k}\mathbf{1} \,. \tag{18}$$

Algorithm 2 presents the proposed algorithm for globally observable games.

5.2. Regret Analysis for Globally Observable Games

With the above algorithm, we can prove the following regret bound for globally observable games.

Theorem 9 *Consider any globally observable partial monitoring game. If we run Algorithm 2 with $c_1 = \Theta\big(\big(c_{\mathcal{G}}^2 \log(T)\log(k_\Pi T)\big)^{1/3}\big)$ and $c_2 = \Theta\big(\sqrt{c_{\mathcal{G}}^2 \log T}\big)$, we have the following bounds. For the adversarial regime with a (Δ, C, T) self-bounding constraint, we have*

$$R_T = O\left(\frac{c_{\mathcal{G}}^2 \log(T)\log(k_\Pi T)}{\Delta_{\min}^2} + \left(\frac{C^2 c_{\mathcal{G}}^2 \log(T)\log(k_\Pi T)}{\Delta_{\min}^2}\right)^{1/3}\right) , \tag{19}$$

and for the adversarial regime, we have

$$R_T = O\big(\big(c_{\mathcal{G}}^2 \log(T)\log(k_\Pi T)\big)^{1/3} T^{2/3}\big) \,,$$

where in the last big-O notation, the terms of $o(\mathrm{poly}(k, c_G)(T\log T)^{2/3})$ are ignored.

Note that (19) with $C = 0$ yields the bound in the stochastic regime. The bound for the adversarial regime is a factor of $(\log(T)\log(k_\Pi T)/\log k)^{1/3}$ worse than the algorithm by Lattimore and Szepesvári (2020b). This comes from the difficulty of obtaining the BOBW guarantee, where we need to aggressively change the learning rate when the environment looks not so much adversarial.

We begin the analysis by decomposing the regret as follows.

Lemma 10 *The regret of Algorithm 2 is bounded as $R_T \leq \mathbb{E}\big[\sum_{t=1}^T \gamma_t + \sum_{t=1}^T \big(\langle \widehat{y}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t)\big) + \sum_{t=1}^T \big(\psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1})\big) + \psi_{T+1}(e_{a^*}) - \psi_1(q_1)\big]$.*

This lemma can be proven based on the fact that we can estimate loss differences between Pareto optimal actions, and boundedness of $\mathcal{L}$, combined with the standard analysis of FTRL given in (7). Note that the first, second, and last terms correspond to the transformation, stability, and penalty terms, respectively. We can bound the stability term on the RHS of Lemma 10 as follows.

Lemma 11 *If ψ_t is given by (5) and b_t is defined by (17), then we have*

$$\mathbb{E}[\langle \widehat{y}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t)] \leq \mathbb{E}\big[2c_{\mathcal{G}}^2 b_t/(\beta_t \gamma_t)\big] \,. \tag{20}$$

Remark 12 Globally observable PM is a generalization of the weakly observable setting in online learning with feedback graphs (Alon et al., 2015). For this online learning problem, the regularizer in the form of $-H(p) - H(\mathbf{1} - p)$ rather than (6) is introduced in Ito et al. (2022a) to make the LHS of (20) easy to bound. However, FTRL with this regularizer requires solving a convex optimization every round. This study shows that the LHS of (20) can be favorably bounded without the regularization of $-H(\mathbf{1} - p)$. The key to the proof of this lemma is that for any $a' \in [k]$ it holds that $\langle \widehat{y}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) = \langle \widehat{y}_t - \widehat{y}_{ta'}\mathbf{1}, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \beta_t \sum_{a=1}^{k} q_{t,a} \xi\left((\widehat{y}_{ta} - \widehat{y}_{ta'})/\beta_t\right)$, which enables us to bound the stability term with b_t in (17), leading to the regret upper bound depending on $Q(a^*)$ in Proposition 13.

Using the definition of β_t and γ_t in (16) with Lemmas 10 and 11, we can bound the regret as follows.

Proposition 13 *Assume β_t and γ_t are given by (16). Then, the regret is bounded as $R_T = O\Big(\mathbb{E}\Big[c_1 B_T^{2/3} + \tilde{c}\sqrt{c_1^2 + (\log k_\Pi + A_T)\left(c_1 + B_T^{1/3}\right)}\Big] + \beta_1 \log k_\Pi\Big)$, where $A_T = \sum_{t=1}^{T} a_t$, $B_T = \sum_{t=1}^{T} b_t$, and $\tilde{c} = O\left(\frac{1}{\sqrt{c_1}}\left(\frac{c_\mathcal{G}^2 \log T}{c_2} + c_2\right)\right) = O\Big(\frac{c_1}{\sqrt{\log(k_\Pi T)}}\Big)$.*

The proof of this lemma is similar to Proposition 2 of Ito et al. (2022a). Now we are ready to prove Theorem 9, whose proof is sketched below and completed in Appendix B.9.

Proof sketch of Theorem 9. We first consider the adversarial regime. In the adversarial regime, Proposition 13 with $A_T \leq T \log k_\Pi$ and $B_T \leq T$ immediately leads to

$$R_T = O\Big(c_1 T^{2/3} + \tilde{c}\sqrt{c_1^2 + (\log k_\Pi + T \log k_\Pi)(c_1 + T^{1/3})}\Big) = O\big((c_1 + \tilde{c}\sqrt{\log k_\Pi})T^{2/3}\big)\,. \quad (21)$$

We next consider the adversarial regime with a self-bounding constraint. Here, we only consider the case of $Q(a^*) > \max\{\mathrm{e}, c_1^3\}$, since otherwise we can easily obtain the desired bound. Note that $A_T \leq Q(a^*) \log(k_\Pi T)$ by Lemma 8 with $Q(a^*) \geq \mathrm{e}$ and $B_T = \sum_{t=1}^{T}(1 - \max_{a \in \Pi} q_{t,a}) \leq \sum_{t=1}^{T}(1 - q_{t,a^*}) = Q(a^*)$. Then, Proposition 13 with these inequalities and $Q(a^*) > c_1^3$ gives

$$R_T \leq O\Big(\mathbb{E}\Big[c_1 Q(a^*)^{2/3} + \tilde{c}\sqrt{\log(k_\Pi T) Q(a^*)^{4/3}}\Big]\Big) \leq O\big((c_1 + \tilde{c}\sqrt{\log(k_\Pi T)})\bar{Q}(a^*)^{2/3}\big)\,. \quad (22)$$

By (21) and (22), there exists $\widehat{c} = O\big(c_1 + \tilde{c}\sqrt{\log(k_\Pi T)}\big)$ satisfying $R_T \leq \widehat{c}\, T^{2/3}$ for the adversarial regime and $R_T \leq \widehat{c}\,\bar{Q}(a^*)^{2/3}$ for the adversarial regime with a self-bounding constraint. Recalling the definitions of c_1 and c_2, we have $\widehat{c} = O\left((c_\mathcal{G}^2 \log(T) \log(k_\Pi T))^{1/3}\right)$, which gives the desired bounds for the adversarial regime. For the adversarial regime with a self-bounding constraint, using $R_T \leq \widehat{c}\,\bar{Q}(a^*)^{2/3}$ and Lemma 4 with $c = 1/2$ for any $\lambda \in (0, 1]$ it holds that

$$R_T = (1 + \lambda)R_T - \lambda R_T \leq (1 + \lambda)\widehat{c} \cdot \bar{Q}(a^*)^{2/3} - \lambda \Delta_{\min} \bar{Q}(a^*)/2 + \lambda C\,. \quad (23)$$

Taking the worst case of this with respect to $\bar{Q}(a^*)$ and taking $\lambda \in (0, 1]$ appropriately gives the desired bound for the adversarial regime with a self-bounding constraint. ■

Acknowledgments

TT was supported by JST, ACT-X Grant Number JPMJAX210E, Japan and JSPS, KAKENHI Grant Number JP21J21272, Japan. JH was supported by JSPS, KAKENHI Grant Number JP21K11747, Japan.

References

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 23–35. PMLR, 2015.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Gábor Bartók. A near-optimal algorithm for finite partial-monitoring games against adversarial opponents. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 696–710. PMLR, 2013.

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *the 24th Annual Conference on Learning Theory*, volume 19, pages 133–154, 2011.

Gábor Bartók, Navid Zolghadr, and Csaba Szepesvári. An adaptive algorithm for finite stochastic partial monitoring. In *the 29th International Conference on Machine Learning*, pages 1–20, 2012.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 42.1–42.23. PMLR, 2012.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

Liad Erez and Tomer Koren. Towards best-of-all-worlds online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 34, pages 28511–28521. Curran Associates, Inc., 2021.

Dean Foster and Alexander Rakhlin. No internal regret via neighborhood watch. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 382–390. PMLR, 2012.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 176–196. PMLR, 2014.

Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85(3):361–391, 2011. ISSN 1573-0565. doi: 10.1007/s10994-011-5257-4.

Jiatai Huang, Yan Dai, and Longbo Huang. Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9173–9200. PMLR, 2022.

Shinji Ito. Hybrid regret bounds for combinatorial semi-bandits and adversarial linear bandits. In *Advances in Neural Information Processing Systems*, volume 34, pages 2654–2667. Curran Associates, Inc., 2021.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022a.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adversarially robust multi-armed bandit algorithm with variance-dependent regret bounds. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1421–1422. PMLR, 2022b.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic MDPs with known transition. In *Advances in Neural Information Processing Systems*, volume 33, pages 16557–16566. Curran Associates, Inc., 2020.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In *Advances in Neural Information Processing Systems 28*, pages 1792–1800. Curran Associates, Inc., 2015.

Fang Kong, Yichi Zhou, and Shuai Li. Simultaneously learning stochastic and adversarial bandits with general graph feedback. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11473–11482. PMLR, 2022.

Tor Lattimore and Csaba Szepesvári. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 529–556. PMLR, 2019a.

Tor Lattimore and Csaba Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2111–2139. PMLR, 2019b.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020a.

Tor Lattimore and Csaba Szepesvári. Exploration by optimisation in partial monitoring. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2488–2515. PMLR, 2020b.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6142–6151. PMLR, 2021.

Tian Lin, Bruno Abrahao, Robert Kleinberg, John Lui, and Wei Chen. Combinatorial partial monitoring game with linear feedback and its applications. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 901–909. PMLR, 2014.

Haipeng Luo and Robert E. Schapire. Achieving all with no parameters: AdaNormalHedge. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1286–1304. PMLR, 2015.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

Shie Mannor and Nahum Shimkin. On-line learning with imperfect monitoring. In *Learning Theory and Kernel Machines*, pages 552–566. Springer, 2003.

Shie Mannor, Vianney Perchet, and Gilles Stoltz. Set-valued approachability and online learning with partial monitoring. *Journal of Machine Learning Research*, 15(94):3247–3295, 2014.

Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and L1 regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 525–533. PMLR, 2011.

Vianney Perchet. Approachability of convex sets in games with partial monitoring. *Journal of Optimization Theory and Applications*, 149(3):665–677, 2011. doi: 10.1007/s10957-011-9797-3.

Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss (extended abstract). In *Computational Learning Theory*, pages 208–223, 2001.

Chloé Rouyer, Dirk van der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

Aldo Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29(1): 224–243, 1999.

Aadirupa Saha and Pierre Gaillard. Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19011–19026. PMLR, 2022.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1743–1759. PMLR, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1287–1295. PMLR, 2014.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 12 1933.

Taira Tsuchiya, Junya Honda, and Masashi Sugiyama. Analysis and design of Thompson sampling for stochastic partial monitoring. In *Advances in Neural Information Processing Systems*, volume 33, pages 8861–8871. Curran Associates, Inc., 2020.

Hastagiri P Vanchinathan, Gábor Bartók, and Andreas Krause. Efficient partial monitoring with prior information. In *Advances in Neural Information Processing Systems 27*, pages 1691–1699. Curran Associates, Inc., 2014.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1263–1291. PMLR, 2018.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7683–7692. PMLR, 2019.

Appendix A. Intermediate Regimes between Stochastic and Adversarial Regimes

This section details the discussion on intermediate regimes between stochastic and adversarial regimes given in Section 2. This section first defines the *stochastically constrained adversarial regime* in PM, and then shows that the stochastic regime, adversarial regime, stochastically constrained adversarial regime, and stochastic regime with adversarial corruptions are indeed adversarial regimes with a self-bounding constraint defined in Definition 3.

The stochastically constrained adversarial regime was initially considered by Wei and Luo (2018) and also discussed in Zimmert and Seldin (2021) in the context of the multi-armed bandit problem. We say that the environment is the stochastically constrained adversarial regime if for any $a \neq a^*$ there exists $\tilde{\Delta}_{a,a^*} > 0$ such that $\mathbb{E}_{x_t \sim \nu^*}[\mathcal{L}_{a x_t} - \mathcal{L}_{a^* x_t} | x_1, \dots, x_{t-1}] \geq \tilde{\Delta}_{a,a^*}$.

Next, we show that the stochastic regime, adversarial regime, stochastically constrained adversarial regime, and stochastic regime with adversarial corruptions are indeed included in the adversarial regime with a self-bounding constraint. We first consider the stochastic regime. Indeed, if outcomes $(x_t)_t$ follow a distribution ν^* independently for $t = 1, 2, \dots, T$, we have $R_T = \max_{a^* \in [k]} \mathbb{E}[\sum_{t=1}^T (\mathcal{L}_{A_t x_t} - \mathcal{L}_{a^* x_t})] = \mathbb{E}[\sum_{t=1}^T \Delta_{A_t}]$, where we define $\Delta \in [0,1]^k$ by $\Delta_a = \mathbb{E}_{x \sim \nu^*}[\mathcal{L}_{ax} - \mathcal{L}_{a^* x}]$. This implies that the stochastic regime is in the adversarial regime with a $(\Delta, 0, T)$ self-bounding constraint. We next consider the stochastic regime with adversarial corruptions. In fact, using the definition of the corruption level C, we have

$$\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{t=1}^T (\mathcal{L}_{A_t x_t} - \mathcal{L}_{a^* x_t})\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \left(\mathcal{L}_{A_t x'_t} - \mathcal{L}_{a^* x'_t}\right)\right] + \mathbb{E}\left[\sum_{t=1}^T \left(\mathcal{L}_{A_t x_t} - \mathcal{L}_{A_t x'_t}\right)\right] + \mathbb{E}\left[\sum_{t=1}^T \left(\mathcal{L}_{a^* x'_t} - \mathcal{L}_{a^* x_t}\right)\right] \\
&\geq \mathbb{E}\left[\sum_{t=1}^T \Delta_{A_t}\right] - 2C \,,
\end{aligned}$$

which implies that the stochastic regime with adversarial corruption with corruption levels C is an adversarial regime with a $(\Delta, 2C, T)$ self-bounding constraint. It is also easy to see that adversarial regimes are the adversarial regime with a $(\Delta, 2T, T)$ self-bounding constraint, and the stochastically constrained adversarial regime are the adversarial regime with a $(\Delta, 0, T)$ self-bounding constraint by defining $\Delta \in [0,1]^k$ by $\Delta_a = \tilde{\Delta}_{a,a^*}$.

Appendix B. Omitted Proofs

B.1. Proof of Lemma 4

Proof Note that the environment is the adversarial regime with a self-bounding constraint with $\Delta \in [0,1]^k$ such that $\Delta_a \geq \Delta_{\min}$ for all $a \in [k] \setminus \{a^*\}$. Hence, the regret is then bounded as

$$\begin{aligned}
R_T &\geq \mathbb{E}\left[\sum_{t=1}^T \Delta_{A_t}\right] - C = \mathbb{E}\left[\sum_{t=1}^T \sum_{a=1}^k p_{t,a} \Delta_a\right] - C \\
&\geq \mathbb{E}\left[\sum_{t=1}^T \sum_{a=1}^k c\, q_{t,a} \Delta_a\right] - C \geq c\, \Delta_{\min} \bar{Q}(a^*) - C \,,
\end{aligned}$$

where the first inequality follows from Definition 3, the equality follows from $A_t \sim p_t$, the second inequality follows from the definition of p_t given in (4), and the last inequality follows from the assumption $p_{t,a} \geq c\, q_{t,a}$ for all $t \in [T], a \in [k]$ and the definition of $\bar{Q}(a^*)$ given in (8). This completes the proof of Lemma 4. ■

B.2. Proof of Lemma 5

Before proving Lemma 5, we review the definition and property of the water transfer operator W_ν introduced by Lattimore and Szepesvári (2019b). We refer to $\mathscr{T} \subset [k] \times [k]$ representing the edges of a directed tree with vertices $[k]$ as *in-tree* with vertex set $[k]$ and define $\mathcal{E} = \{(a,b) \in [k] \times [k] : a \text{ and } b \text{ are neighbors}\}$.

Lemma 14 (Lattimore and Szepesvári, 2019b) *Assume that partial monitoring game $\mathcal{G}$ is non-degenerate and locally observable and let $\nu \in \mathcal{P}_d$. Then there exists a function $W_\nu : \mathcal{P}_k \to \mathcal{P}_k$ such that the following hold for all $q \in \mathcal{P}_k$: (a) $(W_\nu(q) - q)^\top \mathcal{L}\nu \leq 0$; (b) $W_\nu(q)_a \geq q_a/k$ for all $a \in [k]$; and (c) there exists an in-tree $\mathscr{T} \subset \mathcal{E}$ over $[k]$ such that $W_\nu(q)_a \leq W_\nu(q)_b$ for all $(a,b) \in \mathscr{T}$.*

Using this, we prove the generalized version of Proposition 8 of Lattimore and Szepesvári (2020b), where the proof follows a quite similar argument as their proof therein.

Proof of Lemma 5. We define the set of functions that satisfy (2) by

$$\mathcal{H}_\circ = \left\{ G : (e_b - e_c)^\top \sum_{a=1}^{k} G(a, \Phi_{ax}) = \mathcal{L}_{bx} - \mathcal{L}_{cx} \text{ for all } b, c \in \Pi \text{ and } x \in [d] \right\} .$$

Take any $q \in \mathcal{P}_k$. By Sion's minimax theorem, we have

$$\begin{aligned} \mathrm{opt}'_q(\eta) &\leq \min_{G \in \mathcal{H}^\circ,\, p \in \mathcal{P}'_k(q)} \max_{\nu \in \mathcal{P}_d} \left[\frac{1}{\eta}(p-q)^\top \mathcal{L}\nu + \frac{1}{\eta^2} \sum_{x=1}^{d} \nu_x \sum_{a=1}^{k} p_a \left\langle q, \xi\left(\frac{\eta G(a, \Phi_{ax})}{p_a} \right) \right\rangle \right] \\ &= \max_{\nu \in \mathcal{P}_d} \min_{G \in \mathcal{H}^\circ,\, p \in \mathcal{P}'_k(q)} \left[\frac{1}{\eta}(p-q)^\top \mathcal{L}\nu + \frac{1}{\eta^2} \sum_{x=1}^{d} \nu_x \sum_{a=1}^{k} p_a \left\langle q, \xi\left(\frac{\eta G(a, \Phi_{ax})}{p_a} \right) \right\rangle \right] , \end{aligned}$$

where the first inequality follows since we added the constraint that $G \in \mathcal{H}^\circ$, which makes the bias term zero. Take any $\nu \in \mathcal{P}_d$ and let $\mathscr{T}$ be the in-tree over $[k]$. Using these variables, we define the action selection probability vector $p \in \mathcal{P}'_k(q)$ by

$$p = (1-\gamma)u + \frac{\gamma}{k}\mathbf{1} , \quad \text{where} \quad u = W_\nu(q) , \quad \text{and} \quad \gamma = \frac{\eta m k^2}{2} .$$

Here, $W_\nu : \mathcal{P}_k \to \mathcal{P}_k$ is the water operator. It is worth noting that from the assumption that $\eta \leq 1/(mk^2)$, we have $\gamma \leq 1/2$ and $p_a \geq u_a/2 = W_\nu(q)_a/2 \geq q_a/(2k)$, where the last inequality follows from Part (b) of Lemma 14, and this indeed implies $p \in \mathcal{P}'_k(q)$.

We take $G \in \mathcal{H}^\circ$ defined in (3), where we recall that $G(a,\sigma)_b = \sum_{e \in \mathrm{path}_{\mathcal{T}}(b)} w_e(a,\sigma)$. By Lemma 20 of Lattimore and Szepesvári (2020b) and the assumption that $\mathcal{G}$ is non-degenerate, w_e can be chosen so that $\|w_e\|_\infty \le m/2$. Since paths in $\mathcal{T}$ have length at most k, we have $\|G\|_\infty \le km/2$. From the above definitions, for any $x \in [d]$ we have

$$\frac{\eta G(a, \Phi_{ax})}{p_a} \ge -\frac{\eta m k^2}{2\gamma} = -1\,.$$

Hence, using Parts (b) and (c) of Lemma 14, we have

$$\begin{aligned}
\frac{1}{\eta^2}\sum_{a=1}^{k} p_a \left\langle q, \xi\left(\frac{\eta G(a,\Phi_{ax})}{p_a}\right)\right\rangle &\le \sum_{a=1}^{k}\frac{1}{p_a}\sum_{b=1}^{k} q_b \left(G(a,\Phi_{ax})_b\right)^2 \\
&\le 2\sum_{a=1}^{k}\frac{1}{u_a}\sum_{b=1}^{k} q_b \left(G(a,\Phi_{ax})_b\right)^2 \\
&= 2\sum_{b=1}^{k}\sum_{a=1}^{k}\frac{q_b}{u_a}\left(\sum_{e\in \mathrm{path}_{\mathcal{T}}(b)} w_e(a,\Phi_{ax})\right)^2 \\
&\le \frac{m^2}{2}\sum_{b=1}^{k}\sum_{a=1}^{k}\frac{q_b}{u_a}\left(\sum_{e\in \mathrm{path}_{\mathcal{T}}(b)} \mathbb{1}[a\in e]\right)^2 \\
&\le 2m^2k^3\,,
\end{aligned}$$

where the first inequality follows from

$$\xi(x) = \exp(-x) + x - 1 \le x^2 \text{ for } x \ge -1\,, \tag{24}$$

the second inequality follows since $p_a \ge u_a/2$, the third inequality follows since $\|w_e\|_\infty \le m/2$, and the last inequality follows from Part (b) of Lemma 14 to implying that $q_b \le k u_b$ and Part (c) implying that $u_a \ge u_b$ for $a \in \mathrm{path}_{\mathcal{T}}(b)$. Finally,

$$\frac{1}{\eta}(p-q)^\top \mathcal{L}\nu = \frac{1}{\eta}(u-q)^\top \mathcal{L}\nu + \frac{\gamma}{\eta}\left(\frac{1}{k}\mathbf{1} - u\right)^\top \mathcal{L}\nu \le \frac{\gamma}{\eta}\left(\frac{1}{k}\mathbf{1} - u\right)^\top \mathcal{L}\nu \le mk^2\,,$$

where the first inequality follows from Part (a) of Lemma 14. Summing up the above arguments, we have $\mathrm{opt}'_q(\eta) \le 3m^2k^3$, which completes the proof of Lemma 5. ∎

B.3. Proof of Lemma 7

We first analyze the stability term in (7) for ψ_t defined in (5).

Lemma 15 *If ψ_t is given by (5), it holds for any $\ell \in \mathbb{R}^k$ and $p, q \in \mathcal{P}_k$ that*

$$\langle \ell, p - q\rangle - D_t(q,p) \le \beta_t \sum_{a=1}^{k} p_a \xi\left(\frac{\ell_a}{\beta_t}\right)\,,$$

where we recall that $\xi(x) = \exp(-x) + x - 1$.

Proof For any $x, y \in (0,1)$, we let $d(y,x) \geq 0$ be the Bregman divergence over $(0,1)$ induced by $\psi(x) = x \log x$, *i.e.*,

$$d(y,x) = y \log y - x \log x - (\log x + 1)(y - x) = y \log \frac{y}{x} + x - y\,.$$

Using this, the Bregman divergence induced by $\psi_t(p) = (1/\eta_t) \sum_{a=1}^k p_a \log(p_a) = \beta_t \sum_{a=1}^k p_a \log(p_a)$ in (5) can be written as

$$D_t(q,p) = \psi_t(p) - \psi_t(q) - \langle \nabla \psi_t(q), p - q \rangle = \beta_t \sum_{a=1}^k d(q_a, p_a)\,.$$

From this, we have

$$\langle \ell, p - q \rangle - D_t(q,p) \leq \sum_{a=1}^k \left(\ell_a (p_a - q_a) - \beta_t d(q_a, p_a) \right)\,. \tag{25}$$

We show

$$\ell_a(p_a - q_a) - \beta_t d(q_a, p_a) \leq \beta_t p_a \xi \left(\frac{\ell_a}{\beta_t} \right)\,. \tag{26}$$

As $\ell_a(p_a - q_a) - \beta_t d(q_a, p_a)$ is concave in q, its maximum subject to $q \in \mathbb{R}$ is attained when the derivative of it is equal to zero, *i.e.*,

$$\frac{\partial}{\partial q_a} \left(\ell_a(p_a - q_a) - \beta_t d(q_a, p_a) \right) = -\ell_a - \beta_t \left(\log q_a - \log p_a \right) = 0\,.$$

This implies that the maximum is attained when $q_a = q_a^* := p_a \exp\left(-\ell_a/\beta_t\right)$. Hence, we obtain (26) by

$$\begin{aligned}
&\ell_a(p_a - q_a) - \beta_t d(q_a, p_a) \leq \ell_a(p_a - q_a^*) - \beta_t d(q_a^*, p_a) \\
&= \ell_a(p_a - q_a^*) - \beta_t \left(q_a^* \log q_a^* - p_a \log p_a - (\log p_a + 1)(q_a^* - p_a) \right) \\
&= \ell_a p_a - \beta_t \left(q_a^* \log p_a - p_a \log p_a - (\log p_a + 1)(q_a^* - p_a) \right) \\
&= \ell_a p_a + \beta_t (q_a^* - p_a) = \beta_t p_a \left(\exp\left(-\frac{\ell_a}{\beta_t} \right) + \frac{\ell_a}{\beta_t} - 1 \right) \\
&= \beta_t p_a \xi \left(\frac{\ell_a}{\beta_t} \right)\,,
\end{aligned}$$

where the second equality follows from $\log q_a^* = \log p_a - \ell_a/\beta_t$, and the fourth equality follows from $q_a^* = p_a \exp\left(-\ell_a/\beta_t\right)$. Combining (25) and (26) completes the proof. ■

Proof of Lemma 7. Let $a^* = \arg\min_{a \in [k]} \mathbb{E}\left[\sum_{t=1}^T \mathcal{L}_{a x_t}\right] \in \Pi$ be the optimal action in hindsight. We have

$$\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{t=1}^T (\mathcal{L}_{A_t x_t} - \mathcal{L}_{a^* x_t}) \right] = \mathbb{E}\left[\sum_{t=1}^T \sum_{b=1}^k p_{t,b} (\mathcal{L}_{b x_t} - \mathcal{L}_{a^* x_t}) \right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \sum_{b=1}^k (p_{t,b} - q_{t,b}) (\mathcal{L}_{b x_t} - \mathcal{L}_{a^* x_t}) + \sum_{t=1}^T \sum_{b=1}^k q_{t,b} (\mathcal{L}_{b x_t} - \mathcal{L}_{a^* x_t}) \right]\,.
\end{aligned} \tag{27}$$

The first term in (27) is equal to $\mathbb{E}\big[\sum_{t=1}^T (p_t - q_t)^\top \mathcal{L} e_{x_t}\big]$. The second term in (27) can be bounded as

$$\begin{aligned}
\mathbb{E}\left[\sum_{b=1}^k q_{t,b}(\mathcal{L}_{bx_t} - \mathcal{L}_{a^*x_t})\right] &= \mathbb{E}\left[\sum_{b=1}^k q_t^\top \mathcal{L} e_{x_t} - \mathcal{L}_{a^*x_t}\right] \\
&= \mathbb{E}\left[\sum_{b=1}^k q_t^\top \mathcal{L} e_{x_t} - q_t^\top \sum_{a=1}^k G_t(a, \Phi_{ax_t}) + \sum_{a=1}^k G_t(a, \Phi_{ax_t})_{a^*} - \mathcal{L}_{a^*x_t}\right] \\
&\quad + \mathbb{E}\left[q_t^\top \sum_{a=1}^k G_t(a, \Phi_{ax_t}) - \sum_{a=1}^k G_t(a, \Phi_{ax_t})_{a^*}\right] \\
&\le \mathbb{E}[\mathrm{bias}_{q_t}(G; x_t)] + \mathbb{E}\left[q_t^\top \widehat{y}_t - \widehat{y}_{ta^*}\right], \qquad (28)
\end{aligned}$$

where in the last inequality we used the definition in (10) and Lemma 2 with $a^* \in \Pi$ and $q_{t,a} = 0$ for $a \notin \Pi$. The sum over $t \in [T]$ of the last term in (28) can be bounded using (7) and the definition of the regularizer (5) as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \sum_{b=1}^k q_{t,b}(\widehat{y}_{tb} - \widehat{y}_{ta^*})\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) H(q_{t+1}) + \frac{H(q_1)}{\eta_1} + \sum_{t=1}^T \Big(\langle q_t - q_{t+1}, \widehat{y}_t\rangle - D_t(q_{t+1}, q_t)\Big)\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) H(q_{t+1}) + \frac{H(q_1)}{\eta_1} + \sum_{t=1}^T \frac{\langle q_t, \xi(\eta_t \widehat{y}_t)\rangle}{\eta_t}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) H(q_{t+1}) + \frac{H(q_1)}{\eta_1} + \sum_{t=1}^T \frac{1}{\eta_t} \sum_{a=1}^k p_{t,a} \left\langle q_t, \xi\left(\frac{\eta_t G(a, \sigma_t)}{p_{t,a}}\right)\right\rangle\right], \qquad (29)
\end{aligned}$$

where in the second inequality we used the following inequality obtained by Lemma 15:

$$\langle \widehat{y}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) \le \beta_t \sum_{a=1}^k q_{t,a} \xi\left(\frac{\widehat{y}_{ta}}{\beta_t}\right) = \frac{\langle q_t, \xi(\eta_t \widehat{y}_t)\rangle}{\eta_t}.$$

Using the definition of the optimization problem (11) and $V_t' = \max\{0, \mathrm{opt}'_{q_t}(\eta_t)\}$, we have

$$(p_t - q_t)^\top \mathcal{L} e_{x_t} + \mathrm{bias}_{q_t}(G; x_t) + \frac{1}{\eta_t} \sum_{a=1}^k p_{t,a} \left\langle q_t, \xi\left(\frac{\eta_t G(a, \sigma_t)}{p_{t,a}}\right)\right\rangle \le \eta_t V_t'. \qquad (30)$$

Summing up the arguments in (27), (28), (29), and (30), we have

$$R_T \le \mathbb{E}\left[\sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) H(q_{t+1}) + \frac{H(q_1)}{\eta_1} + \sum_{t=1}^T \eta_t V_t'\right],$$

which completes the proof. ■

B.4. Proof of Lemma 8

Proof For any $q \in \mathcal{P}(\Pi)$ and $a^* \in \Pi$, we have

$$\begin{aligned} H(p) &= \sum_{a \in \Pi} q_a \log \frac{1}{q_a} = \sum_{a \in \Pi \setminus \{a^*\}} q_a \log \frac{1}{q_a} + q_{a^*} \log \left(1 + \frac{1 - q_{a^*}}{q_{a^*}} \right) \\ &\leq (k_\Pi - 1) \sum_{a \in \Pi \setminus \{a^*\}} \frac{1}{k_\Pi - 1} q_a \log \frac{1}{q_a} + q_{a^*} \frac{1 - q_{a^*}}{q_{a^*}} \\ &\leq (k_\Pi - 1) \cdot \frac{\sum_{a \in \Pi \setminus \{a^*\}} q_a}{k_\Pi - 1} \log \frac{k_\Pi - 1}{\sum_{a \in \Pi \setminus \{a^*\}} q_a} + q_{a^*} \frac{1 - q_{a^*}}{q_{a^*}} \\ &= (1 - q_{a^*}) \left(\log \frac{k_\Pi - 1}{1 - q_{a^*}} + 1 \right) \leq (1 - q_{a^*}) \log \frac{\mathrm{e} k_\Pi}{1 - q_{a^*}} , \end{aligned} \tag{31}$$

where the first inequality follows from $\log(1 + x) \leq x$ for $x \geq 0$, the last inequality follows from Jensen's inequality, and the last equality follows from $\sum_{a \in \Pi} q_a = 1$. Using (31), for any $a^* \in [k]$ we have

$$\begin{aligned} \sum_{t=1}^T a_t = \sum_{t=1}^T H(q_t) &\leq \sum_{t=1}^T (1 - q_{ta^*}) \log \frac{\mathrm{e} k_\Pi}{1 - q_{ta^*}} \\ &= T \sum_{t=1}^T \frac{1}{T} (1 - q_{t,a^*}) \log \frac{\mathrm{e} k_\Pi}{1 - q_{t,a^*}} \\ &\leq T \left(\sum_{t=1}^T \frac{1}{T} (1 - q_{t,a^*}) \right) \log \frac{\mathrm{e} k_\Pi}{\sum_{t=1}^T \frac{1}{T} (1 - q_{t,a^*})} \\ &= T \frac{Q(a^*)}{T} \log \frac{\mathrm{e} k_\Pi T}{Q(a^*)} = Q(a^*) \left(\log \frac{\mathrm{e} k_\Pi T}{Q(a^*)} \right) , \end{aligned}$$

where in the second inequality we used Jensen's inequality since $f(x) = x \log(1/x)$ is concave, and in the third inequality we used the definition of $Q(a^*)$ in (8). ■

B.5. Proof of Theorem 6

Proof We prove this theorem by bounding the RHS of Lemma 7.

(Bounding the penalty term) Let $t_0 = \min\{t \in [T] : \beta'_t \geq B\}$. Then, the definition of the learning rate (12) gives that

$$\begin{aligned} &\sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) H(q_{t+1}) = \sum_{t=1}^T (\beta_{t+1} - \beta_t) \, H(q_{t+1}) \\ &= \sum_{t=1}^{t_0 - 2} (\beta_{t+1} - \beta_t) \, H(q_{t+1}) + (\beta_{t_0} - \beta_{t_0 - 1}) \, H(q_{t+1}) + \sum_{t=t_0}^T (\beta_{t+1} - \beta_t) \, H(q_{t+1}) \\ &\leq 0 + \left(\beta'_{t_0} - \beta'_{t_0 - 1} \right) H(q_{t+1}) + \sum_{t=t_0}^T \left(\beta'_{t+1} - \beta'_t \right) H(q_{t+1}) \end{aligned}$$

$$\leq \sum_{t=1}^{T} \left(\beta'_{t+1} - \beta'_t\right) H(q_{t+1}) \,,$$

where in the first inequality we used the fact that β'_{t+1} is non-decreasing, $\beta_{t+1} = \beta_t$ for $t \leq t_0 - 1$, $\beta'_t \leq \beta_t$, and $\beta'_t = \beta_t$ for $t \geq t_0$. Using this inequality, we have

$$\begin{aligned}
\sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) H(q_{t+1}) &\leq \sum_{t=1}^{T} \left(\beta'_{t+1} - \beta'_t\right) H(q_{t+1}) \\
&= \sum_{t=1}^{T} \frac{c_1}{\sqrt{1 + (\log k_\Pi)^{-1} \sum_{s=1}^{t} H(q_s)}} \cdot H(q_{t+1}) \\
&= 2c_1 \sqrt{\log k_\Pi} \sum_{t=1}^{T} \frac{H(q_{t+1})}{\sqrt{\log k_\Pi + \sum_{s=1}^{t} H(q_s)} + \sqrt{\log k_\Pi + \sum_{s=1}^{t} H(q_s)}} \\
&\leq 2c_1 \sqrt{\log k_\Pi} \sum_{t=1}^{T} \frac{H(q_{t+1})}{\sqrt{\sum_{s=1}^{t+1} H(q_s)} + \sqrt{\sum_{s=1}^{t} H(q_s)}} \\
&= 2c_1 \sqrt{\log k_\Pi} \sum_{t=1}^{T} \left(\sqrt{\sum_{s=1}^{t+1} H(q_s)} - \sqrt{\sum_{s=1}^{t} H(q_s)} \right) \\
&= 2c_1 \sqrt{\log k_\Pi} \left(\sqrt{\sum_{s=1}^{T+1} H(q_s)} - \sqrt{H(q_1)} \right) \\
&\leq 2c_1 \sqrt{\log k_\Pi} \left(\sqrt{\sum_{s=2}^{T+1} H(q_s)} \right) \leq 2c_1 \sqrt{\log k_\Pi} \sqrt{\sum_{t=1}^{T} H(q_t)} \,, \qquad (32)
\end{aligned}$$

where the second inequality follows from $0 \leq H(q_{t+1}) \leq \log k_\Pi$, the third inequality follows from the inequality $\sqrt{a+b} - \sqrt{b} \leq \sqrt{a}$ that holds for $a, b \geq 0$, and the last inequality follows since $H(q_{T+1}) \leq H(q_1)$.

(Bounding the sum of the transformation and part of stability term) Using the definition of β'_t in (12), we can bound β'_t as

$$\beta'_t = c_1 + \sum_{u=1}^{t-1} \frac{c_1}{\sqrt{1 + (\log k_\Pi)^{-1} \sum_{s=1}^{u} H(q_s)}} \geq \frac{c_1 t}{\sqrt{1 + (\log k_\Pi)^{-1} \sum_{s=1}^{t} H(q_s)}} \,.$$

Using this inequality, we have

$$\sum_{t=1}^{T} \eta_t \leq \sum_{t=1}^{T} \frac{1}{\beta'_t} \leq \sum_{t=1}^{T} \frac{1}{c_1 t} \sqrt{1 + \frac{1}{\log k_\Pi} \sum_{s=1}^{t} H(q_s)} \leq \frac{1 + \log T}{c_1} \sqrt{1 + \frac{1}{\log k_\Pi} \sum_{t=1}^{T} H(q_t)} \,. \qquad (33)$$

Further, we have

$$\sum_{t=1}^{T} \eta_t V'_t \leq \max_{s \in [T]} V'_s \sum_{t=1}^{T} \eta_t = \left(\max_{s \in [T]} \max \left\{0, \, \mathrm{opt}'_*(\eta_s)\right\} \right) \sum_{t=1}^{T} \eta_t \leq 3m^2 k^3 \sum_{t=1}^{T} \eta_t \,, \qquad (34)$$

where in the last inequality we used Lemma 5 with $\eta_t \le 1/(2mk^2)$.

(Summing up the above arguments with a self-bounding technique) By bounding the RHS of Lemma 7 using (32), (33), and (34), we have

$$R_T \le 3m^2k^3 \mathbb{E}\left[\frac{1+\log T}{c_1}\sqrt{1+(\log k_\Pi)^{-1}\sum_{t=1}^T H(q_t)}\right] + 2c_1\sqrt{\log k_\Pi}\,\mathbb{E}\left[\sqrt{\sum_{t=1}^T H(q_t)}\right] + \frac{\log k_\Pi}{\eta_1}$$

$$= O\left(mk^{3/2}\sqrt{\log(T)\sum_{t=1}^T H(q_t)} + mk^{3/2}\sqrt{\log(T)\log k_\Pi}\right) + 2mk^2\log k_\Pi\,, \tag{35}$$

where we set $c_1 = \Theta\left(mk^{3/2}\sqrt{\frac{\log T}{\log k_\Pi}}\right)$.

The desired bound is obtained for the adversarial regime, since $\sum_{t=1}^T H(q_t) \le T\log k_\Pi$. We consider the stochastic regime in the following. If $Q(a^*) \le \mathrm{e}$, Lemma 8 implies $\sum_{t=1}^T H(q_t) \le \mathrm{e}\log(k_\Pi T)$ since $k_\Pi T \ge \mathrm{e}$, and otherwise we have $\sum_{t=1}^T H(q_t) \le Q(a^*)\log(k_\Pi T)$. In the former case, we can trivially obtain the desired bound immediately from (35). For the latter case, using the inequality $\sum_{t=1}^T H(q_t) \le Q(a^*)\log(k_\Pi T)$, (34), and Lemma 4 with $c = 1/(2k)$, we have for any $\lambda > 0$ that

$$\begin{aligned} R_T = (1+\lambda)R_T - \lambda R_T &\le \mathbb{E}\left[(1+\lambda)O\left(mk^{3/2}\sqrt{\log(T)\log(k_\Pi T)Q(a^*)}\right) - \frac{\lambda\Delta_{\min}}{2k}Q(a^*)\right] + \lambda C \\ &\le O\left(\frac{(1+\lambda)^2 m^2k^4\log(T)\log(k_\Pi T)}{\lambda\Delta_{\min}}\right) + \lambda C \\ &= O\left(\frac{m^2k^4\log(T)\log(k_\Pi T)}{\Delta_{\min}} + \lambda\left(\frac{m^2k^4\log(T)\log(k_\Pi T)}{\Delta_{\min}} + C\right) + \frac{1}{\lambda}\frac{m^2k^4\log(T)\log(k_\Pi T)}{\Delta_{\min}}\right)\,, \end{aligned} \tag{36}$$

where the second inequality follows from $a\sqrt{x} - bx/2 \le a^2/(2b)$, which holds for any $a, b, x \ge 0$. Taking

$$\lambda = O\left(\sqrt{m^2k^4\log(T)\log(k_\Pi T)\Big/\left(\frac{m^2k^4\log(T)\log(k_\Pi T)}{\Delta_{\min}} + C\right)}\right)$$

completes the proof. ■

B.6. Proof of Lemma 10

Proof Let $a^* = \arg\min_{a\in[k]}\mathbb{E}\left[\sum_{t=1}^T \mathcal{L}_{ax_t}\right]$ be the optimal action in hindsight, where ties are broken so that $a^* \in \Pi$. Note that since action a with $\dim(\mathcal{C}_a) < d-1$ cannot be uniquely optimal, one can see that we can take action $b \in \Pi$ instead of such a with the same loss. We have

$$
\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{t=1}^{T} (\mathcal{L}_{A_t,x_t} - \mathcal{L}_{a^*,x_t})\right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - e_{a^*}, \mathcal{L} e_{x_t}\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \langle q_t - e_{a^*}, \mathcal{L} e_{x_t}\rangle + \sum_{t=1}^{T} \gamma_t \left\langle \frac{1}{k}\mathbf{1} - q_t, \mathcal{L} e_{x_t}\right\rangle\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T} \langle q_t - e_{a^*}, \mathcal{L} e_{x_t}\rangle + \sum_{t=1}^{T} \gamma_t\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{k} q_{t,a} \left(\mathcal{L}_{a x_t} - \mathcal{L}_{a^* x_t}\right) + \sum_{t=1}^{T} \gamma_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{k} q_{t,a} \left(\widehat{y}_{ta} - \widehat{y}_{ta^*}\right) + \sum_{t=1}^{T} \gamma_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle q_t - e_{a^*}, \widehat{y}_t\rangle + \sum_{t=1}^{T} \gamma_t\right],
\end{aligned}
$$

where the inequality follows from the boundedness of $\mathcal{L}$, the fourth equality follows since $a^* \in \Pi$, $q_{t,a} = 0$ for $a \notin \Pi$, and Lemma 2, and the fifth equality follows from the definitions of $\widehat{y}$ and $q_{t,a} = 0$ for $a \notin \Pi$. Combining the above inequality and (7) completes the proof. ■

B.7. Proof of Lemma 11

Proof We first bound the stability term. Using Lemma 15, for any $a' \in \mathcal{A}$ it holds that

$$
\begin{aligned}
\langle \widehat{y}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) &= \langle \widehat{y}_t - \widehat{y}_{ta'}\mathbf{1}, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) \\
&\le \beta_t \sum_{a=1}^{k} q_{t,a} \xi\left(\frac{\widehat{y}_{ta} - \widehat{y}_{ta'}}{\beta_t}\right).
\end{aligned}
$$

We evaluate the RHS of this inequality. As we define p_t by (18), we have $p_{t,a} \ge \gamma_t/k$ for any $a \in [k]$. We first show that $|(\widehat{y}_{ta} - \widehat{y}_{ta'})/\beta_t| \le 1$ for all $a, a' \in [k]$. Let $\tau = \|G\|_\infty$. Recall that $c_{\mathcal{G}} = \max\{1, k\tau\}$. Then we have

$$
\frac{\widehat{y}_t}{\beta_t} = \frac{G(a, \Phi_{ax})}{\beta_t\, p_{t,A_t}} \ge -\frac{\tau}{\beta_t\, p_{t,A_t}}\mathbf{1} \ge -\frac{1}{2}\mathbf{1},
$$

where the inequalities here are element-wise, the first inequality follows from the definition of τ, and in the last inequality we used $p_{t,a} \ge \gamma_t/k \ge c_{\mathcal{G}}/(2\beta_t k) \ge \tau/(2\beta_t)$ for all $a \in [k]$. In a similar manner we have

$$
\frac{\widehat{y}_t}{\beta_t} = \frac{G(a, \Phi_{ax})}{\beta_t\, p_{t,A_t}} \le \frac{\tau}{\beta_t\, p_{t,A_t}}\mathbf{1} \le \frac{1}{2}\mathbf{1}.
$$

These arguments conclude that $|(\widehat{y}_{ta} - \widehat{y}_{ta'})/\beta_t| \le |\widehat{y}_{ta}/\beta_t| + |\widehat{y}_{ta'}/\beta_t| \le 1$ for all $a, a' \in [k]$. Hence, we have

$$\begin{aligned}\langle \widehat{y}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) &\le \min_{a' \in [k]} \beta_t \sum_{a=1}^{k} q_{t,a} \left(\frac{\widehat{y}_{ta} - \widehat{y}_{ta'}}{\beta_t} \right)^2 \\ &= \frac{1}{\beta_t} \min_{a' \in [k]} \sum_{a=1}^{k} q_{t,a} \left(\widehat{y}_{ta} - \widehat{y}_{ta'} \right)^2 \\ &= \frac{1}{\beta_t} \min_{a' \in [k]} \sum_{a \ne a'} q_{t,a} \left(\widehat{y}_{ta} - \widehat{y}_{ta'} \right)^2 , \end{aligned} \tag{37}$$

where the inequality follows from (24). Now, for any $a \in \mathcal{A}$ we have

$$\mathbb{E}\left[\widehat{y}_{ta}^2\right] = \mathbb{E}\left[\left(\frac{G(A_t, \Phi_{A_t x_t})}{p_{t,A_t}} \right)^2 \right] \le \mathbb{E}\left[\sum_{a=1}^{k} p_{t,a} \frac{\|G\|_\infty^2}{p_{t,a}^2} \right] \le \sum_{a=1}^{k} \frac{k \|G\|_\infty^2}{\gamma_t} = \frac{c_{\mathcal{G}}^2}{\gamma_t} , \tag{38}$$

where the last inequality follows from $p_{t,a} \ge \gamma_t / k$. Hence, using (38) it holds that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\beta_t} \min_{a' \in [k]} \sum_{a \ne a'} q_{t,a} \left(\widehat{y}_{ta} - \widehat{y}_{ta'} \right)^2 \right] &\le \mathbb{E}\left[\frac{2}{\beta_t} \min_{a' \in [k]} \sum_{a \ne a'} q_{t,a} \frac{c_{\mathcal{G}}^2}{\gamma_t} \right] \\ &= \mathbb{E}\left[\frac{2 \min_{a' \in [k]} (1 - q_{ta'}) c_{\mathcal{G}}^2}{\beta_t \gamma_t} \right] = \mathbb{E}\left[\frac{2 c_{\mathcal{G}}^2 b_t}{\beta_t \gamma_t} \right] . \end{aligned} \tag{39}$$

Combining (37) and (39) yields

$$\mathbb{E}[\langle \widehat{y}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t)] \le \mathbb{E}\left[\frac{2 c_{\mathcal{G}}^2 b_t}{\beta_t \gamma_t} \right] ,$$

which completes the proof. ■

B.8. Proof of Proposition 13

Proof Note that the penalty term can be rewritten as

$$\begin{aligned}&\sum_{t=1}^{T} \left(\psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1}) \right) + \psi_{T+1}(e_{a^*}) - \psi_1(q_1) \\ &= \sum_{t=1}^{T} (\beta_t - \beta_{t+1}) \left(-H(q_{t+1}) \right) + \beta_1 H(q_1) = \sum_{t=1}^{T} (\beta_{t+1} - \beta_t)\, a_{t+1} + \beta_1 a_1 , \end{aligned}$$

where we recall that the definition of a_t in (17). Combining this with Lemmas 10 and 11, we have

$$R_T \le \underbrace{\sum_{t=1}^{T} \left(\gamma_t' + \frac{c_{\mathcal{G}}}{2\beta_t} \right)}_{\text{transformation term}} + \underbrace{\sum_{t=1}^{T} \frac{2 c_{\mathcal{G}}^2 b_t}{\beta_t \gamma_t}}_{\text{stability term}} + \underbrace{\sum_{t=1}^{T} \left((\beta_{t+1} - \beta_t) a_{t+1} \right) + \beta_1 a_1}_{\text{penalty term}} , \tag{40}$$

where the first, second, and remaining terms correspond to the transformation, stability, and penalty terms, respectively. We bound each term of the RHS in (40) in the following.

Note that $b_t \leq 1$ and

$$b_t = 1 - \max_{a \in [k]} q_{t,a} \leq - \max_{a \in [k]} q_{t,a} \log \left(\max_{a' \in [k]} q_{t,a'} \right) \leq - \sum_{a \in [k]} q_{t,a} \log q_{t,a} = a_t \leq \log k_\Pi \,, \tag{41}$$

where the first inequality follows from the inequality $1 - x \leq -x \log x$ for $x > 0$. We define $z_t = \frac{a_{t+1} b_t}{\gamma'_t}$ and $Z_t = \sum_{s=1}^t z_s$.

(Bounding the penalty term) From the definition of γ'_t, we can bound z_t from below as

$$z_t = \frac{a_{t+1} b_t}{\gamma'_t} = \frac{4 a_{t+1}}{c_1} \left(c_1 + B_t^{1/3} \right) \geq 4 a_{t+1} \geq 4 b_{t+1} \,, \tag{42}$$

where the second inequality follows from $b_t \leq a_t$ in (41). Further, we can bound z_t from above as

$$z_t = \frac{4 a_{t+1}}{c_1} \left(c_1 + B_t^{1/3} \right) \leq 4 \left(c_1 + B_t^{1/3} \right) \leq 4 \left\{ c_1 + \left(b_1 + \sum_{s=1}^{t-1} z_s \right)^{1/3} \right\} \leq 8 \left(c_1 + Z_{t-1} \right) \,, \tag{43}$$

where the first inequality follows from $a_{t+1} \leq \log k_\Pi$ and $c_1 \geq \log k_\Pi$, and the second inequality follows from $B_t = b_1 + \sum_{s=1}^{t-1} b_{s+1} \leq b_1 + \sum_{s=1}^{t-1} z_s$, and the last inequality follows from $b_1 \leq 1 \leq c_1$. From this, since β_t satisfies $\beta_{t+1} - \beta_t = \frac{z_t}{a_{t+1}} \frac{c_2}{(c_1 + Z_{t-1})^{1/2}}$, we can bound the penalty term in (40) as

$$\begin{aligned} \sum_{t=1}^T (\beta_{t+1} - \beta_t) a_{t+1} &= c_2 \sum_{t=1}^T \frac{z_t}{\sqrt{c_1 + Z_{t-1}}} = 5 c_2 \sum_{t=1}^T \frac{Z_t - Z_{t-1}}{4\sqrt{c_1 + Z_{t-1}} + \sqrt{c_1 + Z_{t-1}}} \\ &< 5 c_2 \sum_{t=1}^T \frac{Z_t - Z_{t-1}}{\sqrt{c_1 + Z_t} + \sqrt{c_1 + Z_{t-1}}} = 5 c_2 \sum_{t=1}^T \left(\sqrt{c_1 + Z_t} - \sqrt{c_1 + Z_{t-1}} \right) \leq 5 c_2 \sqrt{Z_T} \,, \end{aligned} \tag{44}$$

where the first equality follows from the definitions of β_t and z_t, and the first inequality follows since

$$\sqrt{c_1 + Z_t} \leq \sqrt{c_1 + Z_{t-1}} + \sqrt{z_t} < 4 \sqrt{c_1 + Z_{t-1}} \,,$$

where the last inequality follows from (43).

(Bounding the stability term and transformation terms) We define $w_t = \frac{b_t}{\gamma'_t}$ and $W_t = \sum_{s=1}^t w_s$. From the definition of γ'_t, we have

$$w_t = \frac{b_t}{\gamma'_t} = 4 \left(1 + \frac{1}{c_1} B_t^{1/3} \right) \geq 4 \,. \tag{45}$$

Using $b_t \leq 1$, we can confirm that w_t satisfies

$$w_1 \leq 8 \,, \quad w_{t+1} = 4 \left(1 + \frac{1}{c_1} B_{t+1}^{1/3} \right) \leq \left(1 + \frac{1}{c_1} (B_t + 1)^{1/3} \right) \leq 2 w_t \,, \quad w_t \leq 4(1 + t^{1/3}) \,. \tag{46}$$

Then β_t can be bounded as

$$\begin{aligned}\beta_t &\geq c_2 + c_2 \sum_{s=1}^{t-1} \frac{w_s}{\sqrt{c_1 + Z_{s-1}}} \geq \frac{c_2}{\sqrt{c_1 + Z_t}} \left(1 + \sum_{s=1}^{t-1} w_s\right) \\ &= \frac{c_2}{\sqrt{c_1 + Z_t}} \left(1 + W_{t-1}\right) \qquad (47) \\ &\geq \frac{c_2 t}{\sqrt{c_1 + Z_t}} \,, \qquad (48)\end{aligned}$$

where the second inequality follows from (45).

Using the above inequalities, we can bound the stability term in (40) as

$$\begin{aligned}\sum_{t=1}^{T} \frac{b_t}{\gamma_t \beta_t} \leq \sum_{t=1}^{T} \frac{b_t}{\gamma'_t \beta_t} &\leq \sum_{t=1}^{T} \frac{\sqrt{c_1 + Z_t}}{c_2} \frac{w_t}{1 + W_{t-1}} \leq \frac{\sqrt{c_1 + Z_T}}{c_2} \sum_{t=1}^{T} \frac{w_t}{1 + W_{t-1}} \\ &\leq O\left(\frac{\sqrt{c_1 + Z_T}}{c_2} \log\left(1 + W_T\right)\right) \leq O\left(\frac{\sqrt{c_1 + Z_T}}{c_2} \log T\right) \,, \qquad (49)\end{aligned}$$

where the first inequality follows from (47), the last inequality follows from (46), and the fourth inequality can be shown by taking the sum of the following inequality:

$$\log(1 + W_t) - \log(1 + W_{t-1}) = \log \frac{1 + W_t}{1 + W_{t-1}} = \log\left(1 + \frac{w_t}{1 + W_{t-1}}\right) \geq \frac{1}{2} \cdot \frac{w_t}{1 + W_{t-1}} \,,$$

where the inequality follows from the fact that $\log(1 + x) \geq \frac{1}{2}x$ holds for any $x \in [0, 2]$ and that (46) implies $\frac{w_t}{1+W_{t-1}} \leq \frac{w_t}{1+w_t/2} \leq 2$ for all $t \in [T]$.

Using (48), we can bound the second part of the transformation term in (40) as

$$\sum_{t=1}^{T} \frac{1}{\beta_t} \leq \sum_{t=1}^{T} \frac{\sqrt{c_1 + Z_t}}{c_2 t} \leq \frac{\sqrt{c_1 + Z_T}}{c_2} \sum_{t=1}^{T} \frac{1}{t} = O\left(\frac{\sqrt{c_1 + Z_T}}{c_2} \log T\right) \,. \qquad (50)$$

In addition, from the definition of γ'_t, we can bound the remaining part of the transformation term in (40) as

$$\sum_{t=1}^{T} \gamma'_t = \frac{c_1}{4} \sum_{t=1}^{T} \frac{b_t}{c_1 + B_t^{1/3}} \leq \frac{3c_1}{8} \sum_{t=1}^{T} \left(B_t^{2/3} - B_{t-1}^{2/3}\right) \leq \frac{3c_1}{8} B_T^{2/3} \,, \qquad (51)$$

where the first inequality follows from $y^{2/3} - x^{2/3} \geq \frac{2}{3}(y-x)y^{-1/3}$, which holds for any $y \geq x > 0$. Combining (44), (49), (50), and (51), we can bound the right-hand side of (40) as

$$\begin{aligned}
&\sum_{t=1}^{T} \left(\gamma_t + \frac{2c_{\mathcal{G}}^2 b_t}{\gamma_t \beta_t} + (\beta_{t+1} - \beta_t) a_{t+1} \right) + \beta_1 a_1 \\
&= \sum_{t=1}^{T} \left(\gamma_t' + \frac{c_{\mathcal{G}}}{2\beta_t} + \frac{2c_{\mathcal{G}}^2 b_t}{\gamma_t \beta_t} + (\beta_{t+1} - \beta_t) a_{t+1} \right) + \beta_1 a_1 \\
&= O\left(c_1 B_T^{2/3} + \left(\frac{c_{\mathcal{G}}^2 \log T}{c_2} + c_2 \right) \sqrt{c_1 + Z_T} + \beta_1 a_1 \right) \\
&= O\left(c_1 B_T^{2/3} + \left(\frac{c_{\mathcal{G}}^2 \log T}{c_2} + c_2 \right) \sqrt{c_1 + \sum_{t=1}^{T} \frac{a_{t+1}}{c_1} \left(c_1 + B_t^{1/3} \right)} + \beta_1 a_1 \right) \\
&= O\left(c_1 B_T^{2/3} + \frac{1}{\sqrt{c_1}} \left(\frac{c_{\mathcal{G}}^2 \log T}{c_2} + c_2 \right) \sqrt{c_1^2 + (\log k_\Pi + A_T) \left(c_1 + B_T^{1/3} \right)} + \beta_1 \log k_\Pi \right) ,
\end{aligned}$$

where in the third inequality we used (42) and in the last equality we used $a_{T+1} = O(\log k_\Pi)$. ■

B.9. Proof of Theorem 9

Proof We define c_1 and c_2 by

$$c_1 = \Theta\left(\left(c_{\mathcal{G}}^2 \log(T) \log(k_\Pi T) \right)^{1/3} \right) \quad \text{and} \quad c_2 = \Theta\left(\sqrt{c_{\mathcal{G}}^2 \log T} \right) , \tag{52}$$

which implies that $\tilde{c} = c_1 / \sqrt{\log(k_\Pi T)}$. We have

$$B_T = \sum_{t=1}^{T} \left(1 - \max_{a \in \Pi} q_{t,a} \right) \leq \sum_{t=1}^{T} (1 - q_{t,a^*}) = Q(a^*) . \tag{53}$$

We first consider the adversarial regime. Since $A_T \leq T \log k_\Pi$ and $B_T \leq T$, using Proposition 13 we have

$$\begin{aligned}
R_T &= O\left(c_1 T^{2/3} + \tilde{c} \sqrt{c_1^2 + (\log k_\Pi + T \log k_\Pi)(c_1 + T^{1/3})} + \beta_1 \log k_\Pi \right) \\
&= O\left(\left(c_1 + \tilde{c} \sqrt{\log k_\Pi} \right) T^{2/3} + \sqrt{\frac{\log k_\Pi}{\log(k_\Pi T)}} c_1^{3/2} T^{1/2} + \frac{c_1^2}{\sqrt{\log(k_\Pi T)}} + \beta_1 \log k_\Pi \right) .
\end{aligned} \tag{54}$$

We next consider the adversarial regime with a self-bounding constraint. When $Q(a^*) \leq c_1^3$ we can show that the obtained bound is smaller than the desired bound as follows. When $Q(a^*) \leq \mathrm{e} \leq c_1^3$, using Lemma 8 and (53), we have $A_T \leq \mathrm{e} \log(k_\Pi T)$ and $B_T \leq \mathrm{e}$. Hence, from Proposition 13, we have

$$R_T = O\left(c_1 + \tilde{c} \sqrt{c_1^2 + \log(k_\Pi T) c_1} + \beta_1 \log k_\Pi \right) = O\left(\frac{c_1^2}{\sqrt{\log(k_\Pi T)}} + \beta_1 \log k_\Pi \right) = O\left(c_1^3 \right) .$$

When $\mathrm{e} < Q(a^*) \le c_1^3$, using Lemma 8 and (53) we have $A_T \le c_1^3 \log(k_\Pi T)$ and $B_T \le c_1^3$. Hence, from Proposition 13, we have

$$\begin{aligned} R_T &= O\left(c_1^3 + \tilde{c}\sqrt{c_1^2 + \left(\log k_\Pi + c_1^3 \log(k_\Pi T)\right) c_1} + \beta_1 \log k_\Pi \right) \\ &= O\left(c_{\mathcal{G}}^2 \log(T) \log(k_\Pi T) \right) = O\left(c_1^3 \right) . \end{aligned}$$

Hence, we only need to consider the case of $Q(a^*) > c_1^3$ in the following. Since $Q(a^*) \ge \mathrm{e}$ we have $A_T \le Q(a^*) \log(k_\Pi T)$. Using Proposition 13 with this inequality, Lemma 8, and (53), we have

$$\begin{aligned} R_T &= O\left(\mathbb{E}\left[c_1 Q(a^*)^{2/3} + \tilde{c}\sqrt{c_1^2 + \left(\log k_\Pi + Q(a^*) \log(k_\Pi T)\right)\left(c_1 + Q(a^*)^{1/3}\right)} \right] + \beta_1 \log k_\Pi \right) \\ &\le O\left(\mathbb{E}\left[c_1 Q(a^*)^{2/3} + \tilde{c}\sqrt{Q(a^*) \log(k_\Pi T) Q(a^*)^{1/3}} \right] \right) \\ &\le O\left(\left(c_1 + \tilde{c}\sqrt{\log(k_\Pi T)} \right) \bar{Q}(a^*)^{2/3} \right) , \end{aligned} \tag{55}$$

where the first inequality follows from $Q(a^*) > c_1^3$, and the second inequality follows from Jensen's inequality. Hence, by (54) and (55), there exists $\widehat{c} = O\left(c_1 + \tilde{c}\sqrt{\log(k_\Pi T)} \right)$ satisfying and $R_T \le \widehat{c}\,\bar{Q}(a^*)^{2/3}$ for the adversarial regime with a self-bounding constraint and $R_T \le \widehat{c}\,T^{2/3}$ for the adversarial regime.

Now, by recalling the definitions of c_1 and c_2 in (52), we have

$$\begin{aligned} \widehat{c} &= O\left(\left(c_{\mathcal{G}}^2 \log(T) \log(k_\Pi T) \right)^{1/3} + \frac{1}{\sqrt{c_1}} \left(\frac{c_{\mathcal{G}}^2 \log T}{c_2} + c_2 \right) \sqrt{\log(k_\Pi T)} \right) \\ &= O\left(\left(c_{\mathcal{G}}^2 \log(T) \log(k_\Pi T) \right)^{1/3} \right) , \end{aligned} \tag{56}$$

which gives the desired bounds for the adversarial regime.

For the adversarial regime with a self-bounding constraint, from the above inequality $R_T \le \widehat{c}\,\bar{Q}(a^*)^{2/3}$ and Lemma 4 with $c = 1/2 \le 1 - \gamma_t$, we have for any $\lambda \in (0, 1]$ that

$$\begin{aligned} R_T &= (1+\lambda) R_T - \lambda R_T \le (1+\lambda)\widehat{c} \cdot \bar{Q}(a^*)^{2/3} - \frac{\lambda}{2} \Delta_{\min} \bar{Q}(a^*) + \lambda C \\ &\le O\left(\frac{(1+\lambda)^3 \widehat{c}^3}{\lambda^2 \Delta_{\min}^2} \right) + \lambda C = O\left(\left(1 + \frac{1}{\lambda^2} \right) \frac{\widehat{c}^3}{\Delta_{\min}^2} \right) + \lambda C , \end{aligned} \tag{57}$$

where the first inequality follows from the inequality $a x^{2/3} - b(x/2) \le 16a^3/(27b^2)$ for $a, b > 0$, and the last equality follows since $\lambda \in (0, 1]$. Combining (56) and (57), and taking $\lambda = O\left(\frac{c_{\mathcal{G}}^2 \log(T) \log(k_\Pi T)}{C \Delta_{\min}^2} \right)$, we have the desired result for the adversarial regime with a self-bounding constraint. ■

Appendix C. Regret Bounds when the Optimization Problem is Not Exactly Solved

This section discusses the regret bound when the optimization problem (11) is not exactly solved, on which a similar discussion is given in Lattimore and Szepesvári (2020a, Chapter 37). We say that the optimization problem (11) can be solved with precision $\epsilon \geq 0$, if we can obtain $G \in \mathcal{H}$ and $p \in \mathcal{P}'_k(q)$ such that

$$\max_{x \in [d]} \left[\frac{(p-q)^\top \mathcal{L} e_x + \mathrm{bias}_q(G; x)}{\eta} + \frac{1}{\eta^2} \sum_{a=1}^{k} p_a \left\langle q, \xi \left(\frac{\eta G(a, \Phi_{ax})}{p_a} \right) \right\rangle \right] \leq \mathrm{opt}'_q(\eta) + \epsilon \,.$$

Then if we run Algorithm 1 solving (11) with precision ϵ, one can see that we can obtain the following regret bounds. For the adversarial regime with a (Δ, C, T) self-bounding constraint, we have

$$R_T = O\left(\frac{\left(mk^2 + \epsilon^2/(mk)\right)^2 \log(T) \log(k_\Pi T)}{\Delta_{\min}} + \sqrt{\frac{C\left(mk^2 + \epsilon^2/(mk)\right)^2 \log(T) \log(k_\Pi T)}{\Delta_{\min}}} \right) ,$$

and for the adversarial regime, we have

$$R_T = O\left(mk^{3/2} \sqrt{T \log(T) \log k_\Pi} + \epsilon \, \frac{\sqrt{T \log(k_\Pi) \log(T)}}{mk^{3/2}} \right) .$$

Here, we give an overview of the analysis. Considering that the optimization problem in (11) can be solved with precision $\epsilon \geq 0$, the RHS of (30) can be replaced with $3m^2k^3 + \epsilon$. Then a similar analysis as the proof of Theorem 6 leads to

$$R_T \leq O\left(\left(mk^{3/2} + \frac{\epsilon}{mk^{3/2}} \right) \sqrt{\log(T) \sum_{t=1}^{T} H(q_t)} \right) .$$

Using $\sum_{t=1}^{T} H(q_t) \leq T \log k_\Pi$ gives the bound for the adversarial regime. Replacing m^2k^4 with $\left(mk^2 + \frac{\epsilon}{mk}\right)^2$ in (36) and appropriately choose λ (note that we can take λ depending on ϵ), we obtain the desired bound for the adversarial regime with a self-bounding constraint.