EE5940 Project

Akshit Goyal and Jake Roth

12/20/2021 @ 11:59pm

Contents

1 Problem Description

We study a two-player, general-sum repeated game in which each player's rewards depends on both players' actions and are drawn from a (fixed) distribution that is unknown to both players. The setting, which is taken directly from [Tossou et al., 2020], *(i)* generalizes deterministic games by introducing uncertain payoffs and *(ii)* is generalized by so-called "stochastic games" which can be viewed as multi-agent Markov Decision Processes (MDPs).

In the deterministic repeated game setting, it is well known (from Folk Theorems) that any reward profile that is both *individually rational* and *feasible* can be obtained as a Nash equilibrium in repeated game strategies.[1] In particular, payoffs that may not correspond to "solutions" in single-play can correspond to "solutions" in repeated-play, thereby enlarging the space of a attainable reward profiles and solution concepts. This motivates the related problems of

1. Identifying desired equilibria; and
2. Designing (repeated game) strategies that can attain desired equilibria.

1.1 Desirable Equilibria

Regarding the first question, the authors in [Tossou et al., 2020] are concerned with finding an *egalitarian bargaining solution* (EBS) in repeated play. Solution concepts in bargaining games differ from those in non-cooperative games. In non-cooperatives games, agents only consider their own well-being, whereas in bargaining games (or surplus-sharing problems), agents decide collectively how to divide "surplus" payoffs. Notably, this division includes the notion of "walking away from the table," which describe what happens when bargaining breaks down; such points are called *disagreement points* and are crucial to describing precisely what the surplus is measured against.[2]

The EBS is one of several solution concepts that have been proposed for bargaining problems, and its principle can be summarized as: *maximize the minimum of surplus utilities* [Wikipedia, 2021]. The EBS is a Pareto optimal point of the single-play game, which in general is more desirable than a Nash equilibrium of a single-game. The feature that distinguishes an EBS from other solution concepts is its focus on *fairness*. In [Tossou et al., 2020], the disagreement point is taken to be the *safety value* (or *minmax value*) of the game, and hence EBS attempts to give each agent an equal surplus above their respective safety values. Furthermore, the EBS value is unique. Finally, it is notable that the minmax value is chosen as the disagreement point as this ensures the *individual rationality* criterion in various Folk Theorems (see Appendix A) is satisfied; this means that the EBS value can correspond to the payoff profile of some Nash equilibrium in repeated game strategies.

1.2 Finding Desirable Equilibria

In the deterministic setting, [Tossou et al., 2020] show that when agents measure the value of a game by their long-run expected reward over infinitely many plays, then an EBS policy can be implemented by a deterministic, stationary policy over joint actions. In the stochastic setting, [Tossou et al., 2020] approach the problem of finding EBS in self-play from the perspective of *multi-armed bandit* for two-players (**2P-MAB**). Each agent is given control over a set of independent arms, and their payoffs depend not only on the underlying reward distribution but also on the other agent's action. In the model of self-play described in [Tossou et al., 2020], both agents (individually) take on the role of a centralized planner who solves an optimistic UCB-type problem to computing each agent's hypothesized joint action, and then they coordinate joint actions based on the game's history.

1.3 Overview

As with the case of single agent setting, we would like to explore the idea of using Thompson sampling approach to solve repeated games as **2P-MAB** problem. We are interested in finding

[1] See Appendix B.

[2] For example, natural choices include *minmax values* (see Appendix A), a focal Nash equilibrium, or a value derived from *rational threats* (see [Narahari, 2012]).

EBS policy as discussed above. Firstly, we state some definitions and write down the mathematical formulation of our problem. Then we describe Upper Confidence for Repeated Games (UCRG) alogrithm by [Tossou et al., 2020] and our Thompson sampling (TS) approach. Lastly, we conduct computational experiments for two different implementation of our TS approach and compare them to UCRG. We conclude by providing possible future directions for the theoretical analysis of TS approach.

2 Mathemtical formulation

2.1 Definition

1. *Value of a policy*: The value $V^i_M(\pi)$ of policy π for player i in a repeated game M is defined as the infinite horizon undiscounted expected average reward
$$V^i_M(\pi) = \limsup_{T\to\infty} \frac{1}{T}\,\mathbb{E}\left(\sum_{i=1}^{t} r^i_t \,\middle|\, \pi\right)$$
2. *Maximin value*: The maximin policy π^i_{SV} for player i and its value SV^i is defined as
$$\pi^i_{SV} = \arg\max_{\pi_i}\min_{\pi^{-i}} V^i(\pi^i,\pi^{-i}), \quad SV^i = \max_{\pi_i}\min_{\pi^{-i}} V^i(\pi^i,\pi^{-i})$$
3. *Lexicographic maximin ordering*: A lexicographic maximin ordering $\geq_\ell$ on $\mathbb{R}^2$ is defined as
$$x \geq_\ell y \iff \Big(L^1(x) > L^1(y)\Big) \vee \Big(L^1(x) = L^1(y) \wedge L^2(x) \geq L^2(y)\Big)$$
where $L:\mathbb{R}^2 \to \mathbb{R}^2$ is a permutation of $x = (x^1, x^2) \in \mathbb{R}^2$ such that $L^1(x) \leq L^2(x)$.
4. *EBS in repeated games*: A policy π_{Eg} is an EBS if
$$V(\pi_{\mathrm{Eg}}) - SV \geq_\ell V(\pi) - SV \;\forall\pi.$$
5. *EBS value*: The EBS value is referred as the value $V_{\mathrm{Eg}} = V(\pi_{\mathrm{Eg}})$ and $V_+(\pi_{\mathrm{Eg}}) = V_{\mathrm{Eg}} - SV$ as the egalitarian advantage.

2.2 Problem of finding EBS policy in 2P-MAB

The goal is to find a game $\tilde{M}_k$ and a policy $\tilde{\pi}_k$ whose EBS value is near-optimal simultaneously for both players. Let true unknown game is given by M then we want to find $\tilde{M}_k$ and $\tilde{\pi}_k$ such that

$$\Pr\Big\{V_{\tilde{M}_k}(\tilde{\pi}_k) \geq V_M(\pi_{\mathrm{Eg}}) - (\epsilon_k, \epsilon_k)\Big\} = 1 \tag{1}$$

where $\epsilon_k > 0$ is a small configurable error. This is required as game $\tilde{M}_k$ may not achieve the highest EBS value simultaneously for both players.

3 Solution Approaches

3.1 UCRG

Construction of plausible set

Denote by $\mathcal{A}$ the set of joint actions of both players and by $\bar{r}^i_k(a)$ the empirical average reward of player i corresponding to joint action $a \in \mathcal{A}$. [Tossou et al., 2020] construct a set $\mathcal{M}_k$ at epoch k

containing all possible games such that,

$$\mathcal{M}_k = \left\{r : \mathbb{E}\, r^i(a) - \bar{r}^i_k(a) \le C_k(a) \& \,\mathbb{E}\, r^i(a) \le 1,\ \forall i, a\right\}$$

$$C_k(a) = \sqrt{\frac{\ln 1/\delta_k}{1.99 N_{t_k}(a)}}$$

We emphasize that here a refers to the joint action of both players.

Problem of finding optimistic EBS policy in 2P-MAB

The plausible set is used to define the upper and lower bounds of the game

$$\hat{r}^i_k = \bar{r}^i_k(a) + C_k(a), \qquad \check{r}^i_k = \bar{r}^i_k(a) - C_k(a).$$

Denote $\hat{M}$ the game with rewards $\hat{r}$ and $\check{M}$ the game with $\check{r}$. The goal is to find $\tilde{M}_k$ and $\tilde{\pi}_k$ such that

$$V_{\tilde{M}_k}(\tilde{\pi}_k) \ge_\ell V_{M'}(\pi')\ \forall \pi',\ \mathcal{M}' \in \mathcal{M}_k \mid \Pr\left\{V_{M'}({\pi'}_k) \ge V_M(\pi_{\mathrm{Eg}}) - (\epsilon_k, \epsilon_k)\right\} = 1 \tag{2}$$

To solve the above problem, at each epoch $\tilde{M}_k$ is set as the optimistic game $\hat{M}$ with rewards $\hat{r}$. In section 3.2.3, [Tossou et al., 2020] describe the computation of 'optimistic' maximin value and policy.

Computing an EBS policy

Given optimistic game $\hat{M}$ and optimistic maximin value, calculate optimistic advantage game by subtracting maximin value of each player from their respective rewards. Denote advantage game reward by $\hat{r}^i_+(a)$ for player i and joint action a. Then EBS policy is calculated as

1. For any two joint actions a and a', compute $\mathbf{score}(a, a') = \min_{i\in\{1,2\}} w(a, a') \cdot \hat{r}^i_+(a) + (1 - w(a, a')) \cdot r^i_+(a')$ with w as follows:

$$w(a, a') = \begin{cases} 0, \text{ if } \hat{r}^i_+(a) \le \hat{r}^{-i}_+(a) \text{ and } \hat{r}^i_+(a') \le \hat{r}^{-i}_+(a') \\ 1, \text{ if } \hat{r}^i_+(a) \ge \hat{r}^{-i}_+(a) \text{ and } \hat{r}^i_+(a') \ge \hat{r}^{-i}_+(a') \\ \frac{\hat{r}^{-i}_+(a') - \hat{r}^i_+(a')}{(\hat{r}^i_+(a) - \hat{r}^i_+(a')) + (\hat{r}^{-i}_+(a') - r^{-i}_+(a))}, \quad \text{otherwise.} \end{cases}$$

2. Compute

$$a_{\mathrm{Eg}}, a'_{\mathrm{Eg}} = \underset{a\in\mathcal{A}, a'\in\mathcal{A}}{\arg\max}\, \mathbf{score}(\mathrm{a}, \mathrm{a}')$$

3. The policy $\tilde{\pi}_{k,\mathrm{Eg}}$ is given by

$$\tilde{\pi}_{k,\mathrm{Eg}}(a_{\mathrm{Eg}}) = w(a_{\mathrm{Eg}}, a'_{\mathrm{Eg}}), \quad \tilde{\pi}_{k,\mathrm{Eg}}(a'_{\mathrm{Eg}}) = 1 - w(a_{\mathrm{Eg}}, a'_{\mathrm{Eg}})$$

3.2 Thompson Sampling (TS)

Denote by $\theta^{*(i)} \in \mathbb{R}^{m\times n}$ the true mean payoff matrices for $i = 1, 2$. At round t, the payoff for player i is given by

$$X^{(i)}_t = \mathrm{vec}(\mathcal{U}_t)^\top \mathrm{vec}\big(\theta^{*(i)}\big) + W^{(i)}_t$$

where $\mathcal{U}_t \in \mathbb{R}^{m\times n}$ is a joint action matrix such that $\mathcal{U}_t[a_1, a_2] = 1$ if joint action $a = (a_1, a_2)$ is played at round t, and $W^{(i)}_t \sim \mathcal{N}\left(0, \sigma^2_{(i)}\right)\ \forall t$.

Prior distribution

For each player i, we set our prior belief on $\theta^{*(i)}$ to be Gaussian i.e.

$$\theta^{*(i)} \sim \mathcal{N}\left(\mu_1^{(i)}, P_1^{(i)}\right)$$

where $\mu_1^{(i)} \in \mathbb{R}^{m \cdot n}$ is the initial mean and $P_1^{(i)} \in \mathbb{R}^{m \cdot n \times m \cdot n}$ the initial covariance matrix.

Updates for posterior distribution

After observing rewards X_k at each round $t \geq 2$, the player i updates its mean vector and covariance matrix as follows,

$$\begin{aligned} \mu_t^{(i)} &= \mu_{t-1}^{(i)} + \frac{X_t^{(i)} - \operatorname{vec}(\mathcal{U}_t)^\top \mu_{t-1}^{(i)}}{\operatorname{vec}(\mathcal{U}_t)^\top P_{t-1}^{(i)} \operatorname{vec}(\mathcal{U}_t) + \sigma_{(i)}^2} P_{t-1}^{(i)} \operatorname{vec}(\mathcal{U}_t) \\ P_t^{(i)} &= P_{t-1}^{(i)} - \frac{P_{t-1}^{(i)} \operatorname{vec}(\mathcal{U}_t) \operatorname{vec}(\mathcal{U}_t)^\top P_{t-1}^{(i)}}{\operatorname{vec}(\mathcal{U}_t)^\top P_{t-1}^{(i)} \operatorname{vec}(\mathcal{U}_t) + \sigma_{(i)}^2} \end{aligned}$$

Computing an EBS policy

At epoch k, we sample a game with mean payoffs $\theta_k^{(i)}$ for player i as $\theta_k^{(i)} \sim \mathcal{N}\left(\mu_{t_k}^{(i)}, P_{t_k}^{(i)}\right)$ and then compute the EBS policy exactly as outlined in the previous section but for the advantage game with rewards corresponding to the sampled game. t_k denotes the # of rounds played **up to** epoch k.

The key difference compared to UCRG is what game is used to compute the EBS policy while playing the game repeatedly. Also, at every round we compute the 'exact' maximin value assuming sampled game as the true game in contrast to the 'optimistic' maximin policy.

Action Implementation

We consider two styles of game play, which we call (i) Thompson sampling with randomized updates (*TSR*) and (ii) Thompson sampling with empirical updates (*TSE*). The procedure for TSR is as follows:

1. Sample a game from the posterior
2. Compute an approximate EBS policy and value for sampled game
3. Both players agree to play one of the (at most) two joint actions with probability specified by the approximate EBS policy
4. Observe rewards for the single joint action and update the posterior

The procedure for TSE is identical to that of TSR except for the following modification in Step 3:

$3'$. Both players agree to play the joint action that minimizes the distance between the empirical frequency of each action and the estimated EBS policy

This step in TSE amounts to adding one to the current observed joint action counts, comparing the result to the current estimated EBS policy, and choosing the action that minimizes the difference.

Batched Play

In addition, we consider "batched" play where an estimated policy is played for T rounds before being re-estimated.

3.3 Algorithm

Notations

- $N_k(a)$: # of rounds action a has been in episode k.
- N_k: # of rounds episode k has lasted.
- t_k: # of rounds played **up to** episode k.
- $N_{t_k}(a)$: # of rounds action a has been played up to round t_k.
- $\bar{r}^i_k(a)$: empirical average of player i for joint action a up to round t_k.

Below we summarize the two algorithms. In Algorithm 1, the function PLAY_UCRG causes player i to play the action that brings the empirical frequency of the observed plays most in line with the estimated policy at that time (which is symmetric for player $-i$ in self play and is similar to selecting a single joint action), except if three "bad-event" conditions occur:

- If there is a player whose advantage is better than that under the estimated EBS, then play the action that yields the better value.
- If potential errors on the EBS value are too large, then play an action that has large EBS value uncertainty.
- If potential error on the optimistic maximum value is too large, then play an action that has large optimistic maximum uncertainty.

On the other hand, we did not consider making modifications to the Thompson sampling procedure for any of these cases in this project. This simplified procedure is described in Algorithm 2. It is important to note that in our implementation of TS, the action chosen is a joint action whereas player i's UCRG strategy specifies an action for player i (though it is similar to choosing a joint action in self-play). For simplicity, we did not consider implementing separate strategies in TS but rather a single strategy over joint actions.

4 Experiments

Here we summarize experimental results from our Thompson sampling implementation based on the same game as in [Tossou et al., 2020]. The game is given by the following payoff matrices

$$A = \begin{array}{c c} & \begin{array}{cc} 1 & 2 \end{array} \\ \begin{array}{c} 1 \\ 2 \end{array} & \begin{bmatrix} 0.8 & 0.1 \\ 1.8 & 0.3 \end{bmatrix} \end{array}, \qquad B = \begin{array}{c c} & \begin{array}{cc} 1 & 2 \end{array} \\ \begin{array}{c} 1 \\ 2 \end{array} & \begin{bmatrix} 0.8 & 1,8 \\ 0.0 & 0.3 \end{bmatrix} \end{array}. \tag{3}$$

In Section 4.1 and Section 4.2 below, we describe two sets of numerical experiments. In Section 4.3, we compare our results to UCRG.

4.1 Action Implementation

In the top panel of Fig. 1, we plot the representative trajectories of the regret attained by our Thompson sampling implementations, where we define the N-period regret for each player $i = A, B$

Algorithm 1 UCRG

```
Input: Number of episodes $K$, and for each episode $k - \epsilon_k, \delta_k$.
Initialization: Let $t \leftarrow 1$. Set $\mu_1^{(i)} \leftarrow \mu_i$, $P_1^{(i)} \leftarrow P_i$ for each player $i$. Set $N_k, N_k(a), N_{t_k}(a)$ to zero $\forall a \in \mathcal{A}$.
for $k = 1, 2, \ldots$ do
    $t_k \leftarrow t$
    $N_{t_k+1}(a) \leftarrow N_{t_k}(a)\ \forall a$
    $\hat{r}_k^i(a) = \bar{r}_k^i + C_k(a)$, $\check{r}_k^i(a) = \bar{r}_k^i - C_k(a)$, $\forall a, i$
    $\tilde{\pi}_k \leftarrow \text{OptimisticEgalitarianPolicy}(\bar{r}_t, \hat{r}_t, \check{r}_t, \epsilon_k)$[3]
    Execute policy $\tilde{\pi}_k$:
    do
        Let $a_t \leftarrow \text{PLAY\_UCRG}(\tilde{\pi}_k)$, play it and for each $i$ observe $r_t^i(a)$.
        $N_k \leftarrow N_k + 1 \quad N_k(a_t) \leftarrow N_k(a_t) + 1$
        $N_{t_k+1}(a_t) \leftarrow N_{t_k+1}(a_t)$ and
        Update $\bar{r}_k^i(a_t)\ \forall i$
        $t \leftarrow t + 1$
    while $N_k(a_t) \leq \max\{1, N_{t_k}(a)\}$
end for
```

Algorithm 2 TS(E/R)

```
Input: Number of episodes $K$ and batchsize $T$.
Initialization: Let $t \leftarrow 1$. Set $\mu_1^{(i)} \leftarrow \mu_i$, $P_1^{(i)} \leftarrow P_i$ for each player $i$. Set $N_t(a)$ to zero $\forall a \in \mathcal{A}$.
for $k = 1, 2, \ldots$ do
    $\theta_k^{(i)} = [r_k^i(a)]_{a \in \mathcal{A}} \sim \mathcal{N}\left(\mu_{t_k}^{(i)}, P_{t_k}^{(i)}\right)\ \forall i$
    $\tilde{\pi}_k \leftarrow \text{EgalitarianPolicy}\left(\theta_k^{(i)}\right)$
    for $t = 1, 2, \ldots, T$ do
        Execute policy $\tilde{\pi}_k$:
        Let $a_t \leftarrow \text{PLAY\_TS}(\tilde{\pi}_k)$, play it and for each $i$ observe $X_t^{(i)}$.
        $N_r(a) \leftarrow N_r(a) + 1$
        Update $\mu_t^{(i)}, P_t^{(i)}\ \forall i$
    end for
end for
```

to be

$$\text{Regret}_N^{(i)} := \sum_{t=1}^{N} \left(\text{EBS} - R_t^{(i)}\right).$$

We find that the results are encouraging; in some trajectories, the regret appears to grow much slower than linearly. In the middle panel of Fig. 1, we plot a single trajectory of the reward estimation error, which measures the absolute deviation of the posterior mean from the true mean. We find that the game estimation fails to improve after some time. In the bottom panel of Fig. 1, we plot a single trajectory of the policy error, which measures the absolute deviation of the estimated EBS policy on the sampled game for each joint action with the EBS policy for the true game. We find that both TS schemes are able to make good initial progress in finding an approximate EBS policy, but ultimately, the policy errors fail to improve beyond a certain level of precision. This is perhaps not too surprising given the limited precision in the posterior mean estimation.

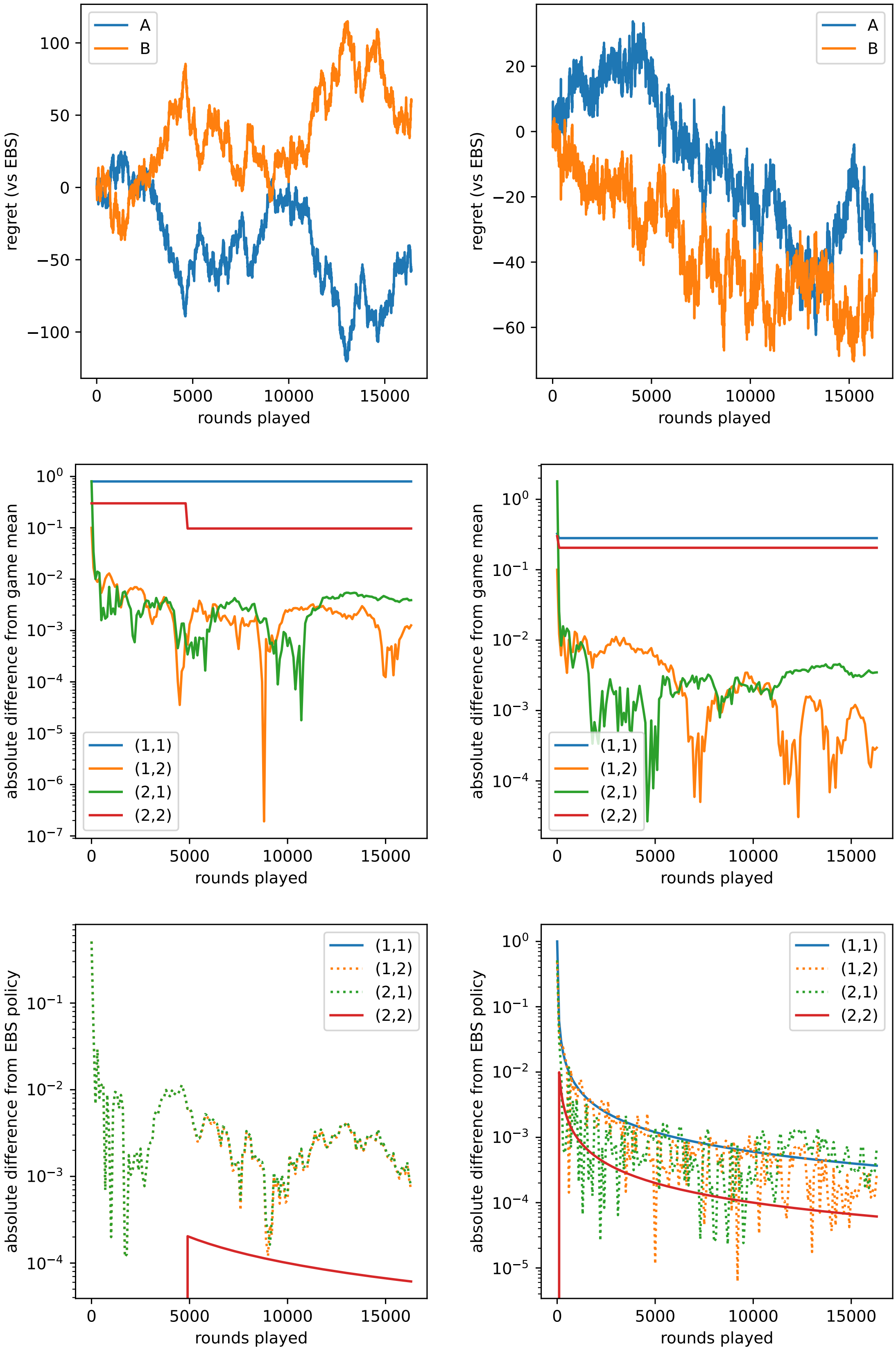

Figure 1: **Regret** (top), **reward estimation error** (middle) and **policy error** (bottom) over 2^{14} rounds of play for a single trajectory.*Left*: TSR. *Right*: TSE.

4.2 Batched Play

To explore whether there might be a simple method for achieving increased precision, we repeat the above experiments for 2^{14} total rounds of play but divide up the policy implementation step into "batches" of constant size, ranging from 2^1 to $2^7 = 128$. In Fig. 2, we display results for batch size of 64. We see that regret can grow linearly in the TSR scheme, while in this trajectory instance, regret appears to grow slower for TSE. Comparing with other batch sizes, however, we were not able to find a clear pattern and will need to simulate over multiple trajectories in each case.

4.3 Comparison to UCRG

In Figs. 1 and 2, we see that the behavior of TS can be highly dependent on the sequence of observed randomness. In some trajectories, regret grows linearly, while in others, regret remains almost constant. We study this behavior over 50 sample trajectories and compare TSE and TSR with batch size 1 to UCRG over 100,000 rounds. In Fig. 3, we find that that both TSR and TSE both indicate that the ensemble average of regret trajectories grows linearly. On the other hand, the minimum regret panel is encouraging. We suspect that the simple implementation of TS that does not check the three "bad-event" conditions in UCRG are partially to blame.

5 Future Ideas and Questions

- Provide regret bounds for the TS approach.
  1. As in the single agent case, the Bayesian regret bound for TS in 2P-MAB case can be derived using the steps in regret bound analysis for UCRG since similar terms need to be bounded. In UCRG analysis, an event E defined as union of 4 events is considered and the regret is upper bounded conditioned on whether E occurs or not. The regret bound for the case when E occurs is bounded using Hoeffding's inequality. For the remaining case, when event E is false, the analysis is reduced to upper bounding $N_k(a)$ which is the # of times joint action a has been played up to episode k. This is the most difficult piece of analysis wherein the bound follows by upper bounding N_k when considering separately that each of event in E doesn't occur.
  2. Empirically, we find that TS approach has linear average regret; we suspect that considering the three "bad-event" cases will help resolve this issue and plan to explore this. Resolving this issue could then help in the analysis.
- UCRG appears to have some connection to fictitious play (FP), since it plays joint action aligned with empirical frequency. It is known that when the FP empirical play distributions converge, then the result is a Nash equilibria. However, there still is the issue of correlation amongst the agents to ensure that the Nash payoff is actualized. Viewing the problem in this way may allow for splitting off the "correlation" part from the "game-estimation" part of the analysis.
- Is **2P-MAB** over joint actions for the EBS-equilibrium a feasibilty problem for some single player?

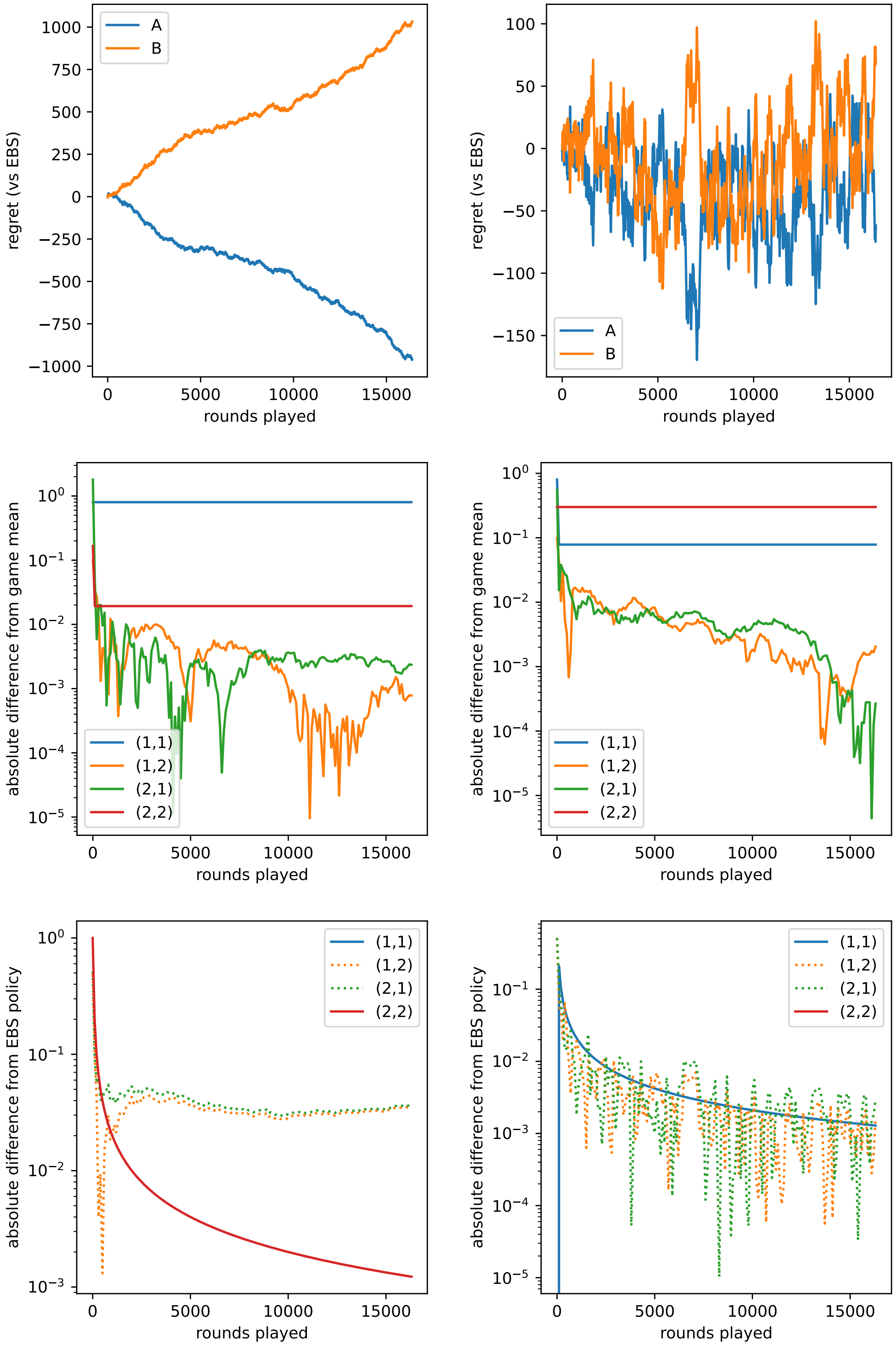

Figure 2: **Regret** (top), **reward estimation error** (middle) and **policy error** (bottom) over 2^{14} rounds of play with batches of size 16 for a single trajectory.*Left*: TSR. *Right*: TSE.

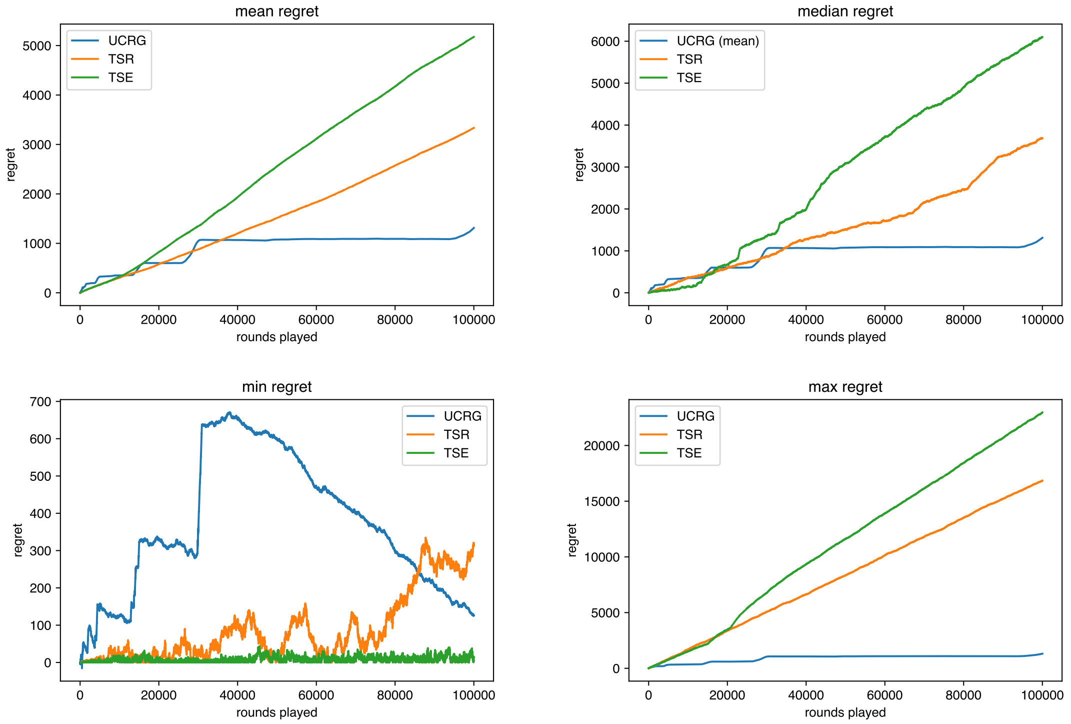

Figure 3: Regret (maximum across players) over 100,000 rounds.

References

[Narahari, 2012] Narahari, Y. (2012). Cooperative game theory: The two person bargaining problem. https://gtl.csa.iisc.ac.in/gametheory/ln/web-cp2-bargaining.pdf.

[Tossou et al., 2020] Tossou, A. C., Dimitrakakis, C., Rzepecki, J., and Hofmann, K. (2020). A novel individually rational objective in multi-agent multi-armed bandits: Algorithms and regret bounds. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1395–1403.

[Wikipedia, 2021] Wikipedia (2021). Cooperative bargaining. https://en.wikipedia.org/wiki/Cooperative_bargaining.