"El saber de mis hijos hará mi grandeza"

UNIVERSIDAD DE SONORA

División de Ciencias Exactas y Naturales

Departamento de Matemáticas

Partially Observable Markov Control Models with Random Discount Factors

T E S I S

Que para obtener el grado académico de:

Maestro en Ciencias (Matemáticas)

Presenta:

Edgar Everardo Martinez Garcia

Director de Tesis:

Dr. Jesús Adolfo Minjárez Sosa

Hermosillo, Sonora, México, 28 de Julio 2021

SINODALES

Dra. Carmen Geraldi Higuera Chan
Universidad de Sonora

Dr. Jesús Adolfo Minjárez Sosa
Universidad de Sonora

Dr. Oscar Vega Amaya
Universidad de Sonora

Dr. Yofre Hernán García Gómez
Universidad Autónoma de Chiapas

Agradecimientos

Quiero comenzar agradeciendo a mis padres y hermanos, que desde que me propuse este objetivo y a pesar de estar lejos en ocasiones, siempre me hicieron sentir su apoyo incondicional.

A Katerine, gracias por tu paciencia y el apoyo que me diste desde la licenciatura sin ti tal vez no hubiera podido concluir mis estudios con éxito, gracias por haber estado ahí.

A Janeth, que aunque ella no lo entienda aún, ha sido una parte muy importante en mi formación y de la persona en la que me he convertido, gracias por todo tu cariño.

A mi profesor y Director de Tesis, Dr. Adolfo Minjárez, gracias por su paciencia y por las horas dedicadas a la preparación y revisión de esta tesis.

A mis sinodales, por los consejos y el tiempo dedicado a la revisión de este trabajo.

Por último quiero agradecer a la Universidad de Sonora y al Consejo Nacional de Ciencia y Tecnología (CONACyT) por el apoyo económico brindado y que me permitió concluir mis estudios de posgrado, apoyo sin el cual no hubiera sido posible.

Edgar Everardo Martinez Garcia
Hermosillo, Sonora
Julio 2021

Contents

Introduction

Three elements are needed to define an optimal control problem (OCP): (1) a decision or control model; (2) a set of admissible control policies, and (3) a performance index. So, the OCP is to find a control policy that minimizes the performance index. We can classify the OCPs in a variety of ways, for instance: deterministic or stochastic; continuous or discrete time; finite or infinite horizon; discounted or average cost performance index; with partial or complete state-system information, among others.

In this work we focus on the study of OCPs considering partially observable discrete-time stochastic systems, under a discounted optimality criterion, but unlike the standard case, we assume random discount factors.

Specifically, we study the OCP associated to partially observable Markov decision processes (POMDPs) in the following scenarios:

1. Observable random discount factors and unknown distribution.
2. Unobservable random discount factors and unknown distribution.
3. Partially observable random discount factors.

For Problem 1, we propose an estimation and control procedure (see e.g. [14–16, 21]) to prove the existence of asymptotically discounted optimal policies. Instead, because of the non-observability of the discount factors, we model the Problem 2 as a minimax control problem. That is, we assume that the controller has an opponent who selects the unknown distribution at each stage, and the objective is to minimize the maximum cost generated precisely by such an opponent. This class of problem is viewed as *game against nature* (see e.g. [17, 21]).

Finally, Problem 3 is analized as a two coupled partially observable stochastic systems, namely, the discount process and the state process. This defines, in a natural way, an extended POMDP which is treated accordingly.

In general terms, the POMDPs associated to the Problems 1-3 will be studied following a standard procedure (see e.g. [7,10,19,23]) which consists to transform the corresponding OCP into a complete observable optimal control problem using a filtering process defined on a suitable space of probability measures. This standard procedure has been applied to analyze specific problems in several fields as inventory systems, queueing systems, economic and financial models (see e.g. [2–5, 11, 18]), where the dynamic of the state's process $\{x_t\}$ is given by a stochastic difference equation of the form

$$x_{t+1} = F(x_t, a_t, \xi_t), \quad t \in \mathbb{N}_0 \tag{I.1}$$

In this case F is a known function, a_t represents the control selected by the controller or decision-maker, and $\{\xi_t\}$ is a sequence of independent and identically distributed random variables known as the disturbance process. Moreover, the observation process $\{y_t\}$ is given as

$$y_t = G(a_t, x_t, \eta_t), \tag{I.2}$$

where G is a known function and $\{\eta_t\}$ is a sequence of i.i.d. random variables, independent on $\{\xi_t\}$.

For example, in an inventory system, let x_t and a_t the stock and the stock ordered at the beginning of the $t-$th stage, and ξ_t the demand during the $t-$th stage. Then, the equation (I.1) takes the form

$$x_{t+1} = \max\{x_t + a_t - \xi_t, 0\}.$$

Consider the situation of a large store where there is no counter serving the customer directly. In this case, the demand of the article is not completely observed, but rather what is observed are the sales thorough the cash register. Hence, equation (I.2) takes the form

$$y_t = \min\{x_t + a_t, \xi_t\}$$

with $\eta_t = \xi_t$.

Another example with partially observable state process is a queueing system with controlled service rate where x_t and u_t are the waiting time and the service rate of the $t-$th customer, respectively, and ξ_t is the interarrival time between the $t-$th and $(t+1)-$th customers. As is shown in [11], if γ_t represents a random "base" service time of the $t-$th customer and a_t is the control defined as $a_t = 1/u_t$, the process $\{x_t\}$ evolves as

$$x_{t+1} = \max\{x_t + a_t\gamma_t - \xi_t, 0\}, \quad t \in \mathbb{N}_0.$$

Consider that the waiting time is partially observable because only is observed when $x_t = 0$, which means that the controller only checks when the customer does not queue or arrives directly at the server.

In other words, the controller cannot register the waiting time of a customer when the server is busy. Hence, the observation process $\{y_t\}$ is defined as

$$y_t = 1_{[x_t=0]}, t \in \mathbb{N}.$$

On the other hand, assuming random discount factors, we generalize the standard case where it is considered constant. This fact is important both from the theoretical point of view and from applications. The later includes for example, financial models where the discount factors usually depend on interest rates, which, in turn, are random.

As is well known, there is great uncertainty surrounding interest rates in financial markets. This uncertainty is due either to the behavior and decisions of investors, to political decisions of governments, weather issues, among many others. Therefore, it is feasible to assume that the random variable that defines the discount factor is partially observable in the sense of equations (I.1) and (I.2). That is, we assume that the discount factor is a stochastic process $\{\alpha_t\}$, where α_t represents the discount factor at time t, evolving as

$$\alpha_{t+1} = G_1(\alpha_t, \xi_t^{(1)})$$

with observation process

$$\beta_t = G_2(\alpha_t, \xi_t^{(2)})$$

where G_1 and G_2 are known functions and $\{\xi_t^{(1)}\}$ and $\{\xi_t^{(2)}\}$ are independent sequences random variables.

In [14] is introduced an example of a consumption-investment problem where the discount factor process $\{\alpha_t\}$ evolves according to an autoregresive process of the form

$$\alpha_{t+1} = h\alpha_t + \xi_t^{(1)},$$

for some suitable $h > 0$. In this example, an observation process could be

$$\beta_t = I_{[\alpha_t \leq \overline{\alpha}]},$$

which means that the discount factor only is observed while it remains below a threshold $\overline{\alpha} \in (0, 1)$.

Although POMDP and control problems with random discount factors (see e.g. [12–16, 21]) have both studied separately, to the best of our knowledge, this is the first work dealing with these two problems together, which constitutes the main motivation and objective in this thesis.

The work is structured in the following manner: In Chapter 1, we introduce basic definitions and results on COMDPs and POMDPs, and present general conditions under which there exist an optimal policy. In Chapter 2, we present the OCP for POMDP with observable random discount factors and then prove the existence of an asymptotically random-discounted policy. In chapter 3, we show the case with unobservable random discount factors and prove the existence of a minimax policy. In chapter 4, we address the problem of the partially observable discount factors and show the existence of an optimal policy.

Chapter 1

Preliminaries on Partially Observable Markov Control Models

1.1 Introduction

In this chapter we will stablish the most important elements on the theory of partially observable Markov control models. These include the techniques and results that we will apply in the development of the following chapters.

In particular the class of partially observable optimal control problems we are interested will be analyzed by applying standard approach which consists of transforming them into new control problems that are completely observable but defined in appropiate measure spaces. In this sense, for the sake of organizing the theory, we first introduce the main facts related with the standard completely observable Markov decision processes. These include the corresponding control model, sets of control policies, and the discounted performance index.

1.2 Completely observable Markov decision processes

In this section we introduce the completely observable Markov decision process and the optimal control problem with infinite horizon.

Definition 1. *A discrete time completely observable control model* consists of five objects:

$$\mathcal{M}_{CO} = (X, A, Q, \underline{\nu}, c), \tag{1.2.1}$$

where X and A are the state and action spaces, respectively, and both are assumed to be Borel spaces; Q, the transition law, is a stochastic kernel of X given $X \times A$; $\underline{\nu} \in \mathbb{P}(X)$ is a probability measure called the initial distribution; and c a bounded measurable function from $X \times A$ to $\mathbb{R}$ called cost-per-stage function.

The control model $\mathcal{M}_{CO}$ represents a controlled stochastic system in which the states are completely observable at times $t \in \mathbb{N}_0$. The dynamics of the system can be described as follows: at time $t = 0$, the state x_0 is a random variable with distribution $\underline{\nu}$. Now if at time t, the state of the system is $x_t = x \in X$, the control $a_t = a \in A$ is applied. Then two things occur: (1) a cost $c(x, a)$ is generated, and (2) the system moves to a new state x_{t+1} according to the probability distribution $Q(\cdot|x, a)$ on X. If the new state is $x_{t+1} = x'$, a control $a_{t+1} = a'$ is chosen and the process is repeated, over and over again.

There are situations where the dynamics of the system is defined by a stochastic difference equation of the form

$$x_{t+1} = F(x_t, a_t, \xi_t), \quad t \in \mathbb{N}_0, \tag{1.2.2}$$

where $\{\xi_t\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with values in a space S, with common distribution θ, independent of the initial state x_0 and known as the *state-disturbance process*. In this case if $x_0 = x$ the initial distribution $\underline{\nu}$ is the probability measure concentrated at $x \in X$. Further, the transition law is given by

$$\begin{aligned} Q(B|x,a) &= P(F(x_t, a_t, \xi_t) \in B | x_t = x, a_t = a) \\ &= \theta(\{s \in S | F(x,a,s) \in B\}) \\ &= \int_S 1_B[F(x,a,s)]\theta(ds), B \in \mathcal{B}(X). \end{aligned} \tag{1.2.3}$$

The actions $a \in A$ are chosen according to rules known as *control policies*. Furthermore the dynamics of the system defines a vector, called *t-history*, $h_t \in \mathbb{H}_t := (X \times A)^t \times X$, of the form

$$h_t := (x_0, a_0, ..., x_{t-1}, a_{t-1}, x_t), \quad t \in \mathbb{N}_0.$$

Specifically, the definition of a control policy is as follows:

Definition 2. A *control policy* is a sequence of stochastic kernels $\pi = \{\pi_t\}$, $t \in \mathbb{N}_0$ on A given $\mathbb{H}_t$. A control policy is *Markovian (deterministic)* if exists a sequence of measurable functions $f_t : X \to A$, such that for all $h_t \in \mathbb{H}_t$, $C \in \mathcal{B}(A)$ and $t \in \mathbb{N}_0$

$$\pi_t(C|h_t) = 1_C[f_t(x_t)],$$

and *stationary* if there exists a measurable function $f : X \to A$, such that for all $h_t \in \mathbb{H}_t$ y $t \in \mathbb{N}_0$

$$\pi_t(C|h_t) = 1_C[f(x_t)],$$

where 1_C is the indicator function of the set C.

We denote by $\mathbb{F}$ the set of all decision functions (or selectors), i.e., measurable functions $f : X \to A$. As usual, a stationary policy is denoted by f^∞, taking the form $f^\infty := \{f\}$.

We denote by Π the set of all control policies and by Π_M the set of Markov policies. Following a standard convention, we denote by $\mathbb{F}$ the set of stationary policies. Hence $\mathbb{F} \subset \Pi_M \subset \Pi$.

Now, we are ready to establish the COMDP. Let $(\Omega, \mathcal{F})$ be a measurable space where $\Omega := (X \times A)^\infty$ and $\mathcal{F} := \mathcal{B}(\Omega) = \mathcal{B}((X \times A)^\infty)$ the corresponding product $\sigma-$algebra. Note that the elements of Ω are of the form

$$\omega = (x_0, a_0, x_1, a_1, ...).$$

Let $\pi \in \Pi$ be a control policy and $\underline{\nu}$ be an arbitrary probability measure on X. Then by Ionescu Tulcea Theorem (see Poposition C.2), there exists a unique probability measure $P^\pi_{\underline{\nu}}$ on $(\Omega, \mathcal{F})$, such that for all $B \in \mathcal{B}(X)$, $C \in \mathcal{B}(A)$ and $h_t \in \mathbb{H}_t$

$$\begin{aligned} P^\pi_{\underline{\nu}}(x_0 \in B) &= \nu(B) \\ P^\pi_{\underline{\nu}}(a_t \in C | h_t) &= \pi_t(C|h_t) \\ P^\pi_{\underline{\nu}}(x_{t+1} \in B | h_t, a_t) &= Q(B|x_t, a_t) \end{aligned} \tag{1.2.4}$$

Definition 3. The stochastic process $(\Omega, \mathcal{F}, P^\pi_{\underline{\nu}}, \{x_t\})$ is called a COMDP.

1.3 Optimal control problem for COMDP

In order to define the optimal control problem (OCP) for COMPD with infinite horizon, we consider the control model (1.2.1). Thus for a control policy $\pi \in \Pi$ and initial distribution $\underline{\nu} \in \mathbb{P}(X)$ we define the *total expected discounted cost* as

$$V(\pi, \underline{\nu}) := E^{\pi}_{\underline{\nu}}\left[\sum_{t=0}^{\infty} \beta^t c(x_t, a_t)\right], \tag{1.3.1}$$

where $E^{\pi}_{\underline{\nu}}$ is the expectation with respect to the probability measure $P^{\pi}_{\underline{\nu}}$, induced by π and $\underline{\nu}$, and $\beta \in (0,1)$ is the so-called *discount factor*. Furthermore, the *optimal cost function or value function* is given by

$$V^*(\underline{\nu}) := \inf_{\pi \in \Pi} V(\pi, \underline{\nu}), \quad \underline{\nu} \in \mathbb{P}(X).$$

So we can define the OCP for a COMDP as follows:

Definition 4. Given the control model $\mathcal{M}_{CO}$, introduced in (1.2.1), a family of control policies Π, and the performance index V, the OCP is to find a policy $\pi^* \in \Pi$ such that

$$V(\pi^*, \underline{\nu}) = V^*(\underline{\nu}), \quad \forall \underline{\nu} \in \mathbb{P}(X). \tag{1.3.2}$$

In this case π^* is called a discounted optimal policy.

1.4 Solution to OCP for the COMDP

For simplicity, we consider that the initial distribution $\underline{\nu} \in \mathbb{P}(X)$ is concentrated in $x_0 = x$. So, from (1.3.1), the total expected discount cost takes the form

$$V(\pi, x) := E^{\pi}_{x}\left[\sum_{t=0}^{\infty} \beta^t c(x_t, a_t)\right].$$

Then, by (1.3.2), the optimal control problem is to find a policy $\pi^* \in \Pi$ such that

$$V(\pi^*, x) = V^*(x) := \inf_{\pi \in \Pi} V(\pi, x), \quad \forall x \in X.$$

For the purposes of this work, we introduce the following weaker optimality criterion (weaker in the sense that optimality implies the next optimality criterion):

Definition 5. A policy $\pi \in \Pi$ is *asymptotically discount optimal* if, for every $x \in X$,

$$\left|V_n(\pi, x) - E^{\pi}_{x}[V^*(x_n)]\right| \to 0, \quad \text{as } n \to \infty$$

where,

$$V_n(\pi, x) := E^{\pi}_{x}\left[\sum_{t=n}^{\infty} \beta^{t-n} c(x_t, a_t)\right]$$

is the total discounted cost from stage n onward.

In order to show the existence of an optimal policy, we impose the following conditions on our model. Let $\mathcal{C}(X)$ be the Banach space of bounded and continuous functions on X with the supremum norm

$$||v|| = \sup_{x \in X} |v(x)|.$$

Assumption 6.

1. A is a compact set.
2. $|c(x,a)| \leq b$ for all $(x,a) \in X \times A$ and continuous in $a \in A$. Therefore, the total discounted cost is uniformly bounded: $|V(\pi,x)| \leq b/(1-\beta)$, for all policy π and initial state x.
3. The kernel Q is weakly continuous, that is, $\int_X u(x')Q(dx'|x,a)$ is a bounded and continuous function on $a \in A$, for each $x \in X$ and each function $u \in \mathcal{C}(X)$.

The solution of the OCP for the CO case can be obtained by applying the following *contraction mapping approach* (see [19]):

1. We define a suitable operator $T : \mathcal{C}(X) \to \mathcal{C}(X)$ and show that T is a *contraction operator* on $\mathcal{C}(X)$.
2. Hence, by the Banach's Fixed-Point Theorem for contraction operators (see Proposition A.1), there exists a unique function $u^* \in \mathcal{C}(X)$ such that $u^* = Tu^*$.
3. Finally, one shows that u^* equals the optimal discounted cost V^*, so that V^* is the unique bounded solution to the optimality equation.

In order to develope the above approach, for a function $u : X \to \mathbb{R}$ in $\mathcal{C}(X)$, define the *dynamic programming operator* by

$$Tu(x) = \min_{a \in A} \left\{ c(x,a) + \beta \int_X u(x')Q(dx'|x,a) \right\}, \quad x \in X \tag{1.4.1}$$

and, for $f^\infty \in \mathbb{F}$ the operator

$$T_f u(x) = c(x, f(x)) + \beta \int_X u(x')Q(dx'|x, f(x)). \tag{1.4.2}$$

Lemma 7. Under Assumption 6, we have:

1. $Tu(x) \in \mathcal{C}(X)$ for all $u \in \mathcal{C}(X)$.
2. The operators T and T_f are contraction operators modulus β.
3. There exist a unique function $u^* \in \mathcal{C}(X)$ and a unique function $u_f^* \in \mathcal{C}(X)$ such that

$$Tu^* = u^* \quad \text{and} \quad Tu_f^* = u_f^*,$$

and moreover, for any function $u \in \mathcal{C}(X)$

$$||T^n u - u^*|| \to 0 \quad \text{and} \quad ||T^n u_f - u_f^*|| \to 0, \quad \text{as } n \to \infty,$$

where T^n is defined recursively by $T^n := T(T^{n-1})$, for all $n = 1, 2, ...$ where T^0 is the identity.

Proof.

1. Let $u \in \mathcal{C}(X)$. It is clear that from Assumption 6.2 and 6.3

$$v(x,a) := c(x,a) + \beta \int_X u(x')Q(dx'|x,a)$$

is continuous on $a \in A$. Then, from Proposition D.2, and Assumption 6.1, we have that $\min_{a \in A} v(x,a)$ is continuous on $x \in X$. Hence, $Tu(x) \in \mathcal{C}(X)$.

2. Let $u, u' \in \mathcal{C}(X)$. Then,

$$\begin{aligned}
&\left|Th(x) - Th'(x)\right| \\
&\qquad = \left| \min_{a\in A}\left\{c(x,a) + \beta \int_X h(x')Q(dx'|x,a)\right\} - \min_{a\in A}\left\{c(x,a) + \beta \int_X h'(x')Q(dx'|x,a)\right\}\right| \\
&\qquad \leq \max_{a\in A}\left|c(x,a) + \beta \int_X h(x')Q(dx'|x,a) - c(x,a) - \beta \int_X h'(x')Q(dx'|x,a)\right| \\
&\qquad = \beta \max_{a\in A}\left| \int_X \Big(h(x') - h'(x')\Big)Q(dx'|x,a)\right| \\
&\qquad \leq \beta \max_{a\in A} \int_X |h(x') - h'(x')|Q(dx'|x,a) \\
&\qquad \leq \beta ||h - h'||
\end{aligned}$$

Therefore,

$$||Th - Th'|| \leq \beta ||h - h'||.$$

Similarly we prove

$$||T_f h - T_f h'|| \leq \beta ||h - h'||.$$

3. This part follows by applying the Banach Fixed Point Theorem to the operator T and T_f (see Proposition A.1).

Then, Lemma 7 has been proved. □

Now, we will prove that the fixed point u_f^* is indeed $V(f^\infty, x)$.

Lemma 8. Under Assumption 6, we have:

1. For every $x \in X$, $u_f^*(x) = V(f^\infty, x)$ for an arbitrary stationary policy $f^\infty \in \mathbb{F}$.
2. A policy π^* is optimal if and only if its total expected cost satisfies $TV(\pi^*, x) = V(\pi^*, x)$ for every $x \in X$.

Proof.

1. By the uniqueness of the fixed point it is enough to prove that $TV(f^\infty, x) = V(f^\infty, x)$. To this end, we write $V(f^\infty, x)$ as

$$V(f^\infty, x) = E_x^{f^\infty}\left[\sum_{t=0}^{\infty} \beta^t c(x_t, a_t)\right] = c(x, f(x)) + \beta E_x^{f^\infty}\left[\sum_{t=1}^{\infty} \beta^{t-1} c(x_t, a_t)\right]. \tag{1.4.3}$$

Then, by standard properties of the expectation operator, we have

$$\begin{aligned} E_x^{f^\infty}\left[\sum_{t=1}^{\infty}\beta^{t-1}c(x_t,a_t)\right] &= E_x^{f^\infty}\left[E_x^{f^\infty}\left(\sum_{t=1}^{\infty}\beta^{t-1}c(x_t,a_t)\middle| h_1\right)\right] \\ &= E_x^{f^\infty}\left[E_{x_1}^{f^\infty}\left(\sum_{t=1}^{\infty}\beta^{t-1}c(x_t,a_t)\right)\right] \\ &= E_x^{f^\infty}\left[E_{x_1}^{f^\infty}\left(\sum_{t=0}^{\infty}\beta^{t}c(x_{t+1},a_{t+1})\right)\right] \\ &= E_x^{f^\infty}[V(f^\infty,x_1)] \\ &= \int V(f^\infty,x')Q(dx'|x,f(x)), \end{aligned} \tag{1.4.4}$$

where the last equation comes from (1.2.4).

Hence, combining (1.4.3) and (1.4.4) we obtain

$$V(f^\infty,x) = c(x,f(x)) + \beta\int V(f^\infty,x')Q(dx'|x,f(x)) = TV(f^\infty,x), \quad \forall x\in X,$$

which proves part 1.

2. Let π^* be a policy such that $V(\pi^*,x') = TV(\pi^*,x')$, i.e.,

$$V(\pi^*,x) = \min_{a\in A}\left\{c(x,a) + \beta\int V(\pi^*,x')Q(dx'|x,a)\right\}, \quad \forall x\in X.$$

To prove that π^* is optimal we need to show that $V(\pi^*,x') \le V(\pi,x)$ for every policy π and initial state $x\in X$. For a history $h_t\in\mathbb{H}_t$, it follows from the Markov property (1.2.4) that

$$\begin{aligned} E_x^\pi[\beta^{t+1}V(\pi^*,x_{t+1})|h_t,a_t] &= \beta^{t+1}\int V(\pi^*,x')Q(dx'|x_t,a_t) \\ &= \beta^t\left\{c(x_t,a_t) + \beta\int V(\pi^*,x')Q(dx'|x_t,a_t)\right\} - \beta^t c(x_t,a_t) \\ &\ge \beta^t V(\pi^*,x_t) - \beta^t c(x_t,a_t). \end{aligned}$$

Equivalently

$$-E_x^\pi[\beta^{t+1}V(\pi^*,x_{t+1})|h_t,a_t] + \beta^t V(\pi^*,x_t) \le \beta^t c(x_t,a_t).$$

Taking expectations E_x^π on both sides of this inequality and summing over $t = 0,1,...,n$ we obtain (telescoping series),

$$\sum_{t=0}^{n}\left\{E_x^\pi[\beta^t V(\pi^*,x_t)] - E_x^\pi[\beta^{t+1}V(\pi^*,x_{t+1})]\right\} \le E_x^\pi\left[\sum_{t=0}^{n}\beta^t c(x_t,a_t)\right].$$

which yields

$$V(\pi^*,x) - \beta^{n+1}E_x^\pi[V(\pi^*,x_{n+1})] \le E_x^\pi\left[\sum_{t=0}^{n}\beta^t c(x_t,a_t)\right].$$

Now, since $V(\pi, x)$ is bounded for all $\pi \in \Pi$, letting $n \to \infty$, we get $V(\pi^*, x) \leq V(\pi, x)$, the desired conclusion, i.e., π^* is optimal.

Lastly, we prove the sufficient condition. Let π^* be an optimal policy. We will show that $V(\pi^*, x)$ satisfies: (1) $V(\pi^*, x) \geq TV(\pi^*, x)$ and (2) $V(\pi^*, x) \leq TV(\pi^*, x)$, hence $V(\pi^*, x)$ satisfies the optimality equation. We express $V(\pi^*, x)$ as

$$V(\pi^*, x) = E_x^{\pi^*}\left[\sum_{t=0}^{\infty} \beta^t c(x_t, a_t)\right] = \int_A \left\{c(x, a) + \beta \int_X V(\pi^{*(1)}, x')Q(dx'|x, a)\right\}\pi_0^*(da|x),$$

where $\pi^{*(1)} = \{\pi_t^{*(1)}\}$ denotes the "1-shifted" policy, i.e., with $x_0 = x$ and $a_0 = a$,

$$\pi_t^{*(1)}(\cdot|h_t) := \pi_{t+1}^*(\cdot|x_0, a_0, h_t), \quad t = 0, 1, ...$$

Thus, since π^* is optimal,

$$\begin{aligned} V(\pi^*, x) &\geq \int_A \left\{c(x, a) + \beta \int_X V(\pi^*, x')Q(dx'|x, a)\right\}\pi_0^*(da|x) \\ &\geq \min_{a \in A}\left\{c(x, a) + \beta \int_X V(\pi^*, x')Q(dx'|x, a)\right\} \\ &= TV(\pi^*, x) \end{aligned}$$

To prove (2), let $f \in \mathbb{F}$ be an arbitrary stationary policy, and let $\pi' := (f, \pi^*)$ be the policy that uses f at the initial stage $t = 0$, and the optimal policy π^* from stage $t = 1$ onwards. Because π^* is optimal we have

$$V(\pi^*, x) \leq V(\pi', x) = c(x, f(x)) + \beta \int V(\pi^*, x')Q(dx'|x, f(x)), \quad \forall x \in X.$$

hence, since $f \in \mathbb{F}$ is arbitrary, we finally obtain

$$V(\pi^*, x) \leq \min_{a \in A}\left\{c(x, a) + \beta \int V(\pi^*, x')Q(dx'|x, a)\right\} = TV(\pi^*, x).$$

□

Theorem 9. *Suppose that Assumption 6 holds. Then:*

(a) The optimal discounted cost function $V^* : X \to \mathbb{R}$ is the unique solution in $\mathcal{C}(X)$ of the optimality equation, that is

$$V^*(x) = \min_{a \in A}\left\{c(x, a) + \beta \int_X V^*(x')Q(dx'|x, a)\right\}, \quad x \in X.$$

(b) There exists $f^* \in \mathbb{F}$ such that

$$V^*(x) = c(x, f^*(x)) + \beta \int_X V^*(x')Q(dx'|x, f^*(x)), \quad x \in X \tag{1.4.5}$$

Moreover, the stationary policy $f_*^\infty = \{f^*\}$ is an optimal control policy.

Proof.

(a) Let $\pi^* \in \Pi$ be an optimal policy, that is $V^*(x) = V(\pi^*, x)$, for all $x \in X$. Then, from Lemma 8.2 we have $TV^*(x) = V^*(x)$, for all $x \in X$. Finally from the uniqueness of the fixed point of the operator T, we prove that V^* is the unique solution to the optimality equation.

(b) The existence of $f^* \in \mathbb{F}$ satisfaying (1.4.5), follows from Proposition D.2. On the other hand iterating (1.4.5) we have

$$\begin{aligned} V^*(x) &= E_x^{f_*^\infty}\left[\sum_{t=0}^{n} \beta^t c(x_t, a_t)\right] + \beta^{n+1} E_x^{f_*^\infty} V^*(x_{n+1}) \\ &\geq E_x^{f_*^\infty}\left[\sum_{t=0}^{n} \beta^t c(x_t, a_t)\right], \quad \forall n \in \mathbb{N}, x \in X. \end{aligned}$$

Thus, letting $n \to \infty$ we obtain

$$V^*(x) \geq V(f_*^\infty, x), \quad x \in X.$$

However, since $V^*(x) = \inf_{\pi \in \Pi} V(\pi, x)$, we conclude that

$$V^*(x) = V(f_*^\infty, x), \quad x \in X,$$

which proves that f_*^∞ is an optimal control policy.

□

Remark 10. Let v_t a sequence of functions in $\mathcal{C}(X)$ defined as $v_0 = v$ and

$$v_t(x) = T^t v(x) = T v_{t-1}(x) := \min_{a \in A}\left\{c(x, a) + \beta \int_X v_{t-1}(x') Q(dx'|x, a)\right\}, \quad t \geq 1, \quad x \in X.$$

Note that by the contraction property of T (see Lemma 7),

$$||v_t - u^*|| = ||Tv_{t-1} - Tu^*|| \leq \beta ||v_{t-1} - u^*||,$$

so that $||v_t - u^*|| \leq \beta^t ||v_0 - u^*||$, for all $t \geq 0$. Therefore, from Theorem 9 we have

$$||v_t - V^*|| \to 0, \quad \text{as} \quad t \to \infty. \tag{1.4.6}$$

The functions v_t are called *value iteration functions*.

We conclude this section presenting sufficient conditions for asymptotic optimality, introduced in Definition 5.

Lemma 11. Under Assumption 6, a policy $\pi \in \Pi$ is asymptotically optimal for the control model (1.2.1) if for all $x \in X$

$$\lim_{t \to \infty} E_x^\pi[\Phi(x_t, a_t)] = 0,$$

where $\Phi : X \times A \to \mathbb{R}$ is called *discrepancy function*, and is defined as

$$\Phi(x, a) = c(x, a) + \beta \int_X V^*(x') Q(dx'|x, a) - V^*(x).$$

Proof. First note that $\Phi \geq 0$. Now, for each $x \in X$, $\pi \in \Pi$ and $t \geq 0$, from (1.2.4),

$$\Phi(x_t, a_t) = E_x^\pi[c(x_t, a_t) + \beta V^*(x_{t+1}) - V^*(x_t)|h_t, a_t].$$

Next, we multiply by β^{t-n} to obtain

$$\beta^{t-n}\Phi(x_t, a_t) = \beta^{t-n}E_x^\pi[c(x_t, a_t) + \beta V^*(x_{t+1}) - V^*(x_t)|h_t, a_t],$$

and taking expectation E_x^π we have

$$\beta^{t-n}E_x^\pi[\Phi(x_t, a_t)] = \beta^{t-n}E_x^\pi[c(x_t, a_t) + \beta V^*(x_{t+1}) - V^*(x_t)].$$

Summing over all $t \geq n$ we obtain

$$\begin{aligned}
\sum_{t=n}^{\infty} \beta^{t-n}E_x^\pi[\Phi(x_t, a_t)] &= \sum_{t=n}^{\infty} \beta^{t-n}E_x^\pi[c(x_t, a_t) + \beta V^*(x_{t+1}) - V^*(x_t)] \\
&= E_x^\pi[\sum_{t=n}^{\infty} \beta^{t-n}c(x_t, a_t)] + E_x^\pi[\sum_{t=n}^{\infty} \beta^{(t+1)-n}V^*(x_{t+1}) - \beta^{t-n}V^*(x_t)] \\
&= V_n(\pi, x) + E_x^\pi[\lim_{N\to\infty} \sum_{t=n}^{N} \beta^{(t+1)-n}V^*(x_{t+1}) - \beta^{t-n}V^*(x_t)] \\
&= V_n(\pi, x) + E_x^\pi[\lim_{N\to\infty} \{\beta^N V^*(x_{n+N}) - V^*(x_n)\}] \\
&= V_n(\pi, x) - E_x^\pi[V^*(x_n)] + E_x^\pi[\lim_{N\to\infty} \beta^N V^*(x_{n+N})].
\end{aligned}$$

Since V^* is a bounded function (see Assumption 6.2), we have $E_x^\pi[\beta^N V^*(x_{n+N})] \to 0$, as $N \to \infty$. Hence,

$$\sum_{t=n}^{\infty} \beta^{t-n}E_x^\pi[\Phi(x_t, a_t)] = V_n(\pi, x) - E_x^\pi[V^*(x_n)].$$

Finally, if $E_x^\pi[\Phi(x_t, a_t)] \to 0$, we get

$$|V_n(\pi, x) - E_x^\pi[V^*(x_n)]| \to 0, \text{ as } \quad n \to \infty,$$

that is, π is asymptotically optimal. □

1.5 Partially observable Markov decision process

We now describe the class of controlled processes that we are interested in this work, the so-called *partially observable Markov Decision Processes* (POMDPs).

Unlike COMDPs, in POMDPs the controller only have partial information of the state of the system x_t through a observation process y_t. Specifically, similarly to (1.2.2), we consider that the dynamics of the system is given by

$$\begin{aligned}
x_{t+1} &= F(x_t, a_t, \xi_t), \quad t \in \mathbb{N}_0 \\
y_t &= G(a_{t-1}, x_t, \eta_t), \quad t \in \mathbb{N} \\
y_0 &= G_0(x_0, \eta_0),
\end{aligned}$$

where F, G and G_0 are known functions; x_t, a_t and y_t are the state, control, and the observation at time t, respectively. Further, $\{\xi_t\}$ is a sequence of i.i.d random variables that take values in a Borel space S_1, with common distribution θ_1; and $\{\eta_t\}$ is sequence of i.i.d. random variables that take values in a Borel space S_2, with common distribution θ_2, known as the *observation noise or measurement.*

A discrete-time partially observed Markov decision model consist of eight objects:

$$\mathcal{M}_{PO} = (X, Y, A, Q, \underline{\nu}, K, K_0, \tilde{c}) \tag{1.5.1}$$

where, X, Y, and A represent the state, observation, and control spaces, respectively, all of them are assumed to be Borel spaces. The following stochastic kernels are defined similarly to (1.2.3), that is, $Q(dx'|x,a)$, the transition law among states, is a stochastic kernel on X given $X \times A$ defined as

$$Q(B|x,a) = \int_{S_1} 1_B[F(x,a,s^{(1)})]\theta_1(ds^{(1)}); \quad B \in \mathcal{B}(X), x \in X, a \in A;$$

$K(dy|a,x)$, the observation kernel, is a stochastic kernel on Y given $A \times X$ defined by

$$K(C|a,x) = \int_{S_2} 1_C[G(a,x,s^{(2)})]\theta_2(ds^{(2)}); \quad C \in \mathcal{B}(Y), x \in X, a \in A;$$

$\underline{\nu} \in \mathbb{P}(X)$ is the distribution (a priori) of x_0; K_0, the initial observation, is a stochastic kernel on Y given X, and if $x_0 = x$, it is defined by

$$K_0(C|x) = \int_{S_2} 1_C[G_0(x,s^{(2)})]\theta_2(ds^{(2)}) \quad C \in \mathcal{B}(Y), x \in X.$$

Finally $\tilde{c} \in \mathcal{B}(X \times A)$ is the cost-per-stage function.

The control model (1.5.1) has the following interpretation: at time $t = 0$ the initial state x_0 has a given distribution $\underline{\nu}$, then an observation y_0 is generated according to the initial observation kernel $K_0(\cdot|x_0)$. Next the controller selects an action $a_0 \in A$, the cost $\tilde{c}(x_0, a_0)$ is incurred, and the system moves to a new state x_1 according to the law $Q(\cdot|x_0, a_0)$. In general, at time t, once the action $a_t \in A$ is selected, the following happen: (1) a cost $\tilde{c}(x_t, a_t)$ is generated, (2) the system moves to the state x_{t+1} according to the transition law $Q(dx_{t+1}|x_t, a_t)$, and (3) the observation y_{t+1} is generated by the observation kernel $K(dy_{t+1}|a_t, x_{t+1})$, and the process is repeated.

Hence, the dynamics of the system define a vector called the *observed history* defined as

$$h_0 := (\underline{\nu}, y_0) \in \mathbb{H}_0 := \mathbb{P}(X) \times Y$$

$$h_t := (\underline{\nu}, y_0, a_0, ..., y_{t-1}, a_{t-1}, y_t) \in \mathbb{H}_t := \mathbb{H}_{t-1} \times A \times Y, \quad t \geq 1.$$

Definition 12. A *control policy* for the PO control model is a sequence of stochastic kernels $\pi = \{\pi_t\}$, $t \in \mathbb{N}_0$ on A given $\mathbb{H}_t$.

Markov policies and stationary policies are defined similarly as in Definition 2. We will remain using the same notation Π, Π_M, and $\mathbb{F}$, for the family of all policies, Markov policies, and stationary policies, respectively. The context itself will determine which one we refer to.

Now, we are ready to establish the POMDP. Let $\Omega := (X \times Y \times A)^\infty$ and $\mathcal{F} = \mathcal{B}(\Omega)$ the corresponding product σ-algebra. Note that the elements of Ω are of the form

$$\omega = (x_0, y_0, a_0, x_1, y_1, a_1, ...)$$

Then, each policy $\pi \in \Pi$ and initial distribution $\underline{\nu} \in \mathbb{P}(X)$, together with the stochastic kernels Q, K and K_0, determine, on the space Ω, a probability measure $P^\pi_{\underline{\nu}}$ given by

$$P^\pi_{\underline{\nu}}(dx_0, dy_0, da_0, dx_1, dy_1, da_1, ...) = \\ \underline{\nu}(dx_0)K_0(dy_0|x_0)\pi_0(da_1|q, y_0)Q(dx_1|x_0, a_0)K(dy_1|a_0, x_1)\pi_1(da_1|q, y_0, a_0, y_1)...$$

The conditional expectation with respect to this probability measure is denoted by $E^\pi_{\underline{\nu}}$.

Definition 13. The stochastic process $(\Omega, \mathcal{F}, P^\pi_{\underline{\nu}}, \{x_t, y_t\})$ is called a POMDP.

We consider the control model (1.5.1). For each policy $\pi \in \Pi$ and initial distribution $\underline{\nu} \in \mathbb{P}(X)$, we define the *total expected discounted cost* as

$$J(\pi, \underline{\nu}) := E^\pi_{\underline{\nu}}\left[\sum_{t=0}^{\infty} \beta^t \tilde{c}(x_t, a_t)\right], \tag{1.5.2}$$

where $\beta \in (0, 1)$ is a given discount factor.Then, similarly as Definition 4, if

$$J^*(\underline{\nu}) := \inf_{\Pi} J(\pi, \underline{\nu}), \quad \underline{\nu} \in \mathbb{P}(X),$$

is the PO-value function, the PO optimal control problem is to find a policy $\pi^* \in \Pi$ such that

$$J(\pi^*, \underline{\nu}) = J^*(\underline{\nu}), \quad \underline{\nu} \in \mathbb{P}(X). \tag{1.5.3}$$

1.6 Reduction of a PO control model to a CO control model

We analyze the solution of the partially observable optimal control problem (PO-OCP) following a standard approach (see e.g., [7, 19, 23]), which consists to transform it into a completely observable optimal control problem (CO-OCP) defined on $\mathbb{P}(X)$. That is, we introduce a controlled CO process $\{\nu_t\} \subset \mathbb{P}(X)$, called the *filtering process* (see e.g., [10]), which evolves according to a difference equation

$$\nu_0 = H_0(\underline{\nu}, y_0) \quad \text{and} \quad \nu_{t+1} = H(\nu_t, a_t, y_{t+1}), \tag{1.6.1}$$

where the functions H_0 and H are known, $\underline{\nu}$ is the a priori distribution of x_0, and $\{y_t\}$ is the observation process.

Lemma 14. Let X, Y and W be Borel spaces and let $R(d(x, y)|w)$ be a stochastic kernel on $X \times Y$ given W. Then, there exist stochastic kernels $H'(dx|w, y)$ and $R'(dy|w)$ on X given $W \times Y$ and on Y given W, respectively, such that

$$R(B \times C|w) = \int_C H'(B|w, y)R'(dy|w), \quad \forall B \in \mathcal{B}(X), C \in \mathcal{B}(Y), w \in W$$

where $R'(dy|w)$ is the marginal distribution of $R(d(x, y)|w)$ on Y, i.e.

$$R'(dy|w) := R(X \times C|w), \quad C \in \mathcal{B}(Y).$$

The functions H_0 and H are stochastic kernels obtained by Lemma 14 on decomposition of probability measures on a product space (see e.g., [7, 9, 10, 19, 23] for more details).

We proceed to obtain the function H_0 and H by applying Lemma 14 in the following sense. Let X and Y be the state and observation spaces, respectively, and $W := \mathbb{P}(X) \times A$. In addition, let $R(d(x,y)|\nu,a)$ a stochastic kernel on $X \times Y$ given W defined as

$$R(B \times C|\nu,a) := \int_X \int_B K(C|a,x')Q(dx'|x,a)\nu(dx), \quad B \in \mathcal{B}(X), C \in \mathcal{B}(Y),$$

where Q and K are as in (1.5.1). Then, from Lemma 14, there exists a stochastic kernel $H'(dx|\nu,a,y)$ on X given $W \times Y = \mathbb{P}(X) \times A \times Y$ such that for each $B \in \mathcal{B}(X), C \in \mathcal{B}(Y)$ and $(\nu,a) \in W$,

$$R(B \times C|\nu,a) = \int_C H'(B|\nu,a,y)R'(dy|\nu,a),$$

where $R'(C|\nu,a) := R'(X \times C|\nu,a)$ is the marginal of $R(\cdot|\nu,a)$ on Y taking the form

$$R'(C|\nu,a) = \int_X \int_X K(C|a,x')Q(dx'|x,a)\nu(dx), \quad C \in \mathcal{B}(Y), (\nu,a) \in W. \tag{1.6.2}$$

Since H' is a stochastic kernel the function $H : \mathbb{P}(X) \times A \times Y$ defined as

$$H(\nu,a,y) := H'(\cdot|\nu,a,y) \tag{1.6.3}$$

is measurable, which in turn yields that

$$k(D|\nu,a) := \int_Y 1_D[H(\nu,a,y)]R'(dy|\nu,a), \quad D \in \mathcal{B}(\mathbb{P}(X)), (\nu,a) \in \mathbb{P}(X) \times A, \tag{1.6.4}$$

defines a stochastic kernel on $\mathbb{P}(X)$ given $\mathbb{P}(X) \times A$.

Following similar arguments as the previous procedure, we prove that there exists a stochastic kernel $H_0'(dx|\underline{\nu},y)$ on X given $\mathbb{P}(X) \times Y$ such that the function $H_0 : \mathbb{P}(X) \times Y \to \mathbb{P}(X)$ defined as

$$H_0(\underline{\nu},y) := H_0'(\cdot|\underline{\nu},y) \tag{1.6.5}$$

is measurable. Thus,

$$k_0(D|\underline{\nu}) := \int_Y 1_D[H_0(\underline{\nu},y)]R_0'(dy|\underline{\nu}), \quad D \in \mathcal{B}(\mathbb{P}(X)), \underline{\nu} \in \mathbb{P}(X); \tag{1.6.6}$$

defines a stochastic kernel on $\mathbb{P}(X)$ given $\mathbb{P}(X)$.

As is shown in [7, 19], from (1.6.1), we have that, for a control policy $\pi \in \Pi$ and initial distribution $\underline{\nu} \in \mathbb{P}(X)$, and for each $B \in \mathcal{B}(X)$,

$$\nu_0(B) = P_{\underline{\nu}}^{\pi}[x_0 \in B|h_0] = H_0(\underline{\nu},y_0)(B) = H_0(h_0)(B), \quad P_{\underline{\nu}}^{\pi} - a.s. \tag{1.6.7}$$

and

$$\nu_{t+1}(B) = P_{\underline{\nu}}^{\pi}[x_{t+1} \in B|h_{t+1}] = H(\nu_t,a_t,y_{t+1})(B), \quad P_{\underline{\nu}}^{\pi} - a.s. \tag{1.6.8}$$

Hence ν_t can be interpreted as the a posteriori distribution of the unobservable state x_t given the observable history h_t.

Taking into account the previous elements, and considering the PO control model (PO-CM) introduced in (1.5.1), we define the following CO control model (CO-CM)

$$\mathcal{M}_{CO} = (\mathbb{P}(X), A, k, k_0, c), \tag{1.6.9}$$

with state space $\mathbb{P}(X)$, control space A, transition law k defined in (1.6.4), initial distribution k_0 defined in (1.6.6) and one-stage cost function $c : \mathbb{P}(X) \times A \to \mathbb{R}$ defined as

$$c(\nu, a) := \int_X \tilde{c}(x, a)\nu(dx).$$

From (1.6.4) and (1.6.6), k and k_0 represent the transition kernels of the process $\{\nu_t\}$ (see (1.6.1), (1.6.7), (1.6.8)) corresponding to the constructed CO-CM (1.6.9). In order to introduce a suitable set of policies for $\mathcal{M}_{CO}$ we define the *information vector*

$$i_t := (\nu_0, a_0, ..., \nu_{t-1}, a_{t-1}, \nu_t) \in \mathbb{I}_t, \quad \text{where} \quad \mathbb{I}_t := (\mathbb{P}(X) \times A)^t \times \mathbb{P}(X)$$

According to the above, similarly as Definition 2, we define the corresponding control policies which are called *information policies*.

Definition 15. An *information policy* (or i-policy) is a sequence of stochastic kernels $\delta = \{\delta_t\}$ on A given $\mathbb{I}_t$, i.e., of the form $\delta_t(da|i_t)$. We denote Δ the set of all i-policies.

The Markov policies and the stationary ones are defined accordingly. Now, related to the equivalence of the models (1.5.1) and (1.6.9), note that $\Delta \subset \Pi$. Indeed if we take an arbitrary i-policy $\delta = \{\delta_t\}$, this defines a policy $\pi^\delta = \{\pi_t^\delta\} \in \Pi$ given by

$$\pi_t^\delta(\cdot|h_t) := \delta_t(\cdot|i_t(h_t)), \quad \forall h_t \in \mathbb{H}_t, t \geq 0,$$

where, $i_t(h_t) \in \mathbb{I}_t$ is the information vector given by the observable history h_t. Then, δ and π^δ are equivalents in the sense that for all $t \geq 0$, π^δ assigns the same conditional probability on A as δ_t, for all $h_t \in \mathbb{H}_t$. In fact we have the following result.

Lemma 16. For all $\pi \in \Pi$ there exists an i-policy $\delta \in \Delta$, such that (see (1.5.2))

$$J(\delta, \underline{\nu}) = J(\pi, \underline{\nu}), \quad \underline{\nu} \in \mathbb{P}(X).$$

Proof. See e.g., [23]. □

As previous sections, an i-policy $\delta \in \Delta$ and $\underline{\nu} \in \mathbb{P}(X)$ defines a probability measure $P_{\underline{\nu}}^\delta$ on the space $(\mathbb{P}(X) \times A)^\infty$

Now, consider the control model (1.6.9) and the set of i-policies Δ. We define the total discounted cost as

$$V(\delta, \underline{\nu}) := E_{\underline{\nu}}^\delta \left[\sum_{t=0}^{\infty} \beta^t c(\nu_t, a_t) \right], \quad \forall \delta \in \Delta, \underline{\nu} \in \mathbb{P}(X)$$

where $\beta \in (0, 1)$ is the discount factor and $E_{\underline{\nu}}^\delta$ is the expectation operator asociated to $P_{\underline{\nu}}^\delta$. Hence the optimal cost function is

$$V^*(\underline{\nu}) := \inf_{\delta \in \Delta} V(\delta, \underline{\nu}), \quad \underline{\nu} \in \mathbb{P}(X).$$

Then, the OCP is to find an i-policy $\delta^* \in \Delta$ such that

$$V(\delta^*, \underline{\nu}) = V^*(\underline{\nu}), \quad \underline{\nu} \in \mathbb{P}(X). \tag{1.6.10}$$

Furthermore, the original PO-OCP (1.5.3) and the CO-OCP (1.6.10) are equivalent, (see e.g., [23]) as is stated in the following result.

Lemma 17. $V(\delta, \underline{\nu}) = J(\delta, \underline{\nu}), \quad \forall \delta \in \Delta, \underline{\nu} \in \mathbb{P}(X)$.

Given the solution to the optimal control problem (OCP) for the completely observable control model (CO-CM) (see Theorem 9) and knowing that we can reduce the POMDP to a COMDP, our objective now is to impose conditions on the original POMDP (1.5.1) in order to ensure a solution in the CO-CM (1.6.9). These conditions are the following,

Assumption 18.

1. A is a compact set.
2. $\tilde{c} \in \mathcal{C}(X \times A)$.
3. The state transition law Q and the observation kernel K are continuous stochastic kernels.
4. The function H and H_0, defined in (1.6.7) and (1.6.8), are continuous on $\mathbb{P}(X) \times A \times Y$ and $\mathbb{P}(X) \times Y$, respectively.

Now, we show that under those conditions we get the suficient conditions in Assumption 6, corresponding to the CO-CM (1.6.9).

Lemma 19. If Assumption 18 holds, then the CO-CM (1.6.9) satisfies:

1. A is a compact set.
2. $c \in \mathcal{C}(\mathbb{P}(X) \times A)$
3. The stochastic kernel k and k_0 are weakly continuous.

Proof. Observe that 1 is the same as Assumption 18. Then we proceed to prove 2 and 3.

Since $\tilde{c}$ is a bounded and continuous function in $X \times A$ and

$$c(\nu, a) = \int_X \tilde{c}(x, a)\nu(dx),$$

Proposition C.1 yields, $c \in \mathcal{C}(\mathbb{P}(X) \times A)$. On the other hand, let $v \in \mathcal{C}(\mathbb{P}(X))$. To prove that k is weakly continuous we need to prove that the function

$$a \mapsto \int_{\mathbb{P}(X)} v(\nu')k(d\nu'|\nu, a), \quad \nu \in \mathbb{P}(X),$$

is continuous.

From (1.6.2) and (1.6.4) we have

$$\begin{aligned} v'(\nu, a) &:= \int_{\mathbb{P}(X)} v(\nu')k(d\nu'|\nu, a) \\ &= \int_Y v[H(\nu, a, y)]R'(dy|\nu, a) \\ &= \int_X \int_X \int_Y v[H(\nu, a, y)]K(dy|a, x')Q(dx'|x, a)\nu(dx). \end{aligned}$$

Now, because H is continuous, from Proposition C.1(b)

$$\int_Y v[H(\nu, a, y)]K(dy|a, x')$$

is continuous. Likewise, applying repeatedly Proposition C.1(b) we get that

$$\int_{\mathbb{P}(X)} v(\nu')k(d\nu'|\nu, a)$$

is continuous. Similarly is proved that k_0 is continuous. □

Summarizing, if we combine Theorem 9 and Lemma 19 we obtain the following result.

Theorem 20. *Suppose that Assumption 18 holds, that is, the CO-CM (1.6.9) satisfies Assumption 6. Then:*

(a) The optimal discounted cost function $V^* : \mathbb{P}(X) \to \mathbb{R}$ is the unique solution in $\mathcal{C}(\mathbb{P}(X))$ of the optimality equation, that is

$$V^*(\nu) = \min_{a \in A} \left\{ c(\nu, a) + \beta \int_X V^*(\nu')k(d\nu'|\nu, a) \right\}, \quad \nu \in \mathbb{P}(X). \tag{1.6.11}$$

(b) There exists $f^* : \mathbb{P}(X) \to A$ such that

$$V^*(\nu) = c(\nu, f^*(\nu)) + \beta \int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu, f^*(\nu)), \quad \nu \in \mathbb{P}(X). \tag{1.6.12}$$

Moreover, the stationary i-policy $f_*^\infty = \{f^*\}$ is an optimal control i-policy.

Chapter 2

POMDP with Observable Random Discount Factors and Unknown Distribution

2.1 Introduction

In this chapter we study a class of partially observable Markov decision processes (POMDP) with random discount factors of the form $\tilde{\alpha}(\xi_t)$, where $\{\xi_t\}$ is a sequence of observable i.i.d. random variables. The role of the discount factors $\tilde{\alpha}(\xi_t)$, $t \in \mathbb{N}_0$, during the evolution of the PO system is the following. At time $t = 0$, the initial x_0 has a given distribution $\underline{\nu}$ and an observation y_0 is generated according to an observation kernel. Then the controller chooses a control $a_0 \in A$ and a cost $c(x_0, a_0)$ is incurred. Next the system moves to a new state x_1 according to a transition law, the random disturbance ξ_1 comes in and a new observation y_1 is generated. Now, the controller selects an action $a_1 \in A$ and incurs a discounted cost $\tilde{\alpha}(\xi_1)c(x_1, a_1)$. Next the system moves to a new state x_2 and the process is repeated. In general, at time t, on the record of the random disturbances, the controller incurs the discounted cost

$$\widetilde{\alpha}(\xi_1)\widetilde{\alpha}(\xi_2)\dots\widetilde{\alpha}(\xi_t)c(x_t, a_t). \tag{2.1.1}$$

Hence, assuming that ξ_t has unknown distribution our objective is to study the partially observable optimal control problem (PO-OCP) under the performance index defined by the accumulation of the discounted costs (2.1.1). Our approach is to combine statistical estimation methods for the distribution θ with the PO procedure studied in the previous chapter.

2.2 The partially observable system

We consider a general POMDP with the dynamics of the system given by a pair of difference equations:

$$x_{t+1} = F(x_t, a_t, w_t^{(1)}), \quad t \in \mathbb{N}_0 \tag{2.2.1}$$

and

$$y_t = G(x_t, w_t^{(2)}), \quad t \in \mathbb{N}_0, \tag{2.2.2}$$

where x_t, a_t and y_t represent the state, the action and the observation at time t, with values in X, A, and Y, respectively; $\{w_t^{(1)}\}$ and $\{w_t^{(2)}\}$ are independent sequences of i.i.d. random variables with values in S_1 and S_2, distributions $\theta_1 \in \mathbb{P}(S_1)$ and $\theta_2 \in \mathbb{P}(S_2)$, respectively. We assume that the state space X, the observation space Y, and the disturbance spaces S_1, S_2, are Borel spaces, while the control space A is a compact metric space. Moreover, the cost-per-stage function, $\tilde{c} : X \times A \to \mathbb{R}$, is bounded and continuous function; and we consider an initial distribution $\underline{\nu} \in \mathbb{P}(X)$. Finally, $\tilde{\alpha} : S \to (0, 1)$ is a function of a random variable ξ_t, which represents the discount factor, where

$\{\xi_t\}$ is a sequence of observable i.i.d. random variables with values in Borel space S, with unknown distribution $\theta \in \mathbb{P}(S)$, that is

$$\theta(B) = P[\xi_t \in B], \quad B \in \mathcal{B}(S). \tag{2.2.3}$$

We call $\tilde{\alpha}$ the discount factor function. Hence, for the stage $t \geq 1$, the discounted cost is as (2.1.1).

We define the corresponding PO control model as

$$\mathcal{M}_{PO} = (X, Y, A, Q, \underline{\nu}, K, \tilde{\alpha}, \tilde{c}), \tag{2.2.4}$$

where the transition law Q and the observation kernel K are defined by the function F and G as follows:

$$Q(B|x_t, a_t) := \int_{S_1} 1_B[F(x_t, a_t, w^{(1)})]\theta_1(dw^{(1)}), B \in \mathcal{B}(X)$$

and

$$K(C|x_t) := \int_{S_2} 1_C[G(x_t, w^{(2)})]\theta_2(dw^{(2)}), C \in \mathcal{B}(Y).$$

Remark 21. Note that, unlike the model (1.5.1), the observations do not depend on the actions (see equation (2.2.2)), then the initial observation kernel is K.

To define the corresponding PO-OCP, we follow the procedure stated in Section 1.5. Indeed, let $H_0 := \mathbb{P}(X) \times Y$ and for $t \in \mathbb{N}$, $\mathbb{H}_t := \mathbb{H}_{t-1} \times A \times S \times Y$ be the space of observable histories. An element $h_t \in \mathbb{H}_t$ takes the form

$$h_t = (\underline{\nu}, y_0, a_0, y_1, a_1, \xi_1, y_2, a_2, \xi_2, ...a_{t-1}, \xi_{t-1}, y_t).$$

The control policies are defined similarly as Definition 2 (see Definition 12). Furthermore, if $\Omega := (X \times Y \times A \times S)^\infty$ and $\mathcal{F} = \mathcal{B}(\Omega)$, then for each $\pi \in \Pi$ and initial distribution $\underline{\nu} \in \mathbb{P}(X)$, there exists a probability measure $P^\pi_{\underline{\nu}}$ satisfying similar properties as in (1.2.4), together with

$$P^\pi_{\underline{\nu}}[\xi_{t+1} \in B|h_t, a_t] = \theta(B), \quad B \in \mathcal{B}(S).$$

Again $E^\pi_{\underline{\nu}}$ denotes the expectation operator with respecto to $P^\pi_{\underline{\nu}}$.

Taking into account (2.1.1), the costs are accumulated in an infinite horizon under the following discounted optimality criterion. For each policy $\pi \in \Pi$ and an initial distribution $\underline{\nu} \in \mathbb{P}(X)$ we define

$$V(\pi, \underline{\nu}) := E^\pi_{\underline{\nu}}\left[\sum_{t=1}^{\infty} \widetilde{\Gamma}_t \tilde{c}(x_t, a_t)\right],$$

where

$$\widetilde{\Gamma}_t := \prod_{k=0}^{t-1} \widetilde{\alpha}(\xi_{k+1}).$$

Hence, the partially observed optimal control problem (PO-OPC) is to find a policy $\pi^* \in \Pi$ such that

$$V^*(\underline{\nu}) := \inf_{\pi \in \Pi} V(\pi, \underline{\nu}) = V(\pi^*, \underline{\nu}), \quad \forall \underline{\nu} \in \mathbb{P}(X). \tag{2.2.5}$$

On the other hand, let

$$\alpha(\theta) := \int_S \widetilde{\alpha}(s)\theta(ds) \quad \text{and} \quad \Gamma_t = \prod_{k=0}^{t-1} \alpha(\theta_{k+1}) = [\alpha(\theta)]^t, \quad \Gamma_0 = 1. \tag{2.2.6}$$

Observe that

$$E^\pi_{\underline{\nu}}[\widetilde{\Gamma}_t \tilde{c}(x_t, a_t)] = E^\pi_{\underline{\nu}}\{[\alpha(\theta)]^t \tilde{c}(x_t, a_t)\} = \alpha(\theta)^t E^\pi_{\underline{\nu}}\{\tilde{c}(x_t, a_t)\}, \quad t \in \mathbb{N}_0, \tag{2.2.7}$$

since from the begining, $w_t^{(1)}$, $w_t^{(2)}$ and ξ_t are independent. Therefore

$$V(\pi, \underline{\nu}) = E^\pi_{\underline{\nu}}\left[\sum_{t=0}^{\infty} \widetilde{\Gamma}_t \tilde{c}(x_t, a_t)\right] = E^\pi_{\underline{\nu}}\left[\sum_{t=0}^{\infty} [\alpha(\theta)]^t \tilde{c}(x_t, a_t)\right] \tag{2.2.8}$$

2.3 The completely observable system

The analysis of the problem (2.2.5) is based on the reduction of the partially observed control model (PO-CM) (2.2.4) to a completely observable, as in Section 1.6. To fix ideas, for each $\pi \in \Pi$ and $\underline{\nu} \in \mathbb{P}(X)$, we consider the filtering process $\{\nu_t\} \subset \mathbb{P}(X)$ defined, for $B \in \mathcal{B}(X)$, as

$$\nu_0(B) := P^\pi_{\underline{\nu}}[x_0 \in B | h_0] \tag{2.3.1}$$

and

$$\nu_{t+1}(B) := P^\pi_{\underline{\nu}}[x_{t+1} \in B | h_{t+1}], \quad t \geq 0. \tag{2.3.2}$$

Furthermore, by Lemma 14 (see (1.6.7) and (1.6.8)), there exist measurable functions, that we will denote by $\Psi_0 : \mathbb{P}(X) \times Y \to \mathbb{P}(X)$ and $\Psi : \mathbb{P}(X) \times A \times Y \to \mathbb{P}(X)$, such that the filtering process (2.3.1)-(2.3.2) satisfies recursive equations of the form

$$\nu_0 = \Psi_0(\underline{\nu}, y_0) \quad \text{and} \quad \nu_{t+1} = \Psi(\nu_t, a_t, y_{t+1}), \quad t \geq 0. \tag{2.3.3}$$

Hence, we get the model (see 1.6.9)

$$\mathcal{M}_{CO} = (\mathbb{P}(X), A, k, k_0, \alpha, c), \tag{2.3.4}$$

where, $\mathbb{P}(X), A$ and c are as the model (1.6.9) in Section 1.6, k is a transition law given by (see (1.6.2) and (1.6.4))

$$k(d\nu'|\nu, a) = \int_Y 1_D[\Psi(\nu, a, y)] R'(dy|\nu, a),$$

and $k_0(\cdot|\underline{\nu})$, the initial distribution, defined by (see 1.6.6)

$$k_0(D|\underline{\nu}) := \int_Y 1_D[\Psi_0(\underline{\nu}, y)] R'_0(dy|\underline{\nu}), \quad D \in \mathcal{B}(\mathbb{P}(X)), \underline{\nu} \in \mathbb{P}(X).$$

In addition, $c : \mathbb{P}(X) \times A \to \mathbb{R}$ is the one stage cost function

$$c(\nu, a) := \int_X \tilde{c}(x, a)\nu(dx)$$

then, performance index for the CO-CM (2.2.8) can be written as

$$V(\delta, \underline{\nu}) = E^{\delta}_{\underline{\nu}}\left[\sum_{t=0}^{\infty}[\alpha(\theta)]^t c(\nu_t, a_t)\right], \tag{2.3.5}$$

while the expected total discounted cost from stage $n \geq t$ onwards as

$$V_n(\delta, \underline{\nu}) = E^{\delta}_{\underline{\nu}}\left[\sum_{t=n}^{\infty}[\alpha(\theta)]^{t-n} c(\nu_t, a_t)\right].$$

Thus the OCP for the CO-CM is to find an i-policy δ^* such that

$$V(\delta^*, \underline{\nu}) = V^*(\underline{\nu}) := \inf_{\delta \in \Delta} V(\delta, \underline{\nu}), \quad \forall \underline{\nu} \in \mathbb{P}(X). \tag{2.3.6}$$

In the scenario of Theorem 20 the following condition ensure the existence of a solution of the completely observable optimal control problem (CO-OCP) (see Assumption 18).

Assumption 22.

1. A is a compact set.
2. $\tilde{c} \in \mathcal{C}(X \times A)$.
3. The state transition law Q and the observation kernel K are continuous stochastic kernels.
4. The functions Ψ_0 and Ψ are continuous on $\mathbb{P}(X) \times Y$ and $\mathbb{P}(X) \times A \times Y$, respectively.

Hence, letting $\beta = \alpha(\theta)$ in Theorem 20, we obtain the following results.

Theorem 23. *Under Assumption 22,*

(a) The optimal discounted cost function $V^* : \mathbb{P}(X) \to \mathbb{R}$ is the unique solution in $\mathcal{C}(\mathbb{P}(X))$ of the optimality equation, that is

$$V^*(\nu) = \min_{a \in A}\left\{c(\nu, a) + \alpha(\theta)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu, a)\right\}, \quad \nu \in \mathbb{P}(X). \tag{2.3.7}$$

(b) There exists $f^* : \mathbb{P}(X) \to A$ such that

$$V^*(\nu) = c(\nu, f^*(\nu)) + \alpha(\theta)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu, f^*(\nu)), \quad \nu \in \mathbb{P}(X).$$

Moreover, the stationary i-policy $f^\infty_* = \{f^*\}$ is an optimal control i-policy.

2.4 Empirical estimation and control

Since θ is unknown, the solution given by Theorem 23 is not accesible to the controller. Therefore, being that θ is unknown and the random disturbance process $\{\xi_t\}$ is observable, we assume that the controller uses the empirical distribution to get an estimate $\hat{\theta}_t$ of θ. That is, $\{\hat{\theta}_t\} \subset \mathbb{P}(S)$ is obtained by the process

$$\hat{\theta}_t(B) := \frac{1}{t}\sum_{i=0}^{t-1} 1_B(\xi_i), \forall t \in \mathbb{N}, B \in \mathcal{B}(S). \tag{2.4.1}$$

It is well known that $\hat{\theta}_t$ converges weakly to θ almost surely (see e.g., [8]). Hence,

$$\int_S \tilde{\alpha}(s)\hat{\theta}_t(ds) \to \int_S \tilde{\alpha}(s)\theta(ds) \text{ a.s.} \quad \text{as } t \to \infty.$$

That is, as $t \to \infty$,

$$\alpha(\hat{\theta}_t) \overset{a.s.}{\to} \alpha(\theta) \tag{2.4.2}$$

Observe that the discounted cost defined in (2.3.5) depends strongly on the controls selected in the first stages, exactly where the information given by the empirical estimation about θ is poor. This discordance implies that we can not ensure the existence of optimal policies when we apply estimation and control procedures. Thus, in this case, the optimality will be studied in an asymptotic sense, as in Definition 5.

The construction of asymptotically discount optimal policies is based in the following variant of the value iteration scheme.

Let $\{v_t\}$ be a sequence of functions in $\mathcal{C}(\mathbb{P}(X))$ defined as

$$v_0 = 0 \quad \text{and} \quad v_t(\nu) = \min_{a\in A}\left\{c(\nu,a) + \alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} v_{t-1}(\nu')k(d\nu'|\nu,a)\right\}, \quad t \geq 1. \tag{2.4.3}$$

Since we know the value of $\hat{\theta}_t$ for all t, we apply the Theorem 23(b), for each t, to obtain that there exists $\hat{f}_t : \mathbb{P}(X) \to A$, such that

$$v_t(\nu) = c(\nu, \hat{f}_t) + \alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} v_{t-1}(\nu')k(d\nu'|\nu,\hat{f}_t). \tag{2.4.4}$$

Hence we apply the estimation process $\hat{\theta}_t$, given by (2.4.1) and (2.4.2), to show that the estimated i-policy $\hat{\delta} = \{\hat{f}_t\}$ is asymptotically discount optimal. That is from Lemma 11 our objective is to prove

$$\lim_{t\to\infty} E^{\hat{\delta}}_{\underline{\nu}}[\Phi(\nu_t, a_t)] = 0, \quad \underline{\nu} \in \mathbb{P}(X),$$

where,

$$\Phi(\nu,a) = c(\nu,a) + \alpha(\theta)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a) - V^*(\nu).$$

Theorem 24. *Under Assumption 22,*

(a) $||v_t - V^*|| = \sup_{\nu\in\mathbb{P}(X)} |v_t(\nu) - V^*(\nu)| \to 0$ *a.s., as* $t \to \infty$.

(b) The i-policy $\hat{\delta} = \{\hat{f}_t\}$ *is asymptotically discounted optimal, i.e.,*

$$E^{\hat{\delta}}_{\underline{\nu}}[\Phi(\nu_t, a_t)] \to 0, \quad \text{as } t \to \infty.$$

Proof. (a) From (2.3.7) and (2.4.3), for each $\nu \in \mathbb{P}(X)$ and $t \in \mathbb{N}$,

$$\begin{aligned}
&\left|v_t(\nu) - V^*(\nu)\right| \\
&= \left|\min_{a\in A}\left\{c(\nu,a)+\alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} v_{t-1}(\nu')k(d\nu'|\nu,a)\right\} - \min_{a\in A}\left\{c(\nu,a)+\alpha(\theta)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a)\right\}\right| \\
&\leq \sup_{a\in A}\left|\alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} v_{t-1}(\nu')k(d\nu'|\nu,a) - \alpha(\theta)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a)\right|
\end{aligned}$$

$$= \sup_{a\in A}\Big|\alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} v_{t-1}(\nu')k(d\nu'|\nu,a) - \alpha(\theta)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a)$$
$$+ \alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a) - \alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a)\Big|$$

$$= \sup_{a\in A}\Big|\alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} \big(v_{t-1}(\nu') - V^*(\nu')\big)k(d\nu'|\nu,a) + \big(\alpha(\hat{\theta}_t) - \alpha(\theta)\big)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a)\Big|$$

$$\leq \sup_{a\in A}\left\{\Big|\alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} \big(v_{t-1}(\nu') - V^*(\nu')\big)k(d\nu'|\nu,a)\Big| + \Big|\big(\alpha(\hat{\theta}_t) - \alpha(\theta)\big)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a)\Big|\right\}$$

$$\leq \sup_{a\in A}\Big|\alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} \big(v_{t-1}(\nu') - V^*(\nu')\big)k(d\nu'|\nu,a)\Big| + \sup_{a\in A}\Big|\big(\alpha(\hat{\theta}_t) - \alpha(\theta)\big)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a)\Big|$$

$$\leq \sup_{a\in A}\alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} \big|v_{t-1}(\nu') - V^*(\nu')\big|k(d\nu'|\nu,a) + \sup_{a\in A}\big|\alpha(\hat{\theta}_t) - \alpha(\theta)\big|\int_{\mathbb{P}(X)} \big|V^*(\nu')\big|k(d\nu'|\nu,a)$$

$$\leq \alpha(\hat{\theta}_t)\big|\big|v_{t-1} - V^*\big|\big| + \big|\alpha(\hat{\theta}_t) - \alpha(\theta)\big|\big|\big|V^*\big|\big|. \tag{2.4.5}$$

Now, if b is a bound of the cost, observe that (see Assumption 6) $||V^*|| \leq b/(1-\alpha(\theta))$. Then, taking $\sup_{\nu\in\mathbb{P}(X)}$ on both sides of (2.4.5) we have

$$\big|\big|v_t - V^*\big|\big| \leq \alpha(\hat{\theta}_t)\big|\big|v_{t-1} - V^*\big|\big| + \big|\alpha(\hat{\theta}_t) - \alpha(\theta)\big|\frac{b}{1-\alpha(\theta)}.$$

Let $L := \limsup ||v_t - V^*||$. Since v_t and V^* are bounded, we have that $L < \infty$. Then by (2.4.2) and taking lim sup as $t \to \infty$, we obtain

$$L \leq \alpha(\theta)L, \tag{2.4.6}$$

but as $0 < \alpha(\theta) < 1$, we have $L = 0$. Therefore

$$\lim_{t\to\infty}\big|\big|v_t - V^*\big|\big| = 0 \quad \text{a.s.}$$

(b) We start by defining the *approximate discrepancy function*

$$\Phi_t(\nu,a) = c(\nu,a) + \alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} v_{t-1}(\nu')k(d\nu'|\nu,a) - v_t(\nu), \quad t \in \mathbb{N}_0.$$

Observe that by (2.4.4)

$$\Phi_t(\nu, \hat{f}_t) = 0, \forall t \in \mathbb{N}_0, \nu \in \mathbb{P}(X). \tag{2.4.7}$$

Hence, if $\{(\nu_t, a_t)\}$ is the sequence of state-actions pairs corresponding to the application of the i-policy $\hat{\delta}$, observe that from (2.4.7)

$$\begin{aligned}\Phi(\nu_t,a_t) &= |\Phi(\nu_t,a_t) - \Phi_t(\nu_t,a_t)| \\ &\leq \sup_{a\in A}|\Phi(\nu_t,a) - \Phi_t(\nu_t,a)| \\ &\leq \sup_{(\nu,a)\in\mathbb{P}(X)\times A}|\Phi(\nu,a) - \Phi_t(\nu,a)| := \mathcal{S}_t\end{aligned}$$

Hence, the remainder of the proof consists in showing that

$$\lim_{t\to\infty} E_{\underline{\nu}}^{\hat{\delta}}[\mathcal{S}_t] = 0. \tag{2.4.8}$$

To this end, for $(\nu, a) \in \mathbb{P}(X) \times A$ we have

$$\begin{aligned}
&\big|\Phi(\nu,a) - \Phi_t(\nu,a)\big| \\
&= \left|c(\nu,a)+\alpha(\theta)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a) - V^*(\nu) - c(\nu,a) - \alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} v_{t-1}(\nu')k(d\nu'|\nu,a) + v_t(\nu)\right| \\
&= \Bigg|\alpha(\theta)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a) - V^*(\nu) - \alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} v_{t-1}(\nu')k(d\nu'|\nu,a) + v_t(\nu) \\
&\qquad + \alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a) - \alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a)\Bigg| \\
&= \Bigg|\big(\alpha(\theta) - \alpha(\hat{\theta}_t)\big)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a) + \big(v_t(\nu) - V^*(\nu)\big) \\
&\qquad + \alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} \big(V^*(\nu') - v_{t-1}(\nu')\big)k(d\nu'|\nu,a)\Bigg| \\
&\le \left|\big(\alpha(\theta) - \alpha(\hat{\theta}_t)\big)\int_{\mathbb{P}(X)} V^*(\nu')k(d\nu'|\nu,a)\right| + \big|v_t(\nu) - V^*(\nu)\big| \\
&\qquad + \left|\alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} \big(V^*(\nu') - v_{t-1}(\nu')\big)k(d\nu'|\nu,a)\right| \\
&\le \big|\alpha(\theta) - \alpha(\hat{\theta}_t)\big|\int_{\mathbb{P}(X)} \big|V^*(\nu')\big|k(d\nu'|\nu,a) + \big|v_t(\nu) - V^*(\nu)\big| \\
&\qquad + \alpha(\hat{\theta}_t)\int_{\mathbb{P}(X)} \big|V^*(\nu') - v_{t-1}(\nu')\big|k(d\nu'|\nu,a) \\
&\le \big|\alpha(\theta) - \alpha(\hat{\theta}_t)\big|\big|\big|V^*\big|\big| + \big|\big|v_t - V^*\big|\big| + \big|\big|V^* - v_{t-1}\big|\big|.
\end{aligned}$$

This implies,

$$\begin{aligned}
\mathcal{S}_t &\le \big|\alpha(\theta) - \alpha(\hat{\theta}_t)\big|\big|\big|V^*\big|\big| + \big|\big|v_t - V^*\big|\big| + \alpha(\hat{\theta}_t)\big|\big|V^* - v_{t-1}\big|\big| \\
&\le \big|\alpha(\theta) - \alpha(\hat{\theta}_t)\big|\frac{b}{1-\alpha(\theta)} + \big|\big|v_t - V^*\big|\big| + \alpha(\hat{\theta}_t)\big|\big|V^* - v_{t-1}\big|\big|.
\end{aligned}$$

Then, by (2.4.2) and part (a) of this theorem, letting $t \to \infty$ we get

$$\mathcal{S}_t \overset{a.s.}{\to} 0 \text{ as } t \to \infty.$$

Now, since $\mathcal{S}_t$ is bounded we obtain

$$\lim_{t\to\infty} E_{\underline{\nu}}^{\hat{\delta}}[\mathcal{S}_t] = 0,$$

that is the i-policy $\hat{\delta}$ is asymptotically discounted optimal.

□

Chapter 3

POMDP with Unobservable Random Discount Factors and Unknown Distribution

3.1 Introduction

We are now interested in study the model introduced in Chapter 2 when the random variable ξ_t really represents a random noise which is imposible to observe and, furthermore, its distribution may change from stage to stage. In opposite to the previous situation, in this case the controller cannot estimate by means of statistical methods the unknown distribution. Hence we analize the problem as a class of minimax systems called *game against nature*. That is, we assume that the controller has an opponent, the nature, who selects the distribution θ_t for ξ_t at each time $t \in \mathbb{N}$. Hence, the objective of the controller is to select actions directed to minimize the maximum cost generated by the nature.

3.2 The transformed minimax control problem

We consider a minimax control model of the form

$$\mathcal{M}_{\max}^{\min} = (\mathbb{P}(X), A, \Theta, k, k_0, c, \alpha), \tag{3.2.1}$$

where $\mathbb{P}(X), A, k, k_0, c, \alpha$ are as the completely observable control model (CO-CM) (2.3.4), which comes from the partially observable control model (2.2.4), and $\Theta \subset \mathbb{P}(S)$ is a Borel subset of the space of probability measures on S, which represents the opponent action space. We suppose that $\{\xi_t\}$ is a sequence of independent and possibly non-observable random variables on $(\Omega, \mathcal{F}, P)$ taking values on S, with corresponding distribution $\theta_t \in \Theta$. That is,

$$\theta_t(B) := P[\xi_t \in B], \quad t \in \mathbb{N}, B \in \mathcal{B}(S).$$

The model (3.2.1) represents a controlled stochastic system which can be seen as a *game against nature* whose evolution is as follows. At time $t \in \mathbb{N}$ the system is in state $\nu_t \in \mathbb{P}(X)$, the controller chooses an action $a_t \in A$ and the opponent, the "nature", picks a distribution $\theta_t \in \Theta$ for the random disturbance ξ_t. Then, on the record of the previous distributions $\theta_1, \theta_2, ..., \theta_{t-1}$, the controller incurs a discounted cost

$$\alpha(\theta_1) \cdots \alpha(\theta_t) c(\nu_t, a_t). \tag{3.2.2}$$

Next, the process moves to a new state according to the transition law k and the process is repeated. Thus, the goal of the controller is to minimize the maximum cost incurred by nature.

As usual the actions are selected according to the control policies, which, in this minimax scenario, are defined as follows. Let

$$\mathbb{H}_0 := \mathbb{P}(X), \quad \text{and} \quad \mathbb{H}_0' := \mathbb{P}(X) \times A$$

and, for all $t \in \mathbb{N}_0$, let

$$\mathbb{H}_t := (\mathbb{P}(X) \times A \times \Theta)^t \times \mathbb{P}(X) \quad \text{and} \quad \mathbb{H}_t' := (\mathbb{P}(X) \times A \times \Theta)^t \times \mathbb{P}(X) \times A.$$

Elements of $\mathbb{H}_t$ and $\mathbb{H}_t'$ take the form,

$$h_t := (\underline{\nu}, a_0, \theta_1, ..., \nu_{t-1}, a_{t-1}, \theta_{t-1}, \nu_t), \quad \text{and} \quad h_t' := (h_t, a_t).$$

Hence, an i-policy for the controller is a sequence $\delta = \{\delta_t\}$ of stochastic kernels on A given $\mathbb{H}_t$ such that $\delta_t(A|h_t) = 1$ for all $h_t \in \mathbb{H}_t$ and $t \in \mathbb{N}_0$. Markov and stationary policies for the controller are defined similarly as in Definition 15. We denote by Δ the set of all i-policies for the controller, and by $\Delta_M \subset \Delta$ the subset of Markov i-policies. The policies for the nature are defined considering the space $\mathbb{H}_t'$. Indeed, a i-policy for the nature (the opponent) is a sequence $\delta' = \{\delta_t'\}$ of stochastic kernels on Θ given $\mathbb{H}_t'$ such that $\delta_t' := (\Theta|h_t') = 1$ for all $h_t' \in \mathbb{H}_t'$ and $t \in \mathbb{N}_0$. We denote by Δ_Θ the set of all i-policies for the nature.

For each pair of policies $(\delta, \delta') \in \Delta \times \Delta_\Theta$ and initial distribution $\underline{\nu} \in \mathbb{P}(X)$, we define the total expected discounted cost as

$$V(\delta, \delta', \underline{\nu}) := E_{\underline{\nu}}^{\delta\delta'} \left[\sum_{t=0}^{\infty} \Gamma_t c(\nu_t, a_t) \right], \tag{3.2.3}$$

where as previous chapters $E_{\underline{\nu}}^{\delta\delta'}$ is the expectation operator corresponding to the probability measure $P_{\underline{\nu}}^{\delta\delta'}$ induced by (δ, δ') and $\underline{\nu}$. In addition,

$$\Gamma_t := \prod_{k=0}^{t-1} \alpha(\theta_{k+1}) \quad \text{if } t \in \mathbb{N}, \quad \Gamma_0 = 1. \tag{3.2.4}$$

Thus, the minimax control problem (MMCP) to the controller is to find an i-policy $\delta^* \in \Delta$ such that

$$V^*(\underline{\nu}) := \inf_{\delta \in \Delta} \sup_{\delta' \in \Delta_\Theta} V(\delta, \delta', \underline{\nu}) = \sup_{\delta' \in \Delta_\Theta} V(\delta^*, \delta', \underline{\nu}), \quad \underline{\nu} \in \mathbb{P}(X). \tag{3.2.5}$$

In this case, the i-policy δ^* is said to be *minimax*, whereas V^* is the *minimax value function*. As well as in previous chapters, we need to impose certain conditions either on the control model (2.2.4) or directly on the minimax control model (3.2.1) (see Lemma 19). We focus on the model (3.2.1).

Assumption 25.

1. A is a compact set.
2. $\Theta \subset \mathbb{P}(S)$ is a compact set.
3. $c \in \mathcal{C}(\mathbb{P}(X) \times A)$.

4. The stochastic kernels k and k_0 are weakly continuous.

5. The function $\tilde{\alpha}(s)$ is continuous on S and $\alpha^* := \sup_{s\in S} \tilde{\alpha}(s) < 1$.

Remark 26. Remember that $\widetilde{\alpha} : S \to (0,1)$ is the discount factor function and $\alpha(\theta) = \int_S \tilde{\alpha}(s)\theta(ds)$. Then from Assumption 25.5

$$\max_{\theta\in\Theta} \alpha(\theta) = \max_{\theta\in\Theta} \int_S \tilde{\alpha}(s)\theta(ds) \leq \alpha^* < 1$$

Also note that the function $\alpha : \Theta \to (0,1)$ is continuous. Indeed, let $\{\theta_t\} \in \Theta$ be a sequence converging to θ, that is $\theta_t \to \theta$ weakly. Then

$$\int_S \tilde{\alpha}(s)\theta_t(ds) \to \int_S \tilde{\alpha}(s)\theta(ds).$$

That is,

$$\alpha(\theta_t) \to \alpha(\theta), \quad \text{as } t \to \infty.$$

3.3 Existence of minimax policies

The solution to the MMCP can be obtained by applying the *contraction mapping approach* similar to that defined in previous chapters. We begin introducing the operators; for any function $u : \mathbb{P}(X) \to \mathbb{R}$ in $\mathcal{C}(\mathbb{P}(X))$, we define the *minimax dynamic programming operator* as

$$Tu(\nu) = \min_{a\in A} \max_{\theta\in\Theta} T_{(a,\theta)}u(\nu), \tag{3.3.1}$$

where

$$T_{(a,\theta)}u(\nu) := c(\nu, a) + \alpha(\theta) \int u(\nu')k(d\nu'|\nu, a), \quad \nu \in \mathbb{P}(X).$$

In addition, for an arbitrary stationary policy $f^\infty \in \Delta_M$, we define

$$T_f u(\nu) = \max_{\theta\in\Theta} T_{(f,\theta)}u(\nu) \tag{3.3.2}$$

where

$$T_{(f,\theta)}u(\nu) := c(\nu, f(\nu)) + \alpha(\theta) \int u(\nu')k(d\nu'|\nu, f(\nu)).$$

Remark 27. Let $u \in \mathcal{C}(\mathbb{P}(X))$, by Assumption 25.4, $\int_{\mathbb{P}(X)} u(\nu')k(d\nu'|\nu, a)$ is continuous with respec to ν and a. Hence $v(\nu, a, \theta) := c(\nu, a) + \alpha(\theta) \int_{\mathbb{P}(X)} u(\nu')k(d\nu'|\nu, a)$ is bounded and continuous. Then, from Proposition D.2 we have that $\max_{\theta\in\Theta} v(\nu, a, \theta)$ is continuous in $\mathbb{P}(X) \times A$. Furthermore, there exists $f^* : \mathbb{P}(X) \to A$ such that

$$\max_{\theta\in\Theta} v(\nu, f^*(\nu), \theta) = \min_{a\in A} \max_{\theta\in\Theta} v(\nu, a, \theta), \quad \forall \nu \in \mathbb{P}(X).$$

That is

$$\max_{\theta\in\Theta} T_{(f^*,\theta)}u(\nu) = Tu(\nu), \quad \nu \in \mathbb{P}(X). \tag{3.3.3}$$

Accordingly we get the following result.

Lemma 28. If Assumption 25 holds, then:

1. $Tu(\nu) \in \mathcal{C}(\mathbb{P}(X))$ for all $u \in \mathcal{C}(\mathbb{P}(X))$.
2. The operators T and T_f are contractions on $\mathcal{C}(\mathbb{P}(X))$ with modulus α^*.
3. There exists unique functions u^* and $u_f^* \in \mathcal{C}(\mathbb{P}(X))$ such that

$$Tu^* = u^* \quad \text{and} \quad T_f u_f^* = u_f^*,$$

and moreover, for any function $u \in \mathcal{C}(\mathbb{P}(X))$

$$||T^n u - u^*|| \to 0, \quad \text{and} \quad ||T_f^n u - u_f^*|| \to 0, \text{ as } n \to \infty,$$

where $T^n u = T(T^{n-1}u)$, $n \in \mathbb{N}$.

Proof.

1. This part is consequence of Remark 27.
2. Let $u, u' \in \mathcal{C}(\mathbb{P}(X))$. Then for $\nu \in \mathbb{P}(X)$ we have

$$\begin{aligned}
&\big|Tu(\nu) - Tu'(\nu)\big| \\
&= \left| \min_{a\in A} \max_{\theta\in\Theta} \left\{ c(\nu,a) + \alpha(\theta) \int_{\mathbb{P}(X)} u(\nu') k(d\nu'|\nu,a) \right\} - \min_{a\in A} \max_{\theta\in\Theta} \left\{ c(\nu,a) + \alpha(\theta) \int_{\mathbb{P}(X)} u'(\nu') k(d\nu'|\nu,a) \right\} \right| \\
&\leq \max_{a\in A} \left| \max_{\theta\in\Theta} \left\{ c(\nu,a) + \alpha(\theta) \int_{\mathbb{P}(X)} u(\nu') k(d\nu'|\nu,a) \right\} - \max_{\theta\in\Theta} \left\{ c(\nu,a) + \alpha(\theta) \int_{\mathbb{P}(X)} u'(\nu') k(d\nu'|\nu,a) \right\} \right| \\
&\leq \max_{a\in A} \max_{\theta\in\Theta} \left| c(\nu,a) + \alpha(\theta) \int_{\mathbb{P}(X)} u(\nu') k(d\nu'|\nu,a) - c(\nu,a) + \alpha(\theta) \int_{\mathbb{P}(X)} u'(\nu') k(d\nu'|\nu,a) \right| \\
&\leq \max_{a\in A} \max_{\theta\in\Theta} \left| \alpha(\theta) \int_{\mathbb{P}(X)} \big(u(\nu') - u'(\nu')\big) k(d\nu'|\nu,a) \right| \\
&\leq \max_{a\in A} \max_{\theta\in\Theta} \alpha(\theta) \int_{\mathbb{P}(X)} \big|u(\nu') - u'(\nu')\big| k(d\nu'|\nu,a) \\
&\leq \alpha^* \max_{a\in A} \int_{\mathbb{P}(X)} |u(\nu') - u'(\nu')| k(d\nu'|\nu,a) \\
&\leq \alpha^* ||u - u'||.
\end{aligned}$$

   Therefore, taking $\sup_{\nu\in\mathbb{P}(X)}$ we obtain

$$||Tu - Tu'|| \leq \alpha^* ||u - u'||.$$

   Similarly it is proved that T_f is a contraction operator.
3. This part follows from the Banach Fixed Point Theorem (see Proposition A.1).

Hence, Lemma 28 has been proved. □

We define the sequence $\{v_n\} \subset \mathcal{C}(\mathbb{P}(X))$ of minimax value iteration functions as

$$v_0 = 0 \quad \text{and} \quad v_n(\nu) = Tv_{n-1}(\nu) = T^n v(\nu), \quad n \in \mathbb{N}. \tag{3.3.4}$$

Theorem 29. *Under Assumption 25 the following holds:*

(a) The minimax value function satisfies

$$V^*(\nu) = TV^*(\nu), \quad \nu \in \mathbb{P}(X). \tag{3.3.5}$$

(b) As $n \to \infty$, $||v_n - V^*|| \to 0$, where $\{v_n\}$ is the sequence of functions defined in (3.3.4).

(c) There exists $f^* : \mathbb{P}(X) \to A$ such that

$$V^*(\nu) = \sup_{\theta \in \Theta} T_{(f^*,\theta)} V^*(\nu), = \quad \nu \in \mathbb{P}(X). \tag{3.3.6}$$

Moreover the stationary i-policy $f_*^\infty = \{f^*\}$, is a minimax i-policy, that is,

$$V^*(\nu) = \sup_{\delta' \in \Delta_\Theta} V(f_*^\infty, \delta', \nu), \quad \nu \in \mathbb{P}(X)$$

Proof. (a) - (b) From Lemma 28.3, it is sufficient to prove that $u^* = V^*$.

Let $f : \mathbb{P}(X) \to A$ such that

$$u^*(\nu) := Tu^*(\nu) = \sup_{\theta \in \Theta} T_{(f,\theta)} u^*(\nu), \quad \nu \in \mathbb{P}(X),$$

which exists from Remark 27. Then

$$u^*(\nu) \geq c(\nu, f(\nu)) + \alpha(\theta) \int u^*(\nu') k(d\nu'|\nu, f(\nu)), \quad \forall \nu \in \mathbb{P}(X), \theta \in \Theta. \tag{3.3.7}$$

Let $\delta' \in \Delta_\Theta$ be an arbitrary policy for the opponent. Then, iterating inequality (3.3.7) we get, for $f^\infty = \{f\}$,

$$\begin{aligned}
u^*(\nu) &\geq E_\nu^{f^\infty \delta'} \left[c(\nu_0, f(\nu_0)) + \sum_{t=1}^{n-1} \prod_{k=0}^{t-1} \alpha(\theta_{k+1}) c(\nu_t, f(\nu_t)) + \prod_{k=0}^{n-1} \alpha(\theta_{k+1}) u^*(\nu_n) \right] \\
&= E_\nu^{f^\infty \delta'} \left[c(\nu_0, f(\nu_0)) + \sum_{t=1}^{n-1} \prod_{k=0}^{t-1} \alpha(\theta_{k+1}) c(\nu_t, f(\nu_t)) \right] + E_\nu^{f^\infty \delta'} \left[\prod_{k=0}^{n-1} \alpha(\theta_{k+1}) u^*(\nu_n) \right] \\
&= E_\nu^{f^\infty \delta'} \left[\sum_{t=0}^{n-1} \Gamma_t c(\nu_t, f(\nu_t)) \right] + E_\nu^{f^\infty \delta'} \left[\Gamma_n u^*(\nu_n) \right].
\end{aligned} \tag{3.3.8}$$

Observe that from Assumption 25.5 and (3.2.4) we have

$$E_\nu^{f^\infty \delta'} \left[\Gamma_n u^*(\nu_n) \right] \leq (\alpha^*)^n ||u^*||. \tag{3.3.9}$$

Hence letting $n \to \infty$ in (3.3.8), since $\alpha^* \in (0, 1)$, from (3.3.9) and (3.2.3) we get

$$u^*(\nu) \geq V(f^\infty, \delta', \nu), \quad \forall \nu \in \mathbb{P}(X). \tag{3.3.10}$$

Since $\delta' \in \Delta_\Theta$ is arbitrary, (3.2.5) yields

$$u^*(\nu) \geq V^*(\nu), \quad \nu \in \mathbb{P}(X). \tag{3.3.11}$$

On the other hand, since α is a continuous function in Θ, from Lemma 28, the compactness of Θ, and Proposition D.2, for each $(\nu, a) \in \mathbb{P}(X) \times A$, there exists $g : \mathbb{P}(X) \times A \to \Theta$ such that $g(\nu, a) \in \Theta$ satisfies

$$\begin{aligned} u^*(\nu) &= \min_{a \in A} \left\{ c(\nu, a) + \alpha(g) \int_{\mathbb{P}(X)} u^*(\nu') k(d\nu' | \nu, a) \right\} \\ &\leq c(\nu, a) + \alpha(g) \int_{\mathbb{P}(X)} u^*(\nu') k(d\nu' | \nu, a), \quad \forall \nu \in \mathbb{P}(X), a \in A. \end{aligned} \tag{3.3.12}$$

For an arbitrary policy $\delta \in \Delta$, iterating (3.3.12) we obtain

$$u^*(\nu) \leq V(\delta, g^\infty, \nu), \quad \forall \nu \in \mathbb{P}(X). \tag{3.3.13}$$

where $g^\infty = \{g\} \in \Delta_{M_\theta}$.

On the other hand, because $\Delta_{M_\theta} \subset \Delta_\Theta$, observe that

$$V(\delta, g^\infty, \nu) \leq \sup_{\delta' \in \Delta_{M_\theta}} V(\delta, \delta', \nu) \leq \sup_{\delta' \in \Delta_\Theta} V(\delta, \delta', \nu).$$

Then, from (3.3.13) we conclude (because δ is arbitrary)

$$u^*(\nu) \leq \inf_{\delta \in \Delta} \sup_{\delta' \in \Delta_\Theta} V(\delta, \delta', \nu) = V^*(\nu), \quad \nu \in \mathbb{P}(X). \tag{3.3.14}$$

Therefore, combining (3.3.11) and (3.3.14) we prove that $V^*(\nu) = u^*(\nu)$, $\nu \in \mathbb{P}(X)$.

(c) The existence of $f^* : \mathbb{P}(X) \to A$ follows from Remark 27. Now, similar as in (3.3.10) we have that for an arbitrary $\delta' \in \Delta_\Theta$

$$V^*(\nu) \geq V(f_*^\infty, \delta', \nu), \quad \forall \nu \in \mathbb{P}(X),$$

which implies that

$$V^*(\nu) = \sup_{\delta' \in \Delta_\Theta} V(f_*^\infty, \delta', \nu), \quad \nu \in \mathbb{P}(X).$$

Hence Theorem 29 has been proved. □

Chapter 4

POMDP with Partially Observable Random Discount Factors

4.1 Introduction

In this chapter we study a partially observable Markov decision process (POMDP) with partially observable (PO) random discount factors. Specifically, we assume that the costs are exponentially discounted with accumulative random discount rate at stage t of the form $e^{-\sum_{k=0}^{t-1}\alpha_k}$ where $\{\alpha_t\}$ is a stochastic process which is assumed to be partially observable. In this case, a cost c incurred at stage t is equivalent to a cost $e^{-\sum_{k=0}^{t-1}\alpha_k}c$ at time $t=0$. Under this setting, we will take advantage of the theory developed in previous chapters in order to propose an appropriate model.

4.2 The partially observable control problem

We consider the a partially observable Markov decision process (POMDP) evolving according to the system equation

$$\begin{aligned} x_{t+1} &= F_1(x_t, a_t, w_t^{(1)}) \quad t \in \mathbb{N}_0 \\ y_t &= F_2(x_t, w_t^{(2)}), \quad t \in \mathbb{N}_0, \end{aligned} \tag{4.2.1}$$

where F_1 and F_2 are known functions, and, as before, x_t, a_t and y_t are the state, control and observation at time t, taking values in Borel spaces X, A and Y, respectively. The set A is assumed to be compact. In addition, $\{w_t^{(1)}\}$ and $\{w_t^{(2)}\}$ are independent sequences of i.i.d. random variables with values in Borel spaces S_1 and S_2 with distributions $\theta_1 \in \mathbb{P}(X)$ and $\theta_2 \in \mathbb{P}(X)$, respectively. We assume that x_0 has distribution $\underline{\nu} \in \mathbb{P}(X)$.

Let Q and K be the transition kernel, and the observation kernel defined by F_1 and F_2 as follows:

$$Q(B|x_t,a_t) := \int_{S_1} 1_B[F_1(x_t,a_t,w^{(1)})]\theta_1(dw^{(1)}) \text{ and } K(B'|x_t) := \int_{S_2} 1_{B'}[F_2(x_t,w^{(2)})]\theta_2(dw^{(2)}), \tag{4.2.2}$$

with $B \in \mathcal{B}(X)$ and $B' \in \mathcal{B}(Y)$.

On the other hand, we consider a stochastic process $\{\alpha_t\}$, representing the discount factor, which is partially observable in the following sense:

$$\begin{aligned} \alpha_{t+1} &= G_1(\alpha_t, \xi_t^{(1)}), \quad t \in \mathbb{N}, \\ \beta_t &= G_2(\alpha_t, \xi_t^{(2)}), \quad t \in \mathbb{N}, \end{aligned} \tag{4.2.3}$$

where $\alpha_t \in \Gamma := (\epsilon_0, \infty)$ for some $\epsilon_0 > 0$, $\beta_t \in \Sigma := (0, \infty)$ represent the observation, $\{\xi_t^{(1)}\}$ and $\{\xi_t^{(2)}\}$ are independent sequences of i.i.d., random variables with values in Borel spaces R_1 and R_2 with distribution $\rho_1 \in \mathbb{P}(R_1)$ and $\rho_2 \in \mathbb{P}(R_2)$, respectively. Furthermore, the functions $G_1 : \Gamma \times R_1 \to \Gamma$ and $G_2 : \Gamma \times R_2 \to \Sigma$ are known. We assume that α_0 has a distribution $\underline{\eta} \in \mathbb{P}(\Gamma)$.

Similarly as (4.2.2), let Q' and K' be the discount factor transition kernel and the corresponding observation kernel defined by G_1 and G_2 as:

$$Q'(C|\alpha_t) := \int_{R_1} 1_C[G_1(\alpha_t, \xi^{(1)})]\rho_1(d\xi^{(1)}) \text{ and } K'(C'|\alpha_t) := \int_{R_2} 1_{C'}[G_2(\alpha_t, \xi^{(2)})]\rho_2(d\xi^{(2)}), \quad (4.2.4)$$

with $C \in \mathcal{B}(\Gamma)$ and $C' \in \mathcal{B}(\Sigma)$.

The discount factor process plays the following role. Let $\tilde{c} : X \times A \to \mathbb{R}$ the one-stage cost function and define $\tilde{\epsilon}(\alpha) := e^{-\alpha}$, with $\alpha \in \Gamma$. Then the discounted cost incurred at time t is $c(x_0, a_0)$ for $t = 0$, and for $t \in \mathbb{N}$,

$$\tilde{\epsilon}(\alpha_0) \cdots \tilde{\epsilon}(\alpha_{t-1})\tilde{c}(x_t, a_t), \quad t \geq 1. \quad (4.2.5)$$

Putting this elements all together, we can define the following extended PO control model

$$\mathcal{M}_{EPO} = (X, Y, A, Q, K, \underline{\nu}, \Gamma, \Sigma, Q', K', \underline{\eta}, \tilde{\epsilon}, \tilde{c}), \quad (4.2.6)$$

The control model (4.2.6) represents a controlled system in which the state and the discount factor are partially observable, whose dynamics can be described as follows: at time $t = 0$ the initial state x_0 has a given distribution $\underline{\nu}$. Then, a state-observation y_0 is generated according to the observation kernel K. Now, a control $a_0 \in A$ is applied and a cost $\tilde{c}(x_0, a_0)$ is generated. Then the system moves to state x_1 through the transition law Q and the discount-factor α_1 has a given distribution $\underline{\eta}$. Then, a state-observation y_1 is generated by K and a discount-factor-observation β_1 is generated according to K'. Now, a control $a_1 \in A$ is applied and a discounted cost $\epsilon(\alpha_1)\tilde{c}(x_1, a_1)$ is generated. Furthermore, if the system is in state x_t and the discount factor is α_t at time $t \geq 1$, two observations are generated; the state-observation y_t generated by the stochastic kernel K and the discount-factor-observation β_t generated by the stochastic kernel K'; next, a control $a_t \in A$ is applied, then (1) a discounted cost $\tilde{\epsilon}(\alpha_1) \cdots \tilde{\epsilon}(\alpha_t)\tilde{c}(x_t, a_t)$ is incurred; (2) the system moves to state x_{t+1} according to the transition law Q, and the discount factor moves to α_{t+1} according to the stochastic kernel Q'; and (3) the observations y_{t+1} and β_{t+1} are generated by the observations kernels K and K', respectively, and the process is repeated.

In order to define the class of control policies for the model (4.2.6), we consider histories of the form:

$$h_0 := (\underline{\nu}, y_0) \in \mathbb{H}_0 := \mathbb{P}(X) \times Y,$$

$$h_1 := (\underline{\nu}, y_0, a_0, \underline{\eta}, y_1, \beta_1) \in \mathbb{H}_1 := \mathbb{H}_0 \times A \times \mathbb{P}(\Gamma) \times Y \times \Sigma$$

$$h_t := (\underline{\nu}, y_0, a_0, \underline{\eta}, y_1, \beta_1, a_1, y_2, \beta_2, ..., y_{t-1}, \beta_{t-1}, a_{t-1}, y_t, \beta_t) \in \mathbb{H}_t := \mathbb{H}_{t-1} \times A \times Y \times \Sigma, \quad t \geq 2.$$

Hence

$$h_t := (h_{t-1}, a_{t-1}, y_t, \beta_t) \in \mathbb{H}_t := \mathbb{H}_{t-1} \times A \times Y \times \Sigma, \quad t \geq 2.$$

Now, control policies are defined as follows.

Definition 30. A control policy is a sequence of stochastic kernels $\pi = \{\pi_t\}$, $t \in \mathbb{N}_0$ on A given $\mathbb{H}_t$. A control policy is *Markovian* if there are measurable functions $f_t : X \times \Gamma \to A$, such that for all $h_t \in \mathbb{H}_t$ and $t \in \mathbb{N}_0$,

$$\pi_t(C|h_t) = 1_C[f_t(x_t, \alpha_t)], \quad C \in \mathcal{B}(A),$$

and *stationary*, if there is a measurable function $f : X \times \Gamma \to A$, such that for all $h_t \in \mathbb{H}_t$ and $t \in \mathbb{N}_0$

$$\pi_t(C|h_t) = 1_C[f(x_t, \alpha_t)], \quad C \in \mathcal{B}(A).$$

Again, we denote by Π the set of all control policies, Π_M the set of Markov policies.

Keeping in mind (4.2.5), the cost are accumulated in an infinite horizon under the following discounted optimality criterion. For each policy $\pi \in \Pi$ and initial distributions $\underline{\nu} \in \mathbb{P}(X)$ and $\underline{\eta} \in \mathbb{P}(\Gamma)$ we define

$$W(\pi, \underline{\nu}, \underline{\eta}) := E^{\pi}_{\underline{\nu},\underline{\eta}}\left[\sum_{t=0}^{\infty} \widetilde{\Lambda}_t \tilde{c}(x_t, a_t)\right]$$

where

$$\tilde{\Lambda}_t := \prod_{k=0}^{t-1} \tilde{\epsilon}(\alpha_k). \quad \text{if} \quad t \geq 1 \quad \text{and} \quad \tilde{\Lambda}_0 = 1,$$

and $E^{\pi}_{\underline{\nu},\underline{\eta}}$ is the expectation with respect to the probability measure $P^{\pi}_{\underline{\nu},\underline{\eta}}$, induced by $\pi, \underline{\nu}, \underline{\eta}$. Hence the extended PO-OCP is to find a policy $\pi^* \in \Pi$ such that

$$W^*(\underline{\nu}, \underline{\eta}) := \inf_{\pi \in \Pi} W(\pi, \underline{\nu}, \underline{\eta}) = W(\pi^*, \underline{\nu}, \underline{\eta}), \quad \underline{\nu} \in \mathbb{P}(X), \underline{\eta} \in \mathbb{P}(\Gamma). \tag{4.2.7}$$

In this case, the extended POMDP takes the form $(\Omega, \mathcal{F}, P^{\pi}_{\underline{\nu},\underline{\eta}}, \{x_t, \alpha_t, y_t, \beta_t\})$, where $\Omega = (X \times Y \times A \times \Gamma \times \Sigma)^{\infty}$ and $\mathcal{F} = \mathcal{B}(\Omega)$.

4.3 The extended completely observable control model

Following the standard approach applied in previous chapters, the analysis of the partially observed optimal control problem (PO-OCP) (4.2.7) is based on its transformation into a completely observable optimal control problem (CO-OCP) by means of a suitable filtering process. Under the present scenario, because we have two independent partially observable processes, namely, the state process and the discount process, we introduce two filtering procedures, which will then be coupled in the corresponding optimality equation.

Specifically, by applying the same procedure given (1.6.2)-(1.6.8), for each $\pi \in \Pi$, we have that there are functions $\Psi_0 : \mathbb{P}(X) \times Y \to \mathbb{P}(X)$ and $\Psi : \mathbb{P}(X) \times A \times Y \to \mathbb{P}(X)$, defining the filtering process $\{\nu_t\} \subset \mathbb{P}(X)$ as

$$\begin{aligned} \nu_0(B) &= P^{\pi}_{\underline{\nu},\underline{\eta}}[x_0 \in B|h_0], \quad B \in \mathcal{B}(X), \\ \nu_{t+1}(B) &= P^{\pi}_{\underline{\nu},\underline{\eta}}[x_{t+1} \in B|h_{t+1}], \quad B \in \mathcal{B}(X), \end{aligned} \tag{4.3.1}$$

and

$$\nu_0 = \Psi_0(\underline{\nu}, y_0), \quad \text{and} \quad \nu_{t+1} = \Psi(\nu_t, a_t, y_{t+1}), t \in \mathbb{N}. \tag{4.3.2}$$

Let k and k_0 the corresponding stochastic kernels (see (1.6.4) and (1.6.5)) defined as

$$k(D_1|\nu, a) := \int_Y 1_{D_1}[\Psi(\nu, a, y)] R'(dy|\nu, a), \quad D_1 \in \mathcal{B}(\mathbb{P}(X)), (\nu, a) \in \mathbb{P}(X) \times A, \tag{4.3.3}$$

or equivalently

$$k(D_1|\nu, a) := \int_X \int_X \int_Y 1_{D_1}[\Psi(\nu, a, y)] K(dy|x') Q(dx'|x, a) \nu(dx);$$

and

$$k_0(D_1|\underline{\nu}) := \int_Y 1_{D_1}[\Psi_0(\underline{\nu}, y)]R_0'(dy|\underline{\nu}), \quad D_1 \in \mathcal{B}(\mathbb{P}(X)), \underline{\nu} \in \mathbb{P}(X), \tag{4.3.4}$$

or equivalently

$$k_0(D_1|\underline{\nu}) := \int_X \int_Y 1_{D_1}[\Psi_0(\underline{\nu}, y)]K(dy|x)\underline{\nu}(dx).$$

On the other hand, proceding similarly, we can obtain a filtering procedure for the discount factor process. That is, for each $\pi \in \Pi$, there exists functions $\Phi, \Phi_0 : \mathbb{P}(\Gamma) \times \Sigma \to \mathbb{P}(\Sigma)$ defining the discount filtering process $\{\eta_t\} \subset \mathbb{P}(\Gamma)$ as

$$\begin{aligned} \eta_0(B) &= P^{\pi}_{\underline{\nu},\underline{\eta}}[\alpha_0 \in B|h_0], \quad B \in \mathcal{B}(\Gamma), \\ \eta_{t+1}(B) &= P^{\pi}_{\underline{\nu},\underline{\eta}}[\alpha_{t+1} \in B|h_{t+1}], \quad B \in \mathcal{B}(\Gamma), \end{aligned} \tag{4.3.5}$$

and

$$\nu_0 = \Phi_0(\underline{\eta}, \beta_0), \quad \text{and} \quad \eta_{t+1} = \Phi(\eta_t, \beta_{t+1}), t \in \mathbb{N}. \tag{4.3.6}$$

Furtheremore, let $\tilde{k}$ and $\tilde{k_0}$ be the stochastic kernels

$$\tilde{k}(D_2|\eta) := \int_\Sigma I_{D_2}[\Phi(\eta, \beta)]R'(d\beta|\eta), \quad D_2 \in \mathcal{B}(\mathbb{P}(\Gamma)), \eta \in \mathbb{P}(\Gamma), \tag{4.3.7}$$

or equivalently

$$\tilde{k}(D_2|\eta) := \int_\Gamma \int_\Gamma \int_\Sigma 1_{D_2}[\Phi(\eta, \beta)]K'(d\beta|\alpha')Q'(d\alpha'|\alpha)\eta(d\alpha)$$

and

$$\tilde{k}_0(D_2|\underline{\eta}) := \int_\Sigma 1_{D_2}[\Phi_0(\underline{\eta}, \beta)]R_0'(d\beta|\underline{\eta}), \quad D_2 \in \mathcal{B}(\mathbb{P}(\Gamma)), \tag{4.3.8}$$

or equivalently

$$\tilde{k}_0(D_2|\underline{\eta}) := \int_\Gamma \int_\Sigma 1_{D_2}[\Phi_0(\underline{\eta}, \beta)]K'(d\beta|\alpha)\underline{\eta}(d\alpha).$$

Now, we define the one-stage cost function $c : \mathbb{P}(X) \times A \to \mathbb{R}$ as

$$c(\nu, a) := \int_X \tilde{c}(x, a)\nu(dx).$$

In addition, we define the function $\epsilon : \mathbb{P}(\Gamma) \to (0, \infty)$ as

$$\epsilon(\eta) := \int_\Gamma \tilde{\epsilon}(\alpha)\eta(d\alpha) \tag{4.3.9}$$

and

$$\Lambda_t := \prod_{k=0}^{t-1} \epsilon(\eta_k), \quad t \geq 1 \quad \text{and} \quad \Lambda_0 = 1.$$

All the previous elements define the following extended completely observable control model (CO-CM):

$$\mathcal{M}_{ECO} = (\mathbb{P}(X), A, k, k_0, \mathbb{P}(\Gamma), \tilde{k}, \tilde{k}_0, \epsilon, c). \tag{4.3.10}$$

In order to define the performance index corresponding to the control model (4.3.10), it is necessary to define the class of information policies, which in this case, we call it *extended information policies*.

To this end, let

$$\bar{i}_0 := (\nu_0) \in \bar{\mathbb{I}}_0 := \mathbb{P}(X),$$
$$\bar{i}_t := (\nu_0, a_0, \nu_1, \eta_1, a_1, ..., \nu_{t-1}, \eta_{t-1}, a_{t-1}, \nu_t, \eta_t) \in \bar{\mathbb{I}}_t := \bar{\mathbb{I}}_{t-1} \times A \times \mathbb{P}(X) \times \mathbb{P}(\Gamma), \quad t \in \mathbb{N}.$$

Definition 31. An *extended information policy* (or $\bar{i}$-policy) is a sequence of stochastic kernels $\overline{\delta} = \{\overline{\delta}_t\}$ on A given $\bar{\mathbb{I}}_t$, that is to say, $\overline{\delta}_t(da|\bar{i}_t)$. We denote $\overline{\Delta}$ as the set of all $\bar{i}$-policies. The class of *Markov $\bar{i}$-policies* and the class of *stationary $\bar{i}$-policies* are defined in a similar way as in Definition 30, and are denoted as $\overline{\Delta}_M$ and $\overline{\mathbb{F}}$, respectively. Note that, as before, $\overline{\mathbb{F}} \subset \overline{\Delta}_M \subset \overline{\Delta}$.

Now, for each $\bar{i}$-policy $\overline{\delta} \in \overline{\Delta}$ and initial pair of distributions $(\underline{\nu}, \underline{\eta}) \in \mathbb{P}(X) \times \mathbb{P}(\Gamma)$ we define the performance index as

$$V(\overline{\delta}, \underline{\nu}, \underline{\eta}) := E^{\overline{\delta}}_{\underline{\nu},\underline{\eta}}\left[\sum_{t=0}^{\infty} \Lambda_t c(\nu_t, a_t)\right], \tag{4.3.11}$$

where $E^{\overline{\delta}}_{\underline{\nu},\underline{\eta}}$ is the expectation with respect to the probability measure $P^{\overline{\delta}}_{\underline{\nu},\underline{\eta}}$, induced by $\pi, \underline{\nu}, \underline{\eta}$. Hence the extended CO-OCP is to find an $\bar{i}$-policy $\overline{\delta}_* \in \overline{\Delta}$ such that

$$V^*(\underline{\nu}, \underline{\eta}) := \inf_{\Pi} V(\pi, \underline{\nu}, \underline{\eta}) = V(\overline{\delta}_*, \underline{\nu}, \underline{\eta}), \quad \underline{\nu} \in \mathbb{P}(X), \underline{\eta} \in \mathbb{P}(\Gamma). \tag{4.3.12}$$

4.4 Solution of the extended completely observable control problem

As before, in order to ensure a solution of the completely observable optimal control problem (4.3.10), we impose conditions on the model (4.3.10). In particular, we follow the *contraction mapping approach*. Turns out that such conditions are similar to Assumption 25 (see also Lemma 19).

Assumption 32.

1. A is a compact set.
2. $c(\nu, a) \in \mathcal{C}(\mathbb{P}(X) \times A)$.
3. The stochastic kernels k, $\tilde{k}$, k_0 and $\tilde{k}_0$ are weakly continuous.

Now, we define the *dynamic programming operators*. For $u \in \mathcal{C}(\mathbb{P}(X) \times \mathbb{P}(\Gamma))$ and $\overline{f} \in \overline{\mathbb{F}}$, let

$$Tu(\nu, \eta) := \min_{a \in A} \left\{ c(\nu, a) + \epsilon(\eta) \int_{\mathbb{P}(\Gamma)} \int_{\mathbb{P}(X)} u(\nu', \eta') k(d\nu'|\nu, a) \tilde{k}(d\eta'|\eta) \right\} \tag{4.4.1}$$

and

$$T_{\overline{f}} u(\nu, \eta) := c(\nu, \overline{f}) + \epsilon(\eta) \int_{\mathbb{P}(\Gamma)} \int_{\mathbb{P}(X)} u(\nu', \eta') k(d\nu'|\nu, \overline{f}) \tilde{k}(d\eta'|\eta). \tag{4.4.2}$$

Remark 33. Let $u(\nu, \eta) \in \mathcal{C}(\mathbb{P}(X) \times \mathbb{P}(\Gamma))$, then, $\int_{\mathbb{P}(\Gamma)} \int_{\mathbb{P}(X)} u(\nu', \eta') k(d\nu'|\nu, \overline{f}) \tilde{k}(d\eta'|\eta)$ are bounded and continuous by Assumption 32.3; and by Assumptiton 32.2, c is bounded and continuous. Then,

$$v(\nu, \eta, a) := c(\nu, a) + \epsilon(\eta) \int_{\mathbb{P}(\Gamma)} \int_{\mathbb{P}(X)} u(\nu', \eta') k(d\nu'|\nu, \overline{f}) \tilde{k}(d\eta'|\eta)$$

is bounded and continuous on $a \in A$ for all $(\nu, \eta) \in \mathbb{P}(X) \times \mathbb{P}(\Gamma)$. Next, from Proposition D.2, there exists a measurable selector $f^* : \mathbb{P}(X) \times \mathbb{P}(\Gamma) \to A$ such that

$$v(\nu, \eta, \overline{f}_*(\nu, \eta)) = \min_{a \in A} v(\nu, \eta, a), \quad \forall (\nu, \eta) \in \mathbb{P}(X) \times \mathbb{P}(\Gamma).$$

The following Lemma is consequence of Assumption 32.

Lemma 34. Under Assumption 32 we have:

1. $Tu(\nu,\eta) \in \mathcal{C}(\mathbb{P}(X)\times\mathbb{P}(\Gamma))$ for all $u \in \mathcal{C}(\mathbb{P}(X)\times\mathbb{P}(\Gamma))$.
2. The operators T and $T_{\overline{f}}$, are contraction operators modulus $\epsilon(\eta)$.
3. There exist a unique function $u^* \in \mathcal{C}(\mathbb{P}(X)\times\mathbb{P}(\Gamma))$ and a unique function $u^*_{\overline{f}} \in \mathcal{C}(\mathbb{P}(X)\times\mathbb{P}(\Gamma))$ such that
$$Tu^* = u^* \quad \text{and} \quad Tu^*_{\overline{f}} = u^*_{\overline{f}},$$
and moreover, for any function $u \in \mathcal{C}(\mathbb{P}(X)\times\mathbb{P}(\Gamma))$
$$||T^n u - u^*|| \to 0 \quad \text{and} \quad ||T^n u_{\overline{f}} - u^*_{\overline{f}}|| \to 0, \quad \text{as } n \to \infty$$

Proof.

1. This part is consequence of Remark 33.
2. Let $u, u' \in \mathcal{C}(\mathbb{P}(X)\times\mathbb{P}(\Gamma))$. Then for $(\nu,\eta) \in \mathbb{P}(X)\times\mathbb{P}(\Gamma)$ we have
$$\big|Tu(\nu,\eta) - Tu'(\nu,\eta)\big|$$
$$= \left| \min_{a\in A}\left\{ c(\nu,a) + \epsilon(\eta)\int_{\mathbb{P}(\Gamma)}\int_{\mathbb{P}(X)} u(\nu',\eta')k(d\nu'|\nu,a)\tilde{k}(d\eta'|\eta)\right\}\right.$$
$$\left. - \min_{a\in A}\left\{ c(\nu,a) + \epsilon(\eta)\int_{\mathbb{P}(\Gamma)}\int_{\mathbb{P}(X)} u'(\nu',\eta')k(d\nu'|\nu,a)\tilde{k}(d\eta'|\eta)\right\}\right|$$
$$\leq \max_{a\in A}\left| c(\nu,a) + \epsilon(\eta)\int_{\mathbb{P}(\Gamma)}\int_{\mathbb{P}(X)} u(\nu',\eta')k(d\nu'|\nu,a)\tilde{k}(d\eta'|\eta)\right.$$
$$\left. - c(\nu,a) - \epsilon(\eta)\int_{\mathbb{P}(\Gamma)}\int_{\mathbb{P}(X)} u'(\nu',\eta')k(d\nu'|\nu,a)\tilde{k}(d\eta'|\eta)\right|$$
$$= \max_{a\in A}\left|\epsilon(\eta)\int_{\mathbb{P}(\Gamma)}\int_{\mathbb{P}(X)} u(\nu',\eta')k(d\nu'|\nu,a)\tilde{k}(d\eta'|\eta) - \epsilon(\eta)\int_{\mathbb{P}(\Gamma)}\int_{\mathbb{P}(X)} u'(\nu',\eta')k(d\nu'|\nu,a)\tilde{k}(d\eta'|\eta)\right|$$
$$= \epsilon(\eta)\max_{a\in A}\left|\int_{\mathbb{P}(\Gamma)}\int_{\mathbb{P}(X)}\Big(u(\nu',\eta') - u'(\nu',\eta')\Big)k(d\nu'|\nu,a)\tilde{k}(d\eta'|\eta)\right|$$
$$\leq \epsilon(\eta)\max_{a\in A}\int_{\mathbb{P}(\Gamma)}\int_{\mathbb{P}(X)}\big|u(\nu',\eta') - u'(\nu',\eta')\big|k(d\nu'|\nu,a)\tilde{k}(d\eta'|\eta)$$
$$\leq \epsilon(\eta)||u - u'||.$$
Therefore, taking $\sup_{(\nu,\eta)\in\mathbb{P}(X)\times\mathbb{P}(\Gamma)}$ we obtain
$$||Tu - Tu'|| \leq \epsilon(\eta)||u - u'||.$$
Similarly is proved that $T_{\overline{f}}$ is a contraction operator.

3. This part follows from the Banach Fixed Point Theorem (see Proposition A.1).

Hence, Lemma 34 has been proved. □

Now, we relate the fixed point $u^*(\nu,\eta)$ to $V(\overline{\delta}_*,\nu,\eta)$.

Theorem 35. *Under Assumption 32 the following holds:*

(a) The optimal discounted cost function $V^*:\mathbb{P}(X)\times\mathbb{P}(\Gamma)\to\mathbb{R}$ satisfies the optimality equation, that is

$$V^*(\nu,\eta)=\min_{a\in A}\left\{c(\nu,a)+\epsilon(\eta)\int_{\mathbb{P}(\Gamma)}\int_{\mathbb{P}(X)}V^*(\nu',\eta')k(d\nu'|\nu,a)\tilde{k}(d\eta'|\eta)\right\},\quad(\nu,\eta)\in\mathbb{P}(X)\times\mathbb{P}(\Gamma). \tag{4.4.3}$$

(b) There exists $\overline{f}_*:\mathbb{P}(X)\times\mathbb{P}(\Gamma)\to A$ such that

$$V^*(\nu,\eta)=c(\nu,\overline{f}_*)+\epsilon(\eta)\int_{\mathbb{P}(\Gamma)}\int_{\mathbb{P}(X)}V^*(\nu',\eta')k(d\nu'|\nu,\overline{f}_*)\tilde{k}(d\eta'|\eta),\quad(\nu,\eta)\in\mathbb{P}(X)\times\mathbb{P}(\Gamma). \tag{4.4.4}$$

Moreover, the stationary policy $\overline{f}_*^\infty=\{\overline{f}_*\}$ is an optimal control $\overline{i}$-policy, that is,

$$V^*(\nu,\eta)=V(\overline{f}_*^\infty,\nu,\eta),\quad(\nu,\eta)\in\mathbb{P}(X)\times\mathbb{P}(\Gamma).$$

Proof.

(a) By Lemma 34.3, we just need to prove that $u^*=V^*$. According to (4.4.1) we have that

$$u^*(\nu,\eta)=Tu^*(\nu,\eta).$$

Therefore,

$$u^*(\nu,\eta)\le c(\nu,a)+\epsilon(\eta)\int_{\mathbb{P}(\Gamma)}\int_{\mathbb{P}(X)}u^*(\nu',\eta')k(d\nu'|\nu,a)\tilde{k}(d\eta'|\eta),\ \nu\in\mathbb{P}(X),\eta\in\mathbb{P}(\Gamma),a\in A. \tag{4.4.5}$$

Now, for an arbitrary policy $\overline{\delta}\in\overline{\Delta}$ iteration of the inequality (4.4.5) yields

$$\begin{aligned}u^*(\nu,\eta)&\le E^{\overline{\delta}}_{\underline{\nu},\underline{\eta}}\left[c(\nu_0,a_0)+\sum_{t=1}^{n-1}\prod_{k=0}^{t-1}\epsilon(\eta_{k+1})c(\nu_t,a_t)+\prod_{k=0}^{n-1}\epsilon(\eta_{k+1})u^*(\nu_n,\eta_n)\right]\\&=E^{\overline{\delta}}_{\underline{\nu},\underline{\eta}}\left[c(\nu_0,a_0)+\sum_{t=1}^{n-1}\prod_{k=0}^{t-1}\epsilon(\eta_{k+1})c(\nu_t,a_t)\right]+E^{\overline{\delta}}_{\underline{\nu},\underline{\eta}}\left[\prod_{k=0}^{n-1}\epsilon(\eta_{k+1})u^*(\nu_n,\eta_n)\right]\\&=E^{\overline{\delta}}_{\underline{\nu},\underline{\eta}}\left[\sum_{t=0}^{n-1}\Lambda_t c(\nu_t,a_t)\right]+E^{\overline{\delta}}_{\underline{\nu},\underline{\eta}}\left[\Lambda_n u^*(\nu_n,\eta_n)\right].\end{aligned} \tag{4.4.6}$$

Now, note that because $\alpha>\epsilon_0>0$ we have that $\tilde{\epsilon}(\alpha):=e^{-\alpha}<e^{-\epsilon_0}<1$. Hence, from (4.3.9),

$$\epsilon(\eta)=\int_\Gamma\tilde{\epsilon}(\alpha)\eta(d\alpha)<e^{-\epsilon_0}<1,\quad\eta\in\mathbb{P}(\Gamma),$$

which in turns implies

$$\Lambda_n=\prod_{k=0}^{n-1}\epsilon(\eta_k)\le(e^{-\epsilon_0})^n. \tag{4.4.7}$$

Therefore

$$E^{\overline{\delta}}_{\underline{\nu},\underline{\eta}}\left[\Lambda_n u^*(\nu_n,\eta_n)\right] \leq (e^{-\epsilon_0})^n ||u^*||. \tag{4.4.8}$$

Hence, letting $n \to \infty$ in (4.4.6), from (4.4.8) we get

$$u^*(\nu,\eta) \leq V(\overline{\delta},\nu,\eta), \quad (\nu,\eta) \in \mathbb{P}(X)\times\mathbb{P}(\Gamma), \overline{\delta} \in \overline{\Delta}.$$

Therefore, since $\overline{\delta} \in \overline{\Delta}$ is arbitrary, (4.3.12) yields

$$u^*(\nu,\eta) \leq V^*(\nu,\eta), \quad (\nu,\eta) \in \mathbb{P}(X)\times\mathbb{P}(\Gamma). \tag{4.4.9}$$

On the other hand, now we will prove that $u^*(\nu,\eta) \geq V^*(\nu,\eta)$. By (4.4.2) and Remark 33, we know that

$$u^*(\nu,\eta) = T_{\overline{f}} u^*(\nu,\eta)$$

therefore,

$$u^*(\nu,\eta) = c(\nu,\overline{f}) + \epsilon(\eta)\int_{\mathbb{P}(\Gamma)}\int_{\mathbb{P}(X)} u^*(\nu',\eta')k(d\nu'|\nu,\overline{f})\tilde{k}(d\eta'|\eta), \quad (\nu,\eta)\in\mathbb{P}(X)\times\mathbb{P}(\Gamma), a \in A. \tag{4.4.10}$$

Now, iteration of the inequality (4.4.10) yields

$$\begin{aligned} u^*(\nu,\eta) &= E^{\overline{f}^\infty}_{\underline{\nu},\underline{\eta}}\left[c(\nu_0,a_0) + \sum_{t=1}^{n-1}\prod_{k=0}^{t-1}\epsilon(\eta_{k+1})c(\nu_t,a_t) + \prod_{k=0}^{n-1}\epsilon(\eta_{k+1})u^*(\nu_n,\eta_n)\right] \\ &= E^{\overline{f}^\infty}_{\underline{\nu},\underline{\eta}}\left[c(\nu_0,a_0) + \sum_{t=1}^{n-1}\prod_{k=0}^{t-1}\epsilon(\eta_{k+1})c(\nu_t,a_t)\right] + E^{\overline{f}^\infty}_{\underline{\nu},\underline{\eta}}\left[\prod_{k=0}^{n-1}\epsilon(\eta_{k+1})u^*(\nu_n,\eta_n)\right] \\ &= E^{\overline{f}^\infty}_{\underline{\nu},\underline{\eta}}\left[\sum_{t=0}^{n-1}\Lambda_t c(\nu_t,a_t)\right] + E^{\overline{f}^\infty}_{\underline{\nu},\underline{\eta}}\left[\Lambda_n u^*(\nu_n,\eta_n)\right]. \end{aligned} \tag{4.4.11}$$

Again, we get that

$$E^{\overline{f}^\infty}_{\underline{\nu},\underline{\eta}}\left[\Lambda_n u^*(\nu_n,\eta_n)\right] \leq (e^{-\epsilon_0})^n ||u^*||.$$

Hence, letting $n \to \infty$ in (4.4.11)

$$u^*(\nu,\eta) = V(\overline{f}_\infty,\nu,\eta), \quad (\nu,\eta)\in\mathbb{P}(X)\times\mathbb{P}(\Gamma), \overline{f}_\infty \in \overline{\mathbb{F}}. \tag{4.4.12}$$

Therefore, since $\overline{F} \subset \overline{\Delta}$,

$$u^*(\nu,\eta) \geq \inf_{\overline{f}_\infty\in\overline{\mathbb{F}}} V(\overline{f}_\infty,\nu,\eta) \geq \inf_{\overline{\delta}\in\overline{\Delta}} V(\overline{\delta},\nu,\eta), \quad (\nu,\eta)\in\mathbb{P}(X)\times\mathbb{P}(\Gamma).$$

Hence,

$$u^*(\nu,\eta) \geq V^*(\nu,\eta), \quad (\nu,\eta)\in\mathbb{P}(X)\times\mathbb{P}(\Gamma). \tag{4.4.13}$$

Finally, combining (4.4.9) and (4.4.13) we prove part (a).

(b) The existence of $\overline{f}_* : \mathbb{P}(X)\times\mathbb{P}(\Gamma) \to A$ follows from Remark 33. On the other hand, similar as in (4.4.12) we have that $\overline{f}_*^\infty$ is $\overline{i}$-optimal.

Hence Theorem 35 has been proved. □

Appendix A

Spaces of functions, functions and contraction operators

A *Borel Space* is a Borel subset of a complete separable metric space. A Borel space is always endowed with the Borel σ-algebra $\mathcal{B}(X)$, that is, the smallest σ-algebra of subsets of X that contains all the open sets in X. In this sense, "measurable", for either sets or functions, means "Borel measurable".

For a Borel space X, we define the following spaces:

1. $\mathbb{B}(X)$, Banach space of real bounded measurable functions on X with the supremum norm

$$||v|| := \sup_{x \in X} |v(x)|$$

2. $\mathcal{C}(X) \subset \mathbb{B}(X)$, Banach space of bounded and continuous functions on X with the supremum norm.

Let (V, d) be a metric space. A functional mapping T from V into itself is said to be a *contraction operator* if for some β satisfying $0 < \beta < 1$ (called the modulus of T) one has

$$d(Tu, Tv) \leq \beta d(u, v), \quad \forall u, v \in V.$$

An element $v^* \in V$ is called a fixed point of T if $Tv^* = v^*$. For a functional $T : V \to V$, the functional T^n is defined recursively by $T^n := T(T^{n-1})$, for all $n = 1, 2, ...$ where T^0 is the identity.

Proposition A.1 (Banach's Fixed Point Theorem)**.** *If T is a contraction operator mapping a complete metric space (V, d) into itself, then T has unique fixed point, say v^*. Furthermore, for any $v \in V$ and $n \geq 0$,*

$$d(T^n v, v) \leq \beta^n d(v, v^*).$$

The function $v_n := Tv_{n-1} = T^n v$ are called successive approximations.

Proof. See [19] and references therein. □

Proposition A.2. *Let X be an arbitrary non-empty set, and let u and v be functions from X to $\mathbb{R}$ bounded from above (so that $\sup u$ and $\sup v$ are finite). Then,*

$$|\sup_x u(x) - \sup_x v(x)| \leq \sup_x |u(x) - v(x)|, \quad \text{and} \quad |\inf_x u(x) - \inf_x v(x)| \leq \sup_x |u(x) - v(x)|$$

Proof. See [19] and references therein. □

Appendix B

Probability Measures

Let X be a Borel space and $\mathcal{B} = \mathcal{B}(X)$ its Borel σ-algebra. We denote by $\mathbb{P}(X)$ the space of probability measures on X. Sea $P, P_1, P_2, ...$ be probability measures on X. It is said that P_n *converges weakly* to P if, as $n \to \infty$,

$$\int f dP_n \to \int f dP, \quad \forall f \in \mathcal{C}(X),$$

where $\mathcal{C}(X)$ is the space of real valued, bounded and continuous functionson X, with the sup norm. Equivalentemente P_n *converges weakly* to P if

$$\int u dP_n \to \int u dP$$

for every unifromly continuous function $u \in \mathcal{C}(X)$. We will always understand $\mathbb{P}(X)$ as a topological space with the topology of weak convergence. In such case, since X is a Borel space, $\mathbb{P}(X)$ is also a Borel space (see e.g. [8]).

Appendix C

Stochastic Kernels

Let X and Y be Borel spaces. A *stochastic kernel* on X given Y is a function $q(dx|y)$ such that for each $y \in Y$, $q(\cdot|y)$ is a probability measure on X, and for each Borel set $B \in \mathcal{B}(X)$, $q(B|\cdot)$ is a measurable function from Y to $[0,1]$. Equivalently, a collection of probability measures $q(dx|y)$ on X parametrized by $y \in Y$ is a stochastic kernel, if and only if, the function $h : Y \to \mathbb{P}(X)$ defined by

$$h(y) := q(\cdot|y)$$

is measurable. A stochastic kernel $q(dx|y)$ on X given Y is said *continuous* if the function h is continuous, that is, $q(\cdot|y_n)$ converges weakly to $q(\cdot|y)$ whenever y_n converges to y. Then, the stochastic kernel $q(dx|y)$ is continuous if $\int v(x)q(dx|y)$ is a continuous function of $y \in Y$ for every function $v \in \mathcal{C}(X)$.

Proposition C.1. *Let $q(dx|y)$ be a stochastic kernel on X given Y, $f(x,y)$ be a real valued measurable function on $X \times Y$, and $f' : Y \to \mathbb{R}$ be the function defined by*

$$f'(y) := \int f(x,y)q(dx,y)$$

whenever the integral exist. Then

- *If $f \in B(X \times Y)$, then $f' \in B(Y)$.*
- *If $q(dx|y)$ is continuous and $f \in \mathcal{C}(X \times Y)$, then $f' \in \mathcal{C}(Y)$.*

Proof. See Lemma 5.2 in [22]. □

Proposition C.2 (Ionescu Tulcea)**.** *Let $X_1, X_2, ...$ be a sequence of Borel spaces, and define $Y_n := X_1 \times \ldots \times X_n$, and $Y := X_1 \times X_2 \times \ldots$. Let $p \in \mathbb{P}(X_1)$ be a given probability measure and for $n = 1, 2, ...$, let $q_n(dx_{n+1}|y_n)$ be a stochastic kernel on X_{n+1} given Y_n. Then, for each $n \geq 2$, there exists a unique probability measure $r_n \in \mathbb{P}(Y_n)$ such that for all $B_i \in \mathcal{B}(X_i)$, where $i = 1, ..., n$*

$$r_n(B_1 \times \ldots \times B_n) = \int_{B_1} p(dx_1) \int_{B_2} q_1(dx_2|x_1) \ldots \int_{B_n} q_{n-1}(dx_n|x_1, ..., x_{n-1})$$

Mmoreover, if a measurable function f on Y_n is r_n-integrable, then

$$\int_{Y_n} f dr_n = \int_{X_1} p(dx_1) \int_{X_2} q_1(dx_2|x1) \ldots \int_{X_n} f(x_1, ..., x_n) q_{n-1}(dx_n|x_1, ..., x_{n-1}).$$

Finally, there exists a unique probability measure r on Y, sometimes writen as

$$r = pq_1q_2 \ldots$$

such that for each n, the marginal of r on Y_n is r_n.

Proof. See [7] and references therein. □

Appendix D

Multifunctions and Measurable Selectors

Throughout the following, X and A denote Borel spaces. A mapping D which associates with each $x \in X$ a non-empty subset $D(x)$ of A is called a *multifunction* (or set-valued function) from X to A. In the text, X and A denote, respectively, the state space and the action set in a Markov decision model, and, in our case, $D(x) = A(x) = A$.

A multifunction D from X to A is said to be

1. *Borel measurable* if $D^{-1}[G]$ is a Borel subset of X for every open set $G \subset A$;
2. *upper semicontinuous (u.s.c.)* if $D^{-1}[F]$ is closed in X for every closed set $F \subset A$;
3. *lower semicontinuous (l.s.c.)* if $D^{-1}[G]$ is open in X for every open set $G \subset A$;
4. *continuous* if it is both u.s.c. and l.s.c.

A multifunction is said to be *compact-valued* if $D(x)$ is a compact set for all $x \in X$.

Given a Borel measurable multifunction D from X to A, and $\mathbb{F}$ denotes the set of measurable functions $f : X \to A$ with $f : X \to A$ with $f(x) \in D(x)$ for all $x \in X$. A function f is called a *selector* for the multifunction D. Moreover, for a given measurable function $v : X \times A \to \mathbb{R}$ we denote

$$v^*(x) := \inf_{D(x)} v(x, a), \quad x \in X.$$

Proposition D.1. *Suppose that D is compact valued. If $v(x, \cdot)$ is l.s.c. on $D(x)$ for every $x \in X$, then there exist a selector $f^* \in \mathbb{F}$ such that*

$$v(x, f^*(x)) = v^*(x) = \min_{D(x)} v(x, a), \quad x \in X$$

and v^* is measurable.

Let $\mathcal{K}(A)$ be the collection of all non-empty compact subsets of A topologized by the Hausdorff metric H. If (A, d) is separable, then $(\mathcal{K}(A), H)$ is a separable metric space (see e.g., [19] and references therein).

Proposition D.2. *Let $D : X \to \mathcal{K}(A)$ be a Borel measurable multifunction, and let $v(x, a)$ be a real valued measurable function on $X \times A$ such that $v(x, a)$ is upper semicontinuous (u.s.c.) in $a \in D(x)$, for each $x \in X$, Then:*

1. *There exists a selector $f : X \to A$ for D, such that*

$$v(x, f(x)) = \max_{a \in D(x)} v(x, a), \forall x \in X,$$

and the function $v^*(x) := \max_{a \in D(x)} v(x, a)$ is measurable.

2. if D is continuous and v is continuous and bounded, then v^* is continuous and bounded.

Proof. See e.g., [19], [7] and references therein. □

Appendix E

Notation

E.1 Abbreviations

- a.s.: almost surely (i.e., with probability 1).
- CO: completely observable.
- CO-CM: completely observable control model.
- COMDP: completely observable Markov decision process.
- ECO: extended completely observable.
- EPO: extended partially observable.
- i.i.d.: independent and identically distributed.
- MDP Markov decision process.
- OCP: optimal control problem.
- PO: partially observable.
- PO-CM: partially observable control model.
- POMDP: partially observable Markov decision process.

E.2 Symbols

- $\mathbb{N}$: set of positive integer numbers.
- $\mathbb{N}_0$: set of non-negative integer numbers.
- $\mathbb{R}$: set of real numbers with usual topology.
- $\mathcal{B}(X)$: Borel σ-algebra of X.
- $\mathcal{C}(X)$: Banach space of bounded and continuous functions on X.
- $\mathbb{P}(X)$: space of probability measures on X.
- $\mathbb{H}_t$: space of histories h_t up to time t.
- Π: set of control policies.
- Π_M: set of Markov policies.

- $\mathbb{F}$: set of stationary policies.
- $||v|| = \sup_{x \in X} |v(x)|$: supremum norm.
- P_ν^π: probability measure when the policy $\pi \in \Pi$ is used and the initial state x_0 has distribution $\nu \in \mathbb{P}(X)$.
- E_ν^π: expectation operator with respect to the probability measure P_ν^π.

Bibliography

[1] M.L. Arcega-López. Procesos de control de markov parcialmente observables con aplicaciones a sistemas de inventarios. Master's thesis, Universidad de Sonora, 2013.

[2] A. Bensoussan, M. Cakanyildirim, J.A. Minjarez-Sosa, and S.P. Sethi. Inventory problems with partially observed demands and lost sales. *Journal of Optimization Theory and Applications*, 136:321–340, 2008.

[3] A. Bensoussan, M. Cakanyildirim, J.A. Minjarez-Sosa, S.P. Sethi, and R. Shi. Partially observed inventory systems: the case of rain checks. *SIAM Journal on Control and Optimization*, 47:2490–2519, 2008.

[4] A. Bensoussan, M. Cakanyildirim, J.A. Minjarez-Sosa, S.P. Sethi, and R. Shi. An incomplete information inventory model with presence of inventories or backorders as only observations. *Journal of Optimization Theory and Applications*, 146:544–580, 2010.

[5] A. Bensoussan, M. Cakanyildirim, and S.P. Sethi. Partially observed inventory systems: the case of zero-balance walk. *SIAM Journal on Control and Optimization*, 46:176–209, 2007.

[6] A. Bensoussan, M. Cakanyldirim, and S.P. Sethi. Partially observed inventory systems: the case of zero-balance walk. *SIAM Journal on Control and Optimization*, 46:176–209, 2007.

[7] D.P. Bertsekas and S.E. Shreve. *Stochastic Optimal Control: The Discrete Time Case*. Academic Press, New York, 1 edition, 1978.

[8] P. Billingsley. *Probability and Measure*. John Wiley and Sons, New York, 3 edition, 1995.

[9] E.B Dynkin and A.A Yushkevich. *Controlled Markov Processes*. Springer-Verlag, New York, 1 edition, 1979.

[10] R.J Elliot, L. Aggoun, and J.B. Moore. *Hidden Markov Models: Estimation and Control*. Springer-Verlag, New York, 1 edition, 1994.

[11] Y.H. Garcia, S. Diaz-Infante, and J.A. Minjarez-Sosa. Partially observable queueing systems with controlled service rates under a discounted optimality criterion. *Kybernetika*, To appear 2021.

[12] J. Gonzalez-Hernández, R.R López-Martínez, J.A Minjarez-Sosa, and J.R. Gabriel-Arguelles. Constrained markov control process with randomized discounted cost criteria: occupation measures and extremal points. *Risk and Decision Analysis*, 4:163–176, 2013.

[13] J. Gonzalez-Hernández, R.R López-Martínez, J.A Minjarez-Sosa, and J.R. Gabriel-Arguelles. Constrained markov control process with randomized discounted rate: infinite linear programming approach. *Optimal Control Applications and Methods*, 35:575–591, 2014.

[14] J. González-Hernández, R.R López-Martínez, and J.A. Minjarez-Sosa. Adaptive policies for stochastic systems under a randomized discounted criterion. *Boletín de la Sociedad Matemática Mexicana*, 14:149–163, 2008.

[15] J. González-Hernández, R.R López-Martínez, and J.A. Minjarez-Sosa. Approximation, estimation and control of stochastic systems under a randomized discounted cost criterion. *Kybernetyka*, 45:737–754, 2009.

[16] J. González-Hernández, R.R López-Martínez, and R. Pérez-Hernández. Markov control processes with randomized discounted cost in borel space. *Mathematical Methods of Operation Research*, 65:27–44, 2007.

[17] T.J. González-Trejo, O. Hernandez-Lerma, and L.F. Hoyos-Reyes. Minimax control of discrete-time stochastic systems. *SIAM Journal on Control and Optimization*, 41:1626–1659, 2003.

[18] O. Hernandez-Lerma and R.Romera. Limiting discounted-cost control of partially observable stochastic system. In *Proceedings of the 41st IEEE Conference on Decision and Control*, volume 2, pages 2189–2194, Las Vegas, Nevada, 2002.

[19] O. Hernández-Lerma. *Adaptive Markov Control Processes*. Springer-Verlag, New York, 1989.

[20] J.A. Minjarez-Sosa. Approximation and estimation in markov control processes under discounted criterion. *Kybernetyka*, 40:681–690, 2004.

[21] J.A. Minjarez-Sosa. Markov control models with unknown random state- action-dependent discount factors. *TOP*, 23:743–772, 2015.

[22] Nowak A. S. Semcontinuous nonstationary stochastic games ll. *Mathematical Analysis and Applications*, 148:22–43, 1990.

[23] A.A. Yushkevich. Reduction of a controlled markov model with incomplete data to a problem with complete information in the case of borel state and control spaces. *Theory of Probability and its Applications*, 21:153–158, 1976.