Instance-Dependent Regret Bounds for Learning Two-Player Zero-Sum Games with Bandit Feedback

Shinji Ito SHINJI@MIST.I.U-TOKYO.AC.JP
The University of Tokyo; RIKEN AIP

Haipeng Luo HAIPENGL@USC.EDU
University of Southern California

Taira Tsuchiya TSUCHIYA@MIST.I.U-TOKYO.AC.JP
The University of Tokyo; RIKEN AIP

Yue Wu WU.YUE@USC.EDU
University of Southern California

Editors: Nika Haghtalab and Ankur Moitra

Abstract

No-regret self-play learning dynamics have become one of the premier ways to solve large-scale games in practice. Accelerating their convergence via improving the regret of the players over the naive $O(\sqrt{T})$ bound after T rounds has been extensively studied in recent years, but almost all studies assume access to exact gradient feedback. We address the question of whether acceleration is possible under bandit feedback only and provide an affirmative answer for two-player zero-sum normal-form games. Specifically, we show that if both players apply the Tsallis-INF algorithm of Zimmert and Seldin (2021), then their regret is at most $O(c_1 \log T + \sqrt{c_2 T})$, where c_1 and c_2 are game-dependent constants that characterize the difficulty of learning —— c_1 resembles the complexity of learning a stochastic multi-armed bandit instance and depends inversely on some gap measures, while c_2 can be much smaller than the number of actions when the Nash equilibria have a small support or are close to the boundary. In particular, for the case when a pure strategy Nash equilibrium exists, c_2 becomes zero, leading to an optimal instance-dependent regret bound as we show. We additionally prove that in this case our algorithm also enjoys last-iterate convergence and can identify the pure strategy Nash equilibrium with near-optimal sample complexity.

Keywords: two-player zero-sum games, bandit feedback, instance-dependent regret

1. Introduction

Since the early studies that reveal the fundamental connection between online learning and game theory (Foster and Vohra, 1997; Freund and Schapire, 1999; Hart and Mas-Colell, 2000), no-regret uncoupled learning dynamics have been heavily studied and become one of the most efficient ways for robustly learning in games and finding equilibria. Indeed, they are the foundation for recent AI breakthroughs such as superhuman AI for poker (Bowling et al., 2015; Moravčík et al., 2017; Brown and Sandholm, 2018, 2019), human-level AI for Stratego (Perolat et al., 2022) and Diplomacy (FAIR et al., 2022), and even alignment of large language models (Jacob et al., 2024; Munos et al., 2024).

The most basic result in this area states that a no-regret self-play learning dynamic converges to some equilibrium, with the convergence rate governed by the time-averaged regret, which is usually $O(1/\sqrt{T})$ after T rounds if one directly uses worst-case $O(\sqrt{T})$ regret bounds from the adversarial online learning literature. However, there is an extensive body of research on accelerating the convergence rate by exploiting the self-play nature and game structures to achieve lower regret,

© 2025 S. Ito, H. Luo, T. Tsuchiya & Y. Wu.

from earlier studies on two-player zero-sum games (Daskalakis et al., 2011; Rakhlin and Sridharan, 2013) to more recent ones on multi-player general-sum games (Syrgkanis et al., 2015; Chen and Peng, 2020; Daskalakis et al., 2021; Farina et al., 2022; Anagnostides et al., 2022b,a). Importantly, these works all assume access to exact gradient feedback.

In contrast, far less effort has been dedicated to the more realistic setting with *bandit feedback* only (that is, each player only observes their noisy payoff in each round), partly because improving over $O(\sqrt{T})$ regret now becomes impossible in the worst case. To get around this barrier, researchers either relax the feedback model (Rakhlin and Sridharan, 2013; Wei and Luo, 2018) or consider different notions of regret (O'Donoghue et al., 2021; Maiti et al., 2023b).

Nevertheless, the aforementioned barrier does not mean that one cannot achieve better than $O(\sqrt{T})$ regret *in all instances*. Indeed, in stochastic m-armed bandits, a problem that can be seen as a special case with one player only, even though $O(\sqrt{mT})$ regret is also unavoidable in the worst case, there are plenty of studies on obtaining *instance-dependent* $o(\sqrt{T})$ regret, via e.g., the Upper Confidence Bound (UCB) algorithm (Auer, 2002). This begs the question: *can we also achieve good instance-dependent regret bounds for self-play learning dynamics with bandit feedback*?

In this work, we provide an affirmative and comprehensive answer to this question for the case of two-player zero-sum normal-form games. Importantly, our results are built on the recent advances on *best-of-both-worlds* for multi-armed bandits (MAB) that use surprisingly simple algorithms and analysis to simultaneously achieve the optimal worst-case regret in the adversarial setting and the optimal instance-dependent regret in the stochastic setting (see e.g., Zimmert and Seldin, 2021). Our work shows that it is possible to extend such techniques to the game setting and achieve similar best-of-both-worlds phenomena, where the two worlds here refer to the case when playing against an arbitrary opponent and the case when playing against the same algorithm (self-play). More specifically, we consider an uncoupled learning dynamic where both players simply apply the Tsallis-INF algorithm of Zimmert and Seldin (2021), a well-known best-of-both-worlds algorithm for MAB, and show the following guarantees.

- For general zero-sum games with possibly mixed Nash equilibria (NE), each player's regret can be bounded by two terms: an $O(c_1 \log T)$ term where c_1 is a game-dependent constant that characterizes the difficulty of learning in this game in a way analogous to stochastic MAB, and an $O(\sqrt{c_2 T})$ term where c_2 is also game-dependent and can be much smaller than the number of actions (the trivial bound) — for example, c_2 is small when the support of an NE is small or when all NE are close to the boundary of the strategy space. See Section 3 for the exact bounds. We also construct an example where our algorithm provably enjoys $o(\sqrt{T})$ regret and verify it empirically in Section 5.

- When specifying our results to the special case with a pure strategy NE (PSNE), our regret bound only contains the $O(c_1 \log T)$ term. In fact, the regret bound is quantitatively similar to playing two stochastic MAB instances with the expected payoff vectors being the row and the column of the game respectively that contain the PSNE. We further prove that no reasonable algorithms can improve over this bound. Moreover, we also show that, somewhat surprisingly, our algorithm enjoys not only average-iterate convergence but also *last-iterate* convergence, a much more preferable convergence guarantee when one cares about the day-to-day behavior of the learning dynamic. See Section 4 for details.

- As a by-product of our regret guarantee, we also prove that in the case with a PSNE, after running Tsallis-INF for a certain number of rounds, the pair of the most frequently selected action of each player is the PSNE with a constant probability, which can boosted to $1-\delta$ by repeating the procedure $\log(1/\delta)$ times. Although the sample complexity of our algorithm could be $\sqrt{m}$ factor larger than that of the optimal algorithm by Maiti et al. (2024), we emphasize that their algorithm controls both players in a centralized and coupled manner, while ours is a decentralized and uncoupled learning dynamic that additionally enjoys no-regret guarantees (even when the opponent deviates and plays arbitrarily).

Related work A line of work that is closely related to ours studies instance-dependent sample complexity (instead of regret) for finding an NE via querying entries of a zero-sum matrix game and obtaining noisy samples (Maiti et al., 2023a, 2024). This can be seen as a generalization of instance-dependent sample complexity from the best-arm identification problem (e.g., Jamieson and Nowak, 2014) to the game setting, while our work is a generalization of instance-dependent regret from stochastic MAB (e.g., Auer, 2002; Garivier and Cappé, 2011) to the game setting. We note that sample complexity generally does not imply any regret guarantees, but the latter can be translated to the former in some cases, and we discuss such translations and compare our bounds to Maiti et al. (2023a, 2024) in Sections 3 and 4.3.

Another closely related topic is dueling bandits (Yue et al., 2012), which in fact can be seen as a special case of playing a skew-symmetric zero-sum game. The idea of sparring, first proposed by Ailon et al. (2014), is equivalent to the self-play dynamic considered here, and the so-called "strong regret" in dueling bandits coincides with the sum of the individual regret of the two players (so all our results apply directly). Most work in dueling bandits assumes the existence of a Condorcet winner, which is equivalent to the existence of a PSNE, and some develops instance-dependent regret that is quantitatively similar to those for stochastic MAB (Yue and Joachims, 2011; Yue et al., 2012; Zoghi et al., 2014). Using Tsallis-INF algorithm to achieve both instance-dependent regret bounds and worst-case robustness has also been studied in Zimmert and Seldin (2021); Saha and Gaillard (2022); Saad et al. (2024), and our bound for the case with a PSNE is similar to theirs and can be seen as a generalization. For the general case when a Condorcet winner might not exist, Dudík et al. (2015) propose the concept of von Neumann winners, which is essentially the same as mixed NE. Assuming a unique von Neumann winner, Balsubramani et al. (2016) provide an instance-dependent regret in the form of $O(\sqrt{sT})$ (ignoring additive terms that are problem-dependent), where s is the size of the support of the von Neumann winner. This bound is closely related to one instantiation of our bound for general zero-sum games, but there are several other advantages of our methods, such as a much simpler algorithm, no requirement on uniqueness, and the fact that the bound holds for both player's individual regret instead of their sum only.

As mentioned, our results are built on the recent line of work on using simple Follow-the-Regularized-Leader algorithm to achieve best-of-both-worlds for MAB, an idea first proposed by Wei and Luo (2018) and later improved and extended to various settings (e.g., Rouyer and Seldin, 2020; Jin and Luo, 2020; Jin et al., 2021; Zimmert and Seldin, 2021; Ito, 2021; Erez and Koren, 2021; Rouyer et al., 2021; Ito et al., 2022; Amir et al., 2022; Masoudian et al., 2022; Tsuchiya et al., 2023; Jin et al., 2023a,b). Even though Zimmert and Seldin (2021); Saha and Gaillard (2022); Saad et al. (2024) already extend the idea to a special case of dueling bandit (which itself is a special case of zero-sum games), our work is the first to extend it to general zero-sum games, which requires new ideas and sheds light on how to further extend these techniques to more challenging settings (e.g., multi-player general-sum games).

There is a surge of studies on understanding the last-iterate convergence of self-play learning dynamics for zero-sum games, especially for the setting with exact gradient feedback (Daskalakis and Panageas, 2019; Mertikopoulos and Zhou, 2019; Golowich et al., 2020; Wei et al., 2021; Hsieh et al., 2021; Gorbunov et al., 2022; Cai et al., 2022, 2024a,b). For the bandit setting, the convergence rate is usually much slower; see e.g., Cai et al. (2023) where an $\tilde{O}(1/T^{\frac{1}{6}})$ rate is obtained. We show that for the special case with a PSNE, the simple Tsallis-INF algorithm already achieves $\tilde{O}(1/\sqrt{T})$ last-iterate convergence (albeit with some instance-dependent constant). Unlike previous analysis, ours is solely based on analyzing the regret, which might be of independent interest.

2. Preliminaries

In this section, we formally describe concepts related to two-player zero-sum games, self-play learning dynamics, and our main algorithm.

Notations Throughout this paper, we will use $\log(\cdot)$ to denote base-2 logarithm, $\ln(\cdot)$ to denote base-e logarithm, and use $\log_+ x = \max\{1, \log x\}$. We use $\tilde{O}(\cdot)$ to hide logarithmic factors; formally, $f(x) = \tilde{O}(g(x))$ means that there exists a positive integer k such that $f(x) = O(g(x) \log^k g(x))$.

Two-Player Zero-Sum Normal-Form Games A two-player zero-sum normal-form game is defined via a payoff matrix $A \in [-1, 1]^{m \times n}$, where m and n are the number of actions for the row player and the column player respectively. When the row player plays action i and the column player plays action j, the entry $A(i, j) \in [-1, 1]$ is the expected reward for the row player and also the expected loss for the column player (hence zero-sum).

The players also have the option to play according to a probability distribution over their actions, or a *mixed strategy*. Let $\mathcal{P}_m = \{x \in [0, 1]^m \mid \|x\|_1 = 1\}$ be the probability simplex of size m. Given mixed strategies $x \in \mathcal{P}_m$ and $y \in \mathcal{P}_n$ of the row and column players, the expected reward for the row player is given by $x^\top Ay$, which is also the expected loss for the column player.

A pair of mixed strategies $(x_\star, y_\star) \in \mathcal{P}_m \times \mathcal{P}_n$ is a *Nash equilibrium* (NE) if $x^\top Ay_\star \leq x_\star^\top Ay_\star \leq x_\star^\top Ay$ hold for all $x \in \mathcal{P}_m$ and $y \in \mathcal{P}_n$. The celebrated Minimax theorem (von Neumann, 1928) implies that $(x_\star, y_\star)$ is an NE if and only if $x_\star \in \mathcal{X}_\star = \arg\max_x \{\min_y x^\top Ay\}$ and $y_\star \in \mathcal{Y}_\star = \arg\min_y \{\max_x x^\top Ay\}$.

A pure-strategy Nash equilibrium (PSNE) is a Nash equilibrium $(x_\star, y_\star)$ where both players choose a pure strategy, i.e., $x_\star \in \{0, 1\}^m$ and $y_\star \in \{0, 1\}^n$. A PSNE is also denoted by $(i_\star, j_\star)$ where $i_\star \in [m]$ and $j_\star \in [n]$ are the indices of the non-zero entries of $x_\star$ and $y_\star$, respectively. The *duality gap* for $(\hat{x}, \hat{y}) \in \mathcal{P}_m \times \mathcal{P}_n$ is defined by

$$\mathrm{DGap}(\hat{x}, \hat{y}) = \max_{x \in \mathcal{P}_m, y \in \mathcal{P}_n} \left\{ x^\top A\hat{y} - \hat{x}^\top Ay \right\} \geq 0, \tag{1}$$

which measures how far $(\hat{x}, \hat{y})$ is from a Nash equilibrium. Indeed, $(x_\star, y_\star)$ is a Nash equilibrium if and only if $\mathrm{DGap}(x_\star, y_\star) = 0$.

Learning Dynamics with Bandit Feedback We consider a realistic setting where both players have no prior information about the game and repeatedly play the game with bandit feedback for T rounds. Specifically, in each round $t = 1, 2, \ldots, T$, the row player chooses a mixed strategy $x_t \in \mathcal{P}_m$, and the column player chooses $y_t \in \mathcal{P}_n$. They each draw their action $i_t \in [m]$ and $j_t \in [n]$ from their mixed strategy, independently from each other. The nature then draws an outcome $r_t \in [-1, 1]$ with expectation $\mathbf{E}[r_t \mid i_t, j_t] = A(i_t, j_t)$ and reveals it to the row player as their realized reward and to the column player as their realized loss.[1] Note that this is a strongly uncoupled learning dynamic as defined by Daskalakis et al. (2011), where the players do not need to know the mixed strategy or the realized action of the opponent (in fact, not even their existence). This property sets us apart from previous works such as Zhou et al. (2017); O'Donoghue et al. (2021) that use the realized action of both players to gain insight about the matrix A.

From each player's perspective, they are essentially facing an MAB problem with time-varying loss vectors: $\ell_t = -Ay_t$ for the row player and $\ell'_t = A^\top x_t$ for the column player, with noisy feedback for the coordinate they choose. The standard performance measure in MAB is the (pseudo-)regret, defined for the row player and the column player respectively as

$$\begin{aligned} \mathrm{Reg}_T &= \max_{x \in \mathcal{P}_m} \mathrm{Reg}_T(x), \quad \text{where } \ \mathrm{Reg}_T(x) = \mathbf{E}\left[\sum_{t=1}^{T} (x - x_t)^\top A y_t\right], \\ \mathrm{Reg}'_T &= \max_{y \in \mathcal{P}_n} \mathrm{Reg}'_T(y), \quad \text{where } \ \mathrm{Reg}'_T(y) = \mathbf{E}\left[\sum_{t=1}^{T} x_t^\top A (y_t - y)\right]. \end{aligned} \tag{2}$$

We say that an algorithm achieves no-regret if Reg_T and Reg'_T grow sublinearly as $o(T)$, which has an important game-theoretic implication, since the duality gap of the average-iterate strategy $(\bar{x}_T, \bar{y}_T)$ where $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$ and $\bar{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t$ is equal to the average regret:

$$\mathrm{DGap}(\mathbf{E}[\bar{x}_T], \mathbf{E}[\bar{y}_T]) = \max_{x \in \mathcal{P}_m, y \in \mathcal{P}_n} \mathbf{E}\left[x^\top A\Big(\frac{1}{T}\sum_{t=1}^{T} y_t\Big) - \Big(\frac{1}{T}\sum_{t=1}^{T} x_t\Big)^\top A y\right] = \frac{1}{T}\left(\mathrm{Reg}_T + \mathrm{Reg}'_T\right).$$

Therefore, the average-iterate strategy converges to a Nash equilibrium, with the convergence rate governed by the average regret. By simply deploying standard adversarial MAB algorithms such as Exp3 (Auer et al., 2002b), one can obtain a convergence rate of $\tilde{O}(\sqrt{(m+n)/T})$, which is not improvable in the worst case even in this game setting (Klein and Young, 1999). The goal of this work is thus to improve the regret/convergence rate in an instance-dependent manner.

Tsallis-INF Algorithm Throughout the paper, we let both players apply the $\frac{1}{2}$-Tsallis-INF algorithm (Zimmert and Seldin, 2021), which is based on the Follow-the-Regularized-Leader (FTRL) framework and chooses its strategy by solving the following optimization problem:

$$x_t = \operatorname*{arg\,min}_{x \in \mathcal{P}_m} \left\{\sum_{s=1}^{t-1} \hat{\ell}_s^\top x + \frac{1}{\eta_t}\psi(x)\right\}, \quad y_t = \operatorname*{arg\,min}_{y \in \mathcal{P}_n} \left\{\sum_{s=1}^{t-1} \hat{\ell'}_s^\top y + \frac{1}{\eta_t}\psi(y)\right\}, \tag{3}$$

where $\eta_t = \frac{1}{2\sqrt{t}}$ is the learning rate, $\psi(x) = -2\sum_{i=1}^{m} \sqrt{x(i)}$ (or $-2\sum_{j=1}^{n} \sqrt{y(j)}$ for the column player, with a slight abuse of the notation) is the $\frac{1}{2}$-Tsallis entropy regularizer, and $\hat{\ell}_s$ and $\hat{\ell}'_s$ are

[1] Our results hold for the more general setting where the observations for the two players are two different samples with mean $A(i_t, j_t)$.

importance-weighted (IW) unbiased estimators for the loss vector ℓ_s and ℓ'_s respectively, defined via[2]

$$\hat{\ell}_t(i) = \frac{\mathbf{1}[i_t = i](1 - r_t)}{x_t(i)} - 1, \quad \hat{\ell}'_t(j) = \frac{\mathbf{1}[j_t = j](1 + r_t)}{y_t(j)} - 1. \tag{IW}$$

Tsallis-INF is an algorithm that achieves the optimal instance-dependent bound in stochastic MAB and simultaneously the optimal worst-case bound in adversarial MAB. Directly applying its guarantee for adversarial MAB shows that both players enjoy $\sqrt{T}$-type regret always, *even when their opponent behaves arbitrarily*. We summarize this in the following theorem and omit further mention in the rest of the paper. On the other hand, note that one cannot directly apply the guarantee of Tsallis-INF for stochastic MAB since the players are not facing a stochastic MAB instance with a fixed expected loss vector.[3] Instead, we will utilize an immediate regret bound, also summarized in the theorem below, along with the self-play nature and the zero-sum game structure to prove our results. For completeness, we provide the proof of this theorem in Appendix B.

Theorem 1 (Zimmert and Seldin, 2021) *For any $x \in \mathcal{P}_m$, the pseudo-regret of the Tsallis-INF algorithm against x is bounded as follows for the row player (and similarly for the column players):*

$$\mathrm{Reg}_T(x) \le \min_{i^* \in [m]} \left\{ \mathbf{E}\left[C_1 \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \sum_{i \in [m] \setminus \{i^*\}} \sqrt{x_t(i)} - C_2 \sqrt{T} \cdot D(x, x_{T+1}) \right] \right\}, \tag{4}$$

where C_1 and C_2 are positive universal constants and $D(x', x) = \sum_{i=1}^{m} \frac{1}{\sqrt{x(i)}} (\sqrt{x'(i)} - \sqrt{x(i)})^2$ is the Bregman divergence associated with the $\frac{1}{2}$-Tsallis entropy. In particular, we always have $\mathrm{Reg}_T = O(\sqrt{mT})$ even if the opponent behaves arbitrarily.

We will show in Appendix B that Theorem 1 holds with $C_1 = 19$ and $C_2 = 2$. It is worth noting that by using a more refined analysis similar to Zimmert and Seldin (2021), the values of C_1 and C_2 could be further improved. Additionally, replacing the IW estimator (IW) with their more sophisticated reduced variance estimator could further improve the values of C_1 and C_2. However, such precise analysis introduces extra terms like $O(m \log T)$, which unnecessarily complicates the upper bound. To avoid such unnecessary complexity and to handle noisy observations r_t, we provide an analysis that differs from theirs.

3. Instance-Dependent Regret Bounds for General Zero-Sum Games

We now provide and discuss our main theorem for general zero-sum games, which states two regret bounds both in the form of $c_1 \log T + \sqrt{c_2 T}$ for some game-dependent constants c_1 and c_2.

Theorem 2 *If both players follow the Tsallis-INF algorithm, then for any $x \in \mathcal{P}_m$ and $y \in \mathcal{P}_n$, the following two upper bounds simultaneously hold for the quantity:*

$$\max\Big\{ \mathrm{Reg}_T(x) + \sqrt{T} C_2 \mathbf{E}\big[D(x, x_{T+1})\big], \mathrm{Reg}'_T(y) + \sqrt{T} C_2 \mathbf{E}\big[D(y, y_{T+1})\big] \Big\} \tag{$\star$}$$

[2] Shifting the loss values uniformly does not affect the behavior of the algorithm, so the -1 in this equation can be removed in implementation. We add it here just to ensure that these indeed serve as unbiased estimators of ℓ_t, ℓ'_t.

[3] In fact, Zimmert and Seldin (2021) provide instance-dependent regret in a setting more general than the standard stochastic setting, but still, that does not directly apply to the game setting, especially when a PSNE does not exist.

- $$(\star) = O\Big(\sqrt{T(|I|+|J|-2)} + \omega \log_+ \frac{mT}{\omega^2} + \omega' \log_+ \frac{nT}{\omega'^2}\Big),$$

  where I and J are the support of $(x_\star, y_\star)$, an NE with maximum support, and $\omega = \sum_{i \notin I} \frac{1}{\Delta(i)}$, $\omega' = \sum_{j \notin J} \frac{1}{\Delta'(j)}$ with $\Delta = \left(x_\star^\top A y_\star\right) \mathbf{1} - A y_\star$ and $\Delta' = A^\top x_\star - \left(x_\star^\top A y_\star\right) \mathbf{1}$;

- $$(\star) = O\Big(\sqrt{T}\Big(\gamma \sqrt{\log_+ \frac{m}{\gamma^2}} + \gamma' \sqrt{\log_+ \frac{n}{\gamma'^2}}\Big) + \frac{m+n}{c} \log T\Big),$$

  where $\gamma = \max_{x_\star \in \mathcal{X}_\star} \sum_{i \in [m]} \sqrt{x_\star(i)} - 1$, $\gamma' = \max_{y_\star \in \mathcal{Y}_\star} \sum_{j \in [n]} \sqrt{y_\star(j)} - 1$, and $c > 0$ is a game-dependent constant such that $\max_{x \in \Delta_n} \geq c \min_{x_\star \in \mathcal{X}_\star} \|x - x_\star\|_1 + c \min_{y_\star \in \mathcal{Y}_\star} \|y - y_\star\|_1$ holds for all $(x, y) \in \mathcal{P}_m \times \mathcal{P}_n$ (which always exists).

While the key of the proof also relies on the self-bounding technique that is common in the analysis of Tsallis-INF, some new ideas are required for the game setting. The core technical innovation involves unifying the two types of instance-dependent parameters into a gap bound formulation that relates the regret and the players' probabilities on non-NE actions. This allows bounding the regret from below, and together with Theorem 1, omits a self-bounding inequality. Resolving this inequality yields the final bound; see details in Appendix D.

We note that Eq. ($\star$) is an upper bound on $\max\{\mathrm{Reg}_T(x), \mathrm{Reg}'_T(x)\}$ since Bregman divergence is non-negative, and we include the Bregman divergence terms in Eq. ($\star$) because they are crucial for proving the last-iterate convergence result in Section 4.2.

In both bounds of Theorem 2, the coefficients for $\sqrt{T}$ are smaller than the trivial bound $\max\{\sqrt{m}, \sqrt{n}\}$ and reflect the proximity of the NE to a pure strategy; specifically, $\sqrt{|I| + |J| - 2} = \gamma = \gamma' = 0$ when the game has a unique PSNE. This case will be further elaborated in Section 4. More generally, the coefficient $\sqrt{|I| + |J| - 2}$ in the first bound is small when the support of the NE is small, and this bound can be seen as a generalization of that in Balsubramani et al. (2016) for the special case of dueling bandits. On the other hand, the coefficients γ and γ' in the second bound are small when the NE are close to the boundary so that some actions have much larger weight than others. This kind of problem dependence resembles that of Maiti et al. (2023a) who study sample complexity of finding approximate NE in the special case of $2 \times n$ games. Indeed, their sample complexity to reach ε duality gap is at a high-level of order $1/\varepsilon^2$ multiplied with a qualitatively similar problem-dependent constant, which exactly corresponds to our $\sqrt{T}$ regret term.

The inverse coefficients for the $\log T$ term, Δ (Δ') and c, quantify the relative suboptimality of alternative actions compared to the NE. In particular, Δ and Δ' are exactly the standard suboptimality gaps for a stochastic MAB instance with loss vector $-Ay_\star$ and $A^\top x_\star$ respectively. Very roughly speaking, this $\log T$ term can then be interpreted as the overhead of finding the non-support of the NE, which is relatively small and is as if playing an MAB with the opponent fixed to a minimax or maximin strategy. On the other hand, the meaning of the inverse coefficient c is less clear, but its existence is guaranteed by Wei et al. (2021, Theorem 5), and we also refer the reader to their work for more details on this constant. It only approaches zero when a strategy sufficiently different from the NE has a disproportionately small duality gap. We demonstrate this with an example:

$$A = \begin{bmatrix} 0 & 3\varepsilon \\ 1-\varepsilon & 2\varepsilon \end{bmatrix}, \tag{5}$$

where $0 < \varepsilon < \frac{1}{3}$. This game has a unique NE $x_\star = (1 - 3\varepsilon, 3\varepsilon), y_\star = (\varepsilon, 1 - \varepsilon)$. Direct calculation shows that $c = \varepsilon$ satisfies the requirement for c. When ε approaches zero, $\gamma \approx \sqrt{\varepsilon}$ vanishes while

$\frac{1}{c} = \frac{1}{\varepsilon}$ explodes. In particular, when $\varepsilon \approx 1/T^{1/3}$, our regret bound is of order $T^{1/3}$, thus provably smaller than the worst-case $\sqrt{T}$ regret. We will revisit this example in numerical experiments in Section 5.

4. Games with Pure-Strategy Nash Equilibria

In this section, we further discuss the case with a unique PSNE denoted as $(i_\star, j_\star)$. Using the first bound in Theorem 2, we immediately obtain the following regret bound since $|I| = |J| = 1$.

Corollary 3 *For a game with a unique PSNE, if both players follow the Tsallis-INF algorithm, then the following regret bound holds:*

$$\max\{\mathrm{Reg}_T, \mathrm{Reg}'_T\} = O\Big(\omega \log_+ \frac{mT}{\omega^2} + \omega' \log_+ \frac{nT}{\omega'^2}\Big) = O\big((\omega + \omega') \log T\big), \tag{6}$$

where $\omega = \sum_{i \neq i_\star} \frac{1}{\Delta(i)}$, $\omega' = \sum_{j \neq j_\star} \frac{1}{\Delta'(j)}$, $\Delta(i) = A(i_\star, j_\star) - A(i, j_\star)$, *and* $\Delta'(j) = A(i_\star, j) - A(i_\star, j_\star)$.

This is a generalization of the standard instance-dependent regret bound for stochastic MAB and also similar to those from the dueling bandit literature (e.g., Yue et al., 2012; Zoghi et al., 2014; Saha and Gaillard, 2022). We next show that this bound is asymptotically optimal in Section 4.1. After that, we present two other results: the last-iterate convergence behavior of our algorithm (Section 4.2) and using our algorithm to identify the PSNE with high probability (Section 4.3).

4.1. Regret lower bound

In this section, we show that the regret bound in Corollary 3 is tight up to some constants. In fact, for any Δ and Δ', there exists a problem instance such that $\liminf_{T\to\infty} \frac{\mathrm{Reg}_T + \mathrm{Reg}'_T}{\log T} = \Omega(\omega + \omega')$ for any *consistent* algorithms (Lattimore and Szepesvári, 2020, Definition 16.1). Note that this lower bound is also applicable to *coupled* algorithms, i.e., this applies to situations where a single algorithm determines both i_t and j_t based on the observation of $\{r_s\}_{s<t}$.

We consider problem instances in which r_t follows a Bernoulli distribution over $\{-1, 1\}$, i.e., $r_t \sim \mathrm{Ber}^\pm(A(i_t, j_t))$ given (i_t, j_t), where $\mathrm{Ber}^\pm(a)$ for a parameter $a \in [-1, 1]$ is a distribution that takes values 1 and -1 with probability $(1 + a)/2$ and $(1 - a)/2$, respectively. Fix an algorithm for choosing (i_t, j_t) given the observations of $\{r_s\}_{s<t}$.

Let $\Delta \in [0, 1/4]^m$ and $\Delta' \in [0, 1/4]^n$ be such that $\Delta_{i_\star} = 0$ and $\Delta'_{j_\star} = 0$ for some $i_\star \in [m]$ and $j_\star \in [n]$. Suppose A is given by

$$A = \mathbf{1}_m {\Delta'}^\top - \Delta \mathbf{1}_n^\top. \tag{7}$$

Then, $(i_\star, j_\star)$ is a Nash equilibrium of the game with payoff matrix A as we have $A(i_\star, j_\star) - A(i, j_\star) = \Delta(i) \geq 0$ and $A(i_\star, j) - A(i_\star, j_\star) = \Delta(j) \geq 0$ for all $i \in [m]$ and $j \in [n]$. Let $N_{T,i}(A)$ and $N'_{T,j}(A)$ denote the expected numbers of times the i-th row and j-th column are chosen:

$$N_{T,i}(A) = \mathbf{E}\Bigg[\sum_{t=1}^T \mathbf{1}[i_t = i]\Bigg] = \mathbf{E}\Bigg[\sum_{t=1}^T x_t(i)\Bigg], \quad N'_{T,j}(A) = \mathbf{E}\Bigg[\sum_{t=1}^T \mathbf{1}[j_t = j]\Bigg] = \mathbf{E}\Bigg[\sum_{t=1}^T y_t(j)\Bigg].$$

We then have the following lower bound:

Theorem 4 *Suppose that there exist a function $g(m,n) > 0$ and a constant $c \in (0,1)$ such that $\mathrm{Reg}_T + \mathrm{Reg}'_T \leq g(m,n)T^{1-c}$ for any $\hat{A} \in [-1,1]^{m \times n}$. Then, if A is given by (7), we have*

$$N_{T,i}(A) = \Omega\Big(\frac{1}{(\Delta(i))^2}\log\Big(\frac{\Delta(i)T^c}{4g(m,n)}\Big)\Big), \quad N'_{T,j}(A) = \Omega\Big(\frac{1}{(\Delta'(j))^2}\log\Big(\frac{\Delta(j)T^c}{4g(m,n)}\Big)\Big)$$

for any $i \in [m]$ and $j \in [n]$ such that $\Delta_i \neq 0$ and $\Delta'_j \neq 0$. Consequently, we have

$$\liminf_{T\to\infty} \frac{\mathrm{Reg}_T + \mathrm{Reg}'_T}{\log T} = \Omega\Big(\sum_{\substack{i\in[m] \\ \Delta(i)>0}} \frac{c}{\Delta(i)} + \sum_{\substack{j\in[n] \\ \Delta'(j)>0}} \frac{c}{\Delta'(j)}\Big) = \Omega\big(c \cdot (\omega + \omega')\big).$$

Remark 5 For the special case in which A is skew-symmetric and $\Delta = \Delta'$, the regret lower bound for the dueling bandit problem (Komiyama et al., 2015, Theorem 2) leads to the same asymptotic lower bound as Theorem 4 above. Our Theorem 4 is more general in that it relaxes this symmetry condition; however, the underlying idea used in their proofs are shared. That said, while Komiyama et al. (2015) follow the proof structure of Lai and Robbins (1985, Theorem 1), our proof, provided in Appendix E, adopts a simplified analytical approach based on the Bretagnolle-Huber inequality (see, e.g., Lattimore and Szepesvári, 2020, Chapter 17).

Remark 6 Under the assumption that bandit algorithms are *minimax optimal*, i.e., if there exists a universal constant C such that $\mathrm{Reg}_T + \mathrm{Reg}'_T \leq C\sqrt{(m+n)T}$ holds for all $\hat{A} \in [-1,1]^{m\times n}$, which corresponds to $g(m,n) = C\sqrt{m+n}$ and $c = 1/2$, we obtain the following finite-time lower bound:

$$R_T(A) = \Omega\Big(\sum_{\substack{i\in[m] \\ \Delta(i)>0}} \frac{1}{\Delta(i)} \log \frac{(\Delta(i))^2 T}{16C^2(m+n)} + \sum_{\substack{j\in[n] \\ \Delta'(j)>0}} \frac{1}{\Delta'(j)} \log \frac{(\Delta'(j))^2 T}{16C^2(m+n)}\Big).$$

This matches upper bound in (6) up to a constant factor, under the conditions that $n = \Theta(m)$ and that the values of non-zero $\Delta(i)$'s and $\Delta'(j)$'s are equivalent up to a constant factor.

4.2. Last-iterate convergence

Somewhat surprisingly, we show that Tsallis-INF also ensures the following last-iterate convergence guarantee.

Proposition 7 *For a game with a unique PSNE, if both players use the Tsallis-INF algorithm, the output distributions converge to the PSNE (in expectation) as follows: for any t,*

$$\mathbf{E}[D(x_\star, x_t) + D(y_\star, y_t)] = O\Big(\frac{1}{\sqrt{t}}\big(\omega \log_+ \frac{mt}{\omega^2} + \omega' \log_+ \frac{nt}{\omega'^2}\big)\Big), \tag{8}$$

where $D(\cdot,\cdot)$ represents the Bregman divergence associated with the $(1/2)$-Tsallis entropy. Consequently, we have

$$\mathbf{E}\Big[\sqrt{\mathrm{DGap}(x_t, y_t)}\Big] = O\Big(\frac{1}{\sqrt{t}}\big(\omega \log_+ \frac{mt}{\omega^2} + \omega' \log_+ \frac{nt}{\omega'^2}\big)\Big). \tag{9}$$

Even though $\mathbf{E}[\sqrt{\mathrm{DGap}(x_t, y_t)}] = O(1/\sqrt{t})$ (ignoring other factors) only imply $\mathbf{E}[\mathrm{DGap}(x_t, y_t)] = O(1/\sqrt{t})$ but not necessarily $\mathbf{E}[\mathrm{DGap}(x_t, y_t)] = O(1/t)$ (so the last-iterate convergence might be slower than the average-iterate convergence), this rate is already much better than the generic $O(1/t^{1/6})$ rate of Cai et al. (2023) for general zero-sum games. Our proof is also particularly simple and is in fact a simple corollary of the regret bound of Theorem 2.

Proof Fix arbitrary $T \in \mathbb{N}$. From the first bound of Theorem 2, we have

$$\mathrm{Reg}_T(x_\star) + \mathrm{Reg}'_T(y_\star) + C_2\sqrt{T}\,\mathbf{E}\left[D(x_\star, x_{T+1}) + D(y_\star, y_{T+1})\right] = O\Big(\omega \log_+ \frac{mT}{\omega^2} + \omega' \log_+ \frac{nT}{\omega'^2}\Big).$$

Since $(x_\star, y_\star)$ is a Nash equilibrium, we know that $\mathrm{Reg}_T(x_\star) + \mathrm{Reg}'_T(y_\star) \geq 0$. This implies

$$\mathbf{E}\left[D(x_\star, x_{T+1}) + D(y_\star, y_{T+1})\right] = O\Big(\frac{1}{\sqrt{T}}\big(\omega \log_+ \frac{mT}{\omega^2} + \omega' \log_+ \frac{nT}{\omega'^2}\big)\Big),$$

which completes the proof of (8). The Bregmann divergence associated with $(1/2)$-Tsallis entropy is bounded as

$$D(x_\star, x_t) = \sum_{i=1}^{m} \frac{1}{\sqrt{x_t(i)}}\Big(\sqrt{x_\star(i)} - \sqrt{x_t(i)}\Big)^2 \geq \sum_{i\in[m]\setminus\{i_\star\}} \sqrt{x_t(i)} \geq \frac{1}{2}\sqrt{\|x_t - x_\star\|_1}.$$

As DGap is a 1-Lipschitz function w.r.t. the L^1 norm (Lemma 15 in Appndix C), we have

$$\begin{aligned}\mathbf{E}\Big[\sqrt{\mathrm{DGap}(x_t, y_t)}\Big] &\leq \mathbf{E}\Big[\sqrt{\mathrm{DGap}(x_\star, y_\star) + \|x_t - x_\star\|_1 + \|y_t - y_\star\|_1}\Big] \\ &\leq 2\,\mathbf{E}[D(x_\star, x_t) + D(y_\star, y_t)].\end{aligned}$$

From this and (8), we obtain (9) as desired. ■

4.3. Sample complexity of identifying PSNE

While the main focus of our work is regret minimization, we show that our algorithm can also find the exact PSNE with high probability, again using its regret guarantee. Specifically, define $\Delta_{\min} = \min\big\{\min_{i\in[m]\setminus\{i_\star\}} \Delta(i), \min_{j\in[n]\setminus\{j_\star\}} \Delta'(j)\big\}$. We prove the following.

Theorem 8 *For output sequences $\{i_t\}_{t=1}^T$ and $\{j_t\}_{t=1}^T$ generated by the Tsallis-INF algorithm, let $\hat{i}_T$ and $\hat{j}_T$ be the most frequently chosen arms in these sequences, i.e., $\hat{i}_T \in \arg\max_{i\in[m]}\big|\{t \in [T] \mid i_t = i\}\big|$ and $\hat{j}_T \in \arg\max_{j\in[n]}\big|\{t \in [T] \mid j_t = j\}\big|$. Then, there exists a universal constant $\alpha > 0$ such that $(\hat{i}_T, \hat{j}_T) = (i_\star, j_\star)$ holds with probability at least $3/4$ for $T \geq \alpha\frac{\omega+\omega'}{\Delta_{\min}}$.*

Proof From the definition of $\hat{i}_T$ and Markov's inequality, we have

$$\begin{aligned}\Pr\big[\hat{i}_T \neq i_\star\big] &\leq \Pr\Big[\sum_{t=1}^{T}\mathbf{1}[i_t = i_\star] \leq \frac{T}{2}\Big] = \Pr\Big[T - \sum_{t=1}^{T}\mathbf{1}[i_t = i_\star] \geq \frac{T}{2}\Big] \\ &\leq \frac{2}{T}\,\mathbf{E}\Big[T - \sum_{t=1}^{T}\mathbf{1}[i_t = i_\star]\Big] = 2 - \frac{2}{T}\,\mathbf{E}\Big[\sum_{t=1}^{T} x_t(i_\star)\Big] = 2(1 - \bar{x}_T(i_\star)).\end{aligned} \qquad (10)$$

As we have $\bar{x}_T \cdot \Delta \geq \sum_{i \neq i_\star} \bar{x}_t(i)\Delta(i) \geq \sum_{i \neq i_\star} \bar{x}_T(i)\Delta_{\min} = \Delta_{\min}(1 - \bar{x}_T(i_\star))$, by combining this with (10), we obtain $\Pr\left[\hat{i}_T \neq i_\star\right] \leq 2(1 - \bar{x}_T(i_\star)) \leq \frac{2}{\Delta_{\min}} \bar{x}_T \cdot \Delta$. As a similar bound holds for $\Pr\left[\hat{j}_T \neq j_\star\right]$, we have

$$\Pr\left[(\hat{i}_T, \hat{j}_T) \neq (i_\star, j_\star)\right] \leq \Pr\left[\hat{i}_T \neq i_\star\right] + \Pr\left[\hat{j}_T \neq j_\star\right] \leq \frac{2}{\Delta_{\min}}\left(\bar{x}_T \cdot \Delta + \bar{y}_T \cdot \Delta'\right) \tag{11}$$

From the definition of Δ and Theorem 2, we have

$$\begin{aligned}
\frac{2}{\Delta_{\min}}\left(\bar{x}_T \cdot \Delta + \bar{y}_T \cdot \Delta'\right) &\leq \frac{2}{\Delta_{\min}} \frac{\mathrm{Reg}_T + \mathrm{Reg}'_T}{T} \\
&\leq 2C_3\left(\frac{\omega}{\Delta_{\min}T} \log_+ \frac{mT}{\omega^2} + \frac{\omega'}{\Delta_{\min}T} \log_+ \frac{nT}{\omega'^2}\right) \\
&\leq 2C_3\left(\frac{\omega}{\Delta_{\min}T} \log_+ \frac{\Delta_{\min}T}{\omega} + \frac{\omega'}{\Delta_{\min}T} \log_+ \frac{\Delta_{\min}T}{\omega'}\right), \\
&\leq 2C_3\left(\frac{1}{\alpha} \log_+ \alpha + \frac{1}{\alpha} \log_+ \alpha\right) \leq \frac{1}{4},
\end{aligned}$$

where C_3 is the contant factor hidden by the $O(\cdot)$ symbol in (6), and in the third inequality we used the fact that $\omega \geq m\Delta_{\min}$ and $\omega' \geq n\Delta_{\min}$. The last inequality holds if we take $\alpha = 8C_3 + 4$. By combining this with (11), we obtain $\Pr\left[(\hat{i}_T, \hat{j}_T) \neq (i_\star, j_\star)\right] \leq 1/4$, which completes the proof. ■

From this, we can further boost the confidence and identify the PSNE with probability at least $(1 - \delta)$ with $O\left(\frac{\omega + \omega'}{\Delta_{\min}} \log(1/\delta)\right)$ samples. More concretely, consider repeating $S > 1$ independent trials of calculating $\hat{i}_T$. Let $\hat{i}_{T,s}$ be the result for the s-th trial. Let $\tilde{i}_{T,S} \in \arg\max_{i \in [m]} \left|\{s \in [S] \mid \hat{i}_{T,s} = i\}\right|$ be the arm most frequently chosen in these S trials. We then have

$$\begin{aligned}
\Pr\left[\tilde{i}_{T,S} \neq i_\star\right] &\leq \Pr\left[\sum_{s=1}^{S} \mathbf{1}\left[\hat{i}_{T,s} = i_\star\right] \leq S/2\right] \\
&\leq \Pr\left[\sum_{s=1}^{S} X_s \leq S/2 \mid X_s \sim \mathrm{Ber}(3/4),\ \text{i.i.d. for } s \in [S]\right] \leq \exp(\Omega(-S)).
\end{aligned}$$

Hence, for any $\delta \in (0, 1)$, by setting $S = \Theta(1/\delta)$, we have $\Pr\left[(\tilde{i}_{T,S}, \tilde{j}_{T,S}) = (i_\star, j_\star)\right] \geq 1 - \delta$. We note that, to perform this procedure, it is necessary to know an approximate value of $\frac{\omega + \omega'}{\Delta_{\min}}$.

We also note that due to Lemma 11, our sample complexity is at most $O\left(\sqrt{\max\{n, m\}}\right)$ times the information-theoretic optimal, which is $O\left(\sum_{i \in [m] \setminus \{i_\star\}} \frac{1}{\Delta_i^2} + \sum_{j \in [n] \setminus \{j_\star\}} \frac{1}{\Delta_j'^2}\right)$ and is achieved by the Midsearch algorithm of Maiti et al. (2024). However, their algorithm is coupled; that is, Midsearch requires the algorithm to control both players at the same time, while our algorithm is a no-regret uncoupled learning dynamic.

5. Numerical Experiments

To validate our theoretical results, we conduct a few numerical experiments.

The first experiment compares Tsallis-INF against two baselines in terms of the regret: the classical UCB1 (Auer et al., 2002a) and Exp3 (Auer et al., 2002b) algorithms, which are known

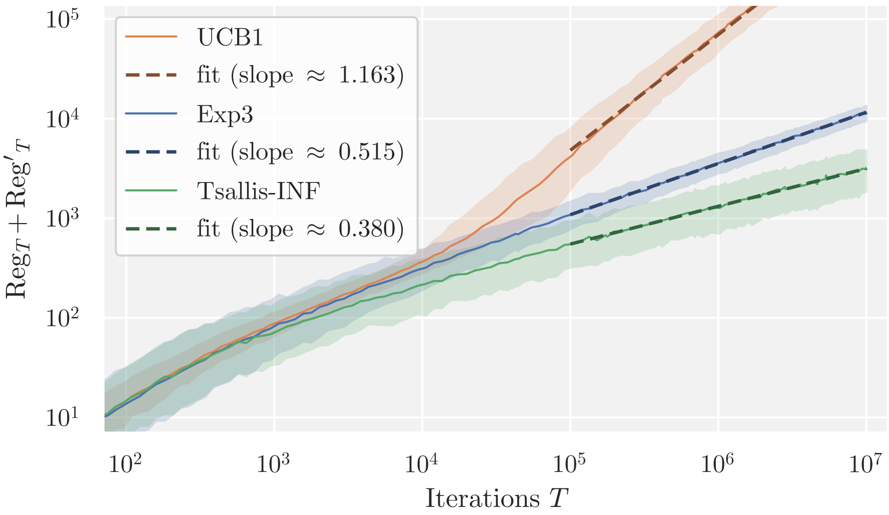

Figure 1: Regret scaling for Tsallis-INF and two other bandit algorithms. Each configuration (T) is run for 512 trials. The interval between the 10th and 90th percentile is overlaid. The thicker dashed line represents a linear fit on the $T \geq 10^5$ subset of the log-log data.

to have $O(T)$ and $\tilde{O}(\sqrt{T})$ regret bounds respectively in the adversarial setting. We compare them on the game associated with A defined in (5), with varying T and $\varepsilon = T^{-1/3}$, where feedback r_t follows a Bernoulli distribution over $\{-1, 1\}$ such that $\mathbf{E}[r_t \mid i_t, j_t] = A(i_t, j_t)$. As discussed, Theorem 2 predicts a regret of $\tilde{O}(T^{1/3})$ for Tsallis-INF. The result of the experiment agrees with all these bounds in Figure 1, where the asymptotic slope in the log-log plot (shown with a linear fit on the $T \geq 10^5$ region) is close to the theoretical prediction.

We have discussed in Section 4.3 that Tsallis-INF needs $\frac{\omega+\omega'}{\Delta_{\min}}$ iterations to identify the PSNE of a game. To validate our theoretical bounds, we conduct our second experiment using the following hard instance introduced by Maiti et al. (2024):

$$A = \begin{bmatrix} 0 & 2\Delta_{\min} & 2\Delta_1 & \cdots & 2\Delta_1 \\ -2\Delta_{\min} & 0 & 1 & \cdots & 1 \\ -2\Delta_1 & -1 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 1 \\ -2\Delta_1 & -1 & \cdots & -1 & 0 \end{bmatrix}, \tag{12}$$

where the top-left entry is the PSNE. We set the number of actions $n = m = 256$ and the gap $\Delta_1 = 0.1$, and vary the value of $\Delta_{\min}$. Let OPT represent the theoretical optimal bound for identifying PSNE (ignoring log terms), defined as $\mathrm{OPT} = \sum_{i \in [m] \setminus \{i_\star\}} \frac{1}{\Delta_i^2} + \sum_{j \in [n] \setminus \{j_\star\}} \frac{1}{\Delta'^2_j}$, which simplifies to $\frac{1}{2\Delta_{\min}^2} + \frac{m-2}{2\Delta_1^2}$ in this experiment. Maiti et al.'s (2024) achieve the optimal $\tilde{O}(\mathrm{OPT})$ sample complexity, and their Figure 2 suggests that the sample complexity of Tsallis-INF divided by OPT is unbounded as $\Delta_{\min}$ decreases, but our analysis in Section 4.3 disagrees with this trend. As shown in Figure 2, the number of iterations needed to identify the PSNE divided by

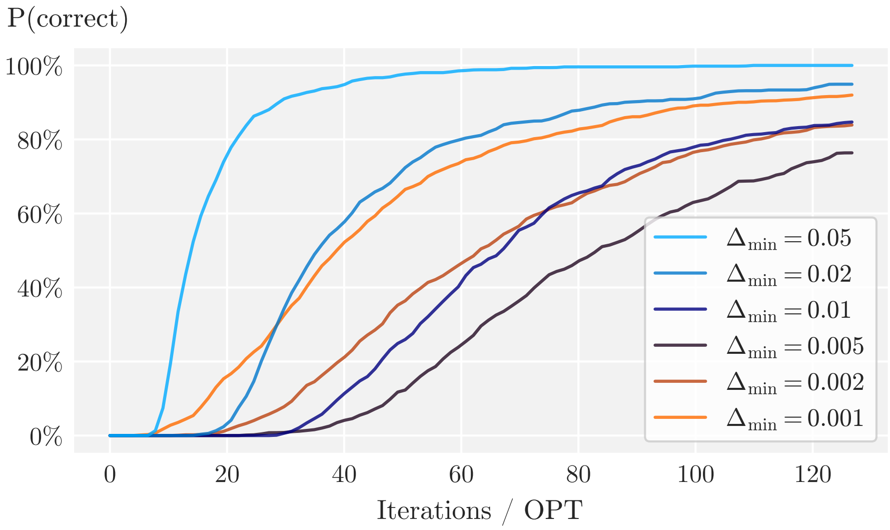

Figure 2: Experimental validation of Tsallis-INF's PSNE identification capability. The plot shows the algorithm's success rate in correctly identifying PSNE against the number of itrations. We use a hard instance of a 256×256 matrix and $\Delta_1 = 0.1$, running 512 trials for each $\Delta_{\min}$ values over a horizon of $128\mathrm{OPT}$ iterations, where OPT is the theoretical lower bound for PSNE identification. The x-axis is scaled by $1/\mathrm{OPT}$.

OPT decreases and then increases as $\Delta_{\min}$ varies. Lemma 11 predicts the minimum ratio occurs when $\frac{\Delta_{\min}}{\Delta_1} = \frac{1}{\sqrt{m}+1} = 1/17$, and among the values we tested, the minimum is reached when $\frac{\Delta_{\min}}{\Delta_1} = 0.005/0.1 = 1/20$, closely matching the prediction. This supports our derived bound of $\tilde{O}\big(\sqrt{m} \cdot \mathrm{OPT}\big)$.

The code for reproducing the experiments is available on https://github.com/EtaoinWu/instance-dependent-game-learning.

6. Conclusions

Prior work on learning in games has primarily focused on the full-information setting, where each player perfectly learns their gradient. In this paper, we investigate the more realistic, partial information setting where only a noisy version of a single realized reward is revealed to the players. Although it is impossible to optimize for all inputs, we demonstrated that Tsallis-INF, an existing best-of-both-worlds optimal bandit algorithm, enjoys improvements by exploiting easier instances with larger gaps. These improvements cover three aspects: regret minimization that bounds long-term average performance, last-iterate convergence guarantees that ensure day-to-day behavior for myopic agents, and a simple way to identify PSNE.

Several important questions remain open. A natural step forward is generalizing our work to general-sum multiplayer games, which is not yet explored due to the added intricacy from the misaligned incentives of players. Equally important is extending our results to extensive-form games and continuous games, each of which introducing extra challenges with their structural complexity. Our algorithm leaves a $O(\sqrt{\max\{m,n\}})$ gap in sample complexity for pure strategy Nash equilibrium identification, and whether this gap is unavoidable in uncoupled learning dynamics remains unknown. More broadly, our work suggests that learning to play games under uncertainty may be more achievable with instance-dependent improvements, opening up new possibilities in other game-theoretic environments.

Acknowledgments

SI was supported by JSPS KAKENHI Grant Number JP25K03184. HL was supported by NSF award IIS-1943607. TT was supported by JSPS KAKENHI Grant Number JP24K23852.

References

Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning*, pages 856–864. PMLR, 2014.

Idan Amir, Guy Azov, Tomer Koren, and Roi Livni. Better best of both worlds bounds for bandits with switching costs. In *Advances in Neural Information Processing Systems*, volume 35, pages 15800–15810. Curran Associates, Inc., 2022.

Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 2022a.

Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Tuomas Sandholm. Uncoupled learning dynamics with $O(\log T)$ swap regret in multiplayer games. In *Advances in Neural Information Processing Systems*, volume 35, pages 3292–3304. Curran Associates, Inc., 2022b.

P Auer, Paul Fischer, and N Cesa-Bianchi. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(3):235–256, 2002a.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

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

Akshay Balsubramani, Zohar Karnin, Robert E. Schapire, and Masrour Zoghi. Instance-dependent regret bounds for dueling bandits. In *29th Annual Conference on Learning Theory*, volume 49, pages 336–360. PMLR, 2016.

Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold'em poker is solved. *Science*, 347(6218):145–149, 2015.

Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, 365(6456): 885–890, 2019.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *Advances in Neural Information Processing Systems*, volume 35, pages 33904–33919. Curran Associates, Inc., 2022.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Uncoupled and convergent learning in two-player zero-sum markov games with bandit feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 36364–36406. Curran Associates, Inc., 2023.

Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng. Fast last-iterate convergence of learning in games requires forgetful algorithms. In *Advances in Neural Information Processing Systems*, volume 37, pages 23406–23434. Curran Associates, Inc., 2024a.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Accelerated algorithms for constrained nonconvex-nonconcave min-max optimization and comonotone inclusion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 5312–5347. PMLR, 2024b.

Clément L Canonne. A short note on an inequality between KL and TV. *arXiv preprint arXiv:2202.07198*, 2022.

Xi Chen and Binghui Peng. Hedging in games: Faster convergence of external and swap regrets. In *Advances in Neural Information Processing Systems*, volume 33, pages 18990–18999. Curran Associates, Inc., 2020.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *10th Innovations in Theoretical Computer Science Conference (ITCS)*, 2019.

Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 235–254. SIAM, 2011.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, pages 563–587. PMLR, 2015.

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

Meta Fundamental AI Research Diplomacy Team FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378 (6624):1067–1074, 2022.

Gabriele Farina, Ioannis Anagnostides, Haipeng Luo, Chung-Wei Lee, Christian Kroer, and Tuomas Sandholm. Near-optimal no-regret learning dynamics for general convex games. In *Advances in Neural Information Processing Systems*, volume 35, pages 39076–39089. Curran Associates, Inc., 2022.

Dean P Foster and Rakesh V Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1-2):40–55, 1997.

Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*, JMLR Workshop and Conference Proceedings, pages 359–376, 2011.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *Advances in Neural Information Processing Systems*, volume 33, pages 20766–20778. Curran Associates, Inc., 2020.

Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. In *Advances in Neural Information Processing Systems*, volume 35, pages 21858–21870. Curran Associates, Inc., 2022.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to Nash equilibrium. In *Conference on Learning Theory*, pages 2388–2422. PMLR, 2021.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Conference on Learning Theory*, pages 2552–2583. PMLR, 2021.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. *Advances in Neural Information Processing Systems*, 35: 28631–28643, 2022.

Athul Paul Jacob, Yikang Shen, Gabriele Farina, and Jacob Andreas. The consensus game: Language model generation via equilibrium search. In *The Twelfth International Conference on Learning Representations*, 2024.

Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th annual conference on information sciences and systems (CISS)*, pages 1–6. IEEE, 2014.

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

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic MDPs with unknown transition. In *Advances in Neural Information Processing Systems*, volume 34, pages 20491–20502, 2021.

Tiancheng Jin, Junyan Liu, and Haipeng Luo. 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, pages 30918–30978, 2023a.

Tiancheng Jin, Junyan Liu, Chloé Rouyer, William Chang, Chen-Yu Wei, and Haipeng Luo. No-regret online reinforcement learning with adversarial losses and transitions. In *Advances in Neural Information Processing Systems*, volume 36, pages 38520–38585. Curran Associates, Inc., 2023b.

Philip Klein and Neal Young. On the number of iterations for Dantzig-Wolfe optimization and packing-covering approximation algorithms. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 320–327. Springer, 1999.

Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on learning theory*, pages 1141–1154. PMLR, 2015.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, NY, 2020. ISBN 978-1-108-57140-1.

Arnab Maiti, Kevin Jamieson, and Lillian Ratliff. Instance-dependent sample complexity bounds for zero-sum matrix games. In *International Conference on Artificial Intelligence and Statistics*, pages 9429–9469. PMLR, 2023a.

Arnab Maiti, Kevin Jamieson, and Lillian J Ratliff. Logarithmic regret for matrix games against an adversary with noisy bandit feedback. *arXiv preprint arXiv:2306.13233*, 2023b.

Arnab Maiti, Ross Boczar, Kevin Jamieson, and Lillian Ratliff. Near-optimal pure exploration in matrix games: A generalization of stochastic bandits & dueling bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2602–2610. PMLR, 2024.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 11752–11762, 2022.

Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173:465–507, 2019.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In *Forty-first International Conference on Machine Learning*, 2024.

Brendan O'Donoghue, Tor Lattimore, and Ian Osband. Matrix games with bandit feedback. In *Uncertainty in Artificial Intelligence*, pages 279–289. PMLR, 2021.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, volume 26, pages 3066–3074. Curran Associates, Inc., 2013.

Chloé Rouyer and Yevgeny Seldin. Tsallis-INF for decoupled exploration and exploitation in multi-armed bandits. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 3227–3249, 2020.

Chloé Rouyer, Yevgeny Seldin, and Nicolo Cesa-Bianchi. An algorithm for stochastic and adversarial bandits with switching costs. In *International Conference on Machine Learning*, pages 9127–9135. PMLR, 2021.

El Mehdi Saad, Alexandra Carpentier, Tomáš Kocák, and Nicolas Verzelen. On weak regret analysis for dueling bandits. In *Advances in Neural Information Processing Systems*, 2024.

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

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, volume 28, pages 2989–2997. Curran Associates, Inc., 2015.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Best-of-both-worlds algorithms for partial monitoring. In *International Conference on Algorithmic Learning Theory*, pages 1484–1515. PMLR, 2023.

John von Neumann. Zur Theorie der Gesellschaftsspiele [On the theory of games of strategy]. *Mathematische Annalen*, 100(1):295–320, December 1928. Original document in German.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR, 2018.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *International Conference on Learning Representations*, 2021.

Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 241–248. Citeseer, 2011.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Yichi Zhou, Jialian Li, and Jun Zhu. Identify the Nash equilibrium in static games with random payoffs. In *Proceedings of the 34th International Conference on Machine Learning*, pages 4160–4169. PMLR, 2017.

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

Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten Rijke. Relative upper confidence bound for the K-armed dueling bandit problem. In *International conference on machine learning*, pages 10–18. PMLR, 2014.

Appendix A. Technical Lemmas

Lemma 9 *For any real numbers a, b, c, if $a > 0, c > 0$, then $a \leq b + \sqrt{ac}$ implies $a \leq 2b + c$.*

Proof The function $g(t) = t - \sqrt{t} - \frac{1}{2}(t-1)$ defined on $\mathbb{R}_{\geq 0}$ is convex and has a minimum of 0 at $t = 1$. This implies that

$$a - \sqrt{ac} - \frac{1}{2}(a-c) = c \cdot g\Big(\frac{a}{c}\Big) \geq 0.$$

Therefore, when $b \geq a - \sqrt{ac}$, we also have $2b \geq a - c$. ■

Lemma 10 *Let $a, b > 0$ and suppose that $0 \leq z_t \leq b$ holds for all $t = 1, 2, \dots, T$. Also, assume that $\sum_{t=1}^{T} z_t^2 \leq a$. We then have*

$$\sum_{t=1}^{T} \frac{z_t}{\sqrt{t}} \leq f(a,b), \quad \text{where} \quad f(a,b) := \min_{s \in [T]} \Big\{ \sqrt{a \log \frac{T}{s}} + 2b\sqrt{s} \Big\}. \tag{13}$$

Note that f is a concave function.

Proof We split the sum into the first n terms and the last $T - s$ terms:

$$\sum_{t=1}^{T} \frac{z_t}{\sqrt{t}} = \sum_{t=1}^{s} \frac{z_t}{\sqrt{t}} + \sum_{t=s+1}^{T} \frac{z_t}{\sqrt{t}}. \tag{14}$$

The first part can be bounded with the fact that $z_t \leq b$:

$$\begin{aligned} \sum_{t=1}^{s} \frac{z_t}{\sqrt{t}} &\leq \sum_{t=1}^{s} \frac{b}{\sqrt{t}} \\ &= b \sum_{t=1}^{s} \frac{1}{\sqrt{t}} \\ &\leq b \cdot 2\sqrt{s}. \end{aligned} \tag{15}$$

The other part can be bounded with the Cauchy-Schwarz inequality:

$$\begin{aligned} \sum_{t=s+1}^{T} \frac{z_t}{\sqrt{t}} &= \sum_{t=s+1}^{T} z_t \frac{1}{\sqrt{t}} \\ &\leq \sqrt{\Big(\sum_{t=s+1}^{T} z_t^2\Big)\Big(\sum_{t=s+1}^{T} \frac{1}{t}\Big)} \\ &\leq \sqrt{a \log \frac{T}{s}}. \end{aligned} \tag{16}$$

Note that our $\log(\cdot)$ is of base 2. Our desired inequality can be obtained by plugging (15) and (16) into (14). ■

Lemma 11 *For $n \geq 2$ and an arbitrary sequence of nonnegative real numbers $x_1, \dots, x_n$, the following inequality holds:*

$$\max\{x_1, \dots, x_n\} \sum_{i=1}^{n} x_i \leq \Big(\frac{1}{2} + \frac{1}{2}\sqrt{n}\Big) \sum_{i=1}^{n} x_i^2.$$

Equality holds when all but one x_i values are equal to $\frac{1}{\sqrt{n}+1}$ times the single outlier.

Proof Without loss of generality we assume that $x_1 \geq x_2 \geq \cdots \geq x_n$. Let $\bar{x} = \frac{1}{n-1} \sum_{i=2}^{n} x_i$ be the average of the last $n-1$ numbers. Due to the strict convexity of the square function, we see from Jensen's inequality that

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + \sum_{i=2}^{n} x_i^2 \geq x_1^2 + (n-1)\bar{x}^2, \tag{17}$$

with equality when all $x_2, \dots, x_n$ are equal to $\bar{x}$. Consider the function $f(k) = \frac{1+(n-1)k}{1+(n-1)k^2}$ defined on $k \in [0,1]$. Note that

$$f(k) = \frac{1+(n-1)k}{1+(n-1)k^2} = 1 + \frac{n-1}{\frac{1}{k} - 1 + \frac{n}{\frac{1}{k}-1} + 2} \leq 1 + \frac{n-1}{2\sqrt{n}+2} = \frac{1}{2}\big(\sqrt{n}+1\big) \tag{18}$$

where the inequality is due to AM-GM, and is tight when $k = \frac{1}{\sqrt{n}+1}$. If we plug in $k = \frac{\bar{x}}{x_1}$, we get

$$1 + (n-1)\frac{\bar{x}}{x_1} \leq \frac{1}{2}\big(\sqrt{n}+1\big)\Big(1 + (n-1)\big(\frac{\bar{x}}{x_1}\big)^2\Big).$$

We multiply both sides by x_1^2 and get

$$x_1(x_1 + (n-1)\bar{x}) \leq \frac{1}{2}\big(\sqrt{n}+1\big)\big(x_1^2 + (n-1)\bar{x}^2\big),$$

which means that

$$x_1 \sum_{i=1}^{n} x_i \leq \frac{1}{2}\big(\sqrt{n}+1\big)\big(x_1^2 + (n-1)\bar{x}^2\big).$$

Together with (17) this completes the proof. Equality holds when both (17) and (18) are tight, i.e., $x_2 = \cdots = x_n = \bar{x} = \frac{1}{\sqrt{n}+1} x_1$. ■

Appendix B. Proof of Theorem 1

This appendix provides the proof of Theorem 1. We include the proof of Theorem 1, as the corresponding proof is not provided in Zimmert and Seldin (2021), and several aspects differ from their setting: the range of the loss is different, noisy feedback r_t such that $\mathbf{E}[r_t \mid i_t, j_t] = A(i_t, j_t)$ is observed, and a negative term is introduced to ensure last-iterate convergence.

We begin by providing the following standard regret upper bound of FTRL. By refining an analysis of FTRL, we obtain a negative term of $-\frac{1}{\eta_{T+1}} D(x^*, x_{T+1})$, which is used to provide the last-iterate convergence result of Proposition 7.

Lemma 12 *Let $\mathcal{X} \in \mathbb{R}^n$ be a non-empty compact convex set. Let ψ be a continuously differentiable convex function over $\mathcal{X}$. Suppose that x_t is given by FTRL with the regularizer function ψ and learning rates $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_{T+1} > 0$, as follows:*

$$x_t \in \operatorname*{arg\,min}_{x \in \mathcal{X}} \left\{ \sum_{s=1}^{t-1} \ell_s^\top x + \frac{1}{\eta_t} \psi(x) \right\}. \tag{19}$$

Then, for any $x^* \in \mathcal{X}$, we have

$$\begin{aligned}\sum_{t=1}^{T} \ell_t^\top (x_t - x^*) \leq \sum_{t=1}^{T} \Big(\ell_t^\top (x_t - x_{t+1}) - \frac{1}{\eta_t} D(x_{t+1}, x_t) \Big) + \sum_{t=1}^{T} \Big(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \Big) (\psi(x^*) - \psi(x_{t+1})) \\ + \frac{1}{\eta_1} (\psi(x^*) - \psi(x_1)) - \frac{1}{\eta_{T+1}} D(x^*, x_{T+1}),\end{aligned} \tag{20}$$

where $D(\cdot, \cdot)$ is the Bregman divergence associated with ψ: $D(y, x) = \psi(y) - \psi(x) - \nabla \psi(x)^\top (y - x)$.

Proof We have

$$\begin{aligned}
&\sum_{t=1}^{T} \ell_t^\top x^* + \frac{1}{\eta_{T+1}} \psi(x^*) \\
&\geq \sum_{t=1}^{T} \ell_t^\top x_{T+1} + \frac{1}{\eta_{T+1}} \psi(x_{T+1}) + \frac{1}{\eta_{T+1}} D(x^*, x_{T+1}) \\
&= \sum_{t=1}^{T-1} \ell_t^\top x_{T+1} + \frac{1}{\eta_T} \psi(x_{T+1}) + \ell_T^\top x_{T+1} + \left(\frac{1}{\eta_{T+1}} - \frac{1}{\eta_T} \right) \psi(x_{T+1}) + \frac{1}{\eta_{T+1}} D(x^*, x_{T+1}) \\
&\geq \sum_{t=1}^{T-1} \ell_t^\top x_T + \frac{1}{\eta_T} \psi(x_T) + \frac{1}{\eta_T} D(x_{T+1}, x_T) + \ell_T^\top x_{T+1} \\
&\qquad + \Big(\frac{1}{\eta_{T+1}} - \frac{1}{\eta_T} \Big) \psi(x_{T+1}) + \frac{1}{\eta_{T+1}} D(x^*, x_{T+1}) \\
&\geq \cdots \\
&\geq \frac{1}{\eta_1} \psi(x_1) + \sum_{t=1}^{T} \left(\frac{1}{\eta_t} D(x_{t+1}, x_t) + \ell_t^\top x_{t+1} + \Big(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \Big) \psi(x_{t+1}) \right) + \frac{1}{\eta_{T+1}} D(x^*, x_{T+1}),
\end{aligned}$$

where each inequality follows from the definition of FTRL (19) and the first-order optimality condition. This immediately leads to the desired inequality. ■

We next provide lemmas to the upper bound the first summation in the RHS of (20) for the $1/2$-Tsallis entropy.

Lemma 13 *Let $\phi(x) = -2\sqrt{x}$ and $D_\phi(y,x) = -2\sqrt{y} + 2\sqrt{x} + (y-x)/\sqrt{x} = (\sqrt{y}-\sqrt{x})^2/\sqrt{x}$ be the Bregman divergence associated with ϕ. Then, for any $x \in (0,1)$ and $a > -1/\sqrt{x}$,*

$$\max_{y \in (0,\infty)} \{a(x-y) - D_\phi(y,x)\} \leq \sqrt{x}\,\xi(a\sqrt{x})$$

for $\xi(z) = z^2/(1+z)$ for $z \geq 0$. If $a \geq -1/(2\sqrt{x})$, then it also holds that

$$\max_{y \in (0,\infty)} \{a(x-y) - D_\phi(y,x)\} \leq 2x^{3/2}a^2.$$

Proof We have

$$\begin{aligned}
a(x-y) - D_\phi(y,x) &= a(x-y) - \frac{(\sqrt{y}-\sqrt{x})^2}{\sqrt{x}} \\
&= (\sqrt{x}-\sqrt{y})\left\{a(\sqrt{x}+\sqrt{y}) - \frac{\sqrt{x}-\sqrt{y}}{\sqrt{x}}\right\} \\
&= (\sqrt{x}-\sqrt{y})\left\{a\left(2\sqrt{x} - (\sqrt{x}-\sqrt{y})\right) - \frac{\sqrt{x}-\sqrt{y}}{\sqrt{x}}\right\} \\
&= 2a\sqrt{x}(\sqrt{x}-\sqrt{y}) - \left(a + \frac{1}{\sqrt{x}}\right)(\sqrt{x}-\sqrt{y})^2 \\
&\leq \frac{(2a\sqrt{x})^2}{4\left(a+\frac{1}{\sqrt{x}}\right)} = \frac{a^2x^{3/2}}{a\sqrt{x}+1} = \sqrt{x}\xi(a\sqrt{x}),
\end{aligned}$$

where we used $c_1 z - c_2 z^2 \leq c_1^2/(4c_2)$ for $c_1 \geq 0$ and $c_2 > 0$ with $a > -1/\sqrt{x}$. The second statement of the lemma follows since $\xi(z) \leq 2z^2$ for any $z \geq -1/2$. This completes the proof. ■

Define

$$\tilde{\ell}_t(i) = \frac{\mathbf{1}[i_t = i](1-r_t)}{x_t(i)} \in \left[0, \frac{2}{x_t(i)}\right].$$

Then, using Lemma 13, we can prove the following lemma, which plays a key role in proving Theorem 1.

Lemma 14 *Suppose that $\eta_t \leq 1/4$. Then it holds that*

$$\tilde{\ell}_t^\top(x_t - x_{t+1}) - \frac{1}{\eta_t}D(x_{t+1},x_t) \leq 4\eta_t \sum_{i=1}^m x_t(i)^{3/2}(1-x_t(i))\tilde{\ell}_t(i)^2.$$

Proof Let $k \in \arg\max_{i\in[m]} x_t(i)$. We then have

$$\begin{aligned}
\tilde{\ell}_t^\top(x_t - x_{t+1}) - \frac{1}{\eta_t}D(x_{t+1},x_t) &= \left(\tilde{\ell}_t - x_t(k)\tilde{\ell}_t(k)\mathbf{1}\right)^\top(x_t - x_{t+1}) - \frac{1}{\eta_t}D(x_{t+1},x_t) \\
&\leq 2\eta_t \sum_{i=1}^m x_t(i)^{3/2}\left(\tilde{\ell}_t(i) - x_t(k)\tilde{\ell}_t(k)\right)^2, \qquad (21)
\end{aligned}$$

where in the inequality we used the second statement in Lemma 13 with

$$\sqrt{x_t(i)}\eta_t \left(\tilde{\ell}_t(i) - x_t(k)\tilde{\ell}_t(k)\right) \geq -\sqrt{x_t(i)}\eta_t x_t(k)\tilde{\ell}_t(k) \geq -\eta_t(1-r_t) \geq -\frac{1}{2}$$

for each $i \in [m]$, which is due to the assumption that $\eta_t \leq 1/4$ and $1 - r_t \in [0, 2]$. We will upper bound the RHS of (21) below. First we have

$$\begin{aligned}
&\sum_{i=1}^m x_t(i)^{3/2} \left(\tilde{\ell}_t(i) - x_t(k)\tilde{\ell}_t(k)\right)^2 \\
&= x_t(k)^{3/2} \left(1 - x_t(k)\right)^2 \tilde{\ell}_t(k)^2 + \sum_{i \neq k} x_t(i)^{3/2} \left(\tilde{\ell}_t(i) - x_t(k)\tilde{\ell}_t(k)\right)^2 . \qquad (22)
\end{aligned}$$

The second term in the last equality is upper bounded by

$$\sum_{i \neq k} x_t(i)^{3/2} \left(\tilde{\ell}_t(i) - x_t(k)\tilde{\ell}_t(k)\right)^2 \leq \sum_{i \neq k} x_t(i)^{3/2}\tilde{\ell}_t(i)^2 + x_t(k)^2\tilde{\ell}_t(k)^2 \sum_{i \neq k} x_t(i)^{3/2}. \qquad (23)$$

The second term in the last inequality is further upper bounded by

$$\begin{aligned}
x_t(k)^2\tilde{\ell}_t(k)^2 \sum_{i \neq k} x_t(i)^{3/2} &\leq x_t(k)^2\tilde{\ell}_t(k)^2 \left(\sum_{i \neq k} x_t(i)\right)^{3/2} \leq x_t(k)^{3/2}\tilde{\ell}_t(k)^2 \left(\sum_{i \neq k} x_t(i)\right) \\
&= x_t(k)^{3/2}\tilde{\ell}_t(k)^2 \left(1 - x_t(k)\right), \qquad (24)
\end{aligned}$$

where the first inequality follows from the superadditivity of $z \mapsto z^{3/2}$ for $z \geq 0$ and the second inequality follows from $\sum_{i \neq k} x_t(i) \in [0, 1]$. Combining (22), (23), and (24), we have

$$\begin{aligned}
&\sum_{i=1}^m x_t(i)^{3/2} \left(\tilde{\ell}_t(i) - x_t(k)\tilde{\ell}_t(k)\right)^2 \\
&\leq x_t(k)^{3/2} \left(1 - x_t(k)\right)^2 \tilde{\ell}_t(k)^2 + \sum_{i \neq k} x_t(i)^{3/2}\tilde{\ell}_t(i)^2 + x_t(k)^{3/2}\tilde{\ell}_t(k)^2(1 - x_t(k)) \\
&\leq 2x_t(k)^{3/2} \left(1 - x_t(k)\right) \tilde{\ell}_t(k)^2 + 2\sum_{i \neq k} x_t(i)^{3/2}(1 - x_t(i))\tilde{\ell}_t(i)^2 \\
&= 2\sum_{i=1}^m x_t(i)^{3/2}(1 - x_t(i))\tilde{\ell}_t(i)^2,
\end{aligned}$$

where the second inequality follows from $1 - x_t(i) \geq 1/2$ for $i \neq k$ since $x_t(i) \leq 1/2$ for $i \neq k$. Finally, combining (21) with the last inequality, we obtain the desired bound. ■

Finally, we are ready to prove Theorem 1. We will show that Theorem 1 holds with $C_1 = 19$ and $C_2 = 2$.

Proof [Proof of Theorem 1] Let $x \in \mathcal{P}_m$ and $T_0 = 4$. When $m = 1$, the LHS and RHS of (4) are 0, and thus we consider the case of $m \geq 2$ below. Recall $\tilde{\ell}_t(i) = \frac{\mathbf{1}[i_t = i](1 - r_t)}{x_t(i)}$. Then, the regret of the row player can be rewritten as

$$\mathrm{Reg}_T(x) = \mathbf{E}\Bigg[\sum_{t=1}^{T}(x_t - x)^\top \ell_t\Bigg] \le \mathbf{E}\Bigg[\sum_{t=T_0+1}^{T}(x_t - x)^\top \ell_t\Bigg] + 2T_0 = \mathbf{E}\Bigg[\sum_{t=T_0+1}^{T}(x_t - x)^\top \hat{\ell}_t\Bigg] + 2T_0$$
$$= \mathbf{E}\Bigg[\sum_{t=T_0+1}^{T}(x_t - x)^\top \tilde{\ell}_t + \sum_{t=T_0+1}^{T}(x_t - x)^\top \left(\hat{\ell}_t - \tilde{\ell}_t\right)\Bigg] + 2T_0 = \mathbf{E}\Bigg[\sum_{t=T_0+1}^{T}(x_t - x)^\top \tilde{\ell}_t\Bigg] + 2T_0, \tag{25}$$

where the second equality follows from the unbiasedness of $\hat{\ell}_t$ and the last equality follows from $\hat{\ell}_t - \tilde{\ell}_t = \mathbf{1}$. From the fact that the outputs of FTRL with loss estimator $\hat{\ell}_t$ and $\tilde{\ell}_t$ are the same and Lemma 12, the inside of the expectation in (25) is upper bounded by

$$\begin{aligned}\sum_{t=T_0+1}^{T} \tilde{\ell}_t^\top (x_t - x) \le \sum_{t=T_0+1}^{T} &\left(\tilde{\ell}_t^\top (x_t - x_{t+1}) - \frac{1}{\eta_t} D(x_{t+1}, x_t)\right) \\ &+ \sum_{t=T_0+1}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\left(\psi(x^*) - \psi(x_{t+1})\right) \\ &+ \frac{1}{\eta_{T_0+1}} (\psi(x^*) - \psi(x_{T_0+1})) - \frac{1}{\eta_{T+1}} D(x^*, x_{T+1}),\end{aligned} \tag{26}$$

We first consider the first term in (26). For $t \ge T_0 = 4$, we have $\eta_t = 1/(2\sqrt{t}) \le 1/4$, and thus from Lemma 14,

$$\tilde{\ell}_t^\top (x_t - x_{t+1}) - \frac{1}{\eta_t} D(x_{t+1}, x_t) \le 4\eta_t \sum_{i=1}^{m} x_t(i)^{3/2}(1 - x_t(i))\tilde{\ell}_t(i)^2. \tag{27}$$

Let $i^* \in [m]$. Then, using $\mathbf{E}_{r_t,i_t,j_t}[\tilde{\ell}_t(i)^2 \mid x_t] \le 4/x_t(i)$, we have

$$\begin{aligned}&\mathbf{E}_{r_t,i_t,j_t}\Bigg[\sum_{i=1}^{m} x_t(i)^{3/2}(1 - x_t(i))\tilde{\ell}_t(i)^2 \mid x_t\Bigg] \le 4\sum_{i=1}^{m} \sqrt{x_t(i)}(1 - x_t(i)) \\ &\le 4\sum_{i \ne i^*} \sqrt{x_t(i)} + 4(1 - x_t(i^*)) \le 8\sum_{i \ne i^*} \sqrt{x_t(i)},\end{aligned} \tag{28}$$

where the last inequality follows from $1 - x_t(i^*) = \sum_{i \ne i^*} x_t(i) \le \sum_{i \ne i^*} \sqrt{x_t(i)}$.

We next consider the second and third terms in (26). We first observe that $1/\eta_{t+1} - 1/\eta_t = 2(\sqrt{t+1} - \sqrt{t}) \le 1/\sqrt{t} \le \sqrt{2/(t+1)}$ and $\psi(x^*) - \psi(x_{t+1}) \le 2\sum_{i=1}^{m} \sqrt{x_{t+1}(i)} - 2 \le 2\sum_{i \in [m]\setminus\{i^*\}} \sqrt{x_{t+1}(i)}$. Using these inequalities, we have

$$\begin{aligned}&\sum_{t=T_0+1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) (\psi(x^*) - \psi(x_{t+1})) \\ &\le 2\sqrt{2}\sum_{t=T_0+1}^{T} \frac{1}{\sqrt{t+1}} \sum_{i \in [m]\setminus\{i^*\}} \sqrt{x_{t+1}(i)} \le 2\sqrt{2}\sum_{t=T_0+2}^{T} \frac{1}{\sqrt{t}} \sum_{i \in [m]\setminus\{i^*\}} \sqrt{x_t(i)} + 2\sqrt{2}\sqrt{\frac{m}{T+1}},\end{aligned} \tag{29}$$

where in the last inequality we used the Cauchy-Schwarz inequality. The remaining term in (26) is at most

$$\frac{1}{\eta_{T_0+1}}\left(\psi(x^*)-\psi(x_{T_0+1})\right) \leq 2\sqrt{2}\sqrt{T_0+1}\sum_{i\neq i^*}\sqrt{x_{T_0+1}(i)} \leq 2\sqrt{10}\sum_{i\neq i^*}\sqrt{x_{T_0+1}(i)}. \tag{30}$$

Finally, by combining (25) with (26), (28), (29), and (30), we obtain that for any $i^* \in [m]$,

$$\mathrm{Reg}_T(x) \leq 2T_0 + 2\sqrt{2}\sqrt{\frac{m}{T+1}} + \mathbf{E}\left[19\sum_{t=T_0+1}^{T}\sum_{i\in[m]\setminus\{i^*\}}\sqrt{\frac{x_t(i)}{t}} - 2\sqrt{T+1}\cdot D(x,x_{T+1})\right]. \tag{31}$$

Since we have $\sum_{t=1}^{T_0}\sum_{i\in[m]\setminus\{i^*\}}\sqrt{x_t(i)} \geq (\sqrt{m}-1)+(T_0-1) = \sqrt{m}+2$, the last inequality implies that the choice of $C_1 = 19$, which is larger than $\frac{2T_0+2\sqrt{2}\sqrt{m/(T+1)}}{\sqrt{m}+2}(\leq 4)$, implies that the first three terms in (31) is upper bounded by

$$2T_0 + 2\sqrt{2}\sqrt{\frac{m}{T+1}} \leq C_1(\sqrt{m}+2) \leq C_1\sum_{t=1}^{T_0}\sum_{i\in[m]\setminus\{i^*\}}\sqrt{x_t(i)}.$$

Combining this inequality with (31) implies that Theorem 1 holds with $C_1 = 19$ and $C_2 = 2$ as desired. ■

Appendix C. Properties of the Duality Gap

Lemma 15 *For any $A \in [-1,1]^{m\times n}$, $\mathrm{DGap}(\hat{x},\hat{y})$ defined in (1) is 1-Lipschitz w.r.t. L^1 norm, i.e., it holds for any $\hat{x},\hat{x}' \in \mathcal{P}_m$ and $\hat{y},\hat{y}' \in \mathcal{P}_n$ that*

$$|\,\mathrm{DGap}(\hat{x},\hat{y}) - \mathrm{DGap}(\hat{x}',\hat{y}')| \leq \|\hat{x}-\hat{x}'\|_1 + \|\hat{y}-\hat{y}'\|_1. \tag{32}$$

Proof We can express $\mathrm{DGap}(\hat{x},\hat{y})$ as follows:

$$\mathrm{DGap}(\hat{x},\hat{y}) = \max_{x\in\mathcal{P}_m}\left\{x^\top A\hat{y}\right\} + \max_{y\in\mathcal{P}_n}\left\{-\hat{x}^\top Ay\right\}. \tag{33}$$

Let $\tilde{x} \in \arg\max_{x\in\mathcal{P}_m}\left\{x^\top A\hat{y}\right\}$. We then have

$$\max_{x\in\mathcal{P}_m}\left\{x^\top A\hat{y}\right\} - \max_{x\in\mathcal{P}_m}\left\{x^\top A\hat{y}'\right\} = \tilde{x}^\top A\hat{y} - \max_{x\in\mathcal{P}_m}\left\{x^\top A\hat{y}'\right\} \leq \tilde{x}^\top A\hat{y} - \tilde{x}^\top A\hat{y}' \tag{34}$$

$$= \tilde{x}^\top A(\hat{y}-\hat{y}') \leq \|A^\top\tilde{x}\|_\infty\|\hat{y}-\hat{y}'\|_1 \leq \|\hat{y}-\hat{y}'\|_1. \tag{35}$$

In a similar way, we can show the following:

$$\max_{y\in\mathcal{P}_n}\left\{-\hat{x}^\top Ay\right\} - \max_{y\in\mathcal{P}_n}\left\{-\hat{x}'^\top Ay\right\} \leq \|\hat{x}-\hat{x}'\|_1. \tag{36}$$

By comibning (33), (35) and (36), we obtain

$$\mathrm{DGap}(\hat{x},\hat{y}) - \mathrm{DGap}(\hat{x}',\hat{y}') \leq \|\hat{x}-\hat{x}'\|_1 + \|\hat{y}-\hat{y}'\|_1. \tag{37}$$

In a similar way, we can show $\mathrm{DGap}(\hat{x}',\hat{y}') - \mathrm{DGap}(\hat{x},\hat{y}) \leq \|\hat{x}-\hat{x}'\|_1 + \|\hat{y}-\hat{y}'\|_1$ as well, which completes the proof. ■

Appendix D. Proof of Theorem 2

In this appendix section, we will prove our main regret bound theorem by first presenting a generalized formulation that encompasses both inequalities stated in the main text. We do this by first defining a unified notation of the gap parameters.

Definition 16 (Admissible (I, Δ, π) and (J, Δ', π')) *Denote $v = \max_{x \in \mathcal{P}_m} \left\{ \min_{y \in \mathcal{P}_n} x^\top A y \right\}$. An action subset $I \subseteq [m]$, a gap vector $\Delta \in \mathbb{R}^m_{\geq 0}$ and a mapping $\pi : \mathcal{P}_m \to \mathcal{X}_\star$ are together called admissible for the row player if*

- *The entries $\Delta(i)$ are positive for every $i \notin I$.*
- *For any $x \in \mathcal{P}_m$, the NE strategy $x_\star = \pi(x) \in \mathcal{X}_\star$ must satisfy:*

$$\mathrm{DGap}(x, y_\star) = v - \min_{y \in \mathcal{P}_n} \left\{ x^\top A y \right\} \geq \Delta \cdot (x - x_\star)_+, \tag{38}$$

  where $y_\star \in \mathcal{Y}_\star$ is an arbitrary NE strategy, and we define $(x)_+ = \max\{x, 0\}$ which applies entrywise to vectors.

  The admissibility for a subset of actions $J \subseteq [n]$, a gap vector $\Delta' \in \mathbb{R}^n_{\geq 0}$, and a mapping $\pi' : \mathcal{P}_n \to \mathcal{Y}_\star$ can be analogously defined for the column player, with

$$\mathrm{DGap}(x_\star, y) = \min_{x \in \mathcal{P}_m} \left\{ x^\top A y \right\} - v \geq \Delta' \cdot (y - y_\star)_+, \tag{39}$$

The following is the full version of our main theorem.

Theorem 17 *If both players follow the Tsallis-INF algorithm, then for any admissible (I, Δ, π) and (J, Δ', π') (Definition 16) such that $I \neq \emptyset$, $J \neq \emptyset$, we have*

$$\begin{aligned} &\max\Big\{ \mathrm{Reg}_T(x) + \sqrt{T} C_2 \mathbf{E}\big[D(x, x_{T+1})\big], \mathrm{Reg}'_T(y) + \sqrt{T} C_2 \mathbf{E}\big[D(y, y_{T+1})\big] \Big\} \\ &= O\Big(\sqrt{T} \big(\sqrt{|I| - 1} + \sqrt{|J| - 1} + \gamma \sqrt{L} + \gamma' \sqrt{L'} \big) + \omega L + \omega' L' \Big) \end{aligned} \tag{40}$$

for any $x \in \mathcal{P}_m$ and $y \in \mathcal{P}_n$, where

$$\begin{aligned} \omega &= \sum_{i \notin I} \frac{1}{\Delta(i)}, \quad \gamma = \max_{x_\star \in \mathcal{X}_\star} \sum_{i \notin I} \sqrt{x_\star(i)}, \quad L = \min\Big\{ \log_+ \frac{T(m - |I|)}{\omega^2}, \log_+ \frac{m - |I|}{\gamma^2} \Big\}, \\ \omega' &= \sum_{j \notin J} \frac{1}{\Delta'(j)}, \quad \gamma' = \max_{y_\star \in \mathcal{Y}_\star} \sum_{j \notin J} \sqrt{y_\star(j)}, \quad L' = \min\Big\{ \log_+ \frac{T(n - |J|)}{\omega'^2}, \log_+ \frac{n - |J|}{\gamma'^2} \Big\}. \end{aligned} \tag{41}$$

We note that the both parts of Theorem 2 are special cases of this theorem. In fact, if $(x_\star, y_\star, I, J, \Delta, \Delta')$ are given by the first condition in Theorem 2, we can verify the admissibility of (I, Δ) by directly plugging the definition $\Delta = \left(x_\star^\top A y_\star \right) \mathbf{1} - A y_\star$ into (38); and similarly admissibility can be proven for (J, Δ').

The first bound in Theorem 2 can be recovered by observing that $\gamma = \gamma' = 0$ as we define I and J to be the support for $x_\star$ and $y_\star$, and by taking the second branch in the definitions of L and L'.

Proof [of Theorem 17] From Equation (38) and (39), we have

$$\begin{aligned}
\mathrm{Reg}_T + \mathrm{Reg}'_T &= \max_{x \in \mathcal{P}_m, y \in \mathcal{P}_n} \mathbf{E}\left[\sum_{t=1}^{T} \left(x^\top A y_t - x_t^\top A y\right)\right] \\
&= \max_{x \in \mathcal{P}_m} \left\{ x^\top A \, \mathbf{E}\left[\sum_{t=1}^{T} y_t\right] \right\} - \min_{y \in \mathcal{P}_n} \left\{ \mathbf{E}\left[\sum_{t=1}^{T} x_t\right] A y \right\} \\
&= T \max_{x \in \mathcal{P}_m} \left\{ x^\top A \, \mathbf{E}[\bar{y}_T] \right\} - T \min_{y \in \mathcal{P}_n} \left\{ \mathbf{E}[\bar{x}_T] A y \right\} \\
&\geq T\Delta \cdot (\mathbf{E}[\bar{x}_T] - x_\star)_+ + T\Delta' \cdot (\mathbf{E}[\bar{y}_T] - y_\star)_+, \qquad (42)
\end{aligned}$$

where we define $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$ and $\bar{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t$.

From Theorem 1, we know that the following bound holds:

$$\mathrm{Reg}_T(x) + C_2\sqrt{T}\,\mathbf{E}\big[D(x, x_{T+1})\big] \leq C_1 \,\mathbf{E}\left[\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \sum_{i \neq i_\star} \sqrt{x_t(i)}\right], \qquad (43)$$

for an arbitrary $i_\star \in I$, and specifically

$$\mathrm{Reg}_T \leq C_1 \,\mathbf{E}\left[\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \sum_{i \neq i_\star} \sqrt{x_t(i)}\right] \overset{\mathrm{def}}{=} C_1 \,\mathbf{E}[S]. \qquad (44)$$

Define S as the summation inside the expectation bracket above. We can split it into two parts:

$$S = \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \sum_{i \in I \setminus \{i_\star\}} \sqrt{x_t(i)} + \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \sum_{i \notin I} \sqrt{x_t(i)}. \qquad (45)$$

The sum within I can be bounded with a Cauchy-Schwarz inequality:

$$\begin{aligned}
\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \sum_{i \in I \setminus \{i_\star\}} \sqrt{1 \cdot x_t(i)} &\leq \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \sqrt{\sum_{i \in I \setminus \{i_\star\}} 1 \sum_{i \in I \setminus \{i_\star\}} \sqrt{x_t(i)}^2} \\
&\leq 2\sqrt{T}\sqrt{|I| - 1}. \qquad (46)
\end{aligned}$$

We define $\bar{x}(i) = \frac{1}{T}\sum_{t=1}^{T} x_t(i)$ as a notational shorthand. To handle the sum outside I, we have due to the Cauchy-Schwarz inequality,

$$\begin{aligned}
\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \sum_{i \notin I} \sqrt{x_t(i)} &= \sum_{t=1}^{s} \frac{1}{\sqrt{t}} \sum_{i \notin I} \sqrt{x_t(i)} + \sum_{i \notin I} \sum_{t=s+1}^{T} \frac{1}{\sqrt{t}} \sqrt{x_t(i)} \\
&\leq \sum_{t=1}^{s} \frac{1}{\sqrt{t}} \sqrt{m - |I|} + \sum_{i \notin I} \sqrt{\left(\sum\nolimits_{t=s+1}^{T} 1/t\right)\left(\sum\nolimits_{t=s+1}^{T} x_t(i)\right)} \\
&\leq 2\sqrt{s(m - |I|)} + \sqrt{T \log(T/s)} \sum_{i \notin I} \sqrt{\bar{x}_T(i)}, \qquad (47)
\end{aligned}$$

in which $s \in [T]$ is a parameter yet to be determined. We bound the last summation above in expectation. Definition 16 guarantees that, for $\mathbf{E}[\bar{x}_T]$, it is possible to find a Nash equilibrium strategy $x_\star$ such that the following holds:

$$\begin{aligned}
\mathbf{E}\Big[\sum_{i \notin I} \sqrt{\bar{x}_T(i)}\Big] &\le \sum_{i \notin I} \sqrt{\mathbf{E}[\bar{x}_T(i)]} \\
&\le \sum_{i \notin I} \sqrt{x_\star(i)} + \sum_{i \notin I} \sqrt{(\mathbf{E}[\bar{x}_T(i)] - x_\star(i))_+} \\
&\le \gamma + \sum_{i \notin I} \sqrt{\frac{1}{\Delta(i)} \Big(\Delta(i) \cdot (\mathbf{E}[\bar{x}_T(i)] - x_\star(i))_+\Big)} \\
&\le \gamma + \sqrt{\Big(\sum_{i \notin I} \frac{1}{\Delta(i)}\Big)\Big(\sum_{i \notin I} \Delta(i) \cdot (\mathbf{E}[\bar{x}_T(i)] - x_\star(i))_+\Big)} \\
&= \gamma + \sqrt{\omega\Delta \cdot (\mathbf{E}[\bar{x}_T] - x_\star)_+}.
\end{aligned}$$

We put (45), (46), (47) together, and take the expectation on both sides to get

$$\mathbf{E}[S] \le 2\sqrt{T}\sqrt{|I|-1} + 2\sqrt{(m-|I|)s} + \sqrt{T\log(T/s)}\big(\gamma + \sqrt{\omega\Delta \cdot (\mathbf{E}[\bar{x}] - x_\star)_+}\big), \tag{48}$$

where the last inequality is due to Jensen's inequality and the concavity of square root. For similarly defined S' and s', we also have a similar bound:

$$\mathbf{E}[S'] \le 2\sqrt{T}\sqrt{|J|-1} + 2\sqrt{(n-|J|)s'} + \sqrt{T\log(T/s')}\big(\gamma + \sqrt{\omega'\Delta' \cdot (\mathbf{E}[\bar{y}] - y_\star)_+}\big). \tag{49}$$

Now, note that from Jensen's inequality, we have

$$\begin{aligned}
&\sqrt{T\log(T/s)\omega}\sqrt{\Delta \cdot (\mathbf{E}[\bar{x}_T] - x_\star)_+} + \sqrt{T\log(T/s')\omega'}\sqrt{\Delta' \cdot (\mathbf{E}[\bar{y}_T] - y_\star)_+} \\
&\le \sqrt{T\big(\log(T/s)\omega + \log(T/s')\omega'\big)\big(\Delta \cdot (\mathbf{E}[\bar{x}_T] - x_\star)_+ + \Delta' \cdot (\mathbf{E}[\bar{y}_T] - y_\star)_+\big)} \\
&\le \sqrt{\big(\omega\log(T/s) + \omega'\log(T/s')\big)\big(\mathrm{Reg}_T + \mathrm{Reg}'_T\big)} && \langle \text{from (42)} \rangle \\
&\le \sqrt{C_1\big(\omega\log(T/s) + \omega'\log(T/s')\big)\big(\mathbf{E}[S] + \mathbf{E}[S']\big)} && \langle \text{from (44) and its counterpart} \rangle
\end{aligned}$$

If we add (48) and (49), we get the following bound:

$$\begin{aligned}
\mathbf{E}[S+S'] &\le 2\sqrt{T}\big(\sqrt{|I|-1} + \sqrt{|J|-1}\big) + 2\sqrt{(m-|I|)s} + 2\sqrt{(n-|J|)s'} \\
&\quad + \gamma\sqrt{T}\sqrt{\log(T/s)} + \gamma\sqrt{T}\sqrt{\log(T/s')} \\
&\quad + \sqrt{C_1\big(\omega\log(T/s) + \omega'\log(T/s')\big)(\mathbf{E}[S] + \mathbf{E}[S'])} \\
&\le 4\sqrt{T}\big(\sqrt{|I|-1} + \sqrt{|J|-1}\big) + 4\sqrt{(m-|I|)s} + 4\sqrt{(n-|J|)s'} \\
&\quad + 2\gamma\sqrt{T}\sqrt{\log(T/s)} + 2\gamma\sqrt{T}\sqrt{\log(T/s')} \\
&\quad + 2C_1\omega\log(T/s) + 2C_1\omega'\log(T/s'),
\end{aligned} \tag{50}$$

where in the last inequality we apply Lemma 10. We take $s = \big\lceil \min\{\frac{T}{2}, \frac{\max\{\omega^2, \gamma^2 T\}}{m-|I|}\}\big\rceil$; since $\Delta_i \le 2$ for every i, we know that $\omega \ge \frac{1}{2}(m-|I|)$, so we have $\frac{\omega^2}{m-|I|} \ge \frac{1}{4}$, and thus the rounding-up increases s by a factor of at most 4. This implies that $4\sqrt{(m-|I|)s} \le 16\gamma\sqrt{T} + 16\omega$.

We also have

$$\begin{aligned} s &\geq \min\left\{\frac{T}{2}, \frac{\max\{\omega^2, \gamma^2 T\}}{m - |I|}\right\}, \\ \frac{T}{s} &\leq \max\left\{2, \min\left\{\frac{T(m-|I|)}{\omega^2}, \frac{m-|I|}{\gamma^2}\right\}\right\}, \\ \log\frac{T}{s} &\leq \min\left\{\log_+ \frac{T(m-|I|)}{\omega^2}, \log_+ \frac{m-|I|}{\gamma^2}\right\} \overset{\text{def}}{=} L. \end{aligned}$$

A similar definition and respective inequalities are omitted for s'. Plugging these bounds into (50) yields

$$\begin{aligned} \mathbf{E}[S + S'] \leq{} & 4\sqrt{T(|I|-1)} + 18\gamma\sqrt{TL} + (2C_1 + 16)\omega L \\ & + 4\sqrt{T(|J|-1)} + 18\gamma'\sqrt{TL'} + (2C_1 + 16)\omega' L'. \end{aligned}$$

Together with (44) and the definition of S and S', this completes the proof. ■

The second part of Theorem 2 is a corollary of the following theorem:

Theorem 18 *Suppose that $c, c' \in (0, 1]$ satisfy*

$$\mathrm{DGap}(x, \hat{y}) \geq c \min_{x_\star \in \mathcal{X}_\star} \|x - x_\star\|_1 + c' \min_{y_\star \in \mathcal{Y}_\star} \|y - y_\star\|_1 \tag{51}$$

for all $x \in \mathcal{P}_m$ and $y \in \mathcal{P}_n$. Define $\gamma \geq 0, \gamma' \geq 0, \rho > 0$ and $\rho' > 0$ by

$$\gamma = \max_{x \in \mathcal{X}_\star}\left\{\sum_{i=1}^{m} \sqrt{x(i)}\right\} - 1, \quad \gamma' = \max_{y \in \mathcal{Y}_\star}\left\{\sum_{j=1}^{n} \sqrt{y(j)}\right\} - 1, \tag{52}$$

$$\rho = \gamma\sqrt{T \log_+\left(\frac{m-1}{\gamma^2}\right)} + \frac{m-1}{c}\log_+\left(\frac{c^2 T}{m-1}\right), \tag{53}$$

$$\rho' = \gamma'\sqrt{T \log_+\left(\frac{n-1}{\gamma'^2}\right)} + \frac{n-1}{c'}\log_+\left(\frac{c'^2 T}{n-1}\right). \tag{54}$$

If both players follow the Tsallis-INF algorithm, we have

$$\begin{aligned} \mathrm{Reg}_T(x) + \sqrt{T}C_2\mathbf{E}[D(x, x_{T+1})] &= O\left(\rho + \sqrt{(\rho + \rho')\frac{m-1}{c}\log_+\left(\frac{c^2 T}{m-1}\right)}\right), \\ \mathrm{Reg}'_T(y) + \sqrt{T}C_2\mathbf{E}[D(y, y_{T+1})] &= O\left(\rho' + \sqrt{(\rho + \rho')\frac{n-1}{c'}\log_+\left(\frac{c'^2 T}{n-1}\right)}\right) \end{aligned}$$

for any $x \in \mathcal{P}_m$ and $y \in \mathcal{P}_n$. Consequently, we have

$$\limsup_{T\to\infty}\frac{\mathrm{Reg}_T}{\sqrt{T}} = O\left(\gamma\sqrt{\log_+\left(\frac{m-1}{\gamma^2}\right)}\right), \quad \limsup_{T\to\infty}\frac{\mathrm{Reg}'_T}{\sqrt{T}} = O\left(\gamma'\sqrt{\log_+\left(\frac{n-1}{\gamma'^2}\right)}\right).$$

From this theorem and the AM-GM inequality, we have

$$\begin{aligned}\max\{\mathrm{Reg}_T, \mathrm{Reg}'_T\} &= O\left(\rho + \rho' + \sqrt{(\rho+\rho')\frac{n-1}{c'}\left(\log_+\left(\frac{c^2 T}{m-1}\right) + \log_+\left(\frac{c'^2 T}{n-1}\right)\right)}\right) \\ &= O(\rho + \rho'),\end{aligned}$$

which implies that the second part of Theorem 2 holds.
Proof From Theorem 1, for any $s \in [T]$ any $i^* \in [m]$, and any $x \in \mathcal{P}_m$, we have

$$\begin{aligned}&\mathrm{Reg}_T(x) + \sqrt{T} C_2 \mathbf{E}[D(x, x_{T+1})] \\ &= O\left(\mathbf{E}\left[\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \sum_{i \in [m]\setminus\{i^*\}} \sqrt{x_t(i)}\right]\right) \\ &= O\left(\mathbf{E}\left[\sum_{t=1}^{s} \frac{1}{\sqrt{t}} \sum_{i \in [m]\setminus\{i^*\}} \sqrt{x_t(i)} + \sum_{t=s+1}^{T} \frac{1}{\sqrt{t}} \sum_{i \in [m]\setminus\{i^*\}} \sqrt{x_t(i)}\right]\right) \\ &= O\left(\sqrt{(m-1)s} + \sum_{i \in [m]\setminus\{i^*\}} \sqrt{\mathbf{E}\left[\sum_{t=s+1}^{T} x_t(i)\right] \log \frac{T}{s}}\right) \\ &= O\left(\sqrt{(m-1)s} + \sqrt{T \log \frac{T}{s}} \sum_{i \in [m]\setminus\{i^*\}} \sqrt{\mathbf{E}\left[\bar{x}_T(i)\right]}\right). \qquad (55)\end{aligned}$$

Denote

$$\tilde{x}_T \in \operatorname*{arg\,min}_{x_\star \in \mathcal{X}_\star} \| \mathbf{E}[\bar{x}_T] - x_\star \|_1, \quad \tilde{y}_T \in \operatorname*{arg\,min}_{y_\star \in \mathcal{Y}_\star} \| \mathbf{E}[\bar{y}_T] - y_\star \|_1. \qquad (56)$$

We then have

$$\begin{aligned}&\sum_{i \in [m]\setminus\{i^*\}} \sqrt{\mathbf{E}[\bar{x}_T(i)]} \\ &\leq \sum_{i \in [m]\setminus\{i^*\}} \sqrt{\tilde{x}_T(i)} + \sum_{i \in [m]\setminus\{i^*\}} \sqrt{| \mathbf{E}[\bar{x}_T(i)] - \tilde{x}_T(i)|} \\ &\leq \sum_{i \in [m]\setminus\{i^*\}} \sqrt{\tilde{x}_T(i)} + \sqrt{(m-1) \sum_{i \in [m]\setminus\{i^*\}} |\mathbf{E}[\bar{x}_T(i)] - \tilde{x}_T(i)|} \qquad \text{(Cauchy-Schwarz)} \\ &\leq \frac{1}{2}\left(\sum_{i=1}^{m} \sqrt{\tilde{x}_T(i)} - 1\right) + \sqrt{(m-1)\|\mathbf{E}[\bar{x}_T] - \tilde{x}_T\|_1} \\ &\leq \frac{1}{2}\gamma + \sqrt{(m-1)\|\mathbf{E}[\bar{x}_T] - \tilde{x}_T\|_1} \qquad \text{(From (52) and (56))} \\ &\leq \frac{1}{2}\gamma + \sqrt{\frac{m-1}{c} \mathrm{DGap}(\mathbf{E}[\bar{x}_T], \mathbf{E}[\bar{y}_T])}. \qquad \text{(From (51) and (56))} \quad (57)\end{aligned}$$

The third inequality can be shown by setting $i^* \in \arg\max_{i\in[m]} \{\tilde{x}_T(i)\}$. From (55) and (57), we have

$$\begin{aligned}&\mathrm{Reg}_T(x) + \sqrt{T} C_2 \mathbf{E}[D(x, x_{T+1})]\\ &= O\left(\sqrt{(m-1)s} + \sqrt{T \log \frac{T}{s}} \left(\gamma + \sqrt{\frac{m-1}{c} \mathrm{DGap}(\mathbf{E}[\bar{x}_T], \mathbf{E}[\bar{y}_T])} \right) \right). \end{aligned} \tag{58}$$

Similarly, for any $s' \in [T]$, we have

$$\begin{aligned}&\mathrm{Reg}'_T(y) + \sqrt{T} C_2 \mathbf{E}[D(y, y_{T+1})]\\ &= O\left(\sqrt{(n-1)s'} + \sqrt{T \log \frac{T}{s'}} \left(\gamma' + \sqrt{\frac{n-1}{c'} \mathrm{DGap}(\mathbf{E}[\bar{x}_T], \mathbf{E}[\bar{y}_T])} \right) \right).\end{aligned}$$

Here, as we have

$$T \cdot \mathrm{DGap}(\mathbf{E}[\bar{x}_T], \mathbf{E}[\bar{y}_T]) = \mathrm{Reg}_T + \mathrm{Reg}'_T,$$

the value of $\mathrm{DGap}(\mathbf{E}[\bar{x}_T], \mathbf{E}[\bar{y}_T])$ is bounded as

$$\begin{aligned}T \cdot \mathrm{DGap}(\mathbf{E}[\bar{x}_T], \mathbf{E}[\bar{y}_T]) = O\Bigg(&\sqrt{(m-1)s} + \sqrt{(n-1)s'} + \gamma \sqrt{T \log \frac{T}{s}} + \gamma' \sqrt{T \log \frac{T}{s'}} \\ &+ \sqrt{\left(\frac{m-1}{c} \log \frac{T}{s} + \frac{n-1}{c'} \log \frac{T}{s'} \right) T \cdot \mathrm{DGap}(\mathbf{E}[\bar{x}_T], \mathbf{E}[\bar{y}_T])} \Bigg)\end{aligned}$$

for any $s, s' \in [T]$, which implies

$$\begin{aligned}&T \cdot \mathrm{DGap}(\mathbf{E}[\bar{x}_T], \mathbf{E}[\bar{y}_T])\\ &= O\left(\sqrt{(m-1)s} + \sqrt{(n-1)s'} + \gamma \sqrt{T \log \frac{T}{s}} + \gamma' \sqrt{T \log \frac{T}{s'}} + \frac{m-1}{c} \log \frac{T}{s} + \frac{n-1}{c'} \log \frac{T}{s'} \right).\end{aligned}$$

By choosing

$$s = \left\lceil \min \left\{ T, \max \left\{ \frac{\gamma^2 T}{m-1}, \frac{m-1}{c^2} \right\} \right\} \right\rceil \quad s' = \left\lceil \min \left\{ T, \max \left\{ \frac{\gamma'^2 T}{n-1}, \frac{n-1}{c'^2} \right\} \right\} \right\rceil \tag{59}$$

we have

$$\begin{aligned}&T \cdot \mathrm{DGap}(\mathbf{E}[\bar{x}_T], \mathbf{E}[\bar{y}_T])\\ &= O\left(\gamma \sqrt{T \log_+ \left(\frac{m-1}{\gamma^2} \right)} + \frac{m-1}{c} \log_+ \left(\frac{c^2 T}{m-1} \right) + \gamma' \sqrt{T \log_+ \left(\frac{n-1}{\gamma'^2} \right)} + \frac{n-1}{c'} \log_+ \left(\frac{c'^2 T}{n-1} \right) \right)\\ &= O\big(\rho + \rho'\big).\end{aligned}$$

From this and (58) with (59), we have

$$\begin{aligned}\mathrm{Reg}_T(x) + \sqrt{T} C_2 \mathbf{E}[D(x, x_{T+1})] &= O\left(\rho + \sqrt{(\rho + \rho') \frac{m-1}{c} \log \frac{T}{s}} \right)\\ &= O\left(\rho + \sqrt{(\rho + \rho') \frac{m-1}{c} \log_+ \left(\frac{c^2 T}{m-1} \right)} \right).\end{aligned}$$

Similarly, we obtain the desired upper bound on $\mathrm{Reg}'_T(y) + \sqrt{T} C_2 \mathbf{E}[D(y, y_{T+1})]$, which completes the proof. ■

Appendix E. Proof of Theorem 4

When A is given by (7), we can see that $(i_\star, j_\star)$ is a Nash equilibrium of the game with payoff matrix A. In fact, if $x_\star$ and $y_\star$ are the indicator vectors of $i_\star$ and $j_\star$, it holds for any $x \in \mathcal{P}_m$ and $y \in \mathcal{P}_n$ that

$$
\begin{aligned}
x^\top A y_\star - x_\star^\top A y &= {\Delta'}^\top y_\star - x^\top \Delta - {\Delta'}^\top y + x_\star^\top \Delta = (x_\star - x)^\top \Delta + (y_\star - y)^\top \Delta' \\
&= -x^\top \Delta - y^\top \Delta' = -\Big(\sum_{i \in [m]} \Delta(i) x(i) + \sum_{j \in [n]} \Delta'(j) y(j) \Big) \leq 0, \qquad (60)
\end{aligned}
$$

which means that $\mathrm{DGap}(x_\star, y_\star) = 0$.

In this section, let $x_t \in \mathcal{P}_m$ and $y_t \in \mathcal{P}_n$ denote indicator vectors of $i_t \in [m]$ and $j_t \in [n]$, respectively. For any fixed algorithm and the true payoff matrix A, we denote the regret of the algorithm as

$$
R_T(A) = \mathrm{Reg}_T(x_\star) + \mathrm{Reg}'_T(y_\star) = \mathbf{E}\Big[\sum_{t=1}^T \big(x_\star^\top A y_t - x_t^\top A y_\star \big) \Big].
$$

Then, if A is given by (7), from (60), we have

$$
R_T(A) = \mathbf{E}\Big[\sum_{t=1}^T \big(x_t^\top \Delta + y_t^\top \Delta' \big) \Big] = \sum_{i=1}^m \Delta(i) N_{T,i}(A) + \sum_{j=1}^n \Delta'(j) N'_{T,j}(A). \qquad (61)
$$

We can show Theorem 4 by using the following lemma:

Lemma 19 *Suppose A is given by (7). Fix an arbitrary $i \in [m] \setminus \{i_\star\}$. Let $\tilde{\Delta} = \Delta - 2\Delta_i \chi_i$ and $\tilde{A} = \mathbf{1}_m {\Delta'}^\top - \tilde{\Delta} \mathbf{1}_n^\top$. We then have*

$$
(\Delta(i))^2 N_{T,i}(A) \geq \frac{1}{5} \ln \frac{T}{2\big(N_{T,i}(A) + T - N_{T,i}(\tilde{A})\big)}.
$$

Proof Note first that, for $p \in [3/8, 1/2]$ and $\delta \in [0, 1/4]$, we have

$$
\begin{aligned}
D_{\mathrm{KL}}(p, p+\delta) &= p \ln \frac{p}{p+\delta} + (1-p) \ln \frac{1-p}{1-p-\delta} \\
&= -p \ln \Big(1 + \frac{\delta}{p} \Big) - (1-p) \ln \Big(1 - \frac{\delta}{1-p} \Big) \\
&\leq p \ln \Big(-\frac{\delta}{p} + \big(\frac{\delta}{p}\big)^2 \Big) + (1-p) \ln \Big(\frac{\delta}{1-p} + \big(\frac{\delta}{1-p}\big)^2 \Big) \\
&= \frac{\delta^2}{p(1-p)} \leq 5\delta^2. \qquad (62)
\end{aligned}
$$

Let P and $\tilde{P}$ be distributions of $\{(i_t, j_t, \ell_t)\}_{t\in[T]}$ for A and $\tilde{A}$, respectively. Then, from the Bretagnolle-Huber inequality (e.g., Canonne, 2022, Corollary 4), we have

$$D_{\mathrm{TV}}(P,\tilde{P}) \leq 1 - \frac{1}{2}\exp(-D_{\mathrm{KL}}(P,\tilde{P})).$$

From the chain rule for the KL divergence (e.g., Lattimore and Szepesvári, 2020, Lemma 15.1), we have

$$\begin{aligned} D_{\mathrm{KL}}(P,\tilde{P}) &= \mathop{\mathbf{E}}_{\{(i_t,j_t,\ell_t)\}\sim P}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}(\mathrm{Ber}^{\pm}(A_{i_t,j_t}), \mathrm{Ber}^{\pm}(\tilde{A}_{i_t,j_t}))\right] \\ &\leq \mathop{\mathbf{E}}_{\{(i_t,j_t,\ell_t)\}\sim P}\left[\sum_{t=1}^{T} \mathbf{1}[i_t = i]\cdot 5(\Delta(i))^2\right] = 5N_{T,i}(A)(\Delta(i))^2, \end{aligned}$$

where the inequality follows from the definition of $\tilde{A}$ and (62). By combining above inequalities, we obtain

$$\frac{1}{T}|N_{T,i}(A) - N_{T,i}(\tilde{A})| \leq D_{\mathrm{TV}}(P,\tilde{P}) \leq 1 - \frac{1}{2}\exp(-D_{\mathrm{KL}}(P,\tilde{P})) \leq 1 - \frac{1}{2}\exp(-5N_{T,i}(A)(\Delta(i))^2),$$

which implies that

$$N_{T,i}(A)(\Delta(i))^2 \geq \frac{1}{5}\ln\frac{T}{2\big(N_{T,i}(A) + T - N_{T,i}(\tilde{A})\big)}.$$

■

Proof [of Theorem 4] If $\tilde{A}$ is given as in Lemma 19, from (61), we have

$$R_T(A) \geq \Delta(i)N_{T,i}(A), \quad R_T(\tilde{A}) \geq \Delta(i)(T - N_{T,i}(\tilde{A})).$$

From this and Lemma 19, we have

$$(\Delta(i))^2 N_{T,i}(A) \geq \frac{1}{5}\ln\frac{T}{2\big(R_T(A)/\Delta(i) + R_T(\tilde{A})/\Delta(i)\big)}.$$

From the assumption that $R_T(\hat{A}) \leq g(m,n)T^{1-\varepsilon}$ for any $\hat{A}$, we have

$$\frac{T}{R_T(A)/\Delta(i) + R_T(\tilde{A})/\Delta(i)} \geq \frac{T}{2g(m,n)T^{1-c}/\Delta(i)} = \frac{\Delta(i)T^c}{2g(m,n)},$$

which implies

$$N_{T,i}(A) \geq \frac{1}{5(\Delta(i))^2}\ln\frac{\Delta_i T^c}{4g(m,n)}.$$

Consequently, we have

$$\begin{aligned}
&\liminf_{T\to\infty} \frac{R_T(A)}{\ln T} = \liminf_{T\to\infty} \frac{1}{\ln T}\Bigg(\sum_{\substack{i\in[m]\\ \Delta(i)>0}} N_{T,i}(A) + \sum_{\substack{j\in[n]\\ \Delta'(j)>0}} \Delta'(j) N'_{T,j}(A)\Bigg) \\
&\geq \liminf_{T\to\infty}\Bigg(\sum_{\substack{i\in[m]\\ \Delta(i)>0}} \frac{1}{5\Delta(i)} + \sum_{\substack{j\in[n]\\ \Delta'(j)>0}} \frac{1}{5\Delta'(j)}\Bigg)\Bigg(c + \frac{1}{\ln T}\ln\frac{\Delta_i}{4g(m,n)}\Bigg) \\
&= \frac{c}{5}\Bigg(\sum_{\substack{i\in[m]\\ \Delta(i)>0}} \frac{1}{\Delta(i)} + \sum_{\substack{j\in[n]\\ \Delta'(j)>0}} \frac{1}{\Delta'(j)}\Bigg),
\end{aligned}$$

which completes the proof. ■