CONTROLLED MARKOV CHAINS: SOME STABILITY PROBLEMS

by

Daniel Felipe Ávila Girardot

Thesis advisor

Ph.D Mauricio Junca

A thesis presented for the degree of
Master in Mathematics

Departamento de Matemáticas
Facultad de Ciencias
Universidad de los Andes
Colombia
2016

Acknowledgements

First of all, I would like to express my deep gratitude to my advisor, Mauricio Junca. He always offered his time, in order to help me and provide me guidance along the development of this work. Without him the completion of this thesis would have been impossible.

Secondly, I would like to thank my classmates and friends: Santiago Pinzón, Andrés Galindo, Jerson Caro, Hernan García, Julian Forero, Gustavo Chaparro, Rodolfo Quintero, Andrés Garcia, Sergio Gil, Andrei Fuentes and Carlos Umana. Who gave me their support along the development of this work.

Finally, I would like to thank my family for supporting me along all this years of study, and in particular for encouraging me to complete my masters studies.

Abstract

Along this work we study some stability-related problems in the context of Controlled Markov chains. As a first problem, we consider a division of the state space, and the goal is to construct a control policy such that the chain stabilize as much as possible in each portion of the division without leaving it.

As a second problem, we characterize the *domain of attraction* and *escape set* of a controlled Markov chain via a function v, which happens to be the solution of a Bellman's equation. The interpretation of v as the solution of a Bellman's equation also provides a way to calculate such function via a linear program.

Finally, under some assumptions, we find a policy that maximize the probability of *reaching* certain set A. Our approach uses certain cost functions and dynamical programming, so it can be solved using a linear program.

Contents

Introduction

Controlled Markov Chains provide a mathematical frame for dynamical systems where outcomes contain certain ambiguity. Such models have been used on inventory control problems, communication models and even on biology models. The goal of this work is to provide control policies that guarantee some kind of *stabilization* for the system.

In [1, 2] the authors treat the problem of finding control policies such that the induced Markov chain has a maximum number of recurrent states, so the chain will stabilize as much as possible. Their approach uses the concept of *entropy* and state a convex program to find such policies. The first problem we consider is to find a policy in such a way that given any partition of the state space, the chain has a maximum number of recurrent states that can't evolve to states that belong to another piece of the partition. Our Approach follows the ideas from those papers and a kind of Lyapunov function.

Following the idea of stabilize the system, what if instead of stabilizing the chain as much as possible, we want to ensure that the chain will eventually reach some desired set? Stated more rigorously, let ν be an initial distribution over the state space. Our objective will be to find a policy π that maximize the quantity,

$$\liminf_{t\to\infty} P_\nu^\pi(X_t \in A)$$

With such problem in mind, we first studied Zubov's method. This method, used in deterministic dynamical systems, allows us to describe the domain of attraction of a stable point (see [7, 11]); that is, allows us to compute the initial states, under which the system will approach a stable point. In this work, we have developed a Zubov's method for controlled Markov chains. Using a function v, which is the solution of a Bellman's equation, we characterize the *domain of attraction* and *escape set*. Moreover, the calculation of v provides a policy such that any state of the domain of attraction has positive probability of reaching A.

Finally, we studied the relation between Zubov's method and the proposed problem. Such method provides a lower bound for the desired quantity. However, using a variation of the function v it's possible to obtain an exact answer. Nevertheless, the Bellman's equations for such variation are more complicated.

Chapter 1

Preliminaries

The objective of this chapter is to introduce the terminology, notation and some important results that will be used along this work. The first section is dedicated to Markov chains. Of particular interest is going to be the discussion concerning invariant distributions, since it will be fundamental in Chapter 2. In second section, controlled Markov chains are going to be treated, and the definition of a *control policy* will be given. Last section is dedicated to a particular case of controlled Markov chains, where a reward function is used.

1.1 Markov Chains

Let $\mathbb{X}$ be a countable set and let $\{X_t\}_{t=0}$ be a $\mathbb{X}$-valued stochastic process defined over a probability space $(\Omega, \mathcal{F}, P)$. We say that the process is a *Markov chain* if for all $k \in \mathbb{N}$

$$P(X_{k+1} = j_{k+1} | X_k = j_k, \ldots, X_0 = j_0) = P(X_{k+1} = j_{k+1} | X_k = j_k).$$

In case $P(X_{k+1} = j | X_k = i)$ does not depend of time, that is on k, we say the Markov chain is *time homogeneous*, in such case we write $P(X_{k+1} = j | X_k = i) = P_{ij}$. In order to simplify notation we will write $P_i(\cdot) := P(\cdot | X_0 = i)$.

Definition 1.1. Let $A \subset \mathbb{X}$, the random variable $\tau_A = \inf\{n \geq 0 \mid X_n \in A\}$ is called the hitting time of A.

With such definition in mind we can decompose the state space via an equivalence relation as follows. First of all we need a way to relate states.

Definition 1.2. Let $i, j \in \mathbb{X}$. If $P_i(\tau_j < \infty) > 0$ we say j is accessible from i, denoted as $i \to j$. Moreover, if $i \to j$ and $j \to i$ we say states i and j communicate, denoted as $i \leftrightarrow j$.

The communication relation just defined is an equivalence relation, see [9] for details, as a consequence we can decompose the state space into disjoint equivalence classes. Using this, we will be able to decompose the state space in a way that will be helpful to understand the long time behavior of the chain. First of all we need a notion that captures such behavior.

Definition 1.3. Given $i \in \mathbb{X}$ define $\tau_i(1) = \inf\{n \geq 1 \mid X_n = i\}$. If $P_i(\tau_i(1) < \infty) = 1$ the state i is called recurrent, otherwise is called transient.

The following lemma contains an important fact, that illustrates why we expect the chain will stabilize in recurrent states.

Lemma 1.4. *Let j be a transient state, then for all $x \in \mathbb{X}$,*

$$\lim_{t\to\infty} P_x(X_t = j) = 0$$

We need another definition that will be fundamental in our work.

Definition 1.5. Let $A \subset \mathbb{X}$. We say A is closed if for all $x \in A$,

$$P_x(\tau_{A^c} < \infty) = 0$$

We give now an important characterization of closed sets.

Lemma 1.6. *A set A is closed if and only if $P_{x,i} = 0$ for all $x \in A, i \in A^c$.*

Proof. Suppose A is closed and let $x \in A, i \in A^c$. The event $\{X_1 = i\}$ is contained in $\{\tau_{A^c} < \infty\}$, so $P_{x,i} \leq P_x(\tau_{A^c} < \infty) = 0$. On the other hand, assume $P_{x,i} = 0$ for all $x \in A, i \in A^c$. Let $n \in \mathbb{N}$, given $x \in A, i \in A^c$ we have that,

$$\begin{aligned} P_x(X_n = i) &= \sum_{i_1\in\mathbb{X},\ldots,i_{n-1}\in\mathbb{X}} P_{x,i_1}\ldots P_{i_{n-1},i} \\ &= \sum_{i_1\in A,\ldots,i_{n-1}\in A} P_{x,i_1}\ldots P_{i_{n-1},i} \\ &= 0 \end{aligned}$$

where the last equality follows from the fact that $i_{n-1} \in A, i \in A^c$ implying $P_{i_{n-1},i} = 0$. Finally, since,

$$\{\tau_{A^c} < \infty\} \subset \bigcup_n \{X_n \in A^c\},$$

then,

$$P_x(\tau_{A^c} < \infty) \leq \sum_n P_x(X_n \in A^c) = 0$$

where the last equality follows from above calculation. □

So intuitively, a closed set is such that if the chain enters in it, it will never escape from it. The state space decomposition is the following, for a proof and further details see [9].

Proposition 1.7. *The state space $\mathbb{X}$ can be decomposed as follows,*

$$\mathbb{X} = T \cup (\cup_i C_i)$$

where T is a set of transient states (not necessarily a class), and the $C_i's$ are closed, disjoint classes of recurrent states. Moreover, reordering of the state space, the transition matrix P can be written as follows,

$$P = \begin{pmatrix} P_1 & 0 & 0 & \dots \\ 0 & P_2 & 0 & \dots \\ \vdots & \vdots & \ddots & \\ Q_1 & Q_2 & \cdots & \cdots \end{pmatrix}$$

where P_i are stochastic matrices for C_i. Furthermore, in case the state space is finite such decomposition can be accomplished as follows,

1. *Decompose $\mathbb{X}$ into equivalence classes.*
2. *The closed classes are recurrent.*
3. *Classes which are not closed consist of transient states.*

Therefore, recurrent states belong to closed classes, and in case the state space is finite, all closed classes will consist of recurrent states. So in some sense, the chain will stabilize in such closed classes. As a consequence, an interesting question is how to characterize those sets. Invariant distributions, help to answer such question.

Definition 1.8. Let ν be a probability distribution over the state space. It's called an invariant or stationary distribution if,

$$\nu^T = \nu^T P$$

The relation between invariant distributions and closed classes is the following, see [9].

Proposition 1.9. *Suppose the chain is irreducible (there is only one class), recurrent and that $\mathbb{X}$ is finite. Then, there exist an unique invariant distribution ν. Moreover, $\nu(x) > 0$ for all $x \in \mathbb{X}$.*

Remark. Note that above proposition implies that whenever all recurrent classes are finite, then there exist an unique invariant distribution fully supported in each recurrent class

As a corollary, we obtain an important theorem, which states that invariant distributions are a convex combinations of invariant distributions corresponding to recurrent classes.

Theorem 1.10. *Let $\mathbb{X}$ be a countable state space, and decompose the state space as $\mathbb{X} = T \cup (\cup_i C_i)$, where T is a set of transient states, and the $C_i's$ are closed, disjoint classes of recurrent states. Suppose all $C_i's$ are finite and let ν_i be the invariant distribution associated to C_i. Then, any invariant distribution ν satisfies,*

$$\nu = \sum_i \alpha_i \nu_i, \qquad \text{where } \sum_i \alpha_i = 1.$$

Proof. Let ν be an invariant distribution. Therefore, for all $t \in \mathbb{N}$,

$$\nu(j) = \sum_{x \in \mathbb{X}} v(x) P_x(X_t = j)$$

Let j be a transient state. So taking limit in above expression, and recalling that,

$$\lim_{t\to\infty} P_x(X_t = j) = 0$$

we obtain, $\nu(j) = 0$. As a consequence, $\nu(x)$ must be supported on the $C_i's$. Recalling that states can be reorganized to obtain,

$$P = \begin{pmatrix} P_1 & 0 & 0 & \dots & 0 \\ 0 & P_2 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & & \\ Q_1 & Q_2 & \cdots & \cdots & \end{pmatrix}$$

we note $\nu^T = \nu^T P$ implies $\nu^T \restriction_{C_i} = \nu^T \restriction_{C_i} P_i$. And because of uniqueness of ν_i we conclude $\nu^T \restriction_{C_i} = \alpha_i \nu_i$ for some α_i. Finally, evaluating last expression in C_i we obtain $\alpha_i = \nu(C_i)$. □

To end this section, we present some useful lemmas. Recall that given a sequence of events $\{E_n \mid n \in \mathbb{N}\}$ we can define the events,

$$\liminf_{n\to\infty} E_n := \bigcup_m \bigcap_{n\geq m} E_n$$
$$\limsup_{n\to\infty} E_n := \bigcap_m \bigcup_{n\geq m} E_n$$

Lemma 1.11. *(see [10]) (Fatou's Lemma for sets)*

$$P\left(\liminf_{n\to\infty} E_n\right) \leq \liminf_{n\to\infty} P(E_n)$$
$$P\left(\limsup_{n\to\infty} E_n\right) \geq \limsup_{n\to\infty} P(E_n)$$

Lemma 1.12. *Let $A \subset \mathbb{X}$ and consider the hitting time of A. Then,*

$$\liminf_{n\to\infty}\{\tau_A \leq n\} = \{\tau_A < \infty\} = \limsup_{n\to\infty}\{\tau_A \leq n\}.$$

Proof. Let's prove the first equality. Let $\omega \in \liminf\{\tau_A \leq n\}$, so there exists $m \in \mathbb{N}$ such that for all $n \geq m$, $\omega \in \{\tau_A \leq n\}$; that is, for all $n \geq m$, $\tau_A(\omega) \leq n$, which implies $\omega \in \{\tau_A < \infty\}$. Now, let $\omega \in \{\tau_A < \infty\}$, so defining $m := \tau_A(\omega)$ we obtain $\omega \in \bigcap_{n\geq m}\{\tau_A \leq n\}$, which implies $\omega \in \liminf\{\tau_A \leq n\}$.

For the second equality, let $\omega \in \{\tau_A < \infty\}$ and let $m \in \mathbb{N}$. Choose i big enough so that $i \geq m$ and $i \geq \tau_A(\omega)$. Therefore, $\omega \in \bigcup_{n\geq m}\{\tau_A \leq n\}$ and since m was arbitrary, $\omega \in \limsup\{\tau_A \leq n\}$. Now, let $\omega \in \limsup\{\tau_A \leq n\}$ so for all $m \in \mathbb{N}$ there exists $n \geq m$ such that $\omega \in \{\tau_A \leq n\}$; that is, $\tau_A(\omega) \leq n < \infty$ so $\omega \in \{\tau_A < \infty\}$. □

Lemma 1.13. *Let A be a closed set and consider the event*

$$E = \{\exists n, m \in \mathbb{N} \text{ such that } n \leq m, X_n \in A, X_m \notin A\}$$

Then, $P_x(E) = 0$ for any $x \in \mathbb{X}$.

Proof. We can write such event as $E = \bigcup_n \bigcup_{m \geq n} \{X_n \in A, X_m \notin A\}$. Therefore,

$$P_x(E) \leq \sum_n \sum_{m \geq n} P_x(X_n \in A, X_m \notin A)$$

Let's see that for all $m \geq n$, $P_x(X_n \in A, X_m \notin A) = 0$. We have that,

$$P_x(X_n \in A, X_m \notin A) = \sum_{j \in A} \sum_{i \notin A} P_x(X_n = j, X_m = i)$$

Now, each term $P_x(X_n = j, X_m = i) = P_x(X_n = j)P_j(X_{m-n} = i)$. Therefore, since A is closed we have $P_j(X_{m-n} = i) = 0$, so that $P_x(X_n \in A, X_{m-n} \notin A) = 0$ and as a consequence $P_x(E) = 0$. □

1.2 Controlled Markov Chains

Controlled Markov chains or Markov decision processes, provide a mathematical scheme for problems where the outcomes contain certain ambiguity. In this section we will shortly introduce the fundamental aspects of controlled Markov Chains in the discrete context, a reference for such topic can be found in [8]. The ideas for the non-discrete context are quite similar however some mathematical issues appear, the interested reader can deepen in [4].

Let $\{(X_t, U_t)\}_{t=0}$ be a stochastic process over a countable set $\mathbb{X} \times \mathbb{U}$. The set $\mathbb{X}$ is the state space of the system, while $\mathbb{U}$ is the set of control actions. Such process is a controlled Markov chain if for all $k \in \mathbb{N}$, $i_{[0,k]} \subset \mathbb{X}$, $u_{[0,k]} \subset \mathbb{U}$,

$$P(X_{k+1} = j | X_{[0,k]} = i_{[0,k]}, U_{[0,k]} = u_{[0,k]}) = \mathbb{T}(X_{k+1} = j | X_k = i_k, U_k = u_k),$$

where $\mathbb{T}$ satisfies, $\sum_{j \in \mathbb{X}} \mathbb{T}(X_{k+1} = j | X_k = i_k, U_k = u_k) = 1$. So the dynamics of the process are defined by the probability of X_{k+1} given the previous state X_k and a control action U_k. Since we will only use homogeneous Markov chains, the following notation makes sense,

$$P_{ij}^u := P(X_{k+1} = j | X_k = i, U_k = u) \text{ where } i, j \in \mathbb{X}, u \in \mathbb{U}.$$

We are going to be interested in finding control policies, or *strategies*, such that the evolution of the process has certain desired property. We will only deal with fully observed controlled Markov chains, which means that for each instant k the controller observes exactly the state x_k. Before giving a precise definition of a control policy we need the following notation, denote $\mathbb{U}(x) \subset \mathbb{U}$ as the set of feasible actions on the state $x \in \mathbb{X}$, and let $\mathbb{K} = \{(x, u) | x \in \mathbb{X}, u \in \mathbb{U}(x)\}$.

Definition 1.14. Let $P(\mathbb{U})$ be the set of probability measures over $\mathbb{U}$. A control policy π is a sequence of functions $(\mu_0, \mu_1, \mu_2, \ldots)$, where each $\mu_k : \mathbb{K}^{k-1} \times \mathbb{X} \to P(\mathbb{U})$ and,

$$\sum_{u \in \mathbb{U}(x_k)} \mu_k(u | x_0, u_0, \ldots, x_{k-1}, u_{k-1}, x_k) = 1$$

for all $x_0, u_0, \ldots, x_{k-1}, u_{k-1}, x_k \in \mathbb{K}^{k-1} \times \mathbb{X}$.

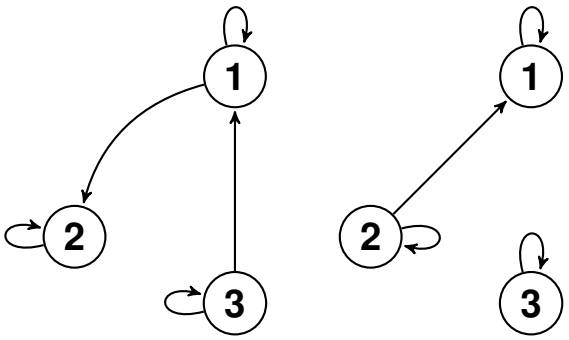


Figure 1.1: Graph representations of transition matrices P^1 and P^2.

Denote Π as the set of all control policies, we have different types of control policies,

1. We say $\pi = (\mu_0, \mu_1, \mu_2, \ldots)$ is *deterministic* if for any $k \in \mathbb{N}$ and any sequence $x_0, u_0, \ldots, x_{k-1}, u_{k-1}, x_k \in \mathbb{K}^{k-1} \times \mathbb{X}$, the measure $\mu_k(.|x_0, u_0, \ldots, x_{k-1}, u_{k-1}, x_k)$ is zero for all but one element of $\mathbb{U}(x_k)$. Note that in such case we can regard μ_k just as a function from $\mathbb{K}^{k-1} \times \mathbb{X}$ to $\mathbb{U}$.

2. We say $\pi = (\mu_0, \mu_1, \mu_2, \ldots)$ is a *Markov policy* if μ_k it doesn't depend on the history; that is, only depends on the current state x_k. Denote $\Pi_M \subset \Pi$ the set of such policies.

3. If π is a Markov policy such that $\pi = (\mu, \mu, \mu, \ldots)$ we say it's a *stationary policy*. Denote $\Pi_S \subset \Pi_M$ the set of such policies.

As we will see in the next section Markov deterministic policies will play an important role on several optimization problems, in particular deterministic stationary policies will be fundamental in infinite horizon problems. We have one first easily proven lemma that tells us that Markov policies are in some sense well-behaved.

Lemma 1.15. *Let $\pi = (\mu_0, \mu_1, \mu_2, \ldots)$ be a Markov policy, the transition probabilities of the Markov chain of the induced probability measure P^π can be computed as,*

$$P^\pi(X_{k+1} = i | X_k = x) = \sum_{u \in \mathbb{U}(x)} \mu_k(u|x) \cdot P(X_{k+1} = i | X_k = x, U_k = u)$$

Moreover, if π is stationary we obtain a homogeneous Markov chain.

Example 1.16. Let $\mathbb{X} = \{1, 2, 3\}$ and $\mathbb{U} = \{1, 2\}$. For each control action we have the transition matrices,

$$P^1 = \begin{bmatrix} 0.3 & 0.7 & 0 \\ 0 & 1 & 0 \\ 0.8 & 0 & 0.2 \end{bmatrix} \qquad P^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0.8 & 0.2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

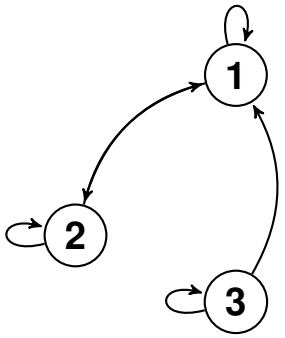
We can represent such matrices as graphs as is shown in figure 1.1. Consider the stationary policy $\pi = (\mu, \mu, \mu, ...)$,

$$\mu = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0.2 & 0.8 \end{bmatrix}$$

So the transition matrix P^π is given by,

$$P^\pi = \begin{bmatrix} 0.3 & 0.7 & 0 \\ 0.8 & 0.2 & 0 \\ 0.16 & 0 & 0.84 \end{bmatrix}$$

As a consequence the induced Markov chain has one recurrent class $\{1, 2\}$ and one transient class $\{3\}$. Note that $\{1, 2\}$ was not even a class with P^1 or P^2.



Graph representation of P^π

In the next section we will briefly discuss how these policies are related, and in particular how they behave with the maximization of costs functions.

1.3 Dynamic programming

In the following section we will discuss the optimization of several cost functions. First of all we need a *reward function*. For all $t \in \mathbb{N}$, let r_t be a function from $\mathbb{X} \times \mathbb{U}$ to $\mathbb{R}$. The objective of the function $r_t(x, u)$ is to define a cost per stage, a cost for, in step or time t, being in state x and selecting a control u.

Relying on the kind of problem there are different cost functions involved. In finite horizon problems, for selecting a policy, the controller gets rewards in the periods of time $1, \ldots, N$. The objective is to find a policy that minimizes or maximizes such rewards. More precisely, given an initial state x we are interested in finding a policy $\pi = (\mu_0, \mu_1, \ldots)$, that minimizes the cost function

$$v_N^\pi(x) := \mathbb{E}_x^\pi \Big[\sum_{t=0}^{N-1} r_t(X_t, U_t) + r_N(X_N) \Big]$$

Since we are going to be interested in the long time behaviour of the chain, this model will not be of interest, for further details of such model see [8] or [3].

On the other hand, on infinite horizon problems there is not finite horizon planning. We will be interested in two kinds of infinite horizon problems: the discounted case and average case. In the first case, the controller receives a penalization $0 < \gamma < 1$ for each period of time. So given an initial state x, we will be interested in finding a policy π that maximizes the cost function,

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

In the average case, we will be interested in finding a policy that maximizes the cost function,

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

On what follows we will briefly discuss whether there exists a solution for the previously presented problems, and computational approaches to solve them.

1.3.1 Discounted Problem

To treat this problem we will make some assumptions,

1. Assume there are stationary rewards $r(x, u)$.
2. Assume there are bounded rewards.
3. Assume the discount factor $\gamma \in (0, 1)$.

Under those hypothesis and using the dominated convergence theorem it's clear that,

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

We are interested in finding a policy π^* such that $v_\gamma^{\pi*} = v_\gamma^*$, where

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

There is one first interesting result, which states that for each $x \in \mathbb{X}$ and any $\pi \in \Pi$ there exists $\pi' \in \Pi_M$ such that $v_\gamma^\pi(x) = v_\gamma^{\pi'}(x)$, see [8] for instance. As a consequence,

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

The existence of an optimal policy is not a trivial issue. Such existence relays on the Bellman's equations, that are defined as follows,

Definition 1.17. The system of equations,

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

are called the Bellman's equation for the discounted system. Denote D as the set of deterministic stationary policies, so Bellman's equations can be expressed in vector notation as,

$$v_\gamma = \sup_{\pi \in D} \left\{ r^\pi + \gamma \cdot P^\pi \cdot v_\gamma \right\}$$

To get a sense of why such equations make sense, let's fix a stationary policy π. Given $x \in \mathbb{X}$ we have that,

$$\begin{aligned}
v^\pi_\gamma(x) &= \mathbb{E}^\pi_x \left[\sum_{t=0}^{\infty} \gamma^t \cdot r(X_t, U_t) \right] \\
&= \mathbb{E}^\pi_x \left[r(X_0, U_0) \right] + \mathbb{E}^\pi_x \left[\sum_{t=1}^{\infty} \gamma^t \cdot r(X_t, U_t) \right] \\
&= \mathbb{E}^\pi_x \left[r(X_0, U_0) \right] + \sum_{j \in \mathbb{X}} \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^t \cdot r(X_t, U_t) \Big| X_1 = j, X_0 = x \right] \cdot P^\pi_{x,j} \\
&= \mathbb{E}^\pi_x \left[r(X_0, U_0) \right] + \gamma \sum_{j \in \mathbb{X}} \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} \cdot r(X_t, U_t) \Big| X_1 = j \right] \cdot P^\pi_{x,j} \\
&= \sum_{u \in \mathbb{U}(x)} r(x,u) \cdot \pi(u|x) + \gamma \sum_{j \in \mathbb{X}} v^\pi_\gamma(j) \cdot P^\pi_{x,j} \\
&= \sum_{u \in \mathbb{U}(x)} \pi(u|x) \cdot \left[r(x,u) + \gamma \sum_{j \in \mathbb{X}} v^\pi_\gamma(j) \cdot P^u_{x,j} \right] \\
&\leq v_\gamma(x)
\end{aligned}$$

So for stationary policies we have $\sup_{\pi \in \Pi_S} v^\pi_\gamma(x) \leq v_\gamma(x)$. In general, we would like to prove $v_\gamma = v^*_\gamma$. To achieve this it's needed to express Bellman's equations in operator notation. Let V denote the space of bounded real-valued functions on $\mathbb{X}$ doted with the supremum norm, and with the partial order relation,

$$\text{Given } u, v \in V \text{ we say } u \geq v \Leftrightarrow u(x) \geq v(x) \text{ for all } x \in \mathbb{X}$$

It's possible to prove that if $v \in V$ then $\sup_{\pi \in D} \left\{ r^\pi + \gamma \cdot P^\pi \cdot v \right\} \in V$. Therefore, makes sense to define an operator $\mathbb{L} : V \to V$,

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

The importance of such operator relies on the following lemma (for details see [8]).

Lemma 1.18. *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* $v = v^*_\gamma$.

Furthermore, the solution v_γ of Bellman's equations is a fixed point of $\mathbb{L}$, so last item of previous lemma is telling us that in case v_γ exists, then $v_\gamma = v^*_\gamma$. To conclude the existence of a solution of Bellman's equation, theorems as the Banach Fixed Point theorem can be applied. Those equations also provide a way to conclude the existence of a policy π such that $v^\pi_\gamma = v^*_\gamma$. To sum up, we have the following, for a proof and further details see [8] or [3].

Theorem 1.19. *1. There exist a solution of Bellman's equation.*

2. *The solution v_γ of Bellman's equations satisfies $v_\gamma = v^*_\gamma$.*
3. *Suppose the state space $\mathbb{X}$ is either countable or finite, and $\mathbb{U}(x)$ is finite for all $x \in \mathbb{X}$. Then there exist a stationary deterministic policy π such that $v^\pi_\gamma = v^*_\gamma$.*
4. *Such optimal policy can be found defining $\pi = (\mu, \mu, \ldots)$ where*

$$\mu(x) \in \arg\max_{u \in \mathbb{U}(x)} \{r(x,u) + \gamma \sum_{j \in \mathbb{X}} P^u_{xj} \cdot v_\gamma(j)\}$$

Remark. There are more general assumptions that can be used in item 3. Since we are not using them on the present work we will not discuss them.

To end this section we present a way to solve such equations via linear programming. Recall that because of the first item of previous lemma, if $v \in V$ satisfies,

$$v \geq \sup_{\pi \in D} \left\{ r^\pi + \gamma \cdot P^\pi \cdot v \right\}$$

then v provides an upper bound for the optimal solution v^*_γ. Therefore, makes sense to approach the optimal solution by the *least large* v that satisfies $v \geq \mathbb{L}(v)$. However, the restriction $v \geq \mathbb{L}(v)$ is non-linear, fortunately it can be can be written as the linear restrictions,

$$v(x) - \sum_{j \in \mathbb{X}} \gamma P^u_{x,j} v(j) \geq r(x,u) \;\text{ for } u \in \mathbb{U}(x), x \in \mathbb{X}$$

The linear program is then as follows. Consider a vector $\nu \in \mathbb{R}^{|\mathbb{X}|}_{\geq 0}$, in case $\sum_{x \in \mathbb{X}} \nu(x) = 1$ we can think of ν as a distribution over the state space.

Primal Linear Program

$$\text{Minimize } \sum_{x\in\mathbb{X}} \nu(x)v_\gamma(x)$$

Subject to

$$v_\gamma(x) - \sum_{j\in\mathbb{X}} \gamma P^u_{x,j} v_\gamma(j) \geq r(x,u) \qquad \text{for } u \in \mathbb{U}(x), x \in \mathbb{X}$$

However, in order to reconstruct the control is more appropriate to consider the dual linear program,

Dual Linear Program

$$\text{Maximize } \sum_{s\in\mathbb{X}} \sum_{u\in\mathbb{U}(s)} r(s,u)x(s,u)$$

Subject to

$$\sum_{u\in\mathbb{U}(j)} x(j,u) - \sum_{s\in\mathbb{X}} \sum_{u\in\mathbb{U}(s)} \gamma P^u_{s,j} x(s,u) = \nu(j)$$

$$x(j,u) \geq 0 \quad \text{for } u \in \mathbb{U}(j), j \in \mathbb{X}$$

Finally, we have the following theorem that ensures that solving above program is equivalent to solving Bellman's equation.

Theorem 1.20. *1. There exist an optimal solution x^* of the dual linear program.*

2. Define a stationary policy $\pi = (\mu, \mu, \dots)$ as,

$$\mu(u^+|s) = \frac{x^*(s,u^+)}{\sum_{u\in\mathbb{U}(s)} x^*(s,u)}$$

Then, v^π_γ is a solution to Bellman's equations.

1.3.2 Average Problem

We need some assumptions,

1. Assume there are stationary rewards $r(x,u)$.
2. Assume there are bounded rewards.

As before, we are going to be interested in finding a policy π^* such that $v^{\pi^*} = v^*$, where,

$$v^*(x) = \sup_{\pi\in\Pi} v^\pi(x) \qquad v^\pi(x) = \limsup_{N\to\infty} \frac{1}{N} \cdot \mathbb{E}^\pi_x\Big[\sum_{t=0}^{N-1} r_t(X_t, U_t)\Big] \quad x \in \mathbb{X}$$

First of all, there is an interesting result relating discounted cost functions with average cost functions (see [8]).

Lemma 1.21. *Let π be a stationary policy, and let v_γ^π be the discounted cost function for some $\gamma \in (0,1)$. Then,*

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

As is shown in [8], for each $x \in \mathbb{X}$ and $\pi \in \Pi$ there exists a Markovian policy π' such that $v^\pi = v^{\pi'}$. So as in the discounted case, it's enough to just consider policies in Π_M. Following the ideas of the discounted case, one would like to define some kind of Bellman's equation that captures the desired cost function.

Definition 1.22. The system of equations,

$$\sup_{u \in \mathbb{U}(x)} \Big\{ \sum_{j \in \mathbb{X}} P_{x,j}^u v(j) - v(x) \Big\} = 0$$

And

$$\sup_{u \in \mathbb{U}(x)} \Big\{ r(x,u) - v(x) + \sum_{j \in \mathbb{X}} P_{x,j}^u h(j) - h(x) \Big\} = 0$$

for $x \in \mathbb{X}$. Are called the optimality equations for the multi-chain model.

Remark. Multi-chain models are those where exist a stationary policy for which the induced Markov chain has at least two recurrent classes. Uni-chain models are easier, in such models there is just one recurrent class, and the v function is constant so there is no need to introduce the first equation.

We have the following theorem that relates the solution of above equations with the average cost function, for a proof see [8].

Theorem 1.23.

- *Suppose above equations have a solution (v, h), then, $v = v^*$.*
- *Suppose the state space $\mathbb{X}$ is finite, and $\mathbb{U}(x)$ is finite for all $x \in \mathbb{X}$. Then, the optimality equations have solution.*
- *Suppose the state space $\mathbb{X}$ is finite, and $\mathbb{U}(x)$ is finite for all $x \in \mathbb{X}$. Then there exist a stationary deterministic policy π such that $v^\pi = v^*$.*
- *Such optimal policy can de found defining $\pi = (\mu, \mu, \ldots)$ where*

$$\mu(x) \in \arg\max_{u \in \mathbb{U}(x)} \{ r(x,u) + \sum_{j \in \mathbb{X}} P_{xj}^u \cdot h(j) \}$$

with the additional condition that $P^\pi v = v$.

Those equations can be solved via linear programming. Consider a vector $\nu \in \mathbb{R}_{\geq 0}^{|\mathbb{X}|}$. The linear program is as follows, see [8] for further details,

Primal Linear Program

$$\text{Minimize } \sum_{x\in\mathbb{X}} \nu(x)v(x)$$

Subject to

$$v(x) \geq \sum_{j\in\mathbb{X}} P^u_{x,j} v(j) \qquad \text{for } u \in \mathbb{U}(x), x \in \mathbb{X}$$

And

$$v(x) \geq r(x,u) + \sum_{j\in\mathbb{X}} P^u_{x,j} h(j) - h(x) \qquad \text{for } u \in \mathbb{U}(x), x \in \mathbb{X}$$

And the dual linear program is given by,

Dual Linear Program

$$\text{Maximize } \sum_{s\in\mathbb{X}} \sum_{u\in\mathbb{U}(s)} r(s,u)x(s,u)$$

Subject to

$$\sum_{u\in\mathbb{U}(j)} x(j,u) - \sum_{s\in\mathbb{X}} \sum_{u\in\mathbb{U}(s)} P^u_{s,j} x(s,u) = 0 \qquad j \in \mathbb{X}$$

And

$$\sum_{u\in\mathbb{U}(j)} x(j,u) + \sum_{u\in\mathbb{U}(j)} y(j,u) - \sum_{s\in\mathbb{X}} \sum_{u\in\mathbb{U}(s)} P^u_{s,j} y(s,u) = \nu(j) \qquad j \in \mathbb{X}$$

$$\text{And } x(j,u) \geq 0, \quad y(j,u) \geq 0 \qquad u \in \mathbb{U}(j), j \in \mathbb{X}$$

The theorem relating above program with multi-chain optimality equations is the following,

Theorem 1.24. *1. There exist an optimal solution (x^*, y^*) of the dual linear program.*

2. Define a stationary policy $\pi = (\mu, \mu, \ldots)$ as,

$$\mu(u^+|s) = \begin{cases} \frac{x^*(s,u^+)}{\sum_{u\in\mathbb{U}(s)} x^*(s,u)} & \text{if } \sum_{u\in\mathbb{U}(s)} x^*(s,u) > 0 \\ \\ \frac{y^*(s,u^+)}{\sum_{u\in\mathbb{U}(s)} y^*(s,u)} & \text{otherwise} \end{cases}$$

Then, $v^\pi = v^*$.

Chapter 2

Recurrent States and Entropy

In this chapter we will study the relation between the entropy function and recurrent states. In [1] the authors treat the problem of finding a control that maximizes the number of recurrent states, in such a way that the recurrent states must avoid certain set of forbidden states $\mathbb{F}$. In a most recent paper [2] the authors generalize such methodology. They treat the problem of finding a control that maximizes the number of recurrent states, subject to a set of convex constraints. Using those ideas we constructed a control such that given any partition of the state space, the control finds the maximum number of recurrent states on each portion of the partition, avoiding states that belong to another piece of the partition. This chapter is organized as follows: First section will be dedicated to briefly discuss the algorithm proposed in [1], while in second section we will treat the proposed problem.

Let $\mathbb{X}$ be a finite state space and $\mathbb{U}$ be a finite set of control actions. Recall that each control u has associated a transition matrix P^u, we will also assume that all control actions are feasible for each state x, that is $\mathbb{U}(x) = \mathbb{U}$. Since will only deal with stationary policies $\pi = (\mu, \mu, \ldots)$ lemma 1.15 implies that transition probabilities can be computed as,

$$P^{\pi}(X_{t+1} = i | X_t = x) = \sum_{u \in \mathbb{U}} \mu(u|x) \cdot P^u_{xi}$$

2.1 Forbidden states

Fix a set of states $\mathbb{F} \subset \mathbb{X}$. Let π be a policy and denote the set of recurrent states under that policy as,

$$\mathbb{X}^{\pi}_R := \{x \in \mathbb{X} \mid P^{\pi}_x(\tau_x(1) < \infty) = 1\}$$

The $\mathbb{F}$-safe recurrent states under policy π are defined as,

$$\mathbb{X}^{\pi}_{R,\mathbb{F}} := \{x \in \mathbb{X}^{\pi}_R \mid P^{\pi}_{x,x^+} = 0 \text{ for all } x^+ \in \mathbb{F}\}$$

So the maximal set of $\mathbb{F}$-safe recurrent states is defined as,

$$\mathbb{X}_{\mathbb{F}} := \bigcup_{\pi} \mathbb{X}^{\pi}_{R,\mathbb{F}}$$

The goal is to find a policy π such that $\mathbb{X}^{\pi}_{R,\mathbb{F}} = \mathbb{X}_{\mathbb{F}}$. Before stating the solution for such problem we require some notation. Denote the set of probability measures over a set $\mathbb{S}$ as $P(\mathbb{S})$. The entropy function is defined as follows,

Definition 2.1. The function $\mathcal{H} : P(\mathbb{S}) \to \mathbb{R}_{\geq 0}$,

$$\mathcal{H}(f) = -\sum_{s \in \mathbb{S}} f(s) \cdot ln(f(s))$$

where the convention $0 \cdot ln(0) = 0$ is used, is called the entropy of f.

The importance of the entropy relies on the following lemma, for a proof see [1].

Lemma 2.2. *Let W be a convex subset of $P(\mathbb{S})$. Define $f^*_{\mathbb{S}} \in \arg\max \mathcal{H}(f_{\mathbb{S}})$ where $f_{\mathbb{S}} \in W$ and let $S_f := \{s \in \mathbb{S} | f(s) > 0\}$. Then, $S_{f_{\mathbb{S}}} \subset S_{f^*_{\mathbb{S}}}$ for all $f_{\mathbb{S}} \in W$.*

Therefore, $f^*_{\mathbb{S}}$ has *least* zeros, and as we will see later this corresponds to a maximum number of recurrent states. The following program solves the desired problem.

$$\begin{aligned}
&\max \mathcal{H}(f_{\mathbb{X},\mathbb{U}}) && f_{\mathbb{X},\mathbb{U}} \in P(\mathbb{X} \times \mathbb{U}) \\
&\text{Subject to} \\
&\sum_{u^+ \in \mathbb{U}} f_{\mathbb{X},\mathbb{U}}(x^+, u^+) = \sum_{x \in \mathbb{X}, u \in \mathbb{U}} P^u_{xx^+} f_{\mathbb{X},\mathbb{U}}(x, u) && \text{for all } x^+ \in \mathbb{X} \\
&\sum_{u \in \mathbb{U}} f_{\mathbb{X},\mathbb{U}}(x, u) = 0 && \text{for all } x \in \mathbb{F}
\end{aligned}$$

Note that between 0 and 1 the entropy is a concave function. Moreover, the constraints are linear so the distributions that satisfy those constraints are a convex subset of $P(\mathbb{X} \times \mathbb{U})$, so above program is convex. The first constraint defines an invariant distribution, and the second constraint is capturing the $\mathbb{F}$- safety. Intuitively above program is just saying that we must take the invariant distribution with least zeros, but those zeros must contain $\mathbb{F}$. So because of Theorem 1.10 the places where the distribution is not zero must be recurrent states. The theorem is as follows.

Theorem 2.3. *Assume the previous problem is feasible and that $f^*_{\mathbb{X},\mathbb{U}}$ is an optimal solution. Use the notation $f^*_{\mathbb{X}}(x) = \sum_{u \in \mathbb{U}} f^*_{\mathbb{X},\mathbb{U}}(x, u)$. Then,*

- $\mathbb{X}_{\mathbb{F}} = S_{f^*_{\mathbb{X}}}$

- $\mathbb{X}_{\mathbb{F}} = \mathbb{X}^{\pi}_{R,\mathbb{F}}$, *where* π *is given by,*

$$\pi(u|x) = \begin{cases} \frac{f^*_{\mathbb{X},\mathbb{U}}(x,u)}{f^*_{\mathbb{X}}(x)} & \text{if } x \in S_{f^*_{\mathbb{X}}} \\ \frac{1}{|\mathbb{U}|} & \text{otherwise} \end{cases}$$

2.2 Partition of the state space

Let's fix a partition of the state space, that is a family of mutually disjoint sets $\mathbb{P}_i$ such that $\cup_i \mathbb{P}_i = \mathbb{X}$. We are interested on states x that are recurrent under some policy π, such that if $x \in \mathbb{P}_i$ then $P^{\pi}_x(X_1 = j) = 0$ for all $j \in \mathbb{P}^c_i$. Recall that because of Proposition 1.7, the state x belongs to a closed class, let C^{π}_x be such class. Therefore, above assertion can be expressed just as $C^{\pi}_x \subset \mathbb{P}_i$. Denote the recurrent states satisfying such property as,

$$\mathbb{X}^{\pi}_{R,\mathbb{P}} := \{x \in \mathbb{X} \mid x \in \mathbb{X}^{\pi}_R \text{ and } C^{\pi}_x \subset \mathbb{P}_i \text{ for some } i\}$$

So the maximal set of recurrent states with such property is defined as,

$$\mathbb{X}_{\mathbb{P}} := \bigcup_{\pi} \mathbb{X}^{\pi}_{R,\mathbb{P}}$$

Then, the problem of interest is to find a policy π such that,

$$\mathbb{X}^{\pi}_{R,\mathbb{P}} = \mathbb{X}_{\mathbb{P}}$$

Remark. One possible approach to solve such problem, would be to fix a piece of the partition $\mathbb{P}_i$, define as a forbidden the set $\mathbb{P}^c_i$ and apply the algorithm proposed in last section. Then, repeat with all the other pieces of the partition and define a new policy using the policies calculated. However, this would require to solve a convex problem for each piece of the partition.

One interesting observation is that policies arising from problems related to maximizing the number of recurrent states, may not be deterministic, as the following example shows.

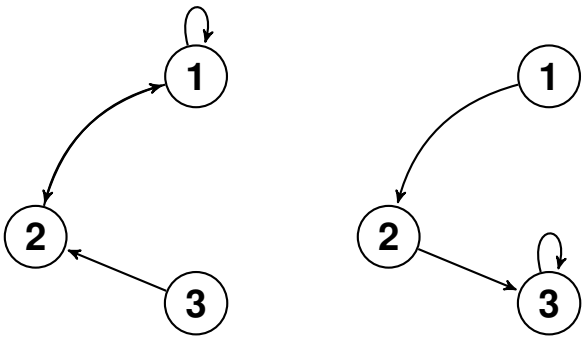
Example 2.4. Let $\mathbb{X} = \{1, 2, 3\}$ and $\mathbb{U} = \{1, 2\}$, where the transition matrices are given by:

$$P^1 = \begin{bmatrix} 0.4 & 0.6 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad P^2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

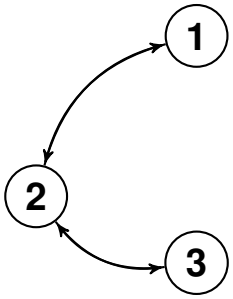
Let's consider a policy π, with corresponding transition matrix P^{π},

$$\pi(u|x) = \begin{bmatrix} 0 & 1 \\ 0.5 & 0.5 \\ 1 & 0 \end{bmatrix} \qquad P^{\pi} = \begin{bmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{bmatrix}$$

Note that the induced Markov chain consist of one class, so there is just one recurrent class. However, all deterministic policies induce at least two classes. To see this, note that starting from



(a) P^1 (b) P^2



(c) P^π

Figure 2.1: Graph representation of transition matrices

state 2, if we select just the first transition matrix, no matters what control action we define on states $1, 3$, there is no way 3 can be accessible from 2. On the other hand, if we select the second transition matrix, no matters what control action we define on states $1, 3$, there is no way 1 can be accessible from 2.

Our approach to solve the proposed problem uses a kind of Lyapunov function. Let $L : \mathbb{X} \to \mathbb{N}$ be defined as,

$$L(\mathbb{P}_i) = i.$$

The idea is to capture the notion of a closed set using such function. Next lemma relates a closed set with the just described function and it will be fundamental to solve the proposed problem.

Lemma 2.5. *Fix a policy π, and let $C^\pi \subset \mathbb{X}$ be a closed set under that policy. Suppose for all $x \in C^\pi$,*

$$\mathbb{E}_x^\pi[L(X_1)] \leq \mathbb{E}_x^\pi[L(X_0)]$$

Then, the value of the L function along C^π is constant.

Proof. Let $A = \{L(x) \mid x \in C\}$. For the sake of contradiction suppose there exist $a, b \in C$ such

that $L(a) > L(b)$, and assume $L(b)$ is the minimum of A. We have that,

$$\begin{aligned}\mathbb{E}_b^\pi[L(X_1)] &= \sum_{x\in\mathbb{X}} L(x)\cdot P_b^\pi(X_1 = x)\\ &= \sum_{x\in C} L(x)\cdot P_b^\pi(X_1 = x)\\ &\leq \mathbb{E}_b^\pi[L(X_0)] = L(b)\end{aligned}$$

Now, since $L(b)$ is the minimum of A we have,

$$L(b)\cdot\sum_{x\in C^\pi} P_b^\pi(X_1 = x) < \sum_{x\in C^\pi} L(x)\cdot P_b^\pi(X_1 = x) \leq L(b)$$

However, since C^π is a closed set $\sum_{x\in C^\pi} P_b^\pi(X_1 = x) = 1$, so above inequality tells us that $L(b) < L(b)$, a contradiction. □

The L function also provides a way to identify closed sets.

Lemma 2.6. *Let $B \subset S$ and assume the Lyapunov function is constant and positive along B and 0 outside B. Suppose that for all $i \in B$ we have*

$$\mathbb{E}_i[L(X_1)] = \mathbb{E}_i[L(X_0)]$$

Then, B is a closed set.

Proof. Let $b \in B$,

$$\begin{aligned}\mathbb{E}_b[L(X_1)] &= \sum_{i\in S} L(i)\cdot P(X_1 = i|X_0 = b)\\ &= \sum_{i\in B} L(i)\cdot P(X_1 = i|X_0 = b)\\ &= L(b)\cdot\sum_{i\in B} P(X_1 = i|X_0 = b)\\ &= \mathbb{E}_b[L(X_0)] = L(b)\end{aligned}$$

That is, $L(b)\cdot\sum_{i\in B} P(X_1 = i|X_0 = b) = L(b)$ and since $L(b) > 0$ we conclude $\sum_{i\in B} P(X_1 = i|X_0 = b) = 1$. Therefore, for all $b \in B$ we have $\sum_{i\in B} P(X_1 = i|X_0 = b) = 1$, so B is a closed set. □

Following the algorithm described in previous section, we state that the following program solves the desired problem.

$$\max \mathcal{H}(f_{\mathbb{X},\mathbb{U}}) \qquad f_{\mathbb{X},\mathbb{U}} \in P(\mathbb{X}\times\mathbb{U})$$

Subject to

$$\sum_{u^+\in\mathbb{U}} f_{\mathbb{X},\mathbb{U}}(x^+,u^+) = \sum_{x\in\mathbb{X},u\in\mathbb{U}} P^u_{xx^+} f_{\mathbb{X},\mathbb{U}}(x,u) \qquad \text{for all } x^+\in\mathbb{X}$$

$$\sum_{u\in\mathbb{U}} f_{\mathbb{X},\mathbb{U}}(x,u)L(x) \geq \sum_{x^+\in\mathbb{X},u\in\mathbb{U}} L(x^+) f_{\mathbb{X},\mathbb{U}}(x,u)P^u_{xx^+} \qquad \text{for all } x\in\mathbb{X}$$

Recall that the entropy is a concave function in $[0,1]$, and since the constraints are linear above program is convex. The first constraint defines an invariant distribution, while second constraint is just the condition presented in lemma 2.5. The theorem is as follows,

Theorem 2.7. *Assume the previous problem is feasible and that $f^*_{\mathbb{X},\mathbb{U}}$ is an optimal solution. Adopt the notation $f^*_{\mathbb{X}}(x) = \sum_{u\in\mathbb{U}} f^*_{\mathbb{X},\mathbb{U}}(x,u)$. Then,*

- $\mathbb{X}_{\mathbb{P}} = S_{f^*_{\mathbb{X}}}$
- $\mathbb{X}_{\mathbb{P}} = \mathbb{X}^{\pi}_{R,\mathbb{P}}$, *where π is given by,*

$$\pi(u|x) = \begin{cases} \frac{f^*_{\mathbb{X},\mathbb{U}}(x,u)}{f^*_{\mathbb{X}}(x)} & \text{if } x\in S_{f^*_{\mathbb{X}}} \\ \frac{1}{|\mathbb{U}|} & \text{otherwise} \end{cases}$$

Proof. Let's begin showing that $\mathbb{X}_{\mathbb{P}} \subset S_{f^*_{\mathbb{X}}}$. Let $b\in\mathbb{X}_{\mathbb{P}}$ so there exist a policy π such that x is recurrent and $C^{\pi}_b \subset \mathbb{P}_i$ for some i. Recall that because of Theorem 1.10 there must be an invariant distribution $f_{\mathbb{X}}$ such that $f_{\mathbb{X}}(b) > 0$ and $f_{\mathbb{X}}(C^{\pi}_b) = 1$. The idea will be to show $S_{f_{\mathbb{X}}} \subset S_{f^*_{\mathbb{X}}}$. Now, let $f_{\mathbb{X},\mathbb{U}} \in P(\mathbb{X}\times\mathbb{U})$ be defined as $f_{\mathbb{X},\mathbb{U}}(x,u) = \pi(u|x)f_{\mathbb{X}}(x)$. Let's see that $f_{\mathbb{X},\mathbb{U}}(x,u)$ is a feasible solution of the program. Because of the invariance of $f_{\mathbb{X}}$ we have,

$$\begin{aligned} f_{\mathbb{X}}(x^+) &= \sum_{x\in\mathbb{X}} f_{\mathbb{X}}(x)P^{\pi}_{x,x^+} \\ &= \sum_{x\in\mathbb{X}} f_{\mathbb{X}}(x)\Big[\sum_{u\in\mathbb{U}} \pi(u|x)P^u_{x,x^+}\Big] \\ &= \sum_{x\in\mathbb{X}}\sum_{u\in\mathbb{U}} f_{\mathbb{X},\mathbb{U}}(x,u)P^u_{x,x^+} \end{aligned}$$

Which corresponds to the first restriction of our program. For the second restriction, let $x\in\mathbb{X}$ so there are two cases: $x\in C^{\pi}_b$ or $x\notin C^{\pi}_b$,

- Assume $x \in C_b^\pi$. Since $C_b^\pi \subset \mathbb{P}_i$, then the value of the function L along C_b^π is constant and equal to $L(b)$. Therefore,

$$\begin{aligned}\sum_{x^+\in\mathbb{X},u\in\mathbb{U}} L(x^+) f_{\mathbb{X},\mathbb{U}}(x,u) P^u_{xx^+} &= \sum_{x^+\in\mathbb{X},u\in\mathbb{U}} L(x^+)\pi(u|x) f_{\mathbb{X}}(x) P^u_{xx^+} \\ &= \sum_{x^+\in\mathbb{X}} L(x^+) f_{\mathbb{X}}(x) P^\pi_{xx^+}\end{aligned}$$

  Since C_b^π is closed under π we have,

$$\begin{aligned}&= \sum_{x^+\in C_b^\pi} L(x^+) f_{\mathbb{X}}(x) P^\pi_{xx^+} \\ &= L(b) f_{\mathbb{X}}(x) \sum_{x^+\in C_b^\pi} P^\pi_{xx^+} \\ &= L(b) f_{\mathbb{X}}(x) \\ &= \sum_{u\in\mathbb{U}} f_{\mathbb{X},\mathbb{U}}(x,u) L(x)\end{aligned}$$

- Assume $x \notin C_b^\pi$. Recall that by construction $f_{\mathbb{X}}(C_b^\pi) = 1$, so we have that $f_{\mathbb{X}}(x) = 0$, then $f_{\mathbb{X},\mathbb{U}}(x,u) = \pi(u|x) f_{\mathbb{X}}(x) = 0$ for all $u \in \mathbb{U}$. So the restriction is satisfied.

Therefore, $f_{\mathbb{X},\mathbb{U}}$ is a feasible solution of the program. Finally, since feasible solutions are a convex subset of $P(\mathbb{X}\times\mathbb{U})$, lemma 2.2 implies $S_{f_{\mathbb{X},\mathbb{U}}} \subset S_{f^*_{\mathbb{X},\mathbb{U}}}$. So that $S_{f_{\mathbb{X}}} \subset S_{f^*_{\mathbb{X}}}$ and since $f_{\mathbb{X}}(b) > 0$ we obtain $b \in S_{f^*_{\mathbb{X}}}$, so $\mathbb{X}_{\mathbb{P}} \subset S_{f^*_{\mathbb{X}}}$.

Let's show that $S_{f^*_{\mathbb{X}}} \subset \mathbb{X}_{\mathbb{P}}$. The idea will be to show that $S_{f^*_{\mathbb{X}}}$ are recurrent states under some policy and then use lemma 2.5. Consider the policy,

$$\pi(u|x) = \begin{cases} \frac{f^*_{\mathbb{X},\mathbb{U}}(x,u)}{f^*_{\mathbb{X}}(x)} & \text{if } x \in S_{f^*_{\mathbb{X}}} \\ \frac{1}{|\mathbb{U}|} & \text{otherwise} \end{cases}$$

Let's see that under such policy $f^*_{\mathbb{X}}$ is an invariant distribution. First of all, note that if $x \notin S_{f^*_{\mathbb{X}}}$ then $f^*_{\mathbb{X},\mathbb{U}}(x,u) = 0$ for all $u \in \mathbb{U}$. Now, let $x^+ \in \mathbb{X}$, so because of the first restriction,

$$\begin{aligned}f^*_{\mathbb{X}}(x^+) &= \sum_{x\in\mathbb{X},u\in\mathbb{U}} P^u_{xx^+} f^*_{\mathbb{X},\mathbb{U}}(x,u) \\ &= \sum_{x\in S_{f^*_{\mathbb{X}}},u\in\mathbb{U}} P^u_{xx^+} f^*_{\mathbb{X}}(x)\pi(u|x) \\ &= \sum_{x\in S_{f^*_{\mathbb{X}}}} f_{\mathbb{X}^*}(x) P^\pi_{xx^+} \\ &= \sum_{x\in\mathbb{X}} f_{\mathbb{X}^*}(x) P^\pi_{xx^+}\end{aligned}$$

So that $f^*_{\mathbb{X}}$ is an invariant distribution under π. Therefore, theorem 1.10 tells us that $S_{f^*_{\mathbb{X}}}$ consist of recurrent states. Now, let $x \in S_{f^*_{\mathbb{X}}}$ and consider the closed class of x under π, C^π_x. In order to apply lemma 2.5 we need to show that for all $b \in C^\pi_x$,

$$\mathbb{E}^\pi_b[L(X_1)] \leq \mathbb{E}^\pi_b[L(X_0)]$$

Let $b \in C^\pi_x$. Because of the second restriction,

$$\sum_{u\in\mathbb{U}} f_{\mathbb{X},\mathbb{U}}(b,u)L(b) \geq \sum_{x^+\in\mathbb{X},u\in\mathbb{U}} L(x^+)f_{\mathbb{X},\mathbb{U}}(b,u)P^u_{bx^+}$$

Since, $f_{\mathbb{X},\mathbb{U}}(b,u) = f^*_{\mathbb{X}}(b)\pi(u|b)$ above expression can be written as,

$$\sum_{u\in\mathbb{U}} f^*_{\mathbb{X}}(b)\pi(u|b)L(b) \geq \sum_{x^+\in\mathbb{X},u\in\mathbb{U}} L(x^+)f^*_{\mathbb{X}}(b)\pi(u|b)P^u_{bx^+}$$

Dividing by $f^*_{\mathbb{X}}(b)$ we obtain $\mathbb{E}^\pi_b[L(X_1)] \leq \mathbb{E}^\pi_b[L(X_0)]$. Finally, using lemma 2.5 we obtain that the value of L along C_x is constant so that $C_x \subset \mathbb{P}_i$ for some i, implying $x \in \mathbb{X}_{\mathbb{P}}$. □

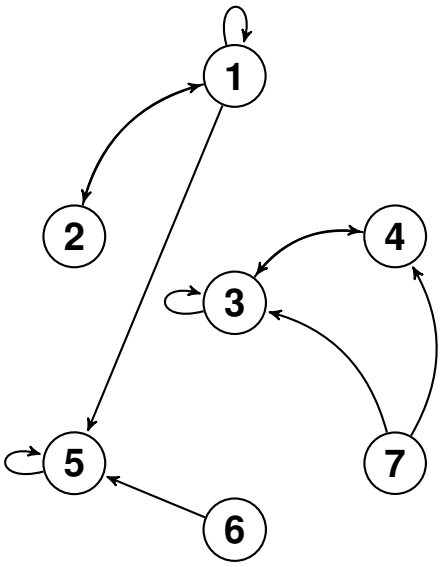
To end this chapter we show an example of the proposed algorithm. Such implementation can be done easily using *Matlab* and a package for convex programming as *cvx*.

Example 2.8. Let $\mathbb{X} = \{1,2,3,4,5,6,7\}$ and $\mathbb{U} = \{1,2\}$, where the transition matrices are given by,

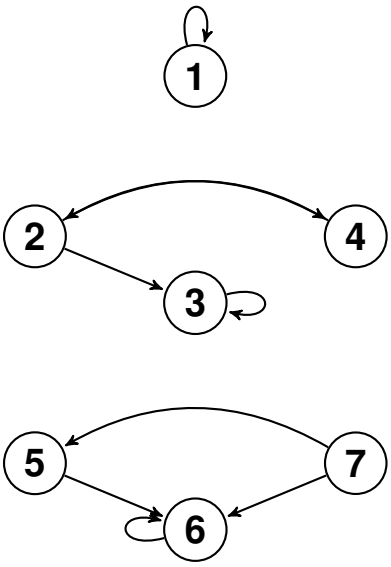
$$P^1 = \begin{bmatrix} 0.4 & 0.3 & 0 & 0 & 0.3 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 & 0 & 0 & 0 \end{bmatrix} \quad P^2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0.8 & 0.2 & 0 \end{bmatrix}$$

Let's consider a partition $\mathbb{P}_1 = \{1\}, \mathbb{P}_2 = \{2,3,4\}, \mathbb{P}_3 = \{5,6,7\}$. Define $L = [1,2,2,2,3,3,3]$. Running the algorithm, we obtain the joint density distribution,

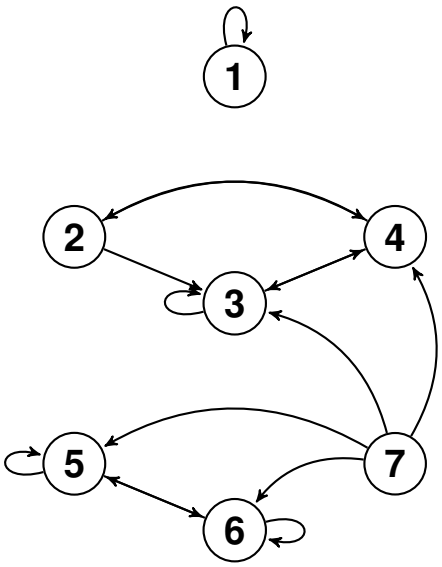
$$f_{\mathbb{X},\mathbb{U}} = \begin{bmatrix} 0 & 0.1015 \\ 0 & 0.0915 \\ 0.1411 & 0.1015 \\ 0.0672 & 0.0915 \\ 0.1015 & 0.1015 \\ 0.1015 & 0.1015 \\ 0 & 0 \end{bmatrix}$$



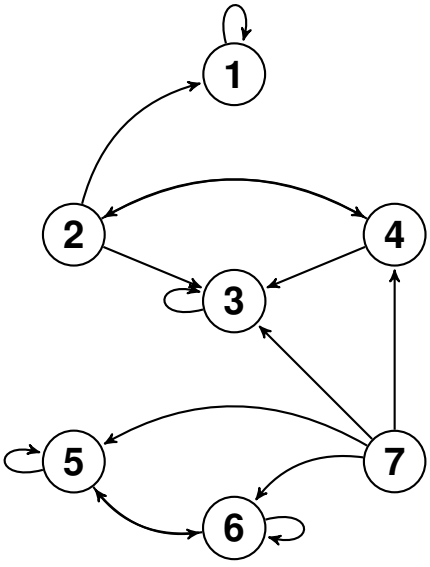
(a) P^1



(b) P^2



(c) Graph representation of P^{π}



(d) Graph representation of P^{π^+}

Figure 2.2: Graph representation of transition matrices

Therefore, using above theorem we obtain that states all states, except 7, are recurrent under policy π, with corresponding transition matrix P^{π},

$$\pi = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0.5818 & 0.4182 \\ 0.4232 & 0.5768 \\ 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0.5408 & 0.4592 \end{bmatrix} \quad P^{\pi} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.5346 & 0.4654 & 0 & 0 & 0 \\ 0 & 0.5768 & 0.4232 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0.1082 & 0.4327 & 0.3673 & 0.0918 & 0 \end{bmatrix}$$

Using above figure we can obtain that under P^{π}, the classes are $\{1\}$,$\{2,3,4\}$,$\{5,6\}$, $\{7\}$. The recurrent classes are $\{1\} \subset \mathbb{P}_1$, $\{2,3,4\} \subset \mathbb{P}_2$, $\{5,6\} \subset \mathbb{P}_3$. So each recurrent class is fully contained in a piece of the partition. Note also that $\{2,3,4\}$, $\{5,6\}$ weren't recurrent classes under P^1 or P^2. As another example consider the partition $\mathbb{P}_1 = \{1,2,3\}$, $\mathbb{P}_2 = \{4,5,6,7\}$. Applying the algorithm we obtain the joint density distribution $f_{\mathbb{X},\mathbb{U}}$ and policy π^+,

$$f_{\mathbb{X},\mathbb{U}} = \begin{bmatrix} 0 & 0.1667 \\ 0 & 0 \\ 0 & 0.1667 \\ 0 & 0 \\ 0.1667 & 0.1667 \\ 0.1667 & 0.1667 \\ 0 & 0 \end{bmatrix} \quad \pi^+ = \begin{bmatrix} 0 & 1 \\ 0.5174 & 0.4826 \\ 0 & 1 \\ 0.3510 & 0.6490 \\ 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0.6325 & 0.3675 \end{bmatrix}$$

Therefore, states $1,3,4,6$ are recurrent under π^+ and transition matrix P^{π^+},

$$P^{\pi^+} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5174 & 0 & 0.2413 & 0.2413 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0.6490 & 0.3510 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0.1265 & 0.5060 & 0.2940 & 0.0735 & 0 \end{bmatrix}$$

So the recurrent classes are $\{1\},\{3\} \subset \mathbb{P}_1$, $\{5,6\} \subset \mathbb{P}_2$. Note that the transient class $\{2,4\}$ is not contained in any piece of the partition.

Chapter 3

Zubov's Method

In this chapter we are going to study Zubov's method in the context of Markov chains. Zubov's method, appears in the context of deterministic dynamical systems when describing the domain of attraction of a stable point, see [7, 11]. In specific, it allows us to characterize the *domain of attraction* and the *escape set* in terms of an appropriate function v. Such function is also characterised as the solution of a differential equation. In [5] Zubov's method is generalized for deterministic controlled systems, and in [6] it's generalized for stochastic differential equations. For the present work we took as guide the constructions made in this last paper.

We will proceed as follows: The first section will deal with the uncontrolled case, our objective will be to describe both, the domain of attraction and the escape set of certain states, in terms of a value function v. In the second section, we will deal with the controlled case, and again, the objective will be to describe those sets in terms of a value function, and compute the Bellman's equation. Finally, in third section we will consider the opposite of the domain of attraction for controlled systems.

3.1 Uncontrolled case

Consider a countable set $\mathbb{X}$, and let $\{X_t\}_{t=0}$ be a $\mathbb{X}$-valued Markov chain defined over a probability space $(\Omega, \mathcal{F}, P)$. Fix a set $A \subset \mathbb{X}$. In order to ensure some kind of stability assume that A is a closed set. The main objects for study are the following,

Definition 3.1. The domain of attraction an escape set of A are defined as,

$$\Lambda = \left\{ x \in \mathbb{X} \;\middle|\; \liminf_{t\to\infty} P_x(X_t \in A) > 0 \right\}$$
$$\Gamma = \left\{ x \in \mathbb{X} \;\middle|\; \liminf_{t\to\infty} P_x(X_t \in A) = 0 \right\}$$

In order to understand those sets we are going to describe them in terms of hitting times. The following lemmas will help us to achieve such objective.

Lemma 3.2. *For all $x \in \mathbb{X}$ the probabilities $P_x(\tau_A \leq t)$ and $P_x(X_t \in A)$ are equal.*

Proof. Define $B := \{\omega \in \Omega | \forall m > \tau_A(\omega), X_m(\omega) \in A\}$. So that,

$$\{\tau_A \leq t\} = (\{\tau_A \leq t\} \cap B) \cup (\{\tau_A \leq t\} \cap B^c)$$

$$\{X_t \in A\} = (\{X_t \in A\} \cap B) \cup (\{X_t \in A\} \cap B^c)$$

Let's see that the events $\{X_t \in A\} \cap B$ and $\{\tau_A \leq t\} \cap B$ are equal. It's clear that $\{X_t \in A\} \cap B \subset \{\tau_A \leq t\} \cap B$. Now, let $\omega \in \{\tau_A \leq t\} \cap B$ so $\tau_A(\omega) \leq t$. Since $\omega \in B$, for all $m \geq \tau_A(\omega)$ we have $X_m(w) \in A$, in particular $X_t(w) \in A$, so that $\{\tau_A \leq t\} \cap B \subset \{X_t \in A\} \cap B$. Therefore,

$$P_x(\{\tau_A \leq t\} \cap B) = P_x(\{X_t \in A\} \cap B)$$

Finally, let's consider the event $E = \{\exists j, m \in \mathbb{N} \text{ such that } j \leq m, X_j \in A, X_m \notin A\}$. Note that, $\{\tau_A \leq t\} \cap B^c \subset E$ and $\{X_t \in A\} \cap B^c \subset E$. Since A is closed, lemma 1.13 implies $P_x(E) = 0$, so that,

$$0 = P_x(\{\tau_A \leq t\} \cap B^c) = P_x(\{X_t \in A\} \cap B^c)$$

Combining both results we obtain the desired equality.

□

Lemma 3.3. *Consider the hitting time of A, τ_A. Then,*

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

Proof. By Lemma 1.12 we know $\liminf\{\tau_A \leq t\} = \{\tau_A < \infty\} = \limsup\{\tau_A \leq t\}$, so by Fatou's lemma (Lemma 1.11),

$$\begin{aligned} P_x(\tau_A < \infty) &= P_x\left(\liminf_{t\to\infty}\{\tau_A \leq t\}\right) \\ &\leq \liminf_{t\to\infty} P_x(\tau_A \leq t) \\ &= \liminf_{t\to\infty} P_x(X_t \in A) \end{aligned}$$

On the other hand, Fatou's Lemma (Lemma 1.11) also implies,

$$\begin{aligned} P_x(\tau_A < \infty) &= P_x\left(\limsup_{t\to\infty}\{\tau_A \leq t\}\right) \\ &\geq \limsup_{t\to\infty} P_x(\tau_A \leq t) \\ &= \limsup_{t\to\infty} P_x(X_t \in A) \end{aligned}$$

So that,

$$\limsup_{t\to\infty} P_x(X_t \in A) \leq P_x(\tau_A < \infty) \leq \liminf_{t\to\infty} P_x(X_t \in A)$$

□

Therefore, the domain of attraction and escape set can be understood as follows,

Corollary 3.4.

$$\Lambda = \{x \in \mathbb{X} \mid P_x(\tau_A < \infty) > 0\}$$

$$\Gamma = \{x \in \mathbb{X} \mid P_x(\tau_A < \infty) = 0\}$$

We have another characterization of those sets, but now in terms of the expectation of τ_A.

Lemma 3.5.

$$\Lambda = \{x \in \mathbb{X} \mid \mathbb{E}_x[e^{-\tau_A}] > 0\}$$

$$\Gamma = \{x \in \mathbb{X} \mid \mathbb{E}_x[e^{-\tau_A}] = 0\}$$

Proof. We have that

$$\mathbb{E}_x[e^{-\tau_A}] = \sum_j e^{-j} P_x(\tau_A = j)$$

Therefore, $\mathbb{E}_x[e^{-\tau_A}] = 0$ if and only if $P_x(\tau_A = j) = 0$ for all j, which happens if and only if $P_x(\tau_A < \infty) = 0$. So using the previous corollary we can conclude the proof. □

The idea now, is to characterize both sets in terms of an appropriate function.

Definition 3.6. Let $v_\gamma : \mathbb{X} \to \mathbb{R}$ be defined as,

$$v_\gamma(x) := \mathbb{E}_x \left[\sum_{t=0}^{\infty} \mathbb{1}_{A^c}(X_t) \cdot \gamma^t \right] \text{ for some } \gamma \in (0, 1).$$

Note that $\sum_{t=0}^{\infty} \mathbb{1}_{A^c}(X_t) \cdot \gamma^t \leq L$, where $L = \sum_{t=0}^{\infty} \gamma^t$. Therefore, $0 \leq v_\gamma(x) \leq L$ so the function is well defined. Before we can state and prove the characterization we need the following lemma, denote $Z = \sum_{t=0}^{\infty} \mathbb{1}_{A^c}(X_t) \cdot \gamma^t$.

Lemma 3.7. *For all $x \in \mathbb{X}$,*

$$P_x(\tau_A < \infty) = P_x(Z < L)$$

Proof. Define $B := \{\omega \in \Omega | \forall m > \tau_A(\omega), X_m(\omega) \in A\}$. So that,

$$\{\tau_A < \infty\} = (\{\tau_A < \infty\} \cap B) \cup (\{\tau_A < \infty\} \cap B^c)$$

$$\{Z < L\} = (\{Z < L\} \cap B) \cup (\{Z < L\} \cap B^c)$$

Let's begin proving $\{Z < L\} \cap B = \{\tau_A < \infty\} \cap B$. Let $\omega \in \{Z < L\} \cap B$ so $\sum_t \mathbb{1}_{A^c}(X_t(\omega)) \cdot \gamma^t < L$. Then, there exist $t \in \mathbb{N}$ such that $g(X_t(\omega)) = 0$. By the definition of g this implies $X_t(\omega) \in A$, so $\tau_A(\omega) < \infty$. Now, let $\omega \in \{\tau_A < \infty\} \cap B$. Then, $\tau_A(\omega) < \infty$ and since ω belongs to B, for all $m \geq \tau_A(\omega)$ we have $X_m(\omega) \in A$, so $g(X_m(\omega)) = 0$ implying $Z(\omega) < L$. Therefore,

$$P_x(\{\tau_A < \infty\} \cap B) = P_x(\{Z < L\} \cap B)$$

Now, let's consider the event $E = \{\exists j, m \in \mathbb{N} \text{ such that } j \leq m, X_j \in A, X_m \notin A\}$. We have that $\{\tau_A < \infty\} \cap B^c \subset E$ and $\{Z < L\} \cap B^c \subset E$. Since A is closed, lemma 1.13 implies $P_x(E) = 0$, so that,

$$0 = P_x(\{\tau_A < \infty\} \cap B^c) = P_x(\{Z < L\} \cap B^c)$$

Both results imply,

$$P_x(\tau_A < \infty) = P_x(Z < L)$$

□

The characterization of the domain an escape set in terms of v is the following,

Theorem 3.8.

$$\Lambda = \{x \in \mathbb{X} \mid v_\gamma(x) < L\}$$

$$\Gamma = \{x \in \mathbb{X} \mid v_\gamma(x) = L\}$$

Proof. Let's consider the partial sums of Z, $S_n = \sum_{i=0}^{n} \gamma^i$. Note that for all $n \in \mathbb{N}$ the event $\{Z \neq S_n\}$ is contained in the event,

$$E = \{\exists j, m \in \mathbb{N} \text{ such that } j \leq m, X_j \in A, X_m \notin A\}$$

Since A is closed Lemma 1.13 implies $P_x(Z \neq S_n) = 0$ for all $n \in \mathbb{N}$. As a consequence,

$$P_x(Z < L) = P_x(Z = 0) + \sum_n P_x(Z = S_n)$$

Let $x \in \Lambda$, then $0 < P_x(\tau_A < \infty)$. Using previous lemma we conclude $0 < P_x(Z < L)$, which implies $P_x(Z = 0) > 0$ or $P_x(Z = S_n) > 0$ for some n, so that,

$$\begin{aligned}
v_\gamma(x) &= \sum_{n=0}^{\infty} S_n P_x(Z = S_n) + L \cdot P_x(Z = L) \\
&\leq P_x(Z = 0) + \sum_{n=0}^{\infty} S_n P_x(Z = S_n) + L \cdot P_x(Z = L) \\
&< L\Big[P_x(Z = 0) + \sum_{n=0}^{\infty} P_x(Z = S_n)\Big] + L \cdot P_x(Z = L) \qquad \text{Since } L \geq 1 \\
&= L \cdot P_x(Z = L) + L \cdot P(Z < L) = L
\end{aligned}$$

On the other hand, if $x \in \Gamma$ then $P_x(\tau_A < \infty) = 0$ so $P_x(Z < L) = 0$. Therefore, $P_x(Z = L) = 1$. As a consequence,

$$\begin{aligned}
v_\gamma(x) &= \sum_{n=0}^{\infty} S_n P_x(Z = S_n) + L \cdot P_x(Z = L) \\
&= L.
\end{aligned}$$

□

3.2 Controlled case

Let $\mathbb{X}$ be a countable state space, and $\mathbb{U}$ a countable set of control actions. Let $\{(X_t, U_t)\}_{t=0}$ be a $\mathbb{X} \times \mathbb{U}$-valued controlled Markov chain, and assume all control actions are feasible for each state; that is $\mathbb{U}(x) = \mathbb{U}$ for all $x \in \mathbb{X}$. Fix a set $A \subset \mathbb{X}$, in order to guarantee some stability assume A is a closed set under some policy $\pi \in \Pi_S$.

Definition 3.9. The domain of attraction an escape set of A are defined as,

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

In order to use the results of previous section we would like to answer if it's enough to just consider polices that makes A a closed set.

Definition 3.10. Let $\Pi_A \subset \Pi_S$ be the set of control polices that make A a closed set. Recall that by assumption such set is non-empty.

The following is fundamental to answer above question,

Lemma 3.11. *Let π be a a policy, and let $\pi_A \in \Pi_A$. Define the policy $\pi_{op} \in \Pi_A$ as,*

$$\pi_{op}(u|x) = \begin{cases} \pi_A(u|x) & \textit{for } x \in A \\ \pi(u|x) & \textit{for } x \notin A \end{cases}$$

Then, for all $x \in A$,

$$\liminf_{t\to\infty} P_x^\pi(X_t \in A) \leq \lim_{t\to\infty} P_x^{\pi_{op}}(X_t \in A)$$
$$\limsup_{t\to\infty} P_x^\pi(X_t \in A) \leq \lim_{t\to\infty} P_x^{\pi_{op}}(X_t \in A)$$

Remark. Since A is closed under policy π_{op}, lemma 3.3 implies $\lim_{t\to\infty} P_x^{\pi_{op}}(X_t \in A)$ exists and is equal to $P_x^{\pi_{op}}(\tau_A < \infty)$.

Proof. Let $x \in A$, since A is closed under policy π_{op} we have $P_x^{\pi_{op}}(X_t \in A) = 1$, as a consequence,

$$\liminf_{t\to\infty} P_x^\pi(X_t \in A) \leq \lim_{t\to\infty} P_x^{\pi_{op}}(X_t \in A)$$
$$\limsup_{t\to\infty} P_x^\pi(X_t \in A) \leq \lim_{t\to\infty} P_x^{\pi_{op}}(X_t \in A)$$

Suppose $x \notin A$. Let's show by induction on t that,

$$P_x^\pi(\tau_A \leq t) = P_x^{\pi_{op}}(\tau_A \leq t)$$

For $t = 1$ we have,

$$\begin{aligned} P_x^\pi(\tau_A \leq 1) &= P_x^\pi(X_0 \in A) + P_x^\pi(X_1 \in A) \\ &= \sum_{j \in A} P_{xj}^\pi \\ &= \sum_{j \in A, u \in \mathbb{U}} \pi(u|x) P_{xj}^u \\ &= \sum_{j \in A, u \in \mathbb{U}} \pi_{op}(u|x) P_{xj}^u \\ &= P_x^{\pi_{op}}(\tau_A \leq 1) \end{aligned}$$

Assume the property is true for t. Let's prove it for $t+1$,

$$\begin{aligned} P_x^\pi(\tau_A \leq t+1) &= P_x^\pi(\tau_A \leq t) + P_x^\pi(X_1 \notin A, \ldots, X_t \notin A, X_{t+1} \in A) \\ &= P_x^{\pi_{op}}(\tau_A \leq t) + \\ &\quad \sum_{j_1 \notin A, \ldots, j_t \notin A, j_{t+1} \in A} P_x^\pi(X_1 = j_1, \ldots, X_t = j_t, X_{t+1} = j_{t+1}) \\ &= P_x^{\pi_{op}}(\tau_A \leq t) + \sum_{j_1 \notin A, \ldots, j_t \notin A, j_{t+1} \in A} P_{x,j_1}^\pi \ldots P_{j_t,j_{t+1}}^\pi \end{aligned}$$

Finally, since $x, j_1, \ldots, j_t \notin A$ then $P_{x,j_1}^\pi \ldots P_{j_t,j_{t+1}}^\pi = P_{x,j_1}^{\pi_{op}} \ldots P_{j_t,j_{t+1}}^{\pi_{op}}$.

Now, note that $\{X_t \in A\} \subset \{\tau_A \leq t\}$, which implies,

$$P_x^\pi(X_t \in A) \leq P_x^\pi(\tau_A \leq t) = P_x^{\pi_{op}}(\tau_A \leq t)$$

Therefore,

$$\begin{aligned} \liminf_{t \to \infty} P_x^\pi(X_t \in A) &\leq \liminf_{t \to \infty} P_x^{\pi_{op}}(\tau_A \leq t) \\ \limsup_{t \to \infty} P_x^\pi(X_t \in A) &\leq \limsup_{t \to \infty} P_x^{\pi_{op}}(\tau_A \leq t) \end{aligned}$$

Finally, since A is closed under π_{op}, lemma 3.2 implies,

$$\begin{aligned} \liminf_{t \to \infty} P_x^\pi(X_t \in A) &\leq \liminf_{t \to \infty} P_x^{\pi_{op}}(\tau_A \leq t) \\ &= \lim_{t \to \infty} P_x^{\pi_{op}}(X_t \in A) \\ \limsup_{t \to \infty} P_x^\pi(X_t \in A) &\leq \limsup_{t \to \infty} P_x^{\pi_{op}}(\tau_A \leq t) \\ &= \lim_{t \to \infty} P_x^{\pi_{op}}(X_t \in A) \end{aligned}$$

□

As a corollary, we obtain the following description of Λ and Γ.

Corollary 3.12.

$$\Lambda = \{x \in \mathbb{X} \mid P_x^\pi(\tau_A < \infty) > 0 \textit{ for some policy } \pi \in \Pi_A\}$$

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

Proof. Clearly,

$$\{x \in \mathbb{X} \mid P_x^\pi(\tau_A < \infty) > 0 \text{ for some policy } \pi \in \Pi_A\} \subset \Lambda$$

Let $x \in \Lambda$, so there exist a policy $\pi \in \Pi_S$ such that,

$$\liminf_{t\to\infty} P_x^\pi(X_t \in A) > 0$$

Now, let $\pi_{op} \in \Pi_A$ be the policy defined in previous lemma. Therefore,

$$0 < \liminf_{t\to\infty} P_x^\pi(X_t \in A) < \lim_{t\to\infty} P_x^{\pi_{op}}(X_t \in A) = \lim_{t\to\infty} P_x^{\pi_{op}}(X_t \in A)$$

□

So that we can focus on policies that make A a closed set. The idea now is to describe a function that characterizes those sets.

Definition 3.13. Let $v_\gamma : \mathbb{X} \to \mathbb{R}$ be defined as,

$$v_\gamma(x) := \inf_{\pi\in\Pi_A} v_\gamma^\pi(x)$$

Where,

$$v_\gamma^\pi(x) = \mathbb{E}_x^\pi\left[\sum_{t=0}^{\infty} \mathbb{1}_{A^c}(X_t)\cdot\gamma^t\right] \text{ for some } \gamma \in (0,1), \text{ and policy } \pi \in \Pi_A.$$

The characterization is as follows,

Theorem 3.14.

$$\Lambda = \{x \in \mathbb{X} \mid v_\gamma(x) < L\}$$

$$\Gamma = \{x \in \mathbb{X} \mid v_\gamma(x) = L\}$$

Proof. Let $x \in \Lambda$, so there exist a policy $\pi \in \Pi_A$ such that $P_x^\pi(\tau_A < \infty) > 0$. Therefore, since A is closed under π, Theorem 3.8 implies $v_\gamma^\pi(x) < L$, as a consequence $v_\gamma(x) < L$.
Let $x \in \Gamma$, so for any policy $\pi \in \Pi_A$ we have $P_x^\pi(\tau_A < \infty) = 0$. As a consequence, Theorem 3.8 implies that for any policy $\pi \in \Pi_A$ the value function $v_\gamma^\pi(x) = L$, so that $v_\gamma(x) = L$. □

Finally, we would like to show how to calculate the value function v_γ. There is an important issue about calculating such function. Note that the optimization must be carried out only on elements of Π_A, so we would have to calculate such set first. Next theorem shows that it's not necessary.

Theorem 3.15. *Define,*

$$v_\gamma^* = \inf_{\pi \in \Pi_S} v_\gamma^\pi$$

Assume $\mathbb{U}$ is finite. Let π^* be an optimal policy (Such policy exists, Theorem 1.19). The set A is closed under π^*, so $v_\gamma = v_\gamma^*$. Moreover, $v_\gamma^{\pi^*}(A) = 0$.

Proof. Let $\pi \in \Pi_S$, given $x \in \mathbb{X}$, dominated convergence theorem implies,

$$\begin{aligned} v_\gamma^\pi(x) &= \mathbb{E}_x^\pi \left[\sum_{t=0}^{\infty} \mathbb{1}_{A^c}(X_t) \cdot \gamma^t \right] \\ &= \lim_{N \to \infty} \mathbb{E}_x^\pi \left[\sum_{t=0}^{N} \mathbb{1}_{A^c}(X_t) \cdot \gamma^t \right] \\ &= \lim_{N \to \infty} \sum_{t=0}^{N} \mathbb{E}_x^\pi \left[\mathbb{1}_{A^c}(X_t) \cdot \gamma^t \right] \\ &= \sum_{t=0} \gamma^t \cdot \mathbb{E}_x^\pi \Big[\mathbb{1}_{A^c}(X_t) \Big] \\ &= \sum_{t=0} \gamma^t \cdot P_x^\pi (X_t \notin A) \end{aligned}$$

Let π^* be an optimal policy and suppose A is not closed under that policy. Let $\pi_A \in \Pi_A$. Since A is closed under π_A, if $x \in A$,

$$P_x^{\pi_A}(X_t \notin A) = 0 \qquad \text{for all } t \geq 0$$

So in view of the previous calculation we obtain, $v_\gamma^{\pi_A}(x) = 0$ for all $x \in A$. However, since A is not closed under π^*, there exist $i \in A$ and $j \notin A$ such that $P_i^{\pi^*}(X_1 = j) > 0$. So that, $v_\gamma^{\pi^*}(i) > 0$, contradicting the fact that $v_\gamma^{\pi^*}(i) = \inf_{\pi \in \Pi_S} v_\gamma^\pi(i)$. □

To end this section we present the bellman's equation for Zubov's Method. Because of previous theorem,

$$\begin{aligned} v_\gamma(x) &= \inf_{\pi \in \Pi_S} \mathbb{E}_x^\pi \left[\sum_{t=0}^{\infty} \mathbb{1}_{A^c}(X_t) \cdot \gamma^t \right] \\ &= \frac{1}{1-\gamma} - \sup_{\pi \in \Pi_S} \mathbb{E}_x^\pi \left[\sum_{t=0}^{\infty} \mathbb{1}_{A}(X_t) \cdot \gamma^t \right] \end{aligned}$$

So by defining,

$$\widehat{v}_\gamma(x) = \sup_{\pi \in \Pi_S} \mathbb{E}_x^\pi \left[\sum_{t=0}^{\infty} \mathbb{1}_{A}(X_t) \cdot \gamma^t \right]$$

we can describe such problem in the same way as the presented in the discounted problem section.

Corollary 3.16. *The Bellman's equation for $\widehat{v}_\gamma$ are,*

$$\widehat{v_\gamma}(x) = \sup_{u\in\mathbb{U}} \Big\{ \mathbb{1}_A(x) + \gamma \sum_{j\in\mathbb{X}} P^u_{x,j} \widehat{v}_\gamma(x) \Big\}$$

Therefore, the linear programs are as follows. Let $\nu \in \mathbb{R}^{|\mathbb{X}|}_{\geq 0}$

Primal Linear Program

$$\text{Minimize } \sum_{x\in\mathbb{X}} \nu(x) \cdot \widehat{v}_\gamma(x)$$

Subject to

$$\widehat{v}(x) - \sum_{j\in\mathbb{X}} \gamma \cdot P^u_{x,j} \cdot \widehat{v}(j) \geq \mathbb{1}_A(x) \qquad \text{for } u \in \mathbb{U}, x \in \mathbb{X}$$

Dual Linear Program

$$\text{Maximize } \sum_{s\in\mathbb{X}} \sum_{u\in\mathbb{U}} \mathbb{1}_A \cdot x(s,u)$$

Subject to

$$\sum_{u\in\mathbb{U}} x(j,u) - \sum_{s\in\mathbb{X}} \sum_{u\in\mathbb{U}} \gamma \cdot P^u_{s,j} \cdot x(s,u) = \nu(j)$$
$$x(j,u) \geq 0 \quad \text{for } u \in \mathbb{U}, j \in \mathbb{X}$$

Such programs can be implemented easily using *MATLAB*, and *CVX*. To end this section we present an example.

Example 3.17. Let $\mathbb{X} = \{1,2,3,4,5,6,7\}$ and $\mathbb{U} = \{1,2\}$, where the transition matrices P^1 and P^2 are the same as in example 2.8. Let $A = \{2,3,4\}$ and recall that in such example we showed that there exist a policy for which A is closed, so that the presented theory applies. Let $\gamma = 0.5$ so $L = 2$. Using the algorithms we obtain that the dual program has an optimal solution,

$$x = \begin{bmatrix} 1.2500 & 0 \\ 0 & 1.8253 \\ 1.5851 & 1.4742 \\ 1.2148 & 1.2756 \\ 1.1120 & 1.1504 \\ 1.0379 & 1.0746 \\ 1 & 0 \end{bmatrix}$$

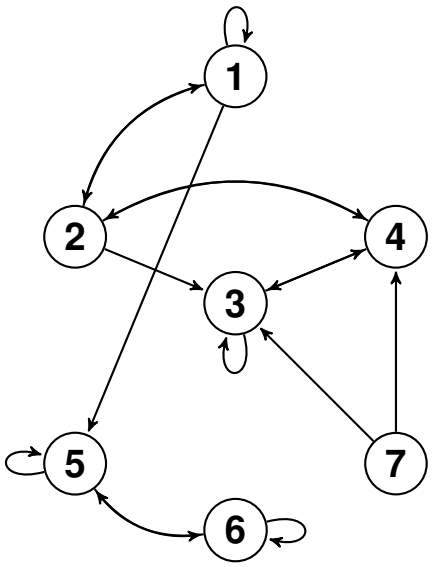


Figure 3.1: Graph representation of P^{π}

Therefore, as is stated in the preliminaries section, an optimal policy is,

$$\pi = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0.5181 & 0.4819 \\ 0.4878 & 0.5122 \\ 0.4915 & 0.5085 \\ 0.4913 & 0.5087 \\ 1 & 0 \end{bmatrix}$$

As a consequence the optimal value function $\widehat{v}$, and the function v are,

$$\widehat{v} = \begin{bmatrix} 0.375 \\ 2 \\ 2 \\ 2 \\ 0 \\ 0 \\ 1 \end{bmatrix} \qquad v = \begin{bmatrix} 1.625 \\ 0 \\ 0 \\ 0 \\ 2 \\ 2 \\ 1 \end{bmatrix}$$

Therefore, the domain of attraction are the states $\{1, 2, 3, 4, 7\}$, while the escape set is $\{5, 6\}$, result that can be checked by inspection on the graphs of P^1 and P^2. Using the policy π, we obtain the

transition matrix,

$$P^{\pi} = \begin{bmatrix} 0.4 & 0.3 & 0 & 0 & 0.3 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.5855 & 0.4145 & 0 & 0 & 0 \\ 0 & 0.5122 & 0.4878 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.4915 & 0.5085 & 0 \\ 0 & 0 & 0 & 0 & 0.4913 & 0.5087 & 0 \\ 0 & 0 & 0.2 & 0.8 & 0 & 0 & 0 \end{bmatrix}$$

Note that with such policy, all elements of the domain of attraction have positive probability of reaching A, as is shown in the above figure.

3.3 Forbidden states

As we saw in previous section, the objective of the domain of attraction is to make the set A as *reachable* as possible. So an interesting question arises: what if we don't want to reach A?, that is; how to construct a policy control that makes A as *forbidden* as possible?. Let $\mathbb{X}$ be a countable state space, and let $\mathbb{U}$ be a finite set of control actions. Fix a set $A \subset \mathbb{X}$, and assume that all control policies make A a closed set. We are going to be interested in the following set.

Definition 3.18.

$$\Gamma_F = \left\{ x \in \mathbb{X} \;\middle|\; \liminf_{t\to\infty} P_x^{\pi}(X_t \in A) = 0 \text{ for some policy } \pi \in \Pi_S \right\}$$

As in the previous sections, we have that

Lemma 3.19.

$$\Gamma_F = \{x \in \mathbb{X} \mid P_x^{\pi}(\tau_A < \infty) = 0 \textit{ for some policy } \pi \in \Pi_S\}$$

We will introduce now a value function characterizing Γ_F.

Definition 3.20. Let $v : \mathbb{X} \to \mathbb{R}$ be defined as,

$$v(x) := \sup_{\pi \in \Pi_S} v^{\pi}(x)$$

Where,

$$v^{\pi}(x) = \mathbb{E}_x^{\pi}\left[\sum_{t=0}^{\infty} \mathbb{1}_{A^c}(X_n) \cdot \gamma^t\right] \text{ for some } \gamma \in (0,1) \text{ and policy } \pi \in \Pi_S.$$

The characterization is the following,

Theorem 3.21.

$$\Gamma_F = \{x \in \mathbb{X} \mid v(x) = L\}$$

Proof. Let $x \in \Gamma_F$, so there exist a policy π such that $P_x^\pi(\tau_A < \infty) = 0$. Therefore, by Theorem 3.8 we have $v^\pi(x) = L$, which implies $v(x) = L$.

On the other hand, assume $x \in \Gamma_F^c$. So for all policies π, $P_x^\pi(\tau_A < \infty) > 0$. In particular for the optimal policy π^* we would have $P_x^{\pi^*}(\tau_A < \infty) > 0$ so by Theorem 3.8 $v(x) = v^{\pi^*}(x) < L$. $\square$

Chapter 4

Reaching a set of states A

In this chapter we will study a particular optimization problem. Let $\mathbb{X}$ denote the set of states and let $\mathbb{U}$ denote the set of control actions. Let ν be an initial distribution over the state space. Our objective will be to find a policy in Π_S, that is a stationary policy, that maximize the probability that the chain will eventually reach a set $A \subset \mathbb{X}$; that is,

$$\sup_{\pi \in \Pi_S} \left\{ \liminf_{t \to \infty} P_\nu^\pi (X_t \in A) \right\}$$

Note that the problem is well defined since for all control policies the quantity,

$$\liminf_{t \to \infty} P_\nu^\pi (X_t \in A)$$

is a positive real number less than or equal to 1. In order to guarantee some kind of stability let's assume that there exist a policy π such that A is a closed set. We will proceed as follows: In the first section, we will show that it's enough to consider controls that make A a closed set. Using this, we will be able to relate the problem with Zubov's method, such approach will lead us to an average cost function. In second section, we will use such average cost function to conclude the existence of an optimal policy, and how to calculate such policy. As an interesting corollary we will also obtain a way to calculate abortion probabilities via dynamic programming, and a way to describe the domain of attraction via average cost functions.

4.1 Relation with Zubov's Method

Recall that in Zubov's method we used a discounted value function to describe the domain of attraction. Therefore, one first approach is whether the calculation of such domain helps us to solve the proposed problem.

We have one first result, which tells us that it's enough to consider control policies that make A a closed set, and a different description to our problem.

Corollary 4.1.

$$\sup_{\pi\in\Pi_S}\left\{\liminf_{n\to\infty} P_\nu^\pi(X_n\in A)\right\} = \sup_{\pi\in\Pi_A}\left\{\lim_{n\to\infty} P_\nu^\pi(X_n\in A)\right\}$$

Moreover, the limit on the right is equal to $P_\nu^\pi(\tau_A<\infty)$.

Proof. Let π be an arbitrary stationary policy. We have that,

$$\begin{aligned}\liminf_{t\to\infty} P_\nu^\pi(X_t\in A) &\le \limsup_{t\to\infty} P_\nu^\pi(X_t\in A)\\ &= \limsup_{t\to\infty}\sum_{x\in\mathbb{X}} P_x^\pi(X_t\in A)\cdot\nu(x)\\ &\le \sum_{x\in\mathbb{X}}\limsup_{t\to\infty} P_x^\pi(X_t\in A)\cdot\nu(x)\end{aligned}$$

Let $\pi_{op}\in\Pi_A$ be defined as in lemma 3.11. Therefore, using such lemma,

$$\sum_{x\in\mathbb{X}}\limsup_{t\to\infty} P_x^\pi(X_t\in A)\cdot\nu(x) \le \sum_{x\in\mathbb{X}}\lim_{t\to\infty} P_x^{\pi_{op}}(X_t\in A)\cdot\nu(x)$$

Now, using of dominated convergence we obtain,

$$\begin{aligned}\sum_{x\in\mathbb{X}}\lim_{t\to\infty} P_x^{\pi_{op}}(X_t\in A)\cdot\nu(x) &= \lim_{t\to\infty}\sum_{x\in\mathbb{X}} P_x^{\pi_{op}}(X_t\in A)\cdot\nu(x)\\ &= \lim_{t\to\infty} P_\nu^{\pi_{op}}(X_t\in A)\end{aligned}$$

Therefore,

$$\liminf_{t\to\infty} P_\nu^\pi(X_t\in A) \le \lim_{t\to\infty} P_\nu^{\pi_{op}}(X_t\in A)$$

Finally, last limit is equal to $P_\nu^\pi(\tau_A<\infty)$ since,

$$\sum_{x\in\mathbb{X}}\lim_{t\to\infty} P_x^{\pi_{op}}(X_t\in A)\cdot\nu(x) = \sum_{x\in\mathbb{X}} P_x^\pi(\tau_A<\infty)\cdot\nu(x) = P_\nu^\pi(\tau_A<\infty)$$

□

Next lemma provides a relation between the proposed problem and the discounted value function characterizing the domain of attraction in Zubov's method.

Lemma 4.2.

$$\sup_{\pi\in\Pi_S}\left\{\liminf_{n\to\infty} P_\nu^\pi(X_n\in A)\right\} \ge 1-(1-\gamma)\sum_{x\in\mathbb{X}} v_\gamma^{\pi^*}(x)\nu(x)$$

where $\pi^*\in\Pi_A$ *is an optimal policy for the discounted problem.*

Proof. Because of previous corollary,

$$\sup_{\pi\in\Pi_S}\Big\{\liminf_{n\to\infty} P_\nu^\pi(X_n\in A)\Big\} = \sup_{\pi\in\Pi_A}\Big\{\lim_{n\to\infty} P_\nu^\pi(X_n\in A)\Big\}$$

Let $\pi\in\Pi_A$, and define $Z=\sum_{t=0}^{\infty}\mathbb{1}_{A^c}(X_t)\cdot\gamma^t$, $L=\frac{1}{1-\gamma}$. Because of Markov's inequality,

$$\begin{aligned} v_\gamma^\pi(x) &= \mathbb{E}_x^\pi[Z] \\ &\geq L\cdot P_x^\pi(Z\geq L) \\ &= L\cdot P_x^\pi(Z=L) \end{aligned}$$

Therefore,

$$\begin{aligned} (1-\gamma)v_\gamma^\pi(x)\nu(x) &\geq P_x^\pi(Z=L)\nu(x) \\ &= (1-P_x^\pi(Z<L))\nu(x) \end{aligned}$$

Reordering and recalling that because of lemma 3.7, $P_x^\pi(\tau_A<\infty)=P_x^\pi(Z<L)$, we obtain,

$$P_x^\pi(\tau_A<\infty)\nu(x)\geq\nu(x)\Big[1-(1-\gamma)v_\gamma^\pi(x)\Big]$$

So that,

$$P_\nu^\pi(\tau_A<\infty)\geq 1-(1-\gamma)\sum_{x\in\mathbb{X}}v_\gamma^\pi(x)\nu(x)$$

Therefore,

$$\sup_{\pi\in\Pi_A}\Big\{P_\nu^\pi(\tau_A<\infty)\Big\}\geq 1-(1-\gamma)\inf_{\pi\in\Pi_A}\Big\{\sum_{x\in\mathbb{X}}v_\gamma^\pi(x)\nu(x)\Big\}$$

Finally, since $v^{\pi^*}(x)\leq v^\pi(x)$ and $\pi^*\in\Pi_A$ we obtain,

$$\sum_{x\in\mathbb{X}}v_\gamma^{\pi^*}(x)\nu(x)=\inf_{\pi\in\Pi_A}\Big\{\sum_{x\in\mathbb{X}}v_\gamma^\pi(x)\nu(x)\Big\}$$

□

So we have shown a relation between the discounted value function and the proposed problem. Recall that because of lemma 1.21,

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

where v^π is the average cost function,

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

Therefore, in view of the previous lemma, the average cost function could provide another bound for such problem. Next section deals with such idea.

4.2 Average Cost Approach

We have the following important lemma, that relates the average cost function with hitting times.

Lemma 4.3. *Let $x \in \mathbb{X}$ and $\pi \in \Pi_A$. Then,*

$$\begin{aligned} v^\pi(x) &= 1 - \lim_{t\to\infty} P_x^\pi(X_t \in A) \\ &= 1 - P_x^\pi(\tau_A < \infty) \end{aligned}$$

Proof.

$$\begin{aligned} \frac{1}{N} \cdot \mathbb{E}_x^\pi\Big[\sum_{t=0}^{N-1} \mathbb{1}_{A^c}(X_t)\Big] &= \frac{1}{N} \cdot \sum_{t=0}^{N-1} \mathbb{E}_x^\pi\Big[\mathbb{1}_{A^c}(X_t)\Big] \\ &= \frac{1}{N} \cdot \sum_{t=0}^{N-1} P_x^\pi(X_t \notin A) \\ &= 1 - \frac{1}{N} \cdot \sum_{t=0}^{N-1} P_x^\pi(X_t \in A) \end{aligned}$$

Therefore, taking $\limsup$ on both sides,

$$\begin{aligned} v^\pi(x) &= \limsup_{t\to\infty} \Big[1 - \frac{1}{N} \cdot \sum_{t=0}^{N-1} P_x^\pi(X_t \in A)\Big] \\ &= 1 - \liminf_{t\to\infty} \frac{1}{N} \cdot \sum_{t=0}^{N-1} P_x^\pi(X_t \in A) \end{aligned}$$

Now, recall that given a sequence $\{s_t\}_{t=0}$ such that the limit exist,

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

then, the limit of the average exist and is equal to L, that is,

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

Since A is closed under π, we know

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

Therefore,

$$\lim_{N\to\infty} \frac{1}{N} \cdot \sum_{t=0}^{N-1} P_x^\pi(X_t \in A) = \lim_{t\to\infty} P_x^\pi(X_t \in A) = P_x^\pi(\tau_A < \infty)$$

□

Remark. Note that above lemma shows that the quantities $P_x(\tau_A < \infty)$ can be computed using dynamic programming. In case A is a closed class such quantities are called absorption probabilities.

As a corollary we obtain that the domain of attraction and escape set can be understood using average cost functions. Define,

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

Corollary 4.4.

$$\Lambda = \left\{ x \in \mathbb{X} \,\middle|\, v^*(x) < 1 \right\}$$
$$\Gamma = \left\{ x \in \mathbb{X} \,\middle|\, v^*(x) = 1 \right\}$$

Proof. Let $x \in \Lambda$, then corollary 3.12 implies $P_x^\pi(\tau_A < \infty) > 0$ for some policy $\pi \in \Pi_A$. In view of previous lemma $v^\pi(x) < 1$, and as a consequence $v^*(x) < 1$. On the other hand, let $x \in \Gamma$ so that corollary 3.12 implies $P_x^\pi(\tau_A < \infty) = 0$ for all polices $\pi \in \Pi_A$. Therefore, above lemma shows $v^\pi(x) = 1$ for all $\pi \in \Pi_A$ and as a consequence $v^*(x) = 1$. □

As an easy consequence, we also obtain the relation with the optimal average cost function.

Corollary 4.5.

$$\sup_{\pi \in \Pi_S} \left\{ \liminf_{t \to \infty} P_\nu^\pi(X_t \in A) \right\} = 1 - \inf_{\pi \in \Pi_A} \left\{ \sum_{x \in \mathbb{X}} v^\pi(x) \nu(x) \right\}$$

Moreover, in case the space state is finite and there are finite control actions $\mathbb{U}$*,*

$$\begin{aligned} \sup_{\pi \in \Pi_S} \left\{ \liminf_{t \to \infty} P_\nu^\pi(X_t \in A) \right\} &= \lim_{t \to \infty} P_\nu^{\pi^*}(X_t \in A) \\ &= 1 - \sum_{x \in \mathbb{X}} v^{\pi^*}(x) \nu(x) \end{aligned}$$

where π^* *is the optimal policy of the average cost problem (such policy exists, Theorem 1.23). Moreover,* π^* *is independent of the initial distribution* ν*.*

Proof. First of all, recall that because of corollary 4.1,

$$\sup_{\pi \in \Pi_S} \left\{ \liminf_{t \to \infty} P_\nu^\pi(X_t \in A) \right\} = \sup_{\pi \in \Pi_A} \left\{ \lim_{t \to \infty} P_\nu^\pi(X_t \in A) \right\}$$

Let $\pi \in \Pi_A$. Therefore, previous lemma implies,

$$v^\pi(x)\nu(x) = \nu(x) - \lim_{t \to \infty} P_x^\pi(X_t \in A)\nu(x)$$

Reordering terms and summing over $\mathbb{X}$ we obtain,

$$1 - \sum_{x \in \mathbb{X}} v^\pi(x)\nu(x) = \sum_{x \in \mathbb{X}} \lim_{t \to \infty} P_x^\pi(X_t \in A)\nu(x)$$

Using dominated convergence, last quantity it's equal to,

$$\lim_{t\to\infty}\sum_{x\in\mathbb{X}} P_x^\pi(X_t\in A)\nu(x) = \lim_{t\to\infty} P_\nu^\pi(X_t\in A)$$

So taking $\sup$ over Π_A we obtain the result. Finally, in case $\mathbb{X}$, $\mathbb{U}$ are finite, Theorem 1.23 guarantees the existence of an optimal policy π^*. Since the optimization is carried out in Π_A, $\pi^*\in\Pi_A$ so that,

$$\inf_{\pi\in\Pi_A}\left\{\sum_{x\in\mathbb{X}} v^\pi(x)\nu(x)\right\} = \sum_{x\in\mathbb{X}} v^{\pi^*}(x)\nu(x)$$

□

To end this section, we present a way to calculate the function v^*. Note that the optimization must be carried out on elements of Π_A, so we would have to calculate such set. However, this is not necessary as the following theorem shows.

Theorem 4.6. *Assume $\mathbb{X}$ and $\mathbb{U}$ are finite. Consider the average cost problem over Π_S,*

$$\inf_{\pi\in\Pi_S} v^\pi$$

Then, the optimal policy π^* (Theorem 1.23 guarantees such existence) belongs to Π_A. Therefore, $v^{\pi^*}=v^*$.

Proof. Note that under any policy π,

$$\begin{aligned} v^\pi(x) &= \limsup_{N\to\infty}\frac{1}{N}\mathbb{E}_x^\pi\left[\sum_{t=0}^{N-1}\mathbb{1}_{A^c}(X_t)\right] \\ &= \limsup_{N\to\infty}\frac{1}{N}\sum_{t=0}^{N-1}\mathbb{E}_x^\pi\left[\mathbb{1}_{A^c}(X_t)\right] \\ &= \limsup_{N\to\infty}\frac{1}{N}\sum_{t=0}^{N-1}P_x^\pi(X_t\notin A) \end{aligned}$$

Assume A is not closed under π^*. Therefore, there exist $i\in A$ and $j\notin A$ such that $P_i^\pi(X_1=j)>0$, so that $v^{\pi^*}(i)>0$. On the other hand, let $\pi_A\in\Pi_A$. Since A is closed under π_A, $P_i^\pi(X_t\notin A)=0$, which implies $v^\pi(i)=0$. A contradiction since $v^{\pi^*}=\inf_{\pi\in\Pi_S}v^\pi$. □

As a consequence the optimality equations are as follows. First of all, because of previous theorem,

$$\begin{aligned} v^* &= \inf_{\pi\in\Pi_S}\{v^\pi\} \\ &= 1-\widehat{v} \end{aligned}$$

where $\widehat{v}$ is defined as,

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

Therefore, because of the discussion in the average problem section, we have the following linear programs. Consider a vector $\nu \in \mathbb{R}_{\geq 0}^{|\mathbb{X}|}$.

Primal Linear Program

$$\text{Minimize} \sum_{x\in\mathbb{X}} \nu(x)\widehat{v}(x)$$

Subject to

$$\widehat{v}(x) \geq \sum_{j\in\mathbb{X}} P_{x,j}^u \widehat{v}(j) \qquad \text{for } u \in \mathbb{U}, x \in \mathbb{X}$$

And

$$\widehat{v}(x) \geq \mathbb{1}_A(x) + \sum_{j\in\mathbb{X}} P_{x,j}^u h(j) - h(x) \qquad \text{for } u \in \mathbb{U}, x \in \mathbb{X}$$

Dual Linear Program

$$\text{Maximize} \sum_{s\in\mathbb{X}} \sum_{u\in\mathbb{U}} \mathbb{1}_A(s) x(s,u)$$

Subject to

$$\sum_{u\in\mathbb{U}} x(j,u) - \sum_{s\in\mathbb{X}} \sum_{u\in\mathbb{U}} P_{s,j}^u x(s,u) = 0 \qquad j \in \mathbb{X}$$

And

$$\sum_{u\in\mathbb{U}} x(j,u) + \sum_{u\in\mathbb{U}} y(j,u) - \sum_{s\in\mathbb{X}} \sum_{u\in\mathbb{U}} P_{s,j}^u y(s,u) = \nu(j) \qquad j \in \mathbb{X}$$

$$\text{And } x(j,u) \geq 0, \quad y(j,u) \geq 0 \qquad u \in \mathbb{U}, j \in \mathbb{X}$$

Example 4.7. Let $\mathbb{X} = \{1,2,3,4,5,6,7\}$ and $\mathbb{U} = \{1,2\}$, where the transition matrices P^1 and P^2 are the same as in example 2.8. Let $A = \{2,3,4\}$, recall that in that example we proved there exist a policy such that A is closed. Consider an homogeneous initial distribution ν. By solving the dual problem we obtain,

$$X = \begin{bmatrix} 0 & 0 \\ 0 & 1.0373 \\ 1.2830 & 0.6346 \\ 0.5078 & 1.0373 \\ 0.7179 & 0.6244 \\ 0.6244 & 0.5333 \\ 0 & 0 \end{bmatrix} \quad Y = \begin{bmatrix} 1.6667 & 0 \\ 0 & 14.0391 \\ 21.0842 & 0 \\ 10.5654 & 13.5764 \\ 0 & 0.1577 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}$$

As a consequence an optimal policy is,

$$\pi^* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0.6691 & 0.3309 \\ 0.3286 & 0.6714 \\ 0.5348 & 0.4652 \\ 0.5393 & 0.4607 \\ 1 & 0 \end{bmatrix}$$

So that v^* and $\widehat{v}$ are given by,

$$\widehat{v} = \begin{bmatrix} 0.5 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \qquad v^* = \begin{bmatrix} 0.5 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

As a consequence,

$$\sup_{\pi \in \Pi_S} \left\{ \liminf_{t \to \infty} P_\nu^\pi (X_t \in A) \right\} = \lim_{t \to \infty} P_\nu^{\pi^*} (X_t \in A) = 0.6429$$

Note that, as in the discounted case, v^* is characterizing the domain of attraction.

Conclusions and Future Directions

Along this work we solved three related-stability problems, providing a theoretical solution for each one, and a computational approach to calculate them.

- We considered a partition of the state space, and found a policy such that the induced Markov chain has a maximum number of recurrent states avoiding states that belong to another piece of the partition. We also provided a convex program to calculate such policy.
- We provided a Zubov's method for controlled Markov chains. Finding two value functions that characterize the domain of attraction and escape set, a discounted cost function and an average cost function. We provided a linear program to calculate such functions and an optimal policy. The equations that describe the discounted cost function happen to be easier than the ones that describe the average cost function.
- We studied the problem of finding a policy that maximizes the probability of *reaching* certain set. We showed discount cost functions can be used to approximate such value, and how average cost functions solve exactly the problem. Since the calculation of average and discount cost functions can be accomplished using linear programming, we found a linear programming approach to solve the desired problem.

Some questions and problems that arise while doing this work, that would be interesting to solve are:

- In Zubov's Method the hypothesis of the set of states A being closed is fundamental. How to verify such hypothesis? An idea is to use a convex program similar to the one presented in Chapter 2.
- Generalize Zubov's Method for continuous times Markovian processes and more general state spaces. In such case what kind of programs can be used to calculate or at least approximate the value function?
- What if we want to maximize the probability of reaching some set A but minimize the probability of reaching a set B? that is,

$$\sup_{\pi\in\Pi_S}\left\{\liminf_{t\to\infty} P_\nu^\pi(X_t\in A)-\liminf_{t\to\infty} P_\nu^\pi(X_t\in B)\right\}$$

Bibliography

[1] E Arvelo and NC Martins, *Control design for markov chains under safety constraints: a convex approach, 2012*, arXiv preprint arXiv:1209.2883.

[2] Eduardo Arvelo and Nuno C Martins, *Maximizing the set of recurrent states of an mdp subject to convex constraints*, Automatica **50** (2014), no. 3, 994–998.

[3] Dimitri P Bertsekas, *Dynamic programming and optimal control*, vol. 1.

[4] Dimitri P Bertsekas and Steven E Shreve, *Stochastic optimal control: The discrete time case*, vol. 23, Academic Press New York, 1978.

[5] Fabio Camilli, Lars Grüne, and Fabian Wirth, *Control lyapunov functions and zubov's method*, SIAM Journal on Control and Optimization **47** (2008), no. 1, 301–326.

[6] Fabio Camilli and Paola Loreti, *A zubovs method for stochastic differential equations*, Nonlinear Differential Equations and Applications NoDEA **13** (2006), no. 2, 205–222.

[7] Wolfgang Hahn and Arne P Baartz, *Stability of motion*, vol. 138, Springer, 1967.

[8] Martin L Puterman, *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons, 2014.

[9] Sidney I Resnick, *Adventures in stochastic processes*, Springer Science & Business Media, 2013.

[10] David Williams, *Probability with martingales*, Cambridge university press, 1991.

[11] Vladimir Ivanovich Zubov, *Methods of am lyapunov and their application*, P. Noordhoff, 1964.