Gradiente Estocástico y Aproximación Estocástica Aplicados a *Q-learning*

Jose Sebastián Ñungo

Bogotá D.C

26 de agosto de 2020

Contenido

Procesos de Decisión de Markov

Un Proceso de Decisión de Markov (PDM):

1. S un espacio de estados.
2. A un conjunto de acciones.
3. $U(s) \subset A$ el conjunto de acciones admisibles dado un estado $s \in S$.
4. $Q(\,\cdot\,|s, a)$ una medida de probabilidad dado un estado $s \in S$ y $a \in U(s)$.
5. $c : S \times A \to \mathbb{R}$ una función de costo. Dado un estado $s \in S$ el costo de tomar una acción $a \in U(s)$. Usualmente denotamos $c_{s,a} = c(s, a)$ con $s \in S$ y $a \in U(s)$.

Procesos de Decisión de Markov

Nos interesa PDM finitos y en tiempo discreto, es decir,

- S y A son conjuntos finitos (Denotaremos los estados como $S = \{1, ..., n\}$)
- Los estados observados, las acciones tomadas, y los costos son ejecutados en tiempo discreto, i.e, $T = \mathbb{N}$.

Procesos de Decisión de Markov

Para cada tiempo $k \in T$ definimos,

- X_k la v.a. que representa el estado, y $x_k \in S$ el estado actual.
- A_k la v.a. que representa la acción, y $a_k \in U(x_k)$ la acción ejecutada.
- c_{x_k, a_k} es el costo de estar en elestado x_k y tomar la acción a_k.

Observe que,

$$\sum_{s \in S} Q(s|x_k, a_k) = 1.$$

Y, el proceso markoviano indica que $X_{k+1}|x_k, a_k \sim Q(\ \cdot\ |x_k, a_k)$.

Políticas

Sea,

$$\mathcal{M} = \{\mu : S \to A | \mu(i) \in U(i),\ i = 1, ..., n\}$$

Definimos una *política* $\pi = (\mu_k)_{k \in T}$ como una tupla de elementos en $\mathcal{M}$.

Una política es una regla con la que se decide que acción se debe tomar en cada instante de tiempo.

Políticas

Denotamos como $P(\mu_k)$ como la matriz de probabilidad correspondiente al control μ_k, esto es,

$$[P(\mu_k)]_{ij} = Q(\ j\ |i,\ \mu_k(i)) = P(X_{k+1} = j | X_k = i,\ \mu_k(i)),$$

para todo $i, j = 1, ..., n$. Además, dado un estado inicial i, la probabilidad de estar en el estado j en el tiempo k dada la política $\pi = (\mu_k)_{k \in T}$, es,

$$P_\pi(X_k = j | X_0 = i) = [P(\mu_{k-1})P(\mu_{k-2}) \cdots P(\mu_0)]_{ij}$$

Políticas - Estacionarias

Decimos que $\pi = (\mu_k)_{k \in T}$ una política *estacionaria* si

$$\mu = \mu_1 = \mu_2 = \mu_3 = \cdots$$

En este caso, dado un estado inicial i, la probabilidad de estar en el estado j en el tiempo k, dada la política π, es,

$$P_\pi(X_k = j | X_0 = i) = \left[P(\mu)^k \right]_{ij}.$$

Costo

Considere el la política $\pi = (\mu_k)_{k \in T}$ y el estado inicial $i \in S$, definimos el *costo esperado* hasta el tiempo N como,

$$\mathbb{E}_\pi \left[\sum_{k=0}^{N} \beta^k c_k | X_0 = i \right] = \sum_{k=0}^{N} \beta^k \sum_{j=1}^{n} P_\pi(X_k = j | X_0 = i) c_{j,\mu_k(j)}$$

con $\beta \in (0, 1]$ (el *factor de descuento*). El vector que describe el costo esperado hasta el tiempo N, en cualquier estado es

$$V^\pi(N) := \left(\mathbb{E}_\pi \left[\sum_{k=0}^{N} \beta^k c_k | X_0 = i \right] \right)_{i=1}^{n}$$

Objetivo

Definimos el *costo esperado descontado* como,

$$V^{\pi} = \lim_{N \to \infty} V^{\pi}(N),$$

y, el objetivo es encontrar una política π^* tal que,

$$V^{\pi^*} = \min_{\pi} V^{\pi} =: V^*.$$

Decimos que V^* es el *costo óptimo*.

Ecuaciones de óptimalidad

Definimos el *operador de programación dinámica* $L : \mathbb{R}^n \to \mathbb{R}^n$ de la forma,

$$L_i(V) = \min_{u \in U(i)} \left\{ c_{iu} + \beta \sum_{j=1}^{n} P_{ij}(u) V_j \right\} \quad \forall i \in \{1, ..., n\}$$

de forma vectorial,

$$L(V) = \min_{\mu \in \mathcal{M}} \{c(\mu) + \beta P(\mu) V\}, \quad c(\mu) = \begin{pmatrix} c_{1,\mu(1)} \\ \vdots \\ c_{n,\mu(n)} \end{pmatrix}$$

Ecuaciones de óptimalidad

Se puede demostrar que encontrar el costo óptimo es equivalente a encontrar $V^* \in \mathbb{R}^n$ tal que,

$$L(V^*) = V^*.$$

De hecho, la ecuación anterior se denomina *ecuaciones de optimalidad*.

Contracción - Punto Fijo de Banach

Teorema (Punto fijo de Banach)

Sea U un espacio de Banach y $\mathcal{L} : U \to U$ una contracción. Entonces

1. *Existe un único v^* en U tal que $\mathcal{L}(v^*) = v^*$; y*
2. *para $v_0 \in U$ arbitrario, la sucesión definida como,*

$$v_{n+1} = \mathcal{L}(v_{n+1}) = \mathcal{L}^{n+1}(v_0)$$

converge a v^*.

Contracción

Teorema

El operador L de programación dinámica con $0 \leq \beta < 1$ es una contracción en $\mathbb{R}^n$ con respecto a la norma $||\cdot||_\infty$.

Demostración.

Tenemos que $L(v) = (L_i(v))_{i=1}^n$. Entonces, $s \in S$. Sin restricción, $L_s(v) \geq L_s(u)$, y

$$a_s^* \in \arg \max_{a \in U(s)} \{c_{s,a} + \beta \sum_{j=1}^{n} P_{ij}(a) v_j\}$$

Entonces,

$$0 \leq L_s(v) - L_s(u) \leq \beta \sum_{j=1}^{n} P_{ij} \max_{j \in S}(v_j - u_j) = \beta ||v - u||_\infty$$

□

Solución

Teorema

Sea $v \in \mathbb{R}^n$. Se tienen las siguientes implicaciones,

a. *Si $v \leq L(v)$ entonces $v \leq V^*$.*
b. *Si $v \geq L(v)$ entonces $v \geq V^*$.*
c. *Si $v = L(v)$ entonces $v = V^*$.*

Gradiente estocástico

Considere la función objetivo $F : \mathbb{R}^n \to \mathbb{R}$ diferenciable y una función $g(w; \eta)$ que representa un estimador para el gradiente $\nabla F : \mathbb{R}^n \to \mathbb{R}^n$, donde η es una v.a.
La idea del método de descenso de gradiente estocástico (MDGE) es hacer una actualización de $w \in \mathbb{R}^n$ con el fin de solucionar el problema de minimización

$$F_* = \min_{w \in \mathbb{R}^n} F(w)$$

Así, lo que propone el método es que la aproximación del argumento que minimiza la función es,

$$w_{k+1} = w_k - \alpha_k g(w_k; \eta_k) \quad k \in \mathbb{N}$$

Donde $\alpha_k \in [0, 1]$ y $g(w_k; \eta_k)$ es el término que aproxima $\nabla F(w_k)$ con un cierto ruido η_k en la k-ésima actualización.

Gradiente estocástico - Condiciones suficientes

1. Gradiente Lipschitz. Existe $L > 0$ tal que

$$||\nabla F(w) - \nabla F(\bar{w})||_2 \leq L||w - \bar{w}||_2 \quad \forall w, \bar{w} \in \mathbb{R}^n.$$

2. η_{k-1} es $\mathcal{F}_k$-medibles. Y, $\mathbb{E}[g(w_k; \eta_k)|\mathcal{F}_k] = \nabla F(w_k)$
3. $\mathbb{E}\Big[||g(w_k; \eta_k)||_2^2|\mathcal{F}_k\Big] \leq M + M_G||\nabla F(w_k)||_2^2$ con $M \geq 0$ y $M_G \geq 1$.
4. Función fuertemente convexa. Existe c tal que,

$$F(\bar{w}) \geq F(w) + \nabla F(w)^T(\bar{w} - w) + \frac{1}{2}c||\bar{w} - w||_2^2 \quad \forall \bar{w}, w \in \mathbb{R}^d$$

5. $\sum_{k=1}^{\infty} \alpha_k = \infty, \quad \sum_{k=1}^{\infty} \alpha_k^2 \leq C < \infty$

Gradiente estocástico - Lemas

Lema

Considere el el supuesto 1. Entonces,

$$F(w) \leq F(\bar{w}) + \nabla F(\bar{w})^T (w - \bar{w}) + \frac{1}{2} L ||w - \bar{w}||_2^2 \quad \forall w, \bar{w} \in \mathbb{R}^n$$

Lema

Considere los supuestos 1, 2, y 3. Entonces,

$$\mathbb{E}[F(w_{k+1})|\mathcal{F}_k] - F(w_k) \leq -\left(1 - \frac{1}{2}\alpha_k L M_G\right)\alpha_k ||\nabla F(w_k)||_2^2 + \frac{1}{2}\alpha_k^2 L M.$$

Lema

Considere el supuesto 4. Entonces,

$$2c(F(w) - F_*) \leq ||\nabla F(w)||_2^2, \quad \forall w \in \mathbb{R}^n.$$

Gradiente estocástico - Teorema de Convergencia

Teorema

Considere los supuestos 1 al 5. Entonces, $\mathbb{E}[F(w_k) - F_*] \to 0$ *cuando* $k \to \infty$.

Demostración.

Sea $K \in \mathbb{N}$ tal que $\alpha_k \leq \frac{1}{LM_G}$, para todo $k \geq K$. De los lemas anteriores se tiene que,

$$\mathbb{E}[F(w_{k+1}) - F_*] - \mathbb{E}[F(w_k) - F_*] \leq -\alpha_k c \mathbb{E}[F(w_k) - F_*] + \frac{1}{2}\alpha_k^2 LM$$

Entonces,

$$\sum_{k=0}^{\infty} \alpha_k c \mathbb{E}[F(w_k) - F_*] \leq \frac{1}{2} CLM + \mathbb{E}[F(w_0) - F_*]$$

Suponga, hacia contradicción, que $\mathbb{E}[F(w_k) - F_*]$ no converge a 0. □

Ejemplo

$$\min_{w \in \mathbb{R}^n} F(w) = \frac{1}{2}||w||^2.$$

con un estimador de gradiente $g(w; \eta_k) = w + \eta_k$ donde η_k una v.a en cada tiempo k.
$\{\mathcal{F}_k\}$ una filtración crecientes de σ-álgebras tal que η_{k-1} es $\mathcal{F}_k$-medible para cada tiempo k. Además,

- $\mathbb{E}[\eta_k | \mathcal{F}_k] = 0$ y
- $\mathbb{E}\left[(\eta_k^i)^2 | \mathcal{F}_k\right] \leq M$ con $M \geq 0$

Aproximación Estocástica

El algoritmo de aproximación estocástica (AE) consiste en hacer una actualización con ruido de un vector $x \in \mathbb{R}^n$ con el propósito de solucionar una ecuación de la forma $F(x) = x$. Acá, $F : \mathbb{R}^n \to \mathbb{R}^n$ de la forma $F(x) = (F_1(x), ..., F_n(x))$, con mapeos $F_i : \mathbb{R}^n \to \mathbb{R}$ para todo $i = 1, ..., n$ para todo $x \in \mathbb{R}^n$.

$$x_{k+1}^i = x_k^i + \alpha_k \left(F_i(x_k) - x_k^i + \eta_k^i \right) \quad k \in T^i.$$

Con $T^i \subseteq T$ el conjunto de índices de tiempo donde se actualiza la componente x^i, $\alpha_k \in [0, 1]$ denominado como *el tamaño del paso* en el tiempo k y η_k^i es el término de ruido de actualización en el tiempo k para la i-ésima componente.

Aproximación Estocástica - Condiciones suficientes

1. Para todo $k_0 \in T^i$ existe $k \in T^i$ tal que $k > k_0$.
2. (a) x_0 es $\mathcal{F}_0-$medible. (b) $\forall k$ η_k es $\mathcal{F}_{k+1}-$medible.
   (c) $\forall i, k$ $\mathbb{E}\left[\eta_k^i | \mathcal{F}_k\right] = 0$.
   (d) $\exists A, B \in \mathbb{R}$, $\mathbb{E}\left[(\eta_k^i)^2 | \mathcal{F}_k\right] \leq A + B \max_{\tau \leq k} \max_j \left|x_\tau^j\right|^2$
3. $\sum_{k=1}^{\infty} \alpha_k = \infty, \quad \sum_{k=1}^{\infty} \alpha_k^2 \leq C < \infty$
4. $\exists x^* \in \mathbb{R}^n$, $\beta \in [0, 1)$ tal que $||F(x) - x^*||_\infty \leq \beta ||x - x^*||_\infty$ para todo $x \in \mathbb{R}^n$.

Aproximación Estocástica - Teorema de Convergencia

Teorema

Bajo el supuesto 1 al 4 se tiene que

a. *La sucesión* $\{x_k\}$ *está acotado, con probabilidad* 1.

b. x_k *converge a* x^* *cunado* $k \to \infty$, *con probabilidad* 1.

Q-learning

Volviendo a los Procesos de Decisión de Markov, sea $P = \{(i, a) | i \in S,\ a \in U(i)\}$ el conjunto de todos los pares estado-acción admisibles, y $|P| = m$. Considere el proceso de X_{k+1} generado por una política que garantiza el supuesto 1 de AE. Después de k iteraciones tenemos un vector $Q_k \in \mathbb{R}^m$, con componentes Q_k^{ia}, $(i, a) \in P$, que se actualiza aleatoriamente de acuerdo a la fórmula,

$$Q_{k+1}^{ia} = Q_k^{ia} + \alpha_k \left[c_{ia} + \beta \min_{v \in U(X_{k+1})} Q_k^{X_{k+1}, v} - Q_k^{ia} \right]$$

Q-learning - Como Modelo de AE

Consideramos $\{\mathcal{F}_k\}$ la filtración generada por el proceso (X_k, A_k), es decir, el proceso de estados y acciones, definiendo,

- $F_{ia}(Q) := c_{ia} + \beta \sum_{j \in S} P_{ij}(a) \min_{v \in U(j)} Q^{jv}$.
- $\eta_k^{ia} = \beta \min_{v \in U(X_{k+1})} Q_k^{X_{k+1}, v} - \beta \mathbb{E}\left[\min_{v \in U(X_{k+1})} Q_k^{X_{k+1}, v} | \mathcal{F}_k \right]$,

Tenemos que,

$$Q_{k+1}^{ia} = Q_k^{ia} + \alpha_k \left[F_{ia}(Q_k) - Q_k^{ia} + \eta_k^{ia} \right].$$

AE produce Q_* tal que $F(Q_*) = Q_*$. Y, definiendo $V_*^i = \min_{u \in U(i)} Q_*^{iu}$ se tiene el punto fijo del operador de programación dinámica.

Ejemplo

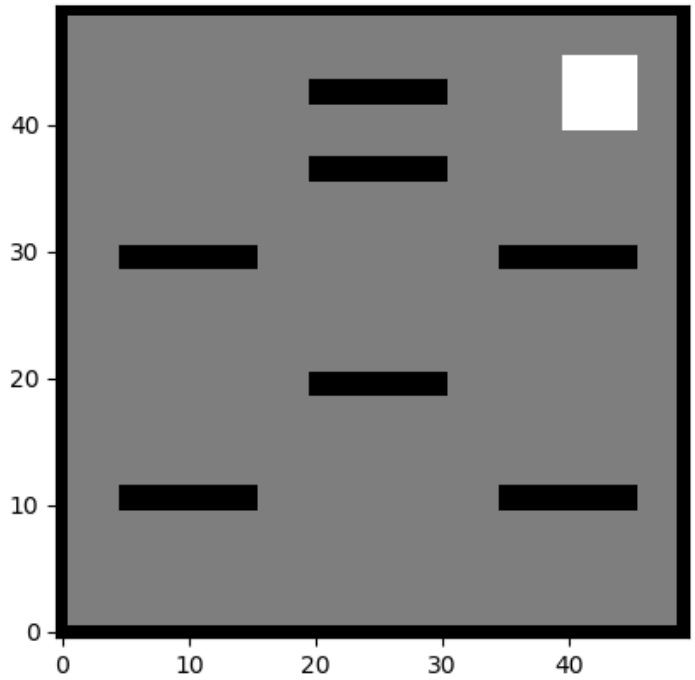

Figura 1: Retícula con los obstáculos. Los objetos negros (incluyendo el marco exterior) corresponden a los obstáculos, y el objeto blanco corresponde a la meta. Los estados que conforman los estados son absorbentes y todo otro estado es transitorio.

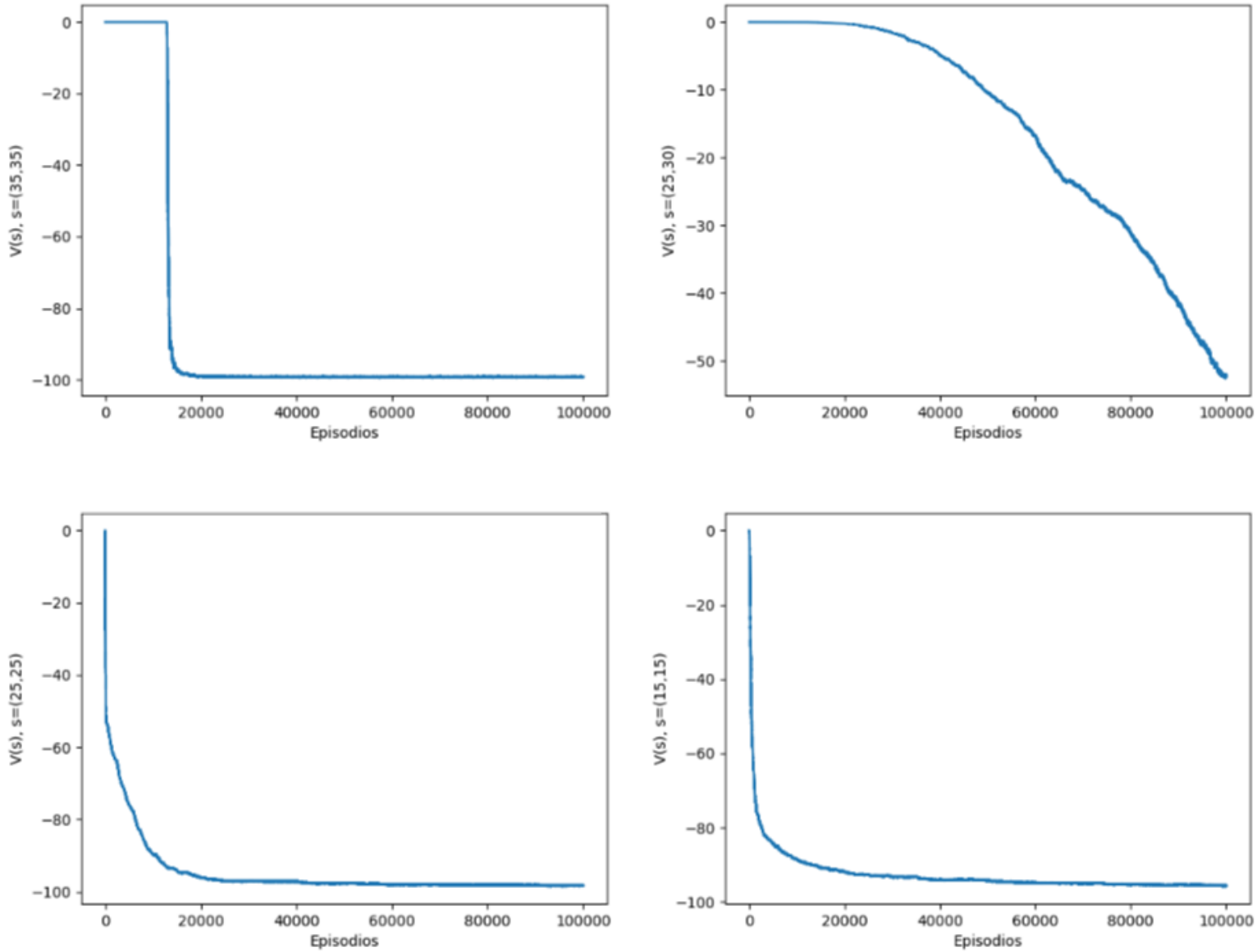

Figura 2: De derecha a izquierda se observa el costo esperado de estados iniciales cada vez más alejados de la meta. Es de mencionar que cada estado utiliza el vector Q de aprendizaje que dejó el estado anterior.

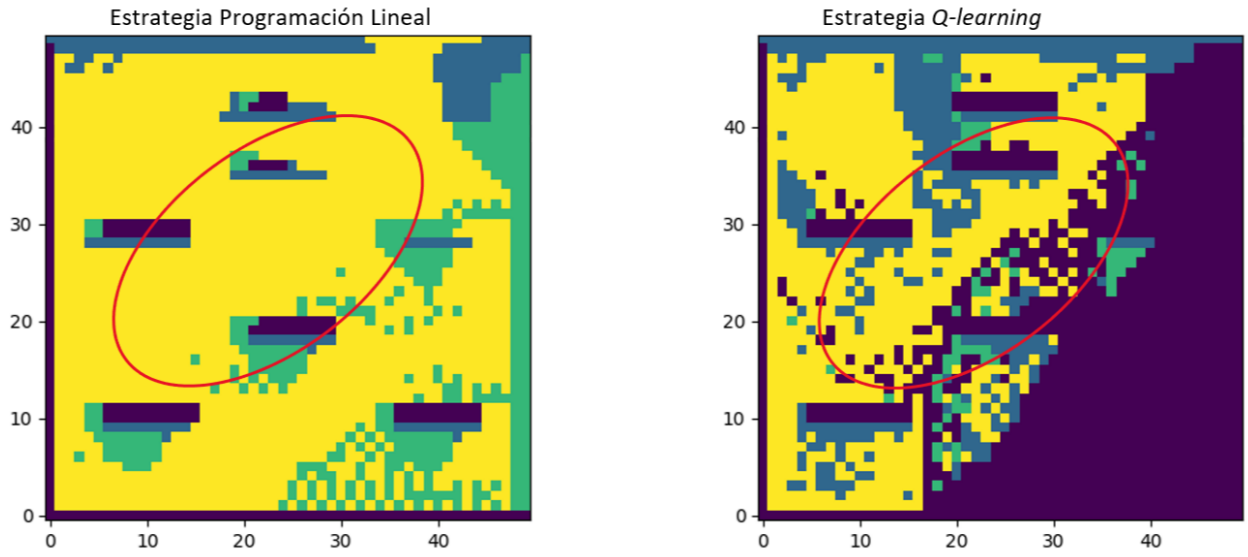

Figura 3: Para referenciar las políticas se ha hecho un mapa de calor que se identifica como: Subir con morado; bajar con azul; derecha con amarillo; verde con izquierda. El mapa de calor de la izquierda corresponde a la política encontrada por el método de Programación Lineal. Y, a la derecha la política encontrada por el algoritmo de *Q-learning*.

Morris

El Morris de tres hombres consta de un tablero como en la figura 4. Cada jugador dispone de tres fichas. Al inicio con el tablero vacío los jugadores toman turnos alternados poniendo sus fichas en las intersecciones. Una vez las seis fichas están en el tablero lo jugadores pueden mover una de sus fichas por las aristas, solo una posición por turno. El objetivo de cada jugador será acomodar sus tres fichas en linea.

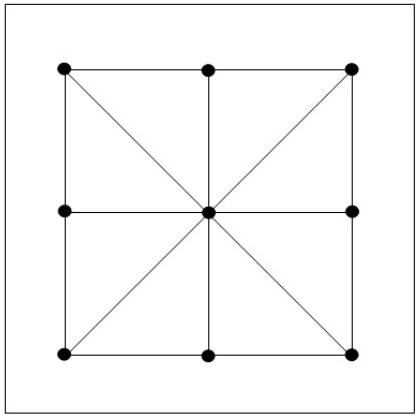

Figura 4: Tablero del Morris de Tres Hombres

Morris

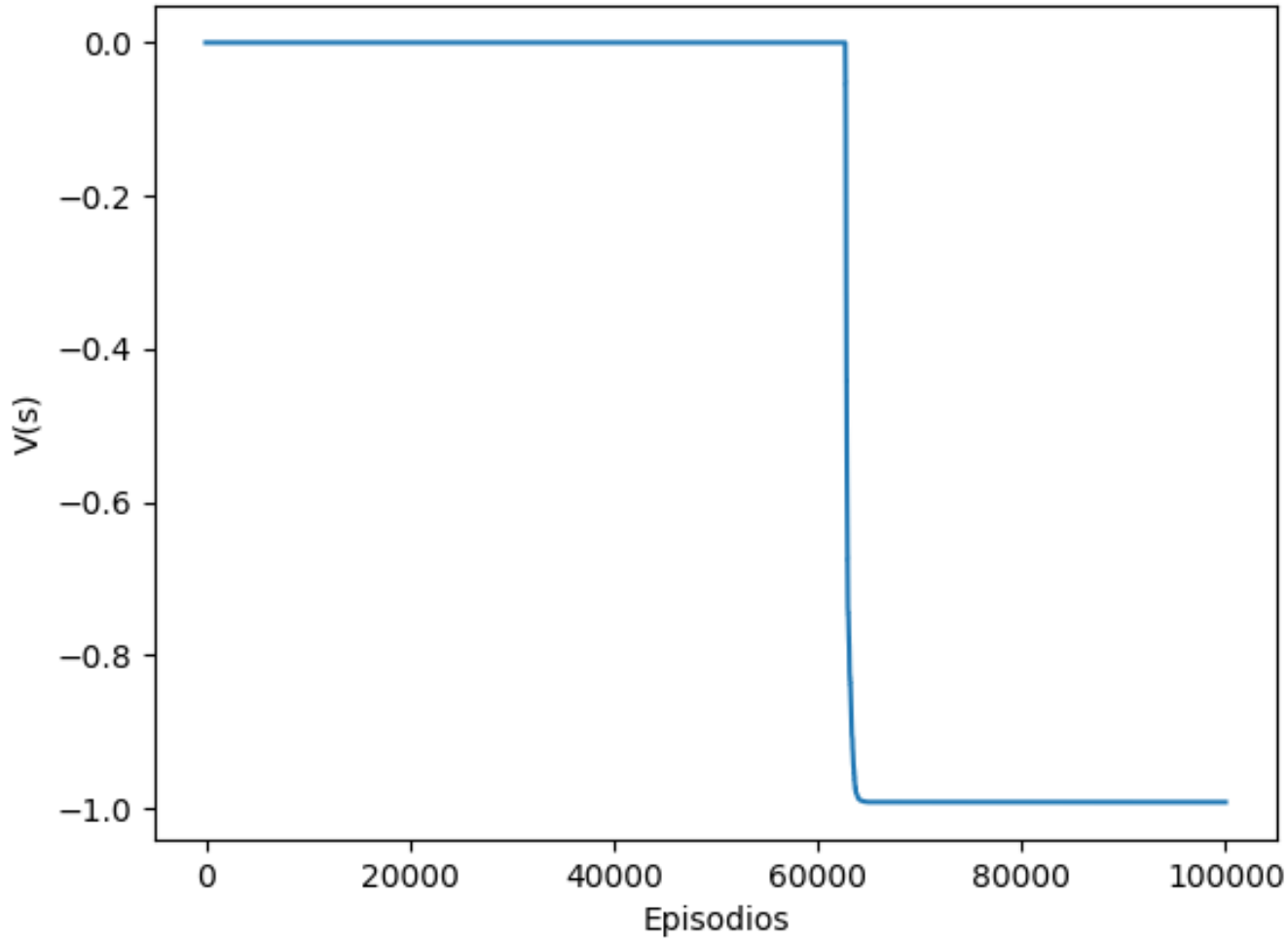

Figura 5: Función de valor para el estado para la primera jugada cuando el tablero está vacío.

Bibliografía


Dimitri P. Bertsekaz.
Dynamic Programming and Optimal Control, volume Volumen I.
Athena Scientific, 2005.


Dimitri P Bertsekas.
Reinforcement learning and optimal control.
Athena Scientific Belmont, MA, 2019.


Martin L. Puterman.
Markove Decision Processes.
Johon Wiley & Sons, Inc., 2005.


Richard S. Sutton and Andrew G. Barto.
Reinforcement Learning.
MIT Press, 2018.


John N Tsitsiklis.
Asynchronous stochastic approximation and q-learning.
Machine learning, 16(3):185–202, 1994.