Near-Optimal Multi-Agent RL with Markov Game Self-Play

Winston Zeng

University of Arizona

November 14 2023

Outline

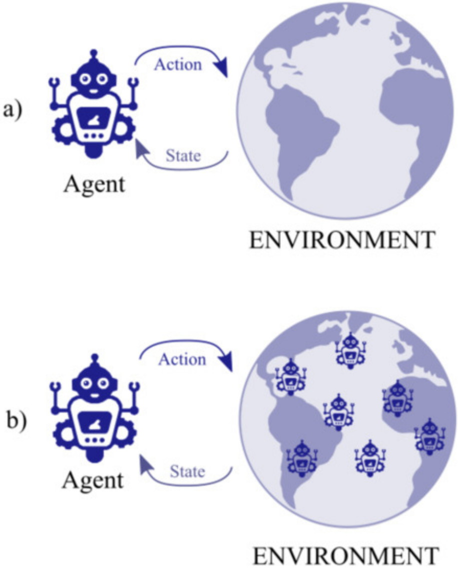
1. Background
   - Multi-Agent RL, Markov Games, Nash Eq - Prisoner's Dilemma
   - 2-Player Zero-Sum Markov Games, Breakdown of 2PZSMG
2. Setup
   - Policy & Value Functions
   - Best Response & Nash Equilibrium
   - Learning Objective
   - Prior Literature Results
   - Model-based vs Model-free
3. Algorithm
   - Optimistic Nash Value Iteration
   - Nash-VI Algorithm
   - Main Improvement
   - Bernstein's Inequality
   - Theoretical Guarantee
4. Other
   - Further Research
   - References

Multi-Agent RL

Multiple agents learn to make decisions in an unknown environment in order to maximize their (own) cumulative rewards

Significant recent success in traditionally hard AI challenges:

- large-scale strategy games
- real-time team-based video games
- behavior learning in complex social scenarios



Markov Games

- Generalization of MDPs into multiplayer setting, and applying game theory
- Focus on two-player zero-sum Markov Games
- Goal: Find a Nash Equilibrium

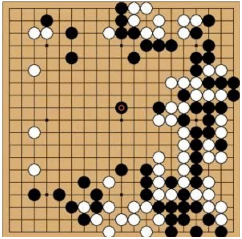


Figure: Go

Figure: Poker

Nash Eq Example - Prisoner's Dilemma

		Player 2	
		confess	don't confess
Player 1	confess	(-6, -6)	(0, -10)
	don't confess	(-10, 0)	(-1, -1)

2-Player Zero-Sum Markov Games (2PZSMG)

- One max player, one min player
- One reward function $R(s, a, b)$
- Zero sum: negative reward is readily applied
- Nash equilibrium: best responses, no gain by changing policies

Breakdown of 2PZSMG

Denoted as $MG(H, S, A, B, \mathbb{P}, r)$

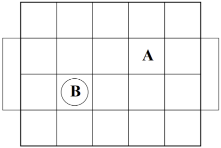
Finite horizon episodic MDPs: $M = (S, A, H, (R_h)_{h=1}^{H}, (P_h)_{h=1}^{H}, \mu)$

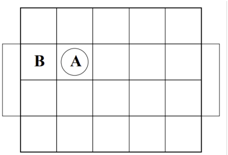
Start at initial state $s_1 \in S$

At each step $h \in [H]$:

- observe state $s_h \in S$
- pick actions $a_h \in A, b_h \in B$ simultaneously
- each player observes actions of opponent
- players receive reward $r_h(s_h, a_h, b_h)$
- game transitions to next state $s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h, b_h)$

Episode ends with state s_{H+1} is reached





Policy & Value Functions

- Markov policy of max player: $\mu = \{\mu_h : S \to \Delta_A\}_{h \in [H]}$
  $\mu_h(s|a) =$ prob. of taking a at s under μ.

Policy & Value Functions

- Markov policy of max player: $\mu = \{\mu_h : S \to \Delta_A\}_{h\in[H]}$
  $\mu_h(s|a) =$ prob. of taking a at s under μ.
- Markov policy of min player: $\nu = \{\nu_h : S \to \Delta_B\}_{h\in[H]}$
  $\nu_h(s|b) =$ prob. of taking b at s under ν.

Policy & Value Functions

- Markov policy of max player: $\mu = \{\mu_h : S \to \Delta_A\}_{h \in [H]}$
  $\mu_h(s|a) =$ prob. of taking a at s under μ.
- Markov policy of min player: $\nu = \{\nu_h : S \to \Delta_B\}_{h \in [H]}$
  $\nu_h(s|b) =$ prob. of taking b at s under ν.
- Value function $V_h^{\mu,\nu} : S \to R$

$$V_h^{\mu,\nu}(s) := \mathbb{E}_{\mu,\nu}[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'})|s_h = s]$$

Policy & Value Functions

- Markov policy of max player: $\mu = \{\mu_h : S \to \Delta_A\}_{h\in[H]}$
  $\mu_h(s|a) =$ prob. of taking a at s under μ.
- Markov policy of min player: $\nu = \{\nu_h : S \to \Delta_B\}_{h\in[H]}$
  $\nu_h(s|b) =$ prob. of taking b at s under ν.
- Value function $V_h^{\mu,\nu} : S \to R$

$$V_h^{\mu,\nu}(s) := \mathbb{E}_{\mu,\nu}[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'})|s_h = s]$$

- Q-value function $Q_h^{\mu,\nu} : S \times A \times B \to R$:

$$Q_h^{\mu,\nu}(s,a,b) := \mathbb{E}_{\mu,\nu}[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'})|s_h = s, a_h = a, b_h = b]$$

- These also help derive the Bellman equations, but are not relevant to theoretical guarantee.

Best Response & Nash Equilibrium

For any policy of the max-player μ, there exists a *best response* of the min-player, which is a policy $\nu^\dagger(\mu)$ satisfying

$$V_h^{\mu,\dagger} := V_h^{\mu,\nu^\dagger(\mu)}(s) = inf_\nu V_h^{\mu,\nu}(s)$$

for any $(s, h) \in S \times [H]$. By symmetry, also define $\mu^\dagger(\nu)$ and $V_h^{\dagger,\nu}$.

Best Response & Nash Equilibrium

For any policy of the max-player μ, there exists a *best response* of the min-player, which is a policy $\nu^{\dagger}(\mu)$ satisfying

$$V_h^{\mu,\dagger} := V_h^{\mu,\nu^{\dagger}(\mu)}(s) = inf_{\nu} V_h^{\mu,\nu}(s)$$

for any $(s,h) \in S \times [H]$. By symmetry, also define $\mu^{\dagger}(\nu)$ and $V_h^{\dagger,\nu}$. There also exist policies μ^*, ν^* that are optimal against opponents' best responses, denoted as

$$V_h^{\mu^*,\dagger}(s) = sup_{\mu} V_h^{\mu,\dagger}(s), V_h^{\dagger,\nu^*}(s) = inf_{\nu} V_h^{\dagger,\nu}(s) \ \forall (s,h)$$

Best Response & Nash Equilibrium

For any policy of the max-player μ, there exists a *best response* of the min-player, which is a policy $\nu^\dagger(\mu)$ satisfying

$$V_h^{\mu,\dagger} := V_h^{\mu,\nu^\dagger(\mu)}(s) = inf_\nu V_h^{\mu,\nu}(s)$$

for any $(s,h) \in S \times [H]$. By symmetry, also define $\mu^\dagger(\nu)$ and $V_h^{\dagger,\nu}$. There also exist policies μ^*, ν^* that are optimal against opponents' best responses, denoted as

$$V_h^{\mu^*,\dagger}(s) = sup_\mu V_h^{\mu,\dagger}(s), V_h^{\dagger,\nu^*}(s) = inf_\nu V_h^{\dagger,\nu}(s) \; \forall (s,h)$$

We call these optimal strategies (μ^*, ν^*) the Nash equilibrium, which satisfies:

$$sup_\mu inf_\nu V_h^{\mu,\nu}(s) = V_h^{\mu^*,\nu^*}(s) = inf_\nu sup_\mu V_h^{\mu,\nu}(s)$$

Best Response & Nash Equilibrium

For any policy of the max-player μ, there exists a *best response* of the min-player, which is a policy $\nu^\dagger(\mu)$ satisfying

$$V_h^{\mu,\dagger} := V_h^{\mu,\nu^\dagger(\mu)}(s) = inf_\nu V_h^{\mu,\nu}(s)$$

for any $(s, h) \in S \times [H]$. By symmetry, also define $\mu^\dagger(\nu)$ and $V_h^{\dagger,\nu}$. There also exist policies μ^*, ν^* that are optimal against opponents' best responses, denoted as

$$V_h^{\mu^*,\dagger}(s) = sup_\mu V_h^{\mu,\dagger}(s), V_h^{\dagger,\nu^*}(s) = inf_\nu V_h^{\dagger,\nu}(s) \ \forall(s, h)$$

We call these optimal strategies (μ^*, ν^*) the Nash equilibrium, which satisfies:

$$sup_\mu inf_\nu V_h^{\mu,\nu}(s) = V_h^{\mu^*,\nu^*}(s) = inf_\nu sup_\mu V_h^{\mu,\nu}(s)$$

The Nash equilibrium policies $V_h^{\mu^*,\nu^*}$ and $Q_h^{\mu^*,\nu^*}$ are abbreviated to V_h^* and Q_h^*. In words, there is no incentive (w.r.t reward) to deviate to a different policy.

Learning Objective

Suboptimality of any pair of policies $(\hat{\mu}, \hat{\nu})$ is the gap b/w their performance and that of the optimal strategies (Nash eq.) when played against the best responses, respectively:

$$V_1^{\dagger,\hat{\nu}}(s_1) - V_1^{\hat{\mu},\dagger}(s_1) = [V_1^{\dagger,\hat{\nu}}(s_1) - V_1^{*}(s_1)] + [V_1^{*}(s_1) - V_1^{\hat{\mu},\dagger}(s_1)] \leq \epsilon$$

Learning Objective

Suboptimality of any pair of policies $(\hat{\mu}, \hat{\nu})$ is the gap b/w their performance and that of the optimal strategies (Nash eq.) when played against the best responses, respectively:

$$V_1^{\dagger,\hat{\nu}}(s_1) - V_1^{\hat{\mu},\dagger}(s_1) = [V_1^{\dagger,\hat{\nu}}(s_1) - V_1^*(s_1)] + [V_1^*(s_1) - V_1^{\hat{\mu},\dagger}(s_1)] \leq \epsilon$$

Regret: Let (μ^k, ν^k) denote the policies deployed by the algorithm in the kth episode. After a total of K episodes, the regret is defined as

$$Reg(K) = \sum_{k=1}^{K} (V_1^{\dagger,\nu^k} - V_1^{\mu^k,\dagger})(s_1)$$

Learning Objective

Suboptimality of any pair of policies $(\hat{\mu}, \hat{\nu})$ is the gap b/w their performance and that of the optimal strategies (Nash eq.) when played against the best responses, respectively:

$$V_1^{\dagger,\hat{\nu}}(s_1) - V_1^{\hat{\mu},\dagger}(s_1) = [V_1^{\dagger,\hat{\nu}}(s_1) - V_1^{*}(s_1)] + [V_1^{*}(s_1) - V_1^{\hat{\mu},\dagger}(s_1)] \leq \epsilon$$

Regret: Let (μ^k, ν^k) denote the policies deployed by the algorithm in the kth episode. After a total of K episodes, the regret is defined as

$$Reg(K) = \sum_{k=1}^{K} (V_1^{\dagger,\nu^k} - V_1^{\mu^k,\dagger})(s_1)$$

One goal of RL is to design algorithms for Markov games that can find an ϵ-approx. Nash eq. within a # of episodes that is small in complexity, i.e. small in dependency on S, A, B, H, and $1/\epsilon$ (PAC sample complexity bound).

Prior Literature Results

Table 1: Sample complexity (the required number of episodes) for algorithms to find ϵ-approximate Nash equlibrium policies in zero-sum Markov games: VI-explore and VI-UCLB (Bai and Jin, 2020), OMVI-SM (Xie et al., 2020), and Nash Q/V-learning (Bai et al., 2020). The lower bound is proved by Jin et al. (2018); Domingues et al. (2020).

	Algorithm	**Task-Agnostic**	**$\sqrt{T}$-Regret**	**Sample Complexity**	**Output Policy**
Model-based	VI-explore	Yes		$\tilde{O}(H^5S^2AB/\epsilon^2)$	a single Markov policy
	VI-ULCB		Yes	$\tilde{O}(H^4S^2AB/\epsilon^2)$	
	OMVI-SM		Yes	$\tilde{O}(H^4S^3A^3B^3/\epsilon^2)$	
	Algorithm 2	Yes		$\tilde{O}(H^4SAB/\epsilon^2)$	
	Algorithm 1		Yes	$\tilde{O}(H^3SAB/\epsilon^2)$	
Model-free	Nash Q-learning			$\tilde{O}(H^5SAB/\epsilon^2)$	a nested mixture of Markov policies
	Nash V-learning			$\tilde{O}(H^6S(A+B)/\epsilon^2)$	
	Lower Bound	-	-	$\Omega(H^3S(A+B)/\epsilon^2)$	-

Nash-VI matches information theoretic lower bound except for a $\tilde{O}(min\{A, B\})$ factor, showing that model-based algorithms can achieve an almost optimal sample complexity.

Model-based vs Model-free

Model-based:

- Using existing data to build an estimate of model
- Run offline planning algorithm on estimated model to get policy
- Play policy in environment

Model-free:

- Directly estimate value/action-value functions
- Play greedy policies w.r.t. estimated value functions

Optimistic Nash-VI

- At high level, standard strategy in majority of model-based RL algorithms
  - Optimistic planning from estimated model
  - Play policy and update model estimate
- Underlies provably efficient model-based algorithms

Nash-VI Algorithm

Algorithm 1 Optimistic Nash Value Iteration (Nash-VI)

1: **Initialize:** for any (s, a, b, h), $\overline{Q}_h(s, a, b) \leftarrow H$, $\underline{Q}_h(s, a, b) \leftarrow 0$, $\Delta \leftarrow H$, $N_h(s, a, b) \leftarrow 0$.
2: **for** episode $k = 1, \ldots, K$ **do**

Nash-VI Algorithm

Algorithm 1 Optimistic Nash Value Iteration (Nash-VI)

Initialize: for any (s,a,b,h), $\overline{Q}_h(s,a,b) \leftarrow H$, $\underline{Q}_h(s,a,b) \leftarrow 0$, $\Delta \leftarrow H$, $N_h(s,a,b) \leftarrow 0$.
for episode $k = 1, \dots, K$ **do**
  for step $h = H, H-1, \dots, 1$ **do**
    for $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ **do**
      $t \leftarrow N_h(s,a,b)$.
      if $t > 0$ **then**
        $\beta \leftarrow \textsc{Bonus}(t, \widehat{\mathbb{V}}_h[(\overline{V}_{h+1} + \underline{V}_{h+1})/2](s,a,b))$.
        $\gamma \leftarrow (c/H)\widehat{\mathbb{P}}_h(\overline{V}_{h+1} - \underline{V}_{h+1})(s,a,b)$.
        $\overline{Q}_h(s,a,b) \leftarrow \min\{(r_h + \widehat{\mathbb{P}}_h \overline{V}_{h+1})(s,a,b) + \gamma + \beta, H\}$.
        $\underline{Q}_h(s,a,b) \leftarrow \max\{(r_h + \widehat{\mathbb{P}}_h \underline{V}_{h+1})(s,a,b) - \gamma - \beta, 0\}$.
    for $s \in \mathcal{S}$ **do**
      $\pi_h(\cdot,\cdot|s) \leftarrow \mathrm{CCE}(\overline{Q}_h(s,\cdot,\cdot), \underline{Q}_h(s,\cdot,\cdot))$.
      $\overline{V}_h(s) \leftarrow (\mathbb{D}_{\pi_h}\overline{Q}_h)(s)$; $\underline{V}_h(s) \leftarrow (\mathbb{D}_{\pi_h}\underline{Q}_h)(s)$.

Nash-VI Algorithm

Algorithm 1 Optimistic Nash Value Iteration (Nash-VI)

Initialize: for any (s, a, b, h), $\overline{Q}_h(s, a, b) \leftarrow H$, $\underline{Q}_h(s, a, b) \leftarrow 0$, $\Delta \leftarrow H$, $N_h(s, a, b) \leftarrow 0$.
for episode $k = 1, \ldots, K$ **do**
  for step $h = H, H-1, \ldots, 1$ **do**
    for $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ **do**
      $t \leftarrow N_h(s, a, b)$.
      if $t > 0$ **then**
        $\beta \leftarrow \textsc{Bonus}(t, \widehat{\mathbb{V}}_h[(\overline{V}_{h+1} + \underline{V}_{h+1})/2](s, a, b))$.
        $\gamma \leftarrow (c/H)\widehat{\mathbb{P}}_h(\overline{V}_{h+1} - \underline{V}_{h+1})(s, a, b)$.
        $\overline{Q}_h(s, a, b) \leftarrow \min\{(r_h + \widehat{\mathbb{P}}_h \overline{V}_{h+1})(s, a, b) + \gamma + \beta, H\}$.
        $\underline{Q}_h(s, a, b) \leftarrow \max\{(r_h + \widehat{\mathbb{P}}_h \underline{V}_{h+1})(s, a, b) - \gamma - \beta, 0\}$.
    for $s \in \mathcal{S}$ **do**
      $\pi_h(\cdot, \cdot | s) \leftarrow \mathrm{CCE}(\overline{Q}_h(s, \cdot, \cdot), \underline{Q}_h(s, \cdot, \cdot))$.
      $\overline{V}_h(s) \leftarrow (\mathbb{D}_{\pi_h} \overline{Q}_h)(s)$; $\underline{V}_h(s) \leftarrow (\mathbb{D}_{\pi_h} \underline{Q}_h)(s)$.
  for step $h = 1, \ldots, H$ **do**
    take action $(a_h, b_h) \sim \pi_h(\cdot, \cdot | s_h)$, observe reward r_h and next state s_{h+1}.
    add 1 to $N_h(s_h, a_h, b_h)$ and $N_h(s_h, a_h, b_h, s_{h+1})$.
    $\widehat{\mathbb{P}}_h(\cdot | s_h, a_h, b_h) \leftarrow N_h(s_h, a_h, b_h, \cdot)/N_h(s_h, a_h, b_h)$.
Output the marginal policies of π^{out}: $(\mu^{\text{out}}, \nu^{\text{out}})$.

Main Improvement

Auxiliary bonus γ in addition to standard bonus β:

- γ is computed by applying the empirical transition matrix to the gap at the next step, whereas

Main Improvement

Auxiliary bonus γ in addition to standard bonus β:

- γ is computed by applying the empirical transition matrix to the gap at the next step, whereas
- Standard bonus β is typically designed according to concentration inequalities: VI-ULCB's $\beta = \tilde{\Theta}(\sqrt{S/t})$

Main Improvement

Auxiliary bonus γ in addition to standard bonus β:

- γ is computed by applying the empirical transition matrix to the gap at the next step, whereas
- Standard bonus β is typically designed according to concentration inequalities: VI-ULCB's $\beta = \tilde{\Theta}(\sqrt{S/t})$
- Aux bonus $\gamma \Rightarrow$ smaller choice of $\beta = \tilde{O}(\sqrt{1/t})$

  $\Rightarrow$ removes S factor from complexity

Main Improvement

Auxiliary bonus γ in addition to standard bonus β:

- γ is computed by applying the empirical transition matrix to the gap at the next step, whereas
- Standard bonus β is typically designed according to concentration inequalities: VI-ULCB's $\beta = \tilde{\Theta}(\sqrt{S/t})$
- Aux bonus $\gamma \Rightarrow$ smaller choice of $\beta = \tilde{O}(\sqrt{1/t})$

  $\Rightarrow$ removes S factor from complexity
- Two choices of bonus function $\beta = BONUS(t, \hat{\sigma}^2)$:

  Hoeffding: $c(\sqrt{H^2\iota/t} + H^2 S\iota/t)$, Bernstein: $c(\sqrt{\hat{\sigma}^2\iota/t} + H^2 S\iota/t)$

Main Improvement

Auxiliary bonus γ in addition to standard bonus β:

- γ is computed by applying the empirical transition matrix to the gap at the next step, whereas
- Standard bonus β is typically designed according to concentration inequalities: VI-ULCB's $\beta = \tilde{\Theta}(\sqrt{S/t})$
- Aux bonus $\gamma \Rightarrow$ smaller choice of $\beta = \tilde{O}(\sqrt{1/t})$

  $\Rightarrow$ removes S factor from complexity
- Two choices of bonus function $\beta = BONUS(t, \hat{\sigma}^2)$:

  Hoeffding: $c(\sqrt{H^2\iota/t} + H^2 S\iota/t)$, Bernstein: $c(\sqrt{\hat{\sigma}^2\iota/t} + H^2 S\iota/t)$
- Bernstein bonus uses a sharper concentration, saving an H factor in sample complexity compared to Hoeffding, which reduces SC to $\tilde{O}(H^3SAB/\epsilon^2)$, matches ITLB in H, S, ϵ factors.

Bernstein's Inequality

Bernstein's inequality:

$X_1, ..., X_n$ iid RVs, $|X_i - \mathbb{E}X_i| \leq R$. $\mathbb{E}[X_i] = \mu$, $Var(X_i) = \sigma^2$

$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2exp(-\frac{n\epsilon^2}{2\sigma^2 + \frac{2}{3}R \cdot \epsilon})$

$\sigma^2 << (B - A)^2$ much sharper than Hoeffding

Recall the suboptimality:

$$V_1^{\dagger,\hat{\nu}}(s_1) - V_1^{\hat{\mu},\dagger}(s_1) = [V_1^{\dagger,\hat{\nu}}(s_1) - V_1^*(s_1)] + [V_1^*(s_1) - V_1^{\hat{\mu},\dagger}(s_1)] \leq \epsilon$$

Recall the alternative objective of designing algorithm with sublinear in K regret. $Reg(K) \leq \sqrt{k}$, divide both sides by k gives $\frac{1}{\sqrt{k}}$.
choosing large enough k gives $\sqrt{k} \leq \epsilon$
uniformly and randomly return μ^k, ν^k
$\mathbb{E}(subopt) = \frac{1}{k}\sum_{k=1}^{K} subopt$= avg regret

Theoretical Guarantee

Nash-VI with Bernstein bonus

For any $p \in (0, 1]$ and letting $\iota = log(SABT/p)$, then with probability $\geq 1 - p$, Nash-VI with Bernstein type bonus (with some absolute $c > 0$) achieves:

- $(V_1^{\dagger,\nu^{out}} - V_1^{\mu^{out},\dagger})(s_1) \leq \epsilon$, if the number of episodes $K \geq \Omega(H^3 SAB\iota/\epsilon^2 + H^3 S^2 AB\iota^2/\epsilon)$
- $Reg(K) = \sum_{k=1}^{K}(V_1^{\dagger,\nu^k} - V_1^{\mu^k,\dagger})(s_1) \leq O(\sqrt{H^2 SABT\iota} + H^3 S^2 AB\iota^2)$

Theoretical Guarantee

Nash-VI with Bernstein bonus

For any $p \in (0, 1]$ and letting $\iota = log(SABT/p)$, then with probability $\geq 1 - p$, Nash-VI with Bernstein type bonus (with some absolute $c > 0$) achieves:

- $(V_1^{\dagger,\nu^{out}} - V_1^{\mu^{out},\dagger})(s_1) \leq \epsilon$, if the number of episodes $K \geq \Omega(H^3 SAB\iota/\epsilon^2 + H^3 S^2 AB\iota^2/\epsilon)$
- $Reg(K) = \sum_{k=1}^{K}(V_1^{\dagger,\nu^k} - V_1^{\mu^k,\dagger})(s_1) \leq O(\sqrt{H^2 SABT\iota} + H^3 S^2 AB\iota^2)$

Compared with

- ITLB: $\Omega(H^3 S(A+B)\iota/\epsilon^2)$
- Regret lower bound $\Omega(\sqrt{H^2 S(A+B)T})$

Nash-VI with Bernstein bonus achieves optimal dependency.

Further Research

- Reward-free learning - VI Zero
- Multiplayer General Sum MGs

  Reward-known and reward-free

References

- Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. ICML 2021.
- Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.
- Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4868–4878, 2018.
- Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. *arXiv preprint arXiv:2002.04017*, 2020.
- Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. (2021, March). Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory* (pp. 578-598). PMLR.