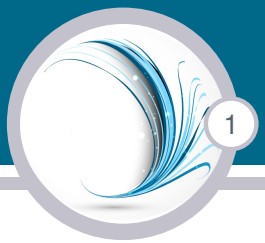
Departamento de Matemáticas, Universidad de los Andes

On controllability of Markov chains: A Markov decision process approach

Daniel Ávila, Mauricio Junca

February 23, 2018

Content



Reaching a Set of States

- We would like to control a system in such a way that the system is taken to a desired place, while some regions are avoided.
- There is uncertainty in the system.

Reaching with Constraints

We would like to control the system in such a way that the system is taken to a desired place, while some regions are avoided with certain tolerance.

Domain of Attraction

We would like to understand the set of initial states which take the system to a certain set.

Controlled Markov Chains

$\mathbb{X}, \mathbb{U}$ are sets corresponding to the state space and control actions respectively, in our work $\mathbb{X}$ and $\mathbb{U}$ are countable.

Consider the measurable space $(\Omega, \mathcal{F})$, where $\Omega := (\mathbb{X} \times \mathbb{U})^{\mathbb{N}}$ and $\mathcal{F}$ is the product sigma algebra.

$\mathbb{U}(x) \subset \mathbb{U}$ is the set of feasible control actions. The set of feasible states and actions is denoted as $\mathbb{K}$, in symbols we can write it as,

$$\mathbb{K} = \{(x, u) | x \in \mathbb{X}, u \in \mathbb{U}(x)\}$$

Q is a stochastic kernel on $\mathbb{X}$ given $\mathbb{K}$, that is, for each $(x, u) \in \mathbb{K}$ $Q(\cdot|x, u)$ is a probability measure on $\mathbb{X}$, and for each $B \subset \mathbb{X}$, the function $Q(B|\cdot)$ is measurable on $\mathbb{K}$.

Control Policies

Let $H_t = \mathbb{K}^t \times \mathbb{X}$ be the set of admissible histories up to time t, so $h_t = (x_0, u_0, \ldots, x_{t-1}, u_{t-1}, x_t)$.

A control policy is a sequence $\pi = \{\mu_t\}_{t \geq 0}$, where each μ_t is a function $\mu_t : H_t \to \mathcal{P}(\mathbb{U})$,

$$\mu_t\left(\mathbb{U}(x_t) | h_t = (x_0, u_0, \ldots, x_{t-1}, u_{t-1}, x_t)\right) = 1 \quad \forall h_t \in H_t \quad t \geq 0$$

We denote by $\overline{\Pi}$ the set of control policies.

Process Induced by a Control Policy

Given a policy $\pi \in \overline{\Pi}$ and an initial distribution ν over $\mathbb{X}$ there exist a unique probability measure P_ν^π on $(\Omega, \mathcal{F})$ and $\mathbb{X} \times \mathbb{U}$-valued process $\{(X_t, U_t)\}_{t \geq 0}$ such that:

- $P_\nu^\pi(\mathbb{K}^\mathbb{N}) = 1$.
- $P_\nu^\pi(X_0 \in B) = \nu(B)$.
- $P_\nu^\pi(U_t \in B | h_t) = \mu_t(B|h_t)$.
- $P_\nu^\pi(X_{t+1} \in B | h_t, U_t = u) = Q(B|x_t, u)$.

Thus given $\pi \in \overline{\Pi}$,

$$P_\nu^\pi(X_{t+1} \in B|h_t) = \sum_{u_t} Q(B|x_t, u_t) \cdot \mu_t(u_t|h_t)$$

The process $\{X_t\}_{t\geq 0}$ may not be Markovian.

Markovian Policy:

It doesn't depend on the history, that is: a policy $\pi = \{\mu_t\}_{t\geq 0}$, where $\mu_t : \mathbb{X} \to \mathcal{P}(\mathbb{U})$. We denote by Π the set of control policies.

Given $\pi \in \Pi$ then $\{X_t\}$ is Markovian,

$$P_\nu^\pi(X_{t+1} \in B|x_t) = \sum_{u_t} Q(B|x_t, u_t) \cdot \mu_t(u_t|x_t)$$

Closed sets

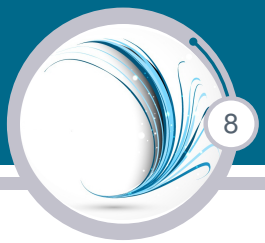
Given a policy $\pi = \{\mu_t\} \in \overline{\Pi}$ and a set $A \subset \mathbb{X}$, we say A is closed under π if:

For all $s, t \in \mathbb{N}, s < t$ such that $x_s \in A, x_t \notin A$ we have,

$$P_\nu^\pi(X_t = x_t | h_{t-1}) = \sum_{u_{t-1} \in \mathbb{U}(x_{t-1})} Q(x_t | x_{t-1}, u_{t-1}) \cdot \mu_t(u_{t-1} | h_{t-1}) = 0$$

In case of a Markovian policy, A is closed if for all $i \in A, j \in A^c$

$$P_\nu^\pi(X_t = j | X_{t-1} = i) = 0$$



Let r be a function $r : \mathbb{K} \to \mathbb{R}$, we call it the reward function.

Given $x \in \mathbb{X}$ and $\pi \in \overline{\Pi}, 0 < \gamma < 1$, we define the discounted cost as:

$$v_\gamma^\pi(x) = \mathbb{E}_x^\pi \Big[\sum_{t=0}^{\infty} \gamma^t \cdot r(X_t, U_t) \Big]$$

The discounted cost problem is to find a policy π^* such that,

$$v_\gamma^{\pi^*}(x) = \sup_{\pi \in \overline{\Pi}} v_\gamma^\pi(x) =: v_\gamma^*(x) \quad \forall x \in \mathbb{X}$$

Optimality Equations

The system of equations,

$$v_\gamma(x) = \sup_{u \in \mathbb{U}(x)} \left\{ r(x,u) + \gamma \sum_{j \in \mathbb{X}} P^u_{xj} \cdot v_\gamma(j) \right\} \quad x \in \mathbb{X}$$

where $P^u_{x,j} := Q(j|x,u)$, are called the Bellman's equation for the discounted system.

To get a sense of why such equations make sense let's fix a Markovian stationary policy π.

$$
\begin{aligned}
v_\gamma^\pi(x) &= \mathbb{E}_x^\pi \left[\sum_{t=0}^{\infty} \gamma^t \cdot r(X_t, U_t) \right] \\
&= \mathbb{E}_x^\pi \left[r(X_0, U_0) \right] + \mathbb{E}_x^\pi \left[\sum_{t=1}^{\infty} \gamma^t \cdot r(X_t, U_t) \right] \\
&= \mathbb{E}_x^\pi \left[r(X_0, U_0) \right] + \sum_{j \in \mathbb{X}} \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^t \cdot r(X_t, U_t) \middle| X_1 = j, X_0 = x \right] \cdot P_{x,j}^\pi \\
&= \mathbb{E}_x^\pi \left[r(X_0, U_0) \right] + \gamma \sum_{j \in \mathbb{X}} \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} \cdot r(X_t, U_t) \middle| X_1 = j \right] \cdot P_{x,j}^\pi \\
&= \mathbb{E}_x^\pi \left[r(X_0, U_0) \right] + \gamma \sum_{j \in \mathbb{X}} v_\gamma^\pi(j) \cdot P_{x,j}^\pi
\end{aligned}
$$

$$= \sum_{u \in \mathbb{U}(x)} r(x,u) \cdot \pi(u|x) + \gamma \sum_{j \in \mathbb{X}} v_\gamma^\pi(j) \cdot \Big[\sum_{u \in \mathbb{U}(x)} P_{x,j}^u \cdot \pi(u|x) \Big]$$

$$= \sum_{u \in \mathbb{U}(x)} \pi(u|x) \cdot \Big[r(x,u) + \gamma \sum_{j \in \mathbb{X}} v_\gamma^\pi(j) \cdot P_{x,j}^u \Big]$$

$$\leq v_\gamma(x) = \sup_{u \in \mathbb{U}(x)} \Big\{ r(x,u) + \gamma \sum_{j \in \mathbb{X}} P_{xj}^u \cdot v_\gamma(j) \Big\}$$

Therefore,

$$\sup_{u \in \mathbb{U}(x)} \Big\{ r(x,u) + \gamma \sum_{j \in \mathbb{X}} P_{xj}^u \cdot v_\gamma(j) \Big\} = \sup_{\pi \in \Pi_S} \Big\{ v_\gamma^\pi(x) \Big\}$$

In general, we would like to prove $v_\gamma = v_\gamma^*$.

Define an operator $\mathbb{L} : V \to V$,

$$\mathbb{L}(v) := \sup_{\pi \in D} \left\{ r^\pi + \gamma \cdot P^\pi \cdot v \right\}$$

Lemma

Suppose there exist $v \in V$, such that,

- $v \geq \mathbb{L}(v)$. Then, $v \geq v_\gamma^*$.
- $v \leq \mathbb{L}(v)$. Then, $v \leq v_\gamma^*$.
- $v = \mathbb{L}(v)$. Then, v is the only element with such property and $v = v_\gamma^*$.

Lemma

For $\mathbb{X}$ countable and $\mathbb{U}(x)$ finite, there exists a solution. It's enough to restrict our attention to Markovian policies, the optimal policy can be found trough a linear program.

Let r be a function $r : \mathbb{K} \to \mathbb{R}$, we call it the reward function.

Given $x \in \mathbb{X}$ and $\pi \in \overline{\Pi}$, we define the long-run average cost as:

$$v^{\pi}(x) = \limsup_{N \to \infty} \frac{1}{N} \mathbb{E}_x^{\pi} \left[\sum_{t=0}^{N-1} r(X_t, U_t) \right]$$

The average cost problem is to find a policy π^* such that,

$$v^{\pi^*}(x) = \sup_{\pi \in \overline{\Pi}} v^{\pi}(x) =: v^*(x) \quad \forall x \in \mathbb{X}$$

Lemma

$$\lim_{\gamma \to 1} (1 - \gamma) v_{\gamma}(x) = v(x)$$

We have the so-called optimality equations:

$$\max_{u \in \mathbb{U}(x)} \left\{ \sum_{j \in \mathbb{X}} P^u_{x,j} v(j) - v(x) \right\} = 0, \quad \forall x \in \mathbb{X}, \tag{MC1}$$

$$\max_{u \in \mathbb{U}(x)} \left\{ r(x,u) - v(x) + \sum_{j \in \mathbb{X}} P^u_{x,j} h(j) - h(x) \right\} = 0, \quad \forall x \in \mathbb{X}, \tag{MC2}$$

where $P^u_{x,j} := Q(j|x,u)$.

Lemma

For $\mathbb{X}$ finite and $\mathbb{U}(x)$ finite, there exists a solution. It's enough to restrict our attention to Markovian policies, the optimal policy can be found trough a linear program.

Plan

Introduction
- Problem of Interest
- Controlled Markov Chains
- Markov Decision Process

Formulation
- Reaching a set of states
- Domain of Attraction
- Reaching with Constraints

Reaching a set of states

We would like to find control policy that maximize the probability of reaching some set $A \subset \mathbb{X}$ while avoiding some set $B \subset \mathbb{X}$.

Let τ_A and τ_B be the hitting times of A and B. Consider an initial distribution ν over the state space.

Our objective will be to find a control policy $\pi \in \overline{\Pi}$ that solves the problem,

$$\max_{\pi} \left\{ P_{\nu}^{\pi}(\tau_A < \tau_B, \tau_A < \infty) \right\} \tag{P1}$$

Idea:

Let's put a reward for passing through the states of A and maximize the expected reward.

How to avoid the set B? Let's change the probability measure so that B is a closed set.

Average Cost vs Hitting time

An average cost function captures the long time run of the process, in our problem, we are interested in the hitting time of A.

Consider the following average cost function for any $\pi \in \overline{\Pi}$,

$$v^{\pi}(x) := \limsup_{N\to\infty} \frac{1}{N} \widehat{\mathbb{E}}_x^{\pi}\Big[\sum_{t=0}^{N-1} 1_A(X_t)\Big]$$

Proposition

Let $x \in \mathbb{X}$ and $\pi \in \overline{\Pi}$. If A is closed under π then,

$$v^{\pi}(x) = P_x^{\pi}(\tau_A < \infty).$$

Proof

Let $\{s_t\}_{t\geq 0}$ be a sequence such that,

$$\lim_{t\to\infty} s_t = L$$

Then,

$$\lim_{N\to\infty} \frac{s_0 + \ldots + s_{N-1}}{N} = L$$

If A is closed under π and P then,

$$\lim_{t\to\infty} P_x^\pi(X_t \in A) = P_x^\pi(\tau_A < \infty).$$

Proof

For any $N \in \mathbb{N}$ we have that

$$\frac{1}{N}\mathbb{E}_x^\pi\Big[\sum_{t=0}^{N-1} \mathbf{1}_A(X_t)\Big] = \frac{1}{N}\sum_{t=0}^{N-1} P_x^\pi(X_t \in A).$$

Taking lim sup on both sides we get

$$v^\pi(x) = \limsup_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1} P_x^\pi(X_t \in A).$$

Therefore,

$$v^\pi(x) = \lim_{t\to\infty} P_x^\pi(X_t \in A) = P_x^\pi(\tau_A < \infty).$$

Change of measure

Thus to use average cost functions we need to modify the probability measure so that A is closed. To avoid B let's force it to be closed.

Let's define $\widehat{Q}$ as follows:

$$\begin{cases} \widehat{Q}(A|x,u) = 1 & \text{if } x \in A \\ \widehat{Q}(B|x,u) = 1 & \text{if } x \in B \\ \widehat{Q}(\cdot|x,u) = Q(\cdot|x,u) & \text{otherwise} \end{cases}$$

The measure induced by $\widehat{Q}$ will be denoted as $\widehat{P}$ and $\widehat{\mathbb{E}}$ will denote the expectation with respect to $\widehat{P}$.

Proposition

Given a policy $\pi \in \overline{\Pi}$, we have that

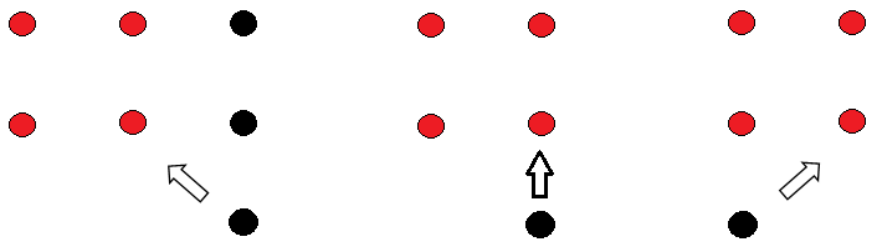
$$P_\nu^\pi(\tau_A < \tau_B, \tau_A < \infty) = \widehat{P}_\nu^\pi(\tau_A < \infty) \tag{1}$$

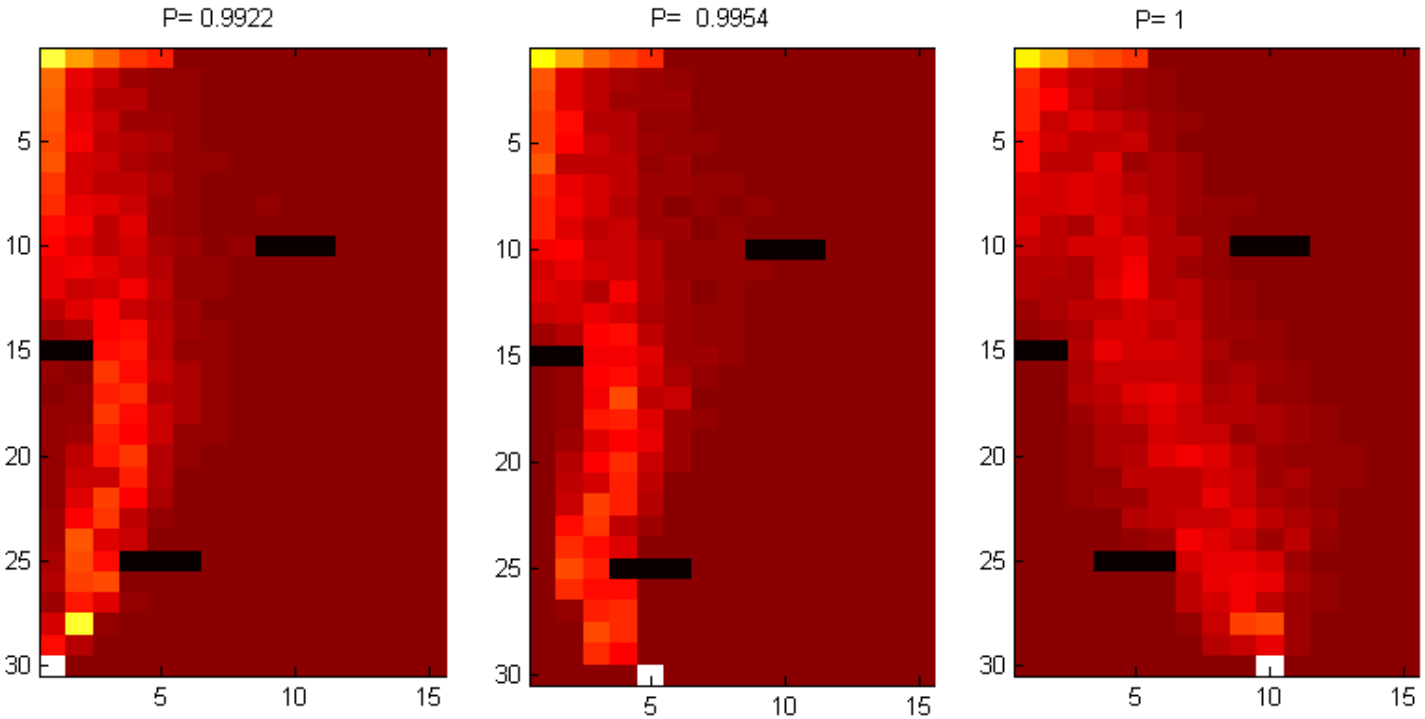
Conclusion

Theorem

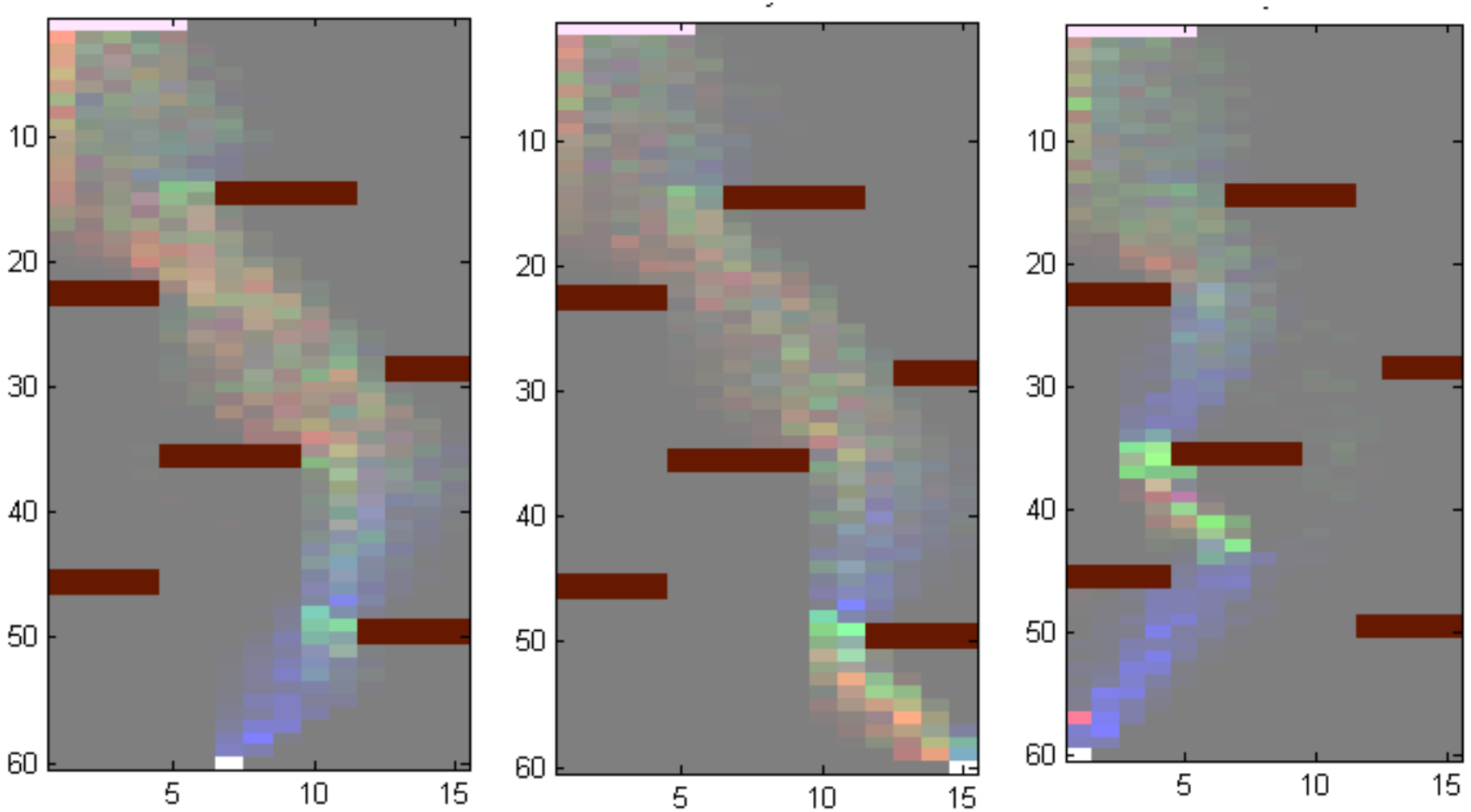
$$\max_{\pi}\left\{P_{\nu}^{\pi}(\tau_A < \tau_B, \tau_A < \infty)\right\} = \max_{\pi}\left\{\widehat{P}_{\nu}^{\pi}(\tau_A < \infty)\right\}$$
$$= \max_{\pi}\left\{v^{\pi}(x)\right\}$$

Example





Example


10
20
30
40
50
60
5
10
15
10
20
30
40
50
60
5
10
15
10
20
30
40
50
60
5
10
15

Domain of Attraction

Definition

$$\Lambda_p = \left\{ x \in \mathbb{X} \,\middle|\, \liminf_{t\to\infty} P_x^\pi (X_t \in A) > p \text{ for some policy } \pi \in \overline{\Pi} \right\}$$

$$\Gamma = \left\{ x \in \mathbb{X} \,\middle|\, \liminf_{t\to\infty} P_x^\pi (X_t \in A) = 0 \text{ for all policies } \pi \in \overline{\Pi} \right\}$$

Assumption

There exists a policy which makes A a closed set. Let Π_A be the set of policies which make A a closed set.

Proposition

$$\Lambda_p = \{x \in \mathbb{X} \mid P_x^\pi(\tau_A < \infty) > p \text{ for some policy } \pi \in \Pi_A\}$$

$$\Gamma = \{x \in \mathbb{X} \mid P_x^\pi(\tau_A < \infty) = 0 \text{ for all policies } \pi \in \Pi_A\}$$

$$v^* = \sup_{\pi \in \Pi_A} v^\pi$$

Theorem

Assume there exists a policy that makes A a closed set,

$$\Lambda_p = \left\{x \in \mathbb{X} \,\middle|\, v^*(x) > p\right\}$$

$$\Gamma = \left\{x \in \mathbb{X} \,\middle|\, v^*(x) = 0\right\}$$

Reaching with Constraints

We would like to solve a constrained version of Problem (P1).

Our objective will be to find a control policy that maximize the probability of reaching A in such a way that the probability of reaching B is less than some $\epsilon > 0$.

The formulation is as follows.

$$\max_{\pi \in \overline{\Pi}} \left\{ P_\nu^\pi(\tau_A < \infty) \right\}$$
$$\text{s.t. } P_\nu^\pi(\tau_B < \infty) \leq \epsilon$$

Non-Markovian

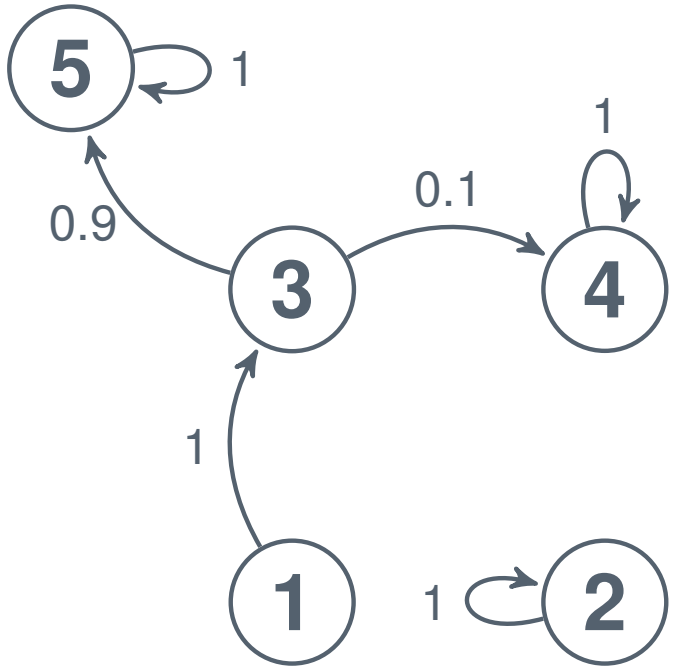
If the chain $\{X_t\}$ has already gone through the set B, then does not matter if it hits the set again.

But instead, if the chain has not gone through B it is better not to reach the set in order to satisfy the constraint.

Therefore an optimal policy may not be Markovian.

Example

$A = \{4\}, B = \{1, 2\}$.



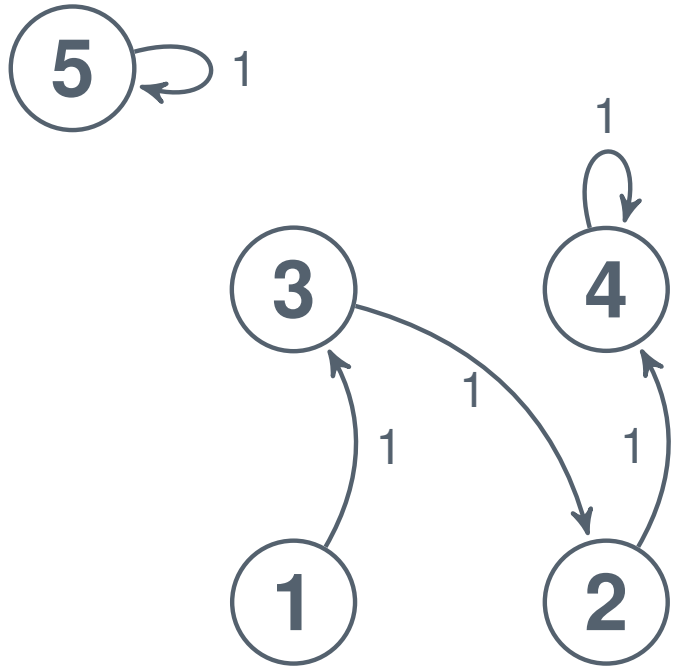


Figure: Control Matrices u_1, u_2

Lagrange Multipliers

$$\max_{\pi \in \overline{\Pi}} \left\{ P_\nu^\pi(\tau_A < \infty) \right\}$$
$$\text{s.t. } P_\nu^\pi(\tau_B < \infty) \leq \epsilon$$

The Lagrangian is:

$$\mathcal{L}(\pi, \lambda) = P_\nu^\pi(\tau_A < \infty) + \lambda \cdot (\epsilon - P_\nu^\pi(\tau_B < \infty))$$

The problem above is equivalent to:

$$\max_{\pi \in \overline{\Pi}} \left\{ \min_{\lambda \geq 0} \mathcal{L}(\pi, \lambda) \right\}$$

Dual Problem

We will approach this problem by solving the dual problem, namely,

$$\min_{\lambda \geq 0} \left\{ \max_{\pi \in \overline{\Pi}} \mathcal{L}(\pi, \lambda) \right\}$$

The idea is to write the Lagrangian as an average cost function.

To achieve this let's enlarge the state space, $\widehat{\mathbb{X}} = \mathbb{X} \times \{0, 1\}$.

Intuitively a state $(x, 0)$ indicates that the process has not reached B, while a state $(x, 1)$ indicates that the chain has already reached B.

Induced Markov Process

We define a stochastic kernel $\widehat{Q}$ on $\widehat{\mathbb{X}}$ given $\widehat{\mathbb{K}} = \widehat{\mathbb{X}} \times \mathbb{U}$ as follows:

$$\begin{cases} \widehat{Q}((y,0)|(x,0),u) = Q(y|x,u), & \text{if } y \notin B \\ \widehat{Q}((y,0)|(x,0),u) = 0, & \text{if } y \in B \\ \widehat{Q}((y,1)|(x,0),u) = 0, & \text{if } y \notin B \\ \widehat{Q}((y,1)|(x,0),u) = Q(y|x,u), & \text{if } y \in B \\ \widehat{Q}((y,0)|(x,1),u) = 0, & \\ \widehat{Q}((y,1)|(x,1),u) = Q(y|x,u). & \end{cases}$$

Lagrangian as an Average Cost Function

For any policy $\widehat{\pi}$ over $\widehat{\mathbb{X}}$, we denote by $\widehat{P}^{\widehat{\pi}}$ the measure induced by the policy for the process $\{(X_t, I_t)\}_{t\geq 0}$.

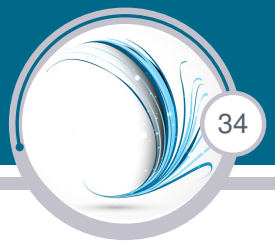
We define the average cost function

$$v^{\widehat{\pi}}(x,i) := \limsup_{N\to\infty} \frac{1}{N}\widehat{\mathbb{E}}^{\widehat{\pi}}_{(x,i)}\Big[\sum_{t=0}^{N-1}[\mathbf{1}_{A\times\{0\}} + \mathbf{1}_{A\times\{1\}} - \lambda\mathbf{1}_{\mathbb{X}\times\{1\}}](X_t, I_t)\Big]. \tag{2}$$

The Lagrangian is:

$$\mathcal{L}_x(\pi,\lambda) = P^{\pi}_x(\tau_A < \infty) + \lambda\cdot(\epsilon - P^{\pi}_x(\tau_B < \infty))$$

Lagrangian as an Average Cost Function



Theorem

Let π be a policy over $\mathbb{X}$ and $\widehat{\pi}$ be its corresponding policy over $\widehat{\mathbb{X}}$. Then,

$$\mathcal{L}_x(\pi, \lambda) = \begin{cases} v^{\widehat{\pi}}(x, 0) + \lambda\epsilon, & \text{if } x \notin B \\ v^{\widehat{\pi}}(x, 1) + \lambda\epsilon, & \text{if } x \in B. \end{cases}$$

Proof

$$v^{\widehat{\pi}}(x,i) := \limsup_{N\to\infty} \frac{1}{N} \widehat{\mathbb{E}}^{\widehat{\pi}}_{(x,i)} \Big[\sum_{t=0}^{N-1} [\mathbf{1}_{A\times\{0\}} + \mathbf{1}_{A\times\{1\}} - \lambda \mathbf{1}_{\mathbb{X}\times\{1\}}](X_t, I_t) \Big]. \tag{3}$$

Lemma

Let $\widehat{\pi}$ be a policy over $\widehat{\mathbb{X}}$. Then

$$v^{\widehat{\pi}}(x,i) = \widehat{P}^{\widehat{\pi}}_{(x,i)} \left(\tau_{A\times\{0\}} < \infty\right) + \widehat{P}^{\widehat{\pi}}_{(x,i)} \left(\tau_{A\times\{1\}} < \infty\right) - \lambda \widehat{P}^{\widehat{\pi}}_{(x,i)} \left(\tau_{X\times\{1\}} < \infty\right).$$

Proof

$$v^{\widehat{\pi}}(x,i) = \widehat{P}^{\widehat{\pi}}_{(x,i)}\left(\tau_{A\times\{0\}} < \infty\right) + \widehat{P}^{\widehat{\pi}}_{(x,i)}\left(\tau_{A\times\{1\}} < \infty\right) - \lambda\widehat{P}^{\widehat{\pi}}_{(x,i)}\left(\tau_{X\times\{1\}} < \infty\right).$$

If $x \notin B$, then

$$\widehat{P}^{\widehat{\pi}}_{(x,0)}(\tau_{X\times\{1\}} < \infty) = P^{\pi}_{x}(\tau_{B} < \infty).$$

If $x \notin B$, then

$$\widehat{P}^{\widehat{\pi}}_{(x,0)}(\tau_{A\times\{1\}} < \infty) = P^{\pi}_{x}(\tau_{A} < \infty, \tau_{B} < \tau_{A}).$$

If $x \notin B$, then

$$\widehat{P}^{\widehat{\pi}}_{(x,0)}(\tau_{A\times\{0\}} < \infty) = P^{\pi}_{x}(\tau_{A} < \infty, \tau_{A} < \tau_{B}).$$

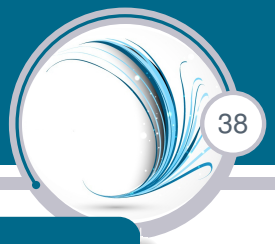
Proof

$$v^{\widehat{\pi}}(x,i) = \widehat{P}^{\widehat{\pi}}_{(x,i)}\left(\tau_{A\times\{0\}} < \infty\right) + \widehat{P}^{\widehat{\pi}}_{(x,i)}\left(\tau_{A\times\{1\}} < \infty\right) - \lambda\widehat{P}^{\widehat{\pi}}_{(x,i)}\left(\tau_{X\times\{1\}} < \infty\right).$$

If $x \in B$, then

$$v^{\widehat{\pi}}(x,1) = P^{\pi}_{x}(\tau_A < \infty) - \lambda$$

Reaching with Constraints



Consequence

- For each $\lambda \geq 0$ the problem

$$\max_{\pi \in \overline{\Pi}} \mathcal{L}(\pi, \lambda).$$

  can be solved using linear programming.
- We can approximate the dual of (P2)

$$\min_{\lambda \geq 0} \left\{ \max_{\pi \in \overline{\Pi}} \mathcal{L}(\pi, \lambda) \right\}$$

Theorem

Suppose $\mathbb{X}, \mathbb{U}$ are finite. (P2) satisfies strong duality, that is,

$$\min_{\lambda \geq 0} \left\{ \max_{\pi \in \overline{\Pi}} \mathcal{L}(\pi, \lambda) \right\} = \max_{\pi \in \overline{\Pi}} \left\{ \min_{\lambda \geq 0} \mathcal{L}(\pi, \lambda) \right\}$$

Future Work

- Use the discounted linear program to obtain strong duality in (P2).
- To solve the linear programs the distributions are required, is there a way to produce a robust method?
- Ergodic Theory?
- Use polynomial optimization in an infinite dimensional setting.

References

- On controllability of Markov chains: A Markov decision process approach, Ávila, Daniel and Junca, Mauricio (In progress)
- Maximizing the probability of attaining a target prior to extinction, Chatterjee, Debasish and Cinquemani, Eugenio and Lygeros, John.
- Markov decision processes: discrete stochastic dynamic programming, Puterman, Martin L.
- Discrete-time Markov control processes: basic optimality criteria, Hernández-Lerma, Onésimo and Lasserre, Jean Bernard.

Gracias!!