Pontificia Universidad Católica de Chile
Departamento de Ingeniería Industrial y Sistemas
ICS3105 - Optimización Dinámica

Tarea 3

2º semestre 2021 - Matías Klapp
Grupo 5
Anaís Martínez, Alexandra Ovalle, Shun Wei Rao

Problema 1

1. a) El vector de valor esperado futuro descontado asociado a la política d' se define igual a:

$$\mathbf{V}^{d'} = \sum_{t=1}^{\infty} \lambda^{t-1} \cdot \mathbf{P}_{d'}^{t-1} \cdot \mathbf{r}_{d'}$$

como es estacionario, $\mathbf{r}_{d'}$ no depende de t

Si se separa el primer término de la sumatoria, la ecuación queda

$$\mathbf{V}^{d'} = \mathbf{r}_{d'} + \sum_{t=2}^{\infty} \lambda^{t-1} \cdot \mathbf{P}_{d'}^{t-1} \cdot \mathbf{r}_{d'} = \mathbf{r}_{d'} + \sum_{t=2}^{\infty} \lambda\lambda^{t-2} \cdot \mathbf{P}_{d'}\mathbf{P}_{d'}^{t-2} \cdot \mathbf{r}_{d'}$$

Si se sacan las constantes de la sumatoria, y se cambia $t = t + 1$ se obtiene,

$$\mathbf{V}^{d'} = \mathbf{r}_{d'} + \lambda\mathbf{P}_{d'} \sum_{t=2}^{\infty} \lambda^{t-2} \cdot \mathbf{P}_{d'}^{t-2} \cdot \mathbf{r}_{d'} = \mathbf{r}_{d'} + \lambda\mathbf{P}_{d'} \underbrace{\sum_{t=1}^{\infty} \lambda^{t-1} \cdot \mathbf{P}_{d'}^{t-1} \cdot \mathbf{r}_{d'}}_{\mathbf{V}^{d'}}$$

Luego,

$$\mathbf{V}^{d'} = \mathbf{r}_{d'} + \lambda\mathbf{P}_{d'}\mathbf{V}^{d'} \tag{1}$$

Para obtener una fórmula exacta para $\mathbf{V}^{d'}$ se comienza con la ecuación (1)

$$\mathbf{V}^{d'} = \mathbf{r}_{d'} + \lambda \mathbf{P}_{d'} \mathbf{V}^{d'}$$

$$\mathbf{V}^{d'} - \lambda \mathbf{P}_{d'} \mathbf{V}^{d'} = \mathbf{r}_{d'}$$

$$(I - \lambda \mathbf{P}_{d'}) \mathbf{V}^{d'} = \mathbf{r}_{d'} \Rightarrow \mathbf{V}^{d'} = (I - \lambda \mathbf{P}_{d'})^{-1} \mathbf{r}_{d'}$$

b)

$$L_{d'}(\mathbf{V}) = \mathbf{r}_{d'} + \lambda \mathbf{P}_{d'} \mathbf{V}$$

Luego $V^{n+1} = L_{d'}(V^n)$

Usando lo anterior,

$$||L_{d'}(\mathbf{V}^{n'+1}) - L_{d'}(\mathbf{V}^{n'})|| = ||\mathbf{r}_{d'} + \lambda \mathbf{P}_{d'} \mathbf{V}^{n'} - (\mathbf{r}_{d'} + \lambda \mathbf{P}_{d'} \mathbf{V}^{n'-1})|| = ||\lambda \mathbf{P}_{d'} \mathbf{V}^{n'} - \lambda \mathbf{P}_{d'} \mathbf{V}^{n'-1}||$$

$$\begin{aligned} ||\lambda \mathbf{P}_{d'} \mathbf{V}^{n'} - \lambda \mathbf{P}_{d'} \mathbf{V}^{n'-1}|| &= ||\lambda \mathbf{P}_{d'} (\mathbf{V}^{n'} - \mathbf{V}^{n'-1})|| \\ &= \lambda ||\mathbf{P}_{d'}|| \cdot ||\mathbf{V}^{n'} - \mathbf{V}^{n'-1}|| \leq \lambda ||\mathbf{V}^{n'} - \mathbf{V}^{n'-1}|| \end{aligned} \tag{2}$$

Con esto, se obtiene

$$||L_{d'}(\mathbf{V}^{n'+1}) - L_{d'}(\mathbf{V}^{n'})|| \leq \lambda ||\mathbf{V}^{n'} - \mathbf{V}^{n'-1}|| \leq \lambda^{n'} ||\mathbf{V}^1 - \mathbf{V}^0||$$

Se cumple entonces que $\mathbf{V}^{m+1} \geq \mathbf{V}^m$, $\forall m \geq n'$.

2. a)

Si $\mathbf{V} \leq L(\mathbf{V})$, entonces, como $L(\mathbf{V})$ es un máximo, debe existir al menos una regla de decisión, tal que, componente a componente,

$$\mathbf{V} \leq \mathbf{r}_{d'} + \lambda \mathbf{P}_{d'} \mathbf{V}$$

$$\mathbf{V} - \lambda \mathbf{P}_{d'} \mathbf{V} \leq \mathbf{r}_{d'}$$

$$(I - \lambda \mathbf{P}_{d'}) \mathbf{V} \leq \mathbf{r}_{d'}$$

$$\mathbf{V} \leq \underbrace{(I - \lambda \mathbf{P}_{d'})^{-1} \mathbf{r}_{d'}}_{\mathbf{V}^{d'}}$$

Así, $\mathbf{V} \leq \mathbf{V}^{d'} \leq \mathbf{V}^{\star}$, es una cota inferior.

b) Como $L(\mathbf{V})$ está definido como el máximo, y usando la misma factorización de la norma que en ecuación (2),

$$||L(\mathbf{V}^{n'+1}) - L(\mathbf{V}^{n'})|| \geq \lambda ||\mathbf{V}^{n'} - \mathbf{V}^{n'-1}|| \geq \lambda^{n'} ||\mathbf{V}^{1} - \mathbf{V}^{0}||$$

De lo que se obtiene, que $\mathbf{V}^{n'} \leq \mathbf{V}^{m} \leq \mathbf{V}^{\star}$

c)

$$\frac{\epsilon(1-\lambda)}{2\lambda)} = \frac{10^{-2}0{,}05}{2 \cdot 0{,}95} = 0{,}0002$$

Problema 2

1. A continuación se modela el MDP de horizonte infinito del estudiante, donde las probabilidades y transiciones de estados están representadas gráficamente, ver Figura 1. Las ecuaciones de optimalidad de Bellman con la función *value-to-go* son:

$$V(0) = \text{máx} \left\{ 1 + \lambda \cdot V(0), 2 + 5 \cdot \lambda \cdot \frac{(V(0) + V(1))}{10} \right\} [2mm]$$

$$V(1) = \text{máx} \left\{ 1 + 5 \cdot \lambda \cdot \frac{(V(0) + V(1))}{10}, 2 + 5 \cdot \lambda \cdot \frac{(V(1) + V(2))}{10} \right\}$$

$$V(2) = 0 + \lambda \cdot V(2)$$

Donde $\lambda = \frac{4}{5}$

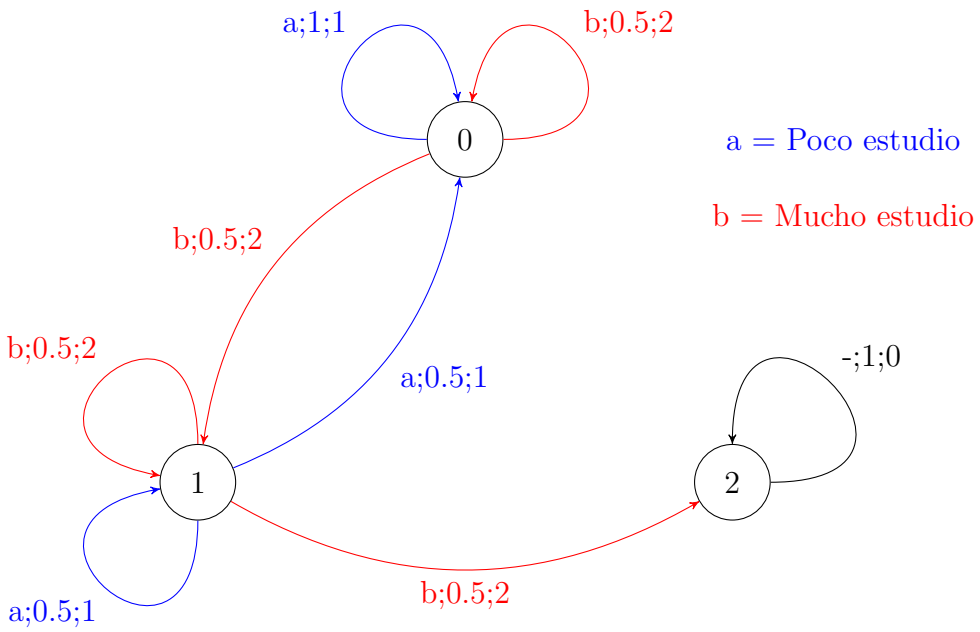

Figura 1: Estados y transiciones del estudiante

2. Para evaluar la política π : Estudiar poco, independiente del estado, se usa el método de invertir, $V^d = (I - \lambda P_d)^{-1} r_d$, donde $d = a$ para todos los estados, excepto 2 que no tiene decisión.

Así, El retorno de la política:

$$r_d = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \tag{3}$$

La matriz probabilidad de transición:

$$P_d = \begin{bmatrix} 1 & 0 & 0 \\ 0{,}5 & 0{,}5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

De modo que,

$$V^d = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - 4/5 \begin{bmatrix} 1 & 0 & 0 \\ 0{,}5 & 0{,}5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 5 \\ 5 \\ 0 \end{bmatrix} \tag{5}$$

Así, es valor de la política π, vale cinco para los estados 0 y 1, y vale cero para el estado 2. En los puntos 3 y 4, que se desarrollarán a continuación, se obtendrán las decisiones óptimas, las cuales muestran que en un estado de descansado (0) convendría estudiar mucho. A diferencia de esta política de estudiar poco en todos los estados, obtiene valores esperados de ocho para el estado 0 y siete para el estado 1.

3. Al resolver el problema a optimalidad, a través de iteración de políticas desde la política estudiar poco, basándonos en el código compartido por el profesor adaptado a este problema (ver detalle en archivos adjuntos), se obtienen los resultados del Cuadro 1, donde s: estado, X: acción y V: valor esperado.

Tabla 1: Resultados iteración de políticas

s	X	V
0	b	8
1	a	7
2	-	0

Los resultados del tabla 1, se obtuvieron en 0.00096 segundos y convergieron a las 2 iteraciones.

4. Al resolver el problema a optimalidad, a través de iteración de valor desde valores nulos, basándonos en el código compartido por el profesor adaptado a este problema (ver detalle en archivos adjuntos), se obtienen los resultados del Cuadro 2, donde s: estado, X: acción y V: valor esperado.

Tabla 2: Resultados iteración de valor

s	X	V
0	b	8
1	a	7
2	-	0

Los resultados del tabla 2, se obtuvieron en 0.000995 segundos y convergieron a las 95 iteraciones.

Por lo tanto, la solución iteración de valor es menos inteligente que la iteración de políticas, dado que no evalúa cada política garantizando una solución mejor en cada iteración. Esta diferencia de complejidades se ve reflejada en el tiempo total de cómputo (0.000995 y 0.00096 segundos, respectivamente) y cantidad de iteraciones necesarias para converger (95 y 2 iteraciones, respectivamente), donde el método iteración de políticas requirió mucho más tiempo en hacer cada iteración, pero, a su vez, fueron "jugadas estratégicas"que

requirieron menos cantidad de iteraciones para converger, lo cual puede ser de gran ayuda en problemas grandes, en los cuales se notaría una mayor diferencia en el cómputo final de tiempo de ejecución de las soluciones.

Problema 3

1. Estados: $s = \{1, .., 100\}$

Acciones: $P = \{p_1, p_2, p_3, p_4\}$

Probabilidades de transición:

$$p(s'|s, p_i) = \begin{cases} p_i(1-p_i)^{\sigma} & \text{si } s' > s \\ 1 - \sum_{\sigma=0}^{100-s-1} p_i(1-p_i)^{\sigma} & \text{si } s' = 1 \\ 0 & eoc \end{cases}$$

donde $\sigma = s' - s - 1$

Las ecuaciones de Bellman:

$$V(s) = \max_{p} \left\{ -1 + \lambda \sum_{j} p(j|s, p) V(j) \right\}$$

Se emplea el método de iteración de valor partiendo de $\mathbf{V}^0 = 0$ y se itera hasta que la diferencia sea de 10^{-10}. El costo inmediato de subir un piso es de 1 mientras que se fija un valor de 100 al llegar al piso 100. Finalmente en la figura 2 se muestra la política óptima de acción que se toma en cada estado.

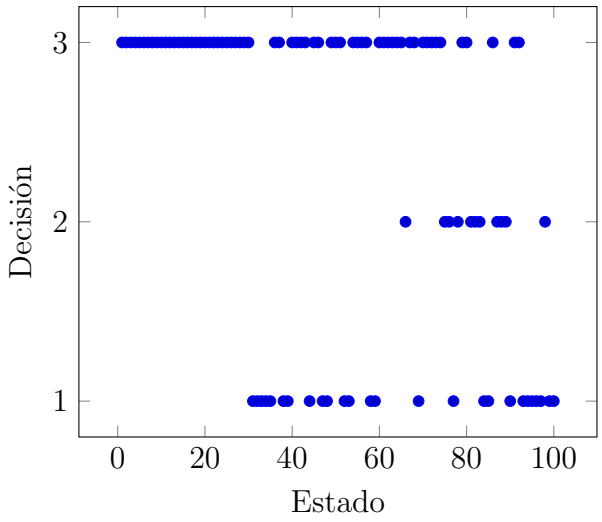

Figura 2: Decisión óptima a tomar en cada estado

Problema 4

1. Etapas por cada semana $t \in T = [1, \ldots, \infty]$
Estado como cantidad en cada semana $y_t = [1, \ldots, Q]$
Acciones como cuanto reponer cada semana $x_t(y_t) = [1, \ldots, Q - y_t]$

El costo de tener y, reponer x es:

$$r(y, x) = K\mathbb{I}_{x>0} + cx + h\mathbb{E}(y + x - d)^+ + q\mathbb{E}(d - y - x)^+$$

$D_t \sim Unif(30, 60)$ por lo que las probabilidades de transición son:

$$p(y'|y, x) = \begin{cases} \frac{1}{31} & \text{si } y' \in [0, y + x] \\ 1 - \frac{y+x}{31} & \text{si } y' = 0 \\ 0 & eoc \end{cases}$$

2. Se comienza con un vector inicial $\mathbf{V}^0 = 0$ y se itera hasta alcanzar la cota de precisión. El siguiente vector se obtiene como:

$$\mathbf{V}^{n+1}(s) = \max_x \left\{ r(y, x) + \lambda \sum_j p(j|y, x)\mathbf{V}^n(j) \right\}$$

Y el criterio de parada es:

$$||\mathbf{V}^{n+1} - \mathbf{V}^n||_\infty < 10^{-4}$$

Finalmente la política de decisión es:

$$d(s) = \arg\max_x \left\{ r(y,x) + \lambda \sum_j p(j|y,x)\mathbf{V}^{n+1}(j) \right\}$$

3. Por otro lado la iteración de política se empieza con cualquier política inicial y se calcula V^n asociado. La siguiente política se escoge como:

$$d_{n+1} = \arg\max_x \left\{ r(y,x) + \lambda \sum_j p(j|y,x)\mathbf{V}^{n}(j) \right\}$$

Se para cuando $d_n = d_{n+1}$

4. El problema lineal es:

$$\begin{aligned} &\max_f \sum_y \sum_x r(y,x) f_{y,x} \\ \text{sa} \quad &\sum_x f_{j,x} - \sum_y \sum_z \lambda p(j|y,z) f_{j,z} = \alpha_j \quad \forall j \end{aligned}$$

El problema está maldito por la dimensionalidad y la licencia estudiantil de Gurobi no permite resolver grandes tamaños de problemas.

Se resume en la tabla 3 la política óptima según el λ. También se muestra el tiempo y cantidad de iteraciones empleadas con los diferentes métodos. Del código es posible notar que la política tiene estructura (s, S), es decir, ordenar hasta S cuando se tiene s o menos stock, ya que desde el estado 100 hasta $s+1$ se ordena 0 y después de s se ordena con una diferencia de una unidad hasta llegar a 0 donde se pide S.

Tabla 3: Resumen métodos de iteración para distintos λ

λ	Política	Tiempo VI	Iteraciones VI	Tiempo PI	Iteraciones PI
0.90	(27,148)	29	188	2.5	5
0.95	(29,186)	60	398	3.4	4
0.99	(31,195)	328.9	2182	26	7

De la Tabla 3, es posible notar que a mayor λ el Tiempo PI aumenta, esto se podría deber a que con un mayor λ se le da más peso al futuro, de modo que analiza muchos más escenarios. La iteración de política converge más rápido que la de valor dado que cada iteración de la iteración de política es más inteligente.

5. El problema con *lead time*