# Exploration by Optimization with Hybrid Regularizers: Logarithmic Regret with Adversarial Robustness in Partial Monitoring

Taira Tsuchiya [1] [2]   Shinji Ito [1] [2] [3]   Junya Honda [4] [2]

## Abstract

Partial monitoring is a generic framework of online decision-making problems with limited feedback. To make decisions from such limited feedback, it is necessary to find an appropriate distribution for exploration. Recently, a powerful approach for this purpose, *exploration by optimization* (ExO), was proposed, which achieves optimal bounds in adversarial environments with follow-the-regularized-leader for a wide range of online decision-making problems. However, a naive application of ExO in stochastic environments significantly degrades regret bounds. To resolve this issue in locally observable games, we first establish a new framework and analysis for ExO with a hybrid regularizer. This development allows us to significantly improve existing regret bounds of best-of-both-worlds (BOBW) algorithms, which achieves nearly optimal bounds both in stochastic and adversarial environments. In particular, we derive a stochastic regret bound of $O(\sum_{a \neq a^*} k^2 m^2 \log T / \Delta_a)$, where $k$, $m$, and $T$ are the numbers of actions, observations and rounds, $a^*$ is an optimal action, and $\Delta_a$ is the suboptimality gap for action $a$. This bound is roughly $\Theta(k^2 \log T)$ times smaller than existing BOBW bounds. In addition, for globally observable games, we provide a new BOBW algorithm with the first $O(\log T)$ stochastic bound.

## 1. Introduction

This paper investigates partial monitoring (PM), a comprehensive model for dealing with online decision-making

[1]Department of Mathematical Informatics, The University of Tokyo, Tokyo, Japan [2]RIKEN AIP, Tokyo, Japan [3]NEC Corporation, Kanagawa, Japan (affiliation upon submission) [4]Department of Systems Science, Kyoto University, Kyoto, Japan. Correspondence to: Taira Tsuchiya <tsuchiya@mist.i.u-tokyo.ac.jp>.

problems with limited feedback (Rustichini, 1999; Piccolboni & Schindelhauer, 2001). A PM game with $k$-actions and $d$-outcomes, denoted by $\mathcal{G} = (\mathcal{L}, \Phi)$, is defined by a loss matrix $\mathcal{L} \in [0,1]^{k \times d}$ and feedback matrix $\Phi \in \Sigma^{k \times d}$, where $\Sigma$ is a set of feedback symbols.

This game unfolds in a sequential manner, involving interactions between a learner and an adversary for $T$ rounds. At the start of the game, the learner and adversary observe $\mathcal{L}$ and $\Phi$. At every round $t \in [T] := \{1, \dots, T\}$, the adversary chooses an outcome $x_t \in [d]$. The learner then chooses an action $A_t \in [k]$ without knowing $x_t$, incurs an unobserved loss $\mathcal{L}_{A_t x_t}$, and receives a feedback symbol $\sigma_t = \Phi_{A_t x_t}$, where $A_{ij}$ is the $(i,j)$-element of matrix $A$. The learner does not have access to outcome and loss, but only to feedback symbols. The learner aims to minimize the cumulative loss over all the rounds with such limited feedback.

To measure the performance of the learner, we consider the (pseudo-)regret $\mathrm{Reg}_T$, which is defined as the difference between the cumulative loss of the learner and the single optimal action $a^*$, which is determined in hindsight. More specifically, letting $\mathrm{Reg}_T(a) = \mathbb{E}\big[\sum_{t=1}^{T}(\mathcal{L}_{A_t x_t} - \mathcal{L}_{a x_t})\big]$, we define the regret by $\mathrm{Reg}_T = \mathrm{Reg}_T(a^*)$ for optimal action $a^* = \arg\min_{a \in [k]} \mathbb{E}\big[\sum_{t=1}^{T} \mathcal{L}_{a x_t}\big]$.

The difficulty of PM games is determined based on its loss matrix and feedback matrix (Bartók et al., 2011; Lattimore & Szepesvári, 2019a). PM games fall into trivial, easy, hard, and hopeless games, for which its minimax regrets are $0$, $\Theta(\sqrt{T})$, $\Theta(T^{2/3})$, and $\Theta(T)$, respectively. PM games for which the regret can be $O(\sqrt{T})$ and $O(T^{2/3})$ are called locally and globally observable, respectively.

PM has been mainly studied in stochastic environments and adversarial environments. In stochastic environments, a sequence of outcomes $x_1, \dots, x_T \in [d]$ is drawn independently and identically from an unknown distribution $\nu^*$. In contrast, in adversarial environments, outcomes are generated in an arbitrary manner and $x_t$ can depend on the past history of actions $(A_s)_{s=1}^{t-1}$.

For the adversarial environments, many algorithms have been developed to achieve the aforementioned minimax regrets and most of the algorithms are based on *follow-*

*Table 1.* Regret upper bounds for partial monitoring. Here the number of rounds is denoted by $T$, the number of actions by $k$, the number of outcomes by $d$, the maximum number of distinct symbols per action by $m$, the minimum suboptimality gap by $\Delta$, the game-dependent constant by $c_{\mathcal{G}}$, and the corruption level by $C$. The variables $\mathcal{R}^{\text{loc}}$ and $\mathcal{R}^{\text{glo}}$ are the regret upper bounds of the proposed algorithms in stochastic environments for locally and globally observable games, respectively. TSPM is the bound by Tsuchiya et al. (2020) for the linearized version of PM; refer to the paper for the definition of $\Lambda'$. ExpPM is by Lattimore & Szepesvári (2020b). PM-DEMD is by Komiyama et al. (2015), where $D(\nu^*)$ is a distribution-dependent constant. ExpPM-BOBW is by Tsuchiya et al. (2023a).

| Observability | Algorithm | Stochastic | Adversarial | Corrupted |
|---|---|---|---|---|
| Locally observable | TSPM | $O\left(\dfrac{k^2 m d \log T}{\Lambda'^2}\right)$ | – | – |
| | ExpPM | – | $O\left(k^{3/2} m \sqrt{T \log k}\right)$ | – |
| | ExpPM-BOBW | $O\left(\dfrac{k^4 m^2 \log T \log(kT)}{\Delta}\right)$ | $O\left(k^{3/2} m \sqrt{T \log T \log k}\right)$ | $\mathcal{R}^{\text{loc}} + \sqrt{C\mathcal{R}^{\text{loc}}}$ |
| | **Ours (Theorem 5)** | $O\left(\displaystyle\sum_{a \neq a^*} \dfrac{k^2 m^2 \log T}{\Delta_a}\right)$ | $O\left(k^{3/2} m \sqrt{T \log T}\right)$ | $\mathcal{R}^{\text{loc}} + \sqrt{C\mathcal{R}^{\text{loc}}}$ |
| Globally observable | PM-DMED | $O(D(\nu^*) \log T)$ | – | – |
| | ExpPM | – | $O\left((c_{\mathcal{G}} T)^{2/3} (\log k)^{1/3}\right)$ | – |
| | ExpPM-BOBW | $O\left(\dfrac{c_{\mathcal{G}}^2 \log T \log(kT)}{\Delta^2}\right)$ | $O\left((c_{\mathcal{G}} T)^{2/3} (\log T \log(kT))^{1/3}\right)$ | $\mathcal{R}^{\text{glo}} + (C^2 \mathcal{R}^{\text{glo}})^{1/3}$ |
| | **Ours (Theorem 8)** | $O\left(\dfrac{c_{\mathcal{G}}^2 k \log T}{\Delta^2}\right)$ | $O\left((c_{\mathcal{G}} T)^{2/3} (\log T)^{1/3}\right)$ | $\mathcal{R}^{\text{glo}} + (C^2 \mathcal{R}^{\text{glo}})^{1/3}$ |

*the-regularized-leader* (FTRL) (Piccolboni & Schindelhauer, 2001; Cesa-Bianchi et al., 2006; Lattimore & Szepesvári, 2020b). In particular, Lattimore & Szepesvári (2020b) propose a strong approach called *exploration by optimization* (ExO). They prove that FTRL with the (negative) Shannon entropy (corresponding to the well-known Exp3 algorithm by Freund & Schapire 1997; Auer et al. 2002) combined with ExO achieves regret bounds of $O(k^{3/2} m \sqrt{T \log k})$ for (non-degenerate) locally observable games and $O((c_{\mathcal{G}} T)^{2/3} (\log k)^{1/3})$ for globally observable games. Here, $m \leq \min\{|\Sigma|, d\}$ is the maximum number of distinct observations per action and $c_{\mathcal{G}}$ is a game-dependent constant defined in Section 5. In the stochastic environments, there are also several algorithms; Starting from Bartók et al. (2012), they include algorithms with optimal $O(\log T)$ regret upper bounds (Komiyama et al., 2015; Tsuchiya et al., 2020).

Although there are many algorithms for stochastic and adversarial environments, it is often the case in real-world problems that the underlying environment is unknown. Hence, algorithms that achieve optimality in both stochastic and adversarial environments without knowing the underlying environment are desired. Such algorithms are called *best-of-both-worlds* (BOBW) algorithms (Bubeck & Slivkins, 2012), which have been widely investigated in recent years in a variety of online decision-making problems, *e.g.,* in Gaillard et al. (2014); Wei & Luo (2018); Zimmert & Seldin (2021). BOBW algorithms based on

FTRL perform well also in stochastic environments with adversarial corruptions, an intermediate regime between stochastic and adversarial environments. Extending the BOBW guarantee to this environment is standard and detailed in Appendix C.

The development of BOBW algorithms has recently extended from bandit problems to the PM problem (Tsuchiya et al., 2023a;c). Tsuchiya et al. (2023a) consider both locally observable and globally observable games. For locally observable games, they construct a BOBW algorithm by extending the ExO approach. As ExO was originally developed for adversarial environments, one could not derive a favorable regret lower bound that is needed for the *self-bounding technique* (Gaillard et al., 2014; Zimmert & Seldin, 2021), a standard technique for proving BOBW guarantees in FTRL (detailed in Section 2). To address this issue, the authors introduce *a restriction of a feasible region* of action selection probability in ExO to derive a lower bound of the regret. This technique allows us to provide an algorithm with a BOBW regret bound of $O(k^{3/2} m \sqrt{T \log k \log T})$ in adversarial environments and of $O(k^4 m^2 \log T \log(kT)/\Delta)$ for the minimum suboptimality gap $\Delta = \min_{a \neq a^*} \Delta_a$, where $\Delta_a = \mathbb{E}_{x_t \sim \nu^*}[\mathcal{L}_{a x_t} - \mathcal{L}_{a^* x_t}]$ is the suboptimality gap for action $a \in [k]$, in stochastic environments.

However, the bound for stochastic environments by their BOBW algorithm is significantly worse than the bounds solely for stochastic environments. First of all, their bound

scales with $\log^2 T$ rather than $\log T$, which seems to be due to the use of the Shannon entropy. In fact, there is no known algorithm that can achieve an $O(\log T)$ bound with the Shannon entropy (see *e.g.,* Jin et al. 2023). Furthermore, their bound exhibits significant dependence on the number of actions $k$ and dependence on the smallest suboptimality gap $\Delta$ rather than action-wise suboptimality gap $(\Delta_a)_{a \neq a^*}$. (Due to space constraints, a further discussion of related work is deferred to Appendix B.)

**Contributions of this Paper**   In order to improve these highly suboptimal dependencies on $k$, $T$, and $\Delta$, we introduce two major techniques to the PM problem.

The first is the introduction of FTRL with a hybrid regularizer consisting of the log-barrier and complement negative Shannon entropy. In contrast to the Shannon entropy regularizer, the log-barrier regularizer is known to provide $O(\log T)$ regret bounds (Wei & Luo, 2018; Ito, 2021b), and combining it with the complement negative Shannon entropy (Ito, 2021a; Ito et al., 2022b; Tsuchiya et al., 2023a) yields the regret bound in terms of action-wise suboptimality gaps $(\Delta_a)_{a \neq a^*}$, in relatively simple settings such as multi-armed bandits and combinatorial semi-bandits.

This hybrid regularizer, however, cannot be employed for PM in a naive manner. This is because PM algorithms need to deal with complex feedback and suffer a very large regret without the ExO framework (Lattimore & Szepesvári, 2020b), and existing ExO analysis heavily relies on the specific properties of the Shannon entropy regularizer. Hence, as a second technique, we analyze the optimal value of ExO when combined with the hybrid regularizer. A new analysis to bound the optimal value allows us to optimize the ExO objective over the feasible region that is independent of the number of actions $k$. This improves the existing suboptimality in terms of $k$ (see Lemma 4).

Consequently, we significantly improve the regret upper bound in locally observable games. In particular, we prove that the regret of the proposed algorithm is bounded by

$$\mathsf{Reg}_T = O\left(\sum_{a \neq a^*} \frac{k^2 m^2 \log T}{\Delta_a}\right)$$

in stochastic environments and is bounded by $\mathsf{Reg}_T = O(k^{3/2}m\sqrt{T \log T})$ in adversarial environments. The bound in stochastic environments is roughly $\Theta(k^2 \log T)$ times smaller than the existing regret bound with the BOBW guarantee. The bound for adversarial environments is $O(\sqrt{\log k})$ times smaller than the existing regret bound with the BOBW guarantee. In addition, for globally observable games, we propose the first BOBW algorithm that achieves an $O(\log T)$ regret upper bound in stochastic environments by making a sacrifice of dependence on $k$.

A comparison of our bounds with existing regret bounds in PM is summarized in Table 1. Note that comparisons with existing algorithms designed exclusively for stochastic environments, TSPM and PM-DMED in Table 1, are difficult for the following reasons. The upper bound of TSPM is for linearized variants of PM, called linear PM. The upper bound of PM-DMED should be better than our algorithm as it is asymptotically optimal, but $D(\nu^*)$ is expressed by a complex optimization problem, and its dependencies on $k$, $m$, and $d$ are unknown.

## 2. Preliminaries

This section provides basic concepts for PM and introduces FTRL, the core framework of proposed algorithms.

**Notation**   Let $e_i \in \{0, 1\}^k$ be the $i$-th standard basis in $\mathbb{R}^k$ and $\mathbf{1}$ be the all-one vector. A scalar $x_i$ is the $i$-th element of vector $x$, and $A_{ij}$ is the $(i, j)$-element of matrix $A$. For vector $x$, let $\|x\|_p$ for $p \in [1, \infty]$ be the $\ell_p$-norm and write $\|x\| = \|x\|_2$. For matrix $A$, we let $\|A\|_\infty = \max_{i,j}|A_{ij}|$ be the maximum norm. The $(k-1)$-dimensional probability simplex is denoted by $\mathcal{P}_k = \{p \in [0, 1]^k \colon \|p\|_1 = 1\}$. Let $\mathrm{int}(A)$ be the interior of set $A$ and $\mathrm{dom}\,\psi$ be the domain of function $\psi$. Given differentiable convex function $\psi$, the Bregman divergence induced by $\psi$ from $y$ to $x$ is defined by $D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle$, and the convex conjugate of $\psi$ by $\psi^*(x) = \sup_{a \in \mathbb{R}^k}\{\langle x, a \rangle - \psi(a)\}$. Table 2 in Appendix A summarizes the notation used in this paper.

We assume that the optimal action $a^*$ is unique in stochastic environments, and thus $\Delta_a > 0$ for all $a \neq a^*$. This assumption is standard in the literature to develop BOBW algorithms (Gaillard et al., 2014; Wei & Luo, 2018; Zimmert & Seldin, 2021; Tsuchiya et al., 2023a).

### 2.1. Basic Concepts in Partial Monitoring

Here, we introduce basic concepts used in PM. We consider a PM game $\mathcal{G} = (\mathcal{L}, \Phi)$, and denote the maximum number of distinct symbols in a single row of $\Phi \in \Sigma^{k \times d}$ over all rows by $m \leq |\Sigma|$. We use $\ell_a = e_a^\top \mathcal{L} \in \mathbb{R}^d$ to denote the $a$-th row of $\mathcal{L}$.

Two 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* is the set of probability vectors in $\mathcal{P}_d$ for which action $a$ is optimal, that is, $\mathcal{C}_a = \{u \in \mathcal{P}_d \colon \max_{b \in [k]}(\ell_a - \ell_b)^\top u \leq 0\}$. Each cell is a convex closed polytope. We denote the dimension of the affine hull of $\mathcal{C}_a$ by $\dim(\mathcal{C}_a)$. Action $a$ is *dominated* if $\mathcal{C}_a = \emptyset$. Non-dominated action $a$ is *Pareto optimal* if $\dim(\mathcal{C}_a) = d - 1$ and *degenerate* if $\dim(\mathcal{C}_a) < d - 1$. We denote the set of Pareto optimal

actions by $\Pi$. Two Pareto optimal actions $a, b \in \Pi$ are *neighbors* if $\dim(\mathcal{C}_a \cap \mathcal{C}_b) = d - 2$. The notion of neighboring will be used to classify PM games based on their difficulty in Section 2.1.1. Using this neighborhood relation, we can construct the undirected graph that is known to be connected (Lattimore & Szepesvári 2020a, Lemma 37.7). This fact will be used for loss difference estimations between Pareto optimal actions in Section 2.1.2.

A PM game is *non-degenerate* if the game does not contain degenerate actions. From hereon, we assume that PM game $\mathcal{G}$ is non-degenerate and contains no duplicate actions.

### 2.1.1. OBSERVABILITY

The above notion of neighborhood relations is used to define *observability* that characterizes the difficulty of PM games.

**Definition 1.** Neighbouring actions $a$ and $b$ are *globally observable* if there exists $w_{e(a,b)} \colon [k] \times \Sigma \to \mathbb{R}$ such that

$$\sum_{c=1}^{k} w_{e(a,b)}(c, \Phi_{cx}) = \mathcal{L}_{ax} - \mathcal{L}_{bx} \text{ for all } x \in [d], \quad (1)$$

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

Locally observable games are globally observable, In the remaining sections, we assume that $\mathcal{G}$ is globally observable.

### 2.1.2. ESTIMATING LOSS DIFFERENCES

Using the definitions up to this point, we explain how to estimate losses in PM. In PM it is common to estimate the loss difference between actions rather than estimating the loss itself. Let $\mathcal{H}$ be the set of all functions from $[k] \times \Sigma$ to $\mathbb{R}^k$. Let $\mathcal{T}$ be an in-tree over $\Pi$ induced by the neighborhood relations with an arbitrarily chosen root. Then, from the fact that the undirected graph induced by $\Pi$ is connected, it follows that loss differences between Pareto optimal actions can be estimated as follows.

**Lemma 2** (Lattimore & Szepesvári 2020b, Lemma 4). *For any globally observable game, there exists a function $G \in \mathcal{H}$ such that for any Pareto optimal actions $a, b \in \Pi$,*

$$\forall x \in [d], \sum_{c=1}^{k}(G(c, \Phi_{cx})_a - G(c, \Phi_{cx})_b) = \mathcal{L}_{ax} - \mathcal{L}_{bx}. \quad (2)$$

---

[1]Here, the set $\{a, b\}$ and the edge $e(a, b)$ connecting $a$ and $b$ in an undirected graph generated by the neighborhood relations are regarded as equivalent. We will omit actions $a$ and $b$ connecting the edge when it is clear from the context.

*In particular, the following $G^{\circ}$ satisfies (2):*

$$G^{\circ}(a, \sigma)_b = \sum_{e \in \text{path}_{\mathcal{T}}(b)} w_e(a, \sigma) \text{ for } a \in \Pi, \quad (3)$$

*where $\text{path}_{\mathcal{T}}(b)$ is the set of edges from $b \in \Pi$ to the root on $\mathcal{T}$.*

### 2.1.3. EXAMPLES OF PARTIAL MONITORING GAMES

For an intuitive understanding of the concept introduced above, we present typical examples of PM games.

**Multi-armed bandits** One of the most well-known examples of PM games is multi-armed bandits with finite loss support (Lai & Robbins, 1985; Auer et al., 2002). For example, the $k$-armed Bernoulli bandits can be represented by $k \times 2^k$ loss and feedback matrices. In particular when $k = 2$,

$$\mathcal{L} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad \Phi = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

One can see that in this case, we have $m = 2$ and the game has no duplicate actions and is non-degenerate.

**Matching pennies** Next, we consider the problem known as matching pennies, which is useful for understanding PM games (Lattimore & Szepesvári, 2020a;b). In this setting, the learner needs to classify emails sequentially delivered to a mailbox are positive P (spam) or negative N (ham). The learner has three choices of actions: (1) classify an email as P, (2) classify an email as N, or (3) let a human inspect an email and determine whether it is P or N. Assume that the learner can only confirm if it is spam or ham by choosing the third action and that the inspection cost by a human is $c_q \geq 0$. When P is wrongly classified as N, the learner suffers a loss of $c_{P \to N} > 0$ for misclassification and vice versa. The goal is to choose a sequence of actions so that the cumulative cost is minimized.

This problem can be indeed formulated as a PM game with $k = 3$ and $d = 2$ with

$$\mathcal{L} = \begin{pmatrix} 0 & c_{N \to P} \\ c_{P \to N} & 0 \\ c_q & c_q \end{pmatrix}, \quad \Phi = \begin{pmatrix} \text{None} & \text{None} \\ \text{None} & \text{None} \\ \text{P} & \text{N} \end{pmatrix}.$$

It is known that this problem falls into easy or hard games depending on the parameters in $\mathcal{L}$. The most well-studied instance is with $c_{N \to P} = c_{P \to N} = 1$, and the problem becomes non-degenerate and locally observable if $c_q \in (0, 1/2)$ and globally observable if $c_q > 1/2$, and trivial if $c_q = 0$ (Lattimore & Szepesvári, 2020b).

### 2.2. Follow the Regularized Leader

We rely on follow-the-regularized-leader (FTRL) to develop the proposed algorithms. In FTRL, we compute a

probability vector $q_t \in \mathcal{P}_k$ that minimizes the sum of expected cumulative loss and convex regularizer $\psi_t \colon \mathcal{P}_k \to \mathbb{R}$ as follows:

$$q_t = \arg\min_{q \in \mathcal{P}_\Pi} \left\langle \sum_{s=1}^{t-1} \widehat{y}_s, \, q \right\rangle + \psi_t(q) \,, \tag{4}$$

where the set $\mathcal{P}_\Pi := \{p \in \mathcal{P}_k \colon p_a = 0, \forall a \notin \Pi\}$ is the convex closed polytope on the probability simplex with nonzero elements only at indices in $\Pi$, and $\widehat{y}_t \in \mathbb{R}^k$ is an estimator of the loss (difference) at round $t$. The following lemma is well-known and used when bounding the regret of FTRL.

**Lemma 3.** *Let $q_1, \ldots, q_T$ be determined by (4). Then, for any $u \in \mathcal{P}_k$, $\sum_{t=1}^{T} \langle \widehat{y}_t, q_t - u \rangle$ is bounded from above by*

$$\sum_{t=1}^{T} \big( \psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1}) \big) + \psi_{T+1}(u) - \psi_1(q_1)$$

$$+ \sum_{t=1}^{T} \big( \langle \widehat{y}_t, q_t - q_{t+1} \rangle - D_{\psi_t}(q_{t+1}, q_t) \big) \,. \tag{5}$$

We refer to the first and second lines in (5) as *penalty* and *stability* terms. The proof of this lemma is standard in the literature and can be found *e.g.,* in Lattimore & Szepesvári (2020a, Chapter 28) and Orabona (2019, Chapter 7). We include the proof for completeness in Appendix D.

To prove the BOBW guarantee with FTRL, it is essential to rely on the self-bounding technique (Gaillard et al., 2014; Zimmert & Seldin, 2021). This approach combines upper and lower bounds that depend on FTRL output $(q_t)_t$ (or action selection probability $(p_t)_t$). For example, if we have $\mathsf{Reg}_T \leq c_\mathsf{U} \sqrt{Q \log T}$ and $\mathsf{Reg}_T \geq c_\mathsf{L} Q$ for $Q = \mathbb{E}\big[\sum_{t=1}^{T} \sum_{a \neq a^*} q_{ta}\big]$ hold, then from the inequality $bx - ax^2 \leq b^2/(4a)$ for $a, b > 0$, the regret is bounded as

$$\mathsf{Reg}_T = 2\mathsf{Reg}_T - \mathsf{Reg}_T \leq 2c_\mathsf{U} \sqrt{Q \log T} - c_\mathsf{L} Q \leq \frac{c_\mathsf{U}^2 \log T}{c_\mathsf{L}} \,.$$

Note that the smaller $c_\mathsf{L}$ becomes the larger the regret upper bound becomes, and we will see that deriving the tight *lower* bound contributes to improving the dependence on $k$.

In the problem of estimating losses from indirect feedback like PM, it is common to apply some transformation to FTRL output $q_t$ to reduce the variance of the loss estimator, instead of using $q_t$ as action selection probability $p_t$. In the next section, we will investigate a favorable transformation.

## 3. Exploration by Optimization

This section first reviews the existing techniques of exploration by optimization (ExO), and then proposes and analyzes a novel ExO with a hybrid regularizer, which forms the key basis for our algorithm and regret analysis in locally observable games.

### 3.1. Principles of Exploration by Optimization

ExO determines action selection probability $p_t$ from FTRL output $q_t$ (Lattimore & Szepesvári, 2020b). In PM games, constructing a loss estimator directly from the output of FTRL tends to result in large variances. This issue arises because $q_t$ in (4) does not take into account the amount of information provided by each action. It is common to add a certain amount of exploration to $q_t$, for example by mixing a uniform distribution with $q_t$ when computing $p_t$. Instead, ExO determines $p_t$ from $q_t$ by minimizing a component of a regret upper bound consisting of the transformation term $(p_t - q_t)^\top \mathcal{L} e_{x_t}$ and the stability term in (5), which is linked to the variance of a loss estimator.

It is well-known that the stability term is bounded as follows (see *e.g.,* Lattimore & Gyorgy 2021, Lemma 26). For a strongly convex function $\phi \colon \mathbb{R} \to \mathbb{R}$ and any $p, q \in (0, 1)$ with $q \in \mathrm{int}(\mathrm{dom}\,\phi)$, $\ell \in \mathbb{R}$, and $z \in \mathbb{R}$, it holds that

$$(q - p)z - D_\phi(p, q) \leq D_{\phi^*}(\nabla\phi(q) - z, \nabla\phi(q)) =: \mathcal{S}_q^\phi(z) \,. \tag{6}$$

Note that $\mathcal{S}_q^\phi(z)$ is convex with respect to $z$.

We consider FTRL with regularizer

$$\psi(q) = \sum_{a=1}^{k} \beta_a \phi(q_a) \tag{7}$$

with action-dependent learning rate $\beta \in \mathbb{R}_+^k$ and strongly convex function $\phi \colon \mathbb{R} \to \mathbb{R}$. Let $\mathcal{H}^\circ$ be the set of $G \colon [k] \times [d] \to \mathbb{R}^k$ satisfying (2), that is,

$$\mathcal{H}^\circ = \left\{ G \colon \sum_{c=1}^{k} (G(c, \Phi_{cx})_a - G(c, \Phi_{cx})_b) = \mathcal{L}_{bx} - \mathcal{L}_{cx} \right.$$

$$\left. \text{for all } a, b \in \Pi \text{ and } x \in [d] \right\} \,.$$

Then the objective function of the optimization problem over $p \in \mathcal{P}_k$ and $G \in \mathcal{H}^\circ$ in ExO with the regularizer $\psi$ in (7) is

$$\Omega(p, G; q, (\beta, \phi)) =$$

$$\max_{x \in [d]} \left[ (p - q)^\top \mathcal{L} e_x + \sum_{b=1}^{k} \beta_b \sum_{a=1}^{k} p_a \mathcal{S}_{q_b}^\phi \left( \frac{G(a, \Phi_{ax})_b}{\beta_b p_a} \right) \right] \,. \tag{8}$$

Note that the first and second terms correspond to the transformation and stability terms, respectively. Note also that (8) is convex with respect to $p$ and $G$ since $\mathcal{S}_{q_a}^\phi$ is convex and hence its perspective $(p, g) \mapsto p \mathcal{S}_q^\phi(g/p)$ is also convex (Boyd & Vandenberghe, 2004, Chapter 3).

**Remark 1.** The objective function of ExO in (8) is a slightly generalized version of the existing ExO in the use

of the *action-wise learning rate* $\beta \in \mathbb{R}_+^k$, while existing formulations use the same learning rate for each action (Lattimore & Szepesvári, 2020b; Lattimore & Gyorgy, 2021; Tsuchiya et al., 2023a). This generalization enables us to obtain a $\Delta_a$-wise regret upper bound.

**Original exploration by optimization** The original formulation of ExO by Lattimore & Szepesvári (2020b) employs the (negative) Shannon entropy regularizer $\psi_t(q) = \frac{1}{\eta} \sum_{a=1}^k \phi^{\mathsf{NS}}(q_a) = \frac{1}{\eta} \sum_{a=1}^k q_a \log q_a$ and solve the following optimization problem:

$$(p, G) = \underset{p \in \mathcal{P}_k, \, G \in \mathcal{H}^\circ}{\arg\min} \; \Omega(p, G; q, (\mathbf{1}/\eta, \phi^{\mathsf{NS}})) \,. \quad (9)$$

Note that the feasible region for action selection probability $p$ here is the probability simplex $\mathcal{P}_k$. They prove that the optimal value of (9) is bounded by $3m^2 k^3 \eta$. Roughly speaking, letting (8) be $\mathsf{ST}(p; e_x)$ they construct $p^{\mathsf{NS}}(\nu) \in \mathcal{P}_k$ to prove

$$\min_{p \in \mathcal{P}_k} \Omega(p) \le \max_{\nu \in \mathcal{P}_d} \min_{p \in \mathcal{P}_k} \mathsf{ST}(p; \nu)$$
$$\le \max_{\nu \in \mathcal{P}_d} \mathsf{ST}(p^{\mathsf{NS}}; \nu) \le 3m^2 k^3 \eta \,, \quad (10)$$

where the first inequality follows from Sion's minimax theorem.[2]

**Exploration by optimization for best-of-both-worlds guarantees** Since the original ExO in (9) considers only adversarial environments, the self-bounding technique, which requires a lower bound in terms of FTRL output $q$, is not applicable. In particular, there is a case where the solution of (9) satisfies $p_a = 0$ even when $q_a > 0$. To address this issue, the following optimization problem over the restricted feasible region $\mathcal{P}'(q) \subset \mathcal{P}_k$ was investigated (Tsuchiya et al., 2023a):

$$(p, G) = \underset{p \in \mathcal{P}'(q), \, G \in \mathcal{H}^\circ}{\arg\min} \; \Omega(p, G; q, (\eta, \phi^{\mathsf{NS}})) \quad \text{for}$$
$$\mathcal{P}'(q) = \{p \in \mathcal{P}_k : p_a \ge q_a/(2k) \text{ for all } a \in [k]\} \,. \quad (11)$$

Note that the feasible region for action selection probability $p$ is *truncated* probability simplex $\mathcal{P}'(q)$, which allows to use the self-bounding technique since $p \ge q/(2k)$. Despite the fact that minimization problem (11) restricts the feasible region, the optimal value in (11) is bounded by $3m^2 k^3 \eta$ as in (9). This can be confirmed by observing $p^{\mathsf{NS}} \in \mathcal{P}'(q)$.

The regret upper bound for stochastic environments in Tsuchiya et al. (2023a;c) obtained by solving the optimization problem (11) is $O(k^4 m^2 \log T \log(kT)/\Delta)$. Although this bound is poly-logarithmic in $T$, it is highly suboptimal. First, the dependence on $T$ is $O(\log^2 T)$, which is

suboptimal and comes from the use of the Shannon entropy regularizer. In fact, all but one regret upper bounds based on the Shannon entropy regularizer are at least $O(\log^2 T)$ even for simple settings like multi-armed bandits, motivating us to use other regularizers.[3] The bound also depends on the minimum suboptimality gap $\Delta$ rather than action-wise suboptimality gaps $(\Delta_a)_{a \ne a^*}$. Furthermore, the dependence on $k$ is $k^4$, which is highly suboptimal and comes from the use of probability simplex $\mathcal{P}_k'(q)$ in (11). In order to overcome these issues, we will employ a hybrid regularizer as the main regularizer, which is known to achieve the $O(\log T)$ and $(\Delta_a)_a$-dependent regret bounds in simpler online settings. We then investigate its behavior under the ExO technique.

## 3.2. Exploration by Optimization with Hybrid Regularizer

Here, we present a novel algorithm and analysis of ExO with the hybrid regularizer to derive an improved regret bound in Section 4. Let

$$\phi^{\mathsf{LB}}(x) = -\log x \,, \; \phi^{\mathsf{CNS}}(x) = (1-x)\log(1-x) \quad (12)$$

be the component of the negative Shannon entropy and log-barrier regularizer, respectively. We then consider the regularizer $\psi \colon \mathcal{P}_k \to \mathbb{R}_+$ defined by

$$\psi(p) = \sum_{a=1}^k \beta_a \phi(p_a) \,, \quad \text{with} \quad (13)$$
$$\phi(x) = \left(\phi^{\mathsf{LB}}(x) + x - 1\right) + \gamma\left(\phi^{\mathsf{CNS}}(x) + x\right)$$

for $\gamma = \log T$ and vector $\beta \in \mathbb{R}_+^k$. Note that $\psi_t \ge 0$ since $\phi^{\mathsf{LB}}(x) + x - 1 \ge 0$ for all $x > 0$ and $\phi^{\mathsf{CNS}}(x) + x \ge 0$ for all $x \le 1$. Let

$$\mathcal{R}(q) = \{p \in \mathcal{P}_k : p_a \ge q_a/2 \text{ for all } a \in [k]\} \quad (14)$$

be the probability simplex truncated to have more than half of the probability vector $q$. We then consider the optimization problem of ExO objective in (8) over this set as follows:

$$(p, G) = \underset{p \in \mathcal{R}(q), \, G \in \mathcal{H}^\circ}{\arg\min} \; \Omega(p, G; q, (\beta, \phi)) \quad (15)$$

with $\phi$ in (13). Let $\mathsf{OPT}_q(\beta)$ be the optimal value of (15).

**Remark 2.** Recall that (11) considers the truncation of the feasible region that depends on $k$, which makes the dependence on $k$ in the stochastic regret bound worse. In contrast, $\mathcal{R}(q)$ does not depend on $k$, which will improve the dependence on $k$ in the regret bound as we will see in Section 4. Note that for this change of the truncation, several techniques are needed as detailed in the following.

---

[2]We abuse the notation and ignore several variables.

[3]The only exception, a reduction approach by Dann et al. (2023) is difficult to use in PM. See Section 4 for a detailed discussion.

The optimal value $\mathsf{OPT}_q(\beta)$ is bounded as follows:

**Lemma 4.** *Consider non-degenerate locally observable partial monitoring games. Suppose that $\gamma \geq 2$ and $\beta \geq 4mk$. Then the optimal value $\mathsf{OPT}_q(\eta)$ of the exploration by optimization in (15) with regularizer (13) is bounded as*

$$\mathsf{OPT}_q(\beta) \leq 2m^2 k^2 \sum_{b=1}^{k} \frac{1}{\beta_b} \min\left\{ q_b, \frac{1 - q_b}{\gamma q_b} \right\} . \quad (16)$$

The proof of Lemma 4 can be found in Appendix E. The first benefit of this lemma, compared to existing ones, is that it establishes a bound for the optimization problem over $\mathcal{R}(q)$ (in (15)), a feasible region independent of $k$. This approach leads to an improvement in the final regret upper bound by a factor of $k$ (see also Remark 3), compared to ExO in (11). The second benefit is that the obtained upper bound depends on FTRL output $q$, whereas the existing bound on optimal values of ExO is independent of $q$. This distinction is crucial when applying the self-bounding technique with regularizer in (13). When applying the self-bounding technique, it is essential for $q$ to appear in the upper bound of the stability or penalty term, where the objective of ExO consists of the stability and transformation terms. If we use the log-barrier regularizer then the penalty term is no longer dependent on $q$ unlike when using the Shannon entropy regularizer, but the self-bounding technique is applicable due to Lemma 4, which bounds the stability term (and transformation term) in terms of $q$.

A crux of the proof is to consider feasible solution $p^{\mathsf{H}} = (q + W_\nu(q))/2$ to upper-bound the optimal value of ExO, where $W_\nu(\cdot)$ is called a *water transfer operator*. Here, the water transfer operator $W_\nu$ transforms FTRL output $q$ into a probability vector that has useful properties in locally observable games (see Lemma 13 in Appendix E and Lattimore & Szepesvári 2019b for details). This feasible solution $p^{\mathsf{H}}$ clearly satisfies $p^{\mathsf{H}} \geq q/2$, that is, $p^{\mathsf{H}} \in \mathcal{R}(q)$. It remains to show that $\Omega(p^{\mathsf{H}}, G^\circ)$ is bounded by the RHS of (16). Note that we cannot use the existing analysis mixing uniform distribution[4] $\mathbf{1}/k \in \mathcal{P}_k$ to construct $p^{\mathsf{NS}} = (1 - \gamma)W_\nu(q) + \gamma \mathbf{1}/k$ for $\gamma \in (0, 1/2]$ instead of $p^{\mathsf{H}}$. This is because $p^{\mathsf{NS}}(\nu)$ is not only outside of our feasible region $\mathcal{R}(q)$ but also results in an upper bound that does not depend on $q$, due to the regret coming from the transformation term. To deal with this issue, the proof of Lemma 4 introduces an analysis that bounds the stability term of the log-barrier regularizer without mixing uniform distribution by exploiting the property of the water transfer operator.

**Remark 3.** By mimicking the construction of $p^{\mathsf{H}}$ in the proof of Lemma 4, we can improve the regret bound in the stochastic environment in Tsuchiya et al. (2023a;c)

[4] This in fact is necessary with the Shannon entropy when bounding the stability term.

---

**Algorithm 1** Algorithm for locally observable games

> **for** $t = 1, 2, \ldots$ **do**
>     Compute $q_t$ via FTRL (4) with regularizer (17).
>     Compute $p_t$ and $G_t$ by exploration by optimization
>     in (15) with $q \leftarrow q_t$, $\phi \leftarrow \phi$ in (17), and $\beta \leftarrow \beta_t$
>     in (18).
>     Sample $A_t \sim p_t$, observe $\Phi_{A_t x_t} \in \Sigma$, and compute
>     $\widehat{y}_t = G_t(A_t, \sigma_t)/p_{tA_t}$.
> **end for**

---

by a factor of $k$. In Lattimore & Szepesvári (2020b); Tsuchiya et al. (2023a;c), $p^{\mathsf{NS}}$ is constructed as $p^{\mathsf{NS}} = (1 - \gamma)W_\nu(q) + \gamma \mathbf{1}/k$ with a certain $\gamma \in (0, 1/2]$. Instead, by introducing $q/2$ so that $p^{\mathsf{NS}} = q/2 + ((1 - \gamma)W_\nu(q) + \gamma \mathbf{1}/k)/2$, we can improve the regret upper bounds in Tsuchiya et al. (2023a;c) in the stochastic environments by a factor of $k$. Still, just solely applying this improvement does not lead to the regret bound that we will obtain in this paper.

## 4. Locally Observable Games

This section proposes an improved BOBW algorithm for locally observable games and derives its regret bounds based on the analysis for ExO with the hybrid regularizer provided in Section 3.

### 4.1. Proposed Algorithm

The proposed algorithm is based on ExO developed in Section 3.2. We use the regularizer (13) for FTRL in Section 3: The regularizer $\psi_t : \mathcal{P}_k \to \mathbb{R}_+$ in (4) is defined by

$$\psi_t(q) = \sum_{a=1}^{k} \beta_{ta} \phi(q_a), \tag{17}$$
$$\phi(x) = (\phi^{\mathsf{LB}}(x) + x - 1) + \gamma(\phi^{\mathsf{CNS}}(x) + x),$$

where $\phi^{\mathsf{LB}}$ and $\phi^{\mathsf{CNS}}$ are defined in (12). Note that this regularizer corresponds to that in (13). The learning rate $\beta_t = (\beta_{t1}, \ldots, \beta_{tk})^\top \in \mathbb{R}_+^k$ in (17) are set to

$$\beta_{ta} = \max\{4mk, \beta'_{ta}\}, \quad \beta'_{ta} = c\sqrt{\alpha_0 + \frac{1}{\gamma} \sum_{s=1}^{t-1} \alpha_{sa}},$$

$$\alpha_{ta} = \min\left\{ q_{ta}, \frac{1 - q_{ta}}{\gamma q_{ta}} \right\}, \tag{18}$$

where $\alpha_0 = \frac{1}{\gamma^{3/2}} + \varepsilon$ for $\varepsilon > 0$ and $c = 2mk$.

We compute $p_t$ and $G_t$ by (15) with $q \leftarrow q_t$ and $\beta \leftarrow \beta_t$. Using this $G_t$, we estimate the loss at round $t$ by $\widehat{y}_t = G_t(A_t, \sigma_t)/p_{tA_t}$. The proposed algorithm for locally observable games is summarized in Algorithm 1.

### 4.2. Regret Bounds and Analysis

Here we provide the regret bound and analysis of the proposed algorithm. The proposed algorithm achieves the following bound:

**Theorem 5.** *Consider any non-degenerate partial monitoring games with horizon $T \geq 8$. In stochastic environments, Algorithm 1 with $c = 2mk$ achieves*

$$\mathsf{Reg}_T = O\left(\sum_{a \neq a^*} \frac{m^2 k^2 \log T}{\Delta_a}\right). \tag{19}$$

*In adversarial environments, the same algorithm achieves*

$$\mathsf{Reg}_T = O\left(k^{3/2} m \sqrt{T \log T} + k^2 m \log T + k^3 m\right).$$

*Further, in stochastic environments with adversarial corruptions, the same algorithm achieves $\mathsf{Reg}_T = O(\mathcal{R}^{\mathrm{loc}} + \sqrt{C\mathcal{R}^{\mathrm{loc}}})$ with $\mathcal{R}^{\mathrm{loc}}$ denoting the RHS of (19).*

The above bounds are significantly better than the existing bounds for PM in Tsuchiya et al. (2023a). The bound for the stochastic environments depends on $\log T$ rather than $\log^2 T$, $(\Delta_a)_{a \neq a^*}$ instead of $\Delta_{\min}$, and $k^3$ instead of $k^4$. The bound for the adversarial environment is $O(\sqrt{\log T})$ times smaller than the bound in Tsuchiya et al. (2023a).

**Remark 4.** Recently, Dann et al. (2023) proposed a framework for converting a bandit algorithm in adversarial environments into a BOBW algorithm with an $O(\log T)$ upper bound in stochastic environments in a blackbox manner. However, it is unclear whether this framework can be applied to PM. This is because, in order to use this framework, losses of selected actions must be observable, which does not apply to PM. They consider how to apply their framework to online learning with feedback graphs (a.k.a. graph bandits), a setting in which the selected action is not always observed. However, this approach requires constructing a *surrogate* loss for the problem, and it is highly unclear whether such a construction is possible for PM.

In the following, we sketch the proof of Theorem 5. Define $u = \left(1 - \frac{k}{T}\right) e_{a^*} + \frac{1}{T}$. Using this, we start by decomposing the regret as follows.

**Lemma 6.** *The regret of Algorithm 1 is bounded from above by*

$$\mathbb{E}\left[\sum_{t=1}^{T} \big(\psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1})\big) + \psi_{T+1}(u) - \psi_1(q_1)\right]$$
$$+ \mathbb{E}\left[\sum_{t=1}^{T} \max\left\{0, \mathsf{OPT}_{q_t}(\beta_t)\right\}\right] + k. \tag{20}$$

Note that the first term in (20) corresponds to the penalty term. The proof of this lemma is similar to Lemma 7

of Tsuchiya et al. (2023a). The RHS of (20) is further bounded by the following lemma:

**Lemma 7.** *The penalty term is bounded from above as*

$$\sum_{t=1}^{T} \big(\psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1})\big) + \psi_{T+1}(u) - \psi_1(q_1)$$
$$\leq \gamma \sum_{a=1}^{k} \beta'_{T+1,a} + 4mk^2 \gamma, \tag{21}$$

*and the optimal value of ExO is bounded as*

$$\mathsf{OPT}_{q_t}(\beta_t) \leq 2m^2 k^2 \sum_{a=1}^{k} \frac{\alpha_{ta}}{\beta'_{ta}}. \tag{22}$$

The bound for the penalty term follows from $u \geq 1/T$, which is standard in the literature. The bound for $\mathsf{OPT}_{q_t}(\beta_t)$ is a direct application of Lemma 4.

Now, we are ready to prove Theorem 5. Here, we only sketch the proof and the complete proof can be found in Appendix F.3.

*Proof sketch of Theorem 5.* For notational simplicity, we may ignore terms whose dependence on $T$ is of order less than $\log T$ in this sketch. We first bound $\sum_{a=1}^{k} \frac{\alpha_{ta}}{\beta'_{ta}}$ in Lemma 7 with the definitions of $\beta_t$ and $\alpha_t$. Since $\alpha_0 \geq \frac{1}{\gamma^{3/2}} \geq \frac{1}{\gamma} \alpha_t$, one can prove that $\sum_{t=1}^{T} \frac{\alpha_{ta}}{\beta'_{ta}} \lesssim O\left(\frac{\gamma}{c^2}\left(\beta'_{T+1,a} - \beta'_{1,a}\right)\right)$. Combining this inequality with Lemmas 6 and 7 and $c = 2mk$ gives $\mathsf{Reg}_T = O\big(\mathbb{E}[\gamma \sum_{t=1}^{T} \sum_{a=1}^{k} \beta'_{T+1,a}]\big)$.

We next bound $\beta'_{T+1,a}$ for suboptimal action $a \neq a^*$ and optimal actions $a^*$ separately. From the definition of $\beta'_{ta}$ in (18), we have $\beta'_{T+1,a} \leq c\sqrt{\alpha_0 + \sum_{t=1}^{T} q_{ta}/\gamma}$ for all $a \neq a^*$. For optimal action $a^*$, using the inequality $\min\left\{z, \frac{1-z}{\gamma z}\right\} \leq \frac{2}{\sqrt{\gamma}}(1 - z)$ that holds for any $z \in [0,1]$ due to $\sqrt{\gamma} \geq 2$, we have $\sum_{t=1}^{T} \alpha_{ta^*} \leq \frac{2}{\sqrt{\gamma}} \sum_{t=1}^{T}(1 - q_{ta^*}) = \frac{2}{\sqrt{\gamma}} \sum_{t=1}^{T} \sum_{a \neq a^*} q_{ta}$. This implies that $\beta'_{T+1,a^*}$ is bounded as $\beta'_{T+1,a^*} \leq c\sqrt{\alpha_0 + \frac{2}{\gamma^{3/2}} \sum_{t=1}^{T} \sum_{a \neq a^*} q_{ta}}$. Combining the upper bounds on $(\beta'_{ta})_{a \neq a^*}$ and $\beta'_{ta^*}$ implies that the regret is roughly bounded as $\mathsf{Reg}_T \leq O(\gamma \sum_{a \neq a^*} \mathbb{E}\big[c\sqrt{\sum_{t=1}^{T} \frac{q_{ta}}{\gamma}}\big] + \gamma c\sqrt{\mathbb{E}\big[\sum_{t, a \neq a^*} \frac{q_{ta}}{\gamma^{3/2}}\big]}) \leq O\big(\gamma \sum_{a \neq a^*} c\sqrt{\frac{Q_a}{\gamma}} + \gamma c\sqrt{\sum_{a \neq a^*} \frac{Q_a}{\gamma^{3/2}}}\big)$, where in the last inequality we let $Q_a = \mathbb{E}[\sum_{t=1}^{T} q_{ta}]$.

For adversarial environments, using the Cauchy–Schwarz inequality and $\sum_{a=1}^{k} q_{ta} = 1$ in the last inequality gives that the regret is roughly bounded as $\mathsf{Reg}_T = O\big(\gamma c\sqrt{Tk/\gamma} + \gamma c\sqrt{kT/\gamma^{3/2}}\big) = O(mk^{3/2}\sqrt{T \log T})$.

For stochastic environments, since $p \in \mathcal{R}(q)$, it holds that $\text{Reg}_T \geq \frac{1}{2} \sum_{a \neq a^*} \Delta_a Q_a$. Hence, $\text{Reg}_T = 2\text{Reg}_T - \text{Reg}_T \leq O(\gamma \sum_{a \neq a^*} c\sqrt{\frac{Q_a}{\gamma}} + \gamma c\sqrt{\sum_{a \neq a^*} \frac{Q_a}{\gamma^{\frac{3}{2}}}} - \sum_{a \neq a^*} \Delta_a Q_a) \leq O(\sum_{a \neq a^*} (c\sqrt{\gamma Q_a} - \Delta_a Q_a) + c(\sqrt{\gamma} \sum_{a \neq a^*} Q_a)^{\frac{1}{2}} - \Delta \sum_{a \neq a^*} Q_a)$. Taking the worst case with respect to $(Q_a)_{a \neq a^*}$ in the last inequality completes the proof for stochastic environments. The proof for stochastic environments with adversarial corruptions mostly follows the same arguments. $\square$

## 5. Globally Observable Games

This section presents an improved BOBW algorithm for globally observable games and shows its regret bounds. Let $c_{\mathcal{G}} = \max\{1, k\|G^\circ\|_\infty\}$ be the game-dependent constant.

We use FTRL with regularizer $\psi_t \colon \mathcal{P}_k \to \mathbb{R}$ defined by

$$\psi_t(q) = \beta_t \left( \psi^{\mathsf{LB}}(q) + \psi^{\mathsf{CLB}}(q) \right) \qquad (23)$$

for the log-barrier regularizer $\psi^{\mathsf{LB}}(x) = -\sum_{a=1}^k \log x_a$ and the complement log-barrier regularizer $\psi^{\mathsf{CLB}}(x) = -\sum_{a=1}^k \log(1 - x_a)$. The learning rate $\beta_t \in \mathbb{R}_+$ in (23) is

$$\beta_t = c_1 \left( k + \sum_{s=1}^{t-1} z_s \right)^{2/3} \text{ with } z_t = \sum_{a=1}^k \min\{q_{ta}, 1 - q_{ta}\}$$

for $c_1 > 0$ specified later. The output of FTRL $q_t$ is transformed to action selection probability $p_t \in \mathcal{P}_k$ by

$$p_t = (1 - \gamma_t)q_t + \gamma_t \frac{1}{k} \quad \text{with} \quad \gamma_t = \frac{\sqrt{2}c_{\mathcal{G}} z_t}{\sqrt{\beta_t}} . \qquad (24)$$

We use the loss estimator $\widehat{y}_t = G^\circ(A_t, \Phi_{A_t x_t})/p_{tA_t}$ for $G^\circ$ in (3). The proposed algorithm for globally observable games is summarized in Algorithm 2 in Appendix G.

The above algorithm achieves the following bounds:

**Theorem 8.** *Consider any globally observable partial monitoring games. In stochastic environments, the above algorithm (Algorithm 2 in Appendix G) with $c_1 = (3c_{\mathcal{G}}/(k \log T))^{2/3}$ achieves*

$$\text{Reg}_T = O\left( \frac{c_{\mathcal{G}}^2 k \log T}{\Delta^2} + c_{\mathcal{G}}^{2/3} k^{4/3} (\log T)^{1/3} \right) . \qquad (25)$$

*In adversarial environments, the same algorithm achieves*

$$\text{Reg}_T = O\left( (c_{\mathcal{G}}^2 k \log T)^{1/3} T^{2/3} + (c_{\mathcal{G}}^2 \log T)^{1/3} k \right) .$$

*Further, in stochastic environments with adversarial corruptions, the same algorithm achieves $\text{Reg}_T = O(\mathcal{R}^{\text{glo}} + (C^2 \mathcal{R}^{\text{glo}})^{1/3})$ with $\mathcal{R}^{\text{glo}}$ the RHS of (25).*

The proof of Theorem 8 can be found in Appendix G. This is the first BOBW bound for globally observable games with an $O(\log T)$ stochastic bound. The above result improves the existing BOBW bounds for PM in stochastic environments, where the bound depends on $\log T$ rather than $\log^2 T$, although its dependence on $k$ is $k$-times worse. The bound for adversarial environments is $\Theta(\sqrt{\log(kT)})$-times smaller than the bound in Tsuchiya et al. (2023a).

The aim of this section is to demonstrate the usefulness of hybrid regularizers by establishing the first $O(\log T)$ stochastic bounds. The choice of the regularizer and analysis largely follow Ito (2022) and there is no significant novelty in this section from the technical viewpoint. As a small contribution, we revised the proof of their paper for dealing with the hybrid regularizer and provided the correct proof.[5] Note also that one can introduce ExO by using a similar analysis for the hybrid regularizer introduced in Section 3 in globally observable games, although we do not provide the details as it does not improve the worst-case bound.

## 6. Concluding Remarks

In this paper, we proposed the first BOBW algorithms for PM that achieve $O(\log T)$ regret bounds in stochastic environments for both locally observable and globally observable games. This was accomplished through the use of hybrid regularizers known to be useful in simpler online learning problems, and the new analysis to combine these regularizers with PM problems, in which we need to deal with limited feedback. The most important contribution of this paper is the development of the analysis to upper-bound the optimal value of the ExO objective induced by the hybrid regularizer over restricted feasible region $\mathcal{R}(q)$. This analysis enabled us to significantly improve the existing regret upper bounds in the stochastic environments.

Our regret upper bounds match the known lower bounds in terms of $T$ both in the worst-case sense (Bartók et al., 2011; Lattimore & Szepesvári, 2019a) and in the gap-dependent sense (Komiyama et al., 2015). However, little else is understood about the dependencies on the variables $k$, $d$, and $m$, as mentioned in Lattimore & Szepesvári (2020a, Section 37.9). Investigating these dependencies is an important future work.

---

[5]In Ito (2022, Lemma 4.9, Proposition 4.10), a case distinction based on the output of FTRL $q$ needs to be discussed. This paper addresses and corrects this, as detailed in Appendix G.

## Impact Statement

This work is primarily theoretical and is not expected to have any direct negative ethical or societal consequences. The result could be advantageous for those studying theoretical elements of online learning with limited feedback, such as online learning with feedback graphs and partial monitoring. Additionally, the improved regret bounds may contribute to the frequent use of partial monitoring algorithms in real-world applications such as recommendation systems and dynamic pricing.

## References

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Bartók, G., Pál, D., and Szepesvári, C. Minimax regret of finite partial-monitoring games in stochastic environments. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19, pp. 133–154, 2011.

Bartók, G., Zolghadr, N., and Szepesvári, C. An adaptive algorithm for finite stochastic partial monitoring. In *the 29th International Conference on Machine Learning*, pp. 1–20, 2012.

Boyd, S. P. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Bubeck, S. and Slivkins, A. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pp. 42.1–42.23, 2012.

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

Dann, C., Wei, C.-Y., and Zimmert, J. A blackbox approach to best of both worlds in bandits and beyond. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pp. 5503–5570, 2023.

Foster, D. J., Rakhlin, A., Sekhari, A., and Sridharan, K. On the complexity of adversarial decision making. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35404–35417, 2022.

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

Gaillard, P., Stoltz, G., and van Erven, T. A second-order bound with excess losses. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pp. 176–196, 2014.

Ito, S. Hybrid regret bounds for combinatorial semi-bandits and adversarial linear bandits. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2654–2667, 2021a.

Ito, S. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pp. 2552–2583, 2021b.

Ito, S. Revisiting online submodular minimization: Gap-dependent regret bounds, best of both worlds and adversarial robustness. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 9678–9694, 2022.

Ito, S. and Takemura, K. An exploration-by-optimization approach to best of both worlds in linear bandits. In *Advances in Neural Information Processing Systems*, volume 36, pp. 71582–71602, 2023.

Ito, S., Tsuchiya, T., and Honda, J. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 28631–28643, 2022a.

Ito, S., Tsuchiya, T., and Honda, J. Adversarially robust multi-armed bandit algorithm with variance-dependent regret bounds. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pp. 1421–1422, 2022b.

Jin, T., Liu, J., and Luo, H. Improved best-of-both-worlds guarantees for multi-armed bandits: FTRL with general regularizers and multiple optimal arms. In *Advances in Neural Information Processing Systems*, volume 36, pp. 30918–30978, 2023.

Komiyama, J., Honda, J., and Nakagawa, H. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1792–1800, 2015.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Lattimore, T. and Gyorgy, A. Mirror descent and the information ratio. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pp. 2965–2992, 2021.

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

Lattimore, T. and Szepesvári, C. An information-theoretic approach to minimax regret in partial monitoring. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pp. 2111–2139, 2019b.

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

Lattimore, T. and Szepesvári, C. Exploration by optimisation in partial monitoring. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pp. 2488–2515, 2020b.

Luo, H. and Schapire, R. E. Achieving all with no parameters: AdaNormalHedge. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pp. 1286–1304, 2015.

Lykouris, T., Mirrokni, V., and Paes Leme, R. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, 2018.

Orabona, F. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

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

Rouyer, C., van der Hoeven, D., Cesa-Bianchi, N., and Seldin, Y. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35035–35048, 2022.

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

Tsuchiya, T., Honda, J., and Sugiyama, M. Analysis and design of Thompson sampling for stochastic partial monitoring. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8861–8871, 2020.

Tsuchiya, T., Ito, S., and Honda, J. Best-of-both-worlds algorithms for partial monitoring. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, pp. 1484–1515, 2023a.

Tsuchiya, T., Ito, S., and Honda, J. Further adaptive best-of-both-worlds algorithm for combinatorial semi-bandits. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, pp. 8117–8144, 2023b.

Tsuchiya, T., Ito, S., and Honda, J. Stability-penalty-adaptive follow-the-regularized-leader: Sparsity, game-dependency, and best-of-both-worlds. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47406–47437, 2023c.

Wei, C.-Y. and Luo, H. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pp. 1263–1291, 2018.

Zimmert, J. and Seldin, Y. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 467–475, 2019.

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

Zimmert, J., Luo, H., and Wei, C.-Y. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 7683–7692, 2019.

# A. Notation

Table 2 summarizes the notation used in this paper.

Table 2. Notation

| Symbol | Meaning |
|---|---|
| $\mathcal{P}_k$ | $(k-1)$-dimensional probability simplex |
| $D_\psi(x, y)$ | Bregman divergence induced by $\psi$ from $y$ to $x$ |
| $\psi^*$ | convex conjugate of $\psi$ |
| $T \in \mathbb{N}$ | time horizon |
| $k \in \mathbb{N}$ | number of actions |
| $d \in \mathbb{N}$ | number of outcomes |
| $\mathcal{L} \in [0,1]^{k \times d}$ | loss matrix |
| $\Sigma$ | set of feedback symbols |
| $\Phi \in \Sigma^{k \times d}$ | feedback matrix |
| $m \in \mathbb{N}$ | maximum number of distinct symbols in a single row of $\Phi$ |
| $a^* \in \Pi$ | (unique) optimal action |
| $A_t \in [k]$ | action chosen by learner at round $t$ |
| $x_t \in [d]$ | outcome chosen by adversary at round $t$ |
| $\text{path}_{\mathscr{T}}$ | path over in-tree $\mathscr{T}$ |
| $\mathcal{H}$ | the set of all functions from $[k] \times \Sigma$ to $\mathbb{R}^k$ |
| $\mathcal{H}^\circ$ | the set of $G\colon [k] \times [d] \to \mathbb{R}^k$ satisfying (2) |
| $G^\circ \in \mathcal{H}^\circ$ | an example of $G$ in $\mathcal{H}^\circ$ defined in (3) |
| $c_{\mathcal{G}} > 0$ | game-dependent constant defined in Section 5 |
| $\Delta_a \in [0,1]$ | suboptimality gap of action $a$ |
| $q_t \in \mathcal{P}_k$ | output of FTRL at round $t$ |
| $p_t \in \mathcal{P}_k$ | action selection probability at round $t$ |
| $\psi_t\colon \mathcal{P}_k \to \mathbb{R}$ | regularizer of FTRL at round $t$ |
| $\beta_t > 0$ | learning rate of FTRL at round $t$ |
| $\psi^{\mathsf{NS}}\colon \mathbb{R}_+^k \to \mathbb{R}$ | $\sum_{i=1}^k x_i \log x_i$, negative Shannon entropy |
| $\psi^{\mathsf{CNS}}\colon (-\infty, 1)^k \to \mathbb{R}$ | $\sum_{i=1}^k (1 - x_i) \log(1 - x_i)$, complement negative Shannon entropy |
| $\psi^{\mathsf{LB}}\colon \mathbb{R}_+^k \to \mathbb{R}$ | $\sum_{i=1}^k \log(1/x_i)$, log-barrier |
| $\psi^{\mathsf{CLB}}\colon (-\infty, 1)^k \to \mathbb{R}$ | $\sum_{i=1}^k \log(1/(1 - x_i))$, complement log-barrier |
| $\mathcal{S}_q^\psi\colon \mathbb{R}^k \to \mathbb{R}$ | stability function induced by function $\psi\colon \mathbb{R}^k \to \mathbb{R}$ in (6) |
| $\mathcal{P}'(q) \subset \mathcal{P}_k$ | feasible set used in existing ExO defined in (11) |
| $\mathcal{R}(q) \subset \mathcal{P}_k$ | feasible set used in ExO defined in (14) |
| $\Omega(\cdot\,; q, (\beta, \phi))$ | objective function in ExO with regularizer $\psi(q) = \sum_{a=1}^k \beta_a \phi(q_a)$ in (8) |
| $\mathsf{OPT}_q(\beta)$ | optimal value of ExO objective $\Omega$ minimized over $\mathcal{R}(q)$ |

# B. Missing Related Work

**Exploration by Optimization** Exploration by optimization (ExO), an important technique for achieving a good regret upper bound in locally observable games, was invented by Lattimore & Szepesvári (2020b) and further extended by Lattimore & Gyorgy (2021). Very recent studies have revealed that ExO is closely related to the notion of *(convexified) decision-estimation coefficient* and the optimal regret upper bound in adversarial environments is characterized by the ExO framework (Foster et al., 2022). In the context of BOBW algorithms, ExO was very recently investigated in PM (Tsuchiya et al., 2023a;c) and linear bandits (Ito & Takemura, 2023).

**Best-of-both-worlds Guarantees** The investigation of the BOBW algorithm was first initiated by Bubeck & Slivkins (2012) in multi-armed bandits, and after that, it has been shown that the BOBW guarantees are possible in a wide range of online decision-making problems. Many algorithms have been developed for special cases of PM such as predication with expert advice (Gaillard et al., 2014; Luo & Schapire, 2015), multi-armed bandits (Zimmert & Seldin, 2019; 2021; Ito, 2021b), and online learning with feedback graphs (Rouyer et al., 2022; Ito et al., 2022a). Most BOBW algorithms are based on FTRL (or online mirror descent) because the self-bounding technique invented in Gaillard et al. (2014) and used in many subsequent studies, *e.g.*, Wei & Luo (2018); Zimmert & Seldin (2021), is a strong tool to derive (poly-)logarithmic regret bounds in stochastic environments.

It is worth noting that the very recent work by Dann et al. (2023) has devised a general framework that transforms a bandit algorithm for adversarial environments into a BOBW algorithm in a blackbox manner. However, it is highly nontrivial whether their algorithm can be applied to PM, and a detailed discussion can be found in Section 4.

The use of FTRL with the hybrid regularizer consisting of the log-barrier and complement Shannon entropy is not new: This regularizer has been used to obtain BOBW guarantees in multi-armed bandits and combinatorial semi-bandits (Ito, 2021a; Ito et al., 2022b; Tsuchiya et al., 2023b). The idea of using complement Shannon entropy was originally introduced by Zimmert et al. (2019) in the context of combinatorial semi-bandits. Note, however, that PM is much more complex than these problems, and a non-trivial analysis is required when combining this regularizer with the ExO approach as detailed in Section 3.

## C. Stochastic Environment with Adversarial Corruptions

Here, we introduce a stochastic environments with adversarial corruptions, considered originally in multi-armed bandits (Lykouris et al., 2018) and later investigated also in PM (Tsuchiya et al., 2023a;c). This environment is an intermediate environment between the stochastic environments and adversarial environments.

In this environment, at each time step $t \in [T]$, a temporary outcome $x'_t \in [d]$, which cannot be observed by the learner, is sampled from an unknown distribution $\nu^*$ in an i.i.d. manner. The adversary then corrupts $x'_t$ to $x_t$ without knowing $A_t$, and based on this outcome the loss and feedback symbol observed by the learner is determined. In this environment, the suboptimality gap is defined in terms of $(x'_t)$ as follows: $\Delta_a = \mathbb{E}_{x'_t \sim \nu^*}[\mathcal{L}_{ax'_t} - \mathcal{L}_{a^*x'_t}]$ for $a \in [k]$.

The corruption level $C \geq 0$ in this environment is defined by $C = \mathbb{E}[\sum_{t=1}^{T}\|\mathcal{L}e_{x_t} - \mathcal{L}e_{x'_t}\|_\infty]$. Note that the stochastic environment with adversarial corruptions with $C = 0$ (resp. $C = 2T$) corresponds to the stochastic (resp. adversarial) environment. Note that the proposed algorithms will work without the knowledge of $C$.

In addition to the stochastic environment with adversarial corruptions, there are several intermediate environments, a stochastically constrained adversarial environment, and an adversarial environment with a self-bounding constraint that includes all of these environments as special cases. Detailed definitions can be found in Tsuchiya et al. (2023a). The algorithms and bounds in this paper are directly applicable to these general environments as well. For the sake of brevity, this paper focuses only on the stochastic environments with adversarial corruptions.

Under this assumption, the regret in the stochastic environment with adversarial corruptions is bounded from below as follows:

**Lemma 9.** *Consider any stochastic environment with adversarial corruptions. Then, the regret is bounded from below as*

$$\mathsf{Reg}_T \geq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a \neq a^*}\Delta_a\, p_{ta}\right] - 2\,C\,.$$

This lower bound will be used to derive the logarithmic regret bound for stochastic environments. The proof can be found in Appendix A of Tsuchiya et al. (2023a), and we include the proof for completeness.

*Proof of Lemma 9.* Recall that in the stochastic environment with adversarial corruptions, $x'_1, x'_2, \ldots$ are sampled in an i.i.d. manner and the suboptimality gap is defined by $\Delta_a = \mathbb{E}_{x'_t \sim \nu^*}[\mathcal{L}_{ax'_t} - \mathcal{L}_{a^*x'_t}]$ for $a \in [k]$. Hence, the regret is bounded from below as

$$
\begin{aligned}
\mathsf{Reg}_T &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathcal{L}_{A_tx_t} - \mathcal{L}_{a^*x_t}\right)\right]\\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathcal{L}_{A_tx'_t} - \mathcal{L}_{a^*x'_t}\right)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathcal{L}_{A_tx_t} - \mathcal{L}_{A_tx'_t}\right)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathcal{L}_{a^*x'_t} - \mathcal{L}_{a^*x_t}\right)\right]\\
&\geq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a \neq a^*}\Delta_a\, p_{ta}\right] - 2\,C\,,
\end{aligned}
$$

which completes the proof of Lemma 9. □

13

# D. Common Analysis

## D.1. Standard Analysis of FTRL

*Proof of Lemma 3.* Let $F_t(q) = \sum_{s=1}^{t-1} \langle \widehat{y}_s, q \rangle + \psi_t(q)$ be the objective function of FTRL at time $t$. Then it holds that

$$\sum_{t=1}^{T} \langle \widehat{y}_t, q_t - u \rangle \leq \psi_{T+1}(u) - \psi_1(x_1) + \sum_{t=1}^{T} \left( F_t(q_t) - F_{t+1}(q_{t+1}) + \langle \widehat{y}_t, q_t \rangle \right) . \tag{26}$$

This inequality holds since

$$-\sum_{t=1}^{T} \langle \widehat{y}_t, u \rangle = \psi_{T+1}(u) - F_{T+1}(u) + \sum_{t=1}^{T} \left( F_t(q_t) - F_{t+1}(q_{t+1}) \right) - F_1(q_1) + F_{T+1}(q_{T+1})$$

$$\leq \psi_{T+1}(u) - \psi_1(q_1) + \sum_{t=1}^{T} \left( F_t(q_t) - F_{t+1}(q_{t+1}) \right) ,$$

where the equality holds by the definition of $F_t$. Now it remains to bound $F_t(q_t) - F_{t+1}(q_{t+1}) + \langle \widehat{y}_t, q_t \rangle$ in (26). This is bounded as

$$F_t(q_t) - F_{t+1}(q_{t+1}) + \langle \widehat{y}_t, q_t \rangle$$
$$= F_t(q_t) - F_t(q_{t+1}) + \psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1}) + \langle \widehat{y}_t, q_t - q_{t+1} \rangle$$
$$\leq -D_{F_t}(q_{t+1}, q_t) + \psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1}) + \langle \widehat{y}_t, q_t - q_{t+1} \rangle ,$$

where the inequality follows from the first-order optimality condition. Since $D_{F_t}(q_{t+1}, q_t) = D_{\psi_t}(q_{t+1}, q_t)$ the proof is completed. $\qquad\square$

## D.2. Basic Facts to Bound Stability Terms

This section presents the basic propositions to bound stability terms, which are quite standard in the literature. The stability terms are bounded in terms of the following functions:

$$\xi(x) := \exp(-x) + x - 1 \leq \begin{cases} \frac{1}{2}x^2 & \text{for } x \geq 0 \\ x^2 & \text{for } x \geq -1 , \end{cases} \tag{27}$$

$$\zeta(x) := x - \log(1 + x) \leq x^2 \quad \text{for } x \geq -\frac{1}{2} . \tag{28}$$

Recall that $\phi^{\mathsf{CNS}}(x) = (1 - x)\log(1 - x)$, $\phi^{\mathsf{LB}}(x) = -\log x$, and $\phi^{\mathsf{CLB}}(x) = -\log(1 - x)$ are the components of the complement negative Shannon entropy, log-barrier, and complement log-barrier, respectively.

**Proposition 10** (stability with complement negative Shannon entropy). *For $x \in (0, 1)$,*

$$\max_{y \in (-\infty, 1)} \left\{ a(x - y) - D_{\phi^{\mathsf{CNS}}}(y, x) \right\} = (1 - x)\xi(-a) \quad \text{for} \quad a \in \mathbb{R} ,$$

*where $\phi^{\mathsf{CNS}}(x) = (1 - x)\log(1 - x)$. The last quantity is further bounded by $(1 - x)a^2$ if $a \leq 1$.*

*Proof.* Since $a(x - y) - D_{\phi^{\mathsf{CNS}}}(y, x)$ is concave with respect to $y \in (-\infty, 1)$, the maximizer $y^*$ satisfies $-a - \nabla\phi^{\mathsf{CNS}}(y^*) + \nabla\phi^{\mathsf{CNS}}(x) = 0$, which is equivalent to (since $\nabla\psi^{\mathsf{CNS}}(x) = -\log(1 - x) - 1$)

$$-a + \log(1 - y^*) - \log(1 - x) = 0 .$$

Hence $1 - y^* = e^a(1 - x)$. Hence

$$a(x - y^*) - D_{\phi^{\mathsf{CNS}}}(y^*, x) = D_{\phi^{\mathsf{CNS}}}(x, y^*)$$
$$= (1 - x)\log(1 - x) - (1 - y^*)\log(1 - y^*) + (\log(1 - y^*) + 1)(x - y^*)$$
$$= (1 - x)\log(1 - x) - \log(1 - y^*) + x(\log(1 - y^*) + 1) - y^*$$
$$= -a(1 - x) + x - y^* = (1 - x)(e^a - a - 1) = (1 - x)\xi(-a) .$$

If $-a \geq -1$ holds, then from (27) the last quantity is further bounded by $(1 - x)\xi(-a) \leq (1 - x)a^2$. $\qquad\square$

**Proposition 11** (stability with log-barrier). *For $x \in (0, 1)$,*

$$\max_{y \in \mathbb{R}_+} \left\{ a(x - y) - D_{\phi^{\mathsf{LB}}}(y, x) \right\} = \zeta(ax) \quad \text{for} \quad a \geq -\frac{1}{x},$$

*where $\phi^{\mathsf{LB}}(x) = -\log x$. The last quantity is further bounded by $\zeta(ax) \leq x^2 a^2$ if $a \geq -\frac{1}{2x}$.*

*Proof.* Since $a(x - y) - D_{\phi^{\mathsf{LB}}}(y, x)$ is concave with respect to $y \in \mathbb{R}_+$, the maximizer $y^*$ satisfies $-a - \nabla\phi^{\mathsf{LB}}(y^*) + \nabla\phi^{\mathsf{LB}}(x) = 0$, which is equivalent to

$$-a + \frac{1}{y^*} - \frac{1}{x} = 0 \,.$$

Note that $y^*$ satisfied $y^* > 0$ holds if $a + \frac{1}{x} > 0$, which is equivalent to $a > -\frac{1}{x}$. Hence

$$
\begin{aligned}
a(x - y^*) - D_{\phi^{\mathsf{LB}}}(y^*, x) &= D_{\phi^{\mathsf{LB}}}(x, y^*) \\
&= -\log(x) + \log(y^*) + \frac{x - y^*}{y^*} = -\log\left(\frac{x}{y^*}\right) + \frac{x}{y^*} - 1 \\
&= -\log(1 + ax) + ax = \zeta(ax) \\
&\leq x^2 a^2 \,,
\end{aligned}
$$

where the inequality holds by $-\log(1 + z) + z \leq z^2$ for $z \geq -1/2$ (in (28)) combined with $a \geq -\frac{1}{2x}$. □

**Proposition 12** (stability with complement log-barrier). *For $x \in (0, 1)$,*

$$\max_{y \in (-\infty, 1)} \left\{ a(x - y) - D_{\phi^{\mathsf{CLB}}}(y, x) \right\} = \zeta(-a(1 - x)) \quad \text{for} \quad a < \frac{1}{1 - x},$$

*where $\phi^{\mathsf{CLB}}(x) = -\log(1 - x)$. The last quantity is further bounded by $\xi(-a(1 - x)) \leq (1 - x)^2 a^2$ if $a \leq \frac{1}{2(1-x)}$.*

*Proof.* Since $a(x - y) - D_{\phi^{\mathsf{CLB}}}(y, x)$ is concave with respect to $y \in (-\infty, 1)$, the maximizer $y^*$ satisfies $-a - \nabla\phi^{\mathsf{CLB}}(y^*) + \nabla\phi^{\mathsf{CLB}}(x) = 0$, which is equivalent to

$$-a - \frac{1}{1 - y^*} + \frac{1}{1 - x} = 0 \,.$$

Note that $y^* < 1$ holds if $a < \frac{1}{1-x}$. Hence

$$
\begin{aligned}
a(x - y^*) - D_{\phi^{\mathsf{CLB}}}(y^*, x) &= D_{\phi^{\mathsf{CLB}}}(x, y^*) \\
&= -\log(1 - x) + \log(1 - y^*) - \frac{x - y^*}{1 - y^*} = -\log\left(\frac{1 - x}{1 - y^*}\right) + \frac{1 - x}{1 - y^*} - 1 \\
&= -\log(1 - a(1 - x)) - a(1 - x) = \zeta(-a(1 - x)) \\
&\leq (1 - x)^2 a^2 \,,
\end{aligned}
$$

where the inequality holds by $-\log(1 + z) + z \leq z^2$ for $z \geq -1/2$ (in (28)) combined with $a \leq \frac{1}{2(1-x)}$. □

## E. Deferred Proofs for Exploration by Optimization (Lemma 4, Section 3)

Before going to the proof of Lemma 4, we introduce the water transfer operator.

**Lemma 13** (Lattimore & Szepesvári, 2019b). *Consider any non-degenerate and locally observable partial monitoring games and let $\nu \in \mathcal{P}_d$. Then there exists a function $W_\nu \colon \mathcal{P}_k \to \mathcal{P}_k$ such that the following three statement 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 $\mathcal{T} \subset \mathcal{E}$ over $[k]$ such that $W_\nu(q)_a \leq W_\nu(q)_b$ for all $(a, b) \in \mathcal{T}$.*

Using this water transfer operator, we prove Lemma 4.

*Proof of Lemma 4.* Recall that $\mathcal{H}^\circ = \left\{ G \colon (e_b - e_c)^\top \sum_{a=1}^k G(a, \Phi_{ax}) = \mathcal{L}_{bx} - \mathcal{L}_{cx} \text{ for all } b, c \in \Pi \text{ and } x \in [d] \right\}$. Recall also that $\mathcal{S}_q^\phi(z) = D_{\phi^*}(\nabla\phi(q) - z, \nabla\phi(q))$ and $(p, g) \mapsto p\mathcal{S}_q^\phi(g/p)$ is convex if $p > 0$.

**Step 1. Applying Sion's minimax theorem** Take any $q \in \mathcal{P}_k$. By Sion's minimax theorem,

$$
\mathsf{OPT}_q(\beta) \leq \min_{G \in \mathcal{H}^\circ, \, p \in \mathcal{R}(q)} \max_{\nu \in \mathcal{P}_d} \left[ (p-q)^\top \mathcal{L}\nu + \sum_{x=1}^d \nu_x \sum_{b=1}^k \beta_b \sum_{a=1}^k p_a \mathcal{S}_{q_b}^\phi \left( \frac{G(a, \Phi_{ax})_b}{\beta_b \, p_a} \right) \right]
$$

$$
\leq \max_{\nu \in \mathcal{P}_d} \min_{G \in \mathcal{H}^\circ, \, p \in \mathcal{R}(q)} \left[ (p-q)^\top \mathcal{L}\nu + \sum_{x=1}^d \nu_x \sum_{b=1}^k \beta_b \sum_{a=1}^k p_a \mathcal{S}_{q_b}^\phi \left( \frac{G(a, \Phi_{ax})_b}{\beta_b \, p_a} \right) \right] . \tag{29}
$$

Take any $\nu \in \mathcal{P}_d$ and let $\mathcal{T}$ be the in-tree over $[k]$ induced by water transfer operator $W_\nu \colon \mathcal{P}_k \to \mathcal{P}_k$ by Part (c) of Lemma 13. Then we let $u = W_\nu(q)$. Using this $u$, we define $p \in \mathcal{R}(q)$ by

$$
p = (1 - \lambda)u + \lambda q \quad \text{with} \quad \lambda = 1/2 \, .
$$

By the definition of $p$, it holds that $p_a \geq u_a/2 = W_\nu(q)_a/2$ and $p_a \geq q_a/2$, and it indeed holds that $p \in \mathcal{R}(q)$. We can take $G \in \mathcal{H}^\circ$ such that $G(a, \sigma)_b = \sum_{e \in \mathrm{path}_\mathcal{T}(b)} w_e(a, \sigma)$ with $w_e$ satisfying $\|w_e\|_\infty \leq m/2$ since $\mathcal{G}$ is non-degenerate (Lattimore & Szepesvári, 2020b, Lemma 20). Since paths in $\mathcal{T}$ have length at most $k$, we have $\|G\|_\infty \leq km/2$.

**Step 2. Preliminaries for bounding the stability term** In the following, we first consider the stability term, which corresponds to the second term in (29). Fix $b \in [k]$ and let $z_b = G(a, \Phi_{ax})/(\beta_b \, p_a) \in \mathbb{R}$ for notational simplicity (ignoring index $a$). Taking $p' \in \mathbb{R}$ such that $\nabla \phi(p') = \nabla \phi(q_b) - z_b$, we have

$$
\begin{aligned}
\mathcal{S}_{q_b}^\phi(z_b) &= D_{\phi^*}(\nabla \phi(q_b) - z_b, \nabla \phi(q_b)) \\
&= \phi^*(\nabla \phi(q_b) - z_b) - \phi^*(\nabla \phi(q_b)) - \nabla \phi^*(\nabla \phi(q_b)) \cdot (-z_b) \\
&= -D_{\phi^*}(\nabla \phi(q_b), \nabla \phi(p')) - \nabla \phi^*(\nabla(p')) \cdot (\nabla \phi(q_b) - \nabla \phi(p')) + q_b z_b &&\text{(by } \nabla \phi^* = (\nabla \phi)^{-1}) \\
&= -D_\phi(p', q_b) - p' z_b + q_b z_b &&\text{(by } D_\phi(x, y) = D_{\phi^*}(\nabla \phi(y), \nabla \phi(y))) \\
&= (q_b - p') z_b - D_\phi(p', q_b) \, .
\end{aligned}
$$

Hence, by the linearity of the Bregman divergence and $\psi(q) = \sum_{b=1}^k \beta_b \phi(q_b) = \sum_{b=1}^k \beta_b(\phi^{\mathsf{LB}}(q_b) + \gamma \phi^{\mathsf{CNS}}(q_b))$ in (13),

$$
\begin{aligned}
\mathcal{S}_{q_b}^\phi(z) &= (q_b - p_b') z_b - \beta_b \big( D_{\phi^{\mathsf{LB}}}(p_b', q_b) + \gamma D_{\phi^{\mathsf{CNS}}}(p_b', q_b) \big) \\
&\leq \min \left\{ (q_b - p_b') z_b - \beta_b D_{\phi^{\mathsf{LB}}}(p_b', q_b), \; (q_b - p_b') z_b - \gamma \beta_b D_{\phi^{\mathsf{CNS}}}(p_b', q_b) \right\} \, . \tag{30}
\end{aligned}
$$

We next check the condition to bound the stability term. In particular, we will prove for any $a, b \in [k]$ it holds that

$$
\left| \frac{q_b G(a, \Phi_{ax})_b}{\beta_b p_a} \right| \leq \frac{2mk}{\beta_b} \leq \frac{1}{2} \, , \tag{31}
$$

which is the condition for bounding the stability term induced by the log-barrier regularizer. This inequality holds since

$$
\begin{aligned}
\left| \frac{q_b G(a, \Phi_{ax})_b}{\beta_b p_a} \right| &\leq \frac{2 q_b}{\beta_b u_a} |G(a, \Phi_{ax})_b| &&\text{(by } p_a \geq u_a/2) \\
&= \frac{2 q_b}{\beta_b u_a} \left| \sum_{e \in \mathrm{path}_\mathcal{T}(b)} w_e(a, \Phi_{ax}) \right| \\
&\leq \frac{m q_b}{\beta_b u_a} \left| \sum_{e \in \mathrm{path}_\mathcal{T}(b)} \mathbb{1}[a \in \mathcal{N}_e] \right| &&\text{(by } \|w_e\|_\infty \leq m/2) \\
&\leq \frac{2 m q_b}{\beta_b u_a} \mathbb{1}\big[ a \in \cup_{e \in \mathrm{path}_\mathcal{T}(b)} \mathcal{N}_e \big] &&\text{(since any action is in } \mathcal{N}_e \text{ at most two edges in } \mathrm{path}_\mathcal{T}(b)) \\
&\leq \frac{2 m q_b}{\beta_b u_b} \mathbb{1}\big[ a \in \cup_{e \in \mathrm{path}_\mathcal{T}(b)} \mathcal{N}_e \big] &&\text{(by } u_a \geq u_b \text{ for } a \in \mathrm{path}_\mathcal{T}(b) \text{ since Part (c) of Lemma 13)} \\
&\leq \frac{2 m k}{\beta_b} \mathbb{1}\big[ a \in \cup_{e \in \mathrm{path}_\mathcal{T}(b)} \mathcal{N}_e \big] &&\text{(by Part (b) of Lemma 13)} \\
&\leq \frac{1}{2} \, ,
\end{aligned}
$$

where the last inequality follows by $\beta_b \geq 4mk$.

**Step 3. Bounding the stability term** Hence, taking the worst-case with respect to $p'$ in (30) with inequality (31), we can further bound the part of the second term in (29) as

$$\sum_{b=1}^{k} \beta_b \sum_{a=1}^{k} p_a \mathcal{S}_{q_b}^{\phi} \left( \frac{G(a, \Phi_{ax})_b}{\beta_b \, p_a} \right) \leq \sum_{a=1}^{k} p_a \sum_{b=1}^{k} \min \left\{ \beta_b \, \zeta \left( \frac{q_b G(a, \Phi_{ax})_b}{\beta_b \, p_a} \right), \, \gamma \beta_b (1 - q_b) \, \xi \left( -\frac{G(a, \Phi_{ax})_b}{\gamma \beta_b \, p_a} \right) \right\}, \quad (32)$$

where this holds from Proposition 11 with $\frac{z_b}{\beta_b} = \frac{G(a, \Phi_{ax})_b}{\beta_b p_a} \geq -1/q_b$ for all $b \in [k]$ that holds by (31), and Proposition 10. In the following we consider the inside of the summations in (32).

**Case $\gamma q_b \leq 1$:** First consider the case of $\gamma q_b \leq 1$. Then it holds that $\frac{1 - q_b}{\gamma q_b^2} \geq \frac{1 - 1/\gamma}{\gamma(1/\gamma)^2} = \gamma - 1 \geq 1$ since $\gamma \geq 2$. Using this and (28) with (31), we can bound the part of (32) by

$$\zeta \left( \frac{q_b G(a, \Phi_{ax})_b}{\beta_b \, p_a} \right) \leq \left( \frac{q_b G(a, \Phi_{ax})_b}{\beta_b \, p_a} \right)^2 \quad (33)$$

$$\leq \frac{1 - q_b}{\gamma q_b^2} \left( \frac{q_b G(a, \Phi_{ax})_b}{\beta_b \, p_a} \right)^2 = \gamma(1 - q_b) \left( \frac{G(a, \Phi_{ax})_b}{\gamma \beta_b \, p_a} \right)^2. \quad (34)$$

**Case $\gamma q_b > 1$:** Next we consider the case of $\gamma q_b > 1$. Now we check the stability condition for the complement negative Shannon entropy. Using (31) and $q_b \geq 1/\gamma$,

$$\left| \frac{G(a, \Phi_{ax})}{\gamma \beta_b p_a} \right| = \frac{1}{\gamma q_b} \left| \frac{q_b G(a, \Phi_{ax})}{\beta_b p_a} \right| \leq \frac{2mk}{\gamma \beta_b q_b} \leq \frac{2mk\gamma}{\gamma \beta_b} \leq 1, \quad (35)$$

where the last inequality holds by $\beta_b \geq 4mk$. Hence, using (28) with (31) and (27) with (35), we can bound (32) by

$$\sum_{a=1}^{k} p_a \sum_{b=1}^{k} \beta_b \min \left\{ \left( \frac{q_b G(a, \Phi_{ax})_b}{\beta_b \, p_a} \right)^2, \, \gamma(1 - q_b) \left( \frac{G(a, \Phi_{ax})_b}{\gamma \beta_b \, p_a} \right)^2 \right\}. \quad (36)$$

Summing up the arguments so far in (33), (34), and (36), we can bound (32) by

$$\sum_{a=1}^{k} p_a \sum_{b=1}^{k} \beta_b \min \left\{ \left( \frac{q_b G(a, \Phi_{ax})_b}{\beta_b \, p_a} \right)^2, \, \gamma(1 - q_b) \left( \frac{G(a, \Phi_{ax})_b}{\gamma \beta_b \, p_a} \right)^2 \right\}$$

$$= \sum_{a=1}^{k} \sum_{b=1}^{k} \frac{1}{\beta_b} \min \left\{ \frac{q_b^2}{p_a}, \, \frac{1 - q_b}{\gamma p_a} \right\} G(a, \Phi_{ax})_b^2. \quad (37)$$

**Step 4. Further bounding the stability term based on the property of water transfer operator** Using $p_a \geq u_a/2$, $\|w_e\|_\infty \leq m/2$, and Part (b) of Lemma 13, and recalling that $G(a, \sigma)_b = \sum_{e \in \text{path}_{\mathscr{I}}(b)} w_e(a, \sigma)$, (37) is bounded as

$$\sum_{a=1}^{k} \sum_{b=1}^{k} \frac{1}{\beta_b} \min \left\{ \frac{q_b^2}{p_a}, \, \frac{1 - q_b}{\gamma p_a} \right\} G(a, \Phi_{ax})_b^2$$

$$\leq 2 \sum_{a=1}^{k} \sum_{b=1}^{k} \frac{1}{\beta_b} \frac{1}{u_a} \min \left\{ q_b^2, \, \frac{1 - q_b}{\gamma} \right\} G(a, \Phi_{ax})_b^2 \qquad \text{(by } p_a \geq u_a/2)$$

$$= \sum_{a=1}^{k} \sum_{b=1}^{k} \frac{1}{\beta_b} \frac{1}{u_a} \min \left\{ q_b^2, \, \frac{1 - q_b}{\gamma} \right\} \left( \sum_{e \in \text{path}_{\mathscr{I}}(b)} w_e(a, \Phi_{ax}) \right)^2$$

$$\leq \frac{m^2}{2} \sum_{a=1}^{k} \sum_{b=1}^{k} \frac{1}{\beta_b} \frac{1}{u_a} \min \left\{ q_b^2, \, \frac{1 - q_b}{\gamma} \right\} \left( \sum_{e \in \text{path}_{\mathscr{I}}(b)} \mathbb{1}[a \in \mathcal{N}_e] \right)^2 \qquad \text{(by } \|w_e\|_\infty \leq m/2)$$

17

$$\leq 2m^2 \sum_{a=1}^{k} \sum_{b=1}^{k} \frac{1}{\beta_b} \frac{q_b}{u_a} \min\left\{ q_b, \frac{1-q_b}{\gamma q_b} \right\} \mathbb{1}\left[ a \in \cup_{e \in \text{path}_{\mathscr{G}}(b)} \mathcal{N}_e \right]$$

(since any action is in $\mathcal{N}_e$ at most two edges in $e \in \text{path}_{\mathscr{G}}(b)$)

$$\leq 2m^2 \sum_{a=1}^{k} \sum_{b=1}^{k} \frac{1}{\beta_b} \frac{q_b}{u_b} \min\left\{ q_b, \frac{1-q_b}{\gamma q_b} \right\} \mathbb{1}\left[ a \in \cup_{e \in \text{path}_{\mathscr{G}}(b)} \mathcal{N}_e \right]$$

(by $u_a \geq u_b$ for $a \in \text{path}_{\mathscr{G}}(b)$ since Part (c) of Lemma 13)

$$\leq 2m^2 k \sum_{a=1}^{k} \sum_{b=1}^{k} \frac{1}{\beta_b} \min\left\{ q_b, \frac{1-q_b}{\gamma q_b} \right\} \mathbb{1}\left[ a \in \cup_{e \in \text{path}_{\mathscr{G}}(b)} \mathcal{N}_e \right]$$

(by Part (b) of Lemma 13)

$$= 2m^2 k^2 \sum_{b=1}^{k} \frac{1}{\beta_b} \min\left\{ q_b, \frac{1-q_b}{\gamma q_b} \right\}.$$

**Step 5. Bounding the transformation term**  Owing to the property of the water transfer operator,

$$(p - q)^\top \mathcal{L}\nu = (1 - \lambda)(u - q)^\top \mathcal{L}\nu \leq 0,$$

where the last inequality follows by Part (a) of Lemma 13.

Summing up the arguments for bounding the stability and transformation terms completes the proof of Lemma 4. □

## F. Deferred Proofs for Locally Observable Games (Section 4)

This section provides the deferred proofs for locally observable games. Recall that the learning rate in (18) is defined as

$$\alpha_{ta} = \min\left\{ q_{ta}, \frac{1 - q_{ta}}{\gamma q_{ta}} \right\}, \quad \beta'_{ta} = c\sqrt{\alpha_0 + \frac{1}{\gamma} \sum_{s=1}^{t-1} \alpha_{sa}}, \quad \beta_{ta} = \max\{4mk, \beta'_{ta}\},$$

where $\alpha_0 = \frac{1}{\gamma^{3/2}} + \varepsilon$.

### F.1. Proof of Lemma 6

*Proof of Lemma 6.* Recall that $a^* = \arg\min_{a \in [k]} \mathbb{E}\left[ \sum_{t=1}^{T} \mathcal{L}_{ax_t} \right] \in \Pi$ is the optimal action in hindsight. Then the regret is decomposed as

$$\text{Reg}_T = \mathbb{E}\left[ \sum_{t=1}^{T} (\mathcal{L}_{A_t x_t} - \mathcal{L}_{a^* x_t}) \right] = \mathbb{E}\left[ \sum_{t=1}^{T} \sum_{b=1}^{k} p_{tb} (\mathcal{L}_{bx_t} - \mathcal{L}_{a^* x_t}) \right]$$

$$= \mathbb{E}\left[ \sum_{t=1}^{T} \sum_{b=1}^{k} (p_{tb} - q_{tb})(\mathcal{L}_{bx_t} - \mathcal{L}_{a^* x_t}) + \sum_{t=1}^{T} \sum_{b=1}^{k} q_{tb}(\mathcal{L}_{bx_t} - \mathcal{L}_{a^* x_t}) \right]$$

$$= \mathbb{E}\left[ \sum_{t=1}^{T} \sum_{b=1}^{k} (p_{tb} - q_{tb})\mathcal{L}_{bx_t} + \sum_{t=1}^{T} \sum_{b=1}^{k} q_{tb}(\widehat{y}_{tb} - \widehat{y}_{ta^*}) \right]. \tag{38}$$

The second term of the last equality is bounded from above as

$$\mathbb{E}\left[ \sum_{t=1}^{T} \sum_{b=1}^{k} q_{tb}(\widehat{y}_{tb} - \widehat{y}_{ta^*}) \right] = \mathbb{E}\left[ \sum_{t=1}^{T} \langle q_t - e_{a^*}, \widehat{y}_t \rangle \right] = \mathbb{E}\left[ \sum_{t=1}^{T} \langle q_t - u, \widehat{y}_t \rangle + \sum_{t=1}^{T} \langle u - e_{a^*}, \widehat{y}_t \rangle \right]$$

$$\leq \mathbb{E}\left[ \sum_{t=1}^{T} \langle q_t - u, \widehat{y}_t \rangle \right] + k, \tag{39}$$

18

where the last inequality follows since

$$\mathbb{E}[\langle u - e_{a^*}, \widehat{y}_t \rangle] = \mathbb{E}\left[\left\langle u - e_{a^*}, \sum_{c=1}^{k} G_t(c, \Phi_{cx_t}) \right\rangle\right]$$

$$= \mathbb{E}[\langle u - e_{a^*}, \mathcal{L}e_{x_t} + \delta \mathbf{1} \rangle] = \mathbb{E}\left[\frac{k}{T}\langle e_{a^*} - \mathbf{1}/k, \mathcal{L}e_{x_t} \rangle\right] \leq \frac{k}{T}, \tag{40}$$

where the second equality holds since by Lemma 2 there exists $\delta \in \mathbb{R}$ such that $\sum_{c=1}^{k} G_t(c, \Phi_{cx_t}) = \mathcal{L}e_{x_t} + \delta \mathbf{1}$. Combining (38) and (39) with Lemmas 3 and 4 completes the proof. $\qquad \square$

## F.2. Proof of Lemma 7

*Proof of Lemma 7.* First we consider bounding the sum of the optimal value $\mathsf{OPT}_{q_t}(\beta_t)$. The analysis was done in Section 3.2; By Lemma 4, the first summation is bounded by

$$\max\{0, \mathsf{OPT}_{q_t}(\beta_t)\} \leq 2m^2k^2\sum_{a=1}^{k}\frac{\alpha_{ta}}{\beta_{ta}} \leq 2m^2k^2\sum_{a=1}^{k}\frac{\alpha_{ta}}{\beta'_{ta}}.$$

Next we consider the penalty term. It holds that $\psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1}) \leq 0$ since $\psi_t(p) \leq \psi_{t+1}(p)$ for all $p \in \mathcal{P}_k$. Since $\psi_t \geq 0$ it suffices to bound $\psi_{T+1}(u)$. By the definition of the regularizer $\psi_t$ in (17), we have

$$\psi_t(u) = \sum_{a=1}^{k}\beta_{ta}\phi(u_a) \leq \max_{x \in [1/T,1]}\phi(x)\sum_{a=1}^{k}\beta_{ta} \leq \max\{\phi(1/T), \phi(1)\}\sum_{a=1}^{k}\beta_{ta},$$

where the first inequality follows by $u_i \geq 1/T$ for all $i \in [k]$ and the second inequality holds since $\phi$ is convex. Further, from the definition of $\phi(x) = (\phi^{\mathsf{LB}}(x) + x - 1) + \gamma \cdot (\phi^{\mathsf{CNS}}(x) + x)$ in (17),

$$\max\{\phi(1/T), \phi(1)\} = \max\left\{\log T + \frac{1}{T} - 1 + \gamma\left[\left(1 - \frac{1}{T}\right)\log\left(1 - \frac{1}{T}\right) + \frac{1}{T}\right], \gamma\right\}$$

$$\leq \max\left\{\log T + \frac{1 + \gamma}{T} - 1, \gamma\right\} = \gamma.$$

Hence, $\psi_{T+1}(u) \leq \gamma\sum_{a=1}^{k}\beta_{T+1,a}$. Summing up the arguments so far, we can bound the penalty term by

$$\psi_{T+1}(u) \leq \gamma\sum_{a=1}^{k}\beta_{T+1,a} \leq \gamma\sum_{a=1}^{k}\beta'_{T+1,a} + 4mk^2\gamma,$$

which follows by the definition of $\beta_{ta}$ and completes the proof of Lemma 7. $\qquad \square$

## F.3. Proof of Theorem 5

Finally, we are ready to prove Theorem 5 with Lemmas 6 and 7.

*Proof of Theorem 5.* We first bound the RHS of $\mathsf{OPT}_{q_t}(\beta_t) \leq 2m^2k^2\sum_{a=1}^{k}\frac{\alpha_{ta}}{\beta_{ta}}$ in Lemmas 7 from the definition of $\beta_t$ and $\alpha_t$ in the following. Since $\alpha_0 \geq \frac{1}{\gamma^{3/2}} \geq \frac{1}{\gamma}\alpha_t$,

$$\sum_{t=1}^{T}\frac{\alpha_{ta}}{\beta'_{ta}} = \frac{1}{c}\sum_{t=1}^{T}\frac{\alpha_{ta}}{\sqrt{\alpha_0 + \frac{1}{\gamma}\sum_{s=1}^{t-1}\alpha_{ta}}} \leq \frac{\gamma}{c}\sum_{t=1}^{T}\frac{\alpha_{ta}/\gamma}{\sqrt{\alpha_0 - \frac{1}{\gamma^{3/2}} + \frac{1}{\gamma}\sum_{s=1}^{t}\alpha_{ta}}}$$

$$\leq \frac{2\gamma}{c}\left(\sqrt{\alpha_0 - \frac{1}{\gamma^{3/2}} + \frac{1}{\gamma}\sum_{t=1}^{T}\alpha_{ta}} - \sqrt{\alpha_0 - \frac{1}{\gamma^{3/2}}}\right) \leq \frac{2\gamma}{c^2}\left(\beta'_{T+1,a} - \beta'_{1,a}\right) + \frac{2}{c\gamma^{5/4}}, \tag{41}$$

where the last inequality follows since

$$\frac{2}{c^2}\beta'_{1,a} - \frac{2}{c}\sqrt{\alpha_0 - \frac{1}{\gamma^{3/2}}} = \frac{2}{c}\left(\sqrt{\alpha_0} - \sqrt{\alpha_0 - \frac{1}{\gamma^{3/2}}}\right) \le \frac{2}{c}\frac{1}{\gamma^{3/2}}\frac{1}{\sqrt{\alpha_0} + \sqrt{\alpha_0 - \frac{1}{\gamma^{3/2}}}} \le \frac{2}{c\gamma^{9/4}}.$$

Hence combining Lemmas 6 and 7 with (41) gives

$$\mathsf{Reg}_T \le \mathbb{E}\left[\sum_{t=1}^{T}\gamma\left(1 + \frac{4m^2k^2}{c^2}\right)\sum_{a=1}^{k}\beta'_{T+1,a}\right] + 4mk^2\gamma + \frac{4m^2k^2}{c\gamma^{5/4}} + \frac{mk^3}{2}$$

$$= \mathbb{E}\left[2\gamma\sum_{t=1}^{T}\sum_{a=1}^{k}\beta'_{T+1,a}\right] + 4mk^2\gamma + \frac{4m^2k^2}{c\gamma^{5/4}}, \tag{42}$$

where the equality follows from $c = 2mk$.

In the following, we bound $\beta'_{T+1,a}$ for suboptimal arm $a \neq a^*$ and optimal arms $a^*$ separately. For suboptimal arm $a \neq a^*$, the definition of $\beta'_{ta}$ in (18) implies that

$$\beta'_{T+1,a} \le c\sqrt{\alpha_0 + \frac{\sum_{t=1}^{T}q_{ta}}{\gamma}}. \tag{43}$$

Next we consider optimal action $a^*$. Since $\sqrt{\gamma} \ge 2$ for any $z \in [0,1]$, it holds that

$$\min\left\{z, \frac{1-z}{\gamma z}\right\} \le \begin{cases} z & \text{if } z \le \frac{1}{\sqrt{\gamma}} \\ \frac{1-z}{\sqrt{\gamma}} & \text{otherwise} \end{cases} \le \frac{2}{\sqrt{\gamma}}(1-z).$$

Using this inequality,

$$\sum_{t=1}^{T}\alpha_{ta^*} \le \frac{2}{\sqrt{\gamma}}\sum_{t=1}^{T}(1 - q_{ta^*}) = \frac{2}{\sqrt{\gamma}}\sum_{t=1}^{T}\sum_{a\neq a^*}q_{ta},$$

which implies

$$\beta'_{T+1,a^*} \le c\sqrt{\alpha_0 + \frac{2}{\gamma^{3/2}}\sum_{t=1}^{T}\sum_{a\neq a^*}q_{ta}}. \tag{44}$$

Summing up the arguments so far by using (43) and (44) with (42) and letting $Q_a = \mathbb{E}\left[\sum_{t=1}^{T}q_{ta}\right]$, we can bound the regret as

$$\mathsf{Reg}_T \le 2\gamma\sum_{a\neq a^*}\mathbb{E}\left[c\sqrt{\alpha_0 + \sum_{t=1}^{T}\frac{q_{ta}}{\gamma}}\right] + 4mk^2\gamma + \gamma c\sqrt{\alpha_0 + \frac{2}{\gamma^{3/2}}\sum_{t=1}^{T}\sum_{a\neq a^*}q_{ta}} + \frac{4m^2k^2}{c\gamma^{5/4}}$$

$$\le 2\gamma\sum_{a\neq a^*}c\sqrt{\alpha_0 + \frac{1}{\gamma}\mathbb{E}\left[\sum_{t=1}^{T}q_{ta}\right]} + 4mk^2\gamma + \gamma c\sqrt{\alpha_0 + \frac{2}{\gamma^{3/2}}\sum_{a\neq a^*}\mathbb{E}\left[\sum_{t=1}^{T}q_{ta}\right]} + \frac{4m^2k^2}{c\gamma^{5/4}}$$

$$\le 2\gamma\sum_{a\neq a^*}c\sqrt{\alpha_0 + \frac{Q_a}{\gamma}} + 4mk^2\gamma + \gamma c\sqrt{\alpha_0 + \frac{2}{\gamma^{3/2}}\sum_{a\neq a^*}Q_a} + \frac{4m^2k^2}{c\gamma^{5/4}}. \tag{45}$$

For the adversarial environment, using the Cauchy–Schwarz inequality and $\sum_{a=1}^{k}q_{ta} = 1$ gives

$$\mathsf{Reg}_T \le 2\gamma c\sqrt{k\left(\alpha_0 + \frac{T}{\gamma}\right)} + 4mk^2\gamma + \gamma c\sqrt{k\left(\alpha_0 + \frac{2T}{\gamma^{3/2}}\right)} + \frac{4m^2k^2}{c\gamma^{5/4}}$$

$$\le 6mk^{3/2}\sqrt{T\log T} + 4mk^2\log T + 12mk\sqrt{k\alpha_0\log T} + \frac{2mk}{(\log T)^{5/4}}.$$

---

**Algorithm 2** Algorithm for globally observable games

---
    **for** $t = 1, 2, \ldots$ **do**
        Compute $q_t$ via FTRL (4) with regularizer (23).
        Compute $p_t$ from $q_t$ by (24).
        Sample $A_t \sim p_t$, observe $\Phi_{A_t x_t} \in \Sigma$, and compute $\widehat{y}_t = G^{\circ}(A_t, \sigma_t)/p_{tA_t}$.
    **end for**

---

For the stochastic environment with adversarial corruptions, we have $\mathsf{Reg}_T \geq \frac{1}{2}\sum_{a \neq a^*} \Delta_a Q_a - C$ by $p_t \in \mathcal{R}(q)$. Hence for any $\lambda \in (0, 1]$, we have

$$\mathsf{Reg}_T = (1 + \lambda)\mathsf{Reg}_T - \lambda\mathsf{Reg}_T$$

$$\leq 2(1 + \lambda)\gamma \sum_{a \neq a^*} c\sqrt{\alpha_0 + \frac{Q_a}{\gamma}} + (1 + \lambda)\gamma c\sqrt{\alpha_0 + \frac{2}{\gamma^{3/2}}\sum_{a \neq a^*} Q_a} - \frac{\lambda}{2}\sum_{a \neq a^*} \Delta_a Q_a$$

$$+ (1 + \lambda)mk^2\gamma + O(mk) + \lambda C \qquad\qquad \text{(By (45) and } \mathsf{Reg}_T \geq \frac{1}{2}\sum_{a \neq a^*} \Delta_a Q_a)$$

$$\leq \sum_{a \neq a^*}\left(2(1 + \lambda)c\sqrt{\gamma Q_a} - \frac{\lambda}{4}\Delta_a Q_a\right) + (1 + \lambda)c\sqrt{2\sqrt{\gamma}\sum_{a \neq a^*} Q_a} - \frac{\lambda}{4}\Delta\sum_{a \neq a^*} Q_a$$

$$+ (1 + \lambda)mk^2\gamma + 12mk\sqrt{k\alpha_0 \log T} + O(mk) + \lambda C$$

$$\leq \sum_{a \neq a^*}\frac{64(1 + \lambda)^2 c^2}{\lambda}\frac{\gamma}{\Delta_a} + \frac{32(1 + \lambda)^2 c^2}{\lambda}\frac{\sqrt{\gamma}}{\Delta} + (1 + \lambda)mk^2\gamma + 12mk\sqrt{k\alpha_0 \log T} + O(mk) + \lambda C,$$

where the last inequality follows from $bx - ax^2 \leq b^2/(4a)$ for $a > 0$ and $b \geq 0$. Setting optimal $\lambda$ completes the proof of Theorem 5. $\qquad\qquad\square$

# G. Deferred Proofs for Globally Observable Games (Theorem 8, Section 5)

This appendix provides the proof of Theorem 8. Let $u \in \mathcal{P}_k$ be the projection of $e_{a^*}$ on $[1/T, 1-1/T]^k \cap \mathcal{P}_k$. In particular, define $u \in [1/T, 1-1/T]^k \cap \mathcal{P}_k$ by

$$u = \underset{u' \in [1/T, 1-1/T]^k \cap \mathcal{P}_k}{\arg\min} \|u' - e_{a^*}\|_\infty.$$

We first decompose the regret as follows.

**Lemma 14.** *The regret of Algorithm 2 is bounded as* $\mathsf{Reg}_T \leq \mathbb{E}\big[\sum_{t=1}^T \gamma_t + \sum_{t=1}^T\big(\langle\widehat{y}_t, q_t - q_{t+1}\rangle - D_{\psi_t}(q_{t+1}, q_t)\big) + \sum_{t=1}^T\big(\psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1})\big) + \psi_{T+1}(u) - \psi_1(q_1)\big] + c_{\mathcal{G}}.$

This lemma can be proven in a very similar manner as Lemma 6. Note that the first, second, and third summations correspond to the transformation, stability, and penalty terms, respectively.

*Proof of Lemma 14.* We first decompose the regret as follows:

$$\mathsf{Reg}_T = \mathbb{E}\left[\sum_{t=1}^T (\mathcal{L}_{A_t x_t} - \mathcal{L}_{a^* x_t})\right] = \mathbb{E}\left[\sum_{t=1}^T \langle p_t - e_{a^*}, \mathcal{L}e_{x_t}\rangle\right] = \mathbb{E}\left[\sum_{t=1}^T \langle q_t - e_{a^*}, \mathcal{L}e_{x_t}\rangle + \sum_{t=1}^T \gamma_t\left\langle\frac{1}{k}\mathbf{1} - q_t, \mathcal{L}e_{x_t}\right\rangle\right]$$

$$\leq \mathbb{E}\left[\sum_{t=1}^T \langle q_t - e_{a^*}, \mathcal{L}e_{x_t}\rangle + \sum_{t=1}^T \gamma_t\right] = \mathbb{E}\left[\sum_{t=1}^T\sum_{a=1}^k q_{ta}(\mathcal{L}_{ax_t} - \mathcal{L}_{a^* x_t}) + \sum_{t=1}^T \gamma_t\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^T\sum_{a=1}^k q_{ta}(\widehat{y}_{ta} - \widehat{y}_{ta^*}) + \sum_{t=1}^T \gamma_t\right] = \mathbb{E}\left[\sum_{t=1}^T \langle q_t - e_{a^*}, \widehat{y}_t\rangle + \sum_{t=1}^T \gamma_t\right], \tag{46}$$

where the inequality follows from the fact that $\mathcal{L} \in [0, 1]^{k \times d}$, and the fifth equality follows from the definitions of $\widehat{y}_t$ and

$q_{ta} = 0$ for $a \notin \Pi$. The first term in (46) is further bounded as

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle q_t - e_{a^*}, \widehat{y}_t \rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle q_t - u, \widehat{y}_t \rangle + \sum_{t=1}^{T} \langle u - e_{a^*}, \widehat{y}_t \rangle\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \langle q_t - u, \widehat{y}_t \rangle\right] + k\,, \tag{47}$$

where the last inequality follows by the same argument in (40). Combining (46) and (47) with Lemma 3 completes the proof of Lemma 14. $\qquad\square$

We can bound the penalty and stability terms on the RHS of Lemma 14 as follows.

**Lemma 15.** *The penalty term is bounded by* $\sum_{t=1}^{T}(\psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1})) + \psi_{T+1}(u) - \psi_1(q_1) \leq 2\beta_{T+1}k\log T$. *The stability term is bounded by*

$$\langle q_t - q_{t+1}, \widehat{y}_t \rangle - D_{\psi_t}(q_{t+1}, q_t) \leq \frac{1}{\beta_t}\sum_{b=1}^{k} \min\left\{q_{tb}^2, (1-q_{tb})^2\right\}\widehat{y}_{tb}^2\,.$$

*Proof of Lemma 15.* The penalty term is bounded by

$$\sum_{t=1}^{T}(\psi_t(q_{t+1}) - \psi_{t+1}(q_{t+1})) + \psi_{T+1}(u) - \psi_1(q_1) \leq \psi_{T+1}(u) \leq 2\beta_{T+1}k\log T\,,$$

where the last inequality holds since $u \in [1/T, 1 - 1/T]^k$. Next we consider the stability term. Taking the worst case with respect to $q_{t+1}$,

$$\langle q_t - q_{t+1}, \widehat{y}_t \rangle - D_{\psi_t}(q_{t+1}, q_t) = \sum_{b=1}^{k}\left((q_{tb} - q_{t+1,b})\widehat{y}_{tb} - \beta_t\left(D_{\phi^{\mathsf{LB}}}(q_{t+1,b}, q_{tb}) + D_{\phi^{\mathsf{CLB}}}(q_{t+1,b}, q_{tb})\right)\right)$$

$$\leq \sum_{b=1}^{k} \min\left\{\left((q_{tb} - q_{t+1,b})\widehat{y}_{tb} - \beta_t D_{\phi^{\mathsf{LB}}}(q_{t+1,b}, q_{tb}), (q_{tb} - q_{t+1,b})\widehat{y}_{tb} - \beta_t D_{\phi^{\mathsf{CLB}}}(q_{t+1,b}, q_{tb})\right)\right\}$$

$$\leq \frac{1}{\beta_t}\sum_{b=1}^{k} \min\left\{q_{tb}^2, (1-q_{tb})^2\right\}\widehat{y}_{tb}^2\,.$$

The last inequality can be proven from Propositions 11 and 12 as follows. For fixed $b \in [k]$ and any $q_{tb} \leq 1/2$, we have

$$(q_{tb} - q_{t+1,b})\widehat{y}_{tb} - \beta_t D_{\phi^{\mathsf{LB}}}(q_{t+1,b}, q_{tb}) = \beta_t\left((q_{tb} - q_{t+1,b})\frac{\widehat{y}_{tb}}{\beta_t} - D_{\phi^{\mathsf{LB}}}(q_{t+1,b}, q_{tb})\right)$$

$$\leq \frac{1}{\beta_t}q_{tb}^2\widehat{y}_{tb}^2 \leq \frac{1}{\beta_t}\min\left\{q_{tb}^2, (1-q_{tb})^2\right\}\widehat{y}_{tb}^2\,,$$

where the second inequality holds by $q_{tb} \leq 1/2$ and the first inequality holds from Proposition 11 with

$$\left|\frac{\widehat{y}_{tb}}{\beta_t}\right| \leq \frac{\|G^\circ\|_\infty}{|\beta_t\, p_{tA_t}|} \leq \|G^\circ\|_\infty\left|\frac{k}{\beta_t\gamma_t}\right| \leq \|G^\circ\|_\infty\left|\frac{k}{\sqrt{2\beta_t}c_{\mathcal{G}}z_t}\right| \leq \|G^\circ\|_\infty\left|\frac{k}{\sqrt{2\beta_t}c_{\mathcal{G}}q_{tb}}\right| \leq \frac{1}{2q_{tb}}\,.$$

Here the forth inequality holds due to $z_t \geq \min\{q_{tb}, 1 - q_{tb}\} = q_{tb}$ and the last inequality follows by $\beta_t \geq \frac{2k^2}{c_{\mathcal{G}}^2}$. For fixed $b \in [k]$ and any $q_{tb} > 1/2$ it also holds that

$$(q_{tb} - q_{t+1,b})\widehat{y}_{tb} - \beta_t D_{\phi^{\mathsf{CLB}}}(q_{t+1,b}, q_{tb}) \leq \frac{1}{\beta_t}(1-q_{tb})^2\widehat{y}_{tb}^2 \leq \frac{1}{\beta_t}\min\left\{q_{tb}^2, (1-q_{tb})^2\right\}\widehat{y}_{tb}^2\,,$$

where the last inequality follows by $q_{tb} > 1/2$, and the first inequality holds from Propositions 11 with

$$\left|\frac{\widehat{y}_{tb}}{\beta_t}\right| \leq \frac{\|G^\circ\|_\infty}{|\beta_t\, p_{tA_t}|} \leq \|G^\circ\|_\infty\left|\frac{k}{\beta_t\gamma_t}\right| \leq \|G^\circ\|_\infty\left|\frac{k}{\sqrt{2\beta_t}c_{\mathcal{G}}z_t}\right| \leq \|G^\circ\|_\infty\left|\frac{k}{\sqrt{2\beta_t}c_{\mathcal{G}}(1-q_{tb})}\right| \leq \frac{1}{2(1-q_{tb})}\,.$$

$\qquad\square$

Finally we are ready to prove Theorem 8.

*Proof of Theorem 8.* First, taking expectation with respect to $A_t \sim p_t$, we can further bound the stability term as

$$\mathbb{E}_t\left[\frac{1}{\beta_t}\sum_{b=1}^k \min\left\{q_{tb}^2, (1-q_{tb})^2\right\}\widehat{y}_{tb}^2\right] = \frac{1}{\beta_t}\sum_{a=1}^k \frac{1}{p_{ta}}\sum_{b=1}^k \min\left\{q_{tb}^2, (1-q_{tb})^2\right\} G(a, \Phi_{ax_t})_b^2 \leq \frac{c_{\mathcal{G}}^2}{\beta_t\gamma_t}z_t^2 \,.$$

Combining this inequality with Lemmas 14 and 15 yields

$$\mathsf{Reg}_T \leq \mathbb{E}\left[2\beta_{T+1}k\log T + \sum_{t=1}^T\left(\frac{c_{\mathcal{G}}^2}{\gamma_t\beta_t}z_t^2 + \gamma_t\right)\right] + c_{\mathcal{G}} = \mathbb{E}\left[2\beta_{T+1}k\log T + 2c_{\mathcal{G}}\sum_{t=1}^T \frac{z_t}{\sqrt{\beta_t}}\right] + k\,.$$

From the definition of $\beta_t$,

$$\sqrt{c_1}\sum_{t=1}^T \frac{z_t}{\sqrt{\beta_t}} = \sum_{t=1}^T \frac{z_t}{\left(k + \sum_{s=1}^{t-1} z_s\right)^{1/3}} \leq \sum_{t=1}^T \frac{z_t}{\left(\sum_{s=1}^t z_s\right)^{1/3}} \leq \int_0^{\sum_{t=1}^T z_t} x^{-1/3}\mathrm{d}x = \frac{3}{2}\left(\sum_{t=1}^T z_t\right)^{2/3}\,.$$

Using this inequality and Jensen's inequality,

$$\mathsf{Reg}_T \leq \left(2c_1 k\log T + \frac{3c_{\mathcal{G}}}{\sqrt{c_1}}\right)\left(k + \mathbb{E}\left[\sum_{t=1}^T z_t\right]\right)^{2/3} = O\left(c_{\mathcal{G}}^{2/3}(k\log T)^{1/3}\left(k + \mathbb{E}\left[\sum_{t=1}^T z_t\right]\right)^{2/3}\right)\,.$$

For adversarial environments, since $z_t \leq 1$, $\mathsf{Reg}_T = O\big((c_{\mathcal{G}}^2 k\log T)^{1/3}T^{2/3}\big)$. For stochastic environments, using $z_t \leq \min_{b\in[k]}\sum_{a\neq b} q_{ta} \leq 1 - q_{ta^*}$, for any $\lambda \in (0,1]$,

$$\mathsf{Reg}_T = (1+\lambda)\mathsf{Reg}_T - \lambda\mathsf{Reg}_T$$

$$\leq O\left((1+\lambda)c_{\mathcal{G}}^{2/3}(k\log T)^{1/3}\left(k + \mathbb{E}\left[\sum_{t=1}^T(1-q_{ta^*})\right]\right)^{2/3}\right) - \frac{\lambda}{2}\Delta\mathbb{E}\left[\sum_{t=1}^T(1-q_{ta^*})\right] + \lambda C$$

$$\leq O\left(\frac{(1+\lambda)^3}{\lambda^2}\frac{c_{\mathcal{G}}^2 k\log T}{\Delta^2} + c_{\mathcal{G}}^{2/3}(\log T)^{1/3}k^{4/3}\right) + \lambda C$$

$$\leq O\left(\frac{c_{\mathcal{G}}^2 k\log T}{\Delta^2} + \frac{1}{\lambda^2}\frac{c_{\mathcal{G}}^2 k\log T}{\Delta^2} + c_{\mathcal{G}}^{2/3}(\log T)^{1/3}k^{4/3}\right) + \lambda C\,,$$

where the second inequality follows by $bx^2 - ax^3 \leq \frac{4b^3}{27a^2}$ and the last inequality follows by $\lambda \in (0,1]$. Setting $\lambda = O\left(\left(\frac{c_{\mathcal{G}}^2 k\log T}{C}\right)^{1/3}\right)$ completes the proof of Theorem 8. $\square$