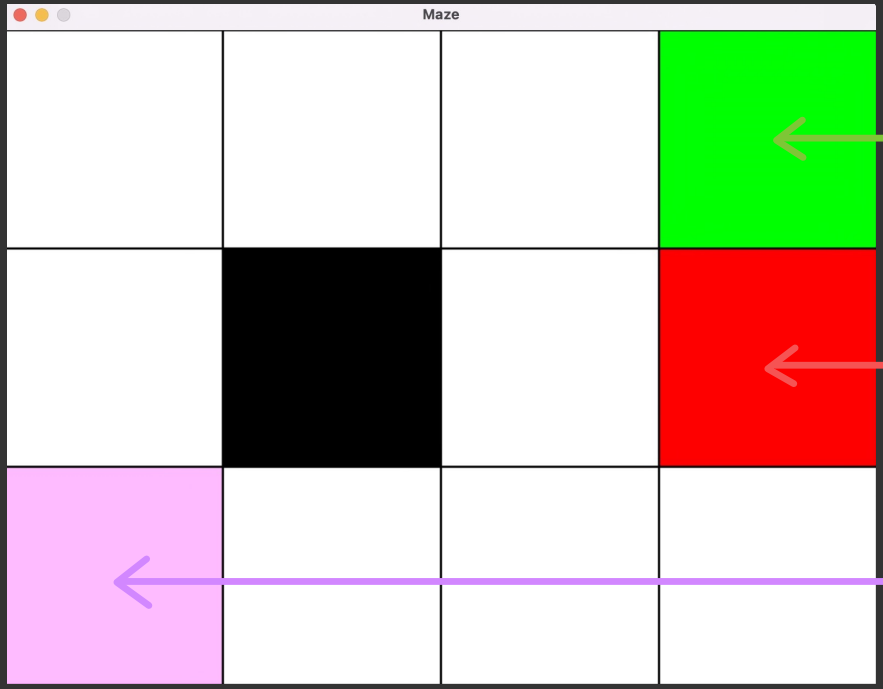
passive reinforcement learning

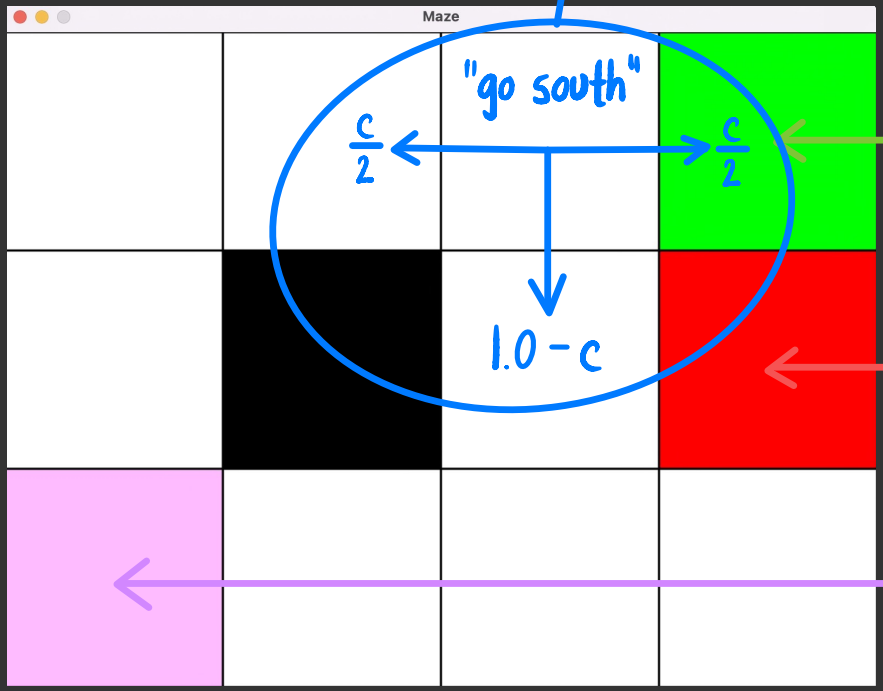
2 nov 2022

CSCI 373



every other state has a "penalty of living" reward of -0.1

the probability of moving in your intended direction is $1.0 - c$



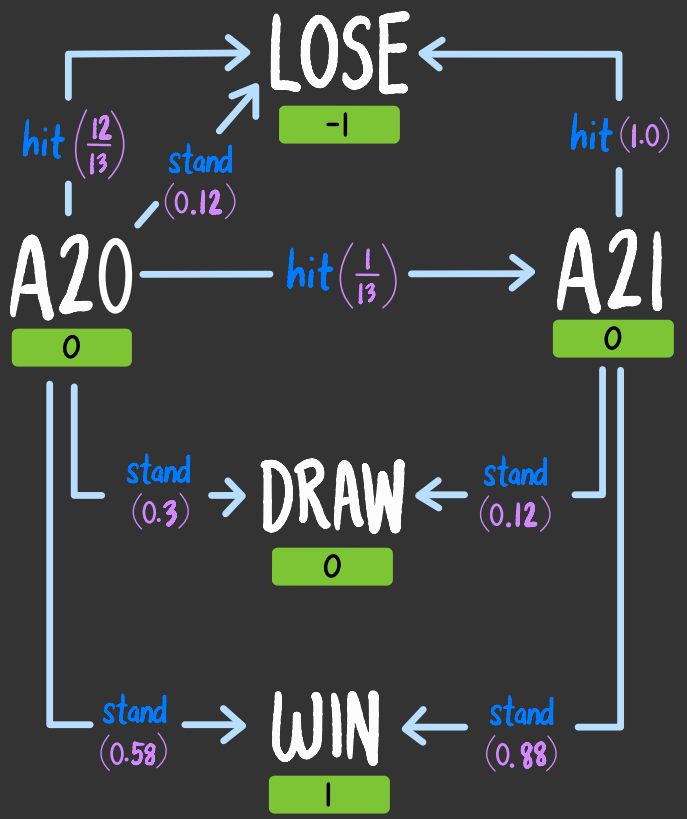
good final state (reward 1)

bad final state (reward -1)

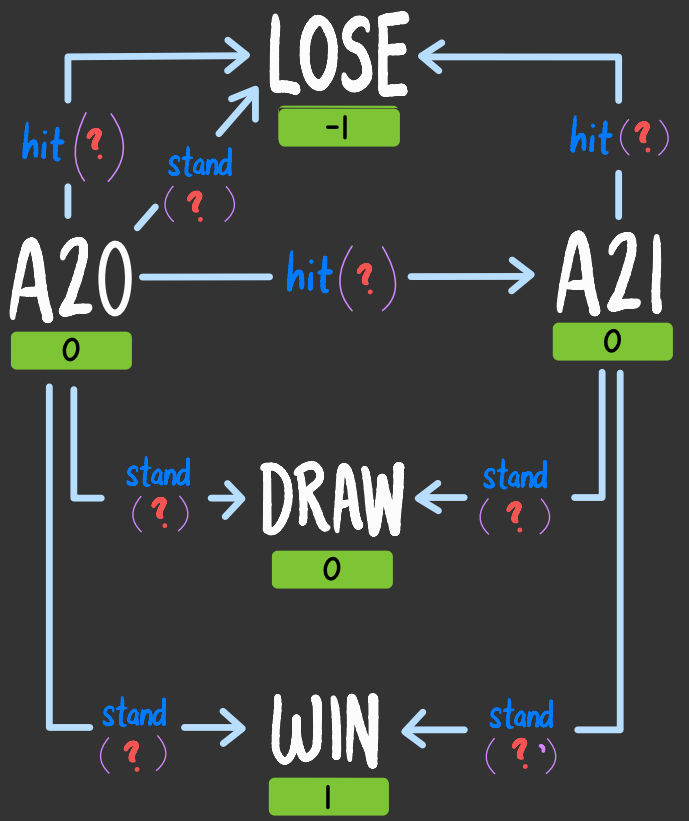
initial state

every other state has a "penalty of living" reward of -0.1

so far we've assumed complete knowledge of the markov decision process



but in a new environment, often we won't know the transition probabilities



or indeed the effects of our actions

let's assume all we know at any given time is:

LOSE
-1

A20
0

A21
0

DRAW
0

WIN
1

actions
- hit
- stand

let's assume all we know at any given time is:

- our current state

let's assume all we know at any given time is:

- our current state
- the reward of our current state

let's assume all we know at any given time is:

- our current state
- the reward of our current state
- the actions we can perform

under these constraints,
how could we compute
the expected utility
of a policy?

i.e. what is U^{π}?

under these constraints,
how could we compute
the expected utility
of a policy?

through experience

i.e. what is U ?

LOSE
-1

A21
0

DRAW
0

WIN
1

actions
- hit
- stand

proposed policy

$\pi = \{A20 \mapsto \text{hit}, A21 \mapsto \text{stand}\}$

LOSE
-1

A20
0

A21
0

what is

$U^{\pi}(A20)$?

DRAW
0

WIN
1

proposed policy

$\pi = \{A20 \mapsto \text{hit}, A21 \mapsto \text{stand}\}$

what is $U^{\pi}(A20)$?

A20
0

LOSE
-1

A21
0

DRAW
0

WIN
1

proposed policy

$\pi = \{A20 \mapsto \text{hit}, A21 \mapsto \text{stand}\}$

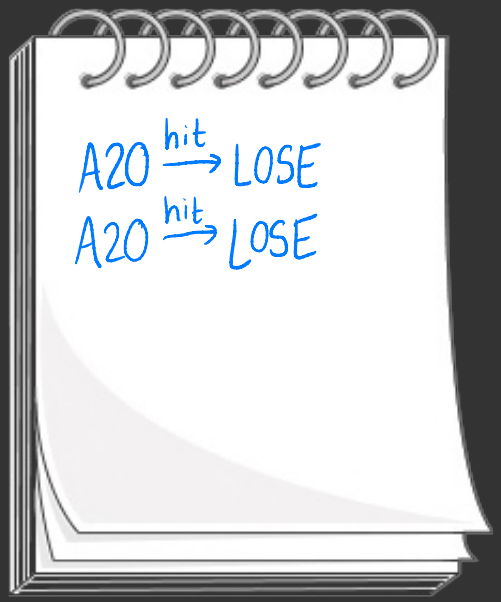
what is

$U^{\pi}(A20)$?

proposed policy

$\pi = \{A20 \mapsto \text{hit}, A21 \mapsto \text{stand}\}$

what is $U^{\pi}(A20)$?



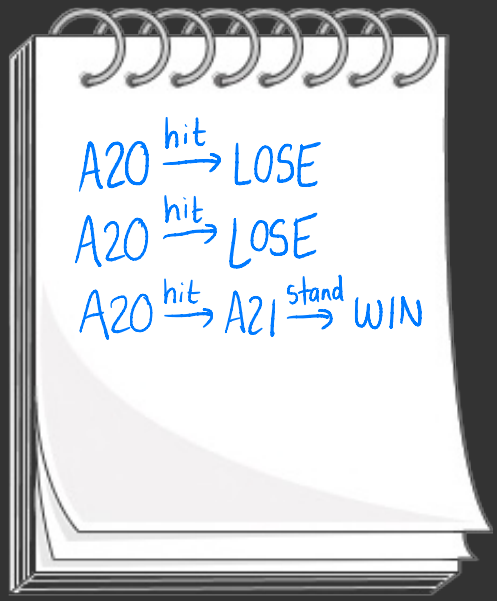
DRAW
0

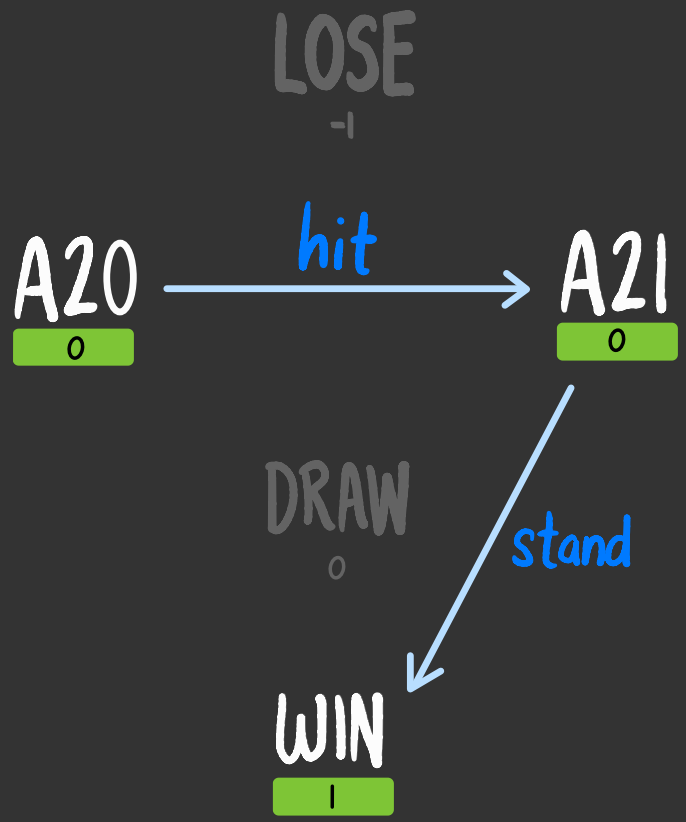
WIN
1

proposed policy

$\pi = \{A20 \mapsto \text{hit}, A21 \mapsto \text{stand}\}$

what is $U^{\pi}(A20)$?

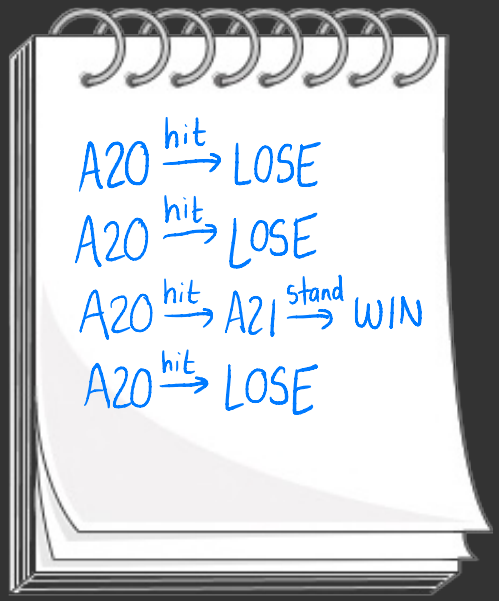




proposed policy

$\pi = \{A20 \mapsto \text{hit}, A21 \mapsto \text{stand}\}$

what is $U^{\pi}(A20)$?



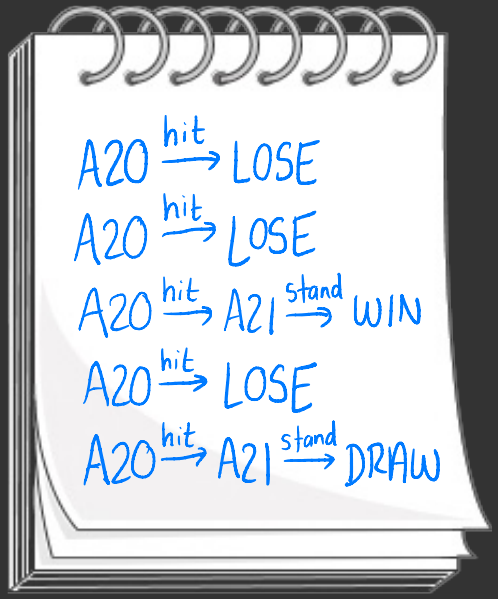
DRAW
0

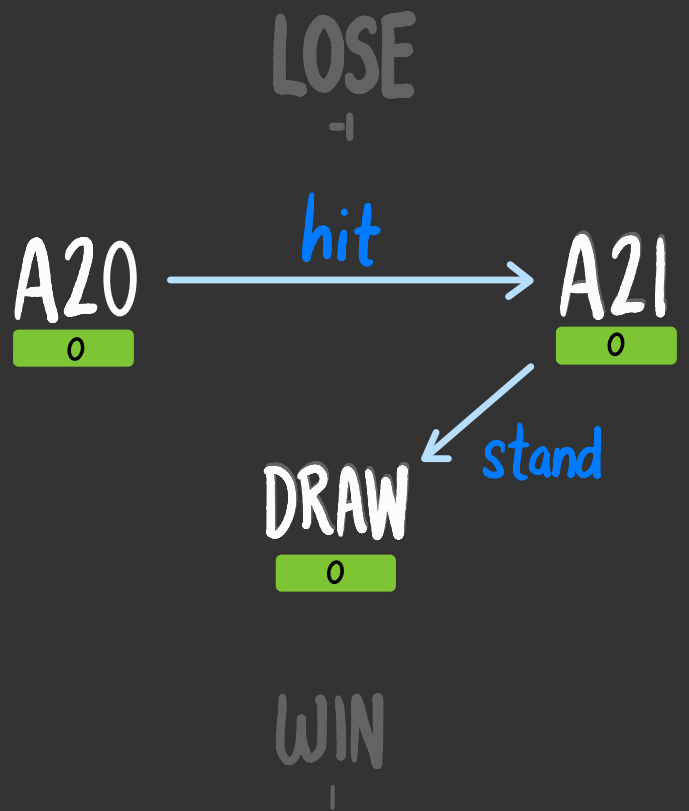
WIN
1

proposed policy

$\pi = \{A20 \mapsto \text{hit}, A21 \mapsto \text{stand}\}$

what is $U^{\pi}(A20)$?

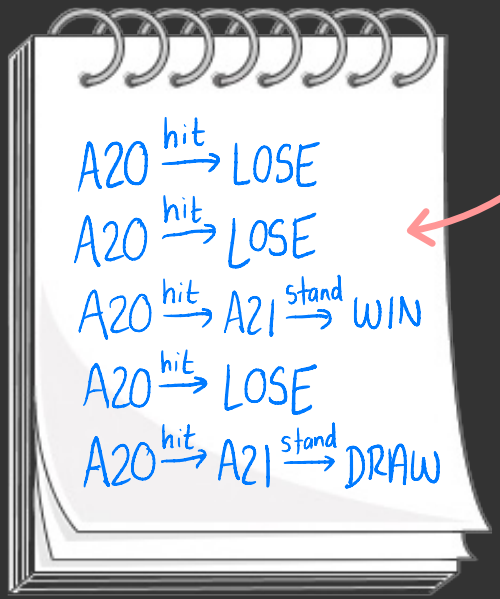




proposed policy

$\pi = \{A20 \mapsto \text{hit}, A21 \mapsto \text{stand}\}$

what is $U^{\pi}(A20)$?



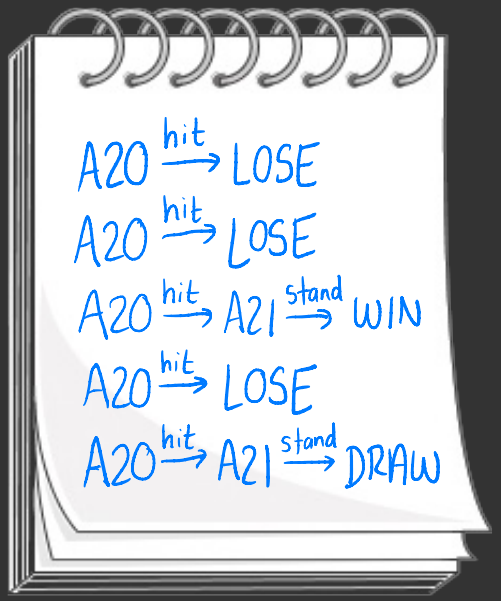
with this data, we can try to estimate the expected utility of our policy

proposed policy

$\pi = \{A20 \mapsto \text{hit}, A21 \mapsto \text{stand}\}$

idea 1: directly estimate the utilities from the data

what is $U^{\pi}(A20)$?



-1
-1
1
-1
0

expected utility

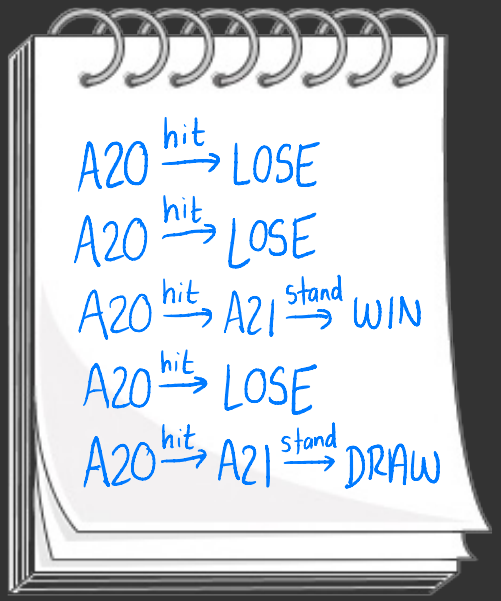
$U^{\pi}(A20) = ?$

proposed policy

$\pi = \{A20 \mapsto \text{hit}, A21 \mapsto \text{stand}\}$

idea 1:
directly estimate the utilities from the data

what is $U^{\pi}(A20)$?



-1
-1
1
-1
0

expected utility

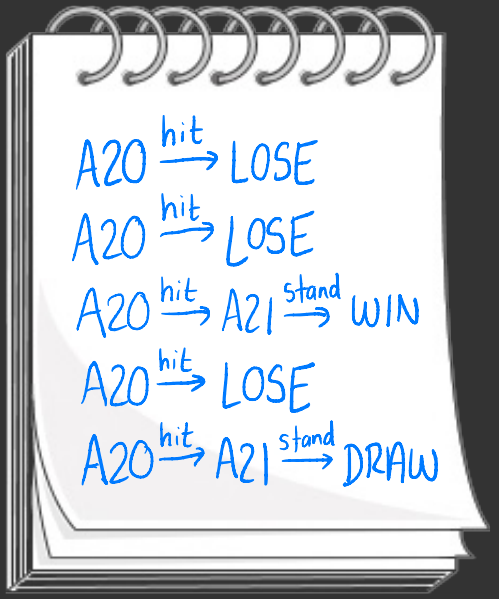
$U^{\pi}(A20) = -0.4$

proposed policy

$\pi = \{A20 \mapsto \text{hit}, A21 \mapsto \text{stand}\}$

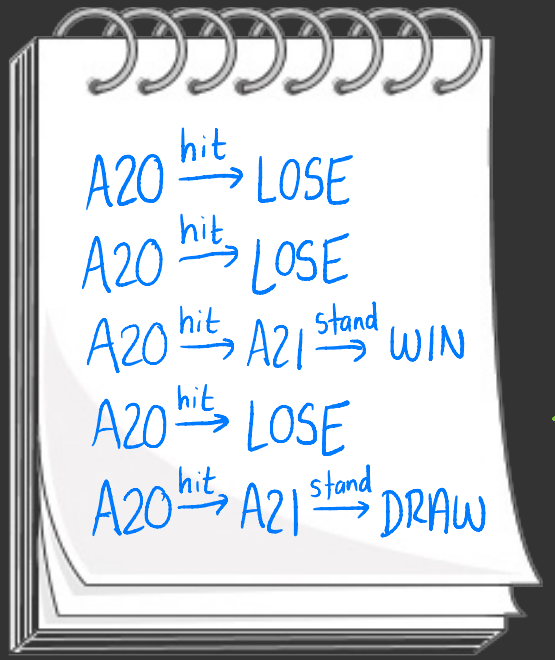
idea 1:
directly estimate the utilities from the data

what is $U^{\pi}(A21)$?



n/a
n/a
1
n/a
0

expected utility
$U^{\pi}(A21) = 0.5$



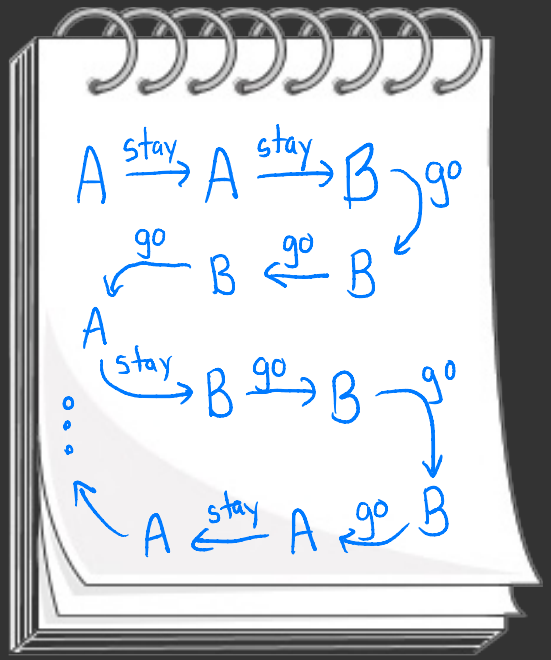
easily obtainable

-1
-1
1
-1
0

expected utility

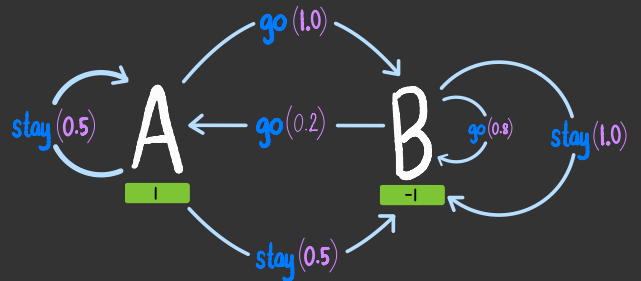
$U^{\pi}(A20) = -0.4$

one complication with this strategy is that the individual utilities might not be easily obtainable

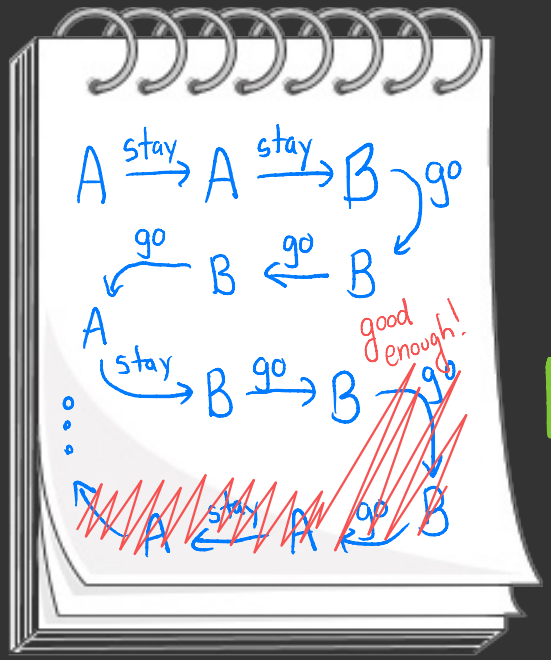


not easily obtainable

?

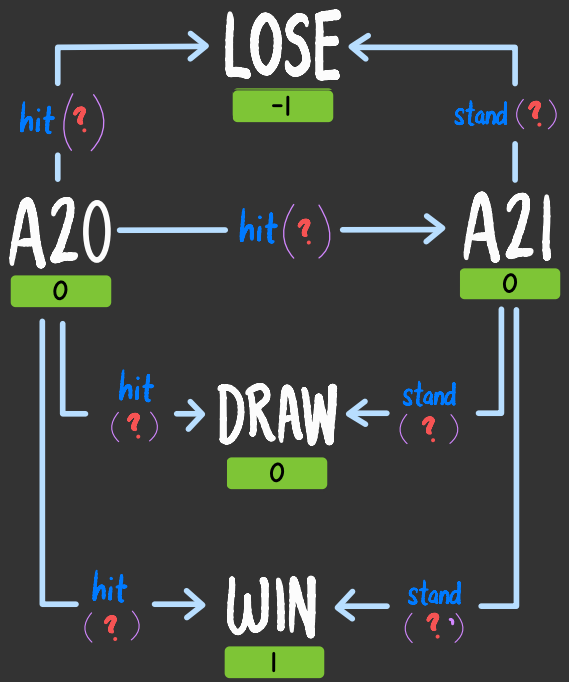


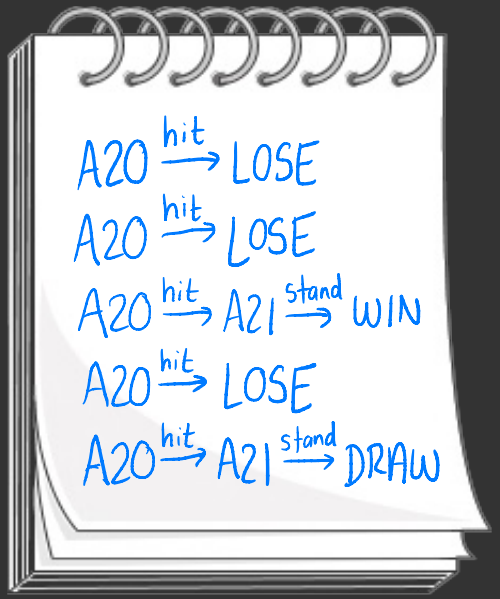
one complication with this strategy is that the individual utilities might not be easily obtainable



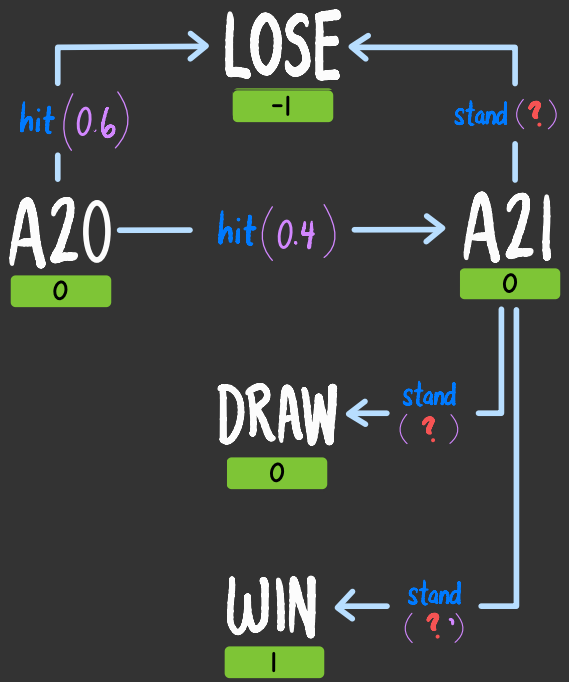
$$1+\frac{1}{2}-\frac{1}{4}-\frac{1}{8}-\frac{1}{16}+\frac{1}{32}-\frac{1}{64}-\frac{1}{128}$$

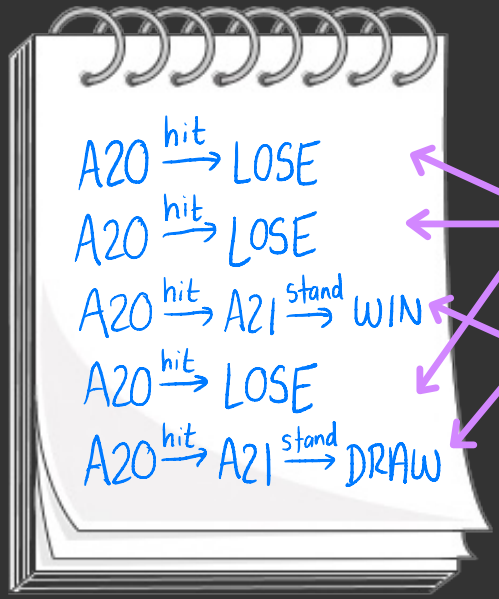
but we can patch this by limiting how far we "roll out" the future





idea 2: estimate the transition probabilities from the data

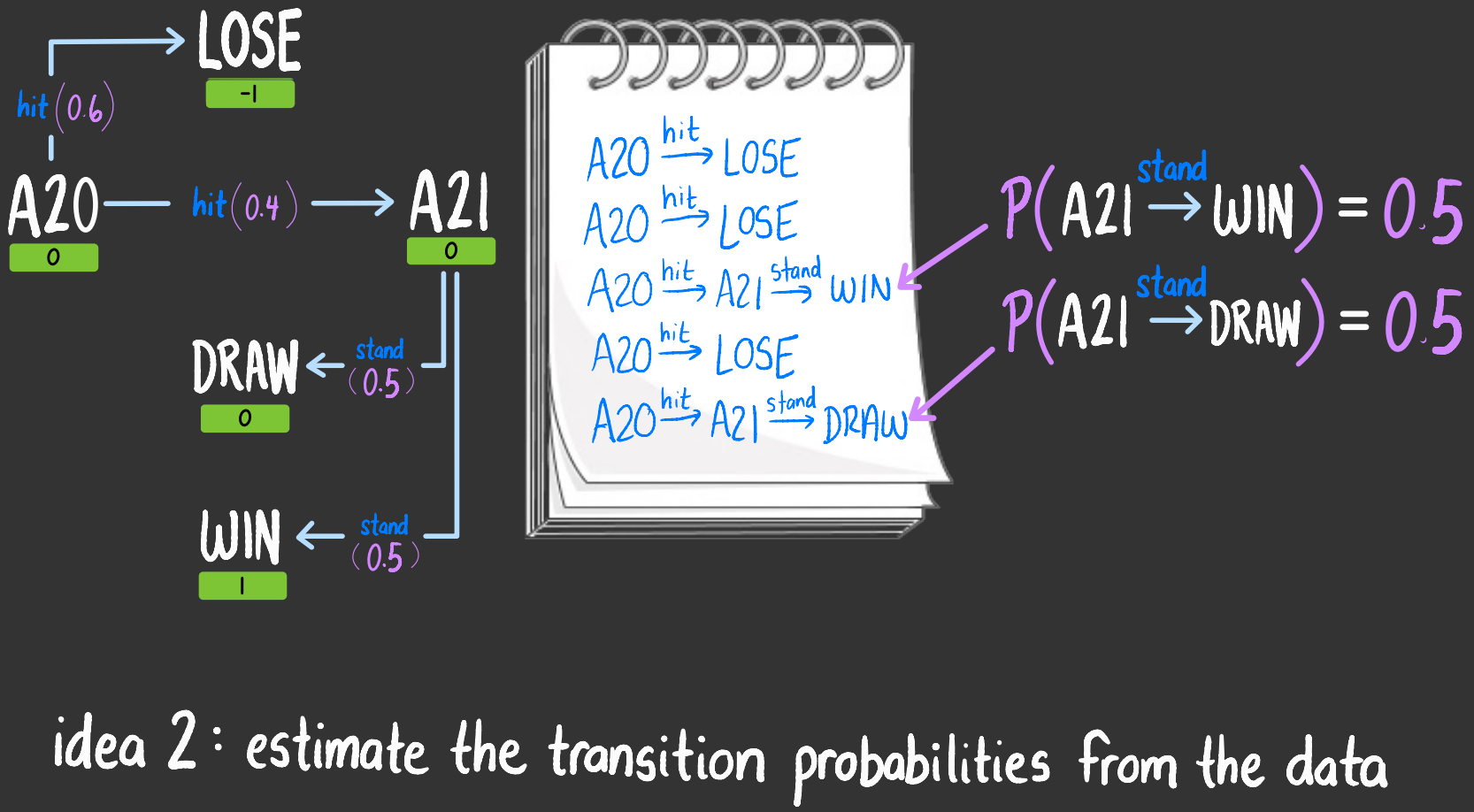




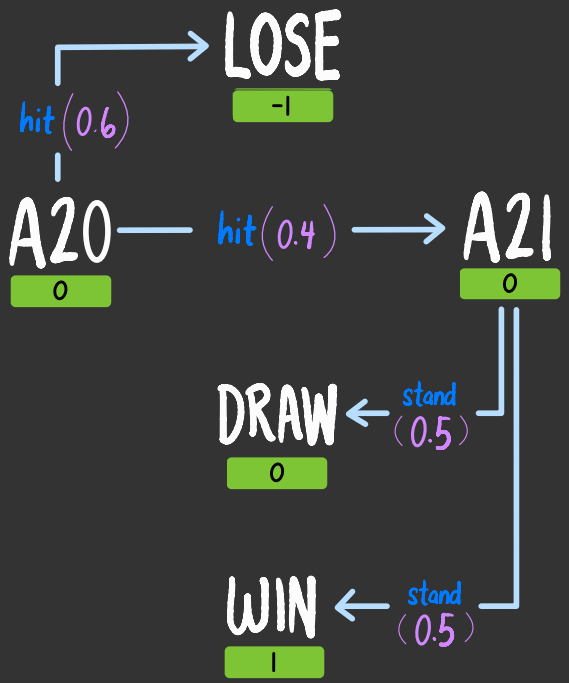
$P(A20 \xrightarrow{hit} LOSE) = 0.6$

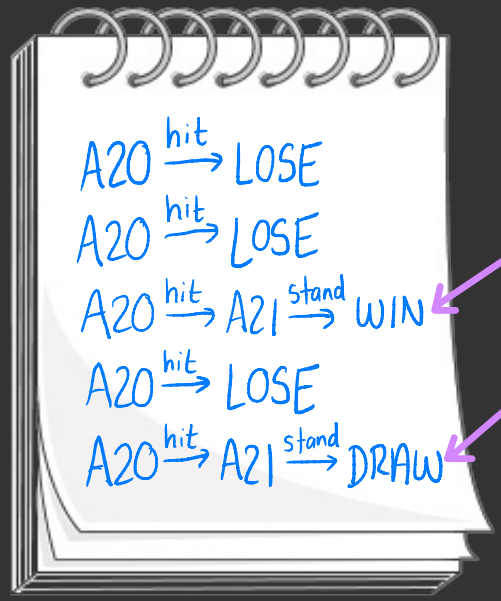
$P(A20 \xrightarrow{hit} A21) = 0.4$

idea 2: estimate the transition probabilities from the data



idea 2: estimate the transition probabilities from the data



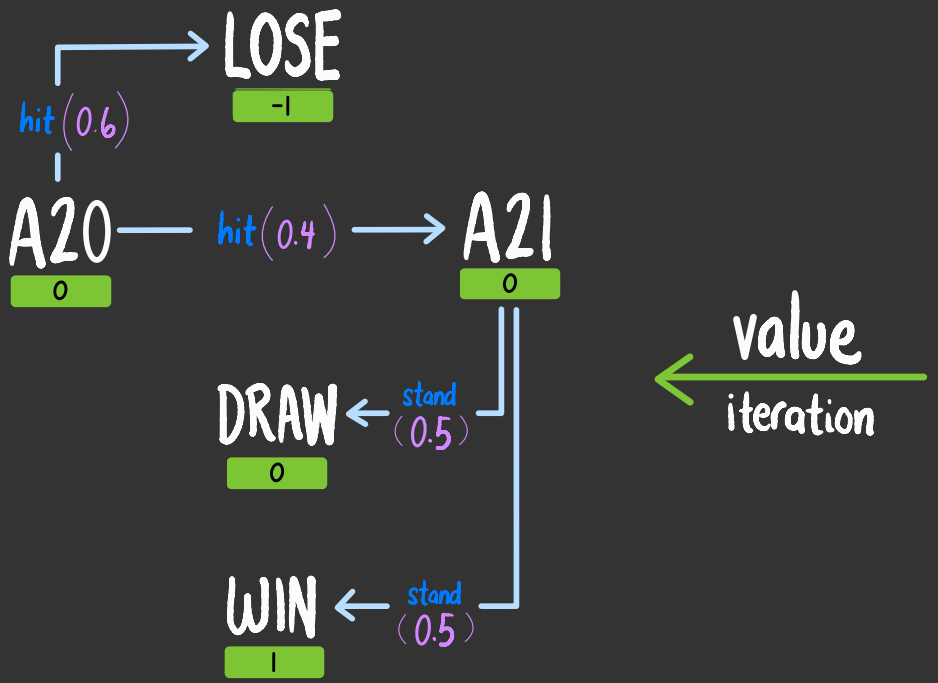


$P(A21 \xrightarrow{stand} WIN) = 0.5$

$P(A21 \xrightarrow{stand} DRAW) = 0.5$

idea 2: estimate the transition probabilities from the data

how can I compute $U^{\pi}(q)$ for the states of the estimated mdp?

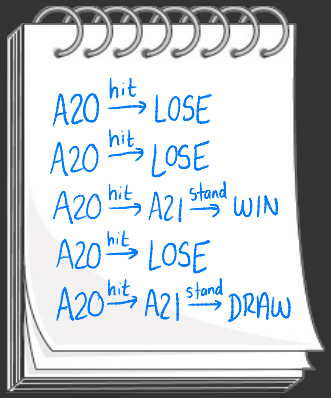


t	U(A)	U(B)
0	1.2500	-1.1300
1	1.6250	-1.3270
2	1.8125	-1.3683
3	1.9063	-1.3661
4	1.9531	-1.3558
5	1.9766	-1.3470
6	1.9883	-1.3411
7	1.9941	-1.3376
8	1.9971	-1.3356
9	1.9985	-1.3345
10	1.9993	-1.3340
11	1.9996	-1.3337
12	1.9998	-1.3335
13	1.9999	-1.3334
14	2.0000	-1.3334
15	2.0000	-1.3334

idea 2: estimate the transition probabilities from the data

then compute expected utilities with value iteration

idea 1:
direct utility estimation



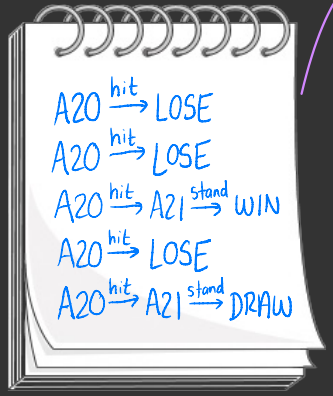
-1
-1
1
-1
0

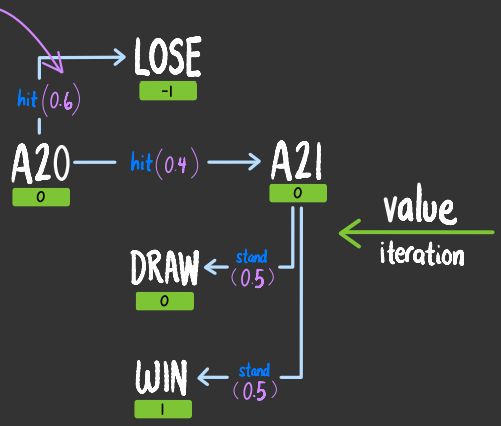
expected utility
$U^{\pi}(A20) = -0.4$

directly estimate the utilities from the data

---

idea 2:
transition probability
estimation



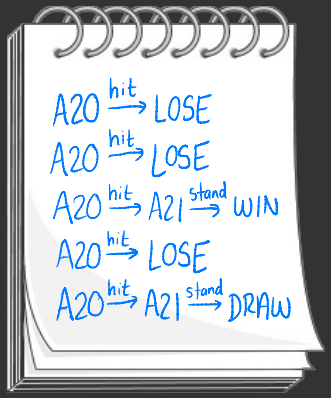


t	U(A)	U(B)
0	1.2500	-1.1300
1	1.6250	-1.3270
2	1.8125	-1.3683
3	1.9063	-1.3661
4	1.9531	-1.3558
5	1.9766	-1.3470
6	1.9883	-1.3411
7	1.9941	-1.3376
8	1.9971	-1.3356
9	1.9985	-1.3345
10	1.9993	-1.3340
11	1.9996	-1.3337
12	1.9998	-1.3335
13	1.9999	-1.3334
14	2.0000	-1.3334
15	2.0000	-1.3334

estimate transition probabilities from the data then do value iteration

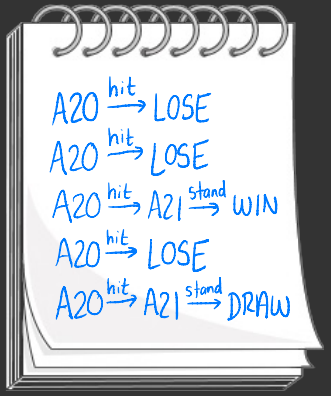
idea 1:

direct utility estimation

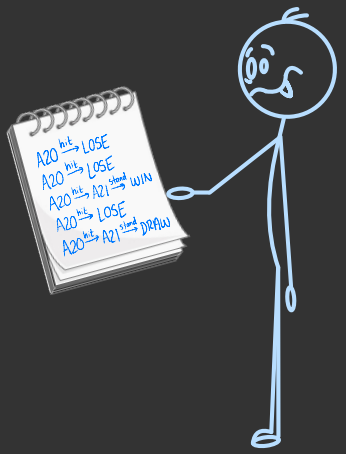


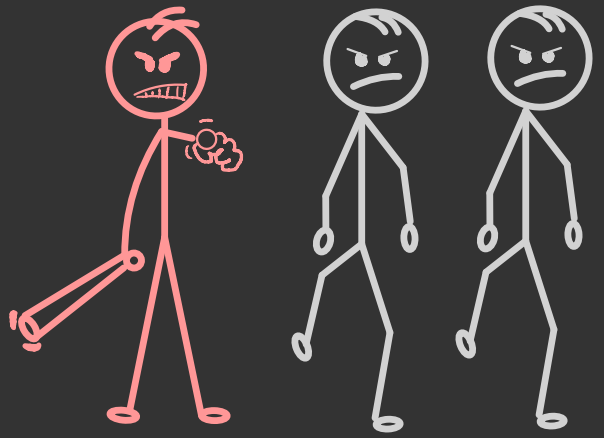
idea 2:

transition probability estimation



what is a drawback to these strategies?



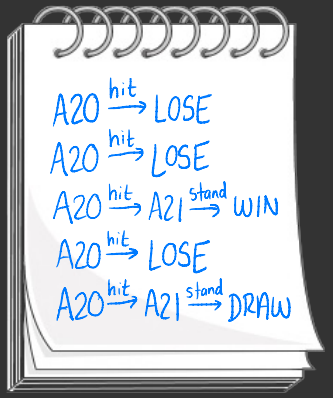


casinos don't like
it when you take

notes

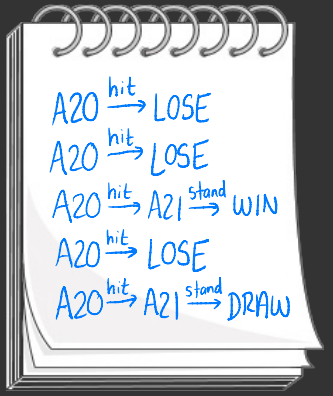
idea 1:

direct utility estimation



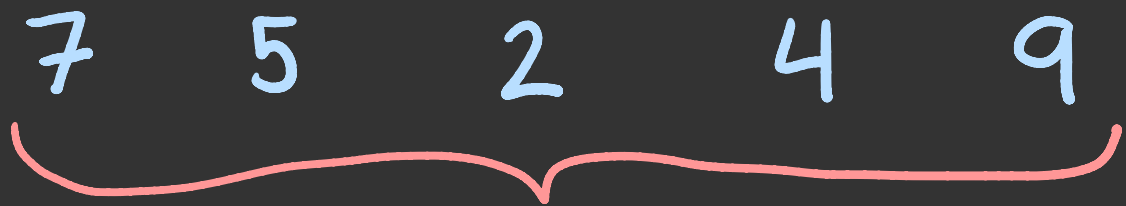
idea 2:

transition probability estimation



what can we do if we don't have a notebook?

averaging a list of numbers is easy if they're written down for us



sum ÷ quantity

27 ÷ 5 = 5.4

but it's not so
hard to do it
in our heads

example numbers?

$$m_n = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{x_n + \sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{\sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} \frac{\sum_{i=1}^{n-1} x_i}{n-1}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} m_{n-1}$$

$$= \frac{x_n}{n} + m_{n-1} - \frac{1}{n} m_{n-1}$$

$$= m_{n-1} + \frac{x_n - m_{n-1}}{n}$$

$$m_n = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{x_n + \sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{\sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} \frac{\sum_{i=1}^{n-1} x_i}{n-1}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} m_{n-1}$$

$$= \frac{x_n}{n} + m_{n-1} - \frac{1}{n} m_{n-1}$$

$$= m_{n-1} + \frac{x_n - m_{n-1}}{n}$$

$$m_n = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{x_n + \sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{\sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} \frac{\sum_{i=1}^{n-1} x_i}{n-1}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} m_{n-1}$$

$$= \frac{x_n}{n} + m_{n-1} - \frac{1}{n} m_{n-1}$$

$$= m_{n-1} + \frac{x_n - m_{n-1}}{n}$$

$$m_n = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{x_n + \sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{\sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} \frac{\sum_{i=1}^{n-1} x_i}{n-1}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} m_{n-1}$$

$$= \frac{x_n}{n} + m_{n-1} - \frac{1}{n} m_{n-1}$$

$$= m_{n-1} + \frac{x_n - m_{n-1}}{n}$$

$$m_n = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{x_n + \sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{\sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} \frac{\sum_{i=1}^{n-1} x_i}{n-1}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} m_{n-1}$$

$$= \frac{x_n}{n} + m_{n-1} - \frac{1}{n} m_{n-1}$$

$$= m_{n-1} + \frac{x_n - m_{n-1}}{n}$$

$$m_n = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{x_n + \sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{\sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} \frac{\sum_{i=1}^{n-1} x_i}{n-1}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} m_{n-1}$$

$$= \frac{x_n}{n} + m_{n-1} - \frac{1}{n} m_{n-1}$$

$$= m_{n-1} + \frac{x_n - m_{n-1}}{n}$$

$$m_n = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{x_n + \sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{\sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} \frac{\sum_{i=1}^{n-1} x_i}{n-1}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} \frac{\sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} m_{n-1}$$

$$= \frac{x_n}{n} + m_{n-1} - \frac{1}{n} m_{n-1}$$

$$= m_{n-1} + \frac{x_n - m_{n-1}}{n}$$

$$m_n = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{x_n + \sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{\sum_{i=1}^{n-1} x_i}{n}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} \frac{\sum_{i=1}^{n-1} x_i}{n-1}$$

$$= \frac{x_n}{n} + \frac{n-1}{n} m_{n-1}$$

$$= \frac{x_n}{n} + m_{n-1} - \frac{1}{n} m_{n-1}$$

$$= m_{n-1} + \frac{x_n - m_{n-1}}{n}$$

to average numbers $x_1, \dots, x_N$
without writing them down:

$$m_n = m_{n-1} + \frac{x_n - m_{n-1}}{n}$$

to average numbers $x_1, \dots, x_N$
without writing them down:

$$m_n = m_{n-1} + \frac{x_n - m_{n-1}}{n}$$

repeat:

$$m = m + \frac{x_n - m}{n}$$

to average numbers $x_1, \dots, x_N$
without writing them down:

$$m_n = m_{n-1} + \frac{x_n - m_{n-1}}{n}$$

repeat:

$$m \mathrel{+}= \frac{x_n - m}{n}$$

to average numbers $x_1, \dots, x_N$ without writing them down:

$m = 0$

for $n = 1$ to N:

$$m \mathrel{+}= \frac{x_n - m}{n}$$

$$m = 0 + \frac{7-0}{1} = 7$$

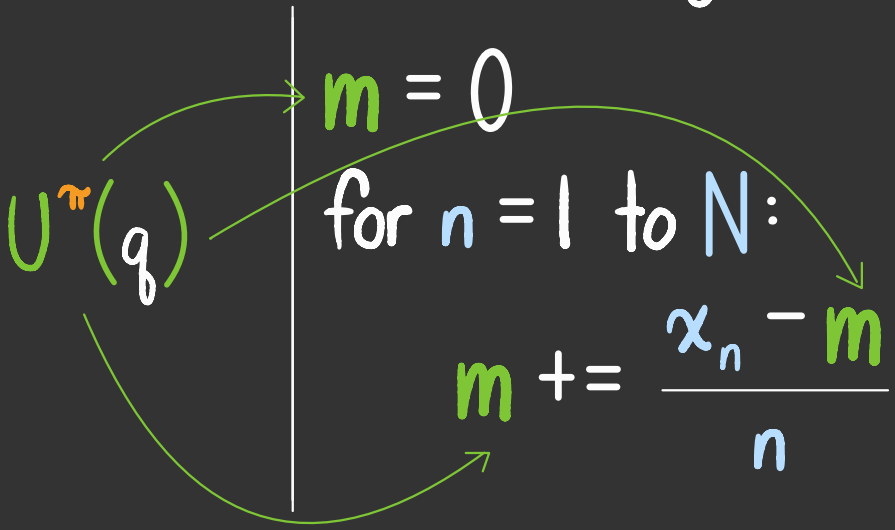
$$m = 7 + \frac{5-7}{2} = 6$$

$$m = 6 + \frac{2-6}{3} = \frac{14}{3}$$

$$m = \frac{14}{3} + \frac{4-\frac{14}{3}}{4} = 4.5$$

$$m = 4.5 + \frac{9-4.5}{5} = 5.4$$

to average numbers $x_1, \dots, x_N$
without writing them down:



to average numbers $x_1, \dots, x_N$
without writing them down:

$$U^\pi(q) = 0$$

for $n = 1$ to N:

$$U^\pi(q) \mathrel{+}= \frac{x_n - U^\pi(q)}{n}$$

to average numbers $x_1, \dots, x_N$
without writing them down:

$U^\pi(q) = 0$

for $n = 1$ to N:

observe transition $q \xrightarrow{\pi(q)} q'$ with utility x_n

$$U^\pi(q) \mathrel{+}= \frac{x_n - U^\pi(q)}{n}$$

but what if?
the mdp is cyclic.

to average numbers $x_1, \ldots, x_N$
without writing them down:

$U^\pi(q) = 0$

for $n = 1$ to N:

observe transition $q \xrightarrow{\pi(q)} q'$ with expected utility $R(q) + \gamma U^\pi(q')$

$$U^\pi(q) \mathrel{+}= \frac{x_n - U^\pi(q)}{n}$$

temporal difference learning

compute an online average of expected utilities

$U^{\pi}(q) = 0$ for every state q

repeat:

observe transition $q \xrightarrow{\pi(q)} q'$ with expected utility $R(q) + \gamma U^{\pi}(q')$

$$U^{\pi}(q) \mathrel{+}= \frac{R(q) + \gamma U^{\pi}(q') - U^{\pi}(q)}{n_q}$$

temporal difference learning

compute an online average of expected utilities

$U^{\pi}(q) = 0$ for every state q

repeat:

observe transition $q \xrightarrow{\pi(q)} q'$ with expected utility $R(q) + \gamma U^{\pi}(q')$

$$U^{\pi}(q) \mathrel{+}= \frac{R(q) + \gamma U^{\pi}(q') - U^{\pi}(q)}{n_q}$$

we need a **separate** denominator for each state q

to the laptop!