Sequential Decision Making: Theory and Practice

Lecture 8: Multi-agent RL

Yuejie Chi

Yale Statistics & Data Science



Fall 2026

Announcement

- First milestone due this Sunday.
- Discuss what you have done, what remaining, disclosure of AI usage.
- Please don't hesitate to schedule meetings with me if you have any challenges questions, or just want to discuss!

Outline

Background: a primer on matrix and Markov games

Statistical perspective: sample complexity

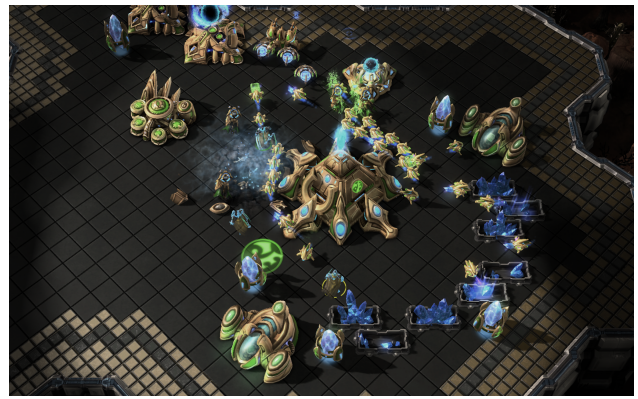
Optimization perspective: last-iterate convergence

Multi-agent reinforcement learning (MARL)





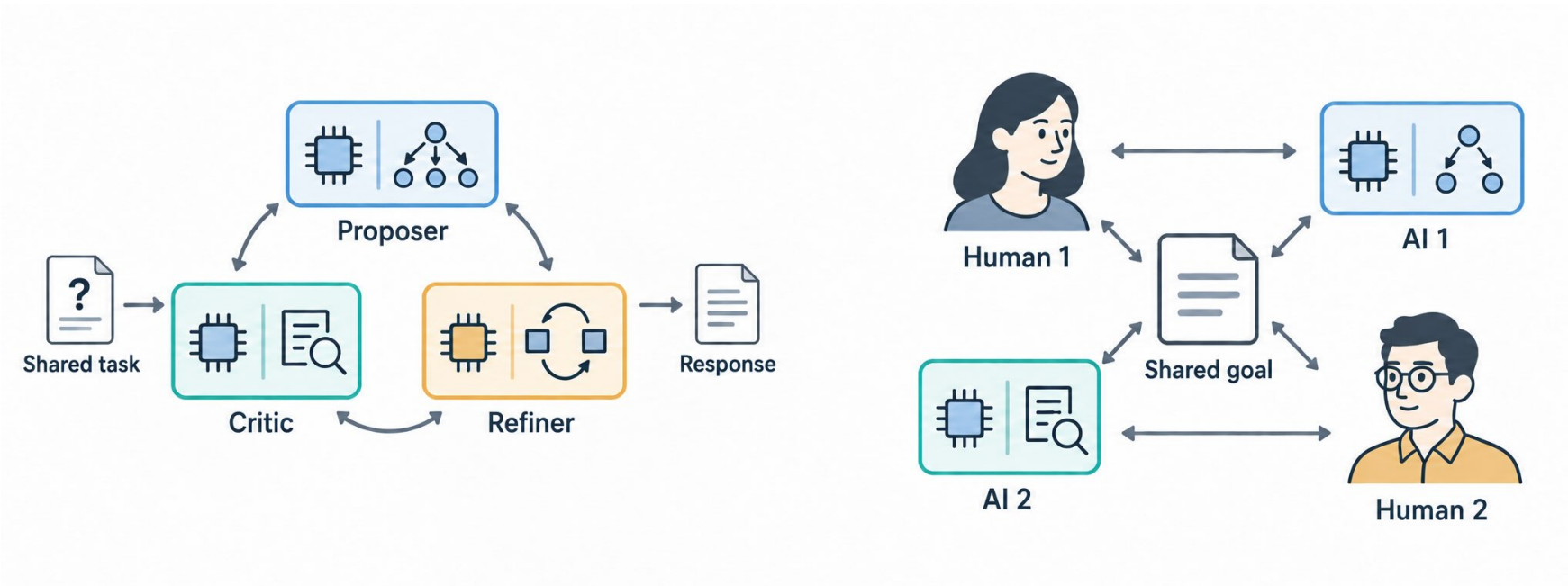




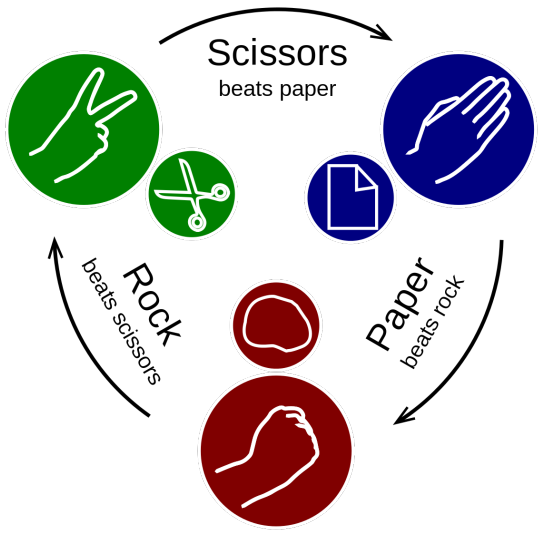
To collaborate or to compete, that is the question.

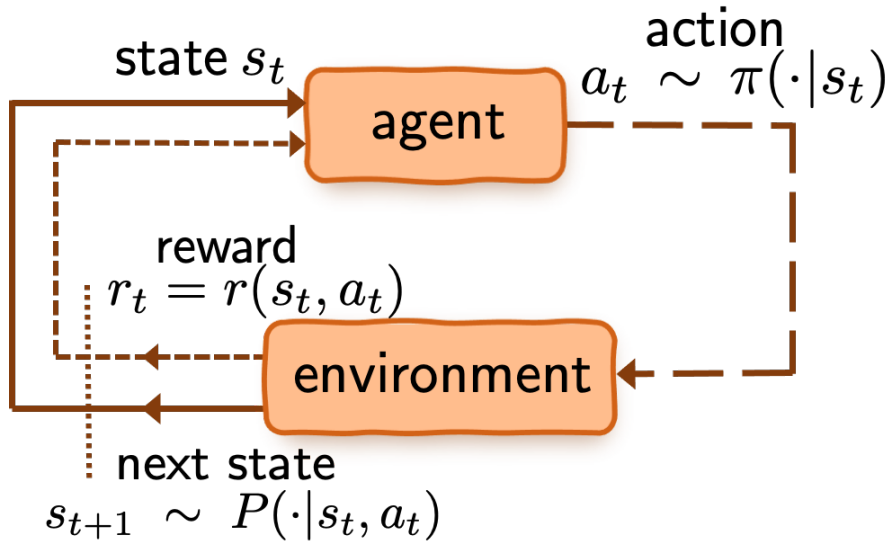
Why multi-agent?

- To decompose a large problem into smaller pieces
- Decentralized decision making across distributed agents
- Multiple agents in a strategic environment

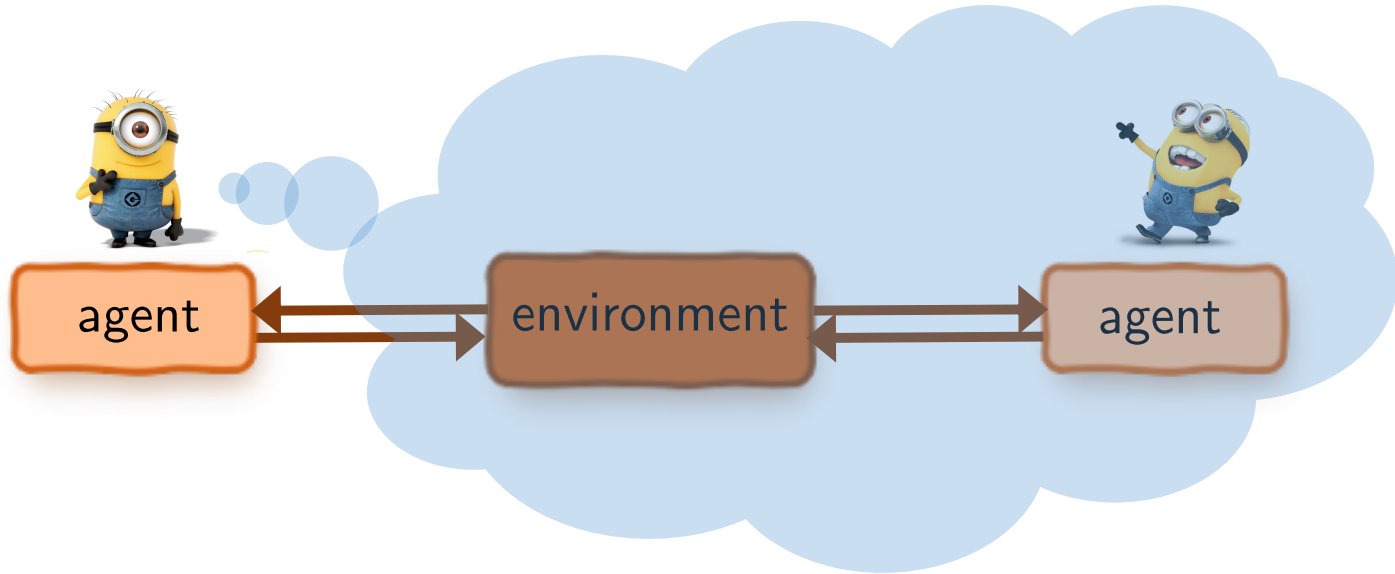


MARL = Game theory + RL



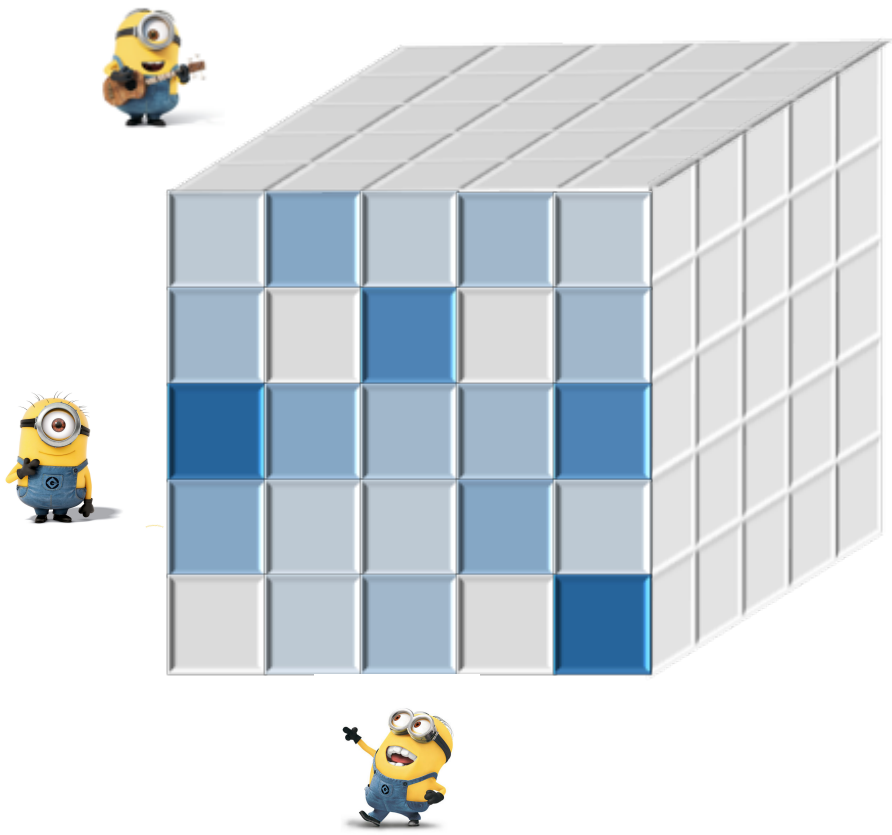


Challenges in MARL: nonstationarity



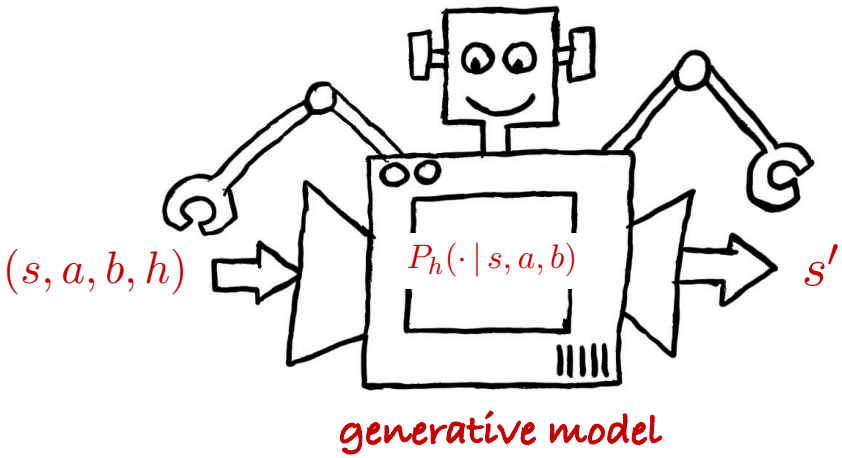
From a single-agent perspective:
the environment is **time-varying** and **nonstationary**!

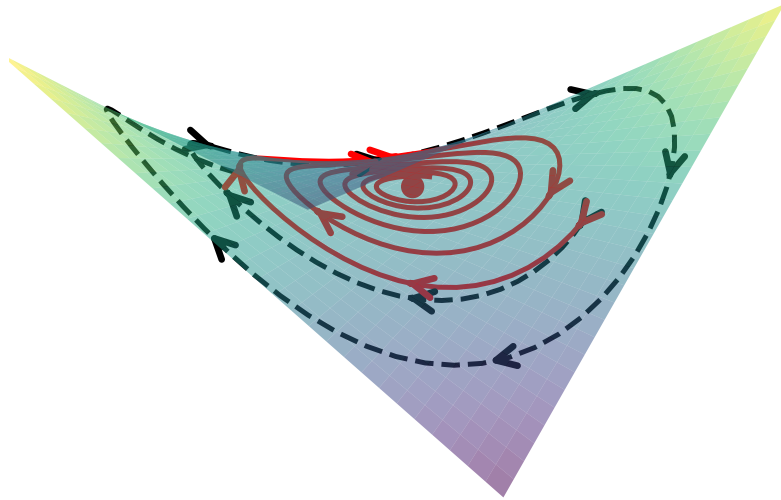
Challenges in MARL: curse of multiple agents



The explosion of choices:
The joint action space grows **exponentially** with the agents!

Statistical versus optimization perspectives





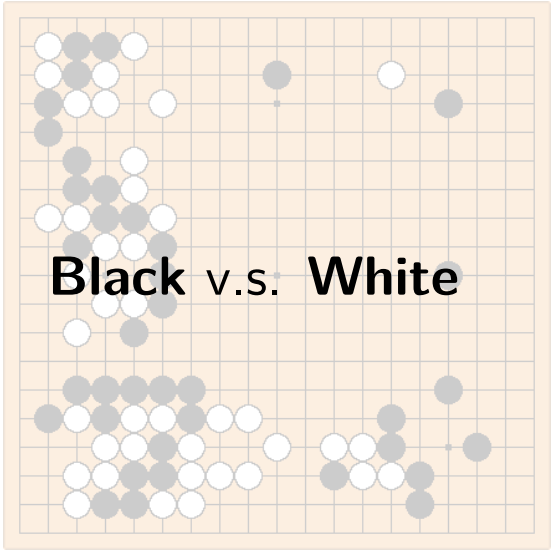
Statistical perspective
sample efficiency

Optimization perspective
computational efficiency

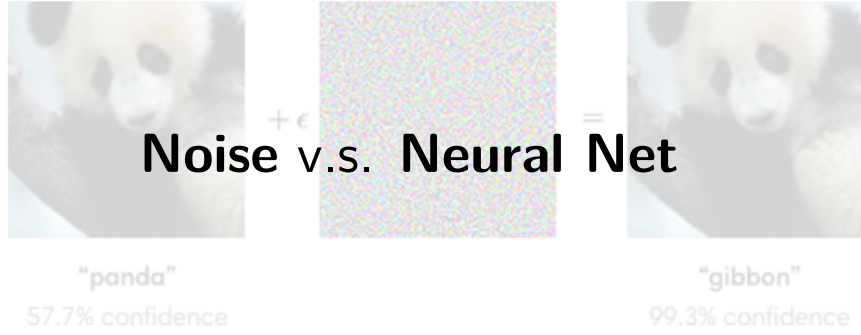
This lecture addresses both perspectives for two-player zero-sum games.

Background:
a primer on matrix and Markov games

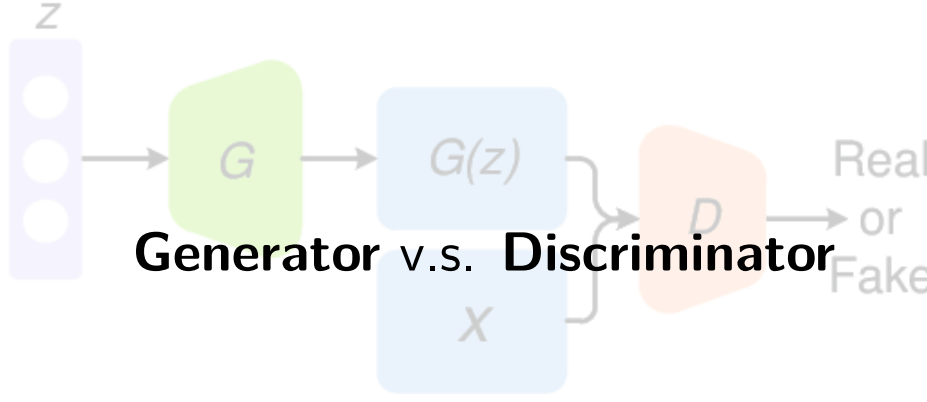
Competitive game



Go



Adversarial Training



Generative Adversarial Networks

Zero-sum two-player matrix game

When the max player takes action i, and the min player takes action j, the payoff received by both players are

$$r_{\max}(i,j) = M_{i,j}, \qquad r_{\min}(i,j) = -M_{i,j},$$

which satisfies

$$\forall\,(i,j): \qquad r_{\max}(i,j) + r_{\min}(i,j) = 0.$$

Zero-sum two-player matrix game

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

- $\mathcal{A}$, $\mathcal{B}$: action space of the two players;
- μ, ν: strategy of the two players; can be pure/mixed.
- $\Delta(\mathcal{A})$, $\Delta(\mathcal{B})$: set of probability distribution over $\mathcal{A}$, $\mathcal{B}$;
- $M \in^{|\mathcal{A}|\times|\mathcal{B}|}$: payoff matrix.

Best response

Against a fixed opponent's strategy, a best response maximizes a player's own expected payoff.

Max player: given $\nu \in \Delta(\mathcal{B})$, the set of best responses is

$$\mathrm{BR}_{\max}(\nu) = \arg\max_{\mu \in \Delta(\mathcal{A})} \mu^\top M \nu.$$

Min player: given $\mu \in \Delta(\mathcal{A})$, the set of best responses is

$$\mathrm{BR}_{\min}(\mu) = \arg\min_{\nu \in \Delta(\mathcal{B})} \mu^\top M \nu.$$

Best responses need not be unique: any mixture supported on best-response actions is also a best response.

Solution concept: Nash equilibrium



John von Neumann



John Nash

Theorem 1 (Neumann's Minimax Theorem)

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

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

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

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

- At a Nash equilibrium, both players are best responding to each other.

Nash equilibrium



John von Neumann



John Nash

Theorem 2 (Neumann's Minimax Theorem)

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

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

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

for all $(\mu, \nu) \in \Delta(\mathcal{A}) \times \Delta(\mathcal{B})$. The Nash gap is defined as

$$\text{NE-gap}(\hat{\mu}^\star, \hat{\nu}^\star) = \max_{\mu} \mu^\top M \hat{\nu}^\star - \min_{\nu} \hat{\mu}^{\star\top} M \nu$$

Solution concept: regret

Repeat the game for T rounds with the sequence of actions

$$(a_t, b_t).$$

The max-player's payoff is $r_t = M_{a_t b_t}$ and the min-player's payoff is $-r_t$.

External regret: compare with the best fixed action in hindsight, against the opponent's recorded actions.

$$R_T^{\max} = \max_{a \in \mathcal{A}} \sum_t M_{a b_t} - \sum_t r_t,$$

$$R_T^{\min} = \sum_t r_t - \min_{b \in \mathcal{B}} \sum_t M_{a_t b}.$$

No-regret play:

$$\frac{R_T^{\max}}{T} \to 0, \qquad \frac{R_T^{\min}}{T} \to 0, \qquad T \to \infty.$$

Implication: No player can obtain a positive long-run average improvement by switching to one fixed action.

How does no regret relate to equilibrium?

Let $\bar{\mu}_T$ and $\bar{\nu}_T$ be the empirical action frequencies:

$$\bar{\mu}_T(a) = \frac{1}{T}\sum_t \mathbb{I}\{a_t = a\}, \qquad \bar{\nu}_T(b) = \frac{1}{T}\sum_t \mathbb{I}\{b_t = b\}.$$

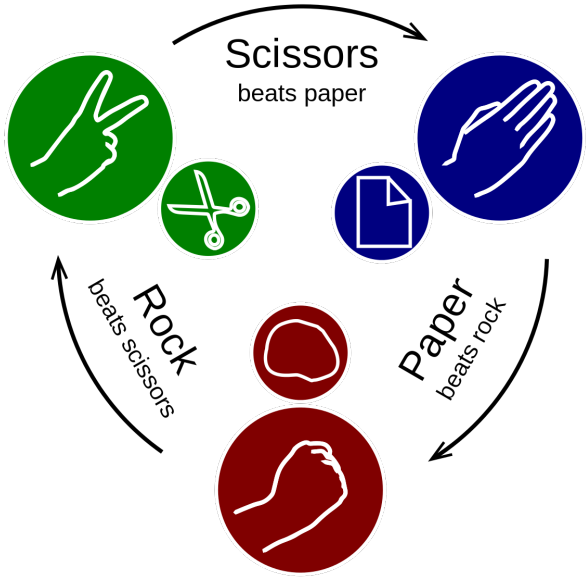
Two-player zero-sum games: the actual payoffs cancel, giving

$$\underbrace{\max_{\mu} \mu^\top M \bar{\nu}_T - \min_{\nu} \bar{\mu}_T^\top M \nu}_{\text{Nash gap}} = \frac{R_T^{\max} + R_T^{\min}}{T}.$$

If both regrets are at most $\epsilon T/2$, the averaged pair $(\bar{\mu}_T, \bar{\nu}_T)$ is an ϵ-NE.

Beyond two-player zero-sum games: No-regret learning need not yield an NE in general-sum games.

Example: rock-paper-scissors

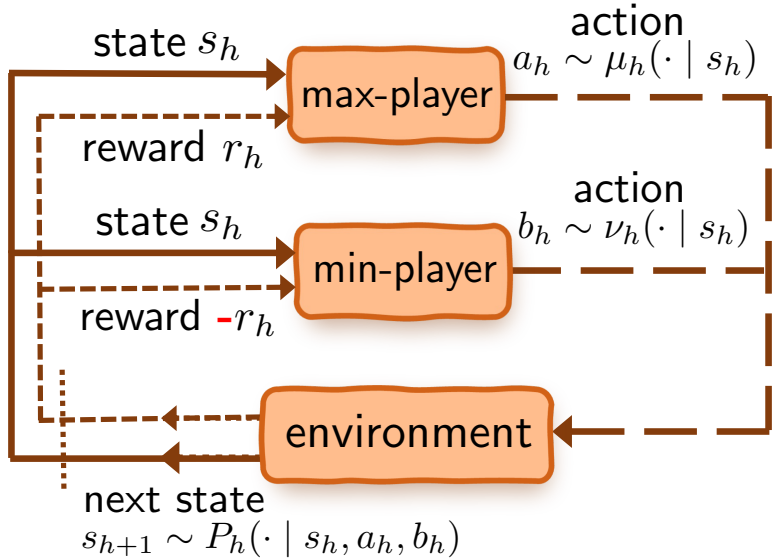


	0	-1	1
	1	0	-1
	-1	1	0

The NE policy for both players is

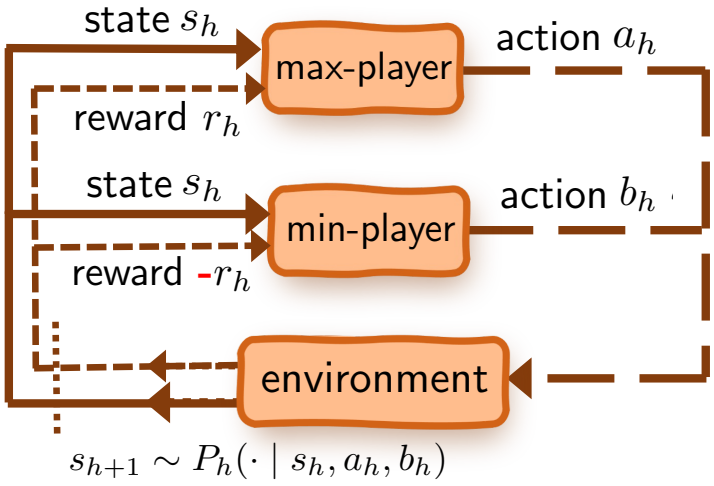
$$\mu^\star = \nu^\star = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right).$$

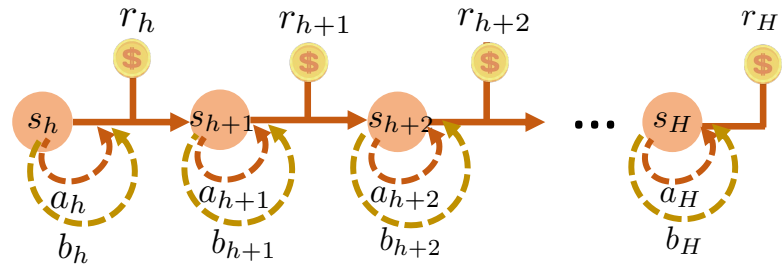
Two-player zero-sum Markov games (finite-horizon)



- $\mathcal{S}$: 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):

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

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

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

Nash equilibrium (NE)



John von Neumann



John Nash

An NE policy pair $(\mu^\star, \nu^\star)$ obeys

$$\max_{\mu} V^{\mu,\nu^\star} = V^{\mu^\star,\nu^\star} = \min_{\nu} V^{\mu^\star,\nu}$$

- no unilateral deviation is beneficial
- no coordination between two agents (they act *independently*)

Nash value iteration (finite-horizon)

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

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

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

- The matrix game can be solved efficiently (we'll return to this later in this lecture).
- Requires knowledge of the transition kernel $P_h(\cdot|s,a,b)$.

Aside: infinite-horizon discounted setting

Value function of policy pair (μ, ν):

$$V^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t, b_t) \,\middle|\, s_0 = s\right]$$

$$Q^{\mu,\nu}(s,a,b) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) \,\middle|\, s_0 = s, a_0 = a, b_0 = b\right]$$

where $\gamma \in [0,1)$ is the discount factor.

Nash value iteration:

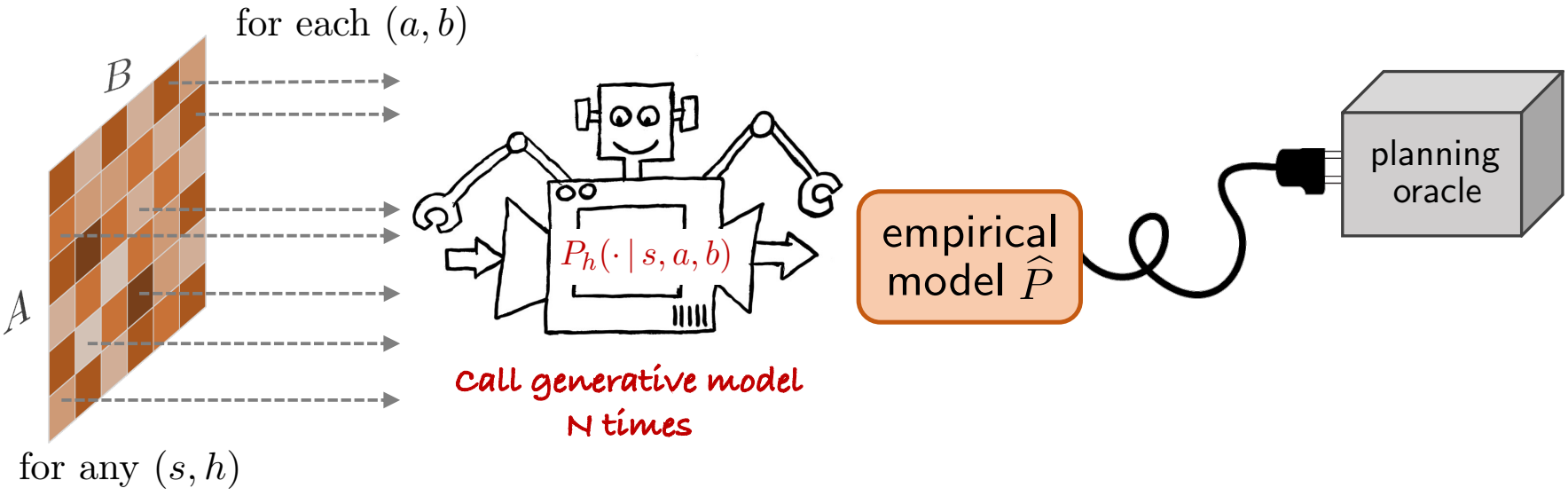
$$Q(s,a,b) \longleftarrow r_h(s,a,b) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a,b)} \left[\underbrace{\max_{\mu(s)} \min_{\nu(s)} \mu(s')^\top Q(s')\nu(s')}_{\text{matrix game}}\right],$$

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

Statistical perspective: sample complexity

A teaser: standard model-based approach

— [Zhang et al., 2020]



1. for each (s, a, b, h), call generative models N times, i.e., sample

$$s'_{(i)} \overset{\text{ind.}}{\sim} P_h(s' \mid s, a, b), \qquad i = 1, \ldots, N$$

2. build empirical model $\widehat{P}$, and run "plug-in" methods (e.g., Nash VI)

Question: *how many samples are sufficient to learn an*

$$\underbrace{\varepsilon\text{-Nash policy pair}}_{\max_\mu V^{\mu,\hat{\nu}} - \varepsilon \le V^{\hat{\mu},\hat{\nu}} \le \min_\nu V^{\hat{\mu},\nu} + \varepsilon} \quad ?$$

- [Zhang et al., 2020] establishes a sample complexity of

$$\frac{H^4 S A B}{\varepsilon^2},$$

  which suffers the **curse of multiagents**!

- Namely, the sample complexity scales with the size of the joint action space AB, rather than $\max\{A, B\}$ or $A + B$.

Can we overcome the curse of multi-agents?

Warm-up: matrix game with bandit feedback

Let $M \in [0,1]^{A \times B}$ be the payoff of the matrix game, that is *unknown*.

Online matrix game learning: in each round,

- Both players play actions according to their current policy

$$a_t \sim \mu_t, \quad b_t \sim \nu_t$$

  and observe the payoff $r_t = M_{a_t b_t}$. The max-player receives r_t; the min-player can use reward $1 - r_t$.

- Update the policies of both players.
  - **important distinction:** centralized versus independent.

Can the players learn the Nash equilibrium efficiently?

First attempt: matrix UCB

— [O'Donoghue et al., 2021]

Observation: view this as a bandit problem over the joint action space

Step 1: build an optimistic estimate of each entry:

$$U_t(a,b) = \underbrace{\overline{M}_t(a,b)}_{\text{empirical mean}} + \underbrace{\sqrt{\frac{2\log(2ABT^2)}{1 \vee N_t(a,b)}}}_{\text{uncertainty bonus}},$$

where $N_t(a,b)$ counts previous plays of (a,b).

Step 2: compute a Nash pair for the optimistic matrix game:

$$(\mu_t, \nu_t) \in \arg \max_{\mu \in \Delta(\mathcal{A})} \min_{\nu \in \Delta(\mathcal{B})} \mu^\top U_t \nu$$

Draw $a_t \sim \mu_t$ and $b_t \sim \nu_t$ independently; observe r_t and update (a_t, b_t).

Theory of matrix UCB

The game value $v^\star = \max_\mu \min_\nu \mu^\top M \nu$ is the payoff that a Nash strategy guarantees against any opponent.

Theorem 3 ([O'Donoghue et al., 2021])

Matrix UCB has expected regret from the game value

$$\mathbb{E}\left[\sum_{t=1}^{T}(v^\star - r_t)\right] \leq \widetilde{O}\left(\sqrt{ABT}\right).$$

- estimates AB entries
- require centralized coordinate and knowing the opponent's action

Can we instead run one ordinary UCB algorithm per player?

Second attempt: independent UCB

Natural idea: each player treats her own actions as arms, and modeling the projected environment by the other player's strategy.

For the max-player,

$$a_t \in \arg\max_a \left\{ \overline{g}_t(a) + \sqrt{\frac{2\log t}{N_t(a)}} \right\}.$$

Here $\overline{g}_t(a)$ averages the rewards observed when playing a, where the expected reward of arm a in round t is

$$g_t(a) = \sum_b \nu_t(b) M_{ab} = (M\nu_t)_a.$$

- **Reward nonstationarity:** M_{ab} is fixed, but the opponent's policy ν_t changes: different arms may be sampled against different ν_t
- Doesn't converge to NE in general.

Fix: an adversarial bandit perspective

Each player faces an **adversarial bandit**:

- an arm's reward can change arbitrarily over time
- only the selected arm's reward is observed
- simultaneous independent draws prevent the opponent from reacting to our current sampled action

Idea: each player runs her own EXP3-type algorithm [Auer et al., 2002]!

Independent EXP3-IX

The opponent's sampled action defines each player's loss vector:

$$\ell_t^{\max}(a) = 1 - M_{a,b_t}, \qquad \ell_t^{\min}(b) = M_{a_t,b}.$$

Initialize p_1 uniformly. Draw $I_t \sim p_t$ and observe $\ell_t(I_t)$: $1 - r_t$ for the max-player, or r_t for the min-player.

Step 1: estimate the loss vector by importance weighting:

$$\widehat{\ell}_t(i) = \frac{\ell_t(I_t)\,\mathbb{I}\{I_t = i\}}{p_t(i) + \gamma}.$$

The extra $\gamma > 0$ limits large estimates when an action is rarely sampled. This is **EXP3-IX** (EXP3 with a bit exploration) [Neu, 2015].

Step 2: update the distribution by exponential weights:

$$p_{t+1}(i) \propto p_t(i) \exp\left(-\eta \widehat{\ell}_t(i)\right).$$

Choose $\eta = 2\gamma = \sqrt{2 \log d/(dT)}$, with $d = A$ or B.

Theory of independent EXP3-IX

Theorem 4 (Corollary of [Neu, 2015], Theorem 1)

With simultaneous independent EXP3-IX learners, with high probability,

$$R_T^{\max} \le \widetilde{O}(\sqrt{AT}), \qquad R_T^{\min} \le \widetilde{O}(\sqrt{BT}).$$

Apply the regret-to-equilibrium identity:

$$\mathrm{Gap}(\bar{\mu}_T, \bar{\nu}_T) = \frac{R_T^{\max} + R_T^{\min}}{T} \le \widetilde{O}\left(\sqrt{\frac{A+B}{T}}\right).$$

Thus an ϵ-NE can be learned in

$$T = \widetilde{O}\left(\frac{A+B}{\epsilon^2}\right) \quad \text{rounds of play.}$$

- each learner uses only its own action and reward
- return empirical action frequencies; current policies may still cycle

V-learning: generalizing to MDPs

Recall: in UCB-VI, we maintain an optimistic estimation of the Q-function, and derive a greedy policy from the Q-function.

V-learning: maintains a separate policy using bandit feedback based on the value function estimate.

At each (s, h), maintain a value $V_h(s)$ and a policy $\pi_h(\cdot \mid s)$. On its tth visit:

1. sample $a_t \sim \pi_h(\cdot \mid s)$; observe r_t and s'_t
2. evaluate the sampled action using the continuation value:

$$y_t = \underbrace{r_t}_{\text{instantaneous reward}} + \underbrace{V_{h+1}(s'_t)}_{\text{next-step value}}$$

3. use y_t to update the **value**, and feed $(a_t, (H - y_t)/H)$ to the bandit learner to update the **policy**

The bandit feedback changes as V_{h+1} is learned.

V-learning: updating the value

— [Jin et al., 2021]

Initialization: $\widetilde{V}_h(s) = V_h(s) = H - h + 1$, $\pi_h(a \mid s) = 1/A$, and $V_{H+1} = 0$.

On the tth visit to (s, h), use the sampled return

$$y_t = r_t + V_{h+1}(s'_t).$$

Optimistic (monotonic) value update:

$$\widetilde{V}_h(s) \leftarrow (1 - \alpha_t)\widetilde{V}_h(s) + \alpha_t \big[y_t + \underbrace{\beta_t}_{\text{bonus}}\big],$$

$$V_h(s) \leftarrow \min\left\{H - h + 1\,,\, \widetilde{V}_h(s)\,,\, V_h(s)\right\},$$

where

$$\alpha_t = \frac{H+1}{H+t}, \qquad \beta_t \asymp \sqrt{\frac{H^3 A}{t}}.$$

V-learning: updating the policy

— [Jin et al., 2021]

Fix (s, h). Let π_t be its policy on visit t and initialize $L_0(a) = 0$.

Step 1: estimate the loss from the sampled return:

$$\widetilde{\ell}_t = \frac{H - y_t}{H}, \qquad \widehat{\ell}_t(a) = \frac{\widetilde{\ell}_t \, \mathbb{I}\{a_t = a\}}{\pi_t(a) + \gamma_t}.$$

Step 2: accumulate the estimated losses with the same α_t as the value:

$$L_t(a) = (1 - \alpha_t) L_{t-1}(a) + \alpha_t \widehat{\ell}_t(a).$$

Step 3: update the policy by exponential weights:

$$\pi_{t+1}(a) \propto \exp\left(-\frac{\eta_t}{\alpha_t} L_t(a)\right), \qquad \eta_t = \gamma_t = \sqrt{\frac{H \log A}{At}}.$$

V-learning in Markov games

Two-player zero-sum game: each player runs independently weighted adversarial bandit learning described above.

Max-player: A actions

$$a_t \sim \mu_h(\cdot \mid s),$$

$$y_t^{\max} = r_t + V_{h+1}^{\max}(s_t').$$

Min-player: B actions

$$b_t \sim \nu_h(\cdot \mid s),$$

$$y_t^{\min} = 1 - r_t + V_{h+1}^{\min}(s_t').$$

- each player updates its own value and policy at (s, h)
- each player only observes the state, its own action and reward, and the next state, not the action of the opponent
- no joint-action value table $Q_h(s, a, b)$ is needed

Theory of V-learning for Markov games

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

For a two-player zero-sum Markov game, monotonic V-learning returns Markov policies $(\widehat{\mu}, \widehat{\nu})$ (defined suitably via mixing past policies) satisfying, with probability at least $1-\delta$,

$$\max_{\mu} V_1^{\mu,\widehat{\nu}}(s_1) - \min_{\nu} V_1^{\widehat{\mu},\nu}(s_1) \leq \widetilde{O}\left(\sqrt{\frac{H^5 S \max\{A,B\}}{K}}\right)$$

after K episodes from initial state s_1.

To learn an ε-NE, it suffices to use

$$\widetilde{O}\left(\frac{H^5 S \max\{A,B\}}{\varepsilon^2}\right) \quad \text{episodes.}$$

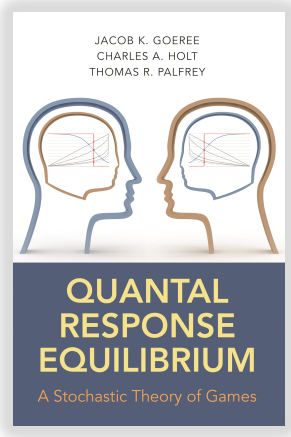
- uses trajectories; no generative model is required
- depends on the larger individual action set, rather than AB

Policy optimization for two-player zero-sum matrix game

Entropy regularization and QRE

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

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



- 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

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

  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

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

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

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

- $\eta > 0$: step size;

Linear convergence of MWU

Theorem 6 ([Sokota et al., 2023])

Let $\zeta^{(t)} = (\mu^{(t)}, \nu^{(t)})$ *and* $\tau > 0$. *Suppose that the initial policies have full support and*

$$0 < \eta \leq \frac{\tau}{\|M\|_\infty^2 + \tau^2}.$$

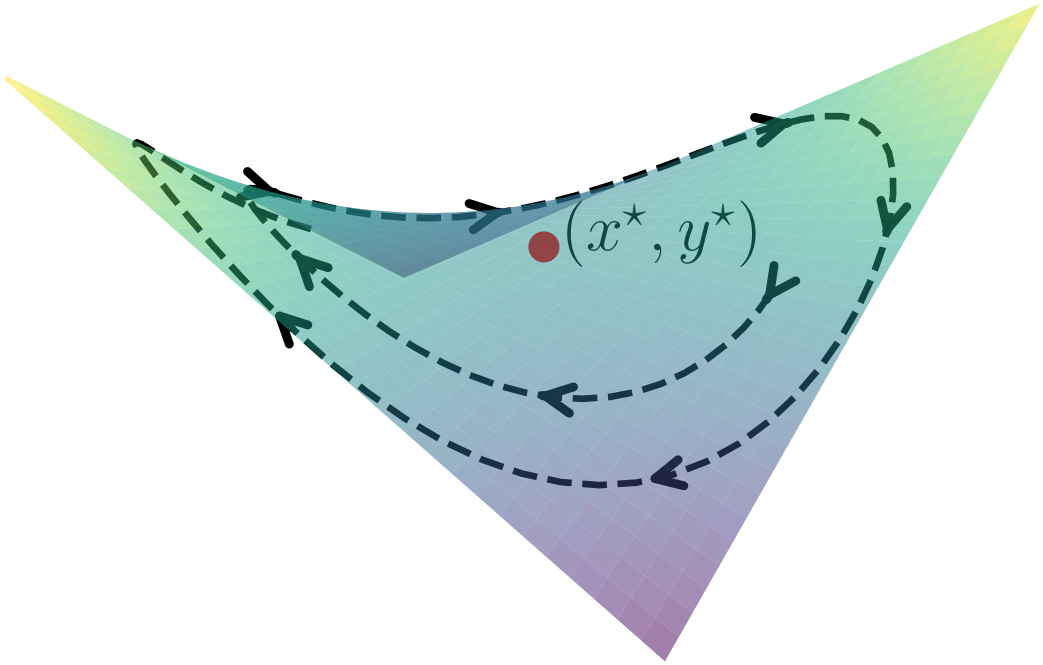
Then, for all $t \geq 0$, *regularized MWU satisfies*

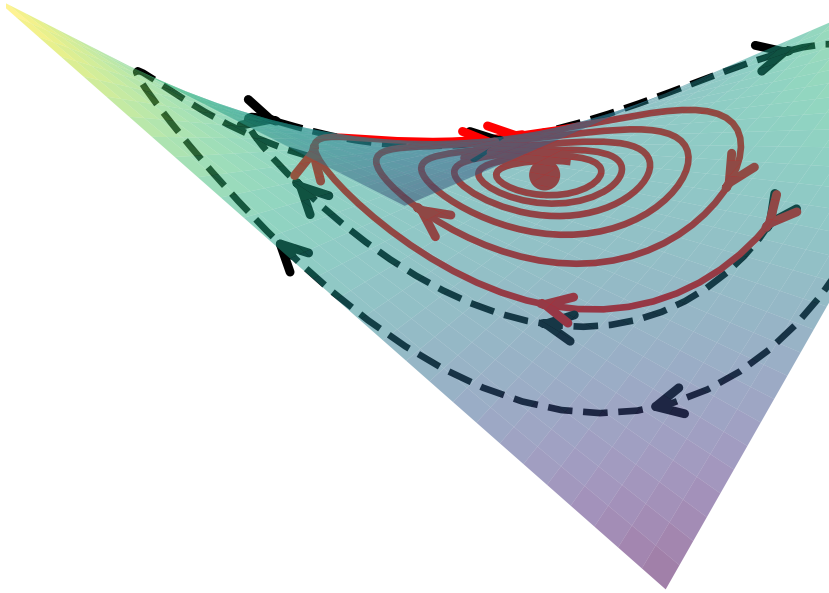
$$\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) \leq (1-\eta\tau)^t \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big),$$

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

- Last-iterate convergence to QRE with a sufficiently small step size.
- Choosing $\eta = \tau/(\|M\|_\infty^2 + \tau^2)$ gives $\widetilde{O}(1 + \|M\|_\infty^2/\tau^2)$ iterations to reach ϵ-accuracy.

Oscillation with larger learning rates





How do we fix the oscillation?

Motivation: an implicit update method

Implicit update (IU) method

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

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

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

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

Linear convergence of IU

For sufficiently small learning rate η, we have

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

Theorem 7 ([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:*

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

2. *update:*

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

The OMWU method

Optimistic multiplicative weights update (OMWU) method

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

1. *extrapolate/predict:*

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

2. *update:*

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

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 8 ([Cen et al., 2021])

Suppose that the learning rates of PU and OMWU satisfy

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

Both methods achieve convergence in

- *KL distance* $\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right) \le \epsilon$,
- *Entrywise distance of log-policies* $\|\log \zeta^{(t)} - \log \zeta_\tau^\star\|_\infty \le \epsilon$,
- *Optimality gap* $\left|f_\tau(\mu^{(t)}, \nu^{(t)}) - f_\tau(\mu_\tau^\star, \nu_\tau^\star)\right| \le \epsilon$,
- *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

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

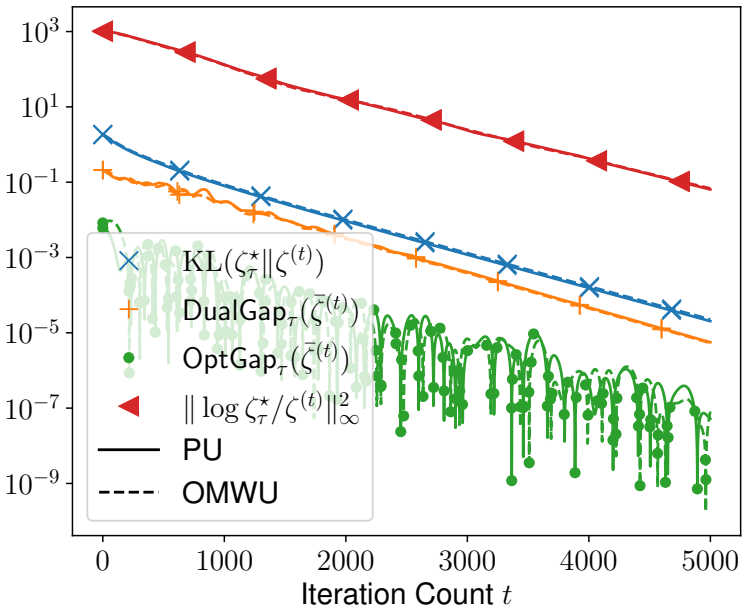
- **Unregularized matrix game:** To get an ϵ-optimal solution to the unregularized problem (ϵ-**NE**), the iteration complexity is at most

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

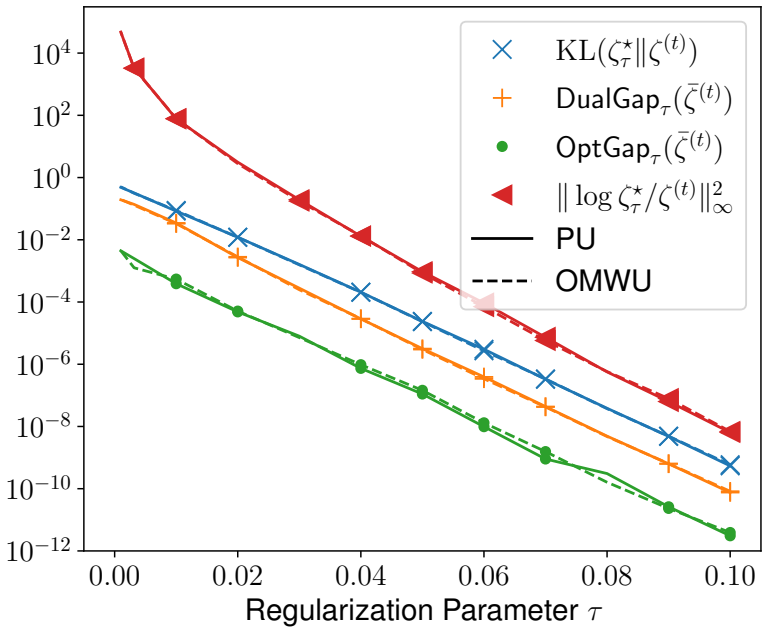
No need to assume unique Nash equilibrium!

Entropy regularization leads to linear convergence

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



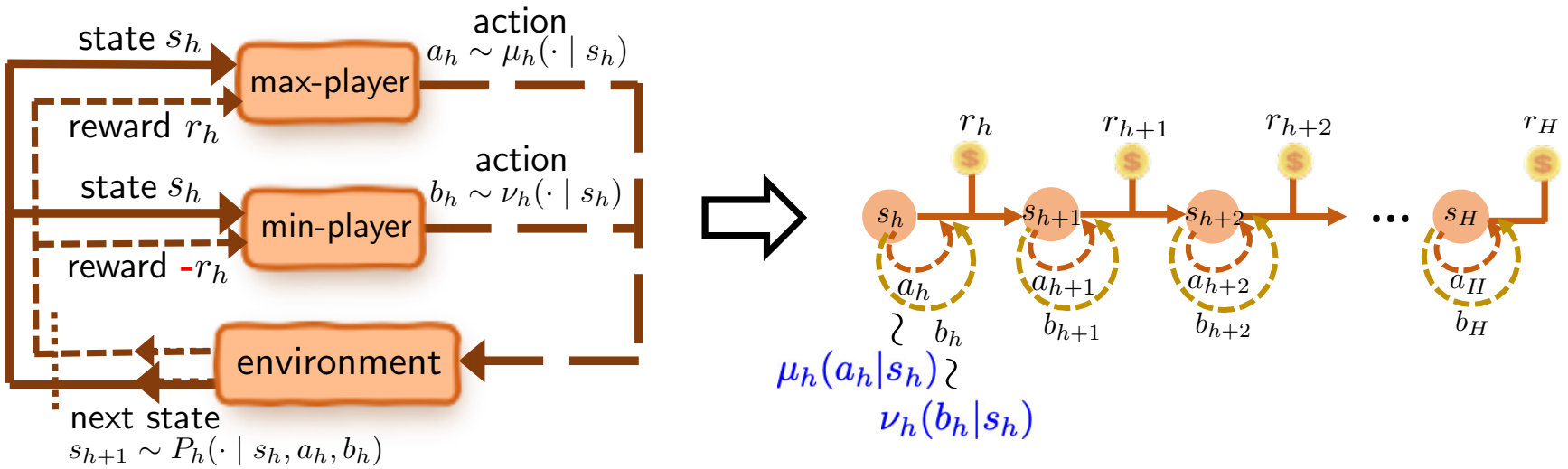
$\tau = 0.01$



#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.

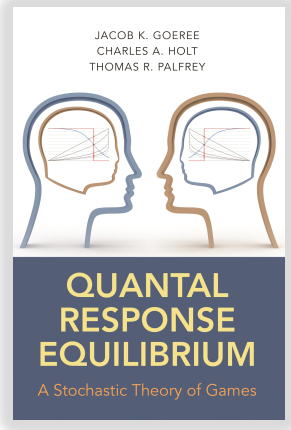
$$\max_{\mu\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu\in\Delta(\mathcal{B})^{|\mathcal{S}|}} V_\tau^{\mu,\nu}(\rho)$$

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

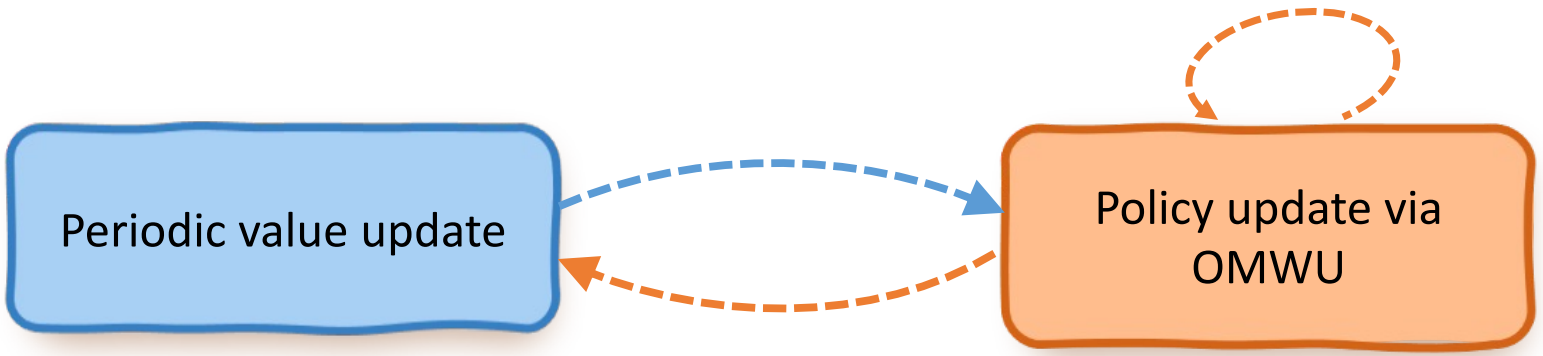
Soft value iteration: for $h = H, \dots, 1$

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

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 9 ([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 $h = H, \ldots, 1$

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

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

Single-loop, two-timescale approach:



Smooth value update

Policy update via OMWU



$Q^{(t+1)} \leftarrow (1-\alpha)Q^{(t)} + \alpha \cdot \texttt{lookahead}$

$(\mu^{(t+1)}, \nu^{(t+1)}) \leftarrow \texttt{OMWU}(Q^{(t)})$

Sublinear convergence in the episodic setting

Theorem 10 ([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 11 ([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


Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002).
The nonstochastic multiarmed bandit problem.
SIAM Journal on Computing, 32(1):48–77.


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.


Jin, C., Liu, Q., Wang, Y., and Yu, T. (2021).
V-learning–a simple, efficient, decentralized algorithm for multiagent RL.
arXiv preprint arXiv:2110.14555.


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.

References II



Neu, G. (2015).

Explore no more: Improved high-probability regret bounds for non-stochastic bandits.
In *Advances in Neural Information Processing Systems*, volume 28.



O'Donoghue, B., Lattimore, T., and Osband, I. (2021).

Matrix games with bandit feedback.
In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 279–289. PMLR.



Sokota, S., D'Orazio, R., Kolter, J. Z., Loizou, N., Lanctot, M., Mitliagkas, I., Brown, N., and Kroer, C. (2023).

A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games.
In *International Conference on Learning Representations*.



Zhang, K., Kakade, S., Basar, T., and Yang, L. (2020).

Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity.
Advances in Neural Information Processing Systems, 33:1166–1178.