ReinforcedLearningClass

Norah Jones

2024-09-09

Tabla de contenidos

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

$$x^2 + 10 = 100 \tag{0.1}$$

Prueba Ecuación 0.1

Referencia ejemplo Ayers (2005) Referencia ejemplo (see Ayers 2005, 52-53; B. Thomas Jr 2010, cap. 1)

Black-Scholes (Ecuación 0.2) is a mathematical model that seeks to explain the behavior of financial derivatives, most commonly options:

$$\frac{\partial \mathrm{C}}{\partial \mathrm{t}} + \frac{1}{2}\sigma^2 \mathrm{S}^2 \frac{\partial^2 \mathrm{C}}{\partial \mathrm{C}^2} + \mathrm{rS}\frac{\partial \mathrm{C}}{\partial \mathrm{S}} = \mathrm{rC} \tag{0.2}$$

1 Tarea 1 (Fecha de Entrega 20 Septiembre 2024 12:00:00)

Ejercicio 1.1. Read (Sec 1.1, pp 1-2 Sutton and Barto 2018) and answer the following. Explain why Reinforcement Learning differs for supervised and unsupervised.

El aprendizaje supervisado requiere de ejemplos de soluciones. Mientras que el reforzado requiere una función de valor.

Ejercicio 1.2. See the first Brunton's youtube about Reinforced Learning. Then acordingly to its presentation explain what is the meaning of the following expression.

$$V_\pi(s) = E\left(\sum_t \gamma^t r_t \mid s_0 = s\right)$$

La expresión presentada en el video Reinforcement Learning.

$$V_\pi(s) = E\left[\sum_t \gamma^t r_t \middle| s_0 = s\right]$$

hace referencia a la función de valor del problema de optimización representada por la recompensa esperada dado la politica π y el estado inicial s. Aquí γ es el factor de descuento y r_t es la recompensa por etapa t.

Ejercicio 1.3. Form (see Sutton and Barto 2018, sec. 1.7) obtain a time line pear year from 1950 to 2012.

```
library(devtools)
```

```
Warning: package 'devtools' was built under R version 4.3.3
```

```
Loading required package: usethis
```

```
Warning: package 'usethis' was built under R version 4.3.3
```

```r
library(milestones)

library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.3.3

Warning: package 'ggplot2' was built under R version 4.3.3

Warning: package 'tibble' was built under R version 4.3.3

Warning: package 'tidyr' was built under R version 4.3.3

Warning: package 'readr' was built under R version 4.3.3

Warning: package 'dplyr' was built under R version 4.3.3

Warning: package 'forcats' was built under R version 4.3.3

Warning: package 'lubridate' was built under R version 4.3.3

-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```r
library(gt)
```

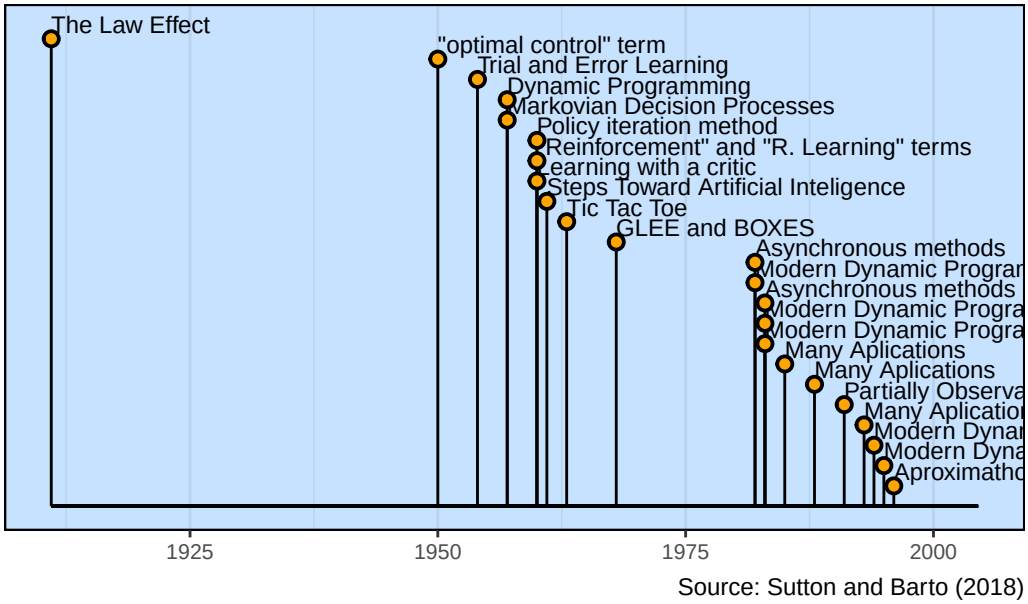
```
Warning: package 'gt' was built under R version 4.3.3
```

```
#library(bibtex)
## Activate the Core Packages
#biblio <- bibtex::read.bib("references.bib")

## Initialize defaults
column <- lolli_styles()

data <- read_csv(col_names=TRUE, show_col_types=FALSE, file='rl_time_line.csv')


data <- data |> arrange(date)

## Build a table
gt(data) |>
  #cols_hide(columns = event) |>
  tab_style(cell_text(v_align = "top"),
            locations = cells_body(columns = date)) |>
  tab_source_note(source_note = "Source: Sutton and Barto (2018)")
```

```
column$color <- "orange"
column$size  <- 15
column$source_info <- "Source: Sutton and Barto (2018)"

## Milestones timeline
milestones(datatable = data, styles = column)
```

date	event	reference
1911	The Law Effect	Thorndike, 1911
1950	"optimal control" term	MR0090477 Bellman, Richard Dynamic programmi
1954	Trial and Error Learning	Minsky,Farley, Clark 1954
1957	Dynamic Programming	MR0090477 Bellman, Richard Dynamic programmi
1957	Markovian Decision Processes	MR0091859 Bellman, Richard A Markovian decisio
1960	Policy iteration method	Ron Howard (1960)
1960	"Reinforcement" and "R. Learning" terms	NA
1960	Learning with a critic	Hoff, 1960
1961	Steps Toward Artificial Inteligence	Minsky, 1961
1963	Tic Tac Toe	Donald Michie, 1961
1968	GLEE and BOXES	Michie and Chambers 1968
1982	Asynchronous methods	Bertsekas, 1982
1982	Modern Dynamic Programming	White 1982
1983	Asynchronous methods	MR0712113 Bertsekas, Dimitri P. Distributed asyn
1983	Modern Dynamic Programming	MR0749232 Ross, Sheldon Introduction to stochasti
1983	Modern Dynamic Programming	White 1983
1985	Many Aplications	MR1295629 White, D. J. Markov decision processes
1988	Many Aplications	MR1295629 White, D. J. Markov decision processes
1991	Partially Observable MDPs	MR1105166 Lovejoy, William S.A survey of algorith
1993	Many Aplications	MR1200993 White, D. J.Markov decision processes:
1994	Modern Dynamic Programming	Puterman, 1994
1995	Modern Dynamic Programming	Bertsekas, 1995
1996	Aproximathon methods	MR1416619 Rust, John Numerical dynamic progra

Source: Sutton and Barto (2018)



Ejercicio 1.4. Consider the following comsuption-saving problem with dynamics

$$x_{k+1} = (1+r)(x_k - a_k), k = 0, 1, \dots, N-1$$

and utility function

$$\beta^N (x_N)^{1-\gamma} + \sum_{k=0}^{N-1} \beta^k (a_k)^{1-\gamma}.$$

Show that the value functions of the DP alghorithm the form

$$J_k(x) = A_k k \beta^k x^{1-\gamma},$$

where $A_N = 1$ and for $k = N-1, \dots, 0$,

$$A_k = \left[1 + \left((1+r)\beta A_{k+1}\right)^{1/\gamma}\right]^{\gamma}.$$

Show also that the optimal policies are $h_k(x) = A^{-1/\gamma} x$, for $k = N-1, \dots, 0$.

Considerando J_N como sigue

$$J_N(x) = \beta^N x^{1-\gamma} K_N,$$

con $K_N = 1$ bajo la hipótesis de que

$$c_k(x,a) = \beta^k a^{1-\gamma}$$

calculamos J_{N-1}.

$$\begin{aligned} J_{N-1}(x) &= \max_{a \in A(x)} \{c_{N-1}(x,a) + J_N((1+i)(x-a))\} \\ &= \max_{a \in A(x)} \left\{\beta^{N-1} a^{1-\gamma} + \beta^N ((1+i)(x-a))^{1-\gamma}\right\} \end{aligned}$$

Definimos el argumento como una función q.

$$\begin{aligned} q(x,a) &= \beta^{N-1} a^{1-\gamma} + \beta^N ((1+i)(x-a))^{1-\gamma} \\ &= C_1 a^{1-\gamma} + C_2 (x-a)^{1-\gamma}, \end{aligned}$$

donde $C_1 = \beta^{N-1}$ y $C_2 = \beta^N (1+i)^{1-\gamma} K_N$. Como q es continua en (x,a). Podemos calcular el máximo mediante el gradiente.

$$\partial_a q = C_1 (1-\gamma) a^{-\gamma} - C_2 (1-\gamma)(x-a)^{-\gamma}.$$

Igualando, $\partial_a q = 0$.

$$\begin{aligned} C_1 a^{-\gamma} &= C_2 (x-a)^{-\gamma} \\ \frac{C_1}{C_2} &= \left(\frac{x-a}{a}\right)^{-\gamma} \\ \left(\frac{C_1}{C_2}\right)^{-\frac{1}{\gamma}} &= \frac{x}{a} - 1 \\ \left(\frac{C_1}{C_2}\right)^{-\frac{1}{\gamma}} + 1 &= \frac{x}{a} \\ a &= \frac{x}{\left(\frac{C_1}{C_2}\right)^{-\frac{1}{\gamma}} + 1} \end{aligned}$$

Finalmente

$$a = h(x) = \frac{x}{(\beta(1+i)^{1-\gamma})^{\frac{1}{\gamma}} + 1}$$

Definiendo $\eta = (\beta(1+i)^{1-\gamma})^{\frac{1}{\gamma}} + 1$, $\eta - 1 = (\beta(1+i)^{1-\gamma})^{\frac{1}{\gamma}}$

entonces

$$h(x) = \frac{x}{\eta},$$

$$\begin{aligned}
J_{N-1}(x) &= \beta^{N-1}\left(\frac{x}{\eta}\right)^{1-\gamma} + \beta^{N}\left((1+i)\left(x - \frac{x}{\eta}\right)\right)^{1-\gamma} \\
&= \beta^{N-1}x^{1-\gamma}\left(\eta^{\gamma-1} + \beta\left(1+i\right)^{1-\gamma}\left(\frac{\eta-1}{\eta}\right)^{1-\gamma}\right) \\
&= \beta^{N-1}x^{1-\gamma}\eta^{\gamma-1}\left(1 + \beta\left(1+i\right)^{1-\gamma}\left(\eta-1\right)^{1-\gamma}\right) \\
&= \beta^{N-1}x^{1-\gamma}\eta^{\gamma-1}\left(1 + \beta\left(1+i\right)^{1-\gamma}\left(\eta-1\right)^{1-\gamma}\right) \\
&= \beta^{N-1}x^{1-\gamma}\eta^{\gamma-1}\left(1 + \beta\left(1+i\right)^{1-\gamma}\left((\beta(1+i)^{1-\gamma})^{\frac{1}{\gamma}}\right)^{1-\gamma}\right) \\
&= \beta^{N-1}x^{1-\gamma}\eta^{\gamma-1}\left(1 + \beta\left(1+i\right)^{1-\gamma}\left(\beta(1+i)^{1-\gamma}\right)^{\frac{1}{\gamma}-1}\right) \\
&= \beta^{N-1}x^{1-\gamma}\eta^{\gamma-1}\left(1 + \beta^{\frac{1}{\gamma}}(1+i)^{(1-\gamma)\left(\frac{1}{\gamma}-1\right)+1-\gamma}\right) \\
&= \beta^{N-1}x^{1-\gamma}\eta^{\gamma-1}\left(1 + \beta^{\frac{1}{\gamma}}(1+i)^{\left(\frac{1}{\gamma}-1\right)}\right) \\
&= \beta^{N-1}x^{1-\gamma}\eta^{\gamma},
\end{aligned}$$

Entonces

$$K_{N-1} = \eta^{\gamma}, h_{k-1}\left(x\right) = \frac{x}{\left(K_{N-1}\right)^{1/\gamma}}$$

Ahora calculamos J_{N-2}

$$\begin{aligned}
J_{N-2}\left(x\right) &= \max_{a \in A(x)}\left\{\beta^{N-2}a^{1-\gamma} + \beta^{N-1}\left[\left(1+i\right)\left(x-a\right)\right]^{1-\gamma}\eta^{\gamma}\right\} \\
&= \max_{a \in A(x)}\left\{q\left(x,a\right)\right\},
\end{aligned}$$

donde

$$q\left(x,a\right) = C_1 a^{1-\gamma} + C_2\left(x-a\right)^{1-\gamma},$$

con $C_1 = \beta^{N-2}$ y $C_2 = \beta^{N-1}(1+i)^{1-\gamma} K_{N-1}$. Obteniendo, por recursividad

$$\begin{aligned} h_{N-2} &= \frac{x}{\left(\frac{C_1}{C_2}\right)^{-\frac{1}{\gamma}} + 1} \\ &= \frac{x}{\left(\frac{1}{\beta(1+i)^{1-\gamma} K_{N-1}}\right)^{-\frac{1}{\gamma}} + 1} \\ &= \frac{x}{\left(\beta(1+i)^{1-\gamma} K_{N-1}\right)^{\frac{1}{\gamma}} + 1} \end{aligned}$$

Entonces, sea

$$\eta' = \left(\beta(1+i)^{1-\gamma} K_{N-1}\right)^{\frac{1}{\gamma}} + 1.$$

Repitiendo, el caso anterior, tenemos que

$$\begin{aligned} J_{N-2}(x) &= \beta^{N-2} x^{1-\gamma} \eta_{,}^{\gamma-1} \left(1 + K_{N-1}\beta(1+i)^{1-\gamma}\left((\beta(1+i)^{1-\gamma}K_{N-1})^{\frac{1}{\gamma}}\right)^{1-\gamma}\right) \\ &= \beta^{N-2} x^{1-\gamma} \eta_{,}^{\gamma-1} \left(1 + K_{N-1}\beta(1+i)^{1-\gamma}\left((\beta(1+i)^{1-\gamma}K_{N-1})^{\frac{1}{\gamma}}\right)^{1-\gamma}\right) \\ &= \beta^{N-2} x^{1-\gamma} \eta_{,}^{\gamma-1} \left(1 + K_{N-1}\beta(1+i)^{1-\gamma}\left(\beta(1+i)^{1-\gamma}K_{N-1}\right)^{\frac{1}{\gamma}-1}\right) \\ &= \beta^{N-2} x^{1-\gamma} \eta_{,}^{\gamma-1} \left(1 + K_{N-1}\beta(1+i)^{1-\gamma}(1+i)^{(1-\gamma)\left(\frac{1}{\gamma}-1\right)} K_{N-1}^{\frac{1}{\gamma}-1}\right) \\ &= \beta^{N-2} x^{1-\gamma} \eta_{,}^{\gamma-1} \left(1 + K_{N-1}\beta^{1/\gamma}(1+i)^{\frac{1}{\gamma}-1} K_{N-1}^{\frac{1}{\gamma}-1}\right) \\ &= \beta^{N-2} x^{1-\gamma} \eta_{,}^{\gamma-1} \left(1 + \beta^{1/\gamma}(1+i)^{\frac{1}{\gamma}-1} K_{N-1}^{\frac{1}{\gamma}}\right) \\ &= \beta^{N-2} x^{1-\gamma} \eta'^{\gamma}, \end{aligned}$$

entonces

$$K_{N-2} = \eta'^{\gamma},$$

y

$$h_{N-2} = \frac{x}{K_{N-2}^{1/\gamma}}$$

Por lo tanto, tenemos que

$$K_n = \left(\beta(1+i)^{1-\gamma} K_{n+1}\right)^{\frac{1}{\gamma}} + 1, n = 0, 1, 2, \ldots, N,$$

con $K_N = 1$.

Obteniendo así

$$J_n(x) = \beta^n x^{1-\gamma} K_n$$
$$h_n(x) = \frac{x}{K_n^{1/\gamma}}$$

Ejercicio 1.5. Consider now the infinite-horizon version of the above compsumption problem.

1. Write down the associated Bellman equation.
2. Argue why a solution to the Bellman equation should be the form

$$v(x) = cx^{1-\gamma},$$

where c is constant. Find the constant c and the stationary optimal policy.

Para el caso infinito. Estamos considerando

$$c(x, a) = a^{1-\gamma}$$

Entonces

$$\nu(x) = \max_{a \in A(x)} \left\{ a^{1-\gamma} + \beta \nu\left((1+i)(x-a) \right) \right\},$$

considerando $\nu(x) = cx^{1-\gamma}$. Entonces

$$\nu(x) = \max_{a \in A(x)} \left\{ a^{1-\gamma} + \beta c \left[(1+i)(x-a) \right]^{1-\gamma} \right\},$$

definimos

$$q(x, a) = a^{1-\gamma} + \beta c \left[(1+i)(x-a) \right]^{1-\gamma},$$

entonces

$$\partial_a q = (1-\gamma) a^{-\gamma} + \beta c (1-\gamma) (1+i)^{1-\gamma} (-1) (x-a)^{-\gamma}.$$

Si $\partial_a q = 0$. Entonces

$$\begin{aligned}
a^{-\gamma} &= \beta c (1+i)^{1-\gamma} (x-a)^{-\gamma} \\
\left(\beta c (1+i)^{1-\gamma} \right)^{-1} &= \left(\frac{x-a}{a} \right)^{-\gamma} \\
\beta^{-1} c^{-1} (1+i)^{\gamma-1} &= \left(\frac{x}{a} - 1 \right)^{-\gamma} \\
\left[\beta^{-1} c^{-1} (1+i)^{\gamma-1} \right]^{-\frac{1}{\gamma}} + 1 &= \frac{x}{a}
\end{aligned}$$

Por lo tanto

$$\begin{aligned} a &= \frac{x}{\left[\beta^{-1}c^{-1}\left(1+i\right)^{\gamma-1}\right]^{-\frac{1}{\gamma}}+1} \\ &= \frac{x}{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1} \end{aligned}$$

Ahora remplazamos en q

$$\begin{aligned} \nu\left(x\right) &= \left(\frac{x}{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1}\right)^{1-\gamma} + \beta c\left[\left(1+i\right)\left(x-\frac{x}{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1}\right)\right]^{1-\gamma} \\ &= x^{1-\gamma}\left(\frac{1}{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1}\right) + x^{1-\gamma}\left(1+i\right)^{1-\gamma}\beta c\left(1-\frac{1}{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1}\right)^{1-\gamma} \\ &= x^{1-\gamma}\left[\left(\frac{1}{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1}\right) + \left(1+i\right)^{1-\gamma}\beta c\left(1-\frac{1}{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1}\right)^{1-\gamma}\right]. \end{aligned}$$

Entonces

$$
\begin{aligned}
c &= \left(\frac{1}{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1}\right)^{1-\gamma}+\left(1+i\right)^{1-\gamma}\beta c\left(1-\frac{1}{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1}\right)^{1-\gamma}\\
&= \left(\frac{1}{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1}\right)^{1-\gamma}+\left(1+i\right)^{1-\gamma}\beta c\left(\frac{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}}{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1}\right)^{1-\gamma}\\
&= \left(\frac{1}{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1}\right)^{1-\gamma}\left(1+\left(1+i\right)^{1-\gamma}\beta c\left(\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}\right)^{1-\gamma}\right)\\
&= \left(\frac{1}{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1}\right)^{1-\gamma}\left(1+\left(1+i\right)^{1-\gamma}\beta c\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}-1}\right)\\
&= \left(\frac{1}{\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1}\right)^{1-\gamma}\left(1+\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}\right)\\
&= \left(\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1\right)^{\gamma-1}\left(1+\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}\right)\\
c &= \left(\left[\beta c\left(1+i\right)^{1-\gamma}\right]^{\frac{1}{\gamma}}+1\right)^{\gamma}
\end{aligned}
$$

Ahora, nos queda despejar c.

$$
\begin{aligned}
c^{\frac{1}{\gamma}} &= \beta^{\frac{1}{\gamma}}c^{\frac{1}{\gamma}}\left(1+i\right)^{\frac{1}{\gamma}-1}+1\\
1 &= \beta^{\frac{1}{\gamma}}\left(1+i\right)^{\frac{1}{\gamma}-1}+c^{-\frac{1}{\gamma}}\\
c^{-\frac{1}{\gamma}} &= 1-\beta^{\frac{1}{\gamma}}\left(1+i\right)^{\frac{1}{\gamma}-1}\\
c &= \left(1-\beta^{\frac{1}{\gamma}}\left(1+i\right)^{\frac{1}{\gamma}-1}\right)^{-\gamma}
\end{aligned}
$$

Ejercicio 1.6. Let $\{\xi_k\}$ be a dynamics of iid random variables such that $E\left[\xi\right]=0$ and $E\left[\xi^2\right]=d$. Consider the dynamics

$$x_{k+1}=x_k+a_k+\xi_k, k=0,1,2,\ldots,$$

and the discounted cost

$$E\left[\sum\beta^k\left(a_k^2+x_k^2\right)\right]$$

1. Write down the associated Bellman equation.
2. Conjecture that the solution to the Bellman equation takes the form $v(x) = ax^2 + b$, where a and b are constant.
3. Determine the constants a and b.
4. Conjecture that the solution to the Bellman equation takes the form $v(x) = ax^2 + b$, where a and b are constant. Determine the constants a and b.

$$\begin{aligned}\nu(x) &= \max_{a\in A(x)} \{c(x,a) + E[\nu(f(x,a))]\} \\ &= \max_{a\in A(x)} \left\{a^2 + x^2 + E[\nu(x+a+\xi)]\right\}\end{aligned}$$

Para $\nu(x) = ax^2 + b$

$$\begin{aligned}\nu(x) &= \max_{a\in A(x)} \{c(x,a) + \beta E[\nu(f(x,a))]\} \\ &= \max_{a\in A(x)} \left\{A^2 + x^2 + \beta\left(E\left[a\left(f^2(x,a)\right)\right] + b\right)\right\} \\ &= \max_{a\in A(x)} \left\{A^2 + x^2 + \beta\left(aE\left[f^2(x,a)\right] + b\right)\right\}\end{aligned}$$

Notemos que

$$\begin{aligned}E\left[f^2(x,A)\right] &= E\left[(x+A+\xi)^2\right] \\ &= E\left[x^2 + A^2 + \xi^2 + 2xA + 2x\xi + 2\xi A\right] \\ &= x^2 + A^2 + E\left[\xi^2\right] + 2xA + 2xE[\xi] + 2AE[\xi] \\ &= x^2 + A^2 + d + 2xA\end{aligned}$$

Entonces

$$\begin{aligned}ax^2 + b &= \max_{a\in A(x)} \left\{A^2 + x^2 + \beta\left[a\left(x^2 + A^2 + d + 2xA\right) + b\right]\right\} \\ &= \max_{a\in A(x)} \left\{A^2 + x^2 + \beta a\left(x^2 + A^2 + d + 2xA\right) + \beta b\right\} \\ &= \max_{a\in A(x)} \left\{A^2 + x^2 + \beta a x^2 + \beta a A^2 + \beta a d + 2\beta a x A + \beta b\right\} \\ &= \max_{a\in A(x)} \left\{A^2(\beta a + 1) + 2\beta a x A + x^2 + \beta a x^2 + \beta a d + \beta b\right\}\end{aligned}$$

Definimos

$$w(x,A) = A^2(\beta a + 1) + 2\beta a x A + x^2 + \beta a x^2 + \beta a d + \beta b,$$

entonces

$$\partial_A w = 2A\left(\beta a + 1\right) + 2\beta a x.$$

Si $\partial_A w = 0$, entonces

$$A = -\frac{\beta a x}{\beta a + 1}$$

Entonces

$$\begin{aligned}\nu\left(x\right) &= \left(\beta a x\right)^2 - 2\frac{\left(\beta a x\right)^2}{\beta a + 1} + x^2 + \beta a x^2 + \beta a d + \beta b \\ &= x^2\left(\left[\beta a\right]^2 - 2\frac{\left(\beta a\right)^2}{\beta a + 1} + 1 + \beta a\right) + \beta a d + \beta b\end{aligned}$$

Entonces

$$\begin{aligned}a &= \left[\beta a\right]^2 - 2\frac{\left(\beta a\right)^2}{\beta a + 1} + 1 + \beta a \\ b &= \beta a d + \beta b,\end{aligned}$$

de forma rapida

$$b = \frac{\beta a d}{1 - \beta},$$

entonces queda pendiente calcular a

$$\begin{aligned}a &= \left[\beta a\right]^2 - 2\frac{\left(\beta a\right)^2}{\beta a + 1} + 1 + \beta a. \\ 0 &= \left(\beta a\right)^2\left(1 - \frac{2}{\beta a + 1}\right) + 1 + \left(\beta - 1\right)a \\ &= \left(\beta a\right)^2\left(\beta a + 1 - 2\right) + \beta a + 1 + \left(a\beta - a\right)\left(\beta a + 1\right) \\ &= \left(\beta a\right)^2\left(\beta a - 1\right) + \beta a + 1 + \left[\left(a\beta\right)^2 + a\beta - \beta a^2 - a\right] \\ &= \left(\beta a\right)^3 + 2a\beta + 1 - \beta a^2 - a \\ &= \beta^3 a^3 - \beta a^2 + \left(2\beta - 1\right)a + 1\end{aligned}$$

Concluyendo que la constante b depende de a y a es una solución, dependiente de β, de la ecuación cúbica que

2 Tarea 2

3 Ejercicio Extra

3.1 Verificar la ecuación de Bellman para el problema de GridWorld de la casilla $s = (2, 2)$

Por la ecuación de Bellman.

$$\begin{aligned} v_\pi(s) &= E_\pi[G_t \mid S_t = s] \\ &= \sum_a \pi(a \mid s) \sum_{s',r} p\left(s', r \mid s, a\right)[r + \gamma v_\pi(s')] \end{aligned}$$

Queremos calcular $v_\pi(s_0)$ donde $s_0 = (2, 2)$ Considerando el $(0, 0)$ la esquina superior izquierda. Comenzaremos revisando $p\left(s', r \mid s, a\right)$. Notemos que en Gridword solo son posibles las recompensas $\{-1, 0, 5, 10\}$ según la posición actual y la acción a. Para nuestro caso, $s = s_0$

$$p(s', r \mid s_0, a) = 0, r \in \{-1, 5, 10\}, \forall a, \forall s'.$$

Por lo anterior la ecuación de Bellman queda como sigue

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} p\left(s', 0 \mid s_0, a\right)[\gamma v_\pi(s')].$$

Definamos la función auxiliar

$$g(a) = \sum_{s'} p\left(s', 0 \mid s_0, a\right)[\gamma v_\pi(s')].$$

Para $s = s_0$ y a fijo.

$$g(a) = \gamma v_\pi(s'),$$

donde s' satisface $\mathcal{P}(s' \mid s, a) = 1$. Para el ejercicio, $v_\pi(s')$ estan dadas para $s' = (1, 2), s' = (2, 1), s' = (2, 3), s' = (3, 2)$, y estamos suponiendo una distribución uniforme en π. Por lo tanto,

$$\pi(a \mid s) = \frac{1}{4}, \forall s.$$

Finalmente,

$$\begin{aligned} v_\pi(s) &= \frac{1}{4}\left(\gamma\left(2.3 + 0.7 + 0.4 - 0.4\right)\right) \\ &= \frac{3}{4}\gamma \approx 0.7 \end{aligned}$$

4 Proyecto: Manejo de Inventario

4.1 Introducción

Dentro del area del Control Estocástico, una de los problemas más conocidos son los problemas de inventario. Donde se presenta una bodega con capacidad máxima K. Cada estapa se extrae una cantidad de mercancía, la que denotaremos como la demanda D_t, y se solicita una cantidad del producto a_t, obteniendo finalmente el nivel de inventario X_t. En general se busca minimizar los costos de la bodega (costos por almacenamiento, costos por pérdida, entre otros).

4.2 Formulación del Proceso de Decisión de Markov.

Para nuestro problema consideraremos un supermercado, centrado en uno de sus pasillos. Suponiendo que en un pasillo se almacena un solo tipo de producto. Definiremos a K la cantidad máxima de producto en el pasillo, X_t a la cantidad del producto disponible para la venta (o la cantidad de producto en el pasillo). Nuestra demanda, o producto solicitado, será denotado por D_t y se considerará una colección de v.a i.i.d. Finalmente, la cantidad recolocada en el pasillo, o producto pedido, será denotada por a_t. Entonces, nuestro conjunto de estados $\mathcal{S}$ está dado por el siguiente conjunto

$$\mathcal{S} = \{s \in \mathbb{Z}^+ : 0 \leq s \leq K\}. \tag{4.1}$$

Nuestro conjunto de acciones $\mathcal{A} = \mathcal{S} = \mathbb{Z}^+$, y para $x \in \mathcal{S}$ nuestro conjunto de acciones admisibles esta dado por

$$\mathcal{A}(x) = \{a \in \mathcal{A} : 0 \leq a \leq K - x\}.$$

4.3 Dinámica del Modelo.

Recordando la fórmula para nuestro modelo.

$$X_{t+1} = f(X_t, a_t), f : \mathcal{S} \times \mathcal{A} \to \mathcal{S}.$$

Entonces el modelo que usaremos esta dado por

$$X_{t+1} = (X_t + a_t - \eta X_t - D_{t+1})^+, \tag{4.2}$$

donde a_t es la cantidad de producto recolodado al final del dia t, η es el factor descomposición, D_t es la demanda del prodcucto en la dia t y $(\cdot)^+ = \max\{\cdot, 0\}$.

4.4 Descripción y Justificación del Modelo.

El modelo Ecuación 4.2 pretende responder a la pregunta que denota el modelo ¿Cuánto producto tendré disponible al dia siguiente?. Lo anterior menciona que nuestras etapas $t \in \mathcal{T} = \{t \in \mathbb{Z}^+ : t \leq T, T \in \mathbb{N}\}$ representaran los dias dentro de un periodo T, t hace referencia al dia actual, y $t+1$ al dia siguiente. Entonces el modelo general esta dado por

$$X_{t+1} = (\text{Today} + \text{In}_t - \text{Out}_{t+1})^+.$$

Esto es, la parte positiva del producto que hay "hoy", es decir, X_t. A eso le agregaremos el producto que entrará hoy al final del dia, en nuestro modelo solo habrá ingreso de producto mediante solicitud (En este caso no consideramos un almacenimiento dentro del supermercado), entonces In_t esta dado por nuestras acciones $\text{In}_t = a_t$.

La parte que saldrá consta de dos elementos. En general consideramos la cantidad de producto que se compró en el dia t. Sin embargo, desconocemos la cantidad requerida, haciendo referencia al dia siguiente. Por lo tanto la demanda está representada por D_{t+1}, la cantidad de producto requerida al dia siguiente. En nuestro modelo también consideramos la salida de producto por considerarse producto no apto para la venta. Entonces

$$\text{Out}_t = D_{t+1} + N_t(X_t).$$

Bajo de la suposición que todos los productos poseen el mismo tiempo de vida con periodos de vida distintos supondremos que cada dia, al final, se retira un factor con respecto a la cantidad actual de producto.

$$N_t = \eta X_t$$

$$\text{Out}_t = D_{t+1} + \eta X_t$$

Finalmente, nos queda definir la función de costo, en nuestro modelo será la ganancia. Al considerar un periodo finito tenemos que la ganancia total G esta dada por

$$G(x_0, \pi) = \sum_{t=0}^{T} G_t(X_t, a_t), X_0 = x_0, X_{t+1} = f(X_t, a_t).$$

donde π es una politica, $\pi = (a_0, a_1, ..., a_{N-1})$. y G_t es la ganancia por etapa, en nuestro caso

$$G_t(x, a) = P_V \min\{x + a, D_t\} - P_S(a - \mathcal{I}_{t=0}x),$$

notemos que en el dia $a = 0$ y $D_0 = 0$, entonces $G_0(x, a) = -P_S x$ donde C es el costo unitario por tener el producto al inicio. Notemos que D_t es una variable aleatoria, entonces la función de valor por estado es la siguiente

$$V^{\pi}(s) = E[G(s, \pi)]$$

Teniendo que la ecuación de Bellman para nuestra función de valor es

$$V^{\pi}(s) = \sum_{a} \pi(a \mid s) \sum_{s'} \mathcal{P}[s' \mid s, a][R(s', a, s) + \gamma V^{\pi}(s')]$$

4.5 Justificación de las acciones.

Ya comentamos que nuestras acciones, será la cantidad de producto que vamos a solicitar. Entonces nuestras acciones serán números enteros y las acciones serán ejecutadas de forma instantea, al momento.

References

Ayers, G. 2005. «Air Pollution and Climate Change: Has Air Pollution Suppressed Rainfall over Australia?» 39 (2): 51-57. https://search.informit.org/doi/10.3316/informit.632702153657460.

B. Thomas Jr, George. 2010. *Cálculo varias variables*. Doceava. Addison-Wesley.