Policy Gradient methods

Sébastien Gros

Cybernetic, NTNU
Elec. Eng., Chalmers

TUM lectures on RL

Outline

Outline

Why another technique?

The approximation problem:

We have seen how to do:

$$Q_{\theta}\left(\mathbf{x},\mathbf{u}\right) \approx Q_{\star}\left(\mathbf{x},\mathbf{u}\right)$$

and defined:

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) := \arg\min_{\mathbf{u}} \; Q_{\boldsymbol{\theta}}\left(\mathbf{x},\mathbf{u}\right)$$

For a "**rich**" parametrization of $Q_{\boldsymbol{\theta}}$

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\pi}_{\star}(\mathbf{x})$$

holds. Otherwise all bets are off

Why another technique?

The approximation problem:

We have seen how to do:

$$Q_\theta(\mathbf{x}, \mathbf{u}) \approx Q_\star(\mathbf{x}, \mathbf{u})$$

and defined:

$$\pi_\theta(\mathbf{x}) := \arg\min_{\mathbf{u}} Q_\theta(\mathbf{x}, \mathbf{u})$$

For a "**rich**" parametrization of Q_θ

$$\pi_\theta(\mathbf{x}) = \pi_\star(\mathbf{x})$$

holds. Otherwise all bets are off

The minimization problem:

Minimization step:

$$\pi_\theta(\mathbf{x}) := \arg\min_{\mathbf{u}} Q_\theta(\mathbf{x}, \mathbf{u})$$

is "easy" only if $Q_\theta(\mathbf{x}, \mathbf{u})$ is convex in $\mathbf{u}$.

However, generic choices of Q_θ are typically not

Why another technique?

The approximation problem:

We have seen how to do:

$$Q_{\theta}(\mathbf{x}, \mathbf{u}) \approx Q_{\star}(\mathbf{x}, \mathbf{u})$$

and defined:

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) := \arg\min_{\mathbf{u}} Q_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{u})$$

For a "**rich**" parametrization of $Q_{\boldsymbol{\theta}}$

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\pi}_{\star}(\mathbf{x})$$

holds. Otherwise all bets are off

The minimization problem:

Minimization step:

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) := \arg\min_{\mathbf{u}} Q_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{u})$$

is "easy" only if $Q_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{u})$ is convex in $\mathbf{u}$.

However, generic choices of $Q_{\boldsymbol{\theta}}$ are typically not

Policy gradient methods build directly π_θ and find the θ such that the policy yields the best closed-loop possible performances on the real system

Policy performance

Recall:

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}}}\left[L(\mathbf{x}, \mathbf{u}) \right] = \sum_{k=0}^{\infty} \int L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \, \mathbb{P}[\mathbf{x}_k \,|\, \mathbf{x}_0] \mathbb{P}[\mathbf{x}_0] \mathrm{d}\mathbf{x}_0 \mathrm{d}\mathbf{x}_k$$

where

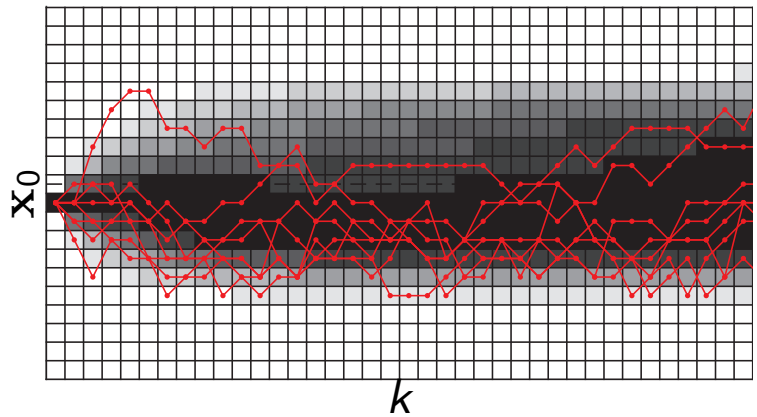
$$\mathbb{P}[\mathbf{x}_k \,|\, \mathbf{x}_0] = \int \prod_{i=0}^{k-1} \mathbb{P}[\mathbf{x}_{i+1} \,|\, \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)] \mathrm{d}\mathbf{x}_1 \ldots \mathrm{d}\mathbf{x}_{k-1}$$

Hence

$$J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}}\left[L(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}}) \right]$$

Note that

$$J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}[V_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}_0)] = \int V_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}_0) \mathbb{P}[\mathbf{x}_0] \mathrm{d}\mathbf{x}_0$$



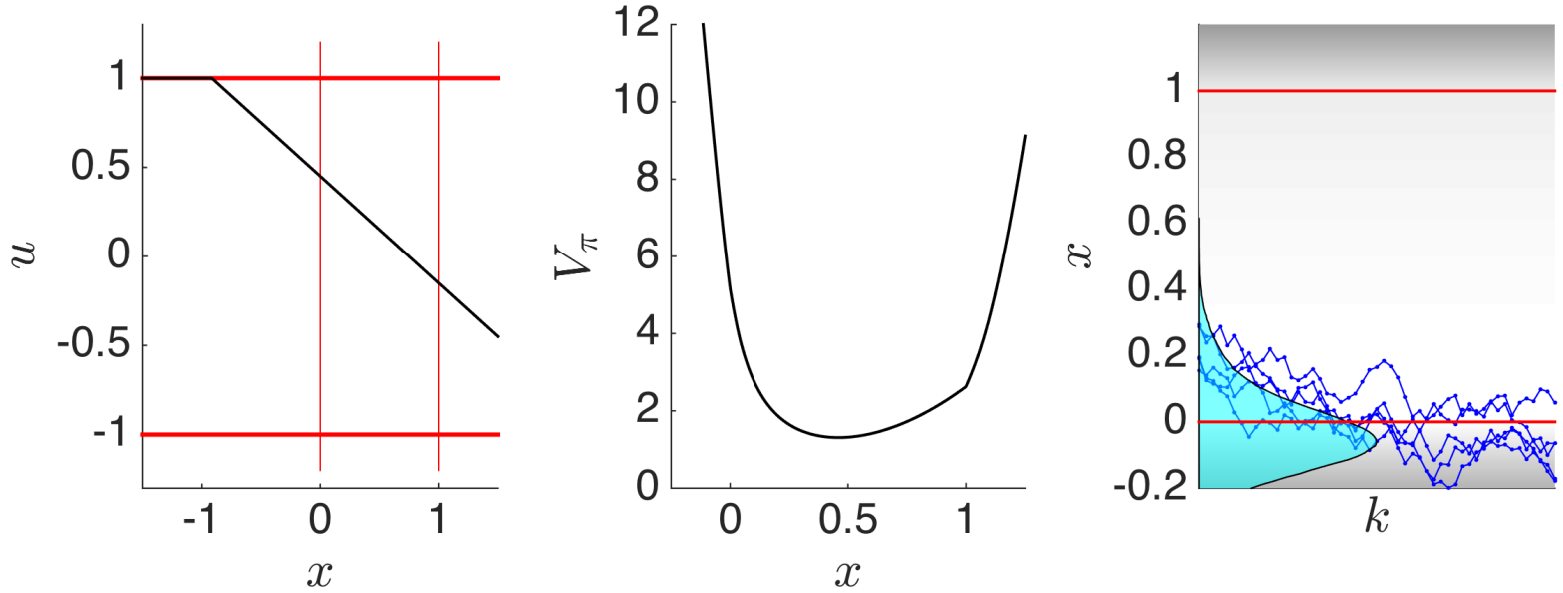
i.e. expectation is taken over initial conditions.

Policy performance - MDP distribution τ_{π}

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_{π}, value function V_{π}

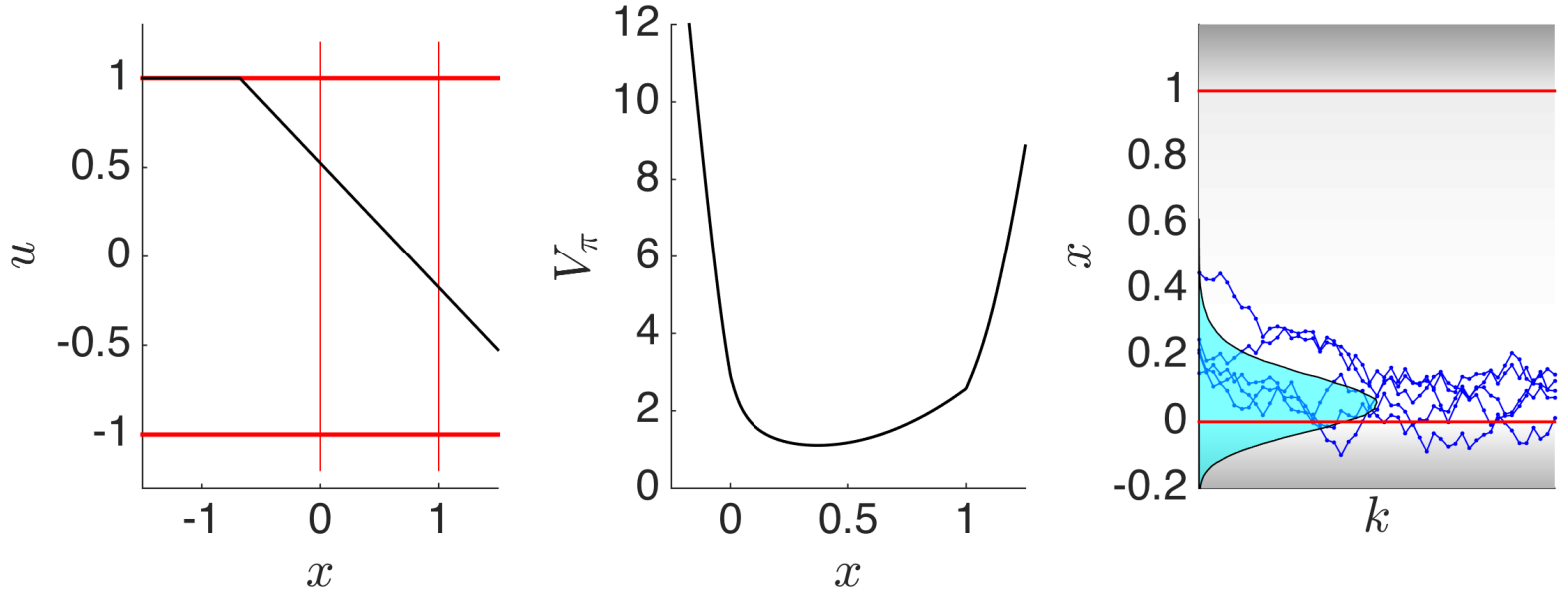
Policy performance - MDP distribution τ_{π}

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_{π}, value function V_{π}



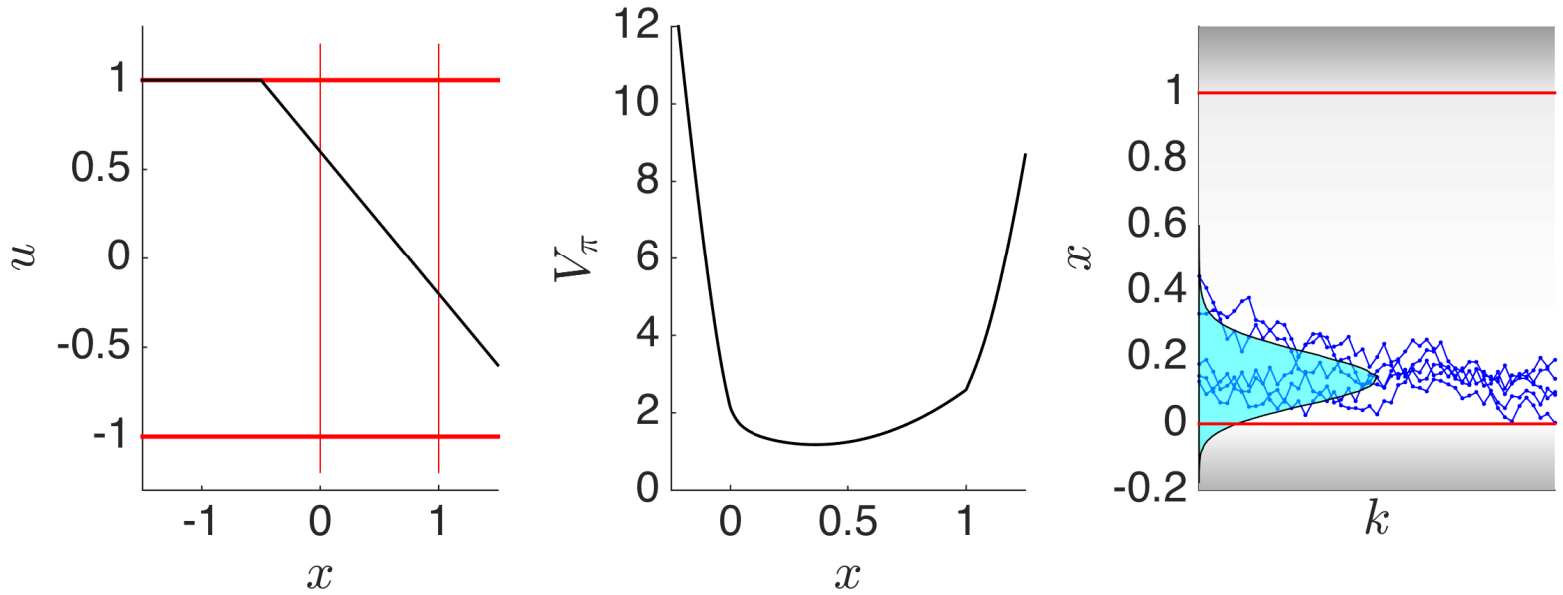
Policy performance - MDP distribution τ_{π}

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_{π}, value function V_{π}



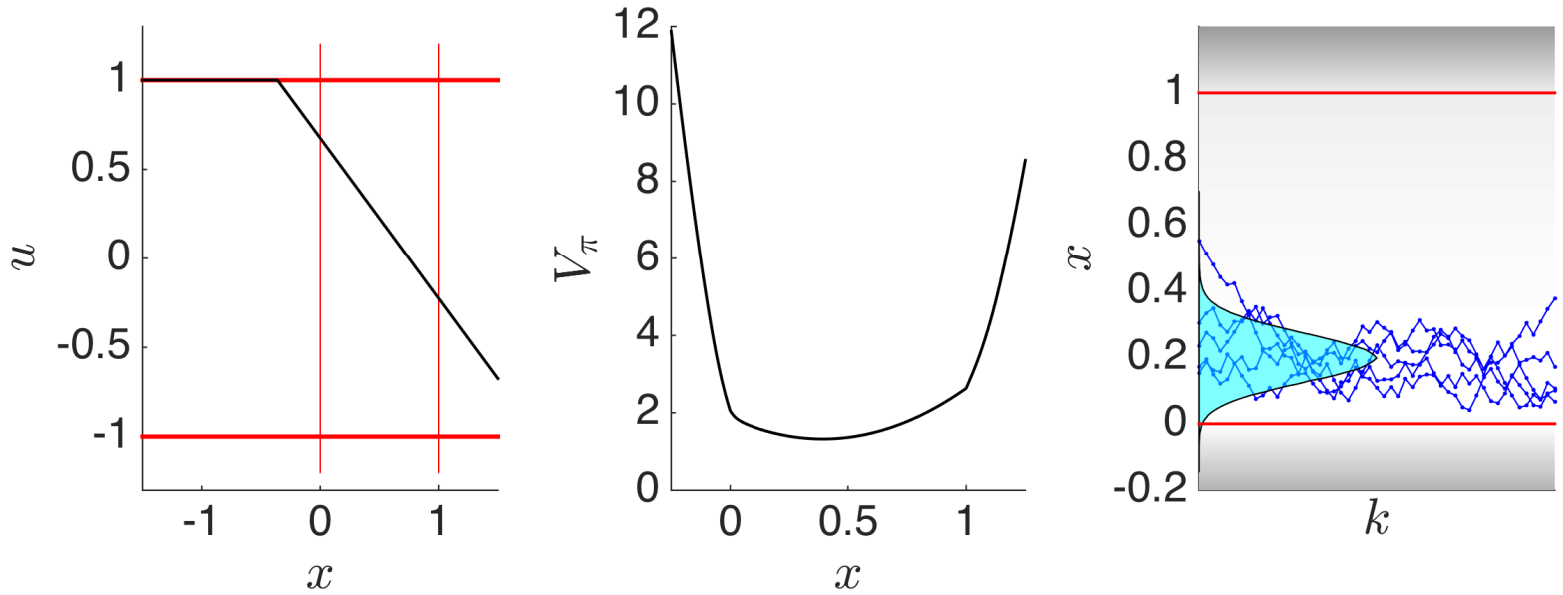
Policy performance - MDP distribution τ_{π}

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_{π}, value function V_{π}



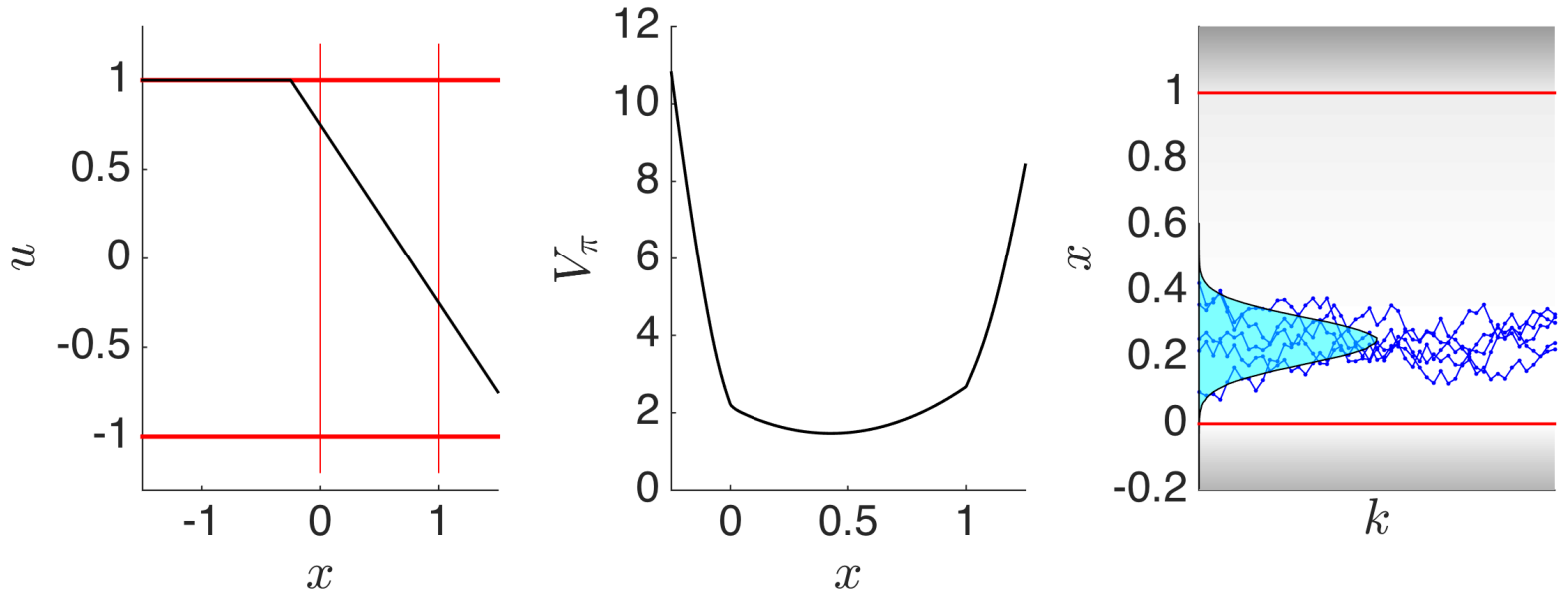
Policy performance - MDP distribution τ_{π}

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_{π}, value function V_{π}



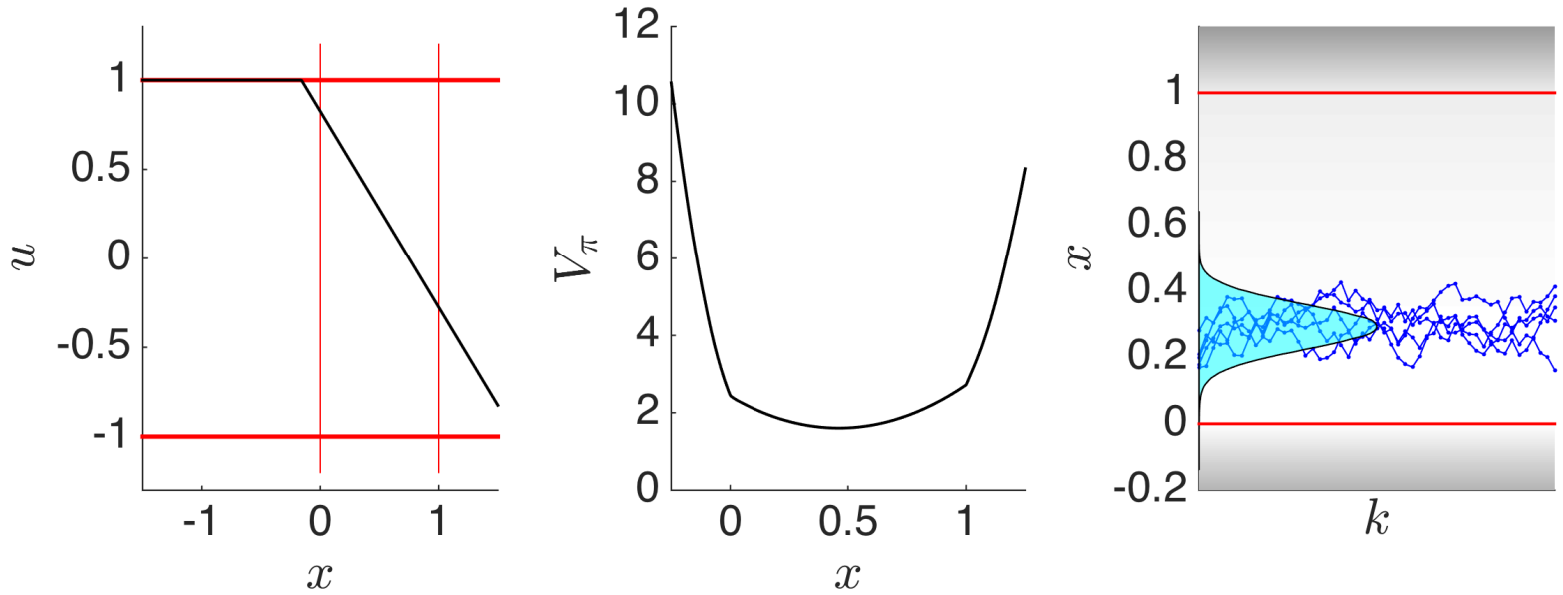
Policy performance - MDP distribution τ_{π}

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_{π}, value function V_{π}



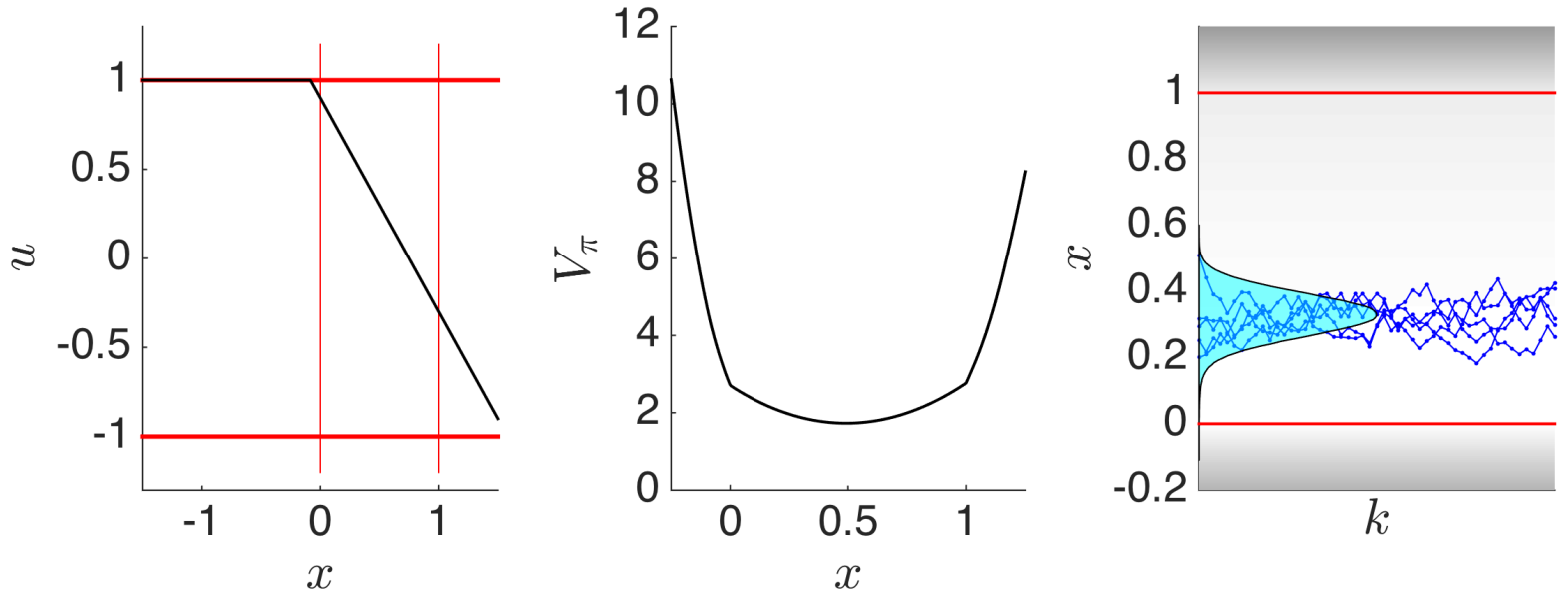
Policy performance - MDP distribution τ_{π}

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_{π}, value function V_{π}



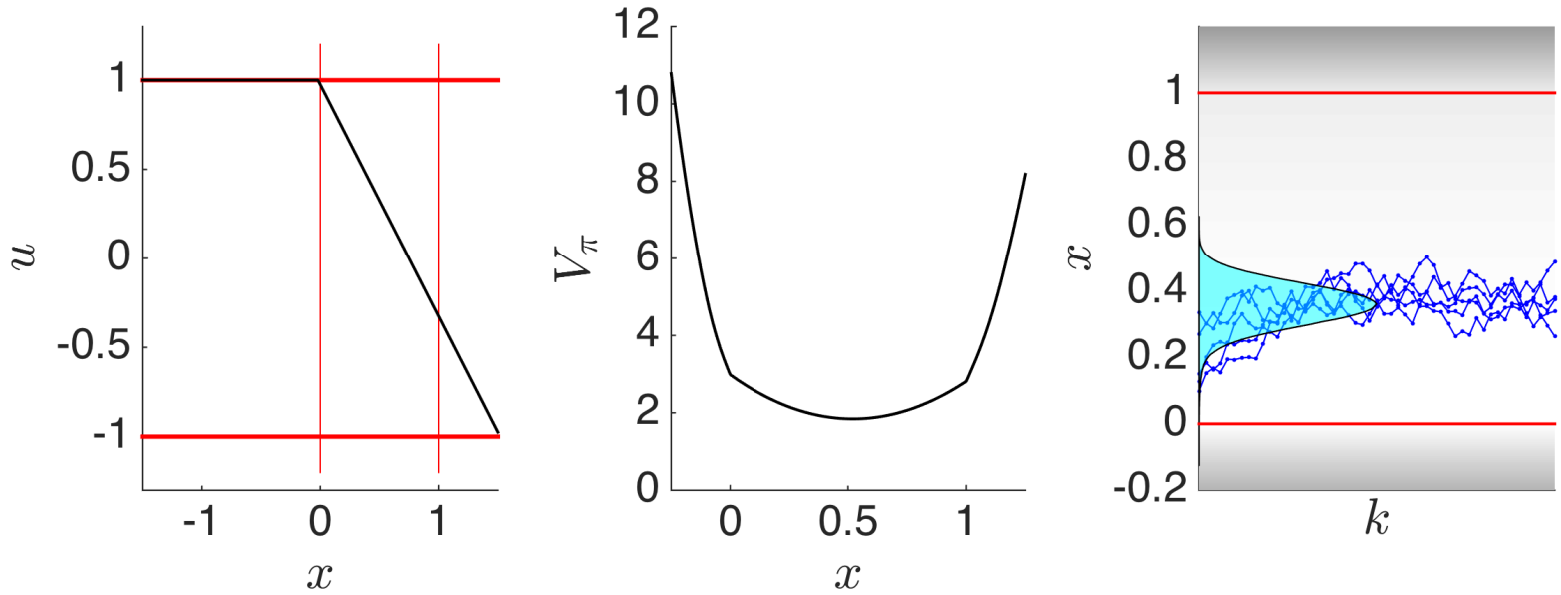
Policy performance - MDP distribution τ_{π}

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_{π}, value function V_{π}



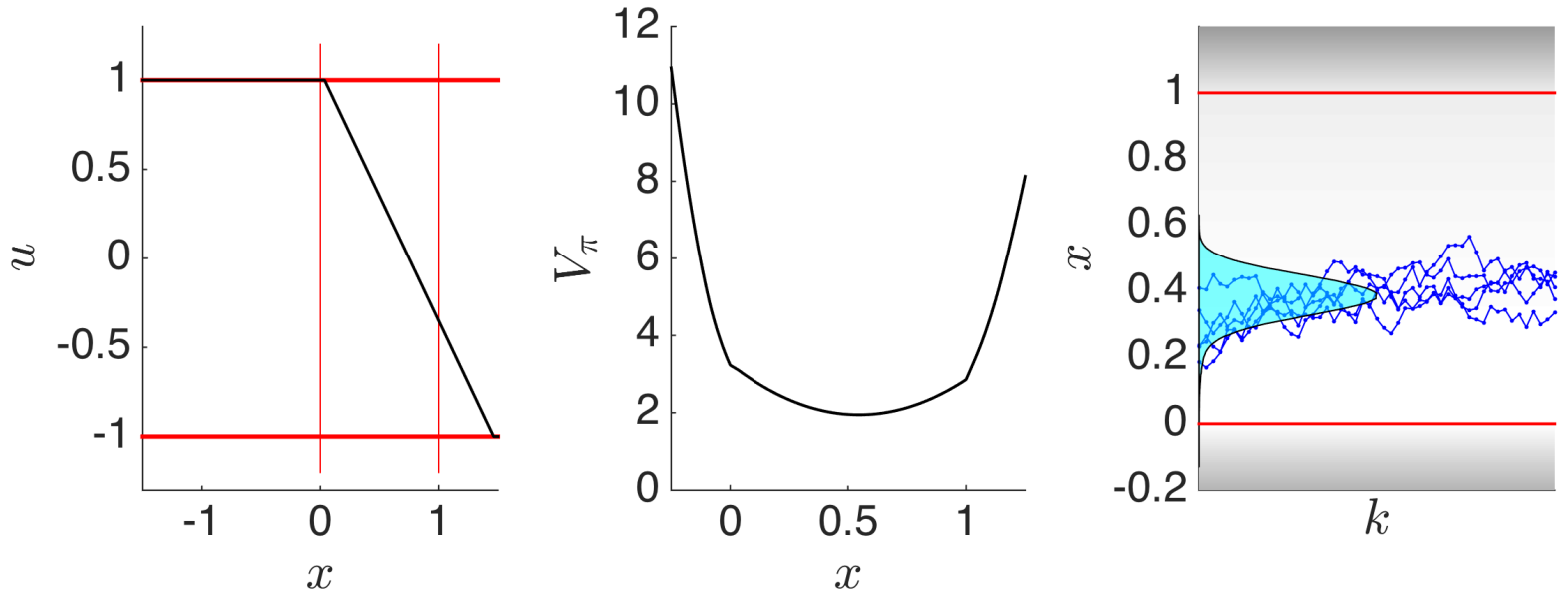
Policy performance - MDP distribution τ_{π}

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_{π}, value function V_{π}



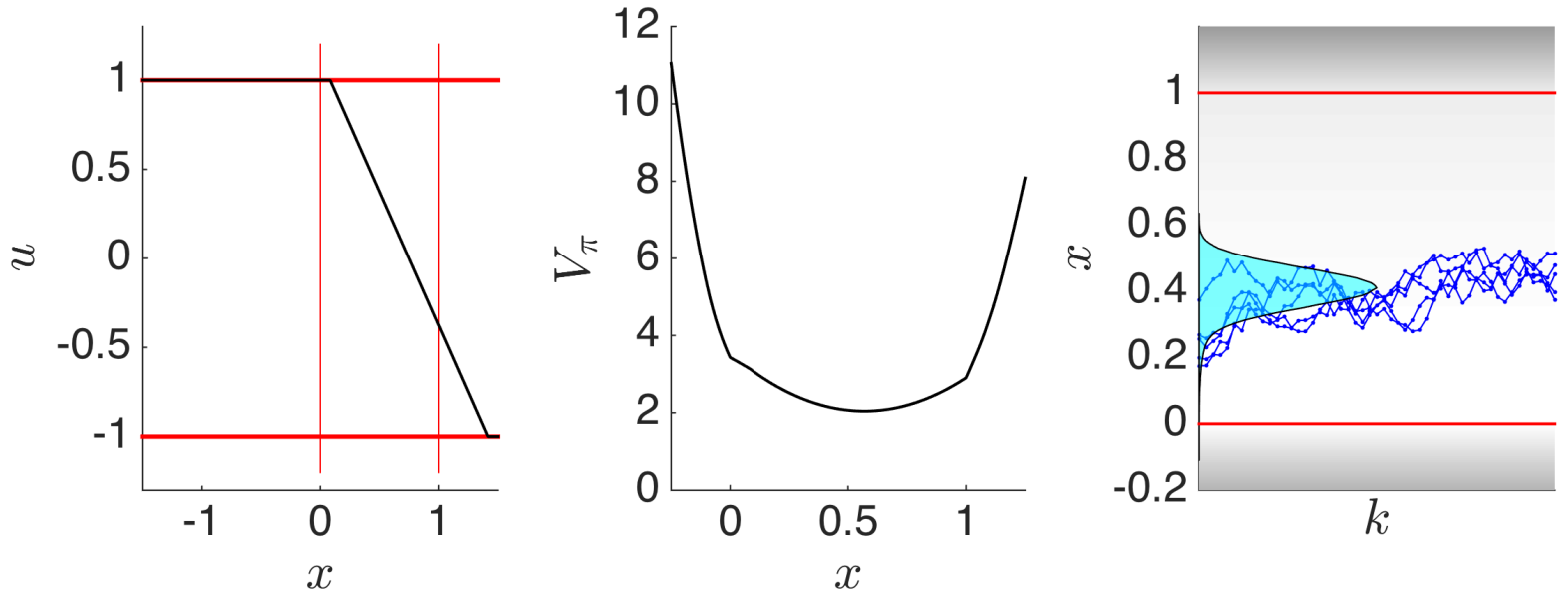
Policy performance - MDP distribution τ_{π}

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_{π}, value function V_{π}



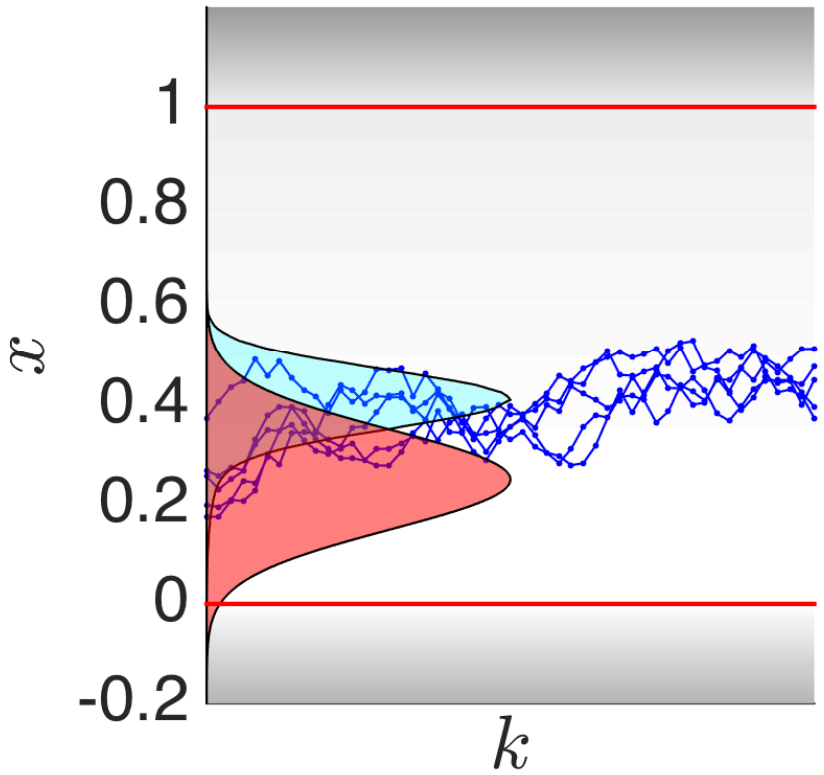
Policy performance - MDP distribution τ_{π}

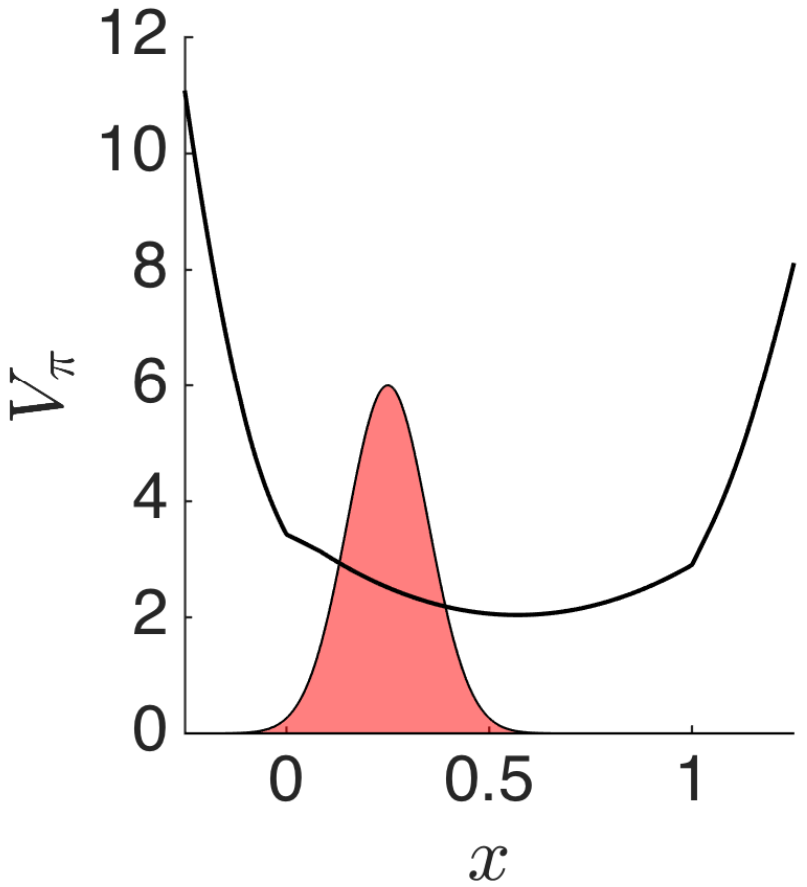
- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_{π}, value function V_{π}



Policy performance - $J(\pi)$

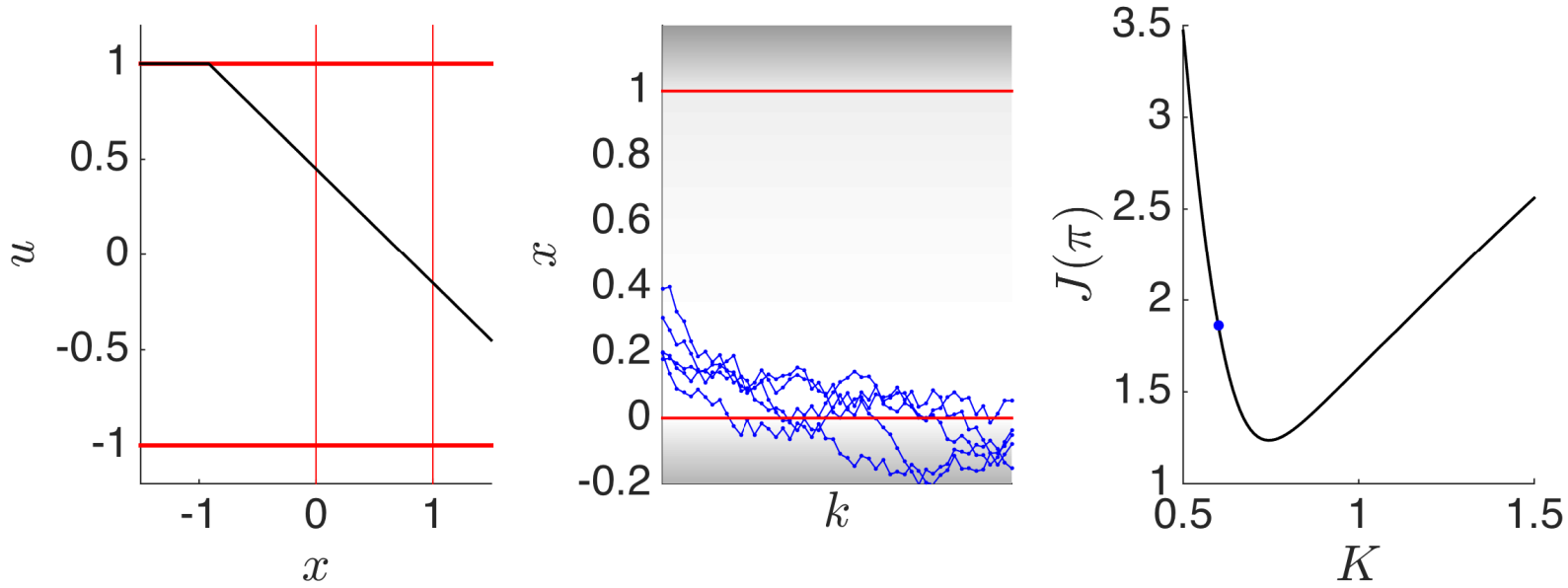
- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_π, value function V_π
- Yields policy performance $J(\pi) = \mathbb{E}\left[V_\pi(\mathbf{x}_0)\right] = \mathbb{E}_{\tau_\pi}\left[L(\mathbf{x}, \pi(\mathbf{x}))\right]$





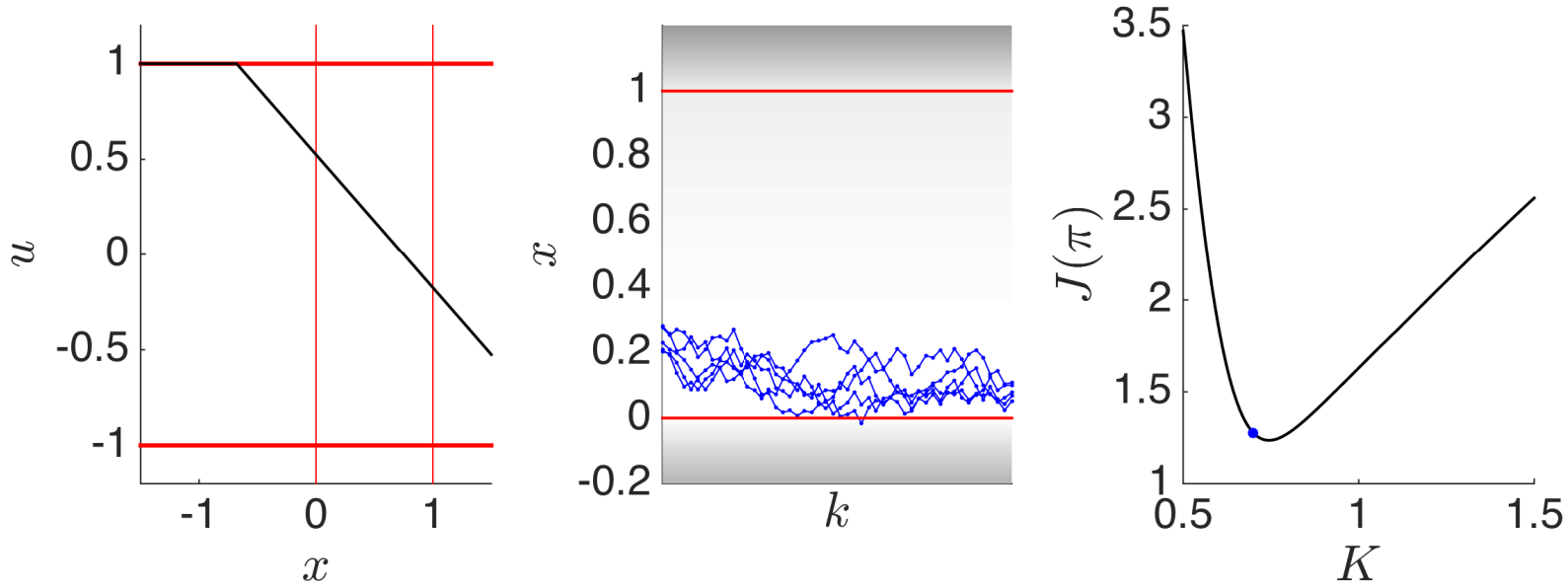
Policy performance - $J(\pi)$

- Policy: $\pi(\mathbf{x}) = \text{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\text{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\text{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_π, value function V_π
- Yields policy performance $J(\pi) = \mathbb{E}\left[V_\pi(\mathbf{x}_0)\right] = \mathbb{E}_{\tau_\pi}\left[L(\mathbf{x}, \pi(\mathbf{x}))\right]$



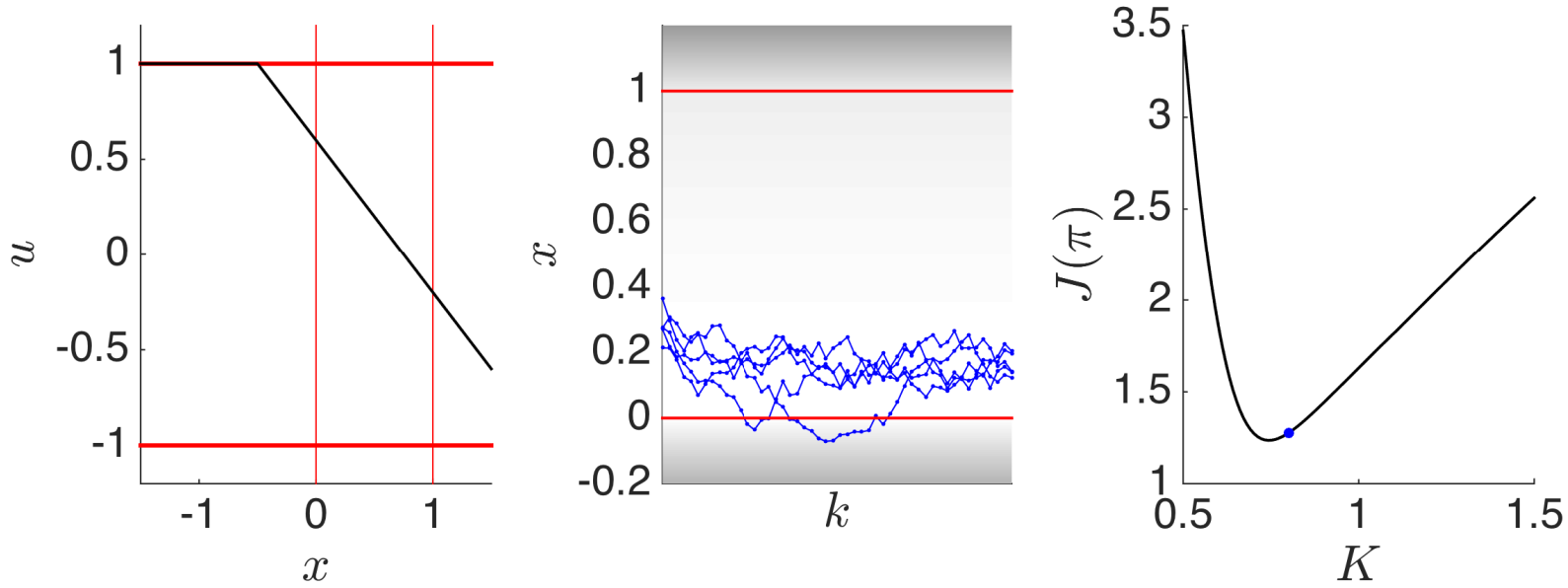
Policy performance - $J(\pi)$

- Policy: $\pi(\mathbf{x}) = \text{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\text{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\text{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_π, value function V_π
- Yields policy performance $J(\pi) = \mathbb{E}[V_\pi(\mathbf{x}_0)] = \mathbb{E}_{\tau_\pi}[L(\mathbf{x}, \pi(\mathbf{x}))]$



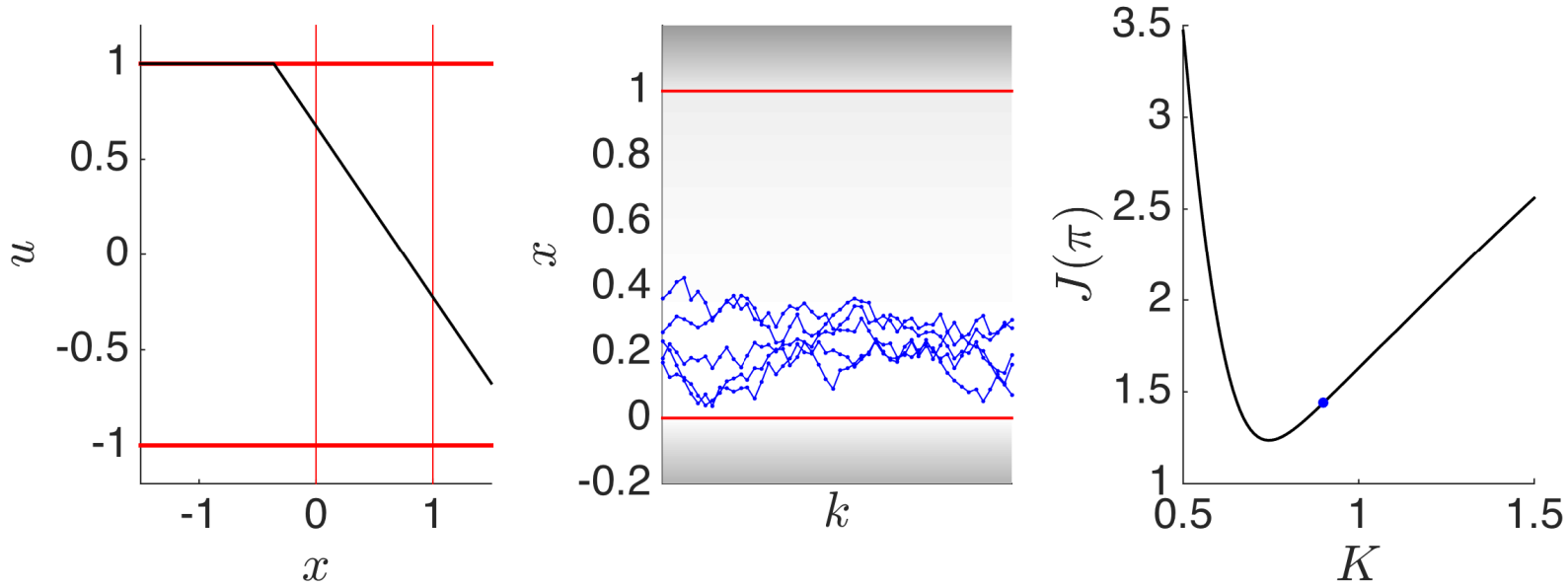
Policy performance - $J(\pi)$

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_π, value function V_π
- Yields policy performance $J(\pi) = \mathbb{E}\left[V_\pi(\mathbf{x}_0)\right] = \mathbb{E}_{\tau_\pi}\left[L(\mathbf{x}, \pi(\mathbf{x}))\right]$



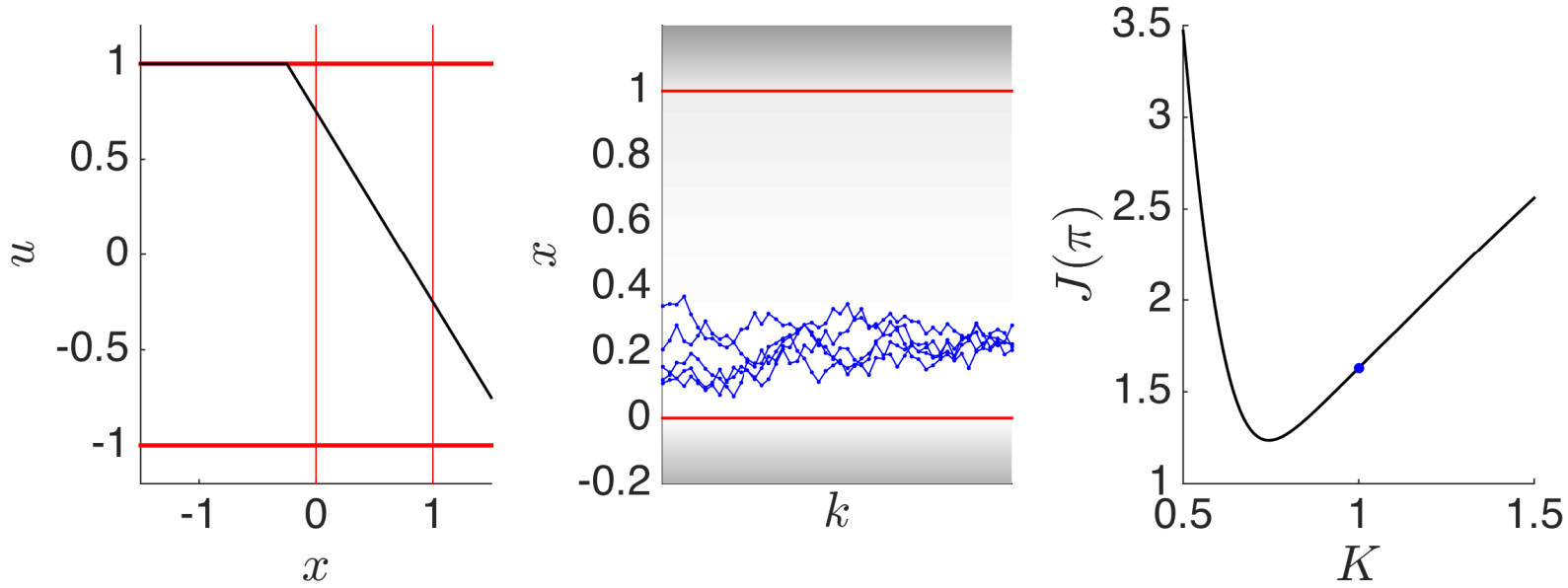
Policy performance - $J(\pi)$

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_π, value function V_π
- Yields policy performance $J(\pi) = \mathbb{E}\left[V_\pi(\mathbf{x}_0)\right] = \mathbb{E}_{\tau_\pi}\left[L(\mathbf{x}, \pi(\mathbf{x}))\right]$



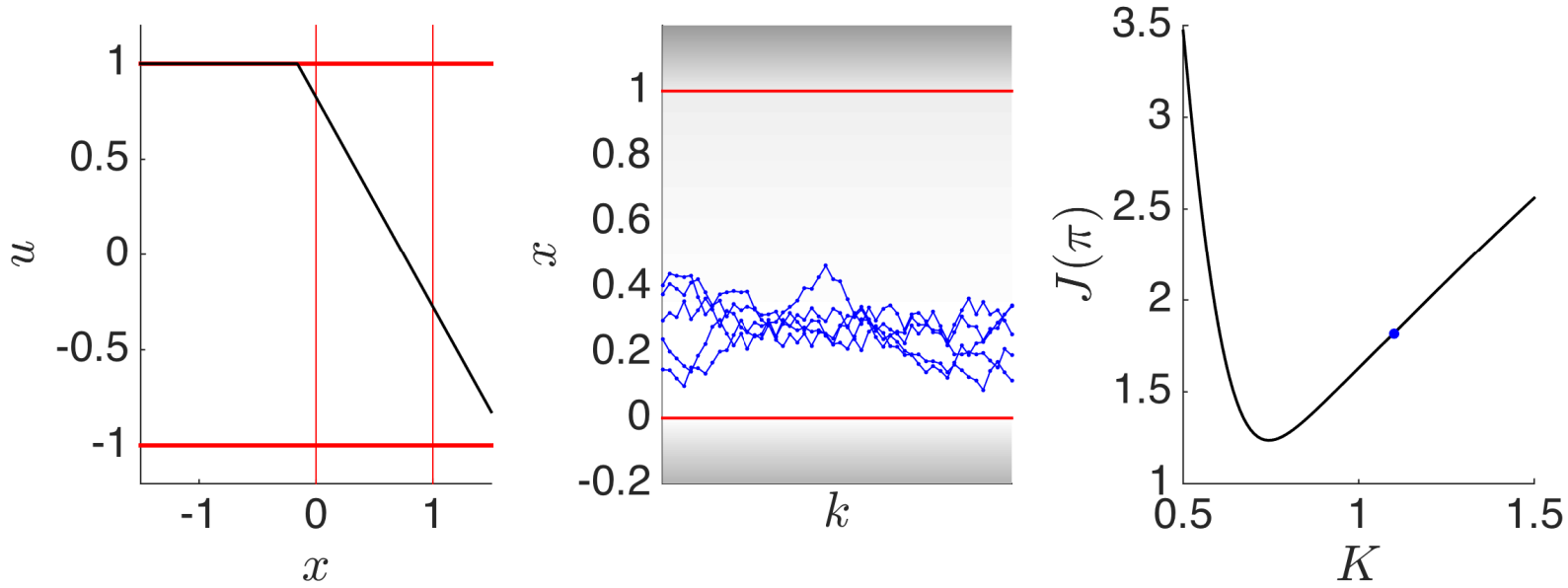
Policy performance - $J(\pi)$

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_π, value function V_π
- Yields policy performance $J(\pi) = \mathbb{E}\left[V_\pi(\mathbf{x}_0)\right] = \mathbb{E}_{\tau_\pi}\left[L(\mathbf{x}, \pi(\mathbf{x}))\right]$



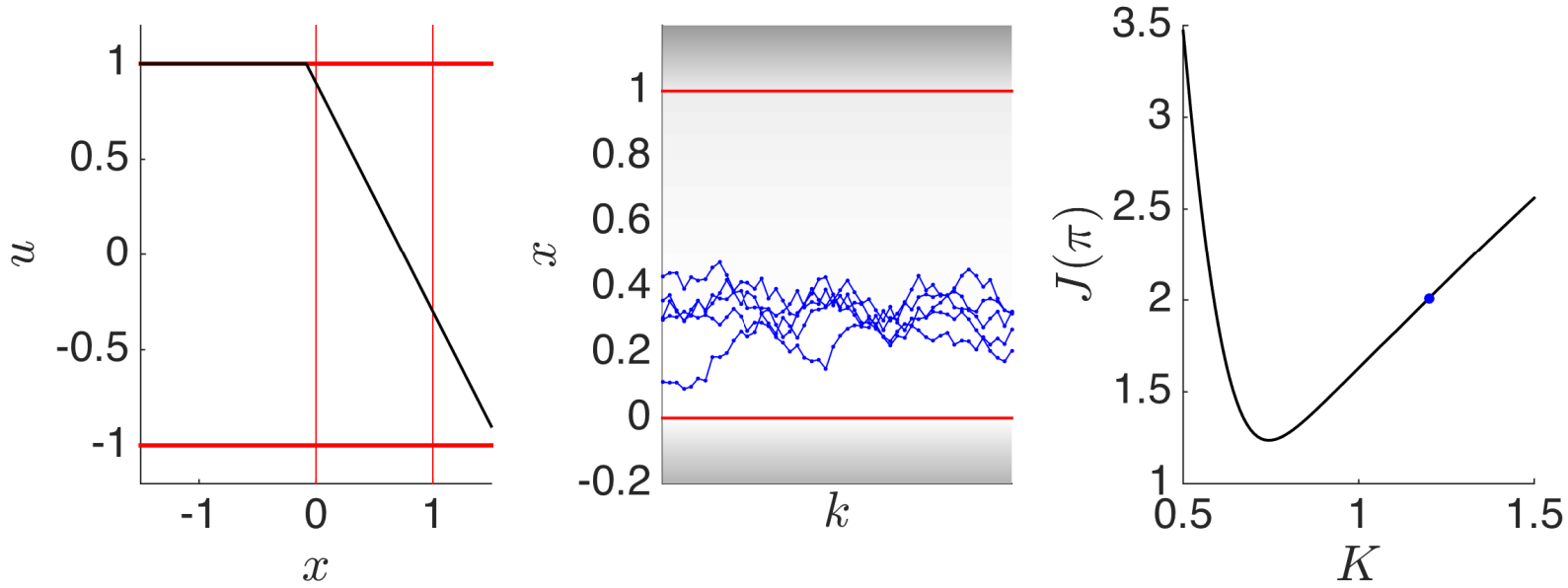
Policy performance - $J(\pi)$

- Policy: $\pi(\mathbf{x}) = \text{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\text{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\text{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_π, value function V_π
- Yields policy performance $J(\pi) = \mathbb{E}[V_\pi(\mathbf{x}_0)] = \mathbb{E}_{\tau_\pi}[L(\mathbf{x}, \pi(\mathbf{x}))]$



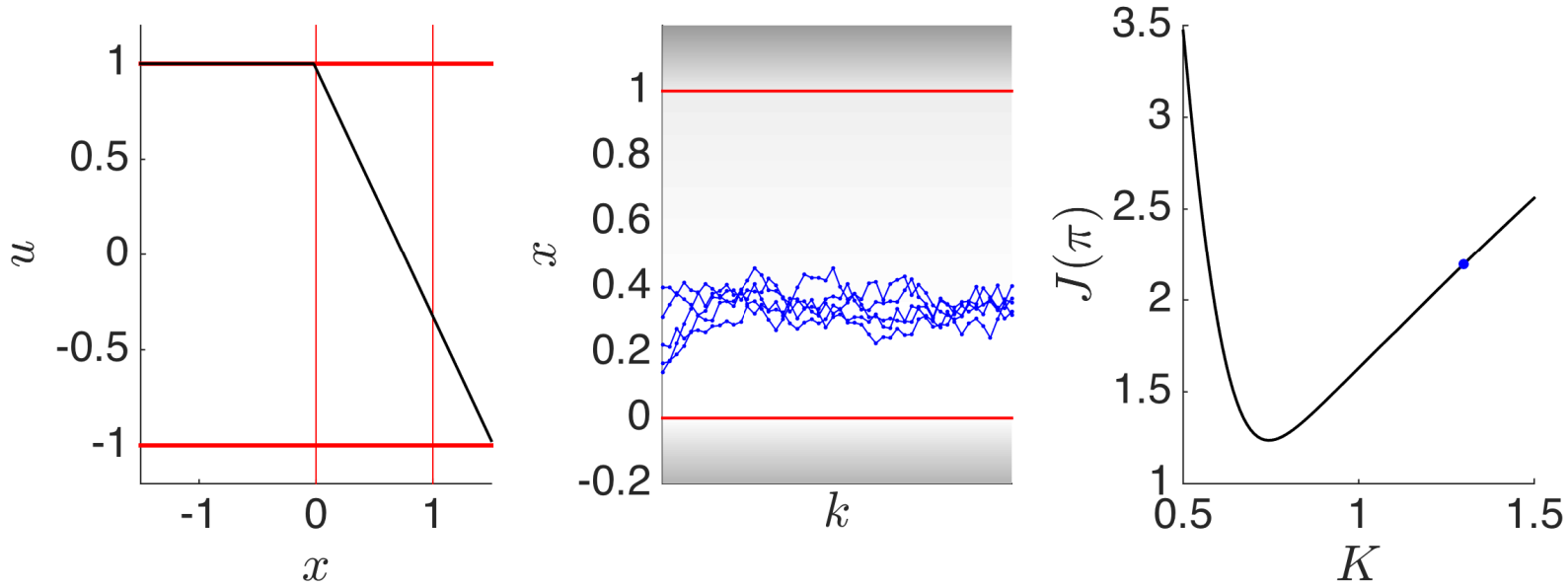
Policy performance - $J(\pi)$

- Policy: $\pi(\mathbf{x}) = \text{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\text{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\text{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_π, value function V_π
- Yields policy performance $J(\pi) = \mathbb{E}[V_\pi(\mathbf{x}_0)] = \mathbb{E}_{\tau_\pi}[L(\mathbf{x}, \pi(\mathbf{x}))]$



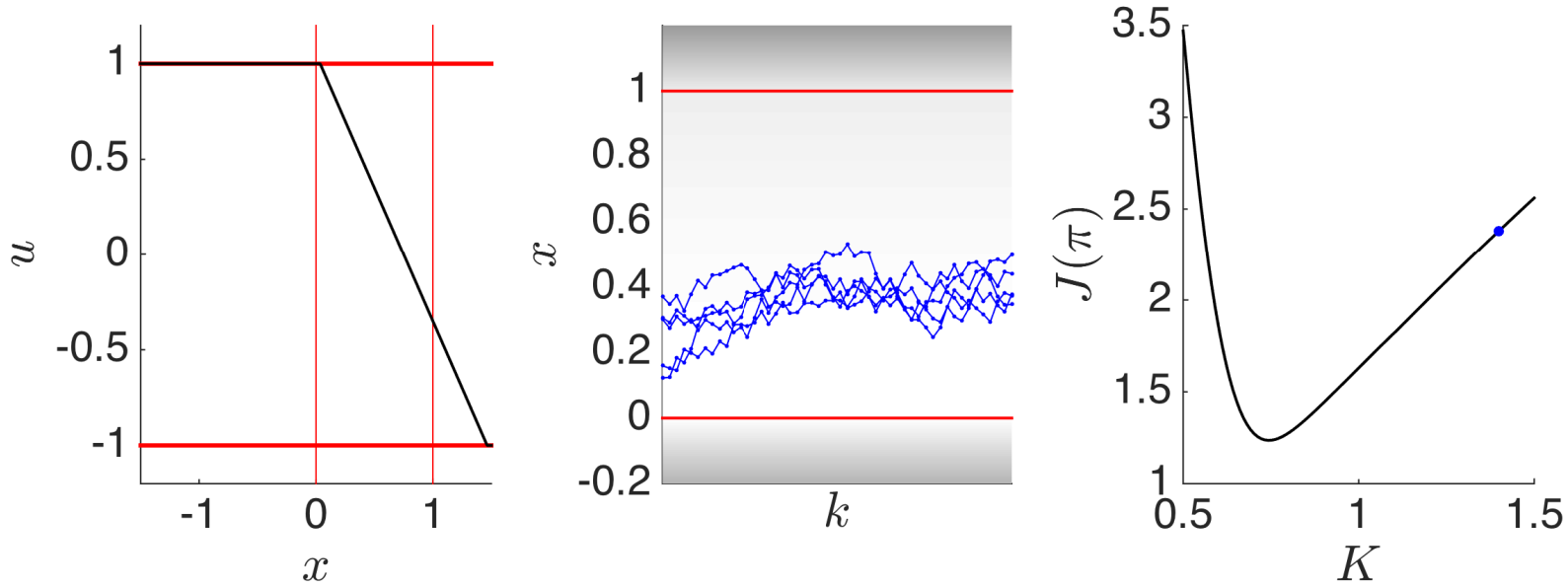
Policy performance - $J(\pi)$

- Policy: $\pi(\mathbf{x}) = \text{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\text{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\text{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_π, value function V_π
- Yields policy performance $J(\pi) = \mathbb{E}\left[V_\pi(\mathbf{x}_0)\right] = \mathbb{E}_{\tau_\pi}\left[L(\mathbf{x}, \pi(\mathbf{x}))\right]$



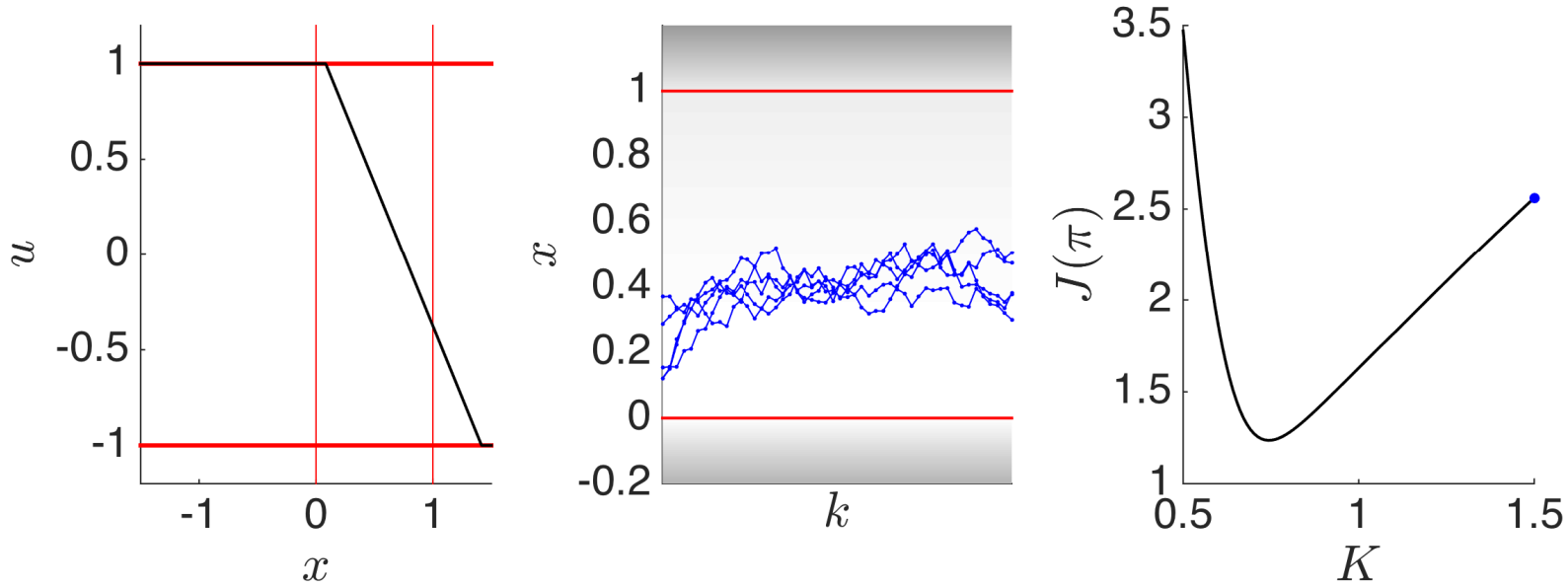
Policy performance - $J(\pi)$

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_π, value function V_π
- Yields policy performance $J(\pi) = \mathbb{E}\left[V_\pi(\mathbf{x}_0)\right] = \mathbb{E}_{\tau_\pi}\left[L(\mathbf{x}, \pi(\mathbf{x}))\right]$



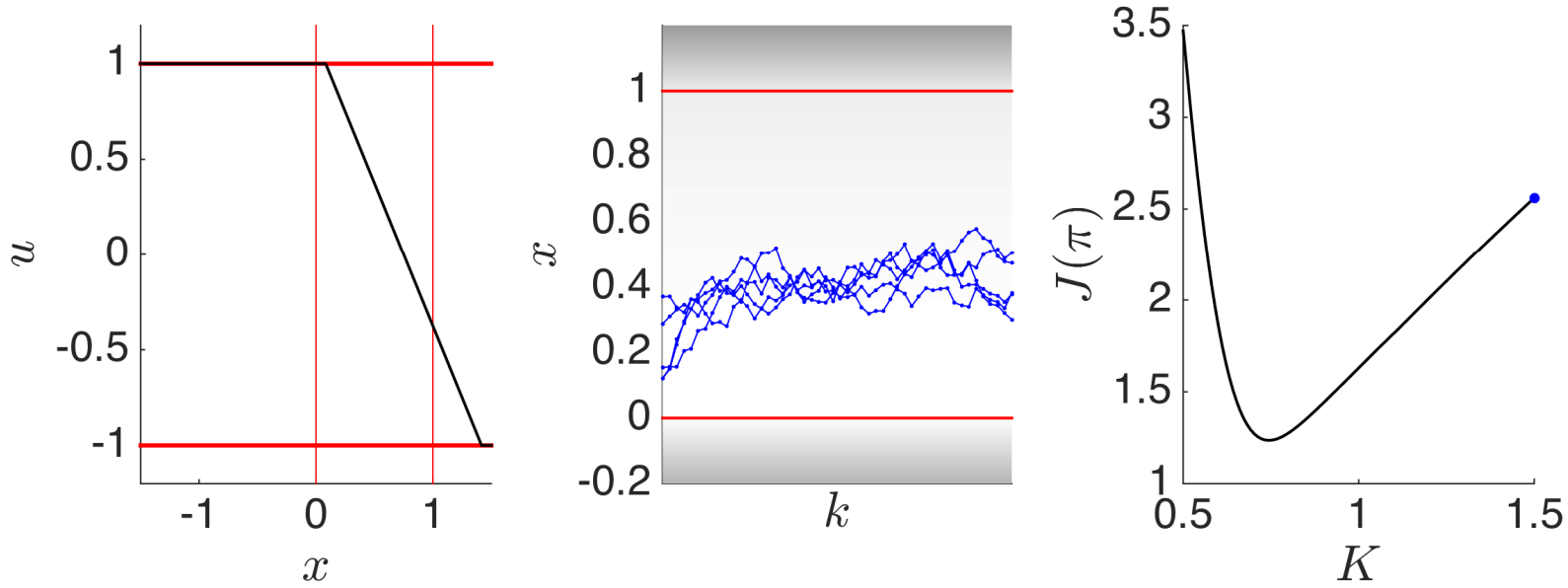
Policy performance - $J(\pi)$

- Policy: $\pi(\mathbf{x}) = \mathrm{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\mathrm{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\mathrm{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_π, value function V_π
- Yields policy performance $J(\pi) = \mathbb{E}[V_\pi(\mathbf{x}_0)] = \mathbb{E}_{\tau_\pi}[L(\mathbf{x}, \pi(\mathbf{x}))]$



Policy performance - $J(\pi)$

- Policy: $\pi(\mathbf{x}) = \text{mid}(\underline{\mathbf{u}}, -K(\mathbf{x} - \mathbf{x}_{\text{ref}}), \bar{\mathbf{u}})$, with $\mathbf{x}_{\text{ref}} = 0.75$
- Initial condition $\mathbf{x}_0 \sim \mathcal{N}(0.25, 0.1)$
- Yields MDP distribution τ_π, value function V_π
- Yields policy performance $J(\pi) = \mathbb{E}\left[V_\pi(\mathbf{x}_0)\right] = \mathbb{E}_{\tau_\pi}\left[L(\mathbf{x}, \pi(\mathbf{x}))\right]$



Optimal $\boldsymbol{\theta}$ must satisfy: $\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) = 0$

Outline

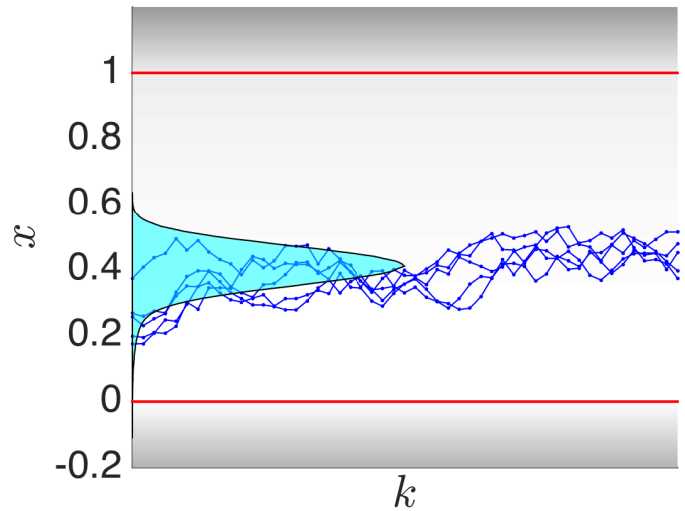
Deterministic Policy gradient theorem

How to compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$?

- changing $\boldsymbol{\theta}$ changes both $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ and $\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$
- policy gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$ is non-trivial...

Reminder:

$$J(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$



Deterministic Policy gradient theorem

How to compute $\nabla_\theta J(\pi_\theta)$?

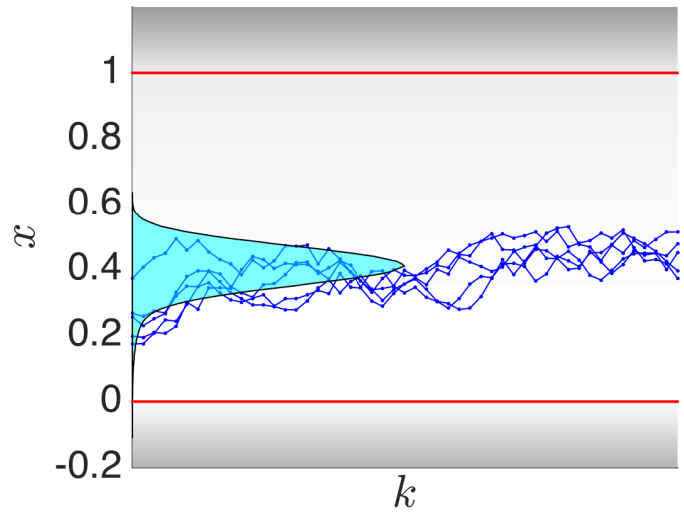
- changing θ changes both π_θ and τ_{π_θ}
- policy gradient $\nabla_\theta J(\pi_\theta)$ is non-trivial...

Policy gradient theorem

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau_{\pi_\theta}} \left[\nabla_\theta \pi_\theta(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\pi_\theta}(\mathbf{x}, \pi_\theta) \right]$$

Reminder:

$$J(\pi) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \pi(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \pi(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$



Deterministic Policy gradient theorem

How to compute $\nabla_\theta J(\pi_\theta)$?

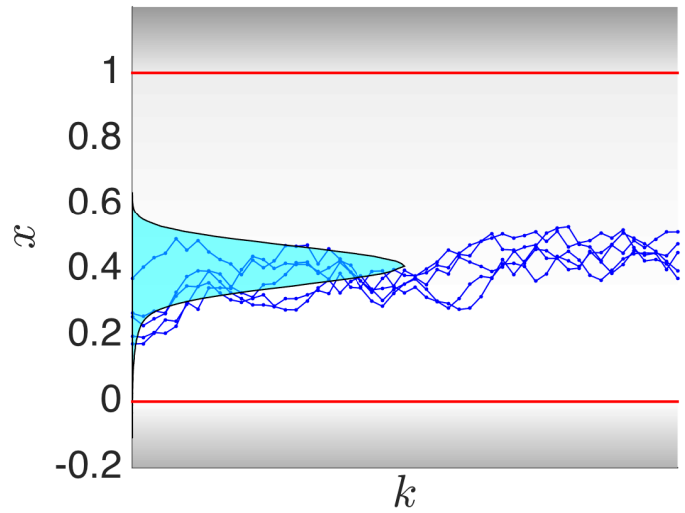
- changing θ changes both π_θ and τ_{π_θ}
- policy gradient $\nabla_\theta J(\pi_\theta)$ is non-trivial...

Policy gradient theorem

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau_{\pi_\theta}} \left[\nabla_\theta \pi_\theta(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\pi_\theta}(\mathbf{x}, \pi_\theta) \right]$$

Reminder:

$$J(\pi) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \pi(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \pi(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$



where $Q_{\pi_\theta}(\mathbf{x}, \mathbf{u})$ is the action-value function associated to policy π_θ i.e.

$$Q_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[V_{\pi_\theta}(\mathbf{x}_+) | \mathbf{x}, \mathbf{u}\right]$$
$$V_{\pi_\theta}(\mathbf{x}) = L(\mathbf{x}, \pi_\theta) + \gamma \mathbb{E}\left[V_{\pi_\theta}(\mathbf{x}_+) | \mathbf{x}, \pi_\theta\right]$$

Deterministic Policy gradient theorem

How to compute $\nabla_\theta J(\pi_\theta)$?

- changing θ changes both π_θ and τ_{π_θ}
- policy gradient $\nabla_\theta J(\pi_\theta)$ is non-trivial...

Policy gradient theorem

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau_{\pi_\theta}} \left[\nabla_\theta \pi_\theta(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\pi_\theta}(\mathbf{x}, \pi_\theta) \right]$$

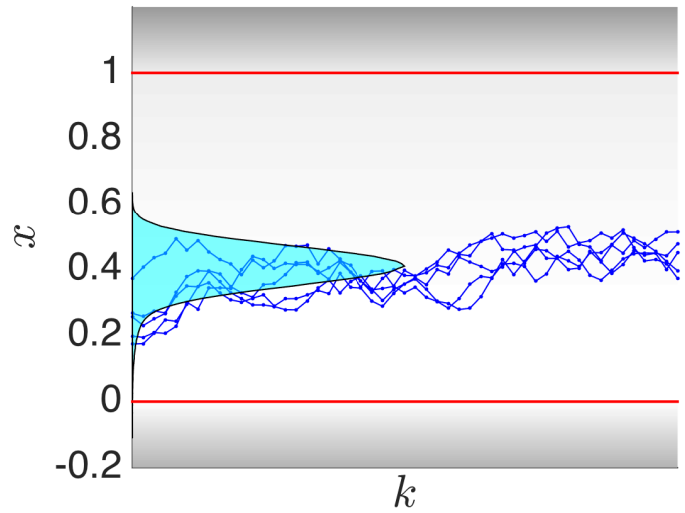
(i.e. we can do a chain-rule-like operation:

$$\nabla_\theta Q_{\pi_\theta}(\mathbf{x}, \pi_\theta) = \nabla_\theta \pi_\theta(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\pi_\theta}(\mathbf{x}, \pi_\theta)$$

and ignore τ_{π_θ})

Reminder:

$$J(\pi) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \pi(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \pi(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$



where $Q_{\pi_\theta}(\mathbf{x}, \mathbf{u})$ is the action-value function associated to policy π_θ i.e.

$$Q_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[V_{\pi_\theta}(\mathbf{x}_+) | \mathbf{x}, \mathbf{u}\right]$$
$$V_{\pi_\theta}(\mathbf{x}) = L(\mathbf{x}, \pi_\theta) + \gamma \mathbb{E}\left[V_{\pi_\theta}(\mathbf{x}_+) | \mathbf{x}, \pi_\theta\right]$$

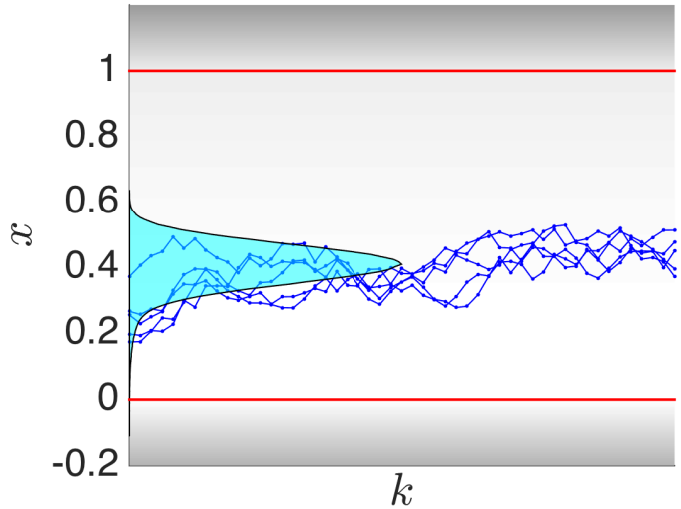
Deterministic Policy gradient theorem

Reminder:

$$J(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$



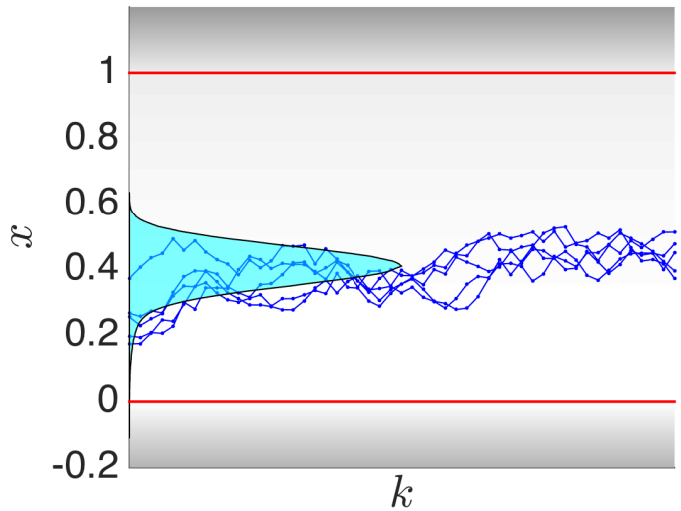
Deterministic Policy gradient theorem

Reminder:

$$J(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}}\right)\right]$$



Remarks

- Need $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ to be (almost everywhere) differentiable w.r.t. $\boldsymbol{\theta}$ (why would it not be?)
- Need $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ to be (almost everywhere) differentiable w.r.t. $\boldsymbol{\theta}$ (why would it not be?)
- Evaluating $\nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}}\right)$ from data requires departing from policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$
- Requires some regularity assumptions (non-trivial, not discussed here)
- Measure theory may kick in (because of $\mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}}$)

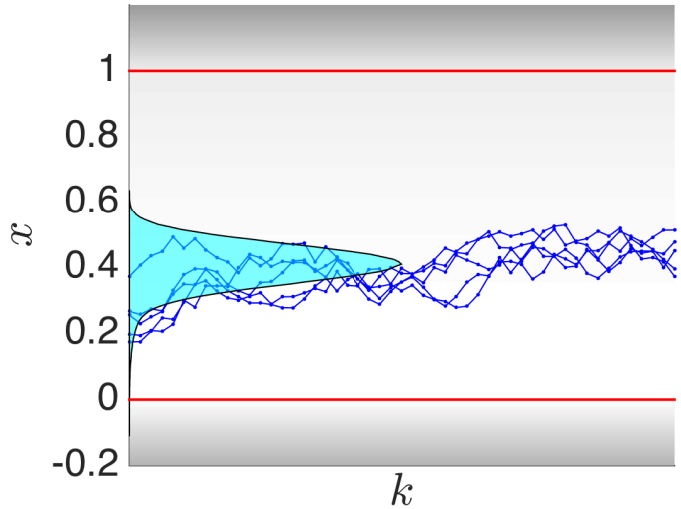
Deterministic Policy gradient theorem

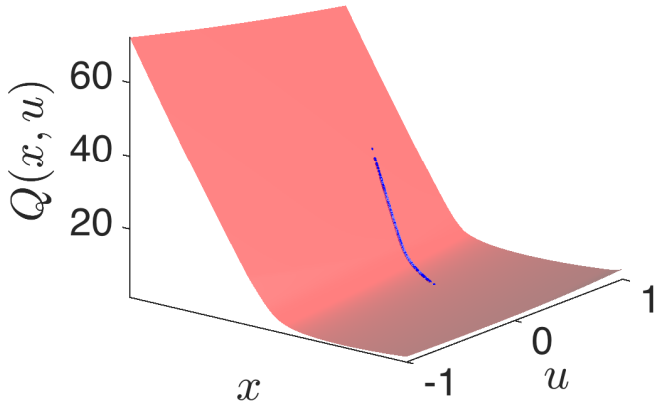
Reminder:

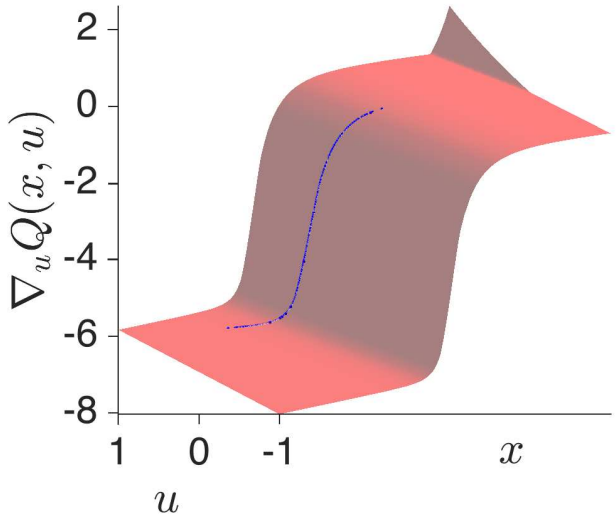
$$J(\pi) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \pi(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \pi(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$







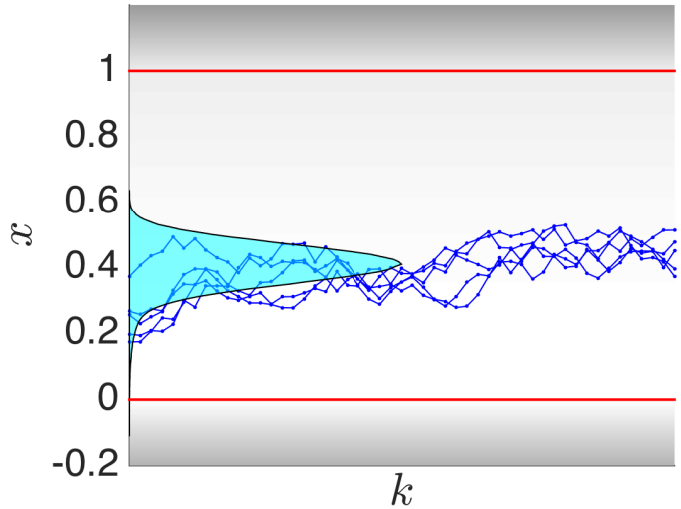
Deterministic Policy gradient theorem

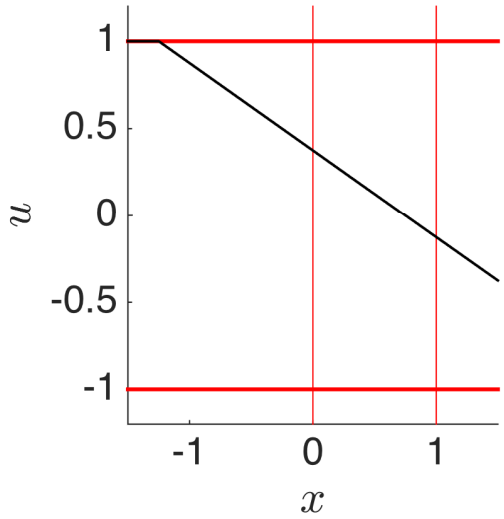
Reminder:

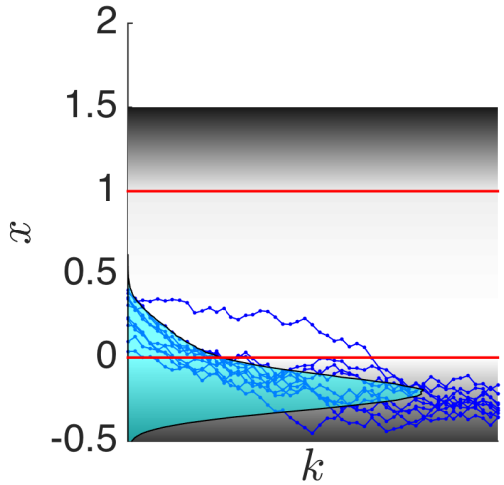
$$J(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

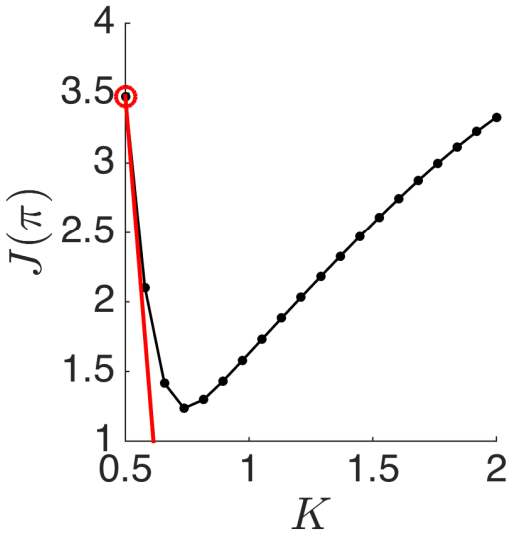
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









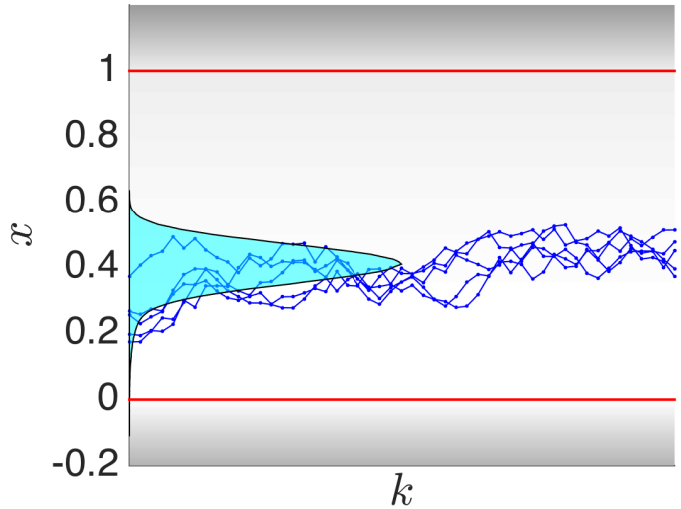
Deterministic Policy gradient theorem

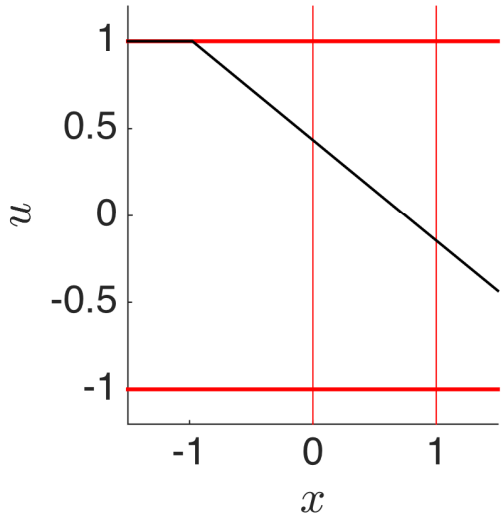
Reminder:

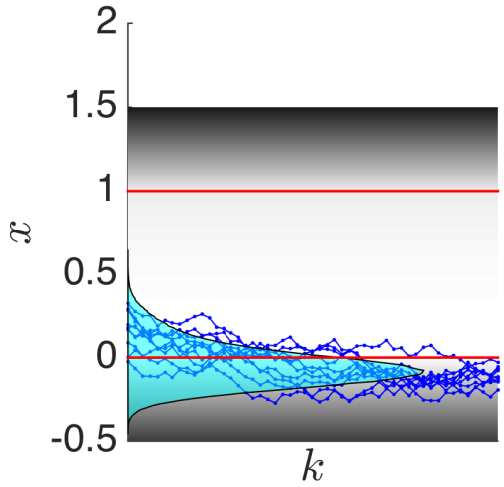
$$J(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

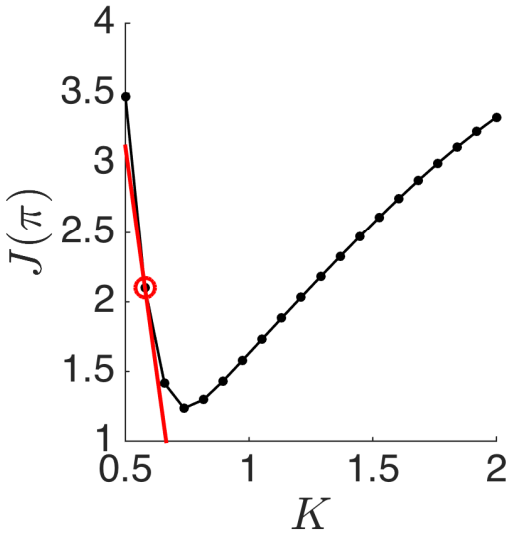
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









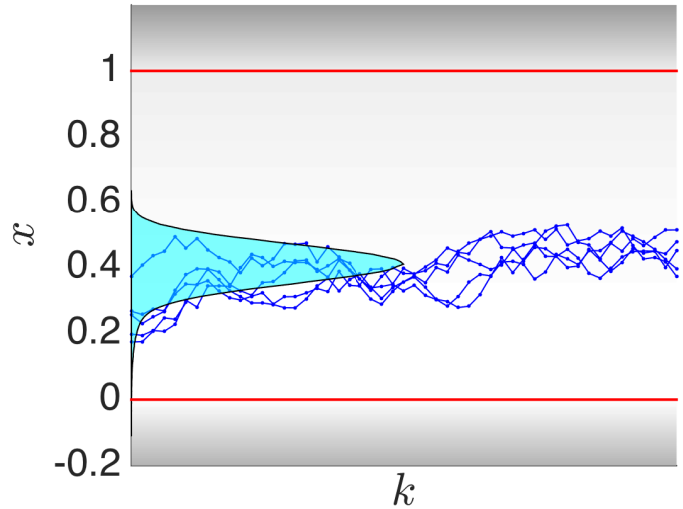
Deterministic Policy gradient theorem

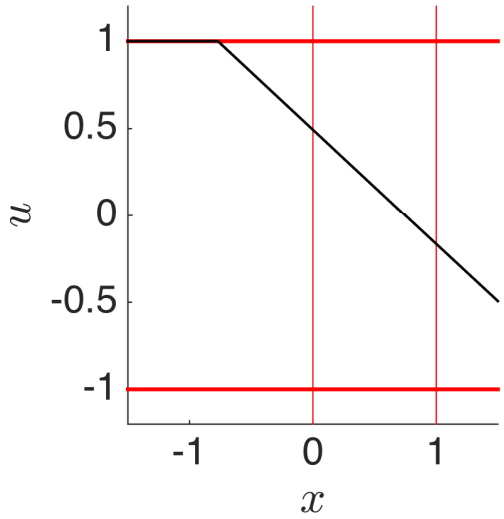
Reminder:

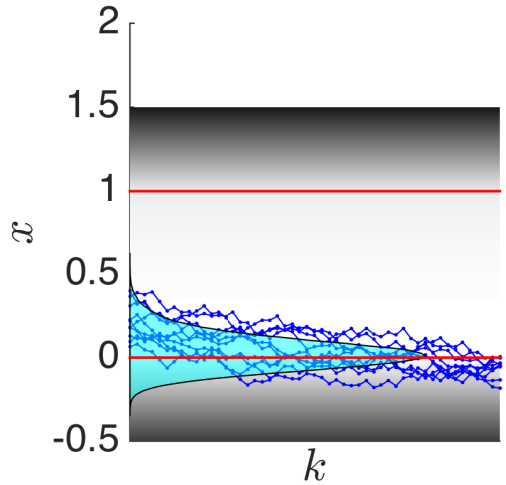
$$J(\pi) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \pi(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \pi(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

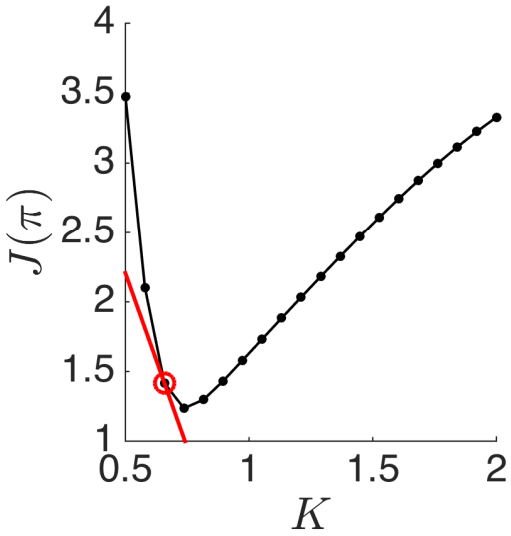
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









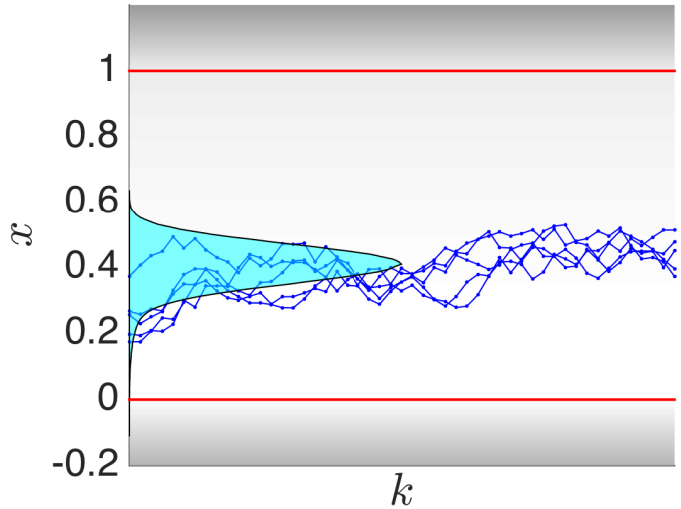
Deterministic Policy gradient theorem

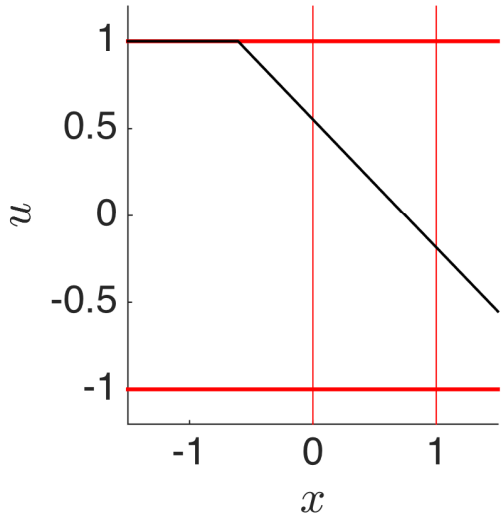
Reminder:

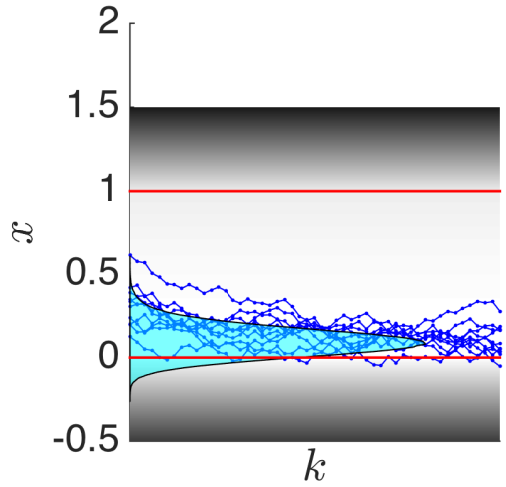
$$J(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

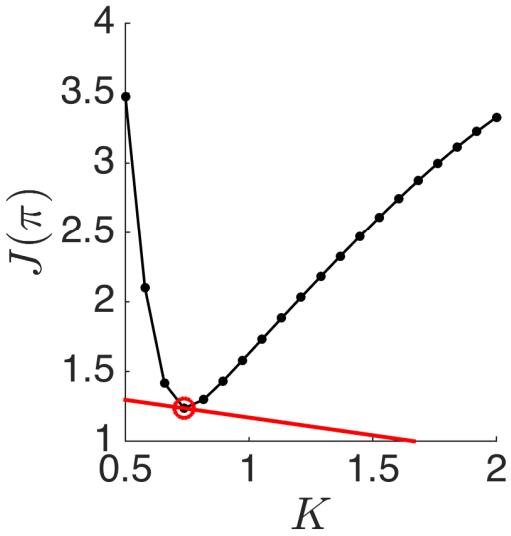
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









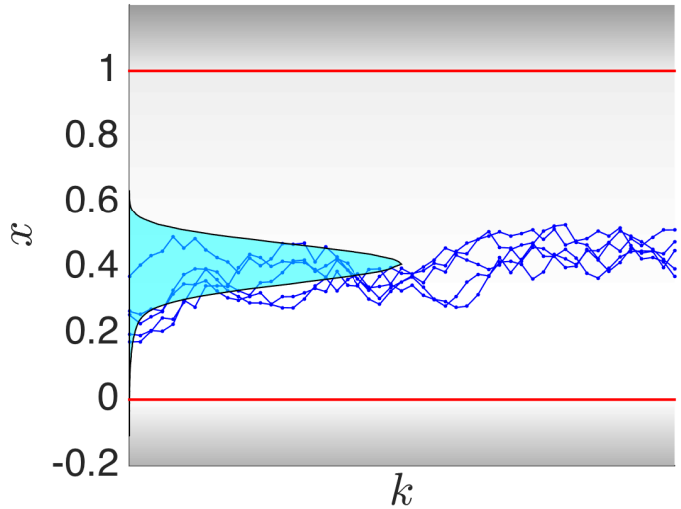
Deterministic Policy gradient theorem

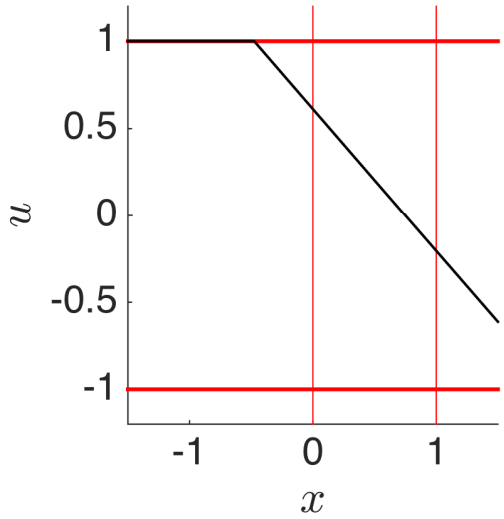
Reminder:

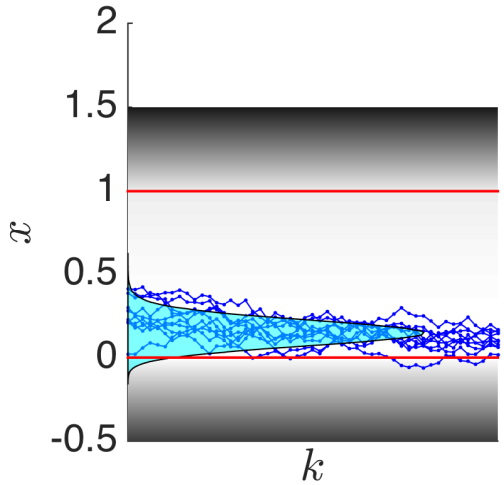
$$J(\pi) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \pi(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \pi(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

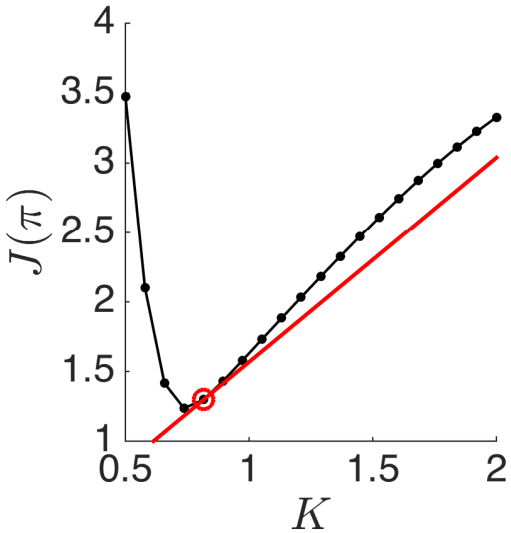
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









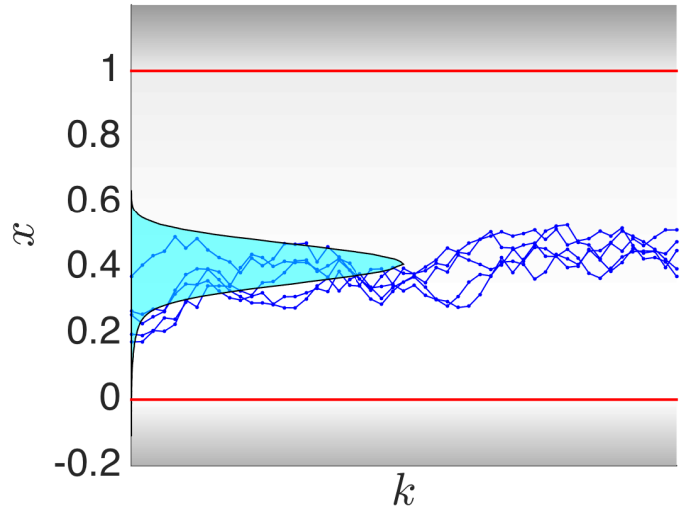
Deterministic Policy gradient theorem

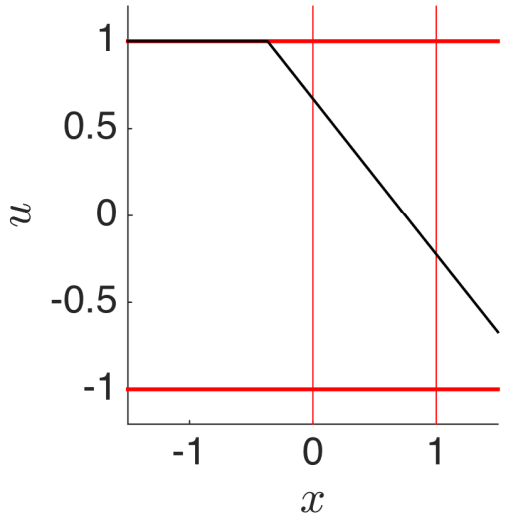
Reminder:

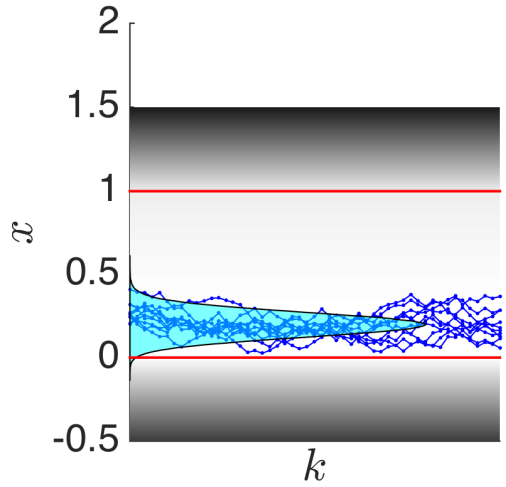
$$J(\pi) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \pi(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \pi(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

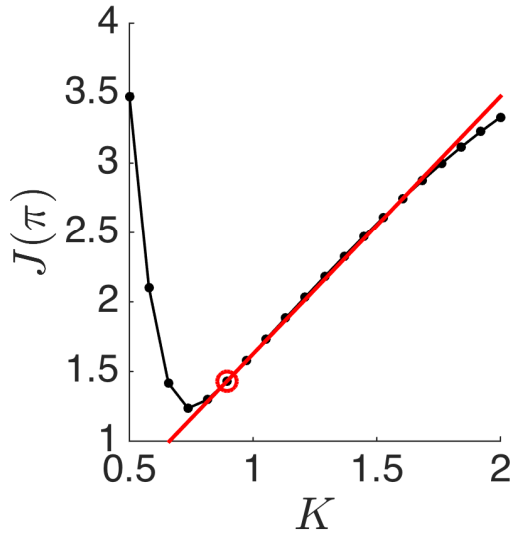
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









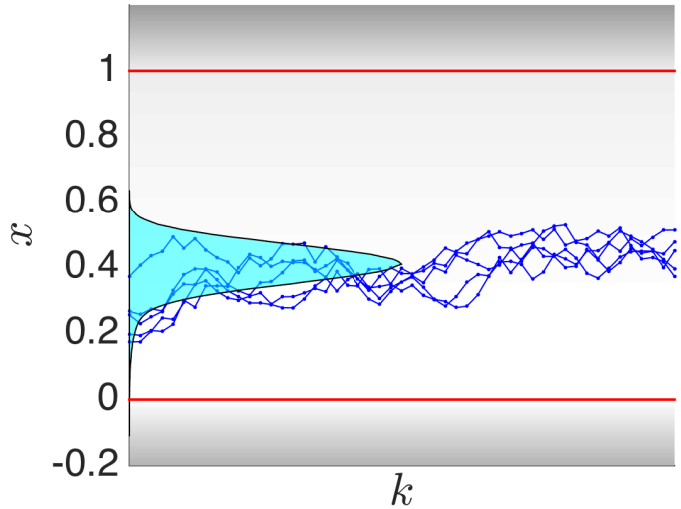
Deterministic Policy gradient theorem

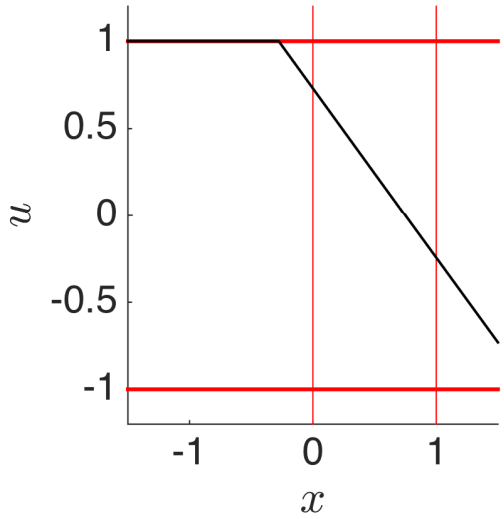
Reminder:

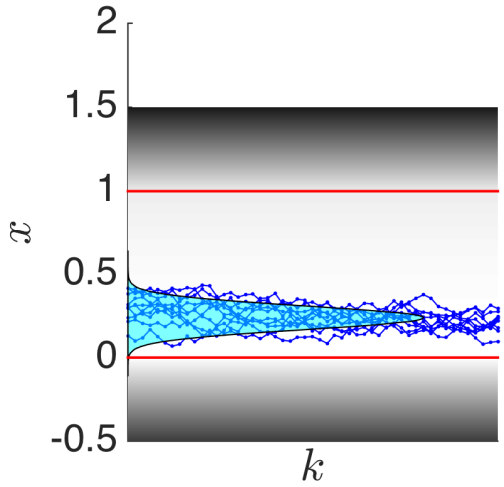
$$J(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

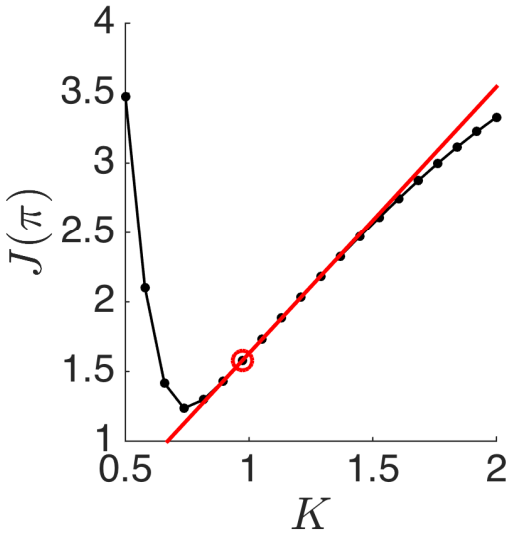
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









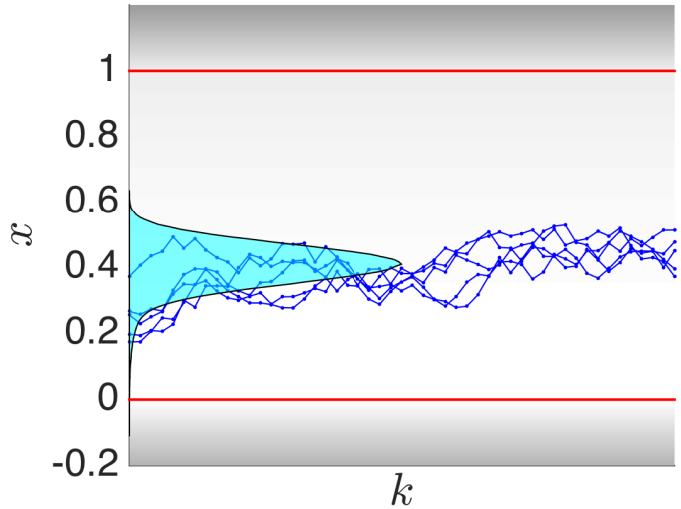
Deterministic Policy gradient theorem

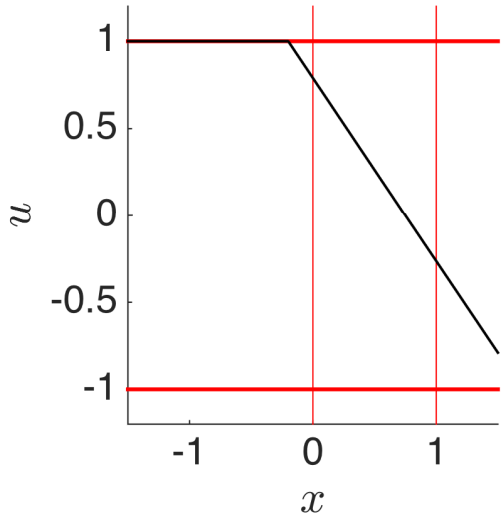
Reminder:

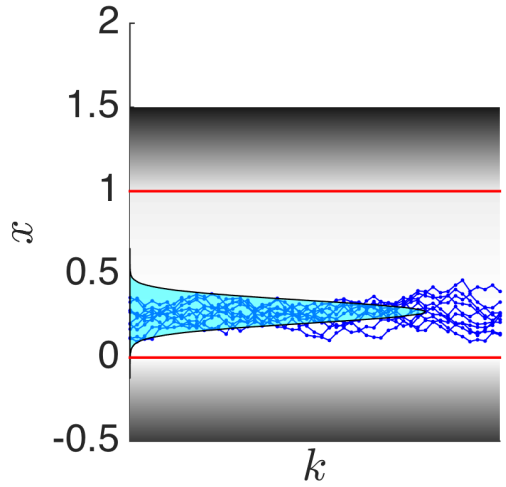
$$J(\pi) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \pi(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \pi(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

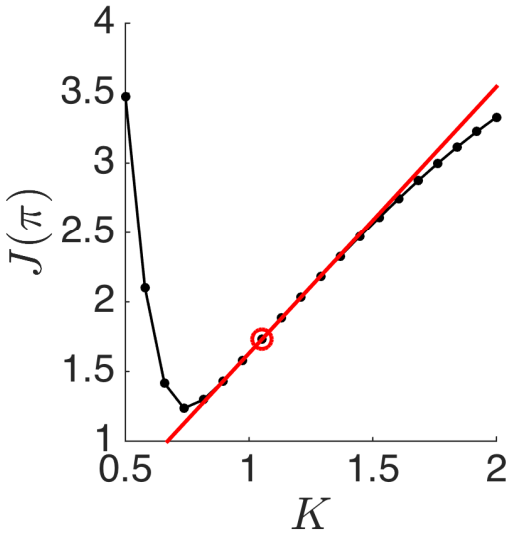
Policy gradient theorem

$$\nabla_{\theta} J(\pi_{\theta}) = \mathbb{E}_{\tau_{\pi_{\theta}}} \left[\nabla_{\theta} \pi_{\theta}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\pi_{\theta}}(\mathbf{x}, \pi_{\theta})\right]$$









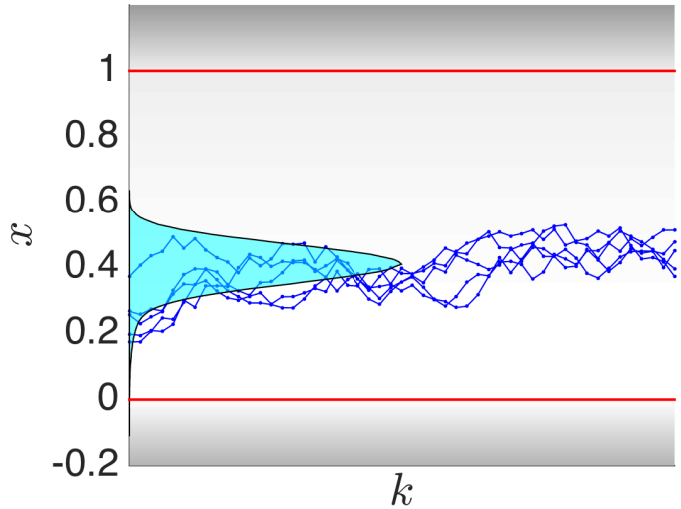
Deterministic Policy gradient theorem

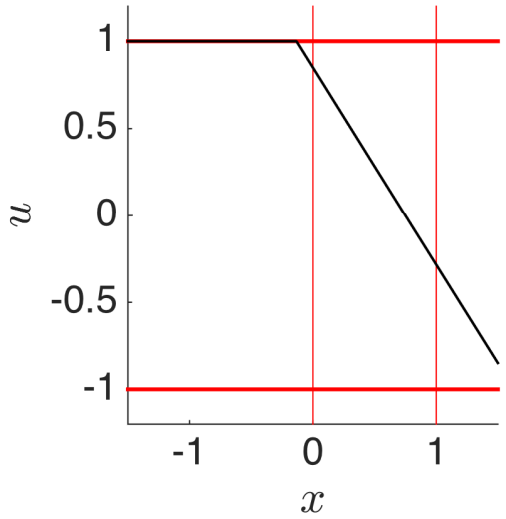
Reminder:

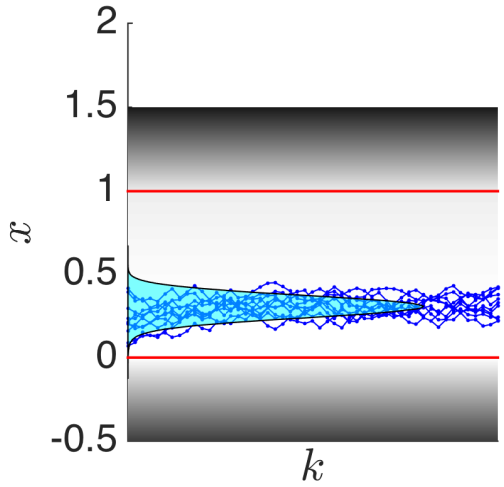
$$J(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

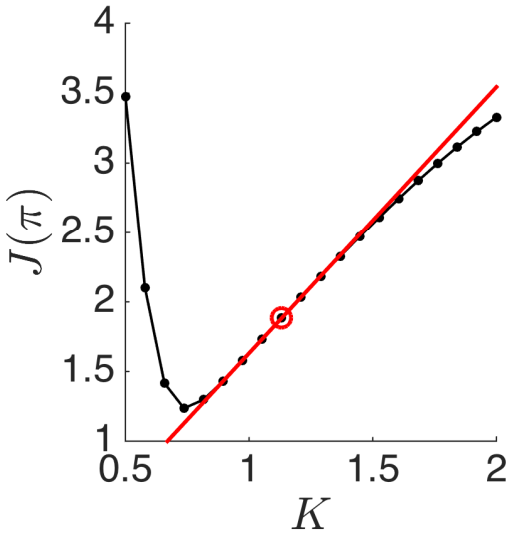
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









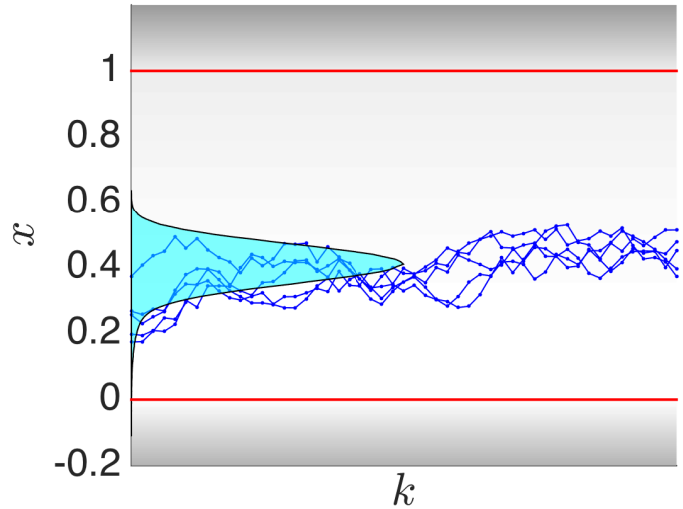
Deterministic Policy gradient theorem

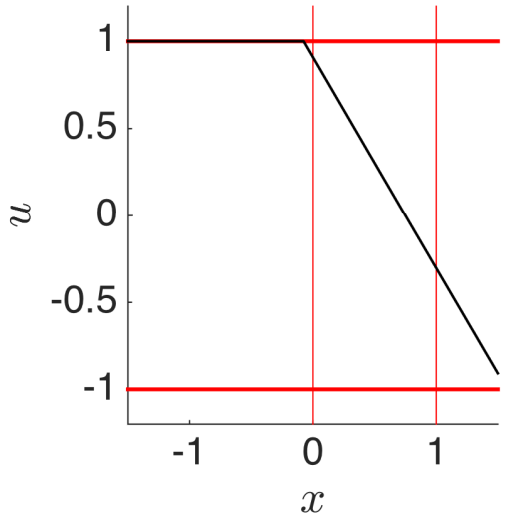
Reminder:

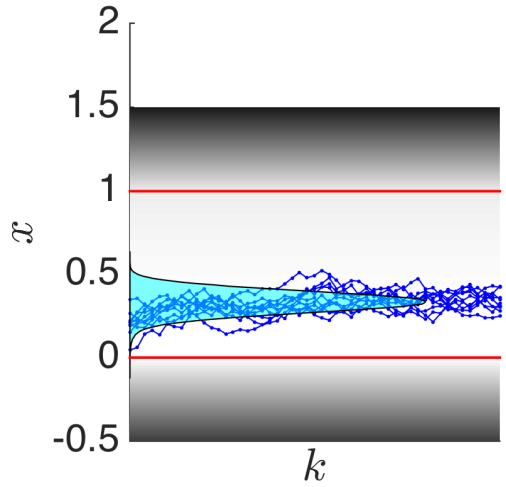
$$J(\pi) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \pi(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \pi(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

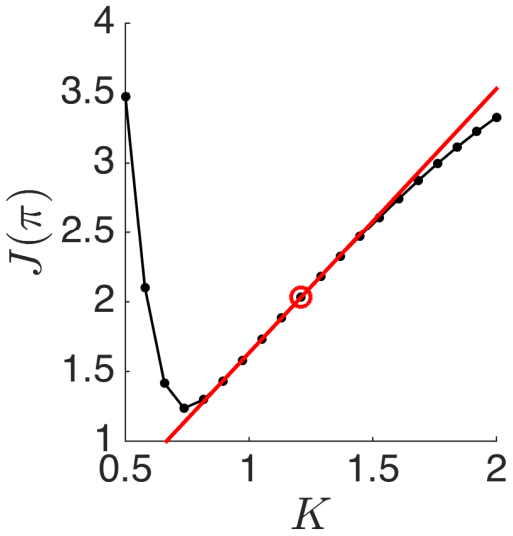
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









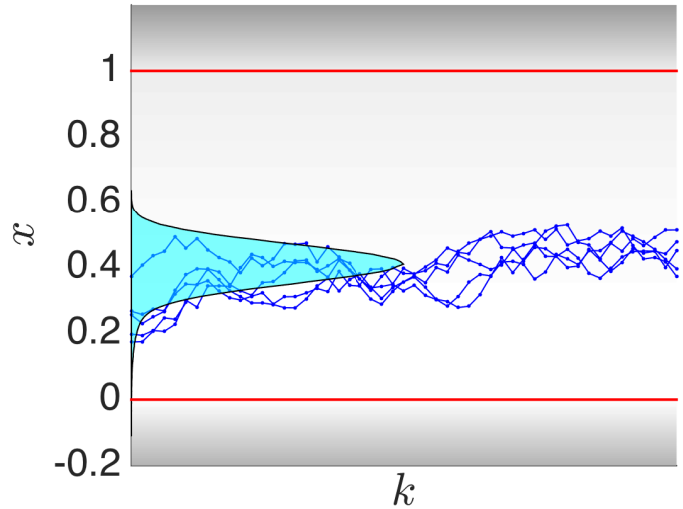
Deterministic Policy gradient theorem

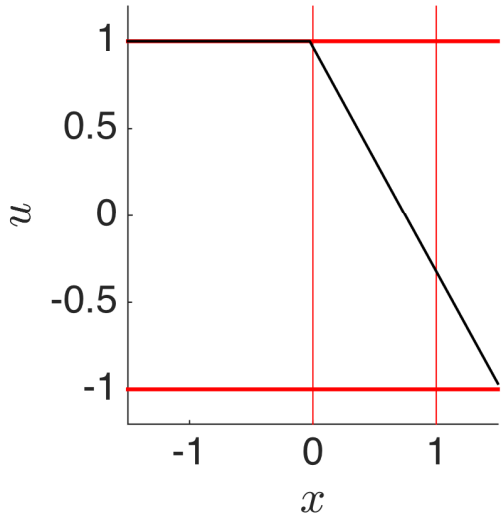
Reminder:

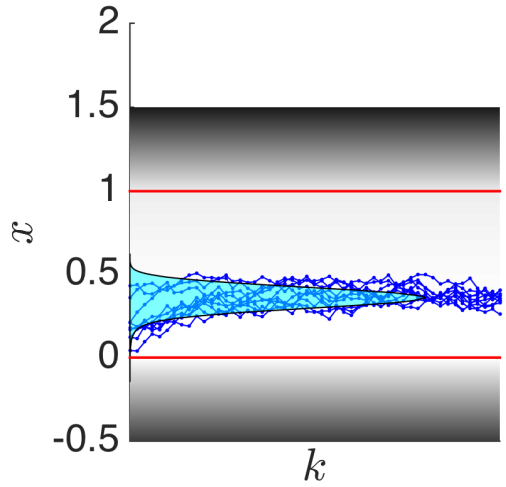
$$J(\pi) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \pi(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \pi(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

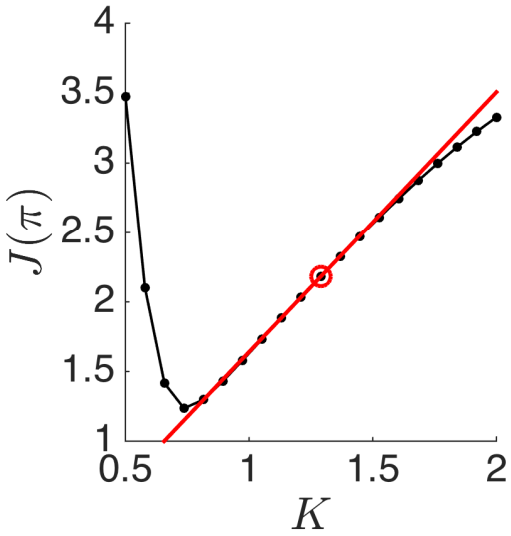
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









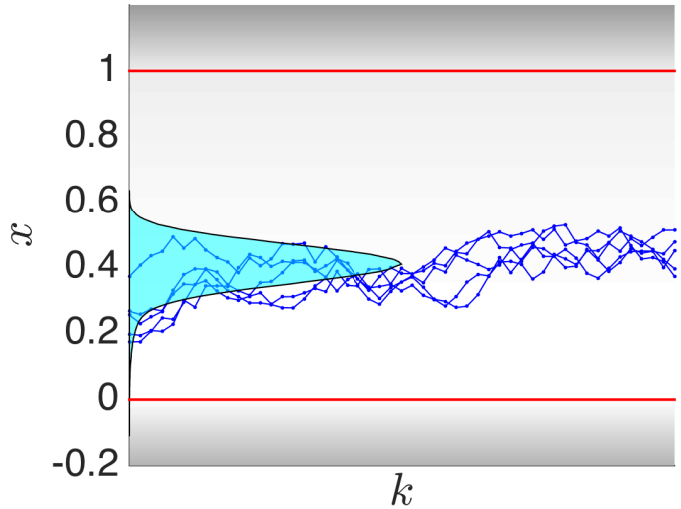
Deterministic Policy gradient theorem

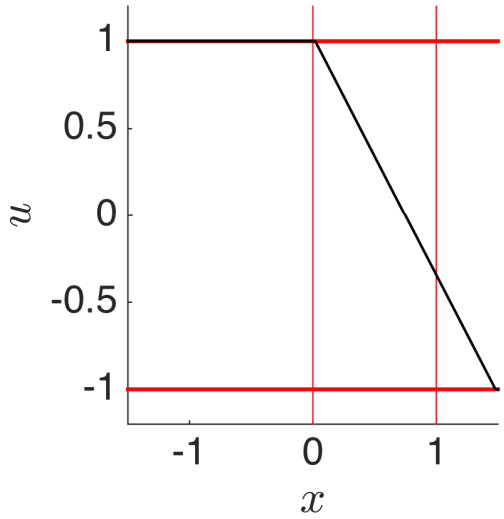
Reminder:

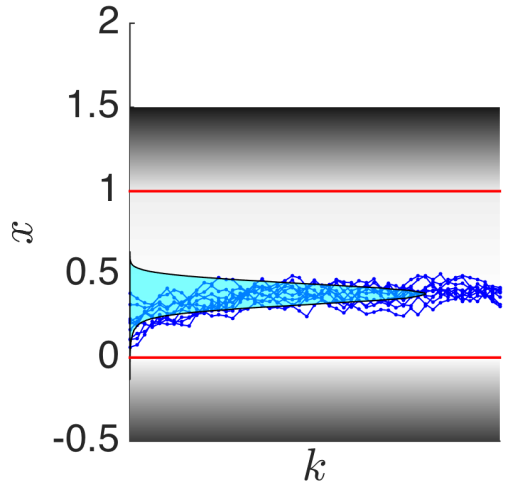
$$J(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

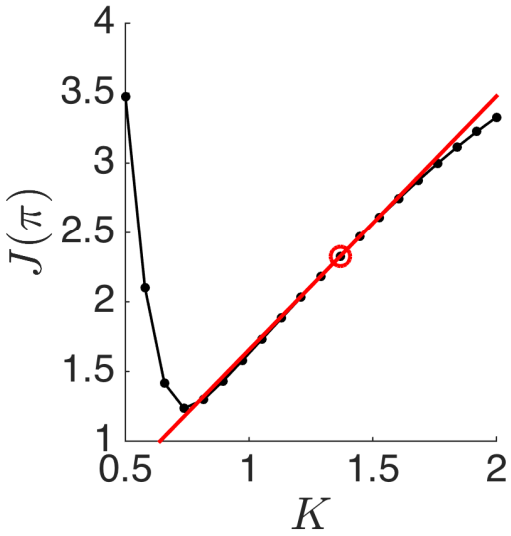
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









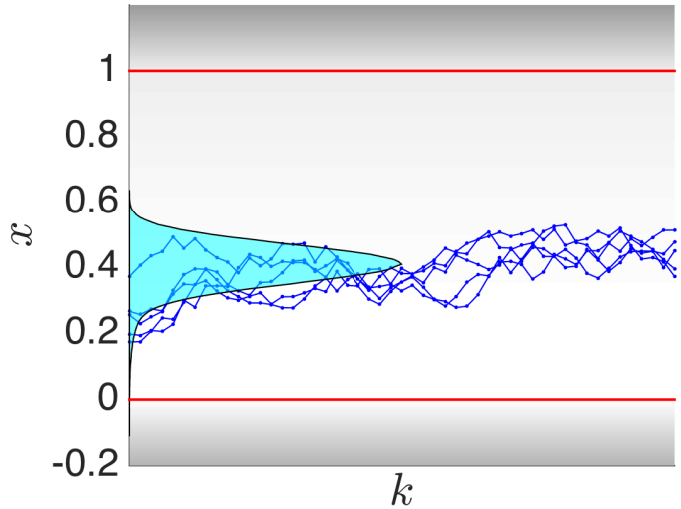
Deterministic Policy gradient theorem

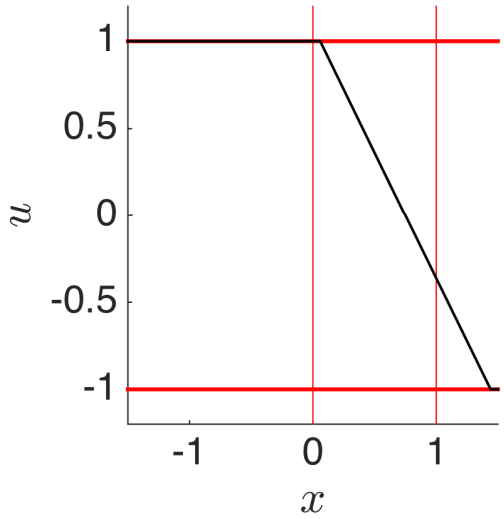
Reminder:

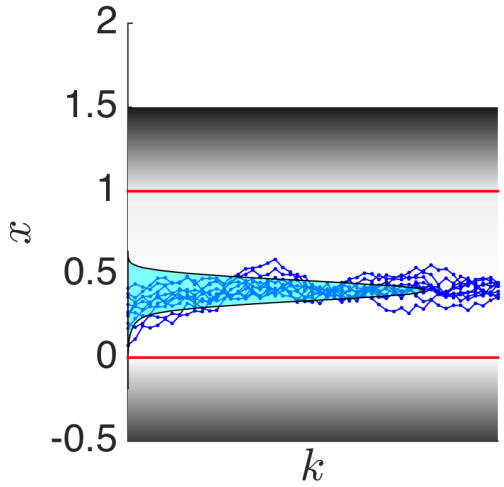
$$J(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

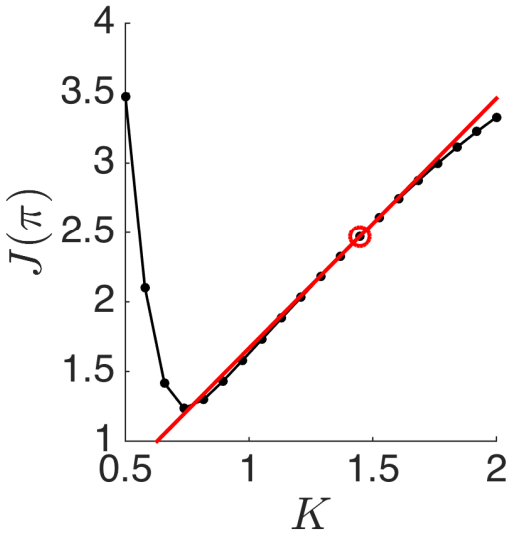
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









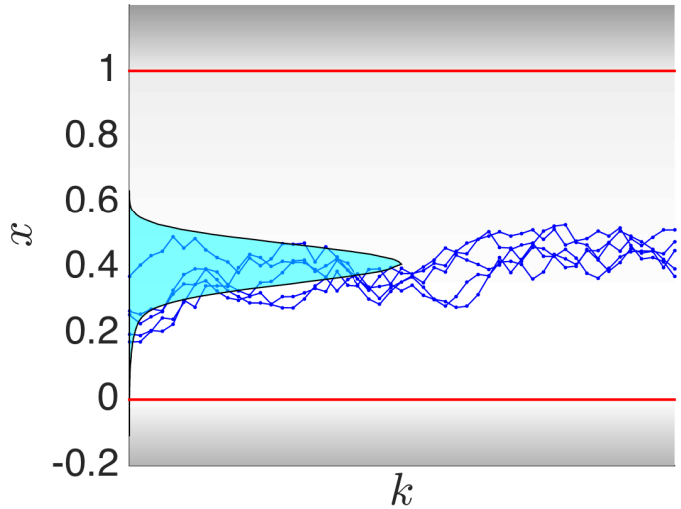
Deterministic Policy gradient theorem

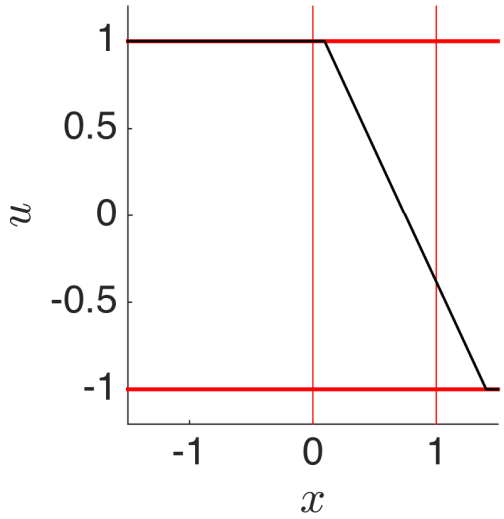
Reminder:

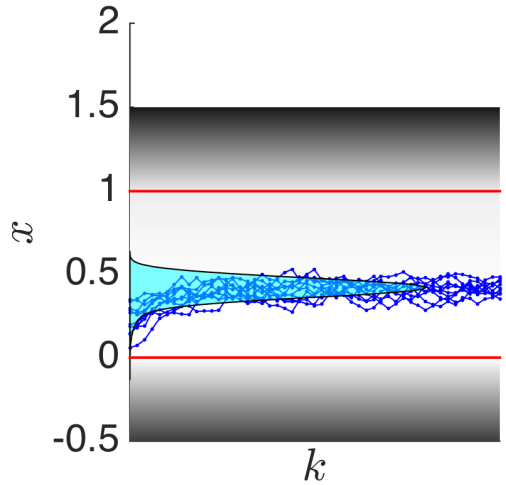
$$J(\pi) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \pi(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \pi(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

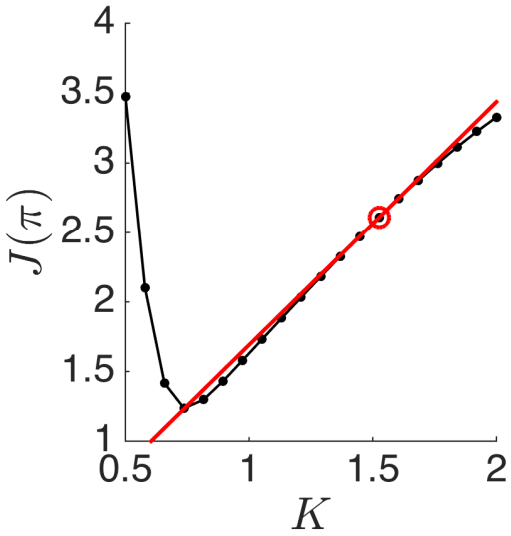
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









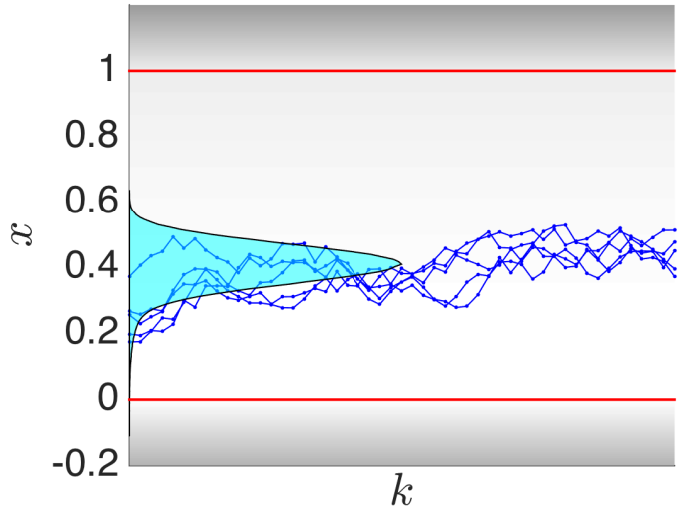
Deterministic Policy gradient theorem

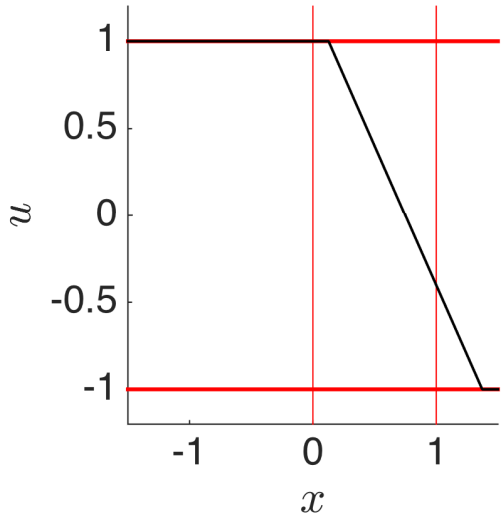
Reminder:

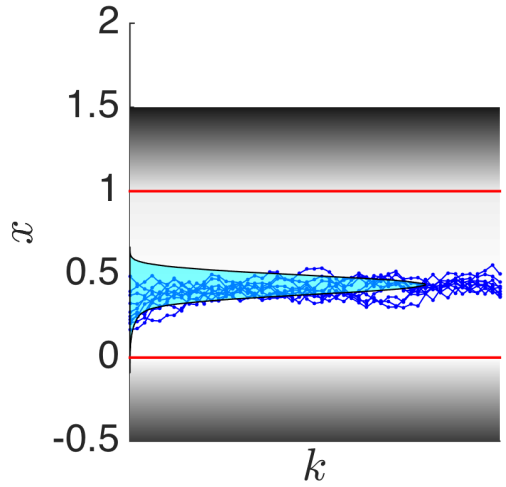
$$J(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

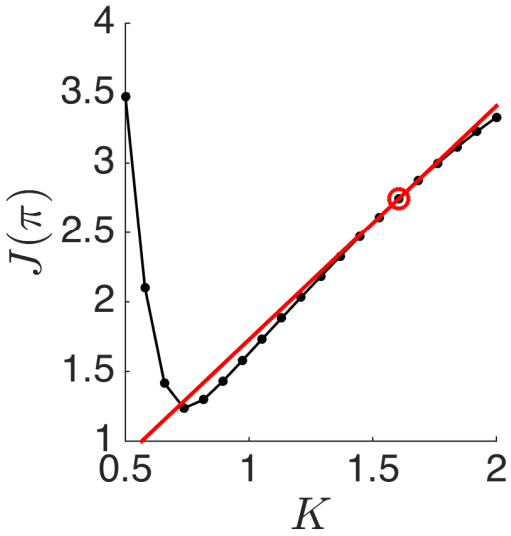
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









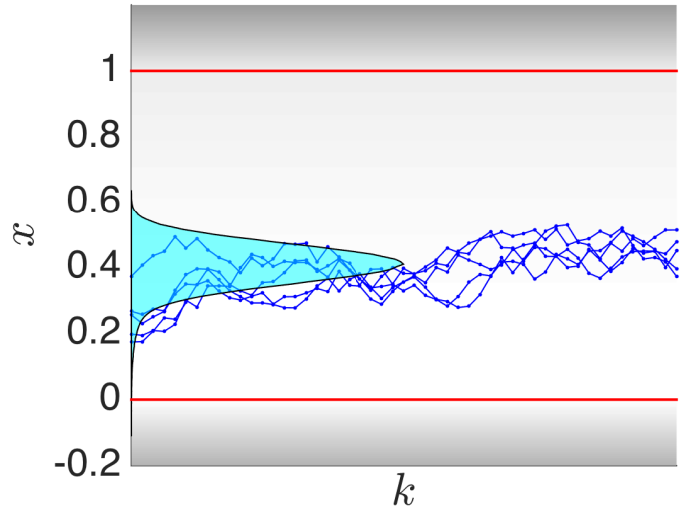
Deterministic Policy gradient theorem

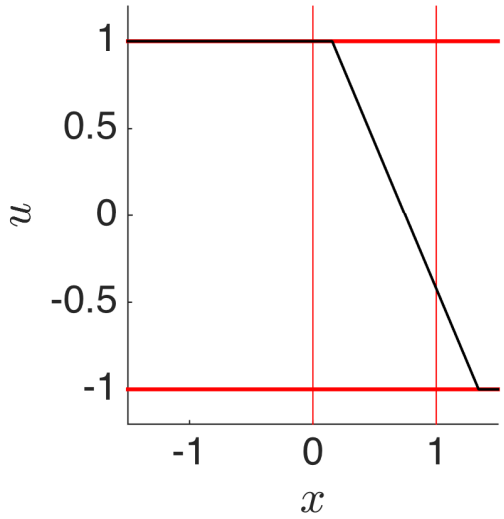
Reminder:

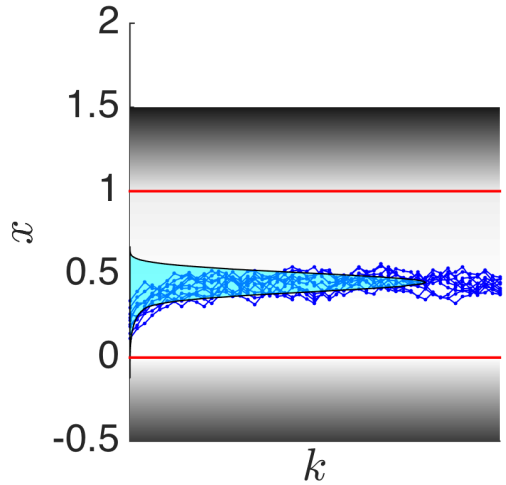
$$J(\pi) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \pi(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \pi(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

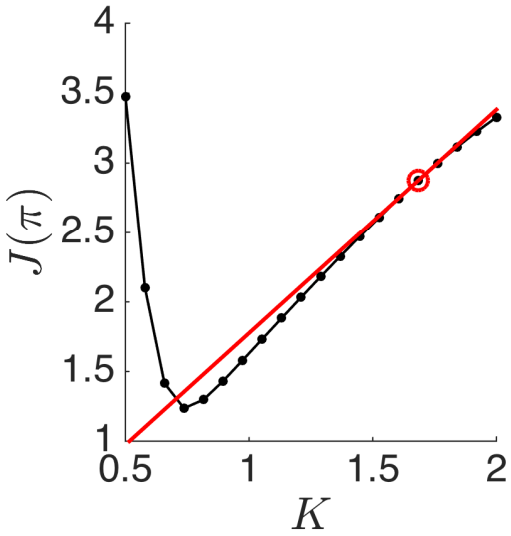
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









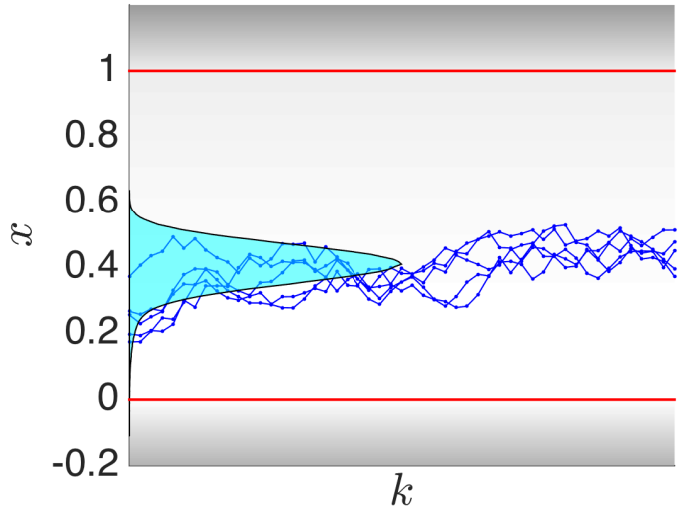
Deterministic Policy gradient theorem

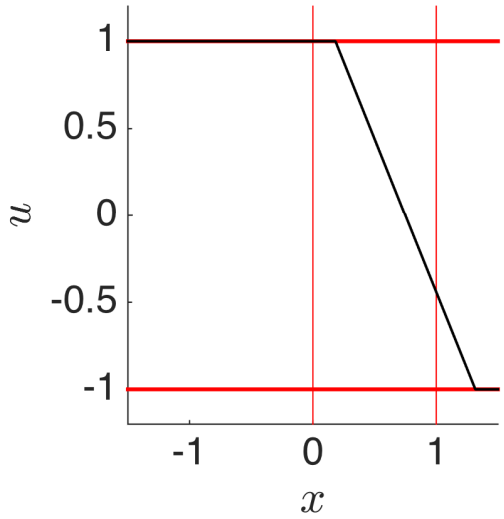
Reminder:

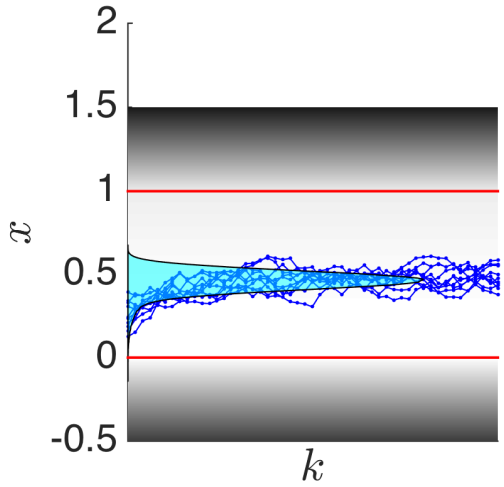
$$J(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

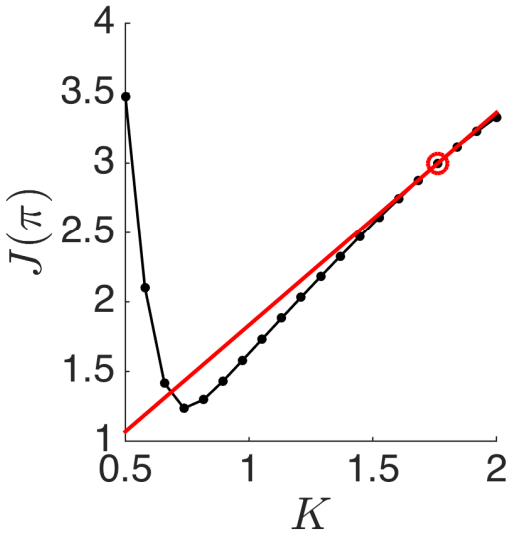
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









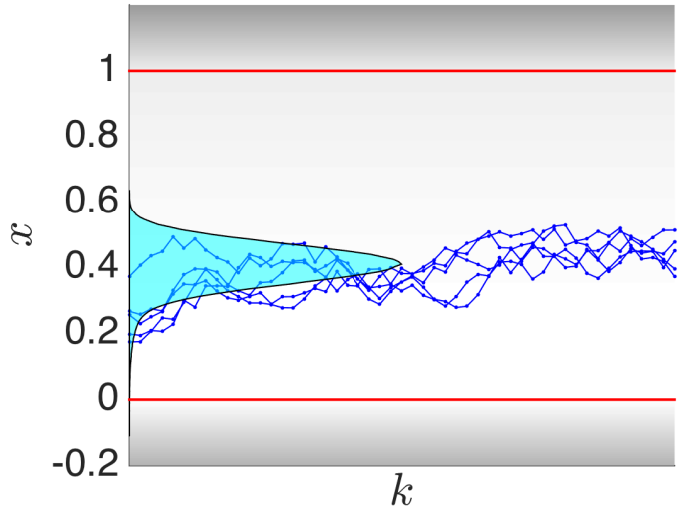
Deterministic Policy gradient theorem

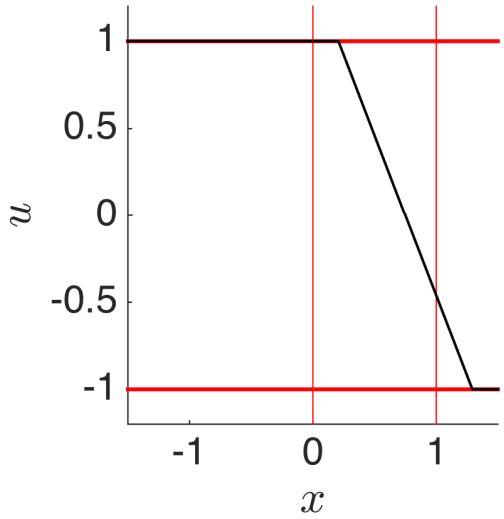
Reminder:

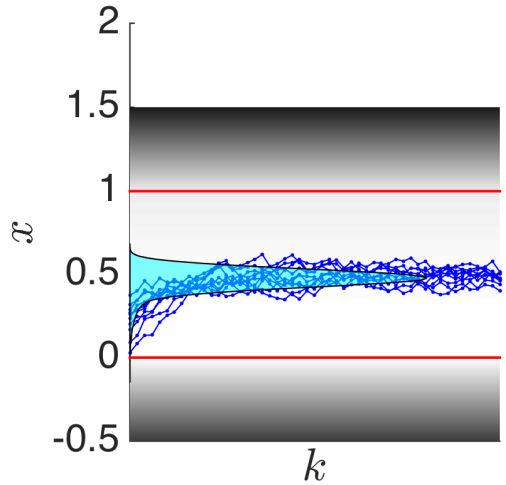
$$J(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

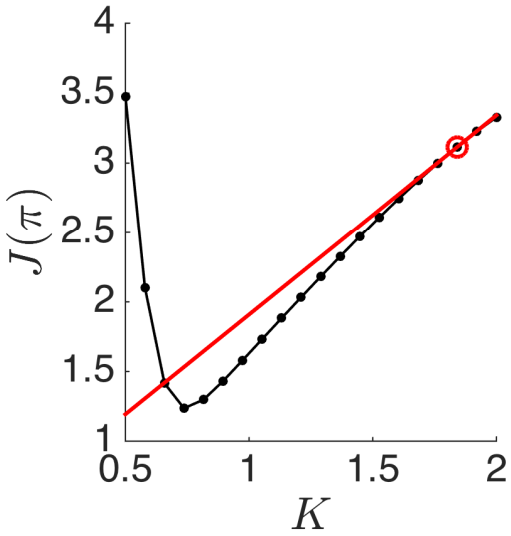
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









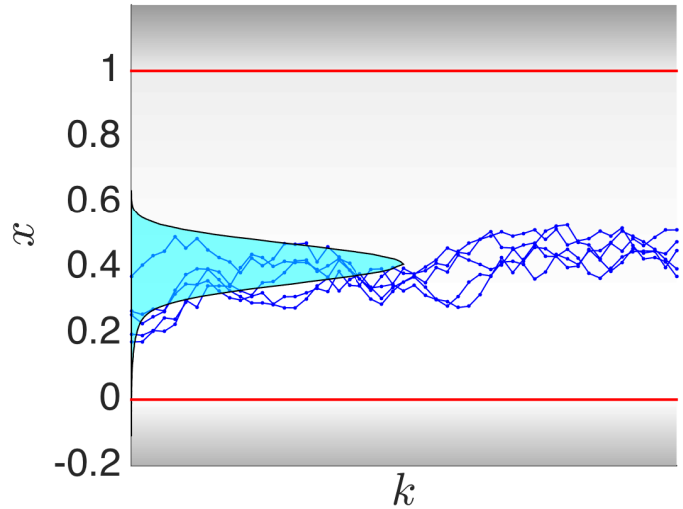
Deterministic Policy gradient theorem

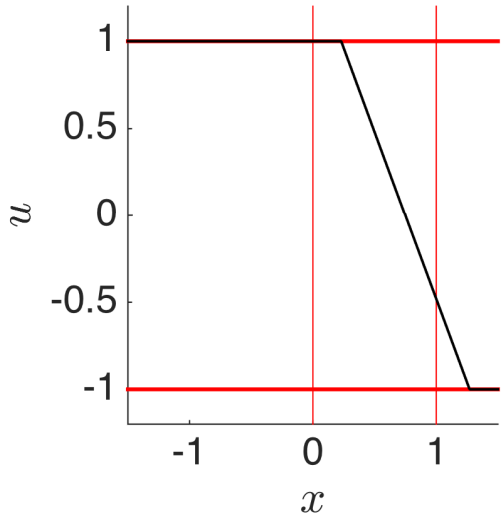
Reminder:

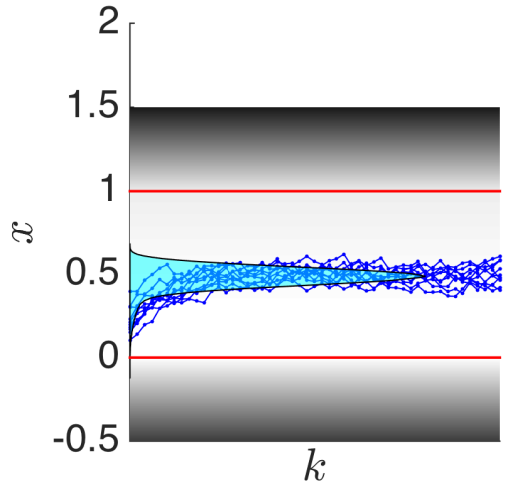
$$J(\pi) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \pi(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \pi(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

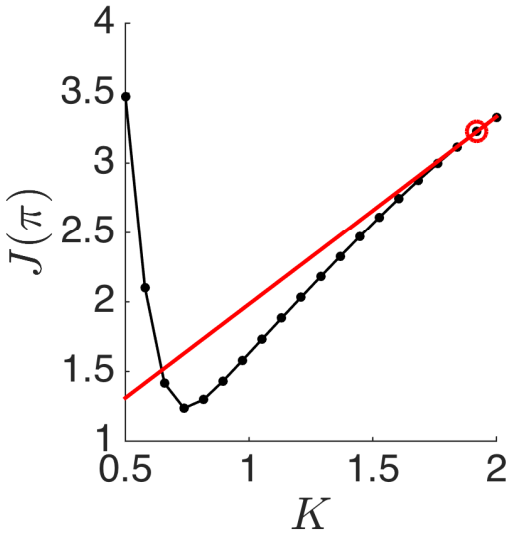
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









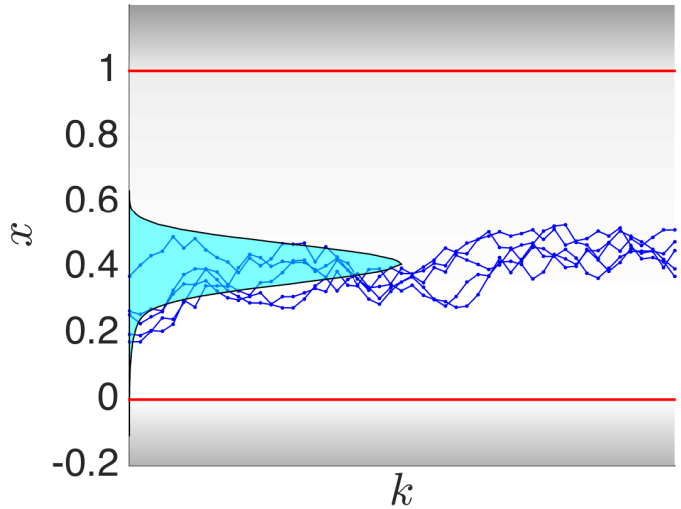
Deterministic Policy gradient theorem

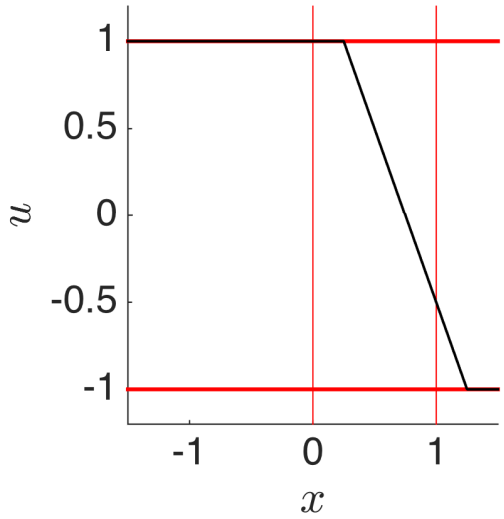
Reminder:

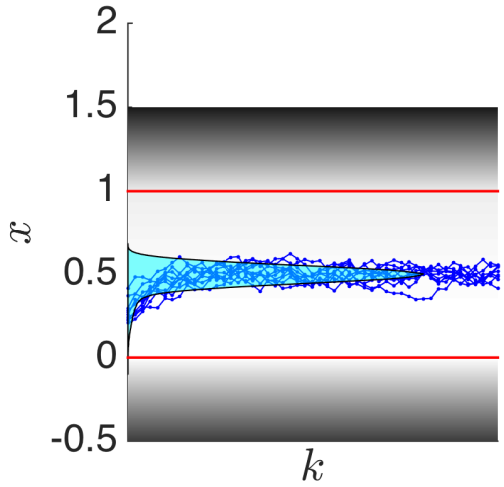
$$J(\boldsymbol{\pi}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i-1} | \mathbf{x}_i, \boldsymbol{\pi}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0\ldots k-1}$$

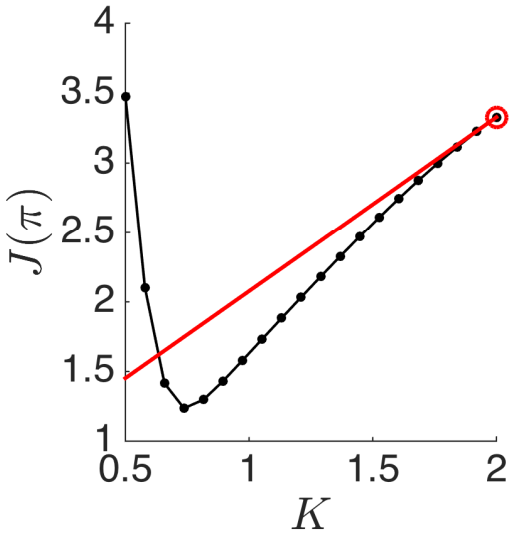
Policy gradient theorem

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$$









Deterministic policy gradient

Algorithm: Policy gradient (prototype)

Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

- Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)
- Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.
- Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}}) \right]$
- Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$

Deterministic policy gradient

Algorithm: Policy gradient (prototype)

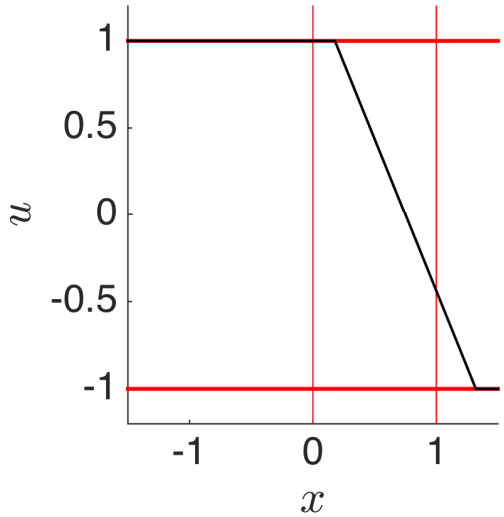
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

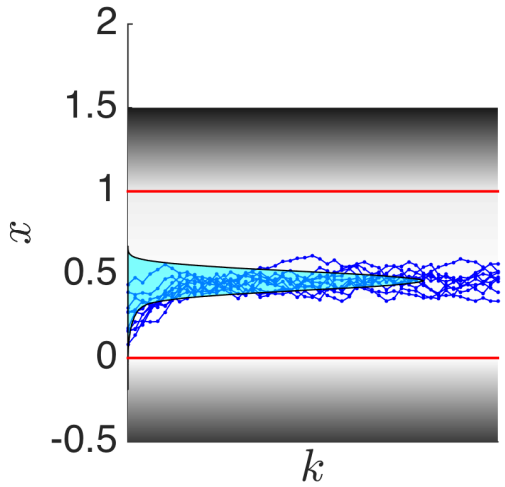
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

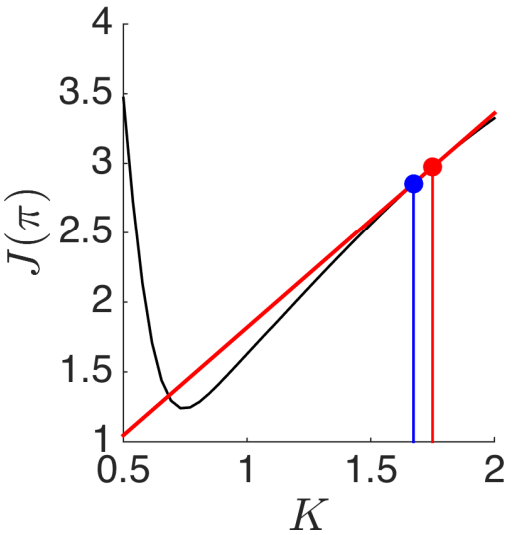
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

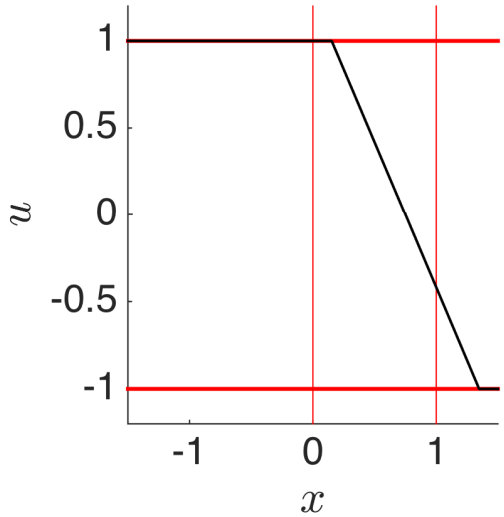
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

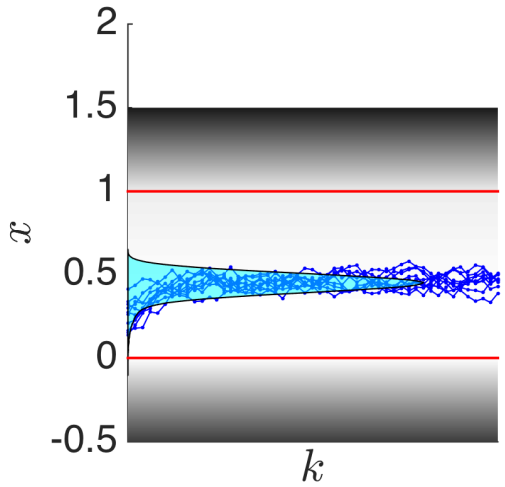
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

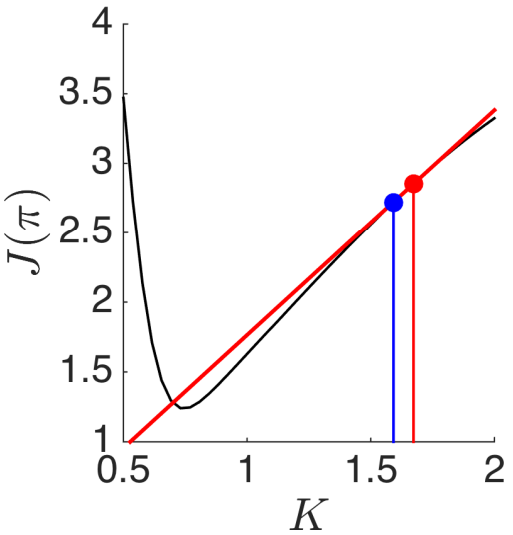
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

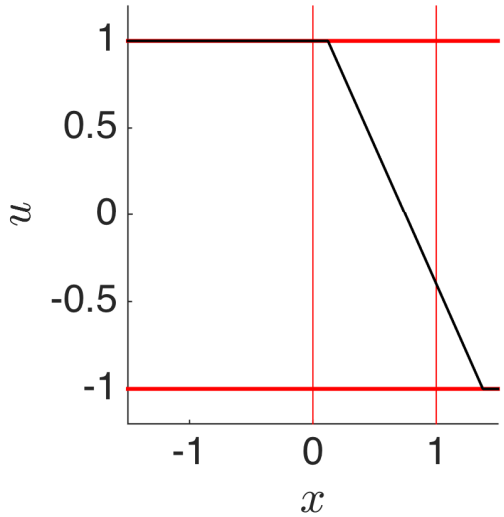
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

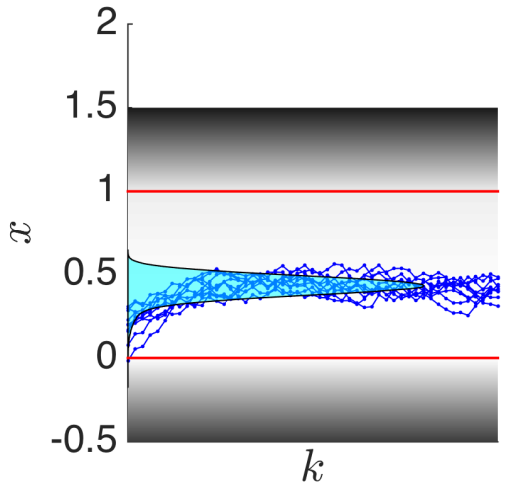
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

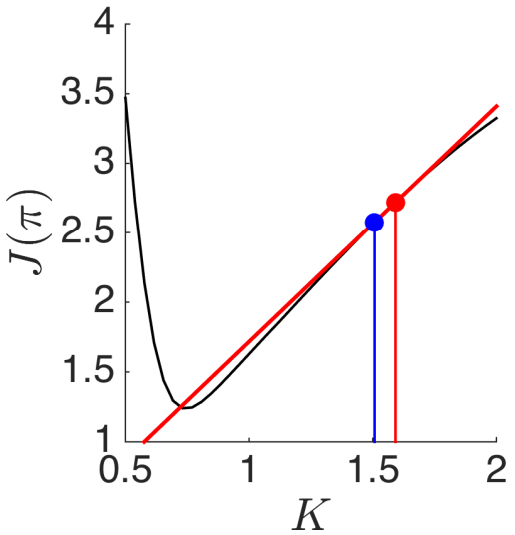
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

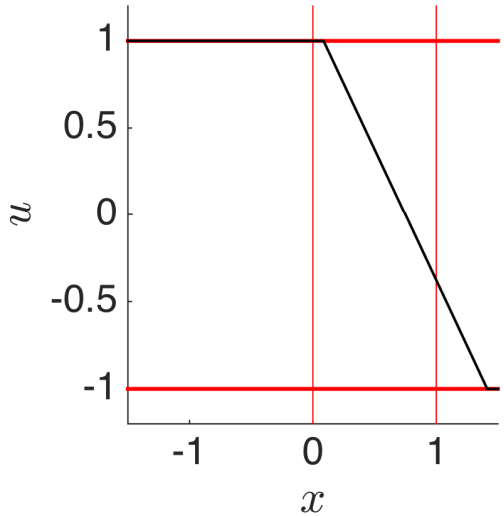
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

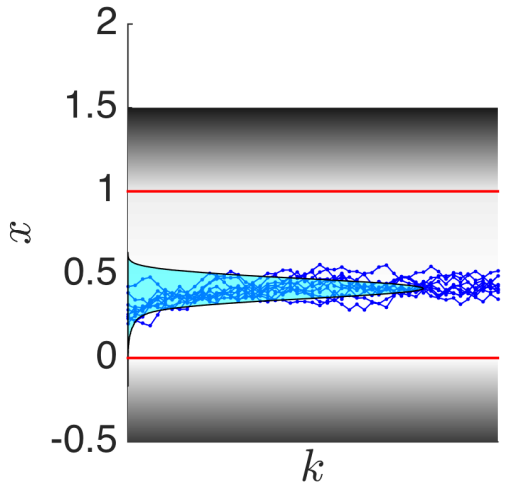
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

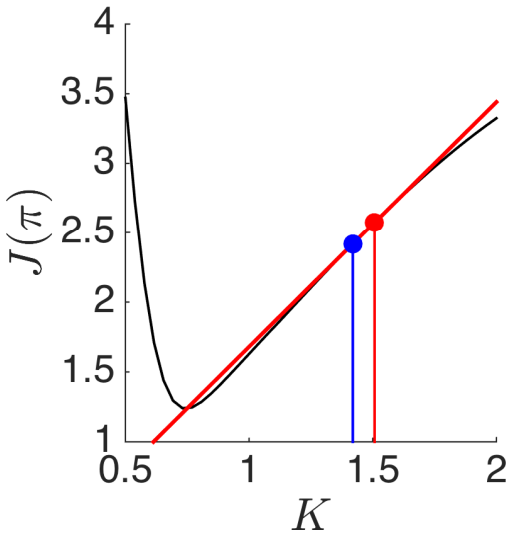
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} [\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

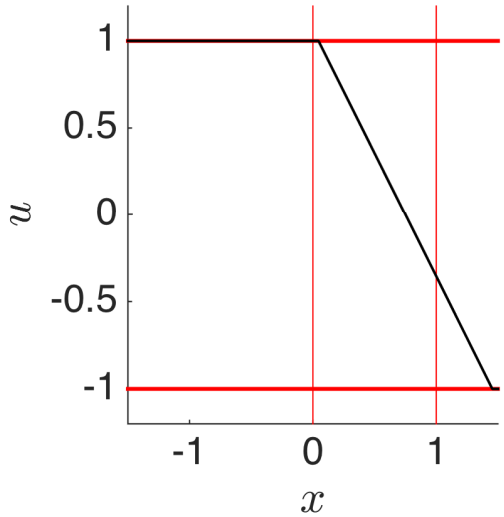
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

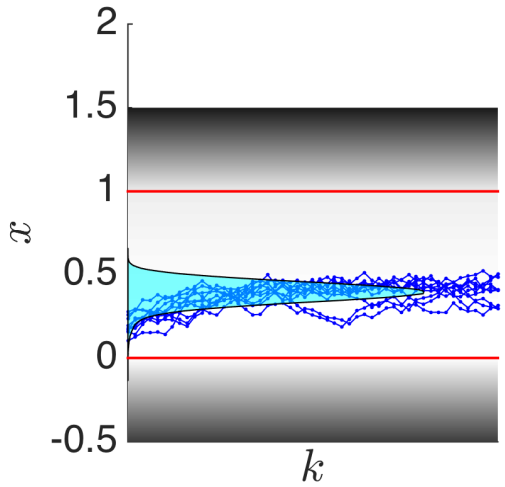
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

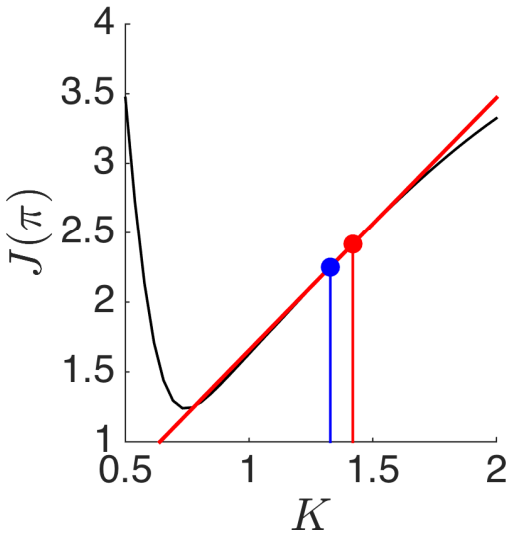
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

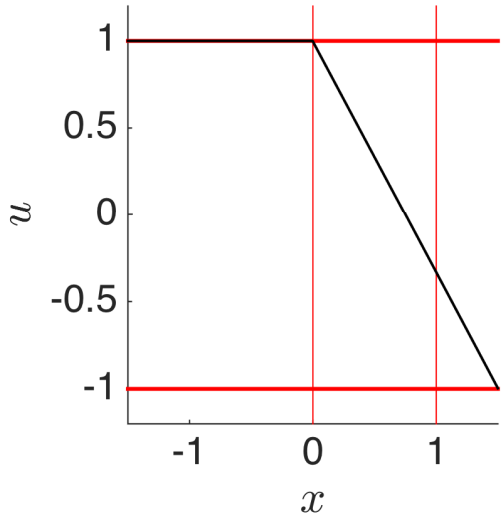
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

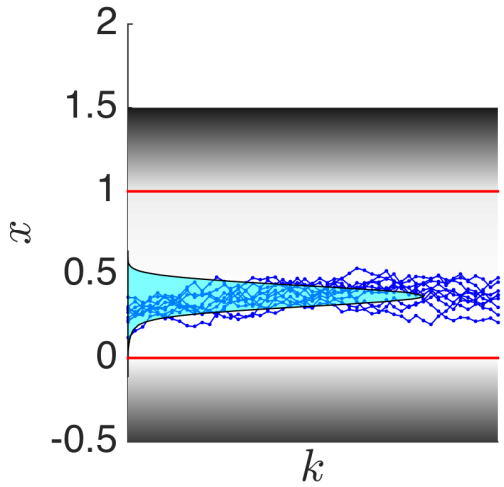
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

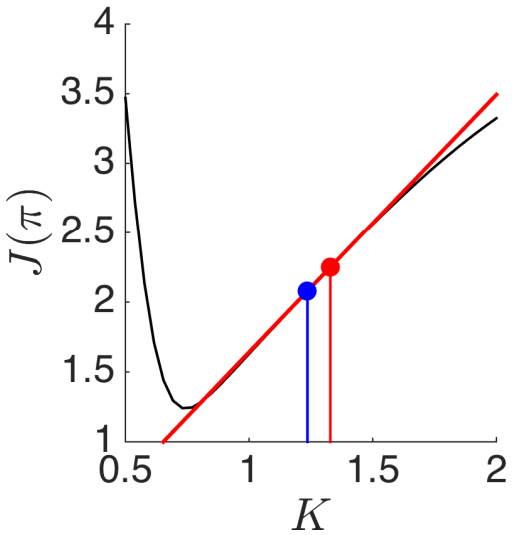
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

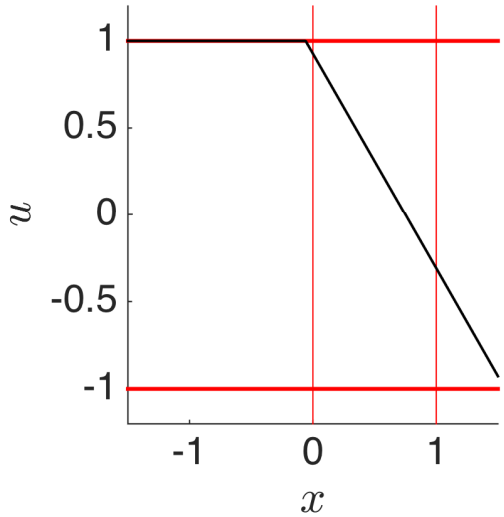
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

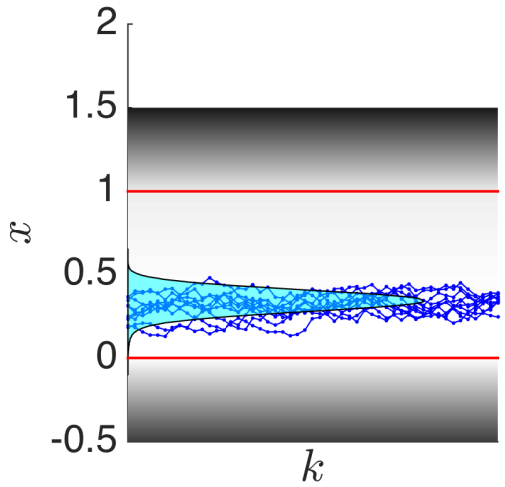
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

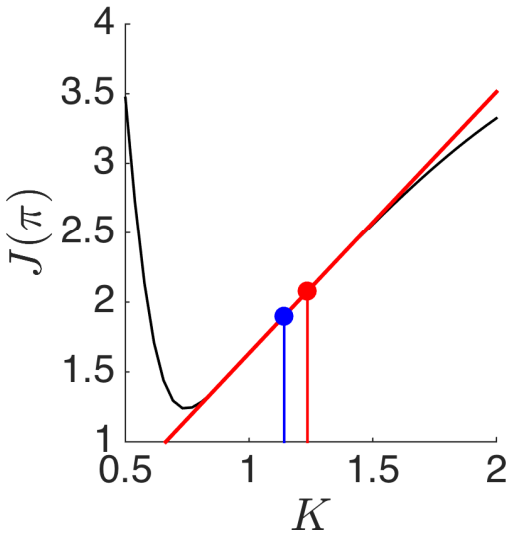
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

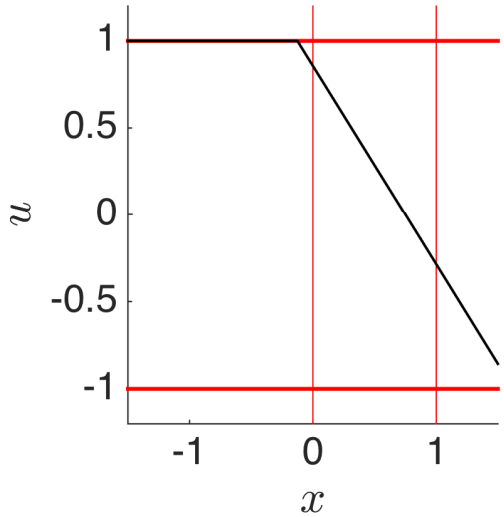
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

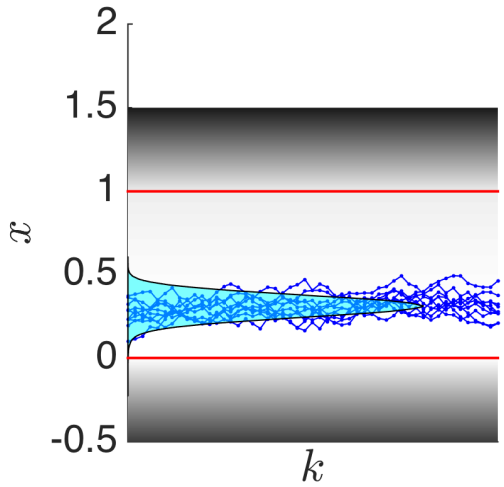
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

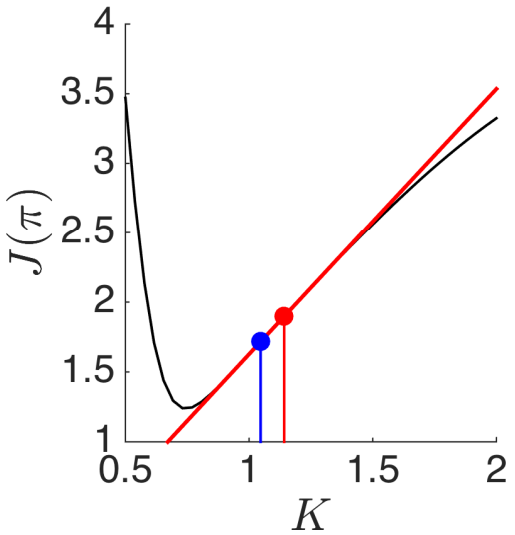
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

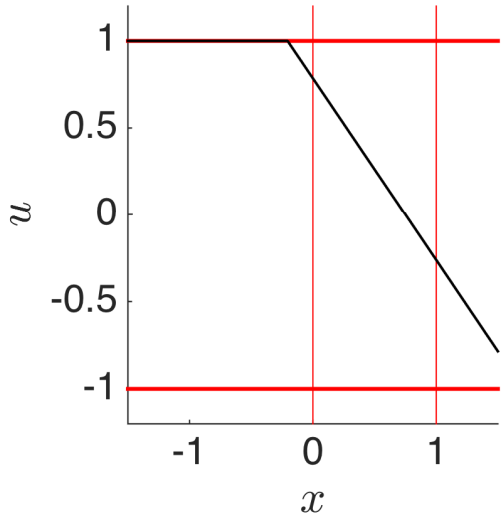
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

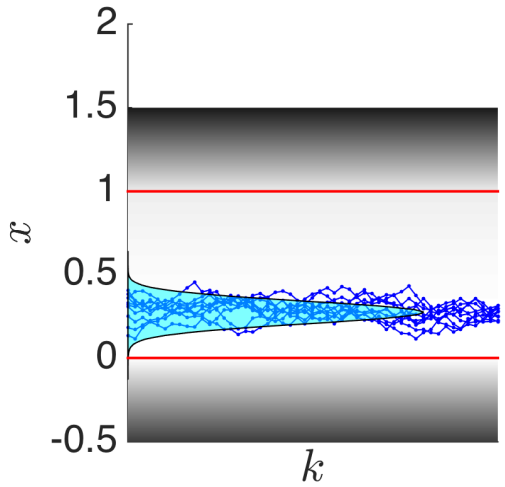
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

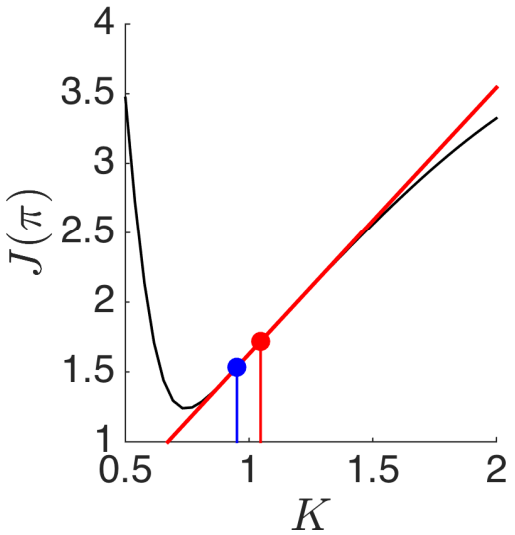
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

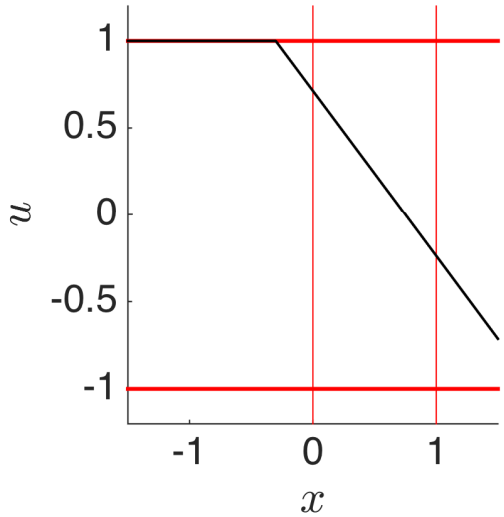
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

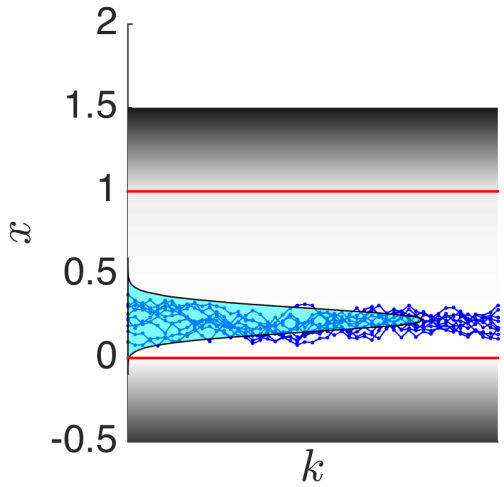
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

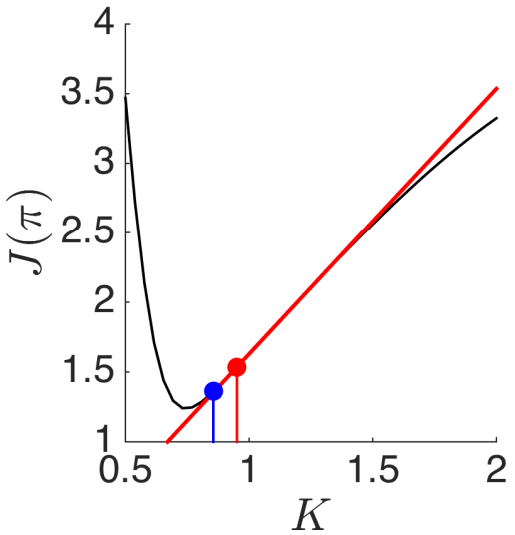
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

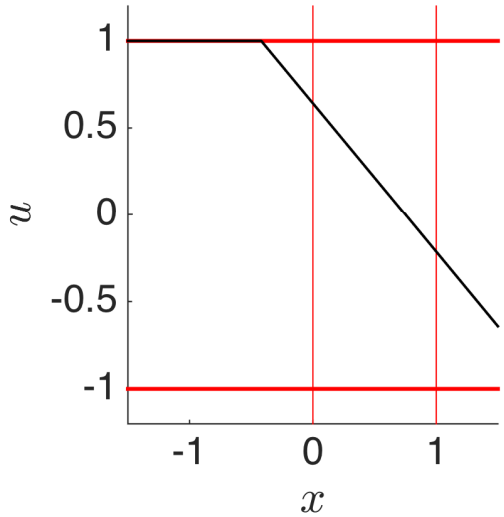
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

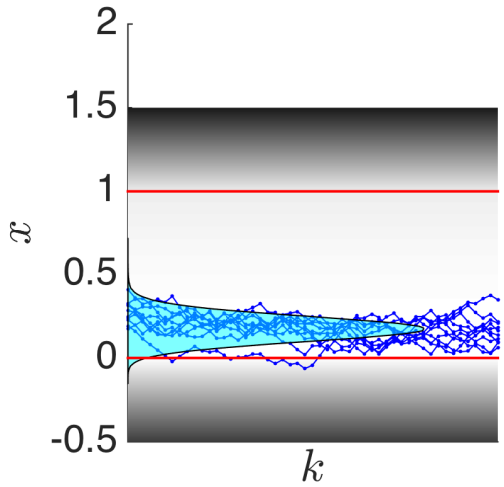
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

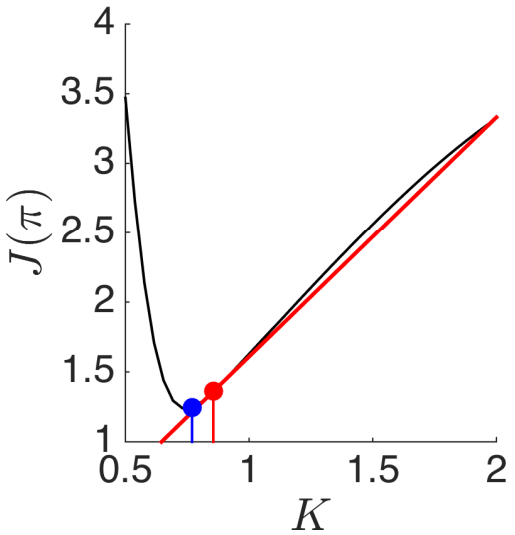
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

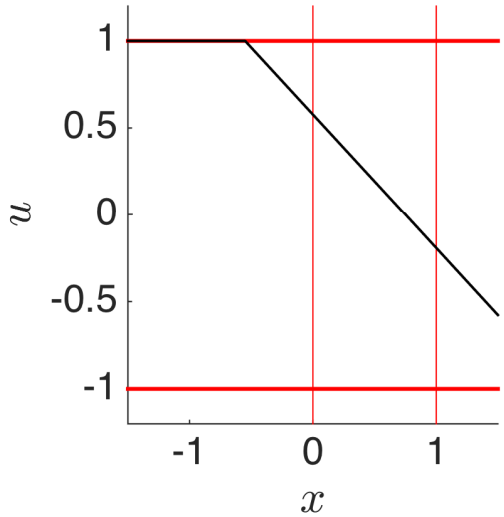
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

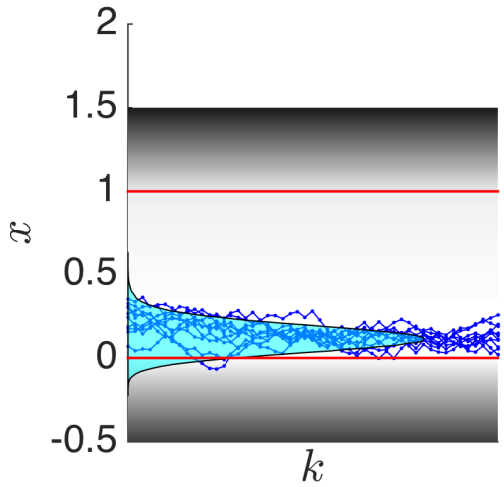
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

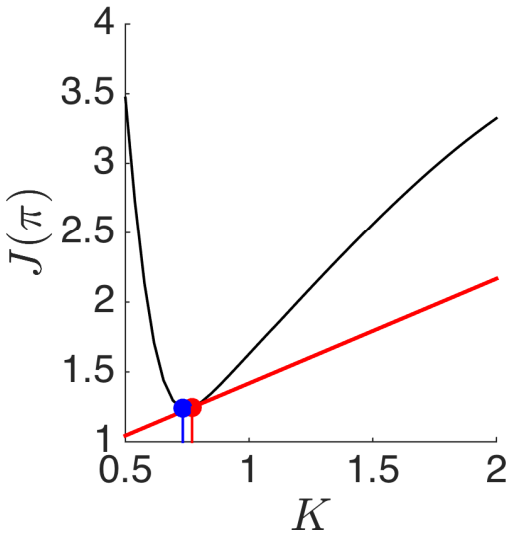
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

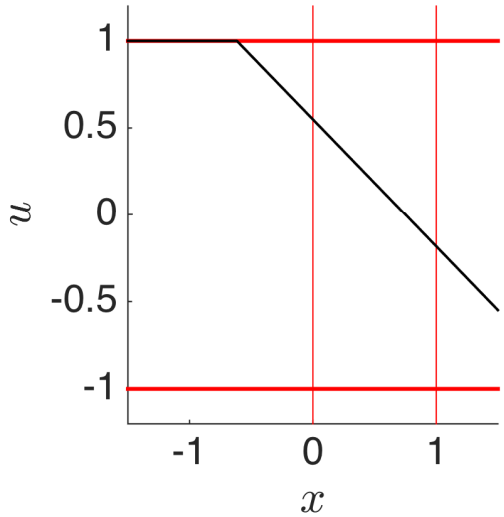
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

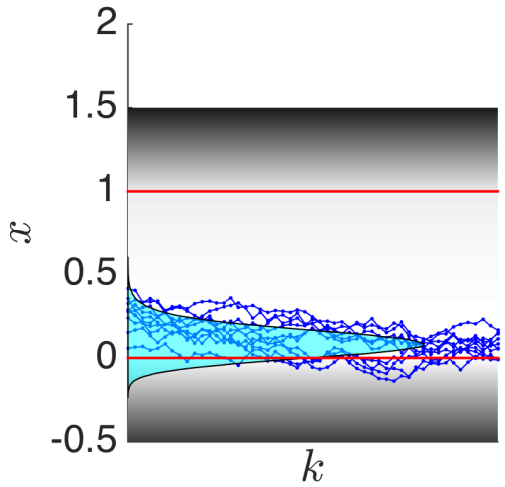
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

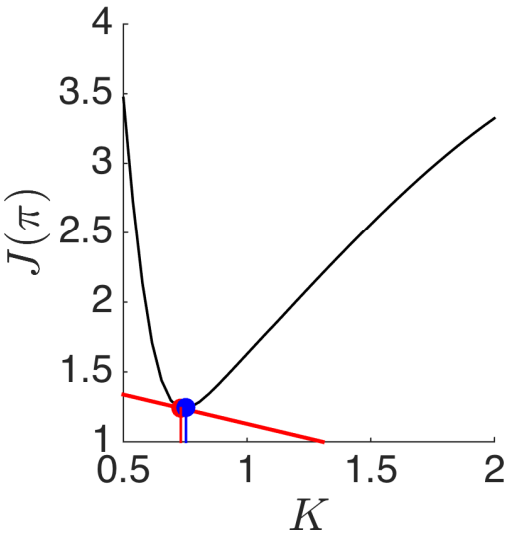
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

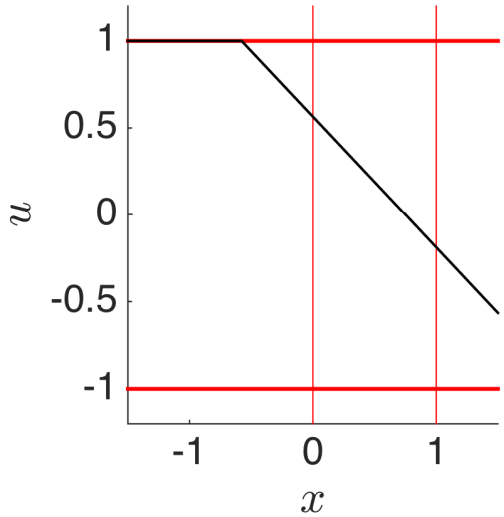
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

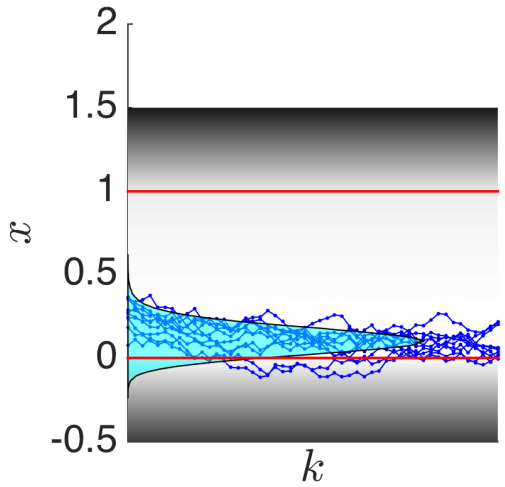
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

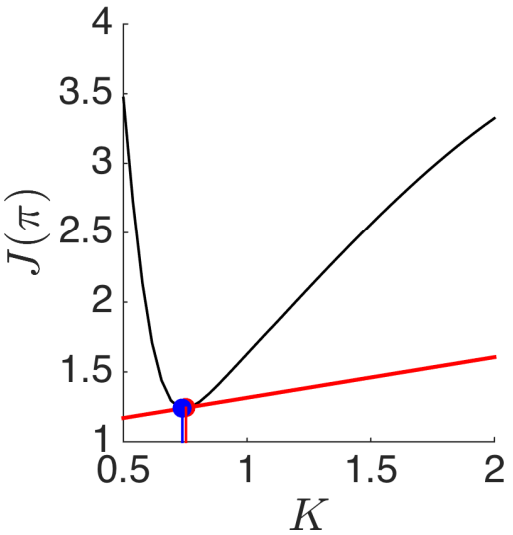
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

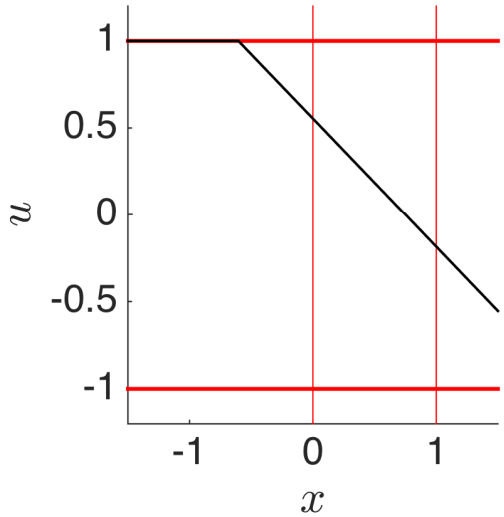
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

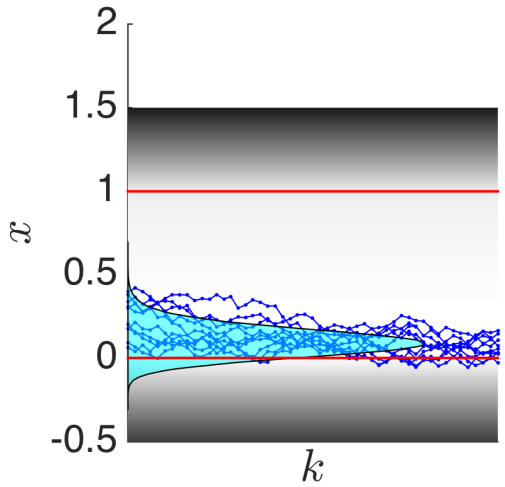
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

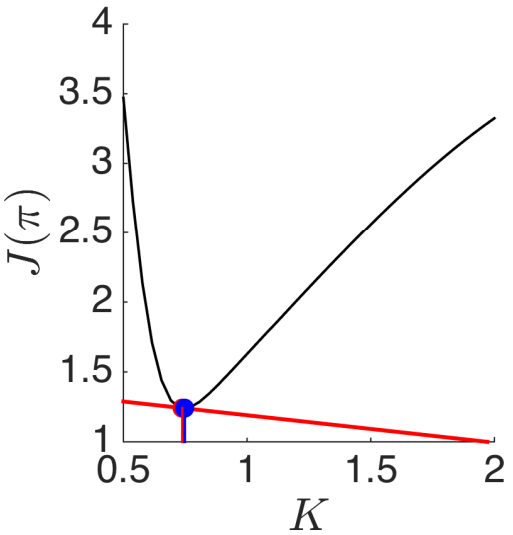
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

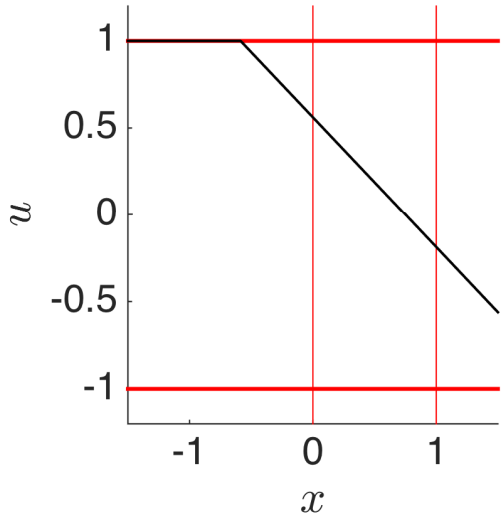
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

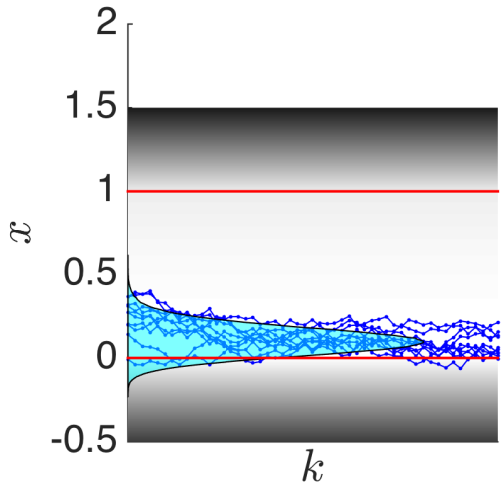
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

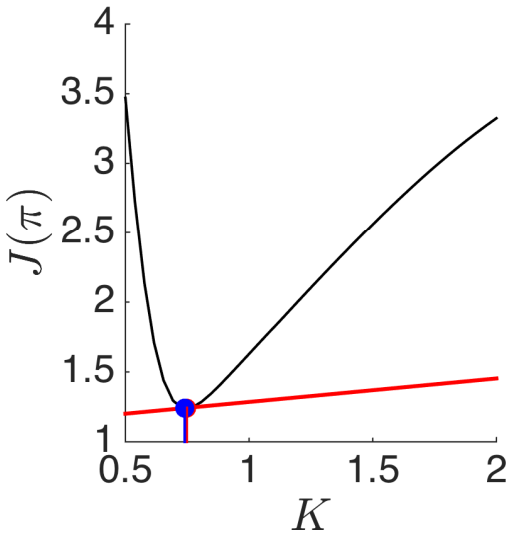
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

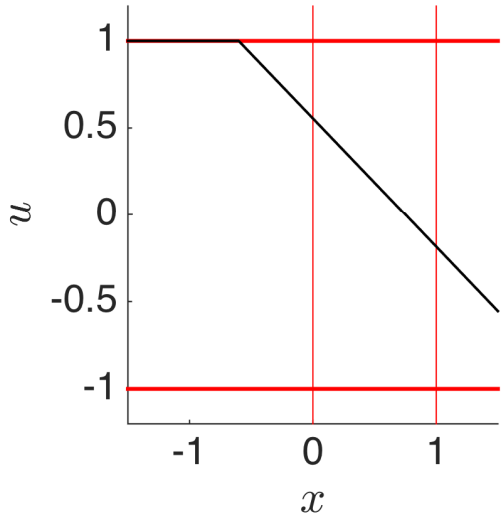
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

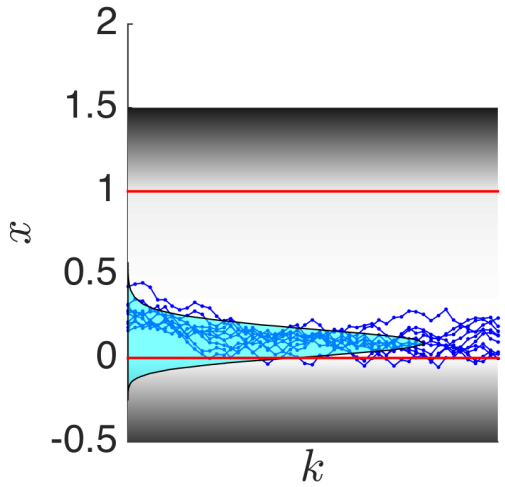
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

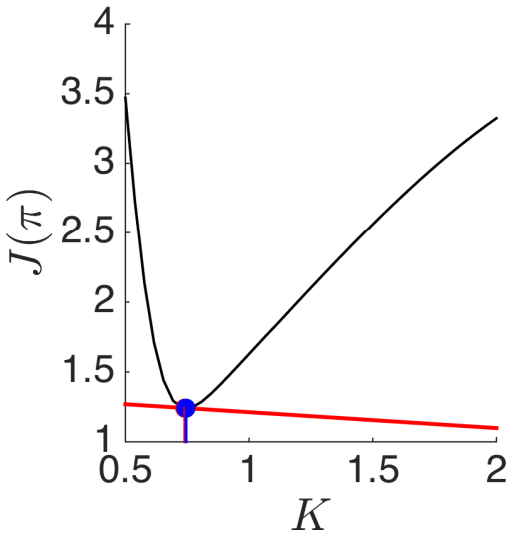
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

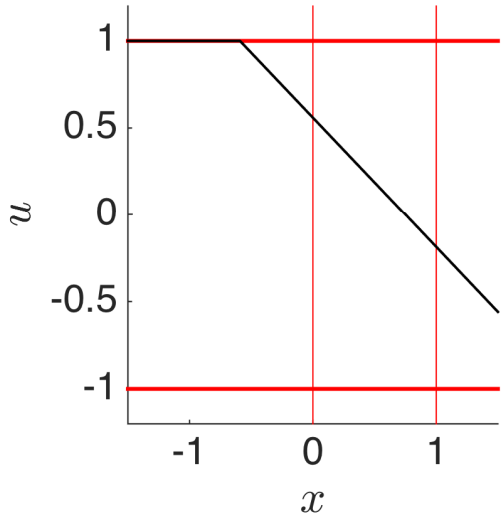
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

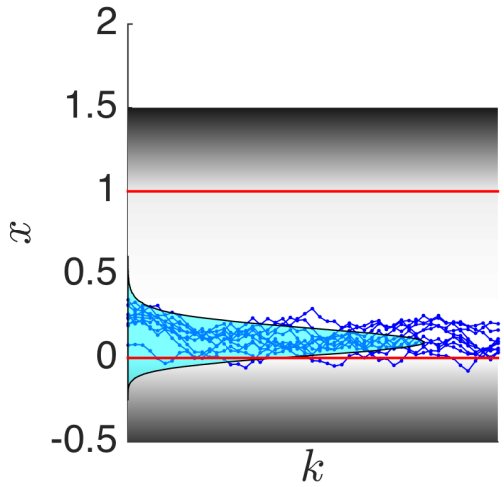
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

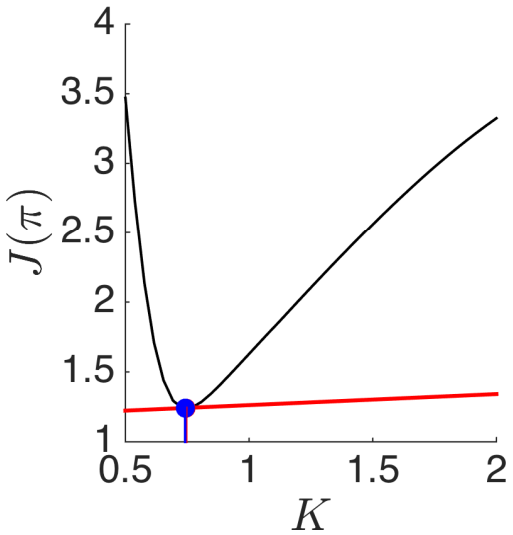
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

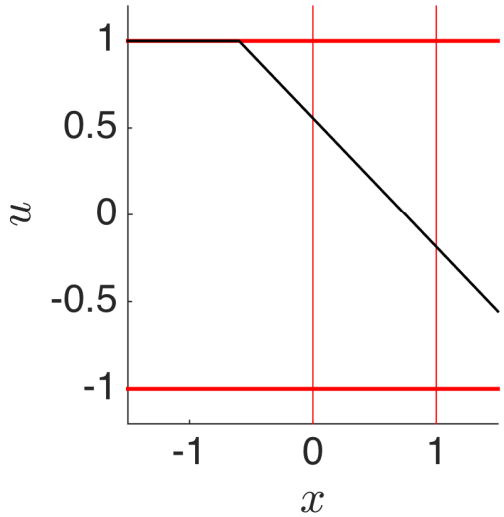
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

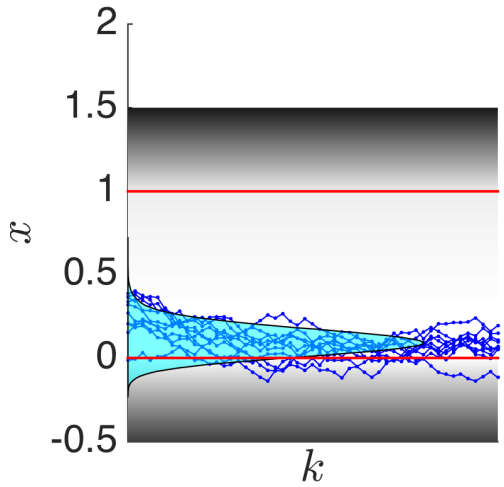
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

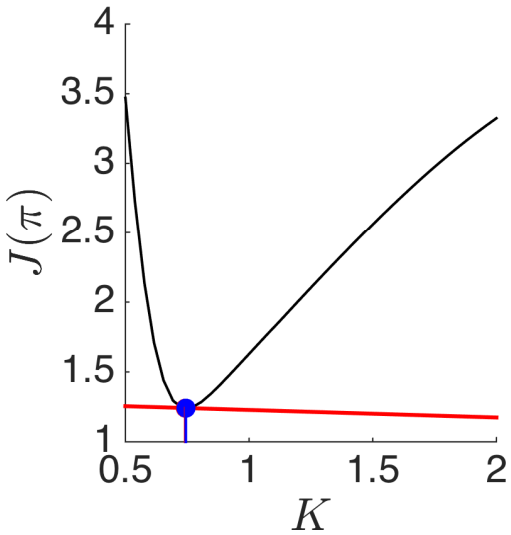
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







Deterministic policy gradient

Algorithm: Policy gradient (prototype)

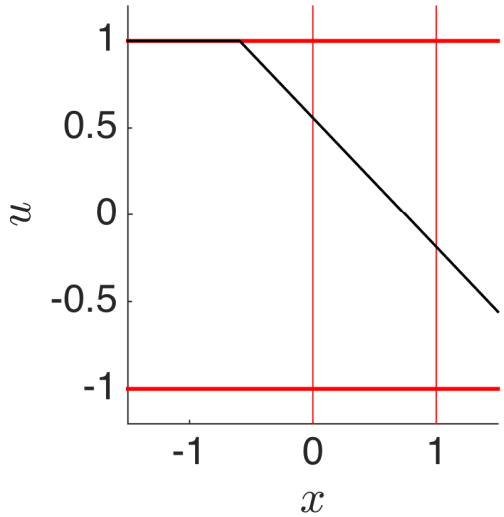
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

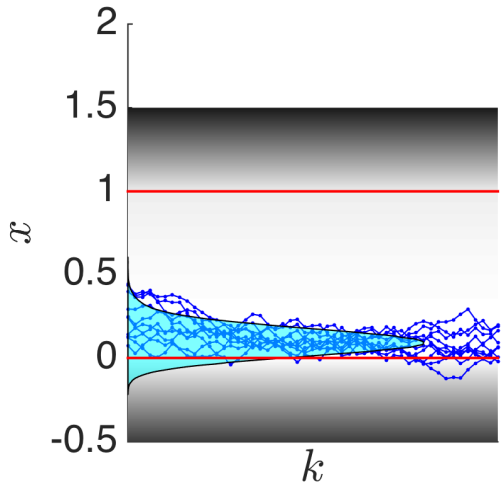
Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)

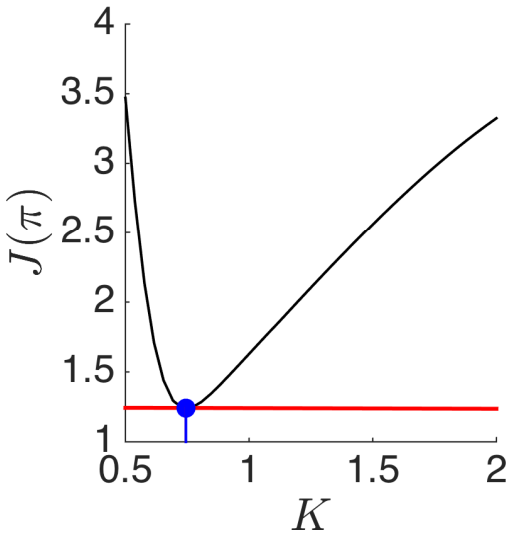
Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.

Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$

Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$







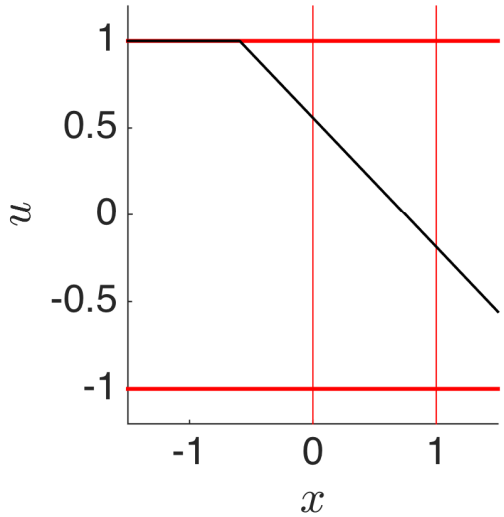
Deterministic policy gradient

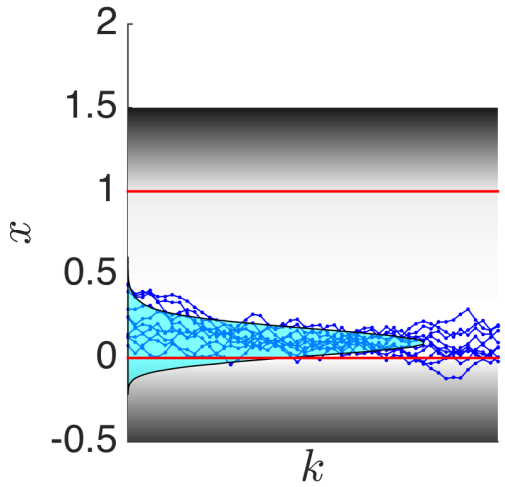
Algorithm: Policy gradient (prototype)

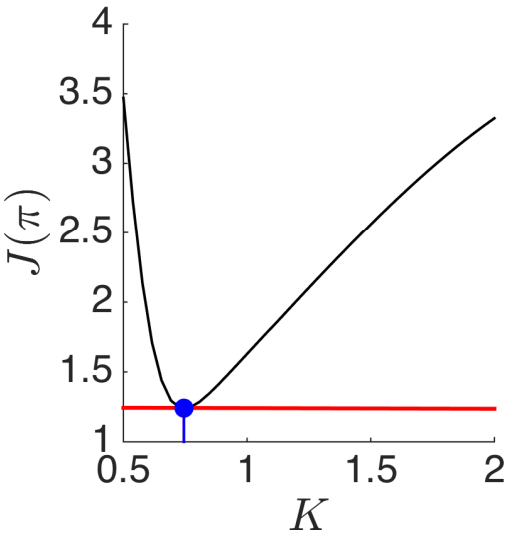
Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

- Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}} + \mathbf{d}$ for $N \to \infty$ samples (yields $\tilde{\tau}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$)
- Compute $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using e.g. TD-learning, MC, etc.
- Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{\mathbf{d}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \boldsymbol{\pi}_{\boldsymbol{\theta}})\right]$
- Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$

On-the-fly versions apply





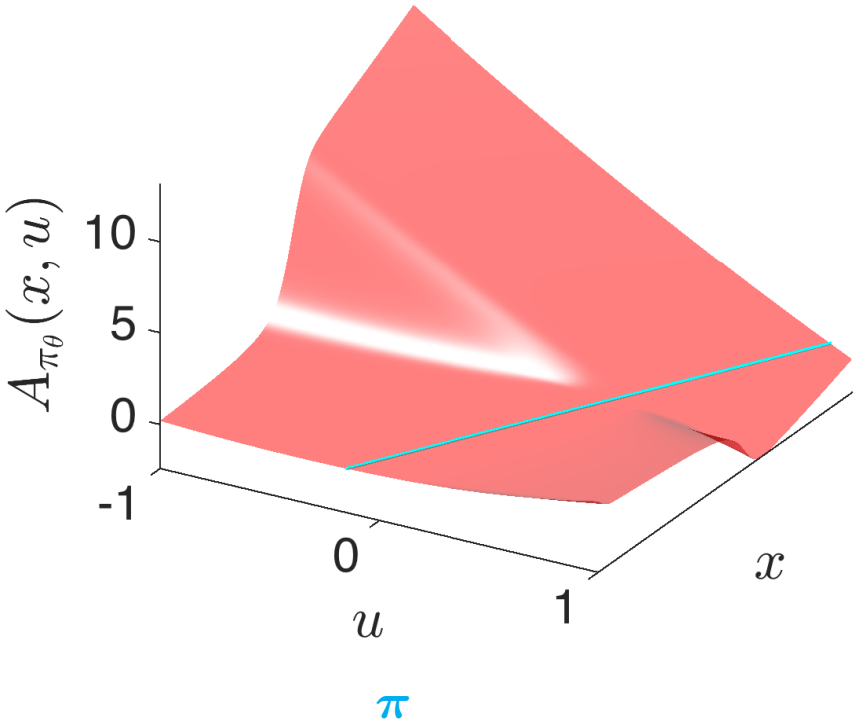


Approximation of Q_{π_θ}

Recall that:

$$Q_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = A_{\pi_\theta}(\mathbf{x}, \mathbf{u}) + V_{\pi_\theta}(\mathbf{x})$$

with $A_{\pi_\theta}(\mathbf{x}, \pi_\theta(\mathbf{x})) = 0$ for all $\mathbf{x}$



Approximation of Q_{π_θ}

Recall that:

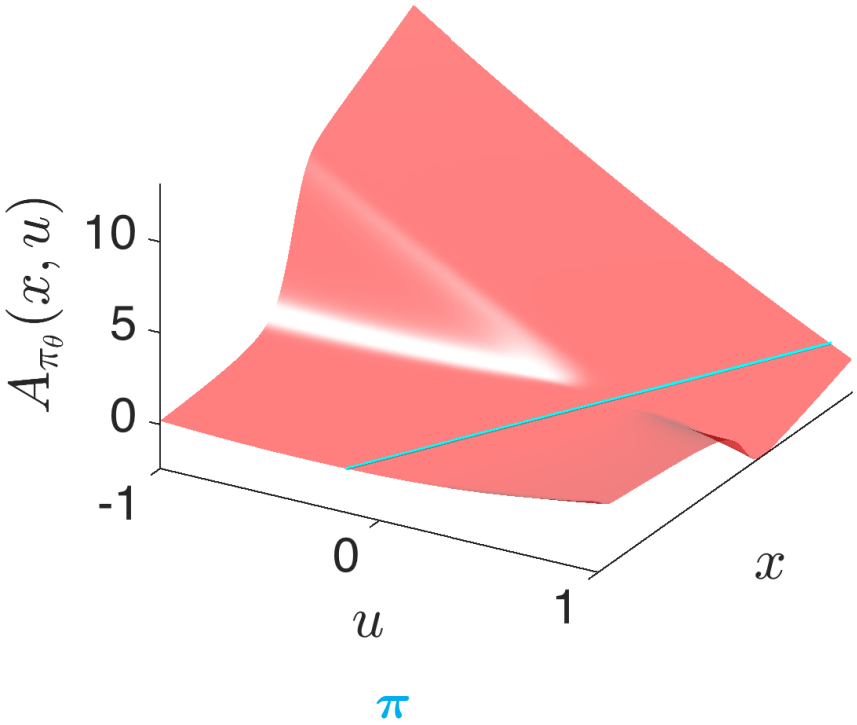
$$Q_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = A_{\pi_\theta}(\mathbf{x}, \mathbf{u}) + V_{\pi_\theta}(\mathbf{x})$$

with $A_{\pi_\theta}(\mathbf{x}, \pi_\theta(\mathbf{x})) = 0$ for all $\mathbf{x}$

For $\mathbf{u} \approx \pi_\theta(\mathbf{x})$, approximate A_{π_θ} with:

$$\hat{A}_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = \mathbf{a}(\mathbf{x})^\top (\mathbf{u} - \pi_\theta(\mathbf{x}))$$

for some state-dependent $\mathbf{a}(\mathbf{x})$



Approximation of Q_{π_θ}

Recall that:

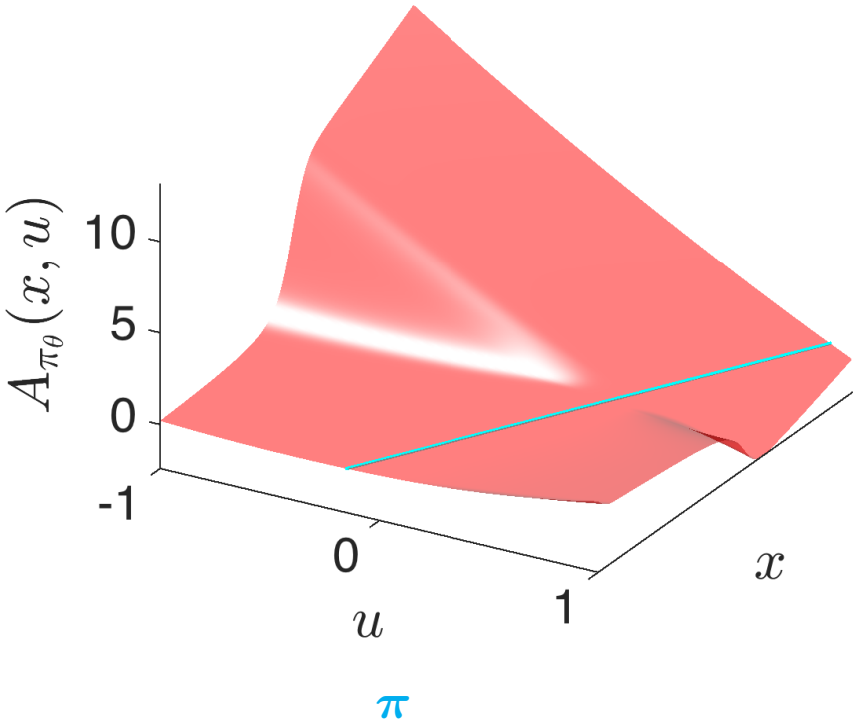
$$Q_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = A_{\pi_\theta}(\mathbf{x}, \mathbf{u}) + V_{\pi_\theta}(\mathbf{x})$$

with $A_{\pi_\theta}(\mathbf{x}, \pi_\theta(\mathbf{x})) = 0$ for all $\mathbf{x}$

For $\mathbf{u} \approx \pi_\theta(\mathbf{x})$, approximate A_{π_θ} with:

$$\hat{A}_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = \mathbf{a}(\mathbf{x})^\top (\mathbf{u} - \pi_\theta(\mathbf{x}))$$

for some state-dependent $\mathbf{a}(\mathbf{x})$



Approximation of Q_{π_θ}

Recall that:

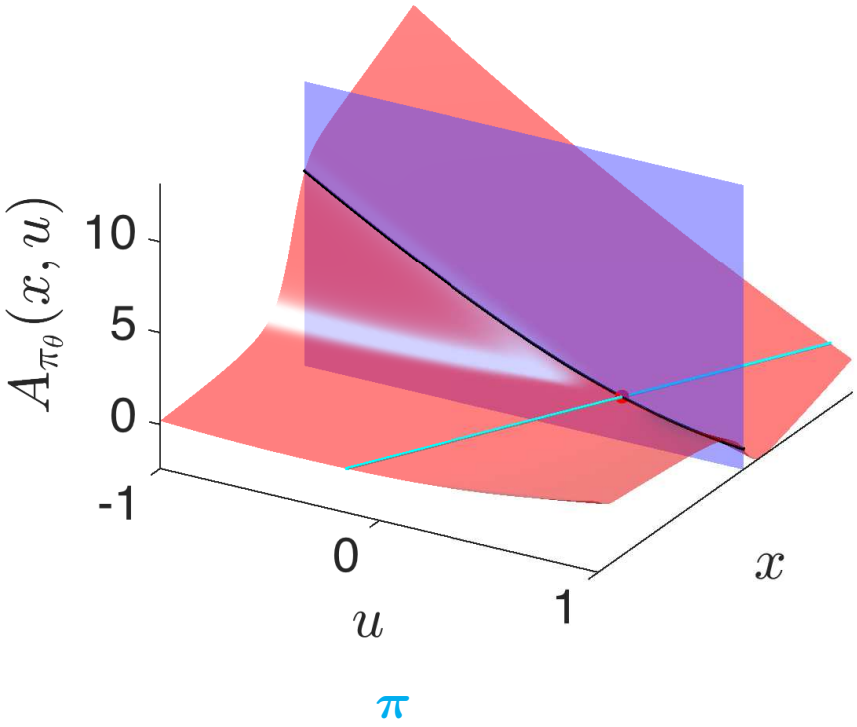
$$Q_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = A_{\pi_\theta}(\mathbf{x}, \mathbf{u}) + V_{\pi_\theta}(\mathbf{x})$$

with $A_{\pi_\theta}(\mathbf{x}, \pi_\theta(\mathbf{x})) = 0$ for all $\mathbf{x}$

For $\mathbf{u} \approx \pi_\theta(\mathbf{x})$, approximate A_{π_θ} with:

$$\hat{A}_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = \mathbf{a}(\mathbf{x})^\top (\mathbf{u} - \pi_\theta(\mathbf{x}))$$

for some state-dependent $\mathbf{a}(\mathbf{x})$



Approximation of Q_{π_θ}

Recall that:

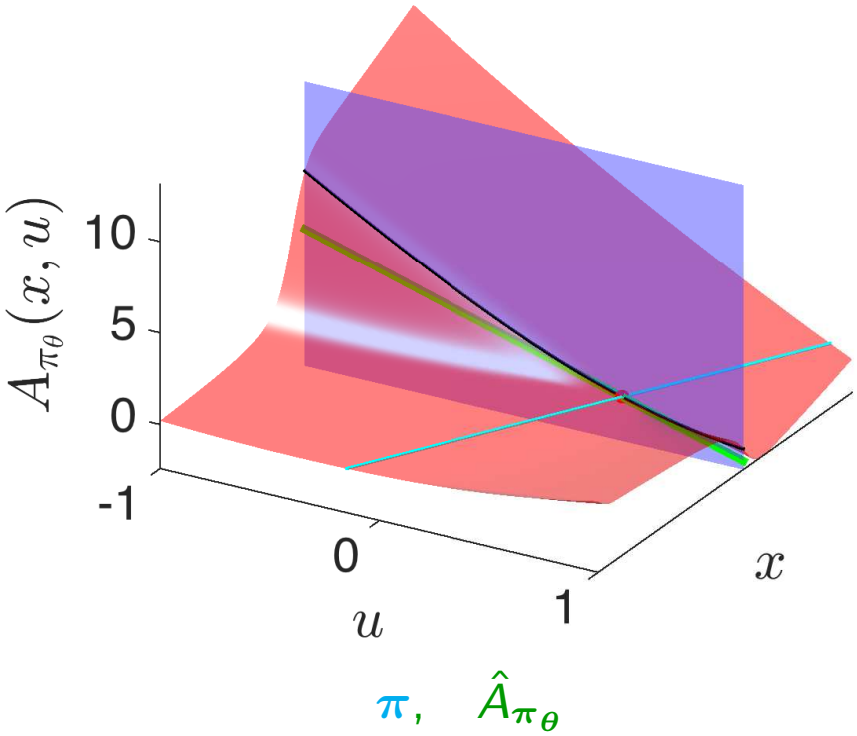
$$Q_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = A_{\pi_\theta}(\mathbf{x}, \mathbf{u}) + V_{\pi_\theta}(\mathbf{x})$$

with $A_{\pi_\theta}(\mathbf{x}, \pi_\theta(\mathbf{x})) = 0$ for all $\mathbf{x}$

For $\mathbf{u} \approx \pi_\theta(\mathbf{x})$, approximate A_{π_θ} with:

$$\hat{A}_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = \mathbf{a}(\mathbf{x})^\top (\mathbf{u} - \pi_\theta(\mathbf{x}))$$

for some state-dependent $\mathbf{a}(\mathbf{x})$



Approximation of Q_{π_θ}

Recall that:

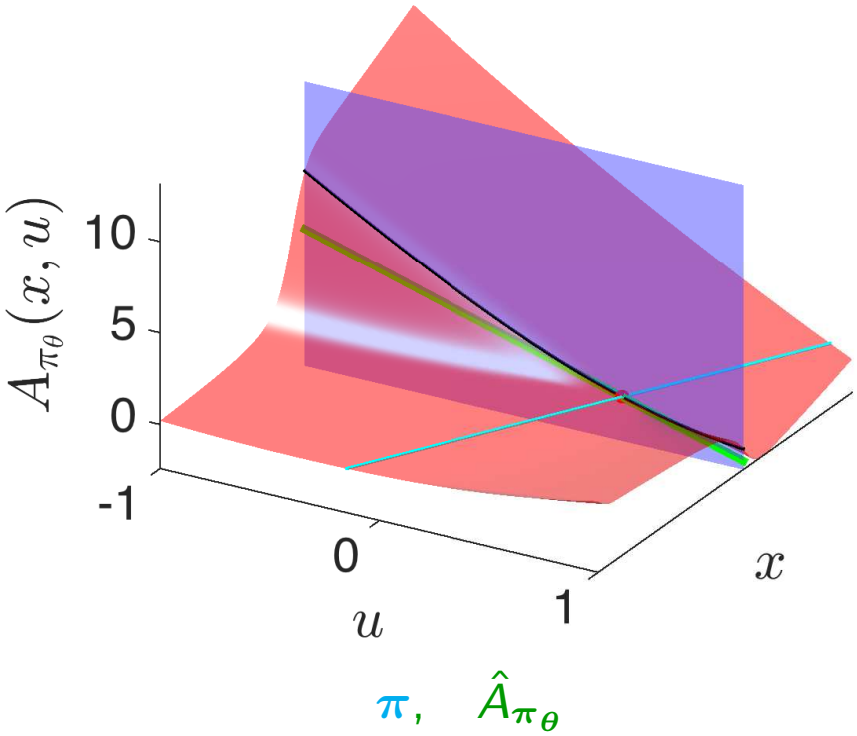
$$Q_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = A_{\pi_\theta}(\mathbf{x}, \mathbf{u}) + V_{\pi_\theta}(\mathbf{x})$$

with $A_{\pi_\theta}(\mathbf{x}, \pi_\theta(\mathbf{x})) = 0$ for all $\mathbf{x}$

For $\mathbf{u} \approx \pi_\theta(\mathbf{x})$, approximate A_{π_θ} with:

$$\hat{A}_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = \mathbf{a}(\mathbf{x})^\top (\mathbf{u} - \pi_\theta(\mathbf{x}))$$

for some state-dependent $\mathbf{a}(\mathbf{x})$



E.g. **compatible function approximation**

$$\hat{A}_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = \mathbf{w}^\top \nabla_\theta \pi_\theta(\mathbf{x}) (\mathbf{u} - \pi_\theta(\mathbf{x}))$$

for some parameters $\mathbf{w}$ of the size of $\boldsymbol{\theta}$

Approximation of Q_{π_θ}

Recall that:

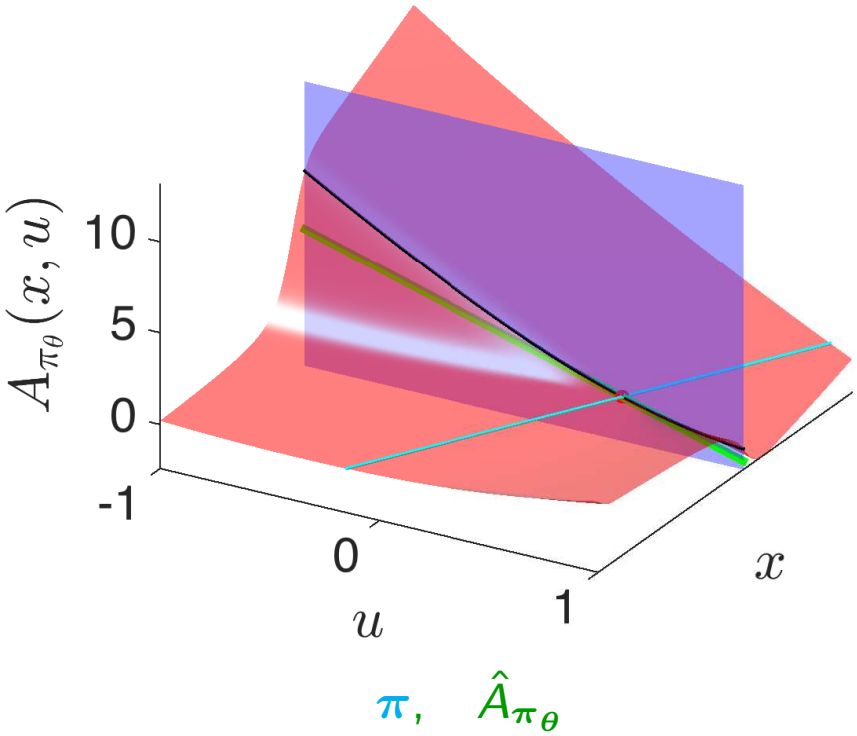
$$Q_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = A_{\pi_\theta}(\mathbf{x}, \mathbf{u}) + V_{\pi_\theta}(\mathbf{x})$$

with $A_{\pi_\theta}(\mathbf{x}, \pi_\theta(\mathbf{x})) = 0$ for all $\mathbf{x}$

For $\mathbf{u} \approx \pi_\theta(\mathbf{x})$, approximate A_{π_θ} with:

$$\hat{A}_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = \mathbf{a}(\mathbf{x})^\top (\mathbf{u} - \pi_\theta(\mathbf{x}))$$

for some state-dependent $\mathbf{a}(\mathbf{x})$



E.g. **compatible function approximation**

$$\hat{A}_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = \mathbf{w}^\top \nabla_\theta \pi_\theta(\mathbf{x}) (\mathbf{u} - \pi_\theta(\mathbf{x}))$$

for some parameters $\mathbf{w}$ of the size of $\boldsymbol{\theta}$

We then use

$$\hat{Q}_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = \hat{A}_{\pi_\theta}(\mathbf{x}, \mathbf{u}) + \hat{V}_{\pi_\theta}(\mathbf{x})$$

with $\hat{V}_{\pi_\theta}(\mathbf{x}) \approx V_{\pi_\theta}(\mathbf{x})$

Approximation of Q_{π_θ}

Recall that:

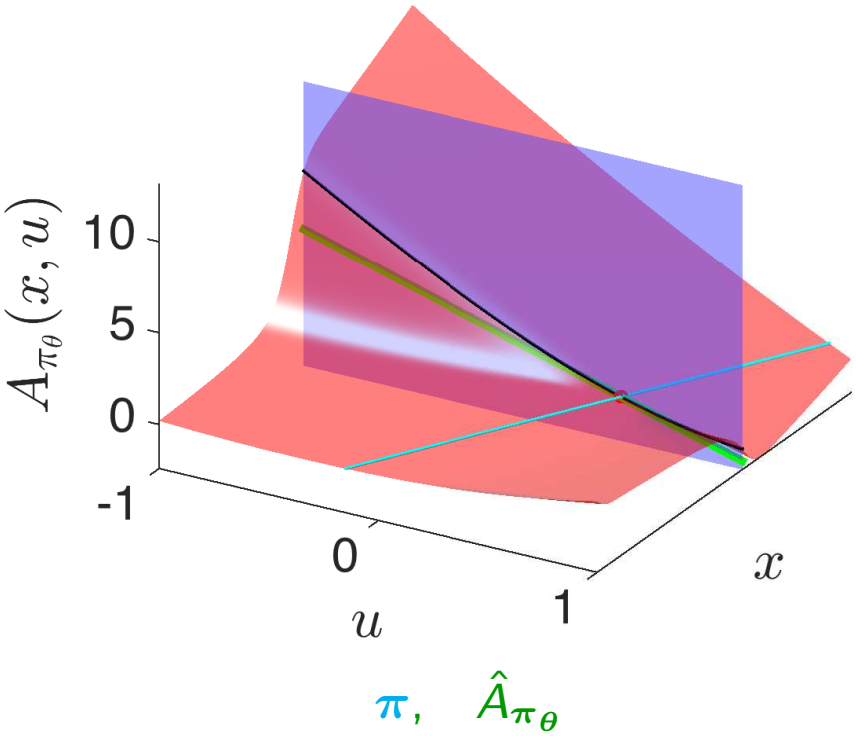
$$Q_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = A_{\pi_\theta}(\mathbf{x}, \mathbf{u}) + V_{\pi_\theta}(\mathbf{x})$$

with $A_{\pi_\theta}(\mathbf{x}, \pi_\theta(\mathbf{x})) = 0$ for all $\mathbf{x}$

For $\mathbf{u} \approx \pi_\theta(\mathbf{x})$, approximate A_{π_θ} with:

$$\hat{A}_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = \mathbf{a}(\mathbf{x})^\top (\mathbf{u} - \pi_\theta(\mathbf{x}))$$

for some state-dependent $\mathbf{a}(\mathbf{x})$



E.g. **compatible function approximation**

$$\hat{A}_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = \mathbf{w}^\top \nabla_\theta \pi_\theta(\mathbf{x}) (\mathbf{u} - \pi_\theta(\mathbf{x}))$$

for some parameters $\mathbf{w}$ of the size of $\boldsymbol{\theta}$

We then use

$$\hat{Q}_{\pi_\theta}(\mathbf{x}, \mathbf{u}) = \hat{A}_{\pi_\theta}(\mathbf{x}, \mathbf{u}) + \hat{V}_{\pi_\theta}(\mathbf{x})$$

with $\hat{V}_{\pi_\theta}(\mathbf{x}) \approx V_{\pi_\theta}(\mathbf{x})$

Then if gradients fit, we get exact policy gradients!

$$\mathbf{w} = \arg\min_{\mathbf{w}} \mathbb{E}_{\tau_{\pi_\theta}} \left[\left\| \nabla_{\mathbf{u}} Q_{\pi_\theta} - \nabla_{\mathbf{u}} \hat{Q}_{\pi_\theta} \right\|^2 \right] \quad \Rightarrow \quad \nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau_{\pi_\theta}} \left[\nabla_\theta \pi_\theta \nabla_{\mathbf{u}} \hat{Q}_{\pi_\theta} \mathbf{w} \right]$$

Gradient fitting?

Fitting the gradients

$$\nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \approx \nabla_{\mathbf{u}} \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$$

using data is very difficult. What about fitting $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \approx \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ (classic $\boldsymbol{\pi}$ evaluation)?

$$\min_{\mathbf{w}} \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\left\| \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \nabla_{\mathbf{u}} \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \right\|^2 \right]$$

- Compatible approximation
- Learning rule?...

vs.

$$\min_{\mathbf{w}} \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\left\| Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \right\|^2 \right]$$

- Can use TD or MC-learning
- $\neq$ fitting gradients...

Gradient fitting?

Fitting the gradients

$$\nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \approx \nabla_{\mathbf{u}} \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$$

using data is very difficult. What about fitting $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \approx \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ (classic $\boldsymbol{\pi}$ evaluation)?

$$\min_{\mathbf{w}} \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\left\| \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \nabla_{\mathbf{u}} \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \right\|^2 \right]$$

- Compatible approximation
- Learning rule?...

vs.

$$\min_{\mathbf{w}} \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\left\| Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \right\|^2 \right]$$

- Can use TD or MC-learning
- $\neq$ fitting gradients...

Fitting $\hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ instead of $\nabla_{\mathbf{u}} \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ yields correct policy gradient if

- Input sequence $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k) + \mathbf{d}_k$ where $\mathbf{d}_k \to 0$ and $\mathbb{E}[\mathbf{d}_k] = 0$

Gradient fitting?

Fitting the gradients

$$\nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \approx \nabla_{\mathbf{u}} \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$$

using data is very difficult. What about fitting $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \approx \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ (classic $\boldsymbol{\pi}$ evaluation)?

$$\min_{\mathbf{w}} \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\left\| \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \nabla_{\mathbf{u}} \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \right\|^2 \right]$$

- Compatible approximation
- Learning rule?...

vs.

$$\min_{\mathbf{w}} \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\left\| Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \right\|^2 \right]$$

- Can use TD or MC-learning
- $\neq$ fitting gradients...

Fitting $\hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ instead of $\nabla_{\mathbf{u}} \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ yields correct policy gradient if

- Input sequence $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k) + \mathbf{d}_k$ where $\mathbf{d}_k \to 0$ and $\mathbb{E}[\mathbf{d}_k] = 0$
- What about input constraints?!?

Gradient fitting?

Fitting the gradients

$$\nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \approx \nabla_{\mathbf{u}} \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$$

using data is very difficult. What about fitting $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \approx \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ (classic $\boldsymbol{\pi}$ evaluation)?

$$\min_{\mathbf{w}} \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\left\| \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \nabla_{\mathbf{u}} \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \right\|^2 \right]$$

- Compatible approximation
- Learning rule?...

vs.

$$\min_{\mathbf{w}} \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\left\| Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \right\|^2 \right]$$

- Can use TD or MC-learning
- $\neq$ fitting gradients...

Fitting $\hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ instead of $\nabla_{\mathbf{u}} \hat{Q}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ yields correct policy gradient if

- Input sequence $\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k) + \mathbf{d}_k$ where $\mathbf{d}_k \to 0$ and $\mathbb{E}[\mathbf{d}_k] = 0$
- What about input constraints?!? Fine if $\mathbf{d}_k$ is in the "border" of the feasible set and centred in the null space of the constraints

Outline

Stochastic policy

Deterministic policy

$$\mathbf{u}_k = \boldsymbol{\pi}\left(\mathbf{x}_k\right)$$

Policy gradient

$$\nabla_{\boldsymbol{\theta}} J = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}}\left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}} \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right]$$

what if inputs are discrete?

Stochastic policy

Deterministic policy

$$\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$$

Policy gradient

$$\nabla_{\boldsymbol{\theta}} J = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} [\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}} \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}]$$

what if inputs are discrete?

Stochastic policy

$$\boldsymbol{\pi}[\mathbf{u}_k \,|\, \mathbf{x}_k]$$

gives input distribution for a given state $\mathbf{x}_k$

Policy gradient?

Stochastic policy

Deterministic policy

$$\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$$

Policy gradient

$$\nabla_{\boldsymbol{\theta}} J = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} [\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}} \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}]$$

what if inputs are discrete?

Stochastic policy

$$\boldsymbol{\pi}[\mathbf{u}_k \,|\, \mathbf{x}_k]$$

gives input distribution for a given state $\mathbf{x}_k$

Policy gradient?

E.g. Gaussian stochastic policy:

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}[\mathbf{u}_k \,|\, \mathbf{x}_k] = \mathcal{N}(\bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}}(\mathbf{x}_k), \Sigma)$$

restricted to feasible $\mathbf{u}$

Covariance Σ can be part of policy parameters $\boldsymbol{\theta}$.

Stochastic policy

Deterministic policy

$$\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$$

Policy gradient

$$\nabla_{\boldsymbol{\theta}} J = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}} \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \right]$$

what if inputs are discrete?

Stochastic policy

$$\boldsymbol{\pi}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right]$$

gives input distribution for a given state $\mathbf{x}_k$

Policy gradient?

E.g. Gaussian stochastic policy:

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right] = \mathcal{N}(\bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}}(\mathbf{x}_k), \Sigma)$$

restricted to feasible $\mathbf{u}$

Covariance Σ can be part of policy parameters $\boldsymbol{\theta}$.

$$\bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}}(\mathbf{x}_k),\ \boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right]$$

Stochastic policy

Deterministic policy

$$\mathbf{u}_k = \boldsymbol{\pi}(\mathbf{x}_k)$$

Policy gradient

$$\nabla_{\boldsymbol{\theta}} J = \mathbb{E}_{\tau_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}} \left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}} \nabla_{\mathbf{u}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \right]$$

what if inputs are discrete?

Stochastic policy

$$\boldsymbol{\pi}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right]$$

gives input distribution for a given state $\mathbf{x}_k$

Policy gradient?

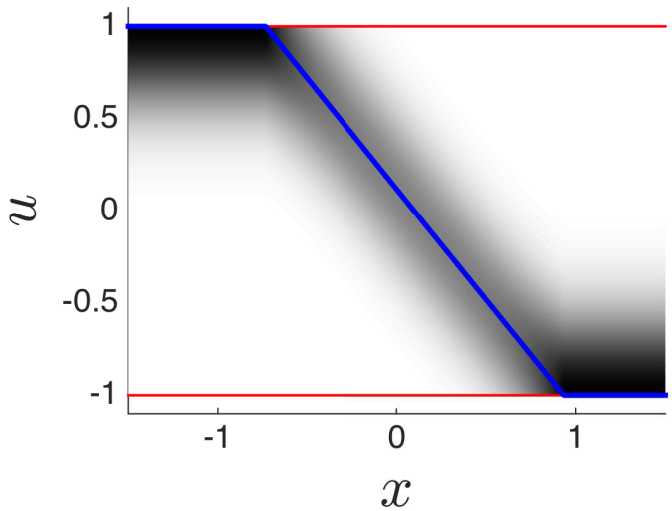
E.g. Gaussian stochastic policy:

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right] = \mathcal{N}(\bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}}(\mathbf{x}_k), \Sigma)$$

restricted to feasible $\mathbf{u}$

Covariance Σ can be part of policy parameters $\boldsymbol{\theta}$.

$$\bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}}(\mathbf{x}_k),\ \boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right]$$



Bellman equation with stochastic policy?

Bellman equations

$$V_{\pi}(\mathbf{x}) = L(\mathbf{x}, \boldsymbol{\pi}(\mathbf{x})) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \boldsymbol{\pi}\right] \qquad V_{\pi}(\mathbf{x}) = Q_{\pi}(\mathbf{x}, \boldsymbol{\pi}(\mathbf{x}))$$
$$Q_{\pi}(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u}\right]$$

become...

Bellman equation with stochastic policy?

Bellman equations

$$V_{\pi}(\mathbf{x}) = L(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \pi\right] \qquad V_{\pi}(\mathbf{x}) = Q_{\pi}(\mathbf{x}, \pi(\mathbf{x}))$$
$$Q_{\pi}(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u}\right]$$

become...

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[L(\mathbf{x}, \mathbf{u}) + \gamma V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u} \sim \pi\right]$$

Bellman equation with stochastic policy?

Bellman equations

$$V_{\pi}(\mathbf{x}) = L(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \pi\right] \qquad V_{\pi}(\mathbf{x}) = Q_{\pi}(\mathbf{x}, \pi(\mathbf{x}))$$
$$Q_{\pi}(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u}\right]$$

become...

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[L(\mathbf{x}, \mathbf{u}) + \gamma V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u} \sim \pi\right]$$
$$Q_{\pi}(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u}\right]$$

Bellman equation with stochastic policy?

Bellman equations

$$V_{\pi}(\mathbf{x}) = L(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \pi\right] \qquad V_{\pi}(\mathbf{x}) = Q_{\pi}(\mathbf{x}, \pi(\mathbf{x}))$$
$$Q_{\pi}(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u}\right]$$

become...

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[L(\mathbf{x}, \mathbf{u}) + \gamma V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u} \sim \pi\right]$$
$$Q_{\pi}(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u}\right]$$
$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[Q_{\pi}(\mathbf{x}, \mathbf{u}) \,|\, \mathbf{u} \sim \pi\right]$$

Bellman equation with stochastic policy?

Bellman equations

$$V_{\pi}(\mathbf{x}) = L(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \pi\right] \qquad V_{\pi}(\mathbf{x}) = Q_{\pi}(\mathbf{x}, \pi(\mathbf{x}))$$
$$Q_{\pi}(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u}\right]$$

become...

$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[L(\mathbf{x}, \mathbf{u}) + \gamma V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u} \sim \pi\right]$$
$$Q_{\pi}(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[V_{\pi}(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u}\right]$$
$$V_{\pi}(\mathbf{x}) = \mathbb{E}\left[Q_{\pi}(\mathbf{x}, \mathbf{u}) \,|\, \mathbf{u} \sim \pi\right]$$

- ... but are otherwise still valid for $\pi\left[\mathbf{u} \,|\, \mathbf{x}\right]$

Bellman equation with stochastic policy?

Bellman equations

$$V_\pi(\mathbf{x}) = L(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}\left[V_\pi(\mathbf{x}_+) \,|\, \mathbf{x}, \pi\right] \qquad V_\pi(\mathbf{x}) = Q_\pi(\mathbf{x}, \pi(\mathbf{x}))$$
$$Q_\pi(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[V_\pi(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u}\right]$$

become...

$$V_\pi(\mathbf{x}) = \mathbb{E}\left[L(\mathbf{x}, \mathbf{u}) + \gamma V_\pi(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u} \sim \boldsymbol{\pi}\right]$$
$$Q_\pi(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}\left[V_\pi(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u}\right]$$
$$V_\pi(\mathbf{x}) = \mathbb{E}\left[Q_\pi(\mathbf{x}, \mathbf{u}) \,|\, \mathbf{u} \sim \boldsymbol{\pi}\right]$$

- ... but are otherwise still valid for $\boldsymbol{\pi}[\mathbf{u} \,|\, \mathbf{x}]$
- Expected values $\mathbb{E}[.]$ over state transitions are taken as:

$$\mathbb{E}\left[\zeta \,|\, \mathbf{x}, \boldsymbol{\pi}\right] = \int \zeta \cdot \mathbb{P}\left[\mathbf{x}_+ \,|\, \mathbf{x}, \mathbf{u}\right] \boldsymbol{\pi}\left[\mathbf{u} \,|\, \mathbf{x}\right] \mathrm{d}\mathbf{x}_+ \mathrm{d}\mathbf{u}$$

Score function

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right] &= \boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right] \frac{\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right]}{\boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right]} \\ &= \boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right] \underbrace{\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right]}_{\text{Score function}}\end{aligned}$$

Score function

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right] = \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right] \frac{\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right]}{\boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right]}$$
$$= \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right] \underbrace{\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right]}_{\text{Score function}}$$

E.g. Gaussian stochastic policy:

$$\boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right] = \frac{1}{\sqrt{2\pi \det \left(\boldsymbol{\Sigma} \right)}} e^{-\frac{1}{2} \left(\mathbf{u}_k - \bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}} \left(\mathbf{x}_k \right) \right)^{\top} \boldsymbol{\Sigma}^{-1} \left(\mathbf{u}_k - \bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}} \left(\mathbf{x}_k \right) \right)}$$

has score function: $\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right] = \left(\mathbf{u}_k - \bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}} \left(\mathbf{x}_k \right) \right)^{\top} \boldsymbol{\Sigma}^{-1} \nabla_{\boldsymbol{\theta}} \bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}} \left(\mathbf{x}_k \right)$

Score function

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right] = \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right] \frac{\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right]}{\boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right]}$$
$$= \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right] \underbrace{\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right]}_{\text{Score function}}$$

E.g. Gaussian stochastic policy:

$$\boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right] = \frac{1}{\sqrt{2\pi \det \left(\boldsymbol{\Sigma} \right)}} e^{-\frac{1}{2} \left(\mathbf{u}_k - \bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}} \left(\mathbf{x}_k \right) \right)^{\top} \boldsymbol{\Sigma}^{-1} \left(\mathbf{u}_k - \bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}} \left(\mathbf{x}_k \right) \right)}$$

has score function: $\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right] = \left(\mathbf{u}_k - \bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}} \left(\mathbf{x}_k \right) \right)^{\top} \boldsymbol{\Sigma}^{-1} \nabla_{\boldsymbol{\theta}} \bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}} \left(\mathbf{x}_k \right)$

Why is the score function useful?

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\varphi(\mathbf{x}, \mathbf{u}) \right] = \nabla_{\boldsymbol{\theta}} \int \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] \varphi(\mathbf{x}, \mathbf{u}) \mathrm{d}\mathbf{u} = \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] \varphi(\mathbf{x}, \mathbf{u}) \right]$$

conversion $\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[. \right] \rightarrow \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{\theta}} . \right]$

Score function

$$\nabla_\theta \pi_\theta [\mathbf{u}_k \,|\, \mathbf{x}_k] = \pi_\theta [\mathbf{u}_k \,|\, \mathbf{x}_k] \frac{\nabla_\theta \pi_\theta [\mathbf{u}_k \,|\, \mathbf{x}_k]}{\pi_\theta [\mathbf{u}_k \,|\, \mathbf{x}_k]}$$
$$= \pi_\theta [\mathbf{u}_k \,|\, \mathbf{x}_k] \underbrace{\nabla_\theta \log \pi_\theta [\mathbf{u}_k \,|\, \mathbf{x}_k]}_{\text{Score function}}$$

Score $\equiv$ relative change:

$$\nabla_\theta \log \pi_\theta = \frac{\nabla_\theta \pi_\theta [\mathbf{u}_k \,|\, \mathbf{x}_k]}{\pi_\theta [\mathbf{u}_k \,|\, \mathbf{x}_k]}$$

E.g. Gaussian stochastic policy:

$$\pi_\theta [\mathbf{u}_k \,|\, \mathbf{x}_k] = \frac{1}{\sqrt{2\pi \det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{u}_k - \bar{\pi}_\theta(\mathbf{x}_k))^\top \Sigma^{-1} (\mathbf{u}_k - \bar{\pi}_\theta(\mathbf{x}_k))}$$

has score function: $\nabla_\theta \log \pi_\theta [\mathbf{u}_k \,|\, \mathbf{x}_k] = (\mathbf{u}_k - \bar{\pi}_\theta(\mathbf{x}_k))^\top \Sigma^{-1} \nabla_\theta \bar{\pi}_\theta(\mathbf{x}_k)$

Why is the score function useful?

$$\nabla_\theta \mathbb{E}_{\pi_\theta} [\varphi(\mathbf{x}, \mathbf{u})] = \nabla_\theta \int \pi_\theta [\mathbf{u} \,|\, \mathbf{x}] \varphi(\mathbf{x}, \mathbf{u}) d\mathbf{u} = \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta [\mathbf{u} \,|\, \mathbf{x}] \varphi(\mathbf{x}, \mathbf{u})]$$

conversion $\nabla_\theta \mathbb{E}_{\pi_\theta} [.] \to \mathbb{E}_{\pi_\theta} [\nabla_\theta .]$

Score function & policy gradient

Deterministic policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}_k)$ performance

$$J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i+1} \,|\, \mathbf{x}_i,\ \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \cdot \mathrm{d}\mathbf{x}_0 \ldots \mathrm{d}\mathbf{x}_{k-1}$$

Stochastic policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right]$ performance

$$J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) =$$

$$= \sum_{k=0}^{\infty} \int \gamma^k \boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right] L(\mathbf{x}_k, \mathbf{u}_k) \prod_{i=0}^{k-1} \boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_i \,|\, \mathbf{x}_i\right] \mathbb{P}\left[\mathbf{x}_{i+1} \,|\, \mathbf{x}_i, \mathbf{u}_i\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0,\ldots,k-1} \mathrm{d}\mathbf{u}_{0,\ldots,k-1}$$

Score function & policy gradient

Deterministic policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}_k)$ performance

$$J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \sum_{k=0}^{\infty} \int \gamma^k L(\mathbf{x}_k, \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}_k)) \prod_{i=0}^{k-1} \mathbb{P}\left[\mathbf{x}_{i+1} \,|\, \mathbf{x}_i,\, \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}_i)\right] \mathbb{P}\left[\mathbf{x}_0\right] \cdot \mathrm{d}\mathbf{x}_0 \ldots \mathrm{d}\mathbf{x}_{k-1}$$

Stochastic policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right]$ performance

$$J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) =$$

$$= \sum_{k=0}^{\infty} \int \gamma^k \boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right] L(\mathbf{x}_k, \mathbf{u}_k) \prod_{i=0}^{k-1} \boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u}_i \,|\, \mathbf{x}_i\right] \mathbb{P}\left[\mathbf{x}_{i+1} \,|\, \mathbf{x}_i, \mathbf{u}_i\right] \mathbb{P}\left[\mathbf{x}_0\right] \mathrm{d}\mathbf{x}_{0,\ldots,k-1} \mathrm{d}\mathbf{u}_{0,\ldots,k-1}$$

Gradient of stochastic policy performance

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u} \,|\, \mathbf{x}\right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})\right]$$

i.e. correlation between score function and action-value function

Policy gradient deterministic vs. stochastic

Gradient of stochastic policy

$$\nabla_\theta J(\boldsymbol{\pi}_\theta) = \mathbb{E}_{\boldsymbol{\pi}_\theta}\left[\nabla_\theta \log \boldsymbol{\pi}_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right] Q_{\boldsymbol{\pi}_\theta}\left(\mathbf{x}, \mathbf{u}\right)\right]$$

Gradient of deterministic policy

$$\nabla_\theta J(\bar{\boldsymbol{\pi}}_\theta) = \mathbb{E}_{\bar{\boldsymbol{\pi}}_\theta}\left[\nabla_\theta \bar{\boldsymbol{\pi}}_\theta\left(\mathbf{x}\right) \nabla_{\mathbf{u}} Q_{\bar{\boldsymbol{\pi}}_\theta}\left(\mathbf{x}, \mathbf{u}\right)\right]$$

Policy gradient deterministic vs. stochastic

Gradient of stochastic policy

$$\nabla_\theta J(\boldsymbol{\pi}_\theta) = \mathbb{E}_{\boldsymbol{\pi}_\theta}\left[\nabla_\theta \log \boldsymbol{\pi}_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right] Q_{\boldsymbol{\pi}_\theta}\left(\mathbf{x}, \mathbf{u}\right)\right]$$

Gradient of deterministic policy

$$\nabla_\theta J(\bar{\boldsymbol{\pi}}_\theta) = \mathbb{E}_{\bar{\boldsymbol{\pi}}_\theta}\left[\nabla_\theta \bar{\boldsymbol{\pi}}_\theta\left(\mathbf{x}\right) \nabla_{\mathbf{u}} Q_{\bar{\boldsymbol{\pi}}_\theta}\left(\mathbf{x}, \mathbf{u}\right)\right]$$

Let Σ be the covariance of $\boldsymbol{\pi}_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right]$ then as stochastic$\rightarrow$deterministic...

$$\lim_{\Sigma \rightarrow 0} \boldsymbol{\pi}_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right] = \delta(\mathbf{u} - \bar{\boldsymbol{\pi}}_\theta\left(\mathbf{x}\right))$$

Policy gradient deterministic vs. stochastic

Gradient of stochastic policy

$$\nabla_\theta J(\boldsymbol{\pi}_\theta) = \mathbb{E}_{\boldsymbol{\pi}_\theta}\left[\nabla_\theta \log \boldsymbol{\pi}_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right] Q_{\boldsymbol{\pi}_\theta}\left(\mathbf{x}, \mathbf{u}\right)\right]$$

Gradient of deterministic policy

$$\nabla_\theta J(\bar{\boldsymbol{\pi}}_\theta) = \mathbb{E}_{\bar{\boldsymbol{\pi}}_\theta}\left[\nabla_\theta \bar{\boldsymbol{\pi}}_\theta\left(\mathbf{x}\right) \nabla_{\mathbf{u}} Q_{\bar{\boldsymbol{\pi}}_\theta}\left(\mathbf{x}, \mathbf{u}\right)\right]$$

Let Σ be the covariance of $\boldsymbol{\pi}_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right]$ then as stochastic$\rightarrow$deterministic...

$$\lim_{\Sigma \rightarrow 0} \boldsymbol{\pi}_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right] = \delta(\mathbf{u} - \bar{\boldsymbol{\pi}}_\theta\left(\mathbf{x}\right))$$

...policy gradients match:

$$\lim_{\Sigma \rightarrow 0} \mathbb{E}_{\boldsymbol{\pi}_\theta}\left[\nabla_\theta \log \boldsymbol{\pi}_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right] Q_{\boldsymbol{\pi}_\theta}\left(\mathbf{x}, \mathbf{u}\right)\right] = \mathbb{E}_{\bar{\boldsymbol{\pi}}_\theta}\left[\nabla_\theta \bar{\boldsymbol{\pi}}_\theta\left(\mathbf{x}\right) \nabla_{\mathbf{u}} Q_{\bar{\boldsymbol{\pi}}_\theta}\left(\mathbf{x}, \mathbf{u}\right)\right]$$

...under some regularity assumptions

Interpretation of stochastic policy gradient?

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right) \right]$$

What are the terms $\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right)$?

Interpretation of stochastic policy gradient?

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right) \right]$$

What are the terms $\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right)$?

- $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is scalar, acts like a "weight"

Interpretation of stochastic policy gradient?

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right) \right]$$

What are the terms $\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right)$?

- $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is scalar, acts like a "weight"
- Change $\boldsymbol{\theta}$ in "direction" $-\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right]$...

  ...decreases the probability of drawing $\mathbf{u}$ in state $\mathbf{x}$

Interpretation of stochastic policy gradient?

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right) \right]$$

What are the terms $\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right)$?

- $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is scalar, acts like a "weight"
- Change $\boldsymbol{\theta}$ in "direction" $-\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right]$...
  ...decreases the probability of drawing $\mathbf{u}$ in state $\mathbf{x}$
- Consider $\boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right] = \mathcal{N}(\bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}} \left(\mathbf{x}_k \right), \Sigma)$

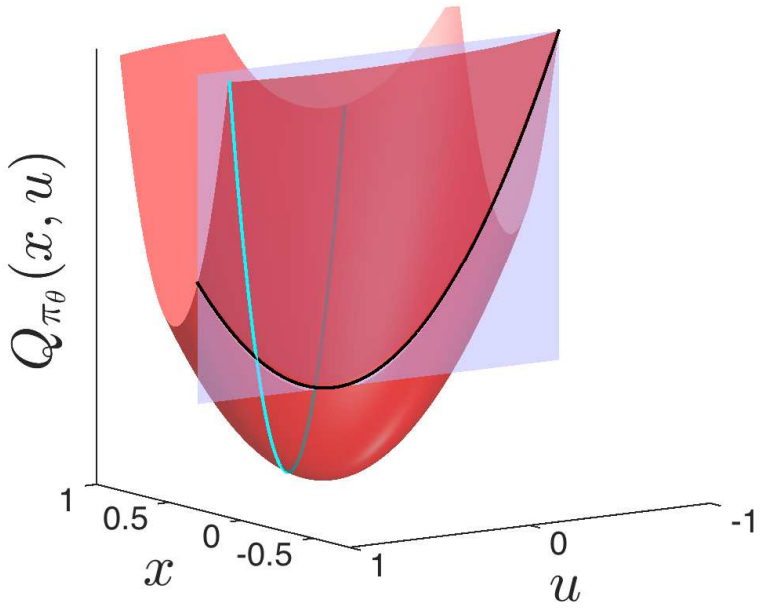
Interpretation of stochastic policy gradient?

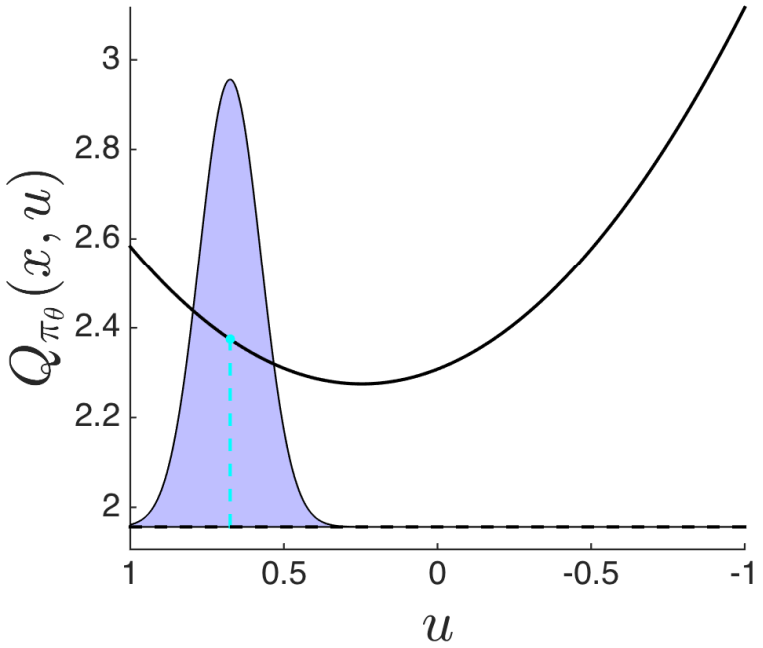
$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right) \right]$$

What are the terms $\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right)$?

- $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is scalar, acts like a "weight"
- Change $\boldsymbol{\theta}$ in "direction" $-\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right]$...

...decreases the probability of drawing $\mathbf{u}$ in state $\mathbf{x}$



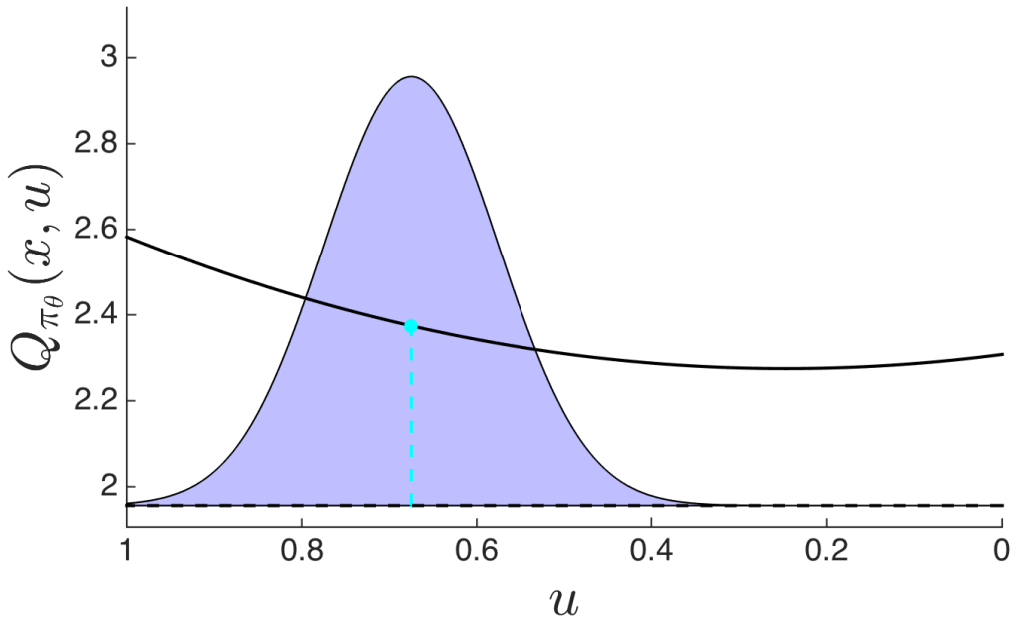


Interpretation of stochastic policy gradient?

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right) \right]$$

What are the terms $\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right)$?

- $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is scalar, acts like a "weight"
- Change $\boldsymbol{\theta}$ in "direction" $-\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right]$...
  ...decreases the probability of drawing $\mathbf{u}$ in state $\mathbf{x}$
- Consider $\boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right] = \mathcal{N}(\bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}} \left(\mathbf{x}_k \right), \Sigma)$

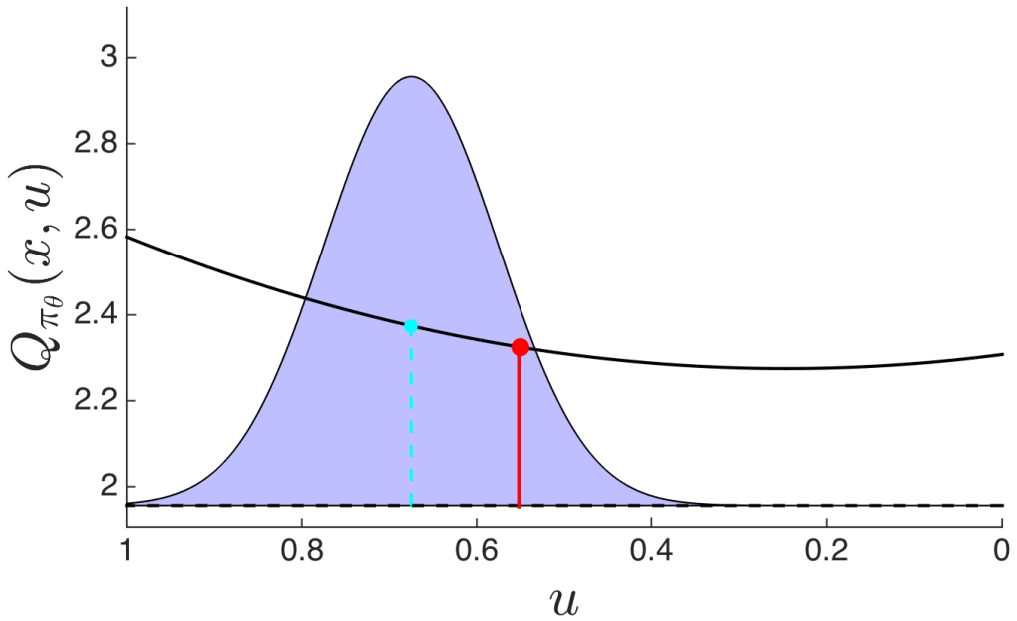


Interpretation of stochastic policy gradient?

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right) \right]$$

What are the terms $\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right)$?

- $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is scalar, acts like a "weight"
- Change $\boldsymbol{\theta}$ in "direction" $-\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right]$...
  ...decreases the probability of drawing $\mathbf{u}$ in state $\mathbf{x}$
- Consider $\boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right] = \mathcal{N}(\bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}} \left(\mathbf{x}_k \right), \Sigma)$



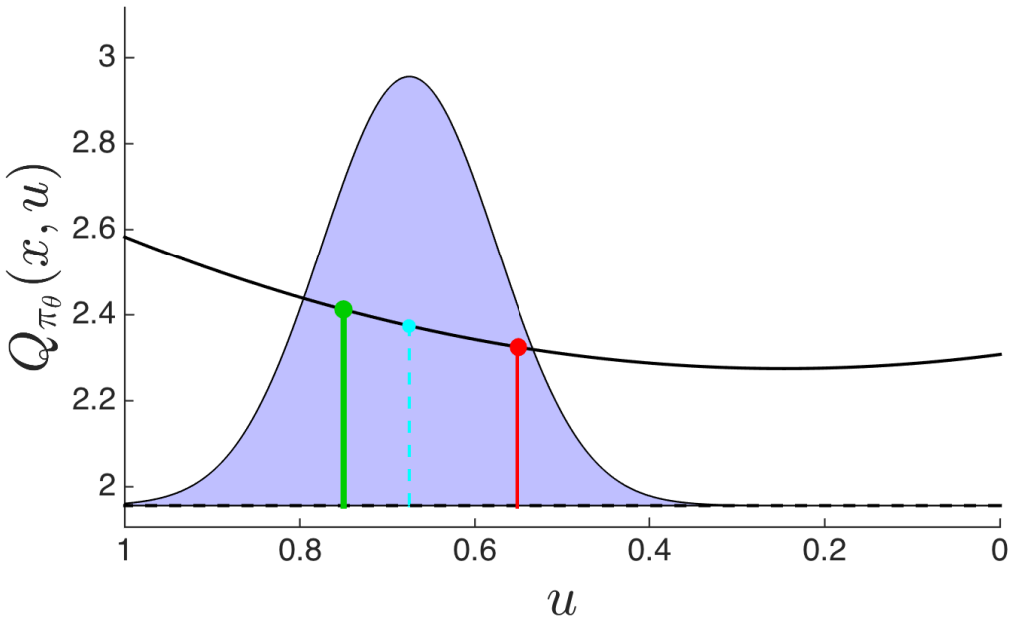
- $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right) > 0$

Interpretation of stochastic policy gradient?

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right) \right]$$

What are the terms $\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right)$?

- $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is scalar, acts like a "weight"
- Change $\boldsymbol{\theta}$ in "direction" $-\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right]$...
  ...decreases the probability of drawing $\mathbf{u}$ in state $\mathbf{x}$
- Consider $\boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u}_k \,|\, \mathbf{x}_k \right] = \mathcal{N}(\bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}} \left(\mathbf{x}_k \right), \Sigma)$



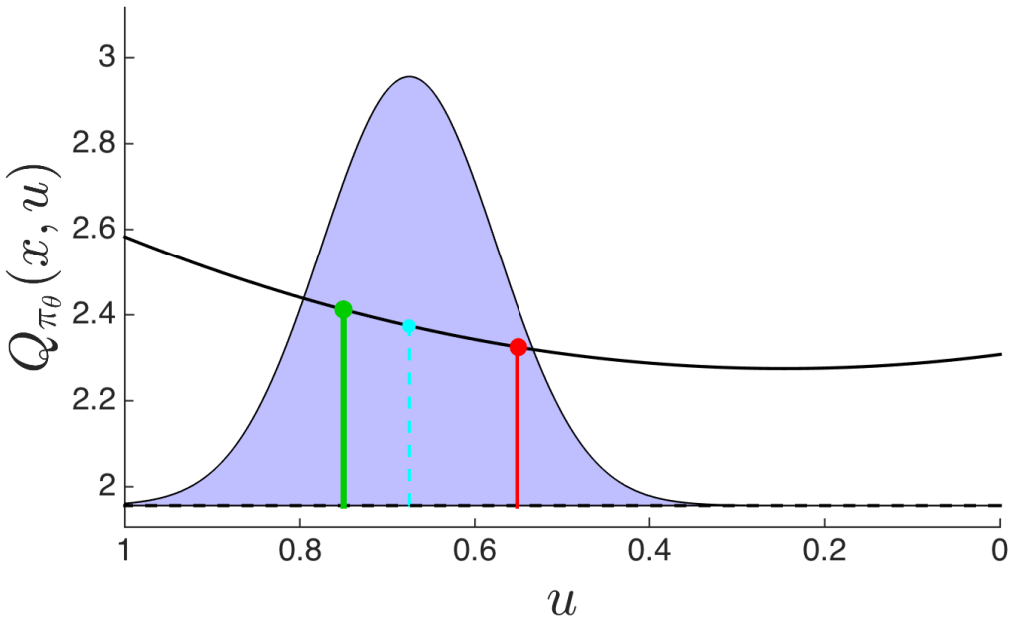
- $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right) > 0$
- $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right) > Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left(\mathbf{x}, \mathbf{u} \right)$

Interpretation of stochastic policy gradient?

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right] Q_{\pi_\theta}\left(\mathbf{x}, \mathbf{u}\right)\right]$$

What are the terms $\nabla_\theta \log \pi_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right] Q_{\pi_\theta}\left(\mathbf{x}, \mathbf{u}\right)$?

- Q_{π_θ} is scalar, acts like a "weight"
- Change θ in "direction" $-\nabla_\theta \log \pi_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right]$...
  ...decreases the probability of drawing $\mathbf{u}$ in state $\mathbf{x}$
- Consider $\pi_\theta\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right] = \mathcal{N}(\bar{\pi}_\theta\left(\mathbf{x}_k\right), \Sigma)$



- $Q_{\pi_\theta}\left(\mathbf{x}, \mathbf{u}\right) > 0$
- $Q_{\pi_\theta}\left(\mathbf{x}, \mathbf{u}\right) > Q_{\pi_\theta}\left(\mathbf{x}, \mathbf{u}\right)$

Step in direction

$$-\nabla_\theta \log \pi_\theta Q_{\pi_\theta}$$

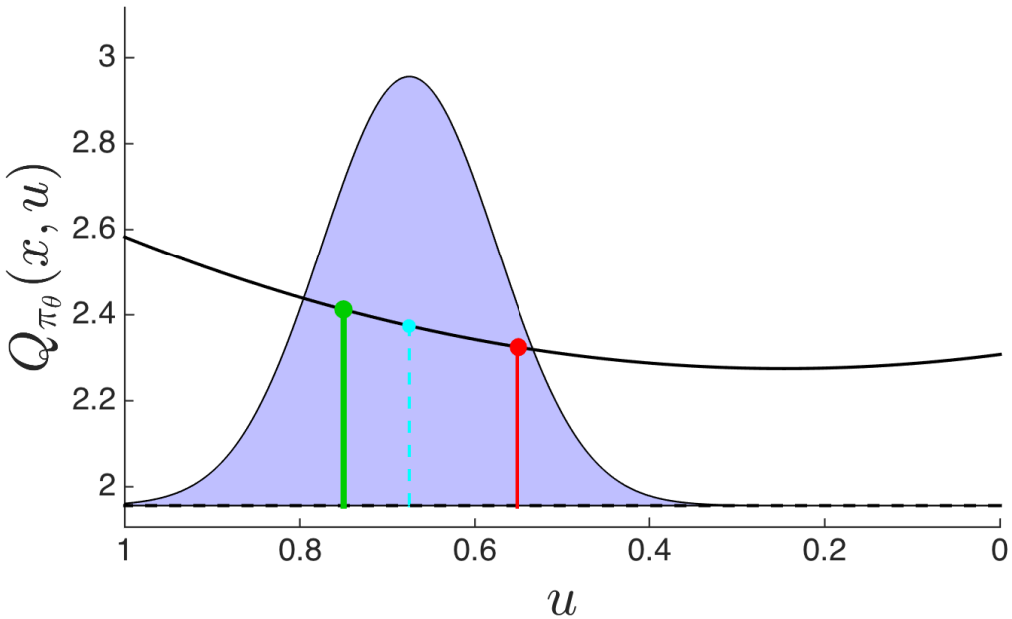
"pushes" $\bar{\pi}_\theta$ away

Interpretation of stochastic policy gradient?

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right] Q_{\pi_\theta}\left(\mathbf{x}, \mathbf{u}\right)\right]$$

What are the terms $\nabla_\theta \log \pi_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right] Q_{\pi_\theta}\left(\mathbf{x}, \mathbf{u}\right)$?

- Q_{π_θ} is scalar, acts like a "weight"
- Change θ in "direction" $-\nabla_\theta \log \pi_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right]$...
  ...decreases the probability of drawing $\mathbf{u}$ in state $\mathbf{x}$
- Consider $\pi_\theta\left[\mathbf{u}_k \,|\, \mathbf{x}_k\right] = \mathcal{N}(\bar{\pi}_\theta\left(\mathbf{x}_k\right), \Sigma)$



- $Q_{\pi_\theta}\left(\mathbf{x}, \mathbf{u}\right) > 0$
- $Q_{\pi_\theta}\left(\mathbf{x}, \mathbf{u}\right) > Q_{\pi_\theta}\left(\mathbf{x}, \mathbf{u}\right)$

Step in direction

$$-\nabla_\theta \log \pi_\theta Q_{\pi_\theta}$$

"pushes" $\bar{\pi}_\theta$ away

Both $\mathbf{u}$ and $\mathbf{u}$ "push" $\bar{\pi}_\theta$ away from them, $\mathbf{u}$ "pushes" harder, $\bar{\pi}_\theta$ in average moves to the right

Stochastic policy gradient - Advantage critic

Baseline:

$$\mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] B \left(\mathbf{x} \right) \right] = 0$$

holds for any baseline function B

Key idea: use baseline to prevent "everyone from pushing"

Stochastic policy gradient - Advantage critic

Baseline:

$$\mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta\left[\mathbf{u}\,|\,\mathbf{x}\right] B\left(\mathbf{x}\right)\right] = 0$$

holds for any baseline function B

Key idea: use baseline to prevent "everyone from pushing"

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta\left[\mathbf{u}\,|\,\mathbf{x}\right] Q_{\pi_\theta}\left(\mathbf{x},\mathbf{u}\right)\right] \\ &= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta\left[\mathbf{u}\,|\,\mathbf{x}\right] \left(Q_{\pi_\theta}\left(\mathbf{x},\mathbf{u}\right) - V_{\pi_\theta}\left(\mathbf{x}\right)\right)\right] \\ &= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta\left[\mathbf{u}\,|\,\mathbf{x}\right] A_{\pi_\theta}\left(\mathbf{x},\mathbf{u}\right)\right]\end{aligned}$$

A_{π_θ} is the Advantage function

Stochastic policy gradient - Advantage critic

Baseline:

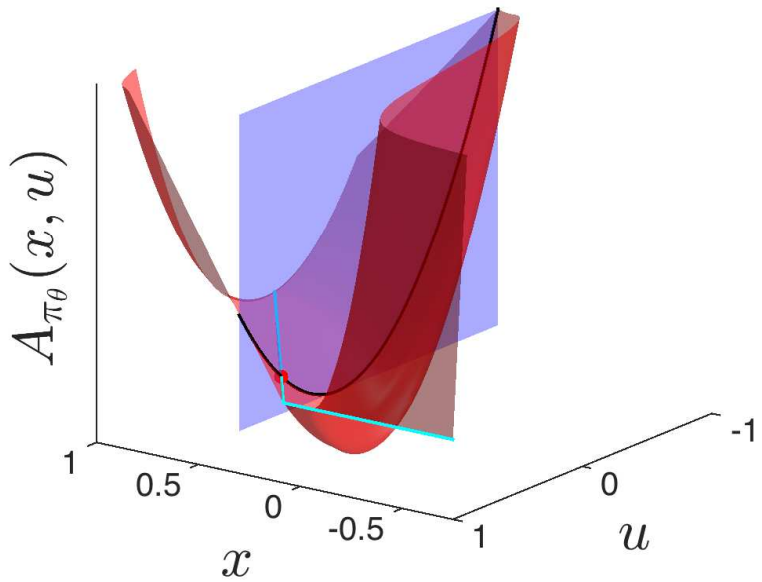
$$\mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right] B\left(\mathbf{x}\right)\right] = 0$$

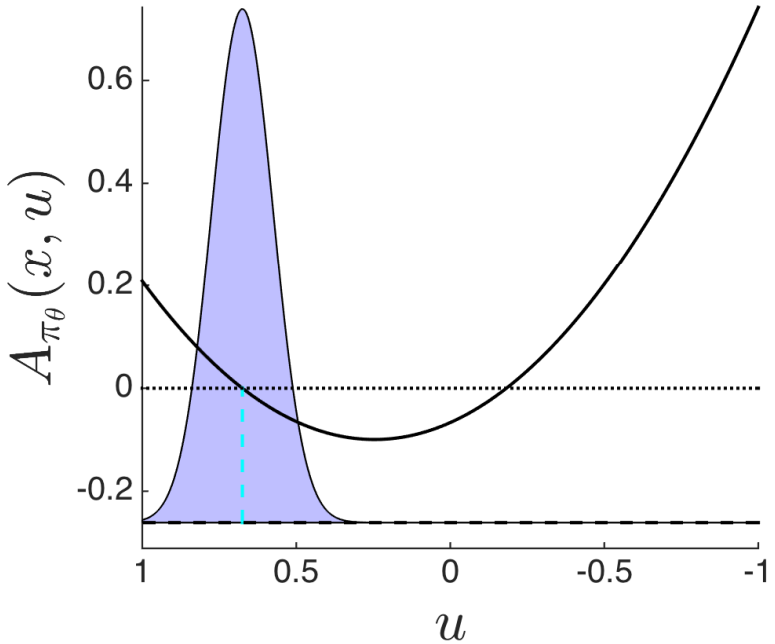
holds for any baseline function B

Key idea: use baseline to prevent "everyone from pushing"

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta\left[\mathbf{u} \,|\, \mathbf{x}\right] A_{\pi_\theta}\left(\mathbf{x}, \mathbf{u}\right)\right]$$

A_{π_θ} is the Advantage function





Stochastic policy gradient - Advantage critic

Baseline:

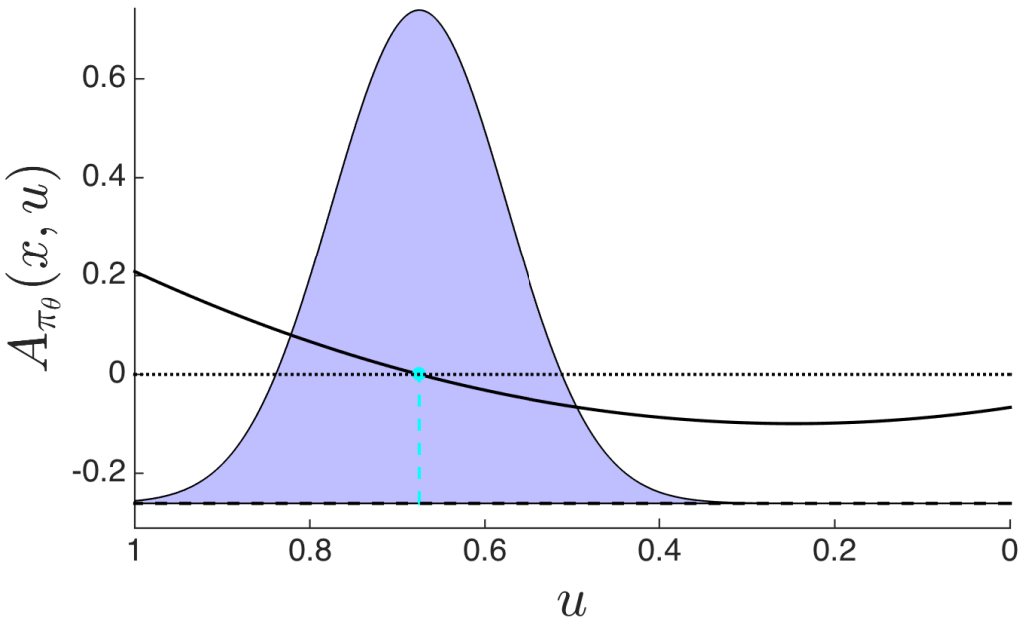
$$\mathbb{E}_{\boldsymbol{\pi_\theta}}\left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi_\theta}\left[\mathbf{u} \,|\, \mathbf{x}\right] B\left(\mathbf{x}\right)\right] = 0$$

holds for any baseline function B

Key idea: use baseline to prevent "everyone from pushing"

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi_\theta}) = \mathbb{E}_{\boldsymbol{\pi_\theta}}\left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi_\theta}\left[\mathbf{u} \,|\, \mathbf{x}\right] A_{\boldsymbol{\pi_\theta}}\left(\mathbf{x}, \mathbf{u}\right)\right]$$

$A_{\boldsymbol{\pi_\theta}}$ is the Advantage function



Step towards $-\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi_\theta} A_{\boldsymbol{\pi_\theta}}$

Stochastic policy gradient - Advantage critic

Baseline:

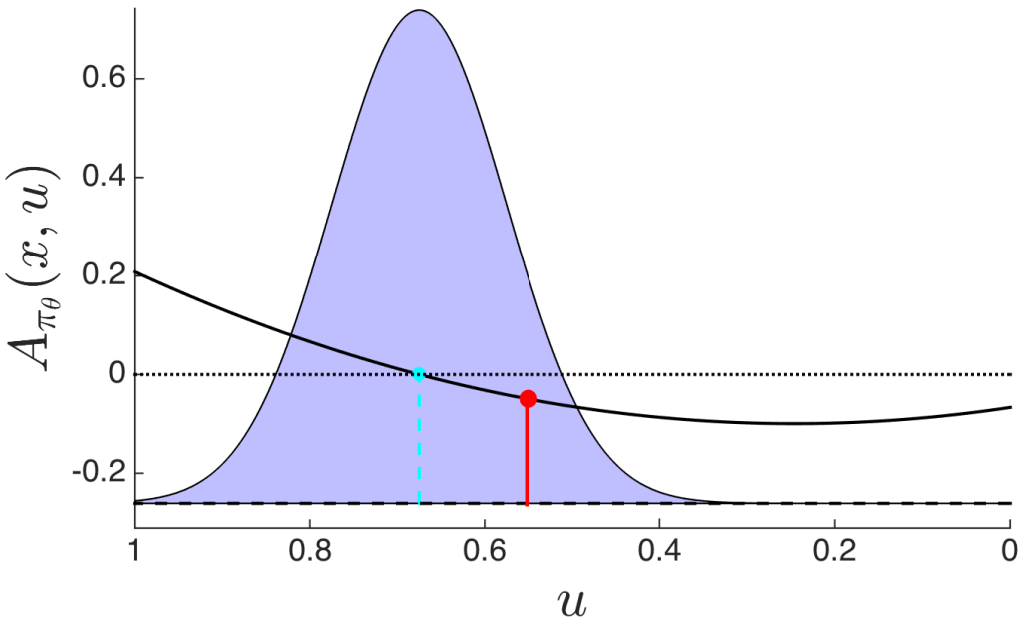
$$\mathbb{E}_{\boldsymbol{\pi_\theta}}\left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi_\theta}\left[\mathbf{u} \,|\, \mathbf{x}\right] B\left(\mathbf{x}\right)\right] = 0$$

holds for any baseline function B

Key idea: use baseline to prevent "everyone from pushing"

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi_\theta}) = \mathbb{E}_{\boldsymbol{\pi_\theta}}\left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi_\theta}\left[\mathbf{u} \,|\, \mathbf{x}\right] A_{\boldsymbol{\pi_\theta}}\left(\mathbf{x}, \mathbf{u}\right)\right]$$

$A_{\boldsymbol{\pi_\theta}}$ is the Advantage function



Step towards $-\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi_\theta} A_{\boldsymbol{\pi_\theta}}$

- $A_{\boldsymbol{\pi_\theta}}\left(\mathbf{x}, \mathbf{u}\right) < 0$ pulls

Stochastic policy gradient - Advantage critic

Baseline:

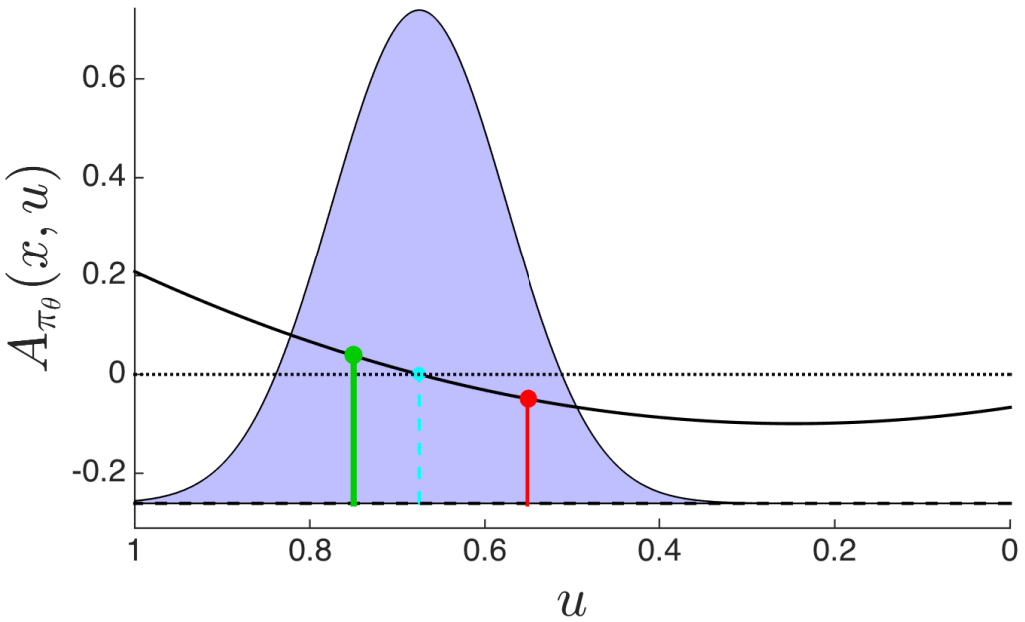
$$\mathbb{E}_{\boldsymbol{\pi_\theta}}\left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi_\theta}\left[\mathbf{u} \,|\, \mathbf{x}\right] B\left(\mathbf{x}\right)\right] = 0$$

holds for any baseline function B

Key idea: use baseline to prevent "everyone from pushing"

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi_\theta}) = \mathbb{E}_{\boldsymbol{\pi_\theta}}\left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi_\theta}\left[\mathbf{u} \,|\, \mathbf{x}\right] A_{\boldsymbol{\pi_\theta}}\left(\mathbf{x}, \mathbf{u}\right)\right]$$

$A_{\boldsymbol{\pi_\theta}}$ is the Advantage function



Step towards $-\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi_\theta} A_{\boldsymbol{\pi_\theta}}$

- $A_{\boldsymbol{\pi_\theta}}\left(\mathbf{x}, \mathbf{u}\right) < 0$ pulls
- $A_{\boldsymbol{\pi_\theta}}\left(\mathbf{x}, \mathbf{u}\right) > 0$ pushes

Stochastic policy gradient - Advantage critic

Baseline:

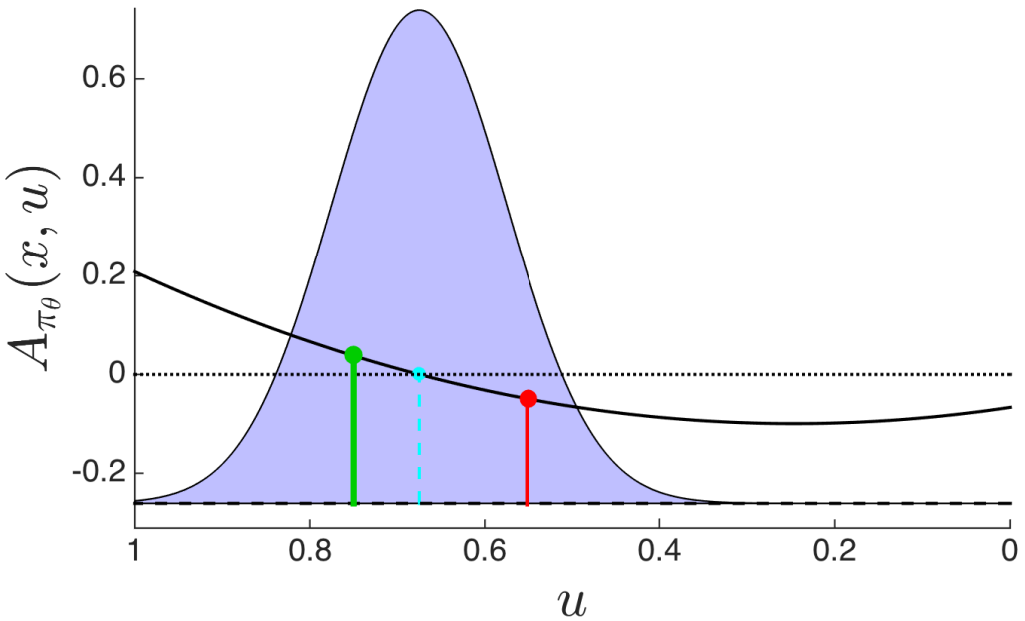
$$\mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u} \,|\, \mathbf{x}\right] B\left(\mathbf{x}\right)\right] = 0$$

holds for any baseline function B

Key idea: use baseline to prevent "everyone from pushing"

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}}\left[\mathbf{u} \,|\, \mathbf{x}\right] A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{x}, \mathbf{u}\right)\right]$$

$A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is the Advantage function



Step towards $-\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$

- $A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{x}, \mathbf{u}\right) < 0$ pulls
- $A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{x}, \mathbf{u}\right) > 0$ pushes

Sample $\mathbf{u}$ "pulls" and $\mathbf{u}$ "pushes" $\bar{\boldsymbol{\pi}}_{\boldsymbol{\theta}}$, drive it to $A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} = 0$

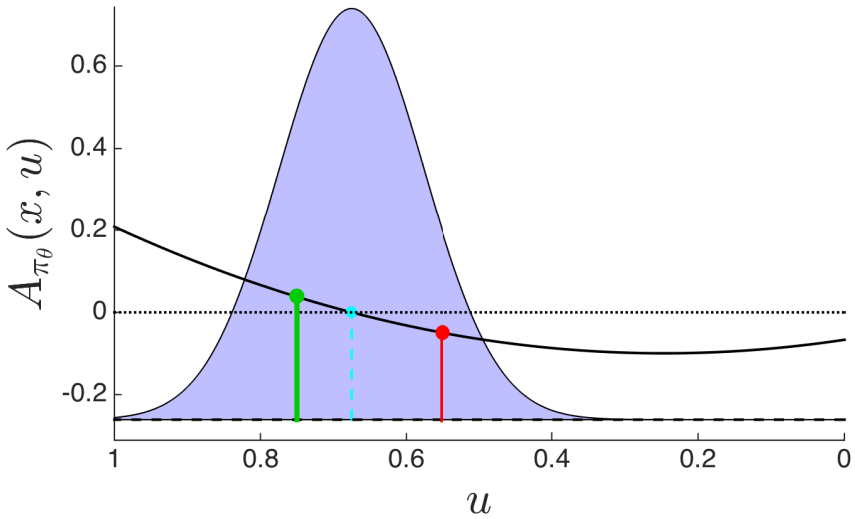
Stochastic policy gradient

Algorithm: RL using stochastic π-gradient

Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

- Run policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ for $N \to \infty$ samples (yields $\mathbf{x}_0, \mathbf{u}_0, \ldots \mathbf{x}_{N-1}, \mathbf{u}_{N-1}$)
- Evaluate $Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using (LS)TD-learning, MC-like, ...
- Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \frac{1}{N} \sum_{k=0}^{N-1} \nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} [\mathbf{u}_k | \mathbf{x}_k] \, Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} (\mathbf{x}_k, \mathbf{u}_k)$
- Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$

return Optimized policy parameters $\boldsymbol{\theta}$



Stochastic policy gradient

Algorithm: RL using stochastic π-gradient

Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

- Run policy $\pi_{\boldsymbol{\theta}}$ for $N \to \infty$ samples (yields $\mathbf{x}_0, \mathbf{u}_0, \ldots \mathbf{x}_{N-1}, \mathbf{u}_{N-1}$)
- Evaluate $Q_{\pi_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using (LS)TD-learning, MC-like, ...
- Compute $\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) = \frac{1}{N} \sum_{k=0}^{N-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}} [\mathbf{u}_k | \mathbf{x}_k] \, Q_{\pi_{\boldsymbol{\theta}}} (\mathbf{x}_k, \mathbf{u}_k)$
- Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}})$

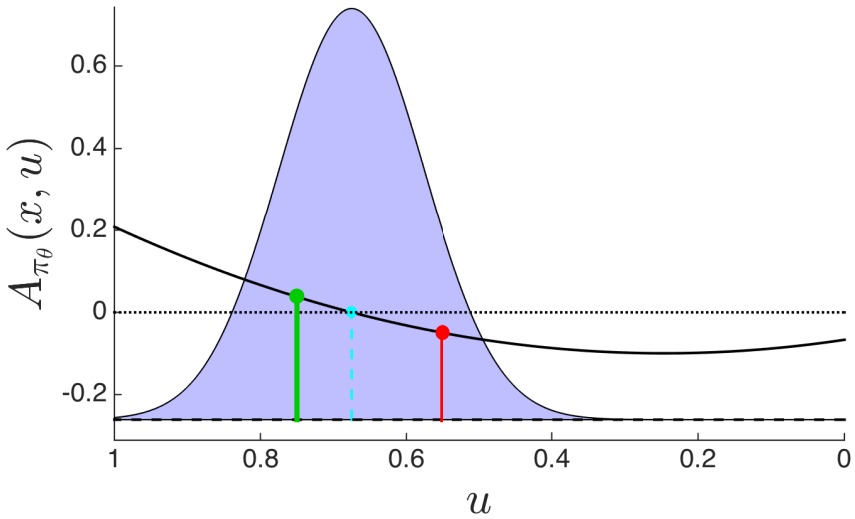
return Optimized policy parameters $\boldsymbol{\theta}$

In practice, $Q_{\pi_{\boldsymbol{\theta}}}$ approximated:

$$Q_{\mathbf{w}} \approx Q_{\pi_{\boldsymbol{\theta}}}$$

Labelled **Actor-Critic**:

- $\pi_{\boldsymbol{\theta}}$ is the "actor"
- $Q_{\mathbf{w}}$ is the "critic"



Stochastic policy gradient

Algorithm: RL using stochastic π-gradient

Input: Initial policy parameters $\boldsymbol{\theta}$, step-size $\alpha > 0$
while $\|\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}})\| > \mathrm{Tol}$ **do**

- Run policy $\pi_{\boldsymbol{\theta}}$ for $N \to \infty$ samples (yields $\mathbf{x}_0, \mathbf{u}_0, \ldots \mathbf{x}_{N-1}, \mathbf{u}_{N-1}$)
- Evaluate $A_{\pi_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{u})$ using (LS)TD-learning, MC-like, ...
- Compute $\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) = \frac{1}{N} \sum_{k=0}^{N-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}} [\mathbf{u}_k | \mathbf{x}_k] A_{\pi_{\boldsymbol{\theta}}} (\mathbf{x}_k, \mathbf{u}_k)$
- Gradient step: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}})$

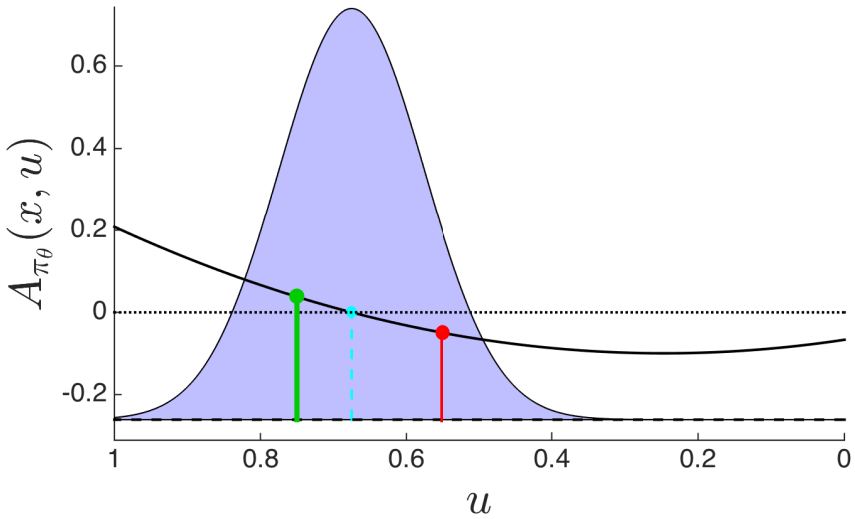
return Optimized policy parameters $\boldsymbol{\theta}$

In practice, $Q_{\pi_{\boldsymbol{\theta}}}$ approximated:

$$Q_{\mathbf{w}} \approx Q_{\pi_{\boldsymbol{\theta}}}$$

Labelled **Actor-Critic**:

- $\pi_{\boldsymbol{\theta}}$ is the "actor"
- $Q_{\mathbf{w}}$ is the "critic"



Stochastic policy gradient - Compatible function approximation

What is the impact of approximation $Q_{\mathbf{w}} \approx Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$?

If the following conditions hold:

- $\nabla_{\mathbf{w}} Q_{\mathbf{w}} = \nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}}$
- $\mathbf{w} = \mathrm{a}\min_{\mathbf{w}} \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\left(Q_{\mathbf{w}} - Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \right)^2 \right]$

then

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}_{\boldsymbol{\theta}} \left[\mathbf{u} \,|\, \mathbf{x} \right] Q_{\mathbf{w}} \left(\mathbf{x}, \mathbf{u} \right) \right]$$

What about $A_{\mathbf{w}} \approx Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} - V_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$? Same story if

$$A_{\mathbf{w}} = Q_{\mathbf{w}} - \hat{V}$$

where $Q_{\mathbf{w}}$ is compatible and $\hat{V}$ evaluated independently of $\mathbf{w}$

Stochastic policy gradient - TD critic

Consider TD error:

$$\delta\left(\mathbf{x}, \mathbf{u}, \mathbf{x}_{+}\right)=L\left(\mathbf{x}, \mathbf{u}\right)+\gamma V_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{x}_{+}\right)-V_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{x}\right)$$

We observe that:

$$\begin{aligned}\mathbb{E}\left[\delta \,|\, \mathbf{x}, \mathbf{u}\right] &= \mathbb{E}\left[L\left(\mathbf{x}, \mathbf{u}\right)+\gamma V_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{x}_{+}\right)\right]-V_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{x}\right) \\ &= Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{x}, \mathbf{u}\right)-V_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{x}\right) \\ &= A_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left(\mathbf{x}, \mathbf{u}\right)\end{aligned}$$

Stochastic policy gradient - TD critic

Consider TD error:

$$\delta(\mathbf{x}, \mathbf{u}, \mathbf{x}_+) = L(\mathbf{x}, \mathbf{u}) + \gamma V_{\pi_\theta}(\mathbf{x}_+) - V_{\pi_\theta}(\mathbf{x})$$

We observe that:

$$\begin{aligned}\mathbb{E}[\delta \,|\, \mathbf{x}, \mathbf{u}] &= \mathbb{E}[L(\mathbf{x}, \mathbf{u}) + \gamma V_{\pi_\theta}(\mathbf{x}_+)] - V_{\pi_\theta}(\mathbf{x}) \\ &= Q_{\pi_\theta}(\mathbf{x}, \mathbf{u}) - V_{\pi_\theta}(\mathbf{x}) \\ &= A_{\pi_\theta}(\mathbf{x}, \mathbf{u})\end{aligned}$$

Hence we have thee possible critics:

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta[\mathbf{u} \,|\, \mathbf{x}] A_{\pi_\theta}(\mathbf{x}, \mathbf{u})] \\ &= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta[\mathbf{u} \,|\, \mathbf{x}] \delta(\mathbf{x}, \mathbf{u}, \mathbf{x}_+)] \\ &= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta[\mathbf{u} \,|\, \mathbf{x}] Q_{\pi_\theta}(\mathbf{x}, \mathbf{u})]\end{aligned}$$

Stochastic policy gradient - TD critic

Consider TD error:

$$\delta(\mathbf{x}, \mathbf{u}, \mathbf{x}_+) = L(\mathbf{x}, \mathbf{u}) + \gamma V_{\pi_\theta}(\mathbf{x}_+) - V_{\pi_\theta}(\mathbf{x})$$

We observe that:

$$\begin{aligned}\mathbb{E}[\delta \mid \mathbf{x}, \mathbf{u}] &= \mathbb{E}[L(\mathbf{x}, \mathbf{u}) + \gamma V_{\pi_\theta}(\mathbf{x}_+)] - V_{\pi_\theta}(\mathbf{x}) \\ &= Q_{\pi_\theta}(\mathbf{x}, \mathbf{u}) - V_{\pi_\theta}(\mathbf{x}) \\ &= A_{\pi_\theta}(\mathbf{x}, \mathbf{u})\end{aligned}$$

- $Q_{\pi_\theta} \rightarrow$ all samples $\mathbf{x}, \mathbf{u}$ "push" the policy, bad samples push harder, policy moves to higher performance

Hence we have thee possible critics:

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta[\mathbf{u} \mid \mathbf{x}] A_{\pi_\theta}(\mathbf{x}, \mathbf{u})] \\ &= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta[\mathbf{u} \mid \mathbf{x}] \delta(\mathbf{x}, \mathbf{u}, \mathbf{x}_+)] \\ &= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta[\mathbf{u} \mid \mathbf{x}] Q_{\pi_\theta}(\mathbf{x}, \mathbf{u})]\end{aligned}$$

Stochastic policy gradient - TD critic

Consider TD error:

$$\delta(\mathbf{x}, \mathbf{u}, \mathbf{x}_+) = L(\mathbf{x}, \mathbf{u}) + \gamma V_{\pi_\theta}(\mathbf{x}_+) - V_{\pi_\theta}(\mathbf{x})$$

We observe that:

$$\begin{aligned}\mathbb{E}[\delta \mid \mathbf{x}, \mathbf{u}] &= \mathbb{E}[L(\mathbf{x}, \mathbf{u}) + \gamma V_{\pi_\theta}(\mathbf{x}_+)] - V_{\pi_\theta}(\mathbf{x}) \\ &= Q_{\pi_\theta}(\mathbf{x}, \mathbf{u}) - V_{\pi_\theta}(\mathbf{x}) \\ &= A_{\pi_\theta}(\mathbf{x}, \mathbf{u})\end{aligned}$$

Hence we have thee possible critics:

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta[\mathbf{u} \mid \mathbf{x}] A_{\pi_\theta}(\mathbf{x}, \mathbf{u})] \\ &= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta[\mathbf{u} \mid \mathbf{x}] \delta(\mathbf{x}, \mathbf{u}, \mathbf{x}_+)] \\ &= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta[\mathbf{u} \mid \mathbf{x}] Q_{\pi_\theta}(\mathbf{x}, \mathbf{u})]\end{aligned}$$

- $Q_{\pi_\theta} \rightarrow$ all samples $\mathbf{x}, \mathbf{u}$ "push" the policy, bad samples push harder, policy moves to higher performance
- $A_{\pi_\theta} \rightarrow$ good samples "pull" the policy towards them, bad samples "push" the policy away

Stochastic policy gradient - TD critic

Consider TD error:

$$\delta(\mathbf{x}, \mathbf{u}, \mathbf{x}_+) = L(\mathbf{x}, \mathbf{u}) + \gamma V_{\pi_\theta}(\mathbf{x}_+) - V_{\pi_\theta}(\mathbf{x})$$

We observe that:

$$\begin{aligned}\mathbb{E}[\delta \mid \mathbf{x}, \mathbf{u}] &= \mathbb{E}[L(\mathbf{x}, \mathbf{u}) + \gamma V_{\pi_\theta}(\mathbf{x}_+)] - V_{\pi_\theta}(\mathbf{x}) \\ &= Q_{\pi_\theta}(\mathbf{x}, \mathbf{u}) - V_{\pi_\theta}(\mathbf{x}) \\ &= A_{\pi_\theta}(\mathbf{x}, \mathbf{u})\end{aligned}$$

Hence we have thee possible critics:

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta[\mathbf{u} \mid \mathbf{x}] A_{\pi_\theta}(\mathbf{x}, \mathbf{u})] \\ &= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta[\mathbf{u} \mid \mathbf{x}] \delta(\mathbf{x}, \mathbf{u}, \mathbf{x}_+)] \\ &= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta[\mathbf{u} \mid \mathbf{x}] Q_{\pi_\theta}(\mathbf{x}, \mathbf{u})]\end{aligned}$$

- $Q_{\pi_\theta} \rightarrow$ all samples $\mathbf{x}, \mathbf{u}$ "push" the policy, bad samples push harder, policy moves to higher performance
- $A_{\pi_\theta} \rightarrow$ good samples "pull" the policy towards them, bad samples "push" the policy away
- $\delta \rightarrow$ noisy version of A_{π_θ}, estimation of Q_{π_θ} not needed