Randomised Bayesian Least-Squares Policy Iteration

Nikolaos Tziortziotis[†], Christos Dimitrakakis[⋆] and Michalis Vazirgiannis[†‡]

[†] DaSciM, LIX, École Polytechnique, France

[⋆] Department of Informatics, University of Oslo, Norway

[‡] Athens University of Economics and Business, Greece

email: {ntziorzi, christos.dimitrakakis, michalis.vazirgiannis}@gmail.com

Contribution

Bayesian Least-Squares Policy Iteration:

✓ Off-policy, model-free, policy iteration algorithm

✓ Bayesian LSTD (BLSTD) for policy evaluation

✓ Quantifies our uncertainty about value function

Randomised Bayesian LSPI:

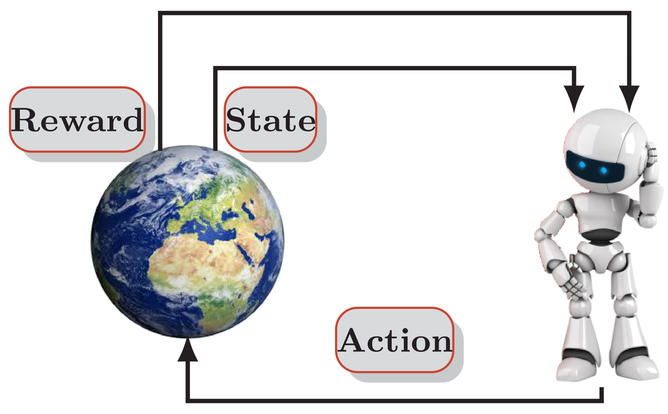
✓ Online variant of BLSPI

✓ Exploration using randomised value functions

✓ Efficient in several representative RL problems

Problem Formulation

Our objective is to design agents able to learn to act in an unknown environment, by interaction and reinforcement.



RL task is folmulated as **MDP**: $\{\mathcal{S}, \mathcal{A}, P, r, \gamma\}$

Policy π: $\mathcal{S} \to \mathcal{A}$ (map states to actions)

Value function:

$$Q^\pi(\boldsymbol{s}, a) \triangleq \mathbb{E}^\pi_\mu \left[\sum_{t=0}^{\infty} \gamma^t r(\boldsymbol{s}_t, a_t) | \boldsymbol{s}_0 = \boldsymbol{s}, a_0 = a \right]$$

Bellman operator:

$$T^\pi Q(\boldsymbol{s}, a) \triangleq r(\boldsymbol{s}, a) + \gamma \int_{\mathcal{S}} Q(\boldsymbol{s}', \pi(\boldsymbol{s}')) dP(\boldsymbol{s}' | \boldsymbol{s}, a)$$

Unique fixed-point of Bellman operator:

$$Q^\pi = T^\pi Q^\pi \Rightarrow Q^\pi = (\boldsymbol{I} - \gamma P^\pi)^{-1} \mathcal{R}$$

☹ The value function cannot be represented in an explicit way (continuous state space)

☹ The model of the MDP is unknown

☺ **Value function approximation:**

$$Q^\pi_{\boldsymbol{\theta}}(\boldsymbol{s}, a) = \boldsymbol{\phi}(\boldsymbol{s}, a)^\top \boldsymbol{\theta} = \sum_{i=1}^{k} \phi_i(\boldsymbol{s}, a) \theta_i,$$

where $\mathcal{F} = \{ f_{\boldsymbol{\theta}} | f_{\boldsymbol{\theta}}(\cdot) = \boldsymbol{\phi}(\cdot)^\top \boldsymbol{\theta} \}$.

☺ **Set of transitions:** $\mathcal{D} = \{(\boldsymbol{s}_i, a_i, r_i, \boldsymbol{s}'_i)\}_{i=1}^N$,

where we define: $\mathcal{R} = [r_i, \dots, r_N]^\top$,

$\tilde{\Phi} = [\phi(\boldsymbol{s}_1, a_1)^\top; \dots; \phi(\boldsymbol{s}_N, a_N)^\top]$, and

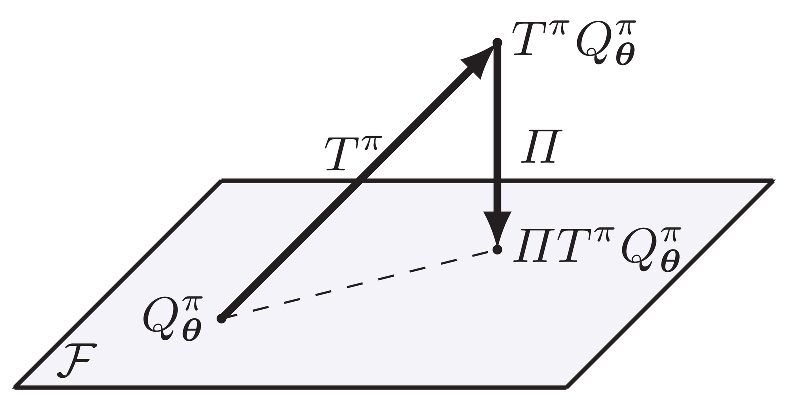
$\tilde{\Phi}' = [\phi(\boldsymbol{s}'_1, \pi(\boldsymbol{s}'_1))^\top; \dots; \phi(\boldsymbol{s}'_N, \pi(\boldsymbol{s}'_N))^\top]$.

Least-Squares Policy Iteration

- Start with an abritrary policy π_k, $k = 0$.

1. **Policy evaluation phase**

Minimise *mean-square projected Bellman error* (MSPBE): $\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^k} \| Q^\pi_{\boldsymbol{\theta}} - \Pi T^\pi Q^\pi_{\boldsymbol{\theta}} \|^2_{\Xi}$.



Nested Optimization Problem

$$\boldsymbol{u}^* = \arg\min_{\boldsymbol{u} \in \mathbb{R}^k} \| \Phi \boldsymbol{u} - T^\pi \Phi \boldsymbol{\theta} \|^2_{\Xi} \quad \textit{(Projection step)}$$

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^k} \| \Phi \boldsymbol{\theta} - \Phi \boldsymbol{u}^* \|^2_{\Xi} \quad \textit{(Fixed-point step)}$$

Solution

$$\boldsymbol{u}^* = \tilde{C}^{-1} \tilde{\Phi} (\mathcal{R} + \gamma \tilde{\Phi}' \boldsymbol{\theta}), \text{ and } \boldsymbol{\theta} = A^{-1} \boldsymbol{b},$$

where $\tilde{C} \triangleq \tilde{\Phi}^\top \tilde{\Phi}$, $A \triangleq \tilde{\Phi}^\top (\tilde{\Phi} - \gamma \tilde{\Phi}')$, and $\boldsymbol{b} \triangleq \tilde{\Phi}^\top \mathcal{R}$.

2. **Policy improvement phase**

$$\pi_{k+1} = \arg\max_{a \in \mathcal{A}} Q^{\pi_k}(\boldsymbol{s}, a)$$

⟳ Repeat until convergence

Bayesian LSPI

- Approximate policy iteration algorithm
- BLSTD-Q is used for policy evalutation

Empirical Bellman operator:

$$\hat{T}^\pi Q^\pi_{\boldsymbol{\theta}} = \mathcal{R} + \gamma P^\pi Q^\pi_{\boldsymbol{\theta}} + N, \quad N \sim \mathcal{N}(0, \beta^{-1} \boldsymbol{I})$$

Given observations $\mathcal{D}$ (LSTD-Q solution):

$$Q^\pi_{\boldsymbol{\theta}} = \hat{\Pi} \hat{T}^\pi Q^\pi_{\boldsymbol{\theta}} \Leftrightarrow \tilde{\Phi}^\top \mathcal{R} = \tilde{\Phi}^\top (\tilde{\Phi} - \gamma \tilde{\Phi}') \boldsymbol{\theta} + \tilde{\Phi}^\top N$$

Linear regression model: $\boldsymbol{b} = A \boldsymbol{\theta} + \tilde{\Phi}^\top N$

Likelihood: $p(\boldsymbol{b} | \boldsymbol{\theta}, \beta) = \mathcal{N}(\boldsymbol{b} | A \boldsymbol{\theta}, \beta^{-1} \tilde{C})$

✓ ML inference corresponds to standard LSTD

Prior distribution over $\boldsymbol{\theta}$: $p(\boldsymbol{\theta} | \alpha) = \mathcal{N}(\boldsymbol{\theta} | 0, \alpha^{-1} \boldsymbol{I})$.

Logarithm of posterior distribution:

$$\ln p(\boldsymbol{\theta} | \mathcal{D}) \propto -\frac{\beta}{2} E_{\mathcal{D}}(\boldsymbol{\theta}) - \frac{\alpha}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta},$$

where $E_{\mathcal{D}}(\boldsymbol{\theta}) = (\boldsymbol{b} - A\boldsymbol{\theta})^\top \tilde{C}^{-1} (\boldsymbol{b} - A\boldsymbol{\theta})$ (MSPBE).

✓ MAP inference refers to ℓ_2-reg. $(\lambda = \alpha / \beta)$

Posterior distribution: $p(\boldsymbol{\theta} | \mathcal{D}) = \mathcal{N}(\boldsymbol{\theta} | \boldsymbol{m}, S)$

$S \triangleq (\alpha \boldsymbol{I} + \beta A^\top \tilde{C}^{-1} A)^{-1}$ and $\boldsymbol{m} \triangleq \beta S A^\top \tilde{C}^{-1} \boldsymbol{b}$.

Predictive distribution:

$$p(Q^\pi_{\boldsymbol{\theta}}(\boldsymbol{s}^*, a^*) | \mathcal{D}) = \int_{\boldsymbol{\theta}} p(Q^\pi_{\boldsymbol{\theta}}(\boldsymbol{s}^*, a) | \boldsymbol{\theta}) dp(\boldsymbol{\theta} | \boldsymbol{b}, \alpha, \beta)$$

$$= \mathcal{N}(Q^\pi_{\boldsymbol{\theta}}(\boldsymbol{s}^*, a^*) | \boldsymbol{\phi}(\boldsymbol{s}^*, a^*)^\top \boldsymbol{m}, \boldsymbol{\phi}(\boldsymbol{s}^*, a^*)^\top S \boldsymbol{\phi}(\boldsymbol{s}^*, a^*)).$$

Randomised Bayesian LSPI

- Online, offpolicy, policy iteration algorithm
- RBLSPI demostrates efficient exploration
- Exploits the advantage of BLSTD to quantify our uncertainty about the value function
- A value function is sampled right after each policy evaluation step
- Actions are selected greedily based on the sampled value function thereafter

RBLSPI Algorithm

Input: Basis ϕ, γ, α, β, K

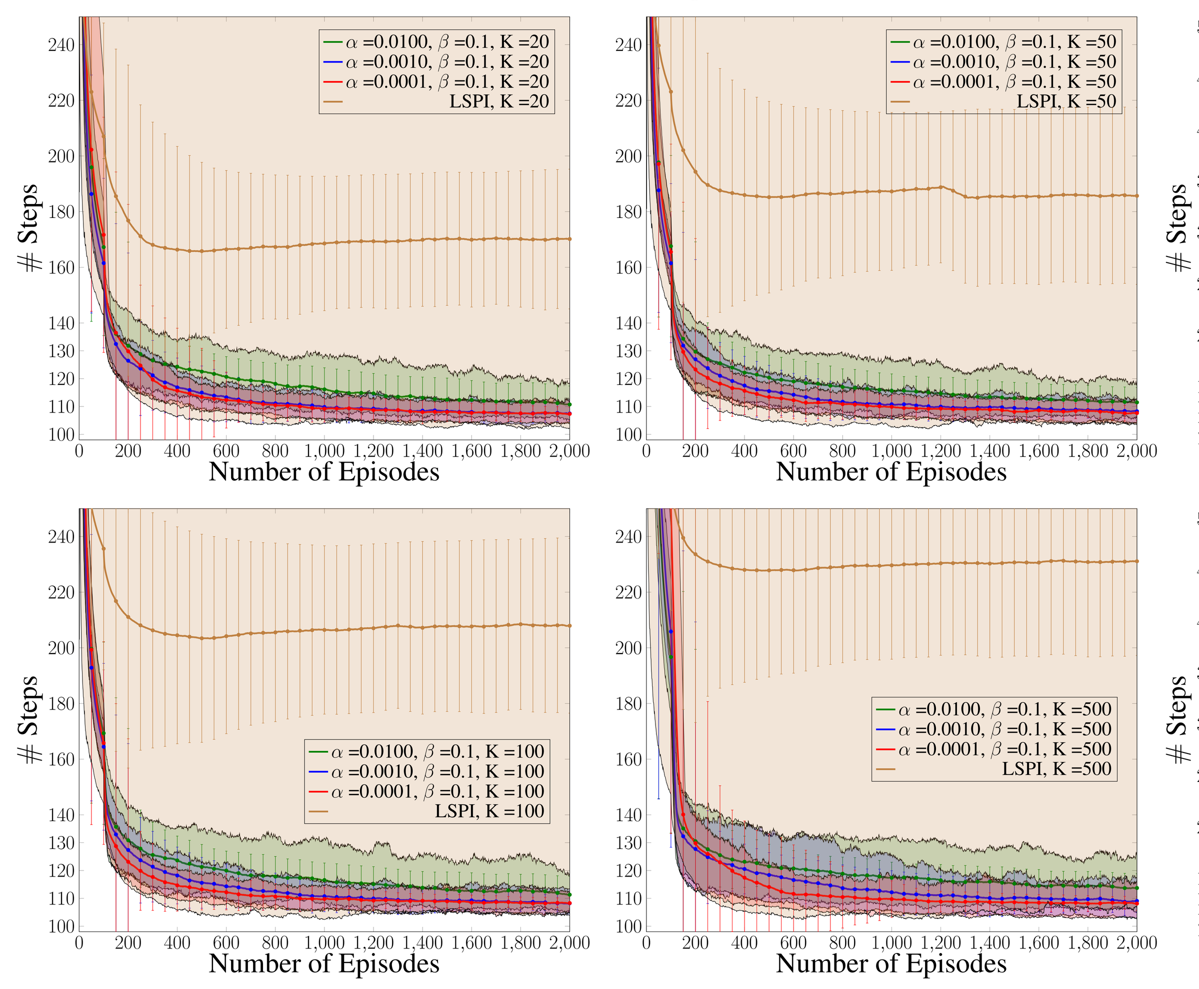
Init: $A \leftarrow \mathbf{0}_{k,k}$, $\tilde{C} \leftarrow \mathbf{0}_{k,k}$, $\boldsymbol{b} \leftarrow \mathbf{0}_k$, $\boldsymbol{m} \sim \mathcal{N}(\mathbf{0}_k, \boldsymbol{I})$, $\tilde{\boldsymbol{\theta}} = \boldsymbol{m}$

```
Observe s_0
for t = 0, ... do
    /* Behavior policy*/
    a_t ∈ arg max_{a∈A} φ(s_t, a)^T θ̃
    Apply action a_t
    Observe r_t, s_{t+1}
    a* = arg max_{a∈A} φ(s_{t+1}, a)^T m
    A ← A + φ(s_t, a_t)(φ(s_t, a_t) − γφ(s_{t+1}, a*))^T
    C̃ ← C̃ + φ(s_t, a_t)φ(s_t, a_t)^T
    b ← b + φ(s_t, a_t) r_t
    if (t mod K == 0) then
        S = (αI + βA^T C̃^{-1} A)^{-1}
        m = βSA^T C̃^{-1} b
        /* Sample posterior distribution */
        θ̃ ~ N(m, S)
    end if
end for
```

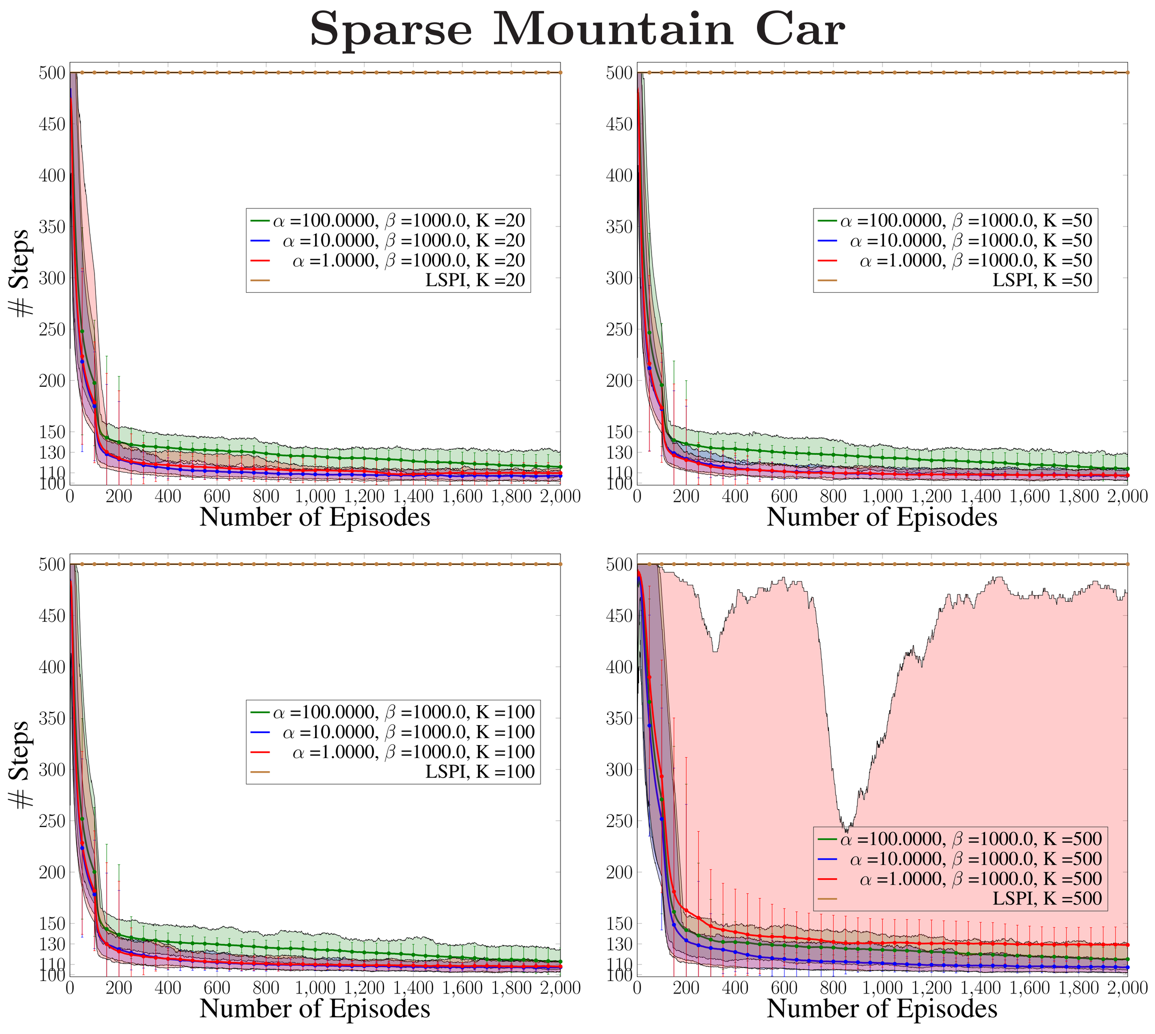
Empirical Results

- Randomised BLSPI vs. online LSPI using ϵ−greedy exploration strategy (ϵ tuned)
- Environments: **i)** (Sparse reward) Mountain Car, **ii)** Cart Pole, **iii)** Pendulum, **iv)** PuddleWorld

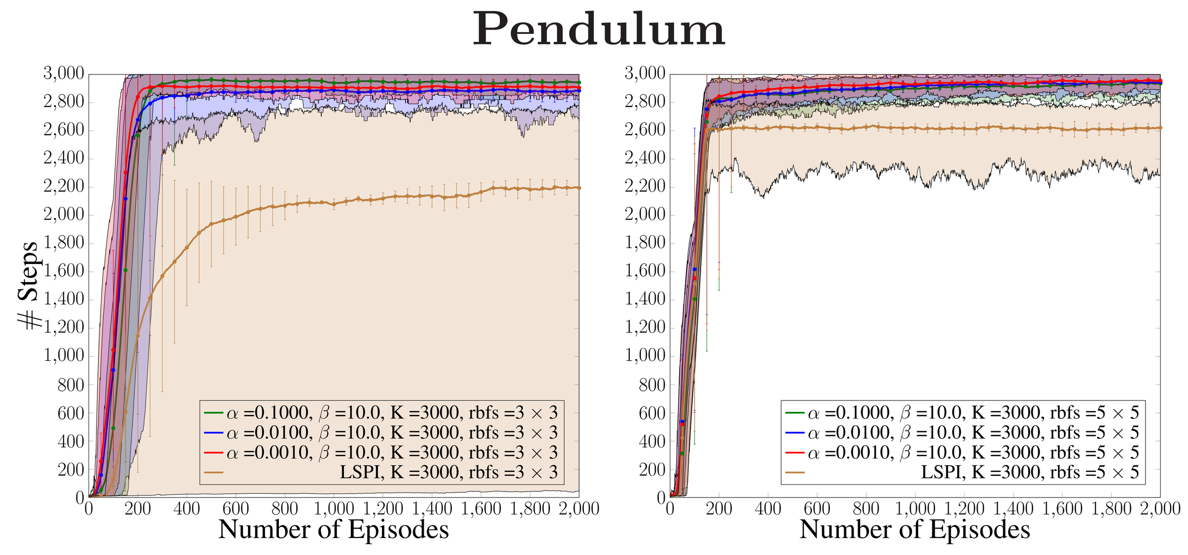
Mountain Car



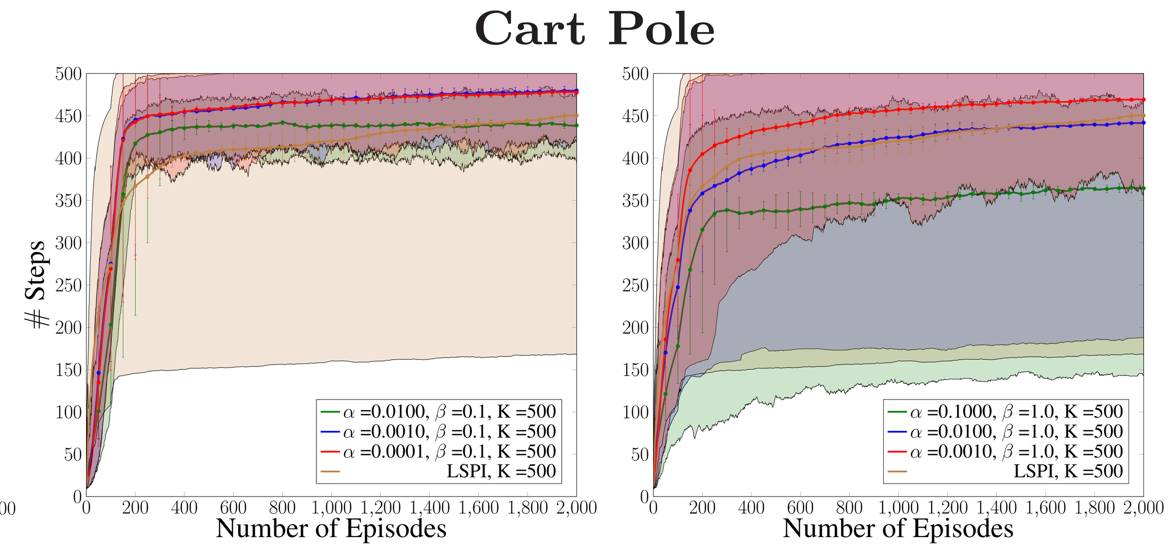
Sparse Mountain Car



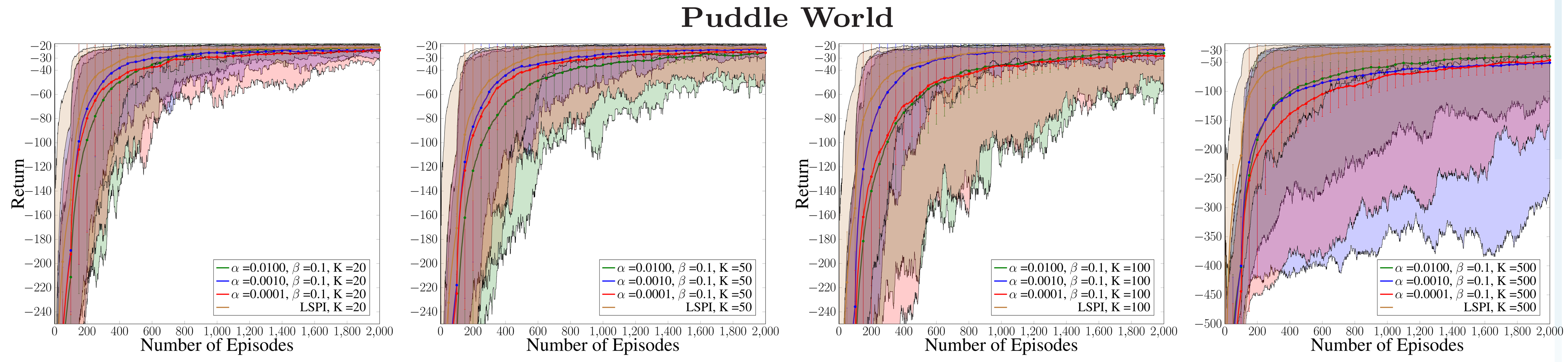
Pendulum



Cart Pole



Puddle World



The mean performance (average return or number of steps over 100 episodes) across 100 independent runs is presented. The error bars show 95% confidence intervals, while the shaded regions show 90% percentiles over 100 runs