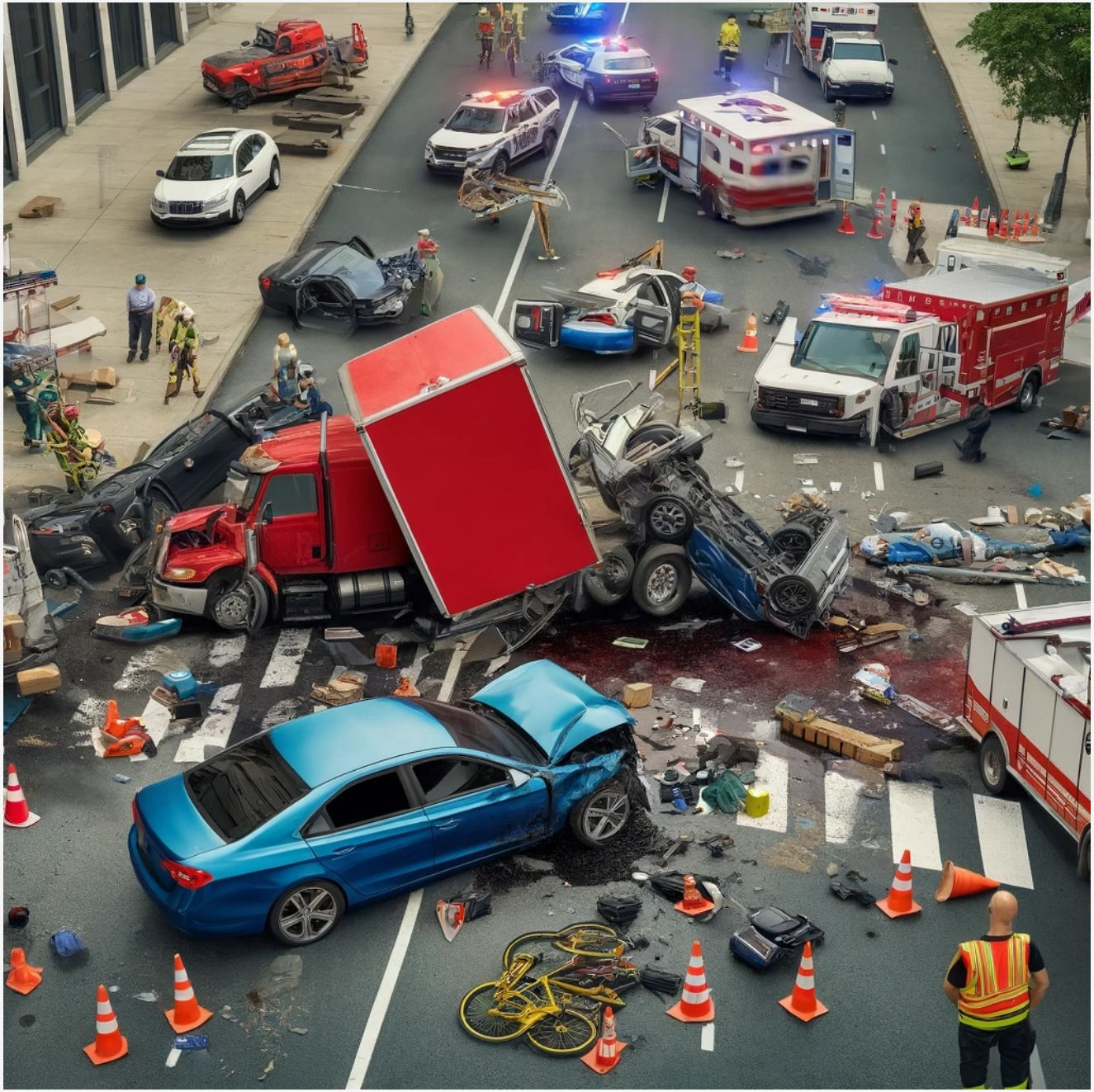
Safe Dynamic Programming and Learning

Rafal Wisniewski*, Manuela Bujorianu**

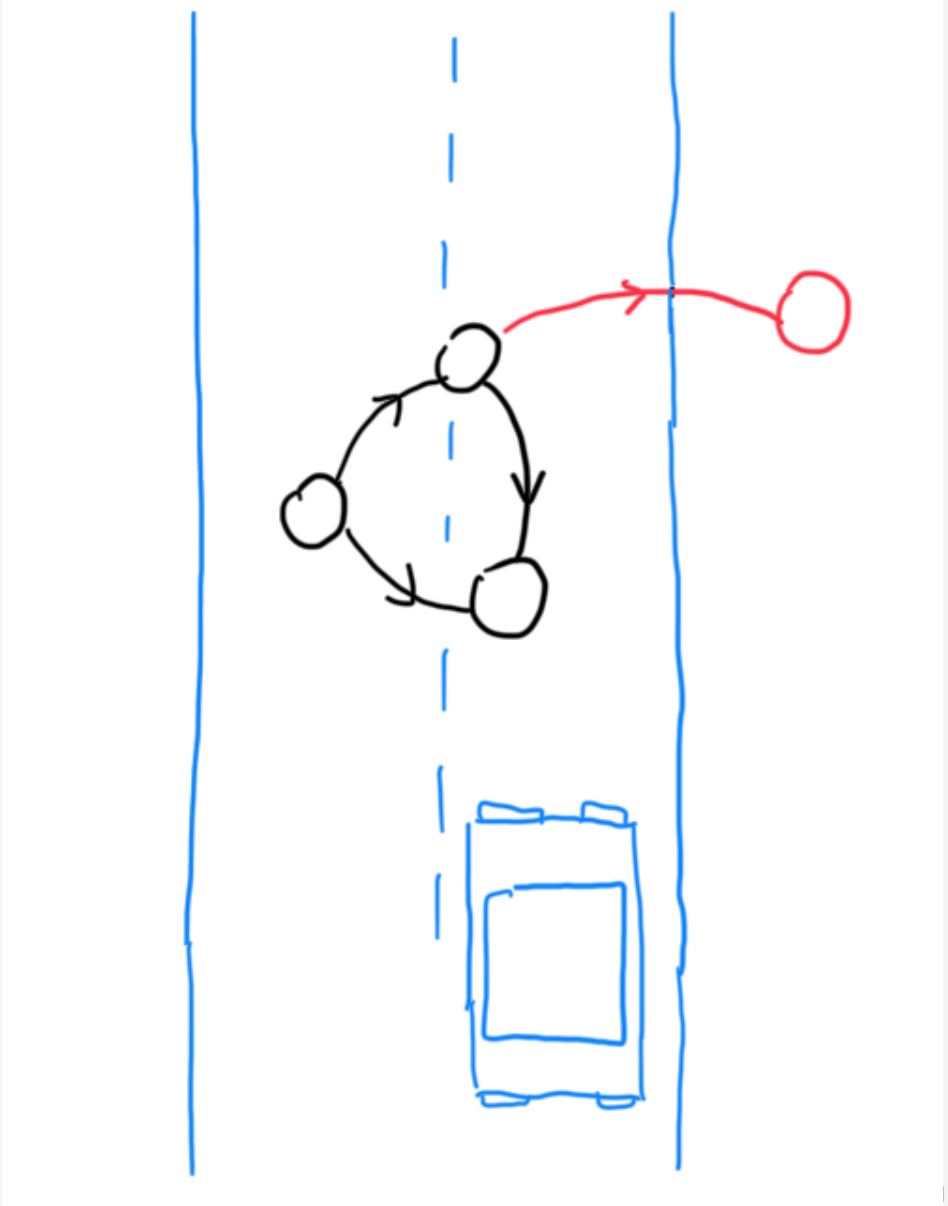
*Automation and Control, Aalborg University

**University College London

July 9, 2024



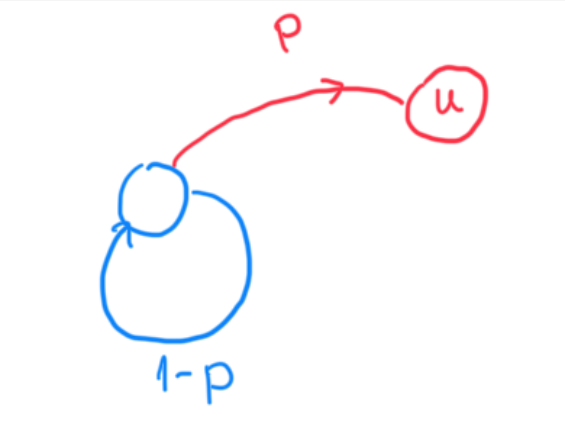
Setting the scene - how to define safety



Setting the scene - a possible definition

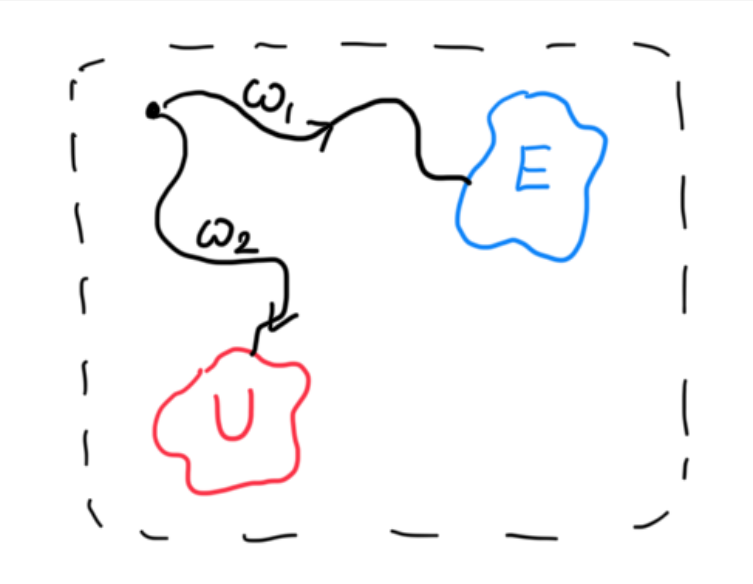
Let p be acceptable level of safety

$$\mathbb{P}[X_t \in U] < p \text{ for } t \in \{0, 1, \ldots, N\}$$



$$\mathbb{P}[X_t \in U \text{ at some } t] = p + (1-p)p + (1-p)^2 p + \ldots = \frac{p}{1-(1-p)} = 1$$

Setting the scene - a definition



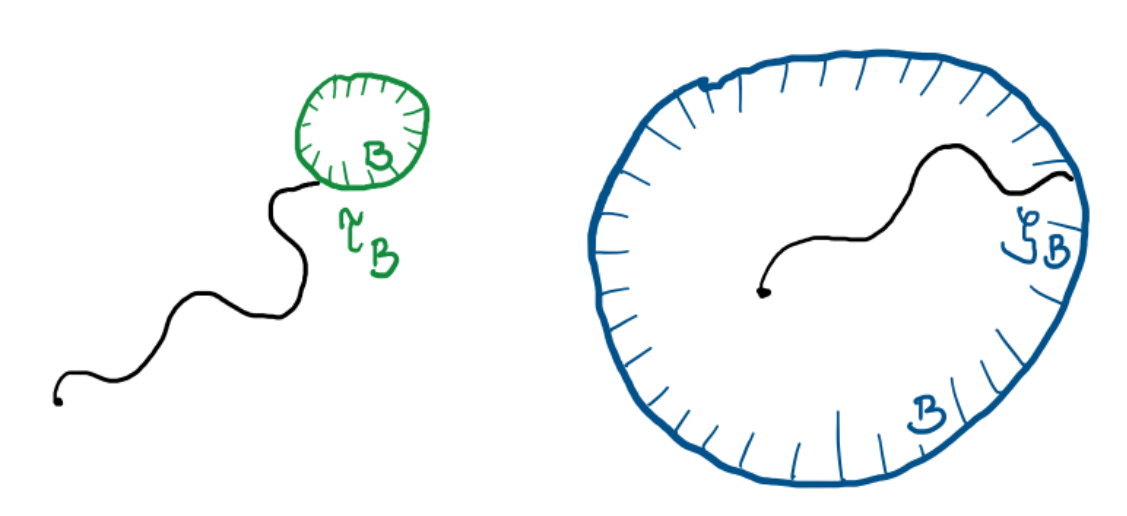
$$\mathbb{P}[X_t \in U \text{ at some } t] < p$$

Setting the scene - a definition

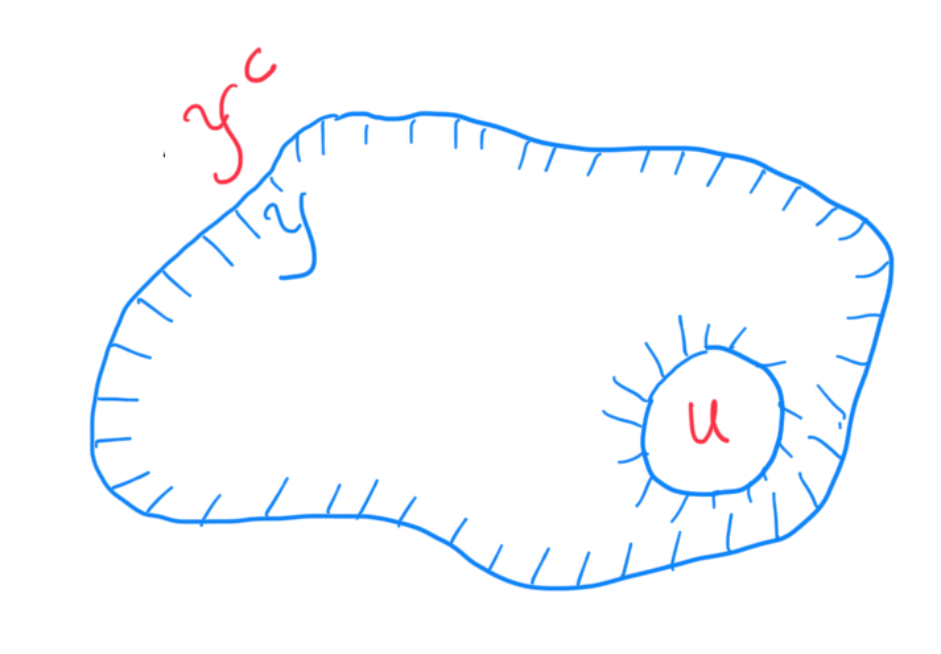
- For a measurable set B, the first hitting time τ_B associated to this set, is

$$\tau_B := \inf\{t \geq 0 | X_t \in B\}$$

- The first exit time from B is $\zeta_B = \tau_{B^c}$ (i.e., the first hitting time of the complement of B).



p-safety



Definition

A point $y \in \mathcal{Y}$ is p-safe if

$$\mathbb{P}^y[\tau_U < \zeta_{\mathcal{Y}}] \leq p$$

Strong p-safety

Definition

Let

$$q(y) \equiv \mathbb{P}^y[\tau_U < \zeta_{\mathcal{Y}}]$$

be a **safety function**.

A Borel subset $A \subset \mathcal{Y}$ is p-**safe** if all points $y \in A$ are p-safe,:
strong safety

$$q(A; U, \mathcal{Y}) \equiv \sup\{q(y) |\ y \in A\} \leq p$$

Weak p-safety

Definition (weak safety)

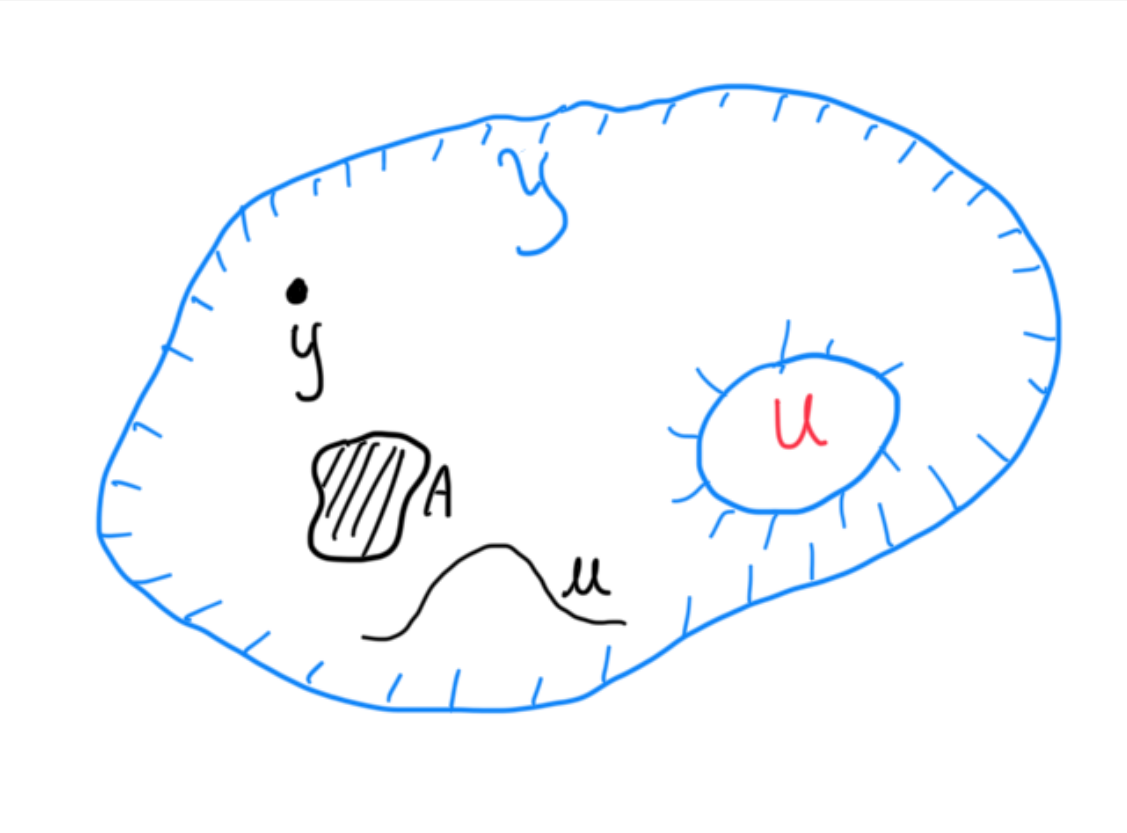
For an initial measure μ, we define

$$\mu q \equiv \int_{\mathcal{Y}} q(y)\mu(dy)$$

We say that the initial **measure μ is p-safe** if

$$\mu q \leq p$$

p-safety, strong and weak



Safety $\Longrightarrow$ Characterisation [1]

[1] *Bujorianu, L. M., R. Wisniewski, E. Boulougouris, Stochastic Safety for Markov Chains, IEEE Control Systems Letters, 5(2), 427-432, 2021*

Markov Chains - the notation

- $(X_t) := (X_t)_{t \in \mathbb{N}}$ is a discrete-time (homogeneous) Markov Chain with **transition probabilities**

$$p_{ij} := \mathbb{P}[X_{t+1} = j | X_t = i] = \mathbb{P}[X_1 = j | X_0 = i]$$

- The **transition probability matrix** P of (X_t) is

$$P := (p_{ij})$$

- The **generator** of Markov chain is

$$\mathcal{L} = P - I$$

(discrete Laplacian $\Delta = -\mathcal{L}$).

Computation of safety function

The safety function

$$q(j) := \mathbb{P}^j[\tau_U < \zeta_{\mathcal{Y}}]$$

is the solution of **Dirichlet problem**

$$\begin{aligned} (\mathcal{L}q)(j) &= 0, \quad \forall j \in \mathcal{Y} \setminus U \\ q(j) &= 1, \quad \forall j \in U \\ q(j) &= 0, \quad \forall j \in \delta\mathcal{Y} \end{aligned}$$

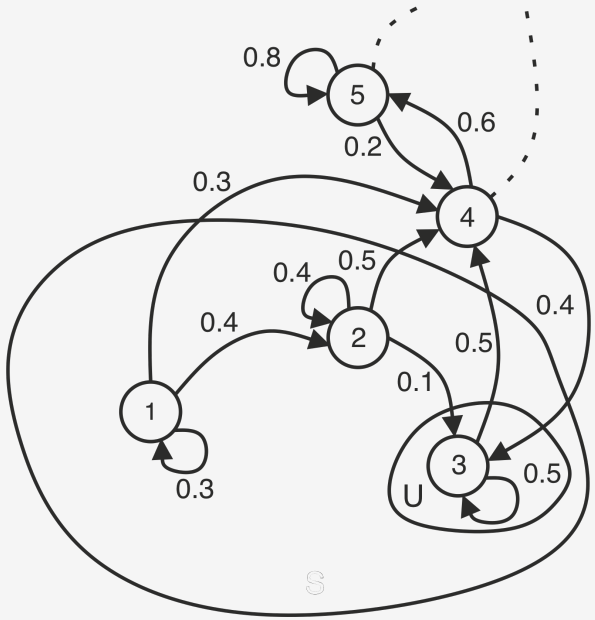


Figure: State-space is $\mathcal{Y} = \{1, 2, 3\}$, and the unsafe set is a singleton $U = \{3\}$.

$$\mathcal{L} = \begin{pmatrix} -0.7 & 0.4 & 0 & 0.3 & 0 & 0 & \dots \\ 0 & -0.6 & 0.1 & 0.5 & 0 & 0 & \dots \\ 0 & 0 & -0.5 & 0.5 & 0 & \dots & \dots \\ 0 & 0 & 0.4 & -1 & 0.6 & \dots & \dots \\ 0 & 0 & 0 & 0.2 & -0.2 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}.$$

Dirichlet problem gives

$$0.7q(1) - 0.4q(2) - 0.3q(4) = 0$$
$$0.6q(2) - 0.1q(3) - 0.5q(4) = 0$$

The boundary conditions give

$$q(3) = 1, q(4) = 0$$

Consequently, the remaining values of the vector q are

$$q(1) = 2/21 \text{ and } q(2) = 1/6$$

Safety $\Longrightarrow$ Optimisation [2]

[2] *Wisniewski R., L.M. Bujorianu, Safety of stochastic systems: An analytic and computational approach, Automatica, Vol. 133, 2021*

Towards stochastic barriers - Excessive functions

A function h is excessive (on the state space $\mathcal{Y}$) if

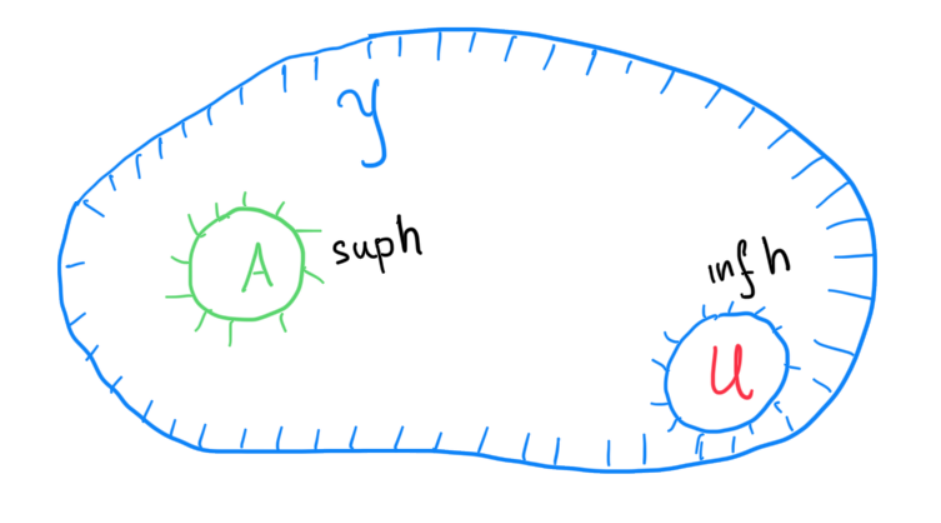
- $h \geq 0$
- $\mathcal{L}h \leq 0 \qquad (Ph \leq h)$

on $\mathcal{Y}$.

Stochastic barrier function

A function h is called a **stochastic barrier function** for the chain (X_n) w.r.t. a triple $(A, U, \mathcal{Y})$ if

- h is an excessive function on $\mathcal{Y}$
- $\inf\{h(u) |\ u \in U\} \geq \sup\{h(a) |\ a \in A\}$



Barrier functions - relation to safety

Proposition

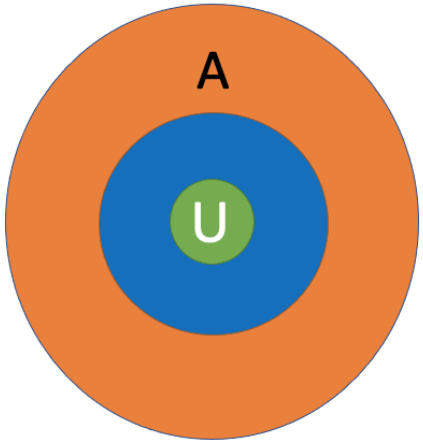
Suppose that there exists a barrier function h. Then

$$q(A; U, \mathcal{Y}) \leq \frac{H_A}{H_U}$$

$$H_A := \sup_{y \in A} h(y), \qquad H_U := \inf_{y \in U} h(y).$$

Example (Random walk)

$S = D_{(0,0)}(10)$, $U = D_{(0,0)}(1)$, and $A = D_{(0,0)}(10) \setminus \mathrm{int}(D_{(0,0)}(5))$.



Suppose $h(x) = 10^2 - x_1^2 - x_2^2$.

Since random walk is a martingale and h is concave, $(h_t) = (h(X_t))$ is a super-martingale, and hence an excessive function.

$$P(A; U, S) \leq \frac{10^2 - 5^2}{10^2 - 1^2} \approx 0.76$$

Properties of barrier certificates

Proposition

Let $\mathcal{C}_b$ be the set of all barrier functions for a Markov chain (X_n) and a triple $(A, U, \mathcal{Y})$. Then:

- *$\mathcal{C}_b$ is a positive cone that contains constant functions.*
- *If $h^1, h^2 \in \mathcal{C}_b$ then $h^1 \wedge h^2 \in \mathcal{C}_b$.*
- *If $\mathcal{C}_b \neq \emptyset$ then there exists a function $h \in \mathcal{C}_b$ and $p \in [0, 1]$ such that:*
  (a) $h \geq 1$ on U
  (b) $h \leq p$ on A.

p-safe computation

Corollary

Let $p \in [0,1]$. *If there exists a function* $h : \mathcal{Y} \to \mathbb{R}$ *such that*

(a) $h \geq 0 \quad$ *on* $\quad \mathcal{Y}$

(b) $Ph \leq h \quad$ *on* $\quad \mathcal{Y}$

(c) $h \geq 1 \quad$ *on* $\quad U$

(d) $h \leq p \quad$ *on* $\quad A$

then strong safety

$$P(A; U, \mathcal{Y}) \leq p$$

Main results on barrier certificates RAF: Perhaps remove

- Let $\mathcal{E}(\mathcal{Y})$ be the set of excessive functions on $\mathcal{Y}$.
  Consider the following set of barrier functions:

$$\mathcal{K}_b := \{h \in \mathcal{E}(\mathcal{Y}) |\ h \geq 1 \text{ on } U\}$$

- Define the value function $H : \mathcal{K}_b \times A \to \mathbb{R}_+$ by

$$H(h, j) := h(j)$$

  Then strong safety

$$q(A; U, \mathcal{Y}) = \max_{j \in A} \inf_{h \in \mathcal{K}_b} H(h, j) = \inf_{h \in \mathcal{K}_b} \max_{j \in A} H(h, j)$$

Optimisation for p-safety

$$q(A; U, \mathcal{Y}) = \inf_{h \in \mathcal{K}_b} \max_{j \in A} H(h, j)$$

translates to

$$q(A; U, \mathcal{Y}) = \inf p$$

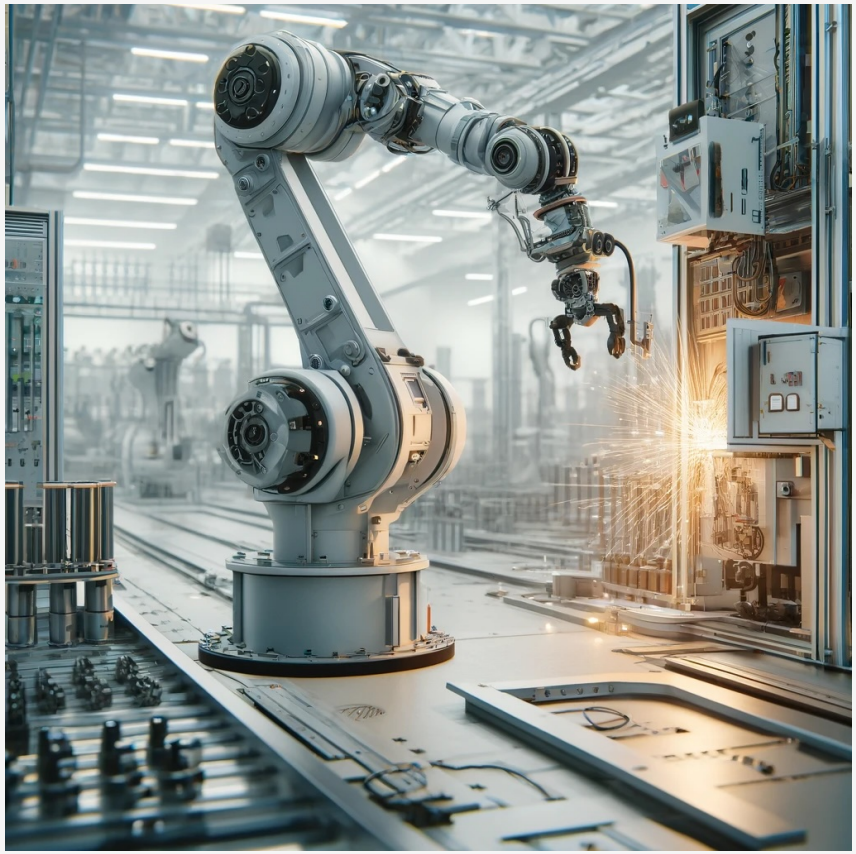
subject to $h \in \mathbb{R}_+^{\mathcal{Y}}$ and $p \in \mathbb{R}_+$ with

$$\begin{aligned} h(j) &\geq 0 && \text{for } j \in \mathcal{Y} \\ (Ph)(j) &\leq h(j) && \text{for } j \in \mathcal{Y} \\ h(j) &\geq 1 && \text{for } j \in U \\ h(j) &\leq p && \text{for } j \in A \end{aligned}$$

Safety $\implies$ Dynamic programming [3]

[3] *R. Wisniewski and M. L. Bujorianu, Probabilistic safety guarantees for Markov decision processes, to appear in IEEE Transactions on Automatic Control.*

Reach-avoidance problem



Markov Decision Processes - the notation

- Transition probabilities for MDP

$$p_{iaj} = \mathbb{P}[X_{t+1} = j | X_t = i,\ A_t = a]$$

where $(i, a, j) \in \mathcal{Y} \times \mathcal{A} \times \mathcal{Y}$, and the transition probability matrix

$$P(a) = [p_{iaj}]_{(i,j)\in\mathcal{Y}^2}$$

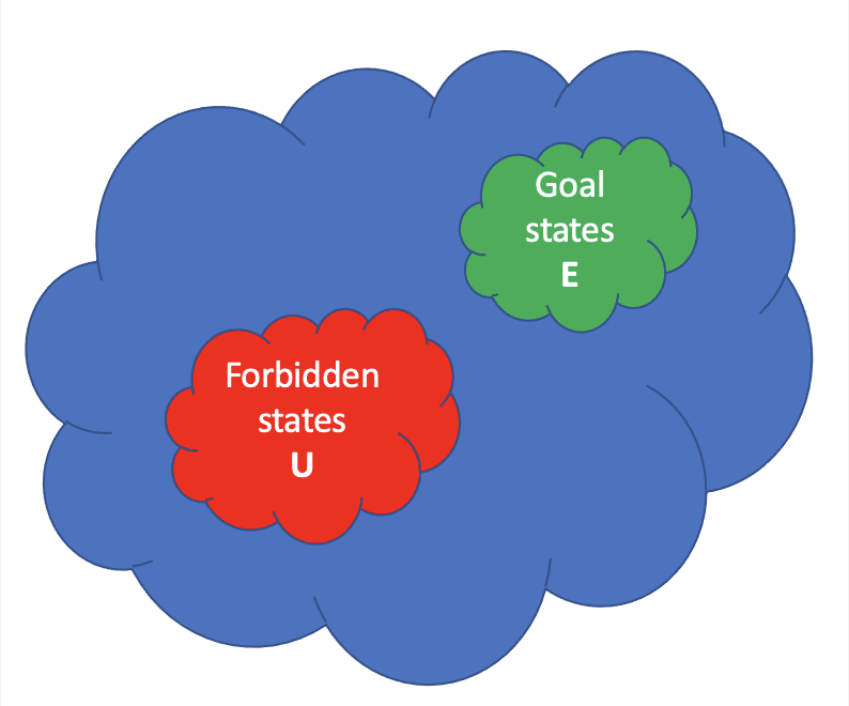
- A stationary Markov policy

$$\pi_{ia} = \mathbb{P}[A_t = a | X_t = i] = \mathbb{P}[A_0 = a | X_0 = i]$$

- A stationary policy π induces Markov chain

$$p_{ij}(\pi) = \sum_{a\in\mathcal{A}} \pi_{ia} p_{iaj}$$

Definitions of the reach-avoidance problem



Markov Chains - a taboo set

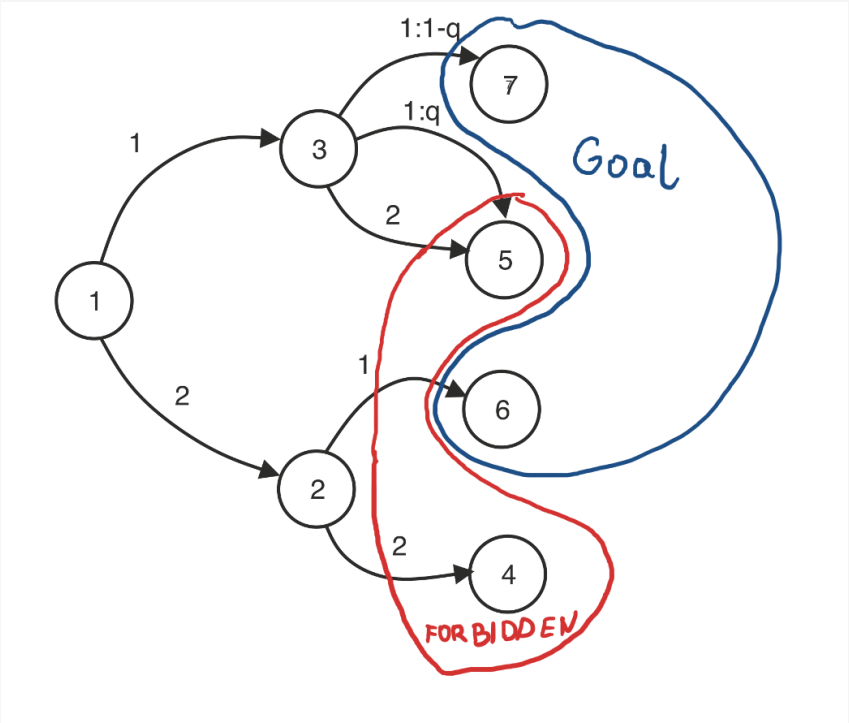


Figure: Goal $\mathbf{E} = \{6, 7\}$, Unsafe $\mathbf{U} = \{4, 5\}$, Taboo $\mathbf{H} = \{1, 2, 3\}$.

A small remark on an interesting object - Green kernel

- Green kernel

$$G := \sum_{k=0}^{\infty} P^k$$

- Then the cost function $V(y) = \mathbb{E}^y \sum_{t=0}^{\infty} \rho(X_t)$

$$V = GR \qquad \text{with } R = (\rho(1), \ldots, \rho(k), \ldots)$$

- $V = PV + R$ (Bellman equation).

A small remark on an interesting object - Green kernel

What to do to get Green kernel well defined:

- To study the resolvent by introducing the forgetting factor $\gamma \in]0,1[$

$$G := \sum_{k=0}^{\infty} \gamma^k P^k$$

Hence to study discounted cost functions:

$$V(y) = \mathbb{E}^y \sum_{t=0}^{\infty} \gamma^t R(X_t)$$

- To restrict the states to transient states (punctuate the state by removing the recurrent states)

Markov Chains

- H is an arbitrary subset of $\mathcal{Y}$ - the taboo set, and let $Q = (p_{ij})_{i,j \in H}$.
- The occupation (Green) operator on H,

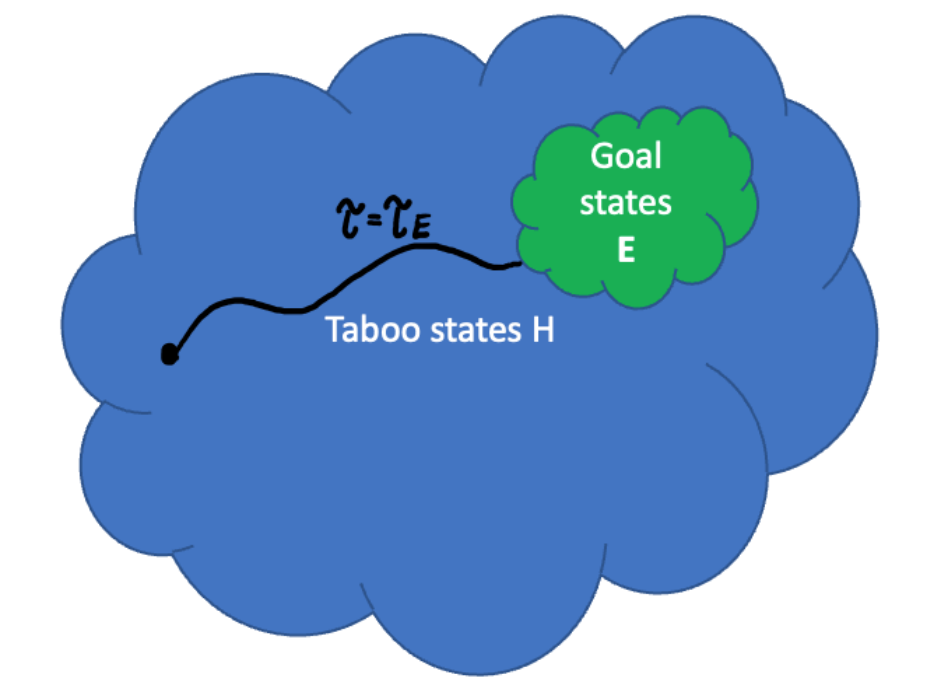
$$G := \sum_{k=0}^{\infty} Q^k$$

- Hence,

$$G = I + QG = I + GQ$$

i.e., $G = (I - Q)^{-1}$
$\mathcal{L} = -G^{-1}$

Markov Decision Problem



- Let E be a target set, and the taboo $H := \mathcal{Y} \setminus E$
- $\tau = \tau_E$ is the first hitting time of E.

Markov Decision Problem

- The cost for the policy π up to time τ is

$$V_\pi(i) := \mathbb{E}_\pi \left[\sum_{t=0}^{\tau-1} \rho(X_t, A_t) \mid X_0 = i \right]$$

- Let $R_\pi := (R_\pi(i))_{i \in H}$ with the components

$$R_\pi(i) = \sum_{a \in \mathcal{A}} \pi_{ia} \rho(i, a)$$

- The cost function V_π is

$$V_\pi = G(\pi) R_\pi$$

where $G(\pi)$ is the Green kernel associated to H.

Immediate reward for safety

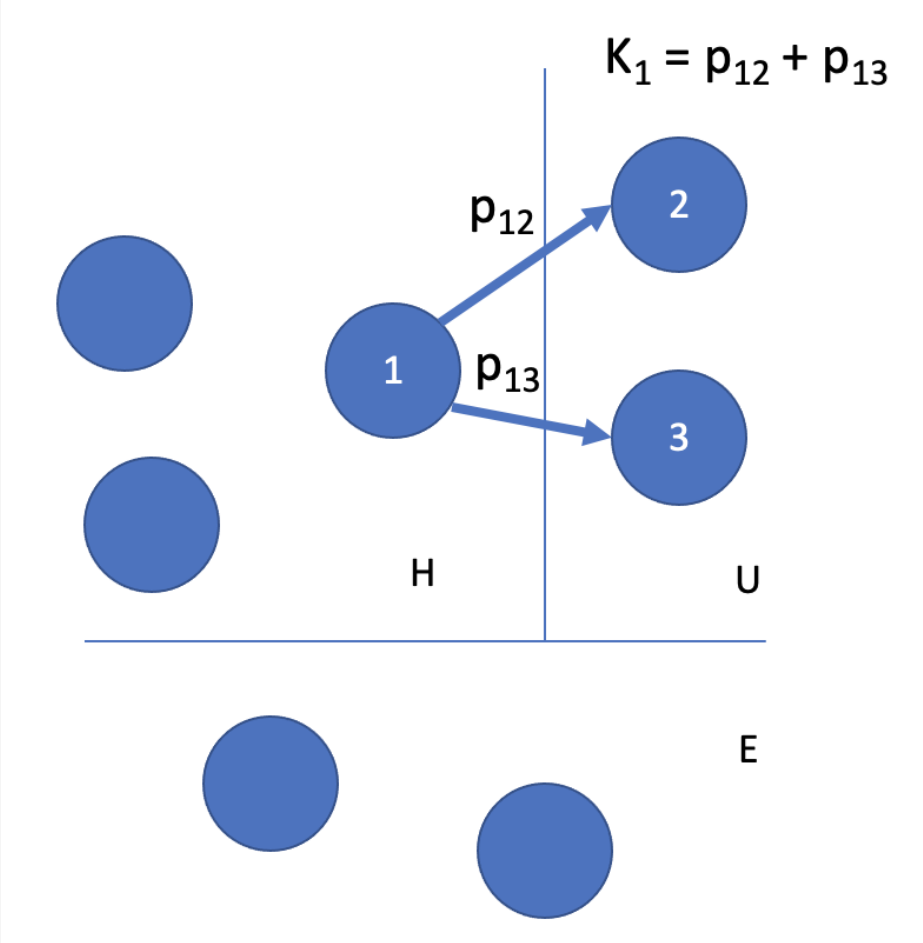


Figure: Immediate reward for safety

Lemma

Suppose μ is the initial distribution of the process (X_t) and the support of μ is in $H = \mathcal{Y} \setminus (E \cup U)$. The safety function $S_\pi(\mu)$ is given by

$$S_\pi^\mu = \mathbb{E}_\pi^\mu \sum_{t=0}^{\tau-1} \kappa(X_t, A_t)$$

where $\tau = \tau_{U \cup E}$, and $\kappa(i,a) = \sum_{j \in U} p_{iaj}$, for $i \in H$.

Lemma

Suppose that the MDP (X_t) with a policy π is transient on H. Let

$$K_\pi := P_H^U(\pi)\mathbb{1}_U$$

where $\mathbb{1}_U$ is the column vector of 1s of length $|U|$.
Then the safety function is given by

$$S_\pi = G(\pi)K_\pi$$

and since $G(\pi) = (I - Q(\pi))^{-1}$, it is the solution of

$$S_\pi = Q(\pi)S_\pi + K_\pi.$$

Furthermore, the sequence (S_π^n) defined by

$$S_\pi^{n+1} = Q(\pi)S_\pi^n + K_\pi$$

for an arbitrary S_π^0 converges point-wise to S_π.

Safety $\Longrightarrow$ Learning [4]

[4] *A. Mazumdar, R. Wisniewski and M. L. Bujorianu, Online learning of safety function for Markov decision processes, European Control Conference, 2023*

Proxy set

Figure: The need for proxy states.

Illustration of a proxy set

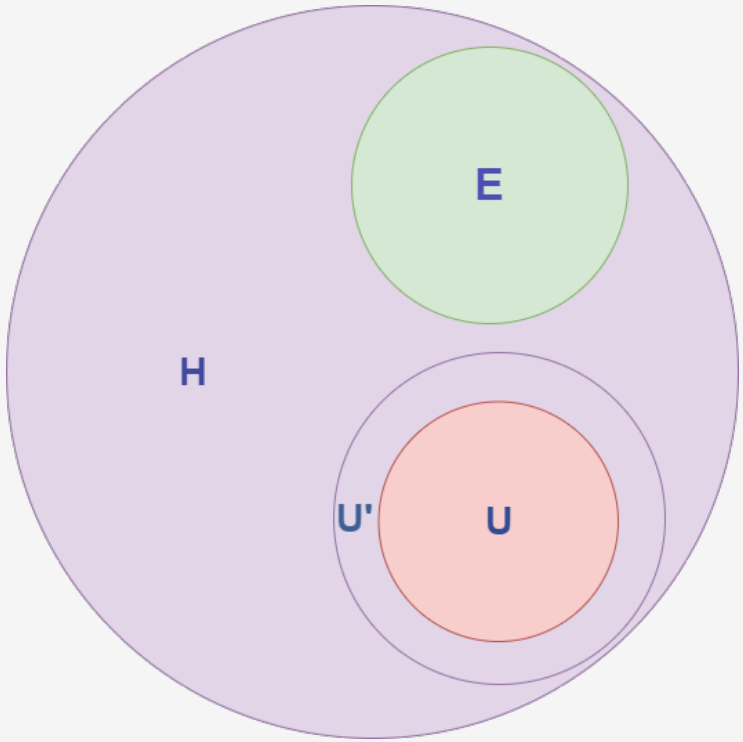


Figure: Illustration of the proxy set, forbidden set, target set, and the taboo set.

Definition

Let π be a policy. Let $U' \subset \mathcal{Y}$. Let q and w are in $[0,1]^{U' \setminus U}$.
The subset U' is a (q,w)-**proxy set** if it has the following properties:

1. U is a proper subset of U'
2. $\tau_{U'} < \tau_U$ almost surely
3. $w(i) \leq \mathbb{P}^i_\pi[\tau_U < \tau_E] \leq q(i)$ for all $i \in U' \setminus U$.

Proxy set

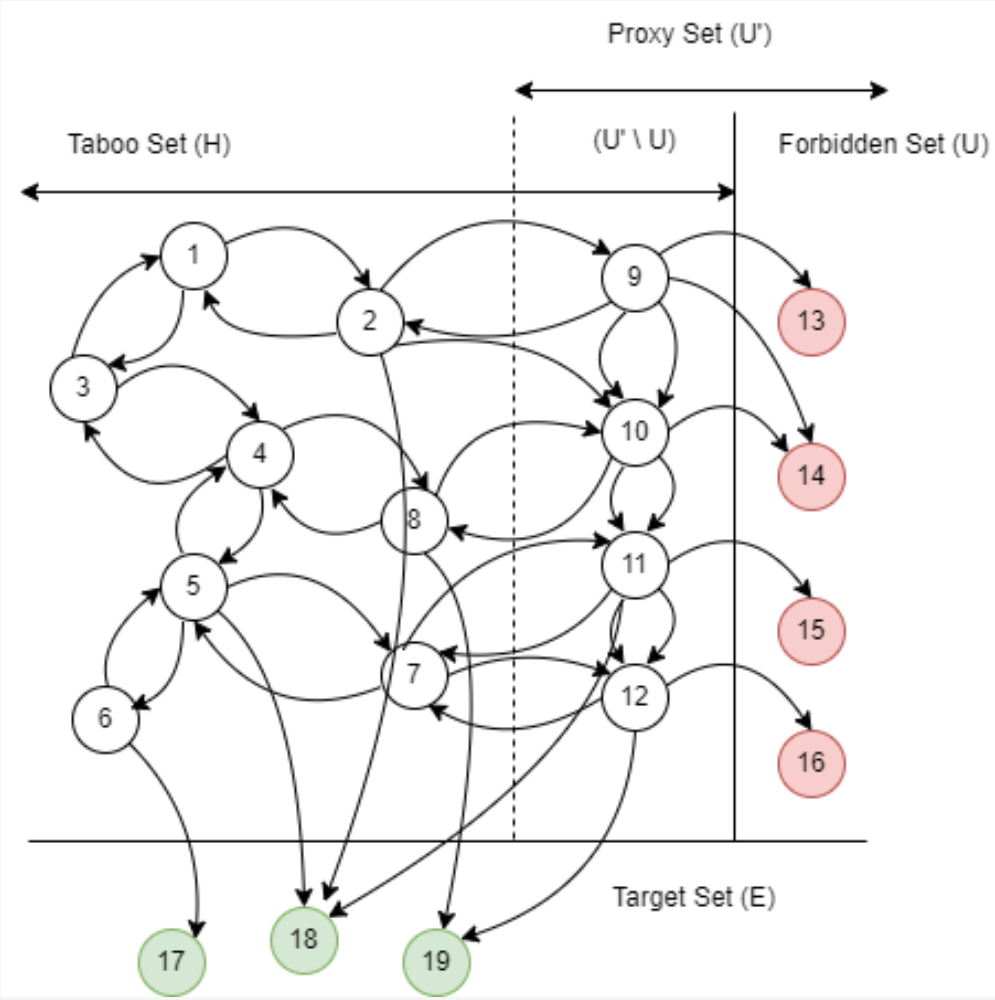


Figure: MDP with the proxy set, forbidden set, target set, and the taboo set.

Approximation of the safety function

Proposition

Suppose U' is a (q, w)-proxy set. The functions q and w are extended to the whole state space by assuming that $q(i) = w(i) = 0$ for $i \in \mathcal{Y} \setminus U'$. Let $\tau' = \tau_{U' \cup E}$. Then

$$\mathbb{E}_\pi^i \sum_{t=0}^{\tau'} w(X_t) \leq S_\pi^\mu \leq \mathbb{E}_\pi^i \sum_{t=0}^{\tau'} q(X_t)$$

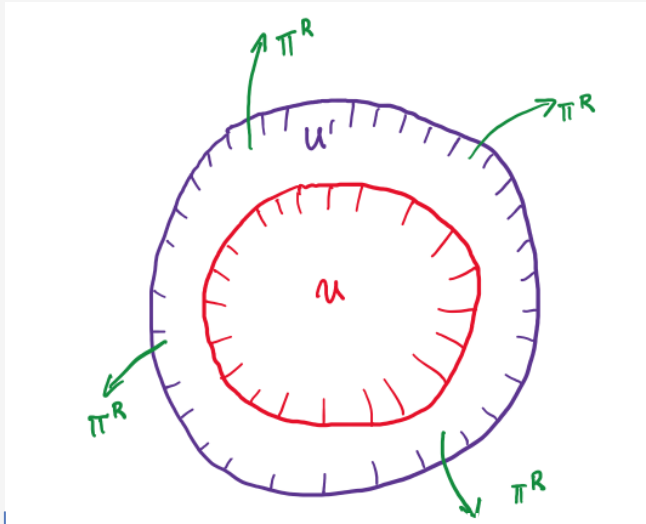
Definition (Repelling policy)

For a policy π, a forbidden set U, a proxy set U', a repelling policy π^R:

$$\mathbb{P}_{\pi^R}[X_{t+1} \in \mathcal{Y} \setminus U' | \; X_t \in U' \setminus U] = 1$$

Resulting policy:

$$\tilde{\pi}(i) := \begin{cases} \pi(i), & \text{for } i \in \mathcal{Y} \setminus U' \\ \pi^R(i), & \text{for } i \in U' \setminus U \end{cases}$$



Learning safety (upper/lower bounds)

Learning upper boud with TD(0) (value iteration)

$$V_{t+1}(X_t) = (1 - \alpha_t(X_t))\, V_t(X_t) + \alpha_t(X_t)\,[q_t + V_t(X_{t+1})]$$

Illustration

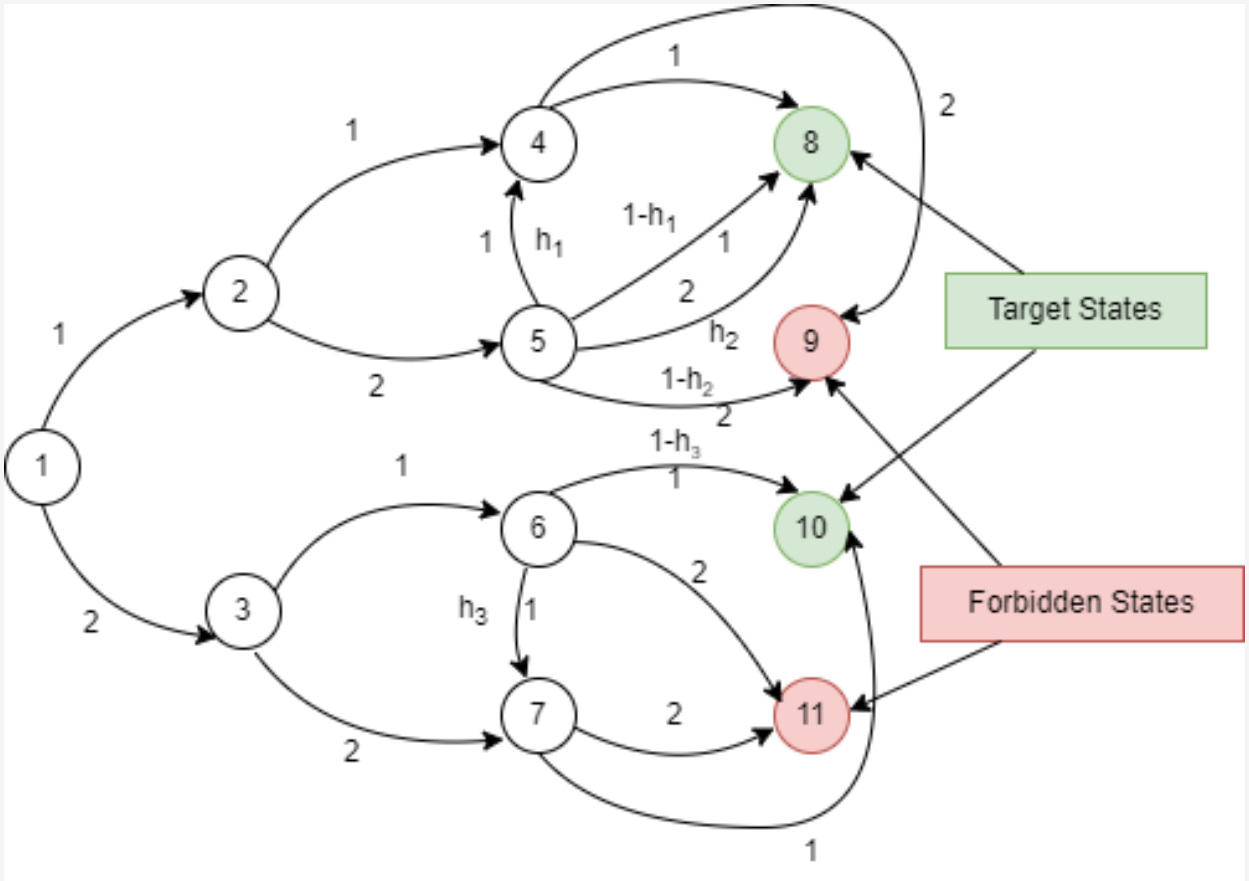
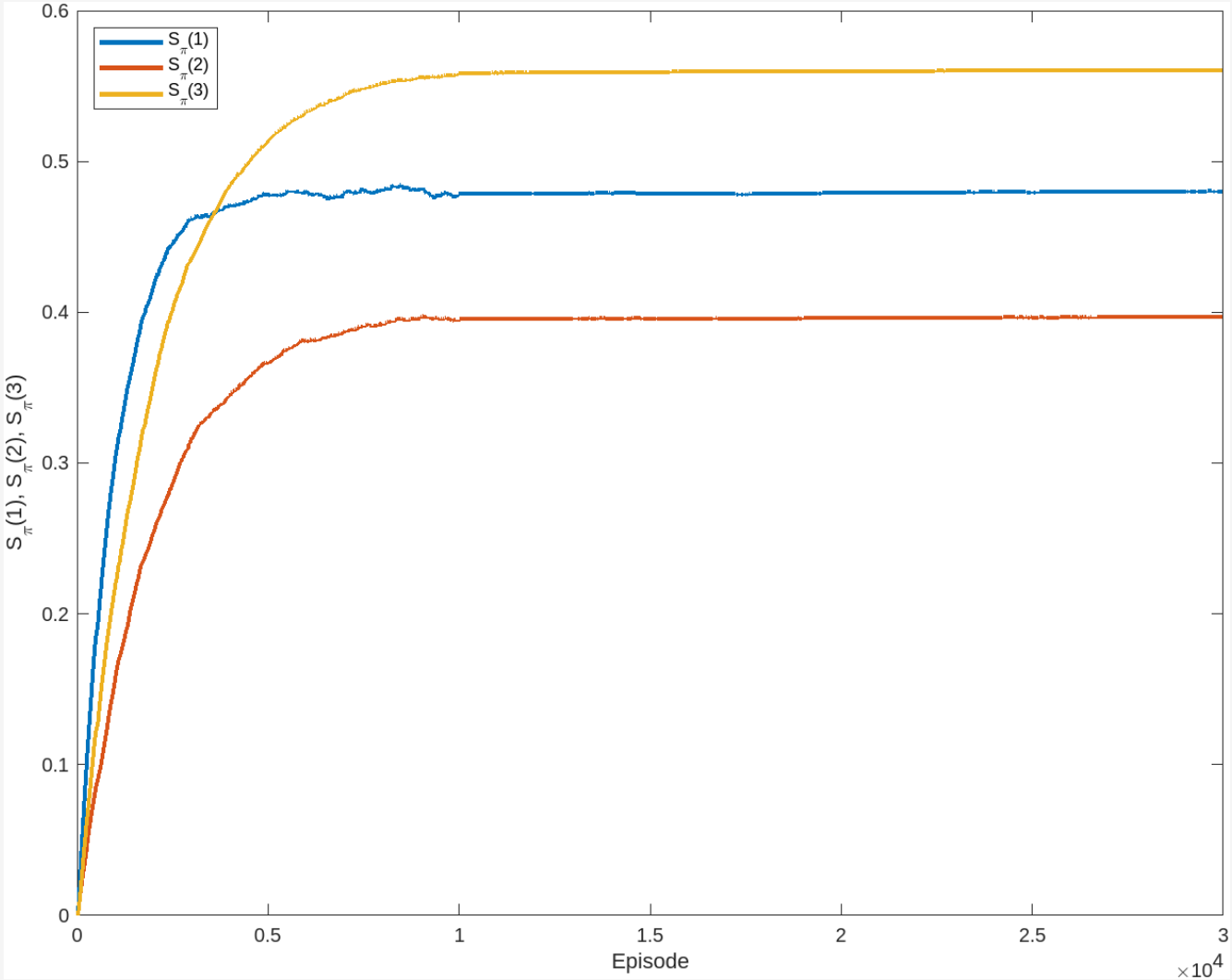


Figure: Policy is uniformly distributed, $\pi_{ia} = 0.5$, $\forall i \in \mathcal{Y}$ and $\forall a \in \{1, 2\}$. We set $h_1 = 0.4$, $h_2 = 0.6$ and $h_3 = 0.5$. Upper bound for $\mathbb{P}^i_\pi[\tau_U < \tau_E]$ is 0.7, 0.5, 0.8 and 0.7 for state $i = 4$, 5, 6 and 7, respectively. The lower bound for $\mathbb{P}^i_\pi[\tau_U < \tau_E]$ is assumed to be 0.3, 0.15, 0.5 and 0.35 for state 4, 5, 6 and 7.

Bounds of the Safety function

State i	$S_\pi(i)$	$S_\pi(i)$ using RL algorithm	Upper bound of $S_\pi(i)$	Lower bound of $S_\pi(i)$
1	0.4813	0.4816	0.6740	0.3253
2	0.4	0.4	0.5967	0.2242
3	0.5625	0.5609	0.7475	0.4248

Learning convergence



Safety $\implies$ MDP with constraints [5]

[5] *R. Wisniewski and M. L. Bujorianu, Probabilistic safety guarantees for Markov decision processes, to appear in IEEE Transactions on Automatic Control.*

Safety guarantees in stochastic optimisation

- We strive to find the minimum V^* of the cost

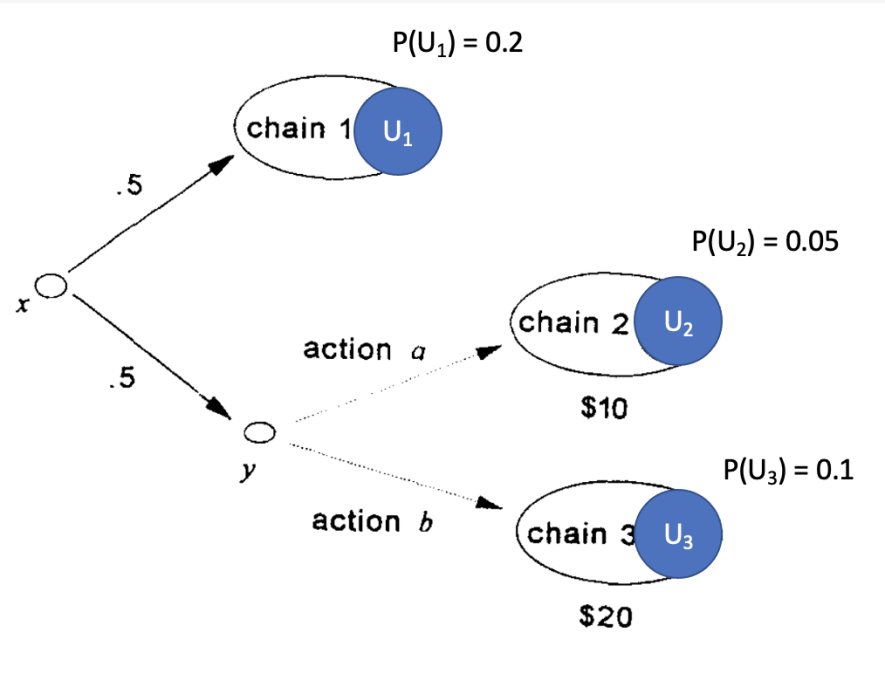
$$V_\pi^\mu := \mathbb{E}_\pi^\mu \left[\sum_{t=0}^{\tau_E - 1} \rho(X_t, A_t) \right]$$

subject to

$$S_\pi^\mu \leq p$$

Haviv's counter-example for principle of optimality [6]

Maximize the total return while keeping the probability of occupying states in $U = U_1 \cup U_2 \cup U_3$ less than or equal to $\mathbf{p} = \mathbf{0.125}$.



$$\mathbb{P}[X_\tau \in U | \mathbf{X_0} = \mathbf{x}, \underline{\mathbf{A_2} = \mathbf{a}}] = 0.5\ 0.2 + 0.5\ 0.05 = \mathbf{0.125}$$

$$\mathbb{P}[X_\tau \in U | \mathbf{X_0} = \mathbf{x}, \mathbf{A_2} = \mathbf{b}] = 0.5\ 0.2 + 0.5\ 0.1 = \mathbf{0.15}$$

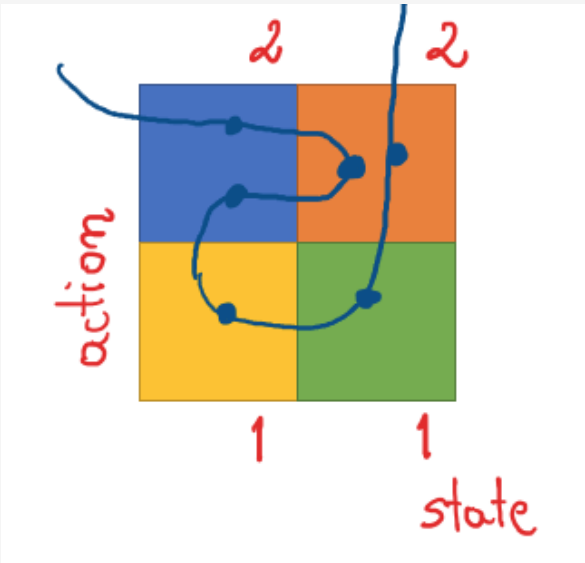
$$\mathbb{P}[X_\tau \in U | \mathbf{X_0} = \mathbf{y}, \mathbf{A_2} = \mathbf{a}] = \mathbf{0.05} \quad \text{and} \quad \mathbb{P}[X_\tau \in U | \mathbf{X_0} = \mathbf{y}, \underline{\mathbf{A_2} = \mathbf{b}}] = \mathbf{0.1}$$

[6] *M. Haviv, On constrained Markov decision processes, Operat. research let. , 1996.*

Linear programming

- A state-action occupation measure

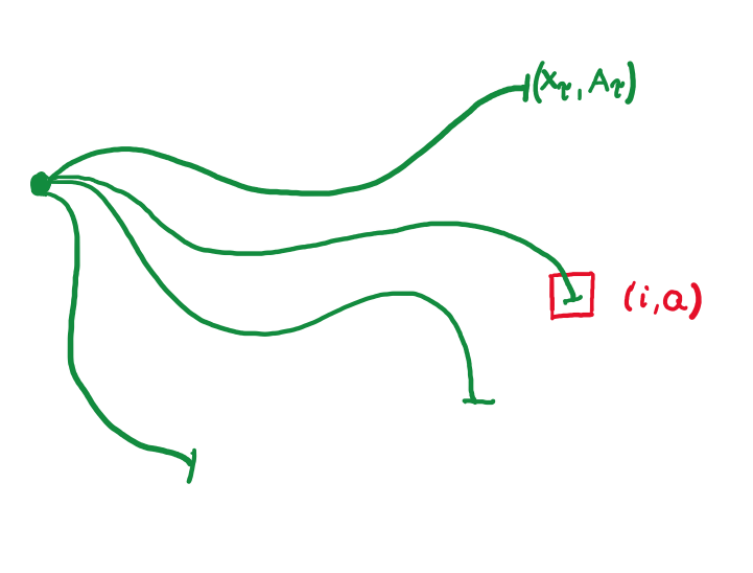
$$\overline{\gamma}^{\mu}(i,a) := \sum_{t=0}^{\infty} \mathbb{P}^{\mu}[X_t = i, A_t = a, t < \tau] = \mathbb{E}^{\mu} \sum_{t=0}^{\tau-1} I_{\{(X_t,A_t)=(i,a)\}}$$



- Then, the expectated value of a function $\rho : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ is:

$$\mathbb{E}^{\mu}\left[\sum_{t=0}^{\tau-1} \rho(X_t, A_t)\right] = \sum_{a} \sum_{i} \rho(i,a)\overline{\gamma}^{\mu}(i,a)$$

Evolution equation



- A state-action hitting measure by

$$\overline{\lambda}^{\mu}_{\tau}(i,a) := \mathbb{P}^{\mu}[X_\tau = i, A_\tau = a]$$

for $(i,a) \in \mathcal{Y} \times \mathcal{A}$.

- The evolution equation for the state-action measures is

$$\sum_{a\in\mathcal{A}} \overline{\lambda}^{\mu}_{\tau}(\cdot,a) = \mu(\cdot) + \sum_{a\in\mathcal{A}} \overline{\gamma}^{\mu}_{<\tau}(\cdot,a)\mathcal{L}(a), \qquad \mathcal{L}(a) = P(a) - I$$

Linear programming for the safety guarantee

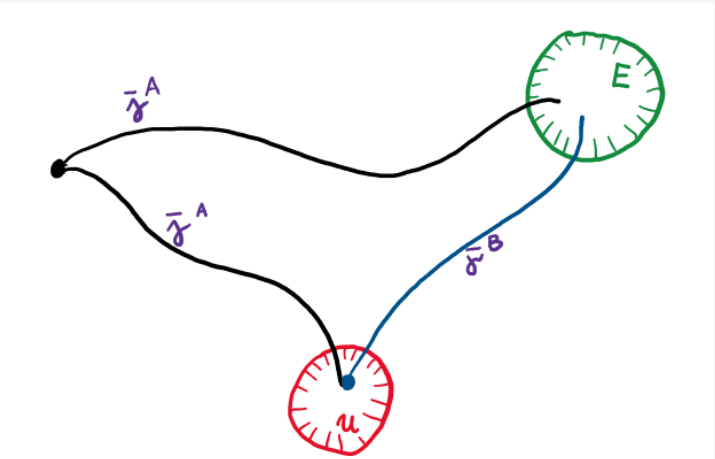
- Two occupation measures

$$\overline{\gamma}^A(i,a) = \sum_{t=0}^{\infty} \mathbb{P}^\mu[X_t = i, A_t = a, t < \tau]$$

$$\overline{\gamma}^B(i,a) = \sum_{t=0}^{\infty} \mathbb{P}^\mu[X_t = i, A_t = a, \tau \leq t < \tau_E]$$

- and the hitting measures, for $(i,a) \in \mathcal{Y} \times \mathcal{A}$

$$\overline{\lambda}^A(i,a) = \mathbb{P}^\mu[X_\tau = i, A_\tau = a], \quad \overline{\lambda}^B(i,a) = \mathbb{P}^\mu[X_{\tau_E} = i, A_{\tau_E} = a].$$



Proposition

The minimum of V_π^μ over stationary policies $\pi \in \mathcal{D}^{\mathcal{Y}\setminus E}$ is the solution of the following linear program

$$\min_{\pi\in\mathcal{D}^{\mathcal{Y}\setminus E}} V_\pi^\mu = \min \sum_{i\in H\cup U}\sum_{a\in\mathcal{A}} (\overline{\gamma}^A(i,a)+\overline{\gamma}^B(i,a))\rho(i,a)$$

subject to

$$\sum_{a\in\mathcal{A}} \overline{\lambda}^A(j,a) = \mu(j) + \sum_{(i,a)\in\mathcal{Y}\times\mathcal{A}} \overline{\gamma}^A(i,a)(p_{iaj}-\delta_j(i))$$

$$\sum_{a\in\mathcal{A}} \overline{\lambda}^B(j,a) = \sum_{a\in\mathcal{A}} \overline{\lambda}^A(j,a) + \sum_{(i,a)\in\mathcal{Y}\times\mathcal{A}} \overline{\gamma}^B(i,a)(p_{iaj}-\delta_j(i))$$

both $\overline{\lambda}^A$ and $\overline{\lambda}^B$ are zero on H, $\overline{\lambda}^B$ is additionally zero on U,

$$\sum_{(i,a)\in H\times\mathcal{A}} \overline{\gamma}^A(i,a)\kappa(i,a) \le p$$

Proposition (cnt.)

The optimal policy is given by

$$\pi_{ia} = \frac{\overline{\gamma}^A(i,a) + \overline{\gamma}^B(i,a)}{\gamma(i)},$$

where

$$\gamma(i) = \sum_{a \in \mathcal{A}} (\overline{\gamma}^A(i,a) + \overline{\gamma}^B(i,a)).$$

What next?

Learning the transition probabilities

$$\forall x \in H,\ \forall a \in A,\ \forall y \in \mathcal{Y}$$

$$\gamma(x, a, y) = \frac{N(x, a, y)}{N(x, a) \vee 1}$$

where, $N(x, a)$ and $N(x, a, y)$ the number of visits to (x, a) and (x, a, y).