MARL in Persuasion Game

1 Repeated Bayesian Persuasion as a turn-based Markov Game

First, consider a classical Bayesian persuasion model:

The state of nature follows a publicly known prior distribution f. The sender (she) observes the realization of the state $\theta \sim f$ and samples a signal σ according to her signaling scheme. Her signal s is sent to the receiver (he), and interpreted as a distribution of the state of nature $\mathbf{Pr}(\theta|\sigma)$. He then takes some action a for optimal reward.

	guilty	innocent
convict	(1,1)	(1,0)
acquit	(0,0)	(0,1)

Table 1: payoff table, row player is the prosecutor (sender), column player is the judge (receiver). The left column is if the state of defendant is indeed innocent ($\theta = 0$), the right column is if the state of defendant is indeed guilty ($\theta = 1$).

Both the prosecutor and the judge want to maximize their utility. If the prior distribution $f(\theta = 1) = 0.3$, then the optimal persuasion for the prosecutor is to set its signal scheme as reporting the defendant as innocent with probability only $\frac{4}{7}$, whenever the defendant is indeed innocent; reporting the defendant as guilty with probability 1, whenever the defendant is indeed guilty. In this way, the judge have to follow the persuasion, to always convict whenever the prosecutor suggests so.

Now, we put this model in a two-player (turn-based) Markov game framework, and the research question is, *can the sender and receiver learn to persuade and act optimally?*

We adopt the notation system of RL. Let the state space be $\mathcal{S}$. Let the action space for player 1 (sender) be $\mathcal{A}$, for player 2 (receiver) be $\mathcal{B}$, the reward function $r_1, r_2 : \mathcal{B} \times \Theta \to [0, 1]$ are defined by the payoff matrix.

We know that the optimal signal scheme can be direct and persuasive. So we can have $\mathcal{A} = \mathcal{B}$, and the state space $\mathcal{S} \equiv \Theta \cup \mathcal{A}$. The learning process goes as for $t \in \{1, 2, \ldots, T\}$

1. At round $2t - 1$, player 1 observes the state $s_{2t-1} \in \Theta$ and takes an action $a_{2t-1} \in \mathcal{A}$ according to the distribution $\pi_1(a|s_{2t-1})$, which can be viewed as her signaling scheme. The action a_{2t-1} solely determines the next state $s_{2t} = a_{2t-1}$. For whatever the action b_{2t-1} player 2 takes, it have not effect on the next state, and the payoff for both player is always 0.
2. At round $2t$, player 2 observes the state $s_{2t} \in \mathcal{A}$, and takes an action $b_{2t} \in \mathcal{B}$ according to the distribution $\pi_2(b|s_{2t})$, which is his policy to follow the persuasion. The next state $s_{2t+1} \sim f$ is a random sample from publicly known prior distribution. For whatever the action a_{2t} player 1 takes, it have not effect on the next state, and the payoff for player 1 and 2 are respectively $r_1(b_{2t}, s_{2t-1}), r_2(b_{2t}, s_{2t-1})$. The player 1 and 2 updates their policy π_1, π_2 accordingly based on the action/feedback tuple $(r_1, a_{2t-1}), (r_2, b_{2t})$

[BJ20] proposed the provably no-regret algorithm to achieve the equilibrium of any turn-based markov game in self-play. Our problem can be formulated as a turn-based Markov game, but it is more challenging to identify the equilibrium with optimal utility.

Preprint. Under review.

2 A simpler case: Learning to persuade against best-responding receiver

Suppose the receiver is perfect, such that he knows the signaling scheme π of the sender and always best response. That is, to solve for the optimization program,

$$\forall s \in \mathcal{A}, y(\pi, s) = \arg\max_{y \in \mathcal{B}} \sum_{\theta} \mathbf{Pr}(\theta|\pi, s) r(\theta, y)$$

So the corresponding expected reward function for the sender with strategy x at state θ is

$$l(x, \theta) = \sum_{s} \mathbf{Pr}(s|x) s(\theta, y(\pi, s))$$

Then the objective for the optimal persuasion can be formulated as,

$$\sum_{\theta} \mathbf{Pr}(\theta) \max_{x \in \Delta_{\mathcal{A}}} l(x, \theta)$$

Meanwhile, the sender is running an exp3 algorithm per state. Then, the objective of the exp3 at each state are not independent from each other.

Then how can the sender use the best responses of receiver to determine the receiver's payoff matrix and thereby compute the optimal signaling scheme?

[PSTZ19] designs the algorithm to iteratively learn the leader's optimal strategies, given the best responding receiver. Using the same technique, we can determine the n partition of feasible regions $\mathcal{P}_i$ and corresponding follower's (i.e., receiver's) best response f_i. Then, to find the optimal signaling scheme given prior distribution p_0, we just need to solve the optimization program,

$$\begin{aligned} &\max \sum_{i \in [n]} u(x_i) p_i \\ s.t. &\sum_{i \in [n]} x_i p_i = p_0 \\ &x_i \in \mathcal{P}_i \end{aligned}$$

Here each $x_i = \sum_j q_j v_{i,j}$ can be represented by a linear combination of the vertex $v_{i,j}$ of the convex region $\mathcal{P}_i$, and therefore $u_i(x_i) = \sum_j q_j u(v_{i,j})$ can be represented by a linear combination of the utility at these vertex. So it can written as a linear program with combined variable $p_i q_j$:

$$\begin{aligned} &\max \sum_{i \in [n]} p_i \sum_j q_j u(v_{i,j}) \\ s.t. &\sum_{i \in [n]} p_i \sum_j q_j v_{i,j} = p_0 \end{aligned}$$

3 Learning to persuade against best-responding receiver with No Regret

In persuasion game, how can the sender design the learning process in a no-regret fashion? Or more simplified setting of Stackelberg game, how can the leader control the exploration in the algorithm that have certain no-regret guarantee?

$$R(T) = \sum_{t=1}^{T} u_1(x_t, f^*(x_t)) - \max_{x \in \Delta_{\mathcal{A}}} u_1(x, f^*(x))$$

We know that the sampling algorithm by [LCM09] can identify the optimal strategy in $O(m^2 n L + V^{-1} n \log n)$, where $m = |\mathcal{A}|, n = |\mathcal{B}|, L, V$ is the number of the leader's and the follower's actions, the representation precision and the volume of the smallest feasible region. In the worst case, V

can be $O(2^{mL})$, which makes the algorithm no better than brutal force. [PSTZ19] improve the time complexity to $O(m^2nL + \text{Ext}(A))$, where $\text{Ext}(A) = n\binom{m+n}{n}$ is the total number of intersecting points of each polygon that corresponds the follows best response of certain action. Therefore, when T is sufficiently large or $T \to \infty$, it is trivial to design the no regret algorithm.

In addition, for a Stackelberg game where the leader has m actions and the follower has two actions. We show there is an $O(mL)$ algorithm to determine the optimal strategy. Specifically, we can formulate the objective of the leader as a linear program,

$$\max_{b_j \in \{0,1\}} \max_{x \in \Delta_{\mathcal{A}}} u_1(x, b_j)$$
$$s.t. \quad u_2(x, b_j) \geq u_2(x, 1 - b_j)$$

In the first m rounds, we observe the follower's best response to each action a_i as well as the payoff $u_1(a_i, f^*(a_i))$, and accordingly partition the action into two sets $\mathcal{A}^0, \mathcal{A}^1$ based whether the follower responds with its action 0 or 1. Now if one of sets is empty, then we are already done here, by taking the action with highest payoff $\max_i u_1(a_i, f^*(a_i))$. Otherwise, the optimal strategy might be a mixed strategy and we show it is possible to find the separating hyperplane by determining the ratio of $\frac{u_2(a_i,0)-u_2(a_i,1)}{u_2(a_j,0)-u_2(a_j,1)}$ for each $i \neq j$ through binary search. Because with these ratios, we can easily determine whether $\sum_{a_i \in \mathcal{A}} x_i[u_2(a_i, 0) - u_2(a_i, 1)]$ is positive or not.

Specifically, pick any action a_i in $\mathcal{A}^0$ and any action a_j in $\mathcal{A}^1$, we construct a mixed strategy that plays a_i with probability ρ and a_j with probability $1 - \rho$. We search for the closest $\rho = \frac{u_2(a_i,0)-u_2(a_i,1)}{u_2(a_j,0)-u_2(a_j,1)}$. Start with $u = 1$ and $l = 0$ $\rho = \frac{u+l}{2}$, we observe the follower's best response on the mixed strategy, if it is 0, we update $l = \rho$, otherwise we update $u = \rho$. As the representation precision is L bits, the search takes at most L rounds. Meanwhile, we only need to search for $m - 1$ pairs to fully determine the ratio for any two actions. In total, it takes $(m - 1)L$ rounds.

It is also possible to apply ellipsoid method to solve for the case when the follower has n actions. We first sample the initial point for each region where the best response is one of the followers' action j. The remaining task is to run the ellipsoid method for each region, and the best response of the follower serves as the membership oracle. However, this method also relies on the fact that the smallest region is large enough, and it takes also $O(m^2nL)$.

Now we propose another approach to this regret minimization problem:

First, consider the case when the follower is trivially best responding with a single action. The naive approach as above is to enumerate all m actions and then keep taking the best action. Clearly, the regret is linear when $T = O(m)$. However, it is possible to run the OSMD with two points feedback to reduce the regret to $O(\sqrt{T})$.

Algorithm 1 The OSMD Algorithm against trivial follower

Input: Number of actions m, explore and exploitation parameter η_t, δ_t
Initialization: $x^0 = \min_{x \in \Delta_{\mathcal{A}}} R(x)$,
Observe the payoff $u^0 \leftarrow u_1(x^0, f^*(x^0))$
for $t = 1 \dots T$ **do**
    Sample unit vector uniformly v_t.
    $x^{t+1/2} \leftarrow \prod(x^t + \eta_t v_t)$
    Observe the payoff $u^t \leftarrow u_1(x, f^*(x^{t+1/2}))$
    if $u^t > u^{t-1}$ **then**
        $x^{t+1} \leftarrow \prod(x^t + \delta_t v_t)$
    else if $u^t < u^{t-1}$ **then**
        $x^{t+1} \leftarrow \prod(x^t - \delta_t v_t)$
    else
        $x^{t+1} \leftarrow \prod(x^{t+1/2})$
    end if
end for

Is it possible for the leader to find the best action by simply solving a single convex optimization problem? This question can be answer by find the conditions about utility matrix A, B, function $f(x) = xAy^T(x)$, where $y(x) = \arg\max_y xBy^T$ is concave. It is easy to see that there is no need for y to randomize, so $y(x)$ must be a unit vector. We can simplify the problem as $f(x) = \sum_i x_i a_{ij}$, where $j = \arg\max_j \sum_i x_i b_{ij}$,

1. Zero sum, because $A = -B$, we can write $f(x) = \min_j \sum_i x_i a_{ij}$, and this piecewise minimum of linear function of x is concave on x.

2. Or more generally, $A = kB$, $k < 0$

Under such condition, the buyer can run a simple online convex optimization algorithm to find best strategy.

3.1 A simplified setting: pull arm for both sender and receiver

Consider the setting in [BJWX21]. We can design online learning algorithm that can control both the leader and follower's action to learn the SSE.

Let $A, B \in [0,1]^{m\times n}$ describe the leader and follower's utility in the Stackelberg game. Denote the leader's utility as $U(x, j; A, B) = xAe_j$, when the leader takes strategy $x \in \Delta_{\mathcal{A}}$ and the follower takes an action $j \in [n]$ (no need to randomize). In addition, we denote $U(x; A, B) = xAe_{j^*}$, where $j^* = \arg\max_j xBe_j$ is the follower's best response.

We define a gap Δ_j^B that measures the maximum difference between a follower's action j and any other action j', that is, $\Delta_j^B \equiv \max_{x\in\Delta_{\mathcal{A}}} \min_{j'\neq j} [xBe_j - xBe_{j'}]$. We denote x_j^B as the corresponding maximizer. And WLOG. we can assume $\Delta_j^B > 0, \forall j \in [n]$, meaning any of the follower's action j could be a best response against the leader's strategy x_j^B. In addition, we define $\Delta^B \equiv \min_{j\in[n]} \Delta_j^B$.

Theorem 1. *Given an estimation of follower's utility $\bar{B}$, such that $\|\bar{B} - B\|_1 \le \epsilon$, and $\Delta^B > 4\epsilon$, we can find an $8\epsilon/\Delta^B-$SSE.*

Proof. Based on the utility estimation $(A, \bar{B})$, we can compute its SSE, $\bar{x}^*, \bar{j}^*$. Now we show that the leader's strategy $\bar{x} \equiv (1 - \frac{4\epsilon}{\Delta})\bar{x}^* + \frac{4\epsilon}{\Delta} x_{\bar{j}^*}^{\bar{B}}$ with the follower's strategy $\bar{j}^*$ is an $\epsilon-$SSE.

Recall that under $x_{\bar{j}^*}^{\bar{B}}$, the follower with utility matrix $\bar{B}$ will take best response $\bar{j}^*$.

For our analysis below, we let x^*, j^* be the SSE of the game (A, B), and a strategy $x \equiv (1 - \frac{4\epsilon}{\Delta})x^* + \frac{4\epsilon}{\Delta} x_{j^*}^B$. So $U(x^*, j^*; A, B)$ is the leader's utility at SSE. We argue that

$$\begin{aligned}
U(\bar{x}; A, B) =& U(\bar{x}, \bar{j}^*; A, B) \\
=& U(\bar{x}, \bar{j}^*; A, \bar{B}) \\
\ge& (1 - \frac{4\epsilon}{\Delta^B}) U(\bar{x}^*, \bar{j}^*; A, \bar{B}) \\
\ge& (1 - \frac{4\epsilon}{\Delta^B}) U(x; A, \bar{B}) \\
=& (1 - \frac{4\epsilon}{\Delta^B}) U(x, j^*; A, \bar{B}) \\
=& (1 - \frac{4\epsilon}{\Delta^B}) U(x, j^*; A, B) \\
\ge& (1 - \frac{4\epsilon}{\Delta^B})^2 U(x^*, j^*; A, B)
\end{aligned}$$

The first and third equality are both implied by Claim 1. The other two equalities are because the leader's utility stay the same, as long as the leader and followers' strategy are the same. The second and the last inequality are by construction of $\bar{x}$ and x respectively. The third inequality is by the definition of SSE, no strategy can achieve better utility than $\bar{x}^*$.

Claim 1. *The follower under utility B will best response to $\bar{x}$ with action $\bar{j}^*$*

Proof.

$$\begin{aligned}
\bar{x}Be_{\bar{j}^*} - \max_{j\neq j'} \bar{x}Be_{j'} \geq & \bar{x}\bar{B}e_{\bar{j}^*} - \max_{j\neq j'} \bar{x}\bar{B}e_{j'} - \bar{x}[\bar{B}-B]e_{\bar{j}^*} + \max_{j\neq j'} \bar{x}[\bar{B}-B]e_{j'} \\
\geq & \bar{x}\bar{B}e_{\bar{j}^*} - \max_{j\neq j'} \bar{x}\bar{B}e_{j'} - \|B-\bar{B}\|_1 \\
= & [(1-\frac{4\epsilon}{\Delta})\bar{x}^* + \frac{4\epsilon}{\Delta}x^{\bar{B}}_{\bar{j}^*}]\bar{B}e_{\bar{j}^*} - \max_{j\neq j'}[(1-\frac{4\epsilon}{\Delta})\bar{x}^* + \frac{4\epsilon}{\Delta}x^{\bar{B}}_{\bar{j}^*}]\bar{B}e_{j'} - \|B-\bar{B}\|_1 \\
\geq & (1-\frac{4\epsilon}{\Delta})[\bar{x}^*\bar{B}e_{\bar{j}^*} - \max_{j\neq j'}\bar{x}^*\bar{B}e_{j'}] + \frac{4\epsilon}{\Delta}[x^{\bar{B}}_{\bar{j}^*}\bar{B}e_{\bar{j}^*} - \max_{j\neq j'} x^{\bar{B}}_{\bar{j}^*}\bar{B}e_{j'}] - \|B-\bar{B}\|_1 \\
\geq & \frac{4\epsilon}{\Delta^{\bar{B}}}\Delta^{\bar{B}} - \|B-\bar{B}\|_1 \\
= & \frac{4\epsilon}{\Delta^{\bar{B}}}(\Delta^B - \epsilon) - \epsilon \\
\geq & 2\epsilon
\end{aligned}$$

The first equality is by the construction of $\bar{x}$. The first and third inequality is by the convexity of maximum function. The second inequality is by Lemma 1 The fourth inequality is by the definition of $x^{\bar{B}}_{\bar{j}^*}$ and the fact that $\bar{j}^*$ is the best response to $\bar{x}^*$, the utility gap should be non-negative. The last equality is by Claim 2. The last inequality is by the fact that $\Delta \geq 4\epsilon$.

□

Claim 2. *Given $\|\bar{B}-B\| \leq \epsilon$, we have $\Delta^{\bar{B}} \geq \Delta^B - \epsilon$*

Proof.

$$\begin{aligned}
\Delta^{\bar{B}} &= \min_{j\in[n]} \max_{x\in\Delta_{\mathcal{A}}} \min_{j'\neq j} \left[x\bar{B}e_j - x\bar{B}e_{j'}\right] \\
&= \min_{j\in[n]} \max_{x\in\Delta_{\mathcal{A}}} \min_{j'\neq j} \left[xBe_j - xBe_{j'} + x(\bar{B}-B)(e_j - e_{j'})\right] \\
&\geq \min_{j\in[n]} \max_{x\in\Delta_{\mathcal{A}}} \min_{j'\neq j} \left[xBe_j - xBe_{j'} - \|\bar{B}-B\|_1\right] \\
&= \min_{j\in[n]} \max_{x\in\Delta_{\mathcal{A}}} \min_{j'\neq j} [xBe_j - xBe_{j'}] - \|\bar{B}-B\|_1 \\
&= \Delta^B - \epsilon
\end{aligned}$$

The first and last equalities are by definition, the second and third equalities are by some linear transformation. The first inequality by Lemma 1. □

Lemma 1. *$-\|B\|_1 < xBy < \|B\|_1$ if $\|x\|_\infty < 1$ and $\|y\|_\infty < 1$*

Proof. Let $b_{i,j}$ be the i,j th entry of matrix B. We can derive $xBy = \sum_{i\in[m]}\sum_{j\in[n]} x_i b_{i,j} y_j \geq -\sum_{i\in[m]}\sum_{j\in[n]} |b_{i,j}| = -\|B\|_1$.

Similarly, $xBy = \sum_{i\in[m]}\sum_{j\in[n]} x_i b_{i,j} y_j \leq \sum_{i\in[m]}\sum_{j\in[n]} |b_{i,j}| = \|B\|_1$

□

□

Corollary 1.1. *Given time horizon T, there exists an explore-then-commit algorithm that can achieve regret $O(T^{\frac{2}{3}})$*

Proof. For the first $T^{\frac{2}{3}}$ rounds, the algorithm explores by sampling each action pairs (i,j) for $T^{\frac{2}{3}}/AB$ times, such that we have $\|\bar{B}-B\|_1 \leq (AB)^{3/2}T^{-1/3}$. Then, according to Theorem 1, we can obtain a $(AB)^{3/2}T^{-1/3}(\Delta^B)^{-1}-$SSE.

□

Can we improve this result with exp3 style exploration?

3.2 The harder setting: pull arm for sender only

Pessimistic strategy x^- for the leader,

$$\begin{aligned} \max_{x} \quad & xAe_j \\ s.t. \quad & xBe_j \geq xBe_{j'} \qquad \forall j \neq j', \quad \forall B \in C(\hat{B}) \end{aligned}$$

Optimistic strategy x^+ for the leader,

$$\begin{aligned} \max_{x} \quad & xAe_j \\ s.t. \quad & xBe_j \geq xBe_{j'} \qquad \forall j \neq j', \quad \exists B \in C(\hat{B}) \end{aligned}$$

If we view x^- as the exploitation, x^+ as the exploration. Can we make a balance by mixing them by some carefully chosen ratio?

References

[BJ20] Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pages 551–560. PMLR, 2020.

[BJWX21] Yu Bai, Chi Jin, Huan Wang, and Caiming Xiong. Sample-efficient learning of stackelberg equilibria in general-sum games. *arXiv preprint arXiv:2102.11494*, 2021.

[LCM09] Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. Learning and approximating the optimal strategy to commit to. In *International symposium on algorithmic game theory*, pages 250–262. Springer, 2009.

[PSTZ19] Binghui Peng, Weiran Shen, Pingzhong Tang, and Song Zuo. Learning optimal strategies to commit to. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2149–2156, 2019.