A Temporal-Difference Approach to Policy Gradient Estimation

Samuele Tosatto, Andrew Patterson, Martha White, A. Rupam Mahmood

International Conference Of Machine Learning 2022

Classic Policy Gradients

Classic Policy Gradients

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t Q^\pi(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

Classic Policy Gradients

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \underbrace{Q^\pi(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t | s_t)}\right]$$

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{g}(s_t, a_t)\right]$$

Classic **Policy Gradients**

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t Q^\pi(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t | s_t)\right]$$

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{g}(s_t, a_t)\right]$$

$\mathbf{g}(s_0, a_0)$

Classic Policy Gradients

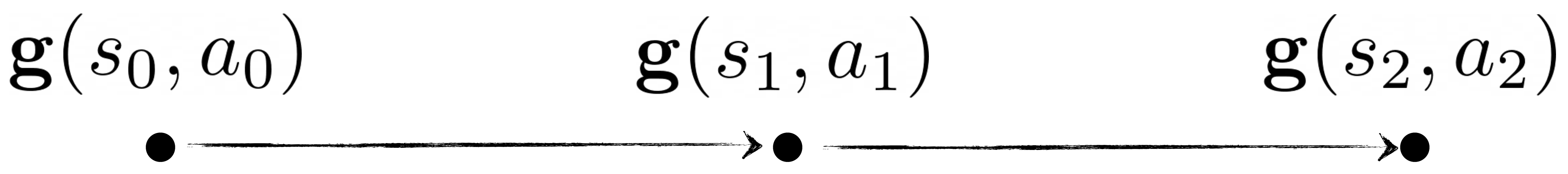
$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t Q^\pi(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{g}(s_t, a_t)\right]$$

Classic Policy Gradients

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t Q^\pi(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{g}(s_t, a_t)\right]$$



Classic Policy Gradients

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t Q^\pi(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{g}(s_t, a_t)\right]$$

Classic Policy Gradients

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \underbrace{Q^\pi(s_t, a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)}\right]$$

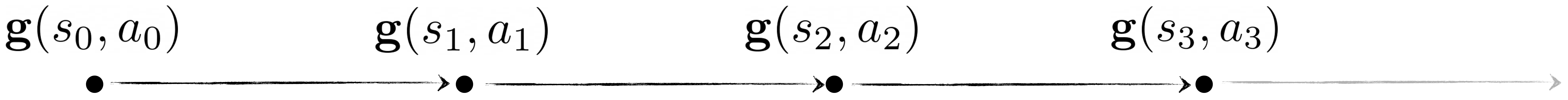
$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{g}(s_t, a_t)\right]$$

$\mathbf{g}(s_0, a_0)$ → $\mathbf{g}(s_1, a_1)$ → $\mathbf{g}(s_2, a_2)$ → $\mathbf{g}(s_3, a_3)$ →

Classic Policy Gradients are Monte-Carlo Estimators

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t Q^\pi(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{g}(s_t, a_t)\right]$$



Classic Policy Gradients are Monte-Carlo Estimators

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \underbrace{Q^\pi(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t | s_t)}\right]$$

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{g}(s_t, a_t)\right]$$

$\mathbf{g}(s_0, a_0) \rightarrow \mathbf{g}(s_1, a_1) \rightarrow \mathbf{g}(s_2, a_2) \rightarrow \mathbf{g}(s_3, a_3) \rightarrow$

High Variance

Classic Off-Policy **Policy Gradients** are biased

Classic policy gradients are biased if importance sampling is not done correctly

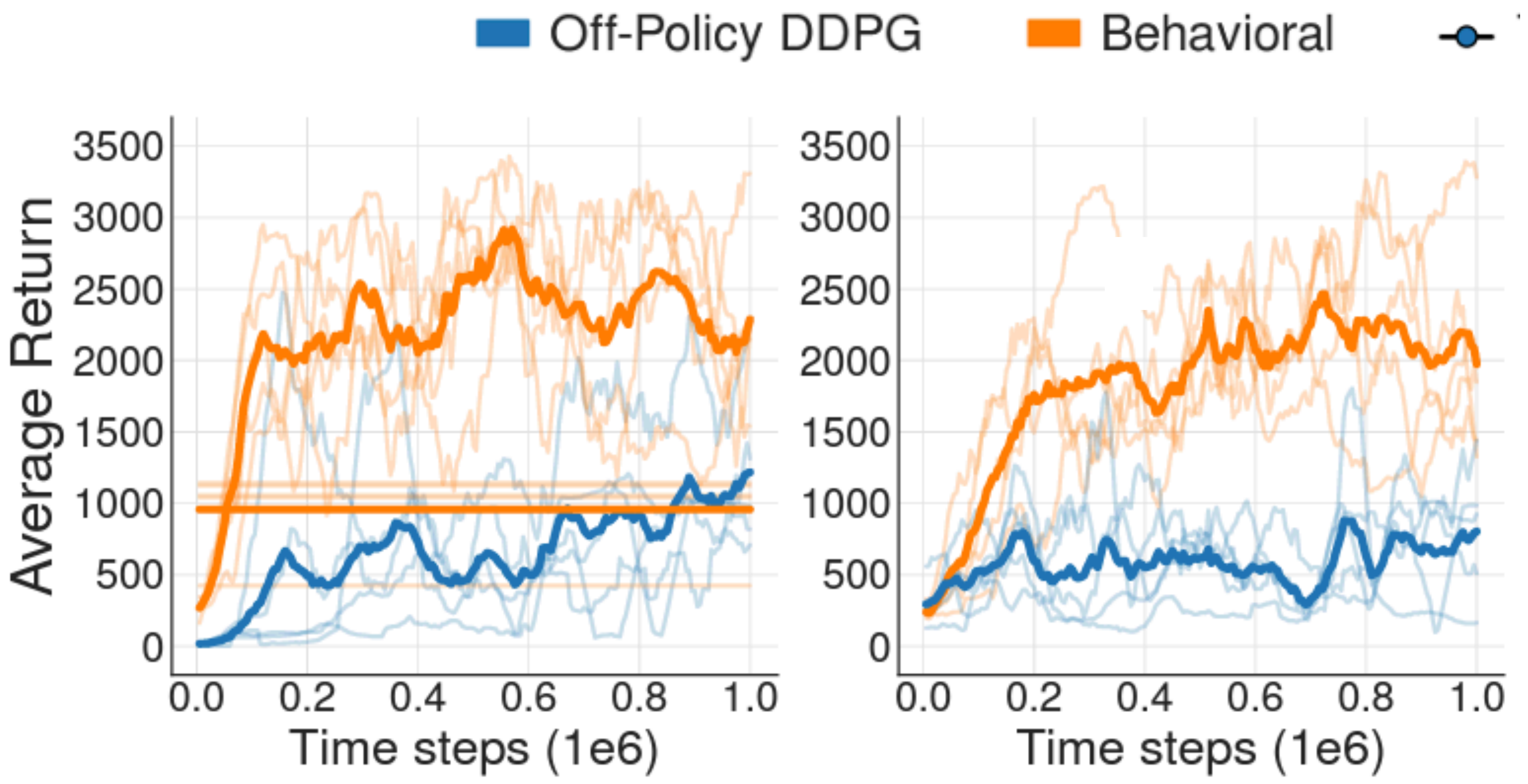
Classic Off-Policy **Policy Gradients** are biased

Classic policy gradients are biased if importance sampling is not done correctly

OffPAC, DDPG, SAC, …

Classic Off-Policy Policy Gradients are biased

Classic policy gradients are biased if importance sampling is not done correctly



(a) Final buffer performance

(b) Concurrent performance

OffPAC, DDPG, SAC, …

"off-policy deep reinforcement learning algorithms are ineffective when learning truly off-policy"
Fujimoto et al. 2019

Main Contribution

Main Contribution

Gradient Bellman Equation:

$$\mathbf{\Gamma}^{\pi}(s,a) = \mathbb{E}_{s',a'}\left[\gamma \mathbf{g}(s',a') + \gamma \mathbf{\Gamma}^{\pi}(s',a')\right]$$

Main Contribution

Gradient Bellman Equation:

$$\mathbf{\Gamma}^{\pi}(s,a) = \mathbb{E}_{s',a'}\left[\gamma \mathbf{g}(s',a') + \gamma \mathbf{\Gamma}^{\pi}(s',a')\right]$$

Gradient Critic

Main Contribution

Gradient Bellman Equation:

$$\boldsymbol{\Gamma}^{\pi}(s,a) = \mathbb{E}_{s',a'}\left[\gamma \mathbf{g}(s',a') + \gamma \boldsymbol{\Gamma}^{\pi}(s',a')\right]$$

Gradient Critic

$$\Gamma^{\pi}(s,a) =$$

Main Contribution

Gradient Bellman Equation:

$$\boldsymbol{\Gamma}^{\pi}(s,a) = \mathbb{E}_{s',a'}\left[\gamma \mathbf{g}(s',a') + \gamma \boldsymbol{\Gamma}^{\pi}(s',a')\right]$$

Gradient Critic

$$\Gamma^{\pi}(s,a) = \quad \mathbf{g}(s_1, a_1)$$

Main Contribution

Gradient Bellman Equation:

$$\mathbf{\Gamma}^{\pi}(s,a) = \mathbb{E}_{s',a'}\left[\gamma \mathbf{g}(s',a') + \gamma \mathbf{\Gamma}^{\pi}(s',a')\right]$$

Gradient Critic

$$\Gamma^{\pi}(s,a) =$$

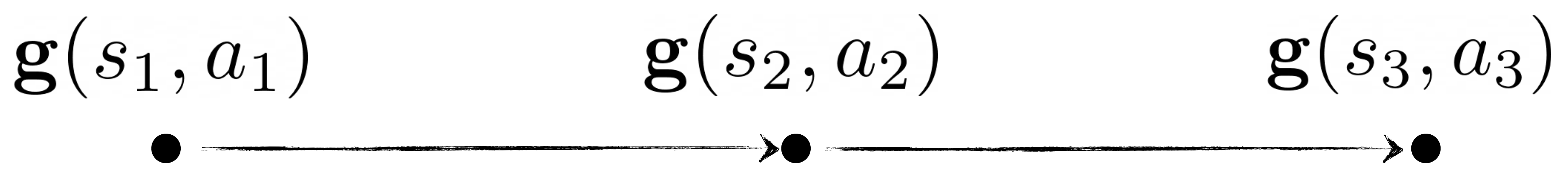
Main Contribution

Gradient Bellman Equation:

$$\boldsymbol{\Gamma}^{\pi}(s,a) = \mathbb{E}_{s',a'}\left[\gamma \mathbf{g}(s',a') + \gamma \boldsymbol{\Gamma}^{\pi}(s',a')\right]$$

Gradient Critic

$$\Gamma^{\pi}(s,a) =$$



Main Contribution

Gradient Bellman Equation:

$$\boldsymbol{\Gamma}^{\pi}(s,a) = \mathbb{E}_{s',a'}\left[\gamma \mathbf{g}(s',a') + \gamma \boldsymbol{\Gamma}^{\pi}(s',a')\right]$$

Gradient Critic

$$\Gamma^{\pi}(s,a) =$$

Main Contribution

Gradient B

$\Gamma^{\pi}(s,a) =$

Use Temporal-Difference!

Gradient Temporal-Difference Learning
with Regularized Corrections
Ghiassan et al. 2020

$\Gamma^{\pi}(s,a) =$

Policy Gradient with a Gradient Critic

Policy Gradient with a Gradient Critic

Low λ $(\lambda \to 0)$ Full use of our Gradient Critic = Bootstrapping

Policy Gradient with a Gradient Critic

Low λ $(\lambda \to 0)$ Full use of our Gradient Critic = Bootstrapping

$$\nabla_\theta J(\theta) = \mathbb{E}_{s_0}\left[\mathbf{g}(s_0, a_0^\pi) + \mathbf{\Gamma}^\pi(s_0, a_0^\pi)\right]$$

Policy Gradient with a Gradient Critic

Low λ $(\lambda \to 0)$ Full use of our Gradient Critic = Bootstrapping

$$\nabla_\theta J(\theta) = \mathbb{E}_{s_0}[\mathbf{g}(s_0, a_0^\pi) + \mathbf{\Gamma}^\pi(s_0, a_0^\pi)]$$

Starting State Distribution

Policy Gradient with a Gradient Critic

Low λ $(\lambda \to 0)$ Full use of our Gradient Critic = Bootstrapping

Low Variance

$$\nabla_\theta J(\theta) = \mathbb{E}_{s_0}[\mathbf{g}(s_0, a_0^\pi) + \mathbf{\Gamma}^\pi(s_0, a_0^\pi)]$$

↑ Starting State Distribution

Policy Gradient with a Gradient Critic

Low λ $(\lambda \rightarrow 0)$ Full use of our Gradient Critic = Bootstrapping

$$\nabla_\theta J(\theta) = \mathbb{E}_{s_0}[\mathbf{g}(s_0, a_0^\pi) + \mathbf{\Gamma}^\pi(s_0, a_0^\pi)]$$

Starting State Distribution

Low Variance

Unbiased in some cases

Policy Gradient with a Gradient Critic

Low $\lambda \quad (\lambda \to 0)$ Full use of our Gradient Critic = Bootstrapping

$$\nabla_\theta J(\theta) = \mathbb{E}_{s_0}[\mathbf{g}(s_0, a_0^\pi) + \mathbf{\Gamma}^\pi(s_0, a_0^\pi)]$$

Starting State Distribution

Low Variance

Unbiased in some cases

High $\lambda \quad (\lambda \to 1)$ Classic Gradient = Monte-Carlo

Policy Gradient with a Gradient Critic

Low λ $(\lambda \to 0)$ Full use of our Gradient Critic = Bootstrapping

$$\nabla_\theta J(\theta) = \mathbb{E}_{s_0}[\mathbf{g}(s_0, a_0^\pi) + \boldsymbol{\Gamma}^\pi(s_0, a_0^\pi)]$$

Starting State Distribution

Low Variance

Unbiased in some cases

High λ $(\lambda \to 1)$ Classic Gradient = Monte-Carlo

$$\nabla_\theta J^\pi(\theta) = \mathbb{E}_\beta\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{g}(s_t, a_t) \frac{\pi(a_t|s_t)}{\beta(a_t|s_t)}\right]$$

Policy Gradient with a Gradient Critic

Low $\lambda \quad (\lambda \to 0)$ Full use of our Gradient Critic = Bootstrapping

$$\nabla_\theta J(\theta) = \mathbb{E}_{s_0}[\mathbf{g}(s_0, a_0^\pi) + \mathbf{\Gamma}^\pi(s_0, a_0^\pi)]$$

Starting State Distribution

Low Variance

Unbiased in some cases

High $\lambda \quad (\lambda \to 1)$ Classic Gradient = Monte-Carlo

$$\nabla_\theta J^\pi(\theta) = \mathbb{E}_\beta\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{g}(s_t, a_t) \frac{\pi(a_t|s_t)}{\beta(a_t|s_t)}\right]$$

High Variance

Policy Gradient with a Gradient Critic

Low λ $(\lambda \to 0)$ Full use of our Gradient Critic = Bootstrapping

$$\nabla_\theta J(\theta) = \mathbb{E}_{s_0}\left[\mathbf{g}(s_0, a_0^\pi) + \mathbf{\Gamma}^\pi(s_0, a_0^\pi)\right]$$

Starting State Distribution

Low Variance

Unbiased in some cases

High λ $(\lambda \to 1)$ Classic Gradient = Monte-Carlo

$$\nabla_\theta J^\pi(\theta) = \mathbb{E}_\beta\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{g}(s_t, a_t) \frac{\pi(a_t|s_t)}{\beta(a_t|s_t)}\right]$$

High Variance

Usually Biased!

Policy Gradient with a Gradient Critic

Low λ ... ping

... ariance

$\nabla_\theta J$ ... sed in some cases

St ...

High λ

We can obtain an unbiased estimator without using importance sampling

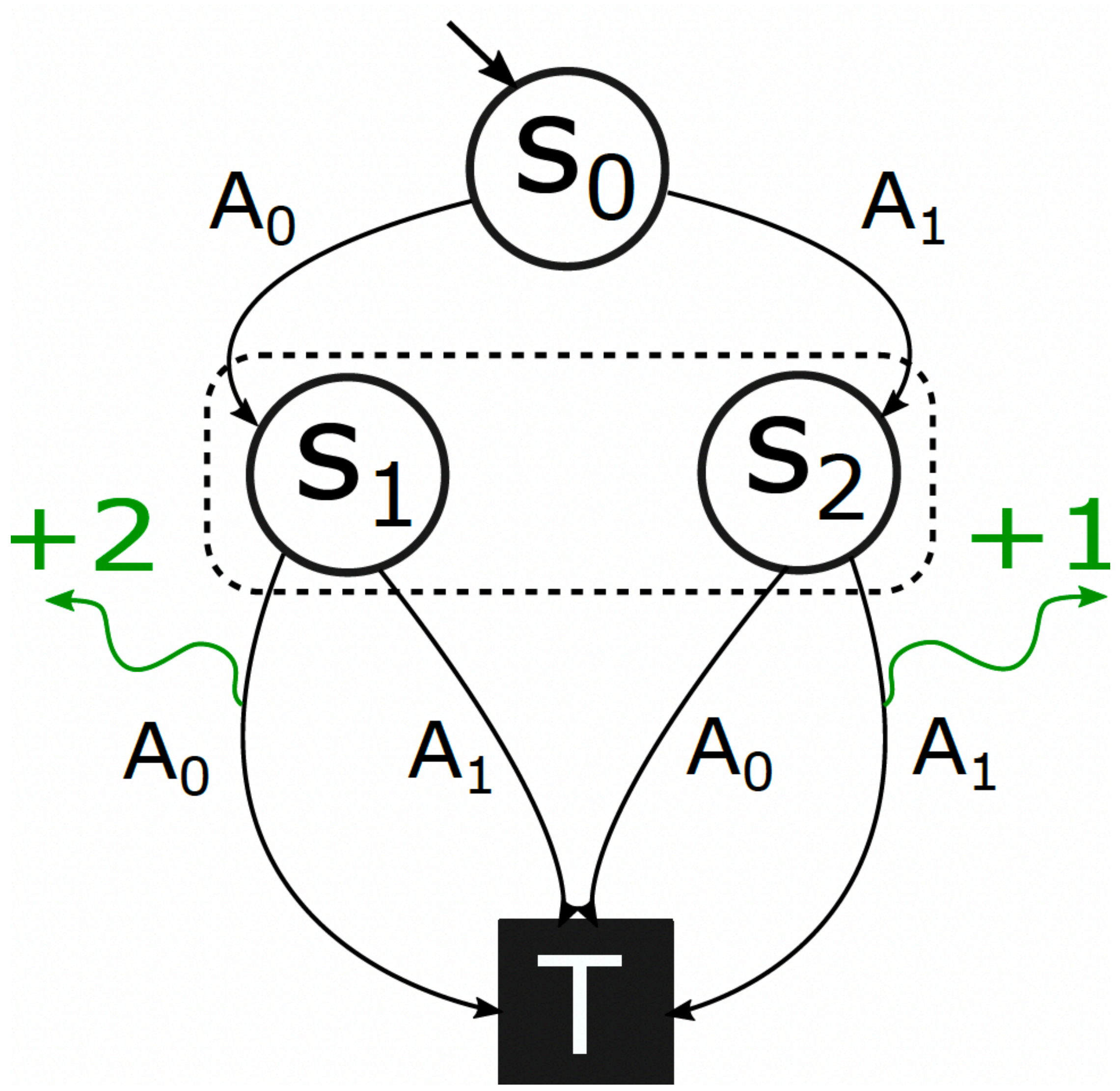
$$\nabla_\theta J^\pi(\theta) = \mathbb{E}_\beta \left[\sum_{t=0}^{\infty} \gamma^t \mathbf{g}(s_t, a_t) \frac{\pi(a_t|s_t)}{\beta(a_t|s_t)} \right]$$

High Variance

Usually Biased!

Empirical Analysis

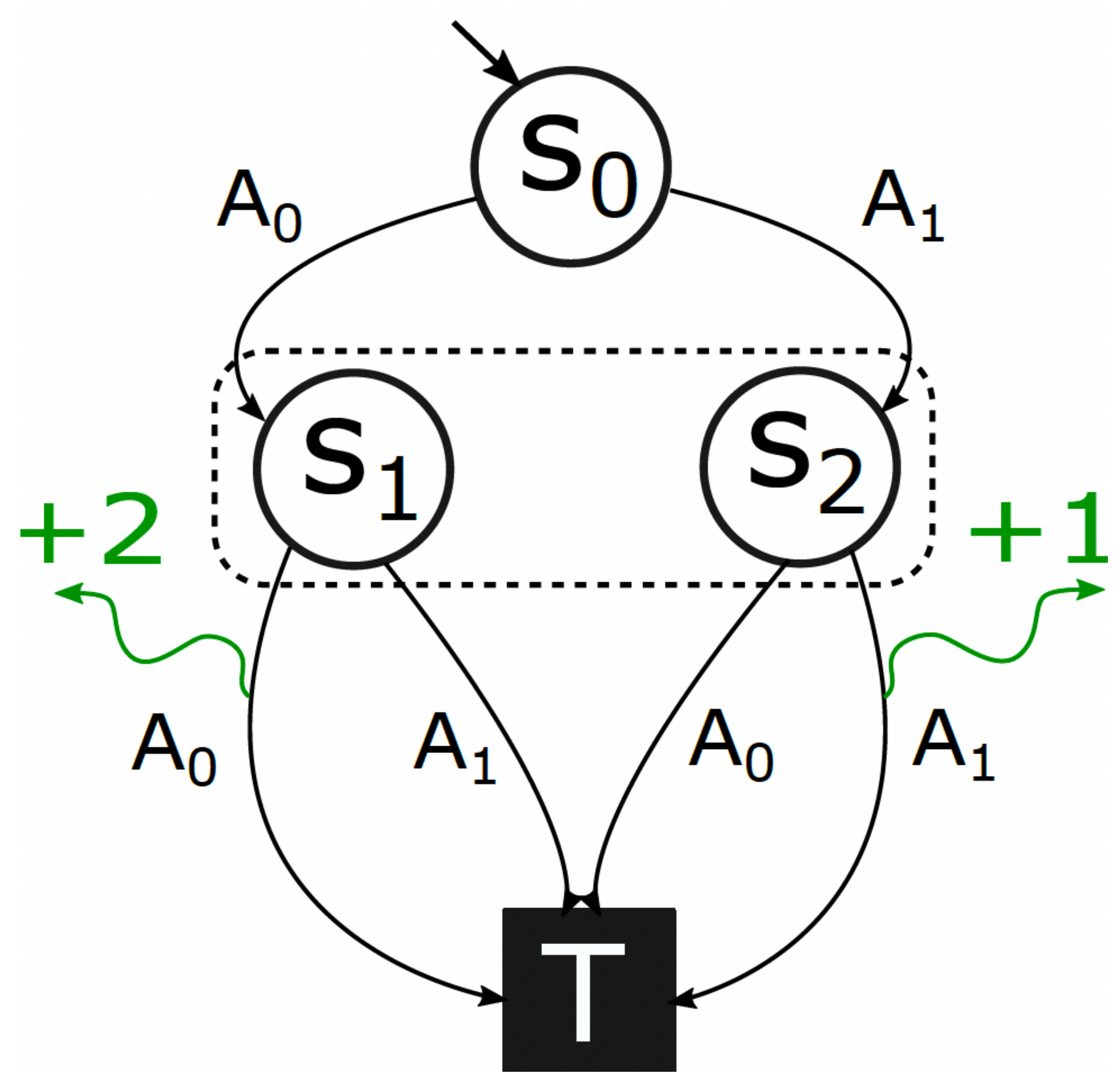
Empirical Analysis



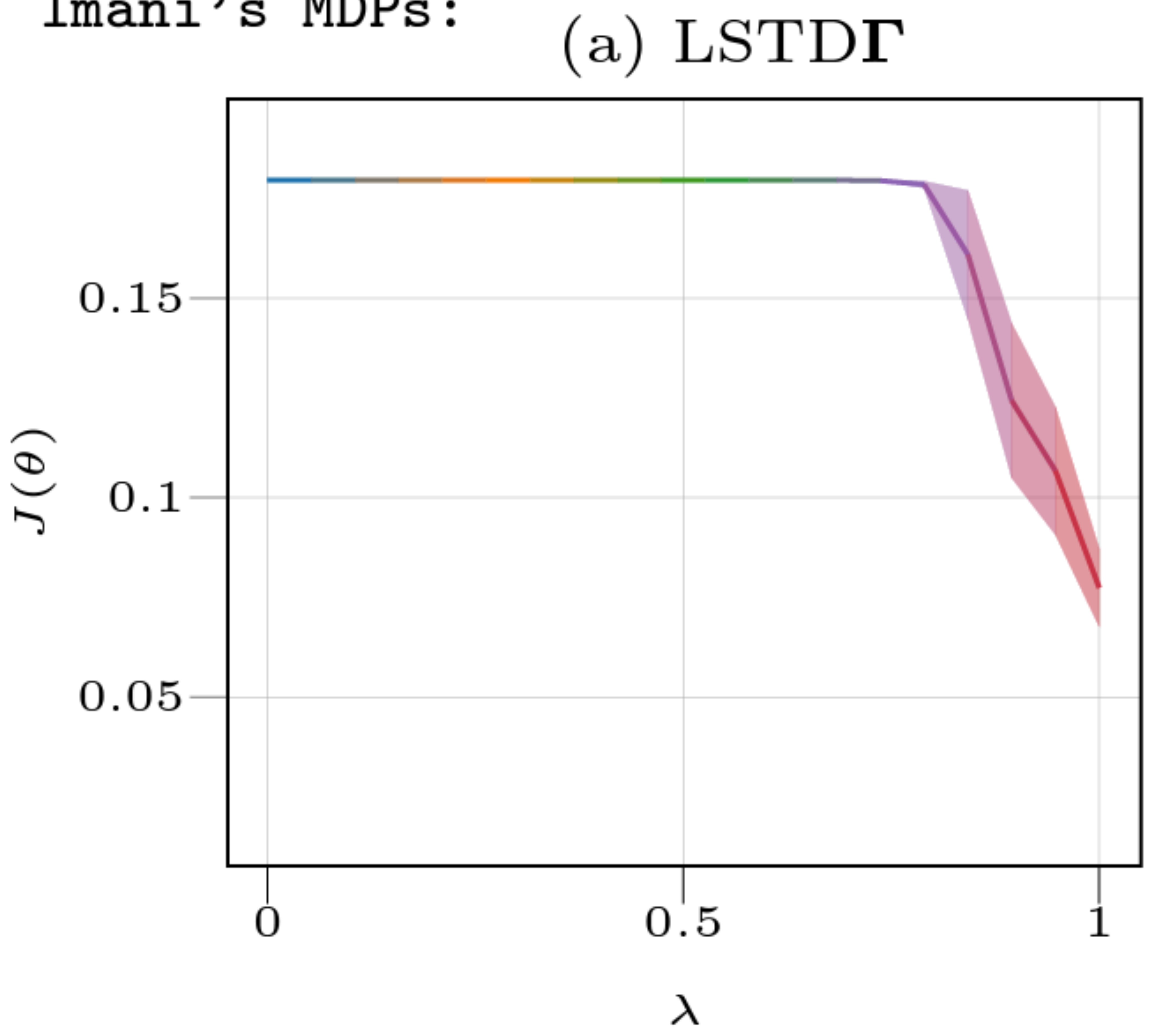
Imani et al. 2018

Empirical Analysis

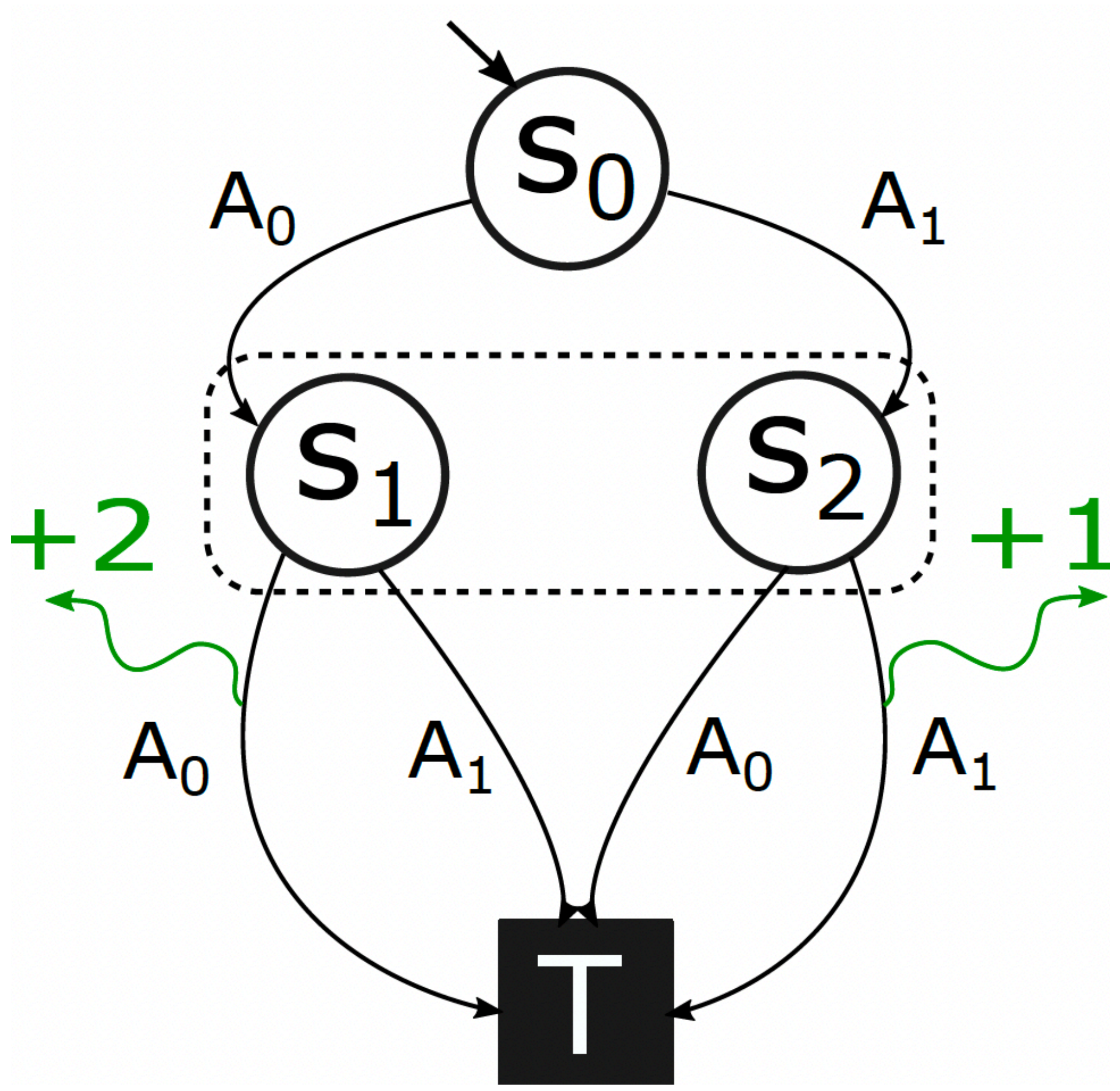
Imani's MDPs:



Imani et al. 2018

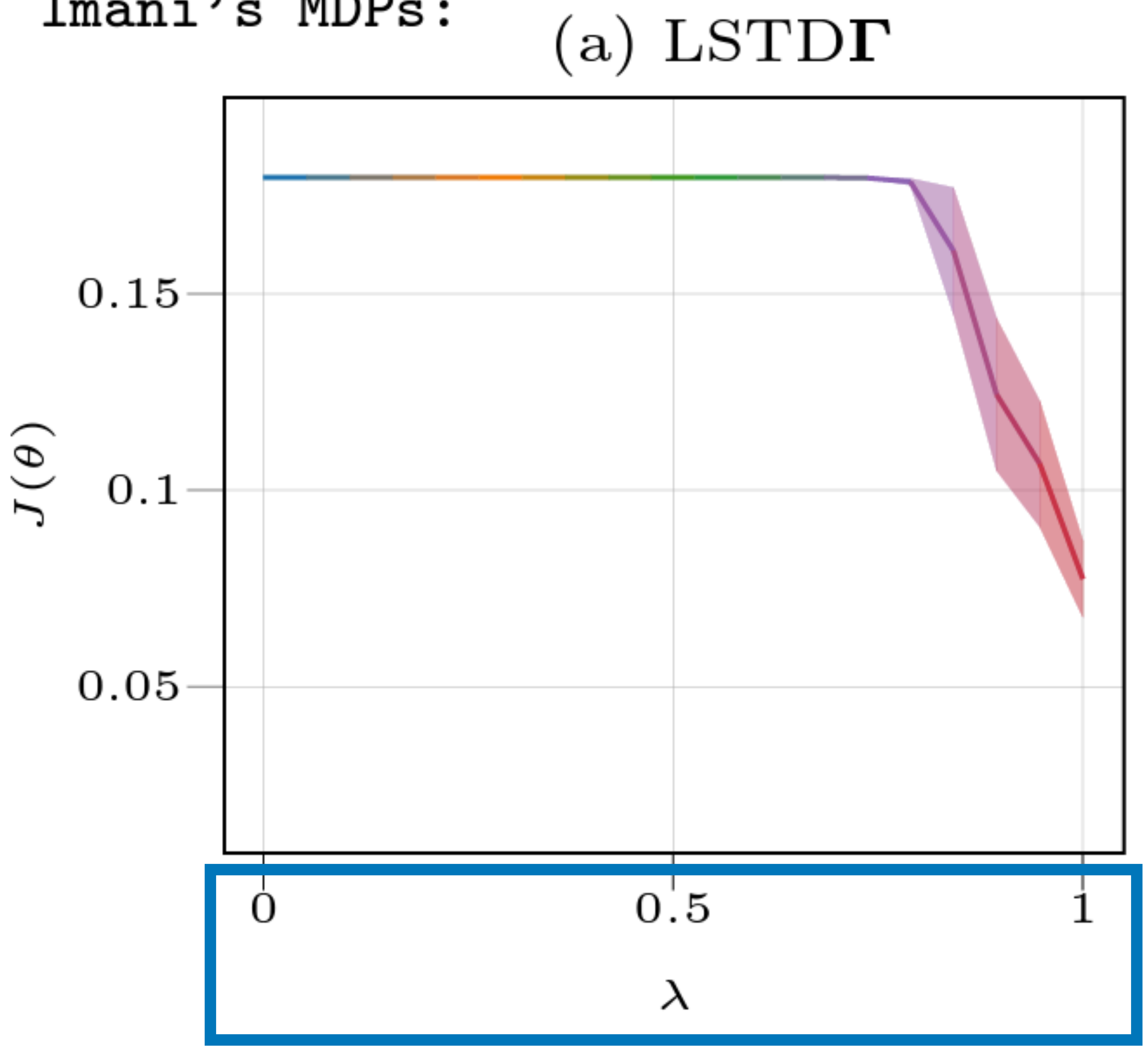


Empirical Analysis

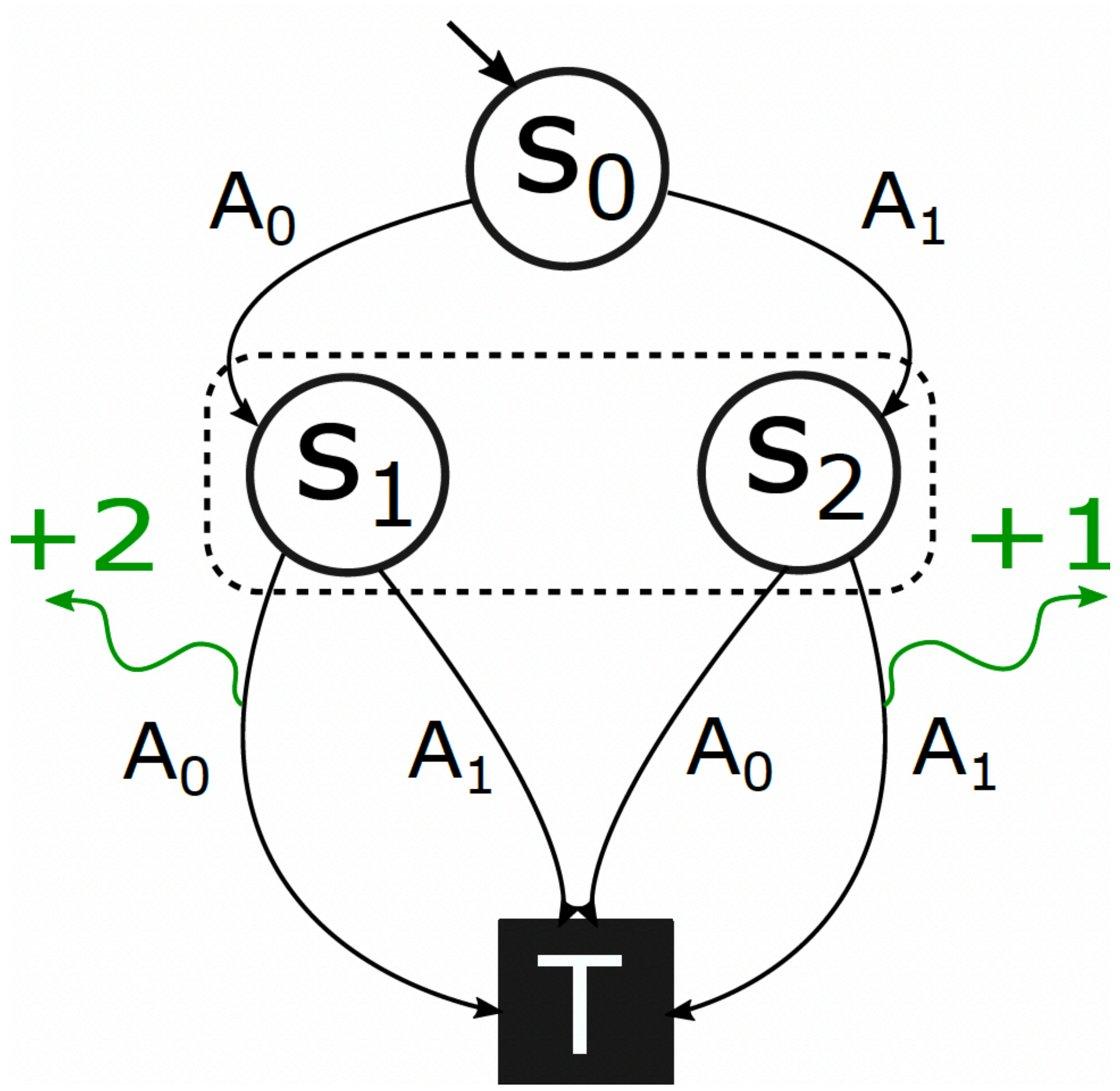


Imani et al. 2018

Imani's MDPs:

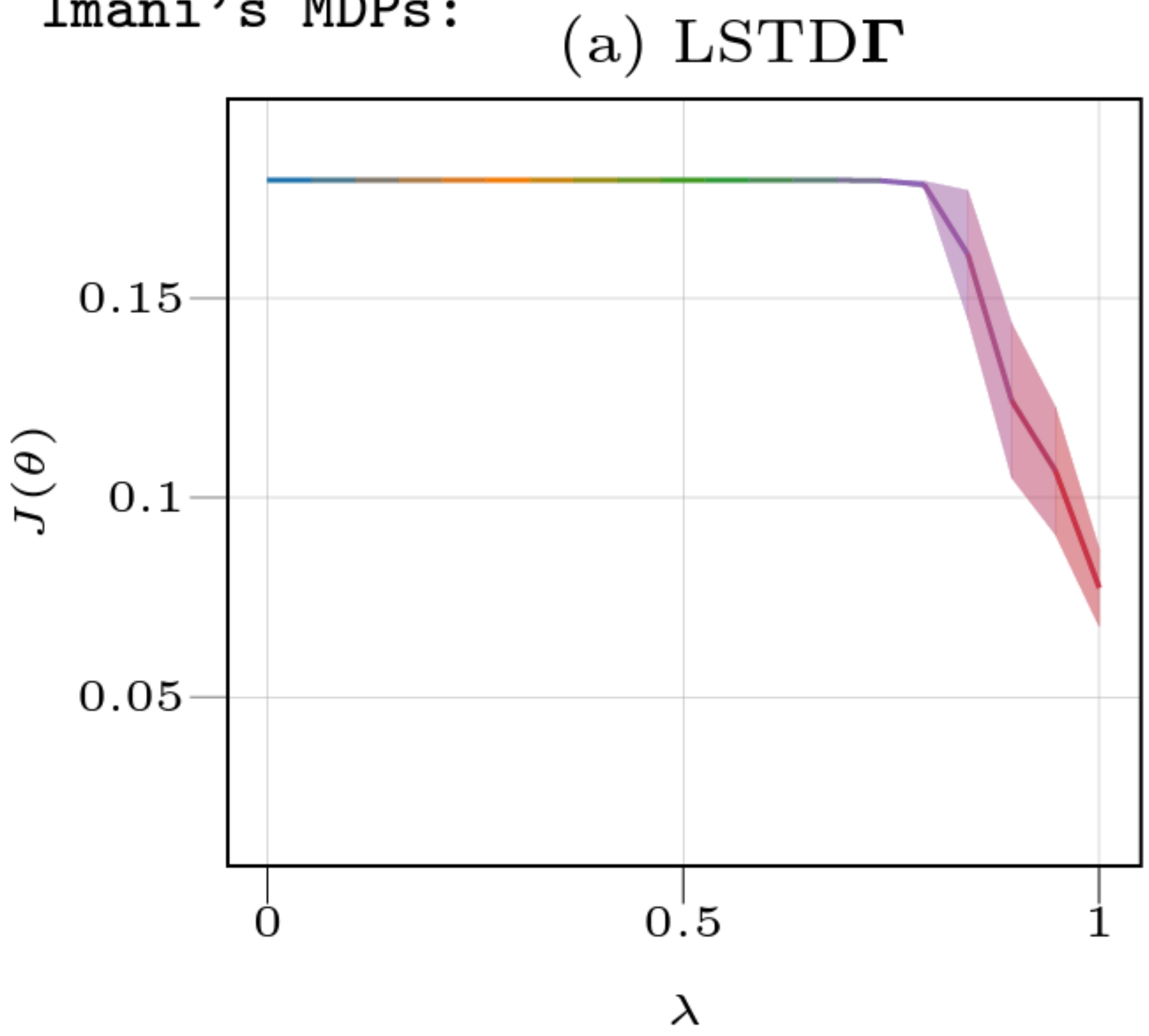


Empirical Analysis

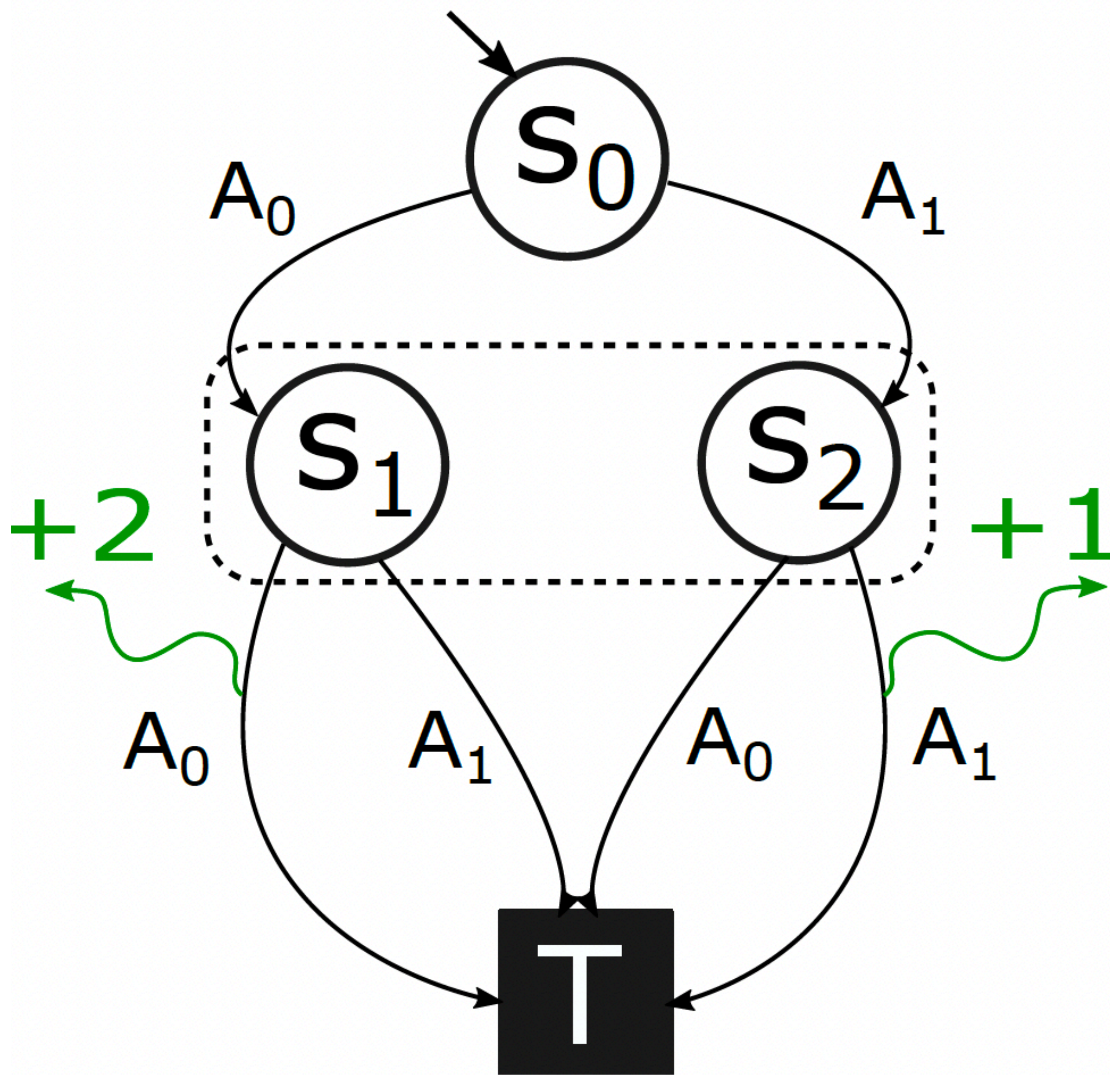


Imani et al. 2018

Imani's MDPs:

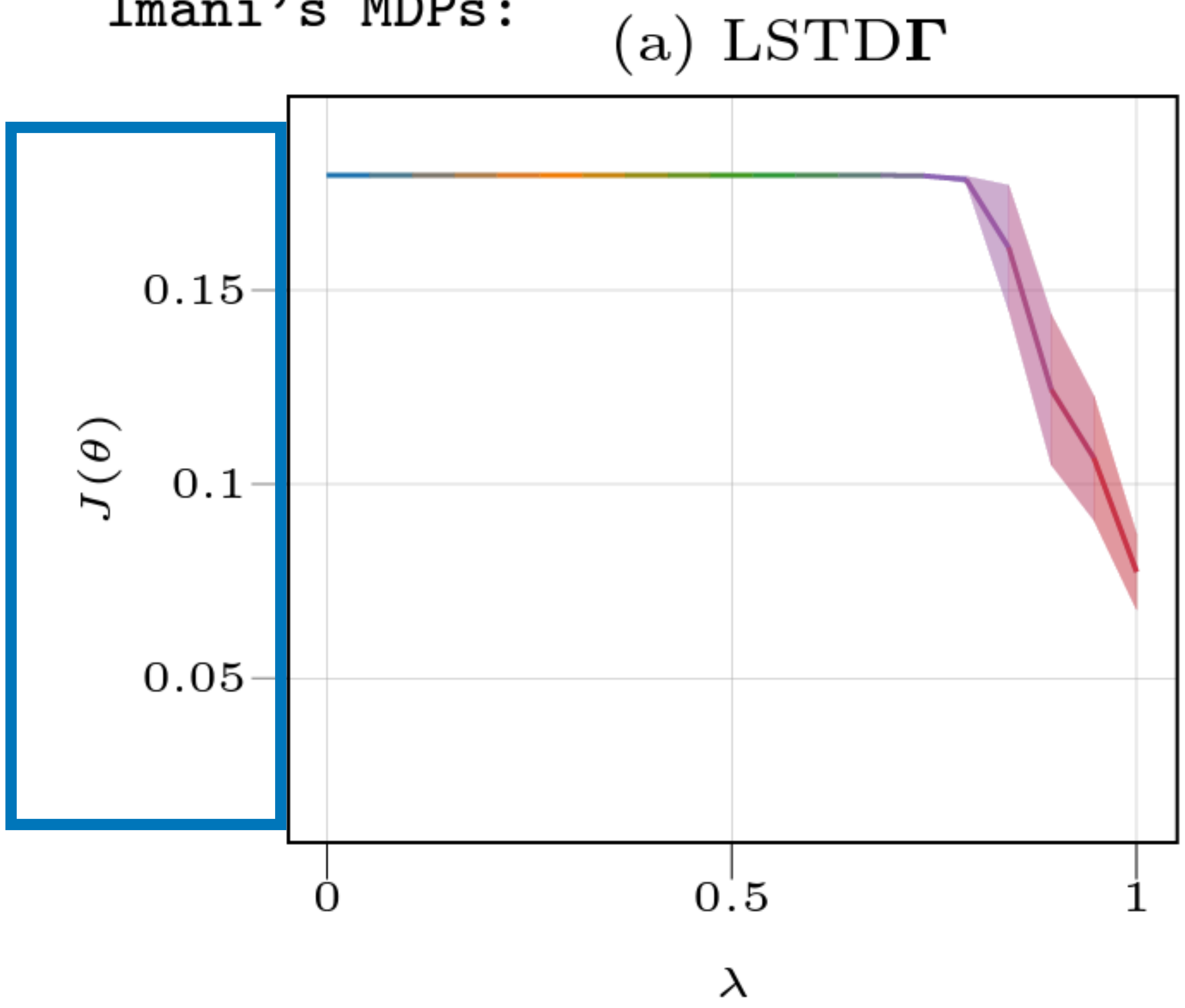


Empirical Analysis

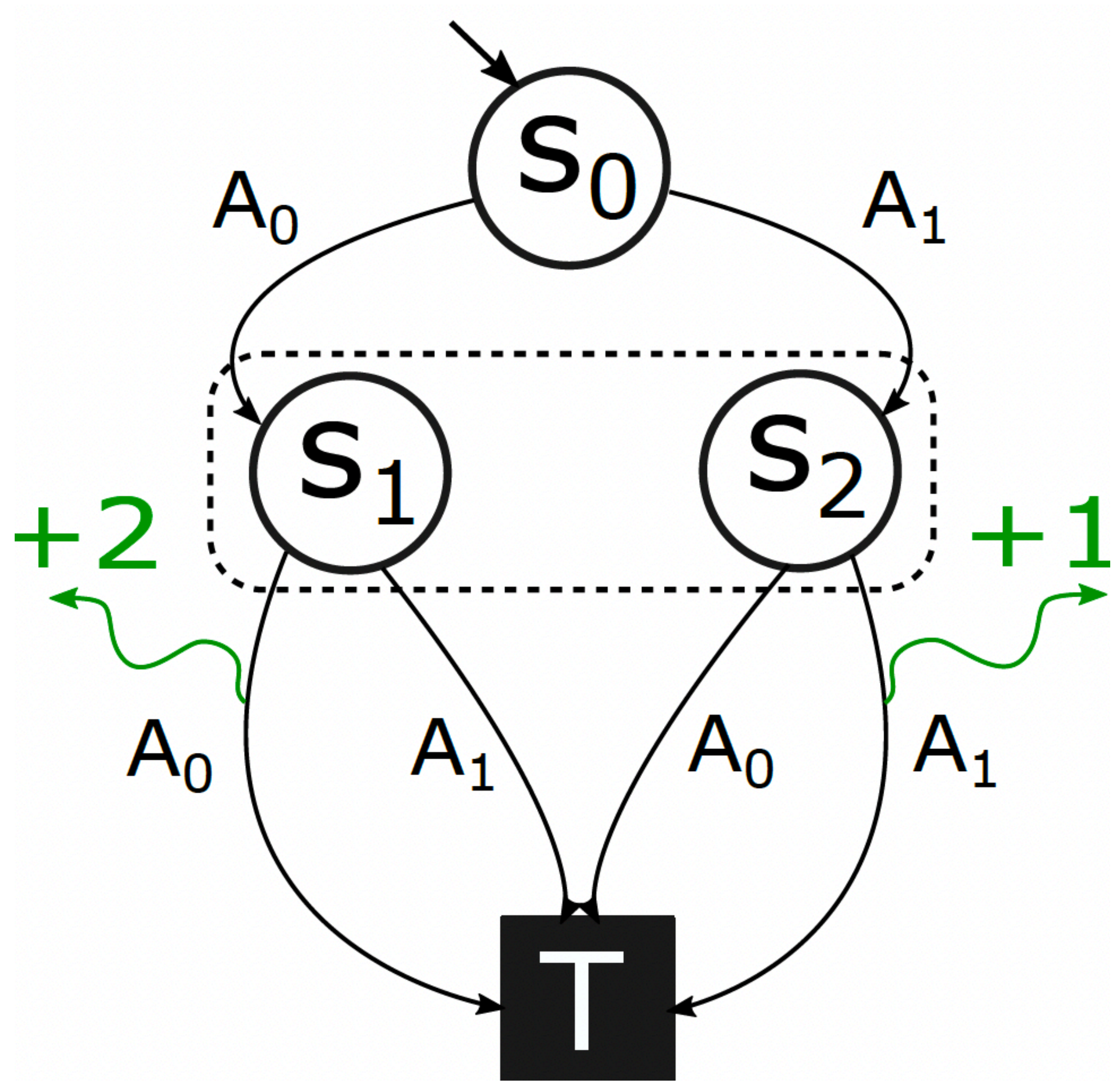


Imani et al. 2018

Imani's MDPs:

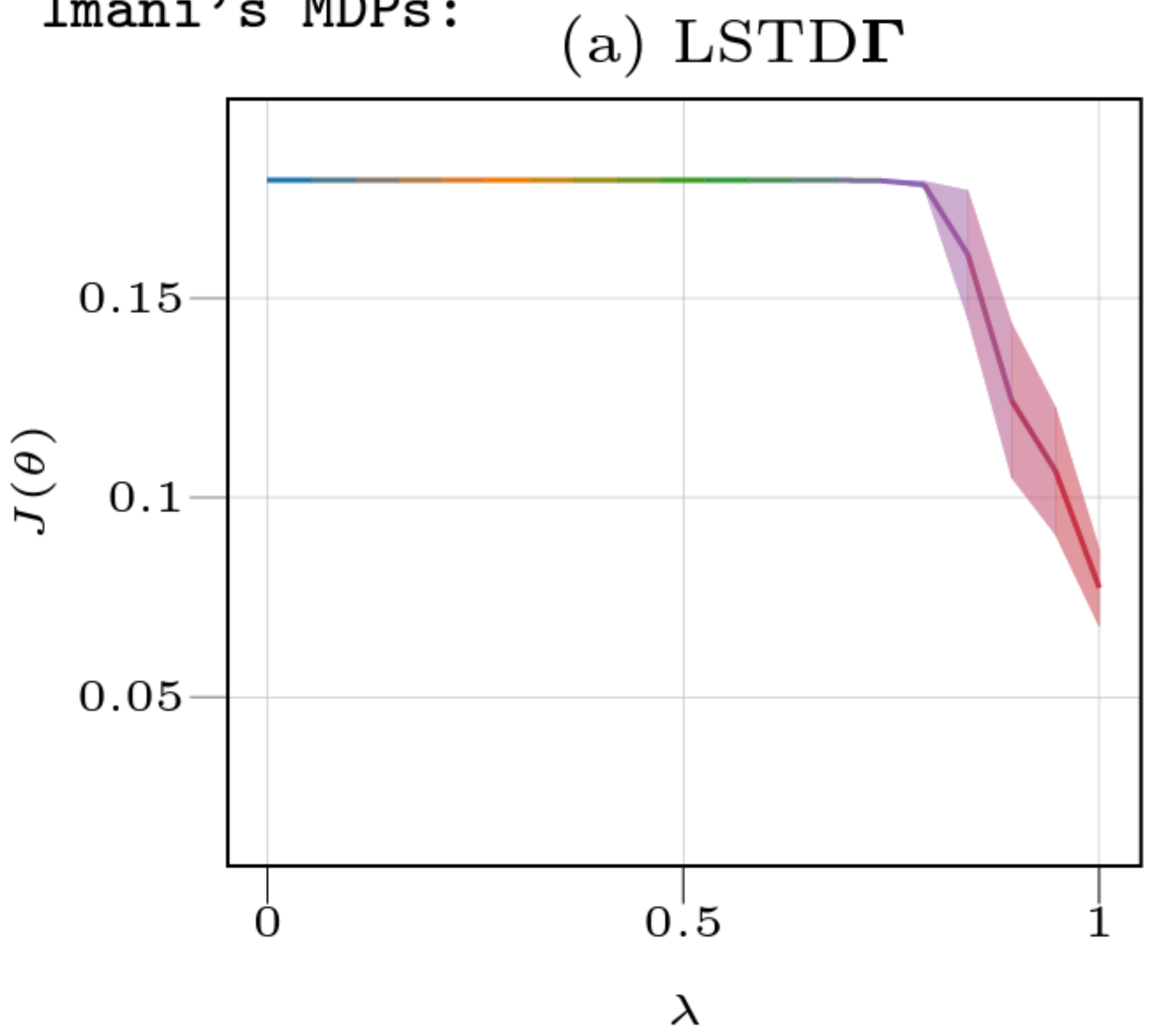


Empirical Analysis



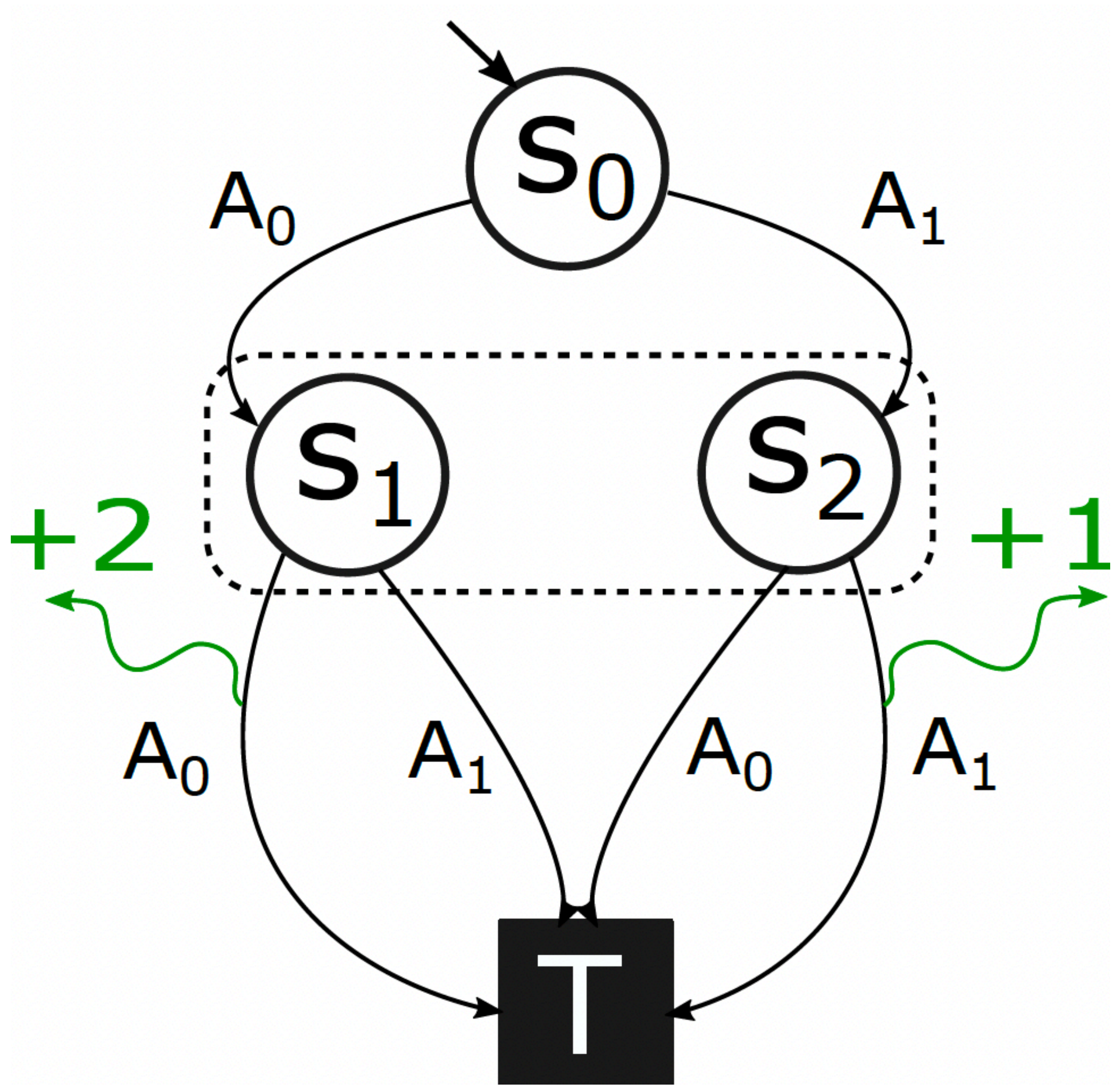
Imani et al. 2018

Imani's MDPs:

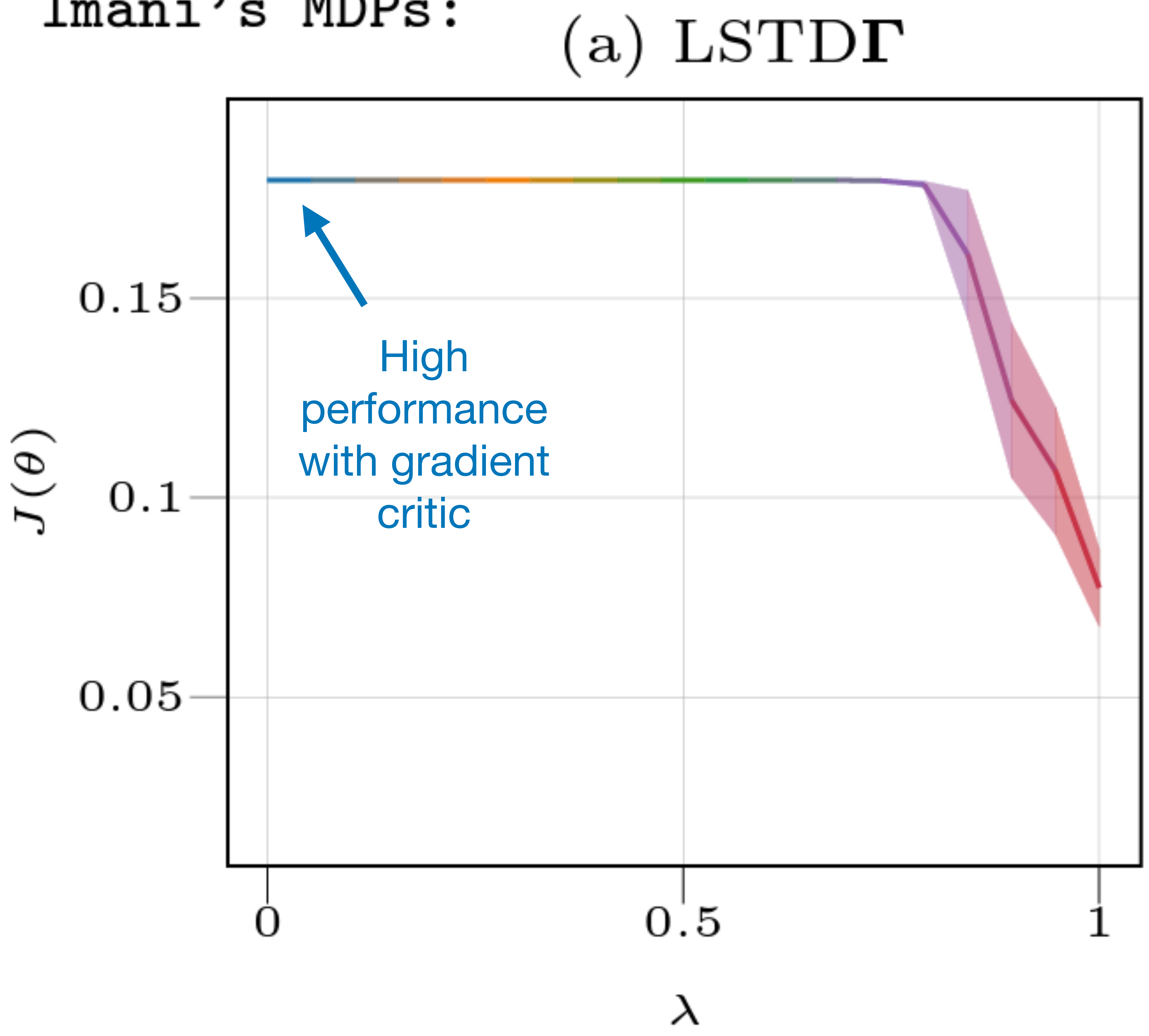


Empirical Analysis

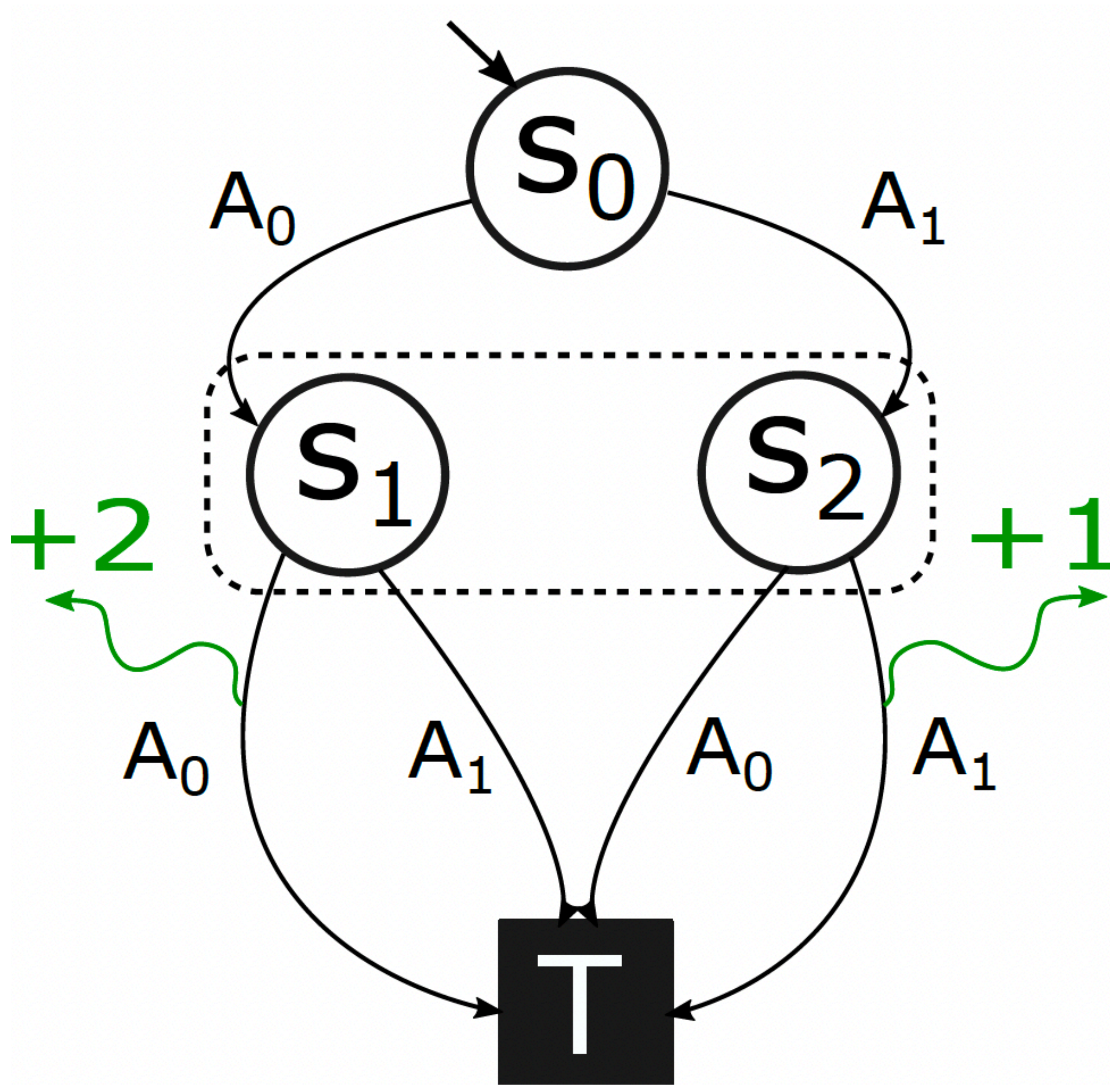
Imani's MDPs:



Imani et al. 2018

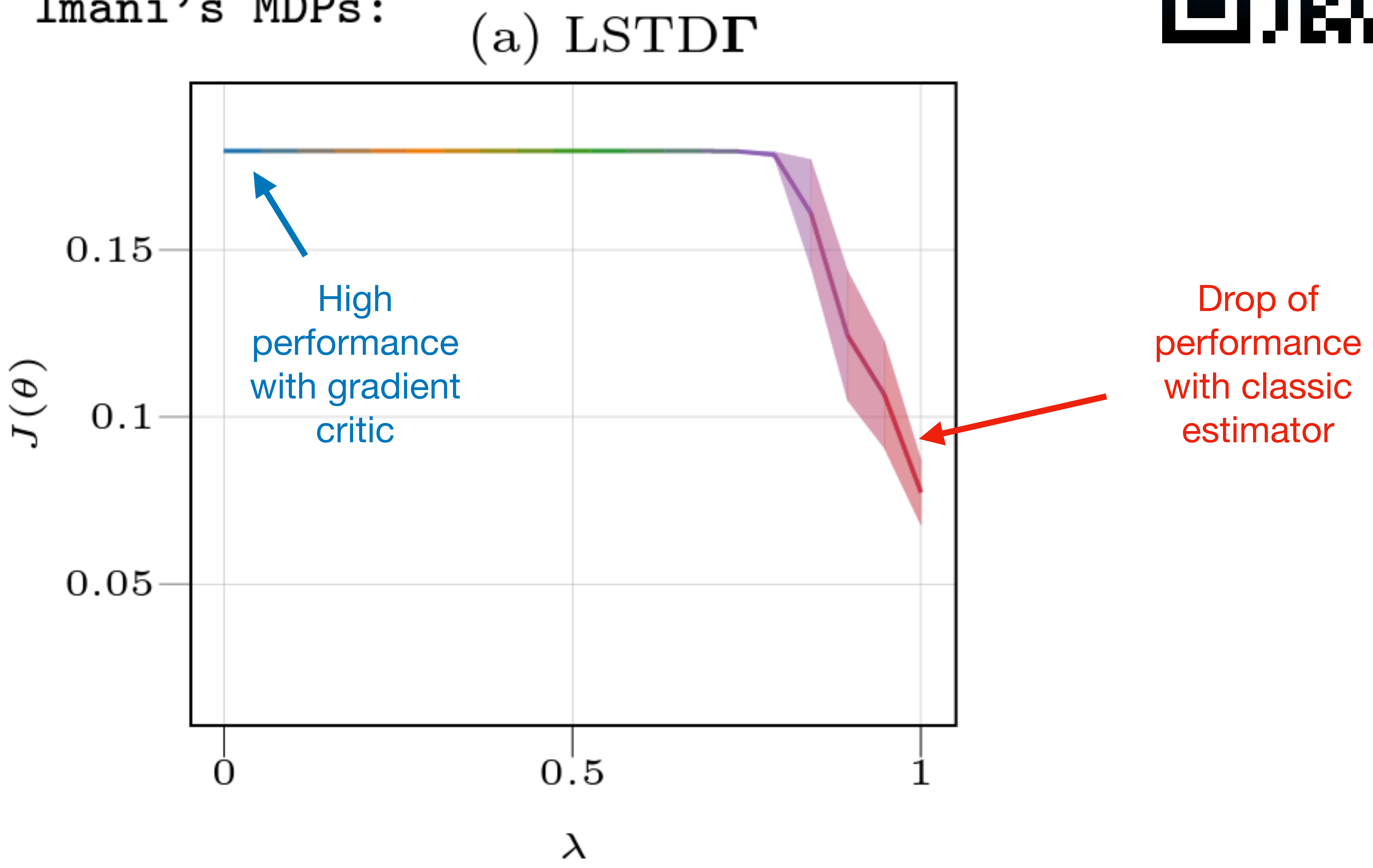


Empirical Analysis

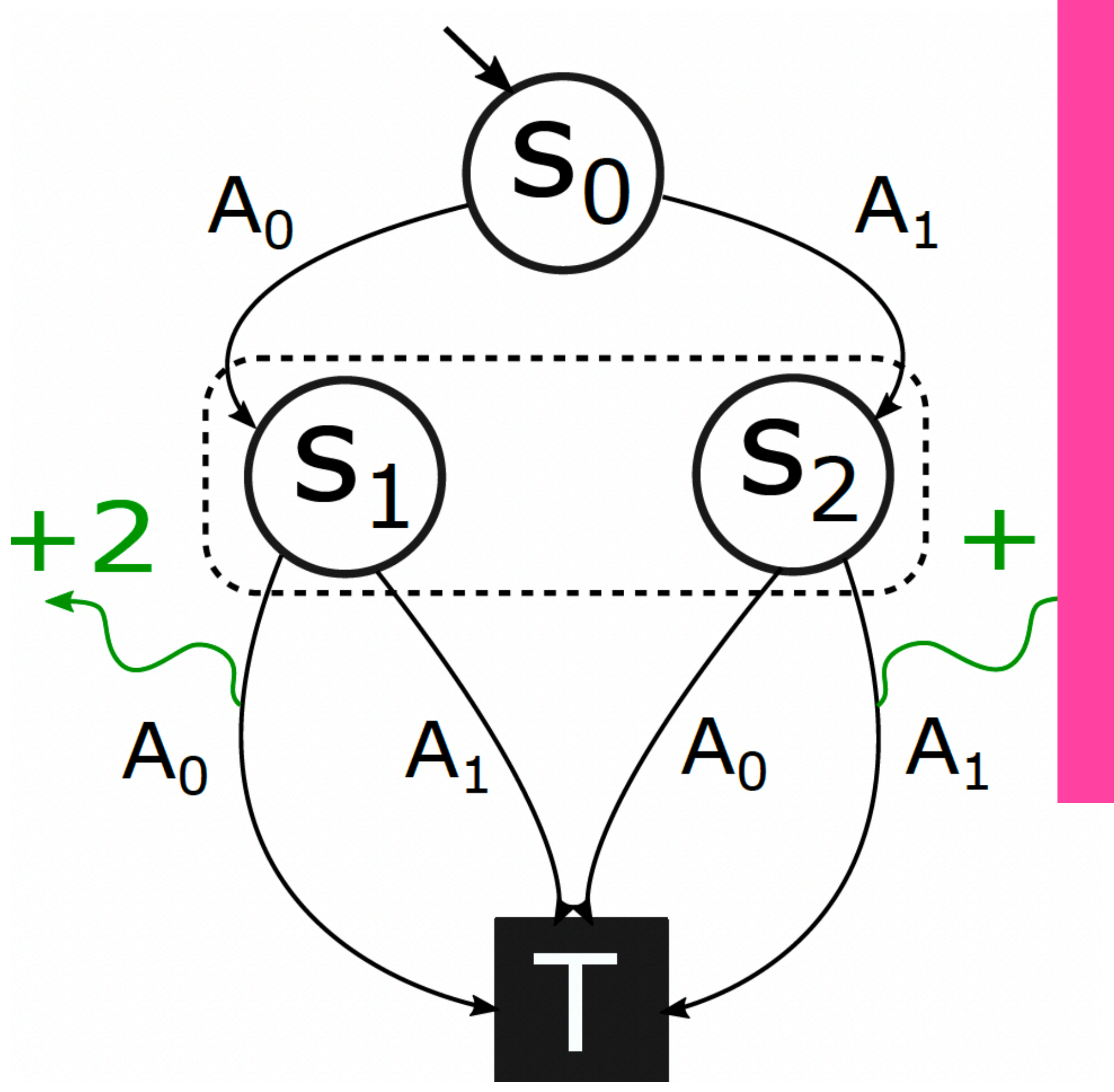


Imani et al. 2018

Imani's MDPs:

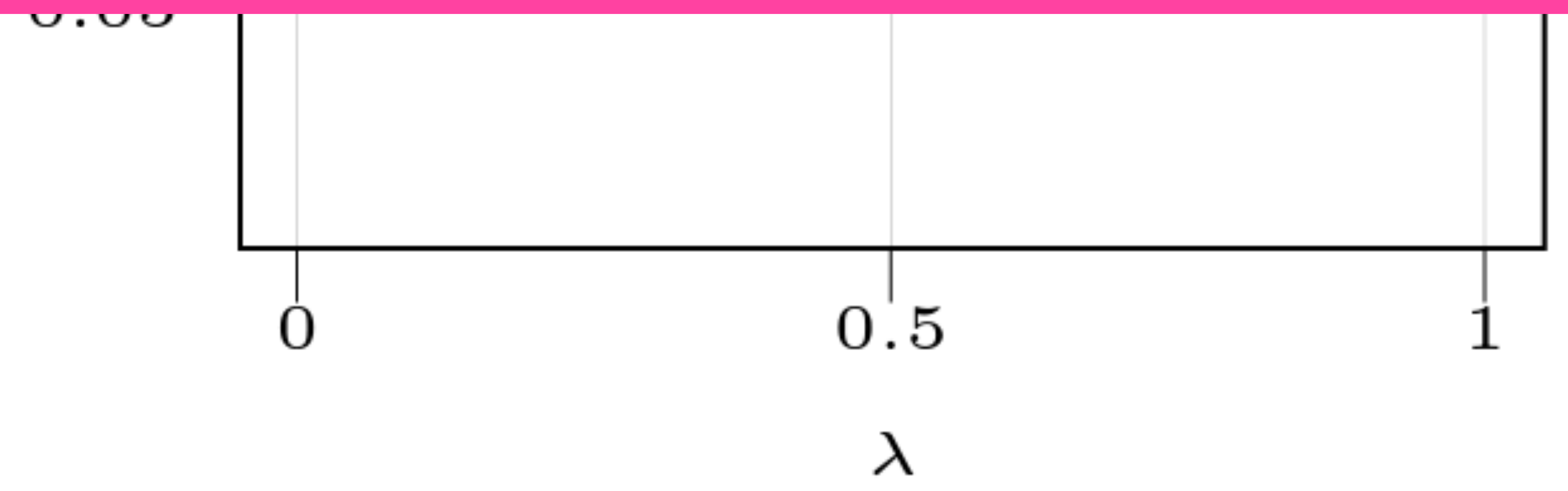


Empirical Analysis



Imani et al. 2018

The gradient critic can help achieving higher performance when samples are off-policy

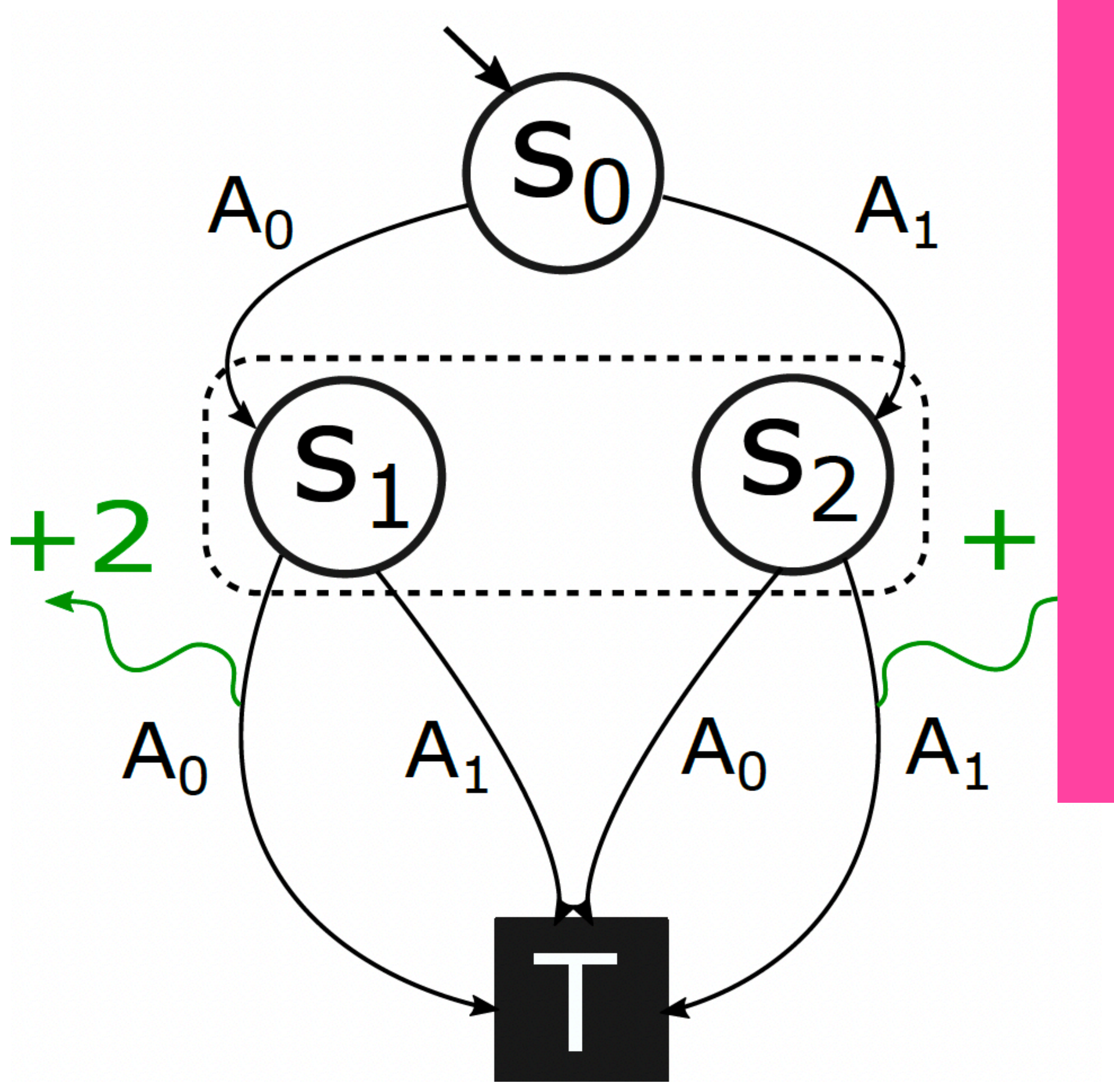


Drop of performance with classic estimator

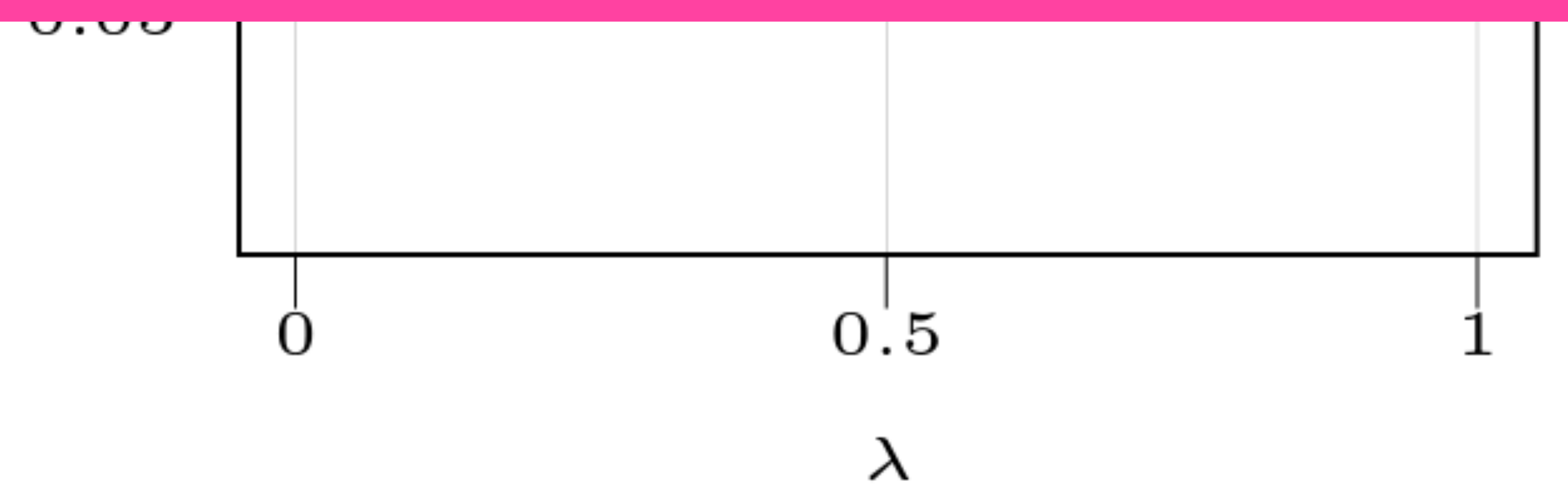
Empirical Analysis

Check our paper for more details!

The gradient critic can help achieving higher performance when samples are off-policy



Imani et al. 2018



Drop of performance with classic estimator