# Tercera Parte:
# Clase 1 – MDP con Horizonte Infinito

Optimización Dinámica - ICS

Mathias Klapp - 2020


- Introducción a MDPs de horizonte infinito
- Retorno descontado: Evaluación de política
- Retorno descontado: Optimización

# ¿Cómo vamos?

- Procesos de Decisión Markoviana (MDP) con horizonte finito.
    - Técnica de solución: Backward DP
    - Aplicaciones: Inventario, ruteo , selección de personal, renovación de equipos.

- **Ahora veremos MDPs de horizonte Infinito**
    - Cuando horizonte es prácticamente infinito.
    - Ejemplos: inventario de coca-cola, control de admisión web, generación eléctrica, Pensiones de cotizante joven.

# Maximización del costo esperado infinito

$$\max_{\pi \in \Pi^{MD}} \left\{ \lim_{T \to \infty} \mathbb{E} \left( \sum_{t=1}^{\mathrm{T}} r_t(S_t, d_t^{\pi}(S_t)) \,|S_1 = s \right) \right\}$$

Desafíos:

1. ¿Cómo lo resolvemos?.
   - Backward DP no sirve sin etapa terminal.
   - ¿Cómo evaluamos una política $\pi$?

2. Valor objetivo puede ser infinito (y pierde sentido).
   - Ejemplo: $A$ retorna $\$1$ a perpetuidad y $B$ retorna $\$1.000.000$ a perpetuidad. Ambas decisiones poseen valor ∞, pero $B$ es mejor.

3. Diferenciar valor del retorno en el tiempo.
   - Ejemplo: $A$ cuesta $\$10^6$ en $t = 1$, luego retorna \$1 a perpetuidad versus $B$ que siempre paga $\$0$. ¿Qué prefiere?

# Criterio de retorno esperado descontado:

$$V^*(s) = \max_{\pi \in \Pi^{MD}} \{V^\pi(s)\},$$

$$V^\pi(s) = \lim_{T \to \infty} \mathbb{E}\left[\sum_{t=1}^{\mathrm{T}} \lambda^{t-1} \cdot r_t(S_t, d_t^\pi(S_t)) \,|S_1 = s\right]$$

Observaciones:

- Optimiza el valor presente del retorno esperado.
- Factor de descuento: $\lambda \in [0,1)$
  - \$1 en el periodo $t$ vale \$$\lambda^{t-1}$ pesos hoy ($t = 1$).
- Privilegia futuro de corto plazo sobre el largo plazo.

# Supuestos

1. **Retorno finito:** $|r_t(s,x)| < \infty$, para $s \in \mathbb{S}, x \in \mathbb{X}_t(s)$

2. **Espacio de estados discreto**

3. **Estacionariedad:**
   - Retorno y probabilidades independientes del tiempo $t$.
     - $r_t(s,x) = r(s,x)$
     - $p_t(s'|s,x) = p(s'|s,x)$

NOTA:

- Si $|r_t(s,x)| < \infty$, entonces $V^*(s) < \infty$.

# Teorema de convergencia acotada

Si:

- $A_t < \infty$ es una serie de variables aleatorias acotadas.
- $A_t$ converge a la v.a. $A$ con probabilidad 1: $\mathbb{P}\left(\lim_{t\to\infty} A_t = A\right) = 1$

entonces:

$$\lim_{t\to\infty} \mathbb{E}(A_t) = \mathbb{E}(A)$$

En nuestro curso implica que:

$$V^{\pi}(s) = \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=1}^{\mathrm{T}} \lambda^{t-1} r(S_t, d_t^{\pi}(S_t))\, |S_1 = s\right] = \mathbb{E}\left[\sum_{t=1}^{\infty} \lambda^{t-1} r(S_t, d_t^{\pi}(S_t))\, |S_1 = s\right]$$


- Introducción a MDPs de horizonte infinito
- Retorno descontado: Evaluación de política
- Retorno descontado: Optimización

# Desafío: ¿Cómo evaluar una política?

$$V^{\pi}(s_1) = \mathbb{E}\left(\sum_{t=1}^{\infty} \lambda^{t-1} r(S_t, d_t^{\pi}(S_t)) \middle| S_1 = s_1\right)$$

**Observación 1:**
El valor de una política $\pi \in MD$ se puede escribir como:

$$V^{\pi}(s_1) = \sum_{t=1}^{\infty}\left(\lambda^{t-1} \cdot \sum_{s_t \in \mathbb{S}} p_{\pi}^{(t-1)}(s_t|s_1) \cdot r\big(s_t, d_t^{\pi}(s_t)\big)\right)$$

, donde $p_{\pi}^{(t)}(s_t|s_1)$: probabilidad de transición a $s_t$ en $t$ etapas desde $s_1$.

Notación vectorial si:

$$V^{\pi} = \sum_{t=1}^{\infty} \lambda^{t-1} \cdot P_{\pi}^{(t-1)} r_{d_t^{\pi}}$$

, donde

- $V^{\pi}, r_{d_t^{\pi}} \in \mathbb{R}^{|\mathbb{S}|}$ son los vectores de retornos y valor para la política
- $P_{\pi}^{(t)} = \prod_{k=1}^{t-1} P_{d_k^{\pi}}{}^{(1)} \in \mathbb{R}^{|\mathbb{S}|\times|\mathbb{S}|}$ matriz de transición asociada a política $\pi$ de $t$ etapas

# ¿Cómo evaluar una política estacionaria?

Si política es estacionaria, es decir $\pi = (d, d, \dots)$, entonces:

$$V^d = \sum_{t=1}^{\infty} \lambda^{t-1} P_d^{t-1} r_d$$

En este caso se cumple que:

$$V^d = r_d + \lambda P_d \sum_{t=2}^{\infty} \lambda^{t-2} P_d^{t-2} r_d$$
$$V^d = r_d + \lambda P_d V^d$$

**Teorema:**

Sea $L_d(V) \coloneqq r_d + \lambda P_d V$.

La **única** solución $V \in \mathbb{R}^{n_{\mathbb{S}}}$ del sistema $V = L_d(V)$ es

$$V^d = (I - \lambda P_d)^{-1} r_d$$

y es el valor $V^d$ de una política estacionaria $\pi = (d, d, \dots)$

# Propiedades

**Propiedades:**

Sean $V, U \in \mathbb{R}^{n_s}$ vectores cualquiera y $d$ una política estacionaria.

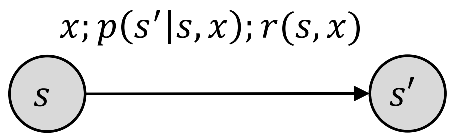
1. Si $V \geq 0$, entonces $(I - \lambda P_d)^{-1} V \geq V$
2. Si $V \geq 0$, entonces $(I - \lambda P_d)^{-1} V \geq 0$
3. Si $V \geq U$, entonces $(I - \lambda P_d)^{-1} V \geq (I - \lambda P_d)^{-1} U$

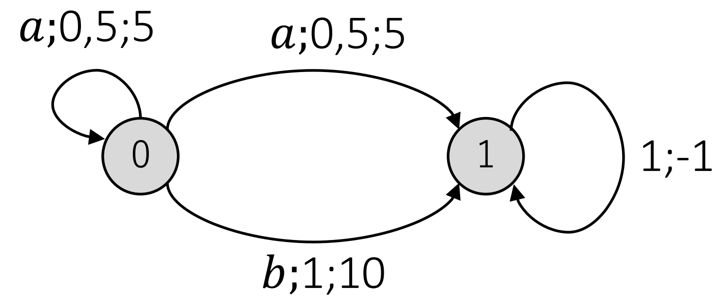
## Prueba

1 y 2: $(I - \lambda P_d)^{-1} V = \sum_{t=1}^{\infty} \lambda^{t-1} P_d^{t-1} V \geq V \geq 0$

3: $(I - \lambda P_d)^{-1}(V - U) \geq 0$ (de 1)

# Ejemplo




Dos posibles políticas estacionarias:

- $d(0) = a$

$$V^d(0) = \frac{5 - \frac{11}{2} \cdot \lambda}{\left(1 - \frac{1}{2}\lambda\right)(1 - \lambda)}, V^d(1) = -\frac{1}{1 - \lambda}$$

- $d'(0) = b$

$$V^{d'}(0) = \frac{10 - 11 \cdot \lambda}{(1 - \lambda)}, V^{d'}(1) = -\frac{1}{1 - \lambda}$$

# Desafío

¿Cómo resolver $L_d(V) = r_d + \lambda P_d V$ sin invertir la matriz de transición?

¡Teorema de punto fijo!

# Teorema de Punto Fijo

## Definición: Contracción

Una función $f: \mathbb{R}^n \to \mathbb{R}^n$ es una contracción si existe $0 < c < 1$ tal que

$$\|f(V) - f(U)\| \leq c \cdot \|V - U\|, \text{ para cualquier } U, V \in \mathbb{R}^n$$

## Teorema de Punto Fijo:

Para una contracción $f: \mathbb{R}^n \to \mathbb{R}^n$ se cumple que:

a. El sistema vectorial $V = f(V)$ posee una sola solución $V^*$.

b. Dado $V^0$, la secuencia $V^{n+1} = f(V^n)$ converge a $V^* = \lim_{n\to\infty} V^n$

# Evaluación numérica de política

**Teorema: $L_d$ es una contracción**

- $L_d(V) = r_d + \lambda P_d V$ es una contracción, por lo que $V^{n+1} = L_d(V^n)$ converge a $V^d$.

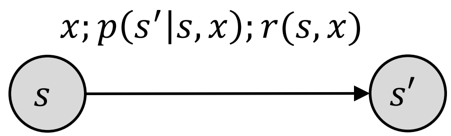
Demostración:

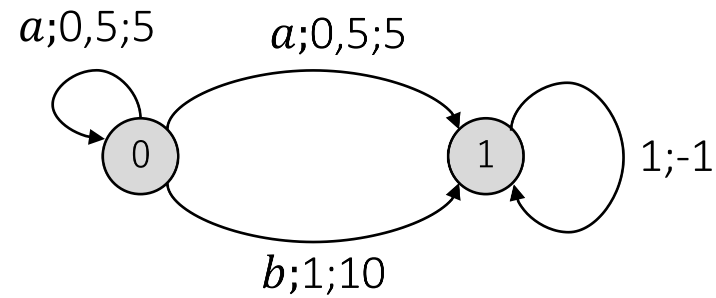
$$\|L_d(V) - L_d(U)\| = \|\lambda P_d V - \lambda P_d U\|$$

$= \lambda\|P_d(V - U)\| \leq \lambda\|V - U\|$ ($P_d$ es una matriz de transición)

Corolario: El valor de la política $d$ se obtiene iterando.

# Calculemos numéricamente el ejemplo ($\lambda = 0{,}8$)




Dos posibles políticas estacionarias:

- $d(0) = a, d(1) = a$
  $$V^d(0) = \frac{5 - 5{,}5\lambda}{(1 - 0{,}5\lambda)(1-\lambda)} = 5, V^d(1) = -\frac{1}{1-\lambda} = -5$$
- $d'(0) = b, d'(1) = a$
  $$V^{d'}(0) = \frac{10 - 11\lambda}{(1-\lambda)} = 6, V^{d'}(1) = -\frac{1}{1-\lambda} = -5$$


- Introducción a MDPs de horizonte infinito

- Retorno descontado: Evaluación de política

- Retorno descontado: Optimización

# ¿Cómo resolver $V^*(s) = \max_{\pi \in \Pi^{MD}} V^\pi(s)$?

Aunque no existe valor terminal, se cumple el principio de recursión:

$$V^*(s) = \max_{x \in \mathbb{X}(s)} \left\{ r(s,x) + \lambda \sum_{j \in \mathbb{S}} p(j|s,x) V^*(j) \right\}$$

Defina la función $L^*(V): \mathbb{R}^{n_s} \to \mathbb{R}^{n_s}$ como:

$$L^*(V) = \left\{ \max_{x \in \mathbb{X}(s)} \left\{ r(s,x) + \sum_{j \in \mathbb{S}} \lambda p(j|s,x) V(j) \right\} \right\}_{s \in \mathbb{S}}$$

El vector valor óptimo $V^* \in \mathbb{R}^{n_s}$ cumple la **ecuación vectorial de Bellman:**

$$\boldsymbol{V^* = L^*(V^*)}$$

# Condiciones de optimalidad

**Teorema:**

Si $V \in \mathbb{R}^{n_s}$ es un vector tal que:

a. $V \geq L^*(V)$ entonces $V \geq V^*$ ($V$ es una cota superior).

b. $V \leq L^*(V)$ entonces $V \leq V^*$ ($V$ es una cota inferior).

c. $V = L^*(V)$ entonces $V = V^*$ ($V$ es el vector óptimo).

Demostración.

# Optimización de MDPs infinitos

Teorema: $L^*$ es una contracción

La función $L(V)$, donde para cada $s \in \mathbb{S}$ por

$$L(V)(s) = \max_{x \in \mathbb{X}(s)} \left\{ r(s,x) + \lambda \sum_{j \in \mathbb{S}} p(j|s,x) V(j) \right\}$$

es una contracción, por lo que la serie $V^{n+1} = L(V^n)$ converge a $V^*$.

Corolario: El valor de la política óptima es único (trivial).

Corolario 2: Existe una política óptima estacionaria.

# Existencia de política óptima estacionaria

**Teorema de existencia:**

Si

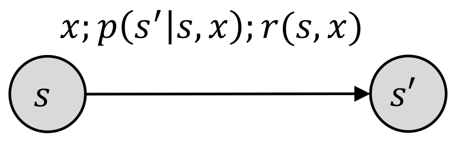
1. $\mathbb{S}$ es contable y
2. $\mathbb{X}$ es finito o

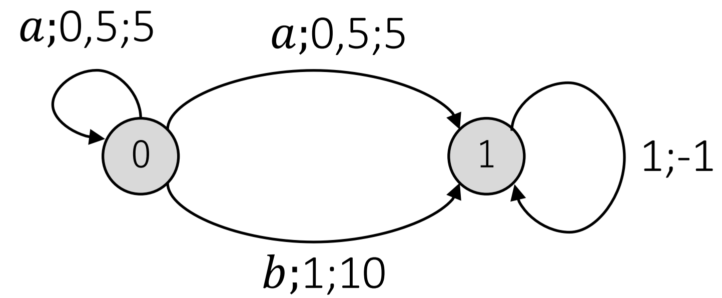
   $\mathbb{X}$ es compacto y $r(s, x), p(j|s, x)$ son continuas en $x$

Entonces existe una política de decisión óptima Markoviana-determinística y **estacionaria**.

La prueba es directa del resultado anterior.

# Optimizemos numéricamente ($\lambda = 0{,}8$)




- $d^*(0) = b$

- $V^* = \begin{bmatrix} 6 \\ -5 \end{bmatrix}$

# Tercera Parte:
# Clase 1 – MDPs con Horizonte Infinito

Optimización Dinámica - ICS

Mathias Klapp - 2020