Coordinate Ascent for Off-Policy RL with Global Convergence Guarantees

Hsin-En Su[1]*, Yen-Ju Chen[1]*, Ping-Chun Hsieh[1], and Xi Liu[2]

[1] Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[2] Applied Machine Learning, Meta AI, Menlo Park, CA, USA

* Equal Contribution

Introduction

- We propose CAPO, an off-policy actor-critic framework which naturally enables direct off-policy policy updates with more flexible use of adaptive behavior policies, without the need for distribution correction or importance sampling correction to the gradient.
- We show that CAPO converges to a globally optimal policy under tabular softmax parameterization for general coordinate selection rules and further characterize the convergence rates of CAPO under multiple popular variants of coordinate ascent.
- Through experiments, we demonstrate that NCAPO achieves comparable or better empirical performance than various popular benchmark methods in the MinAtar.

Coordinate Ascent Policy Optimization (CAPO)

- Off-Policy Actor-Critic (Off-PAC)

$$\theta_{m+1} = \theta_m + \eta \cdot g(\theta)$$

$$g(\theta) = \mathop{\mathbb{E}}_{b}\left[\rho(s_t, a_t) \cdot \psi(s_t, a_t) \cdot Q^{\pi,\gamma}(s_t, a_t)\right]$$

- Coordinate Ascent Policy Optimization (CAPO)

$$\theta_{m+1}(s,a) = \theta_m(s,a) + \underbrace{\alpha_m(s,a)}_{\text{learning rate}} \cdot \underbrace{\mathbb{I}\{(s,a) \in B_m\}}_{\text{coordinate ascent}} \cdot \underbrace{\text{sign}\left(A^{\pi_{\theta_m}}(s,a)\right)}_{\text{update direction}}$$

Asymptotic Global Convergence of CAPO With General Coordinate Selection

Theorem 1:
Consider a tabular softmax parameterized policy π_θ. Under CAPO update with $\alpha_m(s,a) \geq \log\left(\frac{1}{\pi_{\theta_m(a|s)}}\right)$, if Condition $\lim_{M\to\infty} \sum_{m=1}^{M} \mathbb{I}\{(s,a) \in B_m\} \to \infty$ is satisfied, then we have $V^{\pi_m}(s) \to V^*(s)$ as $m \to \infty$, for all $s \in \mathcal{S}$.

Convergence Rates of CAPO With Specific Coordinate Selection Rules

- **Cyclic CAPO**: Under Cyclic CAPO, every state action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ will be chosen for policy update by the coordinate generator cyclically. Specifically, Cyclic CAPO sets $|B_m| = 1$ and $\bigcup_{i=1}^{|\mathcal{S}||\mathcal{A}|} B_{m\cdot|\mathcal{S}||\mathcal{A}|+i} = \mathcal{S} \times \mathcal{A}$.

- **Randomized CAPO**: Under Randomized CAPO, in each iteration, one state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ is chosen randomly from some coordinate generator distribution d_{gen} with support $\mathcal{S} \times \mathcal{A}$ for policy update, where $d_{\text{gen}}(s,a) > 0$ for all (s,a). Our convergence analysis can be readily extended to the case of time-varying d_{gen}.

- **Batch CAPO**: Under Batch CAPO, we let each batch contain all of the state-action pairs, i.e., $B_m = \{(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A}\}$, in each iteration.

Algorithm	Convergence Rate
Policy Gradient	$V^*(\rho) - V^{\pi_m}(\rho) \leq \frac{16\cdot|\mathcal{S}|}{\inf_{m\geq 1} \pi_m(a^*\|s)^2\cdot(1-\gamma)^6} \cdot \left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2 \cdot \left\|\frac{1}{\mu}\right\|_\infty \cdot \frac{1}{m}$
Cyclic CAPO	$V^*(\rho) - V^{\pi_m}(\rho) \leq \frac{2\cdot|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4} \cdot \left\|\frac{1}{\mu}\right\|_\infty \cdot \max\left\{\frac{2}{\min_s \mu(s)}, \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)}\right\} \cdot \frac{1}{m}$
Batch CAPO	$V^*(\rho) - V^{\pi_m}(\rho) \leq \frac{|\mathcal{A}|}{(1-\gamma)^4} \cdot \left\|\frac{1}{\mu}\right\|_\infty \cdot \frac{1}{\min_s\{\mu(s)\}} \cdot \frac{1}{m}$
Randomized CAPO	$\mathop{\mathbb{E}}_{(s,a)\sim d_{\text{gen}}}\left[V^*(\rho) - V^{\pi_m}(\rho)\right] \leq \frac{2}{(1-\gamma)^4} \cdot \left\|\frac{1}{\mu}\right\|_\infty \cdot \frac{1}{\min_{(s,a)}\left\{d_{\text{gen}}(s,a)\cdot\mu(s)\right\}} \cdot \frac{1}{m}$

Neural Coordinate Ascent Policy Optimization (NCAPO)

- To preserve the coordinate update and variable learning rate, we leverage the tabular CAPO to derive target action distributions.

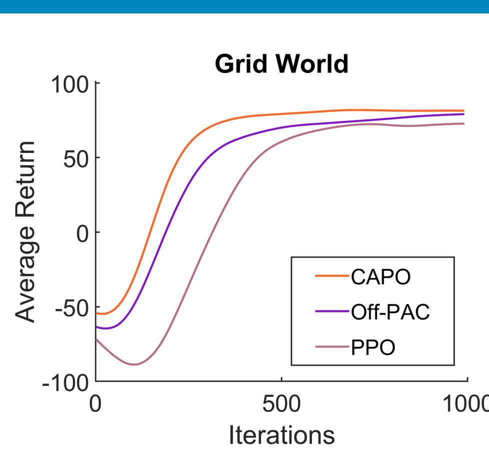
$$\pi_{\theta_m}(a|s) = \frac{\exp(f_{\theta_m}(s,a))}{\sum_{a' \in \mathcal{A}} \exp(f_{\theta_m}(s,a'))}$$

where f_{θ_m} denote the output of the policy network parameterized by θ_m.
- After CAPO update, we'll get the target policy π_{m+1}.
- We update the policy network f_θ by minimizing the NCAPO loss (KL-divergence loss).

$$\mathcal{L}(\theta) = \sum_{s \in B} D_{\text{KL}}\left(\pi_{\theta_m}(\cdot|s) \| \pi_{\theta_{m+1}}(\cdot|s)\right)$$

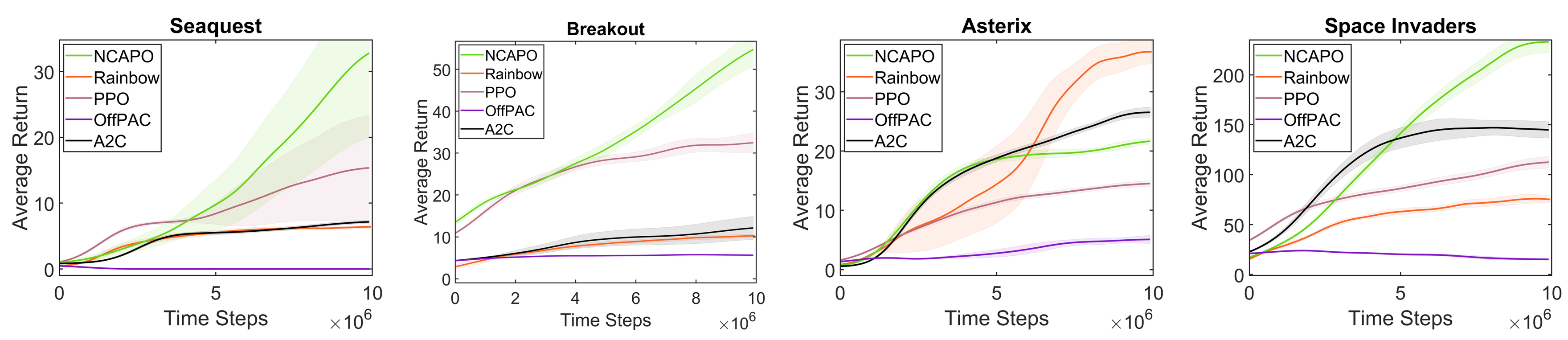
An Ablation Study: Validating Theory

- We validate the theory under a relatively simple and ideal non-atari environment.
- GridWorld: The goal is located at the bottom-right corner with a reward of **100**, the agent moves with a cost of -1.



Experimental Results - Comparison with Benchmarks

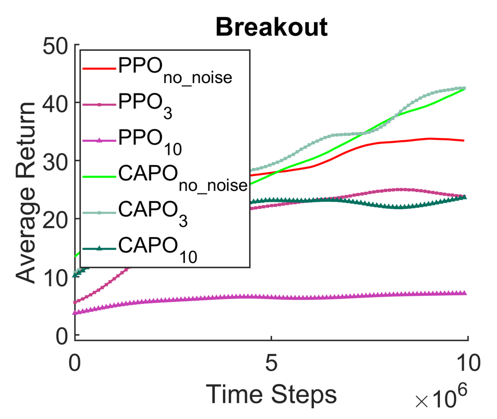
The following figures show the performance of NCAPO and other benchmark methods algorithms in MinAtar. We can observe that NCAPO has the best performance in *Seaquest, Breakout, Space Invaders*. We also see that NCAPO is more robust across tasks than PPO and Rainbow.



(a) Seaquest (b) Breakout (c) Asterix (d) Space Invaders

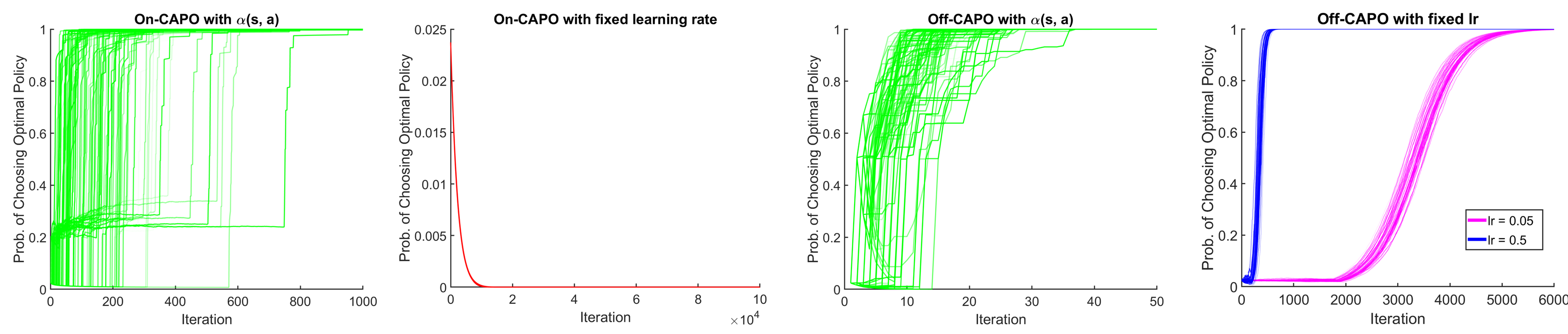
An Ablation Study: CAPO for Low-Fidelity RL Tasks

- CAPO requires only the **sign** of the advantage function, instead of the exact advantage value.
- CAPO could serve as a promising candidate solution for RL tasks with low-fidelity or multi-fidelity value estimation.
- Experiment: We evaluate NCAPO in MinAtar with noisy rewards (for **5%** of steps a large noise $\mathcal{N}(0, \sigma^2)$ is injected).



An Ablation Study: Effect of Learning Rate

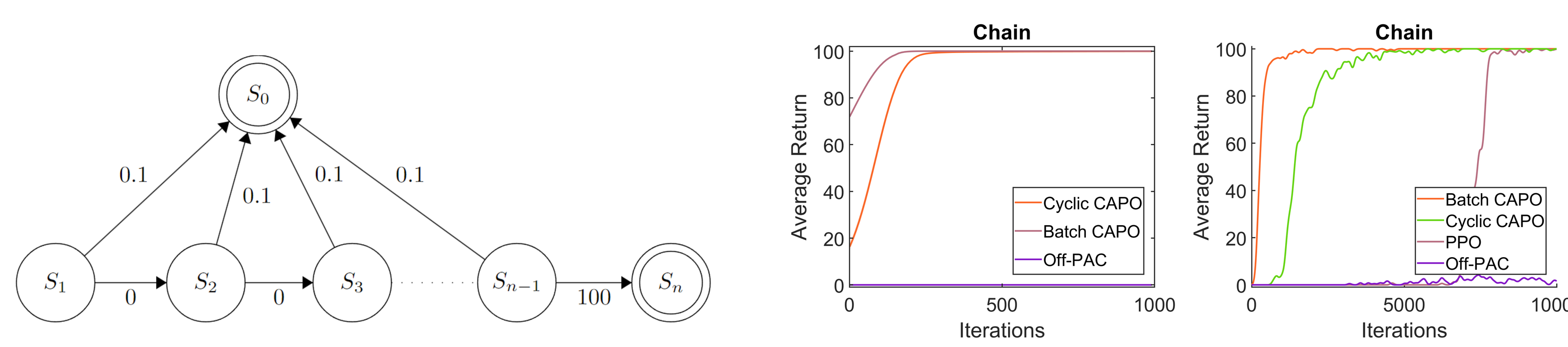
- Environment: 4-arms bandit with $r = [10, 9.9, 9.9, 0]$.
- Policy initialization: $\pi_1 \approx [0.0237, 0.4762, 0.4762, 0.0237]$.
- On-policy CAPO with fixed learning rate can get stuck in a sub-optimal policy due to the skewed policy initialization that leads to insufficient visitation to each action.
- Off-policy CAPO with fixed learning rate learns very slowly.



(a) On-CAPO general lr (b) On-CAPO fixed lr (c) Off-CAPO general lr (d) Off-CAPO fixed lr

An Ablation Study: Exploration Capability & Stochastic Environments

- Chain:
  - ▷ The environment has a total of $n+1$ states, and the agent always starts at S_1.
  - ▷ The agent may choose to move to the terminated state S_0 and receive a reward of **0.1**, or to move one state to the right.
  - ▷ The transition from S_{n-1} to S_n would induce a huge reward of **100**.
  - ▷ A well-performing policy should prefer the delayed reward of **100**.
- Stochastic Chain
  - ▷ When moving right, the stride length to be uniformly random between **0** and **3**.



(a) Chain environment (b) Chain (c) Stochastic Chain

Conclusion

- We propose CAPO, which takes the first step towards addressing off-policy policy optimization by exploring the use of coordinate ascent in RL.
- We show that the general CAPO can attain asymptotic global convergence and establish the convergence rates of CAPO with several popular coordinate selection rules.
- We show that the neural implementation of CAPO can serve as a competitive solution compared to the benchmark RL methods experimentally and thereby demonstrates the future potential of CAPO.

https://github.com/NYCU-RL-Bandits-Lab/CAPO

{littlecrazymouse.cs10,mru.11,pinghsieh}@nycu.edu.tw, xiliu.tamu@gmail.com