Operator Splitting Value Iteration

Amin Rakhsha[1,2]

Andrew Wang[1,2]

Mohammad Ghavamzadeh[3]

Amir-massoud Farahmand[2,1]

[1]University of Toronto [2]Vector Institute [3]Google Research

Problem Definition

P

True Dynamics
(Expensive)

$\hat{P}$

Inaccurate
(No Cost)

?

V^{π} , V^{*}

Value Function

Given:

- P: True dynamics, but expensive to query
- $\hat{P}$: Inaccurate dynamics, but no cost to query

How should we learn the value function V^{π} or V^{*}?

Problem Definition

P

True Dynamics
(Expensive)

$\hat{P}$

Inaccurate
(No Cost)

?

V^{π}, V^{*}

Value Function

Environment	Learned Model	Model-based RL
Accurate Simulator	Approximate Simulator	Robotics

Existing Methods

P

True Dynamics
(Expensive)

$\hat{P}$

Inaccurate
(No Cost)

?

V^{π}, V^{*}

Value Function

1. Value Iteration using P (Model-free algorithms)
   - Accurate, but lots of queries to P, slow rate (Error $\approx \gamma^{k}$)
2. Value Iteration using $\hat{P}$ (Dyna architecture)
   - No query cost to P, but biased value due to model error

Our Contribution

Operator Splitting Value Iteration (OS-VI)

- Fewer queries to P, rate $\propto$ Model Error

$$\text{Error} \approx \left(\frac{\gamma}{1-\gamma} \times \text{Model Error} \right)^{k}$$

This can be significantly faster than VI!

- No bias due to model error

Operator Splitting Dyna (OS-Dyna)

- Sample-based version of OS-VI

Method

- $V_{k+1} \leftarrow$ Solution of an auxiliary MDP $(\hat{P}, \bar{r}_k)$.
- Auxiliary MDP at iteration k:

Dynamics

$\hat{P}$

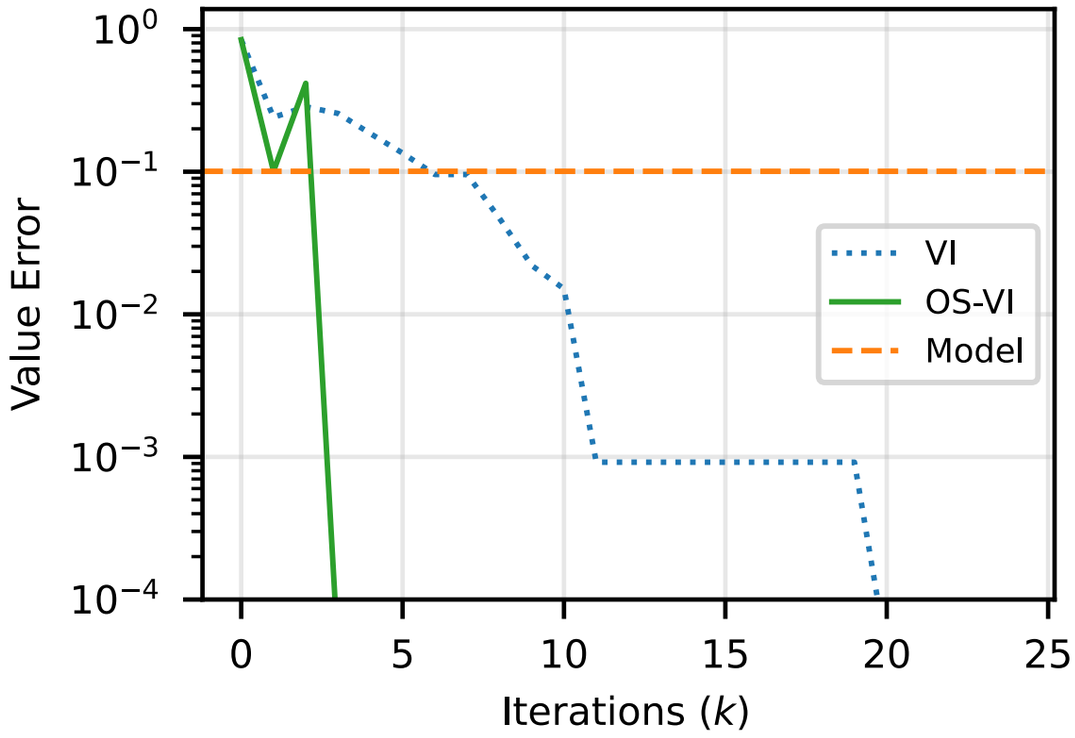
Rewards

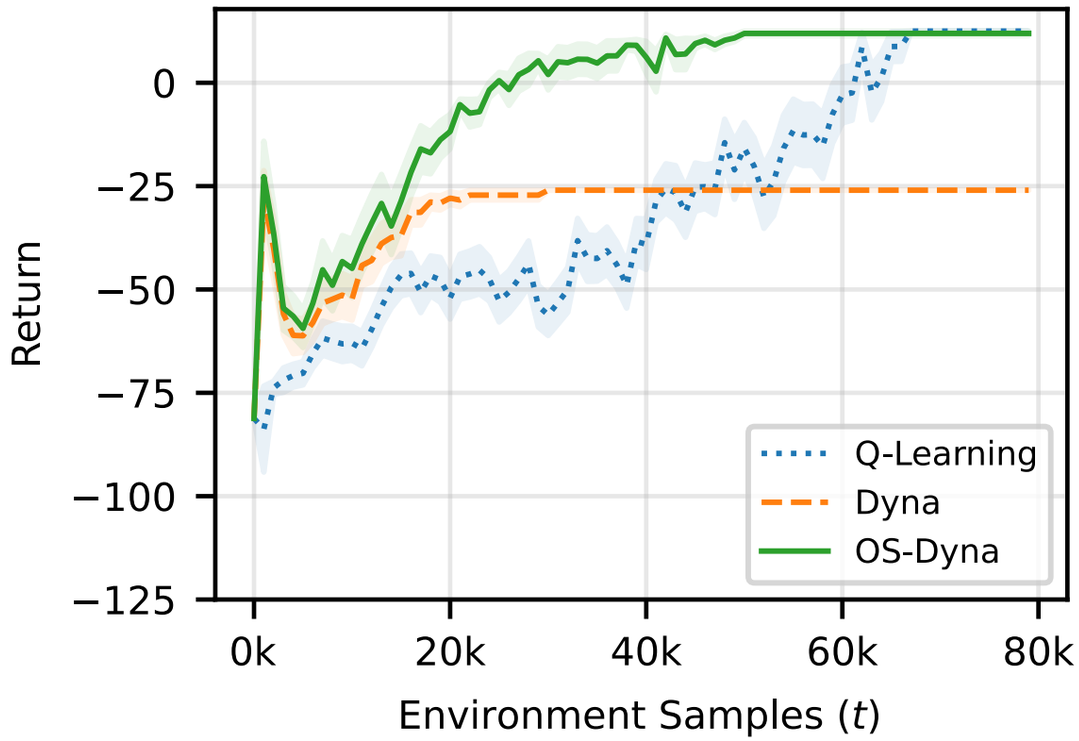
$R + \gamma P V_k - \gamma \hat{P} V_k$

Original Reward

Error Correction

Results





Faster convergence to the true value function!