

الف) در Transition بین state های Reward توسط محیط دریافت می شود که می تواند

Stochastic هم باشد. ولی این کار باعث می شود که ~~نسبت به~~ نسبت به یاداشی های که در دراز مدت کسب خواهیم کرد بی تفاوت باشیم و آن action ای را انتخاب کنیم که Reward لحظاتی بسیاری داشته باشد. ولی در یاداشی تکاملی می توانیم به یاداشی های که در دراز مدت درست خواهیم آورد توجه بسیاری بکنیم که این کار باعث می شود انتخاب درست تری داشته باشیم. این مفهوم معمولاً با یک discount factor همراه است که این factor نشان دهنده میزان اهمیت یاداشی های دور تری بوده است که هرچه در این مقدار به 1 نزدیکتر باشد، یاداشی های بلندمدت دارای ارزش بیشتری می شوند و اگر این مقدار نزدیک صفر باشد یاداشی لحظاتی اهمیت بیشتری پیدا می کند.

$$R_{t:t+K+1} = \sum_{t=t}^{t+K+1} \gamma^t R_t$$

ب) اگر مدل محیط را داشته باشیم (یعنی P, R و S) هر ادایسته باشیم و مدل خیلی بزرگتر باشد بزرگ نباشد از روش مبتنی بر مدل با استفاده از Policy iteration و Value iteration استفاده می کنیم، در غیر این صورت از روش های مستقل استفاده می کنیم.

| | 0 | 1 | 2 | 3 |
|---|---|---|---|----|
| 0 | | | | +3 |
| 1 | | | | -2 |
| 2 | | | | |

$$V[0,2] = 0,6 \times [0 + 0,2 \times 3] = 0,36$$

$$V[1,2] = 0,2 \times [0 + 0,2 \times -2] = -0,08$$

~~0,08~~

$$V[2,3] = 0,2 \times [0 + 0,2 \times -2] = -0,08$$

| 0 | 1 | 2 | 3 |
|---|---|-------|-------|
| 0 | 0 | 0,36 | +3 |
| 0 | 0 | -0,08 | -2 |
| 0 | | 0 | -0,08 |

| | | | |
|---|---|---|---|
| ↖ | → | → | |
| ↖ | ↖ | ↑ | |
| ↑ | | ↖ | ← |

$$V[0,1] = 0,6 \times [0 + 0,2 \times 0,36] = 0,0432$$

$$V[1,1] = 0,2 \times [0 + 0,2 \times -0,08] = -3,2 \times 10^{-3}$$

$$V[2,2] = 0,6 \times [0 + 0,2 \times -0,08] + 0,2 \times [0 + 0,2 \times -0,08] = ~~-0,128~~ -0,0128$$

$$V[0,2] = 0,6 \times [0 + 0,2 \times 3] + 0,2 \times [0 + 0,2 \times -0,08] = 0,3632$$

$$V[1,2] = 0,6 \times [0 + 0,2 \times 0,36] + 0,2 \times [0 + 0,2 \times -2] = -0,0368$$

$$V[2,3] = -0,08$$

| | | | |
|---|-----------------------|---------|-------|
| 0 | 0,0432 | 0,3632 | +3 |
| 0 | $-3,2 \times 10^{-3}$ | -0,0368 | -2 |
| 0 | | -0,0128 | -0,08 |

| | | | |
|---|---|---|---|
| → | → | → | |
| ↑ | ↑ | ↑ | |
| ↑ | | ↑ | ← |

$$V[0,0] = 0,6 \times [0 + 0,2 \times 0,0432] = 5,184 \times 10^{-3}$$

$$V[1,0] = 0,2 \times [0 + 0,2 \times -3,2 \times 10^{-3}] = -1,28 \times 10^{-4}$$

$$\begin{aligned} V[0,1] &= 0,6 \times [0 + 0,2 \times 0,3632] + 0,2 \times [0 + 0,2 \times -3,2 \times 10^{-3}] \\ &= 0,043072 \end{aligned}$$


$$\begin{aligned} V[1,1] &= 0,6 [0 + 0,2 \times 0,0432] + 0,2 \times [0 + 0,2 \times -0,0368] \\ &= 1,984 \times 10^{-3} \end{aligned}$$


$$V[0,2] = 0,3632$$

$$\begin{aligned} V[1,2] &= 0,6 \times [0 + 0,2 \times 0,3632] + 0,2 [0 + 0,2 \times -2] + 0,2 \times [0 + 0,2 \times -3,2 \times 10^{-3}] \\ &= -0,036928 \end{aligned}$$

$$\begin{aligned} V[2,2] &= 0,6 \times [0 + 0,2 \times -0,0368] + 0,2 \times [0 + 0,2 \times -0,08] \\ &= -7,616 \times 10^{-3} \end{aligned}$$

$$\begin{aligned} V[2,3] &= 0,6 \times [0 + 0,2 \times -0,0128] + 0,2 \times [0 + 0,2 \times -2] \\ &= -0,081536 \end{aligned}$$

| | | | |
|---------------------------|---|----------------------------|---------------|
| 5,184 $\times 10^{-3}$ | 0,043 072 | 0,9632 | +3 |
| -1,28 $\times 10^{-4}$ | 1,984 $\times 10^{-3}$ | -0,036 928 | -2 |
| 0 |  | -7,616 $\times 10^{-3}$ | -0,081 536 |

| | | | |
|---|---|---|---|
| → | → | → | |
| → | ↑ | ↑ | |
| ↑ |  | ↑ | ← |