A Sample Complexity Analysis of PPO in RKHS

Liang Ding

Fudan University

Joint work with Shuang Li, Wendy Ren, Lu Zou

Content

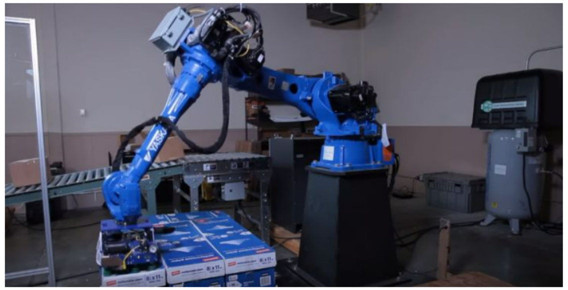
- Background
- Reproducing Kernel Hilbert Space (RKHS)
- Proximal Policy Optimization
- Numerical Experiments

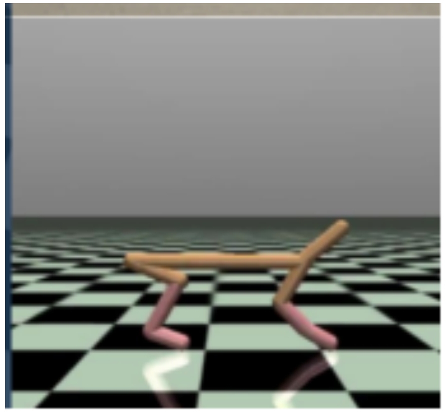
Our Goal: Provable Sample Efficiency

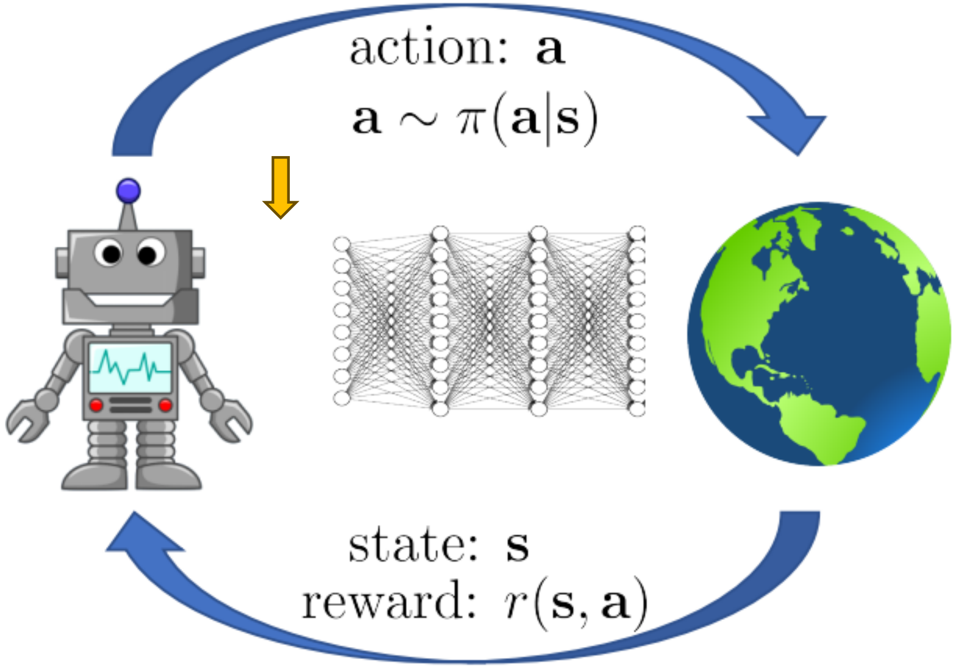
- Design provably sample-efficient RL algorithm
- Sample efficiency & Computational efficiency
- Function approximation setting

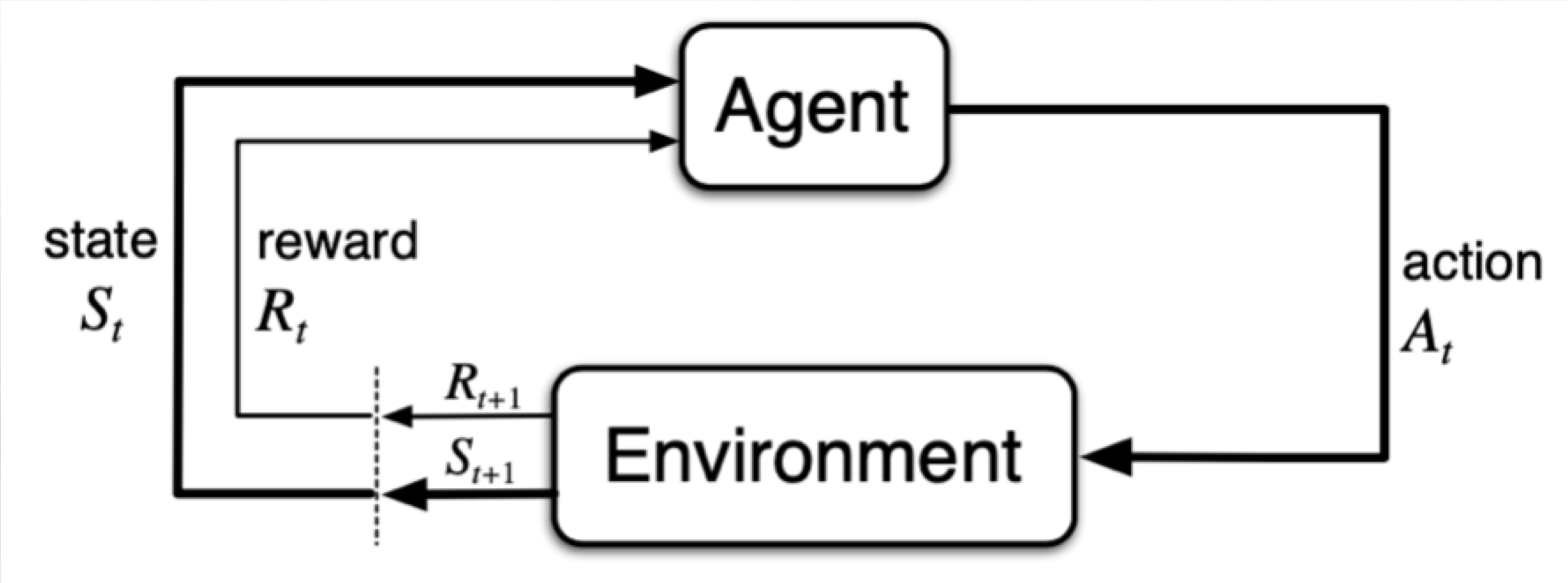
Background: What is RL?

- RL = decision under uncertainty
- RL models the natural learning-based control process.
- The agent progressively improves its behavioral skills (policy) through iterative interactions with the environment and feedback in the form of rewards.









The agent-environment interaction in a MDP

Markov Decision Process (MDP)

- RL operates within a framework called Markov Decision Process

- MDP's: General formulation for decision making under uncertainty

  Defined by: $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathbb{T}, \gamma)$

  $\mathcal{S}$: set of possible states [start state = s_0, optional terminal / absorbing state]

  $\mathcal{A}$: set of possible actions

  $R(s_t, a_t)$: reward given (state, action) tuple

  $\mathbb{p}(s|s_t, a_t)$: transition probability distribution,

  γ : discount factor

- **Markov property**: Current state completely characterizes state of the world

A RL Policy

- The agent of a RL model takes in the current **state** s_t at time t and makes an **action** $a_t \sim \pi_\theta(\cdot\,|s_t)$, where θ are the parameters of the policy.
- Most recent observation is sufficient statistic of the next state

$$s_{t+1} \sim \mathbb{p}(s|s_t, a_t)$$

- Rewards can be calculated from a reward functions (determined by the environment) such that $r_t = R(s_t, a_t)$.

- Following policy π that produces sample trajectories:

$$\cdots s_t\,, a_t, r_t, s_{t+1}, a_{t+1}, r_{t+1}, , \cdots$$

Value Function

- How good is a state?
  - State-value function $V^{\pi}(s_t)$ of a policy π at state s_t : the expected future return of π starting from s_t :

$$V^{\pi}(s_t) = \mathbb{E}_{\pi,P,R}[\sum_{k=0}^{\infty} \gamma^k r(s_{t+k}, a_{t+k}) | s_t].$$

- How good is a state-action pair?
  - Action-value function or the Q-function $Q^{\pi}(s_t, a_t)$: expected future return after performing action a_t :

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{\pi,P,R}[\sum_{k=0}^{\infty} \gamma^k r(s_{t+k}, a_{t+k}) | s_t, a_t].$$

Bellman equations: fixed point

- Bellman Equations:

$$V^{\pi}(s_t) = \mathbb{E}_{\pi,R}[r(s_t, a_t) + \gamma V^{\pi}(s_{t+1}) \ \ |s_t]$$

$$Q^{\pi}(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{\pi,R}[Q^{\pi}(s_{t+1}, a_{t+1}) \ \ |s_t]$$

- γ<1: right-hand side is a contraction mapping, for $\mathcal{S}$ and $\mathcal{A}$ are finite (Tabular RL), can use Temporal-Difference (TD):

$$V^{\pi}(s_t) \leftarrow V(s_t) + h_t[\, r(s_t, a_t) + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)]$$

$$Q^{\pi}(s_t, a_t) \leftarrow Q^{\pi}(s_t, a_t) + h_t[\, r(s_t, a_t) + \gamma Q^{\pi}(s_{t+1}, a_{t+1}) - Q^{\pi}(s_t, a_t)]$$

Objective

- Search policy π to maximize the expected value function

$$\mathbb{E}_{s_0 \sim \nu}[V^{\pi}(s_0)] = \mathbb{E}_{\nu,\pi,R}[r(s_0, a_0) + \gamma V^{\pi}(s_1) \ \ |s_0]$$

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}_{s_0 \sim \nu}[V^{\pi}(s_0)]$$

Proximal Policy Optimization

- Update rule given Q^{π_k}:

$$\pi_{k+1} = arg\max_{\pi} \mathbb{E}_{s\sim D, a\sim\pi}[Q^{\pi_k}(s,a)] - \eta\, KL(\pi||\pi_k)$$

- It has a closed form solution:

$$\pi_{k+1}(a|s) \propto \pi_k(\mathrm{a}|\mathrm{s})\exp[\eta Q^{\pi_k}(s,a)]$$

- Converge to the optimal policy for finite $\mathcal{S}$ and $\mathcal{A}$ (or linear MDP) at sub-linear rate $1/\sqrt{k}$

Problem

- Setting:
  - n i.i.d. samples of initial states $\{s_0^i\}_{i=1}^n$ following a distribution ν
  - State and action spaces $\mathcal{S}$ and $\mathcal{A}$ are large and continuous
- Impossible to sample every $s \in \mathcal{S}$ and $a \in \mathcal{A}$
- Impossible to run TD on every $s \in \mathcal{S}$ and $a \in \mathcal{A}$
- How to generalize TD to an empirical and (nonlinear) functional setting?
- If the generalization exists, what is its convergence property?

Content

- Background
- Reproducing Kernel Hilbert Space (RKHS)
- Proximal Policy Optimization
- Numerical Experiments

Kernel Function

- Let $K: \Omega\times\Omega \to \mathbb{R}$ be a symmetric positive definite kernel function with $\Omega = \mathcal{S}\times\mathcal{A}$, i.e.

$$\mathrm{K}(\omega, w) = \sum_{l} \phi_l(\omega)\phi_l(w) = \boldsymbol{\Phi}^{\mathrm{T}}(\omega)\boldsymbol{\Phi}(w),$$

where $\{\phi_l: \Omega \to \mathbb{R}\}$ are called features.

- Define the linear space: $F_{K,n}(\Omega) = \{ \sum_{i=1}^{n} \beta_i K(\cdot, \omega_i),\ \beta_i \in \mathbb{R}, \omega_i \in \Omega \}$
- Equip this space with the bilinear form:

$$\left\langle \sum_{i=1}^{n} \beta_i K(\cdot, \omega_i), \sum_{i=1}^{n} c_i K(\cdot, \omega_i) \right\rangle_K := \sum_{i,j=1}^{n} \beta_i c_j K\left(\omega_i, \omega_j\right)$$

- RKHS $\mathcal{H}_K(\Omega)$ generated by K: the closure of $F_{K,n}(\Omega)$ under inner product $\langle\cdot,\cdot\rangle_K$ (Example: Sobolev spaces, discrete set,….)

Representer Theorem

- Given data $\{\omega_i, y_i\}$, and the functional minimization

$$\hat{f} = \underset{f \in \mathcal{H}_K(\Omega)}{argmin} \sum_i L(f(\omega_i), y_i) + h(\langle f, f \rangle_K)$$

- The minimizer $\hat{f}$ admits a closed form solution:

$$\hat{f} = \sum_i \beta_i^* K(\omega_i, \cdot)$$

Representer Theorem for Bellman Equation

- If Q resides in a RKHS, given data $\{\omega_i \in \Omega, \omega_i' \in \Omega\}$, where

$$(\omega_i, \omega_i') = (s_0^i, a_0^i, s_1^i, a_1^i) \sim \nu(s_0)\pi(a_0|s_0)\mathbb{p}(s_1|s_0, a_0)\pi(a_1|s_1)$$

- Define the following fixed point KRR

$$\hat{Q}^\pi = \underset{f \in \mathcal{H}_K(\Omega)}{argmin} \frac{1}{n} \sum_i \left(f(\omega_i) - r(\omega_i) - \gamma \hat{Q}^\pi(\omega_i') \right)^2 + \lambda \|f\|_K^2$$

- $\hat{Q}$ has a closed form solution: $\hat{Q}^\pi = \sum_i \beta_i^* K(\omega_i, \cdot)$

- Intuition: use cross-covariance operator to represent the Bellman equation

Representer Theorem for Bellman Equation

- Cross-covariance operator:

$$C_{\omega,\omega} = \mathbb{E}[\mathbf{\Phi}(\omega) \otimes \mathbf{\Phi}(\omega)], \qquad C_{\omega,\omega'} = \mathbb{E}[\mathbf{\Phi}(\omega) \otimes \mathbf{\Phi}(\omega')]$$

- Bellman equation represented by cross-covariance operator (a weak form):

$$C_{\omega,\omega} Q^{\pi} = \left(C_{\omega,\omega} r + \gamma C_{\omega,\omega'} Q^{\pi}\right)$$

- Empirical cross-covariance operator:

$$\hat{C}_{\omega,\omega} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{\Phi}(\omega_i) \otimes \mathbf{\Phi}(\omega_i), \qquad \hat{C}_{\omega,\omega'} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{\Phi}(\omega_i) \otimes \mathbf{\Phi}(\omega'_i)$$

- Empirical Bellman equation with penalty

$$\hat{C}_{\omega,\omega} \hat{Q}^{\pi} = \left(\hat{C}_{\omega,\omega} r + \gamma \hat{C}_{\omega,\omega'} \hat{Q}^{\pi}\right) + \lambda \hat{Q}^{\pi}$$

Representer Theorem for Bellman Equation

- Take the RKHS functional derivative of the fixed Point RKHS

$$J[f] = \frac{1}{n}\sum_i \left(f(\omega_i) - r(\omega_i) - \gamma \hat{Q}^{\pi}(\omega_i')\right)^2 + \lambda \, \|f\|_K^2$$

- By setting $\nabla J[f] = 0$, we can exactly recover the empirical Bellman equation

$$\hat{C}_{\omega,\omega} \, \hat{Q}^{\pi} = \left(\hat{C}_{\omega,\omega} r + \gamma \hat{C}_{\omega,\omega'} \hat{Q}^{\pi}\right) + \lambda \hat{Q}^{\pi}$$

- By representer theorem, $\hat{Q}^{\pi} = \sum_i \beta_i^* K(\omega_i \, ,\cdot\,)$

Kernel Gradient Descent

- Closed form solution $\hat{Q}^{\pi} = \sum_i \beta_i^* K(\omega_i, \cdot)$ where

$$\boldsymbol{\beta}^* = [\mathbf{K} + \lambda n \mathbf{I} - \gamma \mathbf{C}]^{-1} \mathbf{r}$$

$$\mathbf{K} = \left[k\left(\omega_i, \omega_j\right)\right]_{i,\mathrm{j}}, \quad \mathbf{C} = \left[k\left(\omega_i', \omega_j\right)\right]_{i,\mathrm{j}}, \mathbf{r} = [\mathrm{r}(\omega_i)]$$

- Solve $\boldsymbol{\beta}^*$ by Kernel Gradient Descent:

$$\boldsymbol{\beta}_{t+1} = (1 - \alpha_t)\boldsymbol{\beta}_t + \eta_t(\mathbf{K}\boldsymbol{\beta}_t - \boldsymbol{r} - \gamma \mathbf{C}\boldsymbol{\beta}_t)$$

- This is exactly the Temporal-Difference if we parametrized $\hat{Q}$ by $\sum_i \beta_i K(\omega_i, \cdot)$ but replace l^2 inner product by $\langle \boldsymbol{\beta}, \boldsymbol{\beta}' \rangle = \boldsymbol{\beta} \mathbf{K}\, \boldsymbol{\beta}'$ (a preconditioner)

Kernel Gradient Descent

- Superlinear Convergence of Kernel Gradient Descent

$$\begin{aligned}\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^* &= [\mathbf{I} - (\alpha\mathbf{I} + \eta\mathbf{K} - \eta\gamma\mathbf{C})]\left[\boldsymbol{\beta}_t - [\alpha\mathbf{I} + \eta\mathbf{K} - \eta\gamma\mathbf{C}]^{-\mathbf{1}}\mathbf{r}\right] \\ &= [\mathbf{I} - (\alpha\mathbf{I} + \eta\mathbf{K} - \eta\gamma\mathbf{C})][\boldsymbol{\beta}_t - \boldsymbol{\beta}^*] \\ &= [\mathbf{I} - (\alpha\mathbf{I} + \eta\mathbf{K} - \eta\gamma\mathbf{C})]^{\boldsymbol{t+1}}[\boldsymbol{\beta}_{\mathbf{0}} - \boldsymbol{\beta}^*]\end{aligned}$$

- If eigenvalues of $[\mathbf{I} - (\alpha\mathbf{I} + \eta\mathbf{K} - \eta\gamma\mathbf{C})]$ are small, then $\boldsymbol{\beta}_t \to \boldsymbol{\beta}^*$ exponentially fast

Convergence Analysis

- From the empirical Bellman, we have a statistical-approximation error decomposition:

$$\frac{1}{n}\sum_i |D^\pi(\omega_i)|^2 - \gamma D^\pi(\omega_i) D^\pi(\omega_i{}') = \frac{1}{n}\sum_i \epsilon_i D^\pi(\omega_i) - \lambda\langle D^\pi, \hat{Q}^\pi\rangle_K$$

where $D^\pi = \hat{Q}^\pi - Q^\pi$ is the function difference

$\epsilon_i = r(\omega_i) + \gamma Q^\pi(\omega_i') - Q^\pi(\omega_i)$ is the Bellman residual

- We then can use empirical process to prove the convergence rate

Convergence Analysis (Sobolev)

- Suppose Q^{π} is the s-time weak differentiable and $\dim(\Omega) = d$ (Sobolev RKHS embedded on d-dimensional manifold)

- With step size, weight decay, iteration number, and penalty:

$$\eta \asymp n^{-1}, \qquad \alpha = \lambda, \qquad T \geq C \log n, \qquad \lambda \asymp n^{-\frac{d/2}{2s+d}}$$

- We have:

$$||Q^{\pi} - \hat{Q}_T||_{L^2} = O_p(n^{-\frac{s}{2s+d}}||Q^{\pi}||_K)$$

Convergence Analysis (Gaussian)

- Suppose Q^{π} is infinitely many differentiable and $\Omega = [0,1]^{\mathrm{d}}$ (Gaussian RKHS)

- With step size, weight decay, iteration number, and penalty:

$$\eta \asymp n^{-1}, \qquad \alpha = \lambda, \qquad T \geq C \log n, \qquad \lambda \asymp n^{-\frac{1}{2}} |\log n|^{d}$$

- We have:

$$||Q^{\pi} - \hat{Q}_T||_{L^2} = O_p(n^{-\frac{1}{2}} |\log n|^{d} ||Q^{\pi} - \hat{Q}_T||_K)$$

Content

- Background
- Offline Reinforcement Learning
- Proximal Policy Optimization
- Numerical Experiments

Policy Update

- Given Q^{π_k}, π_{k+1} has a closed form solution:

$$\pi_{k+1}(a|s) \propto \pi_k(\mathrm{a}|\mathrm{s})\exp[\Delta_k \hat{Q}_T^{(k)}(s,a)]$$

- From iteration, π_{k+1} can also be represented by a neural network:

$$\pi_{k+1} \propto \exp[f_{\boldsymbol{\theta}_{k+1}}]$$

where $f_{\boldsymbol{\theta}_{k+1}} = \Delta_k \hat{Q}_T^{(k)} + f_{\boldsymbol{\theta}_k}$, Δ_k can be considered as step size

In experiments ,both π_{k+1} and $\Delta_k \hat{Q}_T^{(k)}$ can be represented by neural nets under the framework of Neural Tangent Kernel

A Fundamental Inequality

- A fundamental inequality for the convergence of value function to the optimal:

$$\min_{1\le k\le K} \mathbb{E}_{\nu^*}\left[V^{\pi^*}(s)\right] - \mathbb{E}_{\nu^*}[V^{\pi_k}(s)] \le \frac{\sum_k 2\Delta_k \left\|Q^{\pi_k} - \hat{Q}_T^{(k)}\right\|_\infty + C}{\sum_k \Delta_k}$$

- To achieve the best stochastic sub-linear convergence rate $1/\sqrt{K}$, set

$$\Delta_k = 1/\sqrt{k}$$

and we also need $\left\|Q^{\pi_k} - \hat{Q}_T^{(k)}\right\|_\infty \lesssim \Delta_k$

Sampling Requirement

- Target: $\left\|Q^{\pi_k} - \hat{Q}_T^{(k)}\right\|_\infty \lesssim \Delta_k$
- As k increases Δ_k decreases and π_k may becomes more complicate
- Interpolation inequality from $\left\|Q^{\pi_k} - \hat{Q}_T^{(k)}\right\|_2$ to $\left\|Q^{\pi_k} - \hat{Q}_T^{(k)}\right\|_\infty$ to derive the required sample number $n^{(k)}$ for estimate Q^{π_k}:

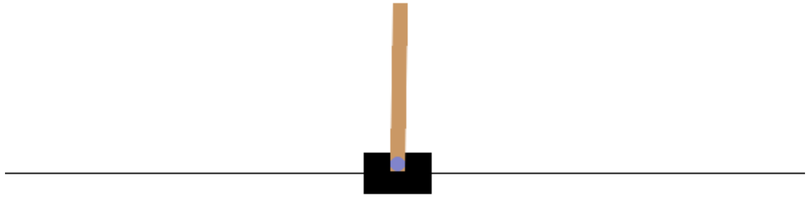
NTK	$\mathcal{O}\left(\frac{\|\pi^k\|_{\mathcal{H}}^{2d} k^d}{(1-c\gamma)^{\frac{3d+1}{d+1}}}\right)$	**Tabular**	$\mathcal{O}\left(\frac{\|\pi^k\|_{\mathcal{H}}^2 k}{(1-c\gamma)^2}\log\frac{\|\pi^k\|_{\mathcal{H}} k}{1-c\gamma} + (\sqrt{k}\|\pi^k\|_{\mathcal{H}})^{\frac{4}{1+\nu}}\right)$
Gaussian	$\mathcal{O}\left(\frac{\|\pi^k\|^{\frac{2}{1-\epsilon}} k^{\frac{1}{1-\epsilon}}}{(1-c\gamma)^2}\log\frac{\|\pi^k\|_{\mathcal{H}} k}{1-c\gamma}\right)$	**Sobolev**	$\mathcal{O}\left(\frac{\|\pi^k\|_{\mathcal{H}}^{\frac{2(2m+d)}{2m-d}} k^{\frac{2m+d}{2m-d}}}{(1-c\gamma)^{\frac{2m+d/2}{m}}}\right)$

Content

- Background
- Offline Reinforcement Learning
- Proximal Policy Optimization
- Numerical Experiments

Cart Pole

Goal: to keep the pole upright for as long as possible.



https://gymnasium.farama.org/environments/classic_control/cart_pole/

Action Space: 2 discrete actions
- 0: Push cart to the left
- 1: Push cart to the right

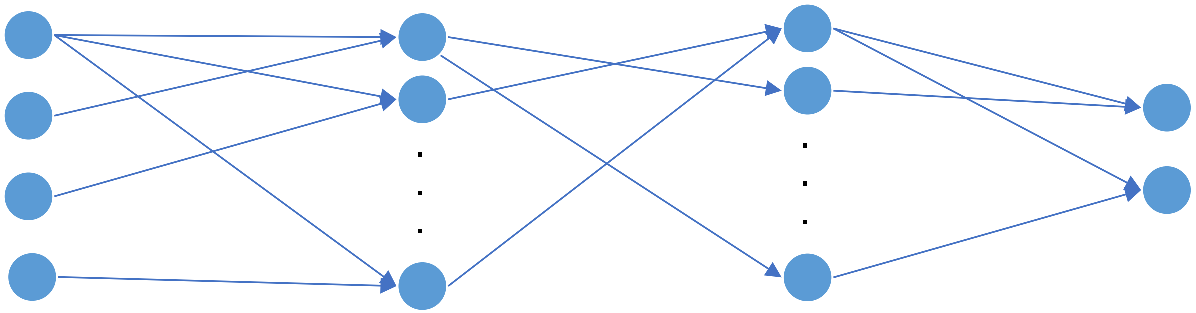
Observation Space/State Space: 4 continuous variables

Num	Observation	Min	Max
0	Cart Position	-4.8	4.8
1	Cart Velocity	-Inf	Inf
2	Pole Angle	~ -0.418 rad (-24°)	~ 0.418 rad (24°)
3	Pole Angular Velocity	-Inf	Inf

Rewards: Since, by default, a reward of +1 is given for every step taken, including the termination step. The default reward threshold is 500.

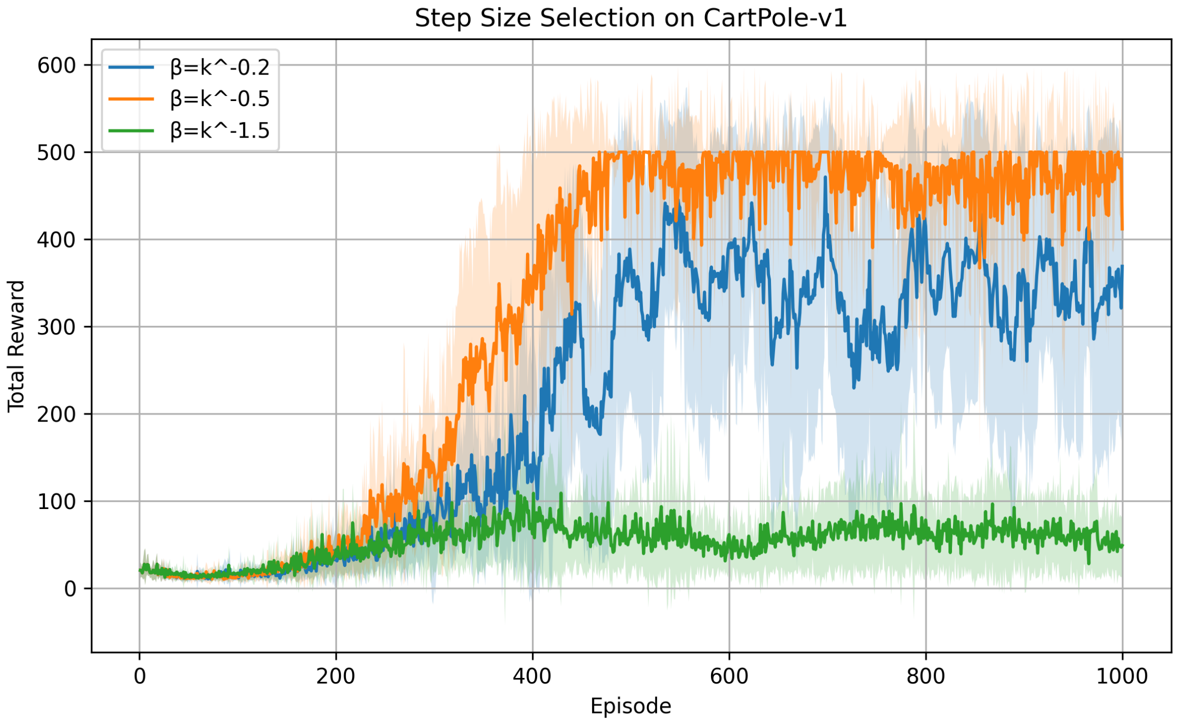
Q network: 3 layers deep neural network

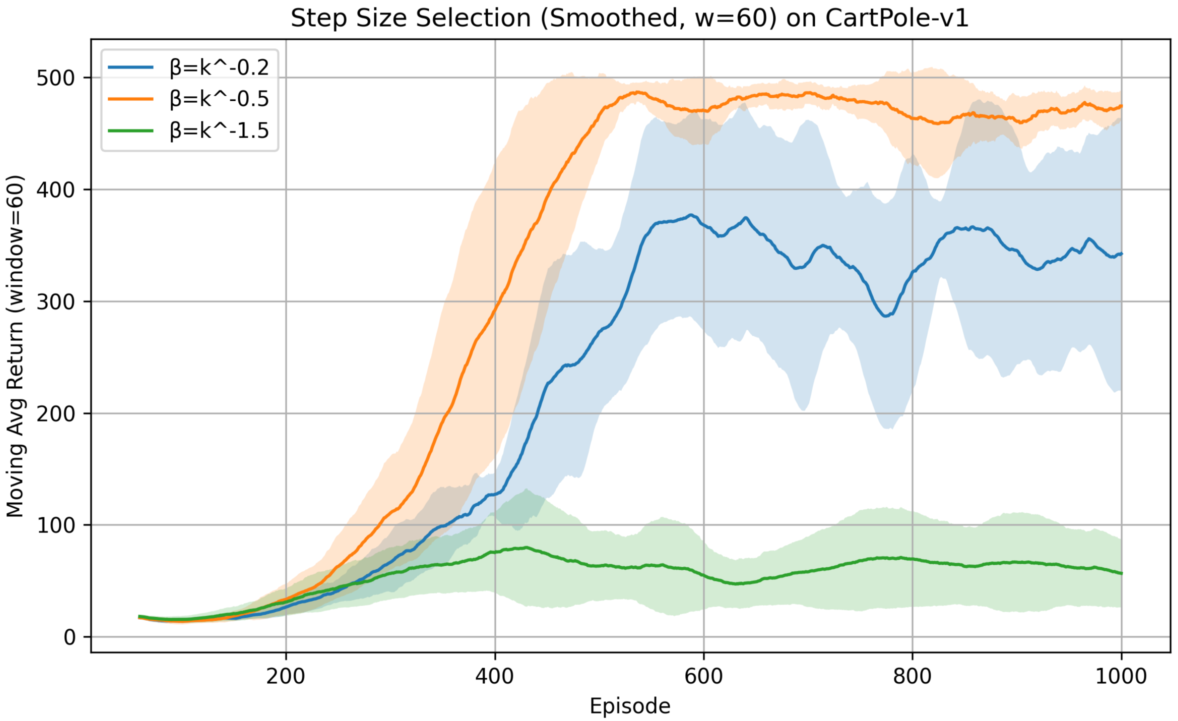
state dim (4) → hidden dim (64) → hidden dim (64) → action dim (2)



Cart Pole

- We test three different step-size schedules.
- Each schedule is run for 10 trials using different random seeds.
- Evaluation Metric: Performance is measured by the total reward per episode. Learning curves show the mean and standard deviation across the 10 seeds. Optimal reward: 500.
- We plot raw returns and moving average returns. The results matches our theoretical proof of step size selection.





Results: $\eta_k = k^{-1.5}$ (green) leads to the failrue of convergence, $\eta_k = k^{-0.5}$ (orange) gets the global convergence, $\eta_k = k^{-0.2}$ (blue) causes global divergence.

Acrobot

Goal: apply torques on the actuated joint to swing the free end of the linear chain above a given height while starting from the initial state of hanging downwards.

https://gymnasium.farama.org/environments/classic_control/acrobot/

Q network: 3 layers deep neural network

state dim (6) → hidden dim (64) → hidden dim (64) → action dim (3)

Action Space: 3 discrete actions

- 0: apply -1 torque to the actuated joint
- 1: apply 0 torque to the actuated joint
- 2: apply 1 torque to the actuated joint

Observation Space/State Space: 6 continuous variables

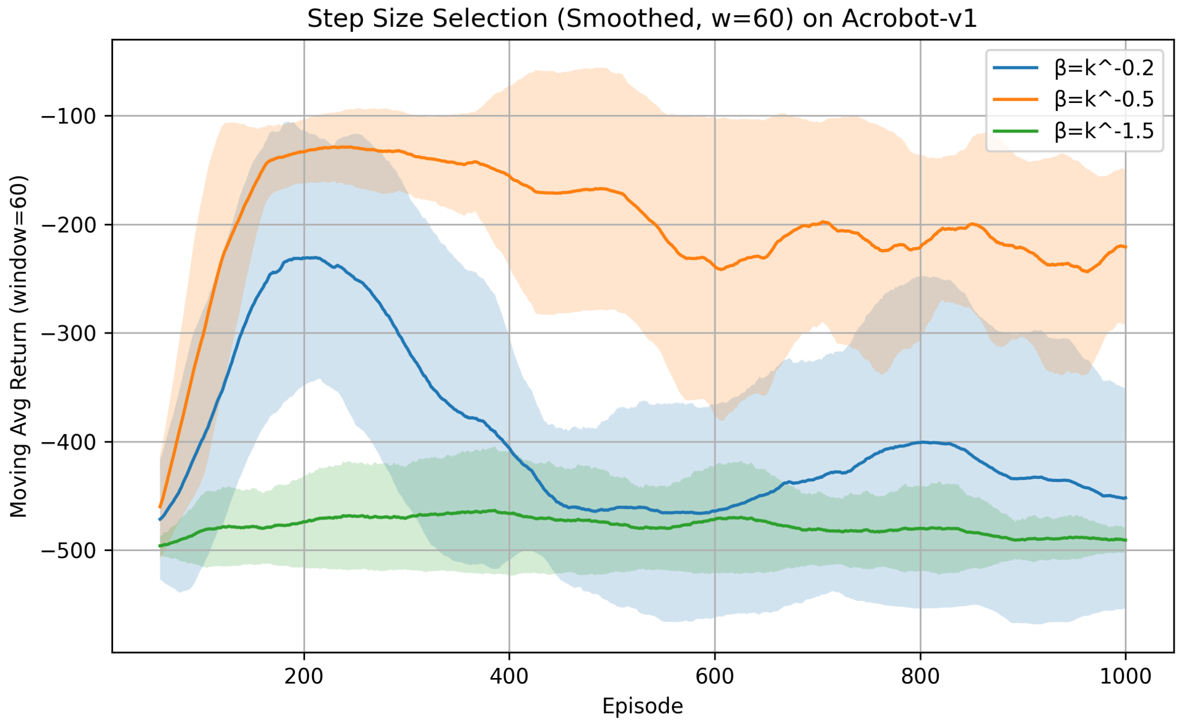
Num	Observation	Min	Max
0	Cosine of theta1	-1	1
1	Sine of theta1	-1	1
2	Cosine of theta2	-1	1
3	Sine of theta2	-1	1
4	Angular velocity of theta1	~ -12.567 (-4 * pi)	~ 12.567 (4 * pi)
5	Angular velocity of theta2	~ -28.274 (-9 * pi)	~ 28.274 (9 * pi)

Rewards

The goal is to have the free end reach a designated target height in as few steps as possible, and as such all steps that do not reach the goal incur a reward of -1. Achieving the target height results in termination with a reward of 0. The reward threshold is -100.

Acrobot

- We test three different step-size schedules.
- Each schedule is run for 10 trials using different random seeds.
- Evaluation Metric: Performance is measured by the total reward per episode. Learning curves show the mean and standard deviation across the 10 seeds. Optimal reward: -100.
- We plot raw returns and moving average returns. The results matches our theoretical proof of step size selection.

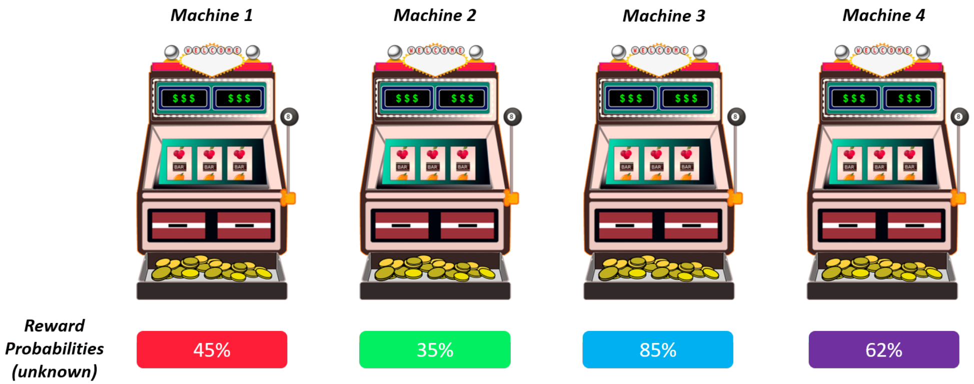


Results: $\eta_k = k^{-1.5}$ (green) leads to the failrue of convergence, $\eta_k = k^{-0.5}$ (orange) gets the global convergence, $\eta_k = k^{-0.2}$ (blue) causes global divergence.

5-armed Bandit

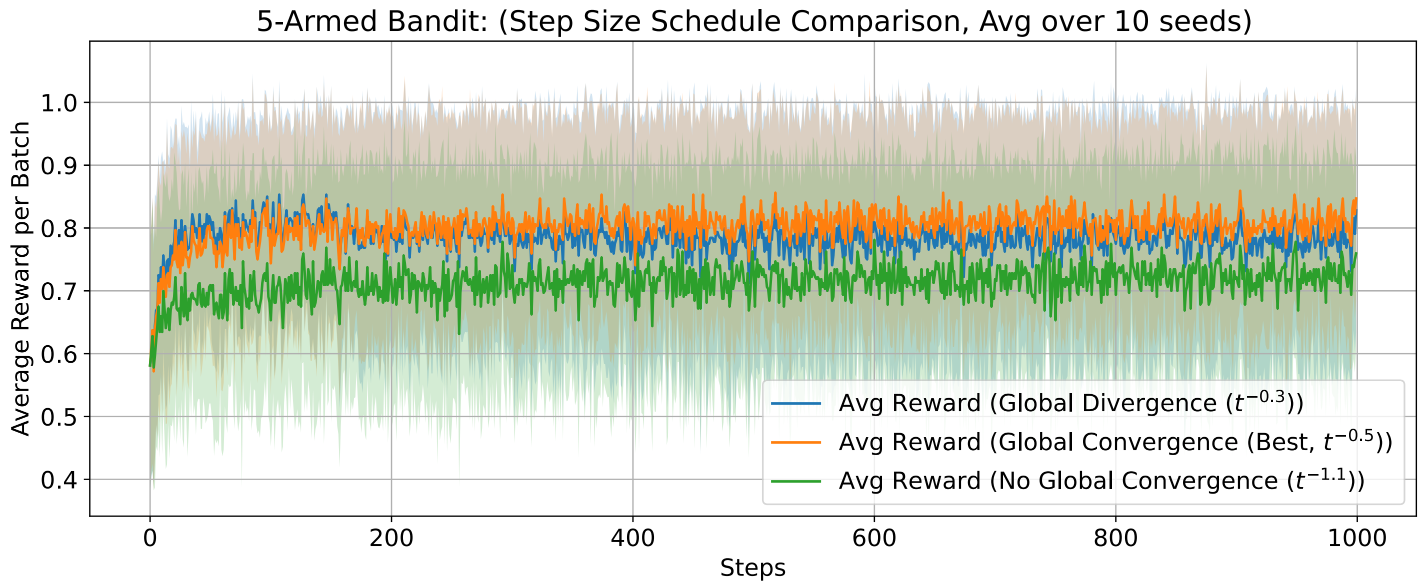
Bernoulli Bandit Env:

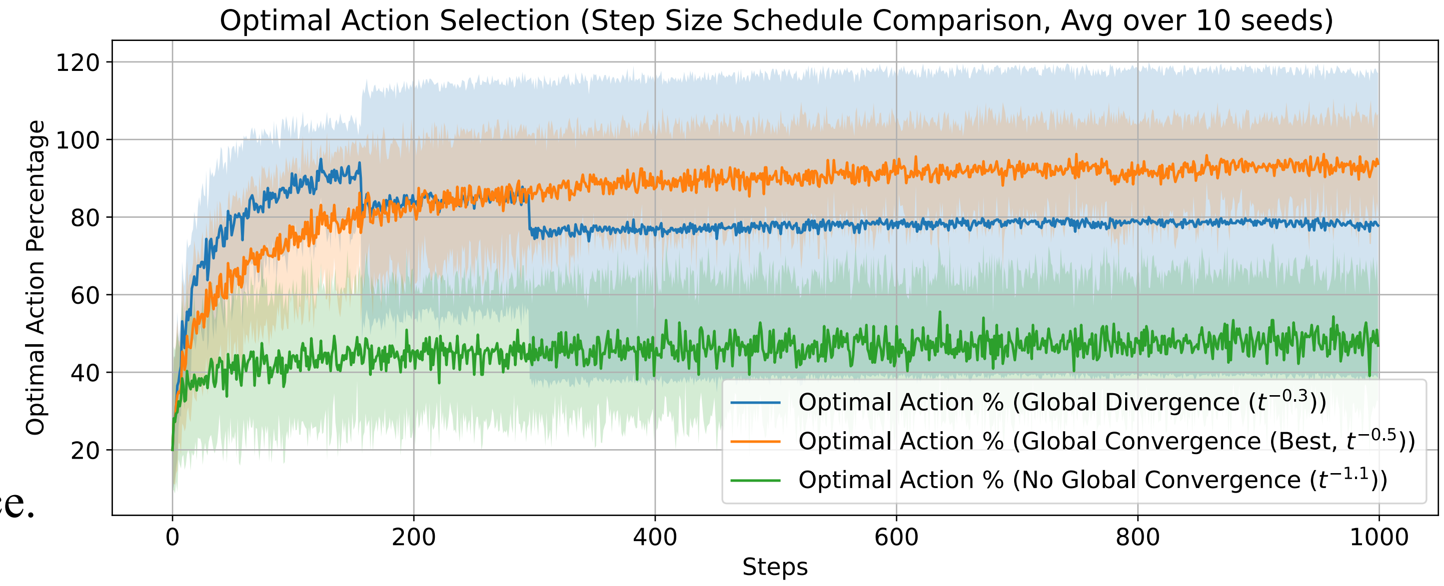
- The reward for pulling each lever follows a Bernoulli distribution. A reward of 1 means you win, and a reward of 0 means you don't win.
- Dummy states with all 1s.



Results:

$\eta_k = k^{-1.1}$ (green) leads to the failrue of convergence.
$\eta_k = k^{-0.5}$ (orange) gets the global convergence.
$\eta_k = k^{-0.3}$ (blue) causes global divergence.





action == bandit_env.optimal_arm

Theoretical Problems Unsolved

- Due to the KL penalty, the convergence of PPO depends on the L_∞ convergence $\left\|Q^{\pi_k} - \hat{Q}_T^{(k)}\right\|_\infty$ instead of the L_2 convergence, the sampling complexity may not be tight. But we DO NOT know how to prove the L_∞ convergence of the fixed point KRR because its structure is quite from classical KRR

- The "complicate" level of policy π_k is highly related to the sampling complexity. We use its RKHS norm $\|\pi_k\|_K$ to indicate its "complicate" level (we think it is better than $\|Q^{\pi_k}\|_K$), but we DO NOT know if we can have a better statistics to reflect this

Thank you for your attention!
Any questions?