Coordinate Ascent for Off-Policy RL with Global Convergence Guarantees

Hsin-En Su, Yen-Ju Chen,
Ping-Chun Hsieh, Xi Liu

AISTATS 2023

Outline

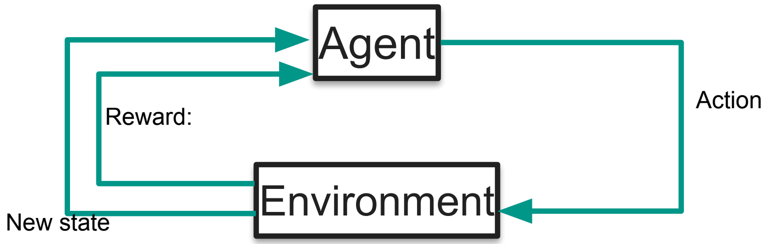
- Introduction
- Methodology
- Experiments
- Conclusion

Introduction

Reinforcement Learning

Objective:

$$J_\mu(\theta) := \mathop{\mathbb{E}}_{s\sim\mu}[V^{\pi_\theta}(s)]$$



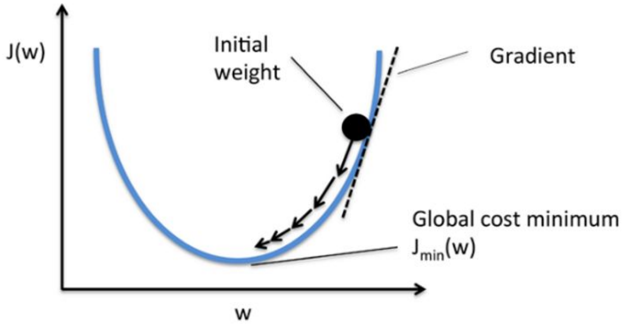
Where:

$$V^{\pi_\theta}(s) = \mathop{\mathbb{E}}_{\pi_\theta}[R_t \mid s_0 = s]$$

Policy Gradient

Update:

$$\theta_{m+1} \leftarrow \theta_m + \nabla_\theta J_\mu(\theta)$$



Where:

$$\nabla_\theta J_\mu(\theta) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_\mu^{\pi_\theta}} \mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a \mid s) A^{\pi_\theta}(s,a)\right]$$

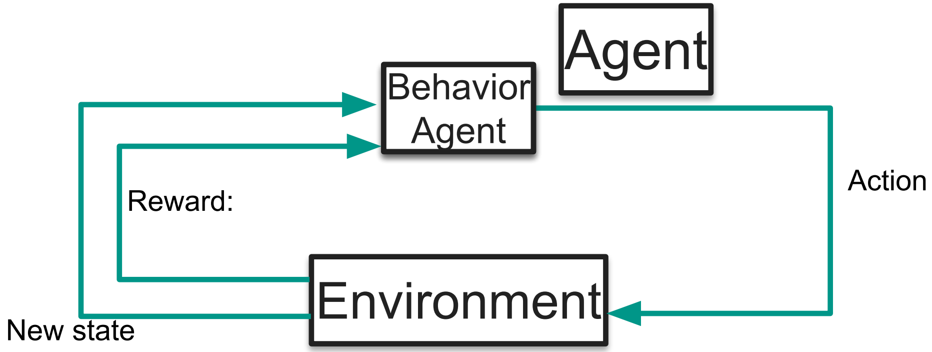
$$d_\mu^{\pi_\theta}(s) := \mathop{\mathbb{E}}_{s_0\sim\mu}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t = s \mid s_0, \pi_\theta)\right]$$

[Sutton et al., NIPS 1999], [Agarwal et al., JMLR 2021]

Motivation

Off-policy learning

- Importance sampling
- i.i.d sample from d^{π_β}



[Sutton et al., MIT press 2018]

Methodology

Coordinate Ascent Policy Optimization

General CAPO

Coordinate aspect:

$$\frac{\partial J_\mu(\theta)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi_\theta(a \mid s)A^{\pi_\theta}(s,a)$$

General CAPO form:

$$\theta_{m+1}(s,a) = \theta_m(s,a) + \mathbb{1}\{(s,a) \in \mathbb{B}_m\} \cdot \alpha(s,a) \cdot sign(A^m(s,a))$$

Learning rate!

$$\begin{bmatrix} \theta_{s_1,a_1} & \theta_{s_1,a_2} & \theta_{s_1,a_3} \\ \theta_{s_2,a_1} & \theta_{s_2,a_2} & \theta_{s_2,a_3} \\ \theta_{s_3,a_1} & \theta_{s_3,a_2} & \theta_{s_3,a_3} \end{bmatrix}$$

Global Convergence (Theorem 1.)

Consider a tabular softmax policy update using CAPO with:

$$\alpha_m(s,a) \geq \log\left(\frac{1}{\pi_{\theta_m}(a \mid s)}\right)$$

and satisfy condition1 (infinite exploration):

$$\forall (s,a),\ \lim_{M\to\infty} \sum_{m=1}^{M} \mathbb{1}\{(s,a) \in \mathbb{B}_m\} \to \infty$$

then we have:

$$V^{\pi_m}(s) \to V^*(s) \ \text{ as } m \to \infty$$

Convergence rate (Theorem 2. 3. 4.)

Algorithm	Convergence Rate
Policy Gradient (Mei et al., 2020)	$V^*(\rho) - V^{\pi_m}(\rho) \leq \frac{16 \cdot \lvert\mathcal{S}\rvert}{\inf_{m\geq 1} \pi_m(a^*\vert s)^2 \cdot (1-\gamma)^6} \cdot \left\lVert \frac{d_\mu^{\pi^*}}{\mu} \right\rVert_\infty^2 \cdot \left\lVert \frac{1}{\mu} \right\rVert_\infty \cdot \boxed{\frac{1}{m}}$
Cyclic CAPO (Theorem 2)	$V^*(\rho) - V^{\pi_m}(\rho) \leq \frac{2 \cdot \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4} \cdot \left\lVert \frac{1}{\mu} \right\rVert_\infty \cdot \max\left\{ \frac{2}{\min_s \mu(s)}, \frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)} \right\} \cdot \boxed{\frac{1}{m}}$
Batch CAPO (Theorem 3)	$V^*(\rho) - V^{\pi_m}(\rho) \leq \frac{\lvert\mathcal{A}\rvert}{(1-\gamma)^4} \cdot \left\lVert \frac{1}{\mu} \right\rVert_\infty \cdot \frac{1}{\min\limits_s\{\mu(s)\}} \cdot \boxed{\frac{1}{m}}$
Randomized CAPO (Theorem 4)	$\mathop{\mathbb{E}}\limits_{(s_m,a_m)\sim d_{\text{gen}}} [V^*(\rho) - V^{\pi_m}(\rho)] \leq \frac{2}{(1-\gamma)^4} \cdot \left\lVert \frac{1}{\mu} \right\rVert_\infty \cdot \frac{1}{\min\limits_{(s,a)}\{d_{\text{gen}}(s,a) \cdot \mu(s)\}} \cdot \boxed{\frac{1}{m}}$

Algorithm

Algorithm 1 Coordinate Ascent Policy Optimization

Initialize policy π_θ, $\theta \in \mathcal{S} \times \mathcal{A}$
for $m = 1, \cdots, M$ **do**
  Generate $|\mathcal{B}|$ state-action pairs $((s_0, a_0), \ldots, (s_{|\mathcal{B}|}, a_{|\mathcal{B}|}))$ from some generator satisfying *Condition* 1.
  for $i = 1, \cdots, |\mathcal{B}|$ **do**
    $\theta_{m+1}(s_i, a_i) \leftarrow \theta_m(s_i, a_i) + \alpha_m(s_i, a_i)\,\mathrm{sign}\,(A^m(s_i, a_i))$
  end for
end for

No need i.i.d sample from d^{π_β}

Importance Sampling

High variance in off-policy PG:

$$\nabla_\theta J(\theta) = \mathop{\mathbb{E}}_{\tau \sim \pi_\beta(\tau)} \left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \boxed{\left(\prod_{t'=1}^{t} \frac{\pi_\theta(a_{t'} \mid s_{t'})}{\pi_\beta(a_{t'} \mid s_{t'})} \right)} \left(\sum_{t'=t}^{T} r(s_{t'}, a_{t'}) \right) \right]$$

CAPO

$$\theta_{m+1}(s,a) = \theta_m(s,a) + \mathbb{1}\{(s,a) \in \mathbb{B}_m\} \cdot \alpha(s,a) \cdot sign(A^m(s,a))$$

RL - Importance Sampling, [Degris et al., ICML 2012]

On-policy CAPO

Does not necessarily achieve infinite visitation

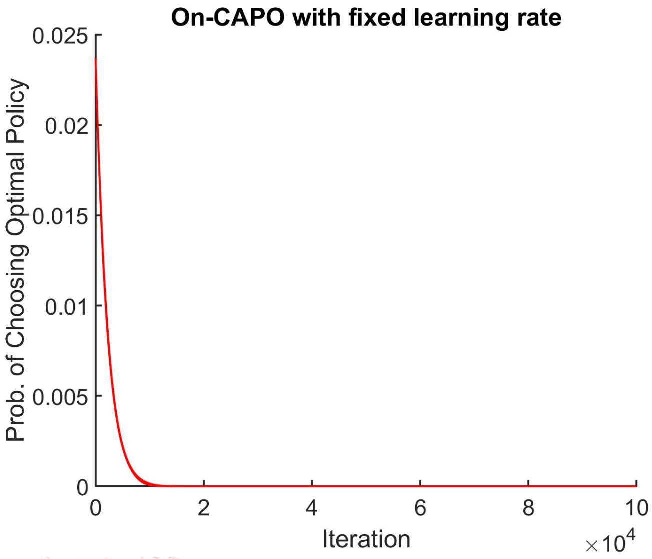
=> special design of learning rate $\Rightarrow$ Global convergence!

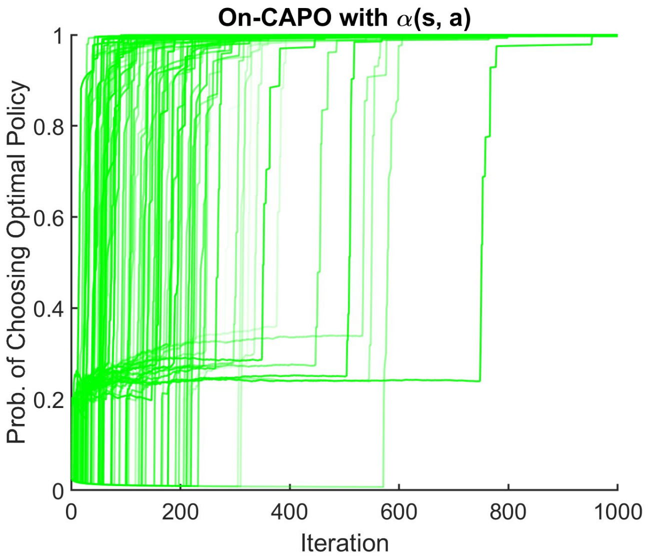
$$\alpha^m(s,a) = \begin{cases} \log\left(\frac{1}{\pi^m(a|s)}\right) & \text{, if } A^m(s,a) \le 0 \\ \log\left(\frac{\beta}{1-\beta} \cdot \frac{1}{\pi^m(a|s)}\right) & \text{, if } A^m(s,a) > 0 \text{ and } \pi^m(a \mid s < \beta) \\ \zeta \log\left(\frac{N^m(s,a)+1}{N^k(s,a)}\right) & \text{, otherwise} \end{cases}$$

$$0 < \beta \le \frac{1}{|\mathcal{A}|+1}, \quad 0 < \zeta \le \frac{1}{|\mathcal{A}|}$$

Importance of Learning Rate

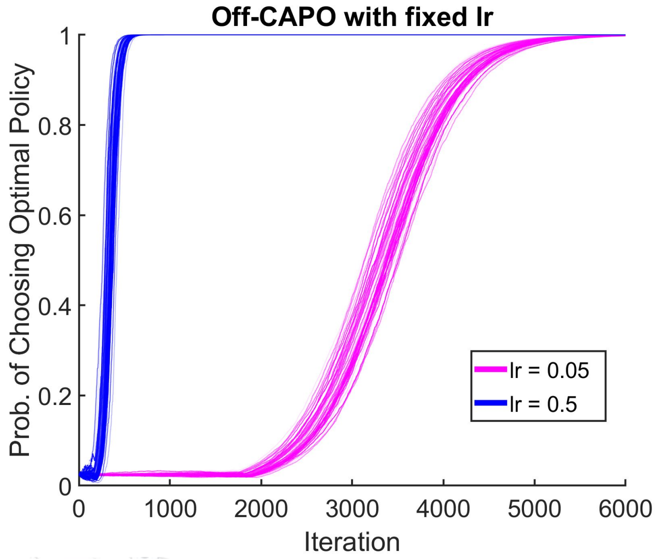
Multi-armed bandit with reward = [10, 9.9, 9.9, 0]

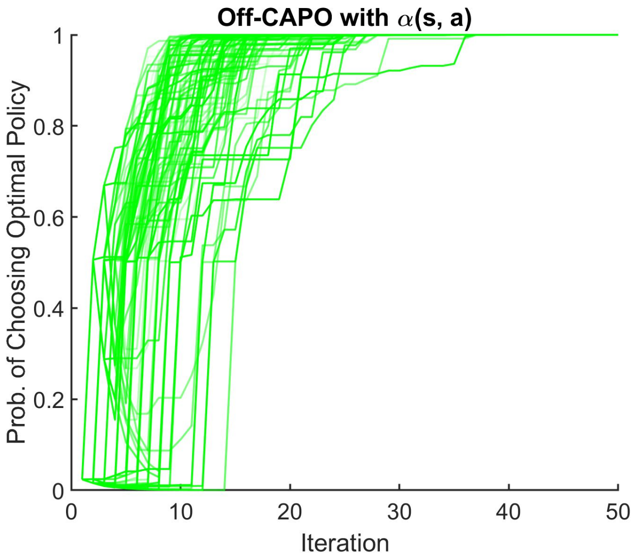




Importance of Learning Rate

Multi-armed bandit with reward = [10, 9.9, 9.9, 0]





Neural CAPO

$\pi_\theta(a|s) = f_\theta(s,a)$?

Original / **Neural Network**

$$\begin{bmatrix} \theta_{s_1,a_1} & \theta_{s_1,a_2} & \theta_{s_1,a_3} \\ \theta_{s_2,a_1} & \theta_{s_2,a_2} & \theta_{s_2,a_3} \\ \theta_{s_3,a_1} & \theta_{s_3,a_2} & \theta_{s_3,a_3} \end{bmatrix} \rightarrow \begin{bmatrix} f_\theta(s_1,a_1) & f_\theta(s_1,a_2) & f_\theta(s_1,a_3) \\ f_\theta(s_2,a_1) & f_\theta(s_2,a_2) & f_\theta(s_2,a_3) \\ f_\theta(s_3,a_1) & f_\theta(s_3,a_2) & f_\theta(s_3,a_3) \end{bmatrix}$$

Neural CAPO

Original

$$\theta_{m+1}(s,a) = \theta_m(s,a) + \mathbb{1}\{(s,a) \in \mathbb{B}_m\} \cdot \alpha(s,a) \cdot sign(A^m(s,a))$$

Neural Network

$$f_{\theta_{m+1}}(s,a) = f_{\theta_m}(s,a) + \mathbb{1}\{(s,a) \in \mathbb{B}_m\} \cdot \alpha(s,a) \cdot sign(A^m(s,a))$$

Neural CAPO

Update the policy with KL-divergence loss:

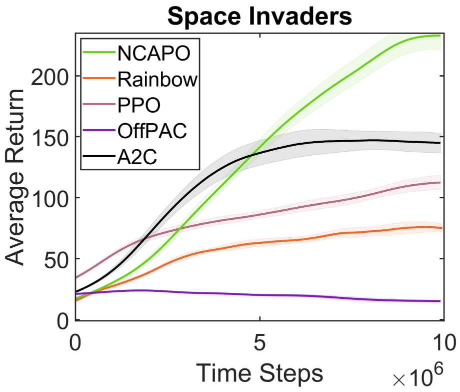
$$L(\theta) = \sum_{s \in \mathcal{B}} D_{KL}(\pi_{f_{\theta_{m+1}}}(\cdot \mid s) \mid\mid \pi_{f_{\theta_m}}(\cdot \mid s))$$

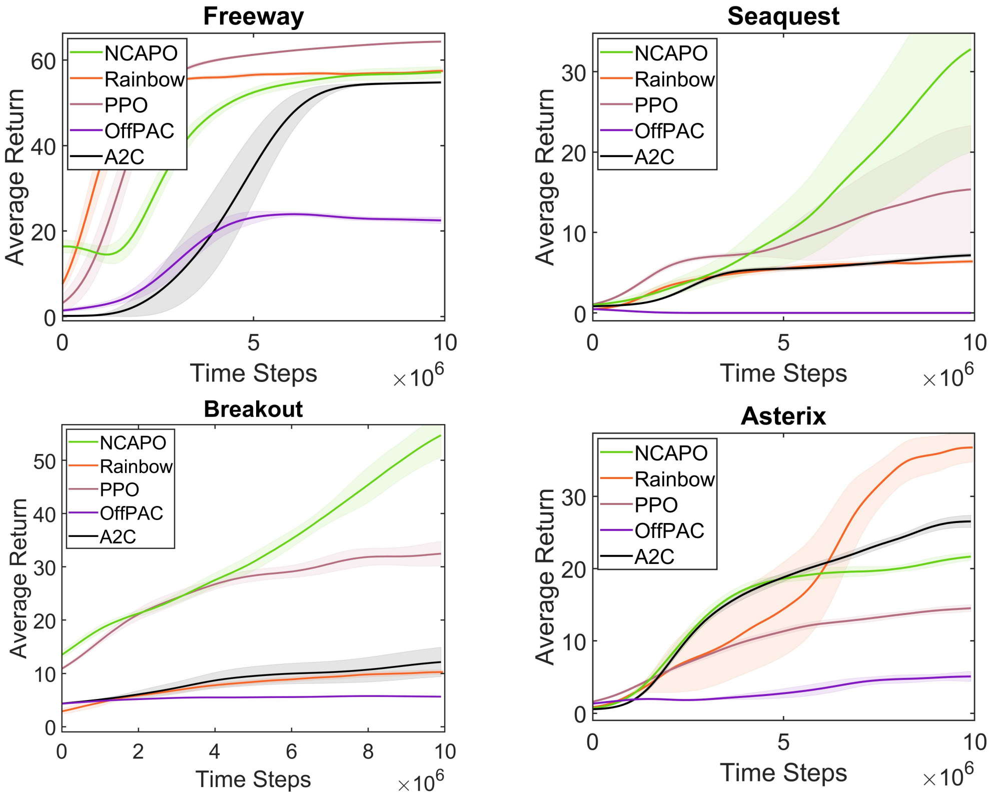
Where:

$$\pi_{f_{\theta_m}}(a \mid s) = \frac{e^{f_{\theta_m}(s,a)}}{\sum_{a' \in \mathcal{A}} e^{f_{\theta_m}(s,a')}}$$

Experiments

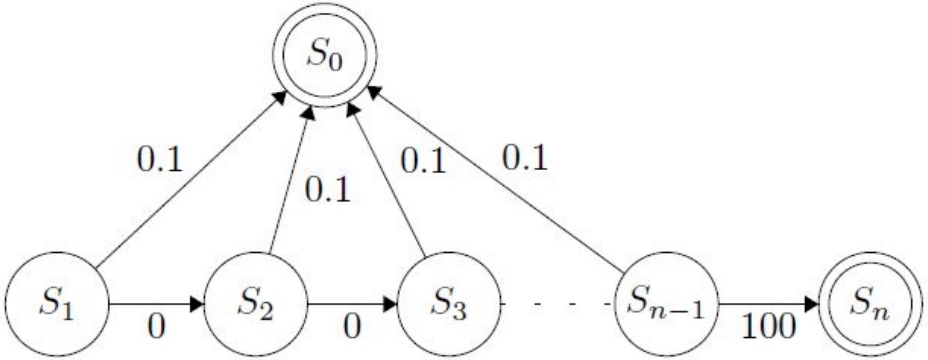
MinAtar

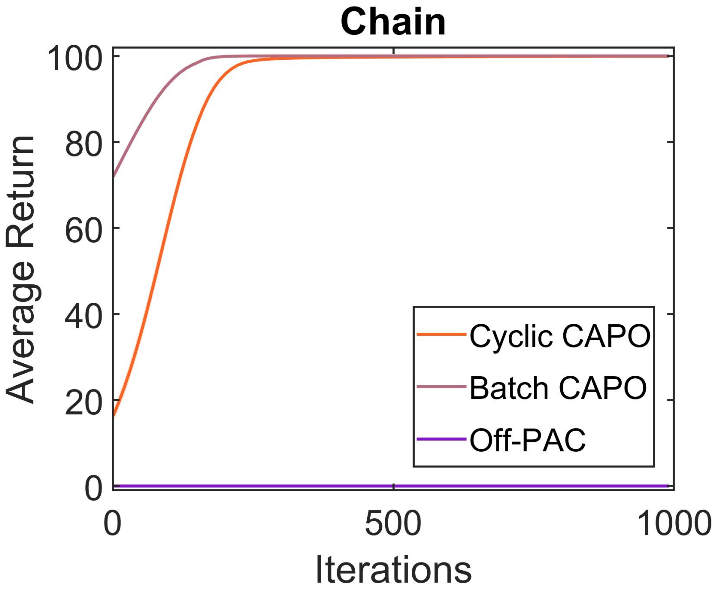




[MinAtar, Young et al., 2019],
[Rainbow, Johan et al., ICML 2021],
[PPO, Schulman et al., ICTAI 2019],
[A2C, Mnih et al., ICML 2016]

Exploration Capability



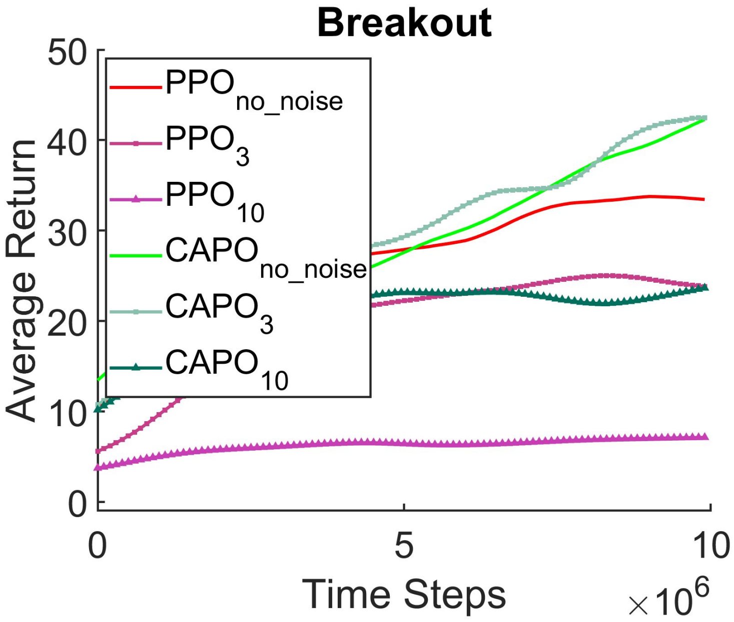


[Liu et al., 2020b]

Magnitude of advantage

Noisy reward

- 5 % of steps
- Noise sample from $N(0, \sigma^2)$



Conclusion

Conclusion

- Better off-policy learning.
- Coordinate ascent in RL.
- Avoid importance sampling & i.i.d sampling from d^{π}.

Thanks!

Thanks!

Any questions?