Optimistic Thompson Sampling for No-Regret Learning in Unknown Games

Yingru Li [1] **Liangqi Liu** [1] **Wenqiang Pu** [2] **Zhi-Quan Luo** [1 2]

Abstract

We propose an Optimism-then-NoRegret learning framework for learning to play a repeated multiplayer game with an unknown reward function and bandit feedback. Our framework encompasses various game algorithms as special cases. It consists of an estimation step for constructing an imagined reward and a no-regret step for playing against an adversary. Thompson Sampling (TS) can be naturally included in the framework, but its effectiveness in this context remains unclear. We demonstrate that TS fails in a class of unknown games. To address this, we propose an optimistic variant of TS combined with suitable full-information adversarial bandit algorithms, achieving sublinear regret in the unknown game. We establish an information-theoretic regret bound for the proposed algorithms. Our analysis highlights that the optimistic variant encourages more exploration than classical TS in unknown games. We evaluate the algorithms on random matrix games and two real-world applications: radar anti-jamming and traffic routing problems. The proposed algorithms outperform baselines substantially.

1. Introduction

Many real-world problems in economics (Fainmesser, 2012), sociology (Skyrms & Pemantle, 2009), transportation (Leblanc, 1975), politics (Ordeshook et al., 1986), signal processing (Song et al., 2011), and other fields (Fudenberg & Tirole, 1991) can be described as *unknown games*, where each player only observes their opponents' actions and the noisy rewards associated with their own selected actions (referred to as *bandit feedback*). The goal of each player is to maximize their individual reward, and the only way to achieve this is to repeatedly play and learn the game structure from the corresponding observed rewards. The challenge in unknown games is how to efficiently learn from bandit feedback. Celebrated no-regret learning algorithms, such as Hedge (Freund & Schapire, 1997) and regret-matching (Hart & Mas-Colell, 2000), ensure sublinear regret guarantees under the *full information* setting, where the rewards of all actions at each round are observable. However, these algorithms cannot handle problems with bandit feedback, which is the case for many real-world problems of interest.

Learning efficiently from bandit feedback presents a significant challenge in unknown games. Traditional no-regret learning algorithms, such as Hedge (Freund & Schapire, 1997) and regret-matching (Hart & Mas-Colell, 2000), guarantee sublinear regret only under the assumption of full information, where rewards of all actions at each round are observable. However, these algorithms are ill-suited for problems with bandit feedback, which are prevalent in many real-world scenarios of interest.

Contribution: To address this challenge, we propose an Optimism-then-NoRegret (OTN) learning framework for playing unknown games. Our framework encompasses various vanilla game algorithms as well as recent works (Sessa et al., 2019; O'Donoghue et al., 2021) that have utilized the upper confidence bound (UCB) technique to exploit the bandit feedback. Besides UCB, Thompson Sampling (TS) can be naturally included in the framework, but its effectiveness in this context remains unclear. In the rest part of this paper, we specifically investigate this and demonstrate that classical TS indeed fails in a specific class of unknown games. To overcome this limitation, we further introduce an optimistic variant of TS (referred as OTS for short) combined with appropriate full information adversarial bandit algorithms. An information-theoretic regret analysis under the OTN framework is presented, which provides sublinear regret bounds of proposed algorithms. Comparison with related existing algorithms is summarized in Table 1, where $\mathcal{A}$ and $\mathcal{B}$ are the sizes of action sets, γ_T is the maximum information gain (introduced in Definition 4.4), and $\beta = \log \mathcal{A}T$.

1.1. Related Works

In the full-information setting, multiplicative-weights (MW) algorithms such as Hedge (Freund & Schapire, 1997) can

*Equal contribution [1]The Chinese University of Hong Kong, Shenzhen, China [2]Shenzhen Research Institute of Big Data. Correspondence to: Yingru Li <yingruli@link.cuhk.edu.cn>.

The Many Facets of Preference Learning Workshop at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023.

Table 1. Regret bounds comparison.

Feedback		Full	Bandit	Bandit + Actions	Bandit + Actions
Imagined Reward		–	IWE	UCB	OTS **[Ours]**
No-Regret Update	Hedge	$\mathcal{O}(\sqrt{T\log\mathcal{A}})$	$\mathcal{O}(\sqrt{T\mathcal{A}\log\mathcal{A}})$	$\mathcal{O}(\sqrt{T\log\mathcal{A}}+\sqrt{\gamma_T\beta T})$	$\mathcal{O}(\sqrt{T\log\mathcal{A}}+\sqrt{\gamma_T\beta T})$
	RM	$\mathcal{O}(\sqrt{T\mathcal{A}})$	$\mathcal{O}(T^{2/3}\mathcal{A}^{2/3})$ **[Ours]**	$\mathcal{O}(\sqrt{T\mathcal{A}}+\sqrt{\gamma_T\beta T})$ **[Ours]**	$\mathcal{O}(\sqrt{T\mathcal{A}}+\sqrt{\gamma_T\beta T})$

achieve optimal regret for adversarial bandit problems. A sequence of works (Daskalakis et al., 2011; Syrgkanis et al., 2015; Chen & Peng, 2020; Hsieh et al., 2021) study no-regret learning algorithms in games. Another common no-regret learning algorithm is regret matching (Hart & Mas-Colell, 2000). Later, it was found that a variation called regret-matching$^+$(RM$^+$) (Tammelin, 2014) leads to significantly faster convergence in practice. Full information feedback requires perfect game knowledge and is unrealistic in many applications. In the more challenging bandit setting, Exp3 (Auer et al., 2002b) is a classical algorithm that utilizes the importance-weighted estimator for reward vector construction. The online algorithms previously referenced can reduce the unknown repeated game into single-agent decision-making by treating the opponents as part of the environment. The environment in this situation is both adversarial and adaptive, where distinct reward functions are selected at every individual time step (Cesa-Bianchi & Lugosi, 2006). Previous literature ignores the fact the agents are playing repeated games in which the reward structure could be exploited. Furthermore, to the application of our interest, the opponent's actions can be observed by the agent. This scenario has received relatively little attention despite its numerous applications (O'Donoghue et al., 2021). Some papers consider a similar setting as ours. (O'Donoghue et al., 2021) seeks to compare the received reward to the Nash value and proposes variants of UCB and K-learning that converge to the Nash value. Our goal is to compete with hindsight's best actions. With this performance metric, we seek to exploit the opponent's strategy instead of only achieving Nash value. Similar to our setting, (Sessa et al., 2019) also uses adversarial regret and focuses on the single-player viewpoint. It utilizes Gaussian Process to exploit the correlations among different game outcomes and obtains a kernel-dependent regret bound with the factor γ_T by an upper confidence bound (UCB) type algorithm. Thompson Sampling (TS) and other randomized exploration variants (Russo et al., 2018; Vaswani et al., 2020) are a strong counterpart of UCB-type algorithm in the reward structure-aware bandit literature. However, no evidence shows these randomized exploration methods can work in our unknown game setting.

2. Repeated Bandit Game

To simplify the presentation, we consider a two-player game scenario involving Alice and Bob. However, our results can be extended straightforwardly to multiplayer games by treating all other players as an *abstract* player.

Protocol. Consider a repeated game between Alice and her opponent Bob, where the action index sets for Alice and Bob are denoted by $\mathcal{A} = 1,\ldots,|\mathcal{A}|$ and $\mathcal{B} = 1,\ldots,|\mathcal{B}|$, respectively.[1] At each time $t = 0,1,\ldots$, Alice selects an action $A_t \in \mathcal{A}$ and Bob simultaneously selects an action $B_t \in \mathcal{B}$. Alice then observes a reward R_{t+1,A_t,B_t} associated with the selected action pair (A_t,B_t), which takes values in $\mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$. Alice's experience up to time t is encoded by the history $H_t = \big(A_0,B_0,R_{1,A_0,B_0},\ldots,A_{t-1},B_{t-1},R_{t,A_{t-1},B_{t-1}}\big)$.

Algorithm. An algorithm $\pi^{\mathrm{alg}} = (\pi_t)_{t\in\mathbb{N}}$ employed by Alice is a sequence of deterministic functions, where each $\pi_t(H_t)$ specifies a probability distribution over the action set $\mathcal{A}$ based on the history H_t. Alice's action A_t is sampled from the distribution π_t, i.e., $\mathbb{P}(A_t \in \cdot \mid \pi_t) = \mathbb{P}(A_t \in \cdot \mid H_t) = \pi_t(\cdot)$.

Environment. The reward R_{t+1,A_t,B_t} revealed by the game environment is a corrupted noisy version of the *mean* reward function $f_\theta(A_t,B_t) : \mathcal{A}\times\mathcal{B} \mapsto [0,1]$, where θ is a random variable taking values from set Θ. The corruption noise $W_t = R_{t+1,A_t,B_t} - f_\theta(A_t,B_t)$ is assumed to be zero-mean Gaussian noise and is independent at each time t.

The above description of the bandit game encompasses various game forms based on the structure of the mean reward function $f_\theta(a,b)$. Several representative game forms such as the matrix game, linear game, and kernelized game are summarized in Table 2 (see Appendix A).

The objective of Alice is to maximize her expected reward $\sum_{t=0}^{T-1}\mathbb{E}\left[R_{t+1,A_t,B_t} \mid \theta\right]$ over some long duration T, regardless of the fixed action sequence $B_{0:T}$[2] Bob chooses. By treating Bob as the adversarial environment, with the best action $A^* = \arg\max_{a\in\mathcal{A}}\sum_{t=0}^{T-1}\mathbb{E}\left[R_{t+1,a,B_t} \mid \theta\right]$ in

[1] We use the shorthand $\mathcal{A}$ to denote the cardinality $|\mathcal{A}|$.

[2] $B_{0:T} = (B_0 = b_0, B_1 = b_1, \ldots, B_{T-1} = b_{T-1})$.

hindsight, the T-period *adversarial regret* is defined by

$$\Re(T, \pi^{\text{alg}}, B_{0:T}, \theta) = \sum_{t=0}^{T-1} \mathbb{E}\left[R_{t+1,A^*,B_t} - R_{t+1,A_t,B_t} \mid \theta\right], \tag{1}$$

where the expectation is taken over the randomness in the actions A_t and the rewards R_{t+1,A_t,B_t}. However, this adversarial regret $\Re(T, \pi^{\text{alg}}, B_{0:T}, \theta)$ is not a suitable regret metric under our game setting since it depends on the specific action sequence $B_{0:T}$ played by Bob. In this paper, we adopt the worse-case regret as the metric. An algorithm π_{alg} is considered *No-Regret* for Alice if, for any $B_{0:T}$, Alice suffers only sublinear regret, i.e.,

$$\Re^*(T, \pi^{\text{alg}}) = \sup_{B_{0:T} \in \mathcal{B}^T} \Re(T, \pi^{\text{alg}}, B_{0:T}) = o(T),$$

where θ is omitted for notation simplicity.

3. Optimism-then-NoRegret Learning

3.1. Full Information Feedback

We will start by providing a brief overview of the full information feedback setting, in which Alice can observe the mean rewards $r_t(a) = f_\theta(a, B_t)$ for all actions $a \in \mathcal{A}$ (details are presented in Appendix A). At time t, Alice picks action $A_t \sim P_{X_t}$ where $P_X(i) = X_i / \sum_{i \in \mathcal{A}} X_i$. Full-information adversarial bandit algorithms, e.g., Hedge (Freund & Schapire, 1997) and Regret Matching (RM) (Hart & Mas-Colell, 2000), can be used to update X_t to $X_{t+1} = g_t(X_t, r_t)$,

$$\begin{cases} \text{Hedge:} & g_{t,a}(X_t, r_t) = X_{t,a} \exp\left(\eta_t r_t(a)\right), \\ \text{RM:} & g_{t,a}(X_t, r_t) = \max\left(0, \sum_{s=0}^{t} r_t(a) - r_t(A_s)\right), \end{cases}$$

where $g_t(\cdot) : \mathbb{R}_+^{\mathcal{A}} \times \mathbb{R}_+^{\mathcal{A}} \mapsto \mathbb{R}_+^{\mathcal{A}}$. Since $r_t(a) = f_\theta(a, B_t)$, the adversarial regret in Equation (1) can be reformulated as the following full-information adversarial regret. For Hedge and RM (Freund & Schapire, 1997), the regrets can be bounded as $\Re_{\text{full}}(T, \text{Hedge}, (r_t)_t) = \mathcal{O}(\sqrt{T \log \mathcal{A}})$ and $\Re_{\text{full}}(T, \text{RM}, (r_t)_t) = \mathcal{O}(\sqrt{T\mathcal{A}})$.

Definition 3.1 (Regret with Full Information)**.** The full information adversarial regret of an algorithm adv for an arbitrary reward sequence $(r_t)_t$ is defined as

$$\Re_{\text{full}}(T, \text{adv}, (r_t)_t) = \max_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=0}^{T-1} r_t(a) - r_t(A_t)\right]. \tag{2}$$

3.2. Bandit Feedback

In the bandit feedback case, Alice can only utilize the information encoded in history H_t up to round t. One important idea is to use bandit feedback information to construct a sequence of full reward vectors $\tilde{R}_t \in \mathbb{R}^{\mathcal{A}}$ in an optimistic sense, which we refer to as the *imagined reward*. Then, we can use a similar update rule g_t to update the probability distribution over actions from X_t to X_{t+1}. This procedure is described in Algorithm 1, termed as Optimism-then-NoRegret (OTN).

Algorithm 1 Optimism-then-NoRegret for Bandit Feedback

Initialize X_1
for round $t = 0, 1, \ldots, T-1$ **do**
  Sample action $A_t \sim P_{X_t}$
  Observe opponent's action B_t and noisy bandit feedback R_{t+1,A_t,B_t}
  Update historical information for player i: $H_{t+1} = (H_t, A_t, B_t, R_{t+1,A_t,B_t})$
  Construct reward vector $\tilde{R}_{t+1}$ with the history and an algorithmic random draw:

$$\tilde{R}_{t+1} = E(H_{t+1}, Z_{t+1}) \in \mathbb{R}^{\mathcal{A}}, \tag{3}$$

  where Z_{t+1} is randomly drawn by the algorithm at time t, independent of history H_{t+1}
  Update: $X_{t+1} = g_t(X_t, \tilde{R}_{t+1})$
end for

The essential part in Algorithm 1 is the construction of the imagined reward vector $\tilde{R}_{t+1}$. A celebrated estimation method is *importance-weighted estimator* (IWE) (Lattimore & Szepesvári, 2020), which combined with Hedge becomes the well known Exp3 algorithm (Auer et al., 2002b) (referred as IWE-Hedge in this paper). Additionally, IWE can also be combined with RM, resulting in IWE-RM for the bandit game. Details of IWE-Hedge and IWE-RM are presented in Appendix C, and their adversarial regrets are provided below.

Proposition 3.2 (Regrets of IWE-Hedge and IWE-RM (Auer et al., 2002b))**.** *Consider play bandit game with IWE-Hedge or IWE-RM, then* $\Re^*(T, \textit{IWE-Hedge}) = \mathcal{O}(\sqrt{T\mathcal{A}\log\mathcal{A}})$.

We also extend the regret matching (RM) (Hart & Mas-Colell, 2000) to our setting under the OTN framework, resulting IWE-RM algorithm (detailed proof can be found in Section C).

Theorem 3.3 (Regret of IWE-RM)**.** *Consider play bandit game with IWE-RM, then* $\Re^*(T, \textit{IWE-RM}) = \mathcal{O}(T^{2/3}\mathcal{A}^{2/3})$.

However, IWE only requires the bandit reward feedback and does not explicitly utilize the information of opponents' action. An important idea is that, with the information of opponent's actions and the knowledge on reward structure f_θ, we can estimate $f_\theta(a, b)$ for all $(a, b) \in \mathcal{A} \times \mathcal{B}$ by

$\tilde{f}_{t+1}(a,b)$ at each time t, and construct imagined reward by $\tilde{R}_{t+1}(a) = \min(\tilde{f}_{t+1}(a,B_t),1)$. By taking Gaussian distribution as the prior, the mean and variance estimation of $f_\theta(a,b)$ up to time t can be recursively updated, denoted as $\mu_t(a,b)$ and $\sigma_t^2(a,b)$. Detailed updating formulas for games in Table 2 are presented in Appendix B. The remaining issue is how to balance exploration and exploitation based on mean and variance estimation.

In bandit environments, the UCB (Auer et al., 2002a) and TS (Thompson, 1933) techniques are widely used to address the exploration-exploitation trade-off. Tailoring to the bandit game considered in this paper, the corresponding constructions using UCB and TS are as follows:

$$\begin{cases} \text{UCB:} & \tilde{f}_{t+1}(a,B_t) \mid H_{t+1} = \mu_t(a,B_t) + \beta_t\sigma_t(a,B_t), \\ & \tilde{R}_{t+1}(a) = \tilde{f}_{t+1}(a,B_t) \wedge 1, \forall a \in \mathcal{A}. \\ \text{TS:} & \tilde{f}_{t+1}(a,B_t) \mid H_{t+1} \sim N(\mu_t(a,B_t), \sigma_t(a,B_t)), \\ & \tilde{R}_{t+1}(a) = \tilde{f}_{t+1}(a,B_t) \wedge 1, \forall a \in \mathcal{A}. \end{cases}$$

The effectiveness of UCB combined with Hedge has been studied in (Sessa et al., 2019). Within the OTN framework, UCB can also be combined with RM, resulting in a sublinear regret based on our analysis (see details in Section 4). The regrets of UCB are provided below:

Proposition 3.4 (Regrets of UCB-Hedge (Sessa et al., 2019) and UCB-RM)**.** *Consider playing the bandit feedback game with UCB-Hedge or UCB-RM, then* $\Re^*(T, \textit{UCB-Hedge}) = \mathcal{O}(\sqrt{T\log\mathcal{A}} + \sqrt{\gamma_T\beta T})$, $\Re^*(T, \textit{UCB-RM}) = \mathcal{O}(\sqrt{T\mathcal{A}} + \sqrt{\gamma_T\beta T})$.

3.3. Failure of TS Estimator

Besides UCB, TS can be naturally included in the OTN framework, but its effectiveness in this context remains unclear. We demonstrate that TS combined with RM fails in a class of bandit games.

Example 3.1 (Matrix Game with Best Response Player)**.** *Consider a class of matrix games with a payoff matrix θ defined as*

$$\theta = \begin{bmatrix} 1 & 1-\Delta \\ 1-\Delta & 1 \end{bmatrix}, \tag{4}$$

where $\Delta \in (0,1)$. At time t, Alice plays action $A_t \sim P_{X_t}$, and Bob is a best response player who can observe X_t and play the best response strategy $Y_t = \max_y X_t^\top \theta y$ by selecting action $B_t \sim P_{Y_t}$. Alice receives the noiseless reward $R_{t+1,A_t,B_t} = \theta_{A_t,B_t}$.

Intuitively, as long as Alice chooses pure strategy, Bob will exploit Alice, resulting in Δ regret for Alice in that round. This divergence fact is formally stated below.

Proposition 3.5 (Divergence)**.** *Let Ω_t denote the event where Alice selects the 2nd row and Bob chooses the 1st column for all time steps $t' \leqslant t$. Assuming uniform initialization for Alice, then $\Re(T) \geqslant 2\mathbb{P}(\Omega_T)\Delta \cdot T$. If the event Ω_t occurs with a constant probability for all $t \geqslant 1$, then Alice experiences linear regret with a constant probability.*

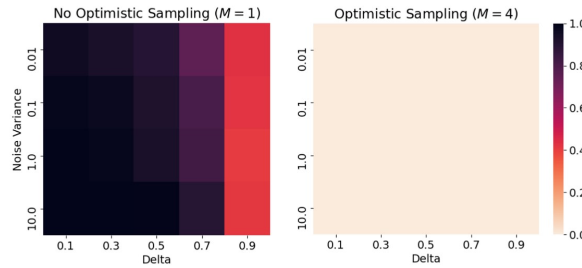


Figure 1. Failure probability for TS-RM and OTS-RM.

As shown in Figure 1, TS-RM indeed fails in different setups for hyperparameter (Δ, σ_n^2). Detailed mathematical justification can be found in Appendix D.

Proposition 3.6 (Failure of TS-RM)**.** *Suppose Alice initializes with a uniform strategy and utilizes Regret Matching with TS estimator (TS-RM). For any $\Delta \in (0,1)$ and $\sigma_n > 0$, there exists a constant $c(\Delta,\sigma_n) > 0$ such that for all rounds $t \geqslant 1$, $\mathbb{P}(\Omega_t) \geqslant c(\Delta,\sigma_n)$.* [3]

3.4. A Simple Fix: Optimistic Sampling

By taking the maximum of multiple functions sampled from the posterior, we can improve the probability of being optimistic for the estimator. At time t, after observing $H_{t+1} = (H_t, A_t, B_t, R_{t+1,A_t,B_t})$, we can perform optimistic sampling by first sampling M_{t+1} independent normal random variables $z_{t+1}^1, \ldots, z_{t+1}^j, \cdots, z_{t+1}^{M_{t+1}} \sim N(0,1)$, and constructing the estimator: $\forall a \in \mathcal{A}$

$$\tilde{f}_{t+1}^{\mathrm{OTS}}(a,B_t) := (\max_{j\in[M]} z_{t+1}^j)\cdot\sigma_t(a,B_t) + \mu_t(a,B_t),$$

$$\tilde{R}_{t+1}^{\mathrm{OTS}}(a) := \mathrm{clip}_{[0,1]}\left(\tilde{f}_{t+1}^{\mathrm{OTS}}(a,B_t)\right).$$

To intuitively understand why optimistic samples help, let us recall Example 3.1. Denote $\tilde{R}_{t+1}^{OTS}(1st)$ and $\tilde{R}_{t+1}^{OTS}(2nd)$ as the optimistic reward for the first and second row actions, respectively. Then the probability of $\tilde{R}_{t+1}^{OTS}(1st) > \tilde{R}_{t+1}^{OTS}(2nd)$ increases as M_{t+1} grows, this implies a decay of $\mathbb{P}(\Omega_t)$ over time t. Detailed mathematical justification can be found in Appendix D. Such a simple optimistic version of TS (referred to as OTS) has the following regret guarantee.

Theorem 3.7 (Regrets of OTS-Hedge and OTS-RM)**.** *For any full information adversarial bandit* adv *working with the imagined reward sequence $\tilde{R}^{\mathrm{est}} = (\tilde{R}_{t+1}, t = 0,1,\ldots)$ constructed by OTS with $M_1 = M_2 = \ldots = M_T = M = \mathcal{O}(\log\mathcal{A}\mathcal{B}T)$, the combined algorithm $\pi = \pi^{\text{adv-OTS}}$ enjoys the regret bound:*

$$\Re^*(T,\pi,\theta) \leqslant \Re_{\mathrm{full}}(T,\mathrm{adv},\tilde{R}^{\mathrm{est}}) + \sqrt{\log(\mathcal{A}T)I(\theta;H_T)T}$$

[3] If $\Delta = 0.1$ and $\sigma_n = 0.1$, we have $c(\Delta,\sigma_n) \approx 0.54$.

where $\mathfrak{R}_{\text{full}}$ is defined in Definition 3.1.

Note that as long as the imagined $\tilde{R}^{\text{est}} = (\tilde{R}_{t+1}, t = 0, 1, \ldots)$ satisfies bounded property $\tilde{R}_t \in [0,1]^{\mathcal{A}}$, then if we take adv to be Hedge, $\mathfrak{R}_{\text{full}}(T, \text{Hedge}, \tilde{R}^{\text{est}}) = \mathcal{O}(\sqrt{T \log \mathcal{A}})$. And if taking adv to be RM, then $\mathfrak{R}_{\text{full}}(T, \text{RM}, \tilde{R}^{\text{est}}) = \mathcal{O}(\sqrt{T\mathcal{A}})$. In addition, Theorem 3.7 also applies to IWE and UCB, resulting in our by-product regret bounds of IWE-RM and UCB-RM.

4. Regret Analysis

Our analysis road map is as follows. First, we derive a general regret decomposition given any imagined reward sequence $\{\tilde{R}_t : t \in \mathbb{Z}_{++}\}$, where each $\tilde{R}_{t+1}$ is constructed using history information H_t with algorithmic randomness.

Proposition 4.1 (Regret decomposition)**.** *For any $a \in \mathcal{A}$, the one-step regret can be decomposed by*

$$\mathbb{E}\left[R_{t+1,a,B_t} - R_{t+1,A_t,B_t} \mid \theta\right] = (I) + (II) + (III)$$

where

$$\begin{aligned}(I) &= \mathbb{E}\left[\tilde{R}_{t+1}(a) - \tilde{R}_{t+1}(A_t) \mid \theta\right], \\ (II) &= \mathbb{E}\left[f_\theta(a, B_t) - \tilde{R}_{t+1}(a) \mid \theta\right], \\ (III) &= \mathbb{E}\left[\tilde{R}_{t+1}(A_t) - f_\theta(A_t, B_t) \mid \theta\right].\end{aligned}$$

Summation of (I) reduces to adversarial regret $\mathfrak{R}_{full}(T, \text{adv}, \tilde{R})$ for a bounded sequence $\tilde{R}_{t+1}$. For the term (II), the following definition is a sufficient condition for $\sqrt{T}$-type regret.

Definition 4.2 (Sufficient optimism)**.** We say the constructed imagined reward sequence $\tilde{R}^{\text{est}} = (\tilde{R}_t, t \in \mathbb{Z}_{++})$ is optimistic if for any action $a \in \mathcal{A}$, $\mathbb{P}(f_\theta(a, B_t) \geqslant \tilde{R}_{t+1}(a) \mid \theta) \leqslant \mathcal{O}(1/\sqrt{T})$.

OTS can satisfy Definition 4.2 by selecting proper M_{t+1} whereas TS cannot satisfy. UCB sequence can also satisfy Definition 4.2. The term (III) can be bounded by one-step information gain $I(\theta; R_{t+1,A_t,B_t} | H_t)$ using differential entropy of Gaussian distribution. Here, we give a general regret bound for any imagined reward sequence.

Theorem 4.3. *For any full information adversarial bandit* adv *working the imagined reward $\tilde{R}^{\text{est}} = (\tilde{R}_t, t \in \mathbb{Z}_{++})$ constructed by estimation algorithm* est*, if the $\tilde{R}^{\text{est}}$ satisfy Definition 4.2, the combined algorithm $\pi = \pi^{\text{adv-est}}$ enjoys the regret*

$$\mathfrak{R}^*(T, \pi, \theta) \leqslant \mathfrak{R}_{\text{full}}(T, \text{adv}, \tilde{R}) + \sqrt{\beta I(\theta; H_T) T}$$

where $\mathfrak{R}_{\text{full}}$ is defined in Definition 3.1 and $\beta = \mathcal{O}(\log \mathcal{A}T)$.

To characterize how much uncertainty reduction is in a particular game structure when observing new information, we define the maximum information gain.

Definition 4.4 (Maximum Information Gain)**.**

$$\gamma_T := \max_{A_{0:T}, B_{0:T}} I(\theta; A_0, B_0, \ldots, A_{T-1}, B_{T-1})$$

where $I(X;Y)$ is the mutual information between random variables X and Y.

Remark 4.5. An important property of the Gaussian distribution is that the information gain does not depend on the observed rewards. This is because the posterior covariance of a multivariate Gaussian is a deterministic function of the sampled points. For this reason, this maximum information ratio γ_T in Definition 4.4 is well defined. That is, $I(\theta; H_T) = I(\theta; A_0, B_0, \ldots, A_{T-1}, B_{T-1}) \leqslant \gamma_T$.

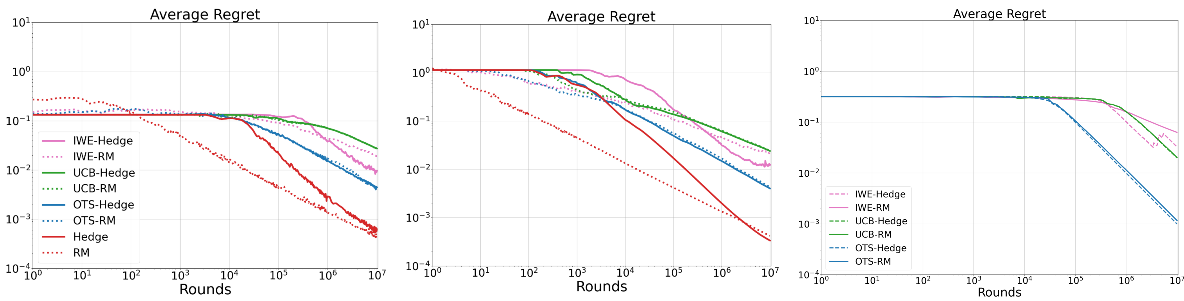
We adopt the results from (Srinivas, Krause, Kakade, and Seeger, 2009) which gives the bounds of γ_T for a range of commonly used covariance functions: finite dimensional linear, squared exponential and Matern kernels, whose details can be found in Appendix E. Utilizing the reward structure can be used to resolve the curse of multi-agent. For example, if the reward structure can be model by squared exponential kernels, the final regret of OTS-Hedge is $\mathcal{O}((\sqrt{\log \mathcal{A}} + \sqrt{\log(\mathcal{A}T)\log(T)^{d+1}})\sqrt{T})$, which has no polynomial dependence on action sizes $\mathcal{A} \times \mathcal{B}$, where $|\mathcal{B}|$ is exponential in the number of opponents.

5. Numerical Studies and Applications

Evaluation of the proposed algorithms on random matrix games and two real-world applications are studied. As baseline algorithms, Hedge and RM, which require full information, are utilized. IWE-Hedge and IWE-RM are used as baselines in the bandit feedback setting. Under the OTN learning framework (Algorithm 1), a total of four algorithms are compared with baselines: OTS-Hedge, OTS-RM, UCB-Hedge, and UCB-RM. Note that GP-MW (Sessa et al., 2019) is UCB-Hedge named here. Average expected regret is utilized as the performance metric. Appendix F.1 provides a detailed definition of the performance metrics and algorithm settings.

5.1. Random Matrix Games

This section evaluates different algorithms on two-player zero-sum matrix games. Each payoff matrix entry is an i.i.d. random variable generated from the uniform distribution $[-1, 1]$, and each player has M actions (payoff matrix is a squared matrix of size M). The total number of rounds is $T = 10^7$. At each round, the two players receive noisy rewards $\pm\tilde{r}_t$, where $\tilde{r}_t = A_{ij} + \epsilon_t$ and $\epsilon_t \sim \mathcal{N}(0, 0.1)$. Different matrix sizes are considered, i.e., $M = 10, 50, 70, 100$, and for each choice of M, 9 independent simulation runs are conducted. The performance, averaged over 9 simulation runs with different opponent models (see details in Appendix F.2) are compared.



(a) Self-play (b) Best-resp. oppo. (c) Stat.oppo.

Figure 2. Averaged regrets for different opponents on 70×70 matrix game.

Convergence curves of the average regret for $M = 70$ are compared in Fig.2a. Under the self-play setting, OTS-based algorithms converge faster than UCB-based algorithms. Furthermore, the average regret of RM decreases earlier than that of Hedge. For a best-response opponent (Fig.2b), all the algorithms in our proposed algorithmic framework outperform the IWE baselines. For a stationary opponent (Fig.2c), results clearly demonstrate that the OTS estimator brings considerable benefits in exploiting this weak opponent, compared to the IWE-based estimators. Comparison with a non-stationary opponent can be found in Appendix F.3, which shows OTS is more robust than IWE.

5.2. Anti-jamming Problem: Linear Game

The anti-jamming problem is an important issue in signal processing literature (Song et al., 2011), which can be formulated as a non-cooperative game between a radar and a jammer. This competition can be modeled in the frequency domain as a linear game, with the signal-to-inference-plus-noise ratio (SINR) serving as the reward function $\text{SINR}(a, b; \theta)$, further information on the anti-jamming game setting can be found in Appendix F.5. We compare the average regret of different algorithms against an adaptive jammer, where the jammer takes action based on the radar's latest 10 actions (see details in Appendix F.5). As illustrated in Fig.3, the OTS-based and UCB-based algorithms outperform the IWE-based algorithms.

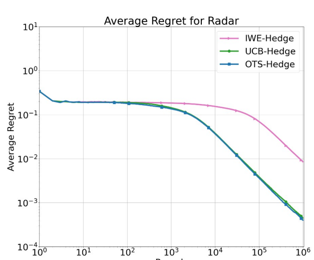


Figure 3. Play against an adaptive jammer.

5.3. Repeated Traffic Routing: Kernelized Game

In this subsection, we consider the traffic routing problem from the transportation literature, which can be modeled as a multi-player game over a directed graph. Each node pair in the graph represents an individual player, and each player seeks to find the best route to send a fixed units from its origin node to its destination node. The travel time serves as the reward, depending on the traversed edges' total occupancy. If one edge is occupied by more players, it incurs more travel time. Each player's action set is the available route in the graph, and the negative travel time of the route is the reward. The Sioux-Falls road network dataset (Bar-Gera, 2015) is used in our experiment and details can be found in Appendix F.6. Fig. 4 demonstrates that OTS-based and UCB-based algorithms outperform the IWE-based algorithms. The three proposed new algorithms, OTS-Hedge, OTS-RM, and UCB-RM, outperform UCB-Hedge (Sessa et al., 2019).

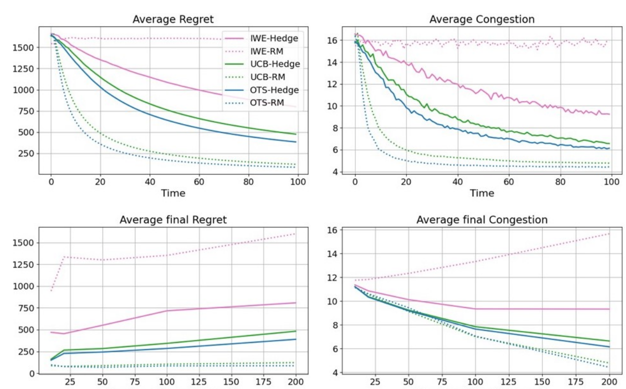


Figure 4. Average regret and congestion in traffic routing problem.

6. Conclusion

In summary, we present an OTN learning framework for playing unknown games, which includes several game algorithms as special cases. To address the limitations of TS, we introduce an optimistic variant of TS that explores the unknown game effectively. The proposed algorithms demonstrate significant performance improvements in both synthetic and real-world scenarios.

References

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

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

Bar-Gera, H. Transportation network test problems (2002). *URL http://www. bgu. ac. il/˜ bargera/tntp*, 2015.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Chen, X. and Peng, B. Hedging in games: Faster convergence of external and swap regrets. *Advances in Neural Information Processing Systems*, 33:18990–18999, 2020.

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

Fainmesser, I. P. Community structure and market outcomes: A repeated games-in-networks approach. *American Economic Journal: Microeconomics*, 4(1):32–69, 2012.

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

Fudenberg, D. and Tirole, J. *Game theory*. MIT press, 1991.

Gordon, R. D. Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941.

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

Hsieh, Y.-G., Antonakopoulos, K., and Mertikopoulos, P. Adaptive learning in continuous games: Optimal regret bounds and convergence to nash equilibrium. In *Conference on Learning Theory*, pp. 2388–2422. PMLR, 2021.

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

Leblanc, L. J. An algorithm for the discrete network design problem. *Transportation Science*, 9(3):183–199, 1975.

Ordeshook, P. C. et al. Game theory and political theory. *Cambridge Books*, 1986.

O'Donoghue, B., Lattimore, T., and Osband, I. Matrix games with bandit feedback. In *Uncertainty in Artificial Intelligence*, pp. 279–289. PMLR, 2021.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Sessa, P. G., Bogunovic, I., Kamgarpour, M., and Krause, A. No-regret learning in unknown games with correlated payoffs. *Advances in Neural Information Processing Systems*, 32, 2019.

Skyrms, B. and Pemantle, R. A dynamic model of social network formation. In *Adaptive networks*, pp. 231–251. Springer, 2009.

Song, X., Willett, P., Zhou, S., and Luh, P. B. The mimo radar and jammer games. *IEEE Transactions on Signal Processing*, 60(2):687–699, 2011.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Syrgkanis, V., Agarwal, A., Luo, H., and Schapire, R. E. Fast convergence of regularized learning in games. *Advances in Neural Information Processing Systems*, 28, 2015.

Tammelin, O. Solving large imperfect information games using cfr+. *arXiv preprint arXiv:1407.5042*, 2014.

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

Vaswani, S., Mehrabian, A., Durand, A., and Kveton, B. Old dog learns new tricks: Randomized ucb for bandit problems. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1988–1998. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/vaswani20a.html.

A. Description of Games and Full Information Game

In this work, we consider three representative game forms: the matrix game, linear game, and kernelized game, as summarized in Table 2. Further details can be found in Appendix A.

Table 2. Game Examples.

	Matrix Game	Linear Game	Kernelized Game
Mean Reward	$f_\theta(a,b)=\theta_{a,b}$ $\theta\in\mathbb{R}^{\mathcal{A}\times\mathcal{B}}$	$f_\theta(a,b)=\phi(a,b)^\top\theta$ $\phi(a,b)\in\mathbb{R}^d$, $\theta\in\mathbb{R}^d$	$f_\theta(a,b)=\theta(a,b)$ $\theta(a,b)\in\mathbb{R}$ is mean of a Gaussian Process

A.1. Different Games

Example A.1 (Matrix games)**.** *In a matrix game, the reward function simplifies to $f_\theta(a,b)=\theta_{a,b}$. In this degenerate setting, θ can be considered as the utility matrix for Alice.*

Example A.2 (Linear games)**.** *In a linear game, a known feature mapping $\phi:\mathcal{A}\times\mathcal{B}\mapsto\mathbb{R}^d$is defined, and the mean reward function is given by $f_\theta(a,b)=\phi(a,b)^\top\theta$, where the reward is linear in the feature. We assume that the random parameter θ follows the normal distribution $N(\mu_p,\Sigma_p)$, and the reward noise $W_{t+1}=R_{t+1,A_t,B_t}-f_\theta(A_t,B_t)$ is normally distributed with with mean zero and variance σ_w^2, independent of (H_t,A_t,B_t,θ).*

In the Section 5.3, the reward structure in the repeated traffic routing problem is modeled using a kernel function, which is referred to as Kernelized games.

Example A.3 (Kernelized games)**.** *In kernelized games, we consider the case where the reward function f_θ is sample from a Gaussian process. The stochastic process $(f_\theta(a,b):(a,b)\in\mathcal{A}\times\mathcal{B})$ follows a multivariate Gaussian distribution, where the mean function is denoted as $\mu(a,b)=\mathbb{E}\left[f_\theta(a,b)\right]$ and covariance (or kernel) function is denoted as $k((a,b),(a',b'))=\mathbb{E}\left[(f_\theta(a,b)-\mu(a,b))(f_\theta(a',b')-\mu(a',b'))\right]$. The kernel function $k((a,b),(a',b'))$ measures the similarity between different action pairs $(a,b),(a',b')\in\mathcal{A}\times\mathcal{B}$ in the game. We assume that the function f_θ is sampled from a Gaussian process prior $GP(0,k((a,b),(a',b')))$, the reward noise $W_{t+1}=R_{t+1,A_t,B_t}-f_\theta(A_t,B_t)$ is independent of (H_t,θ,A_t,B_t), and $(W_t:t\in\mathbb{Z}_{++})$ is an i.i.d sequence following $N(0,\sigma_w^2)$.*

A.2. Full Information Feedback

To introduce the proposed Optimism-then-NoRegret learning framework, we first consider the full information feedback setting where Alice can observe the mean rewards $r_t(a)=f_\theta(a,B_t)$ for all actions $a\in\mathcal{A}$. In this case, the problem can be solved using full-information adversarial bandit algorithms such as Hedge (Freund & Schapire, 1997) and Regret Matching (RM) (Hart & Mas-Colell, 2000) applied to the sequence of adversarial reward vectors $(r_t)_{t\in[T]}\in[0,1]^{\mathcal{A}\times T}$. The procedure is summarized in Algorithm 2, where P_X denotes probability simplex proportional to X, and function $g_t:\Delta^{\mathcal{A}}\times[0,1]^{\mathcal{A}}\mapsto\Delta^{\mathcal{A}}$ in round t is specified as follows:

$$\text{Hedge: } g_{t,a}(X_t,r_t)=X_{t,a}\exp(\eta_t r_t(a)),\quad \text{RM: } g_{t,a}(X_t,r_t)=\max\left(0,\sum_{s=0}^{t}r_t(a)-r_t(A_s)\right).$$

In the full information setting, where $r_t(a)=f_\theta(a,B_t)$, the adversarial regret defined in Equation (1) translates to the following full information adversarial regret:

Definition A.1 (Regret with Full Information)**.** The full information adversarial regret of algorithm adv for arbitrary reward sequence $(r_t)_t$ is defined as

$$\Re_{\text{full}}(T,\mathrm{adv},(r_t)_t)=\max_{a\in\mathcal{A}}\mathbb{E}\left[\sum_{t=0}^{T-1}r_t(a)-r_t(A_t)\right] \tag{5}$$

The following proposition provides the regret bounds for Hedge and RM algorithms in the full information feedback setting (Freund & Schapire, 1997):

Proposition A.2 (Regrets of Hedge and RM). *Consider playing in a full information feedback game with Hedge or RM algorithms. The regrets are bounded as follows:*

$$\Re_{\text{full}}(T, \text{Hedge}, (r_t)_t) = \mathcal{O}(\sqrt{T \log \mathcal{A}}),\ \Re_{\text{full}}(T, \text{RM}, (r_t)_t) = \mathcal{O}(\sqrt{T\mathcal{A}}).$$

Algorithm 2 No Regret Update for Full-Information Feedback

1: Initialize X_1
2: **for** round $t = 0, 1, \dots, T-1$ **do**
3: Sample action $A_t \sim P_{X_t}$,
4: Observe full-information feedback $f_\theta(a, B_t)$ for all $a \in \mathcal{A}$,
5: Update: $X_{t+1} = g_t(X_t, (f_\theta(a, B_t))_{a \in \mathcal{A}})$.
6: **end for**

B. Updating Rules of Different Games

Posterior distribution for Linear Gaussian model (Parametric). Let's consider the linear Gaussian model with a Gaussian prior $N(\mu_p, \Sigma_p)$ and noise likelihood $N(0, \sigma_w^2)$. Here are the key aspects of the model:

- Prior Assumptions: We assume a zero-mean prior with covariance $\Sigma_p = \sigma_p I$, where σ_p is a scalar parameter satisfying $\sigma_p \leqslant 1$.
- Feature Map Assumptions: We assume that the feature map $\phi(a, b)$ satisfies $|\phi(a, b)| \leqslant 1$.
- Covariance Matrix Update: Given the initial covariance matrix $\Sigma_0 = \Sigma_p$, the covariance matrix at time $t+1$ is updated as:

$$\Sigma_{t+1} = \left(\Sigma_t^{-1} + \frac{1}{\sigma_w^2}\phi(A_t, B_t)\phi(A_t, B_t)^\top\right)^{-1}$$

- Mean Vector Update: Given the initial mean vector $\mu_0 = \mu_p$, the mean vector at time $t+1$ is updated as:

$$\mu_{t+1} = \Sigma_{t+1}\left(\Sigma_t^{-1}\mu_t + \frac{R_{t+1, A_t, B_t}}{\sigma_w^2}\phi(A_t, B_t)\right)$$

- $\sigma_t(a, b) = \|\phi(a, b)\|_{\Sigma_t}$ and $\mu_t(a, b) = \phi(a, b)^\top \mu_t$
- In the case where the feature $\phi(a, b) = e_{a,b}$ is a one-hot vector, we denote $n_t(a, b)$ as the counts of occurrences of (a, b) up to time t, the posterior variance is given by:

$$\sigma_t(a, b) = \sqrt{\frac{\sigma_w^2}{\sigma_w^2/\sigma_p^2(a, b) + n_t(a, b)}}$$

Posterior distribution for Gaussian Process (Non-paramatric). In the non-parametric case of a Gaussian Process (GP), the posterior distribution remains Gaussian as well. Here are the relevant details:

- Notation: We define the vector $\mathbf{k}_t((a, b))$ and $\mathbf{R}_t$, and the matrix $\mathbf{K}_t$ as follows:

$$\begin{aligned}
\mathbf{k}_t(a, b) &= [k((A_0, B_0), (a, b)), \dots, k((A_{t-1}, B_{t-1}), (a, b))]^\top \\
\mathbf{R}_t &= [R_{1, A_0, B_0}, \dots, R_{t, A_{t-1}, B_{t-1}}]^\top \\
\mathbf{K}_t(i, j) &= k((A_i, B_i), (A_j, B_j))
\end{aligned}$$

- Variance Assumption: We assume that the variance satisfies $k(x, x) \leqslant 1$ for all $x \in \mathcal{X}$.

- Posterior Variance: The posterior variance at time t is given by: $\sigma_t^2(a,b) = k((a,b),(a,b)) - \mathbf{k}_t((a,b))^\top(\mathbf{K}_t + \sigma^2 \boldsymbol{I}_t)\mathbf{k}_t(a,b)$
- Posterior Mean: The posterior mean at time t is given by: $\mu_t(a,b) = \mathbf{k}_t((a,b))^\top(\mathbf{K}_t + \sigma^2 \boldsymbol{I}_t)^{-1}\mathbf{R}_t$
- Relationship to Linear Gaussian Model: If the kernel $k((a,b),(a,b)) = \phi(a,b)^\top \Sigma_p \phi(a,b)$ is composed of basis functions, the GP reduces to the linear Gaussian model with a prior covariance matrix Σ_p. This ensures coherence between the linear Gaussian model and the kernel model assumptions.

C. Details for Importance Weighted Estimator in Section 3

Importance-weighted estimator. For any measurable function h and probability distribution $(X_a)_{a\in\mathcal{A}}$ over a finite support $\mathcal{A}$, we construct importance weighted estimator

$$\tilde{h}(a) = \mathbb{I}_{A=a}\frac{h(A)}{X_a}, \forall a \in \mathcal{A}$$

which is an unbiased estimator:

$$\mathbb{E}\left[\tilde{h}(a)\right] = \mathbb{E}\left[\mathbb{I}_{A=a}\right] h(a)/X_a = h(a).$$

Exp3: Hedge with importance weighted estimator In this work, we sometimes call Exp3 as IWE-Hedge. The celebrated Exp3 algorithm construct an estimate of reward vector as

$$\tilde{R}_{t+1}(a) = 1 - \frac{\mathbb{I}_{A_t=a}(1 - R_{t+1,A_t,B_t})}{X_{t,a}}$$

We can observe that $\tilde{R}_{t+1}(a)$ is unbiased conditioned on history H_t. Given that X_t is H_t-measurable and A_t is conditionally independent with W_{t+1} and B_t given H_t, and using the fact $\mathbb{I}_{A_t=a}R_{t+1,A_t,B_t} = \mathbb{I}_{A_t=a}R_{t+1,a,B_t}$, we have:

$$\mathbb{E}_t\left[\tilde{R}_{t+1}(a)\right] = 1 - \mathbb{E}_t\left[\mathbb{I}_{A_t=a}\frac{1 - R_{t+1,a,B_t}}{X_{t,a}}\right] = 1 - \mathbb{E}_t\left[\mathbb{I}_{A_t=a}\right]\frac{1 - \mathbb{E}_t\left[f_\theta(a,B_t)\right]}{X_{t,a}} = \mathbb{E}_t\left[f_\theta(a,B_t)\right].$$

Exp3(Auer et al., 2002b) updates the strategy using $X_{t+1,a} \propto X_{t,a}\exp(\eta_t \tilde{R}_{t+1}(a))$.

Regret Matching with importance weighted estimator. Using the importance-weighted estimator, we can obtain an unbiased estimator for the regret $\Re_t \in \mathbb{R}^{\mathcal{A}}$ at round t:

$$\tilde{\Re}_{t,a} = \frac{\mathbb{I}_{A_t=a}R_{t+1,A_t,B_t}}{X_{t,a}} - R_{t+1,A_t,B_t}\frac{\hat{X}_{t,A_t}}{X_{t,A_t}} \tag{6}$$

and update the strategy as follows: Let the cumulative estimated reward be $\tilde{C}_{t,a} = \sum_{s=0}^{t}\tilde{\Re}_{s,a}$,

$$\hat{X}_{t+1,a} = \begin{cases} \tilde{C}^+_{t,a}/\sum_{a\in\mathcal{A}}\tilde{C}^+_{t,a}, & \text{if } \sum_{a\in\mathcal{A}}\tilde{C}^+_{t,a} > 0, \\ \text{arbitrary vector on simplex, e.g. } 1/\mathcal{A}, & \text{otherwise} \end{cases} \tag{7}$$

Here, the sampling distribution $X_{t,a}$ is mixed with uniform distribution

$$X_{t,a} = (1-\gamma_t)\hat{X}_{t,a} + \gamma_t(1/\mathcal{A}), \forall a \in \mathcal{A} \tag{8}$$

The detailed algorithm for Importance-weighted estimator Regret Matching (IWE-RM) is as follows:

Algorithm 3 Importance-weighted estimator with Regret Matching (IWE-RM)

Input: init $X_1 = \hat{X}_1$ as uniform probability vector over $\mathcal{A}$ and sequence $(\gamma_t)_{t\geqslant 0}$
for round $t = 0, 1, \ldots, T-1$ **do**
  Sample action $A_t \sim P_{X_t}$
  Observe noisy bandit feedback R_{t+1,A_t,B_t}.
  Construct regret estimator $\tilde{\Re}_t$ with importance weighted estimation by Equation (6).
  Update $\hat{X}_{t+1}$ and X_{t+1} with Equations (7) and (8) and γ_t.
end for

Fact C.1. *Importance weighted estimator $\tilde{R}_{t+1}$ at round t is $\sigma(H_t, A_t, R_{t+1,A_t,B_t})$-measurable.*

Remark C.1. For any $t \in \mathbb{N}$, the imagined reward vector $\tilde{R}_{t+1}$ constructed by importance weighted estimator satisfies $\mathbb{E}\left[\tilde{R}_{t+1}(a) \mid H_t, \theta\right] = \mathbb{E}\left[f_\theta(a, B_t) \mid H_t, \theta\right] = \mathbb{E}\left[g_\theta(e_a, Y_t) \mid H_t, \theta\right]$ for any $a \in \mathcal{A}$. Therefore, we have $\mathbb{E}\left[\text{pess}_{t+1} \mid \theta\right] = \mathbb{E}\left[\text{est}_{t+1} \mid \theta\right] = 0$.

Lemma C.2. *For all real a, define $a^+ = \max\{a, 0\}$. For all a, b, it is the case that*

$$\left((a+b)^+\right)^2 \leqslant (a^+)^2 + 2(a^+)b + b^2$$

Proof. $(a+b)^+ \leqslant (a^+ + b)^+ \leqslant |a^+ + b|$. □

Lemma C.3. *For all vector $v \in \mathbb{R}^{\mathcal{A}}$, define $v^+ = (v_a^+)_{a\in\mathcal{A}}$. Following Algorithm 3, we have the important observation*

$$\left\langle \tilde{C}_{t-1}^+, \tilde{\mathfrak{R}}_t \right\rangle \leqslant 0$$

Proof. Suppose at round t, Alice choose $A_t \sim P_{X_t}$ and receive the feedback R_{t+1,A_t,B_t}. By algorithm 3 and Equations (6) to (8), If $\sum_a \tilde{C}_{t-1,a}^+ \leqslant 0$, then obviously $\tilde{C}_{t-1,a}^+ = 0$ for all action $a \in \mathcal{A}$. Then, the lemma trivially holds.

Otherwise, we have

$$\begin{aligned}
\left\langle \tilde{C}_{t-1}^+, \tilde{\mathfrak{R}}_t \right\rangle &= R_{t+1,A_t,B_t}\left(\frac{\tilde{C}_{t-1,A_t}^+}{X_{t,A_t}} - \frac{\hat{X}_{t,A_t}}{X_{t,A_t}}\sum_a \tilde{C}_{t-1,a}^+\right) \\
&= R_{t+1,A_t,B_t}\left(\frac{\tilde{C}_{t-1,A_t}^+}{X_{t,A_t}} - \frac{\tilde{C}_{t-1,A_t}^+/\sum_a \tilde{C}_{t-1,a}^+}{X_{t,A_t}}\sum_a \tilde{C}_{t-1,a}^+\right) = 0
\end{aligned}$$

□

Lemma C.4. *Following algorithm 3, we have an important inequality*

$$\sum_a \left(\tilde{C}_{T,a}^+\right)^2 \leqslant \sum_{t=1}^T \sum_a \left(\tilde{\mathfrak{R}}_{t,a}\right)^2$$

Proof. Since from the update rule,

$$\tilde{C}_{T,a} = \tilde{C}_{T-1,a} + \tilde{\mathfrak{R}}_{T,a},$$

by Lemma C.2 and C.3:

$$\begin{aligned}
\sum_a \left(\tilde{C}_{T,a}^+\right)^2 &\leqslant \sum_a \left(\left(\tilde{C}_{T-1,a}^+\right)^2 + 2\tilde{C}_{T-1,a}^+\tilde{\mathfrak{R}}_{T,a} + \left(\tilde{\mathfrak{R}}_{T,a}\right)^2\right) \\
&\leqslant \sum_a \left(\tilde{C}_{T-1,a}^+\right)^2 + \sum_i (\tilde{\mathfrak{R}}_{T,a})^2
\end{aligned}$$

By telescoping series, we conclude the lemma. □

Proof of Theorem 3.3. Notice that the action A_t selected by Alice and the action B_t selected by Bob is independent conditioned on history H_t and θ. According to the algorithm 3, the conditional expectation of the estimated immediate regret is

$$\mathbb{E}\left[\tilde{\mathfrak{R}}_{t,a} \mid \theta, H_t\right] = \mathbb{E}\left[f_\theta(a, B_t) \mid \theta, H_t\right] - \sum_a \hat{X}_{t,a}\mathbb{E}\left[f_\theta(a, B_t) \mid \theta, H_t\right]$$

Step 1 (Bounding the bias of estimated regret.) Recall the definition of immediate regret at time t conditioned on history is

$$\Re_t(a) := \mathbb{E}\left[R_{t+1,a,B_t} - R_{t+1,A_t,B_t} \mid \theta, H_t\right] = \mathbb{E}\left[f_\theta(a, B_t) \mid \theta, H_t\right] - \sum_a X_{t,a}\mathbb{E}\left[f_\theta(a, B_t) \mid \theta, H_t\right]$$

In the following, we use short notation $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid H_t, \theta]$. For any $a \in \mathcal{A}$, the difference with the estimated regret under conditional expectation is

$$\begin{aligned}
\Re_t(a) - \mathbb{E}_t\left[\tilde{\Re}_{t,a}\right] &= \sum_a (\hat{X}_{t,a} - X_{t,a}) \sum_b Y_{t,b} f_\theta(a, b) \\
&= \sum_a (\gamma_t \hat{X}_{t,a} - \gamma_t/\mathcal{A}) \sum_b Y_{t,b} f_\theta(a, b) \\
&\leqslant \sum_a \left|\gamma_t \hat{X}_{t,a} - \gamma_t/\mathcal{A}\right| \\
&\leqslant \sum_a \left(\left|\gamma_t \hat{X}_{t,a}\right| + |\gamma_t/\mathcal{A}|\right) = 2\gamma_t
\end{aligned}$$

Step 2 (Bounding the potential.) For any $a \in \mathcal{A}$,

$$\mathbb{E}_0\left[\tilde{C}_{T,a}\right] \leqslant \mathbb{E}_0\left[\tilde{C}^+_{T,a}\right] = \mathbb{E}_0\left[\sqrt{(\tilde{C}^+_{T,a})^2}\right] \leqslant \mathbb{E}_0\left[\sqrt{\sum_a \left(\tilde{C}^+_{T,a}\right)^2}\right],$$

where the last inequality is due to the Jensen inequality. By lemma C.4 and taking expectation,

$$\mathbb{E}_0\left[\sqrt{\sum_a \left(\tilde{C}^+_{T,i}\right)^2}\right] \leqslant \mathbb{E}_0\left[\sqrt{\sum_{t=1}^T \sum_a \left(\tilde{\Re}_{t,a}\right)^2}\right]$$

The RHS of the above inequality can be bounded as

$$\begin{aligned}
\mathbb{E}_0\left[\sum_{t=1}^T \sum_a \left(\tilde{\Re}_{t,a}\right)^2\right] &= \mathbb{E}_0\left[\sum_{t=1}^T \sum_a \left(\frac{\mathbb{I}_{A_t=a} R_{t+1,A_t,B_t}}{X_{t,a}} - R_{t+1,A_t,B_t}\frac{\hat{X}_{t,A_t}}{X_{t,A_t}}\right)^2\right] \\
&= \mathbb{E}_0\left[\sum_{t=1}^T \sum_a R^2_{t+1,A_t,B_t}\left(\left(\frac{\mathbb{I}_{A_t=a}}{X_{t,A_t}}\right)^2 - \frac{2\hat{X}_{t,A_t}}{X^2_{t,A_t}}\mathbb{I}_{A_t=a} + \frac{\hat{X}^2_{t,A_t}}{X^2_{t,A_t}}\right)\right] \\
&= \mathbb{E}_0\left[\sum_{t=1}^T R^2_{t+1,A_t,B_t}\left(\frac{1}{X^2_{t,A_t}} - \frac{2\hat{X}_{t,A_t}}{X^2_{t,A_t}} + |\mathcal{A}|\frac{\hat{X}^2_{t,A_t}}{X^2_{t,A_t}}\right)\right] \\
&= \mathbb{E}_0\left[\sum_{t=1}^T \mathbb{E}_t\left[R^2_{t+1,A_t,B_t}\left(\frac{1}{X^2_{t,A_t}} - \frac{2\hat{X}_{t,A_t}}{X^2_{t,A_t}} + |\mathcal{A}|\frac{\hat{X}^2_{t,A_t}}{X^2_{t,A_t}}\right)\right]\right],
\end{aligned}$$

where we have the following derivation by the fact

$$\mathbb{E}_t\left[R^2_{t,a,b}\right] := \mathbb{E}_t\left[(f_{\theta]}(a,b) + W_{t+1})^2\right] = \mathbb{E}_t\left[f_{\theta]}(a,b)^2 + W^2_{t+1}\right] \leqslant 1 + \sigma^2_w$$

and the fact $0 \leqslant \hat{X}_{t,a}/X_{t,a} \leqslant 1/(1-\gamma_t)$ for all $a \in \mathcal{A}$,

$$\begin{aligned}
&\mathbb{E}_t\left[R^2_{t+1,A_t,B_t}\left(\frac{1}{X^2_{t,A_t}} - \frac{2\hat{X}_{t,A_t}}{X^2_{t,A_t}} + |\mathcal{A}|\,\frac{\hat{X}^2_{t,A_t}}{X^2_{t,A_t}}\right)\right] \\
&= \sum_a \frac{\mathbb{E}_t\left[R^2_{t+1,a,B_t}\right]}{X_{t,a}} + \sum_a \frac{\hat{X}_{t,a}}{X_{t,a}}\mathbb{E}_t\left[R^2_{t+1,A_t,B_t}\right]\left(|\mathcal{A}|\,\hat{X}_{t,a} - 2\right) \\
&\leqslant (1+\sigma_w^2)\left(\sum_a \frac{1}{X_{t,a}} + \sum_a \frac{\hat{X}_{t,a}}{X_{t,a}}\left(|\mathcal{A}|\,\hat{X}_{t,a} - 2\right)\right) \\
&\leqslant (1+\sigma_w^2)\left(\sum_a \frac{1}{X_{t,a}} + \sum_a \frac{\hat{X}_{t,a}}{X_{t,a}}\left(|\mathcal{A}| - 2\right)\right) \\
&\leqslant (1+\sigma_w^2)\left(\sum_a \frac{|\mathcal{A}|}{\gamma_t} + \min\!\left(\frac{|\mathcal{A}|}{\gamma_t}, \frac{|\mathcal{A}|}{1-\gamma_t}\right)(|\mathcal{A}|-2)\right) \leqslant \frac{2(1+\sigma_w^2)\,|\mathcal{A}|^2}{\gamma_t}
\end{aligned}$$

Then, we derive one important relationship

$$\mathbb{E}_0\left[\sqrt{\sum_a \left(\tilde{C}^+_{T,i}\right)^2}\right] \leqslant \mathbb{E}_0\left[\sqrt{\sum_{t=1}^T \sum_a \left(\tilde{\Re}_{t,a}\right)^2}\right] \leqslant \sqrt{\sum_{t=1}^T \frac{2(1+\sigma_w^2)\,|\mathcal{A}|^2}{\gamma_t}}$$

Step 3 (Put all together.)

$$\mathbb{E}_0\left[\sum_{t=1}^T \Re_t(a)\right] = \mathbb{E}_0\left[\sum_{t=1}^T \mathbb{E}_t\left[\tilde{\Re}_{t,a}\right] + \sum_{t=1}^T 2\gamma_t\right] = \mathbb{E}_0\left[\tilde{C}_{T,a} + \sum_{t=1}^T 2\gamma_t\right] \leqslant \sqrt{\sum_{t=1}^T \frac{2(1+\sigma_w^2)\,|\mathcal{A}|^2}{\gamma_t}} + 2\gamma_t$$

When $\gamma_t = \gamma$,

$$\mathbb{E}_0\left[\sum_{t=1}^T \Re_t(a)\right] \leqslant \sqrt{T}\sqrt{\frac{2(1+\sigma_w^2)\,|\mathcal{A}|^2}{\gamma}} + 2\gamma T,$$

Taking $\gamma = \sqrt[3]{((1+\sigma_w^2)\,|\mathcal{A}|^2)/2T}$, we have

$$\mathbb{E}_0\left[\sum_{t=1}^T \Re_t(a)\right] \leqslant 2^{4/3}(1+\sigma_w^2)^{1/3}\,|\mathcal{A}|^{2/3}\,T^{2/3}.$$

□

D. Analysis of the failure for Thompson sampling estimator in Section 3.3

D.1. Basic setting of the counter example

Consider a class of matrix games with a payoff matrix θ defined as

$$\theta = \begin{bmatrix} 1 & 1-\Delta \\ 1-\Delta & 1 \end{bmatrix}, \tag{9}$$

where $\Delta \in (0,1)$. At time t, Alice plays action $A_t \sim P_{X_t}$, and Bob is a best response player who can observe X_t and play the best response strategy $Y_t = \max_y X_t^\top \theta y$ by selecting action $B_t \sim P_{Y_t}$. Alice receives the noiseless reward $R_{t+1,A_t,B_t} = \theta_{A_t,B_t}$. From the example θ, we can observe the following:

- Observation 1: When Alice uses a pure strategy, she suffers a regret of Δ at that round due to Bob's best-response strategy.

- Observation 2: The best-response strategy for a uniform strategy is also a uniform strategy.

Let's define the following terms:

- X_t and reg_t: Alice's strategy and instantaneous regret at time t.
- $\tilde{R}_t$ and m_t: $\tilde{R}_t = \left[\tilde{R}_{t1}, \tilde{R}_{t2}\right]$ is the estimated reward vector by TS-RM estimator; $m_t = \tilde{R}_{t1} - \tilde{R}_{t2}$ represents the difference between two rewards.

Now, let's consider the following remark regarding initialization:

Remark D.1 (Initialization). The TS-RM algorithm for Alice, initialized with a uniform strategy, will always result in a pure strategy in X_2.

Proof. In the regret-matching algorithm, the instantaneous regret reg_1 at $t = 1$ can be represented as

$$reg_1 = \tilde{R}_1 - \tilde{R}_1^T X_1 \cdot \mathbf{1}, \tag{10}$$

Since Alice and Bob are both initialized with a uniform strategy, i.e., $X_1 = [0.5, 0.5]$, it can be observed that if $m_1 \neq 0$, the two elements in reg_1 will always have opposite signs. If $m_1 = 0$, $X_2 = X_1$ and can still be regarded as an initialization step.According to the regret-matching updating rule, $X_2 \propto \max(reg_1, 0)$, which means X_2 must be a pure strategy. □

Based on Remark D.1, we can draw the following conclusions regarding the counter example:

- The choice of uniform initialization for the TS-RM algorithm does not affect the divergence result.
- This result holds regardless of the specific action chosen by Alice at $t = 1$, indicating that two symmetric conditions arise depending on whether $X_2 = [1, 0]$ or $X_2 = [0, 1]$.

D.2. TS-RM suffers linear regret

According to Remark D.1, without loss of generality (w.l.o.g.), let us define the following events:

- Event ω_t: Alice picks the second row and the best-response opponent chooses the first column at time t.
- Event Ω_t: Alice picks the second row and the best-response opponent chooses the first column for all time $t' \leqslant t$.

The occurrence of event Ω_t implies that Alice experiences linear regret until time t. However, the actual convergence probability is greater than $\mathbb{P}(\Omega_t)$ since even if Ω_t does not occur (i.e., Alice occasionally chooses the optimal result 1), there is still a probability that TS-RM fails. Quantifying this probability is challenging. If we can demonstrate that Ω_t occurs with a constant probability c, then the divergence probability of TS-RM should be greater than c. Specifically, due to the symmetric property of the example θ, we obtain the following propositions:

Proposition D.2. *If Alice initializes with a uniform strategy,*

$$\Re(T) \geqslant 2\mathbb{P}(\Omega_T)\Delta \cdot T$$

If Ω_t happens with constant probability for all t $\geqslant 1$, then Alice suffers linear regret.

Proposition D.3 (Failure of TS-RM)**.** *Suppose Alice initializes with a uniform strategy and utilizes Regret Matching with Thompson Sampling estimator (TS-RM). For any $\Delta \in (0, 1)$ and $\sigma_n > 0$, there exists a constant $c(\Delta, \sigma_n) > 0$ such that for all rounds $t \geqslant 1$, $\mathbb{P}(\Omega_t) \geqslant c(\Delta, \sigma_n)$.*

To investigate the divergence behavior of the TS-RM algorithm, an experiment is conducted using the counter example in . Different values of Δ and σ_n^2 are considered, and 200 independent simulation runs are performed for each combination. The averaged divergence results across these runs are shown in Fig.5, which illustrates that the probability of divergence decreases as Δ and σ_n^2 increase, consistent with our proposition (Prop.D.3). In the following, we provide a detailed proof for the divergence of the TS-RM algorithm.

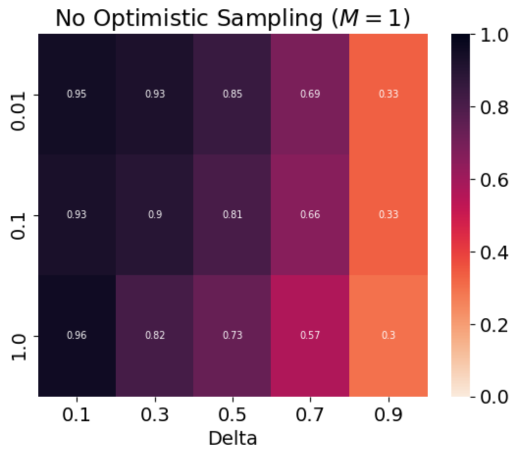


Figure 5. Divergence probability for TS-RM

Proof. Our goal is to prove that there exists a constant $c \geqslant 0$ such that $\mathbb{P}(\Omega_t) \geqslant c$ for all t. The probability $\mathbb{P}(\Omega_t)$ can be expressed as

$$\mathbb{P}(\Omega_t) = \mathbb{P}(\omega_1)\mathbb{P}(\omega_2|\Omega_1)\dots\mathbb{P}(\omega_t|\Omega_{t-1}) \tag{11}$$

Referring to Remark D.1, X_2 is a pure strategy. Since Alice chooses the second row according to Ω_2, we get $X_2 = [0, 1]$ and $m_1 < 0$. Following the event Ω_t, we get $X_t = [0, 1], \forall t \geqslant 2$, which indicates that for the TS-RM estimator $\tilde{R}_t$, only the posterior distribution of $\tilde{R}_{t2}$ is updated. Since no noise is considered in received rewards, by Bayesian rule, we have

$$\tilde{R}_t = \left[z_t, \frac{t}{t+\sigma_n^2}(1-\Delta) + \sqrt{\frac{\sigma_n^2}{\sigma_n^2+t}} z'_t \right],$$

where $z_t, z'_t \sim \mathcal{N}(0, 1)$ are two independent *r.v.*s, and σ_n^2 is the noise variance. Therefore, we can express the regret as:

$$reg_t = \tilde{R}_t - X_t^T \tilde{R}_t \cdot \mathbf{1} = [m_t, 0], \quad \forall t \geqslant 2 \tag{12}$$

where $m_t = z_t - \frac{t}{t+\sigma_n^2}(1-\Delta) - \sqrt{\frac{\sigma_n^2}{t+\sigma_n^2}} z'_t = z_t - \sqrt{\frac{\sigma_n^2}{t+\sigma_n^2}} z'_t - \frac{t}{t+\sigma_n^2}(1-\Delta)$.

The cumulative regret can be represented as

$$\Re_t = [0.5 m_1 + \sum_{k=2}^{t} m_k, -0.5 m_1] \tag{13}$$

According to updating rule of TS-RM, we have:

$$\begin{aligned}
\mathbb{P}(\omega_t|\Omega_{t-1}) &= \mathbb{P}(0.5 m_1 + \sum_{k=2}^{t} m_k \leqslant 0 | \Omega_{t-1}) \\
&\geqslant \mathbb{P}(\sum_{k=2}^{t} m_k \leqslant 0 | \Omega_{t-1}) \\
&= \mathbb{P}(\sum_{k=2}^{t} m_k \leqslant 0 | \Omega_{t-1}) \left(\mathbb{P}(\Omega_{t-1}) + \mathbb{P}(\bar{\Omega}_{t-1}) \right) \\
&\geqslant \mathbb{P}(\sum_{k=2}^{t} m_k \leqslant 0 | \Omega_{t-1}) \mathbb{P}(\Omega_{t-1}) + \mathbb{P}(\sum_{k=2}^{t} m_k \leqslant 0 | \bar{\Omega}_{t-1}) \mathbb{P}(\bar{\Omega}_{t-1}) \\
&= \mathbb{P}(\sum_{k=2}^{t} m_k \leqslant 0)
\end{aligned} \tag{14}$$

where the first inequality is because $m_1 \leqslant 0$ (conditioned on Ω_{t-1}), and the second one is due to $\mathbb{P}(\sum_{k=2}^{t} m_k \leqslant 0|\Omega_{t-1}) \geqslant \mathbb{P}(\sum_{k=2}^{t} m_k \leqslant 0|\bar{\Omega}_{t-1})$.

Define

$$N_t = \sum_{k=2}^{t} m_k \sim \mathcal{N}\left(-\sum_{k=2}^{t} \frac{k}{k+\sigma_n^2}(1-\Delta),\ \sum_{k=2}^{t}(1+\frac{\sigma_n^2}{k+\sigma_n^2})\right) \triangleq \mathcal{N}(\mu_t, \sigma_t^2).$$

As $t \to \infty$, we have:

$$\begin{aligned}
\lim_{t\to\infty} \log \mathbb{P}(\Omega_t) &= \log \prod_{t=1}^{\infty} \mathbb{P}(\omega_t|\Omega_{t-1}) = \sum_{t=1}^{\infty} \log \mathbb{P}(\omega_t|\Omega_{t-1}) \\
&\geqslant \sum_{t=1}^{\infty} \log \mathbb{P}(N_t \leqslant 0) \\
&= \sum_{t=1}^{\infty} \log \mathbb{P}(\mu_t + \sigma_t Z \leqslant 0), \quad Z \sim \mathcal{N}(0,1) \\
&= \sum_{t=1}^{\infty} \log \mathbb{P}(Z \leqslant -\frac{\mu_t}{\sigma_t}) \\
&= \sum_{t=1}^{\infty} \log \Phi(-\frac{\mu_t}{\sigma_t}) \\
&= \sum_{t=1}^{\infty} \log \Phi\left(\frac{\sum_{k=2}^{t} \frac{k}{k+\sigma_n^2}(1-\Delta)}{\sqrt{\sum_{k=2}^{t}(1+\frac{\sigma_n^2}{k+\sigma_n^2})}}\right) \\
&\geqslant \sum_{t=1}^{\infty} \log \Phi\left(\frac{(1-\Delta)\left(t-\sigma_n^2 \ln(t+\sigma_n^2)+2\sigma_n^2 \ln \sigma_n - 1/(\sigma_n^2+1)\right)}{\sqrt{t+\sigma_n^2 \ln(t+\sigma_n^2) - 2\sigma_n^2 \ln \sigma_n - (\sigma_n^2+2)/(\sigma_n^2+1)}}\right)
\end{aligned} \tag{15}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and the last inequality is due to $\sum_{k=1}^{t} \frac{1}{k+a} \leqslant \int_0^t \frac{1}{k+a} dk$. Define

$$f_t(\Delta, \sigma_n) = \frac{(1-\Delta)\left(t-\sigma_n^2 \ln(t+\sigma_n^2)+2\sigma_n^2 \ln \sigma_n - 1/(\sigma_n^2+1)\right)}{\sqrt{t+\sigma_n^2 \ln(t+\sigma_n^2) - 2\sigma_n^2 \ln \sigma_n - (\sigma_n^2+2)/(\sigma_n^2+1)}}, \tag{16}$$

Referring to the lower bound of the standard Gaussian distribution, we can continue to derive:

$$\begin{aligned}
\lim_{t\to\infty} \log \mathbb{P}(\Omega_t) &\geqslant \sum_{t=1}^{\infty} \log \Phi(f_t(\Delta, \sigma_n)) \\
&\geqslant \sum_{t=1}^{\infty} \log\left(1 - \frac{1}{\sqrt{2\pi} f_t(\Delta,\sigma_n) e^{f_t^2(\Delta,\sigma_n)/2}}\right) \\
&\geqslant \sum_{t=1}^{\infty} \left(-\frac{1}{\sqrt{2\pi} f_t(\Delta,\sigma_n) e^{f_t^2(\Delta,\sigma_n)/2}}\right) \\
&> -\infty,
\end{aligned} \tag{17}$$

This shows that there exists a constant $c' > 0$ such that:

$$\lim_{t\to\infty} \log \mathbb{P}(\Omega_t) \geqslant \log c' > -\infty \tag{18}$$

In other words, we have:

$$\lim_{t\to\infty} \mathbb{P}(\Omega_t) \geqslant c > 0, \tag{19}$$

where $c = e^{c'}$. Moreover, $\mathbb{P}(\Omega_t) \geqslant \lim_{t\to\infty} \mathbb{P}(\Omega_t) \geqslant c$ for a finite sequence. Specifically, let $\Delta = 0.1$ and $\sigma_n^2 = 0.1$, we can get $c' = -0.62$, and $c = 0.54$. $\square$

Moreover, the function $f_t(\Delta, \sigma_n^2)$, combined with the derivation above, demonstrates that the divergence probability $\mathbb{P}(\Omega_t)$ decreases as Δ and σ_n^2 increase, which is consistent with our proposition (Proposition D.3) and the experiments.

The above argument is based on the frequentist setting where the underlying instance is fixed, and the agent does not access the right noise likelihood function of the environment. We conjecture that under the Bayesian setting where the prior and likelihood in the game environment are available to the agent, with the exact Bayes posterior, the TS-RM still suffers linear Bayesian adversarial regret.

D.3. Why Optimistic variant of TS can converge?

Assume that Alice chooses the wrong action until time t. By Prop D.3, the TS-RM algorithm will continue to choose the wrong action with a constant probability c. Different from TS-RM, we will prove that even if Alice chooses the wrong action until time t, OTS-RM will eventually yield a sub-linear regret with high probability.

Proof. Unlike TR-RM, the OTS-RM algorithm uses M samples for optimistic sampling in each round. Under the assumption that Alice takes the wrong action until time t, we have:

$$\begin{cases} x_{ti} & \sim \mathcal{N}_i(0,1), \quad i = 1, \dots, M \\ y_{ti} & \sim \mathcal{N}_i\left(\frac{t}{t+\sigma_n^2}(1-\Delta), \frac{\sigma_n^2}{\sigma_n^2+t}\right), \quad i = 1, \dots, M \end{cases} \tag{20}$$

Let $\tilde{R}_{t1} = \max_i x_{ti}$ and $\tilde{R}_{t2} = \max_i y_{ti}$, which are just $\tilde{R}_t(1nd)$ and $\tilde{R}_t(2nd)$ mentioned above, respectively.

The proof will first show that $\tilde{R}_{t1} \geqslant \tilde{R}_{t2}$ with high probability. As a result, $\Re_{t1}$ will decrease, and eventually Alice's strategy will change from a pure strategy $[0,1]$ to a mixed strategy, indicating a decay of $\mathbb{P}(\Omega_t)$ over time t.

According to the anti-concentration property in Lemma E.11, we have:

$$\mathbb{P}(\max_i y_{ti} \leqslant \frac{t}{t+\sigma_n^2}(1-\Delta) + \sqrt{\frac{2\sigma_n^2 \log(M/\delta_1)}{t+\sigma_n^2}}) \geqslant 1 - \delta_1 \tag{21}$$

Thus, the first step in the proof can be written as:

$$\begin{aligned} \mathbb{P}(\tilde{R}_{t1} \geqslant \tilde{R}_{t2}) &= \mathbb{P}(\max_i x_i \geqslant \max_i y_i) \\ &\geqslant (1-\delta_1)\mathbb{P}(\max_i x_i \geqslant \max_i y_i \mid \epsilon) \\ &\geqslant (1-\delta_1)\mathbb{P}(\max_i x_{ti} \geqslant \frac{t}{t+\sigma_n^2}(1-\Delta) + \sqrt{\frac{2\sigma_n^2 \log(M/\delta_1)}{t+\sigma_n^2}} \mid \epsilon) \\ &\geqslant 1 - \delta_1 - (1-\delta_1)\Phi^M\left(\frac{t}{t+\sigma_n^2}(1-\Delta) + \sqrt{\frac{2\sigma_n^2 \log(M/\delta_1)}{t+\sigma_n^2}}\right) \end{aligned} \tag{22}$$

Here, ϵ represents the event where the anti-concentration property occurs.

Let $f(t, \Delta, \sigma_n) = \frac{t}{t+\sigma_n^2}(1-\Delta) + \sqrt{\frac{2\sigma_n^2 \log(M/\delta_1)}{t+\sigma_n^2}}$. Then, we have:

$$\begin{aligned} \mathbb{P}(\tilde{R}_{t1} \geqslant \tilde{R}_{t2}) &\geqslant 1 - \delta_1 - (1-\delta_1)\Phi^M(f) \\ &\geqslant 1 - \delta_1 - (1-\delta_1)\left(1 - \frac{f}{\sqrt{2\pi}(f^2+1)e^{f^2/2}}\right)^M \\ &\geqslant 1 - \delta_1 - (1-\delta_1)\exp\left(\frac{-Mf}{\sqrt{2\pi}(f^2+1)e^{f^2/2}}\right) \end{aligned} \tag{23}$$

where $\Phi(x) \leqslant 1 - \frac{x}{\sqrt{2\pi}(x^2+1)e^{x^2/2}}$ (Gordon, 1941) and $(1-x)^M \leqslant e^{-Mx}$. $\square$

To obtain further insights into the relationship between M and t, we have depicted a figure in Figure 6 that corresponds to the inequality in Equation 23. Based on the analysis, we draw the following conclusions:

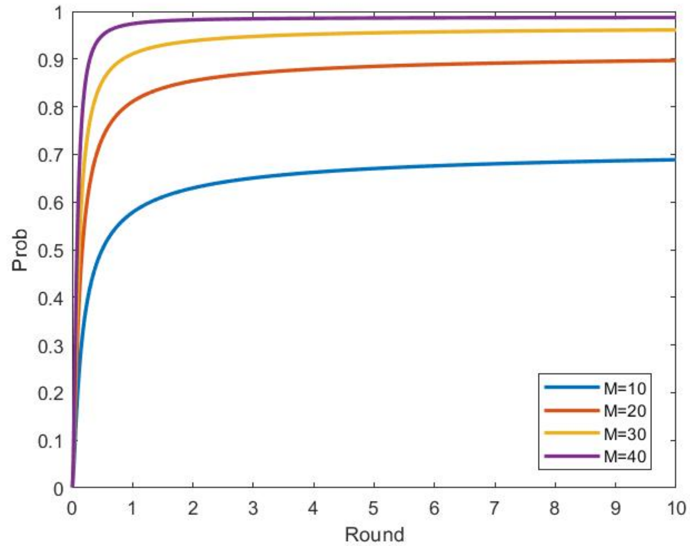


Figure 6. Relationship between t and M (fix Δ and σ_n)

- When Δ, σ_n and t are fixed, $\mathbb{P}(\tilde{R}_{t1} \geqslant \tilde{R}_{t2})$ increases with M. Additionally, the value of t has a significant influence on $\mathbb{P}(\tilde{R}_{t1} \geqslant \tilde{R}_{t2})$.
- When Δ, σ_n and M are fixed, the probability $\mathbb{P}(\tilde{R}_{t1} \geqslant \tilde{R}_{t2})$ increases with time t. Moreover, $\mathbb{P}(\tilde{R}_{t1} \geqslant \tilde{R}_{t2})$ will quickly reach a region close to the maximum in just a few rounds.
- When σ_n, M and t are fixed, $\mathbb{P}(\tilde{R}_{t1} \geqslant \tilde{R}_{t2})$ initially increases with Δ and subsequently decreases with Δ.

These findings provide valuable insights into the behavior of the OTS-RM algorithm and support our claim that $\mathbb{P}(\tilde{R}_{t1} \geqslant \tilde{R}_{t2})$ increases as M grows. This increasing probability implies that Alice's strategy will transition from a pure strategy $[0, 1]$ to a mixed strategy, indicating a decay of $\mathbb{P}(\Omega_t)$ over time t. Consequently, the algorithm achieves sub-linear regret with high probability.

E. Technical details in Section 4

Our analysis road map is as follows. First, in Appendix E.1, as described in Proposition 4.1, we derive a general regret decomposition in given any imagined reward sequence $\{\tilde{R}_t : t \in \mathbb{Z}_{++}\}$, where each $\tilde{R}_{t+1}$ is constructed using history information H_t with algorithmic randomness. Then, to further upper bound the regret, we introduce the generic upper confidence bound (UCB) sequence and lower confidence bound (LCB) sequence as in Definition E.4. With these sequences, we have a general regret bound in Proposition E.8 with generic UCB and LCB sequences. Next, as described in Remark E.9, we specify the so-called information-theoretic confidence bound in Appendix E.2 and show that the imagined reward sequence has good properties when compared with specified sequences in Appendix E.3, finally yielding the information-theoretic regret upper bound in Appendix E.4.

E.1. General regret bound

Proposition E.1 (Restate regret decomposition in Proposition 4.1)**.** *For any $a \in \mathcal{A}$, the one-step regret can be decomposed by*

$$\begin{aligned}
&\mathbb{E}\left[R_{t+1,a,B_t} - R_{t+1,A_t,B_t} \mid \theta\right] = \mathbb{E}\left[f_\theta(a, B_t) - f_\theta(A_t, B_t) \mid \theta\right] \\
&= \underbrace{\mathbb{E}\left[\tilde{R}_{t+1}(a) - \tilde{R}_{t+1}(A_t) \mid \theta\right]}_{(I)} + \underbrace{\mathbb{E}\left[f_\theta(a, B_t) - \tilde{R}_{t+1}(a) \mid \theta\right]}_{(II)} + \underbrace{\mathbb{E}\left[\tilde{R}_{t+1}(A_t) - f_\theta(A_t, B_t) \mid \theta\right]}_{(III)}
\end{aligned}$$

Proof. Since

$$\mathbb{E}\left[R_{t+1,a,B_t} - R_{t+1,A_t,B_t} \mid \theta\right] = \mathbb{E}\left[f_\theta(a, B_t) - f_\theta(A_t, B_t) \mid \theta\right]$$

We have introduced the imagined time-varying sequence $\tilde{R} := \{\tilde{R}_t : t \in \mathbb{Z}_{++}\}$, where each $\tilde{R}_t$ is constructed using history information H_t and takes value in $[0, 1]^{\mathcal{A}}$.

$$f_\theta(a, B_t) - f_\theta(A_t, B_t) = \underbrace{\tilde{R}_{t+1}(a) - \tilde{R}_{t+1}(A_t)}_{(I)\ \ \mathrm{adv}_{t+1}(a)} + \underbrace{f_\theta(a, B_t) - \tilde{R}_{t+1}(a)}_{(II)\ \ \mathrm{pess}_{t+1}(a)} + \underbrace{\tilde{R}_{t+1}(A_t) - f_\theta(A_t, B_t)}_{(III)\, \mathrm{est}_{t+1}} \tag{24}$$

□

Reduction to full-information adversarial regret. Any algorithm $\pi = \pi^{\text{adv-est}}$ constructs the imagined reward sequence $\tilde{R}$ and use adv algorithm for no-regret update will lead to,

$$\begin{aligned}
&\sum_{t=0}^{T-1} \mathbb{E}\left[f_\theta(a, B_t) - f_\theta(A_t, B_t) \mid \theta\right] \\
&\leqslant \mathfrak{R}_{\text{full}}(a; T, \mathrm{adv}, \tilde{R}) + \sum_{t=0}^{T-1} \mathbb{E}\left[\mathrm{pess}_{t+1}(a) \mid \theta\right] + \sum_{t=0}^{T-1} \mathbb{E}\left[\mathrm{est}_{t+1} \mid \theta\right]
\end{aligned} \tag{25}$$

where

$$\mathfrak{R}_{\text{full}}(a; T, \mathrm{adv}, \tilde{R}) = \mathbb{E}\left[\sum_{t=0}^{T-1} \tilde{R}_{t+1}(a) - \tilde{R}_{t+1}(A_t)\right].$$

Recall the Bayesian adversarial regret is

$$\mathfrak{R}(T, \pi^{\text{alg}}, \pi^B) = \mathbb{E}\left[\mathfrak{R}(T, \pi^{\text{alg}}, \pi^B, \theta)\right]$$

where the expectation is taken over the prior distribution of θ. From Equation (25), we have

$$\begin{aligned}
\mathfrak{R}(T, \pi^{\text{alg}}, \pi^B) &= \mathbb{E}\left[\max_{a \in \mathcal{A}} \sum_{t=0}^{T-1} \mathbb{E}\left[R_{t+1,a,B_t} - R_{t+1,A_t,B_t} \mid \theta\right]\right] \\
&\leqslant \max_{a \in \mathcal{A}} \mathfrak{R}_{\text{full}}(a; T, \mathrm{adv}, \tilde{R}) + \mathbb{E}\left[\max_{a \in \mathcal{A}} \sum_{t=0}^{T-1} \mathbb{E}\left[\mathrm{pess}_{t+1}(a) \mid \theta\right]\right] + \sum_{t=0}^{T-1} \mathbb{E}\left[\mathrm{est}_{t+1}\right] \\
&= \mathfrak{R}_{\text{full}}(T, \mathrm{adv}, \tilde{R}) + \mathbb{E}\left[\max_{a \in \mathcal{A}} \sum_{t=0}^{T-1} \mathbb{E}\left[\mathrm{pess}_{t+1}(a) \mid \theta\right]\right] + \sum_{t=0}^{T-1} \mathbb{E}\left[\mathrm{est}_{t+1}\right]
\end{aligned} \tag{26}$$

Remark E.2. If each imagined reward $\tilde{R}_t$ in the sequence $\tilde{R}$ takes bounded value in $[-c, c]^{\mathcal{A}}$, any suitable full information adversarial algorithm adv will give satisfied bound for $\mathfrak{R}_{\text{full}}(a; T, \mathrm{adv}, \tilde{R})$ for any $a \in \mathcal{A}$. Specifically, Hedge suffers $\mathfrak{R}_{\text{adv}}(a; T, \text{Hedge}, \tilde{R}) = \mathcal{O}(2c\sqrt{T \log \mathcal{A}})$ and RM suffers $\mathfrak{R}_{\text{full}}(a; T, \text{RM}, \tilde{R}) = \mathcal{O}(2c\sqrt{T\mathcal{A}})$.

Next, we focus on the pessimism term and estimation term. We now focus on the case generalized from the optimistic Thompson sampling.

Assumption E.3 (Restriction on the imagined reward sequence $\tilde{R}$)**.** In the following context, the imagined reward sequence $\tilde{R} = (\tilde{R}_1, \ldots, \tilde{R}_{t+1}, \ldots)$ satisfies that (1) $\tilde{R}_{t+1}(a) \in [0, C]$ for all $a \in \mathcal{A}$ and (2) $\tilde{R}_{t+1}(a)$ is random only through its dependence on Z_{t+1} given the history H_t, B_t for all $a \in \mathcal{A}$. To clarify, $\tilde{R}_{t+1}(a)$ has no dependence on A_t and $\tilde{R}_{t+1,A_t,B_t}$.

Definition E.4 (UCB and LCB sequence)**.** UCB sequence $U = (U_t \mid t \in \mathbb{N})$ LCB sequence $L = (L_t \mid t \in \mathbb{N})$ are two sequences of functions where each $0 \leqslant L_t \leqslant U_t \leqslant C$ are both deterministic given history H_t.

Definition E.5 (Optimistic Event)**.** For any imagined reward sequence $\tilde{R}_{t+1}$ in Assumption E.3 and any upper confidence sequence U in Definition E.4, we define the event

$$\mathcal{E}_t(\tilde{R}, U, B_t) := \{\tilde{R}_{t+1}(a) \geqslant U_t(a, B_t), \forall a \in \mathcal{A}\}.$$

Fact E.1. *The event $\mathcal{E}_t^o(\tilde{R}, U, B_t)$ is random only through its dependence on Z_{t+1} given H_t, B_t. The pessimism term can be decomposed according to the event $\mathcal{E}_t^o(\tilde{R}, U, B_t)$*

$$\tilde{R}_{t+1}(a) - f_\theta(a, B_t) = \mathbb{1}_{\mathcal{E}_t^o(\tilde{R},U,B_t)}(\tilde{R}_{t+1}(a) - f_\theta(a, B_t)) + (1 - \mathbb{1}_{\mathcal{E}_t^o(\tilde{R},U,B_t)})(\tilde{R}_{t+1}(a) - f_\theta(a, B_t))$$

Consider the case where f_θ takes values in $[0, C]$ and $\tilde{R}_{t+1} \in [0, C]$ by Assumption E.3, for all $a \in \mathcal{A}$,

$$f_\theta(a, B_t) - \tilde{R}_{t+1}(a) \leqslant \mathbb{1}_{\mathcal{E}_t^o(\tilde{R},U,B_t)}(f_\theta(a, B_t) - U_t(a, B_t)) + C(1 - \mathbb{1}_{\mathcal{E}_t^o(\tilde{R},U,B_t)}). \tag{27}$$

Definition E.6 (Concentration Event)**.** For any imagined reward sequence $\tilde{R}_{t+1}$ in Assumption E.3 and any upper confidence sequence U' in Definition E.4, we define the event

$$\mathcal{E}_t^c(\tilde{R}, U', A_t, B_t) := \{\tilde{R}_{t+1}(A_t) \leqslant U'_t(A_t, B_t)\}.$$

Fact E.2. *Consider the case where f_θ takes values in $[0, C]$ and $\tilde{R}_{t+1} \in [0, C]$ by Assumption E.3, the event $\mathcal{E}_t(\tilde{R}, U', A_t, B_t)$ is random only through its dependence on Z_{t+1} given H_t, A_t, B_t. The estimation term then becomes,*

$$\tilde{R}_{t+1}(A_t) - f_\theta(A_t, B_t) \leqslant (\mathbb{1}_{\mathcal{E}_t^c(\tilde{R},U',A_t,B_t)})(U'_t(A_t, B_t) - f_\theta(A_t, B_t)) + C(1 - \mathbb{1}_{\mathcal{E}_t^c(\tilde{R},U',A_t,B_t)}). \tag{28}$$

Definition E.7 (Confidence event)**.** Define the confidence event at round t as

$$\mathcal{E}_t(f_\theta, B_t) := \{\forall a \in \mathcal{A}, f_\theta(a, B_t) \in [L_t(a, B_t), U_t(a, B_t)]\}.$$

Fact E.3. *The event $\mathcal{E}_t^c(f_\theta, B_t)$ is deterministic conditioned on history H_t, B_t and θ.*

Based on the definition of the confidence event, the pessimism term from Equation (27) becomes

$$f_\theta(a, B_t) - \tilde{R}_{t+1}(a) \leqslant C(1 - \mathbb{1}_{\mathcal{E}_t^o(\tilde{R},U,B_t) \cap \mathcal{E}_t^c(f_\theta,B_t)}), \quad \forall a \in \mathcal{A}.$$

The estimation term from Equation (28) becomes

$$\begin{aligned} \tilde{R}_{t+1}(A_t) - f_\theta(A_t, B_t) \leqslant\ & \mathbb{1}_{\mathcal{E}_t^c(\tilde{R},U',A_t,B_t) \cap \mathcal{E}_t^c(f_\theta,B_t)}(U'_t(A_t, B_t) - L_t(A_t, B_t)) \\ & + C\left(1 - \mathbb{1}_{\mathcal{E}_t^c(\tilde{R},U,A_t,B_t)} \mathbb{1}_{\mathcal{E}_t^c(f_\theta,B_t)}\right) \end{aligned}$$

Denote the complement event of $\mathcal{E}$ as $\neg\mathcal{E}$.

$$\begin{aligned} \mathbb{E}\left[\max_{a\in\mathcal{A}} \sum_{t=0}^{T-1} \mathbb{E}\left[\text{pess}_{t+1}(a) \mid \theta\right]\right] &\leqslant \mathbb{E}\left[\max_{a\in\mathcal{A}} \sum_{t=0}^{T-1} \mathbb{E}\left[C(1 - \mathbb{1}_{\mathcal{E}_t^o(\tilde{R},U,B_t) \cap \mathcal{E}_t^c(f_\theta,B_t)}) \mid \theta\right]\right] \\ &= \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{E}\left[C(1 - \mathbb{1}_{\mathcal{E}_t^o(\tilde{R},U,B_t) \cap \mathcal{E}_t^c(f_\theta,B_t)})\right]\right] \\ &= C \sum_{t=0}^{T-1} \mathbb{P}\left(\neg\mathcal{E}_t^o(\tilde{R}, U, B_t) \cup \neg\mathcal{E}_t^c(f_\theta, B_t)\right) \end{aligned} \tag{29}$$

and by assuming $U'_t \geqslant L_t$

$$\begin{aligned} \mathbb{E}\left[\sum_{t=0}^{T-1} \text{est}_{t+1}\right] \leqslant\ & \sum_{t=0}^{T-1} \mathbb{E}\left[(U'_t(A_t, B_t) - L_t(A_t, B_t))\right] \\ & + C \sum_{t=0}^{T} \mathbb{P}(\neg\mathcal{E}_t^c(\tilde{R}, U', A_t, B_t) \cup \neg\mathcal{E}_t^c(f_\theta, B_t)) \end{aligned} \tag{30}$$

Proposition E.8 (General Regret Bound with confidence sequence.)**.** *Given sequences $U' \geqslant U \geqslant L$, we upper bound the Bayesian adversarial regret with*

$$\begin{aligned} \Re(T, \pi^{\text{alg}}, \pi^B) \leqslant\ & \Re_{\text{full}}(T, \text{adv}, \tilde{R}) + c \\ & + C \sum_{t=0}^{T-1} \mathbb{P}(\neg\mathcal{E}_t^c(\tilde{R}, U', A_t, B_t)) + 2\mathbb{P}(\neg\mathcal{E}_t^c(f_\theta, B_t)) + \mathbb{P}\left(\neg\mathcal{E}_t^o(\tilde{R}, U, B_t)\right) \end{aligned} \tag{31}$$

Proof. This is a direct consequence of Equations (26), (29) and (30). $\square$

Remark E.9. In the following sections, we will discuss the specific choice of the UCB and LCB sequence U', U, L. In Appendix E.2, we will show that with information-theoretic confidence bound, the probability $\mathbb{P}(\neg\mathcal{E}_t^c(f_\theta, B_t))$ of the function f_θ not covered in the confidence region is small. In Appendix E.3, we will show the imagined reward sequence constructed by OTS and UCB will lead to small will stay within the sequence U' and U with high probability, i.e., the probability $\mathbb{P}\left(\neg\mathcal{E}_t^o(\tilde{R}, U, B_t)\right)$ and $\mathbb{P}(\neg\mathcal{E}_t^c(\tilde{R}, U', A_t, B_t))$ is small enough. In Appendix E.4, we show that $\sum_{t=0}^{T-1}\mathbb{E}\left[(U_t'(A_t, B_t) - L_t(A_t, B_t))\right]$ can be bounded by the mutual information $I(\theta; H_T)$ with the information-theoretic confidence bound defined in Appendix E.2.

E.2. Information-theoretic confidence bound

Fact E.4 (Chernoff bound)**.** *Suppose X is normal distributed $N(\mu, \sigma^2)$, the optimized chernoff bound for X is*

$$\mathbb{P}(X - \mu \geqslant c) \leqslant \min_{t>0} \frac{\exp(\sigma^2 t^2/2)}{\exp(tc)} = \exp(-c^2/\sigma^2)$$

Lemma E.10. *Conditioned on H_t and B_t, Define the set $\mathcal{F}_t$ as*

$$\mathcal{F}_t := \left\{f_\theta : |f_\theta(a, B_t) - \mu_t(a, B_t)| \leqslant \sqrt{\beta_t'}\sigma_t(a, B_t), \forall a \in \mathcal{A}\right\}$$

Then,

$$\mathbb{P}(f_\theta \in \mathcal{F}_t) \geqslant 1 - 2\mathcal{A}\exp(-\beta_t'/2).$$

Proof. Since $f_\theta(a, b) \mid H_t$ is distributed as $N(\mu_t(a, b), \sigma_t(a, b))$, by Fact E.4,

$$\mathbb{P}(|f_\theta(a, b) - \mu_t(a, b)| \geqslant \sqrt{\beta_t'}\sigma_t(a, b) \mid H_t) \leqslant 2\exp\left(-\frac{\beta_t'}{2}\right) \tag{32}$$

By union bound, we have

$$\mathbb{P}(|f_\theta(a, b) - \mu_t(a, b)| \geqslant \sqrt{\beta_t'}\sigma_t(a, b), \forall a \in \mathcal{A} \mid H_t) \leqslant 2\mathcal{A}\exp(-\beta_t'/2),$$

for any fixed $b \in \mathcal{B}$, We observe that conditioned on H_t, the opponent's action B_t is independent of $f_\theta(a, b)$ for all $(a, b) \in \mathcal{A} \times \mathcal{B}$. Therefore, we further derive

$$\mathbb{P}(|f_\theta(a, B_t) - \mu_t(a, B_t)| \geqslant \sqrt{\beta_t'}\sigma_t(a, b), \forall a \in \mathcal{A} \mid H_t) \leqslant 2\mathcal{A}\exp(-\beta_t'/2).$$

Taking expectations on both sides, we prove the lemma. $\square$

Let the UCB and LCB sequences be

$$U = (\mu_t(a, b) + \sqrt{\beta_t'}\sigma_t(a, b) : t \in \mathbb{N}), \quad L = (\mu_t(a, b) - \sqrt{\beta_t'}\sigma_t(a, b) : t \in \mathbb{N}).$$

Let $\beta_t' = 2\log\mathcal{A}\sqrt{T}$, by Lemma E.10, we can see the probability introduced in Definition E.7 is $\mathbb{P}(\neg\mathcal{E}_t^c(f_\theta, B_t)) \leqslant 2\mathcal{A}\exp(-\beta_t'/2) = 2/\sqrt{T}$.

Relate σ_t to information-theoretic quantity.

Fact E.5 (Mutual information of Guassian distribution)**.** *If $f(a) \sim N(\mu(a), \sigma(a))$ and $o = f(a) + w$ with fixed a and $w \sim N\left(0, \sigma_w^2\right)$, then*

$$I(\theta; o) = \frac{1}{2}\log\left(1 + \sigma_w^{-2}\sigma(a)\right)$$

For an fixed action pair (a, b), by the Fact E.5,

$$I_t\left(\theta; R_{t+1, A_t, B_t} \mid A_t = a, B_t = b\right) = \frac{1}{2}\log\left(1 + \frac{\sigma_t^2(a, b)}{\sigma_w^2}\right)$$

Let the width function w_t be

$$w_t(a,b) = \sqrt{\beta_t I_t\left(\theta; R_{t+1,A_t,B_t} \mid A_t = a, B_t = b\right)} \quad \text{with} \quad \beta_t = \frac{2\beta_t'}{\log(1+\sigma_w^{-2})}.$$

Expanding $w_t(a,b)$ leads to

$$w_t(a,b)^2 = \beta_t I_t(\theta; R_{t+1,A_t,B_t} \mid A_t = a, B_t = b) = \frac{\beta_t' \log(1+\sigma_w^{-2}\sigma_t^2(a,b))}{\log(1+\sigma_w^{-2})} \geqslant \beta_t'\sigma_t^2(a,b). \tag{33}$$

The last inequality follows from the fact that $\frac{x}{\log(1+x)}$ monotonically increases for $x > 0$, and that $\sigma_t^2(a,b) \leqslant k((a,b),(a,b)) \leqslant 1$, leading to the result $\sigma_t^2(a,b) \leqslant \frac{1}{\log(1+\sigma_w^{-2})}\log(1+\sigma_w^{-2}\sigma_t^2(a,b))$.

E.3. Anti-concentration behavior of optimistic Thompson sampling

Let the UCB sequences be

$$U = (\mu_t(a,b) + \sqrt{\beta_t'}\sigma_t(a,b) : t \in \mathbb{N}), \quad U' = (\mu_t(a,b) + \sqrt{\log M\beta_t'}\sigma_t(a,b) : t \in \mathbb{N}).$$

In this section, we show the probability $\mathbb{P}\left(\neg\mathcal{E}_t^o(\tilde{R}, U, B_t)\right)$ and $\mathbb{P}(\neg\mathcal{E}_t^c(\tilde{R}, U', A_t, B_t))$ is small enough. The following two lemmas are important.

Lemma E.11. *Consider a normal distribution $N\left(0,\sigma^2\right)$ where σ is a scalar. Let $\eta_1, \eta_2, \ldots, \eta_M$ be M independent samples from the distribution. For any $w \in \mathbb{R}_+$,*

$$\mathbb{P}\left(\max_{j\in[M]} \eta_j \geqslant w\right) = 1 - \left[\Phi\left(\frac{w}{\sigma}\right)\right]^M$$

Proof. By the fact of Normal distribution, $\mathbb{P}\left(\eta_j \leqslant w\right) = \mathbb{P}(\eta_j/\sigma \leqslant w/\sigma) = \Phi(w/\sigma)$. We have

$$\mathbb{P}(\max_{j\in[M]} \eta_j \geqslant w) = 1 - \mathbb{P}(\max_{j\in[M]} \eta_j \leqslant w) = 1 - \mathbb{P}(\forall j \in [M], \eta_j \leqslant w) = 1 - \left[\Phi\left(\frac{w}{\sigma}\right)\right]^M.$$

□

Proposition E.12. *Let the UCB sequence be*

$$U = (\mu_t(a,b) + \sqrt{\beta_t'}\sigma_t(a,b) : t \in \mathbb{N}).$$

Let $M = \frac{\log(\mathcal{A}\sqrt{T})}{\log\frac{1}{\Phi^2(1)}}$,

$$\mathbb{P}(\neg\mathcal{E}_t(\tilde{R}, U, B_t)) \leqslant \frac{1}{\sqrt{T}}.$$

Proof. Recall that the optimistic Thompson sampling estimator $\tilde{R}$ is generated by

$$\tilde{f}_{t+1}^{\mathrm{OTS}}(a, B_t) := (\max_{j\in[M]} z_{t+1}^j) \cdot \sqrt{\beta_t'}\sigma_t(a,B_t) + \mu_t(a,B_t) \text{ and } \tilde{R}_{t+1}(a) = \mathrm{clip}_{[-c,c]}\left(\tilde{f}_{t+1}^{\mathrm{OTS}}(a,B_t)\right), \forall a \in \mathcal{A}$$

For any fixed $a \in \mathcal{A}$,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_t(\tilde{R}, U, B_t) \mid H_t, B_t) &= \mathbb{P}(\tilde{R}_{t+1}(a) \geqslant U_t(a,B_t), \forall a \in \mathcal{A} \mid H_t, B_t) \\
&\overset{(i)}{\geqslant} \mathbb{P}(\tilde{R}_{t+1}(a) \geqslant \min\{\mu_t(a,B_t) + \sqrt{\beta_t'}\sigma_t(a,B_t), c\}, \forall a \in \mathcal{A} \mid H_t, B_t) \\
&\overset{(ii)}{\geqslant} \mathbb{P}\left(\left(\max_{j\in[M]} z_{t+1}^j\right)\sqrt{\beta_t'}\sigma_t(a,B_t) \geqslant \sqrt{\beta_t'}\sigma_t(a,B_t), \forall a \in \mathcal{A} \mid H_t, B_t\right) \\
&\overset{(iii)}{=} 1 - \mathcal{A}\Phi\left(1\right)^M,
\end{aligned} \tag{34}$$

where (i) is by the definition of $\mathcal{E}(c)$ and $\mathcal{F}_t$ and the fact B_t is conditionally independent of f_θ given H_t; (ii) is due to $\mathrm{clip}_{[-c,c]}(x) \geqslant \min(x,c)$ and the function $min(x,c)$ is non-decreasing function on x; (iii) is by the fact $\{z_{t+1}^j\}_{j\in[M]}$ is independent of $H_{t+1} = (H_t, A_t, B_t, R_{t+1,A_t,B_t})$, the fact that $\sigma_t(a,b)$ and $w_t(a,b)$ are deterministic given H_t and Lemma E.11. Solve $\mathcal{A}\Phi(1)^M = 1/\sqrt{T}$, we have $M = \frac{\log(\mathcal{A}\sqrt{T})}{\log\frac{1}{\Phi(1)}}$. □

Lemma E.13 (Anti-concentration property of maximum of Gaussian R.V.)**.** *Consider a normal distribution $N\left(0,\sigma^2\right)$ where σ is a scalar. Let $\eta_1, \eta_2, \ldots, \eta_M$ be M independent samples from the distribution. Then for any $\delta > 0$*

$$\mathbb{P}\left(\max_{j\in[M]} \eta_j \leqslant \sqrt{2\sigma^2 \log(M/\delta)}\right) \geqslant 1-\delta.$$

Proposition E.14. *Let the UCB sequences be*

$$U' = (\mu_t(a,b) + \sqrt{2\log(M\sqrt{T})\beta_t'}\sigma_t(a,b) : t \in \mathbb{N}).$$

Then

$$\mathbb{P}(\neg\mathcal{E}_t^c(\tilde{R}, U', A_t, B_t)) \leqslant \frac{1}{\sqrt{T}}.$$

Proof. This is a direct consequence of Lemma E.13 under conditional probability given H_t, A_t, B_t with setting $\delta = 1/\sqrt{T}$. □

E.4. Bounding Estimation Regret via Information-theoretic quantity

To characterize the property of game environments and how much information algorithm can acquire about the environment at each round, we define the information ratio of algorithm $\pi = (\pi_t)_{t\in\mathbb{N}}$,

From Equation (33), we have

$$\beta_t I_t(\theta; R_{t+1,A_t,B_t} \mid A_t = a, B_t = b) \geqslant \beta_t' \sigma_t^2(a,b). \tag{35}$$

where

$$\beta_t = \frac{2\beta_t'}{\log(1+\sigma_w^{-2})}.$$

Then,

$$\begin{aligned}
(U_t'(A_t,B_t) - L_t(A_t,B_t)) &\leqslant \left(\sqrt{2\log(M\sqrt{T})}+1\right)\sqrt{\beta_t'}\sigma_t(A_t,B_t) \\
&\leqslant \left(\sqrt{2\log(M\sqrt{T})}+1\right)\sqrt{\beta_t I_t(\theta; R_{t+1,A_t,B_t} \mid A_t, B_t)} \\
&= \left(\sqrt{2\log(M\sqrt{T})}+1\right)\sqrt{\beta_t I_t(\theta; A_t, B_t, R_{t+1,A_t,B_t})}
\end{aligned}$$

Then immediately from Proposition E.8 and Cauchy–Bunyakovsky–Schwarz inequality.

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T-1} U_t'(A_t,B_t) - L_t(A_t,B_t)\right] &\leqslant \mathbb{E}\left[\sum_{t=0}^{T-1}\sqrt{\left(\sqrt{2\log(M\sqrt{T})}+1\right)^2 \beta_t I_t(\theta; A_t, B_t, R_{t+1,A_t,B_t})}\right] \\
&\leqslant \sqrt{\mathbb{E}\left[\sum_{t=0}^{T-1}\left(\sqrt{2\log(M\sqrt{T})}+1\right)^2\beta_t\right]}\sqrt{\mathbb{E}\left[\sum_{t=0}^{T-1} I_t(\theta; A_t, B_t, R_{t+1,A_t,B_t})\right]}
\end{aligned}$$

Let $Z_t = (A_t, B_t, R_{t+1,A_t,B_t}))$, then

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T-1} I_t(\theta; A_t, B_t, R_{t+1,A_t,B_t})\right] &= \mathbb{E}\left[\sum_t I_t(\theta; Z_t)\right] = \sum_{t=0}^{T-1} I(\theta; Z_t \mid Z_0, \ldots, Z_{t-1}) \\
&= I(\theta; Z_0, \ldots, Z_{T-1}) = I(\theta; H_T)
\end{aligned}$$

Then,

$$\mathbb{E}\left[\sum_{t=0}^{T-1} U_t'(A_t, B_t) - L_t(A_t, B_t)\right] \leqslant \sqrt{T\left(\sqrt{2\log(M\sqrt{T})}+1\right)^2 \beta_t I(\theta; H_T)}$$

Plugin the choice of $M = \frac{\log(\mathcal{A}\sqrt{T})}{\log \frac{1}{\Phi(1)}}$ and $\beta_t = 2\log \mathcal{A}\sqrt{T}$ and we have final result

$$\left(\sqrt{2\log(M\sqrt{T})}+1\right)^2 \beta_t = \mathcal{O}(\log \mathcal{A}T(\log T + \log\log \mathcal{A}T)).$$

E.4.1. Bounds on the information gain $I(\theta; H_T)$.

Remark E.15. An important property of the Gaussian distribution is that the information gain does not depend on the observed rewards. This is because the posterior covariance of a multivariate Gaussian is a deterministic function of the points that were sampled. For this reason, this maximum information ratio γ_T in Definition 4.4 is well defined. That is, $I(\theta; H_T) = I(\theta; A_0, B_0, \ldots, A_{T-1}, B_{T-1}) \leqslant \gamma_T$.

We adopt the results from (Srinivas, Krause, Kakade, and Seeger, 2009) which gives the bounds of γ_T for a range of commonly used covariance functions: finite dimensional linear, squared exponential and Matern kernels.

Example E.1 (Finite dimensional linear kernels)**.** *Finite dimensional linear kernels have the form $k(x, x') = x^\top x'$. GPs with this kernel correspond to random linear functions $f(x) = \theta^\top x, \theta \sim N(0, \sigma_0 I)$.*

Example E.2 (Squared exponential kernel)**.** *The Squared Exponential kernel is $k(x, x') = \exp(-(2l^2)^{-1} \|x - x'\|^2 s)$, l is a lengthscale parameter. Sample functions are differentiable to any order almost surely.*

Example E.3 (Matern kernel)**.** *The Matern kernel is given by $k(x, x') = (2^{1-\nu}/\Gamma(\nu)) r^\nu B_\nu(r)$ and $r = (\sqrt{2\nu}/l)\|x - x'\|$, where ν controls the smoothness of sample paths (the smaller, the rougher) and B_ν is a modified Bessel function. Note that as $\nu \to \infty$, appropriately rescaled Matern kernels converge to the Squared Exponential kernel.*

Table 3. Maximum information gain γ_T.

Kernel	Linear	Squared exponential	Materns ($\nu > 1$)
γ_T	$\mathcal{O}(d \log T)$	$\mathcal{O}((\log T)^{d+1})$	$\mathcal{O}(T^{d(d+1)/(2\nu+d(d+1))}(\log T))$

F. Experiments

F.1. Performance metric

Two performance metrics are utilized to evaluate the algorithms: average regret and KL divergence to Nash equilibrium.

Average regret: The expected regret of player i over T time steps is defined as:

$$\text{Regret}^i(T) = \frac{1}{T} \max_{a \in \Delta^{\mathcal{D}(\mathcal{A}^i)}} \mathbb{E}\left[\sum_{t=1}^{T} \phi\left(a, x_t^{-i}\right) - \phi\left(x_t^i, x_t^{-i}\right)\right], \tag{36}$$

where a and x_t^i represent the strategy for player i and x_t^{-i} denotes the strategies of all other players. $\mathcal{D}(\mathcal{A}^i)$ is the action set for player i. The expectation is taken over the randomness of the algorithm and the environment.

Duality Gap & **KL divergence to Nash:** The duality gap for a strategy pair (x, y) is defined as

$$\text{Gap}(x, y) = \max_{(x', y') \in \Delta} \mathbb{E}\left[\phi\left(x', y\right) - \phi\left(x, y'\right)\right]$$

The duality gap provides a measure of how close a solution pair is to a Nash equilibrium. If a solution pair (x, y) has a duality gap of ϵ, it is considered an ϵ-Nash Equilibrium.

After T iterations, the average-iterate strategies are defined as:

$$\overline{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t, \quad \overline{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t.$$

The KL divergences $KL(\overline{x}_T, x^\rangle$ and $KL(\overline{y}_T, y^\rangle$ are also used as performance metrics for comparing the average-iterate solution pair $(\overline{x}_T, \overline{y}_T)$ with a Nash equilibrium pair $(x^\cdot y^\rangle$.

F.2. Different opponents in matrix game

Four types of opponents in random matrix games are introduced, and the performance of different algorithms against these opponents is compared.

Self-play opponent: The opponent uses the same algorithm as the player.

Best-response opponent: The strategy for a best-response opponent is defined as:

$$y^* = \arg\min_{y\in\Delta} y^T(Ax), \tag{37}$$

which implies that the opponent knows matrix A and the player's strategy x at each round.

Stationary opponent: The stationary opponent always samples an action from a fixed strategy.

Non-stationary opponent: A non-stationary opponent changes its strategy randomly every round.

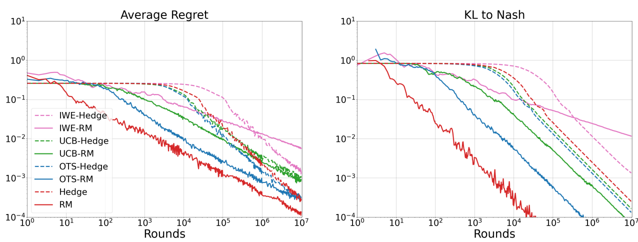
F.3. Additional results for random matrix games

This section presents additional evaluations of different algorithms on two-player zero-sum matrix games. The experiments consider payoff matrices where each entry is an i.i.d. random variable generated from the uniform distribution $[-1, 1]$. Each player has M actions, resulting in a squared payoff matrix of size M. The total number of rounds is set to $T = 10^7$. In each round, the players receive noisy rewards $\pm\tilde{r}_t$, where $\tilde{r}t = Aij + \epsilon_t$, and $\epsilon_t \sim \mathcal{N}(0, 0.1)$. The experiments investigate different matrix sizes, specifically $M = 5, 10, 20, 50, 70, 100$, and for each choice of M, 100 independent simulation runs are conducted. The performance of the algorithms is averaged over these simulation runs.

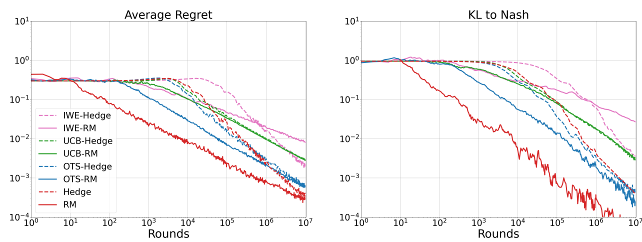
Self-play opponent First, we compare different algorithms under the self-play setting, where both players employ the same algorithm. Convergence curves of two performance metrics (see Appendix F.1) are shown in Figure 7, where each subplot (a)-(f) corresponds to a different matrix size M. The results indicate that algorithms exploiting the game structure outperform the two IWE baselines, particularly for smaller matrix sizes. As the matrix dimension increases, the performance gap between the proposed algorithms and the baselines diminishes. Among the algorithms, those based on the OTS method exhibit faster convergence than the ones based on UCB. Additionally, the average regret of RM decreases earlier than that of Hedge.

Best-response opponent In this subsection, we introduce a best-response opponent (see Appendix F.2), while the player continues to use the various algorithms described above. Figure 8 presents the results for different matrix sizes. We observe that all the algorithms in our proposed framework outperform the IWE baselines. Once again, the OTS-based algorithms demonstrate a faster convergence behavior compared to UCB.

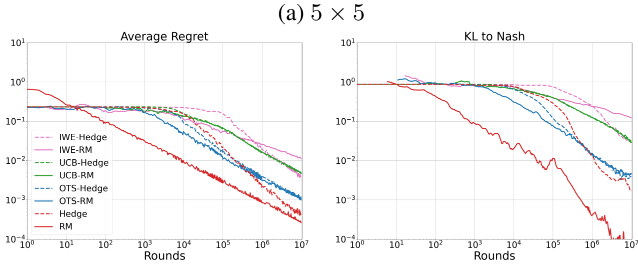
Stationary opponent Here, we consider a stationary opponent whose strategy remains fixed as a probability simplex over the action space, with values generated from a uniform distribution. The average regret in this scenario reflects the ability to exploit the opponent's weakness. Convergence curves of the two performance metrics are compared in Figure 9. The results clearly demonstrate that the OTS estimator provides a significant advantage in exploiting this weak opponent, in comparison to the IWE-based estimators.



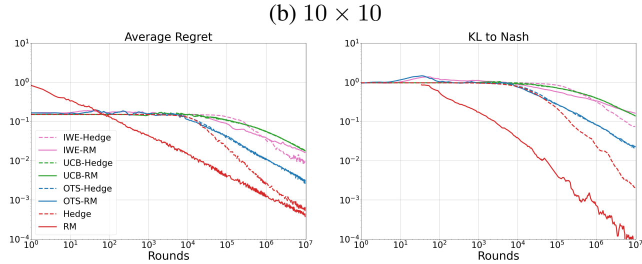
(a) 5×5



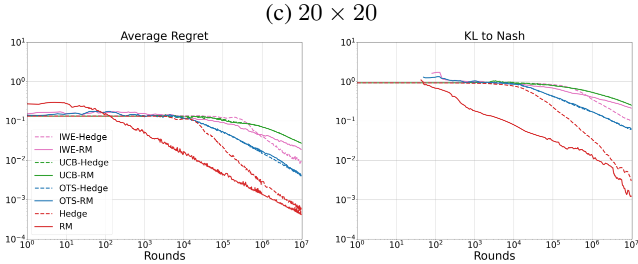
(b) 10×10



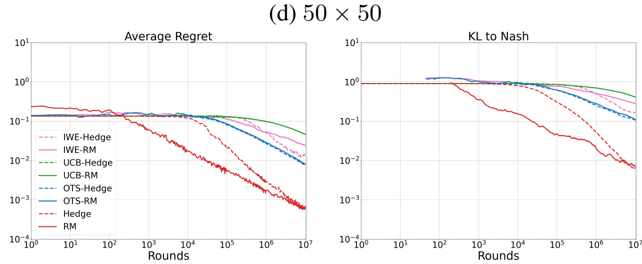
(c) 20×20



(d) 50×50

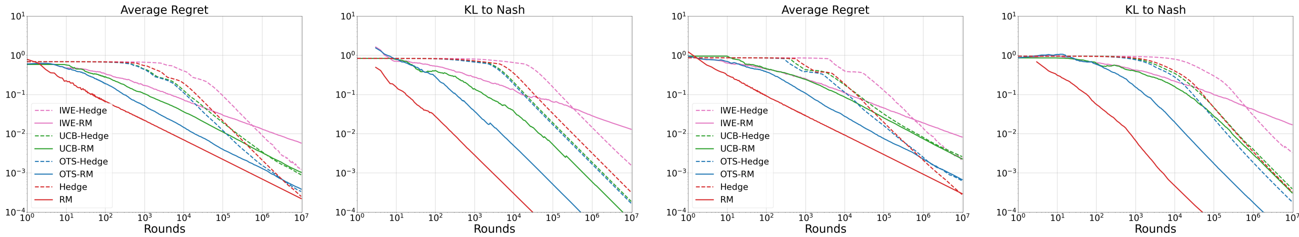


(e) 70×70

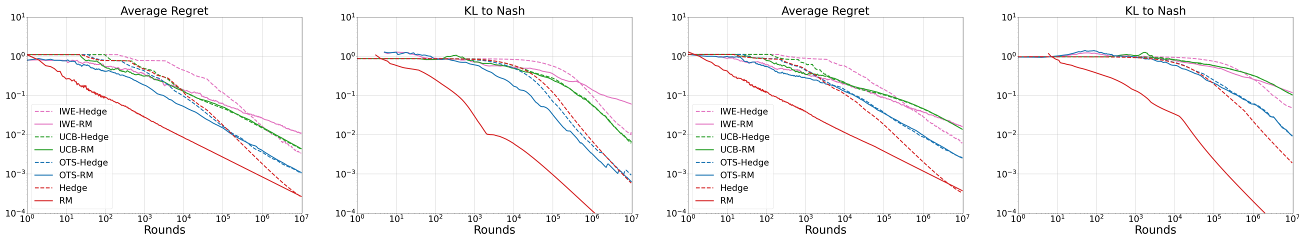


(f) 100×100

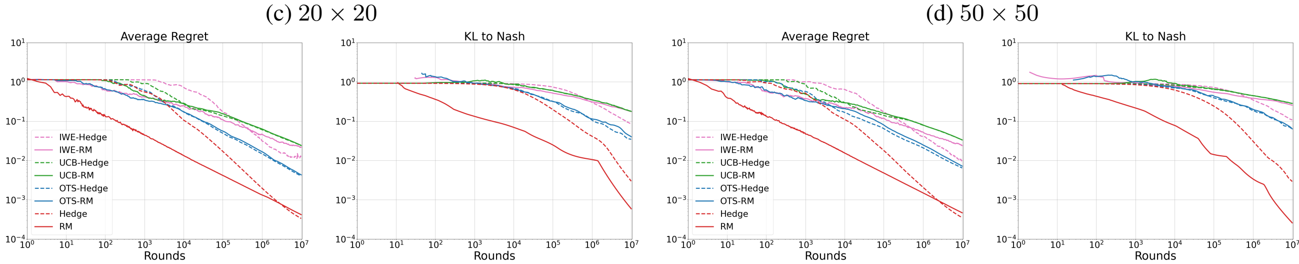
Figure 7. Self-play on different random matrix.



(a) 5×5 (b) 10×10

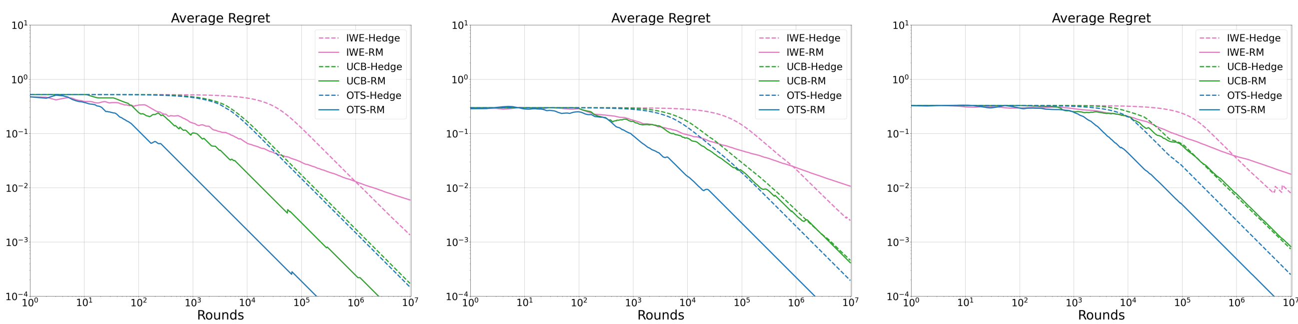


(c) 20×20 (d) 50×50

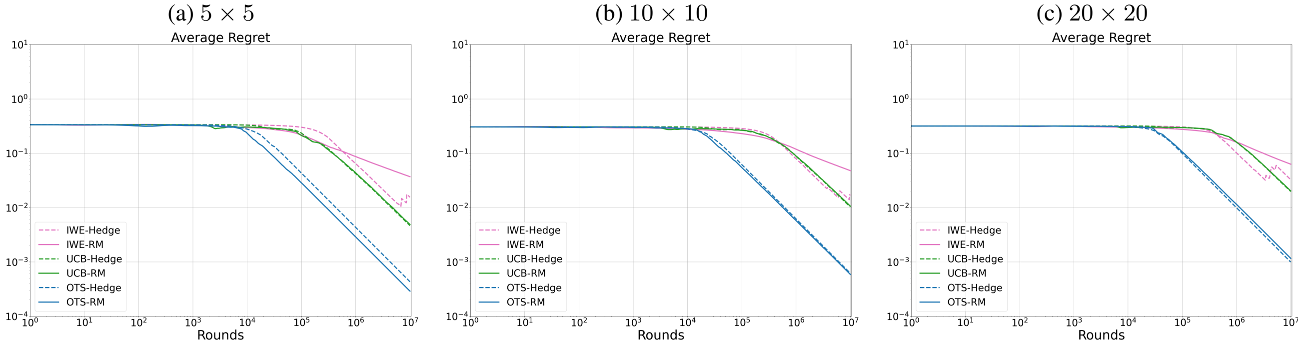


(e) 70×70 (f) 100×100

Figure 8. Best-response opponent on different random matrix.



(a) 5×5 (b) 10×10 (c) 20×20



(d) 50×50 (e) 70×70 (f) 100×100

Figure 9. Stationary opponent on different random matrix.

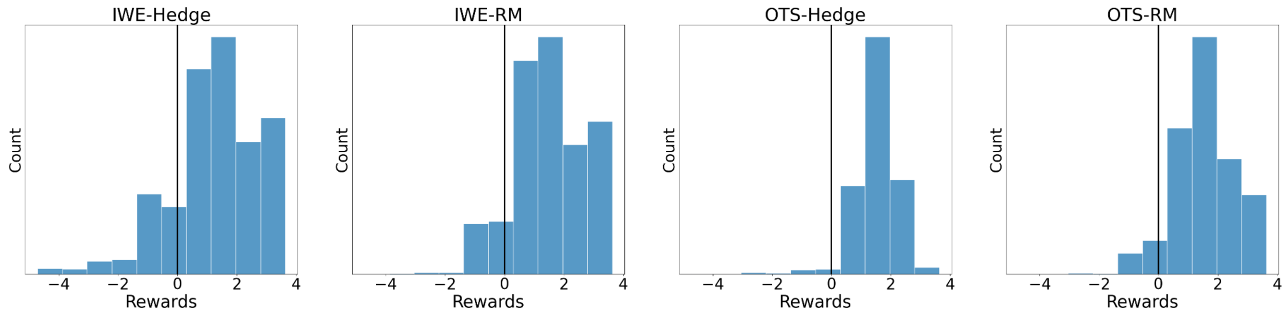


Figure 10. Reward histograms of different algorithms against a non-stationary opponent.

Non-stationary opponent In this subsection, we introduce a non-stationary opponent, requiring the player to develop a robust strategy against all possible opponent's strategies. Specifically, a game matrix $A \in \mathbb{R}^{10\times 5}$ is generated, with each element sampled from $\mathcal{N}(0.5, 2.0)$. The opponent's actions are drawn from a fixed strategy that randomly changes every 50 rounds. Each algorithm is evaluated over 1000 rounds, and 100 simulation runs are conducted. Figure 10 presents histograms of rewards over all rounds and simulation runs, while Table 4 summarizes the percentage of negative rewards and the mean reward values. The results show that OTS has a smaller percentage of negative rewards and achieves higher mean rewards compared to IWE, indicating its superior robustness.

	return< 0	mean return
IWE-Hedge	19.4%	1.24
IWE-RM	12.6%	1.50
OTS-Hedge	2.5%	1.55
OTS-RM	8.8%	1.55

Table 4. Returned rewards of IWE and OTS against a non-stationary opponent.

F.4. Convergence rate related with dimensions

The average regret bounds for IWE-Hedge and the proposed algorithms are $\tilde{\mathcal{O}}(\sqrt{(M+N)/T})$ and $\tilde{\mathcal{O}}(\sqrt{MN/T})$, respectively. Taking OTS-Hedge as an example, for a fixed iteration T, IWE-Hedge implies $\log(\text{average regret}) \propto (1/2)\log(M+N)$, while OTS-Hedge indicates $\log(\text{average regret}) \propto \frac{1}{2}\log(MN)$. When $M = N$, the logarithmic average regret for IWE-Hedge and TS-Hedge should increase with respect to M at the rates of $1/2$ and 1, respectively. The experimental results shown in Figure 11, where M ranges from 2 to 100 with a fixed $T = 10^7$, match the theoretical predictions almost exactly. These empirical results provide strong support for our regret analysis.

F.5. Radar anti-jamming problem

The competition between radar and jammer is an important issue in modern electronic warfare, which can be viewed as a non-cooperative game with two players. This competition occurs at the signal level, where both the radar and jammer can change parameters of their transmitted signals. One representative game form is playing in the frequency domain, as the signal with different carrier frequencies is disjoint.

In our example, we consider the pulse radar and the noise-modulated jammer. The radar transmits pulse signals one by one, with a waiting time interval between consecutive pulses. At the beginning of each pulse, both the radar and jammer transmit their own signals. After a short signal propagation time, each player receives their opponent's signal and obtains a reward. In our setting, both the radar and jammer have three candidate carrier frequencies, denoted as $\mathcal{F} = \{f_1, f_2, f_3\}$. The radar player has three sub-pulses in each radar pulse, and each sub-pulse can choose a different carrier frequency. The action set of the radar is denoted as $\mathcal{A}_R = \mathcal{F}\times\mathcal{F}\times\mathcal{F}$, which has a total of 27 different actions. On the other hand, the jammer player can choose one carrier frequency to transmit the jamming signal and change the carrier frequency for different radar pulses. The action set of the jammer is denoted as $\mathcal{A}_J = \mathcal{F}$. After each iteration between the radar and jammer, the radar obtains a signal-inference-and-noise ratio (SINR), which serves as the reward for the radar in that round.

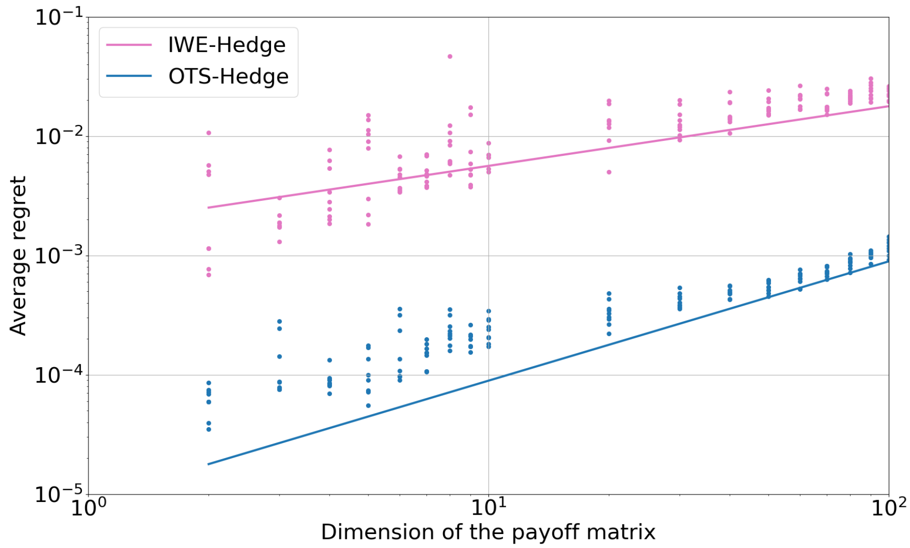


Figure 11. Convergence rate with different dimensions.

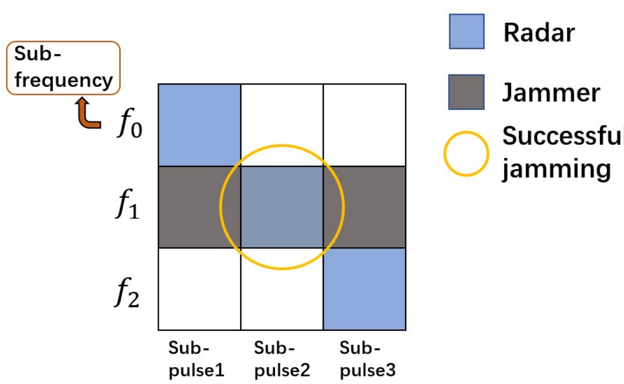


Figure 12. Illustration of the game between radar and jammer for one pulse.

We can observe that both the radar and jammer's actions are related to the frequency set $\mathcal{F} = \{f_1, f_2, f_3\}$. Therefore, the reward in each round can be further defined as a linear function, making the anti-jamming scenario a linear game for the radar.

Definition F.1. The reward function in the anti-jamming problem is defined as follows:

$$\text{SINR}(a, b; \theta) = \phi(a, b)^T \theta$$

Here, $\phi(a, b) = P_a(\theta)/(P_{n_0} + P_b \mathbb{1}(a = b))$ represents a known feature mapping that maps the actions of the radar and the jammer, denoted as (a, b), into the frequency domain related with θ. The parameter θ corresponds to the radar cross section (RCS) associated with the frequencies in the anti-jamming scenario. $P_a(\theta)$ is radar's received power related with θ, P_{n_0} is the noise power and P_b is received jammer's power. The indicator function $\mathbb{1}(a = b)$ evaluates to 1 if the radar and jammer transmit on the same carrier frequency $(a = b)$ and 0 otherwise.

This definition captures the essence of the anti-jamming problem, allowing us to evaluate the performance of different strategies and algorithms based on the signal-to-interference-plus-noise ratio (SINR), and construct the anti-jamming scenario as a linear game.

Adaptive jammer In the case of an adaptive jammer, it takes actions according to the radar's latest 10 pulses. Specifically, it counts the numbers of different carrier frequencies ($f_1, f_2, f_3 \in \mathcal{F}$) appearing in the radar's last 10 pulses, denoted as

Table 5. Parameters of FA radar and jammer.

Parameter	Value
radar transmitter power P_T	30kW
radar transmit antenna gain G_T	32dB
radar initial frequency f_0	3GHz
bandwidth of each subpulse B	2MHz
distance between the radar and the jammer	100km
false alarm rate p_f	1×10^{-4}
jammer transmitter power P_J	100W
jammer transmit antenna gain G_J	15dB

N_1, N_2, and N_3. The jammer then takes action f_i with a probability proportional to N_i for $i = 1, 2, 3$.

F.6. Repeated traffic routing problem

This section considers the traffic routing problem in the transportation literature, which is defined over a directed graph $(\mathcal{V}, \mathcal{E})$ ($\mathcal{V}$ and $\mathcal{E}$ are vertices and edges sets respectively) and modeled as a multi-player game. Each node pair (referred as one origin node and one destination node) in the graph is treated as an individual player and every player i seeks to find the 'best' route (consists of several edges) to send U^i units from its origin node to its destination node. The quality of the chosen route is measured by the travel time, which depends on the total occupancy of the traversed edges. If one edge is occupied by more players, more travel time of this edge is. Specifically, the travel time t_e of edge $e \in \mathcal{E}$ is a function of the total units u traversing e. One common choice is the Bureau of Public Roads function (Leblanc, 1975)

$$t_e(u) = c_e(1 + 0.5(\frac{u}{C_e})^4),$$

where c_e and C_e are free-flow and capacity of edge e respectively. The action of player i is denoted as $a^i \in \mathcal{A}^i \subset \mathbb{R}^{|\mathcal{E}|}$ and the component corresponding to edge e is denoted as $[a^i]_e$. If edge e belongs to the route, $[a^i]_e = U^i$ and otherwise $[a^i]_e = 0$. Further, let $a^{-i} \in \mathcal{A}^{-i}$ be the action of other players, the total occupancy of edge e is $[a^i]_e + [g(a^{-i})]_e$, where $g(a^{-i}) = \sum_{j \neq i} a^j$. This way, the total travel time of a joint action $\boldsymbol{a} = (a^i, a^{-i})$ for player i can be expressed as

$$\ell^i(a^i, a^{-i}) = \sum_{e \in \mathcal{E}} [a^i]_e t_e([a^i]_e + [g(a^{-i})]_e).$$

The reward function of player i is $r^i(a^i, a^{-i}) = -\ell^i(a^i, a^{-i})$. Note that the mathematical form of the reward function is unknown to players, only values of $r^i(a^i, a^{-i})$ and actions (a^i, a^{-i}) can be observed by players.

In our experiment, the Sioux-Falls road network data set (Bar-Gera, 2015) is used and we set c_e and C_e (Bar-Gera, 2015). This network is a directed graph with 24 nodes and 76 edges and there are in total $N = 528$ players. Each player i's action space $\mathcal{A}^i$ is specified by the 5 shortest routes and any route that more than three times longer than the shortest route is further removed from $\mathcal{A}^i$. To exploit the correlations among actions (a^i, a^{-i}) in the reward function, the composite kernel proposed in (Sessa et al., 2019) is used. For player i, let $a^i, b^i \in \mathcal{A}^i$ and $a^{-i}, b^{-i} \in \mathcal{A}^{-i}$, then the kernel function associating (a^i, a^{-i}) and (b^i, b^{-i}) is

$$K^i((a^i, a^{-i}), (b^i, b^{-i})) = k_{\mathrm{L}}(a^i, b^i) k_{\mathrm{P}}(a^i + g(a^{-i}), b^i + g(b^{-i})),$$

where $k_{\mathrm{L}}(\cdot, \cdot)$ and $k_{\mathrm{P}}(\cdot, \cdot)$ are linear and polynomial kernels respectively. The hyperparameters of kernels are set the same as in (Sessa et al., 2019). Details on GP for estimating reward functions can be found in Section 2 of (Sessa et al., 2019) and we do not repeated them here[4]. Except the regret, the congestion of edge e is also included as a performance metric for a joint action $\boldsymbol{a}$, which is computed as $0.15(\sum_{i=1}^{N} [a^i]_e / C_e)^4$. The averaged congestion is obtained by averaging congestion over all edges.

[4]The implementation refers the code released by authors of (Sessa et al., 2019) at `https://github.com/sessap/stackelucb`