A Bayesian Learning Algorithm for Unknown Zero-sum Stochastic Games with an Arbitrary Opponent

Mehdi Jafarnia-Jahromi, Rahul Jain and Ashutosh Nayyar

USC University of Southern California

Overview

Motivations

Zero-sum stochastic games, pivotal in competitive reinforcement learning, epitomize complex decision-making challenges under uncertainty, mirroring real-world scenarios from economics to sports. The difficulty of mastering these games lies in optimizing strategies against an opponent that takes arbitrary time-adaptive history-dependent strategies.

- The main challenge in online learning in a Stochastic Game (SG) is the opponent's non-stationarity and uncontrollability.

Question: Is there an online learning algorithm for efficient exploration in unknown zero-sum stochastic games with an arbitrary time-adaptive history-dependent opponent?

Zero-sum Stochastic Games:

Stochastic Game $M = (\mathcal{S}, \mathcal{A}, r, \theta_*)$, only $\theta_* \sim \mu_1$ is unknown.

Learning Protocol

```
for $t = 1, \dots, T$ do
    The players observe state $s_t$ and simultaneously take actions $a_t = (a_t^1, a_t^2)$.
    The agent (maximizer) receives reward $r(s_t, a_t)$ from the opponent.
    The environment decides the next state $s_{t+1} \sim \theta(\cdot|s_t, a_t)$.
end
The agent receives $\sum_{t=1}^T r(s_t, a_t)$.
```

Goal: achieve low regret $R_T := \sup_{\pi^2 \in \Pi^{\mathsf{HR}}} \mathbb{E}\left[\sum_{t=1}^T (J(\theta_*) - r(s_t, a_t))\right]$, where $J(\theta_*)$ is the maximin average reward obtained by the agent and Π^{HR} is the space of history-dependent randomized policies and expectation is over $\theta_* \sim \mu_1$, $a_t^1 \sim \pi^1(\cdot|h_t)$, $a_t^2 \sim \pi^2(\cdot|h_t)$ and state dynamics.

Assumption (Finite-Diameter): There exists $D \geq 0$ such that for any stationary randomized policy $\pi^2 \in \Pi^{\mathsf{SR}}$ of the opponent and any $s, s' \in \mathcal{S} \times \mathcal{S}$, there exists a stationary randomized policy $\pi^1 \in \Pi^{\mathsf{SR}}$ of the agent, such that the expected time of reaching s' starting from s under policy $\pi = (\pi^1, \pi^2)$ does not exceed D, i.e.,

$$\max_{s,s'} \max_{\pi^2 \in \Pi^{\mathsf{SR}}} \min_{\pi^1 \in \Pi^{\mathsf{SR}}} T^{\pi}_{s \to s'} \leq D,$$

where $T^{\pi}_{s \to s'}$ is the expected time of reaching s' starting from s under policy $\pi = (\pi^1, \pi^2)$.

Our Contribution

- The first online RL algorithm ($\mathrm{PSRL\text{-}ZSG}$) that achieves Bayesian regret bound of $\tilde{O}(HS\sqrt{AT})$ in the infinite-horizon zero-sum stochastic games with average-reward criterion. Here H is an upper bound on the span of the bias function, S is the number of states, A is the number of joint actions and T is the horizon.
- This improves the best existing regret bound of $\tilde{O}(\sqrt[3]{DS^2AT^2})$ by Wei et al., 2017 under the same assumption and matches the theoretical lower bound in T.

PSRL-ZSG Algorithm

Algorithm 2 $\mathrm{PSRL\text{-}ZSG}$

```
Input: $\mu_1$
Initialization: $t \leftarrow 1, t_1 \leftarrow 0$
for episodes $k = 1, 2, \cdots$ do
    $T_{k-1} \leftarrow t - t_k$
    $t_k \leftarrow t$
    Generate $\theta_k \sim \mu_{t_k}$ and compute $\pi_k^1(\cdot)$ using Bellman equation.
    while $t \leq t_k + T_{k-1}$ and $N_t(s,a) \leq 2N_{t_k}(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ do
        Choose action $a_t^1 \sim \pi_k^1(\cdot|s_t)$ and observe $a_t^2$, $s_{t+1}$
        Update $\mu_{t+1}(d\theta) \propto \theta(s_{t+1}|s_t, a_t)\mu_t(d\theta)$.
        $t \leftarrow t + 1$
    end
end
```

Explanation:

- $\mathrm{PSRL\text{-}ZSG}$ proceeds in episodes (t_k : start of episode k and T_k : is length of it).
- In the beginning of each episode, the agent draws a sample of the transition kernel from the posterior distribution μ_{t_k}.
- The maximin strategy is then derived for the sampled transition kernel according to the Bellman equation and used by the agent during the episode.
- The first criterion, $t \leq t_k + T_{k-1}$, states that the length of the episode grows at most by 1 if the other criterion is not triggered. This ensures that $T_k \leq T_{k-1} + 1$ for all k.
- The second criterion is triggered if the number of visits to a state-action pair is doubled.
- These stopping criteria balance the trade-off between exploration and exploitation. In the beginning of the game, the episodes are short to motivate exploration since the agent is uncertain about the underlying environment. As the game proceeds, the episodes grow to exploit the information gathered about the environment.

Related Work (Wei et al., 2017)

- Wei et al., 2017 proposes an optimism-based algorithm (UCSG) that ahieves regret bound of $\tilde{O}(\sqrt[3]{DS^2AT^3}$. Our algorithm significantly improves this result and achieves a regret bound of $\tilde{O}(HS\sqrt{AT})$ under the finite-diameter assumption.
- From the analysis perspective, under the finite-diameter assumption, UCSG uses a sequence of finite-horizon SGs to approximate the average-reward SG and that leads to the sub-optimal regret bound of $O(T^{2/3})$. Our analysis avoids the finite-horizon approximation by directly using the Bellman equation in the infinite-horizon SG and achieves near-optimal regret bound.

Main Result

Theorem 2: Under the finite-diameter assumption, Algorithm 2 can achieve regret bound of $\tilde{O}(HS\sqrt{AT})$.

Proof Sketch

Bellman equation: under the finite-diameter assumption, there exist unique $J(\theta) \in \mathbb{R}$ and unique (upto an additive constant) function $v(\cdot, \theta) : \mathcal{S} \to \mathbb{R}$ that satisfy the Bellman equation, i.e., for all $s \in \mathcal{S}$,

$$J(\theta) + v(s, \theta) = \mathsf{val}\left\{r(s, \cdot, \cdot) + \sum_{s'} \theta(s'|s, \cdot, \cdot) v(s', \theta)\right\}. \tag{1}$$

In particular, the Nash equilibrium of the right hand side for each $s \in \mathcal{S}$ yields maximin stationary policies $\pi^* = (\pi^{1*}, \pi^{2*})$ such that

$$J(\theta) + v(s, \theta) = \max_{q^1 \in \Delta_{A^1}} \left\{r(s, q^1, \pi^{2*}(\cdot|s)) + \sum_{s'} \theta(s'|s, q^1, \pi^{2*}(\cdot|s)) v(s', \theta)\right\},$$
$$J(\theta) + v(s, \theta) = \min_{q^2 \in \Delta_{A^2}} \left\{r(s, \pi^{1*}(\cdot|s), q^2) + \sum_{s'} \theta(s'|s, \pi^{1*}(\cdot|s), q^2) v(s', \theta)\right\}.$$

Regret decomposition:

$$\begin{aligned} R_T(\pi^2) &= \mathbb{E}\left[TJ(\theta_*) - \sum_{t=1}^T r(s_t, a_t)\right] \\ &= \mathbb{E}\left[TJ(\theta_*) - \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} J(\theta_k)\right] + \mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} (J(\theta_k) - r(s_t, a_t))\right]. \end{aligned}$$

- The first term is bounded by $\mathbb{E}[K_T]$ where K_T is the number of episodes by time T. This uses a standard proof (using the property of thompson sampling and the first episode termination criterion). It can be proved that $\mathbb{E}[K_T] \leq \sqrt{2SAT \log T}$ by episode termination criteria.
- The second term is bounded by $H\mathbb{E}[K_T] + \tilde{O}(HS\sqrt{AT})$ and is the main part of the regret analysis. A key observation for proving this is:

Key observation: The policy π_k^1 used by the agent at episode k is the solution of the Nash equilibrium. Thus, for $t_k \leq t \leq t_{k+1} - 1$ and any $s \in \mathcal{S}$, the Nash equilibrium implies that

$$J(\theta_k) + v(s, \theta_k) \leq r(s, \pi_k^1(\cdot|s), q^2) + \sum_{s'} \theta_k(s'|s, \pi_k^1(\cdot|s), q^2) v(s', \theta_k),$$

for any distribution $q^2 \in \Delta_{A^2}$. Let $\pi^2 = (\pi_1^2, \pi_2^2, \cdots) \in \Pi^{\mathsf{HR}}$ be an arbitrary history-dependent randomized strategy for the opponent. Note that for any $t \geq 1$, π_t^2 is h_t-measurable. Replacing s by s_t and q^2 by $\pi_t^2(\cdot|h_t)$ implies that

$$\begin{aligned} &J(\theta_k) - r(s_t, \pi_k^1(\cdot|s_t), \pi_t^2(\cdot|h_t)) \\ &\leq \sum_{s'} \theta_k(s'|s_t, \pi_k^1(\cdot|s_t), \pi_t^2(\cdot|h_t)) v(s', \theta_k) - v(s_t, \theta_k). \end{aligned}$$