Recursive Constraints to Prevent Instability in Constrained Reinforcement Learning

Jaeyoung Lee*, Sean Sedwards*, Krzysztof Czarnecki

University of Waterloo, Canada

* contributed equally

Presented at MoDEM Workshop, 2021

Introduction

- Given Markov decision processes and safety constraints

  Find a policy $\hat{\pi}$ that is $\mathbb{P}(\text{reaching a failure state}) \leq \theta$

  deterministic

  uniformly constrained optimal i.e. $\begin{cases} \textit{safe and optimal} \text{ in each state possible} \\ \textit{least unsafe} \text{ in each of the other states} \end{cases}$

- Motivations

  Safety-critical systems e.g. autonomous driving

  No adequate existing solution

- Main focus

  1. instability issue with reinforcement learning

  2. solution: the idea of recursive constraints

Finite Markov Decision Process (MDP)

- $(\mathcal{S}^+, \mathcal{A}^+, \mathcal{T}, \gamma, \mathcal{R})$ where

    (finite) state space $\mathcal{S}^+ = \mathcal{S} \cup \mathcal{S}_\perp$ ($\mathcal{S}_\perp$: set of all terminal states)

    (finite) action space $\mathcal{A}^+$

    ➡ $\mathcal{A}(s)$: set of all actions $\in \mathcal{A}^+$ available from $s \in \mathcal{S}^+$

    next-state distribution $\mathcal{T}(s, a)$, given action $a \in \mathcal{A}(s)$ at state $s \in \mathcal{S}$

    discount rate $\gamma \in [0,1]$

    reward model $\mathcal{R} : \mathcal{S}^+ \times \mathcal{A}^+ \times \mathcal{S}^+ \to \mathbb{R}$

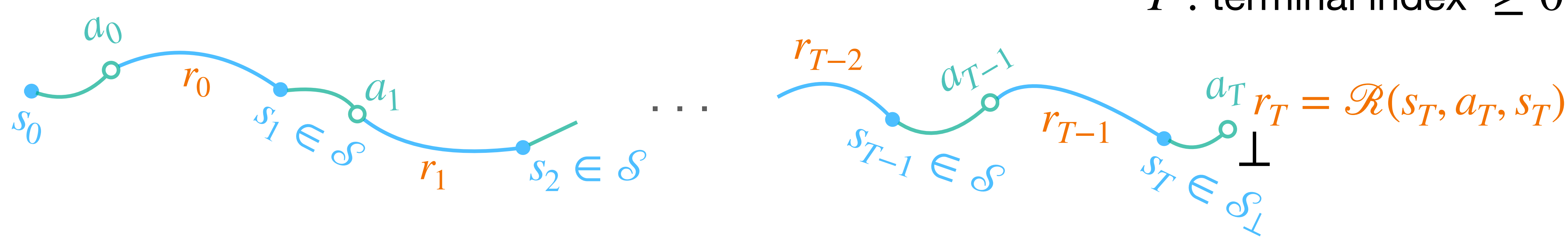
- A *policy* is mapping $\pi : \mathcal{S}^+ \to \mathcal{A}^+$ such that $\pi(s) \in \mathcal{A}(s) \quad \forall s \in \mathcal{S}^+$

States, Actions, Rewards and Value Functions

- Given $s \in \mathcal{S}^+$ (resp. $sa \in \mathcal{S}^+ \times \mathcal{A}(s)$)

  policy π over an MDP generates

T : terminal index ≥ 0



where $\begin{cases} s_0 = s \text{ (resp. } s_0 a_0 = sa) \text{ and } a_t = \pi(s_t) \text{ thereafter} \\ s_{t+1} \sim \mathcal{T}(s_t, a_t) \text{ and } r_t = \mathcal{R}(s_t, a_t, s_{t+1}) \quad \forall t < T \end{cases}$

- Value and Q-functions of policy π

$$V(s \mid \pi) := \mathbb{E}\left(\sum_{t=0}^{T} \gamma^t \cdot r_t \;\middle|\; s_0 = s, \pi \right)$$

$$Q(s, a \mid \pi) := \mathbb{E}\left(\sum_{t=0}^{T} \gamma^t \cdot r_t \;\middle|\; s_0 a_0 = sa, \pi \right)$$

Probabilistic Reachability of Failure States

- Let $\mathscr{F}_\perp \subseteq \mathscr{S}_\perp$ be set of all failure states

- Given policy π

  probabilistic reachability of $\mathscr{F}_\perp$ at state s and state-action sa

$$P(s \mid \pi) := \mathbb{P}\left(s_T \in \mathscr{F}_\perp \,\middle|\, s_0 = s, \pi\right)$$

$$\mathscr{P}(s, a \mid \pi) := \mathbb{P}\left(s_T \in \mathscr{F}_\perp \,\middle|\, s_0 a_0 = sa, \pi\right)$$

- Given threshold $\theta \in [0, 1)$

  partition the state space as $\mathscr{S}^+ = S(\pi) \cup F(\pi)$ where

$$S(\pi) := \{s \in \mathscr{S}^+ \mid P(s \mid \pi) \leq \theta\}$$ **(safe region)**

$$F(\pi) := \{s \in \mathscr{S}^+ \mid P(s \mid \pi) > \theta\}$$ **(unsafe region)**

Desired Properties of Constrained Optimality

- $\hat{\pi}$: assumed existent optimal policy satisfying **P1**—**P4**, associated with θ

$$\hat{S} := S(\hat{\pi}) \text{ and } \hat{F} := F(\hat{\pi})$$

P1 Uniform Optimality ➡ For any policy π

$$P(s \mid \pi) \leq P(s \mid \hat{\pi}) \implies V(s \mid \pi) \leq V(s \mid \hat{\pi}) \quad \forall s \in \hat{S}$$

$$V(s \mid \hat{\pi}) \leq V(s \mid \pi) \implies P(s \mid \hat{\pi}) \leq P(s \mid \pi) \quad \forall s \in \hat{F}$$

P2 Second Uniform Optimality over $\hat{F}$ ➡ For any policy π s.t. $\pi = \hat{\pi}$ over $\hat{S}$

$$P(s \mid \hat{\pi}) \leq P(s \mid \pi) \qquad \forall s \in \hat{F}$$

P3 Monotonicity ➡ If $\vartheta \leq \theta$, then

$$\begin{cases} V(s \mid \hat{\pi}_\vartheta) \leq V(s \mid \hat{\pi}) & \forall s \in \hat{S} \\ P(s \mid \hat{\pi}_\vartheta) \leq P(s \mid \hat{\pi}) & \forall s \in \mathcal{S}^+ \end{cases}$$

Desired Properties of Constrained Optimality

- Policy iteration operator $\mathcal{T}(\pi) := \pi'$ where

$$\pi'(s) \in \begin{cases} \arg\max_{a \in \mathcal{A}(s \mid \pi)} Q(s, a \mid \pi) & \text{if } \mathcal{A}(s \mid \pi) \neq \emptyset \\ \arg\min_{a \in \mathcal{A}(s)} \mathcal{P}(s, a \mid \pi) & \text{otherwise} \end{cases}$$

$$\mathcal{A}(s \mid \pi) := \{\, a \in \mathcal{A}(s) \mid \mathcal{P}(s, a \mid \pi) \leq \theta \,\}$$

P4 Fixed Point Property ➡ $\mathcal{T}(\hat{\pi}) = \hat{\pi}$

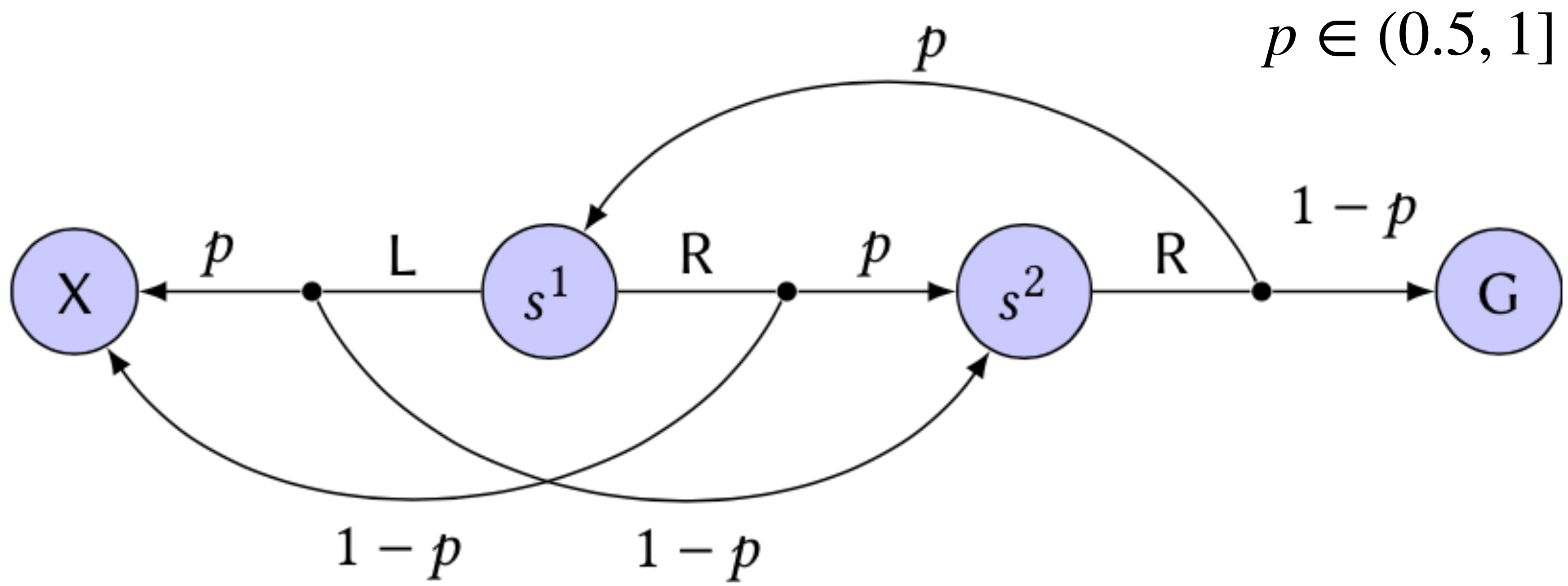
(i) reasonable

(ii) necessary for convergence

- However, we'll show

1. non-existence of such a fixed point of $\mathcal{T}$
2. mismatch between **P1** and **P4**

Counter-MDP

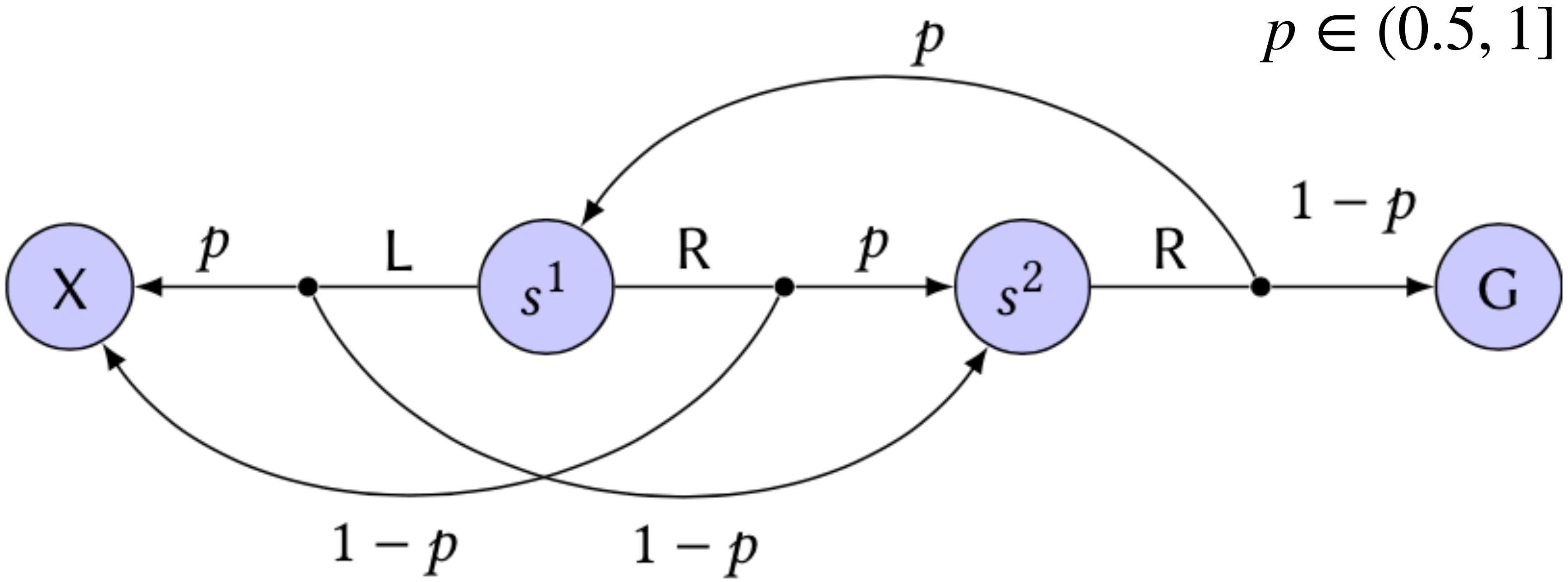


- State space $\mathcal{S}^{+} = \{\mathsf{X}, s^1, s^2, \mathsf{G}\}$ $\begin{cases} \mathcal{S} = \{s^1, s^2\} & \text{(non-terminal states)} \\ \mathcal{S}_{\perp} = \{\mathsf{X}, \mathsf{G}\} & \text{(terminal states)} \\ \mathcal{F}_{\perp} = \{\mathsf{X}\} & \text{(failure state)} \end{cases}$

- Action space $\mathcal{A} = \{\mathsf{L}, \mathsf{R}\}$ $\begin{cases} \mathcal{A}(s^1) = \{\mathsf{L}, \mathsf{R}\} \\ \mathcal{A}(s^2) = \{\mathsf{R}\} \end{cases}$ ⬅ L is not enabled at s^2 for simplicity.

- $p > 0.5$ determines transition probabilities $\mathcal{T}(s, a)(s')$

Counter-MDP



- Only two policies exist $\begin{cases} \pi_{\mathsf{L}} & - \quad \pi_{\mathsf{L}}(s^1) = \mathsf{L} \quad \pi_{\mathsf{L}}(s^2) = \mathsf{R} \\ \pi_{\mathsf{R}} & - \quad \pi_{\mathsf{R}}(s^1) = \mathsf{R} \quad \pi_{\mathsf{R}}(s^2) = \mathsf{R} \end{cases}$

- Reward model : $\mathscr{R}(s, a, s') = -\mathbf{1}(s \notin \mathcal{S}_{\perp})$ with $\gamma = 0.95$

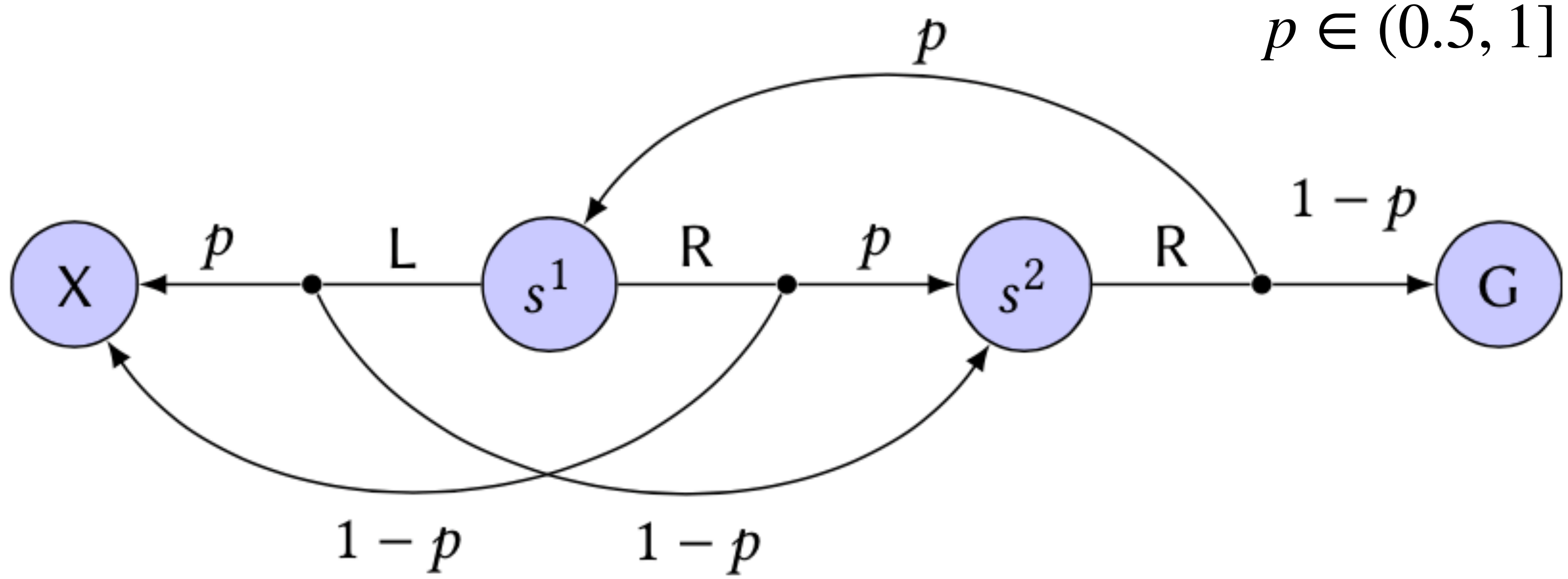
- We investigate π_{L} and π_{R} at state s^1…

$$Q_{a\mathsf{L}} := Q(s^1, a \,|\, \pi_{\mathsf{L}}) \qquad Q_{a\mathsf{R}} := Q(s^1, a \,|\, \pi_{\mathsf{R}})$$

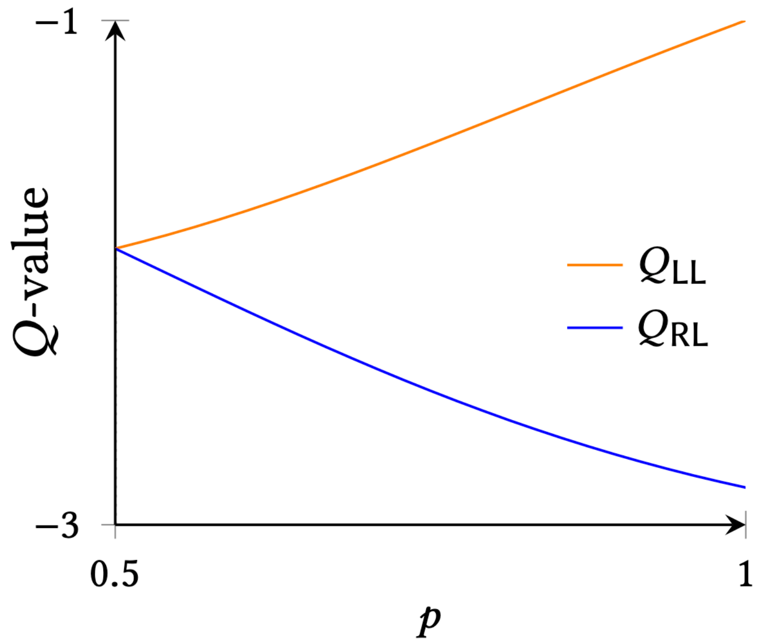
$$\mathscr{P}_{a\mathsf{L}} := \mathscr{P}(s^1, a \,|\, \pi_{\mathsf{L}}) \qquad \mathscr{P}_{a\mathsf{R}} := \mathscr{P}(s^1, a \,|\, \pi_{\mathsf{R}})$$

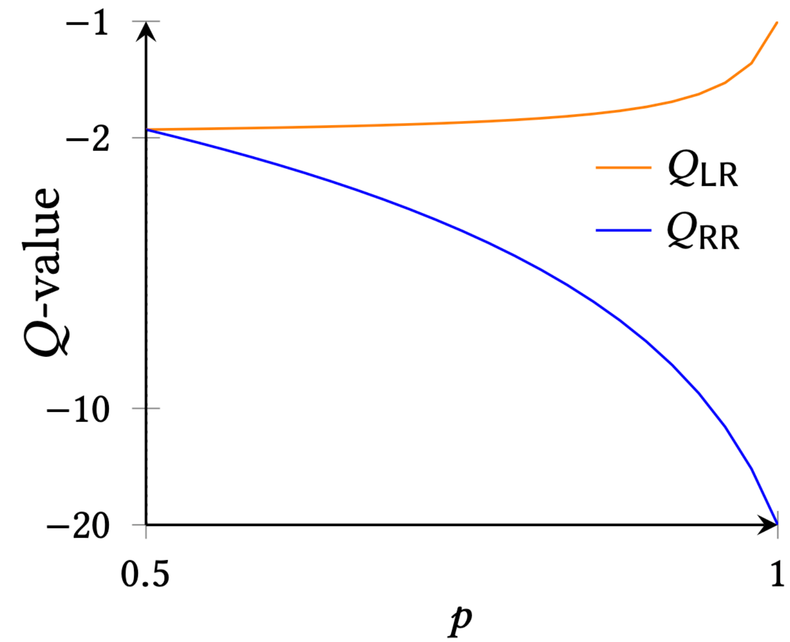
Counter-MDP: Performance

$p \in (0.5, 1]$



$Q_{a\mathrm{L}}$ vs p



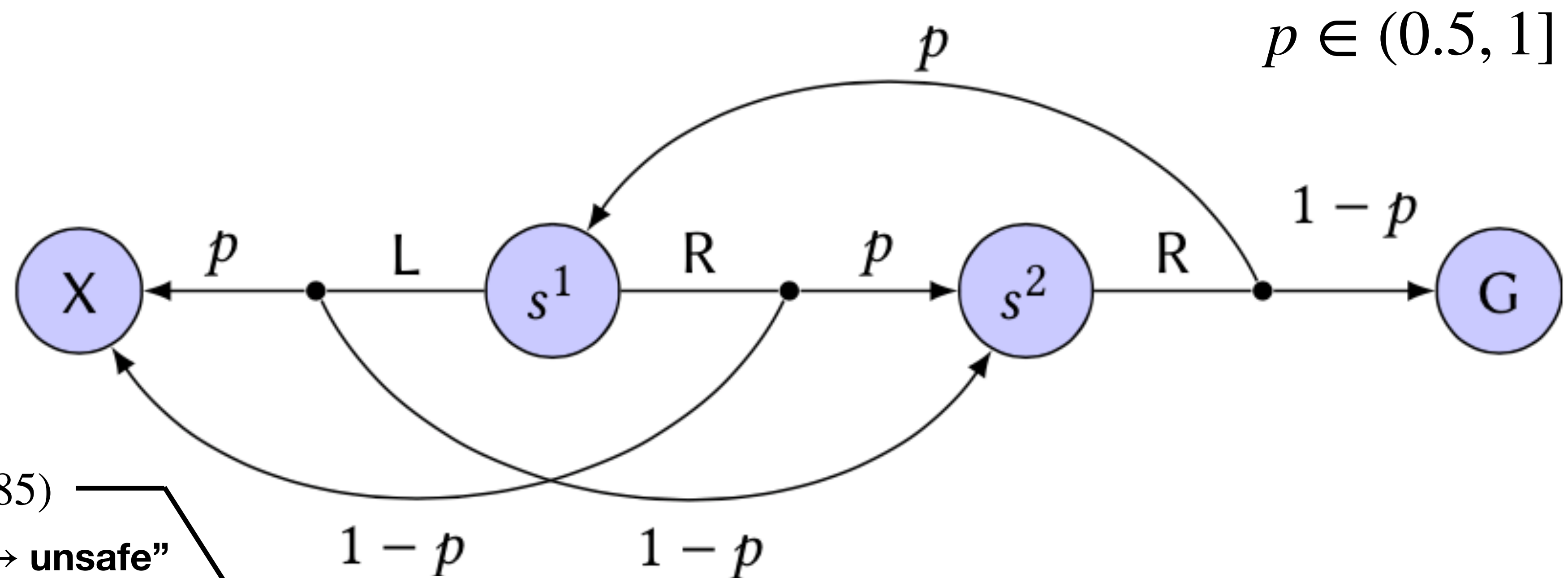


$Q_{a\mathrm{R}}$ vs p

$Q_{\mathrm{RL}} < Q_{\mathrm{LL}}$ $\quad$ $Q_{\mathrm{RR}} < Q_{\mathrm{LR}}$

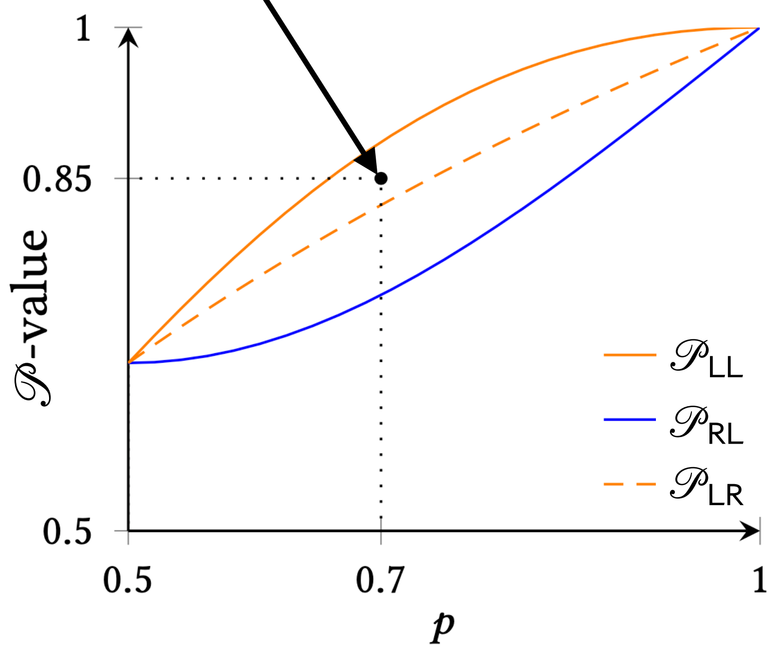
➡ Choosing L at s^1 clearly yields higher Q-values than R

Counter-MDP: Safety

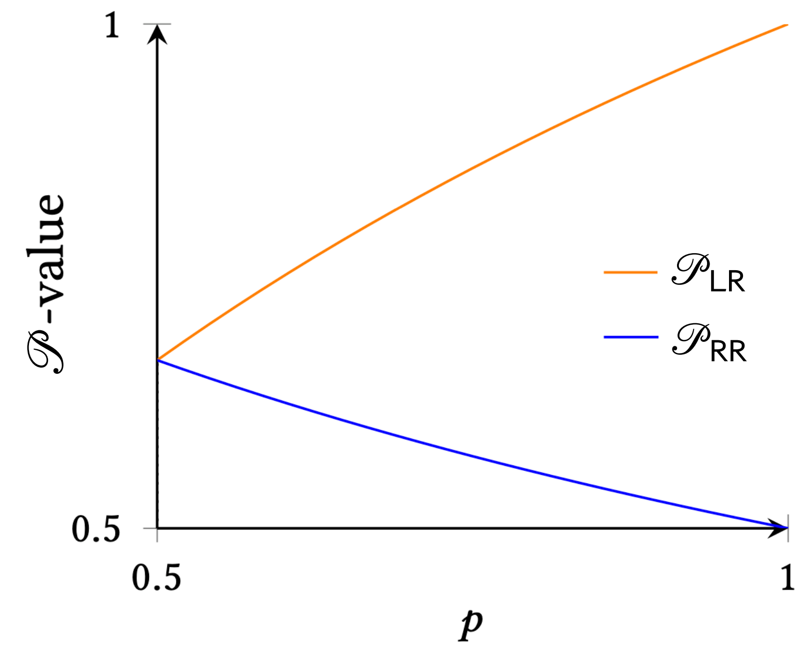


At $(p, \theta) = (0.7, 0.85)$
L alternates "safe ↔ unsafe"

$\mathscr{P}_{a\mathsf{L}}$ vs p



$\mathscr{P}_{\mathsf{RL}} < \mathscr{P}_{\mathsf{LL}}$



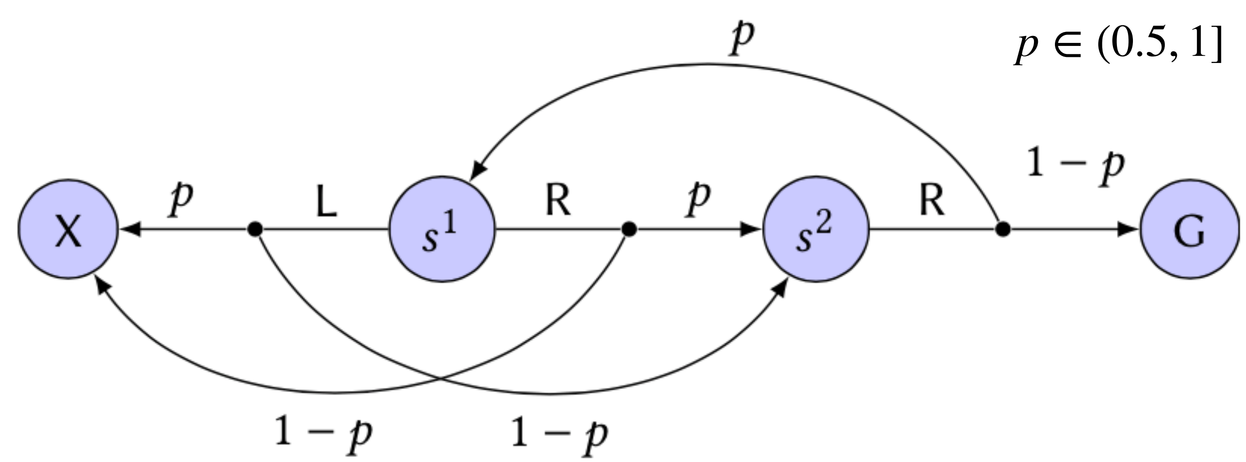
$\mathscr{P}_{a\mathsf{R}}$ vs p

$\mathscr{P}_{\mathsf{RR}} < \mathscr{P}_{\mathsf{LR}}$

➡ Choosing R at s^1 is always safer than L

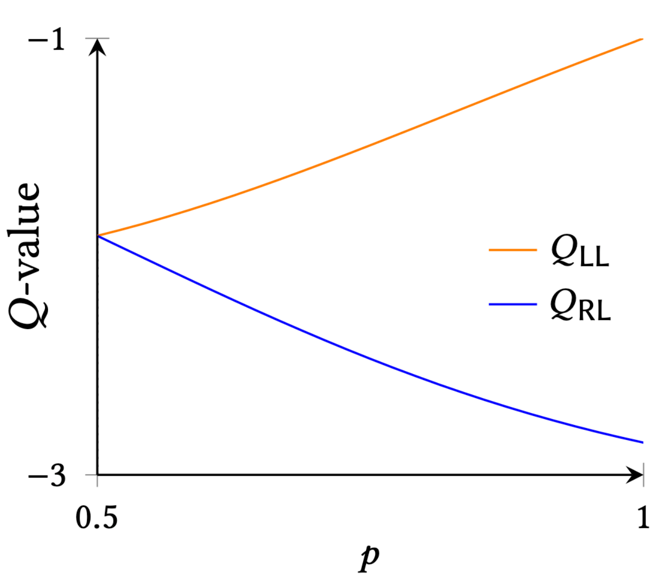
➡ When π_{L} is not safe, L at s^1 can appear safe if π_{R} is followed

Mixing Performance and Safety Causes Oscillations

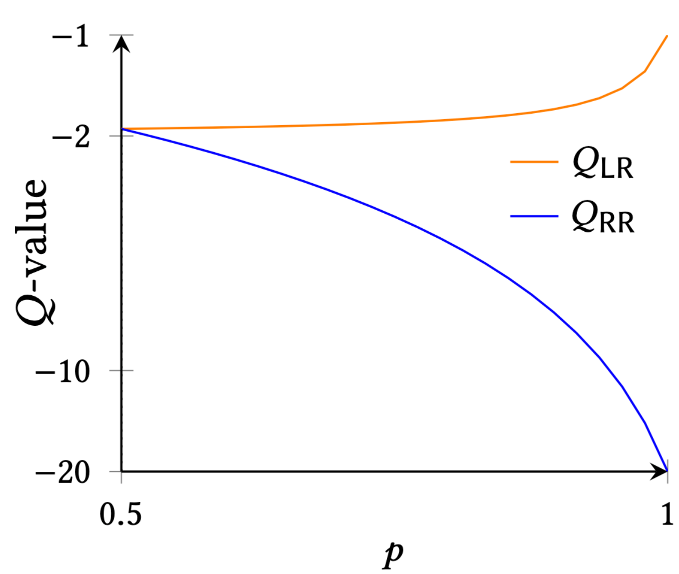


- At $(p, \theta) = (0.7, 0.85)$

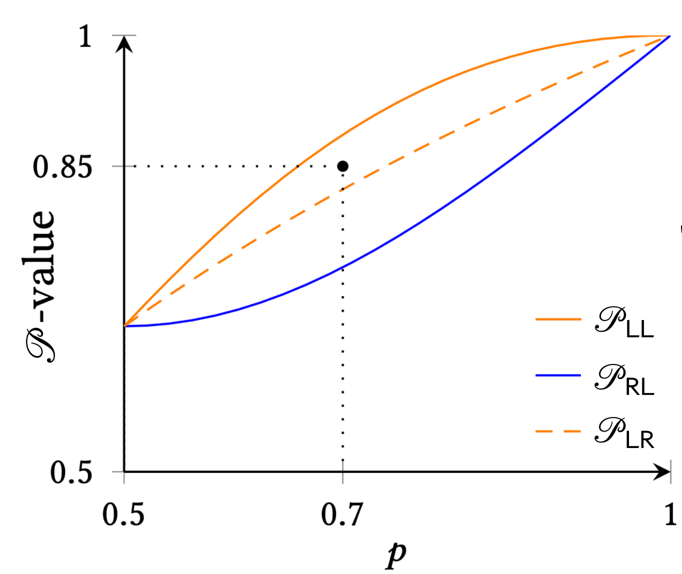
L at s^1 { always yields better performance than R
is always riskier than R
appears safe if π_R is followed while π_L is not safe }



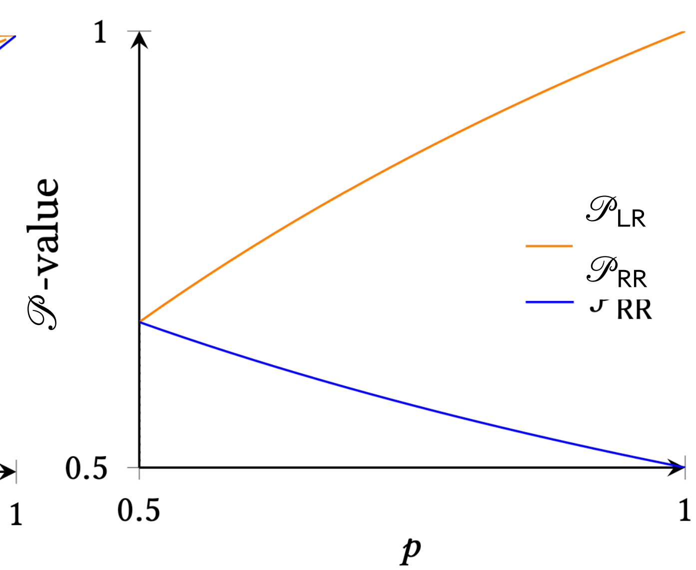
$Q_{a\mathrm{L}}$ vs p



$Q_{a\mathrm{R}}$ vs p

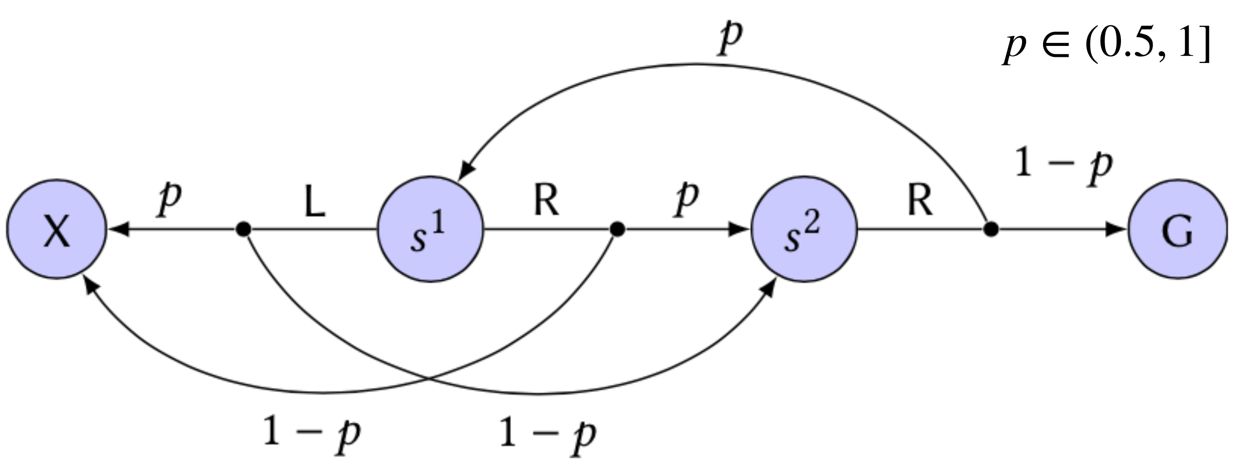


$\mathscr{P}_{a\mathrm{L}}$ vs p



$\mathscr{P}_{a\mathrm{R}}$ vs p

Mixing Performance and Safety Causes Oscillations



- At $(p, \theta) = (0.7,\, 0.85)$

L at s^1 { always yields better performance than R; is always riskier than R; appears safe if π_R is followed while π_L is not safe }

Policy iteration on counter-MDP for $(p, \theta) = (0.7,\, 0.85)$

Iteration i		1	2	3	4	5	⋯
Given policy		π_R	π_L	π_R	π_L	π_R	⋯
Constraints	L	$\mathscr{P}_{LR} \approx 0.82 \leq \theta = 0.85$	$\mathscr{P}_{LL} \approx 0.89 \nleq \theta$	$\mathscr{P}_{LR} \leq \theta$	$\mathscr{P}_{LL} \nleq \theta$	$\mathscr{P}_{LR} \leq \theta$	⋯
	R	$\mathscr{P}_{RR} \approx 0.59 \leq \theta = 0.85$	$\mathscr{P}_{RL} \approx 0.73 \leq \theta$	$\mathscr{P}_{RR} \leq \theta$	$\mathscr{P}_{RL} \leq \theta$	$\mathscr{P}_{RR} \leq \theta$	⋯

➡ Safe actions must be chosen conservatively

What was Wrong with P4 ?

- Suppose $s \in \hat{S}$ i.e. $\mathscr{P}(s, \hat{\pi}(s) \,|\, \hat{\pi}) = P(s \,|\, \hat{\pi}) \leq \theta$

$$\Longrightarrow \begin{cases} \mathscr{A}(s \,|\, \hat{\pi}) := \left\{ a \in \mathscr{A}(s) \,|\, \mathscr{P}(s, a \,|\, \hat{\pi}) \leq \theta \right\} \neq \emptyset \\ \hat{\mathscr{A}}(s) := \left\{ a \in \mathscr{A}(s) \,|\, \mathscr{P}(s, a \,|\, \hat{\pi}) \leq P(s \,|\, \hat{\pi}) \right\} \neq \emptyset \end{cases}$$

- $\hat{\mathscr{A}}(s) \subseteq \mathscr{A}(s \,|\, \hat{\pi})$ ➡ $\hat{\mathscr{A}}(s)$ is more conservative than $\mathscr{A}(s \,|\, \hat{\pi})$

- **P4 Fixed Point Property** $\mathscr{T}(\hat{\pi}) = \hat{\pi}$

$$\Longrightarrow \hat{\pi}(s) \in \underset{a \in \mathscr{A}(s \,|\, \hat{\pi})}{\arg\max}\; Q(s, a \,|\, \hat{\pi}) = \underset{a \in \hat{\mathscr{A}}(s)}{\arg\max}\; Q(s, a \,|\, \hat{\pi})$$

- **P1 Uniform Optimality** $\forall \pi : P(s \,|\, \pi) \leq P(s \,|\, \hat{\pi}) \Longrightarrow V(s \,|\, \pi) \leq V(s \,|\, \hat{\pi})$

$$\Longrightarrow \hat{\pi}(s) \in \underset{a \in \hat{\mathscr{A}}(s)}{\arg\max}\; Q(s, a \,|\, \hat{\pi}) \neq \underset{a \in \mathscr{A}(s \,|\, \hat{\pi})}{\arg\max}\; Q(s, a \,|\, \hat{\pi})$$

What was Wrong with P4 ?

- $\hat{\mathscr{A}}(s) \subseteq \mathscr{A}(s \mid \hat{\pi})$ ➡ $\hat{\mathscr{A}}(s)$ is more conservative than $\mathscr{A}(s \mid \hat{\pi})$
- **P4 Fixed Point Property** $\Longrightarrow \hat{\pi}(s) \in \underset{a \in \mathscr{A}(s \mid \hat{\pi})}{\arg\max}\ Q(s, a \mid \hat{\pi})$

$$\Downarrow \qquad \not\Uparrow$$

- **P1 Uniform Optimality** $\Longrightarrow \hat{\pi}(s) \in \underset{a \in \hat{\mathscr{A}}(s)}{\arg\max}\ Q(s, a \mid \hat{\pi})$

$\mathscr{A}(s \mid \hat{\pi})$ in **P4** has to be more conservative e.g. $\hat{\mathscr{A}}(s)$

➡ True for the counter MDP!

Counter-MDP with Recursive Constraints

Policy iteration on counter-MDP

Iteration i		1	2	3	4	5	$\cdots$
Given policy		π_{R}	π_{L}	π_{R}	π_{L}	π_{R}	$\cdots$
Constraints	L	$\mathscr{P}_{\mathsf{LR}} \approx 0.82 \leq \theta = 0.85$	$\mathscr{P}_{\mathsf{LL}} \approx 0.89 \nleq \theta$	$\mathscr{P}_{\mathsf{LR}} \leq \theta$	$\mathscr{P}_{\mathsf{LL}} \nleq \theta$	$\mathscr{P}_{\mathsf{LR}} \leq \theta$	$\cdots$
	R	$\mathscr{P}_{\mathsf{RR}} \approx 0.59 \leq \theta = 0.85$	$\mathscr{P}_{\mathsf{RL}} \approx 0.73 \leq \theta$	$\mathscr{P}_{\mathsf{RR}} \leq \theta$	$\mathscr{P}_{\mathsf{RL}} \leq \theta$	$\mathscr{P}_{\mathsf{RR}} \leq \theta$	$\cdots$

- Recursive constraints $\mathsf{C}_a(i) \quad (a \in \{\mathsf{L}, \mathsf{R}\})$

$$\mathsf{C}_{\mathsf{L}}(1) = (\mathscr{P}_{\mathsf{LR}} \leq \theta)$$

$$\mathsf{C}_{\mathsf{L}}(2) = (\mathscr{P}_{\mathsf{LL}} \nleq \theta) \wedge \mathsf{C}_{\mathsf{L}}(1) = (\mathscr{P}_{\mathsf{LL}} \nleq \theta) \wedge (\mathscr{P}_{\mathsf{LR}} \leq \theta)$$

$$\mathsf{C}_{\mathsf{L}}(3) = (\mathscr{P}_{\mathsf{LR}} \leq \theta) \wedge \mathsf{C}_{\mathsf{L}}(2) = (\mathscr{P}_{\mathsf{LR}} \leq \theta) \wedge (\mathscr{P}_{\mathsf{LL}} \nleq \theta)$$

$$\mathsf{C}_{\mathsf{L}}(4) = (\mathscr{P}_{\mathsf{LR}} \leq \theta) \wedge \mathsf{C}_{\mathsf{L}}(3) = (\mathscr{P}_{\mathsf{LR}} \leq \theta) \wedge (\mathscr{P}_{\mathsf{LL}} \nleq \theta)$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

Counter-MDP with Recursive Constraints

Policy iteration on counter-MDP

Iteration i		1	2	3	4	5	⋯
Given policy		π_R	π_L	π_R	π_L	π_R	⋯
Constraints	L	$\mathscr{P}_{LR} \approx 0.82 \leq \theta = 0.85$	$\mathscr{P}_{LL} \approx 0.89 \nleq \theta$	$\mathscr{P}_{LR} \leq \theta$	$\mathscr{P}_{LL} \nleq \theta$	$\mathscr{P}_{LR} \leq \theta$	⋯
	R	$\mathscr{P}_{RR} \approx 0.59 \leq \theta = 0.85$	$\mathscr{P}_{RL} \approx 0.73 \leq \theta$	$\mathscr{P}_{RR} \leq \theta$	$\mathscr{P}_{RL} \leq \theta$	$\mathscr{P}_{RR} \leq \theta$	⋯

Policy iteration on counter-MDP, with recursive constraints

Iteration i		1	2	3	4	⋯
Given policy		π_R	π_L	π_R	π_R	⋯
Constraints	L	$C_L \leftarrow (\mathscr{P}_{LR} \leq \theta)$	$C_L \leftarrow (\mathscr{P}_{LL} \nleq \theta) \wedge C_L$	$C_L \leftarrow (\mathscr{P}_{LR} \leq \theta) \wedge C_L$	$C_L \leftarrow (\mathscr{P}_{LR} \leq \theta) \wedge C_L$	⋯
	R	$C_R \leftarrow (\mathscr{P}_{RR} \leq \theta)$	$C_R \leftarrow (\mathscr{P}_{RL} \leq \theta) \wedge C_R$	$C_R \leftarrow (\mathscr{P}_{RR} \leq \theta) \wedge C_R$	$C_R \leftarrow (\mathscr{P}_{RR} \leq \theta) \wedge C_R$	⋯

➡ **Stabilized with recursive constraints!**

Proposed Idea

Policy iteration on counter-MDP, with recursive constraints

Iteration i		1	2	3	4	$\cdots$
Given policy		π_R	π_L	π_R	π_R	$\cdots$
Constraints	L	$\mathrm{C_L} \leftarrow (\mathscr{P}_\mathrm{LR} \leq \theta)$	$\mathrm{C_L} \leftarrow (\mathscr{P}_\mathrm{LL} \nleq \theta) \wedge \mathrm{C_L}$	$\mathrm{C_L} \leftarrow (\mathscr{P}_\mathrm{LR} \leq \theta) \wedge \mathrm{C_L}$	$\mathrm{C_L} \leftarrow (\mathscr{P}_\mathrm{LR} \leq \theta) \wedge \mathrm{C_L}$	$\cdots$
	R	$\mathrm{C_R} \leftarrow (\mathscr{P}_\mathrm{RR} \leq \theta)$	$\mathrm{C_R} \leftarrow (\mathscr{P}_\mathrm{RL} \leq \theta) \wedge \mathrm{C_R}$	$\mathrm{C_R} \leftarrow (\mathscr{P}_\mathrm{RR} \leq \theta) \wedge \mathrm{C_R}$	$\mathrm{C_R} \leftarrow (\mathscr{P}_\mathrm{RR} \leq \theta) \wedge \mathrm{C_R}$	$\cdots$

- Let's extend the idea but except for policy iteration

  initial/early $\mathscr{P}_\mathrm{LR}$ and $\mathscr{P}_\mathrm{RR}$ are typically random and has no information

  those inaccurate constraints will be transferred to all later iterations

- Solution
  1. axis of iteration $i = 1, 2, 3, \ldots$ ➡ axis of horizon $n = 1, 2, \ldots, N$
  2. constraints at stage n : $\mathrm{C}_a(n \mid s) \leftarrow (\mathscr{P}^n(s,a) \leq \theta) \wedge \mathrm{C}_a(n-1 \mid s)$

- $\mathscr{P}^n(s,a)$ is/over-approximates n-bounded probabilistic reachability

$$\mathbb{P}(s_{\min(T,n)} \in \mathscr{F}_\perp \mid s_0 a_0 = sa, \pi) \neq \mathscr{P}(s,a)$$

Proposed Idea: Implementation

➡ Proposed idea can be implemented on top of a naive algorithm

Naive Value Iteration

$\forall (s,a) \in \mathcal{S}^+ \times \mathcal{A}(s)$ **do** /* initialization */

$Q(s,a) \leftarrow \mathbb{E}[r_0 \mid s_0 a_0 = sa]$

$\mathcal{P}(s,a) \leftarrow \mathbb{P}[s_{\min(1,T)} \in \mathcal{F}_\perp \mid s_0 a_0 = sa]$

repeat k **times** /* k number of iters */

$\hat{\mathcal{A}}(s) \leftarrow \{a \in \mathcal{A}(s) \mid \mathcal{P}(s,a) \leq \theta\} \quad \forall s \in \mathcal{S}^+$

$\pi \leftarrow$ GetPolicy($\hat{\mathcal{A}}, Q, \mathcal{P}$)

$(Q, \mathcal{P}) \leftarrow$ Update($\pi, Q, \mathcal{P}$)

return $(Q, \mathcal{P})$ /* $Q \approx Q(\cdot \mid \pi)$ $\mathcal{P} \approx \mathcal{P}(\cdot \mid \pi)$ */

Subroutine GetPolicy($\hat{\mathcal{A}}, Q, \mathcal{P}$)

$\forall s \in \mathcal{S}^+$ **do**

$$\pi(s) \leftarrow a \in \begin{cases} \arg\max_{a \in \hat{\mathcal{A}}(s)} Q(s,a) & \text{if } \hat{\mathcal{A}}(s) \neq \emptyset \\ \arg\min_{a \in \mathcal{A}(s)} \mathcal{P}(s,a) & \text{otherwise} \end{cases}$$

return π

Subroutine Update($\pi, Q, \mathcal{P}$)

$Q' \leftarrow Q, \quad \mathcal{P}' \leftarrow \mathcal{P}$

$\forall (s,a) \in \mathcal{S} \times \mathcal{A}(s)$ **do**

$Q'(s,a) \leftarrow \mathbb{E}[r_0 + \gamma Q(s_1, \pi(s_1)) \mid s_0 a_0 = sa]$

$\mathcal{P}'(s,a) \leftarrow \mathbb{E}[\mathcal{P}(s_1, \pi(s_1)) \mid s_0 a_0 = sa]$

return $(Q', \mathcal{P}')$

Proposed Idea: Implementation

➡ Proposed idea can be implemented on top of a naive algorithm

Value Iteration with Recursive Constraints

$\forall (s,a) \in \mathcal{S}^+ \times \mathcal{A}(s)$ **do** /* initialization */

  $Q^{1:N}(s,a) \leftarrow \mathbb{E}[r_0 \mid s_0 a_0 = sa]$

  $\mathcal{P}^{1:N+1}(s,a) \leftarrow \mathbb{P}[s_{\min(1,T)} \in \mathcal{F}_\perp \mid s_0 a_0 = sa]$ ⟶ $\mathcal{P}^1$ is already accurate.

repeat k **times** /* k number of iters */

  $\hat{\mathcal{A}} \leftarrow \mathcal{A}$

  for $n = 1, 2, \ldots, N$ **do** /* n : horizon */

    $\hat{\mathcal{A}}(s) \leftarrow \{a \in \hat{\mathcal{A}}(s) \mid \mathcal{P}^n(s,a) \leq \theta\} \quad \forall s \in \mathcal{S}^+$ ⟶ Constraints are recursively given $\mathcal{P}^n(s,a) \approx \mathbb{P}(s_{\min(T,n)} \in \mathcal{F}_\perp \mid s_0 a_0 = sa, \pi)$

    $\pi \leftarrow$ GetPolicy$(\hat{\mathcal{A}}, Q^n, \mathcal{P}^n)$

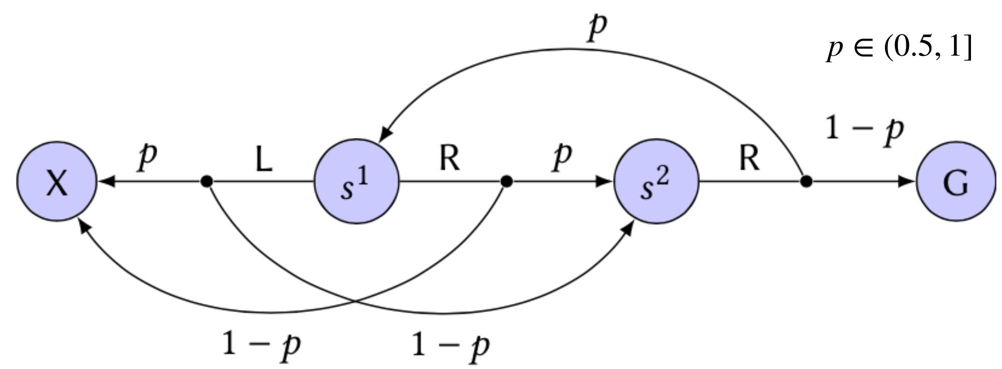
    $(Q^n, \mathcal{P}^{n+1}) \leftarrow$ Update$(\pi, Q^n, \mathcal{P}^n)$ ⟶ $\mathcal{P}^{n+1}$ is updated from $\mathcal{P}^n$ (stable target)

return $(Q^N, \mathcal{P}^{N+1})$ /* $Q^N \approx Q(\cdot \mid \pi)$ $\mathcal{P}^{N+1} \gtrapprox \mathbb{P}(s_{\min(T,N+1)} \in \mathcal{F}_\perp \mid s_0 a_0 = sa, \pi)$ */

Naive v.s. Proposed

- Experiments with CliffWorld

Same states as in counter-MDP

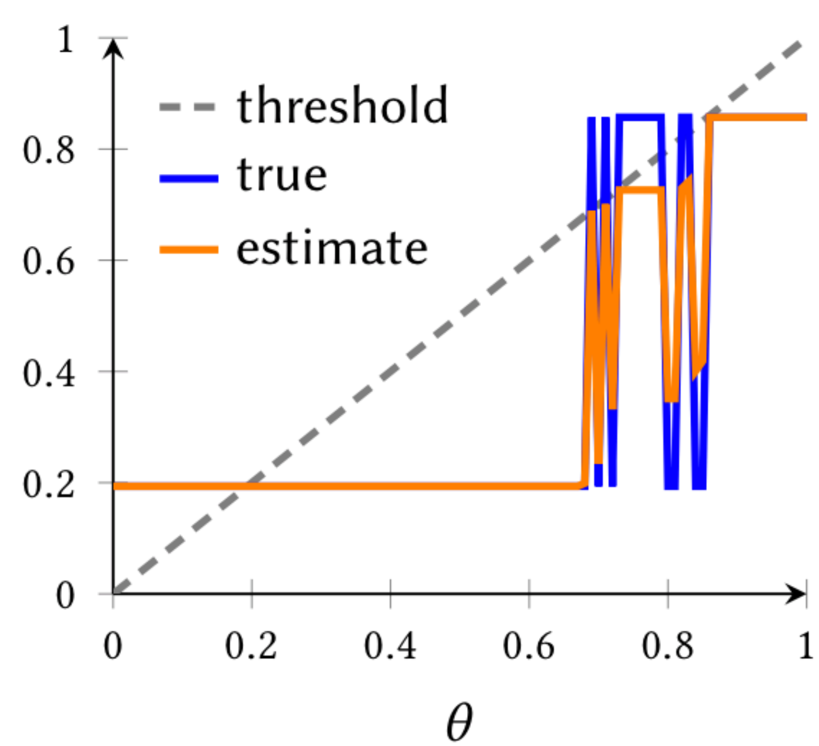


$$\mathscr{A}^+ = \mathscr{A}(s_1) = \mathscr{A}(s_2) = \{\mathsf{L}, \mathsf{R}, \mathsf{U}, \mathsf{D}\}$$

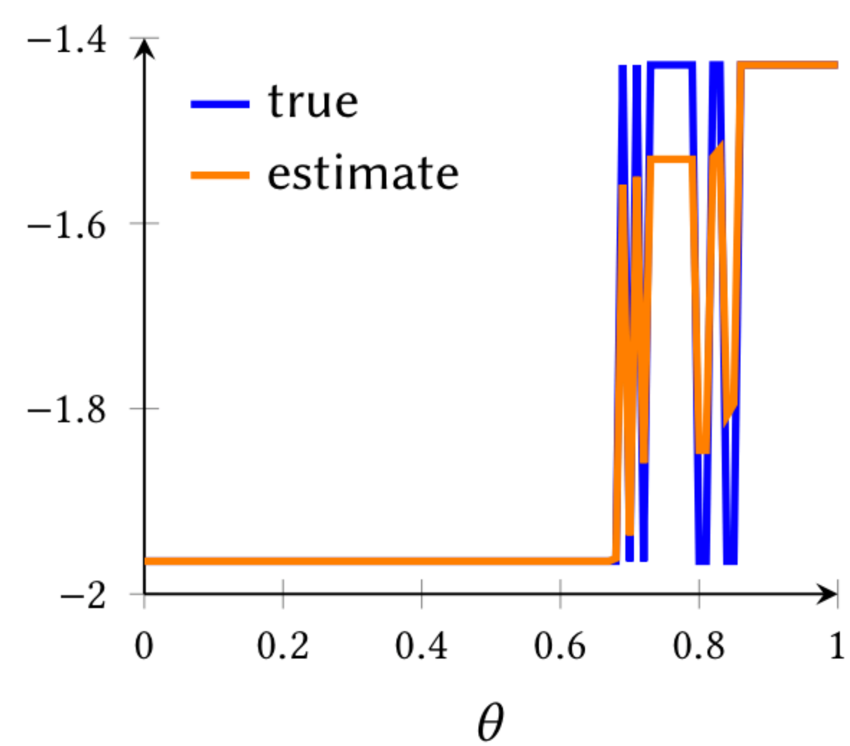
Transitions to $\begin{cases} \text{desired direction (50\%)} \\ \text{random direction (50\%)} \end{cases}$

- Naive value iteration

$k = 50$ iterations



$\mathscr{P}(s, a \mid \hat{\pi})$ v.s. θ

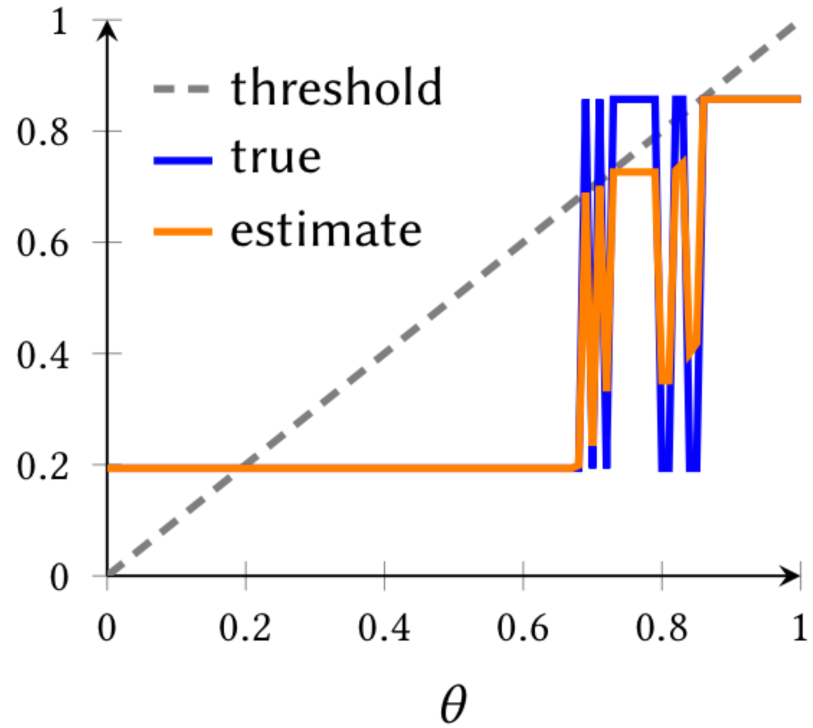


$Q(s, a \mid \hat{\pi})$ v.s. θ

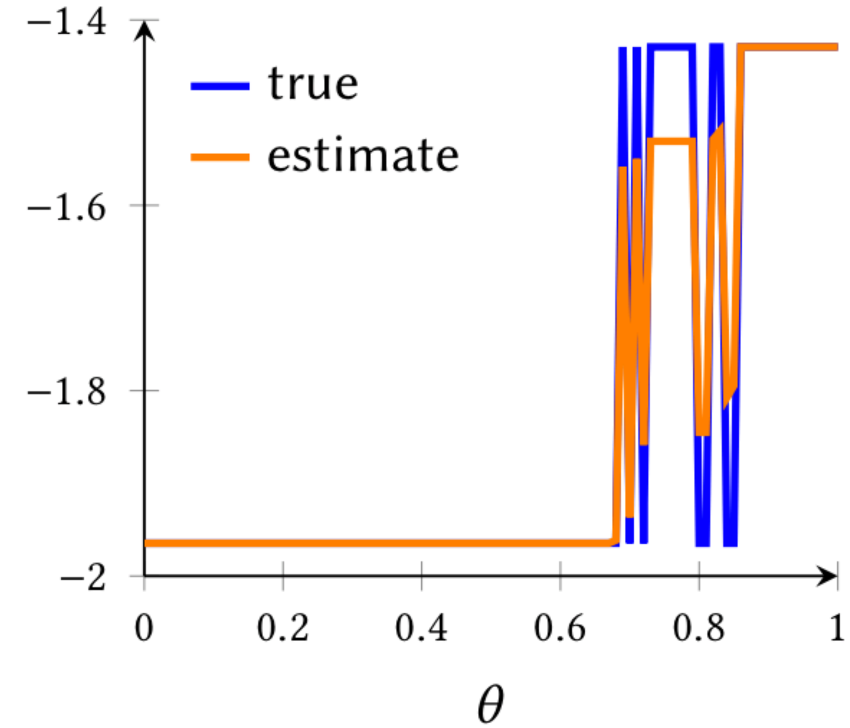
Naive v.s. Proposed

- Naive value iteration

  $k = 50$ iterations



$\mathscr{P}(s, a \,|\, \hat{\pi})$ v.s. θ

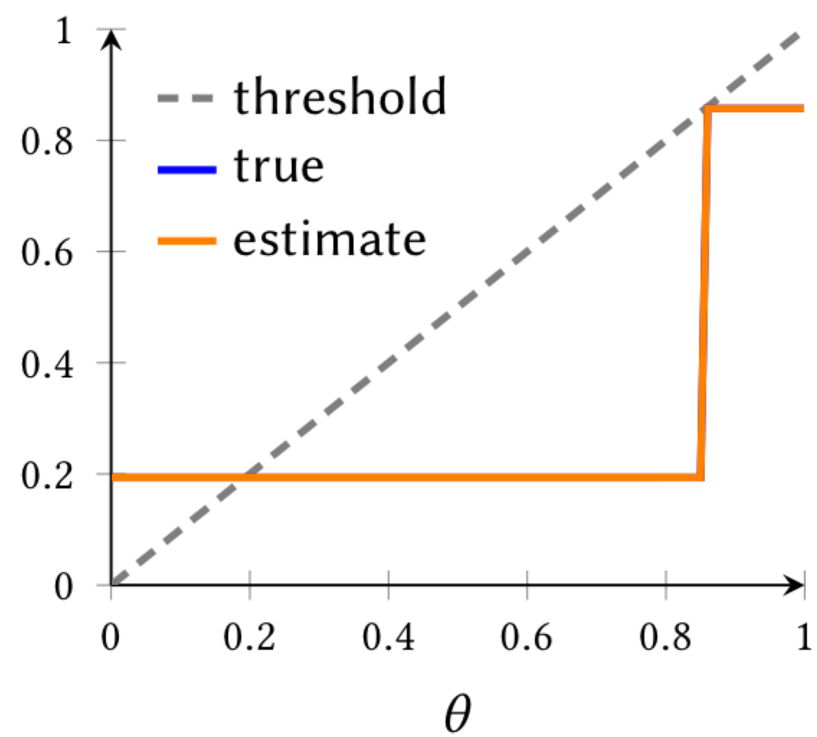


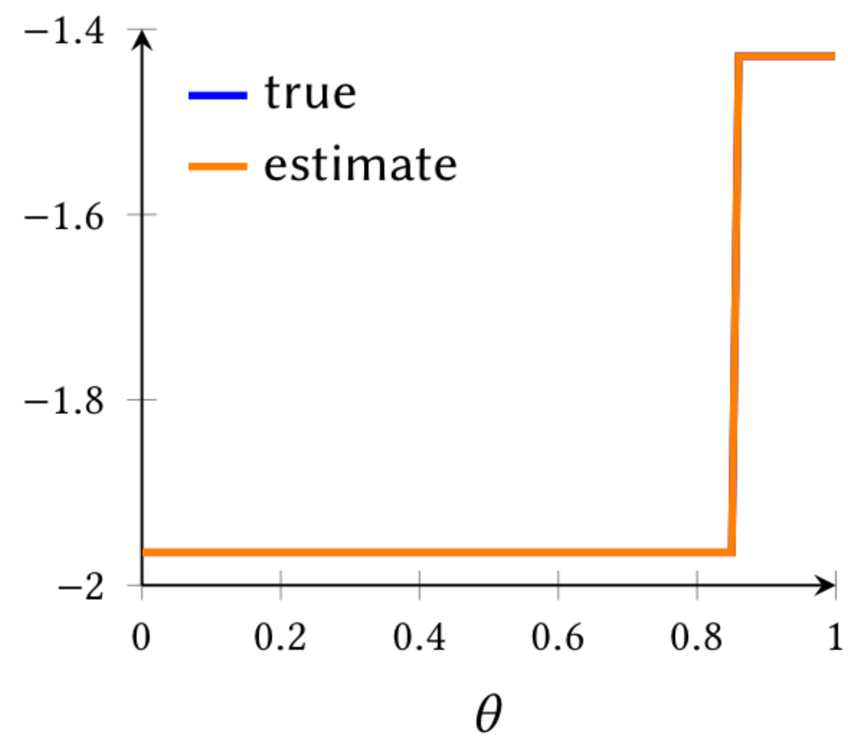
$Q(s, a \,|\, \hat{\pi})$ v.s. θ

- Value iteration with recursive constraints

  $k = 15$ iterations

  $N = 15$ horizon





➡ Instability / violation around $0.7 \le \theta \le 0.9$ has gone, with “true = estimate”

Summary

- **Problem**: difficulty / instability in finding policy $\hat{\pi}$ that is

  deterministic

  *uniformly optimal under safety constraints, in the sense of **P1—P3***

- **Conclusion**: *recursive constraints can solve instability found in naive approaches*

  Policy iteration (counter-MDP)

  Value iteration (CliffWorld experiments)

- **Future work**: extensions to

  reinforcement learning (e.g. Q-learning) with function approximation