UNIVERSIDAD DE SONORA

División de Ciencias Exactas y Naturales

Programa de Posgrado en Matemáticas

Algoritmo de Iteración de Políticas Aproximado en Modelos Markovianos y Semi-Markovianos bajo en Criterio de Costo Descontado

T E S I S

que presenta:

M.C. María Teresa Robles Alcaráz

para obtener el título de:

Doctor en Ciencias **Matemáticas**

Director de tesis: Dr. Oscar Vega Amaya
Dr. Jesús Adolfo Minjárez Sosa

Hermosillo, Sonora, México, 1 de Julio de 2016

SINODALES

Dr. Jorge Álvarez Mena
Universidad Veracruzana, Xalapa, Veracruz, México.

Dr. Fernando Luque Vásquez
Universidad de Sonora, Hermosillo, México.

Dr. Jesús Adolfo Minjárez Sosa
Universidad de Sonora, Hermosillo, México.

Dr. Carlos Gabriel Pacheco González
Centro de Investigación y Estudios Avanzados del I.P.N., Ciudad de México.

Dr. Óscar Vega Amaya
Universidad de Sonora, Hermosillo, México.

UNIVERSIDAD DE SONORA

"El Saber de Mis Hijos
Hará Mi Grandeza"

El saber de mis hijos
hará mi grandeza

MARIA TERESA
ROBLES ALCARAZ
211290159

ACTA DE EXAMEN DE GRADO

En la ciudad de Hermosillo, Sonora, siendo las 11:00 horas del día 1 de julio de 2016, se reunieron en el Auditorio del Departamento de Matemáticas de la Universidad de Sonora, los integrantes del jurado:

DR. JESÚS ADOLFO MINJÁREZ SOSA
DR. OSCAR VEGA AMAYA
DR. CARLOS GABRIEL PACHECO GONZÁLEZ
DR. JORGE ALVAREZ MENA
DR. FERNANDO LUQUE VÁSQUEZ

bajo la presidencia del primero y fungiendo como secretario el último, para realizar el examen de grado del programa de Doctor en Ciencias Matemáticas, a:

MARIA TERESA ROBLES ALCARAZ

quien de acuerdo a la opción de examen de grado presentó un trabajo de investigación titulado

"Algoritmo de Iteración de Políticas Aproximado en Modelos Markovianos y Semi-Markovianos Bajo el Criterio de Costo Descontado"

El jurado, después de debatir entre sí reservada y libremente, emitió el siguiente dictamen:

APROBADA POR UNANIMIDAD

y para constancia se levantó la presente acta.

Acta: 12
Foja: 12
Libro: I

DR. JESUS ADOLFO MINJÁREZ SOSA
Presidente

DR. FERNANDO LUQUE VASQUEZ
Secretario

DR. OSCAR VEGA AMAYA
Sinodal

DR. CARLOS GABRIEL PACHECO GONZÁLEZ
Sinodal externo

DR. JORGE ALVAREZ MENA
Sinodal externo

DR. JESÚS ADOLFO MINJÁREZ SOSA, Coordinador del Programa de Doctor en Ciencias Matemáticas de la Universidad de Sonora, hace constar que las firmas que anteceden corresponden al jurado que intervino en el examen de grado.

Hermosillo, Sonora, a 1 de julio de 2016

DR. JESÚS ADOLFO MINJÁREZ SOSA
Coordinador de programa

DIVISIÓN DE CIENCIAS EXACTAS Y NATURALES, COORDINACION POSGRADO EN MATEMATICAS

Folio: 7541

Índice general

Abstract

The approximate dynamic programming (ADP) is a broad family of algorithms aiming to compute approximated solutions in Markov decision processes (MDPs). It has attracted the attention of many researchers not only because it extended the application field of the MDPs theory, but also because the analysis of the approximate algorithms rises up several theoretical challenges that relate different mathematical fields (approximation theory, stochastic algorithms, statistical estimation, neural networks, etc.)

Concerning to the basic aspects, the departure point of the ADP is the general theory on the classical MDPs algorithms–either for the discounted cost criterion or the average cost criterion. Then, the ADP algorithms can be classified in terms of the algorithm in which they are based, thus defining the subfamilies of procedures generically known as *approximate value iteration*(AVI) algorithm, *approximate policy iteration* (API) algorithm, and *approximate linear programming* (ALP) approach. Roughly speaking, the ADP algorithms interleave an approximation step at each stage of the classical algorithms. This simple idea has produced a vast variety of competing approximate algorithms, but the most part of the literature is concentrated on finite space models. See, for instance, the books [5, 27] and the survey papers [4, 28, 29, 33] for general discussions and comprehensive accounts on the ADP algorithms.

In particular, the API algorithm have been mainly studied for finite spaces models ([4, 5, 8, 27, 29]), but there are some few exceptions ([3, 12, 22, 23, 34, 36, 43]). The present work deals with the API algorithm for a discounted cost criterion with Borel state and action spaces, but differs from these latter papers in several ways.

To begin, in the present work is used a "perturbation approach" for the analysis of the API algorithm. This is done by considering approximation schemes that define a class of operators called averagers. An *averager* is a positive linear operator with norm equal to one that satisfies an additional continuity property. The averagers include several of the most used interpo-

lation schemes (e.g., piecewise linear and multilinear approximations, piecewise constant approximations, kernel-based approximations, certain splines approximations) and some variants of the projected equation method.

The key property of an averager is that allows to see the approximation step as a "perturbation" of the original Markov decision model and the approximate algorithm as the exact algorithm in the "perturbed model". These facts in turn have two important consequences: (i) they allow to establish directly the convergence of API algorithm and to identify the limit functions as the optimal value function on a perturbed model; (ii) they also allow to give finite-time performance bounds in terms of the quality of the approximations given by the averagers for the one-step cost function and the transition law of the original model. Usually the performance bounds are asymptotic and are not tied to the accuracy of the approximation step. It is also worth mentioning that the results in (i) and (ii) are obtained without imposing any other structural condition than the standard continuity and compactness conditions used in the MDP's literature.

On the other hand, the papers [3, 12, 22, 23, 43] follow a free-model approach so their approximations rely on simulation of the controlled processes and prove the convergence either in the mean or with a "high probability"or almost surely under several technical conditions. The papers [3, 12] provide a finite-time bounds which holds with a high probability for models with compact state space and finite action space. Moreover, the bounds are not given in term of the accuracy of the approximation schemes and seem to be quite difficult to estimate. The work [34] is of experimental nature since it focus in comparing the performance of several discretization procedures using some test models. The paper [36] focus on the rate of convergence of the API algorithm for a growth model using a piecewise linear interpolation. The papers [22, 23] prove the almost sure convergence of the API algorithm assuming that the state and control spaces are compact convex subsets of a euclidian space and also that the transition law has a density among other technical conditions. The paper [43] presents results for denumerable space models and claims that the results can be extended to general spaces.

The papers [2, 24, 38] also study the discounted performance criterion for systems with general spaces but using the *approximate value iteration* algorithm. In general terms, they share one or more features mentioned for the references [3, 12, 22, 23, 43].

In the present work the API algorithm is studied following a based-model approach under two settings. In the first one, the one-step cost function and the transition law of the system are completely known. Thus, the performance bounds of the API algorithm are given in terms of the approximation

errors of the costs and of the transition law, and the stopping error of the API iteration algorithm. In the second setting, the controlled system evolves according to a difference equation of the form

$$x_{n+1} = H(x_n, a_n, w_n), \quad n = 0, 1, 2, ...,$$

where H is a given function, x_n and a_n are the state and the control variables at time n, respectively, and the disturbance process $\{w_n\}$ is an observable sequence of independent and identically distributed random vectors with *unknown* density ρ. The density ρ is estimated using an historical record of the random disturbances $\mathbf{w}_t := (w_0, w_1, \ldots, w_{t-1})$ by means of a density estimator $\rho_t(\cdot|\mathbf{w}_t)$, which in turns defines an estimated control model $\mathcal{M}_t$. Then, the API algorithm is applied in the estimated model $\mathcal{M}_t$, yielding a estimated-perturbed model $\tilde{\mathcal{M}}_t$ as well as an estimate-approximate policy iteration (EAPI) algorithm. Clearly, the performance of this new algorithm depends on the accuracy of the model $\mathcal{M}_t$ for estimating the unknown model $\mathcal{M}$, and also on the accuracy of model $\tilde{\mathcal{M}}_t$ for approximating $\mathcal{M}_t$. Then, in addition to the errors in the application of the API algorithm, there is an estimation error whose accuracy is determined by the density estimation process.

The problem of finding optimal policies under unknown disturbance distribution is called adaptive control problem, and has been studied in several contexts (see, e.g., [13, 14, 18, 19, 26]). Typically, in this case, the adaptive policies are obtained by applying the "principle of estimation and control" (see [14]) which consists in substituting the estimates into optimal stationary controls. That is, before choosing the control a_n, the controller gets an estimate ρ_n of the density ρ, and combines this with the history of the system to select a control $a_n = f_n^{\rho_n}(\cdot)$, defining the non-stationary control policy $\pi = \left\{f_n^{\rho_n}(\cdot)\right\}$. Thus, since the discounted criterion depends strongly on the decisions selected at the first stages, precisely when the information about the unknown density ρ is deficient, it is not possible to ensure the optimality of such a policy. This fact implies that the optimality of this class of policies is studied in a weaker sense, the so-called asymptotic discounted optimality. In contrast, the estimate-approximate policy iteration introduced in this paper offers an alternative to numerically approximate optimal stationary policies for control systems with unknown disturbance distribution.

Introducción

En este trabajo se estudian métodos de aproximación para el problema de control óptimo descontado en procesos de decisión markovianos y semi-markovianos con costos no-acotados, y espacios de estados y controles generales.

La importancia de esta clase de procesos se debe a que una diversidad de problemas que aparecen en campos como economía, investigación de operaciones, finanzas, teoría de inventarios, teoría de colas y modelos de mantenimiento-remplazo etc., se formulan como problemas de control óptimo ([14], [16], [20], [30], [32]).

Actualmente se cuenta con una teoría general bien desarrollada para este tipo de problemas; no obstante, la implementación numérica de algoritmos para el cálculo de las soluciones óptimas tiene limitantes importantes. Por un lado, la mayor parte de la literatura se restringe a procesos con espacio de estados finitos, lo cual excluye muchos problemas importantes (e.g.,teoría de inventarios, sistemas de espera, control de pesquerías, problemas de remplazo entre otros). Por otra parte, la mayoría de los algoritmos propuestos son procedimientos heurísticos, es decir, no se tiene garantía de su convergencia o no proporcionan cotas para los errores de aproximación expresadas en términos de cantidades conocidas o que puedan ser calculadas.

Esta situación claramente limita el campo de aplicación de los procesos de decisión markovianos y a su vez, plantea el reto de investigar y proponer algoritmos generales y esquemas de implementación numérica que permitan encontrar soluciones en problemas concretos con la precisión deseada.

El principal enfoque para resolver el problema de control óptimo consiste en establecer que la función de valor óptimo satisface la ecuación conocida como *ecuación de optimalidad*. Esta ecuación es una ecuación funcional que involucra un paso de optimización. Dicha ecuación generalmente no puede resolverse explícitamente, haciéndose necesario el desarrollo de enfoques y métodos numéricos que proporcionen soluciones aproximadas. Entre los principales algoritmos para obtener dichas aproximaciones destacan el

algoritmo de *iteración de valores–o aproximaciones sucesivas–*y el algoritmo de *iteración de políticas*. Ambos algoritmos proporcionan un procedimiento iterativo para aproximar tanto a la función de valor óptimo como a las políticas óptimas. Sin embargo, los aspectos de implementación numérica son particularmente difíciles debido al fenómeno conocido como la *maldición de la dimensión*.

En la práctica, la tarea de implementar cada iteración de estos algoritmos es difícil si la cardinalidad del espacio de estados es muy grande y simplemente imposible si el espacio de estados es infinito. Como una alternativa para enfrentar esta dificultadad surge la *programación dinámica aproximada* (PDA) ([4], [5],[24], [34], [38], [40], [41]) la cual consiste en combinar un algoritmo iterativo–como los antes mencionados–con métodos de aproximación de funciones. En muchos casos, los métodos de aproximación son representados por medio de un operador L que actúa sobre algún espacio de funciones adecuado, de modo que Lu representa la aproximación de la función u. Así, la PDA alterna cada etapa del algoritmo iterativo con un paso de aproximación dado por el operador L para obtener un algoritmo aproximado.

En este trabajo se estudiará el algoritmo de *iteración de políticas aproximado* (IPA) para procesos markovianos y semi-markovianos con respecto al índice de funcionamiento de costo descontado. Para la parte de aproximación se usará una clase de operadores de aproximación denominados *promediadores*.

Los *promediadores* son operadores de aproximación que satisfacen las siguientes propiedades: (a) $L(1) = 1$; (b) L es un operador lineal y monótono; (c) $Lu_n \downarrow 0$ siempre que $u_n \downarrow 0$. Esta clase de aproximadores incluye los esquemas de interpolación más comunes como interpolación lineal y multilineal, interpolaciones con splines de Schoenberg y basados en kérneles, así como interpolaciones con operdores de Bernstein y Hermite-Féjer, y algunas variantes de los métodos de proyección, entre otros ([33], [38]).

Las propiedad clave de un promediador es que puede verse como una probabilidad de transición sobre el espacio de estados del sistema. Esta propiedad permite introducir un *modelo de control perturbado* en el cual el algortimo IPA es el algoritmo de iteración de políticas exacto. Este hecho hace posible, por una parte, establecer la convergencia del algoritmo directamente, así como identificar la naturaleza de la función límite, y por otra, proporcionar cotas de error en términos de la precisión con la que el promediador aproxima a la función de costo por etapa y a la ley de transición del sistema.

Este enfoque de *perturbaciones* basado en promediadores fue estudiado

previamente en [40] y [41] pero enfocados al algoritmo de *iteración de valores* con respecto a los índices en costo descontado y en costo promedio. En el presente trabajo continuamos el estudio de este enfoque para el problema de control óptimo descontado pero considerando ahora al algoritmo de *iteración de politicas*, tanto para procesos markovianos como semi-markovianos. El trabao está estructurado de la siguiente manera.

En el Capítulo 1 se describe el modelo de control markoviano estándar y se presentan los principales resultados sobre el problema de control óptimo con costos acotados que usaremos posteriormente en los Capítulos 2 y 3. Dichos resultados son bien conocidos y un caso particular de los resultados que se presentan en el Capítulo 4, pero se presentan en este capítulo con el ánimo de facilitar la lectura.

En el Capítulo 2 se introducen los promediadores, se presentan algunas de sus propiedades, el modelo perturbado y posteriormente se estudia el algoritmo de iteración de políticas aproximado. En el Capítulo 3 se extiende el análisis a problemas de control con información *incompleta*; específicamente, consideraremos sistemas controlados que evolucionan de acuerdo a una ecuación en diferencias bajo el supuesto de que la densidad del ruido del sistema es *desconocida*. A este tipo de problemas se les conoce como problemas de *control adaptado y* el enfoque usual para abordarlos se basa en el *principio de estimación y control* ([13], [14],[18], [26]). En el presente trabajo damos una alternativa a dicho enfoque que consiste en combinar el enfoque de pertubaciones basado en promediadores con técnicas de estimación estadística para estimar y aproximar políticas óptimas.

El Capítulo 4 está orientado completamente al estudio del problema de control óptimo descontado para procesos semi-markovianos con costos *no-acotados* y espacios de *Borel*. El problema de control óptimo se estudia bajo dos conjuntos de hipótesis: (i) se supone que el modelo de control satisface condiciones usuales de continuidad-compacidad; (ii) la función de costo por etapa y la dinámica del sistema satisfacen una condición de crecimiento. Bajo tales supuestos, se caracteriza a la función de valor óptimo como la única solución de la ecuación de optimalidad en un espacio de funciones con norma ponderada finita y se demuestra la existencia de políticas estaciona rias óptimas. La condición de crecimiento sobre la dinámica del sistema se expresa a través de una condición de Lyapunov, que en nuestra opinión, no había sido considerada previamente en la literatura hasta el momento ([7], [42], [30], [32], [21]).

En el Capítulo 5 se estudia el algoritmo de iteración de políticas aproximado para procesos de control semi-markovinanos. Se parte de los resultados desarrollados en el capítulo anterior para este tipo de procesos. Para efectos

de aproximación numérica se requiere truncar el espacio de estados, con esta finalidad supondremos que es un espacio localmente compacto. Esta condición junto con la condición de Lyapunov nos permitirá restringir el estudio a un conjunto compacto así como obtener cotas de error por truncar el espacio de estados.

Con el mismo enfoque de promediadores se estudia un modelo perturbado combinándolo con un procedimiento de truncamiento.

El algoritmo de iteración de políticas se implementa en un modelo perturbado en un subconjunto compacto del espacio de estados original. Se caracteriza la función de valor óptimo y se asegura la existencia de políticas estacionarias óptimas en el modelo truncado. Además, se proporcionan cotas para el error producido por el truncamiento, entre otros resultados.

Capítulo 1

Problema de Control Óptimo Descontado

1.1. Introducción

En este capítulo se introducen los procesos de control markovianos en tiempo discreto y se presentan los principales resultados sobre el problema de control óptimo descontado con costos acotados. Estos resultados se usaran posteriormente en los Capítulos 2 y 3. La existencia de políticas óptimas se obtiene bajo condiciones usuales de continuidad y compacidad.

1.2. Problema de control óptimo

A lo largo del presente trabajo usaremos la siguiente notación. Dado un espacio de Borel X–es decir, un subconjunto de Borel de un espacio métrico separable y completo– denotaremos por $\mathcal{B}(X)$ a su sigma-álgebra de Borel, esto es, la mínima sigma-álgebra que contiene la familia de subconjuntos abiertos de X. La medibilidad tanto de conjuntos como funciones se entenderá como medibilidad con respecto a la sigma-álgebra de Borel.

Denotaremos por $M(X)$ el espacio de las funciones medibles $u : X \to \mathbb{R}$. El subespacio de las funciones medibles acotadas se denotará como $M_b(X)$ y el subespacio de las funciones continuas acotadas por $C_b(X)$. Para cada $v \in M_b(X)$, la norma del supremo se define como $\|v\|_\infty := \sup_{x\in X} |v(x)|$. Para cada $B \in \mathcal{B}(X)$, denotaremos por I_B a su función indicadora.

Dados dos espacios de Borel X y Y, un *kérnel estocástico* $\gamma(\cdot|\cdot)$ en X dado Y es un mapeo tal que:

(i) $\gamma(\cdot|y)$ es una medida de probabilidad en X para cada $y \in Y$;

(*ii*) $\gamma(B|\cdot)$ es una función medible en Y para cada $B \in \mathcal{B}(X)$.

Denotamos por $\mathbb{N}$ y $\mathbb{N}_0$ a los conjuntos de los enteros positivos y de los enteros no-negativos, respectivamente; por $\mathbb{R}$ y $\mathbb{R}_+$ denotamos a los conjunto de los números reales y de los reales no-negativos, respectivamente.

Como ya se mencionó con anterioridad, para definir el problema de control óptimo (PCO) se requiere un modelo de control, un conjunto de políticas admsibles y un índice de funcionamiento o función objetivo. Estos elementos se introducen a continuación.

Definición 1.2.1 *Un modelo de control markoviano (MCM)*

$$\mathcal{M} = (\mathbf{X}, \mathbf{A}, \{A(x) : x \in \mathbf{X}\}, Q, C), \tag{1.1}$$

consiste en los siguientes elementos:

(a) $\mathbf{X}$ *y* $\mathbf{A}$ *son espacios de Borel y representan el espacio de estados* $\mathbf{X}$ *y el espacio de controles* $\mathbf{A}$*;*

(b) $A(x)$ *es un subconjunto medible de* $\mathbf{A}$ *para cada* $x \in \mathbf{X}$ *y representa el conjunto de controles o acciones admisibles para el estado* $x \in \mathbf{X}$*. Supondrémos que el conjunto de pares estado-acción admisibles*

$$\mathbb{K} := \{(x, a) \in \mathbf{X} \times \mathbf{A} : x \in \mathbf{X}, a \in A(x)\}$$

es un subconjunto medible de $\mathbf{X} \times \mathbf{A}$*.*

(c) la ley de transición de sistema $Q(\cdot|\cdot, \cdot)$ *es un kernel estocástico en* $\mathbf{X}$ *dado* $\mathbb{K}$*;*

(d) la función de costo por etapa $C : \mathbb{K} \to \mathbb{R}$ *es una función medible.*

Un modelo de control markoviano representa un sistema estocástico controlado que evoluciona de la siguiente manera: en el tiempo inicial $n = 0$ se observa el sistema en un estado $x_0 = x \in \mathbf{X}$ y se elige un control $a_0 = a \in A(x)$ con un costo $C(x, a)$. Inmediatamente despues de que se eligió el control, el sistema se mueve a un nuevo estado $x_1 = x' \in \mathbf{X}$ de acuerdo a la medida de probabilidad $Q(\cdot|x, a)$ y se elige un control $a_1 = a' \in A(x')$ con un costo $C(x', a')$, y así sucesivamente. Denotaremos por x_n al estado del sistema en el tiempo $n \in \mathbb{N}_0$ y por a_n al control seleccionado en este tiempo.

Para cada $n \in \mathbb{N}_0$, definimos el espacio de *historias admisibles* hasta tiempo n como

$$\mathbb{H}_n := \mathbb{K}^n \times \mathbf{X}, \quad n \in \mathbb{N}, \quad \text{y} \quad \mathbb{H}_0 := \mathbf{X}$$

Un elemento genérico $h_n \in \mathbb{H}_n$ es un vector de la forma

$$h_n = (x_0, a_0, ..., x_{n-1}, a_{n-1}, x_n),$$

donde $(x_k, a_k) \in \mathbb{K}$ para $k = 0, 1 \ldots, n-1$, y $x_n \in \mathbf{X}$. Observemos que el conjunto $\mathbb{H}_n$ contiene toda la información sobre la evolución del sistema hasta el tiempo n.

Denotemos por $\mathbb{F}$ al conjunto de funciones medibles $f : \mathbf{X} \to \mathbf{A}$ tal que $f(x) \in A(x)$ para toda $x \in \mathbf{X}$. A esta clase de funciones le llamaremos **selectores**.

Definición 1.2.2 *(i) Una política de control admisible es una sucesión $\pi = \{\pi_n\}$ de kérneles estocásticos en $\mathbf{A}$ dado $\mathbb{H}_n$ tal que $\pi_n(A(x_n)|h_n) = 1$ para todo $h_n \in \mathbb{H}_n, n \in \mathbb{N}_0$. Denotemos por Π al conjunto de todas las políticas admisibles.*

(ii) Una política $\pi = \{\pi_n\}$ es determinista si para cada $n \in \mathbb{N}_0$ existe una función medible $f_n : \mathbb{H}_n \to \mathbf{A}$ tal que $\pi_n(D|h_n) = I_D(f_n(h_n))$ para todo $D \in \mathcal{B}(\mathbf{A})$, $h_n \in \mathbb{H}_n$.

(iii) Una política $\pi = \{\pi_n\}$ es markoviana si existe una sucesión $\{f_n\} \subset \mathbb{F}$ tal que $\pi_n(D|h_n) = I_D(f_n(x_n))$ para todo $h_n \in \mathbb{F}, n \in \mathbb{N}_0$.

(iv) Una política $\pi = \{\pi_n\}$ es estacionaria si existe $f \in \mathbb{F}$ tal que $\pi_n(D|h_n) = I_D(f(x_n))$ para todo $D \in \mathcal{B}(\mathbf{A})$, $h_n \in \mathbb{H}_n, n \in \mathbb{N}_0$. En este caso la política $\pi = \{\pi_n\}$ se identifica con el selector f y a la clase de todas las políticas estacionarias con la familia de los selectores $\mathbb{F}$.

Sea $\Omega := (\mathbf{X} \times \mathbf{A})^\infty$ y $\mathcal{F}$ la σ-algebra producto correspondiente. Por el teorema de Ionescu-Tulcea ([1, Theorem 2.7.2, p. 109]) para cada política $\pi \in \Pi$ y cada medida de probabilidad $\nu \in \mathbb{P}(\mathbf{X})$ existe un proceso estocástico $\{(x_n, a_n)\}$ y una medida de probabilidad P_ν^π definidos ambos sobre $(\Omega, \mathcal{F})$ que satisfacen las siguientes propiedades:

(i) $P_\nu^\pi[x_0 \in B] = \nu(B)\ \forall B \in \mathcal{B}(\mathbf{X})$;

(ii) $P_\nu^\pi[a_n \in D|h_n] = \pi_n(D|h_n)\ \forall D \in \mathcal{B}(\mathbf{A}), h_n \in \mathbb{H}_n, n \in \mathbb{N}_0$;

(iii) $P_\nu^\pi[x_{n+1} \in B|h_n, a_n] = Q(B|x_n, a_n), \quad \forall B \in \mathcal{B}(\mathbf{X}), h_n \in \mathbb{H}_n, n \in \mathbb{N}_0$.

Al operador esperanza con respecto a la medida de probabilidad P_ν^π lo denotaremos por E_ν^π. Si ν es una medida concentrada en el estado inicial $x_0 = x$, entonces escribiremos P_x^π y E_x^π en lugar de P_ν^π y E_ν^π, respectivamente.

Nos referiremos a $\{x_n\}$ como el *proceso de estados* y a $\{a_n\}$ como el *proceso de control* inducidos por la política π y la distribución inicial ν. Observe que las propiedades (ii) y (iii) establecen que dichos procesos tiene

una propiedad de pérdida de memoria. En particular, si $\pi = f \in \mathbb{F}$ es una política estacionaria, entonces el proceso de estados $\{x_n\}$ es una cadena de Markov estacionaria con probabilidad de transición $Q(B|x, f(x)), B \in \mathcal{B}(\mathbf{X}), x \in \mathbf{X}$.

Para plantear el problema de control óptimo solamente hace falta especificar el índice de funcionamiento o función objetivo con el cual se medirá el desempeño del sistema bajo las diferentes políticas control. Sea $\alpha \in (0, 1)$ un factor de descuento fijo; se define el *costo descontado* como:

$$V_\pi(x) := E_x^\pi \sum_{t=0}^{\infty} \alpha^n C(x_n, a_n), \quad x \in \mathbf{X}, \pi \in \Pi. \tag{1.2}$$

La *función de valor óptimo* correspondiente se define como

$$V_*(x) := \inf_{\pi \in \Pi} V_\pi(x), \quad x \in \mathbf{X}. \tag{1.3}$$

El *problema de control óptimo* en costo descontado consiste en encontrar una política $\pi^* \in \Pi$ tal que

$$V_*(x) = V_{\pi^*}(x) \ \forall \ x \in \mathbf{X}.$$

En este caso, diremos que π^* es *óptima* en costo descontado o, simplemente, que π^* es *óptima descontada*.

1.3. Existencia de políticas óptimas descontadas

Existen diferentes conjuntos de condiciones que garantizan que las funciones introducidas en (1.2) y (1.3) están bien definidas así como la existencia de políticas óptimas. A continuación presentamos dos de los conjuntos de tales condiciones que son usadas con mayor frecuencia en la literatura.

Hipótesis 1.3.1 ***(a)*** *$C(x, a)$ es una función continua en $\mathbb{K}$ y acotada por una constante $M > 0$;*
(b) *La multifunción $x \to A(x)$ es continua con valores en los conjuntos compactos.*
(c) *$Q(\cdot|x, a)$ es débilmente continua en $\mathbb{K}$, es decir, la función*

$$(x, a) \to \int_{\mathbf{X}} u(y) Q(dy|x, a)$$

es continua para cada $u \in C_b(\mathbf{X})$.

Hipótesis 1.3.2 *(a) $C(\cdot,\cdot)$ es acotada por una constante $M > 0$.*
(b) $C(x,\cdot)$ es una función continua en $A(x)$ para cada $x \in \mathbf{X}$.
(c) La multifunción $x \to A(x)$ toma valores en los subconjuntos compactos de $\mathbf{A}$.
(d) $Q(\cdot|x,\cdot)$ es fuertemente continua en $A(x)$, es decir, la función

$$a \to \int_{\mathbf{X}} u(y)Q(dy|x,a)$$

es continua para cada $u \in M_b(\mathbf{X})$ y cada $x \in \mathbf{X}$.

Para cada función medible $v : \mathbb{K} \to \mathbb{R}$ y cada selector $f \in \mathbb{F}$ definimos

$$v_f(x) := v(x, f(x)), \quad x \in \mathbf{X}.$$

En particular, con esta notación tenemos

$$C_f(x) = C(x, f(x)) \quad \text{y} \quad Q_f(B|x) = Q(B|x, f(x)),$$

para todo $x \in \mathbf{X}$, $B \in \mathcal{B}(\mathbf{X})$.

Para cada $f \in \mathbb{F}$, se define el operador

$$T_f u(x) := C_f(x) + \alpha \int_{\mathbf{X}} u(y)Q_f(dy|x), \quad x \in \mathbf{X},$$

y el *operador de programación dinámica*

$$Tu(x) := \inf_{a \in A(x)} [C(x,a) + \alpha \int_{\mathbf{X}} u(y)Q(dy|x,a)], \quad x \in \mathbf{X},$$

para $u \in M_b(\mathbf{X})$.

Observación 1.3.3 *(a) Notemos que si $C(\cdot,\cdot)$ es acotada, entonces T_f, $f \in \mathbb{F}$, es un operador de contracción con módulo α en el espacio de Banach $(M_b(\mathbf{X}), ||\cdot||_\infty)$. Entonces, por el teorema de punto fijo de Banach, tiene un único u_f en $M_b(\mathbf{X})$.*
(b) Al mismo tiempo bajo la Hipótesis 1.3.1, el operador de programación dinámica T es un operador de contracción con módulo α en el espacio de Banach $(C_b(\mathbf{X}), ||\cdot||_\infty)$, por lo tanto, tienen un único punto fijo en $C_b(\mathbf{X})$. También por teoremas de selección estandar ([14], [31]), para cada función u en $C_b(\mathbf{X})$ existe un selector $f_u \in \mathbb{F}$ tal que $Tu(\cdot) = T_{f_u}(\cdot)$, es decir,

$$Tu(x) = C_{f_u}(x) + \alpha \int_{\mathbf{X}} u(y)Q_{f_u}(dy|x) \quad \forall x \in \mathbf{X}.$$

(c) Del mismo modo notemos que bajo la Hipótesis 1.3.2, se cumplen las mismas conclusiones que en (b) en el espacio de Banach $(M_b(\mathbf{X}), ||\cdot||_\infty)$.
(d) Además es importante resaltar que si en la Hipótesis 1.3.2(b) y (d) cambiamos "continuidad" por "semi-continuidad inferior" se cumplen las mismas conclusiones que en (c) ***(vea [14])***

Aplicando argumentos de programación dinámica se demuestra el siguiente resultado ([6], [14], [15]).

Proposición 1.3.4 *Si se cumple la Hipótesis 1.3.1 entonces:*
(a) la función de valor óptimo V_* *es el único punto fijo en* $C_b(\mathbf{X})$ *del operador* T*, es decir,*

$$V_*(x) = TV_*(x) = \min_{a\in A(x)} \left\{ C(x,a) + \alpha \int_{\mathbf{X}} V_*(y) Q(dy|x,a) \right\}; \qquad (1.4)$$

(b) una política estacionaria $f \in \mathbb{F}$ *es óptima si y sólo si* $V_*(\cdot) = T_f V_*(\cdot)$*;*
(c) existe una política estacionaria f^* *tal que* $V_* = T_{f^*} V_*$*; por lo tanto,* f^* *es óptima.*

Observación 1.3.5 *Bajo la Hipótesis 1.3.2 se cumplen los resultados de la Proposición 1.3.4, pero* V_* *es un punto fijo de* T *en* $M_b(\mathbf{X})$.

Desde un punto de vista teórico, la Proposición 1.3.4 da una solución completa del problema de control óptimo, pues la ecuación (1.4) caracteriza tanto a la función de valor óptimo como a las políticas estacionarias óptimas. A pesar de estos hechos, desde un punto de vista práctico el cálculo de las políticas óptimas es un reto debido a que la función de valor óptimo gene ralmente es desconocida. Para salvar este obstáculo se emplean algoritmos para aproximar la función de valor óptimo, y los procedimientos más comunes son el de iteración de valores e iteración de políticas, que describimos a continuación.

Algoritmo de iteración de valores (IV):

(i) inicio: tome $k = 0$ y elija una función $v_k \in C_b(\mathbf{X})$;

(ii) mejoramiento: calcule la función $v_{k+1} := Tv_k$ y una política f_{k+1} tal que $Tv_k = T_{f_{k+1}} u_k$.

Notemos que bajo la *Hipótesis 1.3.1,* por el teorema de punto fijo de Banach, la sucesión v_k, $k \in \mathbb{N}_0$, converge a V_* uniformemente para cualquier función inicial $v_0 \in C_b(\mathbf{X})$. Además se cumple

$$||V_* - v_k||_\infty \leq \frac{\alpha^k}{1-\alpha}||v_1 - v_0||_\infty \ \forall \ n \in \mathbb{N}_0.$$

Algoritmo iteración de políticas (IP):

(i) inicio: tomamos $k = 0$ y elija una política $g_0 \in \mathbb{F}$;

(ii) evaluación: dada $g_k \in \mathbb{F}$ calculemos $u_k := V_{g_k}$;

(iii) mejoramiento: encuentre $g_{k+1} \in \mathbb{F}$ tal que $Tu_k = T_{g_{k+1}}u_k$.

Note que el algoritmo de IP no está bien definido bajo la Hipótesis *1.3.1* debido a que la existencia de los selectores en el paso de mejoramiento no está garantizada. Este problema no ocurre bajo la Hiótesis 1.3.2.

Por otra parte, si el espacio de estados es muy grande los algoritmos de anteriores se vuelven inviables numéricamente. Para enfrentar esta dificultad surge la Programación Dinámica Aproximada (PDA), la cual consiste en combinar el algoritmo original (por ejemplo, IV IP) con métodos de aproximación de funciones. En este trabajo estudiaremos el algoritmo de iteración de políticas, y cierta clase de esquemas de aproximación representada por operadores que llamaremos promediadores.

Capítulo 2

Algoritmo de Iteración de Políticas Aproximado

2.1. Introducción

El objetivo de éste capítulo es presentar cotas de error por aproximar la política óptima mediante políticas obtenidas por el algoritmo de IP aplicado en un modelo de control perturbado. El modelo perturbado es una aproximación del modelo original, se obtiene por medio de una clase de operadores que llamaremos promediadores (Definición 2.2.1).

El punto importante es que esta clase de operadores definen modelos de control perturbados en los cuales el algoritmo de iteración de políticas es el algoritmo aproximado en el modelo original. En éste capítulo se introducen los operadores de aproximación, un modelo markoviano perturbado y finalmente se presenta el algoritmo de iteración de políticas aproximado.

2.2. Operadores de aproximación

Para un operador $L : M(\mathbf{X}) \to M(\mathbf{X})$, la imagen Lu denotará la aproximación de u en $M(\mathbf{X})$. En particular, para cada $f \in \mathbb{F}$ y $B \in \mathcal{B}(\mathbf{X})$, $LC_f(\cdot)$ y $LQ_f(B|\cdot)$ son aproximaciones de las funciones $C_f(\cdot)$ y $Q_f(B|\cdot)$, respectivamente.

Definición 2.2.1 *Un operador $L : M(\mathbf{X}) \to M(\mathbf{X})$ es un promediador si satisface las siguientes condiciones:*
(a) $L(I_{\mathbf{X}}) = I_{\mathbf{X}}$,

(b) *L es un operador lineal;*
(c) *L es un operador positivo, esto es, $Lu(\cdot) \geq 0$ para cada $u(\cdot) \geq 0$ en $M(\mathbf{X})$;*
(d) *L satisface la siguiente propiedad de continuidad:*

$$v_n(\cdot) \downarrow 0, v_n \in M(\mathbf{X}) \Rightarrow Lv_n(\cdot) \downarrow 0.$$

Definición 2.2.2 *(a) Un promediador L es débilmente continuo si $Lv(\cdot)$ está en $C(\mathbf{X})$ para cada $v(\cdot)$ en $C(\mathbf{X})$.*

(b) El promediador L es fuertemente continuo si $Lv(\cdot)$ está en $C(\mathbf{X})$ para cada $v(\cdot)$ en $M(\mathbf{X})$.

Algunos de los procedimientos de aproximación de funciones más importantes son representados por promediadores. Entre ellos, podemos mencionar a las aproximaciones constantes por secciones, interpolaciones lineales, ciertos métodos basados en "splines", así como interpolaciones con kérneles, entre otros.

Los promediadores son operadores no-expansivos en $M_b(\mathbf{X})$ con respecto a la norma del supremo, es decir,

$$\|Lu - Lv\|_\infty \leq \|u - v\|_\infty \text{ para todo } u, v \in M_b(\mathbf{X}). \tag{2.1}$$

Para verificar esta propiedad primero observe que si L es un promediador, entonces es monótono, es decir, $Lu \geq Lv$ si $u \geq v$. Luego, observe que

$$\begin{aligned} L(u-v) &\leq \|u-v\|_\infty \\ L(v-u) &\leq \|v-u\|_\infty , \end{aligned}$$

lo cual implica (2.1).

La siguiente proposición muestra las propiedades más importantes de los promediadores. De hecho el nombre de promediador se debe a las propiedades (a) y (b) de la siguiente proposición. La demostración sigue argumentos estándar por lo que será omitida.

Proposición 2.2.3 *Sea $L : M(\mathbf{X}) \to M(\mathbf{X})$ un promediador y definimos*

$$L(D|x) := LI_D(x), \quad x \in \mathbf{X}, \ D \in \mathcal{B}(\mathbf{X}).$$

Entonces:
(a) *$L(\cdot|\cdot)$ es una probabilidad de transición en $\mathbf{X}$, esto es, $L(\cdot|x)$ es una medida de probabilidad en $\mathbf{X}$ para cada $x \in \mathbf{X}$ y $L(D|\cdot)$ es una función medible para cada $D \in \mathcal{B}(\mathbf{X})$.*

(b) *Para toda* $v \in M(\mathbf{X})$ *se cumple*

$$Lv(\cdot) = \int_{\mathbf{X}} v(y)L(dy|\cdot).$$

2.3. Modelo markoviano perturbado

Sea L un promediador y un modelo de control $\mathcal{M} = (\mathbf{X}, \mathbf{A}, \{A(x) : x \in \mathbf{X}\}, Q, C)$ fijos. Consideremos el siguiente *modelo de control perturbado*

$$\widetilde{\mathcal{M}} := (\mathbf{X}, \mathbf{A}, \{\ \widetilde{Q}_f\ ,\ \widetilde{C}_f\ ,\ f \in \mathbb{F}\}) \tag{2.2}$$

donde

$$\widetilde{C}_f(x) := LC_f(x) = \int_{\mathbf{X}} C_f(y)L(dy|x), \tag{2.3}$$

$$\widetilde{Q}_f(B|x) := LQ_f(B|x) = \int_{\mathbf{X}} Q_f(B|y)L(dy|x), \tag{2.4}$$

para $x \in \mathbf{X}$, $B \in \mathcal{B}(\mathbf{X})$, $f \in \mathbb{F}$. Las dos igualdades anteriores se deben a la Proposición 2.2.3.

Observe que de (2.4) se sigue que $\widetilde{Q}_f(\cdot|\cdot),\ f \in \mathbb{F}$, es una probabilidad de transición en $\mathbf{X}$. Entonces, por el teorema de Ionescu-Tulcea ([1, Theorem 2.7.2, p. 109]), para cada estado inicial $x_0 = x \in \mathbf{X}$, y cada política $f \in \mathbb{F}$ existe una medida de probabilidad $\widetilde{P}_x^f$ y una cadena de Markov $\{\widetilde{x}_n\}$ definidas en el espacio $(\Omega, \mathcal{F})$ tal que $\widetilde{Q}_f(\cdot|\cdot)$ es la probabilidad de transición en un paso de $\{\widetilde{x}_n\}$. Denotamos $\widetilde{E}_x^f$ el operador esperanza con respecto a la medida de probabilidad $\widetilde{P}_x^f$.

Ahora se define el índice de funcionamiento en el modelo perturbado. Sea $\alpha \in (0, 1)$ un factor de descuento fijo; se define el *costo descontado* como

$$\widetilde{V}_f(x) := \widetilde{E}_x^f \sum_{k=0}^{\infty} \alpha^k \widetilde{C}_f\ (\widetilde{x}_k),\ x \in \mathbf{X},\ f \in \mathbb{F}$$

La *función de valor óptimo* correspondiente se define como

$$\widetilde{V}_*(x) := \inf_{f \in \mathbb{F}} \widetilde{V}_f(x).$$

Una política estacionaria f^*es óptima si satisface la igualdad

$$\widetilde{V}_*(\cdot) = \widetilde{V}_{f*}(\cdot).$$

Ahora, para cada $f \in \mathbb{F}$, se define el operador $\widetilde{T}_f : M_b(\mathbf{X}) \to M_b(\mathbf{X})$ como

$$\widetilde{T}_f u(x) := \widetilde{C}_f(x) + \alpha \int_{\mathbf{X}} u(y)\widetilde{Q}_f(dy|x).$$

Lema 2.3.1 *(a) Para toda función $u \in M_b(\mathbf{X})$*

$$\widetilde{T}_f u(x) = LT_f u(x), \quad \forall x \in \mathbf{X}.$$

(b) Si la función de costo $C_f(\cdot)$ es acotada, entonces $\widetilde{T}_f$ es un operador de contracción en el espacio $(M_b(\mathbf{X}), ||\cdot||_\infty)$ con módulo α.
(c) $\widetilde{V}_f$ es el punto fijo del operador $\widetilde{T}_f$, esto es,

$$\widetilde{V}_f(x) = \widetilde{C}_f(x) + \alpha \int_{\mathbf{X}} \widetilde{V}_f(y)\widetilde{Q}_f(dy|x) = \widetilde{T}_f\widetilde{V}_f(x). \tag{2.5}$$

Demostración. (a) Por la Proposición 2.2.3 tenemos

$$\begin{aligned}\widetilde{C}_f(x) + \alpha \int_{\mathbf{X}} u(y)\widetilde{Q}_f(dy|x) &= LC_f(x) + \alpha \int_{\mathbf{X}}\int_{\mathbf{X}} u(y)Q_f(dy|z)L(dz|x) \\ &= LT_f u(x);\end{aligned}$$

por lo tanto,

$$\widetilde{T}_f u(x) = LT_f u(x).$$

(b) De la definición de $\widetilde{T}_f$ es posible verificar que es un operador de contracción con módulo α en el espacio de Banach $(M_b(\mathbf{X}), ||\cdot||_\infty)$. Este resultado también se sigue notando del inciso anterior que $\widetilde{T}_f$ es la composición de un operador de contracción (T_f) con módulo α y un operador no expansivo (L).

(c) Por el inciso (b) $\widetilde{T}_f$ tiene un único punto fijo $u_f \in M_b(\mathbf{X})$, es decir,

$$u_f(x) = \widetilde{C}_f(x) + \alpha \int_{\mathbf{X}} u_f(y)\widetilde{Q}_f(dy|x),$$

por otra parte, notemos que por propiedades de la esperanza condicional y la porpiedad de Markov se cumple que

$$\begin{aligned}
\widetilde{V}_f(x) &= \widetilde{E}_x^f\left[\widetilde{C}_f(x_0) + \sum_{n=1}^{\infty}\alpha^n\widetilde{C}_f(\widetilde{x}_n)\right] \\
&= \widetilde{C}_f(x) + \widetilde{E}_x^f\left\{\widetilde{E}_x^f\left[\alpha\sum_{n=1}^{\infty}\alpha^{n-1}\widetilde{C}_f(\widetilde{x}_n)|h_1\right]\right\} \\
&= \widetilde{C}_f(x) + \widetilde{E}_x^f\left\{\widetilde{E}_{x_1}^f\left[\alpha\sum_{n=0}^{\infty}\alpha^{n}\widetilde{C}_f(\widetilde{x}_n)\right]\right\} \\
&= \widetilde{C}_f(x) + \widetilde{E}_x^f\left\{\alpha\ \widetilde{V}_f(\widetilde{x}_1)\right\} \\
&= \widetilde{C}_f(x) + \alpha\int_{\mathbf{X}}\widetilde{V}_f(y)\widetilde{Q}_f(dy|x).
\end{aligned}$$

Entonces, $u_f = \widetilde{V}_f$ es el punto fijo del operador $\widetilde{T}_f$, esto es,

$$\widetilde{V}_f(x) = \widetilde{C}_f(x) + \alpha\int_{\mathbf{X}}\widetilde{V}_f(y)\widetilde{Q}_f(dy|x) = \widetilde{T}_f\widetilde{V}_f(x). \qquad \blacksquare$$

Observemos que se cumple

$$\begin{aligned}
||\widetilde{T}_f^n u - \ \widetilde{V}_f||_\infty &= ||\widetilde{T}_f\widetilde{T}_f^{n-1}u - \ \widetilde{T}_f\widetilde{V}_f||_\infty \\
&\leq \alpha||\widetilde{T}_f^{n-1}u - \widetilde{V}_f||_\infty \leq \cdots \leq \alpha^n||u - \widetilde{V}_f||_\infty,
\end{aligned}$$

para toda $u \in M_b(\mathbf{X})$, $n \in \mathbb{N}$.

Ahora definimos el operador de programación dinámica para el modelo perturbado como

$$\widetilde{T}u(x) := \inf_{f\in\mathbf{F}}\left\{\widetilde{C}_f(x) + \alpha\int_{\mathbf{X}}u(y)\widetilde{Q}_f(dy|x)\right\}\ x \in \mathbf{X},\ u \in C_b(\mathbf{X}). \qquad (2.6)$$

El operador $\widetilde{T}$ satisface las siguientes propiedades.

Lema 2.3.2 *Supongamos que se cumple la Hipótesis 1.3.1 y que L es un promediador débilmente continuo. Entonces:*
(a) para toda función $u \in C_b(\mathbf{X})$

$$\widetilde{T}u(x) = LTu(x), \quad \forall x \in \mathbf{X}; \qquad (2.7)$$

(b) $\widetilde{T}$ es es un operador de contracción en el espacio $(C_b(\mathbf{X}), ||\cdot||_\infty)$ con módulo α;
(c) para cada $u \in C_b(\mathbf{X})$ existe un selector medible $f \in \mathbb{F}$ tal que

$$\widetilde{T}u(x) = \widetilde{T}_f u(x) = \widetilde{C}_f(x) + \alpha \int_{\mathbf{X}} u(y)\widetilde{Q}_f(dy|x), \tag{2.8}$$

para cada $x \in \mathbf{X}$.

Demostración**.** (a) Observe que para $u \in C_b(\mathbf{X})$ y $f \in \mathbb{F}$,

$$Tu(x) \leq C_f(x) + \alpha \int_{\mathbf{X}} u(y)Q_f(dy|x),\ x \in \mathbf{X}.$$

Ahora, como L es un operador monótono tenemos que para cada $x \in \mathbf{X}$,

$$\begin{aligned} LTu(x) &\leq L\left\{C_f(x) + \alpha \int_{\mathbf{X}} u(y)Q_f(dy|x)\right\} \\ &= \widetilde{C}_f(x) + \alpha \int_{\mathbf{X}} u(y)\widetilde{Q}_f(dy|x). \end{aligned}$$

Tomando ínfimo sobre $f \in \mathbb{F}$ en la desigualdad anterior

$$\begin{aligned} LTu(x) &\leq \inf_{f\in\mathbf{F}} \left\{\widetilde{C}_f(x) + \alpha \int_{\mathbf{X}} u(y)\widetilde{Q}_f(dy|x)\right\} \\ &= \widetilde{T}u(x). \end{aligned}$$

Para obtener la desigualdad contraria, note que por la Observación 1.3.3(c), para cada u en $C_b(\mathbf{X})$ existe f_u tal que

$$Tu(x) = C_{f_u}(x) + \alpha \int_{\mathbf{X}} u(y)Q_{f_u}(dy|x)\ \forall x \in \mathbf{X}. \tag{2.9}$$

Por otra parte, notemos que para $x \in \mathbf{X}$,

$$\begin{aligned} \widetilde{T}u(x) &\leq \widetilde{C}_{f_u}(x) + \alpha \int_{\mathbf{X}} u(y)\widetilde{Q}_{f_u}(dy|x) \\ &= L\left\{C_{f_u}(x) + \alpha \int_{\mathbf{X}} u(y)Q_{f_u}(dy|x)\right\} \\ &= LTu(x), \end{aligned}$$

donde la última igualdad se sigue de (2.9). Por lo tanto,

$$\widetilde{T}u(x) = LTu(x).$$

(b) Para cada $u \in C_b(\mathbf{X})$, por la Observación 1.3.3(b) se cumple $Tu \in C_b(\mathbf{X})$. Además, como L es un promediador débilmente continuo, $LTu \in C_b(\mathbf{X})$. Entonces, por la parte (a), $\widetilde{T}u \in C_b(\mathbf{X})$, es decir, $\widetilde{T}$ mapea $C_b(\mathbf{X})$ en sí mismo. Ahora, usando que el operador L es no expansivo, se sigue que

$$||\widetilde{T}u - \widetilde{T}v||_\infty = ||LTu - LTv||_\infty \leq ||Tu - Tv||_\infty \leq \alpha||u - v||_\infty,$$

lo cual demuestra la parte (b).
(c) Sea $u \in C_b(\mathbf{X})$. Por (2.9) existe f_u tal que

$$Tu(x) = C_{f_u}(x) + \alpha \int_{\mathbf{X}} u(y) Q_{f_u}(dy|x).$$

Luego por la propiedad de monotonía y linealidad de L se tiene

$$\widetilde{T}u(x) = \widetilde{C}_{f_u}(x) + \alpha \int_{\mathbf{X}} u(y) \widetilde{Q}_{f_u}(dy|x) \ \forall x \in \mathbf{X},$$

lo cual demuestra el resultado deseado.■

Proposición 2.3.3 *Supongamos que se cumple la Hipótesis 1.3.1 y que L es un promediador débilmente continuo. Entonces la función de valor óptimo $\widetilde{V}_*$ es el único punto fijo del operador $\widetilde{T}$ en $C_b(\mathbf{X})$, y existe $f^* \in \mathbb{F}$*

$$\widetilde{V}_*(x) = \widetilde{T}\widetilde{V}_*(x) = \widetilde{T}_{f^*}\widetilde{V}_*(x) \ \forall x \in \mathbf{X}.$$

Además, una política estacionaria $f \in \mathbb{F}$ es óptima si y sólo $\widetilde{T}\widetilde{V}_*(x) = \widetilde{T}_f\widetilde{V}_*(x)$. Por lo tanto, la política f^* es óptima.

Demostración. Por el Lema 2.3.2 existe una única función $u^* \in C_b(\mathbf{X})$ y una política estacionaria $f^* \in \mathbb{F}$ tal que

$$u^*(x) = \widetilde{T}u^*(x) = \widetilde{C}_{f^*}(x) + \alpha \int_{\mathbf{X}} u^*(y) \widetilde{Q}_{f^*}(dy|x).$$

Entonces, u^* es un punto fijo de $\widetilde{T}_{f^*}$, y por (2.5), $u^*(x) = \widetilde{V}_{f^*}(x)$. Por lo tanto, de la definición de $\widetilde{V}_*$, tenemos

$$u^*(x) = \widetilde{V}_{f^*}(x) \geq \widetilde{V}_*(x).$$

Para obtener la desigualdad contraria, notemos que

$$u^*(x) = \widetilde{T}u^*(x) \leq \widetilde{C}_f(x) + \alpha \int_{\mathbf{X}} u^*(y) \widetilde{Q}_f(dy|x), \quad f \in \mathbb{F}.$$

Luego, para cada $x \in \mathbf{X}$ y $h_m \in \mathbb{H}_m$, $m \in \mathbb{N}_0$,se cumple que

$$0 \leq \widetilde{E}_x^f \left\{ \alpha^m \widetilde{C}_f(\widetilde{x}_m) + \alpha^{m+1} u^*(\widetilde{x}_{m+1}) - \alpha^m u^*(\widetilde{x}_m) | h_m \right\}.$$

Tomando valor esperado y sumando se obtiene

$$\begin{aligned} 0 &\leq \widetilde{E}_x^f \left\{ \sum_{m=0}^{n} \alpha^m \widetilde{C}_f(\widetilde{x}_m) + \sum_{m=0}^{n} \alpha^{m+1} u^*(\widetilde{x}_{m+1}) - \sum_{m=0}^{n} \alpha^m u^*(\widetilde{x}_m) \right\} \\ &\leq \widetilde{E}_x^f \left\{ \sum_{m=0}^{n} \alpha^m \widetilde{C}_f(\widetilde{x}_m) + \alpha^{n+1} u^*(\widetilde{x}_{n+1}) - u^*(x_0) \right\}. \end{aligned}$$

Tomando límite cuando $n \to \infty$ obtenemos

$$u^*(x) \leq \widetilde{V}_f(x),$$

lo cual implica

$$u^*(x) \leq \widetilde{V}_*(x).$$

Por lo tanto,

$$\widetilde{V}_*(x) = \widetilde{T}\widetilde{V}_*(x).$$

Finalmente, sea $f^* \in \mathbb{F}$ una política estacionaria óptima, es decir, $\widetilde{V}_*(x) = \widetilde{V}_{f^*}(x)$. Por (2.5), $\widetilde{V}_{f^*}$ es punto fijo de $\widetilde{T}_{f^*}$ y $\widetilde{V}_*$ es un punto fijo de $\widetilde{T}$, de donde se sigue

$$\widetilde{T}\widetilde{V}_*(x) = \widetilde{T}_{f^*}\widetilde{V}_*(x).$$

Ahora suponga que para una política $f \in \mathbb{F}$ se cumple $\widetilde{T}\widetilde{V}_*(\cdot) = \widetilde{T}_f\widetilde{V}_*(\cdot)$. Entonces, $\widetilde{V}_*$ es un punto fijo de $\widetilde{T}_f$, lo cual implica que

$$\widetilde{V}_*(\cdot) = \widetilde{V}_{f^*}(\cdot);$$

por lo tanto, f es óptima. ■

2.4. Algoritmo de iteración de políticas aproximado

En esta sección se introduce el algoritmo de iteración de políticas para el modelo perturbado. El objetivo es calcular cotas de error cuando las políticas obtenidas bajo este algoritmo son evaluadas en el modelo original.

Algoritmo de iteración de políticas aproximado (IPA)

(i) Inicio: sean $k = 0$ y $g_0 \in \mathbb{F}$ arbitraria;

(ii) etapa de evaluación: dado $g_k \in \mathbb{F}$ calcule $u_k(\cdot) := \widetilde{V}_{g_k}(\cdot)$;

(iii) etapa de mejoramiento: calcule la política $g_{k+1} \in \mathbb{F}$ tal que $\widetilde{T}u_k(\cdot) = \widetilde{T}_{g_{k+1}}u_k(\cdot)$ y regrese a la etapa (ii).

Proposición 2.4.1 *Supongamos que se cumple la Hipótesis 1.3.1 y que L es un promediador fuertemente continuo. Entonces la sucesión $\{u_k\}$ converge decrecientemente y en la norma del supremo a la función de valor óptimo $\widetilde{V}_*$. Además,*

$$\|\widetilde{V}_* - u_k\|_\infty \leq \frac{2\alpha}{1-\alpha}\|u_k - u_{k-1}\|_\infty \quad \forall k \in \mathbb{N}.$$

Demostración. Primero notemos que $\widetilde{V}_f$, $f \in \mathbb{F}$, es una función continua y acotada ya que $\widetilde{V}_f = LT_f\widetilde{V}_f$ y L es un promediador fuertemente continuo. Así, para cada $k \in \mathbb{N}_0$, existe una política $g_{k+1} \in \mathbb{F}$ tal que $\widetilde{T}u_k(\cdot) = \widetilde{T}_{g_{k+1}}u_k(\cdot)$; por lo tanto, la etapa de mejoramiento en el algoritmo IPA está bien definida.

Para demostrar la convergencia de la sucesión $\{u_k\}$ a $\widetilde{V}_*$, tomemos $g_0 \in \mathbb{F}$ arbitraria y observemos que

$$\begin{aligned} u_0(x) &:= \widetilde{V}_{g_0}(x) \\ &= \widetilde{C}_{g_0}(x) + \alpha \int_{\mathbf{X}} u_0(y)\widetilde{Q}_{g_0}(dy|x) \\ &\geq \inf_{f\in\mathbf{F}} \{\widetilde{C}_f(x) + \alpha \int_{\mathbf{X}} u_0(y)\widetilde{Q}_f(dy|x)\} \\ &= \widetilde{T}u_0(x). \end{aligned}$$

Por otra parte, existe $g_1 \in \mathbb{F}$ tal que

$$\widetilde{T}u_0(x) = \widetilde{C}_{g_1}(x) + \alpha \int_{\mathbf{X}} u_0(y)\widetilde{Q}_{g_1}(dy|x),$$

lo cual implica

$$u_0(x) \geq \widetilde{C}_{g_1}(x) + \alpha \int_{\mathbf{X}} u_0(y)\widetilde{Q}_{g_1}(dy|x).$$

Iterando la desigualdad anterior obtenemos

$$u_0(x) \geq \widetilde{E}_x^{g_1} \sum_{k=0}^{n-1} \alpha^k \widetilde{C}_{g_1}(\widetilde{x}_k) + \alpha^n \widetilde{E}_x^{g_1} u_0(\widetilde{x}_n).$$

Luego, tomando límite cuando $n \to \infty$, resulta

$$u_0(x) \geq \widetilde{V}_{g_1}(x) = u_1(x).$$

Repitiendo la argumentación anterior se obtiene

$$u_k(x) \geq \widetilde{T} u_k(x) \geq u_{k+1}(x), \ \forall x \in \mathbf{X}, \ k \in \mathbb{N}_0.$$

Entonces la función $u = \lim u_k$ está bien definida. Además, tomando límite cuando $k \to \infty$, por [17, Lemma 3.3], tenemos

$$\lim_{k\to\infty} T u_k(x) = T(\lim_{k\to\infty} u_k(x)) = Tu(x).$$

Luego como $\widetilde{T} = LT$, por el teorema de la convergencia dominada y de la igualdad anterior se cumplen las siguientes igualdades

$$\lim_{k\to\infty} \widetilde{T} u_k(x) = \lim_{k\to\infty} LT u_k(x) = L \lim_{k\to\infty} T u_k(x) = \widetilde{T} u(x).$$

Entonces,

$$u(x) \geq \widetilde{T} u(x) \geq u(x), \ \forall x \in \mathbf{X},$$

lo cual prueba que u es punto fijo del operador $\widetilde{T}$, es decir,

$$u(\cdot) = \widetilde{T} u(\cdot).$$

Además, puesto que L es un promediador fuertemente continuo $u \in C_b(\mathbf{X})$ y por la Proposición 2.3.3, concluimos que

$$u(\cdot) = \widetilde{V}_*(\cdot).$$

Ahora para demostrar la segunda parte de la proposición notemos que

$$\begin{aligned}
\|\widetilde{V}_* - u_k\|_\infty &\leq \|\widetilde{V}_* - \widetilde{T}_{g_k} u_{k-1}\|_\infty + \|\widetilde{T}_{g_k} u_{k-1} - u_k\|_\infty \\
&= \|\widetilde{T}\widetilde{V}_* - \widetilde{T} u_{k-1}\|_\infty + \|\widetilde{T}_{g_k} u_{k-1} - \widetilde{T}_{g_k} u_k\|_\infty \\
&\leq \alpha \|\widetilde{V}_* - u_{k-1}\|_\infty + \alpha \|u_{k-1} - u_k\|_\infty \\
&\leq \alpha \|\widetilde{V}_* - u_k\|_\infty + \alpha \|u_k - u_{k-1}\|_\infty + \alpha \|u_{k-1} - u_k\|_\infty,
\end{aligned}$$

lo cual implica que

$$\|\widetilde{V}_* - u_k\|_\infty \leq \frac{2\alpha}{1-\alpha}\|u_k - u_{k-1}\|_\infty. \blacksquare$$

Cotas para los errores de aproximación. Para obtener cotas para los errores de aproximación del algoritmo IPA, introducimos las siguientes constantes. Sea $\mathbb{F}_0$ la subclase de las políticas estacionarias que contiene las políticas óptimas estacionarias tanto para el modelo original como para el modelo aproximado y las políticas generadas por el algoritmo IPA. Entonces definimos

$$\delta_C := \sup_{f\in\mathbb{F}_0} \|\widetilde{C}_f - C_f\|_\infty \ , \tag{2.10}$$

$$\delta_Q := \sup_{x\in\mathbf{X}, f\in\mathbb{F}_0} \|\widetilde{Q}_f(\cdot|x) - Q_f(\cdot|x)\|_{TV}, \tag{2.11}$$

donde $\|\cdot\|_{TV}$ representa la norma de la variación total para medidas con signo finitas, es decir,

$$\|\lambda\|_{TV} := \sup\left\{\left|\int v(y)\lambda(dy)\right| : v \in M_b(\mathbf{X}), \|v\|_\infty \leq 1\right\},$$

donde λ es una medida con signo en $\mathbf{X}$. Es fácil ver que

$$\left|\int v(y)\lambda(dy)\right| \leq \|\lambda\|_{TV}\|v\|_\infty \ \ \forall v \in M_b(\mathbf{X}).$$

Observemos que δ_C y δ_Q miden la precisión con la que el promediador L aproxima a C y Q. El siguiente teorema nos proporciona cotas de error en términos de estas cantidades.

Teorema 2.4.2 *Supongamos que se cumple la Hipótesis 1.3.1 y que L es un promediador fuertemente continuo. Sea $u_k \in C_b(\mathbf{X})$ y $g_k \in \mathbb{F}$, $k \in \mathbb{N}$, las funciones y políticas definidas por el algoritmo IPA. Entonces:*

(a) $\|\widetilde{V}_* - V_*\|_\infty \leq \dfrac{1}{1-\alpha}\delta_C + \dfrac{\alpha M}{(1-\alpha)^2}\delta_Q$;

(b) *Además,*

$$\|V_* - V_{g_k}\|_\infty \leq \frac{2}{1-\alpha}\delta_C + \frac{2\alpha M}{(1-\alpha)^2}\delta_Q + \frac{2\alpha}{1-\alpha}\|u_k - u_{k-1}\|_\infty,$$

para todo $k \in \mathbb{N}$.

Demostración. (a) Notemos que de (2.5), para cada política $f \in \mathbb{F}$ se cumple

$$\widetilde{V}_f(x) = \widetilde{C}_f(x) + \alpha \int_{\mathbf{X}} \widetilde{V}_f(y)\widetilde{Q}_f(dy|x) \ \forall x \in \mathbf{X}$$

y

$$V_f(x) = C_f(x) + \alpha \int_{\mathbf{X}} V_f(y)Q_f(dy|x).$$

Entonces,

$$\begin{aligned}
\left|\widetilde{V}_f(x) - V_f(x)\right| &\leq \left|\widetilde{C}_f(x) - C_f(x)\right| + \alpha \left|\int_{\mathbf{X}} \widetilde{V}_f(y)\widetilde{Q}_f(dy|x) - \int_{\mathbf{X}} V_f(y)Q_f(dy|x)\right| \\
&\leq \|\widetilde{C}_f - C_f\|_\infty + \alpha \left|\int_{\mathbf{X}} \widetilde{V}_f(y)\widetilde{Q}_f(dy|x) - \int_{\mathbf{X}} V_f(y)\widetilde{Q}_f(dy|x)\right| \\
&+ \alpha \left|\int_{\mathbf{X}} V_f(y)\widetilde{Q}_f(dy|x) - \int_{\mathbf{X}} V_f(y)Q_f(dy|x)\right| \\
&\leq \|\widetilde{C}_f - C_f\|_\infty + \alpha\|\widetilde{V}_f - V_f\|_\infty \\
&+ \alpha\|V_f\|_\infty \sup_{x\in\mathbf{X}} \|\widetilde{Q}_f(\cdot|x) - Q_f(\cdot|x)\|_{TV}.
\end{aligned}$$

De aquí se sigue que

$$\|\widetilde{V}_f - V_f\|_\infty \leq \|\widetilde{C}_f - C_f\|_\infty + \alpha\|\widetilde{V}_f - V_f\|_\infty + \alpha\|V_f\|_\infty \sup_{x\in\mathbf{X}} \|\widetilde{Q}_f(\cdot|x) - Q_f(\cdot|x)\|_{TV}.$$

Ahora, por la Hipótesis 1.3.1(a), obtenemos la desigualdad

$$\|\widetilde{V}_f - V_f\|_\infty \leq \frac{1}{1-\alpha} \sup_{f\in\mathbb{F}_0} \|\widetilde{C}_f - C_f\|_\infty + \frac{\alpha}{(1-\alpha)^2} M \sup_{x\in\mathbf{X}, f\in\mathbb{F}_0} \|\widetilde{Q}_f(\cdot|x) - Q_f(\cdot|x)\|_{TV}, \tag{2.12}$$

la cual es equivalente a

$$-\kappa + V_f \leq \widetilde{V}_f \leq V_f + \kappa,$$

donde

$$\kappa := \frac{1}{1-\alpha}\delta_C + \frac{\alpha}{(1-\alpha)^2} M\delta_Q.$$

Tomando ínfimo sobre la clase de las políticas estacionarias en esta desigualdad llegamos a que

$$-\kappa + V_* \leq \widetilde{V}_* \leq V_* + \kappa.$$

Por lo tanto

$$\|\widetilde{V}_* - V_*\|_\infty \leq \frac{1}{1-\alpha}\delta_C + \frac{\alpha M}{(1-\alpha)^2}\delta_Q. \tag{2.13}$$

De esta manera tenemos que se cumple (a).

(b) Recordemos que $u_k(\cdot) = \widetilde{V}_{g_k}(\cdot)$. Entonces, aplicando (2.12) con g_k en lugar de f,

$$\|\widetilde{V}_{g_k} - V_{g_k}\|_\infty \leq \frac{1}{1-\alpha}\delta_C + \frac{\alpha}{(1-\alpha)^2}M\delta_Q \quad \forall k \in \mathbf{N}_0.$$

Por otra parte,

$$\|V_* - V_{g_k}\|_\infty \leq \|V_* - \widetilde{V}_*\|_\infty + \|\widetilde{V}_* - u_k\|_\infty + \|\tilde{V}_{g_k} - V_{g_k}\|_\infty.$$

Finalmente, (2.13) y la Proposición 2.4.1 implican que

$$\|V_* - V_{g_k}\|_\infty \leq \frac{2}{1-\alpha}\delta_C + \frac{2\alpha}{(1-\alpha)^2}M\delta_Q + \frac{2\alpha}{1-\alpha}\|u_k - u_{k-1}\|_\infty,$$

lo que demuestra la parte (b). ■

Capítulo 3

Estimación y Aproximación en Modelos Markovianos Descontados

3.1. Introducción

En este capítulo estudiaremos sistemas estocásticos controlados cuya evolución del sistema está dada por

$$x_{n+1} = H(x_n, a_n, w_n), \quad n \in \mathbb{N}_0,$$

donde w_n es un ruido aleatorio con densidad desconocida. Supondremos que el ruido aleatorio es observable; esto permitirá obtener un estimador de la densidad desconocida y definir un modelo de control estimado $\mathcal{M}_t$, el cual será perturbado para obtener un modelo aproximado $\widetilde{\mathcal{M}}_t$.

En este caso la política óptima del modelo original será aproximada mediante políticas obtenidas por el algoritmo de IP en el modelo $\widetilde{\mathcal{M}}_t$. Es claro que el buen desempeño de estas políticas en el modelo original dependerá por una parte de la precisión del modelo $\mathcal{M}_t$ para estimar al modelo desconocido $\mathcal{M}$, y por otro lado de la exactitud con la que el modelo $\widetilde{\mathcal{M}}_t$ aproxima al modelo $\mathcal{M}_t$. Se presentan cotas de error por estimación, aproximación y por detener el algoritmo de IP correspondiente.

Finalmente para ilustrar los resultados del capítulo se presenta un ejemplo de control de inventarios.

3.2. Proceso de estimación y control

Consideremos un proceso de control de Markov que evoluciona de acuerdo a la ecuación en diferencias dada por

$$x_{n+1} = H(x_n, a_n, w_n), \quad n \in \mathbb{N}_0, \tag{3.1}$$

donde $H : \mathbb{K} \times \mathbb{R}^m \to \mathbf{X}$ es una función medible, las variables aleatorias $w_n, n \in \mathbb{N}_0$, son independientes e idénticamente distribuidas (i.i.d.) con valo res en el espacio euclidiano $\mathbb{R}^m$ y función de densidad $\rho(\cdot)$, la cual suponemos es desconocida para el controlador. Suponemos además que la función de costo por etapa está dada por

$$C(x, a) := \int_{\mathbb{R}^m} \widehat{c}(x, a, w)\rho(w)dw, \quad (x, a) \in \mathbb{K}, \tag{3.2}$$

donde $\widehat{c} : \mathbb{K} \times \mathbb{R}^m \to \mathbb{R}$ es una función medible.

La ley de transición para el sistema (3.1) toma la siguiente forma

$$Q(B|x, a) = \int_{\mathbb{R}^m} \mathbf{I}_B(H(x, a, w)\rho(w)dw, \tag{3.3}$$

para todo $B \in \mathcal{B}(\mathbf{X}), (x, a) \in \mathbb{K}$, y satisface la igualdad

$$\int_{\mathbf{X}} v(y)Q(dy|x, a) = \int_{\mathbb{R}^m} v((H(x, a, w))\rho(w)dw, \tag{3.4}$$

para cada función medible v en $\mathbf{X}$ siempre que las integrales estén bien definidas.

Hipótesis 3.2.1 ***(a)*** *$H(\cdot, \cdot, w)$ es continua en $\mathbb{K}$ para cada $w \in \mathbb{R}^m$;*
(b) *$\widehat{c}(\cdot, \cdot, w)$ es una función acotada y continua en $\mathbb{K}$ para cada $w \in \mathbb{R}^m$;*
(c) *la multifunción $x \to A(x)$ es continua con valores en la familia de los conjuntos compactos de $\mathbf{A}$.*

Observación 3.2.2 *Bajo la Hipótesis 3.2.1, la función de costo $C(\cdot, \cdot)$ en (3.2) es una función continua y acotada, y la ley de transición $Q(\cdot|\cdot, \cdot)$ en (3.3) es débilmente continua. Por lo tanto, la Hipótesis 3.2.1 implica la Hipótesis 1.3.1 para cualquier densidad $\rho(\cdot)$.*

Supondremos que las variables aleatorias $w_n, n \in \mathbb{N}$, son observables, y denotemos por $\mathbf{w}_t := (w_0, w_1, \ldots, w_{t-1})$ a una muestra observada por el

controlador. Entonces, basado en $\mathbf{w}_t$, el controlador obtiene una densidad estimada

$$\rho_t(\cdot) = \rho_t(\cdot|\mathbf{w}_t)$$

de la densidad desconocida $\rho(\cdot)$. Definamos

$$\eta_t := \int_{\mathbb{R}^m} |\rho(w) - \rho_t(w)|\, dw,\ t \in \mathbb{N}. \tag{3.5}$$

Modelo estimado. Consideremos el modelo estimado

$$\mathcal{M}_t := (\mathbf{X}, \mathbf{A}, \{A(x) : x \in \mathbf{X}\}, Q^{(t)}, C^{(t)}),$$

donde

$$Q^{(t)}(B|x,a) := \int_{\mathbb{R}^m} \mathbf{I}_B(H(x,a,w)\rho_t(w)dw, \tag{3.6}$$

$$C^{(t)}(x,a) := \int_{\mathbb{R}^m} \widehat{c}(x,a,w)\rho_t(w)dw, \tag{3.7}$$

para $(x,a) \in \mathbb{K}$ y $B \in \mathcal{B}(\mathbf{X})$.

Sea $V_f^{(t)}$ el *costo descontado* bajo la política $f \in \mathbb{F}$ para el modelo $\mathcal{M}_t$, esto es,

$$V_f^{(t)}(x) := E_{x,\pi}^{(t)} \sum_{k=0}^{\infty} \alpha^k C_f^{(t)}(x_k^{(t)}),\ x \in \mathbf{X}$$

donde $\{x_k^{(t)}\}$ denota la cadena de Markov inducida por la política $f \in \mathbb{F}$ en el modelo $\mathcal{M}_t$.

Además, la *función de valor óptimo* en el modelo $\mathcal{M}_t$

$$V_*^{(t)}(x) := \inf_{f \in \mathbb{F}} V_f^{(t)}(x).$$

La siguiente proposición provee una cota de error por estimar la función de valor óptimo V_* en el modelo original $\mathcal{M}$ mediante $V_*^{(t)}$.

Proposición 3.2.3 *Bajo la Hipótesis 3.2.1, para cada $t \geq 0$, se cumple*

$$||V_* - V_*^{(t)}||_\infty \leq \left\{ \frac{M}{1-\alpha} + \frac{\alpha M}{(1-\alpha)^2} \right\} \eta_t$$

Demostración. Denotemos por T_t el operador de programación dinámica para el modelo $\mathcal{M}_t$, es decir, para cada $u \in C_b(\mathbf{X})$

$$\begin{aligned} T_t u(x) &:= \min_{a \in A(x)} \left\{ C^{(t)}(x,a) + \alpha \int_{\mathbb{R}^m} u(y) Q^{(t)}(dy|x,a) \right\} \\ &= \min_{a \in A(x)} \left\{ C^{(t)}(x,a) + \alpha \int_{\mathbb{R}^m} u(H(x,a,w)) \rho_t(w) dw \right\}. \end{aligned}$$

Además, de manera análoga a la Proposición 1.3.4, la función $V_*^{(t)}$ es el único punto fijo de T_t. Entonces, la propiedad de contracción de los operadores T and T_t implican que

$$\begin{aligned} |V_*(\cdot) - V_*^{(t)}(\cdot)| &= |TV_*(\cdot) - T_t V_*(\cdot) + T_t V_*(\cdot) - T_t V_*^{(t)}(\cdot)| \\ &\leq |TV_*(\cdot) - T_t V_*(\cdot)| + \alpha |V_*(\cdot) - V_*^{(t)}(\cdot)|, \quad \forall t \in \mathbb{N}, \end{aligned}$$

lo cual, a su vez, implica que

$$|V_*(\cdot) - V_*^{(t)}(\cdot)| \leq \frac{1}{1-\alpha} |TV_*(\cdot) - T_t V_*(\cdot)|. \tag{3.8}$$

Por otra parte,

$$\begin{aligned} &|TV_*(x) - T_t V_*(x)| \\ &\leq \sup_{a \in A(x)} \left\{ \left| C(x,a) - C^{(t)}(x,a) + \alpha \int_{\mathbb{R}^m} V_*(H(x,a,w)) \rho(w) dw \right. \right. \\ &\qquad \left. \left. - \alpha \int_{\mathbb{R}^m} V_*(H(x,a,w)) \rho_t(w) dw \right| \right\} \\ &\leq \sup_{a \in A(x)} \left\{ \int_{\mathbb{R}^m} \hat{c}(x,a,w) |\rho(w) - \rho_t(w)| dw \right. \\ &\qquad \left. + \alpha \int_{\mathbb{R}^m} V_*(H(x,a,w)) |\rho(w) - \rho_t(w)| dw \right\} \\ &\leq M \int_{\mathbb{R}^m} |\rho(w) - \rho_t(w)| dw + \frac{\alpha M}{1-\alpha} \int_{\mathbb{R}^m} |\rho(w) - \rho_t(w)| dw, \quad t \in \mathbb{N}. \end{aligned} \tag{3.9}$$

Finalmente, el resultado se sigue combinando (3.8) y (3.9). ■

Observación 3.2.4 ***(a)*** *Una propiedad deseable de los estimadores es la de consistencia. El método de estimación por kérneles proporciona estimadores con esta propiedad bajo condiciones muy generales.*

Un kérnel $K : \mathbb{R}^m \to \mathbb{R}_+$ es una función medible tal que $\int_{\mathbb{R}^m} K(\mathbf{s})d\mathbf{s} = 1$. El estimador basado en el kérnel K es

$$\rho_t(s) := \frac{1}{tb_t^m} \sum_{i=0}^{t-1} K\left(\frac{s - w_i}{b_t}\right), \quad s \in \mathbb{R}^m,$$

donde $\mathbf{w}_t = (w_0, \ldots, w_{t-1})$ es una muestra aleatoria de la densidad desconocida ρ y b_t, $t \in \mathbb{N}$, es una sucesión de números reales positivos.

(b) *Las siguientes afirmaciones son equivalentes ([10, Capítulo 3] y [9, Capítulo 9]).*

(i) $b_t \to 0$ y $tb_t^m \to \infty$ cuando $t \to \infty$;

(ii) El estimador $\rho_t, t \in \mathbb{N}$, es fuertemente consistente, esto es,

$$\eta_t = \int_{\mathbb{R}^m} |\rho_t(s) - \rho(s)|\, ds \to 0 \quad a.s.;$$

(iii) Para cada $\varepsilon > 0$ existe una constante $r > 0$ y $t_0 \in \mathbb{N}$ tal que

$$P(\eta_t \geq \varepsilon) \leq \exp(-rn) \quad \forall t \geq t_0.$$

Adicionalmente, cada una de las propiedades anteriores implica que

$$\bar{\eta}_t := E \int_{\mathbb{R}^m} |\rho_t(s) - \rho(s)| ds \to 0 \tag{3.10}$$

(c) *Por otra parte, para una clase amplia de densidades y cierta clase de kérneles, la tasa de convergencia en (3.10) es del orden de $O(t^{-2/5})$ cuando $t \to \infty$ [10, Teorema 9.5]. Esto es, existe una constante positiva d y $t_1 \in \mathbb{N}$ tal que $\bar{\eta}_t \leq dt^{-2/5}$ para todo $t \geq t_1$.*

Para el caso paramétrico, esto es, cuando la densidad desconocida ρ pertenece a una familia paramétrica $\{\rho_\theta : \theta \in \Theta\}$, el método de máxima verosimilitud nos provee de estimadores consistentes y asintóticamente normales bajo algunas condiciones de regularidad (vea, [37, Sección 4.2, p. 145]). En algunos casos específicos se puede ver directamente que el estimador propuesto tiene buenas propiedades. Para ilustrar esto veamos el siguiente ejemplo.

Ejemplo 3.2.5 *Supongamos que la densidad desconocida pertenece a la siguiente familia paramétrica*

$$\rho_\theta(s) = \frac{\varphi(s)}{\int_\theta^\infty \varphi(s)ds}\mathbf{I}_{[\theta,\infty)}(s), \quad \theta \in \Theta,$$

donde $\varphi : \mathbb{R} \to \mathbb{R}_+$ *es una función medible tal que* $\int_{-\infty}^{\infty} \varphi(s)ds$ *es finita y* Θ *es un subconjunto de* $\mathbb{R}_+$*. Cálculos directos muestran que*

$$\theta_t := \text{mín}\{w_0, w_1, ..., w_{t-1}\},$$

es el estimador de máxima verosimilitud de θ*. Además*

$$\theta_t \to \theta \ P_\theta - a.s. para \ cada \ \theta \in \Theta.$$

Para demostrar esta afirmación denote por F_θ *y* F_{θ_t} *las funciones de distribución de* θ *y* θ_t *respectivamente. Entonces, para cualquier* $\varepsilon > 0$*, obtenemos*

$$\begin{aligned} P_\theta[|\theta_t - \theta| > \varepsilon] &\leq P_\theta[\theta_t > \theta + \varepsilon] \\ &= 1 - F_{\theta_t}(\theta + \varepsilon) \\ &= (1 - F_\theta(\theta + \varepsilon))^t. \end{aligned}$$

Entonces,

$$\sum_{t=1}^{\infty} P_\theta[|\theta_t - \theta| > \varepsilon] \leq \sum_{t=1}^{\infty}(1 - F_\theta(\theta + \varepsilon))^t < \infty,$$

lo cual combinado con el lema de Borel-Cantelli implica

$$\theta_t \to \theta \ P_\theta - a.s. \ \forall \theta \in \Theta.$$

Ahora notemos que $\rho_{\theta_t}(s) = 0$ *para* $s \leq \theta_t$ *lo que implica*

$$\left|\rho_{\theta_t}(s) - \rho_\theta(s)\right| = \begin{cases} \rho_\theta(s) & si \quad \theta < s \leq \theta_t \\ \\ \rho_{\theta_t}(s) - \rho_\theta(s) & si \quad s > \theta_t \end{cases}.$$

De aquí y notando que $\int_{\theta_t}^{\infty} \rho_{\theta_t}(s)ds = 1$ *se sigue*

$$\begin{aligned}\eta_t &= \int_{\theta}^{\infty} \left|\rho_{\theta_t}(s) - \rho_{\theta}(s)\right| ds \\ &= \int_{\theta}^{\theta_t} \left|\rho_{\theta_t}(s) - \rho_{\theta}(s)\right| ds + \int_{\theta_t}^{\infty} \left|\rho_{\theta_t}(s) - \rho_{\theta}(s)\right| ds \\ &= \int_{\theta}^{\theta_t} \rho_{\theta}(s)ds + \int_{\theta_t}^{\infty} \rho_{\theta_t}(s)ds - \int_{\theta_t}^{\infty} \rho_{\theta}(s)ds \\ &= 2\int_{\theta}^{\theta_t} \rho_{\theta}(s)ds = 2F_{\theta}(\theta_t).\end{aligned}$$

La consistencia fuerte de los estimadores $\{\theta_t\}$ *y la continuidad por la derecha de* F_θ *implican que*

$$\eta_t \to 0 \quad P_\theta - a.s.$$

Por otra parte tenemos

$$\begin{aligned}\bar{\eta}_t &= E\int_{\mathbb{R}} |\rho_t(s) - \rho(s)|ds = 2E(F_\theta(\theta_t)) \\ &= 2\int_{\theta}^{\infty} F_\theta(s)\rho_{\theta_t}(s)ds.\end{aligned}$$

Además notemos que

$$\begin{aligned}\rho_{\theta_t}(s) = F'_{\theta_t}(s) &= (1 - (1 - F_\theta(s))^t)' \\ &= t(1 - F_\theta(s))^{t-1}\rho_\theta(s).\end{aligned}$$

Entonces, de las igualdades anteriores y un cambio de variable adecuado concluimos

$$\begin{aligned}\bar{\eta}_t &= 2\int_{\theta}^{\infty} t\rho_\theta(s)F_\theta(s)(1 - F_\theta(s))^{t-1}ds \qquad (3.11)\\ &= \frac{2}{t+1} \to 0, \quad \textit{cuando} \quad t \to \infty.\end{aligned}$$

3.3. Algoritmo iteración de políticas estimado y aproximado

Para definir el algoritmo iteración de políticas estimado-aproximado (IPEA), el modelo $\mathcal{M}_t$ es perturbado usando un promediador L, obteniendo así un *"modelo estimado-perturbado"* dado por

$$\widetilde{\mathcal{M}}_t = \left(\mathbf{X}, \mathbf{A}, \{\widetilde{Q}_f^{(t)}, \widetilde{C}_f^{(t)} : f \in \mathbb{F}\}\right)$$

donde

$$\widetilde{Q}_f^{(t)} := LQ_f^{(t)}(\cdot|\cdot),\ f \in \mathbb{F}$$

y

$$\widetilde{C}_f^{(t)} := LC_f^{(t)}(\cdot).$$

Notemos que el modelo $\mathcal{M}_t$ satisface la Hipótesis 1.3.1 (vea Observación 3.2.2). De ésta manera, los resultados de las secciones anteriores se cumplen para $\mathcal{M}_t$ y $\widetilde{\mathcal{M}}_t$.

Sea $\widetilde{V}_f^{(t)}$ el *costo descontado* bajo la política $f \in \mathbb{F}$ para el modelo $\widetilde{\mathcal{M}}_t$, esto es,

$$\widetilde{V}_f^{(t)}(x) := \widetilde{E}_{x,f}^{(t)} \sum_{k=0}^{\infty} \alpha^k \widetilde{C}_f^{(t)}(\widetilde{x}_k^{(t)}),\ x \in \mathbf{X}.$$

donde $\{\widetilde{x}_k^{(t)}\}$ denota la cadena de Markov inducida por la política $f \in \mathbb{F}$ en el modelo $\widetilde{\mathcal{M}}_t$.

La *función de valor óptimo* en el modelo $\widetilde{\mathcal{M}}_t$, es

$$\widetilde{V}_*^{(t)}(x) := \inf_{\pi \in f} \widetilde{V}_f^{(t)}(x).$$

Algoritmo Iteración de políticas estimado-aproximado (IPEA)

(i) Inicio: sea $\widehat{g}_0 \in \mathbb{F}$ y $n = 0$;

(ii) evaluación: dado $\widehat{g}_n \in \mathbb{F}$, calcule $v_n(\cdot) := \widetilde{V}_{\widehat{g}_n}^{(t)}(\cdot)$;

(iii) mejoramiento: calcule la política $\widehat{g}_{n+1} \in \mathbb{F}$ tal que

$$\begin{aligned} & \inf_{f \in \mathbb{F}} [\widetilde{C}_f^{(t)}(x) + \alpha \int_{\mathbf{X}} v_n(y) \widetilde{Q}_f^{(t)}(dy|x)] \qquad (3.12) \\ = \ & \widetilde{C}_{\widehat{g}_{n+1}}^{(t)}(x) + \alpha \int_{\mathbf{X}} v_n(y) \widetilde{Q}_{\widehat{g}_{n+1}}^{(t)}(dy|x) \ \ \forall x \in \mathbf{X}. \end{aligned}$$

El siguiente resultado es consecuencia de la Proposición 2.4.1 y de la Observación 3.2.2.

Proposición 3.3.1 *Supongamos que se cumple la Hipótesis 3.2.1 y que L es un promediador fuertemente continuo. Entonces la sucesión $\{v_n(\cdot)\}$ dada*

por el algoritmo IPEA converge decrecientemente en la norma del supremo a la función de valor óptimo $\widetilde{V}_*^{(t)}(\cdot)$. *Además se cumple*

$$\|\widetilde{V}_*^{(t)} - v_n\|_\infty \leq \frac{2\alpha}{1-\alpha}\|v_n - v_{n-1}\|_\infty \quad \forall n \in \mathbb{N}.$$

Para n suficientemente grande se espera que la política $\widehat{g}_n$ generada por el algoritmo IPEA sea una buena política para el modelo $\mathcal{M}$, siempre que el modelo $\widetilde{\mathcal{M}}_t$ sea una buena aproximación de $\mathcal{M}_t$, y que al mismo tiempo que el modelo $\mathcal{M}_t$ sea una buena estimación del modelo $\mathcal{M}$. Para medir la exactitud de esas aproximaciones defina las constantes

$$\delta_{C^{(t)}} := \sup_{f\in\mathbb{F}_1} \|\widetilde{C}_f^{(t)}(\cdot) - C_f^{(t)}(\cdot)\|_\infty, \tag{3.13}$$

$$\delta_{Q^{(t)}} := \sup_{x\in\mathbf{X}, f\in\mathbb{F}_1} \|\widetilde{Q}_f^{(t)}(\cdot|x) - Q_f^{(t)}(\cdot|x)\|_{TV}, \tag{3.14}$$

donde $\mathbb{F}_1$ es una subclase de políticas estacionarias que contiene las políticas óptimas estacionarias para los modelos $\mathcal{M}$, $\mathcal{M}_t$, y $\widetilde{\mathcal{M}}_t$, así como las políticas generadas por el algoritmo IPEA.

Notemos que η_t, definido en (3.5), representa el error de estimación cuando se aproxima el modelo $\mathcal{M}$ con el modelo $\mathcal{M}_t$, mientras que $\delta_{C^{(t)}}$ y $\delta_{Q^{(t)}}$ miden la exactitud de la aproximación que el operador L hace del modelo $\mathcal{M}_t$. Además, recuerde que, el error por detener el algoritmo IPEA está dado por la Proposición 3.3.1.

El siguiente teorema establece las cotas de aproximación del algoritmo IPEA.

Teorema 3.3.2 *Supongamos que se cumple la Hipótesis 1.3.1 y que* L *es un promediador fuertemente continuo. Entonces, para cada* $t, n \in \mathbb{N}$*:*
(a)

$$\|\widetilde{V}_*^{(t)} - V_*\|_\infty \leq \frac{1}{1-\alpha}\delta_{C^{(t)}} + \frac{\alpha M}{(1-\alpha)^2}\delta_{Q^{(t)}} + \left\{\frac{M}{1-\alpha} + \frac{\alpha M}{(1-\alpha)^2}\right\}\eta_t;$$

(b)

$$\begin{aligned}\|V_* - V_{\hat{g}_n}\|_\infty \quad \leq \quad & \frac{2}{1-\alpha}\delta_{C^{(t)}} + \frac{2\alpha M}{(1-\alpha)^2}\delta_{Q^{(t)}} + \left\{\frac{M}{1-\alpha} + \frac{\alpha M}{(1-\alpha)^2}\right\}\eta_t \\ & + \frac{2\alpha}{1-\alpha}\|v_n - v_{n-1}\|_\infty.\end{aligned}$$

La demostración del Teorema 3.3.2 está basada en las Proposiciones 3.3.1 y 3.2.3, y en los Lemas 3.3.3 y 3.3.4 dados a continuación.

Lema 3.3.3 *Bajo la Hipótesis 3.2.1, se tiene*

$$||\widetilde{V}_*^{(t)} - V_*^{(t)}||_\infty \leq \frac{1}{1-\alpha}\delta_{C^{(t)}} + \frac{\alpha M}{(1-\alpha)^2}\delta_{Q^{(t)}} \ \forall t \geq 0.$$

La demostración de este lema se sigue aplicando los mismos argumentos empleados en la demostración del Teorema 2.4.2(a) y por lo tanto será omitida.

Lema 3.3.4 *Si L es un promediador fuertemente continuo y se cumple la Hipótesis 3.2.1, entonces*

$$||\widetilde{V}_f^{(t)} - V_f||_\infty \leq \frac{1}{1-\alpha}\delta_{C^{(t)}} + \frac{\alpha M}{(1-\alpha)^2}\delta_{Q^{(t)}} + \left\{\frac{M}{1-\alpha} + \frac{\alpha M}{(1-\alpha)^2}\right\}\eta_t \ \ \forall f \in \mathbb{F}. \tag{3.15}$$

Demostración. Procediendo como en la demostración de la desigualdad (2.12), se puede demostrar que

$$\begin{aligned} ||\widetilde{V}_f^{(t)} - V_f||_\infty \ &\leq \ \frac{1}{1-\alpha}\sup_{f\in\mathbf{F}}||\widetilde{C}_f^{(t)} - C_f||_\infty \\ &\quad + \frac{\alpha M}{(1-\alpha)^2}\sup_{x\in\mathbf{X}, f\in\mathbf{F}}\|\widetilde{Q}_f^{(t)}(\cdot|x) - Q_f(\cdot|x)\|_{TV}, \end{aligned} \tag{3.16}$$

para todo $t \in \mathbb{N}$ y $f \in \mathbb{F}$. Además,

$$||\widetilde{C}_f^{(t)} - C_f||_\infty \leq ||\widetilde{C}_f^{(t)} - C_f^{(t)}||_\infty + ||C_f^{(t)} - C_f||_\infty, \ \ \forall t \in \mathbb{N}, f \in \mathbb{F}.$$

Entonces, de (3.2), (3.7), (3.13) y (3.5), se obtienen las desigualdades

$$\begin{aligned} \sup_{f\in\mathbf{F}}||\widetilde{C}_f^{(t)} - C_f||_\infty &\leq \sup_{f\in\mathbf{F}}||\widetilde{C}_f^{(t)} - C_f^{(t)}||_\infty + M\int_{\mathbb{R}}|\rho_t(s) - \rho(s)|ds \\ &\leq \delta_{C_t} + M\eta_t. \end{aligned} \tag{3.17}$$

Análogamente, pero usando ahora (3.14), también se cumple que

$$\begin{aligned} &\sup_{x\in\mathbf{X}, f\in\mathbf{F}}||\widetilde{Q}_f^{(t)}(\cdot|x) - Q_f(\cdot|x)||_{TV} \\ &\leq \sup_{x\in\mathbf{X}, f\in\mathbf{F}}\left\{||\widetilde{Q}_f^{(t)}(\cdot|x) - Q_f^{(t)}(\cdot|x)||_{TV} + ||Q_f^{(t)}(\cdot|x) - Q_f(\cdot|x)||_{TV}\right\} \\ &\leq \delta_{Q_t} + \eta_t \end{aligned} \tag{3.18}$$

Combinando (3.16)-(3.18) se demuestra (3.15). ■

Demostración del teorema 3.3.2. Primero notemos que, para cada $t \in \mathbb{N}$, se cumple la desigualdad

$$||\widetilde{V}_*^{(t)} - V_*||_\infty \leq ||\widetilde{V}_*^{(t)} - V_*^{(t)}||_\infty + ||V_*^{(t)} - V_*||_\infty.$$

Entonces, de la Proposición 3.2.3 y el Lema 3.3.3 se sigue la parte (a).

Por otro lado, notemos que

$$\|V_* - V_{\hat{g}_n}\|_\infty \leq ||V_* - \widetilde{V}_*^{(t)}||_\infty + ||\widetilde{V}_*^{(t)} - \widetilde{V}_{\hat{g}_n}^{(t)}||_\infty + ||\widetilde{V}_{\hat{g}_n}^{(t)} - V_{\hat{g}_n}||_\infty.$$

Ahora, como $v_n(\cdot) := \widetilde{V}_{\widehat{g}_n}^{(t)}(\cdot)$, por la parte (a) del teorema, la Proposición 3.3.1 y el Lema 3.3.4 vemos que se cumple la parte (b). ■

Observación 3.3.5 ***(a)*** *Notemos que $\delta_{C^{(t)}}$ y $\delta_{Q^{(t)}}$ pueden ser controladas con la elección de un promediador L lo suficientemente preciso, esto es, por constantes que dependan de L pero no de la muestra aleatoria $\mathbf{w}_t := (w_0, w_1, \ldots, w_{t-1})$.*
(b) *Supongamos que se cumplen las condiciones del Teorema 3.3.2. Además, suponga que $\delta_{C^{(t)}} \leq d_1$ y $\delta_{Q^{(t)}} \leq d_2$ donde d_1 y d_2 son constantes positivas que no dependen de la muestra aleatoria $\mathbf{w}_t := (w_0, w_1, \ldots, w_{t-1})$ y finalmente supongamos también que $\bar{\eta}_t = O(t^{-\gamma})$ cuando $t \to \infty$, esto es, que existen constantes $d > 0$ y $t_0 \in \mathbb{N}$ tales que*

$$\bar{\eta}_t := E \int_{\mathbb{R}} |\rho_t(s) - \rho(s)| ds \leq dt^{-\gamma} \quad \forall t \geq t_0, \tag{3.19}$$

para alguna constante positiva γ. Entonces,

$$E\|\widetilde{V}_*^{(t)} - V_*\|_\infty \leq \tfrac{d_1}{1-\alpha} + \tfrac{\alpha M d_2}{(1-\alpha)^2} + \left\{ \tfrac{M}{1-\alpha} + \tfrac{\alpha M}{(1-\alpha)^2} \right\} dt^{-\gamma}; \tag{3.20}$$

y

$$\begin{aligned} E\|V_* - V_{\hat{g}_{k+1}}\|_\infty \leq{}& \frac{d_1}{1-\alpha} + \frac{\alpha M d_2}{(1-\alpha)^2} + \left\{ \frac{2M}{1-\alpha} + \frac{2\alpha M}{(1-\alpha)^2} \right\} dt^{-\gamma} \\ &+ \frac{2\alpha}{1-\alpha} E\|v_{k+1} - v_k\|_\infty. \end{aligned} \tag{3.21}$$

3.4. Ejemplo: aproximación en un sistema de inventario

Considere un sistema de inventario en tiempo discreto con un único producto y capacidad finita $K > 0$. Sea x_n el nivel de inventario, a_n la cantidad ordenada a la unidad de producción al inicio del período $n \in \mathbb{N}_0$, la cual es inmediatamente abastecida, y w_n es la demanda del producto durante el período $n \in \mathbb{N}_0$. Asumiendo que no hay acumulación de demanda, es decir, la demanda insatisfecha en cada período se pierde, el sistema de inventario evoluciona de acuerdo a la ecuación

$$x_{n+1} = (x_n + a_n - w_n)^+, \quad n \in \mathbb{N}_0, \tag{3.22}$$

donde $y^+ := \text{máx}\{0, y\}$ para cada $y \in \mathbb{R}$ y $x_0 = x$ es el nivel de inventario inicial. La demanda $\{w_n\}$ es una sucesión de variables aleatoria no negativas independientes e idénticamente distribuidas las cuales son también independientes del nivel de inventario inicial. Sea $F(\cdot)$ la función de distribución común del proceso de la demanda y supongamos que admite una función de densidad ρ, esto es,

$$F(s) = \int_0^s \rho(t)dt, \ \forall s \in \mathbb{R}.$$

Así, el espacio de estados y controles son $\mathbf{X} = \mathbf{A} = [0, K]$, mientras que $A(x) = [0, K - x]$ es el conjunto de controles admisibles cuando el nivel de inventario es $x \in \mathbf{X}$. Notemos que la multifunción $x \to A(x)$ es continua con valores en los conjuntos compactos. La ley de transición del sistema está dada por

$$Q(B|x, a) := E_w I_B \left[(x + a - w_0)^+\right] \quad \forall B \in \mathcal{B}(\mathbf{X}), (x, a) \in \mathbb{K},$$

donde E_w representa la esperanza con respecto a la función de distribución $F(\cdot)$. Notemos que la ley de transición satisface la igualdad

$$\int_{\mathbb{R}} v(y)Q(dy|x, a) = E_w v((x + a - w_0)^+),$$

para todo $(x, a) \in \mathbb{K}$ y para cualquier función medible v en $\mathbf{X}$, siempre que la integral este bien definida.

Observemos además que si $v \in C_b(\mathbf{X})$ entonces se cumple la Hipótesis 1.3.1(c).

La función de costo en una etapa está dada por

$$c(x, a, w) := p(w - x - a)^+ + h(x + a) + ca, \tag{3.23}$$

donde p, h, y c son constantes positivas que representan el costo unitario por demanda insatisfecha, el costo unitario por almacenamiento, y el costo unitario de producción, respectivamente. Por lo tanto, la función de costo en un período está dada por

$$C(x,a) := pE_w(w-x-a)^+ + h(x+a) + ca \quad \forall (x,a) \in \mathbb{K}. \tag{3.24}$$

Bajo estas condicones, el problema de inventario satisface la Hipótesis 1.3.1.

Ahora considere el operador de aproximación L definido por el esquema de interpolación lineal con N nodos igualmente espaciados $0 = s_1 < s_2 < \cdots < s_N = K$, y sea $\Delta s := K/N$. Entonces, para cada función medible v en $\mathbf{X}$, el operador L es definido como

$$Lv(x) = \frac{s_{i+1}-x}{s_{i+1}-s_i} v(s_i) + \frac{x-s_i}{s_{i+1}-s_i} v(s_{i+1}), \tag{3.25}$$

donde $x \in [s_i, s_{i+1}]$, $i = 1, 2, \ldots, N-1$.

Claramente, L es un promediador débilmente continuo, esto es, es un promediador que mapea $C_b(\mathbf{X})$ en el mismo (Definición 2.2.1).

Se puede demostrar que para cada $\alpha \in (0,1)$ existe una política estacionaria de umbral óptima [39] de la forma $f(x) = S_\alpha - x$ para $x \in [0, S_\alpha]$, y $f(x) = 0$ en otro caso, donde el punto de re-ordenamiento S_α es una constante no negativa [39]. El punto de re-ordenamiento de una política de umbral óptima $f_{S^*_\alpha}$ que satisface la ecuación

$$F(S^*_\alpha) = \frac{p-h-c}{p-\alpha c},$$

siempre que $p > h + c$. Si $p \leq h + c$ el punto de reordenamiento óptimo es $S^*_\alpha = 0$.

Argumentos similares muestran que existen políticas de umbral óptimas para el modelo $\widetilde{\mathcal{M}}$. Por lo tanto, tomamos $\mathbb{F}_0$ y $\mathbb{F}_1$ como la clase de políticas de umbral.

Para calcular las cantidades δ_C y δ_Q definidas en (2.10) y (2.11), suponemos que el modelo de inventario satisface las siguientes condiciones adicionales:

(a) La densidad ρ es Lipschitz continua con constante $l > 0$;

(b) ρ es acotada por una constante l';

(c) La demanda esperada $\bar{w} := E_w w_0$ es finita.

Bajo estas condiciones adicionales, es posible mostrar que

$$\delta_C = \sup_{f \in \mathbb{F}_0} \|\widetilde{C}_f - C_f\|_\infty \leq (p + 2h + 2c)\Delta s, \tag{3.26}$$

$$\delta_Q = \sup_{x \in \mathbf{X}, f \in \mathbb{F}_0} ||\widetilde{Q}_f(\cdot|x) - Q_f(\cdot|x)||_{TV} \leq (2Kl + 4l')\Delta s. \tag{3.27}$$

Las cotas anteriores combinadas con el Teorema 2.4.2 dejan las siguientes desigualdades:

$$||\widetilde{V}_* - V_*||_\infty \leq \left\{ \frac{p + 2h + 2c}{1 - \alpha} + \frac{\alpha(2Kl + 4l')M}{(1 - \alpha)^2} \right\} \Delta s \tag{3.28}$$

y

$$||V_* - V_{g_k}||_\infty \leq \left\{ \frac{p + 2h + 2c}{1 - \alpha} + \frac{\alpha(2Kl + 4l')M}{(1 - \alpha)^2} \right\} 2\Delta s + \frac{2\alpha}{1 - \alpha} \|u_k - u_{k-1}\|_\infty, \tag{3.29}$$

donde g_k es la política tal que $\widetilde{T} u_{k-1}(\cdot) = \widetilde{T}_{g_k} u_{k-1}(\cdot)$ y M es la cota para el costo por etapa (3.24).

Claramente, $||\widetilde{V}_* - V_*||_\infty$ y $||V_* - V_{g_k}||_\infty$ se pueden hacer arbitrariamente pequeñas tomando una partición suficientemente fina y realizando un número suficientemente grande de iteraciones del algoritmo de iteración de políticas aproximado.

Resultados numéricos. Para ilustrar los resultados numéricos, suponga que la demanda tiene una densidad exponencial,

$$\rho(s) = \lambda \exp(-\lambda s) I_{[0,\infty)}(s), \quad s \in \mathbb{R},$$

con $\lambda = 0.1$. Supongamos además que $c = 1.5$, $h = 0.5$, $p = 3$, $K = 40$. Ahora observe que $\rho(\cdot) \leq l' := \lambda$ y que además es una función de Lipschitz con constante $l := \lambda^2$.

Denotemos por $u_k(\cdot)$, $k \in \mathbb{N}$, a las funciones producidas por el algoritmo IPA y sean $\widetilde{S}_k$ las constantes que representan el punto de reordenamiento de la políticas g_k.

El algoritmo inicia con la política $g(x) := 0$, $x \in \mathbf{X}$, y se detiene cuando la cantidad $||u_k - u_{k-1}||_\infty$ es menor que la tolerancia $\varepsilon = 0.001$. Las Tablas

1 y 2 muestran los resultados numéricos para el algoritmo IPA con factores de descuento $\alpha = 0.6$ y $\alpha = 0.99$, respectivamente, con una partición del espacio de estados con N nodos para $N = 100, 500, 1000, 2000$.

Las Tablas 1 y 2, y las Figuras 1 y 2 muestran que el algoritmo IPA converge muy rápido, y prácticamente identifica los puntos de reordenamiento óptimos $S^*_{0.6} = 6.4663$ y $S^*_{0.99} = 10.79$. En todos los casos el algoritmo IPA se detiene en la cuarta iteración. De hecho, para $N = 2000$, $u_4(\cdot)$ es virtualmente la función de valor óptimo $V^*(\cdot)$ y $\widetilde{S}_4$ es el punto de re-ordenamiento.

La Tabla 3 muestra cotas para δ_C y δ_Q así como cotas de error para $||\widetilde{V}_* - V_*||_\infty$ y $||V_* - V_{g_k}||_\infty$ con factor de descuento $\alpha = 0.6$.

Note que las cotas de error para $||\widetilde{V}_* - V_*||_\infty$ y $||V_* - V_{g_k}||_\infty$ son muy sensibles al factor de descuento α a pesar que las cantidades δ_C y δ_Q pueden controlarse fácilmente.

Tabla 1, α=0.6, S^*_α=6.4663

N	100	500	1000	2000
$\|v_4 - v_3\|_\infty$	$1.635e^{-4}$	$2.018e^{-6}$	$2.761e^{-7}$	$4.184e^{-8}$
$\widetilde{S}_4$	6.463971	6.466142	6.466244	6.46627

Tabla 2, α=0.99, $S^*_{,99}$=10,79

N	100	500	1000	2000
$\|v_4 - v_3\|_\infty$	$1.013e^{-3}$	$1.040e^{-4}$	$1.525e^{-5}$	$1.389e^{-5}$
$\widetilde{S}_4$	10.78305	10.78982	10.7899	10.790

Tabla 3, α=0.6

N	100	500	1000	2000	5000	10000
δ_C	2.8	0.56	0.28	0.14	0.056	0.028
δ_Q	0.48	0.096	0.048	0.024	0.0096	0.0048
$\|\|\widetilde{V}_* - V_*\|\|_\infty$	97	19.4	9.7	4.75	1.93	0.95
$\|\|V_* - V_{g_k}\|\|_\infty$	194	38.8	19.4	9.5	3.86	1.9

Para ilustrar el algoritmo IPEA, consideremos el caso donde la función de densidad de la demanda w pertenece a la siguiente familia paramétrica de densidades

$$\rho_\theta(z) = \lambda \exp(-\lambda(z - \theta)) I_{[\theta,\infty)}(z), \tag{3.30}$$

donde el parámetro θ es *desconocido* pero está en algún intervalo $\Theta = [\theta_1, \theta_2]$ con $0 \leq \theta_1 < \theta_2$.

El objetivo es estimar el parámetro desconocido θ y con base a esta estimación, calcular la política óptima aproximada. Para esto, suponemos que las variables aleatorias w_n, $n \in \mathbb{N}$, son observables y que una muestra $\mathbf{w}_t := (w_0, w_1, \ldots, w_{t-1})$ está disponible para el controlador. Notemos que ρ_θ es un caso particular al presentado en el Ejemplo 3.2.5 tomando $\varphi(z) = \exp(-\lambda z)$. Entonces el estimador de máxima verosimilitud de θ es $\theta_t = \text{mín}\{w_0, w_1, ..., w_{t-1}\}$, y de aquí, un estimador de la densidad está dado por

$$\rho_t(z) = \lambda \exp(-\lambda(z - \theta_t)) I_{[\theta_t, \infty)}(z). \tag{3.31}$$

En este caso tenemos $\theta_t \to \theta$ $P_\theta - c.s.$, y además se cumple

$$\eta_t = \int_\theta^\infty \left| \rho_{\theta_t}(s) - \rho_\theta(s) \right| ds \to 0$$

Las cotas para las cantidades $\delta_{C^{(t)}}$ y $\delta_{Q^{(t)}}$ definidas en (3.13) y (3.14) pueden ser obtenidas análogamente como se obtuvieron δ_C y δ_Q usando $\rho_t(\cdot)$ en lugar de $\rho(\cdot)$, y observando que la densidad $\rho_t(\cdot)$ está acotada por $l'_t := \lambda$ y es Lipschitz continua con módulo $l_t = \lambda^2 \exp(\lambda\theta_t)$. Entonces de (3.26) y (3.27) obtenemos

$$\delta_{C^{(t)}} \leq (p + 2h + 2c)\Delta s$$

y

$$\delta_{Q^{(t)}} \leq (2Kl_t + 4l'_t M)\Delta s.$$

Mas aún, por (3.20) y (3.21), se sigue que

$$\begin{aligned} E\|\widetilde{V}_t^* - V_*\|_\infty \quad \leq \quad & \left\{ \frac{p + 2h + 2c}{1 - \alpha} \Delta s + \frac{\alpha M (2Kl_t + 4\beta_t)}{(1 - \alpha)^2} \right\} \Delta s \\ & + \left\{ \frac{M}{1 - \alpha} + \frac{\alpha M}{(1 - \alpha)^2} \right\} \frac{2}{(t + 1)} \end{aligned} \tag{3.32}$$

y

$$\begin{aligned} E\|V_* - V_{\hat{g}_k}\|_\infty \leq & \left\{ \frac{p + 2h + 2c}{1 - \alpha} + \frac{\alpha M (2Kl_t + 4\beta_t)}{(1 - \alpha)^2} \right\} \Delta s \\ & + \left\{ \frac{M}{1 - \alpha} + \frac{\alpha M}{(1 - \alpha)^2} \right\} \frac{2}{(t + 1)} + \frac{2\alpha}{1 - \alpha} E\|v_k - v_{k-1}\|_\infty, \end{aligned} \tag{3.33}$$

donde

$$\beta_t := El_t = \lambda^2 t e^{\lambda\theta}(t-1)^{-1} \leq \lambda^2 t e^{\lambda\theta_2}(t-1)^{-1}$$

Los resultados numéricos mostrados en la Tabla 4 y Tabla 5 corresponden a los valores $c = 1.5$, $h = 0.5$, $p = 3$, $K = 40$, $\lambda = 0.1$ y factores de descuento $\alpha = 0.6$, y 0.99 respectivamente. Los puntos de reordenamiento óptimos son $S^*_{0.6} = 11.4663$ y $S^*_{0.99} = 15.79$. El algoritmo IPEA fue implementado con muestras $\mathbf{w} = (w_0, \ldots, w_{t-1})$ simuladas con la densidad (3.31) para $t = 10, 20, 50$ suponiendo que el valor verdadero del parámetro es $\theta = 5$ y usando una partición del espacio de estados con $N = 100$ igualmente espaciados.

Tabla 4

$\alpha = 0.6, \quad S^*_{0.6} = 11.4663$

t	10	20	50
$\|v_4 - v_3\|$	$6.064e^{-7}$	$5.984e^{-7}$	$7.331e^{-7}$
$\widetilde{S}_{4,t}$	13.9394	11.6970	11.5385
θ_t	5.3253	5.2307	5.0722

Tabla 5

$\alpha = 0.99, \quad S^*_{0.99} = 15.79$

t	10	20	50
$\|v_4 - v_3\|$	$3.526e^{-5}$	$2.050e^{-5}$	$3.084e^{-5}$
$\widetilde{S}_{4,t}$	16.5738	16.0126	15.8759
θ_t	5.7837	5.2225	5.0859

Capítulo 4

Procesos de Control Semi-Markovianos

4.1. Introducción

A diferencia de los procesos de control markoviano en tiempo discreto estudiados en los capítulos anteriores, los procesos de control semi-Markovianos son una clase de procesos en tiempo continuo que tienen la característica de que el tiempo de permanencia en cada estado es una variable aleatoria δ_n, $n \in \mathbb{N}$, cuya distribución, en un contexto general, dependerá del estado del sistema, del control aplicado y del estado en el siguiente tiempo de decisión.

El objetivo de este capítulo es introducir la teoría básica de los procesos de control semi-markovianos y definir el problema de control óptimo con índice de funcionamiento descontado. Estableceremos condiciones bajo las cuales se garantiza la existencia de políticas óptimas, e introduciremos un ejemplo para ilustrar los resultados. Estos procesos se estudiarán en el contexto de espacios de estado y control de Borel, costos posiblemente no acotados, espacio de controles compactos.

4.2. Modelo de control semi-markoviano

Definición 4.2.1 *Un Modelo de control semi-markoviano es un arreglo de la forma*

$$SM := (\mathbf{X}, \mathbf{A}, \{A(x) : x \in \mathbf{X}\}, Q, C), \tag{4.1}$$

donde $\mathbf{X}$ *y* $\mathbf{A}$ *representan los espacios de estados y controles, respectivamente, los cuales supondremos son espacios de Borel;* $A(x)$ *es el conjunto*

de acciones admisibles para el estado $x \in \mathbf{X}$. *Denotamos*

$$\mathbb{K} := \{(x, a) \in \mathbf{X} \times \mathbf{A} : x \in \mathbf{X}, a \in A(x)\}.$$

Además, $Q(\cdot, \cdot|x, a)$ *es un kernel estocástico en* $\mathbf{X}\times\mathbb{R}_+$ *dado* $\mathbb{K}$, *y* C *es una función medible en* $\mathbb{K}$ *que representa el costo por etapa.*

El modelo SM representa un sistema estocástico controlado en tiempo continuo que evoluciona de la siguiente manera. En la $n-$ésima época de decisión el sistema se encuentra en el estado $x_n = x \in \mathbf{X}$, se elige un control $a_n = a \in A(x)$ y se incurre en un costo $C(x, a)$. El sistema permanece en el estado x un tiempo aleatorio δ_{n+1}, y se mueve a un nuevo estado $x_{n+1} = y \in \mathbf{X}$ de acuerdo a la probabilidad conjunta del evento $x_{n+1} \in B$ y $\delta_{n+1} \leq t$ dada por el kernel $Q(B, [0, t]|x, a)$, donde $B \in \mathcal{B}(\mathbf{X})$ y $t \in \mathbb{R}_+$.

Una vez que el proceso se encuentra en el estado y, se elige un nuevo control y el proceso se repite. Así, para cada $n \in \mathbb{N}_0$, denotaremos por x_n, a_n y δ_{n+1} el estado del sistema inmediatamente después de la n-ésima transición, el control seleccionado y el tiempo que el sistema permanece en el estado x_n, respectivamente.

A las variables aleatorias δ_{n+1}, $n \in \mathbb{N}_0$, se les llama *tiempos de permanencia* y las variables aleatorias

$$T_0 := 0, \quad \text{y} \quad T_n := T_{n-1} + \delta_n \quad \text{para} \quad n \in \mathbb{N}, \tag{4.2}$$

se les llama los *tiempos de transición.*

Las políticas de control son definidas de manera similar que al caso markoviano (Definición 1.2.2) pero considerando ahora que las historias admisibles incluyen los tiempos de permanencia. Es decir, el espacio de historias admisibles es de la forma $\mathbb{H}_0 := \mathbf{X}$, y para cada $n \in \mathbb{N}$, $\mathbb{H}_n := (\mathbb{K}\times\mathbb{R}_+)^n \times \mathbf{X}$.

Un elemento genérico $h_n \in \mathbb{H}_n$ es un vector de la forma

$$h_n = (x_0, a_0, ..., \delta_{n-1}, x_{n-1}, a_{n-1}, \delta_n, x_n).$$

Denotemos por Π al conjunto de todas las políticas y por $\mathbb{F}$ al conjunto de políticas estacionarias.

Procesos controlados semi-markovianos. Sea $(\Omega, \mathcal{F})$ el espacio medible canónico que consiste del espacio muestral $\Omega := (\mathbf{X} \times \mathbf{A}\times\mathbb{R}_+)^\infty$ y su σ-algebra producto $\mathcal{F}$. Entonces, para cada estado inicial $x \in \mathbf{X}$ y política

$\pi \in \Pi$, existe una medida de probabilidad $\mathbb{P}_x^\pi$ en el espacio $(\Omega, \mathcal{F})$ tal que para cada $D \in \mathcal{B}(\mathbf{A}), B \in \mathcal{B}(\mathbf{X})$, $h_n := (x_0, a_0, ..., \delta_{n-1}, x_{n-1}, a_{n-1}, \delta_n, x_n) \in \mathbb{H}_n$, $n \in \mathbb{N}$ ([1, Teorema 2.7.2], [6, Capítulo 7]), se cumple lo siguiente:

- $\mathbb{P}_x^\pi[x_0 = x] = 1$;
- $\mathbb{P}_x^\pi[a_n \in D|h_n] = \pi_n(D|h_n)$;
- $\mathbb{P}_x^\pi[x_{n+1} \in B, \delta_{n+1} \leq t|h_n, a_n] = Q(B, (0,t]|x_n, a_n)$.

El operador esperanza con respecto a $\mathbb{P}_x^\pi$ es denotado por E_x^π. Para $t \in \mathbb{R}_+$ y $(x,a) \in \mathbb{K}$ se definen las distribuciones marginales por

$$G(t|x,a) := Q(\mathbf{X}, [0,t]|x,a) \tag{4.3}$$

$$P(B|x,a) := Q(B, \mathbb{R}_+|x,a). \tag{4.4}$$

El modelo de control semi-markoviano (4.1) representa un proceso que se desarrolla en tiempo continuo con transiciones en los tiempos aleatorios T_n, $n \in \mathbb{N}$, esto es, la evolución está dada por las variables aleatorias

$$Z_t := x_n \quad si \quad T_n \leq t < T_{n+1}, n \in \mathbb{N}.$$

Observe que el proceso Z_t es un proceso que cambia de estado de acuerdo al proceso x_n, y los tiempos donde se dan las transiciones están dados por el proceso T_n.

Un aspecto importante de los procesos es determinar si el proceso es regular, es decir, si únicamente presenta un número finito de transiciones en periodos acotados de tiempo. Para describir este comportamiento se definen las variables aleatorias

$$N(t) := \sup\{n \geq 1 : T_n \leq t\}, \ t \in \mathbb{R}_+,$$

que cuentan las transiciones del proceso en el intervalo de tiempo $[0,t]$. Observe que $N(t) < \infty$ para todo $t \geq 0$ si y solo si $T_n \to \infty$ cuando $n \to \infty$. Para $\pi \in \Pi$, $x \in \mathbf{X}$, diremos que el proceso $\{Z_t, t \geq 0\}$ es *regular en* x si

$$\mathbb{P}_x^\pi[T_n \to \infty] = 1.$$

Si la propiedad anterior se cumple para todo $x \in \mathbf{X}$ diremos que el proceso es regular.

En la siguiente sección se dan condiciones para que el proceso sea regular (vea Observación 4.3.3(e)).

4.3. Hipótesis en el modelo semi-Markoviano

Para definir el criterio de costo descontado supondremos que los costos son continuamente descontados con una tasa $\alpha > 0$. Esto significa que un costo unitario en t unidades de tiempo es equivalente a $exp(-\alpha t)$ unidades en el tiempo presente. De acuerdo a esta consideración tenemos la siguiente definición. Para $\alpha > 0$, se define el *costo descontado semi-markoviano* por

$$V_\pi(x) := E_x^\pi \sum_{n=0}^{\infty} e^{-\alpha T_n} C(x_n, a_n), \ x \in \mathbf{X}, \ \pi \in \Pi. \tag{4.5}$$

La *función de valor óptimo* correspondiente se define como

$$V_*(x) := \inf_{\pi \in \Pi} V_\pi(x), \quad x \in \mathbf{X}.$$

Entonces el problema de control óptimo consiste en encontrar una política π^* tal que

$$V_*(x) = V_{\pi^*}(x), \quad x \in \mathbf{X},$$

a la cual llamaremos óptima.

Las siguientes condiciones garantizan que el costo descontado semi-markoviano está bien definido para todas las políticas, así como la existencia de políticas óptimas.

Hipótesis 4.3.1 ***(a)*** *Para cada $x \in \mathbf{X}$, el conjunto $A(x)$ es compacto.*
(b) *$C(x, a)$ es una función continua en $A(x)$ para cada $x \in \mathbf{X}$.*
(c) *La función*

$$a \to \int_{\mathbf{X} \times \mathbb{R}_+} e^{-\alpha t} u(y) Q(dy, dt|x, a)$$

es continua para cada $u \in M_b(\mathbf{X})$, $x \in \mathbf{X}$.
(d) *Existe una función medible $W : \mathbf{X} \to [\mathbf{1}, \infty)$, y constantes $\beta \in (0, 1)$, y $\overline{c} \in \mathbb{R}$ tales que satisfacen las siguientes propiedades para cada $x \in \mathbf{X}$:*

(d1) $|C(x, a)| \leq \overline{c} W(x)$;

(d2) $\int_{\mathbf{X} \times \mathbb{R}_+} e^{-\alpha t} W(y) Q(dy, dt|x, a) \leq \beta W(x)$;

(d3) la función $a \to \int_{\mathbf{X} \times \mathbb{R}_+} e^{-\alpha t} W(y) Q(dy, dt|x, a)$ *es continua en* $A(x)$.

Ahora se presentan algunas consecuencias de la Hipótesis 4.3.1. Sea $B_W(\mathbf{X})$ el espacio de las funciones medibles $u : \mathbf{X} \to \mathbb{R}$ que satisfacen la siguiente condición:

$$||u||_W = \sup_{x \in \mathbf{X}} \frac{|u(x)|}{W(x)} < \infty.$$

Proposición 4.3.2 *Bajo la Hipótesis 4.3.1, para cada* $x \in \mathbf{X}$, $\pi \in \Pi$, $u \in B_W(\mathbf{X})$, *y* $n \in \mathbb{N}$, *se cumple*

$$E_x^\pi\{e^{-\alpha T_n} W(x_n)\} \leq \beta^n W(x). \tag{4.6}$$

Demostración. Por propiedades de la esperanza condicional se tiene que

$$\begin{aligned}
&E_x^\pi\{e^{-\alpha T_n} W(x_n)\} \\
&= E_x^\pi \left[E_x^\pi \left\{ e^{-\alpha T_{n-1}} e^{-\alpha \delta_n} W(x_n) \right\} | h_{n-1}, a_{n-1} \right] \\
&= E_x^\pi [e^{-\alpha T_{n-1}} E_x^\pi \{ e^{-\alpha \delta_n} W(x_n) | h_{n-1}, a_{n-1} \}] \\
&= E_x^\pi [e^{-\alpha T_{n-1}} \int_{\mathbf{A}} \int_{\mathbf{X} \times \mathbb{R}_+} e^{-\alpha t} W(y) Q(dy, dt | x_{n-1}, a) \pi(da | h_{n-1})] \\
&\leq E_x^\pi [e^{-\alpha T_{n-1}} \int_{\mathbf{A}} \beta W(x_{n-1}) \pi(da | h_{n-1})] \\
&= \beta E_x^\pi [e^{-\alpha T_{n-1}} W(x_{n-1})\}].
\end{aligned}$$

Luego iterando la última desigualdad se obtiene (4.6). ■

Observación 4.3.3 *Bajo la Hipótesis 4.3.1 se cumple lo siguiente.*

(a) Como una consecuencia inmediata de (4.6) tenemos la desigualdad

$$E_x^\pi \sum_{n=0}^{\infty} \{e^{-\alpha T_n} W(x_n)\} \leq \frac{W(x)}{1-\beta}, \quad \forall x \in \mathbf{X}, \ \pi \in \Pi, \ n \in \mathbb{N}.$$

(b) Para cada $u \in B_W(\mathbf{X})$, *de forma análoga como la demostración de (4.6), se demuestra*

$$E_x^\pi e^{-\alpha T_n} u(x_n) \leq W(x) ||u||_W \beta^n, \quad \forall x \in \mathbf{X}, \ \pi \in \Pi. \tag{4.7}$$

(c) Para cada política $\pi \in \Pi$, *se cumple*

$$||V_\pi||_W \leq \frac{\overline{c}}{1-\beta}.$$

En efecto, observe que de (4.6),

$$\begin{aligned} |V_\pi(x)| &\leq E_x^\pi \sum_{n=0}^{\infty} e^{-\alpha T_n} |C(x_n, a_n)| \\ &\leq \overline{c} E_x^\pi \sum_{n=0}^{\infty} e^{-\alpha T_n} W(x_n) \leq \frac{\overline{c}}{1-\beta} W(x); \end{aligned}$$

por lo tanto,

$$||V_\pi||_W \leq \frac{\overline{c}}{1-\beta}.$$

(d) Claramente (c) implica la desigualdad

$$||V_*||_W \leq \frac{\overline{c}}{1-\beta}.$$

(e) Para cada política $\pi \in \Pi$ y estado inicial $x_0 = x \in \mathbf{X}$, el proceso es regular, es decir la sucesión T_n, $n \in \mathbb{N}$, diverge a infinito $P_x^\pi-$casi seguramente. Este hecho se sigue de (4.6) y de la siguiente relación

$$E_x^\pi\{e^{-\alpha T_n}\} \leq E_x^\pi\{e^{-\alpha T_n} W(x_n)\} \leq \beta^n W(x) \to 0,$$

lo cual implica que $T_n \to \infty$ $P_x^\pi-$casi seguramente, para cada estado inicial $x \in \mathbf{X}$.

4.4. Existencia de políticas óptimas

Para $u \in B_W(\mathbf{X})$ se define el operador de programación dinámica para el modelo semi-markoviano como

$$Tu(x) := \inf_{a \in A(x)} \left\{ C(x,a) + \int_{\mathbf{X} \times \mathbb{R}_+} e^{-\alpha t} u(y) Q(dy, dt|x, a) \right\}, \quad x \in \mathbf{X}. \quad (4.8)$$

Proposición 4.4.1 *Supongamos que se cumple la Hipótesis 4.3.1. Entonces:*
(a) Para cada $u \in B_W(\mathbf{X})$, $Tu \in B_W(\mathbf{X})$ y existe $f \in \mathbb{F}$ tal que

$$Tu(x) = C_f(x) + \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} u(y) Q_f(dy, dt|x) \tag{4.9}$$

(b) T es un operador de contracción con módulo β en el espacio de Banach $(B_W(\mathbf{X}), ||\cdot||_W)$.

Demostración (a) De [16, Lema 8.3.7], se sigue que para $u \in B_W(\mathbf{X})$, la función

$$a \to C(x,a) + \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} u(y) Q(dy, dt|x, a)$$

es continua en $A(x)$ para cada $x \in \mathbf{X}$. Entonces, por un teorema de selección medible [16, Lema 8.3.8], Tu es una función medible y existe $f \in \mathbb{F}$ tal que satisface (4.9).

Ahora mostraremos que $Tu \in B_W(\mathbf{X})$. Para esto observemos que

$$\begin{aligned}
|Tu(x)| &\le |C(x,a)| + \left| \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} u(y) Q(dy, dt|x, a) \right| \\
&\le \overline{c} W(x) + ||u||_W \left| \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} W(y) Q(dy, dt|x, a) \right| \\
&\le \overline{c} W(x) + ||u||_W \beta W(x).
\end{aligned}$$

Entonces,

$$||Tu||_W = \sup_{x\in\mathbf{X}} \frac{|Tu(x)|}{W(x)} \le \overline{c} + ||u||_W \beta < \infty,$$

lo cual implica que $Tu \in B_W(\mathbf{X})$.

(b) Para $u, v \in B_W(\mathbf{X})$ se cumple

$$\begin{aligned}
|Tu - Tv| &\le | \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} (u(y) - v(y)) Q(dy, dt|x, a)| \\
&\le ||u - v||_W | \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} W(y) Q(dy, dt|x, a)| \\
&\le \beta ||u - v||_W W(x),
\end{aligned}$$

lo cual implica la desigualdad

$$||Tu - Tv||_W \leq \beta ||u - v||_W.$$

Esto demuestra que T es un operador de contracción con módulo β. ■

Ahora para $u \in B_W(\mathbf{X})$ y $f \in \mathbb{F}$ se define el operador

$$T_f u(x) := C_f(x) + \int\limits_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} u(y) Q_f(dy, dt|x).$$

Análogamente como con el operador T, se puede demostrar que $T_f,\ f \in \mathbb{F}$, es un operador de contracción del espacio Banach $(B_W(\mathbf{X}), ||\cdot||_W)$ en sí mismo con módulo β y su único punto fijo es $V_f(x)$, esto es,

$$V_f(x) = C_f(x) + \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} V_f(y) Q_f(dy, dt|x) = T_f V_f(x). \tag{4.10}$$

Esto último se prueba de la siguiente manera:

$$\begin{aligned}
V_f(x) &= E_x^f[C_f(x_0) + \sum_{n=1}^{\infty} e^{-\alpha T_n} C_f(x_n)] \\
&= C_f(x) + E_x^f\{E_x^f[\sum_{n=1}^{\infty} e^{-\alpha\delta_1} e^{-\alpha T_{n-1}} C_f(x_n)|h_1]\} \\
&= C_f(x) + E_x^f\{e^{-\alpha\delta_1} E_{x_1}^f[\sum_{n=0}^{\infty} e^{-\alpha T_n} C_f(x_n)]\} \\
&= C_f(x) + E_x^f\{e^{-\alpha\delta_1} V_f(x_1)\} \\
&= C_f(x) + \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} V_f(y) Q_f(dy, dt|x).
\end{aligned}$$

Como consecuencia de los resultados previos tenemos la siguiente proposición, la cual garantiza la existencia de políticas óptimas.

Proposición 4.4.2 *Supongamos que se cumple la Hipótesis 4.3.1. Entonces:* ***(a)*** *La función de valor óptimo V_* es el único punto fijo en $B_W(\mathbf{X})$ del operador T, esto es,*

$$V_*(x) = TV_*(x) := \min_{a\in A(x)} \left\{ C(x,a) + \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} V_*(y) Q(dy, dt|x, a) \right\}. \tag{4.11}$$

(b) *Una política estacionaria $f \in \mathbb{F}$ es óptima si y sólo si alcanza el mínimo en (4.11), esto es,*

$$V_*(\cdot) = T_f V_*(\cdot).$$

Demostración. Por la Proposición 4.4.1 existe un punto fijo $u^* \in B_W(\mathbf{X})$ del operador T y existe $f \in \mathbb{F}$ tal que

$$u^*(x) = C_f(x) + \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} u^*(y) Q_f(dy, dt|x).$$

Entonces, por (4.10), u^* es el punto fijo de T_f, y por lo tanto

$$u^*(x) = V_f(x) \geq V_*(x). \tag{4.12}$$

Para obtener la desigualdad contraria, notemos que

$$u^*(x) = Tu^*(x) \leq C(x,a) + \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} u^*(y) Q(dy, dt|x, a).$$

Ahora, sea $\pi \in \Pi$ una política arbitraria. Para cada estado inicial $x \in \mathbf{X}$ y $m = 0, 1, ...,$ se cumple que

$$0 \leq E_x^\pi\{C(x_m, a_m) + e^{-\alpha\delta_{m+1}} u^*(x_{m+1}) - u^*(x_m)|h_m, a_m\},$$

de donde se sigue la desigualdad

$$0 \leq E_x^\pi\{e^{-\alpha T_m} C(x_m, a_m) + e^{-\alpha T_{m+1}} u^*(x_{m+1}) - e^{-\alpha T_m} u^*(x_m)|h_m, a_m\}.$$

Tomando esperanza E_x^π y sumando para $m = 0, 1, ..., n$, obtenemos

$$0 \leq E_x^\pi \left\{ \sum_{m=1}^{n} e^{-\alpha T_m} C(x_m, a_m) + \sum_{m=1}^{n} e^{-\alpha T_{m+1}} u^*(x_{m+1}) - \sum_{m=1}^{n} e^{-\alpha T_m} u^*(x_m) \right\},$$

lo cual implica

$$0 \leq E_x^\pi \left\{ \sum_{m=1}^{n} e^{-\alpha T_m} C(x_m, a_m) + e^{-\alpha T_{m+1}} u^*(x_{m+1}) - u^*(x_0) \right\}.$$

Haciendo tender n a infinito, por (4.7) tenemos que se cumple

$$u^*(x) \leq V_\pi(x),$$

y como $\pi \in \Pi$ es arbitraria, se concluye

$$u^*(x) \leq V_*(x).$$

Combinando (4.12) y la última desigualdad llegamos a que

$$V_* = TV_*.$$

(b) Ahora sea $f^* \in \mathbb{F}$ tal que

$$V_* = TV_* = T_{f^*}V_*.$$

Entonces por (4.10),

$$V_*(x) = V_{f^*}(x),$$

lo cual implica que f^* es una política óptima. Ahora, sea f^* una política óptima, entonces

$$V_*(x) = V_{f^*}(x).$$

Finalmente, de (4.10) y la parte (a) concluimos que

$$V_*(x) = V(x, f^*) = T_{f^*}V_*(x) = T\ V_*(x) \quad \forall x \in \mathbf{X}. \quad \blacksquare$$

4.5. Ejemplo: un problema de remplazo

Consideremos un sistema que es sometido a una sucesión de choques que se producen aleatoriamente en el tiempo. Cada uno de los choques provoca una cantidad aleatoria de daño en el sistema, el cual se acumula conforme pasa el tiempo. Cualquiera de los choques puede provocar que el sistema falle por completo, de tal manera que una falla en el sistema sólo puede ocurrir en el momento de un choque. Supondremos que la probabilidad de falla es $1 - r(\cdot)$, donde $r : \mathbb{R}_+ \to [0, 1]$, es una función del daño no-creciente. A esta función se le llama de supervivencia.

Denotemos por t_n, $n \in \mathbb{N}$, los tiempos en los que ocurren los choques y por $l_n := t_n - t_{n-1}$ los tiempos entre choques consecutivos. Supondremos que la sucesión t_n, $n \in \mathbb{N}$, es estrictamente creciente.

Para cada $n \in \mathbb{N}_0$, denote por w_n la magnitud aleatoria de daño en el tiempo t_n y por x_n el daño acumulado hasta el tiempo t_n. Denotamos por $J(\cdot|x)$ la función de distribución condicional del daño w_n dado $x_n = x$ y por $H(\cdot)$ la función de distribución de los tiempos entre choques consecutivos. Supondremos que estas distribuciones son continuas.

El proceso de remplazo es el siguiente. En el tiempo $t_n = t$ el controlador observa un daño acumulado $x_n = x$ y elige un tiempo de remplazo programado

$a \in \mathbf{A} = A(x) = [\theta_1, \theta_2]$, donde θ_1 y θ_2 son constantes tales que $0 < \theta_1 < \theta_2$. Entonces se presentan los siguientes dos escenarios:

(i) Si $l_{n+1} = l < a$, el sistema es remplazado por falla con probabilidad $1 - r(x + w)$, donde $w = w_{n+1}$ es la magnitud del daño al tiempo t_{n+1}. Además se incurre en un costo fijo K_1.

(ii) Si $l_{n+1} = l \geq a$, el sistema se remplaza en el tiempo programado a, i.e., antes de la falla. En este caso se incurre en un costo $K_2 < K_1$.

Adicionalmente a los costos contemplados en (i) y (ii) supondremos que existe un costo K_3 por operar el sistema, el cual es proporcional al daño acumulado x.

Observemos que el sistema se remplaza al tiempo que se presenta la falla o al tiempo programado, dependiendo cuál sea menor. De esta manera, si definimos la variable aleatoria

$$\delta_n := \text{mín}(l_n, a_n),$$

los tiempos de transición del sistema está n dados por $T_0 = 0$,

$$T_n = T_{n-1} + \delta_n, \quad n \in \mathbb{N}.$$

El proceso de daño acumulado $\{x_n\}$ es un proceso semi-markoviano con espacios de estados $\mathbf{X} = [0, \infty)$ y espacio de control es $\mathbf{A} = A(x) = [\theta_1, \theta_2]$, con tiempos de permanencia $\{\delta_n\}$.

Con todos estos elementos, es fácil ver que el kernel conjunto Q está dado por

$$\begin{aligned} Q(B, [0,t]|x, a) = &\int_0^{\text{mín}(a,t)} H(ds) \int_{(x+w)\in B} r(x+w) J(dw|x) \qquad (4.13) \\ &+ I_B(0) \int_0^{\text{mín}(a,t)} H(ds) \int_0^{\infty} (1 - r(x+w)) J(dw|x) \\ &+ I_B(0) I_{(t>a)} \{1 - H(a)\}, \quad B \in \mathcal{B}(\mathbf{X}), t \in \mathbb{R}_+. \end{aligned}$$

De esto se sigue que la distribución marginal en el tiempo es

$$G(t|x, a) = Q(\mathbf{X}, (0, t]|x, a) = \begin{cases} H(t), & t < a \\ \\ 1 & a \leq t \end{cases}$$

y la distrubución marginal en la variable de estados es

$$\begin{aligned} P(B|x,a) &= Q(B,\mathbb{R}_+|x,a) \\ &= H(a)\int_{(x+w)\in B} r(x+w)J(dw|x) \\ &+ I_B(0)H(a)\int_0^\infty (1-r(x+w))J(dw|x) \\ &+ I_B(0)\{1-H(a)\}. \end{aligned}$$

Ahora defina

$$\widehat{c}(x,a,\delta,y) := \begin{cases} K_1 + K_3 x & si \quad l < a,\ y = 0 \\ K_2 + K_3 x & si \quad a < l,\ y = 0 \\ K_3 x & si \quad l < a,\ y > 0 \end{cases} .$$

Entonces, la función de costo esperado por etapa está dada por

$$\begin{aligned} C(x,a) &= \int \widehat{c}(x,a,\delta,y)Q(dy,d\delta|x,a) \\ &= K_1 H(a)\left\{\int_0^\infty (1-r(x+w))J(dw|x)\right\} + K_2(1-H(a)) + K_3 x. \end{aligned}$$

Ahora notemos que para cada $u \in M_b(\mathbf{X})$

$$\begin{aligned} \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t}u(y)Q(dy,dt|x,a) &= \int_0^a e^{-\alpha t}H(dt)\int_0^\infty u(x+w)r(x+w)J(dw|x) \\ &+ u(0)\int_0^a e^{-\alpha t}H(dt)\int_0^\infty (1-r(x+w))J(dw|x) \\ &+ u(0)\int_a^\infty e^{-\alpha t}H(dt). \end{aligned}$$

Entonces, por la continuidad de H, las funciones

$$a \to C(x,a) \quad \text{y} \quad a \to \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t}u(y)Q(dy,dt|x,a)$$

son continuas, y por lo tanto, se cumplen las Hipótesis 4.3.1 (b), (c).

Para verificar la Hipótesis 4.3.1 (d1) y (d2), se introduce la siguiente hipótesis.

Hipótesis 4.5.1 *Existen constantes $\alpha_0 > 0$ y $q > 0$ tales que*

$$\sup_{a\in[\theta_1,\theta_2]}\{\int_0^a e^{-\alpha_0 t}H(dt)\int_0^\infty (e^{qw}r(w)+(1-r(w))J(dw)+e^{-\alpha_0 a}(1-H(a))\}<1.$$

La Hipótesis 4.3.1 (d1) se cumple considerando la función $W(x) = e^{qx}$ y eligiendo una constante adecuada $\overline{c} \in \mathbb{R}$, tal que

$$|C(x,a)| \leq K_1 + K_2 + K_3 x \leq \overline{c}W(x).$$

Ahora procederemos a verificar la Hipótesis 4.3.1 (d2). Observemos que

$$\begin{aligned}
&\int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t}W(y)Q(dy,dt|x,a)\\
&= \int_0^a e^{-\alpha t}H(dt)\int_0^\infty e^{q(x+w)}r(x+w)J(dw)\\
&+ \int_0^a e^{-\alpha t}H(dt)\int_0^\infty (1-r(x+w))J(dw) + \int_a^\infty e^{-\alpha t}H(dt)\\
&\leq e^{qx}\beta(\alpha,x,a),
\end{aligned}$$

donde

$$\begin{aligned}
\beta(\alpha,x,a) := &\int_0^a e^{-\alpha t}H(dt)\{\int_0^\infty (e^{qw}r(x+w) + e^{-qx}(1-r(x+w)))J(dw)\}\\
&+ \int_a^\infty e^{-\alpha t}H(dt).
\end{aligned}$$

Ahora considere la función

$$I(x) := e^{qw}r(x+w) + e^{-qx}(1-r(x+w))$$

y supongamos que $r(\cdot)$ es derivable entonces,

$$I'(x) = r'(x+w)(e^{qw} - e^{-qx}) - qe^{-qx}(1-r(x+w)).$$

Es fácil ver que $I'(x) < 0$ debido a que $r'(x+w) < 0$ y $(e^{qw} - e^{-qx}) > 0$, lo que implica que $I(x)$ es no creciente. De aquí, el máximo valor de $I(x)$ se alcanza en $x = 0$, esto es,

$$I(0) = 1 + r(w)(e^{qw} - 1) \geq 1,$$

y por lo tanto

$$\begin{aligned}\beta(\alpha,x,a) &\leq \beta(\alpha,0,a)\\ &= \int_0^a e^{-\alpha t}H(dt)\{\int_0^\infty (e^{qw}r(w)+(1-r(w)))J(dw)\}\\ &+ \int_a^\infty e^{-\alpha t}H(dt).\end{aligned}$$

Ahora, por la Hipótesis 4.5.1 tenemos existe $\alpha_0 > 0$ tal que

$$\sup_{a\in[\theta_1,\theta_2]} \beta(\alpha_0,x,a) \leq \sup_{a\in[\theta_1,\theta_2]} \beta(\alpha_0,0,a) < 1.$$

Notemos que si $\alpha_0 < \alpha$ entonces $\beta(\alpha_0,x,a) \geq \beta(\alpha,x,a)$. De esta manera, para $\alpha > \alpha_0$, se cumple

$$\begin{aligned}\int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t}W(y)Q(dy,dt|x,a) &\leq e^{qx}\beta(\alpha,x,a)\\ &\leq e^{qx}\beta(\alpha,0,a) \leq e^{qx}\beta(\alpha_0,0,a)\\ &\leq e^{qx}\sup_{a\in[\theta_1,\theta_2]}\beta(\alpha_0,0,a) \leq \beta W(x),\end{aligned}$$

donde $\beta := \sup_{a\in[\theta_1,\theta_2]}\beta(\alpha_0,0,a) < 1$, lo cual implica la Hipótesis 4.3.1 (d2).

Es importante señalar que la Hipótesis 4.5.1 no es tan restrictiva, observe que para que se cumpla es necesario que exista q tal que

$$\int_0^\infty (e^{qw}r(w)+(1-r(w)))J(dw) < \infty,$$

lo cual se satisface si tomamos por ejemplo $r(x) = e^{-bx},\ b > 0$, con $b > q$. Entonces,

$$O := \int_0^\infty (e^{-w(b-q)}+(1-e^{-bw}))J(dw) < \infty.$$

En este caso, para $\alpha > 0$ y $a \in \mathbf{A} = [\theta_1,\theta_2]$, tenemos

$$\beta(\alpha,0,a) = O\int_0^a e^{-\alpha t}H(dt) + \int_a^\infty e^{-\alpha t}H(dt).$$

Observemos que el primer sumando del lado derecho de la igualdad anterior es creciente como función de a, mientras que el segundo sumando es decreciente. Por lo tanto, para todo $a \in \mathbf{A} = [\theta_1,\theta_2]$, se cumple

$$\beta(\alpha,0,a) \leq O\int_0^{\theta_2} e^{-\alpha t}H(dt) + e^{-\alpha\theta_1}(1-H(\theta_1)).$$

Tomando supremo sobre $a \in \mathbf{A} = [\theta_1, \theta_2]$ en $\beta(\alpha, 0, a)$ obtenemos

$$\sup_{a \in [\theta_1, \theta_2]} \beta(\alpha, 0, a) \leq \overline{O},$$

donde

$$\overline{O} := O \int_0^{\theta_2} e^{-\alpha t} H(dt) + e^{-\alpha \theta_1}(1 - H(\theta_1)).$$

Entonces es posible escoger $\alpha_0 > 0$ tal que

$$O \int_0^{\theta_2} e^{-\alpha_0 t} H(dt) + e^{-\alpha_0 \theta_1}(1 - H(\theta_1)) < 1,$$

y por lo tanto se cumple la Hipótesis 4.5.1, esto es

$$\beta = \sup_{a \in [\theta_1, \theta_2]} \beta(\alpha_0, 0, a) < 1.$$

Tomando $q = 0.1$, $b = 0.2$, $H(\cdot) \approx$Gamma$(3, 0.5)$, $J(\cdot) \approx$Weibull$(3, 1)$, es posible verificar que para $\alpha \geq 0.2$ se cumple la Hipótesis 4.5.1.

Capítulo 5

Aproximación de Modelos Semi-Markovianos

5.1. Introducción

El objetivo de este capítulo es aproximar la solución del problema de control óptimo para modelos semi-markovianos con costos posiblemente no acotados y espacio de estados de Borel localmente compacto. Como un primer paso a la aproximación, se presenta un procedimiento de truncamiento en el espacio de estados. Esto nos permitirá definir un modelo de control res tringido a un conjunto compacto.

Luego se estudia un modelo semi-markoviano perturbado, que se obtiene de perturbar el modelo original (4.1) con promediadores definidos en el espacio de Banach de las funciones medibles $(B_W(X), \|\cdot\|_W)$. Luego combinando los procedimientos de truncamiento y perturbación se presenta un modelo semi-markoviano perturbado-truncado en un conjunto compacto, de tal manera que las propiedades importantes del modelo original se heredan al modelo perturbado-truncado.

La política óptima del modelo original será aproximada por las políticas arrojadas por el algoritmo de IP implementado en el modelo perturbado-truncado. Para medir el desempeño de dichas políticas en el modelo semi-markoviano original, el Teorema 5.4.2 presenta las cotas de error por los diferentes procedimientos implementados en éste capítulo, es decir, por truncamiento, perturbación y por detener el algoritmo de IP.
Finalmente para ilustrar los resultados del capítulo, se aproxima el modelo de remplazo introducido en el capítulo anterior.

5.2. Modelo semi-markoviano truncado

Sea $S\mathcal{M} = (\mathbf{X}, \mathbf{A}, \{A(x) : x \in \mathbf{X}\}, Q, \widehat{c})$ un modelo de control semi-markoviano (Definición 4.2.1) que satisface la Hipótesis 4.3.1. Además impondremos la siguiente condición.

Hipótesis 5.2.1 *El espacio de estados* $\mathbf{X}$ *es un espacio de Borel localmente compacto.*

Observación 5.2.2 *(a) Las propiedades de separabilidad y de compacidad local en el espacio de estados implican que* $\mathbf{X}$ *es un espacio* $\sigma-$*compacto. Por lo tanto, existe una sucesión de subconjuntos abiertos* $O_n, n \in \mathbb{N}$*, tal que* $\overline{O_n} \subset O_{n+1}$ *para todo* $n \in \mathbb{N}$ *y* $\mathbf{X} = \cup_{n=1}^{\infty} O_n$*, [35, Teorema 21, p. 203]*

(b) Bajo las Hipótesis 4.3.1 y 5.2.1, para cada $\epsilon > 0$ *existe un subconjunto compacto* $\mathbf{X}_\epsilon \subset \mathbf{X}$ *tal que*

$$\int_{\mathbf{X}_\epsilon^c \times \mathbb{R}_+} e^{-\alpha t} W(y) Q(dy, dt|x, a) \leq \epsilon W(x) \quad \forall (x, a) \in \mathbb{K} \tag{5.1}$$

donde $\mathbf{X}_\epsilon^c := \mathbf{X} \backslash \mathbf{X}_\epsilon$.

En esta sección nos centraremos en el estudio del problema de control óptimo hasta que el proceso salga del espacio "truncado" $\mathbf{X}_\epsilon$. Para esto definimos

$$\tau = \tau(\epsilon) := \inf\{n \geq 1 : x_n \in \mathbf{X}_\epsilon^c\},$$

donde $\inf \phi := +\infty$.

Para $x \in \mathbf{X}_\epsilon$, $\pi \in \Pi$, se define el índice

$$V_\pi^\epsilon(x) := E_x^\pi \sum_{n=0}^{\infty} e^{-\alpha T_n} C(x_n, a_n) I_{[n<\tau]},$$

y la función de valor óptimo

$$V_*^\epsilon(x) = \inf_{\pi \in \Pi} V_\pi^\epsilon(x).$$

Así el problema de control óptimo restringido al espacio $\mathbf{X}_\epsilon$ consiste en encontrar una política $\pi \in \Pi$ tal que

$$V_*^\epsilon(x) := V_\pi^\epsilon(x), \ x \in \mathbf{X}_\epsilon.$$

Para el estudio de éste problema haremos uso de los siguientes resultados preliminares.

Proposición 5.2.3 *Suponga que se cumplen las Hipótesis 4.3.1 y 5.2.1. Entonces para cada política* $\pi \in \Pi$ *y estado inicial* $x \in \mathbf{X}_\epsilon$

$$\left| E_x^\pi \sum_{n=0}^{\infty} e^{-\alpha T_n} C(x_n, a_n) I_{[n \geq \tau]} \right| \leq \frac{\overline{c}\epsilon}{(1-\beta)^2} W(x).$$

Demostración de la Proposición5.2.3. Para $n \in \mathbb{N}$, y $k < n$, por la Proposición 4.3.2, se sigue

$$\begin{aligned}
&E_x^\pi e^{-\alpha T_n} W(x_n) I_{[\tau=k]} \\
&= E_x^\pi [E_x^\pi e^{-\alpha T_n} W(x_n) I_{[\tau=k]} | h_{n-1}, a_{n-1})] \\
&= E_x^\pi [e^{-\alpha T_{n-1}} I_{[\tau=k]} E_x^\pi (e^{-\alpha \delta_n} W(x_n) | h_{n-1}, a_{n-1})] \\
&\leq \beta E_x^\pi [e^{-\alpha T_{n-1}} I_{[\tau=k]} W(x_{n-1})].
\end{aligned}$$

Entonces,

$$E_x^\pi e^{-\alpha T_n} W(x_n) I_{[\tau=k]} \leq \beta^{n-k} E_x^\pi [e^{-\alpha T_k} I_{[\tau=k]} W(x_k)]. \tag{5.2}$$

Además por la Observación 5.2.2(b) se cumple

$$\begin{aligned}
&E_x^\pi [e^{-\alpha T_k} I_{[\tau=k]} W(x_k)] \\
&= E_x^\pi [E_x^\pi (e^{-\alpha T_k} W(x_k) I_{[\tau=k]} | h_{k-1}, a_{k-1})] \\
&= E_x^\pi [e^{-\alpha T_{k-1}} I_{[x_1 \in \mathbf{X}_\epsilon, .., x_{k-1} \in \mathbf{X}_\epsilon]} E_x^\pi (e^{-\alpha \delta_k} W(x_k) I_{[x_k \in \mathbf{X}_\epsilon^c]} | h_{k-1}, a_{k-1})] \\
&\leq \epsilon E_x^\pi [e^{-\alpha T_{k-1}} I_{[x_1 \in \mathbf{X}_\epsilon, .., x_{k-1} \in \mathbf{X}_\epsilon]} W(x_{k-1})].
\end{aligned}$$

De nuevo por la Proposición 4.3.2,

$$E_x^\pi [e^{-\alpha T_{k-1}} I_{[x_1 \in \mathbf{X}_\epsilon, .., x_{k-1} \in \mathbf{X}_\epsilon]} W(x_{k-1})] \leq \beta^{k-1} W(x). \tag{5.3}$$

Por lo tanto, de las desigualdades (5.2) y (5.3),

$$E_x^\pi e^{-\alpha T_n} W(x_n) I_{[\tau=k]} \leq \epsilon \beta^{n-1} W(x),$$

lo cual a su vez implica

$$E_x^\pi e^{-\alpha T_n} W(x_n) I_{[\tau \le n]} = E_x^\pi \sum_{k=1}^{n} e^{-\alpha T_n} W(x_n) I_{[\tau = k]}$$
$$\le n\epsilon\beta^{n-1} W(x).$$

Finalmente, por la Hipótesis 4.3.1(d1) y la desigualdad anterior

$$E_x^\pi \sum_{n=0}^{\infty} e^{-\alpha T_n} C(x_n, a_n) I_{[\tau \le n]} \le \overline{c} E_x^\pi \sum_{n=0}^{\infty} e^{-\alpha T_n} W(x_n) I_{[\tau \le n]}$$
$$\le \overline{c}\epsilon W(x) \sum_{n=0}^{\infty} n\beta^{n-1} \le \frac{\overline{c} W(x)}{(1-\beta)^2}\epsilon. \quad \blacksquare$$

La siguiente proposición anterior nos provee de una cota para el error producido por truncación. Más aún, prueba que las funciones $V_*^\epsilon(\cdot)$ convergen uniformemente en la norma ponderada a $V_*(\cdot)$ en subconjuntos compactos cuando $\epsilon \to 0$.

Para establecer este resultado denotemos por $B_W(\mathbf{X}_\epsilon)$ al espacio de las funciones medibles $u : \mathbf{X}_\epsilon \to \mathbb{R}$ tales que

$$||u||_W^\epsilon = \sup_{x \in \mathbf{X}_\epsilon} \frac{|u(x)|}{|W(x)|} < \infty.$$

Proposición 5.2.4 *Suponga que se cumplen la Hipótesis 4.3.1 y 5.2.1 . Entonces*

(a) Para cada política $\pi \in \Pi$ y estado inicial $x \in \mathbf{X}_\epsilon$,

$$||V_\pi - V_\pi^\epsilon||_W^\epsilon \le \frac{\overline{c}}{(1-\beta)^2}\epsilon.$$

(b)

$$||V_* - V_*^\epsilon||_W^\epsilon \le \frac{\overline{c}}{(1-\beta)^2}\epsilon.$$

Demostración. (a) Dado $\epsilon > 0$, $x \in \mathbf{X}_\epsilon$, $\pi \in \Pi$, observemos que

$$V_\pi(x) := E_x^\pi \sum_{n=0}^{\infty} e^{-\alpha T_n} C(x_n, a_n) I_{[n<\tau]} + E_x^\pi \sum_{n=0}^{\infty} e^{-\alpha T_n} C(x_n, a_n) I_{[\tau \le n]}$$
$$= V_\pi^\epsilon(x) + E_x^\pi \sum_{n=0}^{\infty} e^{-\alpha T_n} C(x_n, a_n) I_{[\tau \le n]}$$

Entonces, por la Proposición 5.2.3, para cada $x \in \mathbf{X}_\epsilon$ tenemos

$$\begin{aligned} |V_\pi(x) - V_\pi^\epsilon(x)| &= |E_x^\pi \sum_{n=0}^{\infty} e^{-\alpha T_n} C(x_n, a_n) I_{[\tau \leq n]}| \\ &\leq \frac{\overline{c} W(x)}{(1-\beta)^2} \epsilon, \end{aligned}$$

de donde concluimos que

$$||V_\pi - V_\pi^\epsilon||_W^\epsilon \leq \frac{\overline{c}}{(1-\beta)^2} \epsilon.$$

(b) Notemos que

$$\begin{aligned} |V_*(x) - V_*^\epsilon(x)| &= |\inf_{\pi \in \Pi} V_\pi(x) - \inf_{\pi \in \Pi} V_\pi^\epsilon(x)| \\ &\leq \sup_{\pi \in \Pi} |V_\pi(x) - V_\pi^\epsilon(x)| \\ &\leq \frac{\overline{c} W(x)}{(1-\beta)^2} \epsilon \end{aligned}$$

de donde se sigue el resultado.■

Observemos $(B_W(\mathbf{X}_\epsilon), || \cdot ||_W^\epsilon)$ es un espacio de Banach. Ahora, para el estudio del problema de control óptimo restringido en $\mathbf{X}_\epsilon$, definimos los operadores de programación dinámica "truncados" como sigue:

$$T_{\epsilon,f} u(x) := C(x,f) + \int_{\mathbf{X}_\epsilon \times \mathbb{R}_+} e^{-\alpha t} u(y) Q(dy, dt|x, f), \tag{5.4}$$

y

$$T_\epsilon u(x) := \min_{a \in A(x)} \left\{ C(x,a) + \int_{\mathbf{X}_\epsilon \times \mathbb{R}_+} e^{-\alpha t} u(y) Q(dy, dt|x, a) \right\}, \tag{5.5}$$

para $f \in \mathbb{F}$, $x \in \mathbf{X}_\epsilon$ y $u \in B_W(\mathbf{X}_\epsilon)$.

Observación 5.2.5 *(a) Si se cumple la Hipótesis 4.3.1, entonces, para $u \in M_b(\mathbf{X})$, $x \in \mathbf{X}_\epsilon$, la función*

$$\begin{aligned} a \to & \int_{\mathbf{X}_\epsilon \times \mathbb{R}_+} e^{-\alpha t} u(y) Q(dy, dt|x, a) \\ = & \int_{\mathbf{X} \times \mathbb{R}_+} e^{-\alpha t} u(y) I_{\mathbf{X}_\epsilon}(y) Q(dy, dt|x, a); \end{aligned}$$

es continua.

(b) Por argumentos análogos a la Proposición 4.4.1, podemos demostrar que los operadores T_ϵ y $T_{\epsilon,f}$ son de contraccón en el espacio $(B_W(\mathbf{X}_\epsilon), ||\cdot||_W^\epsilon)$ con módulo β. Entonces, existen funciones $u^\epsilon, u_f^\epsilon \in B_W(\mathbf{X}_\epsilon)$ tales que

$$T_\epsilon u^\epsilon = u^\epsilon \quad y \quad T_{\epsilon,f} u_f^\epsilon = u_f^\epsilon. \tag{5.6}$$

(c) Además para cada función $u \in B_W(\mathbf{X}_\epsilon)$ existe $f \in \mathbb{F}$ tal que

$$T_\epsilon u(x) = T_{\epsilon,f} u(x) \quad \forall x \in \mathbf{X}_\epsilon$$

Proposición 5.2.6 *Supongamos que se cumplen las Hipótesis 4.3.1 y 5.2.1. Entonces:*

(a) $T_{\epsilon,f} V_f^\epsilon(\cdot) = V_f^\epsilon(\cdot)$ para toda $f \in \mathbb{F}$.

(b) $V_f^\epsilon(\cdot) = V_*^\epsilon(\cdot)$ si y solo si $V_*^\epsilon(\cdot) = T_{\epsilon,f} V_*^\epsilon(\cdot)$

Demostración (a) Para demostrar que V_f^ϵ es el punto fijo de $T_{\epsilon,f}$, notemos que

$$T_{\epsilon,f}^n u(x) = E_x^f \sum_{k=0}^{n} e^{-\alpha T_k} C_f(x_k) I_{[k<\tau]} + E_x^f [e^{-\alpha T_n} u(x_n) I_{[n<\tau]}] \tag{5.7}$$

para todo $x \in \mathbf{X}_\epsilon$, $f \in \mathbb{F}$, $u \in B_W(\mathbf{X}_\epsilon)$, y $n \in \mathbb{N}$.

Luego, por la desigualdad (4.7) se cumple,

$$\left| E_x^f [e^{-\alpha T_n} u(x_n) I_{[n<\tau]}] \right| \leq W(x) ||u||_W \beta^n \to 0,$$

cuando $n \to \infty$. Ahora por el Teorema de punto fijo de Banach y de (5.7) se sigue

$$u_f^\epsilon(x) = \lim_{n\to\infty} T_{\epsilon,f}^n u(x) = E_x^f \sum_{k=0}^{\infty} e^{-\alpha T_k} C(x_k, a_k) I_{[k<\tau]} = V_f^\epsilon(x)$$

para todo $x \in \mathbf{X}_\epsilon$. Entonces (5.6) demuestra (a).

Notemos que la parte (b) se sigue de la parte (a)■

Proposición 5.2.7 *Suponga que se cumplen las Hipótesis 4.3.1, y 5.2.1. Entonces:*

(a) *La función de valor óptimo V_*^ϵ es el único punto fijo en $B_W(\mathbf{X}_\epsilon)$ del operador T_ϵ, esto es,*

$$V_*^\epsilon(x) = T_\epsilon V_*^\epsilon(x) = \min_{a\in A(x)} \left\{ C(x,a) + \int_{\mathbf{X}_\epsilon\times\mathbb{R}_+} e^{-\alpha t} V_*^\epsilon(y) Q(dy,dt|x,a) \right\}. \tag{5.8}$$

(b) *Existe $f_\epsilon \in \mathbb{F}$ tal que*

$$V_*^\epsilon(\cdot) = T_{\epsilon,,f_\epsilon} V_*^\epsilon(\cdot),$$

y f_ϵ es una política óptima.

Demostración Para la demostración se procede como en la demostración del Teorema 4.4.2. Por la Observación 5.2.5 existe $f_\epsilon \in \mathbb{F}$ tal que

$$u^\epsilon = T_{\epsilon,,f_\epsilon} u^\epsilon.$$

Entonces, u_ϵ^* es el punto fijo de T_{f_ϵ}. Luego de la Proposición 5.2.6(a), se sigue que

$$u_\epsilon = V_{f_\epsilon}^\epsilon(x) \geq V_*^\epsilon.$$

Para la desigualdad contraria, observemos que

$$u_\epsilon^*(x) \leq C(x,a) + \int_{\mathbf{X}_\epsilon\times\mathbb{R}_+} e^{-\alpha t} u_\epsilon^*(y) Q(dy,dt|x,a) \quad \forall x \in \mathbf{X}_\epsilon, a \in A(x).$$

Entonces, para $m = 0,1,2,\ldots$ se cumple

$$0 \leq C(x_m,a_m) I_{[m<\tau]} + \int_{\mathbf{X}_\epsilon\times\mathbb{R}_+} e^{-\alpha t} u_\epsilon^*(y) Q(dy,dt|x_m,a_m) - u_\epsilon^*(x_m) I_{[m<\tau]},$$

de donde se sigue

$$\begin{aligned} 0 \quad \leq \quad & E_x^\pi \left[e^{-\alpha T_m} C(x_m,a_m) I_{[m<\tau]} + e^{-\alpha T_{m+1}} u_\epsilon^*(x_{m+1}) I_{[m<\tau]} \right. \\ & \left. - e^{-\alpha T_m} u_\epsilon^*(x_m) I_{[m<\tau]} \mid h_m, a_m \right], \end{aligned}$$

para todo $x \in \mathbf{X}_\epsilon,\ \pi \in \Pi$.

Sumando $m = 0,1,2,\ldots,n$ y tomando esperanza resulta

$$0 \leq E_x^\pi[\sum_{m=0}^{n} e^{-\alpha T_m} C(x_m, a_m) I_{[m<\tau]} + e^{-\alpha T_{m+1}} u_\epsilon^*(x_{m+1}) I_{[m<\tau]} - u_\epsilon^*(x_0) I_{[m<\tau]}].$$

Por la desigualdad (4.7) y tomando limite cuando $n \to \infty$, se obtiene

$$V_\pi^\epsilon(x) = E_x^\pi[\sum_{m=0}^{\infty} e^{-\alpha T_m} C(x_m, a_m) I_{[m<\tau]} \geq u_\epsilon^*(x),$$

de donde se sigue $V_*^\epsilon(x) \geq u_\epsilon^*(x)$.

Observemos que (b) se sigue de la parte (a) Proposición 5.2.6(b). ■

5.3. Modelo semi-markoviano perturbado

En esta sección introduciremos un modelo semi-markoviano perturbado usando promediadores. Este modelo resulta la aproximación al modelo semi-markoviano introducido en la Definición 4.2.1. Además, por la Hipótesis 5.2.1 impuesta en el espacio de estados, y aplicando el procedimiento desarrollado en la sección anterior, es posible definir un modelo "truncado" en el modelo semi-markoviano perturbado. El objetivo es comparar la función de valor óptimo de este modelo con la función de valor óptimo original (ver Teorema 5.3.6).

Sean $W : \mathbf{X} \to [\mathbf{1}, \infty)$, la función medible y $\beta \in (0, 1)$ la constante de la Hipótesis 4.3.1. Fijemos un promediador L que satisface la siguiente condición adicional.

Hipótesis 5.3.1 *(a)* $\widetilde{W} := L(W) \in B_W(\mathbf{X})$,
(b) $\widetilde{\beta} := \beta \|\widetilde{W}\|_W < 1$.

Notemos que la Hipótesis anterior implica que $Lu \in B_W(\mathbf{X})$ siempre que $u \in B_W(\mathbf{X})$. Esta hipótesis permitirá que el modelo semi-markoviano perturbado herede las propiedades importantes del modelo semi-markoviano original. Por otra parte, con la finalidad de simplificar la exposición supondremos que $A(x) = \mathbf{A}$ $\forall x \in \mathbf{X}$.

Definición 5.3.2 *Se define el modelo de control semi-Markoviano perturbado*

$$\widetilde{SM} := (\mathbf{X}, \mathbf{A}, \widetilde{Q}, \widetilde{C}) \tag{5.9}$$

donde

$$\widetilde{C}(x,a) := LC(x,a) = \int_{\mathbf{X}} C(y,a)L(dy|x) \tag{5.10}$$

y

$$\widetilde{Q}(B,(0,t]|x,a) := LQ(B,(0,t]|x,a), \tag{5.11}$$

para $x \in \mathbf{X}$, $a \in \mathbf{A}$ *y* $B \in \mathcal{B}(\mathbf{X})$.

Notemos que $\widetilde{Q}(\cdot,\cdot|\cdot,\cdot)$ es la composición de las probabilidades de transición Q y L; por lo tanto, $\widetilde{Q}$ es una probabilidad de transición en $\mathbf{X}\times\mathbb{R}_+$ dado $\mathbb{K}$.

Para cada política $\pi \in \Pi$ y estado inicial $x \in \mathbf{X}$, la ley de transición $\widetilde{Q}$ define una medida de probabilidad $\widetilde{P}_{x,\pi}$ y un proceso estocástico $\{\widetilde{x}_n, \widetilde{a}_n, \widetilde{\delta}_n\}$ en el espacio $(\Omega,\mathcal{F}) = (\mathbf{X}\times\mathbf{A}\times\mathbb{R}_+)^\infty, \mathcal{F})$. Sea $\widetilde{E}_{x,\pi}$ el operador esperanza con respecto a la medida de probabilidad $\widetilde{P}_{x,\pi}$.

Ahora se define el índice de funcionamiento correspondiente para el mo delo perturbado. Para $x \in \mathbf{X}$, $\pi \in \Pi$ y $\alpha > 0$, se define el *costo descontado* del modelo semi-Markoviano perturbado por

$$\widetilde{V}_\pi(x) := \widetilde{E}_{x,\pi} \sum_{k=0}^{\infty} e^{-\alpha T_k} \widetilde{C}(\widetilde{x}_k, \widetilde{a}_k),\ x \in \mathbf{X}.$$

La función de valor óptimo se define como

$$\widetilde{V}_*(\cdot) := \inf_\pi \widetilde{V}_\pi(\cdot).$$

Una política π_* es óptima si

$$\widetilde{V}_*(\cdot) = \widetilde{V}_{\pi^*}(\cdot).$$

La siguiente proposición muestra propiedades que se cumplen en el modelo $\widetilde{SM}$.

Proposición 5.3.3 *Bajo las Hipótesis 4.3.1 y 5.3.1 se cumple lo siguiente.*

(a) $|\widetilde{C}(x,a)| \le \widetilde{c}W(x)$ *para toda* $(x,a) \in \mathbb{K}$, *con* $\widetilde{c} := \overline{c}\|\widetilde{W}\|_W$;

(b) $\widetilde{C}(x,\cdot)$ *es una función continua en* $\mathbf{A}$, *para cada* $x \in \mathbf{X}$;

(c) para cada $u \in B_W(\mathbf{X})$ *la función*

$$a \to \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} u(y) \widetilde{Q}(dy,dt|x,a)$$

es continua en $\mathbf{A}$ *para cada* $x \in \mathbf{X}$.

(d) $\int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} W(y)\widetilde{Q}(dy, dt|x, a) \leq \widetilde{\beta} W(x)$ *para cada* $(x, a) \in \mathbb{K}$;

(e) La función $a \to \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} W(y)\widetilde{Q}(dy, dt|x, a)$ *es continua en* $\mathbf{A}$, *para cada* $x \in \mathbf{X}$.

Demostración. (a) Observe que de la Hipótesis 4.3.1 (d1) y la linealidad y monotonía del operador L resulta la desigualdad

$$|\widetilde{C}(x, a)| \leq \overline{c}\widetilde{W}(x) \leq \overline{c}\|\widetilde{W}\|_W W(x),$$

de donde se sigue el resultado.
(b) De Hipótesis 4.3.1(b) y (d1), $C(x, a)$ es continua en $a \in \mathbf{A}$ y acotada por $\overline{c}W(x)$. Sea a_n, $n \in \mathbb{N}$, una sucesión que converge a $a \in \mathbf{A}$, y por el teorema de la convergencia dominada

$$LC(x, a_n) = \int_{\mathbf{X}} C(y, a_n)L(dy|x) \to \int_{\mathbf{X}} C(y, a)L(dy|x) = LC(x, a).$$

por lo tanto $\widetilde{C}(x, a)$ es continua en $a \in \mathbf{A}$.
(c) Sea $u \in B_W(\mathbf{X})$ de [16, Lema 8.3.7], la función

$$a \to \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} u(y)Q(dy, dt|x, a),$$

es continua en $a \in \mathbf{A}$ y de Hipótesis 4.3.1(d2) está acotada por $\beta\|u\|_W W(\cdot)$. Luego por argumentos análogos a los de la demostración en (b) y observando que

$$\int_{\mathbf{X}} \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} u(y)Q(dy, dt|z, a)L(dz|x) = \int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} u(y)\widetilde{Q}(dy, dt|x, a),$$

se sigue el resultado.
(d) Notemos que de la Hipótesis 4.3.1(d2) y de la monotonía del operador L se deduce la desigualdad

$$\int_{\mathbf{X}\times\mathbb{R}_+} e^{-\alpha t} W(y)\widetilde{Q}(dy, dt|x, a) \leq \beta\widetilde{W}(x) \leq \beta\|\widetilde{W}\|_W W(x) = \widetilde{\beta} W(x),$$

para todo $(x, a) \in \mathbb{K}$.

(e) Se sigue por argumentos análogos aplicados en (b).

Observemos que por la Proposición 5.3.3, el modelo semi-markoviano perturbado (5.9) satisface condiciones análogas a las condiciones impuestas por la Hipótesis 4.3.1 en el modelo semi-markoviano original (4.1). Por lo tanto tenemos resultados análogos a los dados por las Proposiciones 4.3.2, 4.4.2 y Observación 4.3.3, para el modelo (5.9). En particular se cumple,

$$\widetilde{E}_{x,\pi}\{e^{-\alpha T_n}u(\widetilde{x}_n)\} \leq \widetilde{\beta}^n u(x)\|u\|_W \quad \forall x \in \mathbf{X}, \pi \in \Pi, u \in B_W(\mathbf{X}). \tag{5.12}$$

Por otra parte, por la Hipótesis 5.2.1 impuesta en el espacio de estados, y repitiendo el procedimiento desarrollado en la sección anterior es posible definir un modelo "truncado" en el modelo semi-markoviano perturbado, lo cual presentamos a continuación.

Sea $\mathbf{X}_\epsilon^c$ el conjunto dado en la Observación 5.2.2 y definamos

$$\widetilde{\tau} = \widetilde{\tau}(\epsilon) := \inf\{n : \widetilde{x}_n \in \mathbf{X}_\epsilon^c\},$$

donde $\inf \phi := +\infty$. Para $x \in \mathbf{X}_\epsilon$, $\pi \in \Pi$, se define

$$\widetilde{V}_\pi^\epsilon(x) := \widetilde{E}_{x,\pi} \sum_{k=0}^{\infty} e^{-\alpha T_k}\widetilde{C}(\widetilde{x}_k, a_k) I_{[n<\widetilde{\tau}]},$$

y la función de valor óptimo

$$\widetilde{V}_*^\epsilon(x) := \inf_{\pi\in\Pi} \widetilde{V}_\pi^\epsilon(x), \ x \in \mathbf{X}_\epsilon.$$

Además para cada $x \in \mathbf{X}_\epsilon$, $f \in \mathbb{F}$ y $u \in B_W(\mathbf{X}_\epsilon)$, se definenen los operadores de programación dinámica análogos a (5.4) y (5.5), respectivamente, como

$$\widetilde{T}_f^\epsilon u(x) := \widetilde{C}(x, f) + \int_{\mathbf{X}_\epsilon\times\mathbb{R}_+} e^{-\alpha t}u(y)\widetilde{Q}(dy, dt|x, f), \tag{5.13}$$

y

$$\widetilde{T}_\epsilon u(x) := \min_{a\in A}\left\{\widetilde{C}(x, a) + \int_{\mathbf{X}_\epsilon\times\mathbb{R}_+} e^{-\alpha t}u(y)\widetilde{Q}(dy, dt|x, a)\right\}. \tag{5.14}$$

Observación 5.3.4 *Observemos que por la Proposición 5.3.3, y por argumentos análogos dados en la Observación 5.2.5 y en la demostración de la Proposición 5.2.6 se cumple lo siguiente:*

(a) Los operadores $\widetilde{T}_f^\epsilon$ y $\widetilde{T}_\epsilon$ son operadores de contracción en el espacio de Banach $(B_W(\mathbf{X}_\epsilon), ||\cdot||_W^\epsilon)$ con módulo $\widetilde{\beta}$.

(b) Para cada función $u(\cdot)$ en $B_W(\mathbf{X}_\epsilon)$ existe un selector medible $f \in \mathbb{F}$ tal que

$$\widetilde{T}_\epsilon u(x) = \widetilde{T}_f^\epsilon u(x).$$

(c) $\widetilde{V}_f^\epsilon$ es el único punto fijo del operador $\widetilde{T}_f^\epsilon$, esto es,

$$\widetilde{V}_f^\epsilon = \widetilde{T}_f^\epsilon \widetilde{V}_f^\epsilon.$$

Como consecuencia de la observación anterior y por argumentos análogos a los de la demostración de la Proposición 5.2.7 tenemos el siguiente resultado.

Proposición 5.3.5 *Supongamos que se cumple la Hipótesis 4.3.1 y que L es un promediador. Entonces la función de valor óptimo $\widetilde{V}_*^\epsilon(\cdot)$ es el único punto fijo de $\widetilde{T}_\epsilon$ en $B_W(\mathbf{X}_\epsilon)$ y existe $f^* \in \mathbb{F}$ tal que*

$$\begin{aligned}\widetilde{V}_*^\epsilon(x) &= \widetilde{T}_\epsilon \widetilde{V}_*^\epsilon(x) = \widetilde{T}_{f^*}^\epsilon \widetilde{V}_*^\epsilon(x) \\ &= \widetilde{C}(x, f^*) + \int_{\mathbf{X}_\epsilon \times \mathbb{R}_+} e^{-\alpha t} \widetilde{V}_*^\epsilon(y) \widetilde{Q}(dy, dt|x, f^*).\end{aligned}$$

Además, una política estacionaria $f^* \in \mathbb{F}$ óptima si y sólo si

$$\widetilde{T}_\epsilon \widetilde{V}_*^\epsilon(\cdot) = \widetilde{T}_{f^*}^\epsilon \widetilde{V}_*^\epsilon(\cdot).$$

Por lo tanto, f^* es óptima.

Con el fin de proporcionar cotas de error por aproximación, introducimos las siguientes constantes:

$$\gamma_C := \sup_{(x,a)\in\mathbb{K}_\epsilon} \frac{1}{W(x)} \left| \widetilde{C}(x,a) - C(x,a) \right| \tag{5.15}$$

y

$$\gamma_Q := \sup_{(x,a)\in\mathbb{K}_\epsilon} \frac{1}{W(x)} \|Q(\cdot,\cdot|x,a) - \widetilde{Q}(\cdot,\cdot|x,a)\|_W \tag{5.16}$$

donde

$$\begin{aligned} & \|Q(\cdot,\cdot|x,a) - \widetilde{Q}(\cdot,\cdot|x,a)\|_W \\ :\ &= \sup_{\|u\|_W \le 1} \left| \int_{\mathbf{X}_\epsilon \times \mathbb{R}_+} e^{-\alpha t} u(y) [Q(dy,dt|x,a) - \widetilde{Q}(dy,dt|x,a)] \right| \end{aligned}$$

Notemos que γ_C y γ_Q miden la precisión de la aproximación del operador L en el modelo truncado. Los siguientes resultados nos proveen de cotas para el error de aproximación en términos de γ_C y γ_Q. Observemos que se cumplen resultados de aproximación análoagos a los obtenidos en las Proposiciones 5.2.3 y 5.2.4. El objetivo de esta sección es encontrar cotas de error por aproximar V_*^ϵ mediante la función $\widetilde{V}_*^\epsilon$.

Teorema 5.3.6 *Supongamos que se cumplen las Hipótesis 4.3.1, 5.2.1 y 5.3.1. Entonces*

$$\|\widetilde{V}_*^\epsilon - V_*^\epsilon\|_W^\epsilon \le \frac{1}{1-\widetilde{\beta}} \gamma_C + \frac{\overline{c}}{(1-\beta)(1-\widetilde{\beta})} \gamma_Q$$

Demostración. Observe que $\widetilde{V}_*^\epsilon(x)$ y $V_*^\epsilon(x)$ son puntos fijos de los operadores $\widetilde{T}_\epsilon$, T_ϵ, respectivamente. Entonces para cada $x \in \mathbf{X}_\epsilon$ se verifica que

$$\begin{aligned} \left|\widetilde{V}_*^\epsilon(x) - V_*^\epsilon(x)\right| &= \left|\widetilde{T}_\epsilon \widetilde{V}_*^\epsilon(x) - T_\epsilon V_*^\epsilon(x)\right| \\ &\le \sup_{a \in \mathbf{A}} \Big\{ \left|\widetilde{C}(x,a) - C(x,a)\right| \\ &+ \left| \int_{\mathbf{X}_\epsilon \times \mathbb{R}_+} e^{-\alpha t} (\widetilde{V}_*^\epsilon - V_*^\epsilon)(y) \widetilde{Q}(dy,dt|x,a) \right| \\ &+ \left| \int_{\mathbf{X}_\epsilon \times \mathbb{R}_+} e^{-\alpha t} V_*^\epsilon(y) (\widetilde{Q}(dy,dt|x,a) - Q(dy,dt|x,a) \right| \Big\} \\ &\le \sup_{a \in \mathbf{A}} \Big\{ \left|\widetilde{C}(x,a) - C(x,a)\right| \\ &+ \|\widetilde{V}_*^\epsilon - V_*^\epsilon\|_W^\epsilon \int_{\mathbf{X}_\epsilon \times \mathbb{R}_+} e^{-\alpha t} W(y) \widetilde{Q}(dy,dt|x,a) \\ &+ \|V_*^\epsilon\|_W^\epsilon \int_{\mathbf{X}_\epsilon \times \mathbb{R}_+} e^{-\alpha t} \frac{V_*^\epsilon(y)}{\|V_*^\epsilon\|_W^\epsilon} (\widetilde{Q}(dy,dt|x,a) - Q(dy,dt|x,a)) \Big\} \\ &\le \sup_{a \in \mathbf{A}} \Big\{ \left|\widetilde{C}(x,a) - C(x,a)\right| + \|\widetilde{V}_*^\epsilon - V_*^\epsilon\|_W^\epsilon \widetilde{\beta} W(x) \\ &+ \|V_*^\epsilon\|_W^\epsilon \|Q(\cdot,\cdot|x,a) - \widetilde{Q}(\cdot,\cdot|x,a)\|_W \Big\} \end{aligned}$$

Entonces,

$$\frac{\left|\widetilde{V}_*^\epsilon(x) - V_*^\epsilon(x)\right|}{W(x)} \leq \sup_{a\in\mathbf{A}}\left\{\frac{\left|\widetilde{C}(x,a) - C(x,a)\right|}{W(x)} + \|\widetilde{V}_*^\epsilon - V_*^\epsilon\|_W^\epsilon\widetilde{\beta} + \|V_*^\epsilon\|_W^\epsilon\frac{1}{W(x)}\|Q(\cdot,\cdot|x,a) - \widetilde{Q}(\cdot,\cdot|x,a)\|_W\right\}.$$

de donde se sigue

$$\|\widetilde{V}_*^\epsilon - V_*^\epsilon\|_W^\epsilon \leq \gamma_C + \|\widetilde{V}_*^\epsilon - V_*^\epsilon\|_W^\epsilon\widetilde{\beta} + \|V_*^\epsilon\|_W^\epsilon\gamma_Q.$$

Por lo tanto,

$$\|\widetilde{V}_*^\epsilon - V_*^\epsilon\|_W^\epsilon \leq \frac{1}{1-\widetilde{\beta}}\gamma_C + \frac{\|V_*^\epsilon\|_W^\epsilon}{1-\widetilde{\beta}}\gamma_Q.$$

Notemos que de Observación 4.3.3(d), la función V_*^ϵ satisface

$$\|V_*^\epsilon\|_W^\epsilon \leq \frac{\overline{c}}{(1-\beta)}.$$

Combinando las desigualdades anteriores se obtiene

$$\|\widetilde{V}_*^\epsilon - V_*^\epsilon\|_W^\epsilon \leq \frac{1}{1-\widetilde{\beta}}\gamma_C + \frac{\overline{c}}{(1-\beta)(1-\widetilde{\beta})}\gamma_Q.\blacksquare$$

Observación 5.3.7 *Una consecuencia inmediata de los Teoremas 5.3.6 y 5.2.4 es la siguiente:*

$$\begin{aligned}\|V_* - \widetilde{V}_*^\epsilon\|_W^\epsilon &\leq \|V_* - V_*^\epsilon\|_W^\epsilon + \|V_*^\epsilon - \widetilde{V}_*^\epsilon\|_W^\epsilon \\ &\leq \frac{\overline{c}}{(1-\beta)^2}\epsilon + \frac{1}{1-\widetilde{\beta}}\gamma_C + \frac{\overline{c}}{(1-\beta)(1-\widetilde{\beta})}\gamma_Q.\end{aligned}$$

5.4. Aproximación de políticas óptimas

En esta sección introduciremos el algoritmo de iteración de políticas para aproximar a $\widetilde{V}_*^\epsilon$, pero recuerde que el objetivo principal es aproximar la política óptima del modelo semi-markoviano original mediante la política óptima aproximada obtenida en la etapa de mejoramiento del algoritmo antes mencionado. Observemos que el Teorema 5.4.2 nos proporciona cotas

de error por operar el sistema bajo las políticas "buenas" obtenidas de esta manera.

Algoritmo de iteración de políticas aproximado-truncado
(i) Inicio: sea $g_0 \in \mathbb{F}$ arbitraria y sea $n = 0$
(ii) Evaluación : dado $g_n \in \mathbb{F}$ calcule $u_n(\cdot) := \widetilde{V}^{\epsilon}_{g_n}(\cdot)$
(iii) Mejoramiento : encuentre $g_{n+1} \in \mathbb{F}$ tal que $\widetilde{T}_{\epsilon}u_n(\cdot) = \widetilde{T}^{\epsilon}_{g_{n+1}}u_n(\cdot)$ y regrese al paso (ii).

La existencia de las políticas en la etapa de mejoramieto se sigue de la Observación 5.3.4(b). Además la demostración del siguiente resultado se sigue de (5.12) y aplicando argumentos análogos a la demostración de la Proposición 2.4.1, por lo que la omitimos .

Proposición 5.4.1 *Supongamos que se satisfacen las Hipótesis 4.3.1 y 5.2.1 y L es un promediador. Entonces, la sucesión $\{u_n(\cdot)\} \in B_W(\mathbf{X}_\epsilon)$ converge decrecientemente y uniformemente en la norma $\|\cdot\|^{\epsilon}_W$ a la función de valor óptimo $\widetilde{V}^{\epsilon}_*(\cdot)$; además, para todo $n \in \mathbb{N}$, se cumple que*

$$\|\widetilde{V}^{\epsilon}_* - u_n\|^{\epsilon}_W \leq \frac{2\widetilde{\beta}}{1-\widetilde{\beta}}\|u_n - u_{n-1}\|^{\epsilon}_W.$$

Finalmente, el resultado principal lo podemos enunciar de la siguiente manera.

Teorema 5.4.2 *Sea $u_n(\cdot) \in B_W(\mathbf{X}_\epsilon)$, $g_n \in \mathbb{F}$, y $n \in \mathbb{N}$, las funciones y políticas definidas en el algoritmo de iteración de políticas aproximado-truncado. Entonces, bajo las Hipótesis 4.3.1 y 5.2.1, se cumple*

$$\begin{aligned}\|V_* - V_{g_n}\|^{\epsilon}_W \leq{}& \frac{2\overline{c}}{(1-\beta)^2}\epsilon \\ &+ \frac{1}{1-\widetilde{\beta}}\gamma_C + \frac{\overline{c}}{(1-\beta)(1-\widetilde{\beta})}\gamma_Q \\ &+ \frac{2\widetilde{\beta}}{1-\widetilde{\beta}}\|u_n - u_{n-1}\|^{\epsilon}_W.\end{aligned}$$

Observemos que la cota del teorema anterior incluye tres fuentes de error: el primer sumando del lado derecho de la desigualdad es un error por truncar el espacio de estados, mientras que el segundo y tercer sumando son errores

por aproximar con el promediador L, y el último sumando es el error por detener el algoritmo de iteración de políticas aproximado-truncado.

Demostración. Para cada $f \in \mathbb{F}$, tenemos que

$$\|V_* - V_f\|_W^\epsilon \leq \|V_* - \widetilde{V}_*^\epsilon\|_W^\epsilon + \|V_f^\epsilon - V_f\|_W^\epsilon + \|\widetilde{V}_*^\epsilon - V_f^\epsilon\|_W^\epsilon \tag{5.17}$$

Note que el primer y segundo término del lado derecho de la desigualdad estan acotados por las cotas proporcionadas en las Proposiciones 5.2.4 y 5.3.7, respectivamenete, mientras que para el tercer sumando tenemos

$$\|\widetilde{V}_*^\epsilon - V_f^\epsilon\|_W^\epsilon \leq \|\widetilde{V}_*^\epsilon - \widetilde{V}_f^\epsilon\|_W^\epsilon + \|\widetilde{V}_f^\epsilon - V_f^\epsilon\|_W^\epsilon. \tag{5.18}$$

Como las funciónes $\widetilde{V}_f^\epsilon$ y V_f^ϵ son puntos fijos de los operadores $\widetilde{T}_f^\epsilon$, y T_f^ϵ, respectivamente, por argumentos análogos a la demostración del Teorema 5.3.6, el segundo sumando del lado derecho de (5.18) es acotado por

$$\|\widetilde{V}_f^\epsilon - V_f^\epsilon\|_W^\epsilon \leq \frac{1}{1-\widetilde{\beta}}\gamma_C + \frac{\overline{c}}{(1-\beta)(1-\widetilde{\beta})}\gamma_Q. \tag{5.19}$$

Ahora tomemos $f := g_n$ y $u_n = \widetilde{V}_f^\epsilon$. Por la Proposición 5.4.1, para el primer sumando del lado derecho de (5.18) tenemos

$$\|\widetilde{V}_*^\epsilon - \widetilde{V}_{g_n}^\epsilon\|_W^\epsilon = \|\widetilde{V}_*^\epsilon - u_n\|_W^\epsilon \leq \frac{2\widetilde{\beta}}{1-\widetilde{\beta}}\|u_n - u_{n-1}\|_W^\epsilon.$$

Así, de (5.19), (5.18) y la última desigualdad

$$\|\widetilde{V}_*^\epsilon - V_f^\epsilon\|_W^\epsilon \leq \frac{1}{1-\widetilde{\beta}}\gamma_C + \frac{\overline{c}}{(1-\beta)(1-\widetilde{\beta})}\gamma_Q + \frac{2\widetilde{\beta}}{1-\widetilde{\beta}}\|u_k - u_{k-1}\|_W^\epsilon. \tag{5.20}$$

Por lo tanto, combinando (5.17), (5.20), y las Proposiciones 5.2.4 y 5.3.7 obtenemos

$$\begin{aligned}\|V_* - V_f\|_W^\epsilon \leq{}& \frac{2\overline{c}}{(1-\beta)^2}\epsilon \\ &+ \frac{1}{1-\widetilde{\beta}}\gamma_C + \frac{\overline{c}}{(1-\beta)(1-\widetilde{\beta})}\gamma_Q \\ &+ \frac{2\widetilde{\beta}}{1-\widetilde{\beta}}\|u_k - u_{k-1}\|_W^\epsilon \qquad \blacksquare\end{aligned}$$

5.5. Ejemplo: aproximación del problema de remplazo

Con el fin de ilustrar los resultados de las dos secciones anteriores consideremos de nuevo el ejemplo del modelo de remplazo de la Sección 4.5. Recordemos que el kernel de transición Q está dado por (4.13), donde $r(\cdot)$ es una función del daño acumulado, J la función de distribución del daño y H la función de distribución de los tiempos entre choques. Supondremos que tanto J como H tienen densidad j y h respectivamente, es decir,

$$H(t) = \int_0^t h(s)ds, \quad J(y) = \int_0^y j(s)ds.$$

Sea $\epsilon > 0$ fijo. Definimos $\mathbf{X}_\epsilon = [0, M]$ para alguna $M > 0$ tal que

$$1 - J(M) = 1 - \int_0^M j(s)ds < \epsilon.$$

Entonces, la desigualdad (5.1) se cumple tomando $W(x) = e^{qx}$ y $r(x) = e^{-bx}$ con $b > q > 0$ como se muestra a continuación.

Para cada $x \in \mathbf{X}_\epsilon = [0, M]$, se cumple que

$$\begin{aligned}
&\int_{\mathbf{X}_\epsilon^c \times \mathbb{R}_+} e^{-\alpha t} W(y) Q(dy, dt|x, a) \\
&= \int_0^a e^{-\alpha t} H(dt) \int_M^\infty e^{q(w+x)} e^{-b(w+x)} J(dw) \\
&= \int_0^a e^{-\alpha t} H(dt) \exp(qx) \int_M^\infty e^{-w(b-q)} e^{-bx} J(dw) \\
&\leq e^{qx} \int_M^\infty J(dw) \leq e^{qx}(1 - J(M)) < W(x)\epsilon.
\end{aligned}$$

Consideraremos el proceso semi-markoviano restringido al *espacio de estados* $\mathbf{X}_\epsilon = [0, M]$, y de acciones $\mathbf{A} = A(x) = [\theta_1, \theta_2]$ para todo $x \in \mathbf{X}_\epsilon$.

Para obtener estimaciones de γ_C y γ_Q es necesario imponer condiciones a la densidad $j(\cdot)$ y $r(\cdot)$. Supongamos que $j(\cdot)$ y $r(\cdot)$ son Lipschitz continuas con módulo b_j and b_r, respectivamente, esto es,

$$|j(x) - j(y)| \leq b_j|y - x|, \ |r(x) - r(y)| \leq b_r|y - x| \ \forall x, y \in [0, M].$$

Observemos que se cumple

$$\begin{aligned}\int_{\mathbf{X}_\epsilon\times\mathbb{R}_+} e^{-\alpha t}u(y)Q(dy,dt|x,a) &= \int_0^a e^{-\alpha t}h(t)dt\int_0^{M-x} u(x+w)r(x+w)j(w)dw\\ &+ u(0)\int_0^a e^{-\alpha t}h(t)dt\int_0^{\infty}(1-r(x+w))j(w)dw\\ &+ u(0)\int_a^{\infty} e^{-\alpha t}h(t)dt.\end{aligned}$$

Consideremos como operador de aproximación L al operador de interpolación lineal con N puntos igualmente espaciados $0 = s_1 < s_2 < \cdots < s_N = M$. Entonces para cada función medible u en $\mathbf{X}_\epsilon$, el operador L está definido para $x \in [s_i, s_{i+1}]$ como

$$Lu(x) = \frac{s_{i+1}-x}{s_{i+1}-s_i}u(s_i) + \frac{x-s_i}{s_{i+1}-s_i}u(s_{i+1}),$$

con $x \in [s_i, s_{i+1}]$, $i = 1, 2, \cdots, N-1$.

Luego

$$\widetilde{Q}(\cdot,\cdot|x,a) = b(x)Q(\cdot,\cdot|s_i,a) + \overline{b}(x)Q(\cdot,\cdot|s_{i+1},a)$$

donde

$$b(x) = \frac{s_{i+1}-x}{s_{i+1}-s_i} \quad \text{y} \quad \overline{b}(x) = \frac{x-s_i}{s_{i+1}-s_i},$$

con $x \in [s_i, s_{i+1}]$, $i = 1, 2, \cdots, N-1$.

Ahora notemos que

$$\begin{aligned}&\int_{\mathbf{X}_\epsilon\times\mathbb{R}_+} e^{-\alpha t}u(y)[\widetilde{Q}(dy,dt|x,a) - Q(dy,dt|x,a)]\\ &= b(x)\{\int_{\mathbf{X}_\epsilon\times\mathbb{R}_+} e^{-\alpha t}u(y)Q(dy,dt|s_i,a) - \int_{\mathbf{X}_\epsilon\times\mathbb{R}_+} e^{-\alpha t}u(y)Q(dy,dt|x,a)\}\\ &+ \overline{b}(x)\{\int_{\mathbf{X}_\epsilon\times\mathbb{R}_+} e^{-\alpha t}u(y)Q(dy,dt|s_{i+1},a) - \int_{\mathbf{X}_\epsilon\times\mathbb{R}_+} e^{-\alpha t}u(y)Q(dy,dt|x,a)\}.\end{aligned}$$

Después de algunos cálculos y observando que $W(y)r(y) = e^{y(q-b)} < 1$, $\|u\|_W \leq 1$, $u(0) \leq W(0) = 1$, obtenemos

$$\gamma_Q \leq 2(|j| + Mb_j + b_r)\Delta s$$

y

$$\gamma_C \leq 2(K_1 b_r + K_3)\Delta s,$$

donde $\Delta s := \frac{M}{N}$

Combinando estos resultados con el Teorema 5.4.2, obtenemos las si guientes cotas

$$\|V_* - V_{g_n}\|_W^\epsilon \leq e_1 + e_2 + e_3$$

donde

$$\begin{aligned} e_1 &: = \frac{2\overline{c}}{(1-\beta)^2}\epsilon, \\ e_2 &: = \frac{1}{1-\widetilde{\beta}}\gamma_C + \frac{\overline{c}}{(1-\beta)(1-\widetilde{\beta})}\gamma_Q, \\ e_3 &: = \frac{2\widetilde{\beta}}{1-\widetilde{\beta}}\|u_n - u_{n-1}\|_W^\epsilon. \end{aligned}$$

Observemos que el error por truncamiento e_1 se puede hacer arbitrariamente pequeño tomando la constante M suficientemente grande. Por otra parte, e_2 es el error por aproximación por usar el promediador L. Por último, e_3 es el error por detener el algoritmo de iteración de políticas.

Resultados numéricos. Para ilustrar los resultados anteriores, suponemos que $j(\cdot)$ es una densidad Weibull$(4, 5)$ y $h(\cdot)$ una densidad Gamma$(5, 2)$. Tomemos el costo de remplazo por falla $K_1 = 15$, el costo de remplazo programado $K_2 = 10$ y costo por unidad de daño acumulado $K_3 = 1$.

Entonces, $\overline{c} = 26$, $W(x) = e^{qx}$ y $r(x) = e^{-bx}$, con $q = 0.03$, $b = 0.05$. En este caso, para con $\alpha = 0.2$, obtenemos $\beta = 0.96$, $b_r = 0.2$, $b_j = 0.42$, y $|j| = 0.45$.

Observemos de la Tabla 5 como se reduce el error por truncamiento eligiendo ϵ lo suficientemente pequeño. En particular, para $\epsilon = 10^{-8}$ obtenemos $\mathbf{X}_\epsilon = [0, 10.358]$, de tal manera que si $M > 10.358$, obtenemos un error $e_1 < 0.000325$.

Por otra parte de la Tabla 6 observamos que tomando $N = 200$ en el operador de aproximación lineal, $\|\widetilde{W}\|_W^\epsilon$ es muy cercano a $\|W\|_W^\epsilon = 1$, lo que significa que β y $\widetilde{\beta}$ son prácticamente iguales.

La Tabla 7 muestra la convergencia del algoritmo de iteración de políticas. En este caso con cinco iteraciones tenemos que el error e_3 es casi cero.

Finalmente la Tabla 8 muestra cotas de error para γ_C y γ_Q. Las cotas de error para $\|V_* - V_f\|_W^\epsilon$ y e_2 son prácticamente las mismas. Notemos e_1 y e_2 son muy sensibles al factor $\beta = 0.96$. Este se puede disminuir tomando un valor de α más grande.

Tabla 5

ϵ	0.0001	0.000005	$1e-8$
e_1	3.25	0.1625	0.000325

Tabla 6

N	100	200	300
$\|\widetilde{W}\|_w^\epsilon$	1.000018	1.000001	1.000001
$\widetilde{\beta}$	0.9600176	0.9600011	0.9600011

Tabla 7

N	100	200	300
$\|u_5 - u_4\|_W$	$3.933603e-15$	$2.242348e-15$	$2.065631e-14$
e_3	$1.888216e-13$	$1.076358e-13$	$9.91515790e-13$

Tabla 8

N	100	500	5000	10000	100000
γ_C	0.35	0.07	0.007	0.0035	0.00035
γ_Q	1.68	0.336	0.033	0.0168	0.00168
$\|V_* - V_f\|_W$	27308	5461	546	260	1.66

Capítulo 6

Comentarios finales y problemas abiertos.

En este trabajo estudiamos el algoritmo de iteración de políticas aproximado (IPA) para aproximar el problema de control óptimo descontado en procesos de decisión markovianos y semi-markovianos con espacios de Borel. El análisis de dicho algoritmo se restringió a una clase de operadores de aproximación denominados *promediadores* (Definición 2.2.1, p. 9). Esta clase de operadores es bastante amplia e incluye a los interpoladores lineales y multilineales, aproximadores constantes por pedazos, splines de Schoenberg, operadores de Bernstein y de Hermite-Féjer, entre otros.

Cuando se usan promediadores, el paso de aproximación en el algoritmo IPA representa una *perturbación* del modelo de decisión original y el algoritmo IPA resulta ser el algoritmo de iteración de valores *exacto* en el modelo perturbado. Este hecho permite probar de forma directa la convergencia del algoritmo IPA bajo condiciones muy generales sobre el modelo original así como dar cotas de error para las soluciones proporcionadas por el algoritmo. Es importante recalcar que dichas cotas son *generales* y que están expresadas en términos de la precisión con la que el promediador aproxima la función de costo por etapa y la ley de evolución del sistema.

En la primera parte del trabajo se estudió el algoritmo IPA para procesos de decisión markovianos con espacios de Borel y *costos acotados*. Para estos procesos se consideraron los dos siguientes casos: (i) la ley de evolución del sistema está dada por un kérnel estocástico, el cual se supone es completamente conocido; (ii) el sistema evoluciona de acuerdo a una ecuación en diferencias bajo el supuesto de que el ruido tiene densidad pero es desconocida. En este último caso, primero se obtiene una estimación de la densidad

desconocida y posteriormente se efectúa el paso de aproximación usando un promediador. Naturalmente, las cotas de desempeño del algoritmo también dependen del error producido por el paso de estimación.

En la segunda parte del trabajo se estudia el problema de control óptimo descontado para procesos de decisión semi-markovianos con espacios de Borel y costos no-acotados. Primero se estudia el problema de control óptimo estándar y se prueba la existencia de políticas estacionarias óptimas, entre otros resultados, bajo condiciones usuales de continuidad-compacidad y una nueva condición de Lyapunov. Posteriormente, suponiendo adicionalmente que el espacio de estados es un espacio de Borel *localmente compacto*, se proporcionan cotas para el error producido por *truncamiento* del espacio de estados. Una vez hecho esto, se estudia el algoritmo IPA para el proceso de decisión semi-markoviano con espacio de estados truncado. Se proporcionan cotas para el desempeño del algortimo IPA en términos de los errores de aproximación producidos por el promediador y por el paso de truncamiento.

Los resultados descritos en la sección anterior, así como los reportados en [40] y [41], muestran que el enfoque de perturbaciones basadas en promediadores es muy flexible y ofrece la posibilidad de estudiar exitosamente una variedad de problemas de control para procesos markovianos y semi-markovianos con costo descontado y costo promedio. Por ejemplo, en nuestra opinión sería interesante estudiar los siguientes problemas.

(1) Convergencia del algoritmo de iteración de políticas aproximado para el índice en costo promedio.

(2) Combinar el enfoque de perturbaciones con el enfoque de programación lineal para obtener aproximaciones al problema de control óptimo.

(3) Estudiar problemas donde la dinámica del sistema sea desconocida. El problema sería combinar el enfoque de perturbaciones, iteración de valores o iteración de políticas, con diferentes esquemas de estimación como un enfoque de verosimilitud, bayesiano o el de la distribución empírica.

(4) Extender los resultados del presente trabajo a juegos estocásticos.

El esquema de aproximación que hemos propuesto es muy general y se puede adaptar casi a cualquier contexto. Creemos que la combinación de los problemas propuestos puede dar lugar a nuevos problemas abiertos.

Bibliografía

[1] Robert B. Ash, *Real Analysis and Probability*, Academic Press, New York, 1972.

[2] A. Almudevar, Approximate fixed point iteration with an application to infinite horizon Markov decision processes, *SIAM Journal on Control and Optimization* 46 (2008), 541-561.

[3] A. Antos, C. Szepesvari, R. Munos, Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path, *Machine Learning* 71 (2008), 89-129.

[4] D. P. Bersetkas, *Approximate policy iteration: a survey and some new methods*, Journal of Control Theory and Applications 9, 2011, 310-335.

[5] D. P. Bertsekas *and* J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1995.

[6] D. P. Bertsekas *and* S.E. Sherve, *Stochastic Optimal Control: The discrete Time Case*, Academic Press, New York, 1978.

[7] Rabi N. Bhattacharya *and* M. Majumdar, *Controlled semi-Markov models-the discounted case,* Journal of Statistical Planning and Inference, 21, 1989, 365-381.

[8] W. L. Cooper, S. G. Henderson *and* M. E. Lewis, *Convergence of simulation-based policy iteration*, Probability in the Engineering and Informational Sciences 17, 2003, 213-234.

[9] L. Devroye *and* L. Györfi, *Nonparametric Density Estimation. The L_1 View*, John Wiley & Sons, New York, 1985.

[10] L. Devroye *and* G. Lugosi, *Combinatorial Methods in Density Estimation*, Springer, New York, 2001.

[11] Pierre L. Ecuyer, *Computing Approximate Solutions to Markov Renewal Programs with Continuous State Spaces*, Draft Version, 1989.

[12] A. Farahmand, M. Ghavamzadeh, C. Szepesvari, S. Mannor, Regularized policy iteration, *Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, 2008, 441-448.

[13] E.I. Gordienko *and* J.A. Minjárez-Sosa, *Adaptive control for discrete-time Markov processes with unbounded costs: discounted criterion*, Kybernetika 34,1998, 217–234.

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

[15] O. Hernández-Lerma *and* J.B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.

[16] O. Hernández-Lerma *and* J.B. Lasserre, *Farther Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.

[17] O. Hernández-Lerma *and* W. Runggaldier, *Monotone approximations for convex stochastic contro problems*, J.Math. Syst., Estimation, and Control 4,1, 1993.

[18] N. Hilgert *and* J.A. Minjárez-Sosa, *Adaptive policies for time-varying stochastic systems under discounted criterion*, Math. Methods Oper. Res. 54, 2001, 491-505.

[19] N. Hilgert *and* J. A. Minjárez-Sosa, *Adaptive control of stochastic systems with unknown disturbance distribution: discounted criteria*, Math. Methods Oper. Res. 63, 2006, 443-460.

[20] M. Kurano, *Semi-Markov Decision Processes and their Applications in Replacement Models*, J. Operations Research Society of Japan, Vol. 28, No. 1, 1985.

[21] Fernando Luque Vásquez, *Modelos de Control Semi-Markoviano en Espacios de Borel,* Tesis Doctoral en Ciencias (Matemáticas), Facultad de Ciencias UNAM, 1997.

[22] J. Ma and W. B. Powell, A convergent recursive least squares approximate policy iteration algorithm for multi-dimensional Markov decision process with continuous state and action spaces, *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, New York, 2009, 66-73.

[23] J. Ma, W. B. Powell, Convergence Analysis of Kernel-based On-policy Approximate Policy Iteration Algorithms for Markov Decision Processes with Continuous Multidimensional States and Actions, Department of Operations Research and Financial Engineering, Princeton University, 2010.

[24] R. Munos, *Performance bounds in L_p-norm for approximate value iteration*, SIAM Journal on Control and Optimization 47, 2007, 2303-2347.

[25] R. Munos, *Error Bounds for Approximate Policy Iteration*, Proccedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[26] J. A. Minjárez-Sosa, *Approximation and estimation in Markov control processes under a discounted criterion*, Kybernetika 40, 2004, 681-690.

[27] W. B. Powell, *Approximate Dynamic Programming. Solving the Curse of Dimensionality* John Wiley & Sons Inc, 2007.

[28] W. B. Powell, Perspectives of approximate dynamic programming, *Ann. Oper. Res.* (2012), 1-38. DOI: 10.1007/s10479-012-1077-6.

[29] W. B. Powell and J. Ma, A review of stochastic algorithms with continuous value function approximation and some new approximate policy iteration algorithms for multidimensional continuous applications, *Journal of Control Theory and Applications* 9 (2011), 336–352.

[30] M.L.Puterman, *Markov Decision Processes.Discrete Stochastic Dynamic Programming*, Wiley, N.Y., 1994.

[31] U. Rieder, *Measurable selection theorems for optimization problems*, Manuscripta Math. 24, 1978, 115–131.

[32] S.M. Ross, *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1970.

[33] J. Rust, *Numerical dynamic programming in economics*, in: Handbook of Computational Economics, vol. 1, H. M. Amman, D. A. Kendrick, J. Rust, J. eds., Elsevier, 1996, 619-728.

[34] J. Rust, *A comparison of policy iteration methods for solving continuous-state, infinite-horizon markovian decision problems using random, quasi-random, and deterministic discretization*, Available from the Economics Working Paper archive: http://econwpa.wustl.edu/eprints/comp/papers/9704/9704001.abs

[35] H.L. Royden, *Real Analysis*, Macmillan Publishing Company, N. Y., 1989.

[36] M. S. Santos *and* J. Rust, *Convergence properties of policy iteration*, SIAM J. Control and Optim. 42, 2004, 2094-2115.

[37] Robert J. Serfling, *Approximation Theorems of Mathematical Statistics*, John-Wiley, 2002.

[38] J. Stachurski, *Continuous state dynamic programming via nonexpansive approximation*, Computational Economics 31, 2008, 141-160.

[39] O. Vega-Amaya *and* R. Montes-de-Oca, *Application of average dynamic programming to inventory systems*, Math. Methods Oper. Res., 47, 1998, 451-471.

[40] O. Vega-Amaya *and* J. López-Borbón, *A performance bound for discounted approximate dynamic programming using averagers*. Reporte Interno, Departamento de Matemáticas, Universidad de Sonora, 2013.

[41] O. Vega-Amaya *and* J. López-Borbón, *A performance Approach to Approximate Value Iteration for Average Cost Markov Decision Process with Borel Spaces and Bounded Costs*. Reporte Interno, Departamento de Matemáticas, Universidad de Sonora, 2013.

[42] Qingda Wei and Xianping Guo, *Semi-Markov Decision processes with Variance Minimization Criterion*, 4OR, Vol.13, Issue1, 2015, 59-79.

[43] X. Xu, D. Hu and X. Lu, Kernel-based least squares policy iteration for reinforcement learning, *IEEE Transactions on Neural Networks* 18 (2007), 973-992.