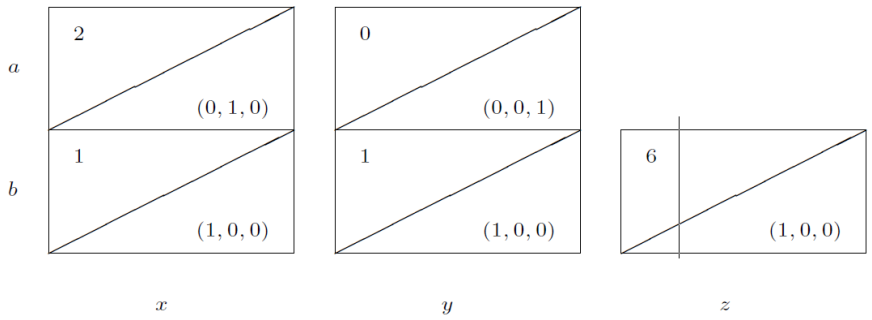

Figure 3.2 Diagrammatic representation of the Markov decision process for Exercise 3.6.

Exercise 3.6

Consider the MDP shown in Figure 3.2. Assuming that this as a discounted infinite-horizon problem with $\delta = \frac{1}{2}$, show that the optimal strategy is $a^*(x) = a$ and $a^*(y) = a$. (Because being in state z gives the highest reward, it seems worth trying a strategy that eventually puts the process in state z starting from any state.) [Hint: solve the dynamic programming equation to find the payoffs for following the specified strategy. Then show that changing the action chosen in any state gives a lower payoff.]

若起始policy是$a(x) = a,\ a(y) = a$:

$$\begin{cases} v(x) = 2 + \frac{1}{2}v(y) \\ v(y) = \frac{1}{2}v(z) \\ v(z) = 6 + \frac{1}{2}v(x) \end{cases}$$

可得:

$$\begin{cases} v(x) = 4 \\ v(y) = 4 \\ v(z) = 8 \end{cases}$$

若令下一步policy是$a(x) = b$, 则:

$$v(x) = 1 + \frac{1}{2}v(x) = 3 < 4$$

若令下一步policy是$a(y) = b$, 则:

$$v(y) = 1 + \frac{1}{2}v(x) = 3 < 4$$

所以policy会一直是$a(x)=a,\ a(y)=a$即收敛。
所以optimal是$a^*(x)=a,\ a^*(y)=a$。

Exercise 3.7

Find the optimal strategy for the previous exercise by starting with $s=\{a(x)=a,a(y)=a\}$ or $s=\{a(x)=a,a(y)=b\}$.

若起始policy是$s=\{a(x)=a,a(y)=a\}$
根据3.6题，optimal是$a^*(x)=a,\ a^*(y)=a$
若起始policy是$s=\{a(x)=a,a(y)=b\}$

$$\begin{cases} v(x)=2+\frac{1}{2}v(y) \\ v(y)=1+\frac{1}{2}v(x) \\ v(z)=6+\frac{1}{2}v(x) \end{cases}$$

可得：

$$\begin{cases} v(x)=\frac{10}{3} \\ v(y)=\frac{8}{3} \\ v(z)=\frac{23}{3} \end{cases}$$

若令下一步policy是$a(x)=b$，则：

$$v(x)=1+\frac{1}{2}v(x)=\frac{8}{3}<\frac{10}{3}$$

若令下一步policy是$a(y)=a$，则：

$$v(y)=\frac{1}{2}v(z)=\frac{23}{6}>\frac{8}{3}$$

所以下一步policy是$a(x)=a,\ a(y)=a$。
根据3.6题，optimal是$a^*(x)=a,\ a^*(y)=a$