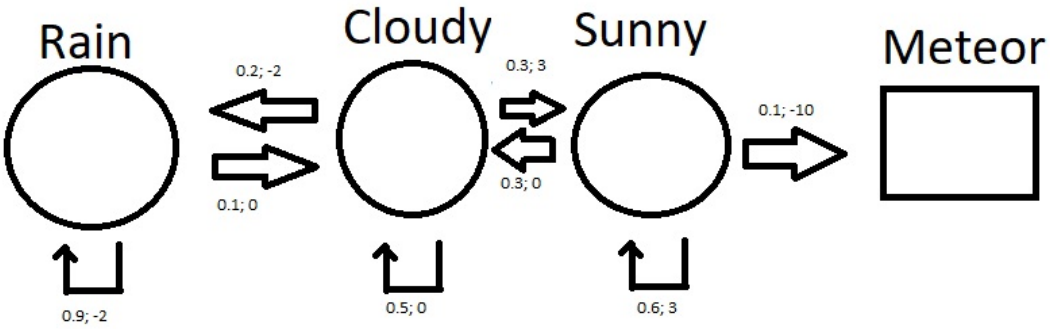
1.1 en 1.2



1.3

Rain Cloudy Sunny Meteor = 0*1 + 3* 1 -10*1 = -7

Rain Cloudy Sunny Cloudy Sunny Sunny Meteor = 0*1 + 3* 1 + 1*0 + 1*3 + 1*3 – 10*1 = -1

1.4

State	V0	V1	V2
Rain	0	-1.8	-3.37
Cloudy	0	0.5	0.316
Sunny	0	0.8	1.375
Meteor	0	0	0

1.5

Een discount van 1 zorgt voor een oneindige loop. Een potentiëlen reward die oneindig ver in de toekost zit telt net zo zwaar mee als een reward nu. Deze reward moet dus ook worden meegenomen en zo kan er geen reward bepaald worden.

Een ander punt is dat in financielen situaties een reward nu kan worden geprefereerd over een reward later omdat er rente kan worden verdient over de reward nu.

Als derde punt. De reward in de toekomst is vaak onzeker omdat er van alles kan verandere waardoor die reward niet exact is zoals wij denken dat hij is.

Als mensen prefereren wij dus reward nu tegenover een reward later. Om dit gedrag na te bootsen in een markov procces gebruiken wij een discount die lager is dan 1 zodat het algoritme ook in bepaalden situaties een reward nu prefereerd over een reward later.

2

Laten we de states s1, s2 en s3 noemen. Laten we beginnen bij s3.

s3 kan geen acties uitvoeren dus de value blijft 0.

S2 kan naar links dit levert -1 + 0 = -1 op of naar rechts dit levert -0.1 + 0 = -0.1 op.

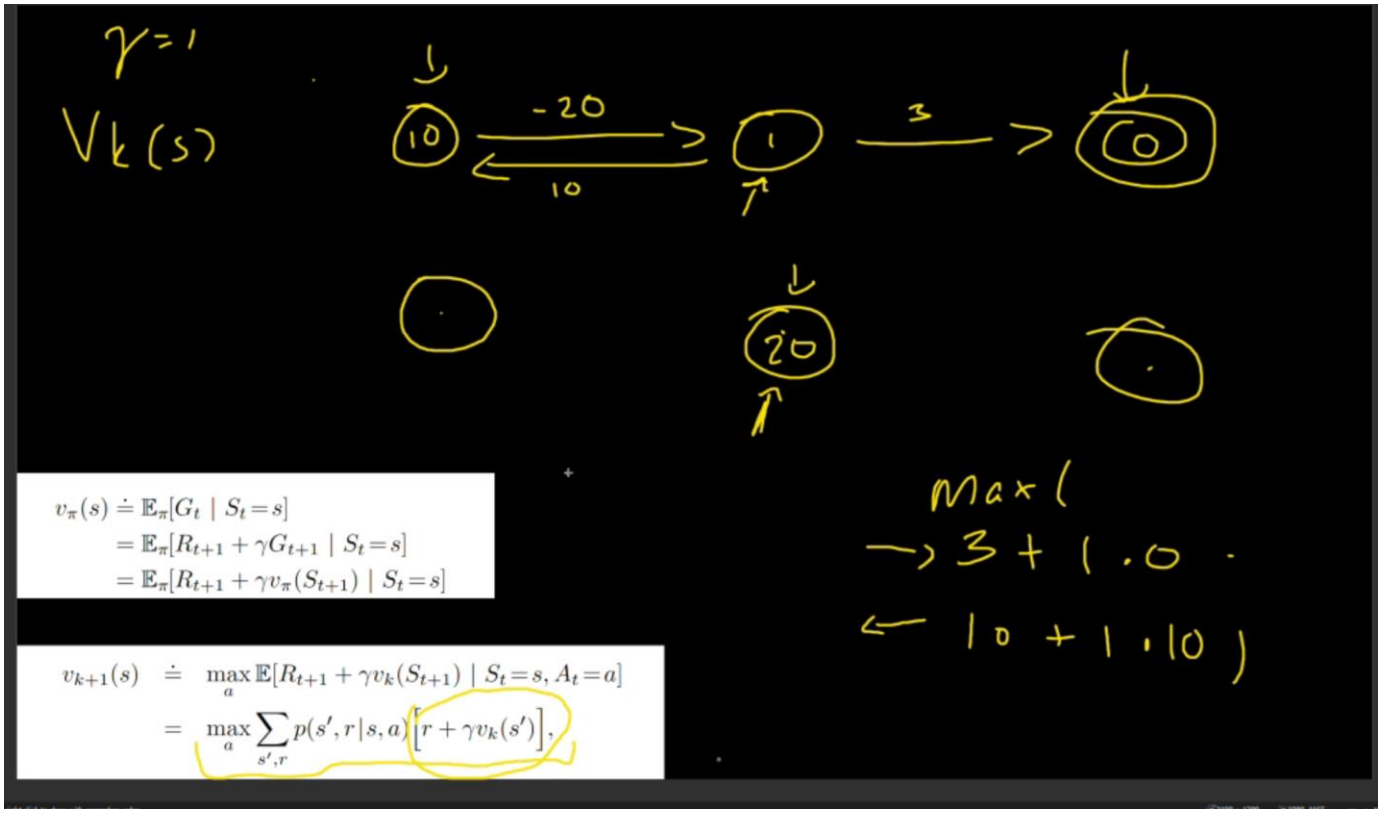
S2 krijgt de value -0.1.

s1 kan naar rechts en dit levert -0.1 + 0 = -0.1 op

s1 wordt -0.1. s2 kan naar links dit levert -1 + 0 = -1 op of naar rechts dit levert -0.1 -0.1 = -0.2. s2 wordt -0.2. Dit gaat even door zie hier onder

Iteratie	S1	S2	S3
0	0	0	-1
1	-0.1	-0.1	-1
2	-0.2	-0.2	-1
3	-0.3	-0.3	-1
4	-0.4	-0.4	-1
5	-0.5	-0.5	-1
6	-0.6	-0.6	-1
7	-0.7	-0.7	-1
8	-0.8	-0.8	-1
9	-0.9	-0.9	-1
10	-1	-1	-1
11	-1.1	-1	-1
12	-1.1	-1	-1

Na iteratie 12 zullen er geen veranderingen meer zijn. Omdat de value van s3 niet meer veranderd en s2 altijd naar s3 zal gaan zal de value van s2 niet meer veranderen. Omdat s1 altijd naar s2 gaat zal s1 ook niet meer veranderen.



$$
\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]
\end{aligned}
$$

$$
\begin{aligned}
v_{k+1}(s) &\doteq \max_a \mathbb{E}[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s, A_t = a] \\
&= \max_a \sum_{s',r} p(s', r \mid s, a)\Big[r + \gamma v_k(s')\Big],
\end{aligned}
$$