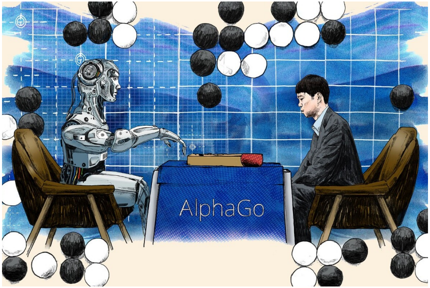
$\mathcal{O}(T^{-1})$ Convergence of Optimistic-Follow-the-Regularized-Leader in Two-Player Zero-Sum Markov Games

Yuepeng Yang

joint work with Cong Ma

University of Chicago, Department of Statistics

Problem setting



- Finding Nash equilibrium in a finite-horizon two-player zero-sum Markov games

Problem setting

1. At each horizon h of the game, the game is at a state s, max player draws an action $a \in \mathcal{A}$ from a policy μ, min player draws an action $b \in \mathcal{B}$ from a policy ν
2. Max player receives $r(s, a, b)$ reward, min player receives $-r(s, a, b)$ reward
3. Then the game goes to a new state s' in horizon $h + 1$. The transition depends on the played actions a, b.
4. Max/min player tries the maximize/minimize total expected reward

Nash equilibrium

Value function (for max player) of policy pair (μ, ν)

$$V(\mu, \nu) = \mathbb{E}_{\mu,\nu}\left[\sum_{i=1}^{H} r(s, a, b)\right]$$

Nash equilibrium: for two-player zero-sum Markov games, there exist policy pair $(\mu^\star, \nu^\star)$ such that

$$\inf_{\nu} \sup_{\mu} V(\mu, \nu) = V(\mu^\star, \nu^\star) = \sup_{\mu} \inf_{\nu} V(\mu, \nu)$$

The problem

- Full information with known reward and transition functions
- The goal is to find a pair of policy in T iterations such that no policy has ϵ-better expected reward V
- We focus on the dependency in T

Q function

- Fix a state s and horizon h
- $Q_h^{\mu,\nu}(a, b)$: expected reward when max player choose a and policy and min player choose b at horizon h and policy μ, ν at horizon $h + 1$ to H.

Q function

- Fix a state s and horizon h
- $Q_h^{\mu,\nu}(a, b)$: expected reward when max player choose a and policy and min player choose b at horizon h and policy μ, ν at horizon $h+1$ to H.
- In a normal form game where there is no state transition, Q is given and independent of μ, ν

  $Q_h^\top \mu_h, Q_h \nu_h$: utility vector for max and min player

Estimating the Q function in Markov games

- We are most interested in the payoff of Nash equilibrium $Q_h^* = Q_h^{\mu^\star,\nu^\star}$

Estimating the Q function in Markov games

- We are most interested in the payoff of Nash equilibrium $Q_h^* = Q_h^{\mu^\star,\nu^\star}$
- In each iteration, we learn a new policy for max and min player. It would be prohibitive to compute the full $Q_h^{\mu,\nu}(a, b)$.

Estimating the Q function in Markov games

- We are most interested in the payoff of Nash equilibrium $Q_h^* = Q_h^{\mu^\star,\nu^\star}$
- In each iteration, we learn a new policy for max and min player. It would be prohibitive to compute the full $Q_h^{\mu,\nu}(a,b)$.
- Alternatively, we maintain $Q_h^i(a,b)$ an estimate of $Q_h^*(a,b)$ and learn the policy in a fixed state s and horizon h as if it is a normal form game

Solving the policies with online learning

- In each iteration, we aim to learn the best policy for each player with respect to the estimates $\{Q_h^i\}_{i=1}^{t-1}$
- Linear loss functions

$$l^i_{\max,h}(\mu) = \left\langle \mu, Q_h^i \nu_h^i \right\rangle$$

$$l^i_{\min,h}(\nu) = \left\langle \nu, {Q_h^i}^\top \mu_h^i \right\rangle$$

$Q_h^\top \mu_h, Q_h \nu_h$: utility vector for max and min player

Optimistic Follow-the-Regularzied Leader

Choose the regularized optimal policy (leader) with respect to reward function Q in previous iterations.

$$\mu_h^t(a) \propto \exp\left(\frac{\eta}{w_t}\left[\sum_{i=1}^{t-1} w_i\left[Q_h^i \nu_h^i\right](a) + \underbrace{w_t\left[Q_h^{t-1}\nu_h^{t-1}\right](a)}_{\text{optimistic term}}\right]\right)$$

$$\nu_h^t(b) \propto \exp\left(\frac{\eta}{w_t}\left[\sum_{i=1}^{t-1} w_i\left[{Q_h^i}^\top \mu_h^i\right](b) + w_t\left[{Q_h^{t-1}}^\top \mu_h^{t-1}\right](b)\right]\right)$$

Learning with optimism

$$\mu^t(a) = \text{argmax}_\mu \left\langle \mu, \left[\sum_{i=1}^{t-1} w_i \boldsymbol{u}^i + \underbrace{w_t \boldsymbol{M}^t}_{\text{optimistic term}} \right] \right\rangle - \frac{R(\mu)}{\eta / w_t}$$

where $\boldsymbol{u}^i = Q^i \nu^i, \boldsymbol{M}^t = Q^{t-1} \nu^{t-1}$

Learning with optimism

$$\mu^t(a) = \mathsf{argmax}_\mu \left\langle \mu, \left[\sum_{i=1}^{t-1} w_i \boldsymbol{u}^i + \underbrace{w_t \boldsymbol{M}^t}_{\text{optimistic term}} \right] \right\rangle - \frac{R(\mu)}{\eta / w_t}$$

where $\boldsymbol{u}^i = Q^i \nu^i, \boldsymbol{M}^t = Q^{t-1} \nu^{t-1}$

- Increasing weight that favors more recent iterations

Learning with optimism

$$\mu^t(a) = \text{argmax}_\mu \left\langle \mu, \left[\sum_{i=1}^{t-1} w_i \boldsymbol{u}^i + \underbrace{w_t \boldsymbol{M}^t}_{\text{optimistic term}} \right] \right\rangle - \frac{R(\mu)}{\eta / w_t}$$

where $\boldsymbol{u}^i = Q^i \nu^i, \boldsymbol{M}^t = Q^{t-1} \nu^{t-1}$

- Increasing weight that favors more recent iterations
- Optimistic term $\boldsymbol{M}^t$ predicts $\boldsymbol{u}^t$

  Rakhlin and Sridharan (2013) *Online Learning with Predictable Sequences*

Performance on normal form games

- FTRL works for normal form games with $\tilde{O}(1/\sqrt{T})$ rate

Performance on normal form games

- FTRL works for normal form games with $\tilde{O}(1/\sqrt{T})$ rate
- OFTRL works for normal form games with $\tilde{O}(1/T)$ rate

  Daskalakis et al. (2021) *Near-Optimal No-Regret Learning in General Games.*

  Anagnostides et al. (2022) *Uncoupled Learning Dynamics with* $O(\log T)$ *Swap Regret in Multiplayer Games*

Algorithm for Markov Games

Zhang et al., (2022) *Policy Optimization for Markov Games: Unified Framework and Faster Convergence*

Algorithm for Markov Games

Zhang et al., (2022) *Policy Optimization for Markov Games: Unified Framework and Faster Convergence*

- Policy update with OFTRL

Algorithm for Markov Games

Zhang et al., (2022) *Policy Optimization for Markov Games: Unified Framework and Faster Convergence*

- Policy update with OFTRL
- Smooth value update:

$$Q_h^t(a,b) = (1-\alpha_t)Q_h^{t-1}(a,b) + \alpha_t\left(r(a,b) + \tilde{Q}_{h+1}^t(a,b)\right)$$

$\tilde{Q}_{h+1}^t(a,b) =$ expected reward of horizon $h+1$ to H after the transition when μ^t, ν^t is played

Algorithm for Markov Games

Zhang et al., (2022) *Policy Optimization for Markov Games: Unified Framework and Faster Convergence*

- Policy update with OFTRL
- Smooth value update:

$$Q_h^t(a,b) = (1-\alpha_t)Q_h^{t-1}(a,b) + \alpha_t \left(r(a,b) + \tilde{Q}_{h+1}^t(a,b)\right)$$

  $\tilde{Q}_{h+1}^t(a,b) =$ expected reward of horizon $h+1$ to H after the transition when μ^t, ν^t is played

- Output mixture policies

$$\hat{\mu}_h(\cdot \mid s) := \sum_{t=1}^{T} \alpha_T^t \mu_h^t(\cdot \mid s)$$

Theoretical challenges in Markov Games

- Except for the last horizon, the Nash equilibrium pay-off matrix $Q^\star$ is not available
- In two-player zero-sum normal form game, the sum of regrets is always non-negative. This fails in Markov game because of approximation in $Q^\star$
- Aggregation of estimation errors and regrets over horizons of the game.

Main theorem

- Quantification of the gap to Nash equilibrium:

$$V(\mu, \nu) = \mathbb{E}_{\mu,\nu}\left[\sum_{i=1}^{H} r(s, a, b)\right]$$

$$\mathrm{NEgap}(\mu, \nu) := \sup_{\mu^\dagger} V(\mu^\dagger, \nu) - \inf_{\nu^\dagger} V(\mu, \nu^\dagger)$$

Main theorem

- Quantification of the gap to Nash equilibrium:

$$V(\mu, \nu) = \mathbb{E}_{\mu,\nu}\left[\sum_{i=1}^{H} r(s, a, b)\right]$$

$$\mathrm{NEgap}(\mu, \nu) := \sup_{\mu^\dagger} V(\mu^\dagger, \nu) - \inf_{\nu^\dagger} V(\mu, \nu^\dagger)$$

- Zhang et al., (2022) *Policy Optimization for Markov Games: Unified Framework and Faster Convergence*:

  $\mathrm{NEgap}(\hat{\mu}, \hat{\nu}) = \tilde{O}(T^{-5/6})$ with empirical evidence for $O(T^{-1})$.

Main theorem

- Quantification of the gap to Nash equilibrium:

$$V(\mu,\nu) = \mathbb{E}_{\mu,\nu}\left[\sum_{i=1}^{H} r(s,a,b)\right]$$

$$\mathrm{NEgap}(\mu,\nu) := \sup_{\mu^\dagger} V(\mu^\dagger,\nu) - \inf_{\nu^\dagger} V(\mu,\nu^\dagger)$$

Theorem 1

For $(\hat{\mu},\hat{\nu})$ the output of the policy optimization algorithm using OFTRL with appropriately choosen stepsize η,

$$\mathrm{NEgap}(\hat{\mu},\hat{\nu}) \lesssim O(H^5/T)$$

Classical analysis framework

- Goal: control the regret of max player

Classical analysis framework

- Goal: control the regret of max player
- Regret bounded in Varition of Utility (RVU property)

$$\underbrace{\max_{\mu^\dagger} \sum_{i=1}^{t} \langle \mu^\dagger - \mu^i, \alpha_t^i \boldsymbol{u}^i \rangle}_{\mathrm{reg}_1^t} \le \alpha + \beta \underbrace{\sum_{i=1}^{t} \|\boldsymbol{u}^i - \boldsymbol{u}^{i-1}\|_*}_{\text{utility}} - \gamma \sum_{i=1}^{t} \|\mu^i - \mu^{i-1}\|$$

Classical analysis framework

- Goal: control the regret of max player
- Regret bounded in Varition of Utility (RVU property)

$$\underbrace{\max_{\mu^\dagger}\sum_{i=1}^{t}\langle \mu^\dagger - \mu^i, \alpha_t^i \boldsymbol{u}^i\rangle}_{\mathrm{reg}_1^t} \leq \alpha + \beta \underbrace{\sum_{i=1}^{t}\|\boldsymbol{u}^i - \boldsymbol{u}^{i-1}\|_*}_{\text{utility}} - \gamma \sum_{i=1}^{t}\|\mu^i - \mu^{i-1}\|$$

- Not avaliable for FTRL

RVU bounds

- Variation in utility is typically controlled by variation of ν, i.e. opponent's policy

$$\mathrm{reg}_1^t \leq \alpha + \beta \sum_{i=1}^{t} \|\nu^i - \nu^{i-1}\|_1 - \gamma \sum_{i=1}^{t} \|\mu^i - \mu^{i-1}\|_1$$

$$\mathrm{reg}_2^t \leq \alpha + \beta \sum_{i=1}^{t} \|\mu^i - \mu^{i-1}\|_1 - \gamma \sum_{i=1}^{t} \|\nu^i - \nu^{i-1}\|_1$$

RVU bounds

- Variation in utility is typically controlled by variation of ν, i.e. opponent's policy

$$\mathrm{reg}_1^t \leq \alpha + \beta \sum_{i=1}^{t} \|\nu^i - \nu^{i-1}\|_1 - \gamma \sum_{i=1}^{t} \|\mu^i - \mu^{i-1}\|_1$$

$$\mathrm{reg}_2^t \leq \alpha + \beta \sum_{i=1}^{t} \|\mu^i - \mu^{i-1}\|_1 - \gamma \sum_{i=1}^{t} \|\nu^i - \nu^{i-1}\|_1$$

- RVU bound for sum of regrets

$$\mathrm{reg}_1^t + \mathrm{reg}_2^t \leq 2\alpha + (\beta - \gamma) \left(\sum_{i=1}^{t} \|\nu^i - \nu^{i-1}\|_1 + \|\mu^i - \mu^{i-1}\|_1 \right)$$

Non-negativity of the sum of regrets

In some settings, sum of regrets of the players are non-negative (two-player zero-sum normal form game, swap regrets, etc.)

$$\sum_{i=1}^{t} \|\nu^i - \nu^{i-1}\|_1 + \|\mu^i - \mu^{i-1}\|_1 \leq \frac{2\alpha}{\gamma - \beta}$$

Now one can control the variation in utility and thus control individual regrets

Analysis in Markov game

- Two parts of NE-gap: payoff regrets and estimation error of $Q^\star$

$$\text{NE-gap} \leq \mathrm{reg}_{1,h}^t + \mathrm{reg}_{2,h}^t + \sum_{t,h} \alpha_T^t \delta_h^t$$

where $\delta_h^t = \|Q_h^t - Q_h^\star\|_\infty$

Analysis in Markov game

- Two parts of NE-gap: payoff regrets and estimation error of $Q^\star$

$$\text{NE-gap} \le \mathrm{reg}_{1,h}^t + \mathrm{reg}_{2,h}^t + \sum_{t,h} \alpha_T^t \delta_h^t$$

where $\delta_h^t = \|Q_h^t - Q_h^\star\|_\infty$

- Last horizon is a normal form game so $\delta_H^t = 0$. Estimation error aggregate through the horizons

$$\delta_h^t \le \sum_{i=1}^{t} \alpha_t^i \delta_{h+1}^i + \max_{s,m=1,2} \mathrm{reg}_{m,h+1}^t$$

Approximate non-negativity

- Non-negativity of sum of regrets fails in Markorv games.

Approximate non-negativity

- Non-negativity of sum of regrets fails in Markorv games.
- Solution: *approximate non-negativity*

$$\mathrm{reg}_{\mu}^{t} + \mathrm{reg}_{\nu}^{t} \geq -2 \sum_{i=1}^{t} \delta_h^t$$

Recall $\delta_h^t = \|Q_h^t - Q_h^{\star}\|_{\infty}$

Aggregation of estimation error and regret

- Interwined error aggregation through the horizons

$$\mathrm{reg}_{\mu}^{t} \leq O(1/t) + \underbrace{\sum_{i=1}^{t} \alpha_t^i \delta_h^t}_{\text{extra term for our analysis}}$$

$$\delta_h^t \leq \sum_{i=1}^{t} \alpha_t^i \delta_{h+1}^i + \max_{\mu,\nu,s} \mathrm{reg}_{h+1}^t$$

- Naive approach of uniform weighting leads to a multiplicative factor of $\log T$ at each horizon

Log-free weights

- Weights $\{w^i\}$ and its normlized version $\{\alpha_t^i\}$ from Jin et al. (2018) *Is Q-learning provably efficient?*
- Increasing weight favors recent development

$$\sum_{i=1}^{T} \frac{1}{T} \cdot \frac{1}{i} \approx \frac{\log T}{T} \quad \text{vs.} \sum_{i=1}^{T} \alpha_T^i \cdot \frac{1}{i} = \left(1 + \frac{1}{H}\right) \frac{1}{T}$$

Log-free weights

- Weights $\{w^i\}$ and its normlized version $\{\alpha_t^i\}$ from Jin et al. (2018) *Is Q-learning provably efficient?*

- Increasing weight favors recent development

$$\sum_{i=1}^{T} \frac{1}{T} \cdot \frac{1}{i} \approx \frac{\log T}{T} \quad \text{vs.} \sum_{i=1}^{T} \alpha_T^i \cdot \frac{1}{i} = \left(1 + \frac{1}{H}\right) \frac{1}{T}$$

- Reduce estimation error of Q_h aggregated over the horizons: $(\log T)^H \to (1 + 1/H)^H$.

Summary

- Use OFTRL to solve Nash equilibrium in a two-player zero-sum Markov game
- Improve the analysis to show that the algorithm finds $O(1/T)$-approximate Nash equilibrium
- Careful treatment of the interwined estimation error and payoff regret aggregating over horizons
    - Approximate non-negativity of sum of regrets
    - Log-free weights

Potential extensions

- OFTRL in multiplayer general-sum Markov games to find correlated equilibrium
- Other notions of regret (interal regret, swap regret, etc.)
- Use of approximate non-negativity in other related problems