Policy Mirror Descent with Lookahead

Kimon Protopapas, Anas Barakat *

ETH zürich

*currently at Singapore University of Technology and Design

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- **What is lookahead?**
  - use multi-step greedy policy improvement instead of 1-step greedy.
  - Idea: applying the Bellman operator multiple times before computing a greedy policy leads to better approximation of optimal value function.

- **1-step greedy policy improvement not necessarily the best choice:**
  - Empirical success: AlphaZero and MuZero
  - Prior theoretical work: lookahead investigated with Policy Iteration e.g. [Efroni et al. 2018] but not with PG.

Main Idea: Policy Gradient Algo + Lookahead

New class of PG algorithms: *h*-PMD bringing together:

1. Policy Mirror Descent (PMD) algorithms
2. Multi-step greedy policy improvement with lookahead depth *h*

Combines benefits of Policy Gradient Methods and Tree Search Methods (e.g. MCTS)

From PMD to PMD with Lookahead

Standard PMD

$$\pi_s^{k+1} \in \mathrm{argmax}_{\pi_s \in \Delta(\mathcal{A})} \left\{ \langle Q^{\pi_k}(s,\cdot), \pi_s \rangle - \frac{1}{\eta_k} D_\phi(\pi_s, \pi_s^k) \right\}$$

$$\Updownarrow$$

$$\pi_{k+1} \in \mathrm{argmax}_{\pi \in \Pi} \left\{ \mathcal{T}^\pi V^{\pi_k} - \frac{1}{\eta_k} D_\phi(\pi, \pi_k) \right\}$$

PMD with Lookahead

$$\pi_{k+1} \in \mathrm{argmax}_{\pi \in \Pi} \left\{ \mathcal{T}^\pi \mathcal{T}^{h-1} V^{\pi_k} - \frac{1}{\eta_k} D_\phi(\pi, \pi_k) \right\}$$

$$\Updownarrow$$

$$\pi_s^{k+1} \in \mathrm{argmax}_{\pi_s \in \Delta(\mathcal{A})} \left\{ \langle Q_h^{\pi_k}(s,\cdot), \pi_s \rangle - \frac{1}{\eta_k} D_\phi(\pi_s, \pi_s^k) \right\}$$

Bellman operators $\mathcal{T}^\pi V = M^\pi(r + \gamma P V)$ $\mathcal{T} V = \max_{\pi \in \Pi} \mathcal{T}^\pi V$

Lookahead values $V_h^\pi = \mathcal{T}^\pi \mathcal{T}^{h-1} V^\pi$ $Q_h^\pi = (r + \gamma P V_h^\pi)$

Convergence and Sample Complexity

- **Setting:** discounted infinite horizon MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$
- **Exact Setting:** improved γ^h-linear convergence rate
- **Inexact Setting**: improved sample complexity
- **Function Approximation Setting:** state space size independent bound
- No dependence on distributional mismatch coefficients

Exact Setting

$$\pi_s^{k+1} \in \mathrm{argmax}_{\pi_s \in \Delta(\mathcal{A})} \left\{ \langle Q_h^{\pi_k}(s,\cdot), \pi_s \rangle - \frac{1}{\eta_k} D_\phi(\pi_s, \pi_s^k) \right\}$$

Theorem 4.1: Under suitable assumptions, iterates of *h*-PMD in the exact setting have a suboptimality gap converging to zero at a linear rate of γ^h:

$$\|V^\star - V^{\pi_k}\|_\infty \le \gamma^{hk} \left(\|V^\star - V^{\pi_0}\|_\infty + \frac{1}{1-\gamma} \sum_{t=1}^{k} \frac{c_{t-1}}{\gamma^{ht}} \right)$$

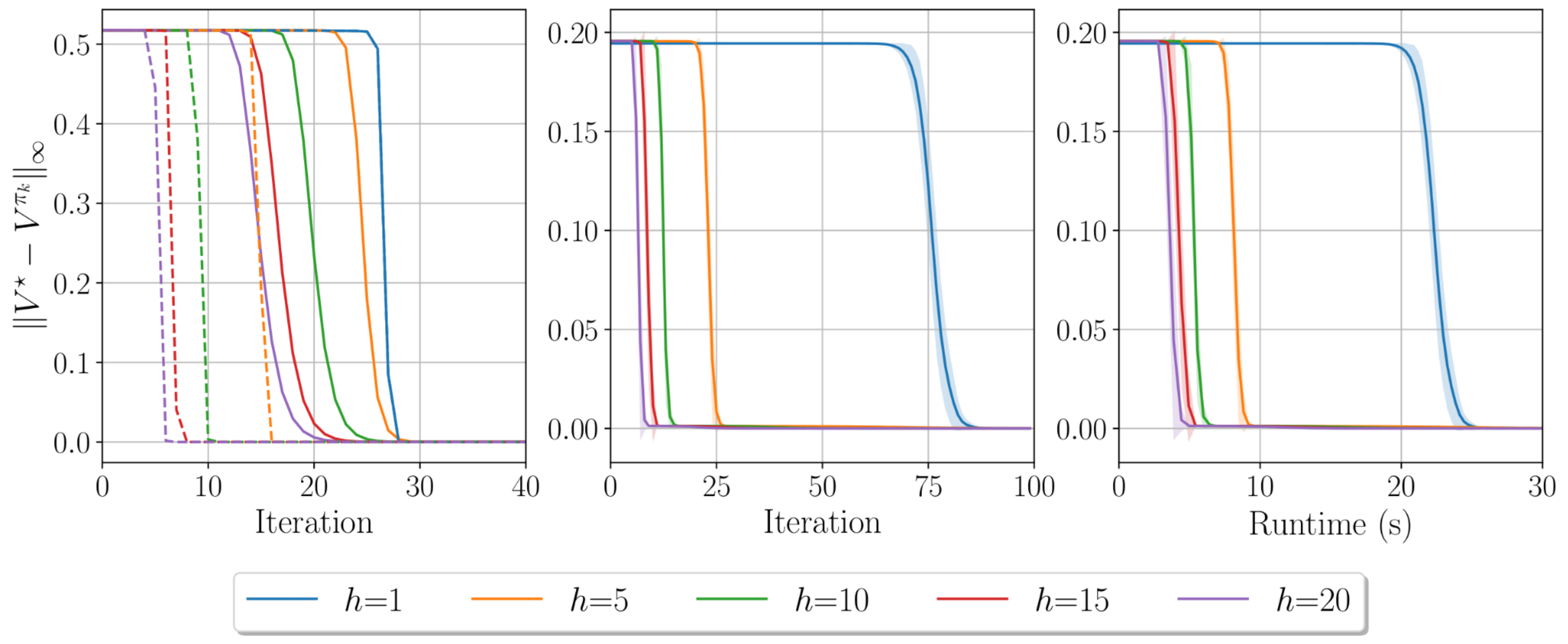
Inexact Setting

$$\pi_s^{k+1} \in \mathrm{argmax}_{\pi_s \in \Delta(\mathcal{A})} \left\{ \langle \hat{Q}_h^{\pi_k}(s,\cdot), \pi_s \rangle - \frac{1}{\eta_k} D_\phi(\pi_s, \pi_s^k) \right\}$$

- **lookahead Q-function estimation via Monte Carlo Planning**

Theorem 5.4: Under suitable assumptions, and using Monte Carlo Planning to estimate lookahead value function, inexact *h*-PMD achieves the following sample complexity:

$$\tilde{O}\left(\frac{\mathcal{S}}{h\epsilon^2(1-\gamma)^6(1-\gamma^h)^2} + \frac{\mathcal{S}\mathcal{A}}{\epsilon^2(1-\gamma)^7} \right)$$



h-PMD in the DeepSea environment from DeepMind's bsuite[2]. From left to right: exact setting, inexact setting (iteration complexity) and inexact setting (time complexity).

Function Approximation Setting

$$\pi_s^{k+1} \in \mathrm{argmax}_{\pi_s \in \Delta(\mathcal{S})} \left\{ \eta_k \langle (\Psi\theta_k)_s, \pi_s \rangle - D_\phi(\pi_s, \pi_s^k) \right\}$$

Assumption 6.1. The feature matrix $\Psi \in \mathbb{R}^{SA \times d}$ where $d \le SA$ is full rank.

Assumption 6.2 (Approximate Universal value function realizability). There exists $\epsilon_{\mathrm{FA}} > 0$ s.t. for any $\pi \in \Pi$, $\inf_{\theta \in \mathbb{R}^d} \|Q_h^\pi - \Psi\theta\|_\infty \le \epsilon_{\mathrm{FA}}$.

Theorem 6.1: Under suitable assumptions, including the assumptions above, the iterates of *h*-PMD using function approximation have a suboptimality gap converging to zero at a linear rate of γ^h, without dependence on state space size:

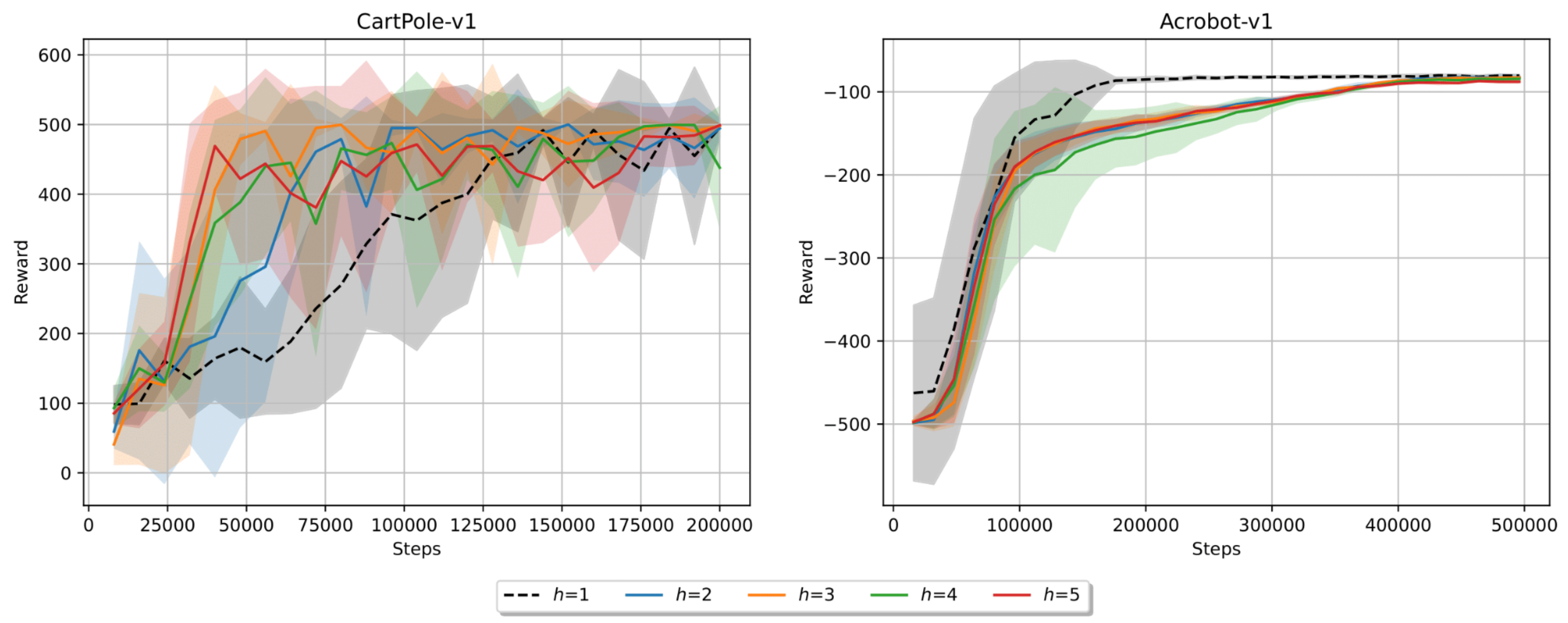
$$\|V^\star - V^{\pi_k}\|_\infty \le \gamma^{hk} \left(\|V^\star - V^{\pi_0}\|_\infty + \frac{1}{1-\gamma} \sum_{t=1}^{k} \frac{c_{t-1}}{1-\gamma} \right) + \frac{2\sqrt{d}\,\epsilon + 2\,(1+\sqrt{d})\epsilon_{\mathrm{FA}}}{(1-\gamma)(1-\gamma^h)}$$

Implies a state-space size independent sample complexity

Continuous Control Simulations

- **Implementation:**
  - *h*-PMD using MCTS for lookahead value function estimation.
  - Uses Deep Mind's MCTS implementation in JAX

- **Message:** lookahead can be beneficial in some environments even in inexact settings



References

[1] E. Johnson, C. Pike-Burke, and P. Rebeschini. Optimal convergence rate for exact policy mirror descent in discounted markov decision processes. NeurIPS 2023.

[2] Y. Efroni, G. Dalal, B. Scherrer, and S. Mannor. Beyond the one-step greedy approach in reinforcement learning. ICML 2018.

[3] J.-B. Grill, F. Altché, Y. Tang, T. Hubert, M. Valko, I. Antonoglou, and R. Munos. Monte-Carlo tree search as regularized policy optimization. ICML 2020.

[4] A. Winnicki and R. Srikant. On the convergence of policy iteration-based reinforcement learning with monte carlo policy evaluation. AISTATS 2023.