No-Regret Learning in Unknown Game with Applications

Yingru Li
yingruli@link.cuhk.edu.cn

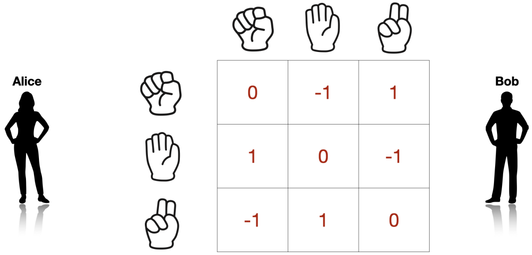
The Chinese University of Hong Kong, Shenzhen, China

August 23, 2022

Outline

Matrix game with known utilities

- Matrix game: Foundation of game theory [Neumann and Morgenstern (44')].
- Traditional goal: find Nash equilibrium
- Known utilities in advance: Linear Programming ; One-shot game.
- Not realistic in many applications.

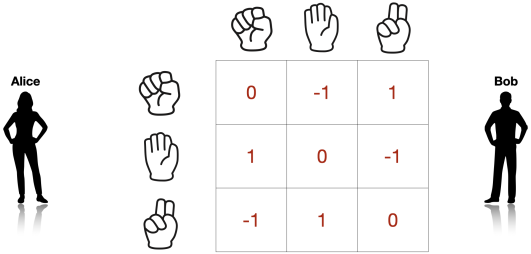


0	-1	1
1	0	-1
-1	1	0

Utility matrix of Alice f_θ

Matrix game with known utilities

- Matrix game: Foundation of game theory [Neumann and Morgenstern (44')].
- Traditional goal: find Nash equilibrium
- Known utilities in advance: Linear Programming ; One-shot game.
- Not realistic in many applications.



Utility matrix of Alice f_θ

Unknown Game: Matrix game with unknown utilities

- Reality: Outcome of the game revealed after playing;
- One-shot game is hopeless.
- Reality: Bob may not be truly adversarial;
- Alice can play better than Nash.
- In repeated play, Alice can hope to learn to play well against the particular opponent (Bob) being faced.

Alice

	b_1	b_2	b_3
a_1	?	?	?
a_2	?	?	?
a_3	?	?	?

Utility matrix of Alice f_θ

Bob

Unknown Game: Matrix game with unknown utilities

- Reality: Outcome of the game revealed after playing;
- One-shot game is hopeless.
- Reality: Bob may not be truly adversarial;
- Alice can play better than Nash.
- In repeated play, Alice can hope to learn to play well against the particular opponent (Bob) being faced.

Alice

	b_1	b_2	b_3
a_1	?	?	?
a_2	?	?	?
a_3	?	?	?

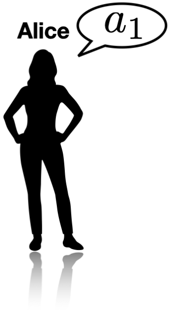
Utility matrix of Alice f_θ

Bob

Repeated Unkonwn Game with full information feedback

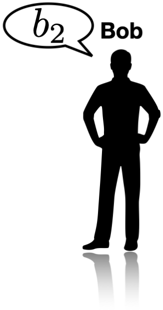
[Freund and Schapire. (99'), Hart and Mas-Colell (00'), Games and Economic Behavior.]

At round t, Alice select a_1 and observe the column b_2 which contains the entries other than a_1.



	b_1	b_2	b_3
a_1	?	$f_\theta(a_1, b_2)$	?
a_2	?	$f_\theta(a_2, b_2)$	?
a_3	?	$f_\theta(a_3, b_2)$	?

Utility matrix of Alice f_θ

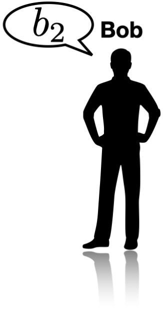


Repeated Unkonwn Game with bandit information feedback

Our practical setup: the only feedback at round t is Noisy bandit feedback + Opponent's action.
Noise W comes from environment's random effect.

	b_1	b_2	b_3
a_1	?	$\boxed{f_\theta(a_1, b_2) + W}$	?
a_2	?	?	?
a_3	?	?	?

Utility matrix of Alice f_θ



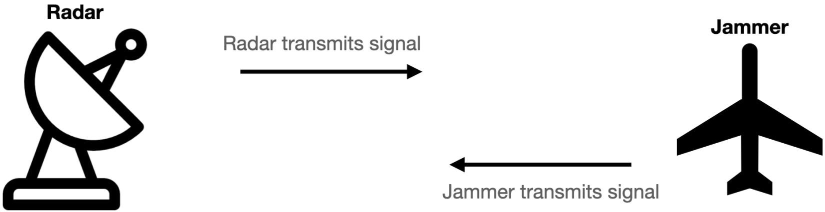
Applictions

- Applications in Signal Processing: e.g. Radar Anti-Jamming

- Applications in Transportation: e.g. Traffic Routing

Application in Signals: Radar Anti-Jamming

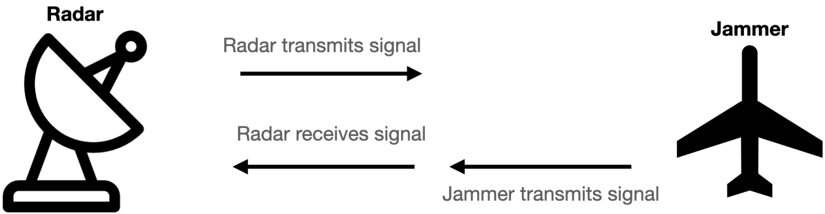
- Scenario: Radar aims to detect the target with sequence of signals (repeated game playing) while jammer aims to prevent.
- Action set: frequencies $\{f_0, f_1, \ldots, f_N\}$.
- Utility of Radar: the detection probability on the target.
- Environment randomness: channel and system noise.



Application in Signals: Radar Anti-Jamming

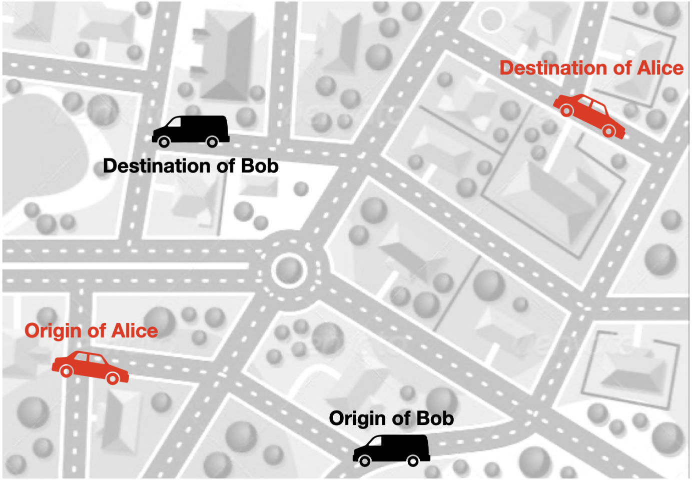
Feedback for Radar:

1. Radar receives echo signal + jamming signal
2. Opponent's action: Frequency of jamming signal extracted from received signals (e.g. FFT)
3. Noisy bandit feedback: Utility can be estimated from received signals



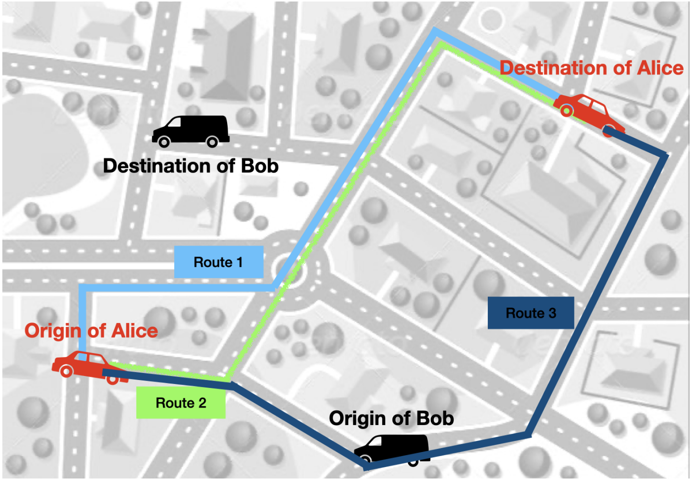
Application in Transportation: Traffic Routing

Simplified scenario: every morning 7:00 in August (repeated games), Alice and Bob choose the routes and start to deliver fix unit of products back and forth from fixed origin to destination.



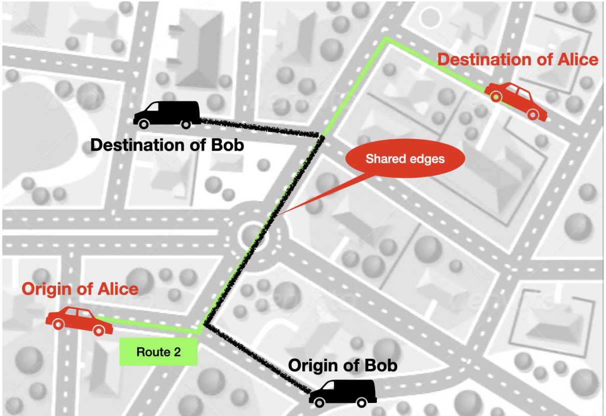
Application in Transportation: Traffic Routing

Action set of Alice: {Route 1, Route 2, Route 3}.



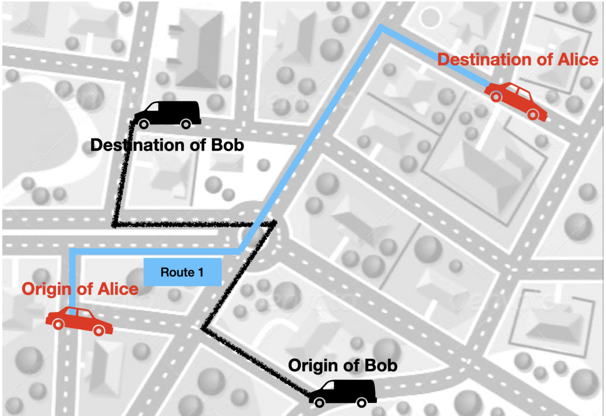
Application in Transportation: Traffic Routing

- Day 1: Alice selects route 2 and bob selects the black route,
- Feedback: incurred total travel time during the day (Noisy bandit feedback) and get informed of bob's chosen route (Opponent's action)
- Because of the shared edge between Alice and Bob's routes, the incurred travel time is long.



Application in Transportation: Traffic Routing

- Day 2: Alice learns from day 1 and selects route 1, and Bob select the black route.
- Feedback: incurred total travel time during the day (Noisy bandit feedback) and get informed of bob's chosen route (Opponent's action)
- Alice suffer less time because of no shared edge in day 2.



Summary and formulation

- Formal protocol: At each round t in the repeated game, Alice selects A_t and Bob selects B_t; Then, Alice received R_{t+1,A_t,B_t} and observe B_t ... (next round)

Round t	Notation	Radar	Traffic
Action	$A_t \in \mathcal{A}$	Frequency	Routes
Others' action	$B_t \in \mathcal{B}$	Frequency	Routes
Bandit feedback for Alice	R_{t+1,A_t,B_t}	Probability of detection	Incurred travel time
Additional observations for Alice	B_t	Jammer's frequency extracted from received signals	Other agent' selected routes
Alice's Objective	$\sum_{t=0}^{T-1} \mathbb{E}[R_{t+1,A_t,B_t}]$	Max Probability of detection in many rounds of game	Min Total travel time in a month

Relation to existing problem setups and popular algorithms

Setup: **Full information game**

- Relation: full column vector instead of just one entry.

- Famous algorithms: Multiplicative-weights [Littlestone and Warmuth, 94'] / Hedge [Freund and Schapire, 97', 99'], Regret Matching [Hart and Mas-Colell, 00']

- Drawback of the setup: Feedback not realistic in applications.

Relation to existing problem setups and popular algorithms

Setup: **Adversarial bandit**

- Relation: if opponent is fully adversarial and we cannot observe opponent's action.
- Famous algorithm: EXP3 [Auer, Cesa-Bianchi, Freund, Schapire. 03' SIAM J. Comput.] and its variants [Bubeck, Lee, Lee, Eldan. 17' STOC]
- Drawback of the setup: Ignore the fact the underlying utility function is static during the game and may have structure among actions.

Relation to existing problem setups and popular algorithms

Setup: **Stochastic bandit**

- Relation: if opponent is stationary and we cannot observe opponent's action.
- Famous algorithm: Thompson sampling (TS) [Thomson, 33'; Russo, Van Roy, Kazerouni, Osband and Wen, 18' Foundations and Trends]
- Drawback of the setup: Opponent is usually smarter than just playing stationarily.

Outline

Quick recap and abstraction: Player-Environment-Player Interface

- For a environment instance indexed by θ,
- At each time $t = 0, 1, \ldots$,
  - Alice executes an action A_t; Simultaneously, Bob executes an action B_t; After execution,
  - Alice observes the reward,

(1) Full feedback (not realistic):

$$R_{t+1,a,B_t} = f_\theta(a, B_t), \quad \forall a \in \mathcal{A}$$

(2) Noisy bandit feedback (realistic):

$$R_{t+1,A_t,B_t} = f_\theta(A_t, B_t) + W_{t+1,A_t,B_t}$$

  - (Optional and realistic) Meanwhile, Alice can observe Bob's selected action B_t.
- Research interest: Noisy bandit feedback + Observe Opponent's action.

No-regret algorithms for full-information feedback

Algorithm No Regret for full-information feedback

1: Initialize $\mathcal{A}$-dim probability vector X_1
2: **for** round $t = 1, 2, \ldots, T$ **do**
3: Sample action A_t from distribution P_{X_t} according to X_t,
4: Observe full-information feedback $f_\theta(a, B_t)$ for all $a \in \mathcal{A}$
5: Update: $X_{t+1} = g_t(X_t, (f_\theta(a, B_t))_{a \in \mathcal{A}})$ with no-regret update g_t
6: **end for**

Two no-regret update algorithm can be applied to this scenario: (require $f_\theta(a, b)$ bounded)

- Hedge (97'): $X_{t+1,a} \propto X_{t,a} \exp(\eta_t f_\theta(a, B_t))$ for all $a \in \mathcal{A}$.
- Regret Matching (00'): see next page.

Regret Matching in full information feedback

- instantaneous regret vector $\mathrm{reg}_{t+1} \in \mathbb{R}^{\mathcal{A}}$

$$\mathrm{reg}_{t+1}(a) = f_\theta(a, B_t) - \sum_a f_\theta(a, B_t) X_{t,a}$$

- Cumulative regret vector Reg_{t+1}

$$Reg_{t+1}(a) = \sum_{s=0}^{t} \mathrm{reg}_{s+1}(a)$$

- Regret Matching update rule: If $\sum_a Reg_{t+1}^{+}(a) = 0$, choose arbitrary probability vector X_{t+1} (usually we choose uniform dist.); otherwise, $\forall a \in \mathcal{A}$,

$$X_{t+1,a} = \frac{Reg_{t+1}^{+}(a)}{\sum_a Reg_{t+1}^{+}(a)}, \text{ where } x^{+} := \max(x, 0).$$

History-dependent Randomized Algorithm for bandit feedack and oppnent's action observation

- Alice's experience through time t is encoded by a history

$$H_t = \left(A_0, B_0, R_{1,A_0,B_0}, \ldots, A_{t-1}, B_{t-1}, R_{t,A_{t-1},B_{t-1}}\right).$$

- An algorithm (randomized policy) employed by Alice is a sequence of deterministic functions,

$$\pi^{\text{alg}} = (\pi_t)_{t\in\mathbb{N}},$$

where $\pi_t(H_t)$ specifies a probability distribution over the action set $\mathcal{A}$,

- Alice select the action according to $A_t \sim \pi_t(H_t)$:

$$\mathbb{P}(A_t \in \cdot \mid \pi_t) = \mathbb{P}(A_t \in \cdot \mid H_t) = \pi_t(\cdot)$$

Objective function from Alice's perspective

- We also allow Bob use alg^B to select his actions $B_0, B_1, \ldots$
- Alice's objective is to maximize expected reward over some long duration T:

$$\sum_{t=0}^{T-1} \mathbb{E}\left[R_{t+1,A_t,B_t} \mid \theta\right]$$

- We compete with the best action in hindsight $A^* = \max_{a \in \mathcal{A}} \sum_{t=0}^{T-1} \mathbb{E}\left[R_{t+1,a,B_t} \mid \theta\right]$
- Naturally, our performance metric is (Adversarial) Regret:

$$\Re(T, \pi^{\mathrm{alg}}, alg^B, \theta) = \max_a \sum_{t=0}^{T-1} \mathbb{E}\left[R_{t+1,a,B_t} - R_{t+1,A_t,B_t} \mid \theta\right]$$

No-regret Learning in Game from Alice's perspective

- We say an algorithm π_{alg} is No-Regret, if for any possible algorithm alg^B used by Bob, Alice can suffer only sublinear regret, i.e.

$$\Re^*(T, \pi^{\text{alg}}, \theta) = \sup_{alg^B} \Re(T, \pi^{\text{alg}}, alg^B, \theta) = o(T),$$

- That is,

$$\lim_{T \to \infty} \frac{\Re^*(T, \pi^{\text{alg}}, \theta)}{T} = 0.$$

Estimation with history and then No-regret update in our setting

Algorithm Estimate-then-NoRegret for bandit feedback

```
Initialize $X_1$
for round $t = 1, 2, \ldots, T$ 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 $H_{t+1} = (H_t, A_t, B_t, R_{t+1})$ for player $i$
    Construct: $\tilde{R}_{t+1} = E(H_{t+1}) \in \mathbb{R}^{\mathcal{A}}$ by estimation function $E$,
    Update: $X_{t+1} = g_t(X_t, \tilde{R}_{t+1})$
end for
```

The price of bandit information compared with full information

Table: Regret bounds comparison.

Feedback		Full	Bandit	Bandit + Actions
Reward vector		direct from feedback	IWE	?
No-Regret Update	Hedge	$\mathcal{O}(\sqrt{T \log \mathcal{A}})$ [97', 99']	$\mathcal{O}(\sqrt{T \mathcal{A} \log \mathcal{A}})$ [00']	?
	RM	$\mathcal{O}(\sqrt{T \mathcal{A}})$ [03']	$\mathcal{O}(T^{2/3} \mathcal{A}^{2/3})$ [L, 20']	?

- Importance-weighted estimator (IWE) do not utilize the information of opponents' actions.
- With additional information of opponent's actions and the knowledge on reward structure $\mathcal{F} = \{f_\rho : \rho \in \Theta\}$, can we have better performance, theoretically and practically?

Outline

Natural attempt: Mean estimator and Thompson Sampling

- With history $H_{t+1} = (H_t, A_t, B_t, R_{t+1,A_t,B_t})$, assume gaussian prior,
- Mean estimator of the full vector by posterior mean μ_t summarizing history H_t

$$\tilde{R}^{\mu}_{t+1}(a) = \mathrm{clip}_{[0,1]}(\mu_t(a, B_t)), \forall a \in \mathcal{A}$$

- Thompson sampling estimator of the full vector by posterior mean μ_t and variance σ_t summarizing H_t

$$\tilde{f}^{TS}_{t+1}(a, B_t) \mid H_{t+1} \sim N(\mu_t(a, B_t), \sigma_t(a, B_t)), \forall a \in \mathcal{A}$$

and

$$\tilde{R}^{TS}_{t+1}(a) = \mathrm{clip}_{[0,1]}(\tilde{f}^{TS}_{t+1}(a, B_t)), \forall a \in \mathcal{A}$$

Simple example showing divergence of RM with Mean or TS

- Consider a class of matrix game instances where $\Delta \in (0, 1)$ is the gap variable

$$\theta = \begin{pmatrix} 1 & 1-\Delta \\ 1-\Delta & 1 \end{pmatrix},$$

- Best-response opponent: Bob have all information of the matrix θ and know Alice's mixed strategy X_t before choosing its own strategy Y_t and select $B_t \sim Y_t$,

$$Y_t = \arg\min_{y \in \Delta} y^T(\theta X_t),$$

- Assume no observation noise.

Proposition 1 (Divergence).

Alice using Regret Matching with Mean Estimator (Mean-RM) or Thompson Sampling Estimator (TS-RM) will suffer linear regret.

Simple example showing divergence of RM with Mean or TS

$$\theta = \begin{pmatrix} 1 & 1-\Delta \\ 1-\Delta & 1 \end{pmatrix}$$

- Observation 1: As long as a pure strategy is used by the Alice, it suffers regret Δ at that round because of the best-response opponent.
- Observation 2: The best-response strategy for the uniform mixed strategy is also the uniform strategy.

Simple example showing divergence of Mean and TS

$$\theta = \begin{pmatrix} 1 & 1-\Delta \\ 1-\Delta & 1 \end{pmatrix}$$

By symmetry, define the following event,

- Event ω_t: Alice picks the 2nd row and Bob chooses the 1st column at time t.
- Event Ω_t: Alice picks the 2nd row and Bob chooses the 1st column for all time $t' \leqslant t$.

Proposition 2.

If Alice initialize with uniform strategy,

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

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

Divergence of Mean-RM and TS-RM

Proposition 3 (Mean-RM).

If Alice initialize with uniform strategy and use Regret Matching with mean estimator (Mean-RM), for all round $t \geqslant 1$, $\mathbb{P}(\Omega_t) = 0.25$.

Proposition 4 (TS-RM).

If Alice initialize with uniform strategy and use Regret Matching with Thompson Sampling estimator (TS-RM), $\forall \Delta \in (0,1), \forall \sigma_w > 0, \exists c(\Delta, \sigma_w) > 0$, *for all round* $t \geqslant 1$,

$$\mathbb{P}(\Omega_t) \geqslant c(\Delta, \sigma_w).$$

Specifically, when $\Delta = 0.1$ *and* $\sigma_w = 0.1$ *(used to update posterior algorithmically), we have* $c(\Delta, \sigma_w) \approx 0.54$.

Divergence of Mean-RM and TS-RM

Proposition 3 (Mean-RM).

If Alice initialize with uniform strategy and use Regret Matching with mean estimator (Mean-RM), for all round $t \geqslant 1$, $\mathbb{P}(\Omega_t) = 0.25$.

Proposition 4 (TS-RM).

If Alice initialize with uniform strategy and use Regret Matching with Thompson Sampling estimator (TS-RM), $\forall \Delta \in (0, 1), \forall \sigma_w > 0, \exists c(\Delta, \sigma_w) > 0$, *for all round* $t \geqslant 1$,

$$\mathbb{P}(\Omega_t) \geqslant c(\Delta, \sigma_w).$$

Specifically, when $\Delta = 0.1$ *and* $\sigma_w = 0.1$ *(used to update posterior algorithmically), we have* $c(\Delta, \sigma_w) \approx 0.54$.

Divergence of Mean-RM and TS-RM

Corollary 1.

*As a corollary of proposition 2, proposition 3 and proposition 4, **Mean-RM** and **TS-RM** would suffer linear regret.*

Divergence of Mean-RM

- $\mathbb{P}(w_1) = 0.25$ by uniform strategy initialization
- Conditioned on ω_1, following the Mean-RM algorithm:
- Each round $t = 0, 1, \ldots,$
  - Alice received $R_{t+1,2nd,1st} = 1 - \Delta$
  - Use the mean estimator to construct imagined reward vector $\tilde{R}^{\mu}_{t+1} = [0,\ 1-\Delta]$
  - and construct the instantaneous and cumulative regret vector

$$\mathrm{reg}_{t+1} = [\Delta - 1,\ 0], \quad Reg_{t+1} = [\underbrace{(0.5 - t - 1)(1-\Delta)}_{\text{Negative}}, 0.5(1-\Delta)]$$

  - By regret matching update rule, Alice' strategy for next time $X_{t+1} = [0, 1]$ is still the pure strategy of 2nd row. This implies that Bob's next strategy is still 1st column.
- As a result, $\mathbb{P}(\Omega_t \mid w_1) = 1, \forall t \geqslant 1$, *i.e.*, Alice will always suffer linear regret.

Divergence of Mean-RM

- $\mathbb{P}(w_1) = 0.25$ by uniform strategy initialization
- Conditioned on ω_1, following the Mean-RM algorithm:
- Each round $t = 0, 1, \ldots,$
  - Alice received $R_{t+1,2nd,1st} = 1 - \Delta$
  - Use the mean estimator to construct imagined reward vector $\tilde{R}^{\mu}_{t+1} = [0,\ 1-\Delta]$
  - and construct the instantaneous and cumulative regret vector

$$\mathrm{reg}_{t+1} = [\Delta - 1,\ 0], \quad Reg_{t+1} = [\underbrace{(0.5 - t - 1)(1-\Delta)}_{\text{Negative}}, 0.5(1-\Delta)]$$

  - By regret matching update rule, Alice' strategy for next time $X_{t+1} = [0, 1]$ is still the pure strategy of 2nd row. This implies that Bob's next strategy is still 1st column.
- As a result, $\mathbb{P}(\Omega_t \mid w_1) = 1, \forall t \geqslant 1$, *i.e.*, Alice will always suffer linear regret.

Divergence of Mean-RM

- $\mathbb{P}(w_1) = 0.25$ by uniform strategy initialization
- Conditioned on ω_1, following the Mean-RM algorithm:
- Each round $t = 0, 1, \ldots,$
  - Alice received $R_{t+1,2nd,1st} = 1 - \Delta$
  - Use the mean estimator to construct imagined reward vector $\tilde{R}^{\mu}_{t+1} = [0,\ 1 - \Delta]$
  - and construct the instantaneous and cumulative regret vector

$$\text{reg}_{t+1} = [\Delta - 1,\ 0], \quad Reg_{t+1} = [\underbrace{(0.5 - t - 1)(1 - \Delta)}_{\text{Negative}}, 0.5(1 - \Delta)]$$

  - By regret matching update rule, Alice' strategy for next time $X_{t+1} = [0, 1]$ is still the pure strategy of 2nd row. This implies that Bob's next strategy is still 1st column.
- As a result, $\mathbb{P}(\Omega_t \mid w_1) = 1, \forall t \geqslant 1$, *i.e.*, Alice will always suffer linear regret.

Divergence of Mean-RM

- $\mathbb{P}(w_1) = 0.25$ by uniform strategy initialization
- Conditioned on ω_1, following the Mean-RM algorithm:
- Each round $t = 0, 1, \ldots,$
  - Alice received $R_{t+1,2nd,1st} = 1 - \Delta$
  - Use the mean estimator to construct imagined reward vector $\tilde{R}^{\mu}_{t+1} = [0,\ 1-\Delta]$
  - and construct the instantaneous and cumulative regret vector

$$\mathrm{reg}_{t+1} = [\Delta - 1,\ 0], \quad Reg_{t+1} = [\underbrace{(0.5 - t - 1)(1-\Delta)}_{\text{Negative}}, 0.5(1-\Delta)]$$

  - By regret matching update rule, Alice' strategy for next time $X_{t+1} = [0, 1]$ is still the pure strategy of 2nd row. This implies that Bob's next strategy is still 1st column.
- As a result, $\mathbb{P}(\Omega_t \mid w_1) = 1, \forall t \geqslant 1$, *i.e.*, Alice will always suffer linear regret.

Divergence of TS-RM

- By regret matching update rule,
- Conditioned on Ω_t, Alice will still use pure strategy of 2nd column if
- the 1st entry of cumulative regret calculated by TS estimator $\leqslant 0$
- We show that there exist some constant $c > 0$ such that $\mathcal{P}(\Omega_t) \geqslant c, \forall t \in \mathbb{Z}_+$ by iteratively calculating the conditional probability,

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

- Specifically, when $\Delta = 0.1$ and $\sigma_w = 0.1$ (used to update posterior algorithmically), we have

$$c \approx 0.54$$

Outline

A simple fix: Optimistic Sampling

- Independently sample M_{t+1} TS estimators

$$\tilde{f}_{t+1}^{TS,j}(a, B_t) \mid H_{t+1} \sim N(\mu_t(a, B_t), \sigma_t(a, B_t)), \forall j = 1, \ldots, M_{t+1}$$

- and taking the maximum,

$$\tilde{f}_{t+1}^{\text{OTS}}(a, B_t) = \max_{j \in [M_{t+1}]} \tilde{f}_{t+1}^{\text{TS},j}(a, B_t)$$

- Construct the imagined reward vector,

$$\tilde{R}_{t+1}^{OTS}(a) = \text{clip}_{[0,1]}(\tilde{f}_{t+1}^{\text{OTS}}(a, B_t)), \forall a \in \mathcal{A}$$

Apply Optimistic Sampling in the Counter example

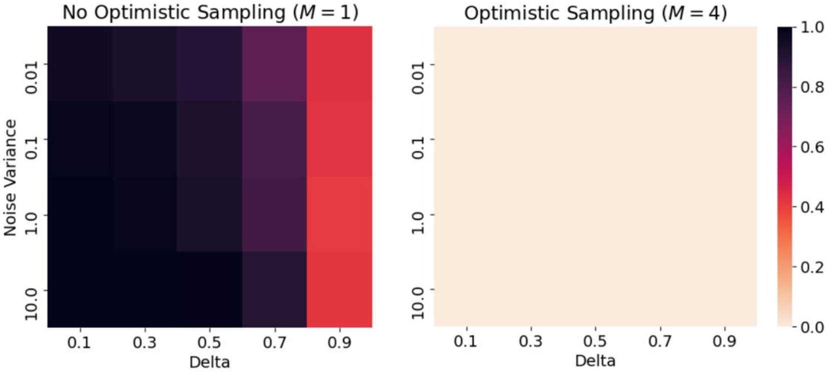


Figure: Failure ratio for TS-RM and OTS-RM with different problem setups (specified by Delta and noise variance).

Why OTS works with larger M_{t+1}?

- Assumed Alice always take the suboptimal action until time t, i.e. Ω_t happens,
- if Alice want to stop taking the suboptimal action and suffering Δ regret, Alice has move from the pure strategy $[0,1]$ to mixed strategy
- by Regret Matching, a sufficient condition for mixed strategy at this situation is to keep $\tilde{R}^{OTS}_{t+1}(1st) > \tilde{R}^{OTS}_{t+1}(2nd)$.
- By optimistic sampling,

$$\begin{cases} \tilde{f}^{TS,j}_{t+1}(1st, 1st) \sim N(0,1), \quad j = 1, \dots, M \\ \tilde{f}^{TS,j}_{t+1}(2nd, 1st) \sim N\left(\frac{t}{t+\sigma_w^2}(1-\Delta), \frac{\sigma_w^2}{\sigma_w^2+t}\right), \quad j = 1, \dots, M \end{cases} \tag{1}$$

then $\tilde{R}^{OTS}_{t+1}(1st) = \max_{j\in[M]} \tilde{f}^{TS,j}_{t+1}(1st, 1st)$ and $\tilde{R}^{OTS}_{t+1}(2nd) = \max_{j\in[M]} \tilde{f}^{TS,j}_{t+1}(2nd, 1st)$.

Key insight of optimistic sampling: anti-concentration

Lemma 2 (Anti-concentration property of maximum of Gaussian R.V.).

Consider a normal distribution $N(0, \sigma^2)$ *where* σ *is a scalar. Let* $\eta_1, \eta_2, \ldots, \eta_M$ *be* M *independent samples from the distribution. Then for any* $\delta > 0$

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

According to the anti-concentration property,

$$\mathbb{P}(\tilde{R}_{t+1}^{OTS}(1st) \leqslant \frac{t}{t + \sigma_w^2}(1 - \Delta) + \sqrt{\frac{2\sigma_w^2 \log(M/\delta_1)}{t + \sigma_w^2}}) \geqslant 1 - \delta_1 \tag{2}$$

The probability of $\tilde{R}_{t+1}^{OTS}(1st) > \tilde{R}_{t+1}^{OTS}(2nd)$

Now, we calculate the probability of

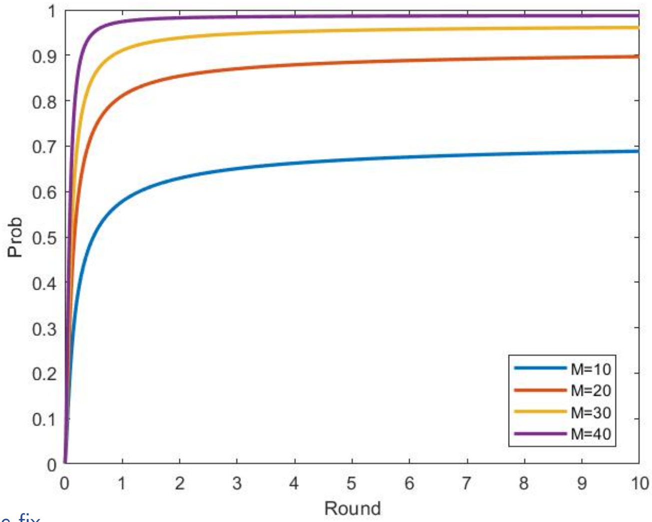
$$\tilde{R}_{t+1}^{OTS}(1st) > \frac{t}{t+\sigma_w^2}(1-\Delta) + \sqrt{\frac{2\sigma_w^2 \log(M/\delta_1)}{t+\sigma_w^2}}$$

Lemma 3.

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* $w \in \mathbb{R}_+$,

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

The probability of $\tilde{R}^{OTS}_{t+1}(1st) > \tilde{R}^{OTS}_{t+1}(2nd)$



Outline

General Regret Bound

- We introduce an imagined reward vector sequence $\tilde{R}_{t+1} \in [0,1]^{\mathcal{A}}$, where each $\tilde{R}_{t+1}$ is constructed using history information H_t with algorithmic randomness.
- For any $a \in \mathcal{A}$, the one-step regret can be decomposed by

$$\begin{aligned}
&\mathbb{E}\left[R_{t+1,a,B_t} - R_{t+1,A_t,B_t} \mid \theta\right] = \mathbb{E}\left[f_\theta(a,B_t) - f_\theta(A_t,B_t) \mid \theta\right] \\
&= \underbrace{\mathbb{E}\left[\tilde{R}_{t+1}(a) - \tilde{R}_{t+1}(A_t) \mid \theta\right]}_{(I)} + \underbrace{\mathbb{E}\left[f_\theta(a,B_t) - \tilde{R}_{t+1}(a) \mid \theta\right]}_{(II)} + \underbrace{\mathbb{E}\left[\tilde{R}_{t+1}(A_t) - f_\theta(A_t,B_t) \mid \theta\right]}_{(III)}
\end{aligned} \tag{3}$$

- Summation of (I) reduces to adversarial regret of Hedge or RM for bounded sequence $\tilde{R}_{t+1}$.
- $(II) \leqslant \mathbb{P}(f_{a,B_t} \geqslant \tilde{R}_{t+1}(a)) \leqslant \mathcal{O}(1/\sqrt{T})$ is small by select proper M_{t+1}. ($M_{t+1} = 1$, which is TS, cannot satisfy. One reason we need modified TS.)
- (III) can be bounded by $\mathcal{O}(\sigma_t(A_t,B_t))$ and further bounded by one-step information gain $I(\theta; R_{t+1,A_t,B_t} | H_t)$ using differential entropy of gaussian distribution.

General Regret Bound

- We introduce an imagined reward vector sequence $\tilde{R}_{t+1} \in [0,1]^{\mathcal{A}}$, where each $\tilde{R}_{t+1}$ is constructed using history information H_t with algorithmic randomness.
- For any $a \in \mathcal{A}$, the one-step regret can be decomposed by

$$\mathbb{E}\left[R_{t+1,a,B_t} - R_{t+1,A_t,B_t} \mid \theta\right] = \mathbb{E}\left[f_\theta(a,B_t) - f_\theta(A_t,B_t) \mid \theta\right]$$
$$= \underbrace{\mathbb{E}\left[\tilde{R}_{t+1}(a) - \tilde{R}_{t+1}(A_t) \mid \theta\right]}_{(I)} + \underbrace{\mathbb{E}\left[f_\theta(a,B_t) - \tilde{R}_{t+1}(a) \mid \theta\right]}_{(II)} + \underbrace{\mathbb{E}\left[\tilde{R}_{t+1}(A_t) - f_\theta(A_t,B_t) \mid \theta\right]}_{(III)} \tag{3}$$

- Summation of (I) reduces to adversarial regret of Hedge or RM for bounded sequence $\tilde{R}_{t+1}$.
- $(II) \leqslant \mathbb{P}(f_{a,B_t} \geqslant \tilde{R}_{t+1}(a)) \leqslant \mathcal{O}(1/\sqrt{T})$ is small by select proper M_{t+1}. ($M_{t+1} = 1$, which is TS, cannot satisfy. One reason we need modified TS.)
- (III) can be bounded by $\mathcal{O}(\sigma_t(A_t,B_t))$ and further bounded by one-step information gain $I(\theta; R_{t+1,A_t,B_t} | H_t)$ using differential entropy of gaussian distribution.

General Regret Bound

- We introduce an imagined reward vector sequence $\tilde{R}_{t+1} \in [0,1]^{\mathcal{A}}$, where each $\tilde{R}_{t+1}$ is constructed using history information H_t with algorithmic randomness.
- For any $a \in \mathcal{A}$, the one-step regret can be decomposed by

$$\mathbb{E}\left[R_{t+1,a,B_t} - R_{t+1,A_t,B_t} \mid \theta\right] = \mathbb{E}\left[f_\theta(a,B_t) - f_\theta(A_t,B_t) \mid \theta\right]$$
$$= \underbrace{\mathbb{E}\left[\tilde{R}_{t+1}(a) - \tilde{R}_{t+1}(A_t) \mid \theta\right]}_{(I)} + \underbrace{\mathbb{E}\left[f_\theta(a,B_t) - \tilde{R}_{t+1}(a) \mid \theta\right]}_{(II)} + \underbrace{\mathbb{E}\left[\tilde{R}_{t+1}(A_t) - f_\theta(A_t,B_t) \mid \theta\right]}_{(III)} \tag{3}$$

- Summation of (I) reduces to adversarial regret of Hedge or RM for bounded sequence $\tilde{R}_{t+1}$.
- $(II) \leqslant \mathbb{P}(f_{a,B_t} \geqslant \tilde{R}_{t+1}(a)) \leqslant \mathcal{O}(1/\sqrt{T})$ is small by select proper M_{t+1}. ($M_{t+1} = 1$, which is TS, cannot satisfy. One reason we need modified TS.)
- (III) can be bounded by $\mathcal{O}(\sigma_t(A_t,B_t))$ and further bounded by one-step information gain $I(\theta; R_{t+1,A_t,B_t} | H_t)$ using differential entropy of gaussian distribution.

General Regret Bound

- We introduce an imagined reward vector sequence $\tilde{R}_{t+1} \in [0,1]^{\mathcal{A}}$, where each $\tilde{R}_{t+1}$ is constructed using history information H_t with algorithmic randomness.
- For any $a \in \mathcal{A}$, the one-step regret can be decomposed by

$$
\begin{aligned}
&\mathbb{E}\left[R_{t+1,a,B_t} - R_{t+1,A_t,B_t} \mid \theta\right] = \mathbb{E}\left[f_\theta(a,B_t) - f_\theta(A_t,B_t) \mid \theta\right] \\
&= \underbrace{\mathbb{E}\left[\tilde{R}_{t+1}(a) - \tilde{R}_{t+1}(A_t) \mid \theta\right]}_{(I)} + \underbrace{\mathbb{E}\left[f_\theta(a,B_t) - \tilde{R}_{t+1}(a) \mid \theta\right]}_{(II)} + \underbrace{\mathbb{E}\left[\tilde{R}_{t+1}(A_t) - f_\theta(A_t,B_t) \mid \theta\right]}_{(III)}
\end{aligned}
\tag{3}
$$

- Summation of (I) reduces to adversarial regret of Hedge or RM for bounded sequence $\tilde{R}_{t+1}$.
- $(II) \leqslant \mathbb{P}(f_{a,B_t} \geqslant \tilde{R}_{t+1}(a)) \leqslant \mathcal{O}(1/\sqrt{T})$ is small by select proper M_{t+1}. ($M_{t+1} = 1$, which is TS, cannot satisfy. One reason we need modified TS.)
- (III) can be bounded by $\mathcal{O}(\sigma_t(A_t,B_t))$ and further bounded by one-step information gain $I(\theta; R_{t+1,A_t,B_t} | H_t)$ using differential entropy of gaussian distribution.

Preview of the bounds

- Problem-dependent quantity: information gain $\gamma_T(\theta) := I(\theta; A_0, B_0, \dots, A_{T-1}, B_{T-1})$ depends on underlying reward structure.
- Define $\gamma_T(\theta, \mathcal{A}, \mathcal{B}) := \min(\gamma_T(\theta), \sqrt{\mathcal{A}\mathcal{B} \log \mathcal{A}\mathcal{B}})$

Table: Regret bounds comparison.

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

Discussion on the bounds

By the results from [Srinivas, TIT09'] which gives the bounds of γ_T for a range of commonly used covariance functions: finite dimensional linear, squared exponential and Matern kernels.

Table: Maximum information gain γ_T.

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

- For example, if using squared exponential bounds, the final regret of OTS-Hedge is

$$\mathcal{O}((\sqrt{\log \mathcal{A}} + \log(T)^{d+1})\sqrt{T}),$$

  which has no polynomial dependence on action sizes $\mathcal{A} \times \mathcal{B}$, similar to full information setting.
- Curse of multi-agent is resolved: $|\mathcal{B}|$ is exponential in the number of opponents

Outline

Type of opponents

- Self-play (regret minimizing) opponent: alg^B can be history-dependent randomized algorithm,
- Best-response opponent: alg^B can access exact information of the mean reward function f_θ and Alice' mixed strategy $\pi_t(\cdot)$ before sampling B_t,
- Stationary opponent: alg^B always select B_t from a stationary distribution,
- Non-stationary opponent: alg^B select B_t from a changing distribution.

Random matrix game: Self-play (regret-minimizing) opponent

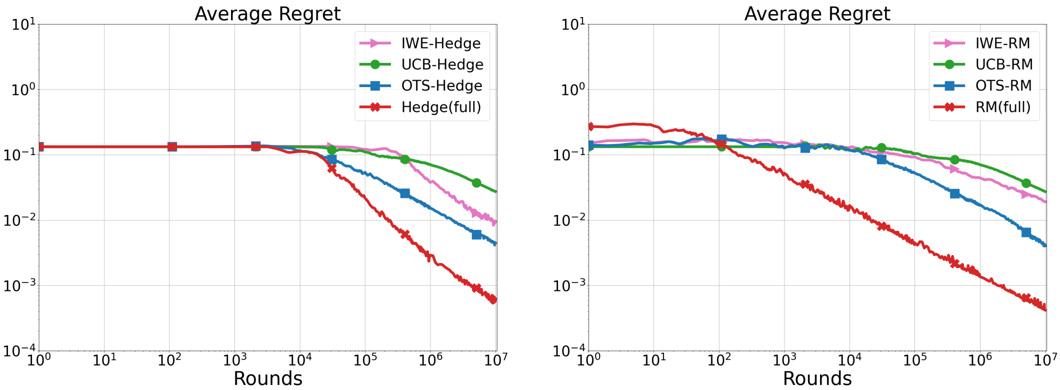


Figure: Matrix size: 70×70. Magnitude advantage of OTS estimator.

Random matrix game: Best-response opponent

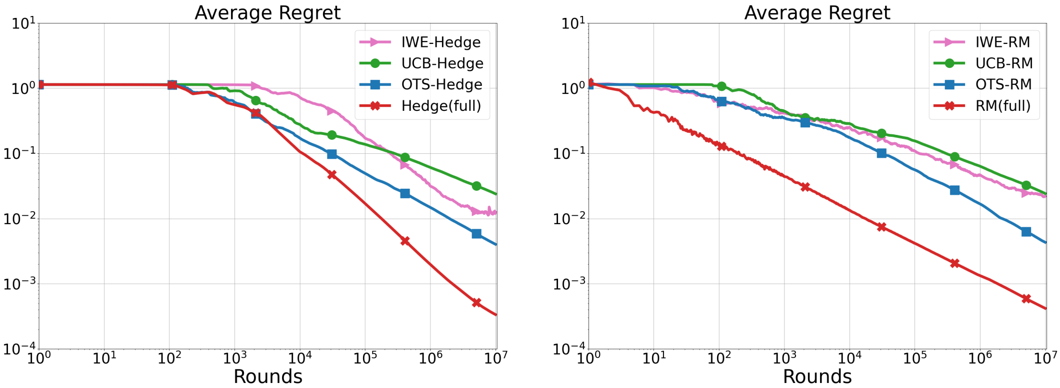


Figure: Matrix size: 70×70. Magnitude advantage of OTS estimator.

Random matrix game: Stationary opponent

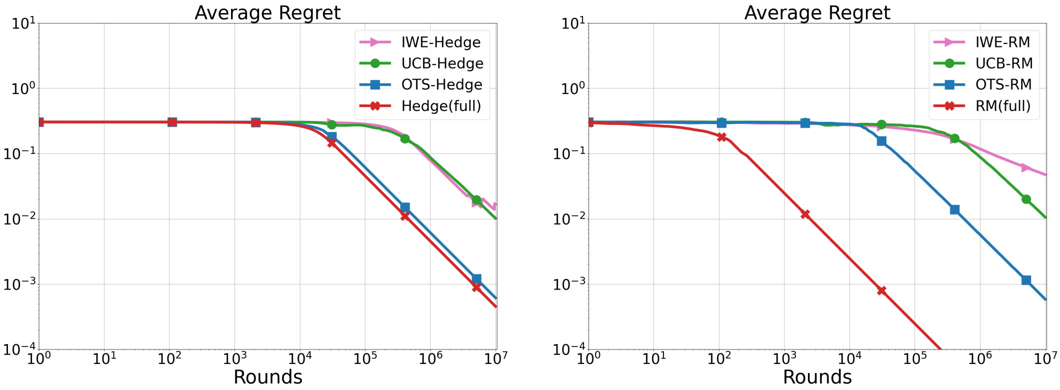
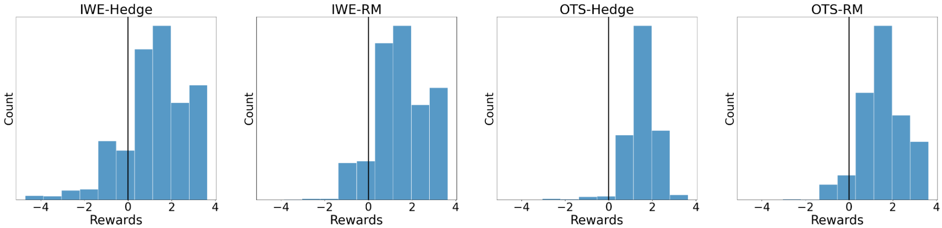


Figure: Matrix size: 70×70. Magnitude advantage of OTS estimator.

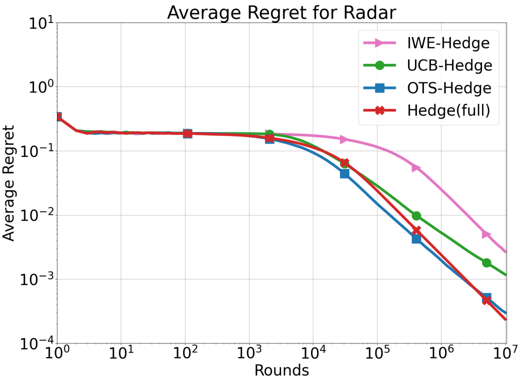
Random matrix game: non-stationary opponent (Robust bandit)

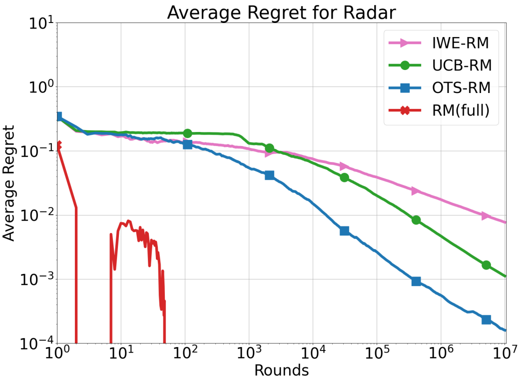
1. A game matrix $\theta \in \mathbb{R}^{10\times 5}$ is generated with each element sampling from $N(0.5, 2.0)$. 2. The opponent's actions are drawn from a fixed strategy that randomly changes every 50 rounds. 3. Each algorithm performs up to 1000 rounds and 100 simulation runs.



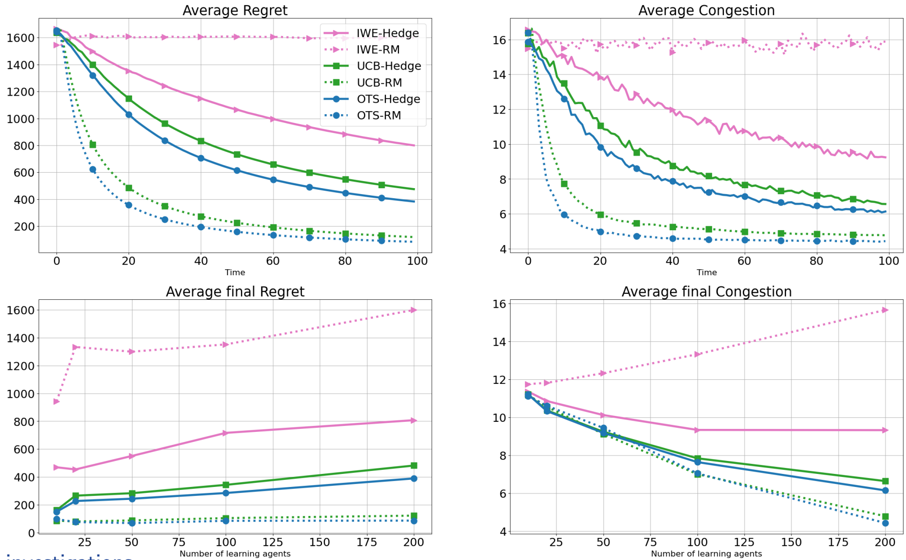
	return< 0	mean return
IWE-Hedge	19.4%	1.24
IWE-RM	12.6%	1.50
OTS-Hedge	2.5%	1.55
OTS-RM	8.8%	1.55

Application: radar anti-jamming





Application: Traffic Routing problem



Outline

Summary

- Considered a realistic unknown game setting with applications. (Rarely studied area but meaningful and increasingly important!)
- Algorithmic framework for considered setting: Reduction to full-information algorithm with any admissable vector estimator of unknown utility function.
- Proved naive application of mean estimator and Thompson Sampling(TS) estimator fails in a carefully constructed class of simple matrices.
- Proved the regret upper bound of simply modified OTS combined with full-information no-regret algorithms. The decomposition of regret in the proof is general.
- Superior empirical performance in playing against different type of opponents and in the real-world applications: radar and traffic problems.

Future works

- In experiments, we observe that the average regret TS-RM/TS-RM+ converge in self-play setting without optimistic modified. Why?
- Dynamic regret given different type of opponents
- Extension to Markov game: need a good motivation

Acknowledgements

- Collaborators: Liangqi Liu (CUHKSZ), Wenqiang Pu (SRIBD), Hao Liang (CUHKSZ)
- Supervisor: Prof. Zhi-Quan Luo
- Early discussion with Prof. Tong Zhang (HKUST)