A Self-Play Posterior Sampling Algorithm for Zero-Sum Markov Games

Wei Xiong (HKUST)

The Hong Kong University of Science and Technology

2022.7.16

Joint work with: Han Zhong (Peking University); Chengshuai Shi, and Cong Shen (University of Virginia); and Tong Zhang (HKUST and Google Research).

Episodic Two-player Zero-sum MGs

$MG(H, \mathcal{X}, \mathcal{A}, \mathcal{B}, \mathbb{P}, r)$.

- Initial state $x^1 \in \mathcal{X}$ is revealed to two players;
- Two players take action $(a^1, b^1) \in \mathcal{A} \times \mathcal{B}$ individually;
- Next state $x^2 \sim \mathbb{P}_1(\cdot|x^1, a^1, b^1)$ and Reward $r^h = r(x^1, a^1, b^1)$;
- Step 2 begins and the episode ends after step H.

Given policy pairs (μ, ν), we define the value function:

$$V_h^{\mu,\nu}(x) = \mathbb{E}_{\mu,\nu}\left[\sum_{h'=h}^{H} r^{h'}\left(x^{h'}, a^{h'}, b^{h'}\right) \mid x^h = x\right]$$

$$Q_h^{\mu,\nu}(x, a, b) = \mathbb{E}_{\mu,\nu}\left[\sum_{h=h}^{H} r^{h'}\left(x^{h'}, a^{h'}, b^{h'}\right) \mid (x^h, a^h, b^h) = (x, a, b)\right],$$

Max-player: maximize $V_h^{\mu,\nu}(x)$; Best response of μ_0:
$V_h^{\mu_0,\dagger} = V_h^{\mu_0,\nu^\dagger(\mu_0)}(x) = \inf_\nu V_h^{\mu_0,\nu}(x)$;
Min-player: minimize $V_h^{\mu,\nu}(x)$: $V_h^{\dagger,\nu_0} = \max_\mu V_h^{\mu,\nu_0}$.

Learning Objective and Function Approximation

- Nash Equilibrium: (μ^*, ν^*) are best response to each other with V_1^*:

$$\mathrm{Reg}(T) := \sum_{t=1}^{T} \left[V_1^* (x_1) - V_1^{\mu_t, \dagger} (x_1) \right],$$

- Approximate Q-value by a function class $\mathcal{F}$ satisfying
  - Realizebility: $Q_h^* \in \mathcal{F}$ and for each f, $Q_h^{\mu_f, \dagger} \in \mathcal{F}$;
  - Completeness: $\mathcal{F}$ is closed under two Bellman operators.
- Bellman Operator: $f : \mathcal{X} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$,

$$(\mathcal{T}_h f)(x, a, b) := \mathbb{E}_{x' \sim \mathbb{P}(\cdot|x,a,b)}[r^h + \underbrace{\max_{\mu \in \Delta_{\mathcal{A}}} \min_{\nu \in \Delta_{\mathcal{B}}} \mu^\top f^h(x, \cdot, \cdot)\nu}_{V_{f,h}(x)}];$$

$$(\mathcal{T}_h^\mu f)(x, a, b) := \mathbb{E}_{x' \sim \mathbb{P}(\cdot|x,a,b)}[r^h + \underbrace{\min_{\nu \in \Delta_{\mathcal{B}}} \mu^\top f^h(x, \cdot, \cdot)\nu}_{V_{f,h}^\mu(x)}].;$$

- Bellman residual:

$$\mathcal{E}_h(f; x, a, b) = f^h(x, a, b) - (\mathcal{T}_h f)(x, a, b);$$
$$\mathcal{E}_h^\mu(f; x, a, b) = f^h(x, a, b) - (\mathcal{T}_h^\mu f)(x, a, b).$$

Algorithm

Distribution shift issue: use (μ_t, ν_t) to estimate $(\mu_t, \nu^\dagger(\mu_t))$.

$$\mathrm{Reg}(T) = \underbrace{\left(\sum_{t=1}^{T} V_1^*(x^1) - V_1^{\mu_t,\nu_t}(x^1)\right)}_{\text{max-player}} + \underbrace{\left(\sum_{t=1}^{T} V_1^{\mu_t,\nu_t}(x^1) - V_1^{\mu_t,\dagger}(x^1)\right)}_{\text{min-player}} :$$

Max-player:

- Optimism in *initial* value: $\tilde{p}_0(f) \propto \exp(\lambda V_{f,1}(x^1)) \prod_{h=1}^{H} p_0^h(f^h)$;
- Likelihood with denominator [DMZZ21]: $\prod_{h=1}^{H} \frac{\exp\left(-\eta L^h(f^h, f^{h+1}; S_t)\right)}{\mathbb{E}_{f^h \sim p_0^h} \exp(-\eta L^h(f^h, f^{h+1}; S_t))}$

$$L^h(f^h, f^{h+1}; S_t) = \sum_{s=1}^{t} \left[f^h(x_s^h, a_s^h, b_s^h) - r_s^h - \max_{\mu \in \Delta_{\mathcal{A}}} \min_{\nu \in \Delta_{\mathcal{B}}} \mu^\top f^h(x_s^{h+1}, \cdot, \cdot)\nu \right]^2$$

Min-player: approximate the best response of μ [JLY21, HLWY21]

- Optimism in *initial* value: $p_0^\mu(g) \propto \exp(-\lambda V_{g,1}^\mu(x^1)) \prod_{h=1}^{H} p_0^h(g^h)$;
- Likelihood with squared loss:

$$L_\mu^h(g^h, g^{h+1}; S_t) = \sum_{s=1}^{t} [g^h(x_s^h, a_s^h, b_s^h) - r_s^h - \min_{\nu \in \Delta_{\mathcal{B}}} \mu^\top f^h(x_s^{h+1}, \cdot, \cdot)\nu]^2,$$

Sketch of the proof

Immediate regret to the Bellman residuals by value decomposition lemma:

$$V^*(x^1) - V_1^{\mu,\nu}(x^1) \le \sum_{h=1}^{H} \mathbb{E}_{\mu,\nu}\mathcal{E}_h(f^h, f^{h+1}; \zeta) + \underbrace{V^*(x^1) - V_{f,1}(x^1)}_{\text{optimism in max-player}};$$

$$V_1^{\mu,\nu}(x^1) - V_1^{\mu,\dagger}(x^1) = -\sum_{h=1}^{H} \mathbb{E}_{\mu,\nu}\mathcal{E}_h^{\mu}(g^h, g^{h+1}, \zeta) + \underbrace{V_{g,1}^{\mu}(x^1) - V_1^{\mu,\dagger}(x^1)}_{\text{optimism in min-player}}.$$

Bellman residuals to *squared* Bellman residuals by decoupling coefficients

$$\sum_{h=1}^{H}\sum_{t=1}^{T}\left[\mathbb{E}_{\pi_t}\left[\mathcal{E}_h^{\mu f_t}\left(g_t; x^h, a^h, b^h\right)\right]\right] \le \mu \sum_{h=1}^{H}\sum_{t=1}^{T}\left[\sum_{s=1}^{t-1}\left[\mathbb{E}_{\pi_s}\mathcal{E}_h^{\mu f_t}\left(g_t; x^h, a^h, b^h\right)\right]^2\right] + \frac{K}{4\mu}.$$

Squared Bellman residual to Likelihood: $\mathbb{E}_{\pi_s} L^h(g_t^h, g_t^{h+1}; \zeta_s) = [\mathbb{E}_{\pi_s}\mathcal{E}_h(g_t;)]^2 + \sigma^2$.
The special likelihood allows us to replace L^h with

$$\Delta L^h(f^h, f^{h+1}; \zeta_s) = L^h(f^h, f^{h+1}; \zeta_s) - (\mathcal{T}_h f^{h+1}(x_s^h, a_s^h, b_s^h) - r_s^h - V_{f^{h+1}}(x_s^{h+1}))^2.$$

Regret bound: $\mathrm{Reg}(T) \le O\left(\sqrt{dc(\mathcal{F}, MG, T)\kappa(\frac{\beta}{T^2})T}\right)$.

- [DMZZ21] Dann, C., Mohri, M., Zhang, T., and Zimmert, J. A provably efficient model-free posterior sampling method for episodic reinforcement learning. Advances in Neural Information Processing Systems, 34, 2021;
- [JLY21] Jin, C., Liu, Q., and Yu, T. The power of exploiter: Provable multi-agent rl in large state spaces. arXiv preprint arXiv:2106.03352, 2021;
- [HLWY21] Huang, B., Lee, J. D., Wang, Z., and Yang, Z. Towards general function approximation in zero-sum markov game. International Conference on Learning Representations, 2021.