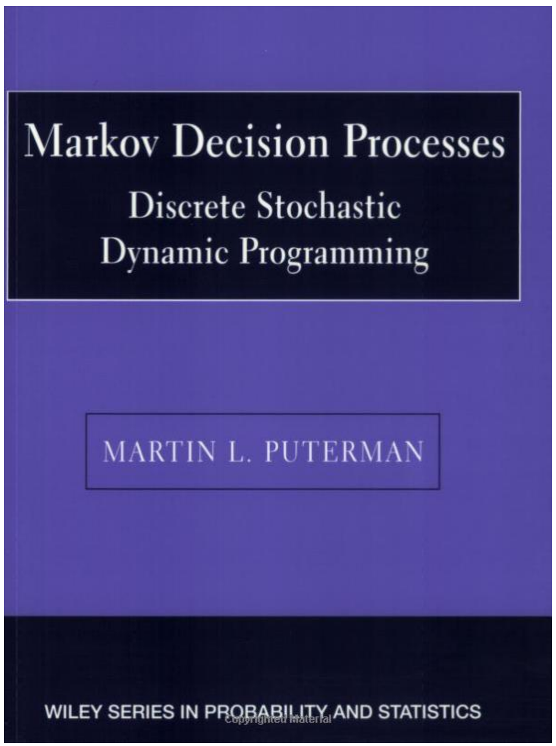
Segunda Parte:
Clase 1 - Procesos de Decisión Markoviana

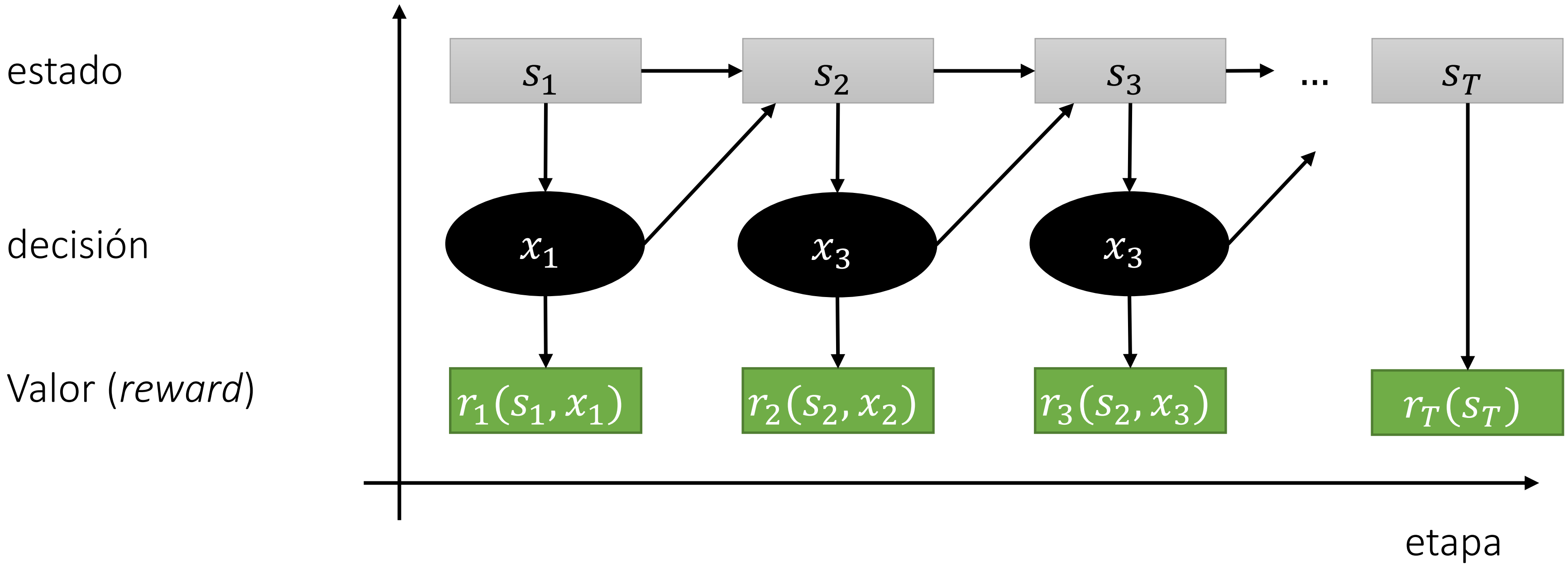
Optimización Dinámica - ICS



Martin L. Puterman

Mathias Klapp

El proceso de decision determinístico:



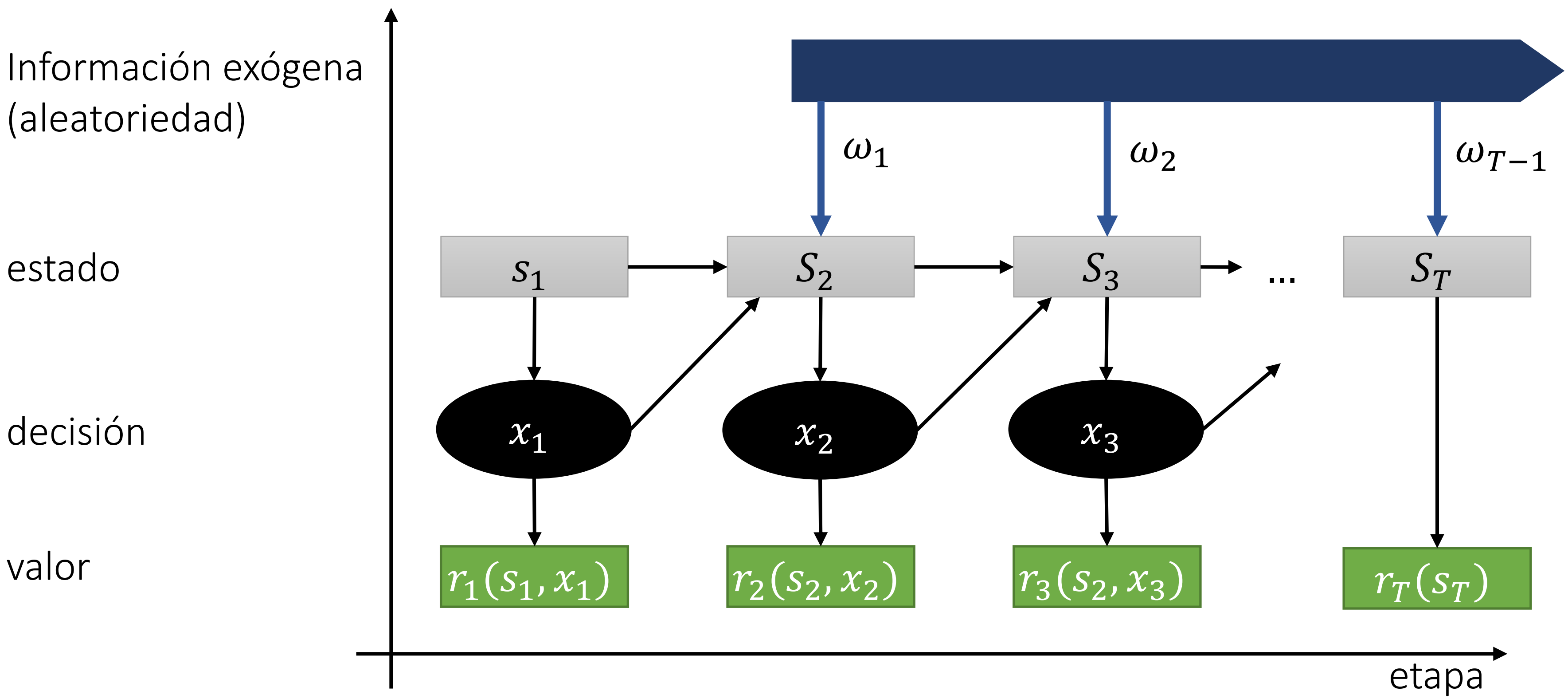
Función de transición de estados:

$$s_{t+1} = f_t(s_t, x_t)$$

Problema:

$$\max_x \sum_{t=1}^{T} r_t(s_t, x_t)$$

Ahora: decisiones bajo incertidumbre



1. El estado del sistema es perturbado por variables aleatorias ω_t:
$$S_{t+1} = f_t(S_t, x_t, \omega_t)$$

2. Objetivo: maximizar **valor esperado**.

Elementos del problema

1. **Etapas de decisión** (finitas):

$$t \in \{1, \dots, T\}$$

2. **Espacio de estados por etapa:**

$$\mathbb{S}_t, \qquad t \in \{1, \dots, T\}$$

- Generalmente asumiremos espacio de estados $\mathbb{S}_t$ contable,
    - **Teoría** aplica para conjuntos $\mathbb{S}_t$ compactos y/o Borelianos.

3. **Espacio de decision** (compacto o contable):

$$\mathbb{X}_t(s_t), \quad s_t \in \mathbb{S}_t, t \in \{1, \dots, T\},$$

Elementos del problema

4. Función de transición estocástica:

Estado futuro $\longrightarrow$ $S_{t+1} = f_t(s_t, x_t, \omega_t)$

Estado

Decisión

Aleatoriedad

Define transición al futuro que depende de una variable exógena aleatoria:

$$\omega_t \sim F_t$$

Equivalente a **probabilidad de transición** $p_t\colon \mathbb{S}_t \times \mathbb{X}_t \to \mathbb{S}_{t+1}$:

$$p_t(s_{t+1}|x_t, s_t) = \mathbb{P}(S_{t+1} = s_{t+1}|x_t, S_t = s_t)$$

Elementos del problema

5. Beneficio inmediato *(reward)* en etapa t:

$$r_t : \mathbb{S}_\mathrm{t} \times \mathbb{X}_\mathrm{t} \rightarrow \mathbb{R}$$

6. Objetivo:

$$\max \mathbb{E}\left(\sum_{t=1}^{\mathrm{T}} r_t(S_t, x_t)\Big|_{s_1}\right)$$

- ¿Cuál es la política de decisión óptima?
    - ¿x_t determinístico o estocástico?
    - ¿Función del estado o historia?: $x_t(S_t)$ v/s $x_t(S_1, x_1, S_2, x_2 \dots, S_t)$

- Proceso de Decisión Markoviana
- Políticas y reglas de decisión
- Optimalidad en MDP
- Evaluación de una política
- Ecuaciones generales Bellman

Proceso de Decisión Markoviana
(MDP: Markov Decision Process)

Un **MDP** es un proceso de decisión secuencial definido por:

1. Etapas de decisión: $1, \dots, T$
2. Espacio de estados: $\mathbb{S}_t$
3. Espacio de decisión: $\mathbb{X}_t(s_t)$
4. Probabilidad de transición: $p_t(s_{t+1}|s_t, x_t)$
5. Valor inmediato: $r_t(s_t, x_t)$, supuesto $|r_t(s_t, x_t)| < \infty$

Objetivo:

$$\max_{\pi \in \Pi} \mathbb{E}\left(\sum_{t=1}^{\mathrm{T}} r_t(S_t, x_t^{\pi}) \,|S_1\right)$$

π: política de decisión

Casos particulares:

- Si probabilidad de transición $p_t(s_{t+1}|s_t, x_t)$ es 1 $\mathbf{o}$ 0, entonces es un **problema de ruta mínima** en grafo acíclico.

- Si espacio de decisión $\mathbb{X}_t(s_t)$ es un *sigleton* (hay una via de acción), entonces es una **cadena de Markov en tiempo discreto** acíclica.

Ejemplo: Control de inventario

- Dinámica de un problema de inventario estocástico
- Bodega de tamaño Q.

En cada periodo $t \in \{1, \ldots, T\}$:

1. Se observa un inventario inicial de s_t unidades.
2. Se decide si reponer x_t unidades a costo $g_t(x_t)$.
3. Se realiza demanda aleatoria $D_t \sim F_t$.
4. Se paga quiebre de stock a $\$q_t$ por unidad.
5. Transición a estado $S_{t+1} = (s_t + x_t - D_t)^+$.
6. Se paga costo de inventario $\$h_t$ por unidad almacenada al siguiente periodo.

Ejemplo: Control de inventario

- Etapas $\{1, \dots, T\}$,
- Estados $\mathbb{S} = \{0, \dots, Q\}$,
- Acciones $\mathbb{X}(s) = \{0, \dots, Q - s\}$

Costo inmediato:

$$r_t(s_t, x_t, D_t) = g_t(x_t) + q \cdot \left((D_t - s_t - x_t)^+\right) + h_t \cdot (s_t + x_t - D_t)^+)$$

en promedio:

$$r_t(s_t, x_t) = g_t(x_t) + q_t \cdot \mathbb{E}_{D_t}\left[\left((D_t - s_t - x_t)^+\right)\right] + h_t \cdot \mathbb{E}_{D_t}[(s_t + x_t - D_t)^+)]$$

Probabilidad de transición de estado:

$$p_t(s_{t+1}|s_t, x_t) = \begin{cases} f_t(s_t + x_t - s_{t+1}) & \text{si} \quad 0 < s_{t+1} < s_t + x_t \\ \sum_{k \geq s_t + x_t} f_t(k) & \text{si} \quad s_{t+1} = 0 \\ 0 & \text{e.o.c.} \end{cases}$$

Ejemplo: Control de inventario

Objetivo:

$$\min_{?} \mathbb{E}_{(S_1,\dots,S_T)} \left(\sum_{t=1}^{\mathrm{T}} r_t(S_t, x_t) \,|S_1 = s_1 \right)$$

¿Cuál es la decision?

- Proceso de Decisión Markoviana
- Políticas y reglas de decisión
- Optimalidad en MDP
- Evaluación de una política
- Ecuaciones generales Bellman

Reglas de Decisión $d_t(\cdot)$

Es un procedimiento para decidir una acción x_t en una etapa t en función de información disponible.

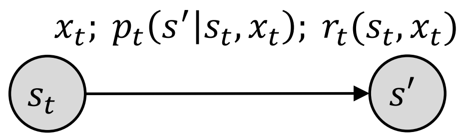
Tipos de regla de decisión:

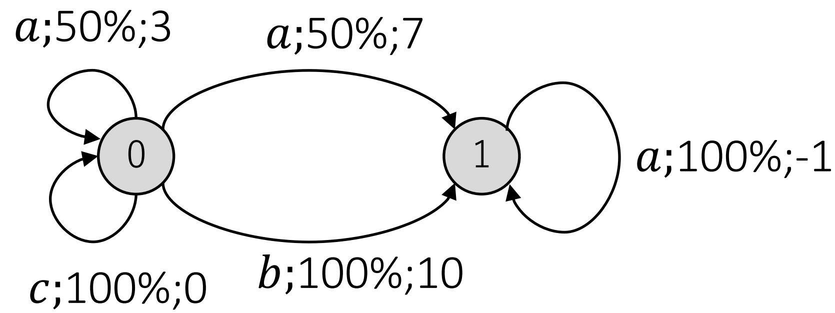
- Regla Markoviana Determinística (MD): Escoge acción $x_t \in \mathbb{X}_t$ de forma determinística en función del estado s_t:
$$d_t: \mathbb{S}_t \to \mathbb{X}_t$$

- Regla Markoviana (M): Escoge una acción aleatoriamente desde distribución de probabilidades $\rho(\mathbb{X}_t)$ sobre $\mathbb{X}_t$. Distribución depende del estado:
$$d_t: \mathbb{S}_t \to \rho(\mathbb{X}_t)$$

- Regla Histórico-dependiente (H): Escoge acción en función de toda la historia (de estados y decisiones) hasta s_t.
  - Puede ser determinística o randomizada (HD o HR).
$$d_t: \mathbb{H}_t \to \mathbb{X}_t$$

Ejemplo: MDP con 2 estados





$$\mathbb{X}(0) = \{a, b, c\}$$
$$\mathbb{X}(1) = \{a\}$$

Veamos un ejemplo de:

- Regla de decisión Markoviana determinística.
- Regla de decisión Markoviana randomizada.
- Regla de decisión Histórico-dependiente randomizada.

Política (revisada)

Una **política** $\pi = (d_1^\pi, d_2^\pi, \dots, d_T^\pi) \in \Pi$ es un vector de reglas de decisión.

- Π : es el conjunto de todas las políticas histórico-dependientes.
- $\Pi^{MD} \subset \Pi^M \subset \Pi$

π se dice **estacionaria** si: $\pi \ = (d^\pi, d^\pi, \dots, d^\pi)$

- $\Pi^S \subset \Pi^{MD}$: es el conjunto de todas las políticas estacionarias.

Pregunta: Para garantizar optimalidad:

- ¿Tenemos que buscar en todo Π?
- ¿Bastaría con Π^M? ¿Π^{HD}? ¿Π^{MD}? ¿Π^S?

- Proceso de Decisión Markoviana
- Políticas y reglas de decisión
- Optimalidad en MDP
- Evaluación de una política
- Ecuaciones generales Bellman

Optimalidad en MDP

Sea $\boldsymbol{V^{\pi}(s)}$ **el valor esperado de una política** $\boldsymbol{\pi}$:

$$V^{\pi}(s) := \mathbb{E}\left[\sum_{k=1}^{T} r_t(S_t, d_t^{\pi}(h_t)) \middle| S_1 = s\right]$$

Una **política** $\boldsymbol{\pi}^*$ es óptima si para todo estado inicial s:

$$V^{\pi^*}(s) \geq V^{\pi}(s), \forall \pi \in \Pi$$

El **valor óptimo del MDP** para el estado inicial s es:

$$V^*(s) = \sup_{\pi \in \Pi} V^{\pi}(s)$$

Suficiencia de políticas determinísticas

Teorema:

Si $|r_t(s_t, x_t)| < \infty$:

La existencia de una política óptima garantiza la existencia de una política óptima **determinística**.

Esta es la intución:

$$\text{Sea } V^*(s) = \sup_{x \in \mathbb{X}} f(x, s)$$

→ valor óptimo sobre soluciones determinísticas

Para todo $s \in \mathbb{S}$ y cualquier distribución de probabilidad $\mathbb{P}$:

$$V^*(s) = \sum_{x \in \mathbb{X}} \mathbb{P}(X = x, s) \cdot V^*(s)$$

$$\geq \sum_{x \in \mathbb{X}} \mathbb{P}(X = x, s) \cdot f(x, s) = \mathbb{E}_X(f(X, s))$$

Consecuencia: La mejor decision determinística es mejor o igual que toda decisión randomizada. Es suficiente buscar en Π^{HD}

- Proceso de Decisión Markoviana
- Políticas y reglas de decisión
- Optimalidad en MDP
- Evaluación de una política
- Ecuaciones generales Bellman

Evaluación recursiva de una política

Sea $\boldsymbol{V_t^{\pi}(h_t)}$ el valor esperado desde la etapa $\boldsymbol{t}$ para una política $\boldsymbol{\pi} = (d_1^{\pi}, \dots, d_T^{\pi})$ dada una historia $h_t = (s_1, x_1, s_2, x_3, \dots, s_t)$ en t:

$$V_t^{\pi}(h_t) = \mathbb{E}\left[\sum_{k=t}^{T} r_k(S_k, d_k^{\pi}(h_k)) \middle| h_t\right]$$

Observaciones:

1. $V^{\pi}(s_1) = V_1^{\pi}(s_1)$
2. $V_t^{\pi}(h_t) = r_t(s_t, d_t^{\pi}(h_t)) + \sum_{s_{t+1} \in \mathbb{S}_{t+1}} p(s_{t+1}|s_t, d_t^{\pi}(h_t)) \cdot V_{t+1}^{\pi}(\overbrace{h_t, d_t^{\pi}(h_t), s_{t+1}}^{h_{t+1}})$

Probemos lo segundo:

$$V_t^{\pi}(h_t) = r_t(s_t, d_t^{\pi}(h_t)) + \mathbb{E}\left[\textstyle\sum_{k=t+1}^{T} r_k(S_k, d_k^{\pi}(h_k)) \middle| h_t\right]$$

$$= r_t(S_t, d_t^{\pi}(h_t)) + \mathbb{E}_{S_{t+1}}\left[\mathbb{E}\left[\textstyle\sum_{k=t+1}^{T-1} r_k(S_k, d_k^{\pi}(h_k)) \middle| h_t, S_{t+1}\right] \middle| h_t\right]$$

$$= r_t(S_t, d_t^{\pi}(h_t)) + \mathbb{E}_{S_{t+1}}[V_{t+1}^{\pi}(h_t, d_t^{\pi}(h_t), S_{t+1})|h_t]$$

Evaluación recursiva de una política

$V_t^\pi(h_t)$ se puede calcular recursivamente:

Para cada $h_T \in \mathbb{H}_T$:

$$V_T^\pi(h_T) = r_T\big(s_T, d_T^\pi(h_T)\big),\ \forall h_t \in \mathbb{H}_t$$

Para cada $t = T-1, \dots, 1$ y cada $h_t \in \mathbb{H}_t$:

$$V_t^\pi(h_t) = r_t\big(s_t, d_t^\pi(h_t)\big) + \sum_{s_{t+1}\in\mathbb{S}_{t+1}} p(s_{t+1}|s_t, d_t^\pi(h_t)) \cdot V_{t+1}^\pi\big((h_t, d_t^\pi(h_t), s_{t+1})\big)$$

Dos preguntas:

- ¿Cómo eliminar dependencia histórica? (cambiar h_t por s_t)

- Proceso de Decisión Markoviana
- Políticas y reglas de decisión
- Optimalidad en MDP
- Evaluación de una política
- Ecuaciones generales Bellman

Ecuaciones generales de Optimalidad (Bellman)

Sea $V_t^*(h_t)$ el ***value-to-go*** (máximo valor esperado) *en* t dada historia h_t:

$$V_t^*(h_t) = \sup_{\pi \in \Pi^{HD}} V_t^\pi(h_t)$$

El principio de recursión es válido:

$$V_t^*(h_t) = \sup_{\pi \in \Pi^{HD}} \{r_t(S_t, d_t^\pi(h_t)) + \mathbb{E}_{S_{t+1}}[V_{t+1}^\pi(h_t, d_t^\pi(h_t), s_{t+1})|h_t]\}$$

$$= \sup_{d_t \in \mathbb{X}_t(s_t)} \{r_t(S_t, x_t) + \sup_{\pi \in \Pi^{HD}} \mathbb{E}_{S_{t+1}}[V_{t+1}^\pi(h_t, x_t, s_{t+1})|h_t]\}$$

$$= \sup_{d_t \in \mathbb{X}_t(s_t)} \{r_t(S_t, x_t) + \textstyle\sum_{s_{t+1} \in \mathbb{S}_{t+1}} p_t(s_{t+1} \,|s_t, x_t) \sup_{\pi \in \Pi^{HD}} V_{t+1}^\pi(h_t, x_t, s_{t+1})\}$$

$$= \sup_{x_t \in \mathbb{X}_t(s_t)} \left\{r_t(s_t, x_t) + \textstyle\sum_{s_{t+1} \in \mathbb{S}_{t+1}} p_t(s_{t+1}|s_t, x_t)\, V_{t+1}^*(h_t, x_t, s_{t+1})\right\}$$

Ecuaciones generales de Bellman

Para cada $h_T \in \mathbb{H}_T$:

$$V_T^*(h_T) = \sup_{x_T \in \mathbb{X}_T(s_T)} r_t(s_T, x_T)$$

Para cada $t = T-1, \dots, 1$ y cada $h_t \in \mathbb{H}_t$:

$$V_t^*(h_t) = \sup_{x_t \in \mathbb{X}_t(s_t)} \left\{ r_t(s_t, x_t) + \sum_{s_{t+1} \in \mathbb{S}_{t+1}} p_t(s_{t+1} | s_t, x_t)\, V_{t+1}^*(h_t, x_t, s_{t+1}) \right\}$$

Finalmente $V^*(s) = V_1^*(s)$

Condición de optimalidad

Teorema:

Si una política π posee valores $\boldsymbol{V_t^{\pi}}$ que cumplen:

1. Para cada $h_T \in \mathbb{H}_T$:

$$V_t^{\pi}(h_T) = \sup_{x_T \in \mathbb{X}_T(s_T)} r_T(s_T, x_T)$$

2. Para cada $t = T-1, \dots, 1$ y cada $h_t \in \mathbb{H}_t$:

$$V_t^{\pi}(h_t) = \sup_{x_t \in \mathbb{X}_t(s_t)} \left\{ r_t(s_t, x_t) + \sum_{s_{t+1} \in \mathbb{S}_{t+1}} p_t(s_{t+1} | s_t, x_t) \, V_{t+1}^{\pi}(h_t, x_t, s_{t+1}) \right\}$$

, entonces π es una **política óptima**.

Condición de optimalidad de una política

Demotración por inducción:

- Es claro que $V_T^\pi(h_T) = V_T^*(h_T)$

- Hipótesis de inducción (HI):
$$V_{t+1}^\pi(h_{t+1}) = V_{t+1}^*(h_{t+1})$$

- Prueba iterativa:
$$V_t^\pi(h_t) = \sup_{x_t \in \mathbb{X}_t(s_t)} \left\{ r_t(s_t, x_t) + \sum_{s_{t+1} \in \mathbb{S}_{t+1}} p_t(s_{t+1} | s_t, x_t)\, V_{t+1}^\pi(h_t, x_t, s_{t+1}) \right\}$$

por HI:
$$V_t^\pi(h_t) = \sup_{x_t \in \mathbb{X}_t(s_t)} \left\{ r_t(s_t, x_t) + \sum_{s_{t+1} \in \mathbb{S}_{t+1}} p_t(s_{t+1} | s_t, x_t)\, V_{t+1}^*(h_t, x_t, s_{t+1}) \right\}$$

Por lo tanto:
$$V_t^\pi(h_t) = V_t^*(h_t)$$

¿Cómo tener independencia de historia?

Suficiencia de política Markoviana Determinística

Versión 1:

Si $\mathbb{S}_t$ es contable, $\mathbb{X}_t(s)$ es finito y $r_t(s,x)$ es acotada,

entonces **existe política óptima MD (Markoviana y determinística)**.

Versión 2:

Si $\mathbb{S}_t$ es contable,

1. $\mathbb{X}_t(s)$ es compacto (cerrado y acotado)
2. $r_t(s,x)$ es acotada y continua en x
3. $p_t(j|s,x)$ es continua en x

entonces **existe política óptima MD (Markoviana y determinística)**.

Suficiencia de decisiones Markovianas

Demostración intuitiva: se requiere probar tres pasos…

1. $V_t^*(h_t)$ sólo depende de h_t a través de s_t, es decir $V_t^*(s_t)$. INDUCCIÓN

2. $\exists x_t^* \in \mathbb{X}_t(s_t)$ que resuelve:

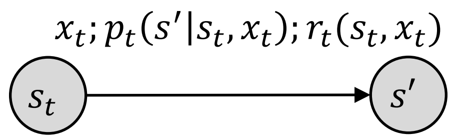
$$V_t^*(s_t) = \sup_{x_t \in \mathbb{X}_t(s_t)} \left\{ r_t(s_t, x_t) + \mathbb{E}_{S_{t+1}}\left(V_{t+1}^*(S_{t+1}) | s_t, x_t\right) \right\}$$

Recordatorio: Teorema Bolzano-Weierstrass

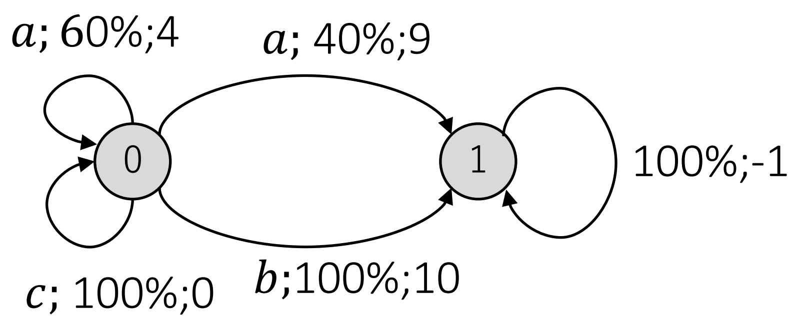
Si problema $\sup\{f(x) : x \in \mathbb{D}\}$ posee $f(x)$ continua sobre $\mathbb{D} \neq \emptyset$ y compacto, entonces **admite al menos una solución óptima**.

3. Objetivo y dominio del problema en la Ecuación del Bellman sólo dependen de s_t, luego $x_t^*(s_t)$. **Esta es la política MD.**

Ejemplo 1: MDP, 2 estados + 2 etapas



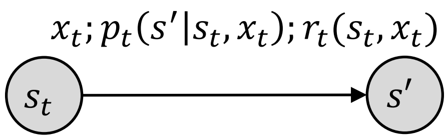
$\mathbb{X}(0) = \{a, b, c\}$



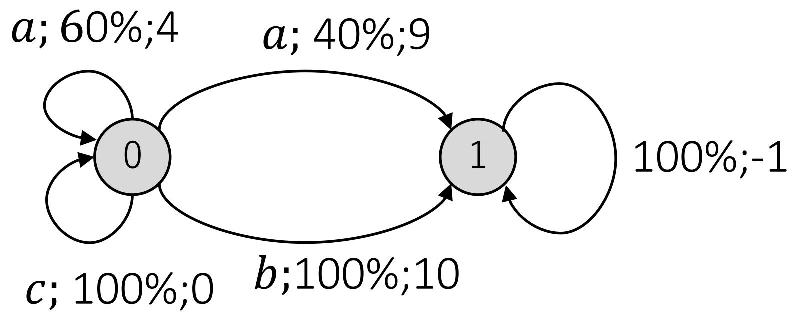
Rewards:

- $r(0, a) = 0{,}6 \cdot 4 + 0{,}4 \cdot 9 = 6$
- $r(0, b) = 10$
- $r(0, c) = 0$
- $r(1) = -1$

Ejemplo 1: MDP, 2 estados + 2 etapas



$\mathbb{X}(0) = \{a, b, c\}$



Etapa 2:

- $V_2^*(0) = \max\{6,10,0\} = 10,$ $\rightarrow d_2^*(0) = b$
- $V_2^*(1) = -1$

Etapa 1:

- $V_1^*(0) = \max\left\{6 + \frac{6V_2^*(0)+4V_2^*(1)}{10}, 10 + V_2^*(1), V_2^*(0)\right\} = 11{,}6, \rightarrow d_1^*(0) = a$
- $V_1^*(1) = -2$

Backward DP para MDP

Valor terminal:

$$V_T^*(s), d_T^*(s) \leftarrow \max_{x \in \mathbb{X}_t(s)} \{r_T(s,x)\}, \qquad \forall s \in \mathbb{S}_T$$

Recursión: para todo $t = T-1, \dots, 1$ y todo $s \in \mathbb{S}_t$:

$$V_t^*(s), d_t^*(s) \leftarrow \max_{x \in \mathbb{X}_t(s)} \left\{r_t(s,x) + \mathbb{E}_{S_{t+1}}[V_{t+1}^*(S_{t+1}) | s, x]\right\}$$

Retornar:

$$\pi^* = (d_1^*, d_2^*, \dots, d_{T-1}^*)$$

1. $\mathcal{O}(T \cdot |\mathbb{S}|)$ problemas.
2. Para cada problema y acción debemos calcular esperanza
3. Caso con estados y decisiones finitas: $\mathcal{O}(T \cdot |\mathbb{S}|^2 \cdot |\mathbb{X}|)$

Ejemplo 2: Control de inventario

Consideremos el problema de control de inventario en bodega de tamaño $Q = 3$ y $T = 3$.

Demanda:

- $P(D = d) = \begin{cases} 25\%, & \text{si } d = 0 \\ 50\%, & \text{si } d = 1, \\ 25\%, & \text{si } d = 2 \end{cases}$ $\qquad \mathbb{E}_D(D) = 1$

- Costos:
  I. Inventario $\$h = 1$
  II. Quiebre $\$q = 15$
  III. Compra $\$g(x) = \begin{cases} 4 + 2 \cdot x & \text{si } x > 0 \\ 0 & \text{en otro caso} \end{cases}$

Resolvamos! Ver planilla..

Ejemplo 2: Control de inventario

- Etapas $\{1, \dots, 3\}$,
- Estados $\mathbb{S} = \{0, \dots, 3\}$,
- Acciones $\mathbb{X}(s) = \{0, \dots, 3 - s\}$

Costo inmediato:

$$r(s, x, D) = 4 \cdot \mathbb{I}_{x>0} + 2 \cdot x + 15 \cdot \left((D - s - x)^{+}\right) + 1 \cdot (s + x - D)^{+})$$

en promedio:

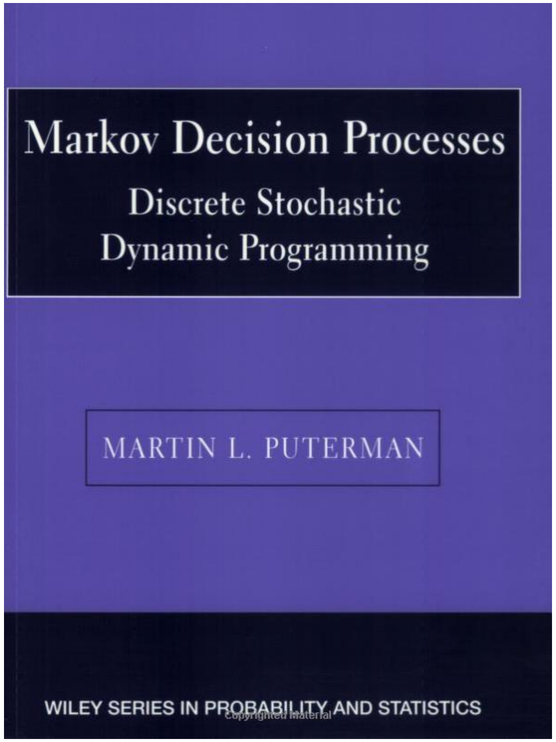
$$r(s, x) = 4 \cdot \mathbb{I}_{x>0} + 2 \cdot x + 15 \cdot \mathbb{E}_D\left[\left((D - s - x)^{+}\right)\right] + 1 \cdot \mathbb{E}_D[(s + x - D)^{+})]$$

Probabilidad de transición de estado:

$$p(s_{t+1} | s + x) = \begin{cases} f(s + x - s_{t+1}) & \text{si} \quad 0 < s_{t+1} \leq s + x \\ \sum_{k \geq s+x} f(k) & \text{si} \quad s_{t+1} = 0 \\ 0 & \text{e. o. c.} \end{cases}$$

Segunda Parte:
Clase 1 - Procesos de Decisión Markoviana

Optimización Dinámica - ICS



Martin L. Puterman

Mathias Klapp