Foundations of Reinforcement Learning

Multi-agent RL: policy optimization

Yuejie Chi

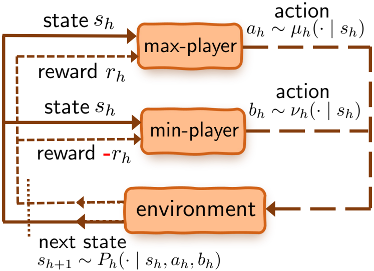
Department of Electrical and Computer Engineering

Carnegie Mellon University

Spring 2023

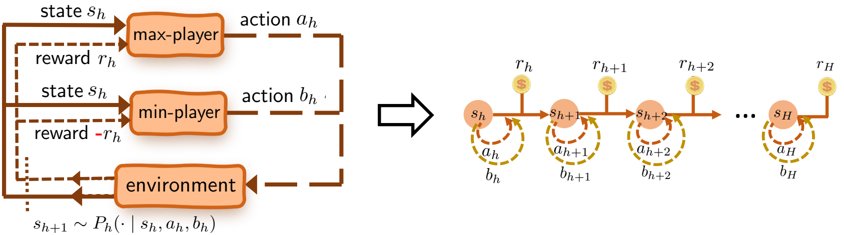
Outline

Two-player zero-sum Markov games (finite-horizon)



- $\mathcal{S}$: shared state space
- H: horizon
- $\mathcal{A} = [A]$: action space of max-player
- $\mathcal{B} = [B]$: action space of min-player
- immediate reward: max-player $r_h(s,a,b) \in [0,1]$
  min-player $-r_h(s,a,b)$
- $\mu = \{\mu_h\}$: policy of max-player; $\nu = \{\nu_h\}$: policy of min-player
- $P_h(\cdot \mid s,a,b)$: unknown transition probabilities

Value function



Value function of policy pair (μ, ν):

$$V_h^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t, b_t) \,\middle|\, s_t = s\right]$$

$$Q_h^{\mu,\nu}(s, a, b) := \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t, b_t) \,\middle|\, s_t = s, a_t = a, b_t = b\right]$$

- $\{(a_t, b_t, s_{t+1})\}$: generated when max-player and min-player execute policies μ and ν *independently (i.e. no coordination)*

Nash value iteration (finite-horizon)

Nash value iteration: for $h = H, \ldots, 1$

$$Q_h(s,a,b) \longleftarrow r_h(s,a,b) + \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a,b)} \Bigg[\underbrace{\max_{\mu(s)} \min_{\nu(s)} \mu(s')^\top Q_{h+1}(s') \nu(s')}_{\text{matrix game}} \Bigg],$$

where $Q_h(s) = [Q_h(s,\cdot,\cdot)] \in \mathbb{R}^{A \times B}$.

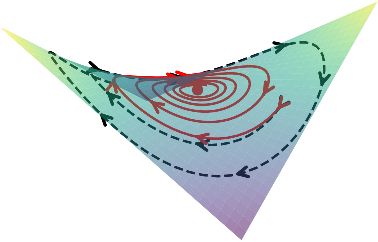
- The matrix game can be solved efficiently (see next lecture).
- Requires knowledge of the transition kernel $P_h(\cdot|s,a,b)$.

Policy optimization: saddle-point optimization

Zero-sum two-player Markov game

Given an initial state distribution $s \sim \rho$, *find policy* π *such that*

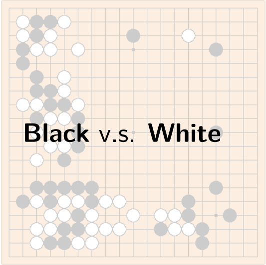
$$\max_{\mu \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu \in \Delta(\mathcal{B})^{|\mathcal{S}|}} V_1^{\mu,\nu}(\rho) := \mathbb{E}_{s \sim \rho}[V_1^{\mu,\nu}(s)]$$



Can we design a policy optimization method that guarantees fast *last-iterate* convergence?

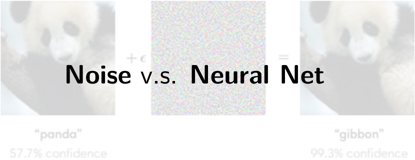
Policy optimization for two-player zero-sum matrix game

Competitive game



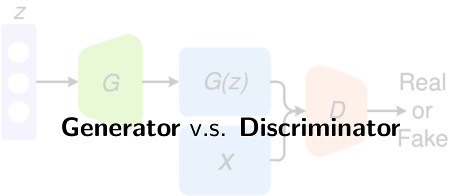
Black v.s. **White**

Go



Noise v.s. **Neural Net**

Adversarial Training

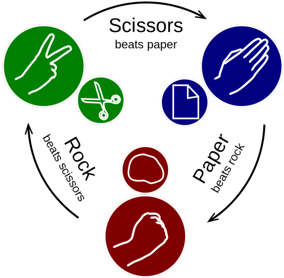


Generator v.s. **Discriminator**

Generative Adversarial Networks

Can we bring some understanding to them?

Zero-sum two-player matrix game



			
	0	-1	1
	1	0	-1
	-1	1	0

Zero-sum two-player matrix game

$$\max_{\mu \in \Delta(\mathcal{A})} \min_{\nu \in \Delta(\mathcal{B})} \mu^\top A \nu$$

- $\mathcal{A}$, $\mathcal{B}$: action space of the two players;
- $\Delta(\mathcal{A})$, $\Delta(\mathcal{B})$: set of probability distribution over $\mathcal{A}$, $\mathcal{B}$;
- $A \in^{|\mathcal{A}| \times |\mathcal{B}|}$: payoff matrix.

Nash equilibrium

John von Neumann

John Nash

Theorem 1 (Neumann's Minimax Theorem)

$$\max_{\mu \in \Delta(\mathcal{A})} \min_{\nu \in \Delta(\mathcal{B})} \mu^\top A \nu = \min_{\nu \in \Delta(\mathcal{B})} \max_{\mu \in \Delta(\mathcal{A})} \mu^\top A \nu$$

A Nash Equilibrium pair $(\mu^\star, \nu^\star)$ satisifies:

$$\mu^\top A \nu^\star \leq {\mu^\star}^\top A \nu^\star \leq {\mu^\star}^\top A \nu,$$

for all $(\mu, \nu) \in \Delta(\mathcal{A}) \times \Delta(\mathcal{B})$.

Nash equilibrium

John von Neumann

John Nash

Theorem 2 (Neumann's Minimax Theorem)

$$\max_{\mu \in \Delta(\mathcal{A})} \min_{\nu \in \Delta(\mathcal{B})} \mu^\top A \nu = \min_{\nu \in \Delta(\mathcal{B})} \max_{\mu \in \Delta(\mathcal{A})} \mu^\top A \nu$$

An ϵ-Nash Equilibrium pair $(\hat{\mu}^\star, \hat{\nu}^\star)$ satisifies:

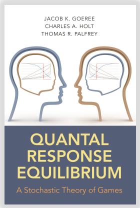
$$\mu^\top A \hat{\nu}^\star - \epsilon \leq \hat{\mu}^{\star\top} A \hat{\nu}^\star \leq \hat{\mu}^{\star\top} A \nu + \epsilon,$$

for all $(\mu, \nu) \in \Delta(\mathcal{A}) \times \Delta(\mathcal{B})$.

Entropy regularization and QRE

Quantal response equilibrium ([McKelvey and Palfrey, 1995])

$$\max_{\mu \in \Delta(\mathcal{A})} \min_{\nu \in \Delta(\mathcal{B})} \mu^\top A \nu + \tau \mathcal{H}(\mu) - \tau \mathcal{H}(\nu)$$



- Unlike NE, QRE assumes bounded rationality: action probability follows the logit function. The **unique** QRE $\zeta_\tau^\star = (\mu_\tau^\star, \nu_\tau^\star)$ satisfying

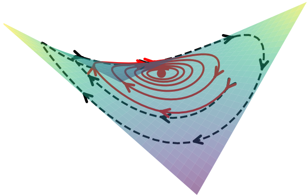
$$\begin{cases} \mu_\tau^\star(a) \propto \exp([A\nu_\tau^\star]_a / \tau), & \forall a \in \Delta(\mathcal{A}) \\ \nu_\tau^\star(b) \propto \exp(-[A^\top \nu_\tau^\star]_b / \tau), & \forall b \in \Delta(\mathcal{B}) \end{cases}$$

  are the best responses in the presence of Gumbel noises.

Translating to an ϵ-NE: setting $\tau \asymp \widetilde{O}(\epsilon)$.

Multiplicative weights update methods

$$\max_{\mu \in \Delta(\mathcal{A})} \min_{\nu \in \Delta(\mathcal{B})} f_\tau(\mu, \nu) := \mu^\top A \nu + \tau \mathcal{H}(\mu) - \tau \mathcal{H}(\nu)$$



How to avoid this?

- Multiplicative Weights Update (**MWU**):

$$\begin{cases} \mu^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta\tau} \exp\left(\eta [A\nu^{(t)}]_a\right) \\ \nu^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta\tau} \exp\left(-\eta [A\mu^{(t)}]_b\right) \end{cases}$$

- $\eta > 0$: step size;
- The trajectory may cycle/diverge!

Motivation: an implicit update method

Implicit update (IU) method

For $t = 0, 1, \cdots,$

$$\begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\big([A\nu^{(t+1)}]/\tau\big)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\big(-[A^\top\mu^{(t+1)}]/\tau\big)^{\eta\tau} \end{cases}$$

This gives

$$\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log\zeta^{(t)} - \eta\tau\log\zeta_\tau^\star, \zeta^{(t+1)} - \zeta_\tau^\star \rangle = 0,$$

which is equivalent to

$$\begin{aligned}(1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) = \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) + \eta\tau\mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \zeta_\tau^\star\big) \\ + (1-\eta\tau)\mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \zeta^{(t)}\big)\end{aligned}$$

Linear convergence of IU

For sufficiently small learning rate η, we have

$$(1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) \geq \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) + \eta\tau\mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \zeta_\tau^\star\big) + (1-\eta\tau)\mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \zeta^{(t)}\big)$$

Theorem 3 ([Cen et al., 2021])

Suppose that $0 < \eta \leq 1/\tau$, *then for all* $t \geq 0$,

$$\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) \leq (1-\eta\tau)^t \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big),$$

where $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) = \mathsf{KL}\big(\mu_\tau^\star \| \mu^{(t)}\big) + \mathsf{KL}\big(\nu_\tau^\star \| \nu^{(t)}\big)$.

Can we make this practical?

The PU method

Predictive update (PU) method

For $t = 0, 1, \cdots,$

1. *extrapolate/predict:*

$$\begin{cases} \bar{\mu}^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\nu^{(t)}]/\tau\right)^{\eta\tau} \\ \bar{\nu}^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \mu^{(t)}]/\tau\right)^{\eta\tau} \end{cases}$$

2. *update:*

$$\begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\bar{\nu}^{(t+1)}]/\tau\right)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \bar{\mu}^{(t+1)}]/\tau\right)^{\eta\tau} \end{cases}$$

The OMWU method

Optimistic multiplicative weights update (OMWU) method

For $t = 0, 1, \cdots,$

1. *extrapolate/predict:*

$$\begin{cases} \bar{\mu}^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\bar{\nu}^{(t)}]/\tau\right)^{\eta\tau} \\ \bar{\nu}^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \bar{\mu}^{(t)}]/\tau\right)^{\eta\tau} \end{cases}$$

2. *update:*

$$\begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\bar{\nu}^{(t+1)}]/\tau\right)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \bar{\mu}^{(t+1)}]/\tau\right)^{\eta\tau} \end{cases}$$

These methods belong to the class of so-called extragradient methods [Korpelevich, 1976].

Linear convergence of PU/OMWU

- Let $\zeta^{(t)} = (\mu^{(t)}, \nu^{(t)})$ and $\bar{\zeta}^{(t)} = (\bar{\mu}^{(t)}, \bar{\nu}^{(t)})$.

Theorem 4 ([Cen et al., 2021])

Suppose that the learning rates of PU and OMWU satisfy

$$\eta_{\mathsf{PU}} \le \frac{1}{\tau + 2\|A\|_\infty}, \text{ and } \eta_{\mathsf{OMWU}} \le \min\left\{\frac{1}{2\tau + 2\|A\|_\infty}, \frac{1}{4\|A\|_\infty}\right\}.$$

Both methods achieve convergence in

- *KL distance* $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) \le \epsilon$,
- *Entrywise distance of log-policies* $\|\log \zeta^{(t)} - \log \zeta_\tau^\star\|_\infty \le \epsilon$,
- *Optimality gap* $\big|f_\tau(\mu^{(t)}, \nu^{(t)}) - f_\tau(\mu_\tau^\star, \nu_\tau^\star)\big| \le \epsilon$,
- *Duality gap* $\max_{\mu' \in \Delta(\mathcal{A})} f_\tau(\mu', \nu^{(t)}) - \min_{\nu' \in \Delta(\mathcal{B})} f_\tau(\mu^{(t)}, \nu') \le \epsilon$

within $\widetilde{O}(\frac{1}{\eta\tau}\log\frac{1}{\epsilon})$ *iterations.*

Last-iterate convergence

PU allows twice as large learning rates than OMWU, at a price of requiring double gradient evaluation per iteration.

- **Entropy-regularized matrix game:** To get an ϵ-optimal solution to the regularized problem (ϵ-**QRE**), the iteration complexity is at most

$$\widetilde{O}\left(\left(1 + \frac{\|A\|_\infty}{\tau}\right)\log\frac{1}{\epsilon}\right).$$

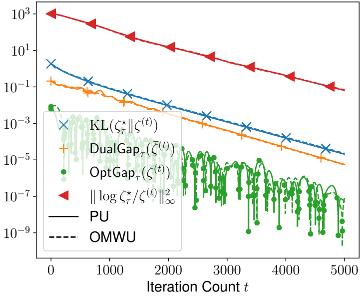
- **Unregularized matrix game:** To get an ϵ-optimal solution to the unregularized problem (ϵ-**NE**), the iteration complexity is at most

$$\widetilde{O}\left(\frac{\|A\|_\infty}{\epsilon}\right).$$

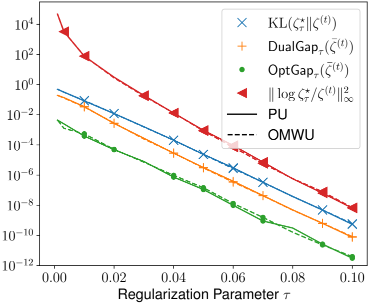
No need to assume unique Nash equilibrium!

Entropy regularization leads to linear convergence

$A \in \mathbb{R}^{100\times 100}$ with $A_{a,b} \sim U([-1,1])$ and $\eta = 0.1$



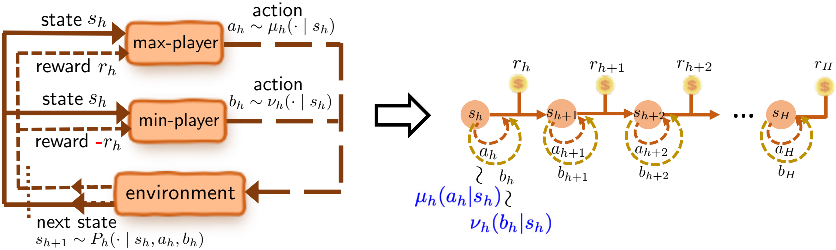
$\tau = 0.01$



#iterations = 1000

Policy optimization for two-player zero-sum Markov game

Entropy regularization in MARL



Promote the stochasticity of the policy pair using the **"soft"** value function:

$$V_\tau^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{h=1}^{H}\Big(r_t + \tau\mathcal{H}(\mu_t(\cdot|s_t) - \tau\mathcal{H}(\nu_t(\cdot|s_t)\Big) \,\Big|\, s_0 = s\right],$$

where $\mathcal{H}$ is the Shannon entropy, and $\tau \geq 0$ is the reg. parameter.

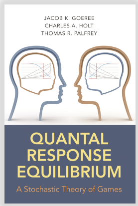
$$\max_{\mu\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu\in\Delta(\mathcal{B})^{|\mathcal{S}|}} V_\tau^{\mu,\nu}(\rho)$$

Quantal response equilibrium (QRE)

Quantal response equilibrium ([McKelvey and Palfrey, 1995])

The quantal response equilibrium (QRE) is the policy pair $(\mu_\tau^\star, \nu_\tau^\star)$ *that is the unique solution to*

$$\max_{\mu \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu \in \Delta(\mathcal{B})^{|\mathcal{S}|}} V_\tau^{\mu,\nu}(\rho).$$



Translating to an ϵ-NE: setting

$$\tau \asymp \widetilde{O}\left(\epsilon/H\right).$$

Soft value iteration via nested-loop OMWU

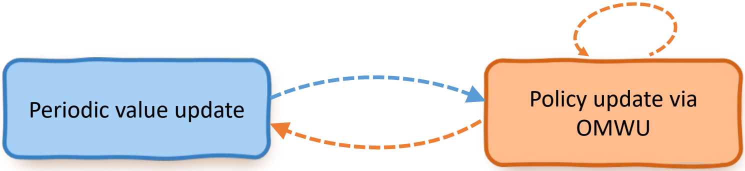
Soft value iteration: for $h = H, \dots, 1$

$$Q_h(s,a,b) \leftarrow r_h(s,a,b) + \cdot \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a,b)} \Bigg[\underbrace{\max_{\mu} \min_{\nu} \mu(s')^\top Q_{h+1}(s') \nu(s') + \tau \mathcal{H}(\mu(s')) - \tau \mathcal{H}(\nu(s'))}_{\text{Entropy-regularized matrix game}} \Bigg],$$

where $Q_h(s) = [Q_h(s,\cdot,\cdot)] \in \mathbb{R}^{A \times B}$.

Nested-loop approach:

$(\mu_h^{(t)}, \nu_h^{(t)}) \leftarrow \texttt{OMWU}(Q_h)$



$Q_h \leftarrow \texttt{SVI}(Q_{h+1})$

Convergence of the nested-loop approach

Theorem 5 ([Cen et al., 2021])

PU/OMWU with value iteration takes no more than

$$\widetilde{O}\left(\frac{H^3}{\epsilon}\right) \text{ iterations}$$

to find an ϵ-approximate NE of the unregularized MG.

- Dimension-free, last-iterate convergence.
- However, might not be easy to implement in practical online setting.

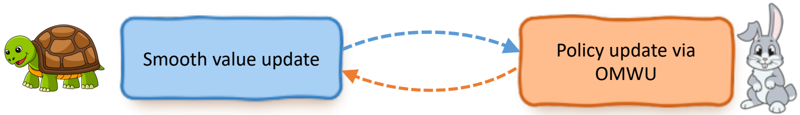
A two-timescale single-loop approach?

Soft value iteration: for $h = H, \dots, 1$

$$Q_h(s,a,b) \leftarrow r_h(s,a,b) + \cdot \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a,b)} \Bigg[\underbrace{\max_{\mu} \min_{\nu} \mu(s')^\top Q_{h+1}(s')\nu(s') + \tau \mathcal{H}(\mu(s')) - \tau \mathcal{H}(\nu(s'))}_{\text{Entropy-regularized matrix game}} \Bigg],$$

where $Q_h(s) = [Q_h(s,\cdot,\cdot)] \in \mathbb{R}^{A \times B}$.

Single-loop, two-timescale approach:



$Q^{(t+1)} \leftarrow (1-\alpha)Q^{(t)} + \alpha \cdot \texttt{lookahead}$ $\quad$ $(\mu^{(t+1)}, \nu^{(t+1)}) \leftarrow \texttt{OMWU}(Q^{(t)})$

Sublinear convergence in the episodic setting

Theorem 6 ([Cen et al., 2022])

The last-iterate of the two-timescale single-loop algorithm finds an ϵ-QRE in

$$\widetilde{O}\left(\frac{H^2}{\tau}\log\frac{1}{\epsilon}\right)$$

iterations, corresponding to $\widetilde{O}\left(\frac{H^3}{\epsilon}\right)$ iterations for finding an ϵ-NE.

- First last-iterate convergence result for the episodic setting.
- **Almost dimension-free:** independent of the size of the state-action space.

Aside: convergence in the discounted setting

Theorem 7 ([Cen et al., 2022])

For the infinite-horizon γ-discounted setting, the last-iterate of the single-loop algorithm finds an ϵ-QRE in

$$\widetilde{O}\left(\frac{S}{(1-\gamma)^4\tau}\log\frac{1}{\epsilon}\right)$$

iterations, and in $\widetilde{O}\left(\frac{S}{(1-\gamma)^5\epsilon}\right)$ iterations for finding an ϵ-NE.

- The analysis is much more involved for the discounted setting.
- Open problem to further fasten the sample complexity especially regarding the size of the state space $\mathcal{S}$.

References I

Cen, S., Chi, Y., Du, S. S., and Xiao, L. (2022).
Faster last-iterate convergence of policy optimization in zero-sum Markov games.
arXiv preprint arXiv:2210.01050.

Cen, S., Wei, Y., and Chi, Y. (2021).
Fast policy extragradient methods for competitive games with entropy regularization.
Advances in Neural Information Processing Systems, 34:27952–27964.

Korpelevich, G. M. (1976).
The extragradient method for finding saddle points and other problems.
Matecon, 12:747–756.

McKelvey, R. D. and Palfrey, T. R. (1995).
Quantal response equilibria for normal form games.
Games and economic behavior, 10(1):6–38.