Programación Dinámica

Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

Octubre de 2024

Contenido

Programación Dinámica: Introducción

- Es una estrategia para resolver el problema de RL cuando se conoce perfectamente el MDP.
- Motiva todos los demás métodos mucho más generales que permiten resolver el problema de RL aún cuando el MDP es conocido parcialmente o desconocido completamente.
- Esto métodos son basados en la observación de la interacción con el ambiente o en simulaciones.
- La base de la PD son las ecuaciones de Bellman.
- Decimos que es un método que hace *bootstraps* (autoreferencia o itera sobre si mismo): utiliza estimaciones de la función de interés (i.e., función valor o de política) para actualizar la misma función.

Contenido

Ecuación de Bellman: Función Valor del Estado

- La ecuación de Bellman de la función valor de estado es ($T = \infty$):

$$\begin{aligned} v_\pi(s) &= E_\pi[R_{t+1} + \gamma G_{t+1} \mid s] \\ &= \sum_a \pi(a \mid s)(E_\pi[R_{t+1} \mid s, a] + \gamma E_\pi[G_{t+1} \mid s, a]) \\ &= \sum_a \pi(a \mid s) \left(\sum_{s', r} p(s', r \mid s, a) r + \gamma E_\pi[G_{t+1} \mid s, a] \right) \end{aligned}$$

- Ahora, consideremos la version muestra de $E_\pi[G_{t+1} \mid s, a]$ para entender que variables entran en consideración.

Ecuación de Bellman: Función Valor del Estado

- $E_\pi[G_{t+1} \mid s, a] = E_\pi[R_{t+2} + \gamma R_{t+3} + ... \mid s, a]$.
- Luego G_{t+1} depende de $(S_{t+1}, A_{t+1}, S_{t+2}, A_{t+2}, ...)$.
- La versión muestral se obtiene simulando:

$$S_t = s, A_t = a \rightarrow S_{t+1} \sim p(\mid S_t, A_t) \rightarrow A_{t+1} \sim \pi(\mid S_{t+1}) \tag{1}$$

$$S_{t+2} \sim p(\mid S_{t+1}, A_{t+1}) \rightarrow A_{t+2} \sim \pi(\mid S_{t+2}) \tag{2}$$

$$... \tag{3}$$

luego, dado un $S_{t+1} = s'$ este determina toda la simulación (se puede ignorar s y a).

Ecuación de Bellman: Función Valor del Estado

- Se sigue que:

$$E_\pi[G_{t+1} \mid s, a] = \sum_{s'} p(s' \mid s, a) E_\pi[G_{t+1} \mid s, a, s'] \quad (4)$$

$$= \sum_{s'} p(s' \mid s, a) E_\pi[G_{t+1} \mid s'] \quad (5)$$

  donde la última igualdad es consecuencia de la propiedad de Markov o puede deducirse de la versón muestral.

- Ahora $p(s' \mid s, a)$ se obtiene de marginalizar $p(s', r \mid s, a)$.
- Recapitulando con lo anterior, finalmente obtenemos la ecuación de Bellman de la función valor del estado:

$$v_\pi(s) = \sum_a \pi(a \mid s) \left(\sum_{r,s'} p(s', r \mid s, a) \left(r + \gamma v_\pi(s') \right) \right)$$

Ecuación de Bellman: Función Valor del Estado

- En resumen la ecuación de Bellman de la función valor de estado (i.e., ecuación de Bellman de expectativas) es:

$$v_\pi(s) = \sum_a \pi(a \mid s) \left(\sum_{r,s'} p(s', r \mid s, a) \left(r + \gamma v_\pi(s') \right) \right)$$

- Dada una función de política y un MDP, la ecuación de Bellman define un sistema de n ecuaciones lineales (i.e., una por cada estado) y n incógnitas (i.e., $v_\pi(s)$).

Cómo estimar la función valor del estado?

- El problema de estimación de la función valor del estado para una política se conoce como *Evaluación de Política o Predicción* (i.e., estimar $v_\pi(s)$):
  1. Fuerza bruta: simular muchos episodios como se mencionó en la introducción (no usa Bellman).
  2. Resolver n-ecuaciones lineales. En la práctica es computacionalemnte muy complejo.
  3. Alternativamente, interpretar el problema como un problema de punto fijo e iterar la funcion de Bellman: Evaluación iterativa de la función de política (*Iterative Policy Evaluation*).

Ejemplo: Proceso Markoviano de Recompensas

- En vez de simular episodios y calcular el promedio de los retornos se puede resolver la ecuación de Bellman de la función valor del estado. Obsérvece que en este ejemplo no es necesario especificar un función de política, no hay acciones.

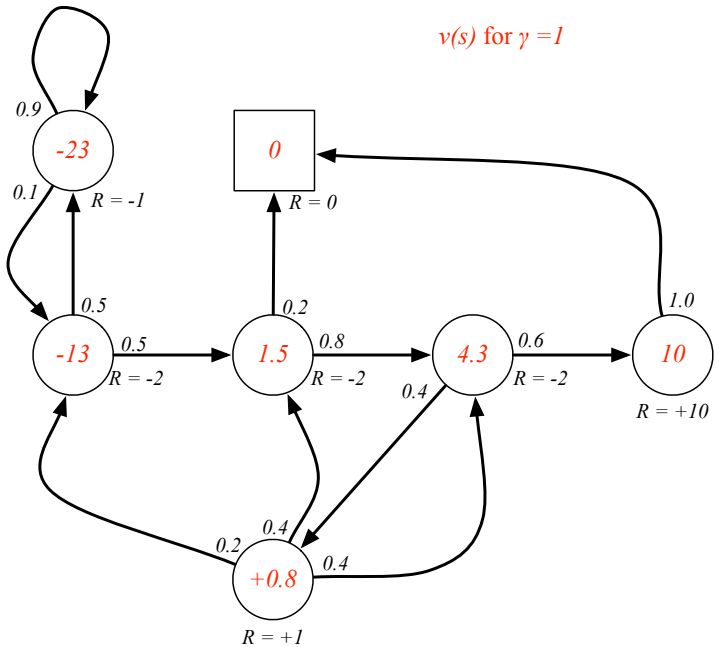


Figura: Tomado de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf). $v(s)$ se estima resolviendo las ecuaciones lineales de Bellman.

Ejemplo: Acciones sobre una grilla

- Considere la grilla de la figura. Las acciones son N,S,E,O. Una acción que implique salirse de la grilla, no cambia el estado y tiene recopensa -1.
- Cualquier otra acción tiene recopensa cero excepto, acciones que saque al agente de A o B. Estas tienen recompensas de 10 y 5 respectivamente y modifican el estado como se muestra en la figura.
- $\gamma = 0{,}9$

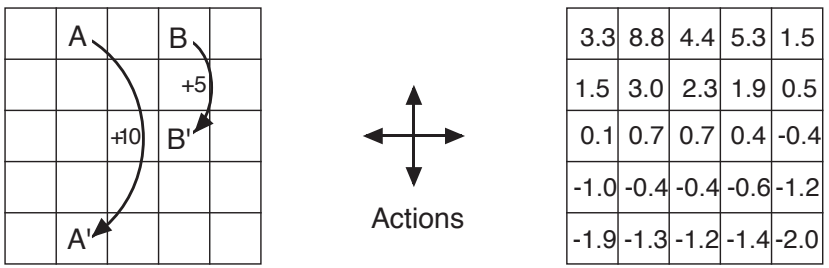


Figure 3.2: Gridworld example: exceptional reward dynamics (left) and state-value function for the equiprobable random policy (right).

Figura: Una acción que saque al agente de A o B además de la recompensa transporta al agente a A', B' respectivamente.

Ecuación de Bellman: Valor del Estado y la Acción

- Un argumento idéntico al utilizado para derivar la ecuacion de Bellman de la función valor nos lleva a:

$$q_\pi(s,a) = \left(\sum_r \left(\sum_{s'} p(s', r \mid s, a) \right) r + \gamma E_\pi[G_{t+1} \mid s, a] \right)$$

- Ahora, por la ecuación 5:

$$E_\pi[G_{t+1} \mid s, a] = \sum_{s'} p(s' \mid s, a) E_\pi[G_{t+1} \mid s'] \tag{6}$$

$$= \sum_{s'} p(s' \mid s, a) \sum_{a'} \pi(a' \mid s') E_\pi[G_{t+1} \mid s', a'] \tag{7}$$

$$= \sum_{s'} p(s' \mid s, a) \sum_{a'} \pi(a' \mid s') q(s', a') \tag{8}$$

Ecuación de Bellman: Valor del Estado y la Acción

- Despues de marginalizar $p(s', r, \mid s, a)$ se obetiene la ecuacion de Bellman de la función valor del estado y la acción:

$$q_\pi(s,a) = \sum_r \sum_{s'} p(s', r \mid s, a) \left(r + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s', a') \right) \tag{9}$$

- De nuevo este es un sistema lineal de ecuaciones en $q_\pi(s, a)$.

Contenido

Preliminares: Mejoramiento de la Política I

Proposición

Supongamos que para $\underline{s}, \underline{a}$ se cumple $q_\pi(\underline{s}, \underline{a}) \geq v_\pi(\underline{s})$, entones para todo s, $v_{\underline{\pi}}(s) \geq v_\pi(s)$, donde $\underline{\pi}$ se igual π para todo $s \neq \underline{s}$ excepto en $s = \underline{s}$ donde $\underline{\pi}(\underline{s}) = \underline{a}$.
Si la primera desigualdad es estricta la segunda también para $\underline{s}$.

$$
\begin{aligned}
v_\pi(\underline{s}) &\leq q_\pi(\underline{s}, \underline{a}) \\
&= E\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid \underline{s}, \underline{a}\right] \\
&= \sum_a \underline{\pi}(a|\underline{s}) E\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid \underline{s}, a\right] \\
&= E_{\underline{\pi}}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid \underline{s}\right] \\
&= E_{\underline{\pi}}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(S_{t+2}) \mid \underline{s}\right) \\
&= E_{\underline{\pi}}\left[G_t \mid \underline{s}\right] = v_{\underline{\pi}}(\underline{s})
\end{aligned}
$$

$$
\underline{\pi}(a|\underline{s}) = \begin{cases} 1, & a = \underline{a} \\ 0, & a \neq \underline{a} \end{cases}
$$

Preliminares: Mejoramiento de la Política I

- Otra forma equivalente de la proposición anterior es:
  Sean π y π' dos funciones de política tales que para todo s, $q_\pi(s, \pi'(s)) \geq v_\pi(s)$, entones para todo s, $v_{\pi'}(s) \geq v_\pi(s)$
  Si la primera desigualdad es estricta en algún estado, entonces la segunda también en ese estado.

Ecuación de Bellman: Función Valor Óptima

- La ecuación de Bellman para la función valor del estado se debe cumplir para $\pi = \pi_*$, luego:

$$v_*(s) = \sum_a \pi_*(a \mid s) \left(\sum_{r,s'} p(s', r \mid s, a) \left(r + \gamma v_*(s') \right) \right)$$

esto implica que:

$$v_*(s) \leq \max_a \left(\sum_{r,s'} p(s', r \mid s, a) \left(r + \gamma v_*(s') \right) \right)$$

Ecuación de Bellman: Función Valor Óptima

- La proposición anterior implica que **no puede ser cierto que**:

$$v_*(s) < \max_a \left(\sum_{r,s'} p(s', r \mid s, a) \left(r + \gamma v_*(s') \right) \right)$$

pues en ese caso existiría una política mejor que π_*.

- De esta forma obtenemos la la ecuación de Bellman de la función valor del estado óptima:

$$v_*(s) = \max_a \left(\sum_{r,s'} p(s', r \mid s, a) \left(r + \gamma v_*(s') \right) \right)$$

Ecuación de Bellman: Función Valor Óptima del Estado

- Ecuación de Bellman para la función valor del estado óptima:

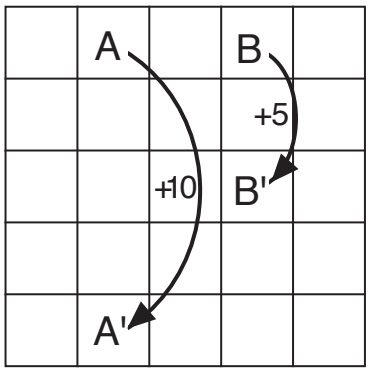
$$v_*(s) = \max_a \left(\sum_{r,s'} p(s', r \mid s, a) \left(r + \gamma v_*(s') \right) \right)$$

- Obérvese que define un sistema de ecuaciones no lineales.
- También puede interpretarse como el punto fijo de un operador en un espacio de funciones. Esto motiva la siguiente estrategia para solucionarla.
- Obsérvese que esta ecuación es equivalente a:

$$v_*(s) = \max_a q_*(s, a) \tag{10}$$

- Esto establece una forma para ir de $q_*(s, a) \rightarrow v_*(s)$.

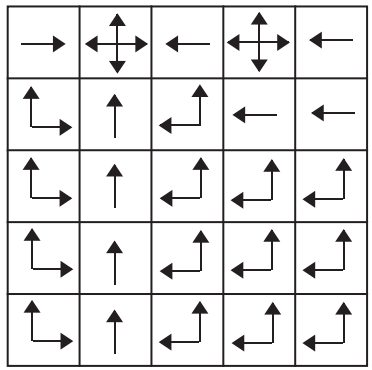
Ejemplo: Acciones sobre una grilla



Gridworld

22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

v_*



π_*

Figure 3.5: Optimal solutions to the gridworld example.

Ecuación de Bellman: Función Valor del Estado y la Acción Óptima

- Ecuación de Bellman para la función valor del estado y la acción óptima:

$$q_*(s,a) = \sum_{r,s'} p(s', r \mid s, a) \left(r + \gamma \max_{a'} q_*(s', a') \right)$$

- Obsérvese que esto es equivalente a:

$$q_*(s,a) = \sum_{r,s'} p(s', r \mid s, a) \left(r + \gamma v_*(s') \right)$$

- Esto establece una forma para ir de $v_*(s) \rightarrow q_*(s,a)$.

Contenido

Preliminares

- El problema de estimación de la función valor del estado para una política se conoce como *Evaluación de Política o Predicción* (i.e., estimar $v_\pi(s)$):
- Este problema es conceptualmente sencillo: resolver n-ecuaciones lineales. En la práctica es computacionalemnte muy complejo.
- Alternativamente interpretar el problema como un problema de punto fijo e iterar la ecuación de Bellman: Evaluación iterativa de la función de política (*Iterative Policy Evaluation*).

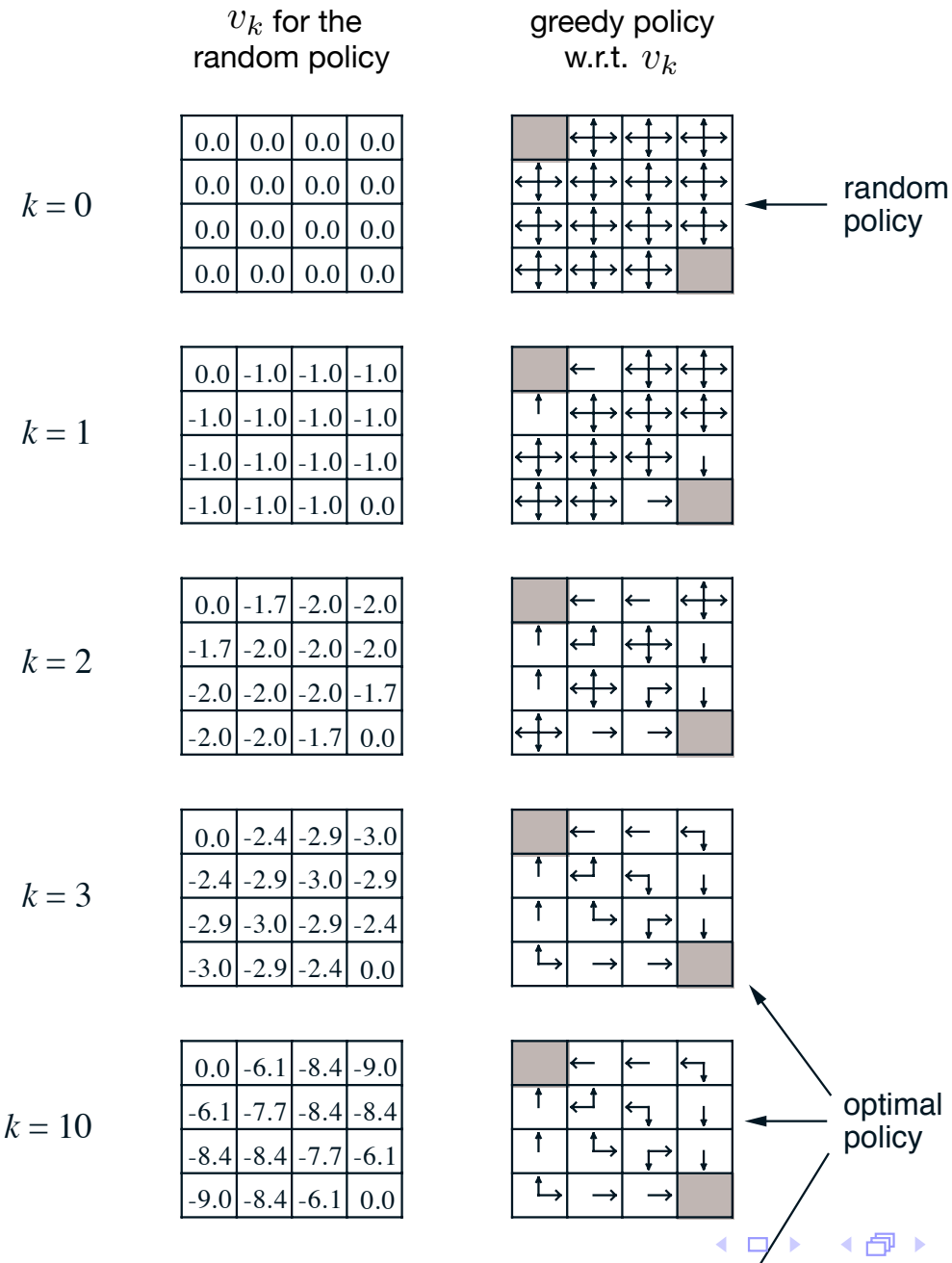
$$v_\pi^0 \rightarrow v_\pi^1 \rightarrow v_\pi^2 ... \rightarrow v_\pi \tag{11}$$

- Obsérvese que en principio se deben hacer muchas iteraciones.

Evalución o Predicción de la Política

- Considere el siguiente ejemplo del robot que debe alcanzar el estado terminal en una grilla. La columna izquierda muestra el resultado de iterar la ecuación de Bellman cuando la función de política es aleatoria uniforme.

Evaluación de Política o Predicción

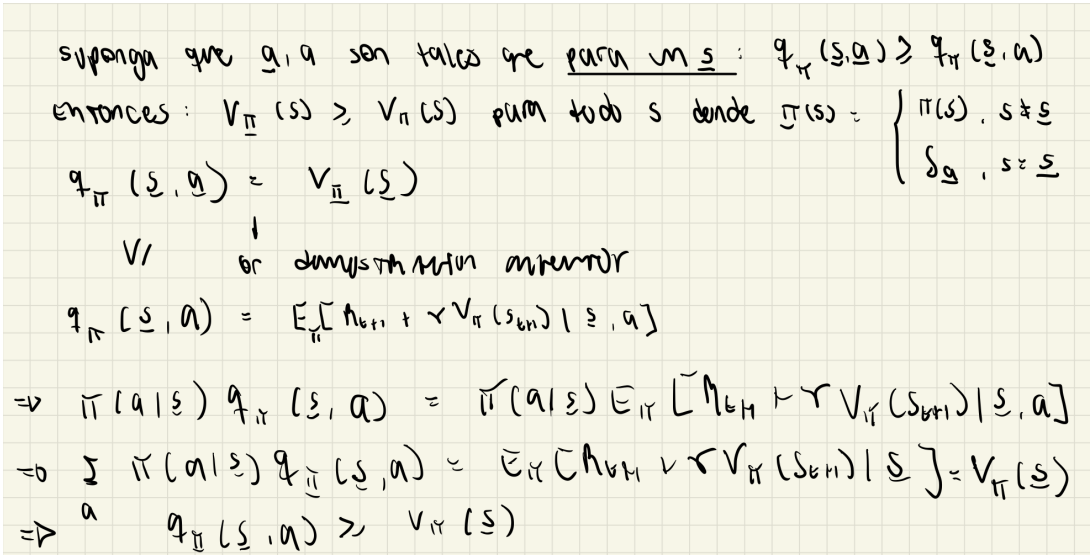


Contenido

Preliminares: Mejoramiento de la Política II

Proposición

Supongamos que para $\underline{s}, \underline{a}$ se cumple $q_\pi(\underline{s}, \underline{a}) \geq q_\pi(\underline{s}, a)$, entones para todo s, $v_{\underline{\pi}}(s) \geq v_\pi(s)$, donde $\underline{\pi}$ se igual π para todo $s \neq \underline{s}$ excepto en $s = \underline{s}$ donde $\underline{\pi}(\underline{s}) = \underline{a}$.
Si la primera desigualdad es estricta la segunda también para $\underline{s}$.



Preliminares: Mejoramiento de la Política II

- La política π' que mejora la política π de forma codiciosa:

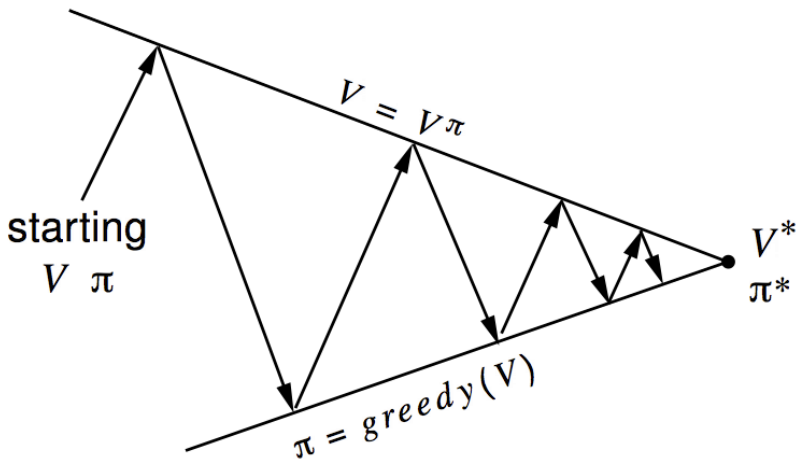
$$\pi'(s) = argmax_a q_\pi(s, a) \tag{12}$$

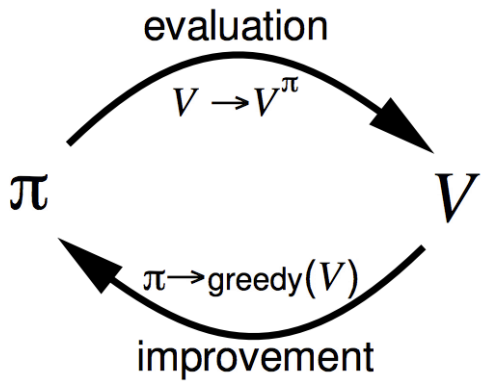
se llama política de mejoramiento.

Iteración de la Función de Política

- *Policy Iteration* consiste en Evaluar, Mejorar, Iterar varias veces:

$$\pi_0 \xrightarrow{\text{E}} v_{\pi_0} \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} v_{\pi_1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \cdots \xrightarrow{\text{I}} \pi_* \xrightarrow{\text{E}} v_*,$$





Policy evaluation Estimate v_π
Iterative policy evaluation

Policy improvement Generate $\pi' \geq \pi$

Iteración de la Función de Política: Ejemplo

- Arrendamiento de autos: ver siguiente diapositiva.

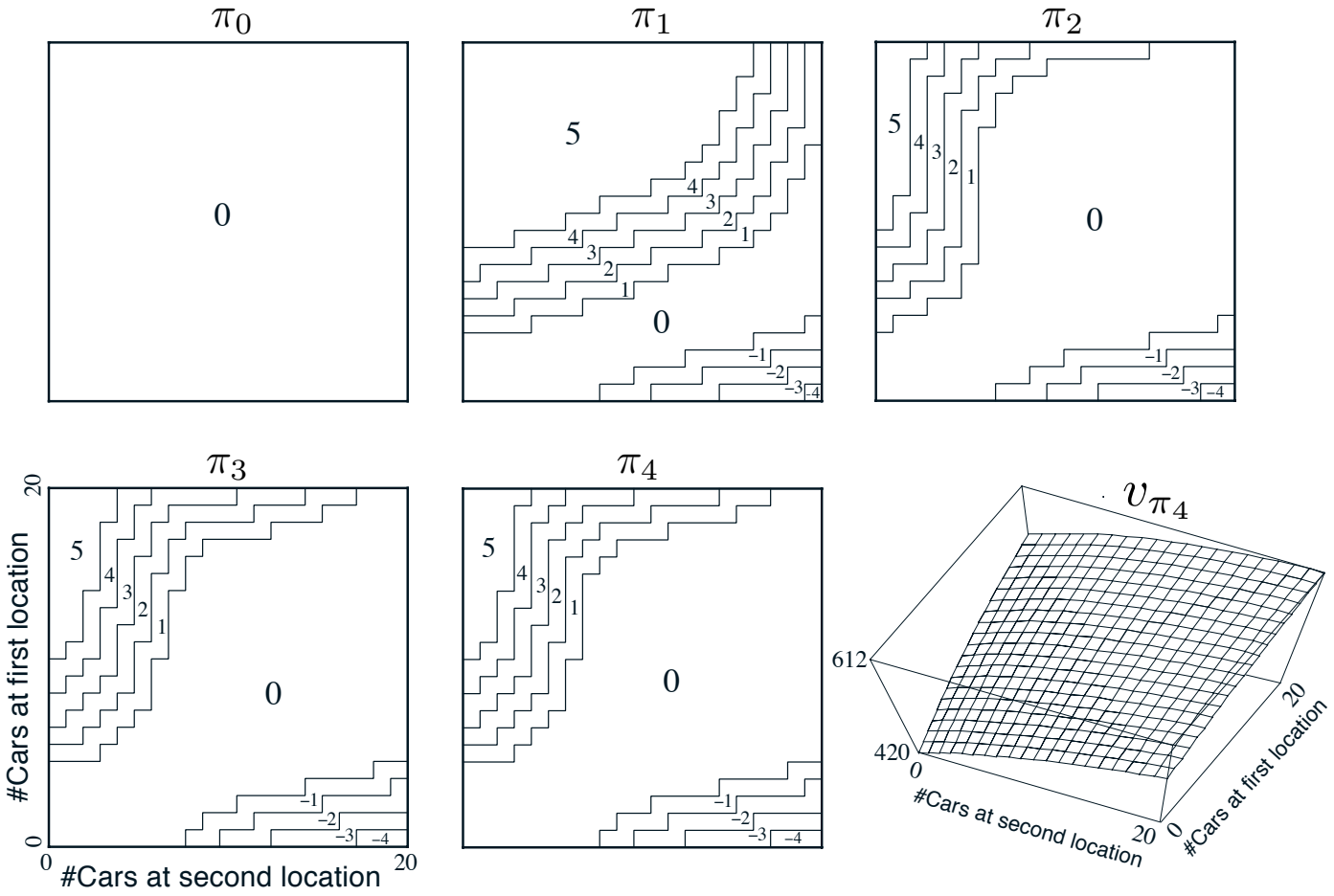


Figure 4.2: The sequence of policies found by policy iteration on Jack's car rental problem, and the final state-value function. The first five diagrams show, for each number of cars at each location at the end of the day, the number of cars to be moved from the first location to the second (negative numbers indicate transfers from the second location to the first). Each successive policy is a strict improvement over the previous policy, and the last policy is optimal. ■

Iteración de la Función de Política: Ejemplo

- Cada auto se arrienda por 10.
- Mover carros entres parqueaderos cuesta 2.
- La solicitud y retorno de carros es una variable aleatoria con distribucion de Poisson (conocida) $p(n) = \frac{\lambda^n}{n!} e^{-\lambda}$.
- Máximo se pueden estacionar 10 autos por cada parqueadero. Máximo mover cinco carros por dia. Los carros solo se pueden usar el día después.
- $\gamma = 0{,}9$.

Proceso Generalizado de Iteración de la Función de Política

- *Generalized Policy Iteration* consiste en Evaluar, Mejorar, Iterar de forma más libre siempre buscando mejorar la política que se tiene y aproximando más la función valor a la óptima (quizás con apenas algunas interaciones en la evaluación antes de hacer mejoramiento de la política).
- GPI refleja una estrategia de cooperación/competencia entre la función valor y la función de política. Cuando se esta evaluando la función valor, están cooperando para ser consitentes. Cuando se esta haciendo una mejora de la función política se está compitiendo.

Iteración de la Función Valor

- Evaluar (una sola iteración), Mejorar, Iterar.
- Es un caso particular del método de iteración de la función política: en el paso de evaluación de la política o predicción, iterar una sola vez.

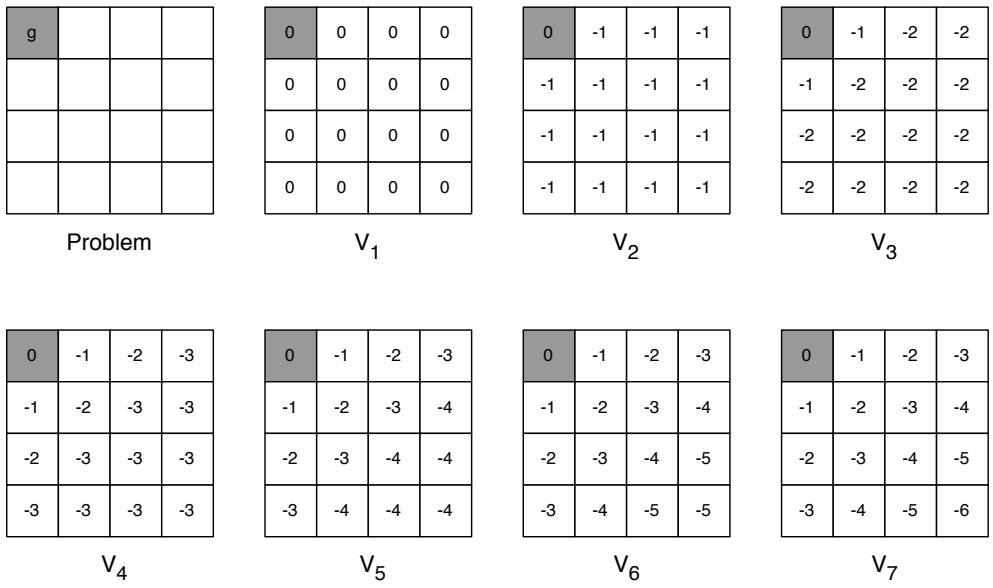


Figura: Tomado de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf)

Programación Dinámica Asincrónica

- Los métodos descritos hasta este punto requieren que se actualice la función valor en todos los estados.
- Tres formas de evitarlo: In-place DP, Prioritize sweeeping y DP en tiempo real.

In Place DP

- Se actualiza la función valor en un estado usando la ultima función valor disponible.

Prioritize sweeeping

- Se actualiza priorizando de acuerdo al estado en el cual el cambio en la función valor es mayor.

Programación Dinámica Resumen

Problem	Bellman Equation	Algorithm
Prediction	Bellman Expectation Equation	Iterative Policy Evaluation
Control	Bellman Expectation Equation + Greedy Policy Improvement	Policy Iteration
Control	Bellman Optimality Equation	Value Iteration

Figura: Tomado de David Silver: Lecture 2 (https://www.davidsilver.uk/wp-content/uploads/2020/03/MDP.pdf)

Contenido

Ejemplo: Modelo básico de crecimiento Brock - Mirman (1972) determinístico

- Sea $u(c_t) = \log(c_t)$, $f(k_t) = k_t^{\alpha}$ donde $\alpha \in (0,1)$, $\delta = 1$ y k_0 es dado.
- El problema de crecimiento óptimo es:

$$\begin{aligned} & \max_{\{c_t\}} \sum_{t=0}^{\infty} \beta^t \log(c_t) \\ s.a \quad : \quad & k_{t+1} = k_t^{\alpha} - c_t, \\ c_t, k_t \quad \geq \quad & 0. \end{aligned}$$

- El problema funcional asociado es:

$$\begin{aligned} v(k) \quad = \quad & \max_{c} \{\log(c) + \beta v(k^{\alpha} - c)\} \\ s.a \quad : \quad & 0 \leq c \leq k^{\alpha} \end{aligned}$$

Ejemplo: Modelo básico de crecimiento Brock - Mirman (1972) determinístico

- Para resolver este problema utilizaremos el método dado por el teorema del punto fijo para contracciones.
- Sea $v_0 = 0$, entonces:

$$\begin{aligned} v_1(k) &= \max_c \{\log(c)\} \\ s.a &: 0 \leq c \leq k^\alpha \end{aligned}$$

este problema tiene la solución de esquina $c = k^\alpha$, luego $v_1(k) = \alpha \log(k)$. Para calcular v_2 resolvemos el problema:

$$\begin{aligned} v_2(k) &= \max_c \{\log(c) + \beta\alpha \log(k^\alpha - c)\} \\ s.a &: 0 \leq c \leq k^\alpha \end{aligned}$$

que lo resuelve $c = \frac{k^\alpha}{1+\beta\alpha}$, luego

$$v_2(k) = \alpha(1 + \beta\alpha) \log(k) + \beta\alpha \log\left(\frac{\beta\alpha}{1 + \beta\alpha}\right) - \log(1 + \beta\alpha).$$

Ejemplo: Modelo básico de crecimiento Brock - Mirman (1972) determinístico

- Ahora, de igual forma podemos deducir que:

$$\begin{aligned} v_3(k) &= \alpha(1+\beta\alpha+\beta^2\alpha^2)\log(k) + \beta^2\alpha\log\left(\frac{\beta\alpha}{1+\beta\alpha}\right) + \\ &\quad (\beta\alpha+\beta^2\alpha^2)\log\left(\frac{\beta\alpha+\beta^2\alpha^2}{1+\beta\alpha+\beta^2\alpha^2}\right) \\ &\quad -\log(1+\beta\alpha+\beta^2\alpha^2) - \beta\log(1+\beta\alpha) \end{aligned}$$

- Luego en general vemos que $v_n(k) = A_n + \left(\alpha\sum_{i=0}^{n-1}(\beta\alpha)^i\right)\log(k)$, donde A_n es una constante que debemos determinar.

Ejemplo: Modelo básico de crecimiento Brock - Mirman (1972) determinístico

- Se sigue que $v(k) = A + \frac{\alpha}{1-\beta\alpha}\log(k)$, donde A es una constante que podemos encontrar simplemente observando que v debe satisfacer:

$$\begin{aligned} v(k) &= \max_c \{\log(c) + \beta v(k^\alpha - c)\} \\ s.a &: 0 \leq c \leq k^\alpha \end{aligned}$$

Es decir,

$$\begin{aligned} A + \frac{\alpha}{1-\beta\alpha}\log(k) &= \max_c \{\log(c) + \beta(A + \frac{\alpha}{1-\beta\alpha}\log(k^\alpha - c))\} \\ s.a &: 0 \leq c \leq k^\alpha \end{aligned}$$

- Con un poco de álgebra se puede mostrar que la solución a este problema es $c = (1-\beta\alpha)k^\alpha$ y $v(k) = \frac{1}{1-\beta}(\log(1-\beta\alpha) + \frac{\beta\alpha}{1-\beta\alpha}\log(\beta\alpha)) + \frac{\alpha}{1-\beta\alpha}\log(k)$.

Ejemplo: Modelo básico de crecimiento Brock - Mirman (1972) determinístico

- De esta forma tenemos que la dinámica de la variable de estado satisface: $k_{t+1} = \beta\alpha k_t^{\alpha}$
- La escogencia óptima para la variable de control, dada la variable de estado, es: $c_t = (1 - \beta\alpha)k_t^{\alpha}$. La función que expresa la escogencia óptima de las variables de control en términos de las variables de estado se llama *función de política*.
- Por tanto, en este caso, la función de política es $c_t = (1 - \beta\alpha)k_t^{a}$.

Ejemplo: Robot que debe alcanzar estado terminal en grilla

- 15 estados.

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	

Table 1: Small Gridworld: A simple example for MDP. In this Small Gridworld, each square with number on it is a non-terminal state. The two shaded squares are the terminal state. The goal is to make a robot move to the terminal state from any non-terminal state.

Ejemplo: Robot que debe alcanzar estado terminal en grilla

Función valor del estado ($\gamma = 1$, recompensa $= -1$ siempre hasta llegar a estado terminal, acciones = N,W,E,S; política aleatoria - si se sale del tablero no cambia el estado).

0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0

Table 2: State-Value Function for a random policy in Small Gridworld. We obtain the value functions for this Small Gridworld by solving linear systems of Bellman equations.

Ejemplo: Robot que debe alcanzar estado terminal en grilla

Función valor de la política aleatoria.

$k = 0$

0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0

$k = 1$

0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0

$k = 2$

0.0	-1.7	-2.0	-2.0
-1.7	-2.0	-2.0	-2.0
-2.0	-2.0	-2.0	-1.7
-2.0	-2.0	-1.7	0.0

$k = 3$

0.0	-2.4	-2.9	-3.0
-2.4	-2.9	-3.0	-2.9
-2.9	-3.0	-2.9	-2.4
-3.0	-2.9	-2.4	0.0

$k = 10$

0.0	-6.1	-8.4	-9.0
-6.1	-7.7	-8.4	-8.4
-8.4	-8.4	-7.7	-6.1
-9.0	-8.4	-6.1	0.0

$k = \infty$

0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0

Table 3: Iterative policy evaluation for Small Gridworld. We show the value functions we get at iterations $k = 0, 1, 2, 3, 10$ of policy evaluation, and the true value functions denoted by $k = \infty$. We could see that the value function are converging to the true value functions as k becomes larger.

Ejemplo: Robot que debe alcanzar estado terminal en grilla

Iteración de ecuación de optimalidad de Bellman.

$k = 0$

0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0

$k = 1$

0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0

$k = 2$

0.0	-1.0	-2.0	-2.0
-1.0	-2.0	-2.0	-2.0
-2.0	-2.0	-2.0	-1.0
-2.0	-2.0	-1.0	0.0

$k = 3$

0.0	-1.0	-2.0	-3.0
-1.0	-2.0	-3.0	-2.0
-2.0	-3.0	-2.0	-1.0
-3.0	-2.0	-1.0	0.0

Table 5: Value Iteration for Small Gridworld. We show iterations $k = 0, 1, 2, 3$ in value iteration. Three steps of value iterations give the optimal value function.

Ejemplo: Robot que debe alcanzar estado terminal en grilla

Función de política óptima

	W	W	W,S
N	E,N	E,S	S
N	N,E	S,E	S
N,E	E	E	

Table 4: Optimal Policy in Small Gridworld. We illustrate the optimal policy obtained by policy iteration in Small Gridworld. The robot will choose the greedy direction towards the terminal state.