Pessimistic Minimax Value Iteration: Provably Efficient Equilibrium Learning from Offline Datasets

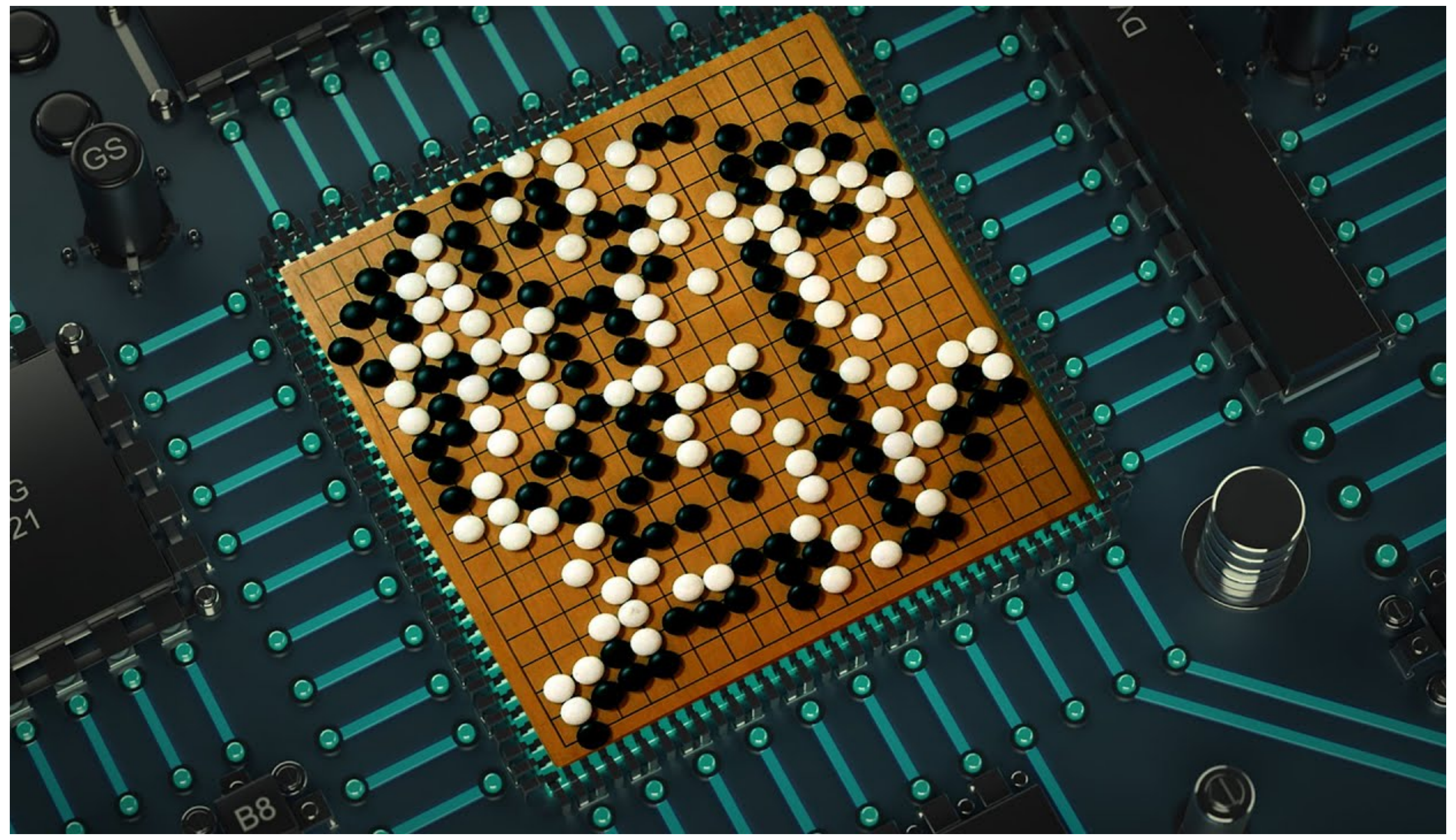
Han Zhong
Peking University

Joint work with Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, Zhuoran Yang

Outline

- Motivation and background
- Formulation and objectives
- Learning two-player zero-sum Markov games with offline datasets
- Conclusion and future directions

Success of RL



Go

Poker

Dota

Multi-agent + Decision-Making

Challenges of RL

Sample Efficiency

$+$

Computational Efficiency

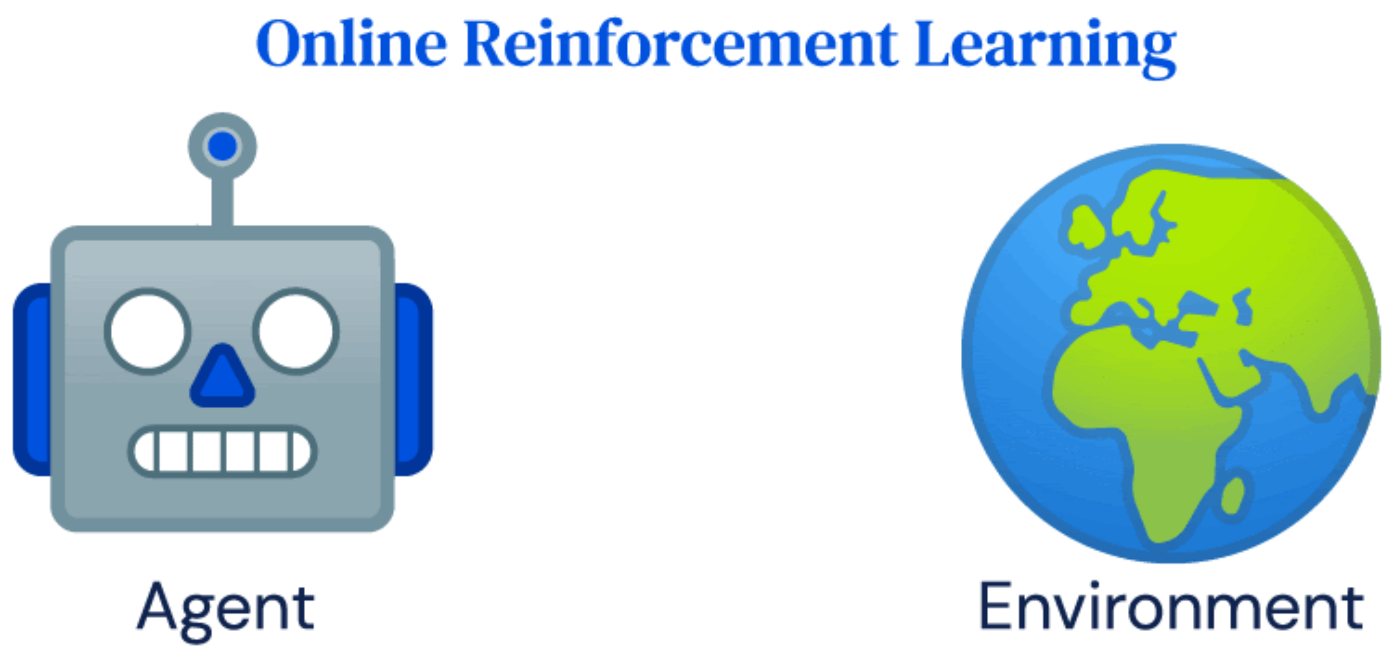
AlphaGo Zero: trained on 3×10^7 games, and took 40 days

Goal: design computationally efficient and sample-efficient learning algorithms

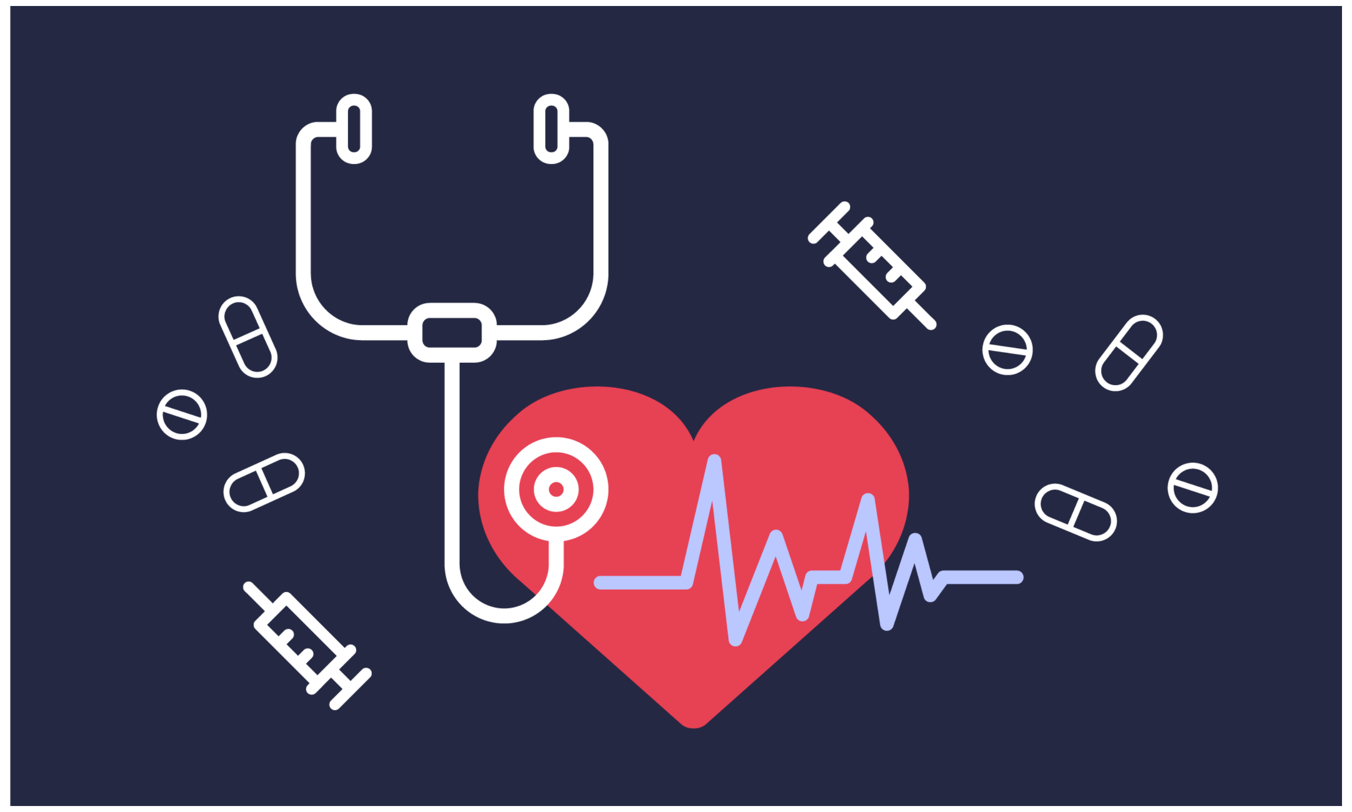
Online RL v.s. Offline RL

Online RL:
Learn from interactions
Exploration v.s. exploitation

Offline RL:
Learn from datasets
Data distribution shift



Online RL v.s. Offline RL



Healthcare

Auto-Driving

In these scenarios, either collecting data is costly and risky, or online exploration is impossible

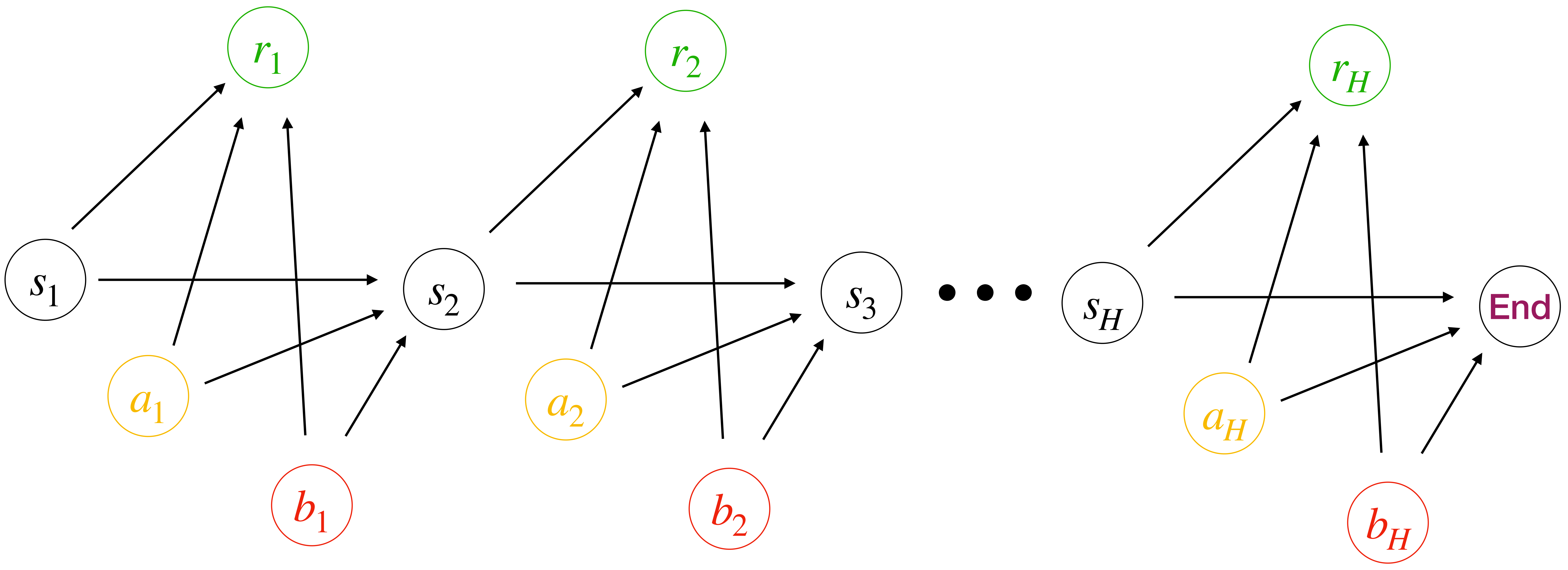
Offline Multi-Agent RL (MARL)

Q1: Can we design sample-efficient equilibrium learning algorithms in offline MARL?

Q2: What is the necessary and sufficient condition for achieving sample efficiency in offline MARL?

Formulation and Objective: Offline Two-player Zero-sum Markov Game

Two player zero-sum Markov Game (MG) $(\mathcal{S}, \mathcal{A}_1, \mathcal{A}_2, H, r, \mathbb{P})$



- $\mathcal{S}$: set of states; $\mathcal{A}_1, \mathcal{A}_2$: set of actions for the max-player/ the min-player
- H: horizon (the length of the game)
- $r_h(s_h, a_h, b_h) \in [0,1]$: reward function at step h
- $\mathbb{P}_h(s_{h+1} \mid s_h, a_h, b_h)$: transition probability at step h

Policy, value function, and Nash equilibria

- Policy: for the max-player: $\pi = \{\pi_h : \mathcal{S} \to \Delta(\mathcal{A}_1)\}$; for the min-player $\nu = \{\nu_h : \mathcal{S} \to \Delta(\mathcal{A}_2)\}$.
- V-function: $V_h^{\pi,\nu}(s_h) := \mathbb{E}_{\pi,\nu}\Big[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'}) \mid s_h\Big]$.
- Q-function: $Q_h^{\pi,\nu}(s_h, a_h, b_h) := \mathbb{E}_{\pi,\nu}\Big[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'}) \mid s_h, a_h, b_h\Big]$.
- Best response: $V_h^{\pi,*} = V_h^{\pi,\mathrm{br}(\pi)} = \inf_{\nu} V_h^{\pi,\nu}, \quad V_h^{*,\nu} = V_h^{\mathrm{br}(\nu),\nu} = \max_{\pi} V_h^{\pi,\nu}$.
- Nash equilibrium (NE): We say (π^*, ν^*) is an NE if π^* and ν^* are the best response to each other.

Metric (Sub-optimality gap): For any (π, ν) and $x \in \mathcal{S}$:

$$\mathrm{SubOpt}((\pi, \nu), x) = V_1^{*,\nu}(x) - V_1^{\pi,*}(x).$$

Data Collection Process

Assumption: The dataset $\mathscr{D} = \{(s_h^\tau, a_h^\tau, b_h^\tau)\}_{\tau,h=1}^{K,H}$ is compliant with the underlying MG:

$$\begin{aligned}
&\mathbb{P}_{\mathscr{D}}\left(r_h^\tau = r, s_{h+1}^\tau = s \,|\, \{(s_h^i, a_h^i, b_h^i)\}_{i=1}^{\tau}, \{(r_h^i, s_{h+1}^i)\}_{i=1}^{\tau-1}\right) \\
&\quad = \mathbb{P}_h\left(r_h = r, s_{h+1} = s \,|\, s_h = s_h^\tau, a_h = a_h^\tau, b_h = b_h^\tau\right),
\end{aligned}$$

for all $h \in [H], s \in \mathcal{S}$, where $\mathbb{P}$ is taken with respect to the underlying MG.

- Markov property + compliant with the underlying MG
- This assumption holds if the dataset is collected by a fixed behavior policy.
- Sequentially adjusted actions (a_h^τ, b_h^τ)

Linear Function Approximation

Linear MG (Xie et al., 2020)

$$r_h(x,a,b) = \phi(x,a,b)^\top \theta_h, \quad \mathbb{P}_h(\,\cdot\mid x,a,b) = \phi(x,a,b)^\top \mu_h(\,\cdot\,)\,.$$

- Q-function admits a linear form: $Q_h^{\pi,\nu}(x,a,b) = \langle \phi(x,a,b), w_h^{\pi,\nu} \rangle$
- Notation: $\phi_h^\tau = \phi(s_h^\tau, a_h^\tau, b_h^\tau), \phi_h = \phi(s_h, a_h, b_h)$

Existing Results for Offline MDP

- Single policy (optimal policy) coverage is the necessary and sufficient condition for achieving sample-efficiency.
- Tabular (Rajaraman et al., 2021, Xie et al., 2021, Uehara and Sun 2021):

$$\sup_{s,a,h} \frac{d_h^{\pi^*}(s,a)}{\mu_h(s,a)}$$

- Linear (Jin et al., 2021, Zenette et al., 2021, Yin et al., 2022):

$$\mathbb{E}_{\pi^*}[\sum_{h=1}^{H} \phi_h^\top \Lambda_h^{-1} \phi_h], \quad \text{where } \Lambda_h = \sum_{k=1}^{K} \phi_h^k (\phi_h^k)^\top + \lambda \cdot I$$

Q: Single policy (NE) coverage is necessary and sufficient?

Single Policy (NE) Coverage is Insufficient

Consider the MGs $\mathscr{M}_1$ and $\mathscr{M}_2$ with payoff matrices:

$$G_1 = \begin{pmatrix} 0.5 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \qquad G_2 = \begin{pmatrix} 0 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{pmatrix}$$

↓

$$\mathrm{SubOpt}_{\mathscr{M}_1}((\hat{\pi}, \hat{\nu}), x) + \mathrm{SubOpt}_{\mathscr{M}_2}((\hat{\pi}, \hat{\nu}), x) \geq 2$$

↓

Either $\mathrm{SubOpt}_{\mathscr{M}_1}((\hat{\pi}, \hat{\nu}), x)$ or $\mathrm{SubOpt}_{\mathscr{M}_2}((\hat{\pi}, \hat{\nu}), x)$ is no less than 1

↓

NE coverage is insufficient

What Coverage Condition is Sufficient?

$\{(\pi, \nu) : (\pi, \nu) \text{ is arbitrary}\}$

$$\begin{pmatrix} 0.5 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}$$

😄

$\{(\pi^*, \nu) : \nu \text{ is arbitrary}\} \cup \{(\pi, \nu^*) : \pi \text{ is arbitrary}\}$

$$\begin{pmatrix} * & -1 & * \\ 1 & 0 & 1 \\ * & -1 & * \end{pmatrix}$$

😄

Ensure that π^* and ν^* are the best response to each other — the definition of NE

Pessimistic Minimax Value Iteration (PMVI)

- Estimate linear coefficients (least-squares regression)
- Estimate Q-functions (pessimism)
- Calculate the output policy pair (NE subroutines)

Estimate Linear Coefficients

- Initialization: Set $\underline{V}_{H+1}(\cdot) = \overline{V}_{H+1}(\cdot) = 0$.
- At h-th step, we estimate the linear coefficients by solving the following least-squares regression problem:

$$\underline{w}_h \leftarrow \mathrm{argmin}_w \sum_{\tau=1}^{K} [r_h^\tau + \underline{V}_{h+1}(x_{h+1}^\tau) - (\phi_h^\tau)^\top w]^2 + \|w\|_2^2,$$

$$\overline{w}_h \leftarrow \mathrm{argmin}_w \sum_{\tau=1}^{K} [r_h^\tau + \overline{V}_{h+1}(x_{h+1}^\tau) - (\phi_h^\tau)^\top w]^2 + \|w\|_2^2.$$

- Solving the above equation gives

$$\underline{w}_h \leftarrow \Lambda_h^{-1}(\sum_{\tau=1}^{K} \phi_h^\tau(r_h^\tau + \underline{V}_{h+1}(x_{h+1}^\tau))),$$

$$\overline{w}_h \leftarrow \Lambda_h^{-1}(\sum_{\tau=1}^{K} \phi_h^\tau(r_h^\tau + \overline{V}_{h+1}(x_{h+1}^\tau))),$$

$$\text{where } \Lambda_h \leftarrow \sum_{\tau=1}^{K} \phi_h^\tau(\phi_h^\tau)^\top + I\,.$$

Estimate Q-functions

- Pessimistic estimators:

$$\underline{Q}_h(\,\cdot\,,\cdot\,,\cdot\,) \leftarrow \Pi_{H-h+1}\{\phi(\,\cdot\,,\cdot\,,\cdot\,)^\top \underline{w}_h - \Gamma_h(\,\cdot\,,\cdot\,,\cdot\,)\},$$
$$\overline{Q}_h(\,\cdot\,,\cdot\,,\cdot\,) \leftarrow \Pi_{H-h+1}\{\phi(\,\cdot\,,\cdot\,,\cdot\,)^\top \overline{w}_h + \Gamma_h(\,\cdot\,,\cdot\,,\cdot\,)\}\,.$$

- Penalty term:

$$\Gamma_h(\,\cdot\,,\cdot\,,\cdot\,) = \beta\sqrt{\phi(\,\cdot\,,\cdot\,,\cdot\,)^\top \Lambda_h^{-1} \phi(\,\cdot\,,\cdot\,,\cdot\,)}$$

Calculate the Output Policy Pair: NE Subroutine

- Solve two normal-form game:

$$(\hat{\pi}_h(\cdot \mid \cdot), \nu'_h(\cdot \mid \cdot)) \leftarrow \mathrm{NE}(\underline{Q}_h(\cdot, \cdot, \cdot)),$$

$$(\pi'_h(\cdot \mid \cdot), \hat{\nu}_h(\cdot \mid \cdot)) \leftarrow \mathrm{NE}(\overline{Q}_h(\cdot, \cdot, \cdot)) .$$

- Calculate V-functions:

$$\underline{V}_h(\cdot) \leftarrow \mathbb{E}_{a \sim \hat{\pi}_h(\cdot|\cdot), b \sim \nu'_h(\cdot|\cdot)} \underline{Q}_h(\cdot, a, b),$$

$$\overline{V}_h(\cdot) \leftarrow \mathbb{E}_{a \sim \pi'_h(\cdot|\cdot), b \sim \hat{\nu}_h(\cdot|\cdot)} \overline{Q}_h(\cdot, a, b) .$$

- Output: $(\hat{\pi} = \{\hat{\pi}_h\}_{h=1}^H, \hat{\nu} = \{\hat{\nu}_h\}_{h=1}^H)$.

Main Results for PMVI

Theorem: Let $\beta = \mathcal{O}(dH \log(dHK/p))$, it holds with probability at least $1 - p$ that

$$\mathrm{SubOpt}\big((\hat{\pi}, \hat{\nu}), x\big) \le 4\beta \cdot \mathrm{RU}(\mathscr{D}, x)\,.$$

- A new notion: Relative Uncertainty:

$$\mathrm{RU}(\mathscr{D}, x) = \inf_{(\pi^*,\nu^*)\text{ is NE}} \left\{ \max \left\{ \sup_{\nu} \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\nu}\left[\sqrt{\phi_h^\top \Lambda_h^{-1} \phi_h} \,\middle|\, s_1 = x\right], \sup_{\pi} \sum_{h=1}^{H} \mathbb{E}_{\pi,\nu^*}\left[\sqrt{\phi_h^\top \Lambda_h^{-1} \phi_h} \,\middle|\, s_1 = x\right] \right\} \right\}.$$

- Data-dependent bound: $\Lambda = \{\Lambda_h\}_{h\in[H]}$ is decided by the offline dataset.

- Only depends on how well $\{(\pi^*, \nu) : \nu \text{ is arbitrary}\} \cup \{(\pi, \nu^*) : \pi \text{ is arbitrary}\}$ are covered - no requirement on coverage of all policy pairs.

- Low relative uncertainty is the sufficient condition for achieving sample-efficiency.

Main Results for PMVI

Sufficient Coverage of Relative Information

$$\Lambda_h \geq I + c_1 \cdot K \cdot \max\left\{ \sup_{\nu} \mathbb{E}_{\pi^*,\nu}\left[\phi_h \phi_h^\top \mid s_1 = x\right], \sup_{\pi} \mathbb{E}_{\pi,\nu^*}\left[\phi_h \phi_h^\top \mid s_1 = x\right] \right\}.$$

$$\text{SubOpt}\big((\hat{\pi}, \hat{\nu}), x\big) \leq \tilde{\mathcal{O}}(d^{3/2} H^2 K^{-1/2})$$

Well-Explored Dataset

Suppose the dataset is collected by a fixed behavior policy pair $(\bar{\pi}, \bar{\nu})$. Moreover

$$\lambda_{\min}(\mathbb{E}_{\bar{\pi},\bar{\nu}}[\phi_h \phi_h^\top]) \geq c, \quad \forall h \in [H].$$

$$\text{SubOpt}\big((\hat{\pi}, \hat{\nu}), x\big) \leq \tilde{\mathcal{O}}(d H^2 K^{-1/2})$$

Proof Sketch

$$\mathrm{SubOpt}\big((\hat{\pi},\hat{\nu}),x\big) = V_1^{*,\hat{\nu}}(x) - V_1^{\hat{\pi},*}(x) = \underbrace{V_1^{*,\hat{\nu}}(x) - V_1^{*}(x)}_{\text{(i)}} + \underbrace{V_1^{*}(x) - V_1^{\hat{\pi},*}(x)}_{\text{(ii)}}$$

$$\text{(i)} = V_1^{*,\hat{\nu}}(x) - V_1^{*}(x)$$

$$\leq \overline{V}_1(x) - V_1^{*}(x)$$ $V_1^{*,\hat{\nu}}(x) \leq \overline{V}_1(x)$ (Pessimism)

$$\leq \overline{V}_1(x) - V_1^{\pi',\nu^*}(x)$$ $V_1^{*}(x) \leq V_1^{\pi',\nu^*}(x)$ (definition of NE)

$$= \sum_{h=1}^{H} \mathbb{E}_{\pi',\nu^*}\Big[\langle \overline{Q}_h(s_h,\cdot,\cdot), \pi'_h(\cdot\,|\,x) \otimes \hat{\nu}_h(\cdot\,|\,x) - \pi'_h(\cdot\,|\,s_h) \otimes \nu^*_h(\cdot\,|\,s_h)\rangle \,|\, s_1 = x\Big]$$ Decomposition Lemma

$$- \sum_{h=1}^{H} \mathbb{E}_{\pi',\nu^*}[\bar{\iota}_h(s_h,a_h,b_h)\,|\,s_1 = x]$$ $\bar{\iota}_h(x,a,b) = \mathbb{E}[r_h(s_h,a_h,b_h) + \overline{V}_{h+1}(s_{h+1})\,|\,(s_h,a_h,b_h) = (x,a,b)] - \overline{Q}_h(x,a,b)$

$$\leq 2\sum_{h=1}^{H} \mathbb{E}_{\pi',\nu^*}[\Gamma_h(s_h,a_h,b_h)\,|\,s_1 = x]$$ Definition of output policy & Pessimism

$$\leq 2\beta \cdot \mathrm{RU}(\mathscr{D},x)$$ Definitions of Γ_h and $\mathrm{RU}(\mathscr{D},x)$

Low Relative Uncertainty is Necessary?

NE Coverage:
😭

Other Conditions:
😐

Low Relative Uncertainty:
😊

Q: Is there a coverage assumption weaker than low relative uncertainty but stronger than NE coverage that empowers efficient offline learning?

Low Relative Uncertainty is Necessary

Minimax Lower Bound:

$$\mathbb{E}_{\mathscr{D}}\left[\frac{\mathrm{SubOpt}\left(\mathrm{Algo}(\mathscr{D});x\right)}{\mathrm{RU}(\mathscr{D},x)}\right] \geqslant C',$$

where C' is an absolute constant and x is the initial state. The expectation is taken with respect to $\mathbb{P}_{\mathscr{D}}$.

$\downarrow$

Low relative uncertainty is necessary

Conclusion and Future Directions

- We propose the first computationally efficient and nearly minimax optimal algorithm for offline linear MGs
- We figure out that low relative uncertainty is the necessary and sufficient condition for achieving sample efficiency in offline linear MGs setup
- General function approximations, offline general-sum MGs…

Thank You!

Paper: https://arxiv.org/abs/2202.07511