Exercícios capítulo 3 do livro do Sutton, 2018

Aluno: Renato Scaroni

1. Exercise 3.8 - Suppose $\gamma = 0.5$ and the following sequence of rewards is received $R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3$, and $R_5 = 2$, with $T = 5$. What are $G_0, G_1, ..., G_5$? Hint: Work backwards.

   Pela definição de G_t, temos que:

$$G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{1}$$

   Como trata-se de uma task episódica, com 5 passos, temos:

$$G_t = \sum_{k=0}^{k+t+1=T} \gamma^k R_{t+k+1} \tag{2}$$

   Assim, pela equação (2), temos que:

$$G_4 = \gamma^0 R_{4+0+1} = R_5 = 2$$

   Portanto tomamos $G_5 = 0$ o que implica que $t < 5$ (2) pode ser escrita como:

$$G_t = R_{t+1} + \gamma G_{t+1}$$

   Assim:

$$G_3 = R_4 + \gamma G_4 = 3 + 0.5 * 2 = 4$$
$$G_2 = R_3 + \gamma G_3 = 6 + 0.5 * 4 = 8$$
$$G_1 = R_2 + \gamma G_2 = 2 + 0.5 * 8 = 6$$
$$G_0 = R_1 + \gamma G_1 = -1 + 0.5 * 6 = 2$$

   Portanto temos

$$\boxed{\begin{array}{l} G_0 = 2 \\ G_1 = 6 \\ G_2 = 8 \\ G_3 = 4 \\ G_4 = 2 \\ G_5 = 0 \end{array}}$$

2. Exercise 3.9 - Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$ followed by an infinite sequence of 7s. What are G_1 and G_0 ?

   Pela definição de G_t, temos que:

$$G_t \doteq \sum_{k=0}^{k=\infty} \gamma^k R_{t+k+1}$$

   Sabendo que $R_t = 7$ para $t \geq 1$, vamos supor que estivéssemos tratando um problema episódico, ou seja, onde G_t é uma soma finita:

$$G_1 = 7 + 7\gamma + 7\gamma^2 + ... + 7\gamma^k$$

   Agora tomando $G_1.\gamma$ e fazendo $G_1.\gamma - G_1$ temos:

$$G_1.(\gamma - 1) = (7\gamma + 7\gamma^2 + ... + 7\gamma^k + 7\gamma^{k+1}) - (7 + 7\gamma + 7\gamma^2 + ... + 7\gamma^k)$$

   Resultando em:

$$G_1 = \frac{\gamma^{k+1}.7 - 7}{\gamma - 1}$$

   Como estamos tratando com uma tarefa contínua e $0 \leq \gamma < 1$, basta tomarmos o limite da expressão acima com k tendendo a infinito:

$$G_1 = \lim_{k \to \infty} \frac{7.(\gamma^{k+1} - 1)}{\gamma - 1} = \frac{7}{1 - \gamma}$$

   Como sabemos que $\gamma = 0.9$, conseguimos calcular o resultado numericamente, o que nos resulta em:

$$G_1 = \frac{7}{1 - 0.9} = \frac{7}{0.1} \Rightarrow G_1 = 70$$

$$G_0 = R_1 + \gamma G_1 = 2 + 0.9 \times 70 \Rightarrow G_0 = 65$$

   Portanto temos:

$$\boxed{\begin{array}{c} G_0 = 65 \\ G_1 = 70 \end{array}}$$

3. Exercise 3.14 The Bellman equation (3.14) must hold for each state for the value function v_π shown in Figure 3.2 (right) of Example 3.5. Show numerically that this equation holds for the center state, valued at $+0.7$, with respect to its four neighboring states, valued at $+2.3$, $+0.4$, 0.4, and $+0.7$. (These numbers are accurate only to one decimal place.)

Sabemos pela definição de v_π, dado um conjunto de estados S de um MDP que:

$$v_\pi(s) \doteq \mathbb{E}[G_t|S_t = s] \Rightarrow$$

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')], \forall s \in S$$

Assim, utilizando como nome do estado sua posição no *gridworld*, por exemplo o estado central seria (2, 2), conseguimos calcular $v_\pi((2, 2))$ como:

$$v_\pi((2, 2)) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')], \forall s \in S$$

$$\begin{aligned} v_\pi((2,2)) &= 0.25 \times p((1,2),0|s,a)[0 + \gamma v_\pi((1,2))] \\ &+0.25 \times p((3,2),0|s,a)[0 + \gamma v_\pi((3,2))] \\ &+0.25 \times p((2,1),0|s,a)[0 + \gamma v_\pi((2,1))] \\ &+0.25 \times p((2,3),0|s,a)[0 + \gamma v_\pi((2,3))] \end{aligned}$$

$$\begin{aligned} v_\pi((2,2)) &= 0,25 \times 1 \times 0,9 \times 2,5 \\ &+0,25 \times 1 \times 0,9 \times (-0,4) \\ &+0,25 \times 1 \times 0,9 \times 0,7 \\ &+0,25 \times 1 \times 0,9 \times 0,4 \end{aligned}$$

$$v_\pi((2, 2)) = 0.5625 - 0.09 + 0,1575 + 0.09$$

De onde concluímos que:

$$\boxed{v_\pi((2, 2)) = 0.72}$$

4. Exercise 3.15 In the gridworld example, rewards are positive for goals, negative for running into the edge of the world, and zero the rest of the time. Are the signs of these rewards important, or only the intervals between them? Prove, using (3.8), that adding a constant c to all the rewards adds a constant, v_c , to the values of all states, and thus does not affect the relative values of any states under any policies. What is v_c in terms of c and γ?

   Por ora, pulemos a primera pergunta do enunciado e partamos direto para a segunda parte: provar que se todas as recompensas forem acrescidas de uma constante c, então v_π será acrescido de uma constante v_c. Para isso, vamos tomar as seguintes definições:

$$v_\pi \doteq \mathbb{E}[G_t|S_t = s]$$

   e

$$G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

   Vamos então tomar um G'_t como sendo uma função retorno de um MDP cujas recompensas sejam dadas pelo conjunto $\mathcal{R}'$ e $R'_i = R_i + c$, $\forall R'_i \in \mathcal{R}'$ e $\forall R_i \in \mathcal{R}$, onde i indica uma interação do agente com o ambiente, $\mathcal{R}$ é o conjunto de todas as recompensas possíveis de um outro MDP arbitrário. Assim:

$$G'_t = \sum_{k=0}^{\infty} (\gamma^k R_{t+k+1} + \gamma^k c)$$

$$= (R_{t+1} + c) + \sum_{k=1}^{\infty} \gamma^k R_{t+k+1} + \sum_{k=1}^{\infty} \gamma^k c$$

$$= R_{t+1} + \sum_{k=1}^{\infty} \gamma^k R_{t+k+1} + \sum_{k=1}^{\infty} \gamma^k c + c$$

$$= G_t + \sum_{k=1}^{\infty} \gamma^k c + c$$

   Como $0 \leq \gamma < 1$, a soma $\sum_{k=1}^{\infty} \gamma^k c$ converge e tem limite bem definido:

$$\lim_{k\to\infty} \sum \gamma^k c = \frac{\gamma c}{1-\gamma}$$

   Portanto:

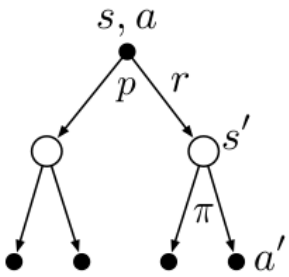
$$G'_t = G_t + v_c$$

   com

$$\boxed{v_c = \frac{\gamma c}{1-\gamma} + c}$$

   Por fim, voltemos a primeira pergunta. Uma vez que os valores de valor de estado se alteram de forma constante quando acrescemos uma constante c a cada recompensa individualmente em um problema de tarefa contínua, a política se manterá, portanto o sinal dos rewards importa, desde que a diferença entre eles se mantenha.

5. Exercise 3.17 What is the Bellman equation for action values, that is, for q_π ? It must give the action value $q_\pi(s, a)$ in terms of the action values, $q_\pi(s', a')$, of possible successors to the state–action pair (s, a). Hint: the backup diagram to the right corresponds to this equation. Show the sequence of equations analogous to (3.14), but for action values.



q_π backup diagram

Vamos analisar o diagrama acima, partindo de baixo para cima. Primeiro, nas subárvores da base, temos que em um estado s',o agente pode escolher dentre algumas opções de a' com probabilidade $\pi(a'|s')$. Portanto precisamos somar os valores esperados de cada par (s', a'), isto é, $q_\pi(s', a')$, ponderados pela probabilidade de se tomar essa ação. Como podemos lidar com casos de tarefas contínuas, multiplicamos cada parcela do somatório pelo fator de desconto γ.

$$\sum_{a'} \pi(a'|s').\gamma.q_\pi(s', a')$$

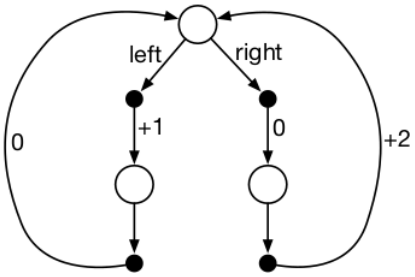
Agora que já temos o valor esperado a partir de s', em seguida somamos o valor r conseguido saindo de s com ação a e ponderamos o resultado pelo valor da probabilidade de se chegar em s', saindo de s com ação a e recebendo recompensa r.

$$p(r, s'|s, a) \left[r + \sum_{a'} \pi(a'|s').\gamma.q_\pi(s', a') \right]$$

Por fim, se somarmos todos os valores acima, tempos a esperança do valor de cada estado s' contado, pois eles serão ponderados pela probabilidade de se chegar em tal estado com recompensa r. Portanto a equação que queremos é:

$$\boxed{q_\pi(s, a) = \sum_{r,s'} p(r, s'|s, a) \left[r + \sum_{a'} \pi(a'|s').\gamma.q_\pi(s', a') \right]}$$

6. Exercise 3.22 Consider the continuing MDP shown below. The only decision to be made is that in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, π_{left} and π_{right}. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$?



Vamos chamar de s_1 o estado em que existe a bifurcação, s_2 o estado da esquerda e e s_3 o da direita. Assim podemos definir os valores de cada estado aplicado a cada política determinística que queremos avaliar seguindo a definição de valor de estado:

$$v_\pi(s) \doteq \mathbb{E}[G_t|S_t = s] \Rightarrow$$

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')], \forall s \in S$$

Vamos considerar primeiro a política π_{left}, que sempre decide pelo ramo da esquerda da figura apresentada. para facilitar a notação, escreveremos as funções valor de estado sob π_{left} como v_{left} e assim teremos:

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

$$v_{left}(s_1) = 1 \times p(s_2, 1|s_1, a)[1 + \gamma v_{left}(s_2)] + 0 \times p(s_3, 0|s_1, a)[0 + \gamma v_{left}(s_3)]$$
$$v_{left}(s_1) = 1 + \gamma v_{left}(s_2)$$

$$v_{left}(s_2) = 1 \times p(s_1, 0|s_2, a)[0 + \gamma v_{left}(s_1)]$$
$$v_{left}(s_2) = \gamma v_{left}(s_1)$$

$$v_{left}(s_3) = 1 \times p(s_3, 2|s_1, a)[2 + \gamma v_{left}(s_1)]$$
$$v_{left}(s_3) = 2 + \gamma v_{left}(s_1)$$

Como sabemos exatamente quais vão ser as recompensas partindo de cada estado, podemos estimar o valor estado s_1 através da função de retorno G_t, como definido no começo do capítulo, onde t é o período no tempo em que atingimos o estado s.

$$G_t \doteq \sum_{k=0}^{k=\infty} \gamma^k R_{t+k+1}$$

Partindo do estado s_1 sob política π_{left}, sabemos que a sequência de recompensas R_t será uma alternância entre 0s e 1s, ou seja:

$$G_t = 1 + \gamma + 0 + \gamma^2 + 0 + \gamma^4 + ...$$
$$G_t = \sum_{k=0}^{\infty} \gamma^{2k}$$

Isso representa uma progressão geométrica de termo inicial $a_1 = 1$ e razão γ^2. Assim, sabemos que a série acima converge pois $0 \leq \gamma < 1$ e seu valor é conhecido. Assim temos que:

$$G_t = \sum_{k=0}^{\infty} \gamma^{2k} = \frac{1}{1-\gamma^2} \Rightarrow$$
$$v_{left}(s_1) = \frac{1}{1-\gamma^2}$$

Analogamente podemos chegar que os valores de v_{right} serão:

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

$$v_{right}(s_1) = 0 \times p(s_2, 1|s_1, a)[1 + \gamma v_{right}(s_2)] + 1 \times p(s_3, 0|s_1, a)[0 + \gamma v_{right}(s_3)]$$
$$v_{right}(s_1) = \gamma v_{right}(s_3)$$

$$v_{right}(s_2) = 1 \times p(s_1, 0|s_2, a)[0 + \gamma v_{right}(s_1)]$$
$$v_{right}(s_2) = \gamma v_{right}(s_1)$$

$$v_{right}(s_3) = 1 \times p(s_3, 2|s_1, a)[2 + \gamma v_{right}(s_1)]$$
$$v_{right}(s_3) = 2 + \gamma v_{right}(s_1)$$

Com $v_{right}(s_1)$ podendo ser calculado da seguinte forma:

$$G_t = 0 + 2\gamma + 0 + 2\gamma^3 + 0 + 2\gamma^5 + ...$$
$$G_t = \sum_{k=0}^{\infty} 2\gamma^{2k+1} \Rightarrow$$
$$v_{right}(s_1) = \frac{2\gamma}{1-\gamma^2}$$

Assim teremos os seguintes valores de ação dependendo do valor de γ:

- $\gamma = 0$:

$$
\begin{aligned}
v_{left}(s_1) &= \frac{1}{1-\gamma^2} = 1 \\
v_{left}(s_2) &= \gamma v_{left}(s_1) = 0 \\
v_{left}(s_3) &= 2 + \gamma v_{left}(s_1) = 2 + 1 = 3
\end{aligned}
$$

$$
\begin{aligned}
v_{right}(s_1) &= \frac{2 \times 0}{1-\gamma^2} = \frac{0}{1-\gamma^2} = 0 \\
v_{right}(s_2) &= \gamma v_{right}(s_1) = 0 \\
v_{right}(s_3) &= 2 + \gamma v_{right}(s_1) = 2
\end{aligned}
$$

- $\gamma = 0.9$:

$$
\begin{aligned}
v_{left}(s_1) &= \frac{1}{1-(0.9)^2} = \frac{1}{0.19} = 5,263157895 \\
v_{left}(s_2) &= 0.9 \times v_{left}(s_1) = 4,736842105 \\
v_{left}(s_3) &= 2 + 0.9 \times v_{left}(s_1) = 2 + 4,736842105 = 6,736842105
\end{aligned}
$$

$$
\begin{aligned}
v_{right}(s_1) &= \frac{2 \times 0.9}{1-(0.9)^2} = \frac{1.8}{0.19} = 9,473684211 \\
v_{right}(s_2) &= 0.9 \times v_{right}(s_1) = 8,526315789 \\
v_{right}(s_3) &= 2 + 0.9 \times v_{right}(s_1) = 2 + 8,526315789 = 10,526315789
\end{aligned}
$$

- $\gamma = 0.5$:

$$
\begin{aligned}
v_{left}(s_1) &= \frac{1}{1-(0.5)^2} = \frac{1}{0.75} = 1.33 \\
v_{left}(s_2) &= 0.5 \times v_{left}(s_1) = 0,666666667 \\
v_{left}(s_3) &= 2 + 0.5 \times v_{left}(s_1) = 2 + 0,666666667 = 2,666666667
\end{aligned}
$$

$$
\begin{aligned}
v_{right}(s_1) &= \frac{2 \times 0.5}{1-(0.5)^2} = \frac{1}{0.75} = 1,33 \\
v_{right}(s_2) &= 0.5 \times v_{right}(s_1) = 0,666666667 \\
v_{right}(s_3) &= 2 + 0.5 \times v_{right}(s_1) = 2 + 0,666666667 = 2,666666667
\end{aligned}
$$

Assim, concluímos que quanto maior o γ, mais a política π_{right} é valorizada e quanto menor, mais a política π_{left} é valorizada.

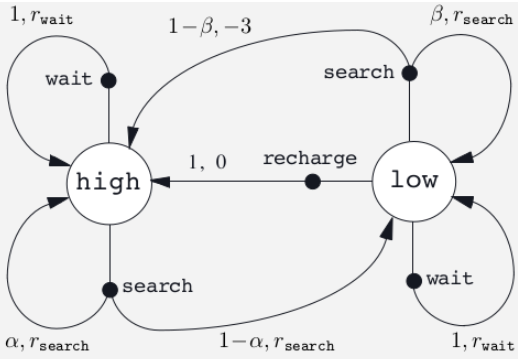
7. Exercise 3.23 - Give the Bellman equation for q_* for the recycling robot.

De acordo com a equação 3.20 na página 63 do livro do Sutton, temos que a equação de otimalidade de Bellman para q_* é:

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma \max_{a'} q_*(s',a')]$$

Vamos agora utilizar como base a tabela abaixo, que descreve o comportamento do MDP para o robô de reciclagem de lixo.

s	a	s'	$p(s' \mid s,a)$	$r(s,a,s')$
high	search	high	α	$r_{\texttt{search}}$
high	search	low	$1-\alpha$	$r_{\texttt{search}}$
low	search	high	$1-\beta$	-3
low	search	low	β	$r_{\texttt{search}}$
high	wait	high	1	$r_{\texttt{wait}}$
high	wait	low	0	-
low	wait	high	0	-
low	wait	low	1	$r_{\texttt{wait}}$
low	recharge	high	1	0
low	recharge	low	0	-



Analisando a tabela vemos que não é dada a probabilidade $p(s',r|s,a)$, porém como $p(s'|s,a) = \sum_r p(s',r|s,a)$ e cada tripla (s, a, s') é mapeada para um único r, então podemos tomar, para esse caso específico, $p(s',r|s,a) = p(s'|s,a)$.

Para facilitar a notação, vamos chamar high de h, search de s, low de l, wait de w e recharge de r. Assim vamos seguir analisando cada par (s, a) possível:

- (high, search):

$$\begin{aligned} q_*(h,s) &= p(h, r(h,s,h)|h,s)[r(h,s,h) + \gamma \max\{q_*(h,s), q_*(h,w)\}] \\ &+ p(l, r(h,s,l)|h,s)[r(h,s,l) + \gamma \max\{q_*(l,s), q_*(l,w), q_*(l,r)\}] \end{aligned}$$

$$\begin{aligned} q_*(h,s) &= \alpha[r_s + \gamma \max\{q_*(h,s), q_*(h,w)\}] \\ &+ (1-\alpha)[r_s + \gamma \max\{q_*(l,s), q_*(l,w), q_*(l,r)\}] \end{aligned}$$

- (low, search):

$$\begin{aligned} q_*(l,s) &= p(h, r(l,s,h)|l,s)[r(l,s,h) + \gamma \max\{q_*(h,s), q_*(h,w)\}] \\ &+ p(l, r(l,s,l)|l,s)[r(l,s,l) + \gamma \max\{q_*(l,s), q_*(l,w), q_*(l,r)\}] \end{aligned}$$

$$\begin{aligned} q_*(l,s) &= (1-\beta)[-3 + \gamma \max\{q_*(h,s), q_*(h,w)\}] \\ &+ \beta[r_s + \gamma \max\{q_*(l,s), q_*(l,w), q_*(l,r)\}] \end{aligned}$$

- (high, wait):

$$q_*(h,w) = p(h, r(h,w,h)|h,w)[r(h,w,h) + \gamma \max\{q_*(h,s), q_*(h,w)\}] \\ + p(l, r(h,w,l)|h,w)[r(h,w,l) + \gamma \max\{q_*(l,s), q_*(l,w), q_*(l,r)\}]$$

$$q_*(h,w) = r_w + [\gamma \max\{q_*(h,s), q_*(h,w)\}]$$

- (low, wait):

$$q_*(l,w) = p(h, r(l,w,h)|l,w)[r(l,w,h) + \gamma \max\{q_*(h,s), q_*(h,w)\}] \\ + p(l, r(l,w,l)|l,w)[r(l,w,l) + \gamma \max\{q_*(l,s), q_*(l,w), q_*(l,r)\}]$$

$$q_*(l,w) = r_w + [\gamma \max\{q_*(l,s), q_*(l,w), q_*(l,r)\}]$$

- (low, recharge):

$$q_*(l,r) = p(h, r(l,r,h)|l,r)[r(l,r,h) + \gamma \max\{q_*(h,s), q_*(h,w)\}] \\ + p(l, r(l,r,l)|l,r)[r(l,r,l) + \gamma \max\{q_*(l,s), q_*(l,w), q_*(l,r)\}]$$

$$q_*(l,r) = \gamma \max\{q_*(h,s), q_*(h,w)\}]$$

Assim resultamos nas seguintes equções de otimalidade de Bellman:

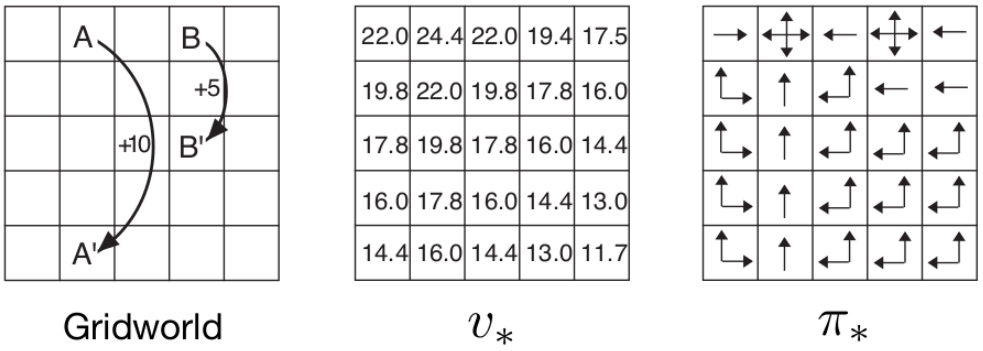
$$\begin{aligned}
q_*(h,s) &= \alpha[r_s + \gamma \max\{q_*(h,s), q_*(h,w)\}] \\
&\quad + (1-\alpha)[r_s + \gamma \max\{q_*(l,s), q_*(l,w), q_*(l,r)\}] \\
q_*(h,w) &= r_w + [\gamma \max\{q_*(h,s), q_*(h,w)\}] \\
q_*(l,s) &= (1-\beta)[-3 + \gamma \max\{q_*(h,s), q_*(h,w)\}] \\
&\quad + \beta[r_s + \gamma \max\{q_*(l,s), q_*(l,w), q_*(l,r)\}] \\
q_*(l,w) &= r_w + [\gamma \max\{q_*(l,s), q_*(l,w), q_*(l,r)\}] \\
q_*(l,r) &= \gamma \max\{q_*(h,s), q_*(h,w)\}
\end{aligned}$$

8. Exercise 3.24 Figure 3.5 gives the optimal value of the best state of the gridworld as 24.4, to one decimal place. Use your knowledge of the optimal policy and (3.8) to express this value symbolically, and then to compute it to three decimal places.

A equação (3.8) é a definição de função retorno, isto é:

$$G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Assim, vamos olhar a politica ótima a que o enunciado se refere:



Calculando então o valor do estado A no instante t a partir da equação (3.8), vemos que $R_{t+1} = 10$, assim, no estado A' só há uma ação possível, de recompensa 0, e o mesmo ocorre nas 4 próximas tomadas de ação, ou seja, $R_{t+1+1} = R_{t+1+2} = R_{t+1+3} = R_{t+1+4} = 0$, o que nos leva novamente ao estado A, e assim sucessivamente. Assim temos que:

$$G_t = \gamma^0 \times 10 + \gamma^5 \times 10 + \gamma^{10} \times 10 + ... = \sum_{k=0}^{\infty} \gamma^{5k} 10$$

Que como sabemos, para $0 \leq \gamma < 1$ converge para o valor:

$$G_t = \sum_{k=0}^{\infty} \gamma^{5k} 10 = \frac{10}{1-\gamma^5}$$

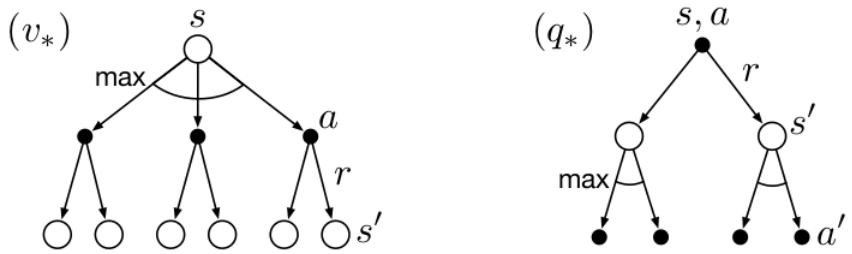
Isso nos permite calcular o valor de G_t do estado A em um tempo t arbitrário utilizando $\gamma = 0.9$, como definido na página 63 do livro do Sutton:

$$G_t = \sum_{k=0}^{\infty} \gamma^{5k} 10 = \frac{10}{1-(0.9)^5} = \frac{10}{1-(0.590)} = \frac{10}{0.41}$$

$$\boxed{G_t = 24.390}$$

9. Exercise 3.25 Give an equation for v_* in terms of q_*

Olhando os diagramas de *backup* para q_* e v_*, temos que v_* escolhe a melhor ação baseado no valor esperado de recompensa dos estados subsequentes, enquanto q_* escolhe, dado que a ação a foi tomada no estado s, a ação a' de maior valor a ser tomada em cada estado s' subsequente possível.



Assim, o valor de cada estado passa a ser o valor da ação ótima de cada estado, de forma que, se temos a equação de otimalidade de Bellman para q_* de acordo com a equação 3.20 na página 63 do livro do Sutton sendo:

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma \max_{a'} q_*(s',a')]$$

E a equação que representa v_* é:

$$v_*(s) = \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma v_*(s')]$$

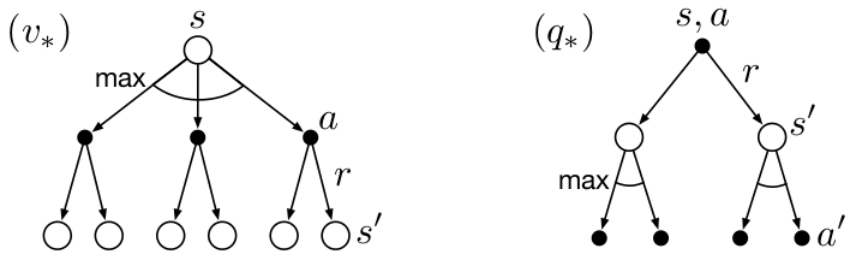
Então, escrevendo as duas equações juntas tempos que:

$$v_*(s) = \max_a q_*(s,a)$$

$$\boxed{v_*(s) = \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma \max_{a'} q_*(s',a')]}$$

10. Exercise 3.26 Give an equation for q_* in terms of v_* and the four-argument p.

Análogo ao exercício anterior, sabemos, pelo diagrama de *backup*, que o que v_* faz é tomar a que maximize q_*.



Portanto, dado que:

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma \max_{a'} q_*(s',a')]$$

Temos que:

$$\boxed{q_*(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma v_*(s')]}$$

11. Exercise 3.27 Give an equation for π_* in terms of q_* .

Por definição, $\pi_*(a|s)$ indica a probabilidade de se tomar a ação a no estado s sob a política ótima π_*, isto é, se $\pi_*(a|s) > 0$ então $q_*(s,a) \geq q_*(s,a')$, $\forall a' \in \mathcal{A}(s)$, $a' \neq a$, onde $\mathcal{A}(s)$ é o conjunto de ações possíveis no estado s.

Assim, seja $\mathcal{A}_{max}(s) = \{a|\, \text{argmax}_a\, q_*(s,a)\}$ o conjunto de todos as ações possíveis em estado s cujo valor é máximo, e sabendo que sob política π_* sempre será escolhida uma ação de valor ótimo, qualquer distribuição de probabilidade que satisfaça a soma $\sum_{a \in \mathcal{A}_{max}(s)} \pi_*(a|s) = 1$ é uma distribuição possível para π_*. portanto, denotando a cardinalidade de um conjunto X como sendo $|X|$, podemos escrever essa probabilidade como sendo:

$$\boxed{\pi_*(a|s) = \begin{cases} \dfrac{1}{|\mathcal{A}_{max}(s)|}, a \in \mathcal{A}_{max}(s) \\ 0, a \notin \mathcal{A}_{max}(s) \end{cases}}$$

12. Exercise 3.28 Give an equation for π_* in terms of v_* and the four-argument p.

Analogamente à ideia do exercício anterior, sabemos pela definição que $\pi_*(a|s)$ é a probabilidade de escolher a ação ótima para o estado s, com a diferença que neste caso o que se quer maximizar é o valor do estado. Assim, tomando a equação:

$$v_*(s) = \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma v_*(s')]$$

Essa equação é apenas uma forma mais conveniente de escrever

$$v_*(s) = \sum_a \pi_*(a|s) \sum_{s',r} p(s', r|s, a)[r+\gamma v_*(s')],\ a \in argmax_a \sum_{s',r} p(s', r|s, a)[r+\gamma v_*(s')]$$

Sendo $A' = argmax_a \sum_{s',r} p(s', r|s, a)[r + \gamma v_*(s')]$ o conjunto de todos as ações possíveis no estado s que maximizam o valor da expressão dada, temos que:

$$\boxed{\pi_*(a|s) = \begin{cases} \dfrac{1}{|A'|}, a \in A' \\ 0, a \notin A' \end{cases}}$$

13. Exercise 3.29 Rewrite the four Bellman equations for the four value functions (v_π , v_* , q_π , and q_*) in terms of the three argument function p (3.4) and the two-argument function r (3.5).

Segundo descrito na página 49 do livro temos que:

$$p(s'|s, a) \doteq \sum_{r \in \mathcal{R}} p(s', r|s, a)$$
$$r(s, a) \doteq \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r|s, a)$$

Assim, vamos partir da definição de v_π em função de $p(s', r|s, a)$ e $r(s, a, s')$:

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')]$$
$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)r + \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\gamma v_\pi(s')$$
$$v_\pi(s) = \sum_a \pi(a|s) \sum_r \sum_{s'} p(s', r|s, a)r + \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)\gamma v_\pi(s')$$
$$v_\pi(s) = \sum_a \pi(a|s) \sum_r r \sum_{s'} p(s', r|s, a) + \sum_a \pi(a|s) \sum_{s'} \gamma v_\pi(s') \sum_r p(s', r|s, a)$$
$$v_\pi(s) = \sum_a \pi(a|s) r(s, a) + \sum_a \pi(a|s) \sum_{s'} \gamma v_\pi(s') p(s'|s, a)$$
$$v_\pi(s) = \sum_a \pi(a|s)[r(s, a) + \sum_{s'} \gamma v_\pi(s') p(s'|s, a)]$$

Pela definição de v_* temos:

$$v_*(s) = \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')]$$

Então podemos dizer que v_* em função de $p(s'|s, a)$ e $r(s, a)$ é:

$$v_*(s) = \max_a [r(s, a) + \sum_{s'} \gamma v_\pi(s') p(s'|s, a)]$$

Quanto à equação de valor da ação, temos:

$$q_\pi(s, a) = \sum_{r,s'} p(r, s'|s, a)[r + \sum_{a'} \pi(a'|s').\gamma.q_\pi(s', a')]$$
$$q_\pi(s, a) = \sum_{r,s'} p(r, s'|s, a)r + \sum_{r,s'} p(r, s'|s, a) \sum_{a'} \pi(a'|s').\gamma.q_\pi(s', a')$$
$$q_\pi(s, a) = \sum_r \sum_{s'} p(r, s'|s, a)r + \sum_r \sum_{s'} p(r, s'|s, a) \sum_{a'} \pi(a'|s').\gamma.q_\pi(s', a')$$
$$q_\pi(s, a) = \sum_r r \sum_{s'} p(r, s'|s, a) + \sum_{s'} \sum_r p(r, s'|s, a) \sum_{a'} \pi(a'|s').\gamma.q_\pi(s', a')$$
$$q_\pi(s, a) = r(s, a) + \gamma. \sum_{s'} p(s'|s, a) \sum_{a'} \pi(a'|s') q_\pi(s', a')$$

Pela definição de q_* temos:

$$q_*(s, a) = \sum_{s',r} p(s', r|s, a)[r + \gamma \max_{a'} q_*(s', a')]$$

Então podemos dizer que v_* em função de $p(s'|s, a)$ e $r(s, a)$ é:

$$q_*(s, a) = r(s, a) + \gamma. \sum_{s'} p(s'|s, a) \max_{a'} [\pi(a'|s') q_\pi(s', a')]$$

Então as equações que queremos são:

$$\begin{array}{rcl} v_\pi(s) & = & \sum_a \pi(a|s)[r(s,a) + \sum_{s'} \gamma v_\pi(s')p(s'|s,a)] \\ v_*(s) & = & \max_a[r(s,a) + \sum_{s'} \gamma v_\pi(s')p(s'|s,a)] \\ q_\pi(s,a) & = & r(s,a) + \gamma. \sum_{s'} p(s'|s,a) \sum_{a'} \pi(a'|s')q_\pi(s',a') \\ q_*(s,a) & = & r(s,a) + \gamma. \sum_{s'} p(s'|s,a) \max_{a'} [\pi(a'|s')q_\pi(s',a')] \end{array}$$