Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Matemáticas

Fundamentos matemáticos del aprendizaje por refuerzo

Autor: Manuel García Madrid

Tutor: Eustasio del Barrio Tellado

2024/25

Resumen

Este trabajo presenta los fundamentos matemáticos del aprendizaje por refuerzo, trabajando previamente el problema del bandido multibrazo. El enfoque principal es el estudio de los procesos de decisión de Markov y de su control estocástico. Se introducen las ecuaciones de Bellman y su optimización en MDPs, lo que permite el desarrollo de métodos tabulares como la iteración de valor y la iteración de política. Además, se comprueba su convergencia y se incluyen implementaciones prácticas de algoritmos en problemas discretos. Estos conceptos proporcionan una base teórica sólida para la planificación y el control en entornos de decisión secuenciales.

Palabras clave: aprendizaje por refuerzo, bandido multibrazo, cota inferior de Lai-Robbins, proceso de decisión de Markov, ecuaciones de Bellman; operadores de optimalidad de Bellman; algoritmo de programación dinámico, iteración de política.

Abstract

This work presents the mathematical foundations of reinforcement learning, first addressing the multi-armed bandit problem. The main focus is the study of Markov decision processes and their stochastic control. Bellman equations and their optimization in MDPs are introduced, enabling the development of tabular methods such as value iteration and policy iteration. Additionally, their convergence is analyzed, and practical algorithm implementations are included for discrete problems. These concepts provide a solid theoretical foundation for planning and control in sequential decision-making environments.

Key words: reinceforcement learning, multiarmed bandit, Lai-Robbins lower bound, Markov decision process, Bellman equations, Bellman optimality operators, dynamic programming algorithm, policy iteration.

A mi familia, por creer en mí en los momentos más difíciles, por apoyarme siempre y por haberme convertido en la persona que soy hoy en día.

A María, por su paciencia, apoyo incondicional y confianza en mí durante todo este proceso. Gracias por estar siempre a mi lado haciendo que todo sea más fácil y bonito.

A mis amigos del colegio y de Macotera, por haber sido parte de mi vida desde el principio, por haberme aportado momentos tan enriquecedores y por haberme enseñado lo que es la amistad.

A mis amigos del grado, por acompañarme en esta etapa tan señalada de nuestras vidas, por apoyarnos día a día durante estos últimos años y por haberme ayudado a hacer que todo pareciera más sencillo.

A mi tutor Tasio, por aceptar dirigirme este trabajo, por su dedicación y atención y por haberme introducido en una materia que me fascina y no conocía.

Índice general

Introducción

El aprendizaje por refuerzo es un enfoque de aprendizaje automático, cuyo objetivo es permitir que un agente aprenda a tomar decisiones óptimas a partir de la interacción directa con su entorno. Este enfoque se ejemplifica con el juego del tres en raya. Existen distintos métodos clásicos para resolver este problema, pero la mayoría consideran el resultado final de cada partida y no aprovechan la información obtenida durante el juego. El aprendizaje por refuerzo propone un algoritmo basado en funciones de valor que permite estimar la probabilidad de ganar desde cada estado del juego. Se construye una tabla donde cada estado del tablero tiene un valor asociado. Inicialmente, los estados ganadores tienen un valor de 1, los perdedores un valor de 0, y el resto un valor intermedio, como 0.5, reflejando una incertidumbre inicial. El aprendizaje ocurre a través de múltiples partidas contra un oponente imperfecto. Durante el juego, el agente selecciona movimientos observando los valores de los estados resultantes. Normalmente, elige el movimiento con mayor valor (estrategia codiciosa), pero ocasionalmente realiza movimientos exploratorios para descubrir nuevas estrategias. Además, se actualizan los valores de los estados durante el juego. Este método tiene ventajas sobre los enfoques clásicos, ya que permite mejorar continuamente la estrategia a partir de la información adquirida en cada jugada, en lugar de esperar múltiples partidas para evaluar una política completa. También es más eficiente, ya que evita dar crédito a movimientos que no ocurrieron realmente.

El aprendizaje por refuerzo no se limita a juegos con episodios discretos y recompensas finales, sino que también puede aplicarse a problemas donde el comportamiento es continuo y las recompensas varían con el tiempo. Con este propósito en mente, el presente trabajo tiene como objetivo exponer sus fundamentos matemáticos desde una perspectiva teórica y computacional. Para ello, se presenta un desarrollo progresivo que proporciona las bases necesa-

rias para la comprensión y aplicación de algoritmos de control en entornos inciertos.

En el primer capítulo se aborda uno de los primeros y más sencillos problemas en este contexto, el del bandido multibrazo. Su nombre proviene de la analogía con una máquina tragaperras de múltiples brazos, donde cada brazo tiene una distribución de recompensa desconocida, y el objetivo del jugador es maximizar la recompensa total mediante la elección óptima de brazos a lo largo del tiempo. En este marco, Herbert Robbins [7] introdujo la idea del equilibrio entre exploración y explotación. La cota logarítmica de la pérdida en el algoritmo *explora-luego-decide* es un importante resultado que respalda teóricamente dicho equilibrio. Asimismo, al final del capítulo se proporciona la cota inferior de Lai-Robbins, resultado válido para todas las estrategias que cumplen una propiedad de consistencia. Este resultado otorga un método capaz de discutir la optimalidad de cualquier estrategia considerada consistente, como se hace para el algoritmo *explora-luego-decide*.

Una vez estudiado el bandido multibrazo, en el segundo capítulo se introducen los procesos de decisión de Markov (MDPs), que son la herramienta idónea para tratar problemas de aprendizaje por refuerzo. Dado un modelo de Markov, la existencia de su correspondiente proceso de decisión de Markov es un resultado central, y permite el desarrollo matemático de una teoría capaz de abarcar numerosos problemas reales. También se introducen las políticas, que son las reglas que guían al agente para tomar decisiones en cada estado del sistema, y se enuncian resultados básicos.

Con estas herramientas, el tercer capítulo trata la optimalidad de políticas en el marco de la teoría de control estocástico. Los fundamentos matemáticos del aprendizaje por refuerzo están íntimamente ligados a las ecuaciones de Bellman, las cuales proporcionan un marco recursivo para el cálculo del valor esperado de una política dada. El capítulo demuestra que las políticas óptimas están asociadas a los operadores de Bellman, que son contracciones. Por tanto, haciendo uso del Teorema del Punto Fijo de Banach, se demuestra la existencia y caracterización de las políticas óptimas. De hecho, la optimización de Bellman en MDPs permite la demostración del teorema del algoritmo de programación dinámica, que respalda la existencia de una política óptima, estacionaria y determinista. Este resultado es fundamental ya que garantiza que se puede encontrar una solución óptima que no varía con el tiempo, y que proporciona una única acción óptima para cada estado del sistema.

Con ayuda de la teoría de control estocástico aplicada a los MDPs, en el cuarto capítulo se describen dos procedimientos para el cálculo de estrategias óptimas: la iteración de valor y la iteración de política. La iteración de valor actualiza iterativamente la función de valor utilizando la ecuación de Bellman hasta la convergencia. Sin embargo, la iteración de política alterna entre evaluar una política fija y mejorarla hasta alcanzar su optimalidad.

La teoría recogida en los cuatro primeros capítulos de este Trabajo de Fin de Grado utiliza como principales referencias tanto el texto de Döring [3], como el de Sutton y Barto [8]. Son dos documentos muy recientes que recogen de manera actualizada la teoría matemática del aprendizaje por refuerzo. En los apéndices finales se recogen resultados utilizados a lo largo del documento que no han sido estudiados exactamente igual en el grado.

Pese a que este trabajo está enfocado en fundamentos matemáticos, en el quinto y último capítulo se pretende hacer una implementación práctica de estos algoritmos para ejemplos muy sencillos. La resolución de una cuadrícula simple mediante el algoritmo de iteración de valor, y de un problema de control de inventario mediante la iteración de política, permiten visualizar cómo los métodos tabulares pueden emplearse en escenarios reales, proporcionando una conexión entre la teoría matemática y su aplicabilidad práctica.

El aprendizaje por refuerzo ha demostrado ser una herramienta poderosa en una amplia variedad de aplicaciones. Desde el control de robots autónomos hasta la gestión eficiente de carteras en finanzas y la mejora de estrategias en videojuegos, los algoritmos de aprendizaje por refuerzo han permitido desarrollar soluciones innovadoras en múltiples campos. Estas aplicaciones reflejan la relevancia práctica del aprendizaje por refuerzo.

Capítulo 1

Problema del bandido multibrazo

En este primer capítulo se expone el problema del bandido multibrazo, y se estudia la base matemática que respalda algunos de sus principales algoritmos, como el algoritmo *explora-luego-decide*. A su vez, se deducen cotas inferiores y superiores que discuten su optimidad. Es un marco clásico de toma de decisiones que representa un punto de partida ideal para introducir los fundamentos del aprendizaje por refuerzo (Reinforcement Learning, RL).

Sin embargo, los resultados teóricos y prácticos logrados en este marco no siempre se traducen directamente al aprendizaje por refuerzo, donde la presencia de estados y transiciones introduce una complejidad significativamente mayor. Este capítulo, por lo tanto, servirá como un primer paso hacia la comprensión de las bases del RL, preparando al lector para abordar los contextos más ricos y dinámicos que se explorarán en capítulos siguientes.

1.1. Contexto

El problema del bandido multibrazo fue establecido por Robbins [7] en 1952 y es un desafío clásico en teoría de decisión y aprendizaje automático. Este consiste en un agente que para cada tiempo t debe seleccionar entre un conjunto de k brazos, cada uno con una recompensa desconocida que varía al

ser jugado. El objetivo es aprender de manera eficiente una estrategia capaz de maximizar la recompensa acumulada.

Mediante una serie de definiciones y resultados, se introducirán a continuación los conceptos básicos alrededor de los que se trabajará en el resto del capítulo.

Definición 1.1. Supongamos que $\mathcal{A} = \{a_1, ..., a_k\}$ es un conjunto finito, cuyos elementos se llamarán brazos, y $v = \{P_a\}_{a\in\mathcal{A}}$ es una familia de distribuciones de probabilidad reales con esperanzas finitas, que se llamarán distribuciones de recompensa. El conjunto v es llamado modelo de bandido estocástico. Se define el valor de acción de un brazo a como $Q_a := \int_{\mathbb{R}} x dP_a(x)$. Se denota a_*, y se llama brazo óptimo, al brazo con mayor valor de acción, es decir,

$$a_* = \text{argmax}_{a\in\mathcal{A}} Q_a.$$

Dado un número natural n, se define una estrategia de aprendizaje para n rondas como un conjunto $(\pi_t)_{t=1,...,n}$ donde:

- π_1 es una distribución inicial sobre $\mathcal{A}$.
- Si $2 \leq t \leq n$, $\pi_t(\cdot; a_1, x_1, ..., a_{t-1}, x_{t-1})$ indica la probabilidad de escoger cada brazo en el momento t conociendo que en el momento i se ha seleccionado el brazo a_i y se ha recibido la recompensa x_i, $\forall i = 1, ..., t-1$.

Definición 1.2. Sea v un modelo de bandido estocástico y $(\pi_t)_{t=1,...,n}$ una estrategia de aprendizaje para n rondas. Un proceso estocástico $(A_t^\pi, X_t^\pi)_{t=1,...,n}$ definido sobre un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ se denomina proceso de bandido estocástico con estrategia de aprendizaje π si se cumplen las siguientes condiciones:

- La acción inicial A_1^π sigue la distribución inicial π_1.
- Para cada $t = 2, \ldots, n$, se cumple que:

  $$\mathbb{P}(A_t^\pi = a \mid A_1^\pi, X_1^\pi, \ldots, A_{t-1}^\pi, X_{t-1}^\pi) = \pi_t(a; A_1^\pi, X_1^\pi, \ldots, A_{t-1}^\pi, X_{t-1}^\pi),$$

  donde π_t determina la probabilidad de elegir la acción a en función del historial observado.
- Para cada $t = 1, \ldots, n$, la recompensa X_t^π satisface:

  $$\mathbb{P}(X_t^\pi \in B \mid A_1^\pi, X_1^\pi, \ldots, A_t^\pi) = P_{A_t^\pi}(B),$$

donde $P_{A_t^\pi}(B) := \sum_{a \in A} P_a(B)\mathbf{1}_{\{A_t^\pi = a\}}$.

Aquí, A_t^π se interpreta como el brazo elegido en el tiempo t, mientras que X_t^π representa la recompensa obtenida al jugar el brazo A_t^π.

Nota. Una construcción sencilla, que se puede consultar en [3], página 8, demuestra que para todo modelo de bandido estocástico y toda estrategia $(\pi_t)_{t=1,\dots,n}$, existe un proceso de bandido estocástico (A_t^π, X_t^π).

Definición 1.3. Dado un modelo de bandido v y una estrategia de aprendizaje $(\pi_t)_{t=1,\dots,n}$, se define la pérdida acumulada de la estrategia π como:

$$R_n(\pi) := nQ_* - \mathbb{E}[\sum_{t=1}^{n} X_t^\pi]. \tag{1.1}$$

Nota. El término de pérdida acumulada de la definición anterior no se refiere a la habitual función de pérdida presente en el contexto de aprendizaje automático. En este caso, se refiere al concepto de *regret* recogido en las referencias correspondientes al problema del bandido multibrazo.

El objetivo de este problema es encontrar de manera óptima una estrategia capaz de minimizar la pérdida acumulada.

Se enuncia y demuestra a continuación un lema que proporciona una expresión de la pérdida que será de gran utilidad a lo largo de todo el capítulo. Para ello, es clave conocer algunas propiedades básicas de la esperanza condicionada, que se recogen en A.10.

Lema 1.4 (descomposición de la pérdida)**.** *Definiendo* $T_a(n) := \sum_{t=1}^{n} \mathbf{1}_{A_t=a}$, *la pérdida de la estrategia* π *tras* n *rondas puede expresarse como*

$$R_n(\pi) = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)],$$

donde las diferencias $\Delta_a := Q_* - Q_a$ *se denominan huecos de recompensa.*

Demostración. Usando en la última igualdad la propiedad (1) de A.10,

$$R_n(\pi) = nQ_* - \mathbb{E}\left(\sum_{t=1}^{n} X_t\right) = \sum_{t=1}^{n} \mathbb{E}(Q_* - X_t) = \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} \mathbb{E}((Q_* - X_t)\mathbf{1}_{\{A_t=a\}})$$

$$= \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} \mathbb{E}\left(\mathbb{E}\left((Q_* - X_t)\mathbf{1}_{\{A_t=a\}} \middle| A_t, X_1, \ldots, A_t\right)\right).$$

Usando la propiedad (3) de A.10 con $g(A_1, X_1, \ldots, A_t) = \mathbf{1}_{\{A_t=a\}}$, se tiene que

$$R_n(\pi) = \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} \mathbb{E}\left(\mathbf{1}_{\{A_t=a\}} \mathbb{E}\left(Q_* - X_t \middle| A_t, X_1, \ldots, A_t\right)\right)$$

$$= \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} \mathbb{E}\left(\mathbf{1}_{\{A_t=a\}} (\mathbb{E}Q_* - \mathbb{E}X_t) \middle| A_t, X_1, \ldots, A_t\right))$$

$$= \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} \mathbb{E}\left(\mathbf{1}_{\{A_t=a\}} (Q_* - Q_a)\right) = \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} \mathbb{E}\left(\mathbf{1}_{\{A_t=a\}} \Delta_a\right),$$

por la linealidad de la esperanza condicionada, y porque si $A_t = a$, X_t queda determinada independientemente de lo ocurrido antes. Por tanto,

$$R_n(\pi) = \sum_{a \in \mathcal{A}} \sum_{t=1}^{n} \Delta_a \, \mathbb{E}(\mathbf{1}_{\{A_t=a\}})$$

$$= \sum_{a \in \mathcal{A}} \Delta_a \, \mathbb{E}\left(\sum_{t=1}^{n} \mathbf{1}_{\{A_t=a\}}\right) = \sum_{a \in \mathcal{A}} \Delta_a \, \mathbb{E}(T_a(n)).$$

□

1.2. Cota superior para el algoritmo *explora-luego-decide*

A la hora de diseñar un algoritmo para el problema del bandido multibrazo, es necesario valorar dos aspectos clave. Es necesario probar diferentes brazos para aprender cuál ofrece las mejores recompensas (lo que se denomina exploración), pero también jugar los brazos ya conocidos para maximizar sus ganancias (explotación). De esta manera, se puede definir un primer algoritmo que es fácil de imaginar, pero efectivo si se utiliza adecuadamente. Para ello, es necesario introducir ciertos conceptos.

Definición 1.5. Si (A_t^π, X_t^π) es un proceso de bandido estocástico para cierta estrategia π, entonces dado $a \in \mathcal{A}$ se define

$$\hat{Q}_a(t) := \frac{1}{T_a(t)} \sum_{j=1}^{t} X_j \mathbf{1}_{\{A_j = a\}} \tag{1.2}$$

como el valor de acción estimado del brazo a hasta el tiempo t.

Nota. La estimación arriba definida es la media de las recompensas que ha aportado el brazo a en las veces que ha sido jugado hasta el momento t.

El algoritmo *explora-luego-decide* para un límite temporal n consiste en jugar cada brazo de manera alternativa m veces, y las $n-km$ veces restantes jugar el brazo con mayor valor de acción estimado. La estrategia π_t puede expresarse como:

$$\pi_t(a; a_1, x_1, \ldots, a_t, x_t) = \begin{cases} 1 & \text{si } a = a_t \mod (k+1) \text{ y } t \leq mk, \\ 1 & \text{si } a = \arg\max_a \hat{Q}_a(mk) \text{ y } t > mk, \\ 0 & \text{en otro caso.} \end{cases}$$

Para poder acotar superiormente la pérdida de este algoritmo, y así valorar su optimalidad, es necesario trabajar previamente con las variables aleatorias subgaussianas y la desigualdad de concentración de Hoeffding, que se presenta a continuación.

Definición 1.6. Una variable aleatoria X en un espacio probabilístico $(\Omega, \mathcal{A}, \mathbb{P})$ es llamada σ-subgaussiana para $\sigma > 0$, si

$$M_{X-\mathbb{E}(X)}(\lambda) = \mathbb{E}(e^{\lambda(X-\mathbb{E}(X))}) \leq e^{\frac{\lambda^2\sigma^2}{2}}, \quad \forall \lambda \in \mathbb{R}.$$

Nota. Intuitivamente, que una variable sea σ-subgaussiana significa que sus colas se dispersan menos de lo que lo haría una normal $N(0, \sigma^2)$.

Proposición 1.7. *Si X es σ-subgaussiana, entonces*

$$\mathbb{P}(X - \mathbb{E}(X) \geq a) \leq e^{-\frac{a^2}{2\sigma^2}} \quad y \quad \mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq 2e^{-\frac{a^2}{2\sigma^2}}$$

para todo $a > 0$.

Demostración. En la demostración se introducirá y se usará el método de Chernoff, una herramienta utilizada para la obtención de desigualdades maximales.

Se define $Z = X - \mathbb{E}X$. Entonces Z es variable aleatoria centrada que cumple que

$$\mathbb{E}(e^{\lambda Z}) \leq e^{\frac{\lambda^2\sigma^2}{2}} \quad \text{con } \lambda \in \mathbb{R}.$$

Se considera $\lambda > 0$. Por la desigualdad de Markov y la monotonía de $f(x) = e^{\lambda x}$,

$$\mathbb{P}(Z \geq a) = \mathbb{P}(e^{\lambda Z} \geq e^{\lambda a}) \leq \frac{\mathbb{E}(e^{\lambda Z})}{e^{\lambda a}} = e^{-(\lambda a - \psi(\lambda))},$$

donde $\psi(\lambda) = \log \mathbb{E}(e^{\lambda Z})$. Se llama $\psi^*(a) := \sup_{\lambda>0}(\lambda a - \psi(\lambda))$ a la función de Chernoff, y de aquí se obtiene que

$$\mathbb{P}(Z \geq a) \leq e^{-\psi^*(a)}, \quad \text{la desigualdad de Chernoff.}$$

Por la condición de σ-subgaussianidad,

$$\psi(\lambda) \leq \frac{\sigma^2\lambda^2}{2},$$

luego

$$\psi^*(a) = \sup_{\lambda>0}(\lambda a - \psi(\lambda)) \geq \sup_{\lambda>0}\left(\lambda a - \frac{\sigma^2\lambda^2}{2}\right) = \sup_{\lambda>0} h(\lambda),$$

donde $h(\lambda)$ es una parábola con coeficiente negativo, por lo que se puede hallar su máximo. Derivando con respecto a λ,

$$h'(\lambda) = a - \lambda\sigma^2 = 0 \implies \lambda = \frac{a}{\sigma^2}.$$

Por tanto, $\psi^*(a) \geq h\left(\frac{a}{\sigma^2}\right) = \frac{a^2}{2\sigma^2}$. Así, se obtiene de la desigualdad de Chernoff que $\mathbb{P}(Z \geq a) \leq e^{-\psi^*(a)} \leq e^{-\frac{a^2}{2\sigma^2}}$.

Como se puede replicar paso a paso lo visto para $-Z$,

$$\mathbb{P}(|Z| \geq a) = \mathbb{P}(Z \geq a \text{ o } -Z \geq a) \leq \mathbb{P}(Z \geq a) + \mathbb{P}(-Z \geq a) \leq 2\, e^{-\frac{a^2}{2\sigma^2}}.$$

□

Lema 1.8.

1. *Toda variable σ-subgaussiana cumple que* $\mathrm{Var}(X) \leq \sigma^2$.
2. *Si X es σ-subgaussiana, entonces cX es $|c|\sigma$-subgaussiana.*
3. *Si X_1 y X_2 son variables aleatorias independientes, X_1 es σ_1-subgaussiana y X_2 es σ_2-subgaussiana, entonces $X_1 + X_2$ es $\sqrt{\sigma_1^2 + \sigma_2^2}$-subgaussiana.*
4. *(Lema de Hoeffding) Si $a \leq X \leq b$, para $a, b \in \mathbb{R}$, entonces X es $\frac{b-a}{2}$-subgaussiana.*

Demostración.

1. Dado $\lambda \in \mathbb{R}$, $M_{X-\mathbb{E}X}(\lambda)$ es la función generadora de momentos de $X - \mathbb{E}X$, y como X es σ-subgaussiana,

$$M_{X-\mathbb{E}(X)}(\lambda) \leq e^{\frac{\lambda^2\sigma^2}{2}}. \tag{1.3}$$

Esta condición afirma que $M_{X-\mathbb{E}(X)}(\lambda)$ está acotada para todo $\lambda \in (-b, b)$ con $b \in \mathbb{R}$. Por tanto f es derivable (C^∞) en $(-b, b)$. Siguiendo la sección 30 de [1] y aplicando el teorema de derivación bajo el signo integral, se obtiene que

$$M'_{X-\mathbb{E}(X)}(\lambda) = \mathbb{E}\left[\frac{d}{d\lambda}e^{\lambda(X-\mathbb{E}(X))}\right] = \mathbb{E}[(X - \mathbb{E}(X))e^{\lambda(X-\mathbb{E}(X))}], \text{ y que}$$

$$\begin{aligned} M''_{X-\mathbb{E}(X)}(\lambda) &= \mathbb{E}\left[\frac{d}{d\lambda}\left((X - \mathbb{E}(X))e^{\lambda(X-\mathbb{E}(X))}\right)\right] \\ &= \mathbb{E}[(X - \mathbb{E}(X))^2 e^{\lambda(X-\mathbb{E}(X))}]. \end{aligned}$$

Por tanto, se pueden expresar ambos lados de (1.3) como polinomio de Taylor en torno a $\lambda = 0$:

$$1 + \lambda(\mathbb{E}(X - \mathbb{E}(X)) + \frac{\lambda^2}{2}\mathbb{E}((X - \mathbb{E}(X))^2) + o(\lambda^2) \leq 1 + \frac{\lambda^2\sigma^2}{2} + o(\lambda^2).$$

Simplificando se obtiene que $Var(X) + 2\,\frac{o(\lambda^2)}{\lambda^2} \leq \sigma^2 + 2\,\frac{o(\lambda^2)}{\lambda^2}$.

Como se debe cumplir esta desigualdad $\forall \lambda \in \mathbb{R}$, si $\lambda \to 0$, se concluye que $\mathrm{Var}(X) \leq \sigma^2$.

2. $$\mathbb{E}\left(e^{\lambda(cX-\mathbb{E}(cX))}\right) = \mathbb{E}\left(e^{c\lambda(X-\mathbb{E}(X)}\right) \leq e^{\frac{(c\lambda)^2\sigma^2}{2}} = e^{\frac{\lambda^2(c\sigma)^2}{2}},$$
lo que muestra que X es $|c|\sigma$-subgaussiana.

3. Utilizando que son independientes,
$$\mathbb{E}\left(e^{\lambda(X_1+X_2+\mathbb{E}(X_1)+\mathbb{E}(X_2))}\right) = \mathbb{E}\left(e^{\lambda(X_1-\mathbb{E}X_1)}e^{\lambda(X_2-\mathbb{E}X_2)}\right)$$
$$= \mathbb{E}\left(e^{\lambda(X_1-\mathbb{E}X_1)}\right)\mathbb{E}\left(e^{\lambda(X_2-\mathbb{E}X_2)}\right) \leq e^{\frac{\lambda^2\sigma_1^2}{2}}e^{\frac{\lambda^2\sigma_2^2}{2}} = e^{\frac{\lambda^2(\sigma_1^2+\sigma_2^2)}{2}},$$
lo que implica que $X_1 + X_2$ sea $\sqrt{\sigma_1^2+\sigma_2^2}$-subgaussiana.

4. Sin pérdida de generalidad, cambiando X por $X - \mathbb{E}X$ si fuera necesario, se puede considerar que $\mathbb{E}X = 0$. Así, la condición del enunciado supondría que $a \leq 0 \leq b$.

   Dado $\lambda \in \mathbb{R}$, la función $e^{\lambda x}$ es convexa en x, luego
   $$e^{\lambda x} \leq \frac{(x-a)e^{\lambda b}}{b-a} + \frac{(b-x)e^{\lambda a}}{b-a}.$$
   Así, tomando esperanzas, y llamando $1 \geq p := \frac{-a}{b-a} \geq 0$,
   $$\mathbb{E}(e^{\lambda X}) \leq \frac{be^{\lambda a}}{b-a} - \frac{ae^{\lambda b}}{b-a} = (1-p)e^{\lambda a} + pe^{\lambda b}$$
   $$= (1-p)e^{-p\lambda(b-a)} + pe^{\lambda b} = (1-p+pe^{\lambda(b-a)})e^{-p\lambda(b-a)}.$$
   Si $u = \lambda(b-a)$, y $\phi(u) = -pu + \log(1-p+pe^u)$, se tiene
   $$\mathbb{E}(e^{\lambda X}) \leq e^{\phi(u)}.$$
   Entonces,
   $$\phi(0) = 0,$$
   $$\phi'(u) = -p + \frac{pe^u}{1-p+pe^u},$$

$$\phi'(0) = 0,$$
$$\phi''(u) = \frac{pe^u(1-p)}{(1-p+pe^u)^2}.$$

Llamando $z = pe^u \geq 0$ y $q = 1-p \geq 0$, por la desigualdad de las medias aritmética y geométrica,

$$\phi''(u) = \frac{zq}{(z+q)^2} \leq \frac{1}{4}.$$

Por tanto, el teorema de Taylor afirma que

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\xi) \leq \frac{1}{8}u^2,$$

para cierto $\xi \in [0, u]$. Por tanto,

$$\mathbb{E}(e^{\lambda X}) \leq e^{\phi(u)} \leq e^{\frac{1}{8}u^2} \leq e^{\frac{\lambda^2}{8}(b-a)^2},$$

es decir, X es $\frac{b-a}{2}$-subgaussiana.

□

Nota. Una variable σ-subgaussiana, también es σ'-subgaussiana $\forall \sigma' \geq \sigma$.

Teorema 1.9 (Desigualdad de Hoeffding)**.** *Supongamos que $X_1, \ldots, X_n$ son variables aleatorias independientes e igualmente distribuidas, con $\mathbb{E}(X_i) = \mu$ y tal que X_i es $\sigma-$subgaussiana.*
Entonces para todo $a > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq a\right) \leq e^{-\frac{na^2}{2\sigma^2}} \quad y \quad \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq a\right) \leq 2e^{-\frac{na^2}{2\sigma^2}}.$$

Demostración. Por definición, $(X_i - \mu)_{i=1}^n$ son v.a.i.i.d. σ-subgaussianas, luego aplicando el tercer apartado del lema 1.8,

$$\sum_{i=1}^{n}(X_i - \mu) \quad \text{es} \quad \sqrt{n}\sigma\text{-subgaussiana.}$$

Ahora, aplicando el segundo apartado del mismo lema,

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \quad \text{es} \quad \frac{\sigma}{\sqrt{n}}\text{-subgaussiana.}$$

Para concluir, basta aplicar directamente la proposición 1.7. □

Ahora, se aplicará la desigualdad de Hoeffding para demostrar una importante cota superior de la pérdida para el algoritmo *explora-luego-decide*. Este resultado es clave porque garantiza que la pérdida no será excesiva, ayuda a elegir cuánto explorar, y permite comparar el algoritmo con otros métodos.

Teorema 1.10 (Cota para el algoritmo *explora-luego-decide*)**.** *Dado un modelo bandido con P_a σ-subgaussiana para todo brazo a, con k brazos y $km \leq n$ para cierto $n \in \mathbb{N}$. Entonces*

$$R_n(\pi) \leq m \sum_{a \in A} \Delta_a + (n - mk) \sum_{a \in A} \Delta_a \exp\left(\frac{-m\Delta_a^2}{4\sigma^2}\right), \tag{1.4}$$

donde π es la estrategia explora-luego-decide.

Demostración. El algoritmo se basa en explorar cada brazo m veces, y luego seleccionar el brazo con mayor recompensa estimada las $(n - mk)$ veces restantes. Por tanto, con ayuda del lema de descomposición de la pérdida 1.4, se obtiene que

$$R_n(\pi) = \sum_{a \in A} m\Delta_a + \sum_{a \in A} \Delta_a (n - mk) \mathbb{P}\left(\hat{Q}_a(mk) \geq \max_{b \in A} \hat{Q}_b(mk)\right),$$

donde interviene el concepto de valor acción estimado (1.2). Bastaría con acotar la probabilidad por $\exp\left(\frac{-m\Delta_a^2}{4\sigma^2}\right)$ para concluir.

Sea a_* el brazo con mayor recompensa esperada, Q_*. Entonces es claro que

$$\mathbb{P}\left(\hat{Q}_a(mk) \geq \max_{b \in A} \hat{Q}_b(mk)\right) \leq \mathbb{P}\left(\hat{Q}_a(mk) \geq \hat{Q}_*(mk)\right)$$

$$= \mathbb{P}\left(\hat{Q}_a(mk) - Q_a - (\hat{Q}_*(mk) - Q_*) \geq \Delta_a\right).$$

Sumando $\Delta_a = Q_* - Q_a$ a cada lado y reagrupando, se obtiene que

$$\hat{Q}_a(mk) - Q_a - \hat{Q}_*(mk) + Q_*$$

$$= \frac{1}{T_a(mk)} \sum_{j=1}^{mk} X_j^{(a)} \mathbf{1}_{A_k = a} - Q_a - \frac{1}{T_a(mk)} \sum_{j=1}^{mk} X_j^{(a_*)} \mathbf{1}_{A_k = a_*} + Q_*$$

$$= \frac{1}{m}\sum_{j=1}^{m} X_j^{(a)} - Q_a - \frac{1}{m}\sum_{j=1}^{m} X_j^{(a_*)} + Q_*$$

$$= \frac{1}{m}\sum_{j=1}^{m}(X_j^{(a)} - X_j^{(a_*)}) - \mathbb{E}\left(\frac{1}{m}\sum_{j=1}^{m}(X_j^{(a)} - X_j^{(a_*)})\right),$$

pues $X_1^{(a)}, \ldots, X_m^{(a)}$ son v.a.i.i.d por la definición de la estrategia y

$$\mathbb{E}\left(\frac{1}{m}\sum_{j=1}^{m} X_j^{(a)} - X_j^{(a_*)}\right) = \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}(X_j^{(a)} - X_j^{(a_*)})$$

$$= \frac{1}{m}\, m\,(Q_a - Q_*) = (Q_a - Q_*).$$

Por tanto, la situación coincide con la del enunciado de la desigualdad de Hoeffding, por lo que la probabilidad queda acotada por

$$\exp\left(\frac{-m\,\Delta_a^2}{2\,(\sqrt{2\,\sigma^2})^2}\right) = \exp\left(\frac{-m\,\Delta_a^2}{4\sigma^2}\right),$$

ya que $X_j^{(a)} - X_j^{(a_*)}$ es $\sqrt{2\,\sigma^2}$- subgaussiana por el tercer apartado del lema 1.8. Por tanto, se concluye. □

Una vez obtenida esta cota, se presenta el dilema de la elección de m, es decir, cuánto hay que explorar para minimizar la pérdida. Para facilitar los cálculos, se considera el problema de $k = 2$ brazos. Sin pérdida de la generalidad, se asume que $a_* = a_1$, luego $\Delta_{a_1} = 0$ y se denota $\Delta = \Delta_{a_2}$. La cota de la pérdida (1.4) en este caso se reduce a

$$R_n(\pi) \leq m\Delta + (n - 2m)\Delta e^{-\frac{m\Delta^2}{4\sigma^2}}$$

$$= \Delta\left(m + (n - 2m)e^{-\frac{m\Delta^2}{4\sigma^2}}\right) \leq \Delta\left(m + ne^{-\frac{m\Delta^2}{4\sigma^2}}\right).$$

Minimizar el lado derecho en m equivale a minimizar

$$f(m) = m + ne^{-\frac{m\Delta^2}{4\sigma^2}}.$$

Se trata de una función convexa y por tanto, en caso de tener un mínimo relativo, se trata de un mínimo absoluto. Pese a que m debe ser un entero

mayor o igual que 1, se tratará el problema como en el caso continuo y luego se elegirá uno de los dos enteros más cercanos al valor obtenido. Por tanto, derivando se obtiene que

$$f'(m) = 1 + n\left(-\frac{\Delta^2}{4\sigma^2}\right)e^{-\frac{m\Delta^2}{4\sigma^2}} = 0 \Longleftrightarrow \frac{n\Delta^2}{4\sigma^2} = e^{\frac{m\Delta^2}{4\sigma^2}}$$

$$\Longleftrightarrow \frac{m\Delta^2}{4\sigma^2} = \log\left(\frac{n\Delta^2}{4\sigma^2}\right) \Longleftrightarrow m = \frac{4\sigma^2}{\Delta^2}\log\left(\frac{n\Delta^2}{4\sigma^2}\right).$$

Como m debe ser un entero mayor o igual que 1, la cota de la pérdida se minimiza en

$$m = \text{máx}\left\{1, \left\lceil \frac{4\sigma^2}{\Delta^2}\log\left(\frac{n\Delta^2}{4\sigma^2}\right)\right\rceil\right\}.$$

Haciendo el cálculo (en el caso continuo),

$$f\left(\frac{4\sigma^2}{\Delta^2}\log\left(\frac{n\Delta^2}{4\sigma^2}\right)\right) = \Delta\left(\frac{4\sigma^2}{\Delta^2}\log\left(\frac{n\Delta^2}{4\sigma^2}\right)\right) + \Delta n e^{-\log\left(\frac{n\Delta^2}{4\sigma^2}\right)}$$

$$= \frac{4\sigma^2}{\Delta}\log\left(\frac{n\Delta^2}{4\sigma^2}\right) + \frac{4\sigma^2}{\Delta} = \frac{4\sigma^2}{\Delta}\left(1 + \log\left(\frac{n\Delta^2}{4\sigma^2}\right)\right),$$

y teniendo en cuenta que si $m = 1$, $R_n(\pi) \leq (n-1)\Delta$, entonces

$$R_n(\pi) \lesssim \text{mín}\{(n-1)\Delta, \Delta + \frac{4\sigma^2}{\Delta}(1 + \text{máx}\{0, \log\left(\frac{n\Delta^2}{4\sigma^2}\right)\})\}.$$

Si $\frac{n\Delta^2}{4\sigma^2} \geq 1$, es decir, si $n \geq \frac{4\sigma^2}{\Delta^2}$, entonces

$$R_n(\pi) \lesssim \Delta + \frac{4\sigma^2}{\Delta} + \frac{4\sigma^2}{\Delta}\log\left(\frac{n\Delta^2}{4\sigma^2}\right) = c_1(\Delta, \sigma^2) + c_2(\Delta, \sigma^2)\log(n). \quad (1.5)$$

Aunque una cota logarítmica en n para la pérdida es un gran avance, es dependiente de parámetros del modelo, como σ^2 y Δ. Mientras que la dependencia de σ^2 pueden solventarse con ciertos algoritmos, la de Δ es muy severa, pues es desconocido. Por tanto los cálculos no permiten calcular m. Sin embargo, a través del teorema de Lai-Robbins, se asegura que asintóticamente en n esta cota superior esta cerca de ser óptima, al menos para modelos gaussianos.

1.3. Cota inferior de Lai-Robbins

En la sección anterior se han deducido estimaciones logarítmicas para la pérdida para modelos con distribuciones de recompensa con colas pequeñas. Ahora, se deducirán cotas inferiores que establecerán la optimalidad de los algoritmos.

Con el objetivo de demostrar la desigualdad de Lai-Robbins, se introducen en el apéndice B nociones de teoría de la probabilidad, destacando el concepto de entropía relativa A.5.

Antes de obtener una cota asintótica como la de Lai-Robbins, es interesante destacar que se pueden obtener cotas para la pérdida que funcionen para varios modelos a la vez, si estos cumplen una serie de condiciones.

Definición 1.11. Si $\xi = \{v^{(i)}\}_{i\in I}$ es una familia de modelos estocásticos de bandido, entonces se llama pérdida minimax de ξ tras n rondas a

$$R_n^*(\xi) = \inf_{\pi} \sup_{v\in\xi} R_n(\pi, v).$$

La pérdida minimax está asociada a la mejor estrategia posible para minimizar la mayor pérdida que se podría sufrir en el peor escenario dentro de una familia de modelos. Esta definición hace referencia al concepto de pérdida de una estrategia dado un modelo, recogido en (1.1).

Lo siguiente se centrará en probar una cota inferior para la pérdida minimax. Para ello se construirán dos modelos de bandido multibrazo v_μ y $v_{\mu'}$, tales que $R_n(\pi, v_\mu) + R_n(\pi, v_{\mu'})$ sea mayor o igual que una función de la entropía relativa entre sus distribuciones asociadas. Esta cota inferior se obtendrá usando el teorema de Bretagnolle-Huber A.9. Para controlar dicha entropía relativa es relevante el lema siguiente. Para su demostración se hace uso del siguiente concepto:

Si μ y ν son medidas en $\mathcal{F}$ y $\mathcal{G}$ respectivamente, $\mu \otimes \nu$ representa la medida en $\mathcal{F} \times \mathcal{G}$, definida por:

$$\mu \otimes \nu(F \times G) = \mu(F)\nu(G), \quad \forall F \in \mathcal{F},\, G \in \mathcal{G}.$$

Lema 1.12 (Descomposición de la entropía)**.** *Sean k el número de brazos y $v = \{P_1, \dots, P_k\}$ y $v' = \{P'_1, \dots, P'_k\}$ dos modelos de bandido con k brazos, cuyas distribuciones de probabilidad tienen densidad. Además, sea π una estrategia de aprendizaje y $\mathbb{P}_\pi$ y $\mathbb{P}'_\pi$ las distribuciones de probabilidad en $(\mathbb{R} \times \{a_1, \dots, a_k\})^n$ que describen el proceso de n pasos para los dos modelos bajo la estrategia de aprendizaje π. Se sigue dicha notación puesto que el espacio de recompensas es $\mathbb{R}$ y $\mathcal{A} = \{a_1, \dots, a_k\}$ es el conjunto de brazos. Entonces se cumple la siguiente descomposición:*

$$D(\mathbb{P}_\pi, \mathbb{P}'_\pi) = \sum_{a \in A} \mathbb{E}_\pi[T_a(n)] D(P_a, P'_a).$$

Demostración. $\mathbb{P}_\pi, \mathbb{P}'_\pi$ son probabilidades en $(\mathcal{A} \times \mathbb{R})^n$, y P_{a_i}, P'_{a_i} con $i = 1, \dots, n$ son probabilidades en $\mathbb{R}$. Si $\mathbb{P}_\pi \not\ll \mathbb{P}'_\pi$, entonces existe $F = (A_1 \times B_1) \times \cdots \times (A_n \times B_n) \subseteq (\mathcal{A} \times \mathbb{R})^n$ con $\mathbb{P}'_\pi(F) = 0$, pero $\mathbb{P}_\pi(F) \neq 0$. Por tanto, existe $i \leq n$ tal que $P'_{A_i}(B_i) = 0$, pero $P_{A_i}(B_i) \neq 0 \implies P_{A_i} \not\ll P'_{A_i}$. Así, $D(\mathbb{P}_\pi, \mathbb{P}'_\pi) = \infty = \sum_{a \in \mathcal{A}} \mathbb{E}_\pi(T_a(n)) D(P_a, P'_a)$. Luego se supone que $\mathbb{P}_\pi \ll \mathbb{P}'_\pi$.

Dado $a \in \mathcal{A}$, sea p_a la función de densidad de P_a. Si ρ es la medida de conteo de $\mathcal{A}$, es decir,

$$\rho(A) = \#(A \cap \mathcal{A}) \quad \forall A \subseteq \mathcal{A},$$

entonces:

$$\mathbb{P}_\pi \ll (\rho \otimes \lambda)^{\otimes n}.$$

Esto se debe a que dado $F = (A_1 \times B_1) \times \cdots \times (A_n \times B_n) \subseteq (\mathcal{A} \times \mathbb{R})^n$, si $(\rho \otimes \lambda)^{\otimes n}(F) = \rho(A_1)\lambda(B_1) \cdots \rho(A_n)\lambda(B_n) = 0$, entonces existe i tal que $\rho(A_i) = 0$ o $\lambda(B_i) = 0$.

- Si $\rho(A_i) = 0 \implies A_i = \emptyset \implies \mathbb{P}_\pi(F) = 0$ trivialmente.
- Si $\lambda(B_i) = 0 \implies \mathbb{P}_\pi(F) = \prod_{t=1}^n \pi_t(A_t; A_1, X_1, \dots, A_{t-1}, X_{t-1}) P_{A_t}(B_t) = 0$, ya que si $\lambda(B_i) = 0$ entonces $P_{A_i}(B_i) = 0$. También se ha usado que $\mathbb{P}_\pi(X_t \in B_t | A_t) = P_{A_t}(B_t)$, y que $(A_1, X_1, \dots, A_n, X_n)$ es el proceso de bandido estocástico correspondiente a π.

Por tanto, si $F = (A_1 \times B_1) \times \cdots \times (A_n \times B_n) \subseteq (\mathcal{A} \times \mathbb{R})^n$ y $p(a_1, x_1, \dots, a_n, x_n) = \prod_{t=1}^n \pi_t(a_t; a_1, x_1, \dots, a_{t-1}, x_{t-1}) p_{a_t}(x_t)$, entonces

$$\mathbb{P}_\pi(F) = \int_F p(a_1, x_1, \dots, a_n, x_n) \; (\rho \otimes \lambda)^{\otimes n} \left(d(a_1, x_1, ..., a_t, x_t)\right),$$

y se puede decir que

$$\frac{d\mathbb{P}_\pi}{d((\rho\otimes\lambda)^{\otimes n})}(a_1,x_1,\ldots,a_n,x_n)=p(a_1,x_1,\ldots,a_n,x_n).$$

De la misma manera se tiene que

$$\frac{d\mathbb{P}'_\pi}{d((\rho\otimes\lambda)^{\otimes n})}=p', \quad \text{donde}$$

$$p'(a_1,x_1,\ldots,a_n,x_n)=\prod_{t=1}^{n}\pi_t(a_t;a_1,x_1,\ldots,a_{t-1},x_{t-1})p'_{a_t}(x_t).$$

Por tanto,

$$\frac{d\mathbb{P}_\pi}{d\mathbb{P}'_\pi}=\frac{d\mathbb{P}_\pi/d((\rho\otimes\lambda)^{\otimes n})}{d\mathbb{P}'_\pi/d((\rho\otimes\lambda)^{\otimes n})}=\frac{p}{p'},$$

luego,

$$\frac{d\mathbb{P}_\pi}{d\mathbb{P}'_\pi}(a_1,x_1,\ldots,a_n,x_n)=\frac{\prod_{t=1}^{n}\pi_t(a_t;a_1,x_1,\ldots,a_{t-1},x_{t-1})p_{a_t}(x_t)}{\prod_{t=1}^{n}\pi_t(a_t;a_1,x_1,\ldots,a_{t-1},x_{t-1})p'_{a_t}(x_t)}$$

$$=\prod_{t=1}^{n}\frac{p_{a_t}(x_t)}{p'_{a_t}(x_t)}.$$

Si se usa la definición de entropía relativa, y que

$$\frac{dP_{a_i}}{dP'_{a_i}}=\frac{p_{a_i}}{p'_{a_i}},$$

entonces, si $(A_1,X_1,\ldots,A_n,X_n)$ es el proceso de bandido estocástico correspondiente a π:

$$D(\mathbb{P}_\pi,\mathbb{P}'_\pi)=\mathbb{E}_\pi\left(\log\frac{\mathbb{P}_\pi}{\mathbb{P}'_\pi}\right)=\mathbb{E}_\pi\left(\log\prod_{t=1}^{n}\frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)}\right)$$

$$=\sum_{t=1}^{n}\mathbb{E}_\pi\left(\log\frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)}\right)=\sum_{t=1}^{n}\mathbb{E}_\pi\left(\mathbb{E}\left(\log\frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)}\mid A_t\right)\right)$$

$$=\sum_{t=1}^{n}\mathbb{E}_\pi\left(\mathbb{E}\left(\log\frac{dP_{A_t}(X_t)}{dP'_{A_t}(X_t)}\mid A_t\right)\right)=\sum_{t=1}^{n}\mathbb{E}_\pi(D(P_{A_t},P'_{A_t}))$$

$$= \sum_{a\in\mathcal{A}} \mathbb{E}_\pi(\sum_{t=1}^{n} \mathbf{1}_{\{A_t=a\}}\, D(P_{A_t}, P'_{A_t})) = \sum_{a\in\mathcal{A}} \mathbb{E}_\pi(T_a(n)) D(P_{A_t}, P'_{A_t}).$$

□

El siguiente teorema establece una cota para la pérdida minimax para recompensas gaussianas. Es una cota más general que la de Lai-Robbins y se usa para evaluar el peor caso sin hacer suposiciones sobre la estrategia. De esta manera se analizan casos donde los algoritmos podrían no ser óptimos en el sentido de Lai-Robbins.

Teorema 1.13. *Sea ξ_k la familia de todos los bandidos con k brazos con recompensas Gaussianas de varianza unitaria y esperanzas $\mu_1, \ldots, \mu_k \in [0,1]$. Entonces*

$$R_n^*(\xi_k) \geq \frac{1}{27}\sqrt{(k-1)n}$$

se cumple para todo $n \geq k$.

Demostración. Si se encuentra un modelo específico $v^* \in \xi_k$ que para cualquier estrategia cumpla que $R_n(\pi, v^*) \geq \frac{1}{27}\sqrt{(k-1)n}$, entonces $\sup_{v\in\xi} R_n(\pi, v)$ también lo cumple, y por tanto $R_n^*(\xi_k)$.

Nótese en primer lugar, que cada modelo queda determinado completamente por un vector de medias $\mu = (\mu_1, \ldots, \mu_k)$.

Sea $\mathcal{A} = (a_1, \ldots, a_k)$ el conjunto de todos los brazos. Dado $\Delta \in [0, 1/2]$, se define $\mu = (\Delta, 0, \ldots, 0)$. Dada una estrategia π cualquiera, sea $a = \arg\min_{a_i\in A} \mathbb{E}_{v_\mu,\pi}[T_{a_i}(n)]$ el brazo que se espera que se juegue menos si se sigue la estrategia π en el modelo v_μ. Se define también $\mu' = (\Delta, 0, \ldots, 0, 2\Delta, 0, \ldots, 0)$ tal que el valor 2Δ se sitúa en la posición del brazo a. Se denotan $\mathbb{P}$ y $\mathbb{P}'$ las probabilidades de los procesos de bandido correspondientes a la estrategia π y los modelos v_μ y v'_μ, respectivamente.

Usando el lema de descomposición de la pérdida 1.4 y que para v_μ el brazo óptimo es a_1,

$$R_n(\pi, v_\mu) = \sum_{a_i\in\mathcal{A}} \mathbb{E}[T_{a_i}(n)]\Delta_i = \sum_{a_i\neq a_1} \mathbb{E}[T_{a_i}(n)]\Delta = \Delta(n - \mathbb{E}[T_{a_1}(n)])$$

$$= \Delta(n - \mathbb{E}[T_{a_1}(n)]) = \Delta(n - \mathbb{E}[T_{a_1}(n)(1_{\{T_{a_1}(n)>\frac{n}{2}\}} + 1_{\{T_{a_1}(n)\leq\frac{n}{2}\}})]).$$

Como $T_{a_1}(n) \leq n$,

$$R_n(\pi, v_\mu) \geq \Delta\Big(n - n\mathbb{P}(T_{a_1}(n) > \frac{n}{2}) - \frac{n}{2}\,\mathbb{P}(T_{a_1}(n) \leq \frac{n}{2})\Big)$$

$$= \Delta\Big(n - n(1 - \mathbb{P}(T_{a_1}(n) \leq \frac{n}{2})) - \frac{n}{2}\,\mathbb{P}(T_{a_1}(n) \leq \frac{n}{2})\Big)$$

$$= \frac{n\Delta}{2}\mathbb{P}(T_{a_1}(n) \leq \frac{n}{2}).$$

De igual manera, como para el modelo v'_μ el brazo a_1 no es óptimo, y su hueco de recompensa es Δ,

$$R_n(\pi, v'_\mu) = \sum_{a_i \in \mathcal{A}} \mathbb{E}'[T_{a_i}(n)]\Delta'_{a_i} \geq \mathbb{E}'[T_{a_1}(n)]\Delta \geq \mathbb{E}'[T_{a_1}(n)(1_{\{T_{a_1}(n) > \frac{n}{2}\}})]\Delta$$

$$= \frac{n\Delta}{2}\mathbb{P}'(T_{a_1}(n) > \frac{n}{2}).$$

Si se llama A al suceso $\{T_{a_1}(n) \leq \frac{n}{2}\}$, se pueden combinar las desigualdades ahora obtenidas, el teorema de Bretagnolle-Huber A.9 y el lema anterior para obtener que

$$R_n(\pi, v_\mu) + R_n(\pi, v'_\mu) \geq \frac{n\Delta}{2}\Big(\mathbb{P}(A) + \mathbb{P}'(A^c)\Big) \geq \frac{n\Delta}{4}e^{-D(\mathbb{P},\mathbb{P}')}$$

$$= \frac{n\Delta}{4}e^{-\sum_{a_i \in \mathcal{A}} \mathbb{E}_\mu[T_{a_i}(n)] D(\mathcal{N}(\mu_i,1),\mathcal{N}(\mu'_i,1))} = \frac{n\Delta}{4}e^{-\mathbb{E}_\mu[T_a(n)]\frac{(2\Delta)^2}{4}},$$

usando que los vectores μ y μ' solo difieren en la posición de a, y el segundo apartado de A.8.

Como si se sigue la estrategia π en el modelo v_μ, a es el brazo con menor esperanza de ser jugado, entonces

$$(k-1)\,\mathbb{E}[T_a(n)] \leq \sum_{a_i \in \mathcal{A}} \mathbb{E}[T_{a_i}(n)] = n.$$

Por tanto,

$$R_n(\pi, v_\mu) + R_n(\pi, v'_\mu) \geq \frac{n\Delta}{4}\, e^{\frac{-2n\Delta^2}{k-1}} = f(\Delta).$$

Como Δ es arbitrario en $[0, 1/2]$, usando que $n \geq k-1$ se puede evaluar la cota en $\Delta = \sqrt{\frac{k-1}{4n}}$.

Entonces se obtiene que

$$R_n(\pi, v_\mu) + R_n(\pi, v'_\mu) \geq f\left(\sqrt{\frac{k-1}{4n}}\right) = \frac{n\sqrt{k-1}}{8\sqrt{n}} e^{-1/2}$$

$$= \frac{1}{8}\sqrt{(k-1)n}\, e^{-1/2} \geq \frac{2}{27}\sqrt{(k-1)n}.$$

Por tanto, alguna de las dos pérdidas debe ser, como mínimo,

$$\frac{1}{27}\sqrt{(k-1)n}.$$

□

Ahora, se busca establecer la cota inferior de Lai-Robbins. Es una cota asintótica en n para la pérdida de una estrategia, que considera un modelo de bandido fijo cualquiera dentro de una clase. Para evitar casos extremos, se considerarán estrategias con crecimiento sublineal para la pérdida.

Definición 1.14. Se dice que una estrategia π es consistente sobre una clase de modelos de bandido $\mathcal{E}$ si para todo modelo $v \in \mathcal{E}$ y cualquier $p > 0$, se cumple que:

$$\lim_{n\to\infty} \frac{R_n(\pi, v)}{n^p} = 0. \tag{1.6}$$

Teorema 1.15 (Cota inferior de Lai-Robbins)**.** *Si π es una estrategia consistente sobre $\mathcal{E} = \mathcal{M}_1 \times ... \times \mathcal{M}_k$, conjunto de modelos de bandido, entonces la pérdida para todos los modelos $v = (P_1, ..., P_k) \in \mathcal{E}$ cumple que:*

$$\liminf_{n\to\infty} \frac{R_n(\pi, v)}{\log n} \geq c^*(v, \mathcal{E}) := \sum_{a\in\mathcal{A}} \frac{\Delta_a}{d_a},$$

donde $d_a = \inf_{P'\in\mathcal{M}_a}\{D(P_a, P') : Q_{P'} > Q_*\}$ y Q_P es la esperanza asociada a P.

Demostración. Se fija π estrategia consistente y $v = (P_1, \ldots, P_k)$ un modelo bandido. Si se demuestra que

$$\liminf_{n\to\infty} \frac{\mathbb{E}[T_a(n)]}{\log n} \geq \frac{1}{d_a},$$

con ayuda del lema de descomposición de la pérdida 1.4 y las propiedades del límite inferior,

$$\liminf_{n\to\infty} \frac{R_n(\pi, v)}{\log n} = \liminf_{n\to\infty} \sum_{a\in\mathcal{A}} \frac{\Delta_a \,\mathbb{E}[T_a(n)]}{\log n} \geq \sum_{a\in\mathcal{A}} \liminf_{n\to\infty} \frac{\Delta_a \,\mathbb{E}[T_a(n)]}{\log n} \geq \sum_{a\in\mathcal{A}} \frac{\Delta_a}{d_a}.$$

Se fija, por tanto, un brazo a y $\epsilon > 0$. Se define v' como el modelo que resulta de sustituir P_a por $P'_a \in \mathcal{M}_a$ tal que $D(P_a, P'_a) \leq d_a + \epsilon$ y $Q'_a > Q_*$ (P'_a existe por la definición de ínfimo).

Hay que destacar que v y v' generan $\mathbb{P}$ y $\mathbb{P}'$, dos probabilidades que describen el proceso de n pasos para cada modelo según π. Hay, por tanto, dos funciones de pérdida $R_n(\pi, v)$ y $R_n(\pi, v')$.

Como el resto de brazos mantienen sus distribuciones de recompensa, por la descomposición de la entropía 1.12 se obtiene que

$$D(\mathbb{P}, \mathbb{P}') = \sum_{j=1}^{K} \mathbb{E}[T_{a_j}(n)] D(P_{a_j}, P'_{a_j}) \leq \mathbb{E}[T_a(n)](d_a + \epsilon). \tag{1.7}$$

Por la descomposición de la pérdida y la definición de esperanza,

$$R_n(\pi, v) = \sum_{j=1}^{K} \mathbb{E}[T_{a_j}(n)]\Delta_{a_j} \geq \Delta_a \,\mathbb{E}[T_a(n)] \geq \Delta_a \,\mathbb{E}[T_a(n)\mathbf{1}_{\{T_a(n) > \frac{n}{2}\}}]$$

$$\geq \Delta_a \frac{n}{2} \mathbb{P}\left(T_a(n) > \frac{n}{2}\right).$$

El mejor brazo para v' es a', ya que $Q_{P'_a} > Q_*$. Si $T_a(n) \leq \frac{n}{2}$, el resto de brazos deben ser jugados al menos $\frac{n}{2}$ veces en total. Luego se tiene que

$$R_n(\pi, v') = \sum_{a_j \neq a} (Q'_a - Q_{a_j}) \mathbb{E}'[T_{a_j}(n)] \geq \sum_{a_j \neq a} (Q'_a - Q_*) \mathbb{E}'(T_a(n))$$

$$\geq \sum_{a_j \neq a} (Q'_a - Q_*) \mathbb{E}'(T_a(n)\mathbf{1}_{\{T_a(n) \leq \frac{n}{2}\}}) \geq \frac{n}{2}(Q'_a - Q_*) \mathbb{P}'\left(T_a(n) \leq \frac{n}{2}\right).$$

Por tanto, si A es el suceso $\{T_a(n) > \frac{n}{2}\}$,

$$R_n(\pi, v) + R_n(\pi, v') \geq \frac{n}{2}\left(\Delta_a \,\mathbb{P}\left(T_a(n) > \frac{n}{2}\right) + (Q'_a - Q_a)\,\mathbb{P}'\left(T_a(n) \leq \frac{n}{2}\right)\right)$$

$$\geq \frac{n}{2} \min\{\Delta_a, Q'_a - Q_*\} \left(P(A) + P'(A^C)\right).$$

Aplicando el teorema de Bretagnolle-Huber A.9, y (1.7),

$$R_n(\pi, v) + R_n(\pi, v') \geq \frac{n}{4} \min\{\Delta_a, Q'_a - Q_*\} e^{-D(\mathbb{P}, \mathbb{P}')}$$

$$\geq \frac{n}{4} \min\{\Delta_a, Q'_a - Q_*\} e^{-\mathbb{E}[T_a(n)](d_a + \epsilon)}.$$

Reordenando esta desigualdad y usando propiedades de límites superiores e inferiores,

$$e^{\mathbb{E}[T_a(n)](d_a+\epsilon)} \geq \frac{n \ \min\{\Delta_a, Q'_a - Q_*\}}{4\left(R_n(\pi, v) + R_n(\pi, v')\right)}$$

$$\Leftrightarrow \mathbb{E}[T_a(n)](d_a + \epsilon) \geq \log n + \log\left(\frac{\min\{\Delta_a, Q'_a - Q_*\}}{4\left(R_n(\pi, v) + R_n(\pi, v')\right)}\right)$$

$$\Leftrightarrow \frac{\mathbb{E}[T_a(n)]}{\log n} \geq \frac{1}{d_a + \epsilon}\left(1 + \frac{\log\left(\frac{\min\{\Delta_a, Q'_a - Q_*\}}{4\left(R_n(\pi, v) + R_n(\pi, v')\right)}\right)}{\log n}\right)$$

$$\Leftrightarrow \liminf_{n\to\infty} \frac{\mathbb{E}[T_a(n)]}{\log n} \geq \frac{1}{d_a + \epsilon}\left(1 + \liminf_{n\to\infty} \frac{-\log\left(R_n(\pi, v) + R_n(\pi, v')\right)}{\log n}\right)$$

$$= \frac{1}{d_a + \epsilon}\left(1 - \limsup_{n\to\infty} \frac{\log\left(R_n(\pi, v) + R_n(\pi, v')\right)}{\log n}\right).$$

Por la propiedad de consistencia (1.6) de π, existe N tal que si $n \geq N$, existen $C, C' > 0$ tal que

$$R_n(\pi, v) \leq C\, n^p \text{ y } R_n(\pi, v') \leq C'\, n^p, \quad \forall p > 0.$$

Por tanto,

$$0 \leq \limsup_{n\to\infty} \frac{\log\left(R_n(\pi, v) + R_n(\pi, v')\right)}{\log n} \leq \limsup_{n\to\infty} \frac{\log\left(C\, n^p + C'\, n^p\right)}{\log n}$$

$$= \limsup_{n\to\infty} \frac{p\, \log(n) + \log(C + C')}{\log n} = p.$$

Haciendo $p \to 0$ se obtiene que el límite superior es 0, y entonces

$$\liminf_{n\to\infty} \frac{\mathbb{E}_\pi[T_a(n)]}{\log n} \geq \frac{1}{d_a + \epsilon}.$$

Como $\epsilon > 0$ es arbitrario, se concluye. □

Esta cota proporciona un límite ineludible para cualquier estrategia consistente. De esta manera, si un algoritmo tiene una cota superior del mismo orden que la cota inferior de Lai-Robbins, entonces se considera óptimo en orden de magnitud. Por tanto se ha obtenido un criterio mediante el que se puede discutir la optimalidad de cualquier algoritmo para el que se tenga una cota superior, como es el caso del algoritmo *explora-luego-decide*. Esta estrategia es consistente, ya que dado $p > 0$, usando la cota (1.5) se obtiene que

$$0 \leq \frac{R_n(\pi)}{n^p} \lesssim \frac{c_1(\Delta, \sigma^2) + c_2(\Delta, \sigma^2)\log(n)}{n^p}.$$

Ahora tomando límites cuando $n \to \infty$,

$$0 \leq \lim_{n\to\infty} \frac{R_n(\pi)}{n^p} \lesssim \lim_{n\to\infty} \frac{c_1(\Delta, \sigma^2) + c_2(\Delta, \sigma^2)\log(n)}{n^p} = 0,$$

por órdenes de infinitud. Entonces, puesto que la cota superior (1.5) es del mismo orden (logarítmico en n) que la cota inferior de Lai-Robbins, se considera un algoritmo óptimo en orden de magnitud.

Capítulo 2

Problemas de decisión de Markov

Tras abordar el problema del bandido multibrazo en el capítulo anterior, donde se analizaron escenarios de decisión con una única etapa y recompensas inmediatas, se amplía ahora el alcance hacia problemas más complejos que involucran decisiones secuenciales. Este capítulo se centrará en los procesos de decisión de Markov (MDPs, por sus siglas en inglés), un marco formal que permite modelar entornos donde las decisiones actuales afectan no solo a las recompensas inmediatas, sino también al estado futuro del sistema.

Se explorará a continuación la estructura matemática de los MDPs, sus elementos principales y las herramientas que ofrecen para la toma de decisiones óptimas en contextos dinámicos e inciertos.

2.1. Contexto

En el campo del aprendizaje por refuerzo, los problemas de decisión de Markov desempeñan un papel central. Estos proporcionan la base matemática para modelar entornos en los que un agente aprende a tomar decisiones óptimas a través de la interacción con su entorno dinámico en el que las decisiones influyen en los estados futuros y las recompensas acumuladas a largo plazo.

Para comprender los MDPs, resulta esencial entender su relación con las cadenas de Markov, un concepto previo que establece la base de los procesos estocásticos en los que se fundamentan. Una cadena de Markov describe un sistema cuya evolución depende únicamente del estado actual y no de la secuencia de estados anteriores, una propiedad conocida como propiedad de Markov.

Esta simplicidad en la dependencia del estado presente hace que las cadenas de Markov sean ideales para modelar la dinámica de los sistemas en Reinforcement Learning. Los MDPs, por su parte, enriquecen este modelo básico al incluir decisiones deliberadas y recompensas asociadas, permitiendo a un agente aprender estrategias que maximicen su desempeño a lo largo del tiempo. De hecho, al final del capítulo se demostrará que los MDPs son en esencia cadenas de Markov extendidas, donde la transición entre estados depende no solo del estado actual, sino también de la acción tomada. Este vínculo entre cadenas de Markov y MDPs establece un puente fundamental para avanzar hacia el estudio de teoría más compleja y la implementación de algoritmos de aprendizaje por refuerzo.

Definición 2.1. $(S_t)_{t\in\mathbb{N}}$ es una cadena de Markov de estados finitos si es un proceso estocástico en tiempo discreto en un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$, con valores en un conjunto finito $\mathcal{S}$, que cumple que

$$\mathbb{P}(S_{t+1} = s_{t+1} | S_0 = s_0, \ldots, S_t = s_t) = \mathbb{P}(S_{t+1} = s_{t+1} | S_t = s_t).$$

Dicha condición es conocida como propiedad de Markov. Además, se dice que la cadena de Markov es homogénea si las probabilidades de transición de estados no cambian con el tiempo. En este caso, se cumple que

$$\mathbb{P}(S_{t+1} = s_{t+1} | S_0 = s_0, \ldots, S_t = s_t) = \mathbb{P}(S_{t+1} = s_{t+1} | S_t = s_t) = p_{t,t+1}.$$

Se conoce por matriz de transición a la matriz $P = (p_{i,j})$, donde $p_{i,j}$ es la probabilidad de pasar del estado s_i al estado s_j. Estas matrices son matrices estocásticas, ya que cumplen que:

- $p_{i,j} \geq 0 \quad \forall i, j$ con $s_i, s_j \in \mathcal{S}$.
- $\sum_j p_{i,j} = 1 \quad \forall i$ con $s_i \in \mathcal{S}$.

Dada una distribución inicial μ en $\mathcal{S}$ y una matriz de transición P, se pueden expresar las probabilidades de la siguiente manera:

$$\mathbb{P}(S_0 = s_0, \ldots, S_n = s_n) = \mu(s_0)\, p_{0,1} \ldots p_{n-1,n}.$$

A raíz de esta fórmula, pueden calcularse otras muchas probabilidades, por ejemplo:

$$\begin{aligned}&\mathbb{P}(S_{t+1}=s_{t+1},\dots,S_{t+k}=s_{t+k}\,|\,S_t=s_t)\\&=\frac{\mathbb{P}(S_t=s_t,S_{t+1}=s_{t+1},\dots,S_{t+k}=s_{t+k})}{\mathbb{P}(S_t=s_t)}\\&=\frac{\sum_{s_0,\dots,s_{t-1}}\mu(s_0)\,p_{0,1}\dots p_{t-1,t}\cdot p_{t,t+1}\cdot\dots\cdot p_{t+k-1,t+k}}{\sum_{s_0,\dots,s_{t-1}}\mu(s_0)\,p_{0,1}\dots p_{t-1,t}}=p_{t,t+1}\cdot\dots\cdot p_{t+k-1,t+k}.\end{aligned}\tag{2.1}$$

Otra manera de expresar una cadena de Markov, si $\mathbb{P}(S_n = s') > 0$, es considerar $\tilde{\mathbb{P}} := \mathbb{P}(\cdot \,|\, S_n = s')$, y en $(\Omega, \mathcal{F}, \tilde{\mathbb{P}})$ tomar $\tilde{S} = (\tilde{S}_t)_{t\in\mathbb{N}} := (S_{t+n})_{t\in\mathbb{N}}$. En estas condiciones, se sabe que $(\tilde{S}_t)_{t\in\mathbb{N}}$ es una cadena de Markov con matriz de transición igual a la de S, pero que comienza en s'.

A una cadena de Markov se le puede añadir otro proceso de recompensas $(R_t)_{t\in\mathbb{N}}$, con valores en un conjunto $\mathcal{R}$, junto a una función de transición/recompensa $p : (\mathcal{R}\times\mathcal{S})\times\mathcal{S} \to [0,1]$. Así, (R_t) se muestrea junto al nuevo estado, coordinado por la función p, de manera que $p(r, s'; s)$ denota la probabilidad de pasar del estado s a s' recibiendo la recompensa r. En este caso, la propiedad de Markov implica que

$$\begin{aligned}&\mathbb{P}_\mu(S_{t+1}=s_{t+1},R_{t+1}=r_{t+1}|S_0=s_0,\dots,S_t=s_t)\\&=\mathbb{P}_\mu(S_{t+1}=s_{t+1},R_{t+1}=r_{t+1}|S_t=s_t).\end{aligned}$$

En esta situación, al igual que antes, se puede construir el proceso $(\tilde{S}_t, \tilde{R}_t)_{t\in\mathbb{N}}$, en el espacio probabilístico $(\Omega, \mathcal{F}, \tilde{\mathbb{P}})$, donde $\tilde{\mathbb{P}} := \mathbb{P}(\cdot \,|\, S_n = s')$, siempre que $\mathbb{P}(S_n = s') > 0$). De manera análoga al caso anterior, $(\tilde{S}, \tilde{R})$ es una cadena de Markov con recompensas, que empieza en s' y tiene las mismas transiciones que (S, R).

Ahora se probará una fórmula que será necesaria más adelante.

Proposición 2.2. *Dada una cadena de Markov con recompensas, $(S_t, R_t)_{t\in\mathbb{N}}$ y $s, s' \in \mathcal{S}$, entonces:*

$$\mathbb{E}_s(f(R_t, R_{t+1}, \dots)|S_t = s') = \mathbb{E}_{s'}(f(R_0, R_1, \dots)) \tag{2.2}$$

Demostración. Se denota por $\mathbb{E}_s$ a la esperanza vinculada al proceso con estado inicial s. Entonces,

$$\mathbb{E}_s(f(R_t, R_{t+1}, \dots)|S_t = s') = \mathbb{E}(f(R_t, R_{t+1}, \dots)|S_0 = s_0, S_t = s')$$

$$
\begin{aligned}
&= \mathbb{E}(f(R_t, R_{t+1}, \ldots)|S_t = s') \\
&= \sum_{r_t, r_{t+1}, \ldots} f(r_t, r_{t+1}, \ldots)\mathbb{P}(R_t = r_t, R_{t+1} = r_{t+1}, \ldots | S_t = s') \\
&= \sum_{r_t, r_{t+1}, \ldots} f(r_t, r_{t+1}, \ldots)\mathbb{P}(\tilde{R}_0 = r_t, \tilde{R}_1 = r_{t+1}, \ldots | \tilde{S}_0 = s') \\
&= \sum_{r_t, r_{t+1}, \ldots} f(r_t, r_{t+1}, \ldots)\mathbb{P}(R_0 = r_t, R_1 = r_{t+1}, \ldots | S_0 = s') \\
&= \mathbb{E}_{s'}(f(R_0, R_1, \ldots)).
\end{aligned}
$$

□

2.2. Procesos de decisión de Markov

Los procesos de decisión de Markov (MDP) amplían este concepto incorporando decisiones al sistema, las cuales influyen en las probabilidades de transición a futuros estados. Además, se asocian recompensas a las transiciones, lo que permite evaluar la calidad de las decisiones tomadas.

Definición 2.3. Un modelo de decisión de Markov discreto es una tupla $(\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$ tal que:

- $\mathcal{S}$ es el espacio de estados.
- Para cada $s \in \mathcal{S}$, $\mathcal{A}_s$ es el espacio de acciones del estado s. Se define el espacio de todas las acciones como $\mathcal{A} = \bigcup_{s \in \mathcal{S}} \mathcal{A}_s$.
- $\mathcal{R} \subset \mathbb{R}$ es el espacio de recompensas.

  Se denomina modelo de Markov discreto porque los conjuntos $\mathcal{S}, \mathcal{A}, \mathcal{R}$ son finitos o numerables. Se denota por $\overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}$ a los conjuntos potencia correspondientes.
- $p : \overline{\mathcal{S}} \otimes \overline{\mathcal{R}} \times (\mathcal{S} \times \mathcal{A}) \to [0, 1]$ es la función de transición/recompensa, que cumple que:
  1. Para cada $s \in \mathcal{S}$ y para cada $a \in \mathcal{A}_s$, $p(\cdot; s, a)$ es una distribución de probabilidad discreta sobre $\mathcal{S} \times \mathcal{R}$.

2. Para cada $s \in \mathcal{S}$ y para cada $a \in \mathcal{A}_s$, se cumple que

$$\sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r; s, a) = 1.$$

La función $p(s', r; s, a)$ representa la probabilidad de obtener la recompensa r y avanzar al estado s' cuando se toma la acción a en el estado s.

Para incluir el proceso de elección de una acción en un estado concreto, es necesario el siguiente concepto.

Definición 2.4. Dado un modelo de decisión de Markov $(\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$, una política $\pi = (\pi_t)_{t \in \mathbb{N}_0}$ es:

- Una distribución discreta inicial en $\mathcal{A}, \pi_0 : \overline{\mathcal{A}} \times \mathcal{S} \to [0, 1]$.
- Una sucesión de distribuciones discretas de probabilidad en $\mathcal{A}$, $(\pi_t)_{t \in \mathbb{N}}$ con $\pi_t : \overline{\mathcal{A}} \times ((\mathcal{S} \times \mathcal{A})^t \times \mathcal{S})$ que cumple que

$$\pi_t(\mathcal{A}_s; s_0, a_0, ..., s_{t-1}, a_{t-1}, s) = 1$$

para todo $(s_0, a_0, ..., s_{t-1}, a_{t-1}, s) \in (\mathcal{S} \times \mathcal{A})^t \times \mathcal{S}$.

El conjunto de todas las políticas se denota por Π.

Con la definición de política, se puede definir un proceso estocástico en $\mathcal{S} \times \mathcal{A} \times \mathcal{R}$ completamente definido por p y una política π.

Teorema 2.5 (Existencia de MDPs discretos)**.** *Dado $(\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$ un modelo de decisión de Markov discreto, una política π, y μ una distribución discreta de probabilidad en $\mathcal{S}$, entonces existe un proceso estocástico $(S_t, A_t, R_t)_{t \geq 0}$ en un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P}^\pi_\mu)$ con valores en $\mathcal{S} \times \mathcal{A} \times \mathcal{R}$, tal que, para todo $t \in \mathbb{N}$:*

$$\mathbb{P}^\pi_\mu(S_0 = s_0, A_0 = a_0) = \mu(s_0)\pi_0(a_0; s_0), \tag{2.3}$$

$$\mathbb{P}^\pi_\mu(A_t = a_t | S_0 = s_0, A_0 = a_0, \ldots, S_t = s_t) = \pi_t(a_t; s_0, a_0, ..., s_t), \tag{2.4}$$

$$\mathbb{P}^\pi_\mu(S_{t+1} = s_{t+1}, R_t = r_t | S_t = s, A_t = a) = p(s_{t+1}, r_t; s, a). \tag{2.5}$$

Demostración. Esta demostración cubre el caso de límite temporal finito. La extensión al caso $T = +\infty$ requiere el Teorema de Extensión de Kolgomorov, y los detalles de la adaptación a este caso se comentan después de la prueba. Sea, por tanto, $T < \infty$ un límite temporal finito. Se considera el espacio

probabilístico $(\Omega_T, \mathcal{F}_T, \mathbb{P}_T)$, donde $\Omega_T = (\mathcal{S} \times \mathcal{A} \times \mathcal{R})^T$ son las trayectorias de longitud T, $\mathcal{F}_T$ es su conjunto potencia, y $\mathbb{P}_T$ es explícitamente:

$$\mathbb{P}_T(\omega) := \mu(s_0)\pi_0(a_0; s_0) \prod_{i=1}^{T} p(s_i, r_{i-1}; s_{i-1}, a_{i-1})\pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i)$$
$$\cdot p(\mathcal{S} \times \{r_T\}; s_T, a_T), \tag{2.6}$$

para todo $\omega = (s_0, a_0, r_0, \ldots, s_T, a_T, r_T) \in \Omega_T$. Es decir, se muestrea el primer estado por μ, la primera acción por π_0, y la primera recompensa se toma como 0. A partir de ahí, cada nuevo estado y acción quedan determinados por p y π_i, respectivamente. El último factor expresa que la trayectoria termina en el estado s_T con la recompensa r_T.

Hay que comprobar que $\mathbb{P}_T$ es una medida de probabilidad discreta sobre Ω_T:

- No negatividad: Cada uno de los términos de la definición es una función no negativa:

$$\mu(s_0) \geq 0 \quad \text{por ser distribución inicial,}$$
$$\pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i) \geq 0 \quad \text{pues las políticas son distribuciones,}$$
$$p(s_i, r_{i-1} | s_{i-1}, a_{i-1}) \geq 0 \quad \text{por expresar probabilidad,}$$
$$\text{al igual que } p(\mathcal{S} \times \{r_T\}; s_T, a_T).$$

- Normalización: Hay que probar que la suma sobre todas las posibles trayectorias de tamaño T tiene probabilidad 1:

$$\sum_{s_0, a_0, r_0, ..., s_T, a_T, r_T} \mathbb{P}_T((s_0, a_0, r_0, \ldots, s_T, a_T, r_T))$$
$$= \sum_{s_0, a_0, r_0} \cdots \sum_{s_T, a_T, r_T} \mu(s_0)\pi_0(a_0; s_0) \prod_{i=1}^{T} p(s_i, r_{i-1}; s_{i-1}, a_{i-1})$$
$$\cdot \pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i)\, p(\mathcal{S} \times \{r_T\}; s_T, a_T).$$

  Reorganizando las sumas para evaluar iterativamente, se tiene que: En el paso $t = T$,

$$\sum_{r_{T-1} s_T, a_T, r_T} p(s_T, r_{T-1}; s_{T-1}, a_{T-1})\pi_T(a_T; s_0, a_0, ..., s_T)p(\mathcal{S} \times \{r_T\}; s_T, a_T)$$

$$= \sum_{a_T} \pi_T(a_T; s_0, a_0, ..., s_T) \sum_{s_T, r_{T-1}} p(s_T, r_{T-1}; s_{T-1}, a_{T-1})$$
$$\cdot \sum_{r_T} p(\mathcal{S} \times \{r_T\}; s_T, a_T) = 1 \cdot 1 \cdot 1 = 1,$$

por ser π_T distribución de probabilidad y por la definición de p.

En el paso $t = T - 1$, se obtiene que

$$\sum_{r_{T-2}, s_{T-1}, a_{T-1},} p(s_{T-1}, r_{T-2}; s_{T-2}, a_{T-2}) \pi_{T-1}(a_{T-1}; s_0, a_0, ..., s_{T-1}) = 1,$$

por las propiedades de p y las políticas π_i.

Continuando de esta manera para $t = T - 2, T - 3, ..., 1$, en cada paso las sumas sobre r_{t-1}, s_i y a_i eliminan términos y producen factores 1. Finalmente, se llega a $\sum_{s_0, a_0} \mu(s_0)\pi_0(a_0; s_0) = 1$, por lo que se puede concluir que $\sum_{\omega \in \Omega} \mathbb{P}_T(\omega) = 1$.

Si se considera la probabilidad $\mathbb{P}_T : \mathcal{F}_T \to [0, 1]$, donde $\mathcal{F}_T$ es el conjunto potencia de Ω_T, entonces dado $A \in \mathcal{F}_T$ se define:

$$\mathbb{P}_T(A) := \sum_{\omega \in A} \mathbb{P}_T(\omega) \in [0, 1].$$

Entonces, $\mathbb{P}_T$ cumple la $\sigma-$ aditividad, ya que

$$\mathbb{P}_T(\bigcup_{i \in I} A_i) = \sum_{\omega \in \bigcup_{i \in I} A_i} \mathbb{P}_T(\omega) = \sum_{i \in I} \sum_{\omega \in A_i} \mathbb{P}_T(\omega) = \sum_{i \in I} \mathbb{P}_T(A_i),$$

por ser una unión disjunta de conjuntos discretos de trayectorias. Es decir, son todo trayectorias distintas y la probabilidad total es la suma de las probabilidades de cada trayectoria.

Solo falta comprobar que para todo $t \in \mathbb{N}$ se cumplen las propiedades del teorema:

1. Ahora se comprobará la condición:

   $$\mathbb{P}_t(S_0 = s_1, A_0 = a_0) = \mu(s_0)\pi_0(a_0; s_0)$$

   se satisface porque la construcción explícita de $\mathbb{P}_t$ asegura que μ y π_0 gobiernan la elección del estado inicial y la acción inicial, respectivamente.

2.

$$\mathbb{P}_t(A_t = a_t | S_0 = s_0, A_0 = a_0, \ldots, S_t = s_t) = \pi_t(a_t; s_0, a_0, \ldots, s_t).$$

Usando la definición (2.6),

$$\mathbb{P}_t(S_0 = s_0, A_0 = a_0, \ldots, S_t = s_t, A_t = a_t) =$$

$$= \mu(s_0)\pi_0(a_0; s_0) \prod_{i=1}^{t} p(\{s_i\} \times \mathcal{R}; s_{i-1}, a_{i-1})\pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i),$$

ya que no hay condición sobre la recompensa. Como

$$\mathbb{P}_t(S_0 = s_0, A_0 = a_0, \ldots, A_t = a_t, S_t = s_t)$$

$$= \mathbb{P}_t(S_0 = s_0, A_0 = a_0, \ldots, S_t = s_t, A_t \in \mathcal{A})$$

$$= \mu(s_0)\pi_0(a_0; s_0) \prod_{i=1}^{t-1} p(\{s_i\} \times \mathcal{R}; s_{i-1}, a_{i-1})\pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i)$$

$$\cdot p(\{s_t\} \times \mathcal{R}; s_{t-1}, a_{t-1}) \sum_{a \in \mathcal{A}} \pi_t(a; s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t),$$

y el último sumatorio vale 1, por la definición de probabilidad condicionada,

$$\mathbb{P}_t(A_t = a_t | S_0 = s_0, A_t = a_t, \ldots, S_t = s_t)$$

$$= \frac{\mathbb{P}_t(S_0 = s_0, A_0 = a_0, \ldots, S_t = s_t, A_t = a_t)}{\mathbb{P}_t(S_0 = s_0, A_0 = a_0, \ldots, A_t = a_t, S_t = s_t)} = \pi_t(a_t; s_0, a_0, \ldots, s_t).$$

3.

$$\mathbb{P}_{t+1}(S_{t+1} = s_{t+1}, R_t = r_t | S_t = s, A_t = a)$$

$$= \frac{\mathbb{P}_{t+1}(S_t = s_t, A_t = a, S_{t+1} = s_{t+1}, R_t = r_t)}{\mathbb{P}_{t+1}(S_t = s, A_t = a)}$$

$$= \frac{\sum_{s_0, a_0, \ldots, s_{t-1}, a_t - 1} \mu(s_0)\pi_0(a_0; s_0) \prod_{i=1}^{t} p(\{s_i\} \times \mathcal{R}; s_{i-1}, a_{i-1})}{\sum_{s_0, a_0, \ldots, s_{t-1}, a_t - 1} \mu(s_0)\pi_0(a_0; s_0) \prod_{i=1}^{t} p(\{s_i\} \times \mathcal{R}; s_{i-1}, a_{i-1})}$$

$$\cdot \frac{\pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i)}{\pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i)} p(s_{t+1}, r_t; s, a) = p(s_{t+1}, r_t; s, a),$$

simplificando de manera análoga a la propiedad anterior.

□

En palabras, dado un estado s, la acción se muestrea según π, teniendo en cuenta las acciones y estados pasados, no las recompensas. El próximo estado y recompensa se muestrea según p, usando el estado actual y la acción tomada según π.

Como se comenta en la demostración, el siguiente paso consiste en extender esta probabilidad para el caso de trayectorias infinitas. Se define

$$\Omega = (\mathcal{S} \times \mathcal{A} \times \mathcal{R})^{\infty}$$

como el conjunto de todas las secuencias infinitas posibles de estados, acciones y recompensas, y también la σ-álgebra $\mathcal{F}$ asociada a Ω como

$$\mathcal{F} = (\overline{\mathcal{S}} \otimes \overline{\mathcal{A}} \otimes \overline{\mathcal{R}})^{\otimes\infty}.$$

El teorema de Kolmogorov (que se puede encontrar en la sección 36 de [1]) asegura, tras comprobar trivialmente que se cumplen sus conciciones, que existe una única medida de probabilidad $\mathbb{P}$ en el espacio de trayectorias infinitas $(\Omega, \mathcal{F})$, tal que $\mathbb{P}_T = \mathbb{P} \circ \Pi_T^{-1}$, es decir, que extiende todas las $\mathbb{P}_T$. Se denota $\mathbb{P}_\mu^\pi := \mathbb{P}$ para indicar la dependencia de μ y π. Esto da lugar a la existencia del espacio probabilístico $(\Omega, \mathcal{F}, \mathbb{P}_\mu^\pi)$, en el que se puede definir $(S_t, A_t, R_t)(\omega) = (s_t, a_t, r_t)$ para todo $t \in \mathbb{N}_0$ y $\omega = (s_0, a_0, r_0, s_1, a_1, r_1, \dots) \in \Omega$. Además, en estas condiciones, la probabilidad $\mathbb{P}_\mu^\pi$ cumple trivialmente las propiedades (2.3), (2.4), (2.5).

Definición 2.6. Dado un modelo de decisión de Markov $(\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$, una política π y una distribución inicial μ en $\mathcal{S}$, el proceso estocástico $\{(S_t, A_t)\}_{t\in\mathbb{N}_0}$ en $(\Omega, \mathcal{F}, \mathbb{P}_\mu^\pi)$ es denominado proceso de decisión de Markov en tiempo discreto. $(R_t)_{t\in\mathbb{N}_0}$ es el correspondiente proceso de recompensas.

Notación:

1. Si la distribución inicial de $\mathcal{S}$ y la política están claras, suelen eliminarse de las expresiones $\mathbb{E}_\mu^\pi$ y $\mathbb{P}_\mu^\pi$.
2. Si de antemano se selecciona un estado inicial $s \in \mathcal{S}$, en lugar de escribir $\mathbb{P}_{\delta_s}^\pi$, se escribe $\mathbb{P}_s^\pi$.

La expresión explícita (2.6) del cálculo de probabilidades de trayectorias de cualquier longitud, permite el cálculo de otras muchas probabilidades, como la calculada en la propiedad siguiente.

Propiedad 2.7. *Sean $(S_t, A_t)_{t\in\mathbb{N}_0}$ un proceso de decisión de Markov en $(\Omega, \mathcal{F}, \mathbb{P})$ y $(R_t)_{t\in\mathbb{N}_0}$ el proceso de recompensas correspondiente, ambos asociados a una política π y una distribución inicial μ. En estas condiciones se cumple que:*

$$\mathbb{P}((S_t, A_t, R_t, \ldots, S_T, A_T, R_T) = (s_t, a_t, r_t, \ldots, s_T, a_T, r_T) \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1})$$

$$= p(\{s_t\} \times \mathcal{R}; s_{t-1}, a_{t-1}) \cdot \pi_t(a_t; s_0, a_0, \ldots, a_{t-1}, s_t)$$

$$\cdot \prod_{i=t+1}^{T} p(s_i, r_{i-1}; s_{i-1}, a_{i-1})\pi_i(a_i; s_0, a_0, \ldots, a_{i-1}, s_i) \cdot p(\mathcal{S} \times \{r_T\}; s_T, a_T).$$

Demostración.

$$\mathbb{P}((S_t, A_t, R_t, \ldots, S_T, A_T, R_T) = (s_t, a_t, r_t, \ldots, s_T, a_T, r_T) \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1})$$

$$= \frac{\mathbb{P}(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_T = s_T, A_T = a_T, R_T = r_T)}{\mathbb{P}(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1})}$$

$$= \frac{\sum_{s_0,a_0,r_0,\ldots,s_{t-2},a_{t-2},r_{t-2},r_{t-1}} \mu(s_0)\pi_0(a_0; s_0) \prod_{i=1}^{T} p(s_i, r_{i-1}; s_{i-1}, a_{i-1})}{\sum_{s_0,a_0,r_0,\ldots,s_{t-2},a_{t-2},r_{t-2}} \mu(s_0)\pi_0(a_0; s_0) \prod_{i=1}^{t-1} p(s_i, r_{i-1}; s_{i-1}, a_{i-1})}$$

$$\cdot \frac{\pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i)p(\mathcal{S} \times \{r_T\}; s_T, a_T)}{\pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i)p(\mathcal{S} \times \mathcal{R}; s_{t-1}, a_{t-1})}$$

$$= \frac{\sum_{s_0,a_0,r_0,\ldots,s_{t-2},a_{t-2},r_{t-2}} \mu(s_0)\pi_0(a_0; s_0) \prod_{i=1}^{t-1} p(s_i, r_{i-1}; s_{i-1}, a_{i-1})}{\sum_{s_0,a_0,r_0,\ldots,s_{t-2},a_{t-2},r_{t-2}} \mu(s_0)\pi_0(a_0; s_0) \prod_{i=1}^{t-1} p(s_i, r_{i-1}; s_{i-1}, a_{i-1})}$$

$$\cdot \frac{\pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i)}{\pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i)} \sum_{r_{t-1}} p(s_t, r_{t-1}; s_{t-1}, a_{t-1})\pi_t(a_t; s_0, a_0, \ldots, a_{t-1}, s_t)$$

$$\cdot \prod_{i=t+1}^{T} p(s_i, r_{i-1}; s_{i-1}, a_{i-1})\pi_i(a_i; s_0, a_0, \ldots, a_{i-1}, s_i) \cdot p(\mathcal{S} \times \{r_T\}; s_T, a_T)$$

$$= p(\{s_t\} \times \mathcal{R}; s_{t-1}, a_{t-1}) \cdot \pi_t(a_t; s_0, a_0, \ldots, a_{t-1}, s_t)$$

$$\cdot \prod_{i=t+1}^{T} p(s_i, r_{i-1}; s_{i-1}, a_{i-1})\pi_i(a_i; s_0, a_0, \ldots, a_{i-1}, s_i) \cdot p(\mathcal{S} \times \{r_T\}; s_T, a_T).$$

□

Esta propiedad es análoga a la fórmula de cálculo de probabilidades para las cadenas de Markov (2.1).

Hay casos prácticos en los cuáles las recompensas están completamente determinadas por el estado y la acción. Por tanto, pese a ser un ligero abuso de notación, las siguientes redefiniciones como probabilidades marginales facilitan la comprensión.

Definición 2.8. Dado un modelo de decisión de Markov $(\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$, se definen las probabilidades de estado-acción-estado, así como las probabilidades de recompensa esperada y de estado-acción de la siguiente manera:

$$\begin{aligned} p(s'; s, a) &:= p(\{s'\} \times \mathcal{R}; s, a), \\ p(r; s, a) &:= p(\mathcal{S} \times \{r\}; s, a), \\ r(s, a) &:= \sum_{r \in \mathcal{R}} r \cdot p(r; s, a), \\ r(s, a, s') &:= \sum_{r \in \mathcal{R}} r \cdot \frac{p(s', r; s, a)}{p(s'; s, a)}, \end{aligned}$$

para cada $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$ cuando el denominador no es cero.

Propiedades 2.9.

1. $p(s', s; s, a) = \mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a)$.
2. $p(r; s, s) = \mathbb{P}(R_t = r; S_t = s, A_t = a)$.
3. $r(s, a) = \mathbb{E}(R_t \mid S_t = s, A_t = a)$.
4. *Si* $p(s'; s, a) > 0$, *entonces* $r(s, a, s') = \mathbb{E}(R_t \mid S_t = s, A_t = a, S_{t+1} = s')$.

Demostración.

1.

$$\mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a) = \sum_{r_t \in \mathcal{R}} \mathbb{P}(S_{t+1} = s', R_t = r_t | S_t = s, A_t = a)$$

$$= \sum_{r_t \in \mathcal{R}} p(s', r_t; s, a) = p(\{s'\} \times \mathcal{R}; s, a) = p(s', s; s, a).$$

2. Análogo a la demostración de la propiedad anterior.

3. Por las propiedades de la esperanza condicionada, y la propiedad 2,

$$\mathbb{E}(R_t \mid S_t = s, A_t = a) = \sum_{r \in \mathcal{R}} r\, \mathbb{P}(R_t = r \mid S_t = s, A_t = a)$$

$$= \sum_{r \in \mathcal{R}} r\, p(r; s, a) = r(s, a).$$

4. Usando la regla de Bayes y la primera propiedad, se obtiene que

$$\mathbb{E}(R_t \mid S_t = s, A_t = a, S_{t+1} = s')$$

$$= \sum_{r_t \in \mathcal{R}} r_t\, \mathbb{P}(R_t = r \mid S_t = s, A_t = a, S_{t+1} = s')$$

$$= \sum_{r_t \in \mathcal{R}} r_t\, \frac{\mathbb{P}(S_{t+1} = s', R_t = r_t \mid S_t = s, A_t = a)}{\mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a)} = \sum_{r_t \in \mathcal{R}} r_t \frac{p(s', r; s, a)}{p(s'; s, a)}$$

$$= r(s, a, s').$$

□

En muchos casos, el proceso finaliza cuando se alcanza un estado en particular. Estos casos se formalizan mediante la siguiente definición.

Definición 2.10. Se dice que un estado $s \in \mathcal{S}$ es terminal si se cumple que $p(s; s, a) = 1$ para todo $a \in \mathcal{A}_s$. El conjunto de todos los estados terminales se denota por Δ, y se asume que $p(s, 0; s, a) = 1$ para todo $s \in \Delta, a \in \mathcal{A}_s$. Es decir, en el momento en el que se alcanza un estado terminal, se deja de recibir recompensas.

2.3. Políticas de Markov y estacionarias

Una política define de qué manera elegir las acciones en función del estado actual. Sin embargo, no todas las políticas son igual de manejables desde el punto de vista computacional. Las políticas de Markov dependen únicamente del estado actual, lo que las hace más prácticas que aquellas que consideran todo el historial pasado. A su vez, si la política no varía con el tiempo, se dice que es estacionaria. Finalmente, una política es determinista si en cada estado elije una única acción con probabilidad 1.

El estudio de políticas que sean simultáneamente de Markov, estacionarias y deterministas es fundamental porque, bajo condiciones generales, existen políticas con estas propiedades que son óptimas, lo que reduce significativamente la complejidad del problema. Se formalizan estos conceptos en la siguiente definición.

Definición 2.11. Se dice que una política $\pi = (\pi_t)_{t\in\mathbb{N}}$ es:

(a) de Markov si existe una sucesión de núcleos $(\phi_t)_{t\in\mathbb{N}_0}$ en $\overline{\mathcal{A}}\times\mathcal{S}$, es decir, $\phi_t(\cdot; s)$ es una distribución de probabilidad en $\mathcal{A}$ para todo $s \in \mathcal{S}$, tal que:

$$\pi_t(\cdot; s_0, a_0, \ldots, s_t) = \phi_t(\cdot; s_t), \quad \forall (s_0, a_0, \ldots, s_t) \in (\mathcal{S}\times\mathcal{A})^{t-1}\times\mathcal{S}.$$

El conjunto de todas las políticas de Markov se denota por Π_M.

(b) estacionaria si existe un núcleo ϕ en $\overline{\mathcal{A}}\times\mathcal{S}$ tal que:

$$\pi_t(\cdot; s_0, a_0, \ldots, s_t) = \phi(\cdot; s_t), \quad \forall (s_0, a_0, \ldots, s_t) \in (\mathcal{S}\times\mathcal{A})^{t-1}\times\mathcal{S}.$$

El conjunto de todas las políticas estacionarias se denota por Π_S.

(c) estacionaria determinista si existe un núcleo ϕ en $\overline{\mathcal{A}}\times\mathcal{S}$ que solo toma valores en $\{0, 1\}$ tal que:

$$\pi_t(\cdot; s_0, a_0, \ldots, s_t) = \phi(\cdot; s_t), \quad \forall (s_0, a_0, \ldots, s_t) \in (\mathcal{S}\times\mathcal{A})^{t-1}\times\mathcal{S}.$$

El conjunto de todas las políticas estacionarias deterministas se denota por Π_S^D.

En todos los casos estacionarios, suele denotarse π en lugar de ϕ. De las definiciones, trivialmente se deduce que:

$$\Pi_S^D \subseteq \Pi_S \subseteq \Pi_M.$$

En los MDPs, el término Markov hace referencia a la propiedad de Markov. En particular, cuando se trabaja con cualquiera de estas políticas, el proceso resultante conserva dicha propiedad, lo que implica que el futuro del sistema depende únicamente del estado actual y no del historial observado. Este principio se extiende tanto al proceso (S, A), que forma una cadena de Markov homogénea, como al proceso completo que incluye las recompensas, (S, A, R). Se formaliza esta idea mediante los siguientes resultados.

Proposición 2.12 (Propiedad de Markov)**.** *Si $\pi \in \Pi_S$, entonces $\{(S_t, A_t)\}_{t\in\mathbb{N}}$ es una cadena de Markov homogénea en $\mathcal{S} \times \mathcal{A}$ con transiciones:*

$$p_{(s,a),(s',a')} = p(s'; s, a) \cdot \pi(a'; s').$$

Demostración. Denotemos $\mathbb{P} = \mathbb{P}_\mu^\pi$. La propiedad de Markov se puede demostrar para todas las políticas $\pi \in \Pi_M$:

$$\mathbb{P}(S_{t+1} = s_{t+1}, A_{t+1} = a_{t+1} \mid S_0 = s_0, A_0 = a_0, \ldots, S_t = s_t, A_t = a_t)$$
$$= \frac{\mathbb{P}(S_0 = s_0, A_0 = a_0, \ldots, S_t = s_t, A_t = a_t, S_{t+1} = s_{t+1}, A_{t+1} = a_{t+1})}{\mathbb{P}((S_0 = s_0, A_0 = a_0, \ldots, S_t = s_t, A_t = a_t)}$$

$$= \frac{\mu(s_0)\pi_0(a_0; s_0)\prod_{i=1}^{t+1} p(\{s_i\} \times \mathcal{R}; s_{i-1}, a_{i-1})\pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i)}{\mu(s_0)\pi_0(a_0; s_0)\prod_{i=1}^{t} p(\{s_i\} \times \mathcal{R}; s_{i-1}, a_{i-1})\pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i)}$$

$$\cdot\frac{p(\mathcal{S} \times \mathcal{R}; s_{t+1}, a_{t+1})}{p(\mathcal{S} \times \mathcal{R}; s_t, a_t)} = p(\{s_{t+1}\} \times \mathcal{R}; s_t, a_t)\pi_{t+1}(a_{t+1}; s_0, a_0, ..., a_t, s_{t+1})$$
$$= p(s_{t+1}; s_t, a_t)\phi_{t+1}(a_{t+1}; s_{t+1})$$
$$= \mathbb{P}(S_{t+1} = s_{t+1}, A_{t+1} = a_{t+1} \mid S_t = s_t, A_t = a_t).$$

Además, si $\pi \in \Pi_S$, esta última igualdad proporciona la probabilidad de transición:

$$\mathbb{P}(S_{t+1} = s_{t+1}, A_{t+1} = a_{t+1} \mid S_t = s_t, A_t = a_t) = p(s_{t+1}; s_t, a_t) \cdot \pi(a_{t+1}; s_{t+1}).$$

□

Proposición 2.13. *Si $\pi \in \Pi_S$, entonces $\{(S_t, A_t, R_t)\}_{t\in\mathbb{N}}$ cumple la propiedad de Markov para procesos con recompensa:*

$$\mathbb{P}((S_{t+1}, A_{t+1}, R_t) = (s_{t+1}, a_{t+1}, r_t) \mid S_0 = s_0, A_0 = a_0, \ldots, S_t = s_t, A_t = a_t)$$

$$= \mathbb{P}((S_{t+1}, A_{t+1}, R_t) = (s_{t+1}, a_{t+1}, r_t) \mid S_t = s_t, A_t = a_t).$$

Además, las probabilidades de transición son:

$$p_{(s,a),(s',a',r)} = p(s', r; s, a) \cdot \pi(a'; s').$$

Demostración.

$$\mathbb{P}((S_{t+1}, A_{t+1}, R_t) = (s_{t+1}, a_{t+1}, r_t) \mid S_0 = s_0, A_0 = a_0, \ldots, S_t = s_t, A_t = a_t)$$

$$= \frac{\mathbb{P}(S_0 = s_0, A_0 = a_0, \ldots, S_t = s_t, A_t = a_t, S_{t+1} = s_{t+1}, R_t = r_t, A_{t+1} = a_{t+1})}{\mathbb{P}((S_0 = s_0, A_0 = a_0, \ldots, S_t = s_t, A_t = a_t)}$$

$$= \frac{\mu(s_0)\pi_0(a_0; s_0)\prod_{i=1}^{t} p(\{s_i\} \times \mathcal{R}; s_{i-1}, a_{i-1})\pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i)}{\mu(s_0)\pi_0(a_0; s_0)\prod_{i=1}^{t} p(\{s_i\} \times \mathcal{R}; s_{i-1}, a_{i-1})\pi_i(a_i; s_0, a_0, ..., a_{i-1}, s_i)}$$

$$\cdot \frac{p(s_{t+1}, r_t; s_t, a_t)\pi_{t+1}(a_{t+1}; s_0, a_0, \ldots, s_{t+1})p(\mathcal{S} \times \mathcal{R}; s_{t+1}, a_{t+1})}{p(\mathcal{S} \times \mathcal{R}; s_t, a_t)}$$

$$= p(s_{t+1}, r_t; s_t, a_t)\phi_{t+1}(a_{t+1}; s_{t+1})$$

$$= \mathbb{P}((S_{t+1}, A_{t+1}, R_t) = (s_{t+1}, a_{t+1}, r_t) \mid S_t = s_t, A_t = a_t).$$

Si $\pi \in \Pi_S$, esta última igualdad proporciona la probabilidad de transición:

$$\mathbb{P}((S_{t+1}, A_{t+1}, R_t) = (s', a', r) \mid S_t = s_t, A_t = a_t) = p(s', r; s, a)\pi(a'; s').$$

□

La propiedad de Markov en los MDPs es fundamental para los problemas de aprendizaje por refuerzo, ya que garantiza que el estado actual contiene toda la información relevante del pasado para la toma de decisiones óptima. Gracias a ello, se pueden demostrar resultados clave sobre la existencia y caracterización de soluciones óptimas. Esta propiedad también permite formular los problemas de manera recursiva, facilitando el uso de técnicas dinámicas que se estudiaran en los siguientes capítulos.

Capítulo 3

Teoría de control estocástico

En el presente capítulo, se abordará la teoría de control estocástico aplicada a los MDPs, que fueron desarrollados en el capítulo anterior. El control estocástico proporciona las bases teóricas necesarias para formular y resolver problemas de decisión secuenciales en entornos inciertos, sentando así los cimientos del aprendizaje por refuerzo desde una perspectiva matemática rigurosa.

En particular, este capítulo está enfocado en la derivación y comprensión de las ecuaciones de Bellman, un conjunto de expresiones recursivas fundamentales en el análisis y resolución de MDPs. Se presentarán también los operadores asociados a estas ecuaciones, cuyo papel es crucial para garantizar la existencia y unicidad de soluciones, así como para guiar los procedimientos computacionales en los que se basa el aprendizaje por refuerzo. La optimización de la función de valor será el objetivo principal de este desarrollo, ya que constituye el criterio clave para identificar una política óptima, es decir, una estrategia que maximice la recompensa esperada a lo largo del tiempo.

Al final del capítulo se presenta una demostración formal del teorema del algoritmo de programación dinámica. Este resultado establece la existencia y caracterización de una política óptima que además es estacionaria y determinista. Esta caracterización es esencial ya que garantiza que se puedan tomar decisiones óptimas basadas únicamente en el estado actual, asignando además una única acción óptima a cada estado. Este teorema sustenta una técnica esencial tanto para resolver MDPs como para implementar algoritmos

de aprendizaje por refuerzo.

3.1. Contexto

En el capítulo anterior, se trabajó con MDPs una vez se había fijado una política π y una función de transición p. Sin embargo, en el ámbito del aprendizaje por refuerzo se busca una política que optimice la recompensa esperada. Para ello hay que formalizar este concepto.

Dado un modelo de recompensa de Markov, y un tiempo terminal T, el objetivo es encontrar una política π que maximice la suma de recompensas futuras:

$$\mathbb{E}^{\pi}\left[\sum_{t=0}^{T} R_t\right].$$

En estas condiciones es vital la elección del tiempo terminal T:

- Dado un estado $s' \in \mathcal{S}$, puede elergirse $T = min\{t : S_t = s'\}$, es decir, se consideran las recompensas esperadas hasta que se alcanza por primera vez el estado s'.
- Puede considerarse $T \sim \text{Geo}(1-\gamma)$, con $\gamma \in (0,1)$, para alguna variable aleatoria geométrica T independiente de (S, A, R). Bajo estas condiciones, la probabilidad de que el proceso acabe es independiente del tiempo en el que se encuentre. Integrando sobre el horizonte temporal independiente, se obtiene que, como R_t es independiente de T, por la definición de esperanza condicionada,

$$\mathbb{E}^{\pi}\left[\sum_{t=0}^{T} R_t\right] = \sum_{k=0}^{\infty} \mathbb{E}^{\pi}\left[\sum_{t=0}^{T} R_t \mid T = k\right] \mathbb{P}^{\pi}(T = k)$$

$$= \sum_{k=0}^{\infty} \mathbb{E}^{\pi}\left[\sum_{t=0}^{k} R_t\right] (1-\gamma)\gamma^k = (1-\gamma) \sum_{k=0}^{\infty} \sum_{t=0}^{k} \mathbb{E}^{\pi}\left[R_t\right] \gamma^k$$

$$= (1-\gamma) \sum_{t=0}^{\infty} \sum_{k=t}^{\infty} \mathbb{E}^{\pi}\left[R_t\right] \gamma^k = (1-\gamma) \sum_{t=0}^{\infty} \mathbb{E}^{\pi}\left[R_t\right] \sum_{k=t}^{\infty} \gamma^k$$

$$= (1-\gamma)\sum_{t=0}^{\infty}\mathbb{E}^{\pi}\left[R_t\right]\gamma^t\frac{1}{1-\gamma} = \sum_{t=0}^{\infty}\gamma^t\,\mathbb{E}^{\pi}\left[R_t\right] = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t R_t\right],$$

usando Fubini/Tonelli en dos ocasiones asumiendo recompensas acotadas y razonando por la convergencia dominada. Esto se hace ya que a lo largo del capítulo se tomará $\mathcal{R}$ finito, como se enunciará más adelante. Esta fórmula demuestra que equivale optimizar la suma de recompensas esperadas hasta un tiempo terminal geométrico finito, que optimizar la suma de recompensas descontadas esperadas.

El parámetro $\gamma \in (0,1)$ se conoce como factor de descuento y representa cómo se valoran las recompensas futuras en comparación con las inmediatas. Dado que una recompensa obtenida en el tiempo t se multimplica por γ^t, un factor más próximo a 1 valorará más las recompensas lejanas que uno próximo a 0. Este capítulo se centrará en esta elección de T, conocida por horizonte temporal infinito con factor de descuento γ.

- Cabe mencionar que también existe el caso en el que el horizonte temporal $T \in \mathbb{N}$ es fijo y determinista. Este se conoce como el problema de control estocástico en tiempo finito. En este caso, el horizonte temporal tiene una duración fija previamente definida, lo que genera políticas óptimas $(\pi_t)_{t\leq T}$ que son no estacionarias. Esto contrasta con el caso de un horizonte temporal geométrico, donde la ausencia de memoria en las variables aleatorias geométricas permite obtener políticas óptimas estacionarias como se demostrará a lo largo del capítulo.

3.2. Introducción a las ecuaciones de Bellman

A partir de este momento se asume que los conjuntos $\mathcal{S}, \mathcal{A}$ y $\mathcal{R}$ son finitos. Además las σ-álgebras consideradas en las distribuciones de probabilidad serán sus conjuntos potencia. También se considera un horizonte temporal infinito con factor de descuento γ.

Se introduce la notación $\mathbb{P}^{\pi}_{s,a}$, que indica la probabilidad del proceso cuyo estado inicial es s y cuya primera acción es a. A su vez, $\mathbb{E}^{\pi}_{s,a}$ representa la esperanza asociada a dicha probabilidad.

El concepto de recompensa esperada descontada se formaliza en la siguiente definición.

Definición 3.1 (Funciones de valor)**.** Dada una política $\pi \in \Pi$ y $\gamma \in (0,1)$ un factor de descuento, se denomina función de valor de estado-acción a la función $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ definida por:

$$Q^\pi(s,a) = \mathbb{E}^\pi_{s,a}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right].$$

Es decir, representa la recompensa esperada total considerando la política π, el estado inicial s y la acción inicial a.

La función de valor de estado se define como

$$V^\pi(s) = \mathbb{E}^\pi_s\left[\sum_{t=0}^{\infty} \gamma^t R_t\right] = \sum_{a \in \mathcal{A}_s} \pi_0(a;s) Q^\pi(s,a). \tag{3.1}$$

En este caso, denota la recompensa esperada total considerando sólo la política π y el estado inicial s.

Las funciones de valor son fundamentales en los problemas de decisión de Markov, ya que evalúan cuantitativamente la calidad de los estados y las acciones. Por este motivo permiten la comparación entre políticas y son la base de algoritmos de aprendizaje por refuerzo.

Si se consideran solo políticas estacionarias, se tiene que las funciones de valor cumplen unos determinados sistemas de ecuaciones, conocidos como ecuaciones de Bellman. Estas ecuaciones proporcionan una relación recursiva que describe cómo se pueden calcular las funciones de valor y las soluciones óptimas, como se verá a lo largo del capítulo.

Proposición 3.2 (Ecuaciones de Bellman)**.** *Dada $\pi \in \Pi_S$, entonces se cumple que:*

$$Q^\pi(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}_{s'}} p(s';s,a)\pi(a';s')Q^\pi(s',a'), \quad s \in \mathcal{S}, a \in \mathcal{A}$$

$$V^\pi(s) = \sum_{a \in \mathcal{A}_s} \pi(a;s)\left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s';s,a)V^\pi(s')\right), \quad s \in \mathcal{S}.$$

Demostración. Como $\pi \in \Pi_S$, por la proposición 2.13 del capítulo anterior, (S, A, R) es una cadena de Markov con recompensas sobre $\mathcal{S} \times \mathcal{A} \times \mathcal{R}$. Por tanto, dados $s, s' \in \mathcal{S}$ y $a, a' \in \mathcal{A}$, usando la fórmula 2.2 obtenemos que:

$$\mathbb{E}^{\pi}_{s,a}\left[R_1 + \gamma R_2 + \gamma^2 R_3 + \cdots \mid S_1 = s', A_1 = a'\right]$$

$$= \mathbb{E}^{\pi}_{s',a'}\left[R_0 + \gamma R_1 + \gamma^2 R_2 + \cdots\right] = \mathbb{E}^{\pi}_{s',a'}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right] = Q^{\pi}(s', a').$$

Usando esta igualdad y la ley de la probabilidad total,

$$Q^{\pi}(s,a) = \mathbb{E}^{\pi}_{s,a}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right] = \mathbb{E}^{\pi}_{s,a}\left[R_0\right] + \mathbb{E}^{\pi}_{s,a}\left[\gamma R_1 + \gamma^2 R_2 + \gamma^3 R_3 + \cdots\right]$$

$$= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}_{s'}} \mathbb{E}^{\pi}_{s,a}\left[R_1 + \gamma R_2 + \gamma^2 R_3 + \cdots \mid S_1 = s', A_1 = a'\right]$$

$$\cdot \mathbb{P}^{\pi}_{s,a}(S_1 = s', A_1 = a') = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}_{s'}} \mathbb{P}^{\pi}_{s,a}(S_1 = s', A_1 = a') Q^{\pi}(s', a')$$

$$= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}_{s'}} p(s'; s, a)\pi(a'; s') Q^{\pi}(s', a'),$$

y se concluye la primera ecuación. Para la segunda, es necesario tener en cuenta la primera ecuación, y la definición de V^{π} (3.1), teniendo en cuenta que la política es estacionaria:

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}_s} \pi(a; s) Q^{\pi}(s, a)$$

$$= \sum_{a \in \mathcal{A}_s} \pi(a; s) \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}_{s'}} p(s'; s, a)\pi(a'; s') Q^{\pi}(s', a') \right)$$

$$= \sum_{a \in \mathcal{A}_s} \pi(a; s) \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) V^{\pi}(s') \right).$$

□

El siguiente resultado permite, en el caso de las políticas estacionarias, escribir Q^{π} como función de V^{π}.

Corolario 3.3. *Dada una política estacionaria $\pi \in \Pi_S$, se puede escribir la función de valor de estado-acción como:*

$$Q^\pi(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s';s,a) V^\pi(s'), \quad s \in \mathcal{S}, a \in \mathcal{A}. \tag{3.2}$$

Demostración. Dados $s \in \mathcal{S}$ y $a \in \mathcal{A}$, a partir de la ecuación de Bellman para Q^π, y la definición de V^π (3.1) en el caso estacionario se obtiene:

$$\begin{aligned} Q^\pi(s,a) &= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}_{s'}} p(s';s,a)\pi(a';s')Q^\pi(s',a') \\ &= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s';s,a) V^\pi(s'). \end{aligned}$$

□

Para afrontar el capítulo, es necesario sentar unas bases previas de análisis funcional que se pueden encontrar en el apéndice C, cuyo resultado principal es el Teorema del Punto Fijo de Banach B.3.

En el caso de los MDPs las contracciones son habituales y proporcionan algoritmos convergentes gracias al Teorema del Punto Fijo de Banach. Con este objetivo se define $X := \{v : \mathcal{S} \to \mathbb{R}\}$ y se considera el espacio $(X, \|\cdot\|_\infty)$. Como $\mathcal{S}$ es finito, es compacto y por tanto es un espacio de Banach. Del mismo modo, $(Y, \|\cdot\|_\infty)$ es un espacio de Banach, donde $Y := \{q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$.

Definición 3.4 (Operadores de esperanza de Bellman)**.** Dado un modelo de decisión de Markov y una política estacionaria $\pi \in \Pi_S$, los operadores $T^\pi : X \to X$ dado por

$$T^\pi(v)(s) := \sum_{a \in \mathcal{A}_s} \pi(a;s) \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s';s,a) v(s') \right), \quad \forall s \in \mathcal{S}$$

y $T^\pi : Y \to Y$ dado por

$$T^\pi(q)(s,a) := r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}_{s'}} p(s';s,a)\pi(a';s')q(s',a'), \quad \forall s \in \mathcal{S}, a \in \mathcal{A},$$

se conocen por operadores de esperanza de Bellman.

Estos operadores de esperanza emergen de manera natural de las ecuaciones de Bellman 3.2. Por lo tanto, si π es una política estacionaria, entonces V^π y Q^π son puntos fijos de los operadores de esperanza de Bellman.

Teorema 3.5. *Ambos operadores de esperanza de Bellman son contracciones con constante de contracción γ. Es decir, para todo $v_1, v_2 \in X$, $q_1, q_2 \in Y$, se cumple que:*

$$\|T^\pi(v_1) - T^\pi(v_2)\|_\infty \leq \gamma \|v_1 - v_2\|_\infty \quad y$$
$$\|T^\pi(q_1) - T^\pi(q_2)\|_\infty \leq \gamma \|q_1 - q_2\|_\infty.$$

Demostración. Sean $v_1, v_2 \in X$. Entonces

$$\|T^\pi(v_1) - T^\pi(v_2)\|_\infty = \max_{s\in\mathcal{S}} |T^\pi(v_1)(s) - T^\pi(v_2)(s)|$$

$$= \gamma \max_{s\in\mathcal{S}} \left| \sum_{a\in\mathcal{A}_s} \pi(a;s) \left(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s';s,a)(v_1(s') - v_2(s')) \right) \right|$$

$$\leq \gamma \|v_1 - v_2\|_\infty \cdot \max_{s\in\mathcal{S}} \left| \sum_{a\in\mathcal{A}_s} \pi(a;s) \sum_{s'\in\mathcal{S}} p(s';s,a) \right| = \gamma \|v_1 - v_2\|_\infty.$$

Para el segundo operador, dados $q_1, q_2 \in Y$,

$$\|T^\pi(q_1) - T^\pi(q_2)\|_\infty = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |T^\pi(q_1)(s,a) - T^\pi(q_2)(s,a)|$$

$$= \gamma \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| \sum_{s'\in\mathcal{S}} p(s';s,a) \sum_{a'\in\mathcal{A}_{s'}} \pi(a';s')(q_1(s',a') - q_2(s',a')) \right|$$

$$\leq \gamma \|q_1 - q_2\|_\infty \cdot \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| \sum_{s'\in\mathcal{S}} p(s';s,a) \sum_{a'\in\mathcal{A}_{s'}} \pi(a';s') \right| = \gamma \|q_1 - q_2\|_\infty.$$

□

Las ecuaciones de Bellman demostradas en la proposición 3.2, aseguran que las funciones de valor son puntos fijos para los operadores de esperanza de Bellman. Por tanto, usando el teorema anterior y el Teorema del Punto Fijo de Banach, dichos puntos fijos son únicos. Es decir, las ecuaciones de Bellman son sistemas de ecuaciones lineales con solución única. En lugar de proceder

usando métodos de álgebra lineal, es más razonable usar la iteración de punto fijo de Banach, esto es, iterar $v_{n+1} = T^\pi v_n$ para algún vector inicial v_0 (o $q_{n+1} = T^\pi q_n$ para alguna matriz inicial q_0). Este razonamiento dará lugar en el próximo capítulo a algoritmos tabulares capaces de converger a una solución óptima.

3.3. Optimización de Bellman para MDPs

Tras analizar las ecuaciones de Bellman para políticas estacionarias fijas, en esta sección se demostrará que es posible encontrar una política óptima, en el sentido de que maximice las funciones de valor.

Definición 3.6. Fijado un modelo de decisión de Markov, entonces:

(I) La función $V^* : \mathcal{S} \to \mathbb{R}$ dada por

$$V^*(s) := \sup_{\pi \in \Pi} V^\pi(s), \quad s \in \mathcal{S},$$

se llama función de valor de estado óptima.

(II) La función $Q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ dada por

$$Q^*(s,a) := \sup_{\pi \in \Pi} Q^\pi(s,a), \quad s \in \mathcal{S},\ a \in \mathcal{A},$$

se llama función de valor de estado-acción óptima.

(III) Una política $\pi^* \in \Pi$ que satisface

$$V^{\pi^*}(s) = V^*(s), \quad s \in \mathcal{S},$$

se llama política óptima.

Cabe destacar que asumiendo un modelo de Markov finito, entonces los superiores de las definiciones son máximos que se alcanzan siempre. Además, a priori, V^* y Q^* no representan las funciones de valor de la mejor política, sino que son las mejores recompensas esperadas posibles punto por punto. Por tanto, no está claro si existe una política que sea óptima para todos los estados iniciales. Sin embargo, los resultados de Bellman demuestran que existe una política estacionaria, e incluso determinista, que puede caracterizarse resolviendo un conjunto de ecuaciones no lineales. Para ello, se introducen los

operadores de optimalidad de Bellman, que operan sobre los espacios X e Y definidos previamente.

Definición 3.7 (Operadores de optimalidad de Bellman)**.** Dado un modelo de Markov, entonces:

(I) El sistema de ecuaciones no lineales

$$v(s) = \max_{a \in \mathcal{A}_s} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) v(s') \right\}, \quad s \in \mathcal{S},$$

se denomina sistema de optimalidad de Bellman. El operador T^*, definido como:

$$(T^*)(v)(s) := \max_{a \in \mathcal{A}_s} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) v(s') \right\}, \quad s \in \mathcal{S},$$

se llama operador de optimalidad de Bellman para funciones de valor de estado.

(II) El sistema de ecuaciones no lineales

$$q(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) \max_{a' \in \mathcal{A}_{s'}} q(s', a'), \quad (s,a) \in \mathcal{S} \times \mathcal{A},$$

se denomina sistema de optimalidad de estado-acción de Bellman. El operador T^*, definido por:

$$(T^*)(q)(s,a) := r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) \max_{a' \in \mathcal{A}_{s'}} q(s', a'), \quad (s,a) \in \mathcal{S} \times \mathcal{A},$$

se llama operador de optimalidad de Bellman para funciones de valor de estado-acción.

Pese a que ambos operadores comparten notación, T^*, se pueden distinguir inequívocamente por el número de argumentos.

Al tratarse de sistemas no lineales, no es sencillo obtener las soluciones. Sin embargo, se probará que los operadores de optimalidad de Bellman están directamente relacionados con las funciones de valor óptimas, y que por tanto resolver estas ecuaciones equivale a encontrar una política óptima.

Proposición 3.8. *Los operadores de optimalidad de Bellman son crecientes.*

Demostración. Sean $v_1, v_2 \in X$ tal que $v_1 \leq v_2$. Dado $s' \in \mathcal{S}$, se tiene que $v_1(s') \leq v_2(s')$. Teniendo en cuenta que $p(s'; s, a) \geq 0 \ \forall s \in \mathcal{S}, a \in \mathcal{A}$, que $\gamma > 0$ y que $r(s, a)$ es finito $\forall s \in \mathcal{S}, a \in \mathcal{A}$, se obtiene que

$$r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) v_1(s') \leq r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) v_2(s').$$

Tomando el máximo sobre todas las acciones $a \in \mathcal{A}_s$ se concluye que $\forall s \in \mathcal{S}$:

$$T^*(v_1)(s) = \max_{a \in \mathcal{A}_s} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) v_1(s') \right\}$$

$$\leq \max_{a \in \mathcal{A}_s} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) v_2(s') \right\} = T^*(v_2)(s).$$

Dados $q_1, q_2 \in Y$ tal que $q_1 \leq q_2$, teniendo en cuenta las mismas consideraciones que antes, se tiene que para todo $s \in \mathcal{S}$ y para todo $a \in \mathcal{A}$,

$$(T^*)(q_1)(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) \max_{a' \in \mathcal{A}_{s'}} q_1(s', a')$$

$$\leq r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) \max_{a' \in \mathcal{A}_{s'}} q_2(s', a') = (T^*)(q_2)(s,a).$$

Por tanto $(T^*)(q_1) \leq (T^*)(q_2)$, y se concluye que T^* es creciente.

□

El siguiente paso será demostrar que los operadores de optimalidad son contracciones. Para ello, como son operadores no lineales que incluyen la función máximo, es necesario introducir el siguiente lema.

Lema 3.9. *Dadas $f, g : \mathcal{A} \to \mathbb{R}$, se cumple la siguiente desigualdad:*

$$|\max_{a \in \mathcal{A}} f(a) - \max_{a \in \mathcal{A}} g(a)| \leq \max_{a \in \mathcal{A}} |f(a) - g(a)|.$$

Demostración. Dado $a' \in \mathcal{A}$, por la definición de valor absoluto se tiene que $f(a') - g(a') \leq |f(a') - g(a')|$. Luego

$$f(a') \leq g(a') + |f(a') - g(a')| \leq \max_{a \in \mathcal{A}} g(a) + \max_{a \in \mathcal{A}} |f(a) - g(a)|.$$

Como esta cota es cierta para todo $a' \in \mathcal{A}$, se cumple que

$$\max_{a\in\mathcal{A}} f(a) - \max_{a\in\mathcal{A}} g(a) \leq \max_{a\in\mathcal{A}} |f(a) - g(a)|.$$

Intercambiando los roles de f y g, se deduce que

$$\max_{a\in\mathcal{A}} g(a) - \max_{a\in\mathcal{A}} f(a) \leq \max_{a\in\mathcal{A}} |g(a) - f(a)| \leq \max_{a\in\mathcal{A}} |f(a) - g(a)|.$$

Por lo tanto, juntando ambas desigualdades,

$$-\max_{a\in\mathcal{A}} |f(a) - g(a)| \leq \max_{a\in\mathcal{A}} f(a) - \max_{a\in\mathcal{A}} g(a) \leq \max_{a\in\mathcal{A}} |f(a) - g(a)|,$$

y se concluye el lema.

□

Teorema 3.10. *Los operadores de optimalidad de Bellman son contracciones con constante de contracción γ. Es decir, para todo $v_1, v_2 \in X$, $q_1, q_2 \in Y$, se cumple que:*

$$\|T^*(v_1) - T^*(v_2)\|_\infty \leq \gamma \|v_1 - v_2\|_\infty \quad y$$
$$\|T^*(q_1) - T^*(q_2)\|_\infty \leq \gamma \|q_1 - q_2\|_\infty.$$

Demostración. Dados $v_1, v_2 \in X$, por el lema 3.9 se tiene que

$$\|T^*(v_1) - T^*(v_2)\|_\infty = \max_{s\in\mathcal{S}} |T^*(v_1)(s) - T^*(v_2)(s)|$$

$$= \max_{s\in\mathcal{S}} |\max_{a\in\mathcal{A}} [r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s';s,a)v_1(s')] - \max_{a\in\mathcal{A}} [r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s';s,a)v_2(s')]|$$

$$\leq \max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \gamma \sum_{s'\in\mathcal{S}} p(s';s,a) \cdot |v_1(s') - v_2(s')|$$

$$\leq \max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \gamma \sum_{s'\in\mathcal{S}} p(s';s,a) \cdot \|v_1 - v_2\|_\infty = \gamma \|v_1 - v_2\|_\infty,$$

puesto que $\sum_{s'\in\mathcal{S}} p(s';s,a) = 1$.

Ahora, dados $q_1, q_2 \in Y$, y aplicando la desigualdad triangular y el lema 3.9 en la primera desigualdad, se tiene que:

$$\|T^*(q_1) - T^*(q_2)\|_\infty = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |T^*(q_1)(s,a) - T^*(q_2)(s,a)|$$

$$= \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\gamma \sum_{s'\in\mathcal{S}} p(s';s,a)[\max_{a'\in\mathcal{A}_{s'}} q_1(s',a') - \max_{a'\in\mathcal{A}_{s'}} q_2(s',a')]|$$

$$\leq \gamma \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{s'\in\mathcal{S}} p(s';s,a) \max_{a'\in\mathcal{A}_{s'}} |q_1(s',a') - q_2(s',a)|$$

$$\leq \gamma \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{s'\in\mathcal{S}} p(s';s,a)\|q_1 - q_2\|_\infty = \gamma \|q_1 - q_2\|_\infty.$$

Por tanto, ambos operadores son contracciones.

□

Por el Teorema del Punto Fijo de Banach, estos operadores tienen puntos fijos únicos. El siguiente resultado, que es el más importante del capítulo, demuestra que dichos puntos fijos son las funciones de valor óptimas.

Teorema 3.11. *Las funciones de valor óptimas son los únicos puntos fijos de los operadores de optimalidad de Bellman. Es decir, $T^*V^* = V^*$ y $T^*Q^* = Q^*$. Además se cumple que $V^*(s) = \max_{a\in\mathcal{A}_s} Q^*(s,a)$.*

Demostración. Se hará la prueba en dos pasos. En primer lugar se probará para $\overline{Q}^*(s,a) = \sup_{\pi\in\Pi_S} Q^\pi(s,a)$ y $\overline{V}^*(s) = \sup_{\pi\in\Pi_S} V^\pi(s)$. Después se trabajará considerando todas las posibles políticas.

- Dados $s \in \mathcal{S}$ y $a \in \mathcal{A}$,

$$\overline{Q}^*(s,a) = \sup_{\pi\in\Pi_S} Q^\pi(s,a) = r(s,a) + \gamma \sup_{\pi\in\Pi_S} \sum_{s'\in\mathcal{S}} p(s';s,a)$$

$$\cdot \sum_{a'\in\mathcal{A}_{s'}} \pi(a';s')Q^\pi(s',a').$$

  Como para toda $\pi \in \Pi_S$ se tiene que

$$\sum_{s'\in\mathcal{S}} p(s';s,a) \sum_{a'\in\mathcal{A}_{s'}} \pi(a';s')Q^\pi(s',a') \leq \sum_{s'\in\mathcal{S}} p(s';s,a) \max_{a'\in\mathcal{A}_{s'}} Q^\pi(s',a')$$

$$\leq \sum_{s'\in\mathcal{S}} p(s';s,a) \max_{a'\in\mathcal{A}_{s'}} \sup_{\pi\in\Pi_S} Q^\pi(s',a'),$$

  entonces

$$\sup_{\pi\in\Pi_S} \sum_{s'\in\mathcal{S}} p(s';s,a) \sum_{a'\in\mathcal{A}_{s'}} \pi(a';s')Q^\pi(s',a') \leq \sum_{s'\in\mathcal{S}} p(s';s,a) \max_{a'\in\mathcal{A}_{s'}} \overline{Q}^*(s',a').$$

Sea π^* una política determinista que dado un estado s' elije con probabilidad 1 cierto $a^*(s') \in \arg\max_{a' \in \mathcal{A}_{s'}} \overline{Q}^*(s', a')$.
Entonces se tiene que para esta política,

$$\sum_{s' \in \mathcal{S}} p(s'; s, a) \sum_{a' \in \mathcal{A}_{s'}} \pi^*(a'; s') Q^{\pi^*}(s', a') = \sum_{s' \in \mathcal{S}} p(s'; s, a) \max_{a' \in \mathcal{A}_{s'}} \overline{Q}^*(s', a').$$

Por tanto, se alcanza la igualdad y

$$\overline{Q}^*(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) \max_{a' \in \mathcal{A}_{s'}} \overline{Q}^*(s', a').$$

- Demostrando que $\overline{V}^*(s) = \max_{a \in \mathcal{A}_s} \overline{Q}^*(s, a)$ para todo $s \in \mathcal{S}$, entonces está claro que

$$T^* \overline{V}^*(s) = \max_{a \in \mathcal{A}_s} \left\{ r(s; a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) \overline{V}^*(s') \right\}$$

$$= \max_{a \in \mathcal{A}_s} \left\{ r(s; a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) \max_{a' \in \mathcal{A}_{s'}} \overline{Q}^*(s', a') \right\}$$

$$= \max_{a \in \mathcal{A}_s} \overline{Q}^*(s, a) = \overline{V}^*(s).$$

Recordando que $\overline{V}^*(s) = \sum_{\pi \in \Pi_S} V^\pi(s)$, entonces dados $\pi \in \Pi_S$ y $s \in \mathcal{S}$ se tiene que

$$V^\pi(s) = \sum_{a \in \mathcal{A}_s} \pi_0(a; s) Q^\pi(s, a) \leq \max_{a \in \mathcal{A}_s} Q^\pi(s, a) \leq \max_{a \in \mathcal{A}_s} \overline{Q}^*(s, a)$$

$$\implies \overline{V}^*(s) \leq \max_{a \in \mathcal{A}_s} \overline{Q}^*(s, a).$$

Repitiendo el argumento del paso anterior, se define π^* como la política determinista que en cada estado s escoge una acción $a^*(s) \in \arg\max_{a \in \mathcal{A}_s} \overline{Q}^*(s, a)$.
Entonces

$$V^{\pi^*}(s) = \sum_{a \in \mathcal{A}_s} \pi_0^*(a; s) Q^{\pi^*}(s, a) = Q^{\pi^*}(s, a^*(s)) = \max_{a \in \mathcal{A}_s} \overline{Q}^*(s, a),$$

y se alcanza la igualdad.

- Se considera ahora $Q^* = \sup_\pi Q^\pi$. Dados $s \in \mathcal{S}$ y $a \in \mathcal{A}$, se recuerda que para π arbitraria, por la fórmula de la probabilidad total,

$$Q^\pi(s,a) = r(s,a) + \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}_{s'}} \mathbb{E}^\pi_{s,a} \left[\sum_{t=1}^{\infty} \gamma^t R_t | S_1 = s', A_1 = a' \right]$$

$$\cdot p(s'; s, a) \pi_0(a'; s').$$

Se define $\tilde{\pi} = (\tilde{\pi}_t)$ donde $\tilde{\pi}_t(a; s_0, a_0, ..., s_t) = \pi_{t+1}(a; s, a, s_0, a_0, ..., s_t)$ para $t \geq 0$. Entonces, si se prueba la igualdad

$$\mathbb{E}^\pi_{s,a} \left[\sum_{t=1}^{\infty} \gamma^t R_t | S_1 = s', A_1 = a' \right] = \gamma \, Q^{\tilde{\pi}}(s', a'),$$

se puede concluir que

$$Q^*(s,a) = r(s,a) + \gamma \sup_{\tilde{\pi}} \sup_{\pi_0} \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}_{s'}} p(s'; s, a) \pi_0(a'; s') Q^{\tilde{\pi}}(s,a)$$

$$= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) \max_{a \in \mathcal{A}_{s'}} Q^*(s,a),$$

usando el mismo argumento que en el paso para políticas estacionarias, ya que es la misma situación.

Para demostrar la igualdad, se define el proceso

$$(\tilde{S}_t, \tilde{A}_t, \tilde{R}_t) = (S_{t+1}, A_{t+1}, R_{t+1}) \quad \text{para}\, t \geq 0,$$

junto a la probabilidad $\tilde{\mathbb{P}} = \mathbb{P}^\pi_{s,a}(\cdot | S_1 = s', A_1 = a')$. Esto es un MDP que empieza en (s', a'), tiene función de transición p y política $\tilde{\pi}$.
Por tanto,

$$\mathbb{E}^\pi_{s,a} \left[\sum_{t=1}^{\infty} \gamma^t R_t | S_1 = s', A_1 = a' \right] = \gamma \, \mathbb{E}^\pi_{s,a} \left[\sum_{t=0}^{\infty} \gamma^t \tilde{R}_t | S_1 = s', A_1 = a' \right]$$

$$= \gamma \, \tilde{\mathbb{E}}^{\tilde{\pi}}_{s',a'} \left[\sum_{t=0}^{\infty} \gamma^t \tilde{R}_t \right] = \gamma \, Q^{\tilde{\pi}}(s', a').$$

- Demostrando que $V^*(s) = \max_{a\in\mathcal{A}_s} Q^*(s,a)$ de la misma forma que se hizo para el caso sobre políticas estacionarias,

$$T^*V^*(s) = \max_{a\in\mathcal{A}_s}\left\{ r(s;a) + \gamma \sum_{s'\in\mathcal{S}} p(s';s,a)V^*(s') \right\}$$

$$= \max_{a\in\mathcal{A}_s}\left\{ r(s;a) + \gamma \sum_{s'\in\mathcal{S}} p(s';s,a) \max_{a'\in\mathcal{A}_{s'}} Q^*(s',a') \right\}$$

$$= \max_{a\in\mathcal{A}_s} Q^*(s,a) = V^*(s).$$

□

Esta prueba demuestra que basta con centrarse en políticas estacionarias a la hora de encontrar una política óptima. De hecho, la siguiente definición establece una política que además es determinista, y que puede obtenerse resolviendo el sistema de optimalidad de Bellman.

Definición 3.12. Dada una función $q : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, la política dada por

$$\pi_q(a;s) := \begin{cases} 1 & \text{si } a = a^*(s), \\ 0 & \text{en caso contrario,} \end{cases} \qquad \text{con} \quad a^*(s) \in \arg\max_{a\in\mathcal{A}_s} q(s,a),$$

se conoce por política codiciosa con respecto a q. Si la política es codiciosa con respecto a q_v, donde

$$q_v(s,a) := r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s';s,a)v(s'), \quad s\in\mathcal{S},\ a\in\mathcal{A}_s,$$

entonces se denota por π_v.

En el caso de que existan varias acciones que maximicen la función de valor de estado-acción q, entonces se fija una de ellas.

Se presenta a continuación un lema que muestra de que manera se relacionan las políticas codiciosas con los operadores de Bellman. Es un resultado fundamental para afrontar la prueba del teorema del algoritmo de programación dinámica.

Lema 3.13. *Si π_q es una política codiciosa con respecto a q, entonces $T^*(q) = T^{\pi_q}(q)$. Si $q = q_v$, también se cumple que $T^*(v) = T^{\pi_v}(v)$.*

Demostración. Sean $s \in \mathcal{S}$ y $a \in \mathcal{A}$. Entonces, por la definición de política codiciosa de π_q, se tiene que:

$$T^{\pi_q}(q)(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}_{s'}} p(s';s,a)\pi_q(a';s')q(s',a')$$

$$= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s';s,a) \max_{a' \in \mathcal{A}_{s'}} q(s',a') = (T^*)(q)(s,a),$$

por la definición del operador de optimalidad. De manera similar, si q_v proviene de cierta $v \in X$, entonces

$$T^{\pi_v}(v)(s) = \sum_{a \in \mathcal{A}_s} \pi_v(a;s) \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s';s,a)v(s') \right)$$

$$= r(s,a^*(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s';s,a^*(s))v(s'),$$

donde $a^*(s) \in \arg\max_{a \in \mathcal{A}_s} \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s';s,a)v(s') \right)$. Por lo tanto, se concluye que

$$T^{\pi_v}(v)(s) = \max_{a \in \mathcal{A}_s} \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s';s,a)v(s') \right) = T^*(v)(s).$$

□

Con ayuda del lema, se obtiene el resultado que formaliza el objetivo del capítulo.

Teorema 3.14 (Algoritmo de programación dinámica)**.** *Existe una política óptima π^* estacionaria y determinista. Esa política es el resultado de resolver el sistema de optimalidad de estado-acción de Bellman, y hacer la política codiciosa con respecto a la solución.*

Demostración. Sea $q \in Y$ una solución del sistema de optimalidad de Bellman de estado-acción. Por la unicidad de puntos fijos del operador T^*, se obtiene $q = Q^*$. Usando el lema anterior en la segunda igualdad,

$$Q^* = T^*(Q^*) = T^{\pi_{Q^*}} Q^* = T^{\pi_q} Q^*.$$

Ahora, la unicidad de puntos fijos de T^{π_q}, establece que $Q^* = Q^{\pi_q}$. Dado que $s \in \mathcal{S}$, usando esta igualdad y la definición de π_q, se obtiene que

$$V^*(s) = \max_{a \in \mathcal{A}_s} Q^*(s,a) = \max_{a \in \mathcal{A}_s} Q^{\pi_q}(s,a) = \sum_{a \in \mathcal{A}_s} \pi_q(a;s) Q^{\pi_q}(s,a) = V^{\pi_q}(s).$$

Por tanto $V^* = V^{\pi_q}$, y por definición, π_q es óptima.

□

Como v se representa mediante un vector y q mediante una matriz, parece más sencillo resolver $T^*(v) = v$, y después obtener q_v mediante la fórmula de transformación arriba expuesta. Por tanto, este teorema establece el fundamento del algoritmo de programación dinámica: resolver el sistema de optimalidad de Bellman y luego actuar de manera codiciosa con respecto a la solución obtenida.

El hecho de que se pueda garantizar la existencia de una política óptima que sea determinista y estacionaria es fundamental en el contexto del aprendizaje por refuerzo. La estacionariedad implica que la regla de decisión no cambia con el tiempo, lo que simplifica la implementación de algoritmos óptimos en problemas a largo plazo. Por otro lado, el determinismo asegura que para cada estado exista una única acción óptima, eliminando la necesidad de introducir aleatoriedad en la toma de decisiones y permitiendo una mejor ejecución de la política. Juntas, estas propiedades garantizan que estos problemas puedan resolverse de manera eficiente, proporcionando soluciones robustas y computacionalmente manejables. Además, el hecho de que la política se obtenga de manera codiciosa es útil para desarrollar algoritmos capaces de obtenerla.

Capítulo 4

Métodos tabulares básicos

En el capítulo anterior, se introdujeron los fundamentos del control estocástico, enfatizando su formulación mediante procesos de decisión de Markov (MDPs). Se estableció que el objetivo principal en este contexto es encontrar una política óptima, es decir, una estrategia de decisión que maximice la recompensa esperada a lo largo del tiempo. Para ello, se presentó el teorema del algoritmo de programación dinámica, el cual garantiza la existencia de una política estacionaria óptima y proporciona la base para los métodos computacionales de solución.

Este capítulo se enfoca en algoritmos tabulares básicos para la resolución de MDPs. En particular, estudia dos enfoques fundamentales dentro de la programación dinámica:

- Iteración de valor: Un método basado en la actualización iterativa de los valores de los estados hasta alcanzar la convergencia a los valores óptimos.
- Iteración de política: Un enfoque que alterna entre la evaluación y la mejora de una política hasta encontrar la óptima.

Ambos algoritmos aprovechan la estructura de los MDPs y la unicidad de puntos fijos de los operadores de Bellman para encontrar soluciones óptimas de manera eficiente en entornos tabulares, es decir, cuando el espacio de estados y acciones es finito.

A lo largo de este capítulo, se desarrollarán las formulaciones matemáticas

de cada algoritmo, se analizarán sus condiciones de convergencia y se comparará su eficiencia computacional. Con ello, quedaran sentadas las bases para métodos más avanzados, como aquellos basados en aproximación de funciones. Además, se habrá proporcionado dos métodos más sencillos capaces de modelar y optimizar numerosos problemas reales, como podrá apreciarse en el quinto capítulo.

4.1. Algoritmo de iteración de valor

Para encontrar la política óptima basta con conocer la función de valor óptima. El objetivo de esta sección será desarrollar un método que aproxime dicha función de valor. El Teorema del Punto Fijo de Banach, además de proporcionar la unicidad de puntos fijos de una contracción T, asegura la convergencia de la sucesión $(T^{n+1}(v_0))$ hacia el punto fijo para cualquier v_0. El algoritmo de iteración de valor está basado en esta convergencia, actualizando el vector V en cada iteración.

Dado $\epsilon > 0$, se define el algoritmo de iteración de valor como sigue:

Algoritmo 1 Iteración de Valor

Entrada: $\epsilon > 0$
Salida: Aproximación $V \approx V^*$, política $\pi \approx \pi^*$.

Inicializar $V = 0$, $V_{\text{nuevo}} = 0$
$\Delta = 1$
Mientras $\Delta > \epsilon$ **hacer**
    $V = V_{\text{nuevo}}$
    Para cada $s \in \mathcal{S}$ **hacer**
        $$V_{\text{nuevo}}(s) = \max_{a \in \mathcal{A}_s} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) V(s') \right\}$$
    Fin Para cada
    $\Delta = \max_{s \in \mathcal{S}} |V_{\text{nuevo}}(s) - V(s)|$
Fin Mientras
Devolver $V = V_{\text{nuevo}}$ y la política codiciosa con respecto a V.

Nótese que como $\mathcal{S}$ es finito, existe una biyección lineal $\Phi : (X, |\cdot\|_\infty) \to (\mathbb{R}^{|\mathcal{S}|}, |\cdot\|_\infty)$ dada por $\Phi(v) = (v(s_1), ..., v(s_{|\mathcal{S}|}))$. Aunque T^* es no lineal, existe

un operador $\tilde{T}^*$ tal que $\tilde{T}^*(v(s_1), ..., v(s_{|\mathcal{S}|})) = \Phi(T^*(v))$. Dicho operador es la versión matricial de T^*, y sigue siendo contracción con mismo coeficiente. Por simplificar la notación, se hará referencia únicamente a T^*, tanto aplicado a elementos de X como de $\mathbb{R}^{|\mathcal{S}|}$.

Teorema 4.1. *El algoritmo de iteración de valor acaba, y el vector V devuelto cumple que $\|V - V^*\|_\infty \leq \frac{\gamma\,\epsilon}{1-\gamma}$.*

Demostración. La sucesión de vectores $(v_n)_{n\in\mathbb{N}_0} = (T^*(v_{n-1}))_{n\in\mathbb{N}}$ es la generada por el algoritmo, considerando $v_0 = 0$ el vector nulo. Como T^* es una contracción, la sucesión converge hacia cierto v^* por el Teorema del Punto Fijo de Banach. Por lo tanto, por la desigualdad triangular se obtiene que

$$\|v_{n-1} - v_n\|_\infty \leq \|v_{n-1} - v^*\|_\infty + \|v^* - v_n\|_\infty \to 0, \quad \text{cuando } n \to \infty.$$

Es decir, dado el $\epsilon > 0$ del algoritmo, existe un $N \in \mathbb{N}$ tal que $\|v_{N-1} - v_N\|_\infty = \Delta < \epsilon$, y por tanto el algoritmo finaliza.

Supongamos que el algoritmo finaliza tras n pasos, es decir, $V = v_n$, y $\|v_n - v_{n-1}\|_\infty < \epsilon$. Entonces dado $m \in \mathbb{N}$, por la desigualdad triangular,

$$\|v_n - v_{n+m}\|_\infty \leq \sum_{i=o}^{m-1} \|v_{n+i} - v_{n+i+1}\|_\infty = \sum_{i=o}^{m-1} \|(T^*)^i(v_n) - (T^*)^i(v_{n+1})\|_\infty.$$

Como T^* es una contracción con coeficiente γ, una sencilla inducción demuestra que $(T^*)^i$ es contracción con coeficiente γ^i. Por tanto,

$$\|v_n - v_{n+m}\|_\infty \leq \sum_{i=o}^{m-1} \gamma^i \|v_n - v_{n+1}\|_\infty.$$

Ahora, la continuidad de la norma, junto con esta desigualdad, establece que

$$\|V - V^*\|_\infty = \|\lim_{m\to\infty}(v_n - v_{n+m})\|_\infty = \lim_{m\to\infty} \|(v_n - v_{n+m})\|_\infty$$

$$\leq \lim_{m\to\infty} \sum_{i=o}^{m-1} \gamma^i \|v_n - v_{n+1}\|_\infty = \frac{1}{1-\gamma}\, \|v_{n+1} - v_n\|_\infty$$

$$= \frac{1}{1-\gamma}\, \|T^*(v_n) - T^*(v_{n-1})\|_\infty \leq \frac{\gamma}{1-\gamma}\, \|v_n - v_{n-1}\|_\infty \leq \frac{\gamma\,\epsilon}{1-\gamma},$$

aplicando la fórmula de la serie geométrica y la propiedad de contracción.

□

El teorema anterior establece una cota para el error de la aproximación que aporta el algoritmo. Sin embargo, la función de valor obtenida no es exactamente la óptima, sino una aproximación. Al transferir la función de valor a una política, es necesario analizar el efecto que esto puede tener.

Definición 4.2. Dado $\epsilon > 0$, una política $\pi \in \Pi$ es ϵ- óptima si

$$V^*(s) \leq V^\pi(s) + \epsilon \quad \forall s \in \mathcal{S}.$$

Una vez obtenida una aproximación V de V^*, se transfiere a una función de valor de estado-acción mediante la fórmula

$$Q(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s';s,a)V(s'),$$

obteniendo así una función Q aproximada. Obtener dicha aproximación iterando en un algoritmo al igual que se hace con V es posible, pero es mucho más costoso al trabajar con matrices. Esto no es necesario salvo que se desconozcan las transiciones $p(s';s,a)$, como en el caso del Q-learning.

Teorema 4.3. *Sea V la función de valor devuelta por el algoritmo, y Q la transformación según la fórmula anterior. Entonces la política codiciosa respecto a Q, π_Q, es $\frac{2\epsilon\gamma}{1-\gamma}$ - óptima.*

Demostración. Dado $s \in \mathcal{S}$, se tiene que por la definición de política codiciosa,

$$T^*(V)(s) = \max_{a \in \mathcal{A}_s} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s';s,a)V(s') \right\} = \max_{a \in \mathcal{A}_s} \{Q(s,a)\}$$

$$= \sum_{a \in \mathcal{A}_s} \pi_Q(a;s)Q(s,a) = \sum_{a \in \mathcal{A}_s} \pi_Q(a;s) \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s';s,a)V(s') \right)$$

$$= T^{\pi_Q}(V)(s).$$

Por tanto, $T^*(V) = T^{\pi_Q}(V)$. Ahora, si el algoritmo finaliza tras n iteraciones, $V = v_n$ y $\|v_n - v_{n-1}\|_\infty < \epsilon$. Entonces, usando la desigualdad triangular y esta igualdad, junto con la definición de contracción se obtiene que

$$\|V^{\pi_Q} - V\|_\infty = \|T^{\pi_Q}(V^{\pi_Q}) - V\|_\infty \leq \|T^{\pi_Q}(V^{\pi_Q}) - T^*(V)\|_\infty + \|T^*(V) - V\|_\infty$$

$$= \|T^{\pi_Q}(V^{\pi_Q}) - T^{\pi_Q}(V)\|_\infty + \|T^*(V) - T^*(v_{n-1})\|_\infty$$
$$\leq \gamma \|V^{\pi_Q} - V\|_\infty + \gamma \|V - v_{n-1}\|_\infty.$$

Despejando en esta desigualdad se obtiene que $\|V^{\pi_Q} - V\|_\infty \leq \frac{\gamma}{1-\gamma} \|v_n - v_{n-1}\|_\infty$.

En la prueba del teorema 4.1, se llegó a que $\|V - V^*\|_\infty \leq \frac{\gamma}{1-\gamma} \|v_n - v_{n-1}\|_\infty$. Juntando ahora ambas expresiones,

$$\|V^{\pi_Q} - V^*\|_\infty \leq \|V^{\pi_Q} - V\|_\infty + \|V - V^*\|_\infty \leq 2\,\frac{\gamma}{1-\gamma} \|v_n - v_{n-1}\|_\infty$$

$$\leq \frac{2\,\gamma\,\epsilon}{1-\gamma},$$

condición equivalente a la $\frac{2\gamma\epsilon}{1-\gamma}$- optimalidad, luego se concluye. □

Dada la convergencia del algoritmo, es importante analizar el grado de convergencia según la siguiente definición.

Definición 4.4. Sea $(M, \|\cdot\|)$ un espacio normado. Una sucesión $(x_n)_{n\in\mathbb{N}}$ en M con límite x^* se dice que tiene orden de convergencia $\alpha > 0$ si exite una constante $K < 1$ tal que

$$\|x_{n+1} - x^*\| \leq K\|x_n - x^*\|^\alpha \quad \forall n \in \mathbb{N}.$$

El caso $\alpha = 1$ se conoce como convergencia lineal.

En el caso del algoritmo de iteración de valor, el Teorema del Punto Fijo de Banach asegura la convergencia lineal. Sin embargo, el siguiente teorema afirma que no se puede mejorar dicha convergencia.

Teorema 4.5. *Para todas las posibles inicializaciones del algoritmo de iteración de valor, se obtiene una convergencia lineal con constante $K = \gamma$. Existe una inicialización para la cual la velocidad de convergencia es exactamente lineal.*

Demostración. Dado $n \in \mathbb{N}$, como V^* espunto fijo de T^*, que es una contracción, entonces

$$\|v_{n+1} - V^*\|_\infty = \|T^*(v_n) - T^*(V^*)\|_\infty \leq \gamma \|v_n - V^*\|,$$

por lo que la sucesión del algoritmo es lineal, con constante $K = \gamma$.

Ahora, sea $v_0 = V^* + k\mathbf{1}$ el vector inicial. Considerando $V^* + k \in X$, dado $s \in \mathcal{S}$ se tiene que

$$T^*(V^* + k)(s) = \max_{a \in \mathcal{A}_s} \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) V^*(s) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) k \right]$$

$$= V^*(s) + \gamma k.$$

Por tanto, abusando de notación al denotar igual a $V^* \in X$ que $V^* \in \mathbb{R}^{|\mathcal{S}|}$, se obtiene que

$$T^*(v_0) = \Phi(T^*(V^* + k)) = \Phi(V^* + \gamma k) = V^* + \gamma k \mathbf{1}.$$

Usando esta igualdad, obtenemos que

$$\|v_1 - V^*\|_\infty = \|T^*(v_0) - T^*(V^*)\|_\infty = \|V^* + \gamma k \mathbf{1} - V^*\|_\infty = \gamma \|v_0 - V^*\|_\infty.$$

Razonando por inducción, se supone que $\|v_n - V^*\|_\infty = \gamma \|v_{n-1} - V^*\|_\infty$. Como $T^*(v_n) = V^* + \gamma^{n+1} k \mathbf{1}$, entonces $\|v_n - V^*\|_\infty = \gamma^n \|v_0 - V^*\|_\infty$ y por tanto,

$$\|v_{n+1} - V^*\|_\infty = \gamma^{n+1} \|v_0 - V^*\|_\infty = \gamma \|v_n - V^*\|_\infty,$$

con lo que puede concluirse que el orden de convergencia de este algoritmo no puede ser mejor que lineal.

□

4.2. Algoritmo de iteración de política

El enfoque de iteración de política es una estrategia utilizada para encontrar una política óptima mediante la mejora progresiva de una política existente. A diferencia de los métodos clásicos de control estocástico, esta técnica es ampliamente empleada en el aprendizaje por refuerzo debido a su efectividad en algoritmos de aproximación. Su procedimiento consiste en repetir dos fases principales: la evaluación de la política, en la que se calcula o estima la función de valor V^π, y la mejora de la política, en la que esta se actualiza seleccionando la mejor acción posible de manera codiciosa.

En contraste con la iteración de valor, actualiza directamente la política en cada iteración. Dentro de este enfoque, el método actor-crítico alterna entre dos componentes: el crítico, que evalúa la calidad de la política midiendo su función de valor, y el actor, que utiliza esta evaluación para ajustarla y mejorarla. Esta interacción permite un proceso de optimización más eficiente al combinar el aprendizaje basado en valores con la actualización directa de la política.

4.2.1. Evaluación de política

Dada una política π, calcular la función de valor V^π puede enfocarse de tres maneras. La primera consiste en aproximar la esperanza mediante el método de Monte Carlo, lo que implica simular múltiples trayectorias y promediar los retornos observados. El segundo enfoque resuelve directamente la ecuación lineal de Bellman mediante álgebra matricial, utilizando técnicas como la inversión de matrices. Finalmente, se puede buscar el punto fijo del operador de Bellman, aplicando el Teorema del Punto Fijo de Banach, que garantiza la convergencia mediante iteraciones sucesivas.

Como se vio en el capítulo anterior, para una política estacionaria $\pi \in \Pi_S$, el operador de esperanza de Bellman para una función real $v : \mathcal{S} \to \mathbb{R}$ es

$$T^\pi(v)(s) := \sum_{a \in \mathcal{A}_s} \pi(a; s) \left(r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) v(s') \right), \quad \forall s \in \mathcal{S},$$

y además, V^π es su único punto fijo. El factor $r(s, a)$ muestra la recompensa inmediata cuando el sistema se encuentra en el estado s y se toma la acción a. Está claro que es dependiente del modelo de Markov, ya que $r(s, a) = \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} r\, p(s', r; s, a)$. Por tanto, puede escribirse también como

$$T^\pi(v)(s) := \sum_{a \in \mathcal{A}_s} \pi(a; s) \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r; s, a)[r + \gamma v(s')], \quad \forall s \in \mathcal{S}.$$

Esto es un sistema lineal de $|\mathcal{S}|$ ecuaciones y $|\mathcal{S}|$ incógnitas, que puede expresarse de forma matricial como:

$$V^\pi = r^\pi + \gamma P^\pi V^\pi,$$

donde $r^\pi = \left(\sum_{a \in \mathcal{A}_s} \pi(a; s) r(s, a) \right)_{s \in \mathcal{S}}$ y $P^\pi = \left(\sum_{a \in \mathcal{A}_s} \pi(a; s) p(s'; s, a) \right)_{(s, s') \in \mathcal{S} \times \mathcal{S}}$.

Proposición 4.6. *El sistema lineal $V^\pi = r^\pi + \gamma P^\pi V^\pi$ tiene una única solución. Esta es V^π y puede obtenerse como*

$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi.$$

Demostración. La unicidad está garantizada previamente, ya que T^π es una contracción. Ahora

$$V^\pi = r^\pi + \gamma P^\pi V^\pi \iff V^\pi - \gamma P^\pi V^\pi = r^\pi \iff (I - \gamma P^\pi) V^\pi = r^\pi$$
$$\iff V^\pi = (I - \gamma P^\pi)^{-1} r^\pi.$$

La matriz $(I - \gamma P^\pi)$ es invertible ya que como P^π es una matriz estocástica, sus autovalores cumplen $|\lambda| \leq 1$, luego los autovalores de $(I - \gamma P^\pi)$ son $1 - \gamma\lambda$, con $|1 - \gamma\lambda| \geq 1 - \gamma > 0$. □

Por lo tanto, la evaluación de una política equivale a la inversión de una matriz. Esto puede ser muy costoso computacionalmente, sobretodo si el conjunto de estados es grande. Por ello, es preferible usar métodos de iteración, siguiendo la propiedad de contracción del operador T^π.

Dada una política estacionaria π, se define el algoritmo síncrono de evaluación de política iterativa como:

Algoritmo 2 Evaluación iterativa de política (síncrono)

Entrada: $\pi \in \Pi_S$, $\epsilon > 0$
Salida: Aproximación $V \approx V^\pi$

Inicializar $V(s) = 0$ para todo $s \in \mathcal{S}$
$V_{\text{nuevo}}(s) = 0$ para todo $s \in \mathcal{S}$
$\Delta = 1$
Mientras $\Delta > \epsilon$ **hacer**
    $V = V_{\text{nuevo}}$
    Para cada $s \in \mathcal{S}$ **hacer**

$$V_{\text{nuevo}}(s) = \sum_{a \in \mathcal{A}_s} \pi(a \mid s) \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a) \big[r + \gamma V(s') \big]$$

    Fin Para cada
    $\Delta = \text{máx}_{s \in \mathcal{S}} \big| V_{\text{nuevo}}(s) - V(s) \big|$
Fin Mientras
Devolver V_{nuevo} como aproximación de V^π

Teorema 4.7. *El algoritmo anterior finaliza, y además* $\|V - V^\pi\|_\infty \leq \frac{\gamma\epsilon}{1-\gamma}$.

Demostración. La prueba es idéntica a la del teorema 4.1, sustituyendo simplemente T^* por T^π, ya que ambas son contracciones con factor γ.

□

Este algoritmo admite una versión asíncrona, donde solo se utiliza una única estructura de almacenamiento para V, y cada coordenada $V(s)$ se actualiza inmediatamente en la misma memoria en cuanto se obtiene su nuevo valor. Este enfoque reduce el uso de memoria, ya que no se necesita una copia auxiliar.

Algoritmo 3 Evaluación iterativa de política (asíncrono)

Entrada: Una política fija $\pi \in \Pi_S$, $\epsilon > 0$
Salida: Una aproximación $V \approx V^\pi$
Inicializar $V(s) = 0$ para todo $s \in \mathcal{S}$
$\Delta = 2\,\epsilon$
Mientras $\Delta > \epsilon$ **hacer**
$\Delta = 0$
Para cada $s \in \mathcal{S}$ **hacer**
$v = V(s)$

$$V(s) = \sum_{a \in \mathcal{A}_s} \pi(a\,;\,s) \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p\big(s', r \mid s, a\big)\,\big[r + \gamma\, V(s')\big]$$

$\Delta = \text{máx}\big\{\Delta,\, \big|v - V(s)\big|\big\}$
Fin Para cada
Fin Mientras
Devolver V como aproximación de V^π

Proposición 4.8. *El algoritmo totalmente asíncrono de evaluación iterativa de política converge hacia* V^π.

Demostración. Sea $\mathcal{S} = \{s_1, ..., s_K\}$. Para cada $s \in \mathcal{S}$, se define el operador

$$T_s^\pi V(s') = \begin{cases} T^\pi V(s'), & \text{si } s' = s, \\ V(s'), & \text{si } s' \neq s. \end{cases}$$

que aplica T^π a la coordenada s de V, y deja el resto igual. De esa manera, cada iteración del algoritmo aplica a V el operador $\overline{T}^\pi = (T^\pi_{s_K} \circ ... \circ T^\pi_{s_1})$. Entonces, dado que T^π es una contracción, obtenemos que dados $s \in \mathcal{S}$, V y W,

$$|\overline{T}^\pi(V)(s) - \overline{T}^\pi(W)(s)| = |(T^\pi_{s_K} \circ ... \circ T^\pi_{s_1})(V)(s) - (T^\pi_{s_K} \circ ... \circ T^\pi_{s_1})(W)(s)|$$

$$= |T^\pi(V)(s) - T^\pi(W)(s)| \leq \|T^\pi(V) - T^\pi(W)\|_\infty \leq \gamma\|V - W\|_\infty.$$

Luego $\|\overline{T}^\pi(V) - \overline{T}^\pi(W)\|_\infty \ \overline{T}^\pi \leq \gamma\|V - W\|_\infty$, y es una contracción.

Ahora, como V^π es un punto fijo de T^π, dados V y $s \in \mathcal{S}$,

$$\overline{T}^\pi(V^\pi)(s) = (T^\pi_{s_K} \circ ... \circ T^\pi_{s_1})(V^\pi)(s) = T^\pi_s(V^\pi)(s) = T^\pi(V^\pi)(s) = V^\pi(s).$$

Por lo tanto V^π también es punto fijo de $\overline{T}^\pi$, y como este es contracción, el Teorema del Punto Fijo de Banach establece la convergencia del algoritmo cuando el número de iteraciones tiende a infinito.

□

4.2.2. Mejora de política

El objetivo que se presenta ahora es el de mejorar la política en términos de que su función de valor sea mayor. Es decir, dada $\pi \in \Pi$ y V^π su función de valor, una mejora de π es $\pi' \in \Pi$ tal que

$$V^\pi(s) \leq V^{\pi'}(s), \quad \forall s \in \mathcal{S}.$$

Se dice que es una mejora estricta si la desigualdad es estricta para algún $s \in \mathcal{S}$.

Teorema 4.9 (Teorema de mejora de política)**.** *Dadas $\pi, \pi' \in \Pi_S$ políticas estacionarias, se cumple que*

(1) si

$$V^\pi(s) \leq \sum_{a \in \mathcal{A}_s} \pi'(a; s) Q^\pi(s, a), \quad \forall s \in \mathcal{S},$$

entonces π' es una mejora de π.

(II) *si*

$$V^{\pi}(s) < \sum_{a \in \mathcal{A}_s} \pi'(a;s) Q^{\pi}(s,a), \quad \textit{para algún } s \in \mathcal{S},$$

entonces la mejora es estricta.

(III) *para toda $\pi \in \Pi_S$, la política codiciosa con respecto a Q^{π} mejora a π.*

Demostración.

(I) Gracias a la definición del operador de esperanza de Bellman, dado $s \in \mathcal{S}$,

$$V^{\pi}(s) \leq \sum_{a \in \mathcal{A}_s} \pi'(a;s) Q^{\pi}(s,a)$$

$$= \sum_{a \in \mathcal{A}_s} \pi'(a;s) \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s';s,a) V^{\pi}(s') \right) = T^{\pi'}(V^{\pi})(s).$$

Como trivialmente $T^{\pi'}$ es un operador creciente, y se acaba de comprobar que $V^{\pi} \leq T^{\pi'}(V^{\pi})$, entonces para todo $s \in \mathcal{S}$ se tiene que

$$V^{\pi}(s) \leq T^{\pi'}(V^{\pi})(s) \leq T^{\pi'}(T^{\pi'}(V^{\pi}))(s) \leq \ldots \leq \lim_{k \to \infty} (T^{\pi'})^k (V^{\pi})(s).$$

Como $T^{\pi'}$ es contracción, el Teorema del Punto Fijo de Banach asegura la convergencia uniforme de $(T^{\pi'})^k(V^{\pi})$ hacia $V^{\pi'}$ cuando $k \to \infty$. Por tanto, se tiene la convergencia puntual y $V^{\pi}(s) \leq V^{\pi'}(s)$ para todo $s \in \mathcal{S}$, es decir, π' es una mejora de π.

(II) Como $T^{\pi'}$ es un operador monótono estricto, la desigualdad anterior es estricta para cierto $s \in \mathcal{S}$, lo que equivale a que π' sea mejora estricta de π.

(III) Sea π' es la política codiciosa con respecto a Q^{π}. Entonces para cada $s \in \mathcal{S}$, π' selecciona con probabilidad 1 una acción $a^*(s) \in \arg\max_{a \in \mathcal{A}_s} Q^{\pi}(s,a)$. Por tanto,

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}_s} \pi(a;s) Q^{\pi}(s,a) \leq \max_{a \in \mathcal{A}_s} Q^{\pi}(s,a) \sum_{a \in \mathcal{A}_s} \pi(a;s)$$

$$= \sum_{a \in \mathcal{A}_s} \pi'(a;s) Q^{\pi}(s,a),$$

y se concluye aplicando directamente (I).

□

El siguiente lema refleja como mejorar políticas hasta llegar a una óptima.

Lema 4.10. *Sea $\pi \in \Pi_S$ y π' la política codiciosa con respecto a Q^π. Entonces*

$$V^\pi = V^{\pi'} \implies \pi, \pi' \text{ son políticas óptimas.}$$

Demostración. Para políticas estacionarias se tiene que $V^\pi = V^{\pi'} \implies Q^\pi = Q^{\pi'}$, gracias a la definición 3.2. Por lo tanto, como π' selecciona para cada $s' \in \mathcal{S}$ una acción $a^*(s') \in \arg\max_{a' \in \mathcal{A}_{s'}} Q^\pi(s', a')$ con probabilidad 1, se tiene que dados $s \in \mathcal{S}$ y $a \in \mathcal{A}$,

$$Q^{\pi'}(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}_{s'}} p(s'; s, a)\pi'(a'; s')Q^{\pi'}(s', a'))$$

$$= r(s,a)+\gamma \sum_{s' \in \mathcal{S}} p(s'; s, a)Q^{\pi'}(s', a^*(s')) = r(s,a)+\gamma \sum_{s' \in \mathcal{S}} p(s'; s, a)Q^{\pi}(s', a^*(s'))$$

$$= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) \max_{a' \in \mathcal{A}_{s'}} Q^\pi(s', a')$$

$$= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) \max_{a' \in \mathcal{A}_{s'}} Q^{\pi'}(s', a'),$$

ya que por hipótesis, $Q^\pi = Q^{\pi'}$. Entonces esto equivale a que

$$T^*(Q^\pi)(s,a) = Q^\pi(s,a) \quad \text{y} \quad T^*(Q^{\pi'})(s,a) = Q^{\pi'}(s,a).$$

Es decir, ambas funciones de estado-acción cumplen la ecuación de optimalidad de Bellman. Como su solución es única, y es Q^*, se concluye que $Q^\pi = Q^* = Q^{\pi'}$. Por tanto π y π' son políticas óptimas. □

Corolario 4.11. *Sea $\pi \in \Pi_S$ una política no óptima, y π' la política codiciosa con respecto a Q^π. Entonces π' es una mejora estricta de π.*

Demostración. Por definición de política codiciosa, se tiene que

$$V^\pi(s) = \sum_{a \in \mathcal{A}_s} \pi(a; s)Q^\pi(s,a) \leq \max_{a \in \mathcal{A}_s} Q^\pi(s,a) = V^{\pi'}(s) \quad \forall s \in \mathcal{S}.$$

Como π no es óptima, por el lema anterior $V^\pi \neq V^{\pi'}$, luego existe un $s_0 \in \mathcal{S}$ tal que $V^\pi(s_0) < V^{\pi'}(s_0)$. Es decir, π' es una mejora estricta de π. □

Estos resultados respaldan un clásico algoritmo de mejora de política, que mejora una política mediante la política codiciosa con respecto a su función de valor:

Algoritmo 4 Mejora de política codiciosa

Entrada: $\pi \in \Pi_{\mathcal{S}}$, Q^{π}
Salida: π', la política codiciosa con respecto a Q^{π}
  Inicializar $\pi' := \pi$
  Para cada $s \in \mathcal{S}$ **hacer**
    Escoger $a^*(s) \in \arg\max_{a \in \mathcal{A}_s} Q^{\pi}(s, a)$
    $\pi'(a^*(s); s) := 1$
    Para cada $a \in \mathcal{A}_s \setminus \{a^*(s)\}$ **hacer**
      $\pi'(a; s) := 0$
    Fin Para cada
  Fin Para cada
  Devolver π' como la política mejorada

4.2.3. Algoritmo de iteración de política codiciosa

La evaluación y la mejora de políticas se combinan dando lugar a lo que se conoce por algoritmo de iteración de política. La idea es bastante clara: dada una política π, se computa su función de valor V^{π}, y se mejora usando la política codiciosa con respecto a Q^{π}.

Teorema 4.12. *El algoritmo de iteración de política codiciosa para MDPs finitos, inicializado con cualquier política, termina en un número finito de iteraciones. Además la política devueta es la política óptima* π^*.

Demostración. Como $\mathcal{S}$ y $\mathcal{A}$ son finitos, entonces el conjunto de políticas determinitas también lo es: $|\Pi_S^D| = |\mathcal{S}|^{|\mathcal{A}|} < \infty$. Por tanto, como el corolario 4.11 asegura que cada iteración proporciona una política estrictamente mejorada, se concluye que el algoritmo finaliza tras un número finito de pasos. Por el lema 4.10, una vez el algoritmo finaliza la política resultante es óptima, luego por su unicidad, esta es π^*. □

Por tanto, el algoritmo completo de iteración de política es:

Algoritmo 5 Iteración de política codiciosa

Entrada: Política inicial $\pi \in \Pi_{\mathcal{S}}$, función de valor inicial V
Salida: Política óptima $\pi^* \in \Pi_{\mathcal{S}}^{D}$
Inicializar arbitrariamente $V_{\text{nuevo}}, \pi_{\text{nuevo}}$
stop $= 0$
Mientras stop $= 0$ **hacer**
Evaluación de política : Obtener V^{π} como en 4.6
Establecer $Q^{\pi}(s,a) = \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r; s, a)\big[r + \gamma V^{\pi}(s')\big]$ para todo $s \in \mathcal{S}$ y $a \in \mathcal{A}$
Mejora de política: Obtener la política mejorada π_{nuevo} a partir de Q^{π}
Si $Q^{\pi_{\text{nuevo}}} = Q^{\pi}$ **entonces**
stop $= 1$
Fin Si
Fin Mientras
Devolver $\pi^* = \pi$

Aunque excede los contenidos de este trabajo, hay que destacar el caso en el que $\mathcal{S}$ y $\mathcal{A}$ no sean finitos. Una iteración de este tipo también converge hacia la política óptima en este caso. La computación exacta de las funciones de valor no es posible, pero mediante aproximaciones puede obtenerse una política óptima aproximada suficientemente buena.

Los algoritmos de iteración de valor e iteración de política son fundamentales en el aprendizaje por refuerzo. Haber demostrado su convergencia permite conocer cómo obtener una política óptima en una gran variedad de problemas prácticos. Además, su solidez teórica ha sentado las bases para métodos más avanzados como Q-learning y métodos de política aproximados, ampliando aún más el alcance del aprendizaje por refuerzo.

Capítulo 5

Implementación práctica de algoritmos

En el último capítulo se busca trasladar los conceptos teóricos desarrollados a lo largo del trabajo a escenarios concretos, concretando la aplicación de los métodos en problemas reales. Tras haber profundizado en la teoría de los MDPs y el control estocástico, y haber demostrado la convergencia de algunos algoritmos básicos, este capítulo sirve como puente entre la abstracción matemática y la resolución práctica de problemas.

Para ello, se presentan dos ejemplos muy recurrentes. En el primero, se aborda el desafío de un robot que debe encontrar la mejor ruta hacia un estado objetivo, mientras evita un área penalizante, de acuerdo con un criterio de optimización. Mediante el algoritmo de iteracion de valor se demuestra cómo alcanzar la solución óptima, evidenciando paso a paso el proceso de convergencia y optimización en un entorno sencillo pero representativo. El segundo ejemplo se centra en un problema de inventario, resuelto mediante el algoritmo de iteración de política, que ilustra de manera clara cómo las decisiones pueden ser optimizadas a través de iteraciones sucesivas de mejora de políticas.

Ambos casos no solo validan la solidez de los métodos estudiados, sino que además ofrecen una perspectiva práctica sobre cómo estas herramientas pueden ser implementadas en aplicaciones reales. Así, el capítulo cierra el trabajo aportando una conexión entre la teoría y la práctica, destacando la eficacia de

los algoritmos previamente estudiados en la solución de problemas de control estocástico y toma de decisiones en entornos inciertos.

5.1. Cuadrícula 2x2

En esta sección se presenta el ejemplo de un MDP con el objetivo de ilustrar el funcionamiento del algoritmo de iteración de valores. Para ello se comenzará definiendo los componentes del MDP, y se mostrará paso a paso cómo se actualizan las funciones de valor en cada estado. Este ejemplo, por su sencillez, permite comprobar los fundamentos del algoritmo de iteración de valor, y sirve como precedente para abordar problemas más complejos.

Este ejemplo esta inspirado en el de la sección 4.1.2 de [9]. En este problema, se modela el movimiento de un robot en una cuadrícula de 2x2, donde cada celda representa un estado. El robot tiene cinco opciones de acción: moverse hacia arriba, hacia la derecha, hacia abajo, hacia la izquierda o quedarse en su posición. El objetivo principal es que el robot aprenda a navegar en la cuadrícula de forma óptima, evitando el estado penalizante y buscando llegar al estado objetivo.

Se define el conjunto de estados como $\mathcal{S} = \{s_1, s_2, s_3, s_4\}$, donde s_2 es el área penalizante y s_4 el área objetivo. Las posibles acciones son $\mathcal{A} = \{a_1, a_2, a_3, a_4, a_5\}$, que corresponden respectivamente a moverse hacia arriba, hacia la derecha, hacia abajo, hacia la izquierda y a no moverse.

Las probabilidades de transición son deterministas, considerando que si se

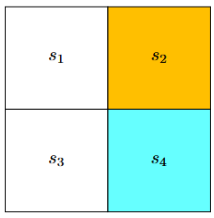

Figura 5.1: Ejemplo de cuadrícula 2×2 con área penalizante y área objetivo.

pretende hacer un movimiento válido (no se sale de la cuadrícula), la probabilidad de llegar al estado contiguo es 1. Sin embargo, si se pretende hacer un

movimiento no válido, se permanece en el estado actual con probabilidad 1. Cabe detacar que acceder a la zona penalizante se considera un movimiento válido.

Las recompensa inmediata relativa a acceder al área penalizante o a intentar salir de la cuadrícula es -1. La recompensa correspondiente a acceder al área objetivo es 1.

El algoritmo de iteración de valor comienza inicializando $V = 0$ para todos los estados. Teniendo en cuenta que buscamos para cada estado s,

$$V_{\text{nuevo}}(s) = \max_{a \in \mathcal{A}_s} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'; s, a) V(s') \right\},$$

se recogen todos los datos en la siguiente tabla:

q(s, a)	a_1	a_2	a_3	a_4	a_5
s_1	-1	-1	0	-1	0
s_2	-1	-1	1	0	-1
s_3	0	1	-1	-1	0
s_4	-1	-1	-1	0	1

Cuadro 5.1: Tabla de los valores $q(s, a)$ en la primera iteración.

Se actualizan los valores de V seleccionando el máximo valor para cada estado, obteniendo que

$$V(s_1) = 0, \quad V(s_2) = 1, \quad V(s_3) = 1, \quad V(s_4) = 1.$$

En la siguiente iteración, se obtienen los valores recogidos en la segunda tabla.

Ahora, se vuelven a actualizar los nuevos valores de acción, y se obtiene que

$$V(s_1) = \gamma 1, \quad V(s_2) = 1 + \gamma 1, \quad V(s_3) = 1 + \gamma 1, \quad V(s_4) = 1 + \gamma 1.$$

Realizando las siguientes iteraciones se comprueba que los nuevos valores de V difieren cada vez menos unos de otros, debido a que $\gamma \in (0, 1)$. Sin embargo, las acciones que dan lugar a dichas funciones de valor de estado-acción son

q(s, a)	a_1	a_2	a_3	a_4	a_5
s_1	-1+γ0	-1+γ1	0+γ1	-1+γ0	0+γ0
s_2	-1+γ1	-1+γ1	1+γ1	0+γ0	-1+γ1
s_3	0+γ0	1+γ1	-1+γ1	-1+γ1	0+γ1
s_4	-1+γ1	-1+γ1	-1+γ1	0+γ1	1+γ1

Cuadro 5.2: Tabla de los valores $q(s, a)$ en la segunda iteración.

siempre las mismas a partir de la segunda iteración. Por tanto estableciendo un criterio de convergencia con cierto $\epsilon > 0$, se finaliza el algoritmo. Haciendo la política codiciosa con respecto a la función de valor de estado-acción final, se obtiene la política óptima determinista siguiente:

$$\pi(s_1) = a_3, \quad \pi(s_2) = a_3, \quad \pi(s_3) = a_2, \quad \pi(s_4) = a_5.$$

Esta política podría deducirse intuitivamente, pero el ejemplo sirve de precedente a cuadrículas más amplias y a problemas de laberintos mucho más complejos.

5.2. Problema del control de inventario

A continuación se presenta un problema clásico de control de inventarios para un solo producto, cuyo objetivo es determinar cuántas unidades solicitar en cada periodo para maximizar los beneficios a largo plazo. El modelo se formula como un proceso de decisión de Markov (MDP), donde cada estado representa el nivel de inventario disponible y cada acción decide cuántas unidades pedir. Al existir una demanda aleatoria en cada período y costos asociados tanto a los pedidos de unidades como al mantenimiento de inventario, es fundamental encontrar una política que, en cada estado, seleccione la acción óptima de manera consistente. Para ello, en esta sección se utiliza la iteración de política como método para calcular la solución buscada en algún ejemplo numérico concreto.

Este problema está inspirado en la sección 3.2 de [6], pero también se apoya en el caso más general de [2]. Antes de formular el modelo, es importante dejar claros ciertos aspectos:

- El inventario tiene una capacidad máxima de M productos.

- La decisión de solicitar productos adicionales para rellenar el inventario se realiza al comienzo de cada mes.
- La demanda de productos se produce a lo largo del mes, pero todos los pedidos se satisfacen el último día del mes.
- Si la demanda excede la capacidad del inventario, el cliente que no reciba su producto se perderá. Es decir, no hay lista de espera para los casos en los que la demanda no pueda ser cubierta.
- Los precios de venta, mantenimiento de inventario y solicitud de stock adicional no varían con el tiempo.

Se considerará un horizonte temporal infinito con factor de descuento $\gamma \in (0, 1)$. El conjunto de estados $\mathcal{S} = \{0, 1, 2, ..., M\}$ representa la cantidad de inventario dsiponible al inicio de cada mes. Para cada $s \in \mathcal{S}$ se define su estado de acciones $\mathcal{A}_s = \{0, 1, 2, ..., M - s\}$, que representa la cantidad de unidades a pedir para rellenar inventario. Por tanto $\mathcal{A} = \bigcup_{s \in \mathcal{S}} \mathcal{A}_s$ es el conjunto total de acciones.

La demanda se representa mediante una variable aleatoria discreta D. Se asume que la demanda tiene una distribución de probabilidad conocida $p_j = P(D = j)$. El estado del inventario en el mes $t+1$, s_{t+1}, puede escribirse como $s_{t+1} = \text{máx}\{s_t + a_t - D_t, 0\}$. Por lo tanto, las probabilidades de transición son

$$P(s'; s, a) = \begin{cases} p_{(s+a-s')} & \text{si } s + a \geq s', \\ 0 & \text{en cualquier otro caso.} \end{cases}$$

Se explican a continuación los parámetros económicos del modelo. El costo de pedir u unidades es $C(u)$. Dicho valor tiene una componente fija $K > 0$ que interviene en todos los pedidos, y una componente que varía en funcion de la cantidad solicitada. Es decir,

$$C(u) = \begin{cases} K + c(u) & \text{si } u > 0, \\ 0 & \text{si } u = 0. \end{cases}$$

El costo de mantenimiento de un inventario de u unidades se representa por la función creciente $h(u)$. Finalmente, si la demanda es de j unidades y puede satisfacerse con el nivel de inventario actual, los ingresos quedan descritos por $f(u)$.

Como la venta efectiva en cada mes es mín$\{s+a, D\}$, la recompensa esperada sigue la siguiente fórmula:

$$r(s,a) = \sum_{j=0}^{s+a-1} p_j\, f(j) + \sum_{j=s+a}^{\infty} p_j\, f(s+a) - C(a) - h(s+a).$$

Una vez descrito el modelo, se plantea su resolución siguiendo el algoritmo de iteración de política. Este procedimiento combina dos fases que se repiten hasta convergencia, como se apreció en el capítulo anterior.

Para comprobar la aplicabilidad del algoritmo de iteración de políticas a este problema, se plantea el siguiente ejemplo concreto:

- Capacidad de inventario de $M = 5$, es decir, 6 posibles estados.
- La demanda puede ser desde 0 hasta 4 productos cada mes, y sigue una distribución con vector de probabilidades $(0{,}05, 0{,}25, 0{,}3, 0{,}35, 0{,}05)$.
- El costo fijo por realizar un pedido es $K = 1{,}0$.
- El costo variable de pedido es $c(u) = u$, es decir, de una unidad monetaria por cada unidad de producto pedido.
- El costo de mantenimiento de inventario es de $h(u) = 0{,}5u$.
- El precio de venta de cada producto es de 5.0 unidades monetarias por producto.
- Se considera el factor de descuento habitual de $\gamma = 0{,}95$.

Se ha programado en Python un código que implementa el algoritmo de iteración de política siguiendo la misma notación que se ha usado en el capítulo anterior. También se ha implementado el ejemplo concreto que se acaba de describir, calculando por tanto las recompensas inmediatas y las probabilidades de transición utilizando la distribución de la demanda. Inicializando el algoritmo con la política inicial que consiste en no pedir productos en ningún estado, es decir, $\pi(s) = 0\, \forall s = 0, 1, ..., 5$, y ejecutando el programa se obtiene el resultado ilustrado en la imagen inferior.

La política sugiere que la estrategia que más ganancias genera en este problema concreto es tener un mínimo de 3 unidades en el inventario, y no pedir más productos a partir de esa cantidad. Este tipo de regla de decisión es muy típico en el problema del inventario, y se conoce como política (Σ, σ), ya que

```
Convergencia en la iteración 2

Política Óptima:
  Estado s=0 -> acción: pedir 3 unidades
  Estado s=1 -> acción: pedir 2 unidades
  Estado s=2 -> acción: pedir 1 unidades
  Estado s=3 -> acción: pedir 0 unidades
  Estado s=4 -> acción: pedir 0 unidades
  Estado s=5 -> acción: pedir 0 unidades

Valor Óptimo de cada estado:
  V*(0) = 114.0000
  V*(1) = 115.0000
  V*(2) = 116.0000
  V*(3) = 118.0000
  V*(4) = 118.8845
  V*(5) = 119.5775
```

Figura 5.2: Ejecución de la iteración de política para el ejemplo de control de inventario.

incita a ampliar el nivel del inventario a Σ unidades siempre que al comienzo del periodo el inventario contenga menos de σ productos.

Se puede apreciar que para este problema la iteración de política es un algoritmo muy apropiado, ya que en tan solo 2 iteraciones se ha conseguido encontrar la política óptima.

Apéndice A

Complementos de probabilidad

A.1. Entropía relativa

Definición A.1. Dadas μ y ν medidas en $(\Omega, \mathcal{A})$, entonces $\mu \ll \nu$ ("μ es absolutamente continua respecto de ν") si $\nu(A) = 0$ implica que $\mu(A) = 0$ para todo $A \in \mathcal{A}$.

Teorema A.2 (Teorema de Radon-Nikodym)**.** *Sean μ, ν medidas en $(\Omega, \mathcal{A})$ con ν σ-finita y $\mu \ll \nu$. Entonces existe $f : \Omega \to \mathbb{R}$ medible tal que $f \geq 0$ y*

$$\mu(A) = \int_A f \, d\nu \quad \forall A \in \mathcal{A}.$$

Además, f es única ν-c.s.

Demostración. [4], página 129. □

Definición A.3. Se denota $\frac{d\mu}{d\nu} := f$, y se llama derivada de Radon-Nikodym de μ con respecto a ν, a la función del teorema de Radon-Nikodym.

Propiedad A.4. *Dada $g : \Omega \longrightarrow \mathbb{R}$ medible, con $\mu \ll \nu$, y $\frac{d\mu}{d\nu}$ la derivada de Radon-Nikodym de μ con respecto a ν, entonces*

$$\int g \, d\mu = \int g \, \frac{d\mu}{d\nu} \, d\nu.$$

Demostración. [4], página 134. □

Una vez definida la derivada de Radon-Nikodym, y enunciado una de sus principales propiedades, se presenta a continuación el concepto de entropía relativa, que es fundamental en el primer capítulo del texto.

Definición A.5. Sean P y Q probabilidades en $(\Omega, \mathcal{F})$. Entonces

$$D(P,Q) = \begin{cases} \int \log\left(\frac{dP}{dQ}\right) dP & \text{si } P \ll Q, \\ \infty & \text{en cualquier otro caso ,} \end{cases}$$

se llama entropía relativa de P respecto de Q.

A raíz de la definición, se enuncian y demuestran ahora algunas de las propiedades de la entropía relativa.

Propiedades A.6.

1. $D(P,Q) \geq 0$.
2. $D(P,Q) = 0 \iff P = Q$.

Demostración.

1. Si $P \not\ll Q$ está claro.

   Si $P \ll Q$, llamamos $\Phi(x) = x\log(x), x \geq 0$ y entonces,

   $$D(P,Q) = \int \log\left(\frac{dP}{dQ}\right)\frac{dP}{dQ}dQ = \int \Phi\left(\frac{dP}{dQ}\right) dQ.$$

   Como $x\log(x) \geq x - 1$, se concluye que:

   $$D(P,Q) \geq \int \left(\frac{dP}{dQ} - 1\right) dQ = 1 - 1 = 0.$$

2. $D(P,Q) = 0 \iff P \ll Q$ y $\log\left(\frac{dP}{dQ}\right) = 0 \iff \frac{dP}{dQ} = 1 \iff P(A) = \int_A dQ = Q(A) \quad \forall A \quad \iff P = Q$.

□

Lema A.7. *$P \ll Q \ll \nu$ implica que $\frac{dP}{dQ}\frac{dQ}{d\nu} = \frac{dP}{d\nu}$.*

Demostración. Gracias a la propiedad A.4, tenemos que dado $B \in \mathcal{A}$:

$$P(B) = \int_B \frac{dP}{dQ}\, dQ = \int_B \frac{dP}{dQ}\frac{dQ}{d\nu}\, d\nu.$$

□

Lema A.8.

1. *Si $P \sim f$ y $Q \sim g$, entonces $\frac{dP}{dQ} = \frac{f}{g}$.*
2. *Si $P \sim \mathcal{N}(\mu_1, \sigma_1^2)$ y $Q \sim \mathcal{N}(\mu_2, \sigma_2^2)$,*

$$D(P,Q) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

Demostración.

1. Se tiene que $P \ll Q \ll \mu$, siendo μ la medida de Lebesgue (cualquier conjunto de medida de Lebesgue nula tiene probabilidad nula). Por lo tanto, por el lema A.7 y la definición de función de densidad,

$$\frac{dP}{dQ} = \frac{\frac{dP}{d\mu}}{\frac{dQ}{d\mu}} = \frac{f}{g}.$$

   La expresión es válida ya que $Q(\{g = 0\}) = 0 = P(\{f = 0\})$.

2. $$D(P,Q) = \int \log\left(\frac{dP}{dQ}\right) dP = \int \log\left(\frac{dP}{dQ}\right)\frac{dP}{d\mu} d\mu$$
$$= \int_{\mathbb{R}} f(x) \log\left(\frac{f(x)}{g(x)}\right) dx = \int_{\mathbb{R}} f(x) \log\left(\frac{\sigma_2}{\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}}\right) dx$$
$$= \int_{\mathbb{R}} f(x) \log\left(\frac{\sigma_2}{\sigma_1}\right) dx + \int_{\mathbb{R}} f(x) \left(\frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}\right) dx$$

$$= \log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{1}{2\sigma_1^2}\int_{\mathbb{R}} f(x)(x-\mu_1)^2 dx + \frac{1}{2\sigma_2^2}\int_{\mathbb{R}} f(x)(x-\mu_2)^2 dx$$

$$= \log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{1}{2\sigma_1^2}\sigma_1^2 + \frac{1}{2\sigma_2^2}(\sigma_1^2 + (\mu_1-\mu_2)^2)$$

$$= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2},$$

donde se ha usado que

$$\mathbb{E}(X-\mu_1) = \sigma_1^2 \quad \text{si} \quad X \sim \mathcal{N}(\mu_1, \sigma_1^2),$$

y que

$$\mathbb{E}(X-\mu_2) = \mathbb{E}[(X-\mu_1)^2] + 2\mathbb{E}[(X-\mu_1)(\mu_1-\mu_2)] + \mathbb{E}[(\mu_1-\mu_2)^2]$$
$$= \sigma_1^2 + (\mu_1-\mu_2)^2$$

en ese mismo caso.

□

El siguiente resultado hace mención al concepto de entropía relativa, y es necesario para el desarrollo del primer capítulo del texto.

Teorema A.9 (de Bretagnolle-Huber)**.** *Si P y Q son probabilidades en $(\Omega, \mathcal{F})$, entonces*

$$P(A) + Q(A^c) \geq \frac{1}{2}e^{-D(P,Q)} \quad \forall A \in \mathcal{F}.$$

Demostración. El caso $D(P,Q) = \infty$ es trivial. Por lo tanto, supongamos que $P \ll Q$. Se define $\nu = P + Q$. Es claro que $P \ll Q \ll \nu$.

Sean $f = \frac{dP}{d\nu}$ y $g = \frac{dQ}{d\nu}$. Como $Q(\{\omega : g(\omega) = 0\}) = 0$, entonces

$$P(\{\omega : g(\omega) = 0\}) = 0.$$

Así, aplicando la desigualdad de Jensen para la función cóncava $\varphi(x) = \log(x)$, se tiene:

$$\exp(-D(P,Q)) = \exp\left(-\int_{\{\omega: g(\omega)>0\}} \log\left(\frac{f}{g}\right) dP\right) = \exp\left(\int \log\left(\frac{g}{f}\right) dP\right)$$

$$\leq \exp\left(2\log\left(\int \sqrt{\frac{g}{f}}\, dP\right)\right) = \exp\left(2\log\left(\int \sqrt{\frac{g}{f}}\, \frac{dP}{d\nu}\, d\nu\right)\right)$$
$$= \left(\int \sqrt{fg}\, d\nu\right)^2.$$

Dado $\omega \in \Omega$, sean $M(\omega) = \text{máx}\{f(w), g(w)\}$ y $m(w) = \text{mín}\{f(w), g(w)\}$. Entonces $f(\omega)g(\omega) = m(\omega)M(\omega) \quad \forall \omega \in \Omega$.

Por tanto,

$$\left(\int \sqrt{fg}\, d\nu\right)^2 = \left(\int \sqrt{Mm}\, d\nu\right)^2 \leq \left(\int M\, d\nu\right)\left(\int m\, d\nu\right),$$

por la desigualdad de Cauchy-Schwarz. Como $f(\omega)+g(\omega) = M(\omega)+m(\omega) \quad \forall \omega \in \Omega$, y $\int f + g\, d\nu = \int f\, d\nu + \int g\, d\nu = P(\Omega) + Q(\Omega) = 2$, entonces

$$\int M\, d\nu = \int f + g\, d\nu - \int m\, d\nu = 2 - \int m\, d\nu \leq 2.$$

Así, se concluye que

$$\frac{1}{2}\exp(-D(P,Q)) \leq \int m\, d\nu = \int_A m\, d\nu + \int_{A^C} m\, d\nu$$
$$\leq \int_A f\, d\nu + \int_{A^c} g\, d\nu = P(A) + Q(A^C).$$

□

A.2. Esperanza condicionada

Sea X una variable aleatoria integrable en $(\Omega, \mathcal{F}, \mathbb{P})$ y $\mathcal{G} \subset \mathcal{F}$ una σ-álgebra. Se define la medida ν en $(\Omega, \mathcal{G})$ dada por

$$\nu(A) = \int_A X\, d\mathbb{P}, \quad A \in \mathcal{G}.$$

Entones, como $\mathbb{P}$ es probabilidad en $(\Omega, \mathcal{G})$, se tiene que $\nu \ll \mathbb{P}$. El teorema de Radon-Nikodym afirma que en estas condiciones existe una función $h : \Omega \to \mathbb{R}$, $\mathcal{G}$- medible e integrable tal que

$$\int_A X\, d\mathbb{P} = \nu(A) = \int_A h\, d\mathbb{P}, \quad \forall A \in \mathcal{G}. \tag{A.1}$$

Se define la esperanza de X condicionada por $\mathcal{G}$ como

$$\mathbb{E}(X|\mathcal{G}) := h = \frac{d\nu}{d\mathbb{P}|_{\mathcal{G}}}.$$

$\mathbb{E}(X|\mathcal{G})$ es una variable aleatoria $\mathcal{G}$- medible e integrable, y es la única ($\mathbb{P}$- c. s.) con estas propiedades tal que

$$\mathbb{E}(XI_A) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})I_A), \quad \forall A \in \mathcal{G}. \tag{A.2}$$

Si $\mathcal{G} = \sigma(\mathcal{Y})$, con Y una variable aleatoria, se utiliza la notación $\mathbb{E}(X|Y)$ en lugar de $\mathbb{E}(X|\mathcal{G})$. En este caso, la propiedad (A.2) implica que $\mathbb{E}(X|Y)$ sea la única variable aleatoria medible de Y e integrable tal que

$$\mathbb{E}(XI_{(Y\in B)}) = \mathbb{E}(\mathbb{E}(X|Y)I_{(Y\in B)}), \quad \forall B \text{ medible}.$$

Las siguientes propiedades son de mucha utilidad a lo largo del texto:

Propiedades A.10. *Sean X, Y variables aleatorias en $(\Omega, \mathcal{F}, \mathbb{P})$ y $\mathcal{G} = \sigma(Y)$ σ-álgebra contenida en $\mathcal{F}$. Entonces:*

1. $\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X)$.
2. $\mathbb{E}(aX_1 + bX_2|Y) = a\mathbb{E}(X_1|Y) + b\mathbb{E}(X_2|Y)$.
3. $\mathbb{E}(g(Y)X|Y) = g(Y)\mathbb{E}(X|Y)$.

Demostración.

1. Como $\Omega \in \mathcal{G}$, entonces por (A.1)
$$\int_\Omega X\, d\mathbb{P} = \int_\Omega \mathbb{E}(X|Y)\, d\mathbb{P},$$
luego
$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)).$$

2. Usando (A.1), para todo $A \in \mathcal{G}$ se cumple que
$$\int_A \mathbb{E}(aX_1 + bX_2|Y)\, d\mathbb{P} = \int_A aX_1 + bX_2\, d\mathbb{P} = a\int_A X_1\, d\mathbb{P} + b\int_A X_2\, d\mathbb{P}$$
$$= a\int_A \mathbb{E}(X_1|Y)\, d\mathbb{P} + b\int_A \mathbb{E}(X_2|Y)\, d\mathbb{P} = \int_A a\mathbb{E}(X_1|Y) + b\mathbb{E}(X_2|Y)\, d\mathbb{P}.$$
Como es una función única ($\mathbb{P}$- c. s.), se concluye.

3. En [1], páginas 447 y 448, se encuentra esta demostración.

□

Apéndice B

Teorema del Punto Fijo de Banach

Definición B.1. Un espacio de Banach es un espacio vectorial normado $(M, \|\cdot\|)$ sobre $\mathbb{K}$ (uno de los cuerpos $\mathbb{R}$ o $\mathbb{C}$), que sea completo, es decir, toda sucesión de Cauchy sea convergente (tomando la distancia asociada a su norma).

Dado un espacio K compacto, el espacio $C(K)$ de todas las funciones continuas $K \to \mathbb{K}$ es un espacio de Banach junto a la norma del supremo $\|\cdot\|_\infty$.

Definición B.2. Dado un espacio vectorial normado $(M, \|\cdot\|)$ sobre $\mathbb{K}$, una contracción de M es una aplicación $T : M \to M$ tal que existe $\lambda \in [0, 1)$ de manera que:

$$\|T(m_1) - T(m_2)\| \leq \lambda \|m_1 - m_2\| \quad \forall m_1, m_2 \in M.$$

El valor λ se conoce como constante de contracción.

El siguiente resultado resulta muy útil para obtener convergencias en espacios de Banach, y es fundamental a lo largo del documento.

Teorema B.3 (del punto fijo de Banach)**.** *Dado un espacio vectorial normado $(M, \|\cdot\|)$ sobre $\mathbb{K}$, y una contracción $T : M \to M$, entonces*

(I) existe un único punto fijo $m^* \in M$, es decir, tal que $T(m^*) = m^*$.

(II) fijando un $m_0 \in M$ arbitrario, la sucesión $(m_t)_{t \in \mathbb{N}_0}$ dada por $m_{t+1} = T(m_t) = T^{t+1}(m_0)$ converge en M hacia m^*.

Demostración. Página 50, teorema 2.4.1. [5] □

Bibliografía

[1] Billingsley, P. *Probability and Measure*, 1 ed. Wiley series in Probability and athematical Statistics. Wiley-Interscience, January 1995.

[2] C. Selvakumar, P. M., and Elango, C. Discrete markov decision process – inventory control in a service facility system. *Revista Investigación Operacional 41*, 6 (2020), 872–881.

[3] Döring, L. The mathematics of reinforcement learning. *University of Manheim* (2023).

[4] Halmos, P. R. *Measure Theory*. The University series in higher mathematics. Springer Science+Business Media New York, 1974.

[5] Kesavan, S. *Functional Analysis*, 2 ed. Texts and Readings in Mathematics. Springer Singapore, February 2023.

[6] Puterman, M. L. *Markov decision processes : discrete stochastic dynamic programming*, 1 ed. Wiley Series in Probability and Statistics. John Wiley Sons, Inc, 1994.

[7] Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society 58*, 5 (1952), 527–535.

[8] Sutton, R. S., and Barto, A. G. *Reinceforcement Learning: An introduction*, 2 ed. Adaptative Computation and Machine Learning series. MIT Press, November 2018.

[9] Zhao, S. *Mathematical Foundations of Reinforcement Learning*. Springer Nature Press and Tsinghua University Press, 2025.