

active
reinforcement
learning
4 nov
2022

CSCI
373

LOSE
-

A20
0

DRAW
0

WIN
1

A21
0

actions
• hit
• stand

temporal difference learning

compute an online average of expected utilities

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

repeat:

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

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

last time we found a way to compute the expected utility of a policy in an environment where we know only our current state, reward, and available actions

LOSE
-

A20
0

DRAW
0

WIN
1

A21
0

actions
• hit
• stand

temporal difference learning

compute an online average of expected utilities

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

repeat:

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

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

rather than evaluating a single policy,
we usually want to know the best policy

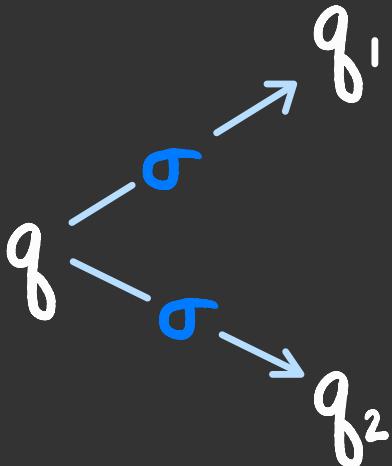
in other words, what's the best action
to perform in each state?

$$U^\pi(q) \rightsquigarrow U(q, \sigma)$$

expected utility of
state q under policy π

expected utility of
action σ in state q

$$R(q) + \gamma p_1 U(q_1)$$

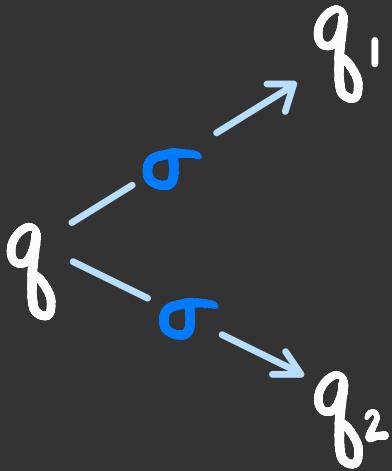


$$+ \gamma p_2 U(q_2)$$

$$U(q, \sigma)$$

expected utility of
action σ in state q

$$R(q) + \gamma \max_{\sigma'} U(q', \sigma')$$



$$U(q, \sigma)$$

$$+ \gamma p_1 \max_{\sigma'} U(q_1, \sigma')$$

expected utility of
action σ in state q

$$\frac{U(q, \sigma) = R(q) + \gamma \sum_{q'} P(q \xrightarrow{\sigma} q') \max_{\sigma'} U(q', \sigma')}{\text{expected utility of action } \sigma \text{ in state } q}$$

expected utility of action σ' in state q'

$$U(q, \sigma) = R(q) + \gamma \sum_{q'} P(q \xrightarrow{\sigma} q') \max_{\sigma'} U(q', \sigma')$$

expected utility of
action σ in state q

expected utility of
action σ' in state q'

let's compute an online average of these

temporal difference learning

compute an online average of expected utilities

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

repeat:

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

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

not

V temporal difference learning

compute an online average of expected utilities

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

repeat:

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

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

not

V temporal difference learning

compute an online average of expected utilities

$U(q, \sigma) = 0$ for every state q , action σ

repeat:

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

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

not

V temporal difference learning

compute an online average of expected utilities

$U(q, \sigma) = 0$ for every state q , action σ

repeat:

no policy now, so we'll choose an action

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

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

not

V temporal difference learning

compute an online average of expected utilities

$U(q, \sigma) = 0$ for every state q , action σ

repeat:

choose action σ

observe transition $q \xrightarrow{\sigma} q'$ with expected utility $R(q) + \gamma \max_{\sigma'} U(q', \sigma')$

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

not

V temporal difference learning

compute an online average of expected utilities

$U(q, \sigma) = 0$ for every state q , action σ

repeat:

choose action σ

observe transition $q \xrightarrow{\sigma} q'$ with expected utility $R(q) + \gamma \max_{\sigma'} U(q', \sigma')$

$$U(q, \sigma) += \frac{R(q) + \gamma \max_{\sigma'} U(q', \sigma') - U(q, \sigma)}{n_{q, \sigma}}$$

qlearning

compute an online average of expected utilities

$U(q, \sigma) = 0$ for every state q , action σ

repeat:

choose action σ

observe transition $q \xrightarrow{\sigma} q'$ with expected utility $R(q) + \gamma \max_{\sigma'} U(q', \sigma')$

$$U(q, \sigma) += \frac{R(q) + \gamma \max_{\sigma'} U(q', \sigma') - U(q, \sigma)}{n_{q, \sigma}}$$

qlearning

compute an online average of expected utilities

$U(q, \sigma) = 0$ for every state q , action σ

repeat:

choose action σ

observe transition $q \xrightarrow{\sigma} q'$ with expected utility $R(q) + \gamma \max_{\sigma'} U(q', \sigma')$

$$U(q, \sigma) += \frac{R(q) + \gamma \max_{\sigma'} U(q', \sigma') - U(q, \sigma)}{n_{q, \sigma}}$$

qlearning

these are called
qvalues

compute an online average of expected utilities

$$U(q, \sigma) = 0 \text{ for every state } q, \text{ action } \sigma$$

repeat:

choose action σ

observe transition $q \xrightarrow{\sigma} q'$ with expected utility $R(q) + \gamma \max_{\sigma'} U(q', \sigma')$

$$U(q, \sigma) += \frac{R(q) + \gamma \max_{\sigma'} U(q', \sigma') - U(q, \sigma)}{n_{q, \sigma}}$$

qlearning

compute an online average of expected utilities

how do we do
this?

$U(q, \sigma) = 0$ for every state q , action σ

repeat:

choose action σ

observe transition $q \xrightarrow{\sigma} q'$ with expected utility $R(q) + \gamma \max_{\sigma'} U(q', \sigma')$

$$U(q, \sigma) += \frac{R(q) + \gamma \max_{\sigma'} U(q', \sigma') - U(q, \sigma)}{n_{q, \sigma}}$$



to the laptop!

qlearning

compute an online average of expected utilities

$$U(q, \sigma) = 0 \text{ for every state } q, \text{ action } \sigma$$

repeat:

$$\sigma = \begin{cases} \text{random } \sigma & \text{with prob } \epsilon \quad ("exploration") \\ \underset{\sigma}{\operatorname{argmax}} \ U(q, \sigma) & \text{otherwise} \quad ("exploitation") \end{cases}$$

observe transition $q \xrightarrow{\sigma} q'$ with expected utility $R(q) + \gamma \max_{\sigma'} U(q', \sigma')$

$$U(q, \sigma) += \frac{R(q) + \gamma \max_{\sigma'} U(q', \sigma') - U(q, \sigma)}{n_{q, \sigma}}$$

exploration

VS.

exploitation



exploration

VS.

exploitation



exploration

VS.

exploitation



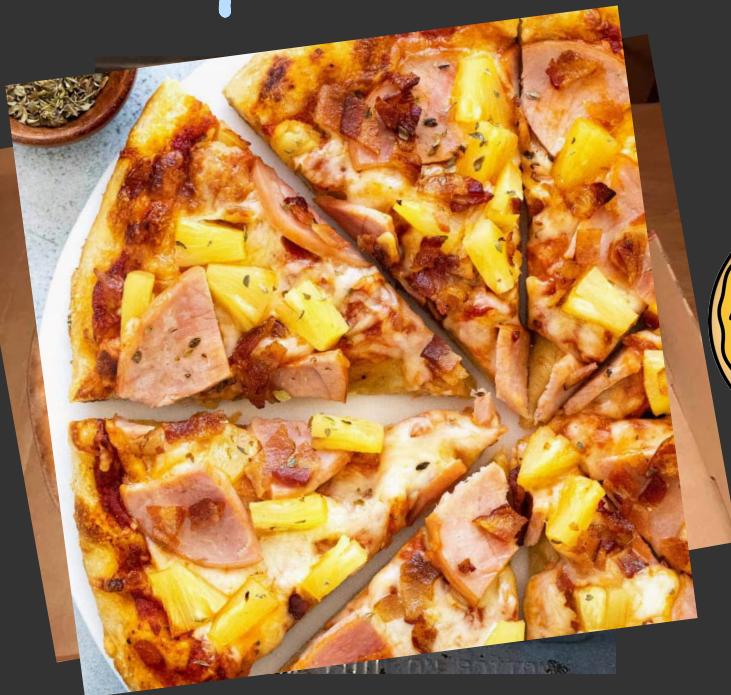
exploration

VS.

exploitation



exploration

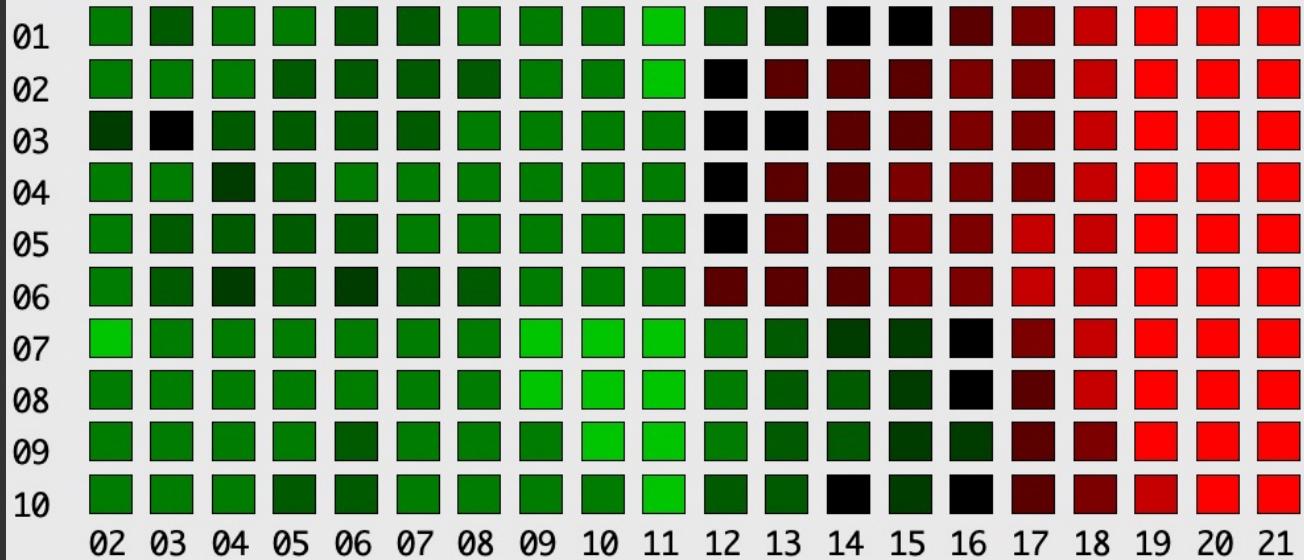


VS.

exploitation



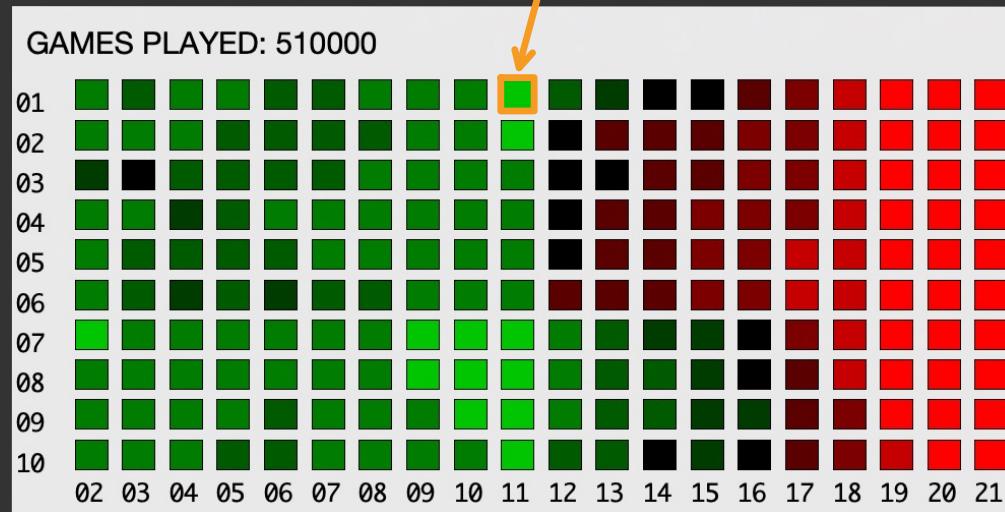
GAMES PLAYED: 510000



to the laptop!

■ if $\mathbb{U}(q, \text{hit}) - \mathbb{U}(q, \text{stand}) > 0$
■ if $\mathbb{U}(q, \text{hit}) - \mathbb{U}(q, \text{stand}) < 0$

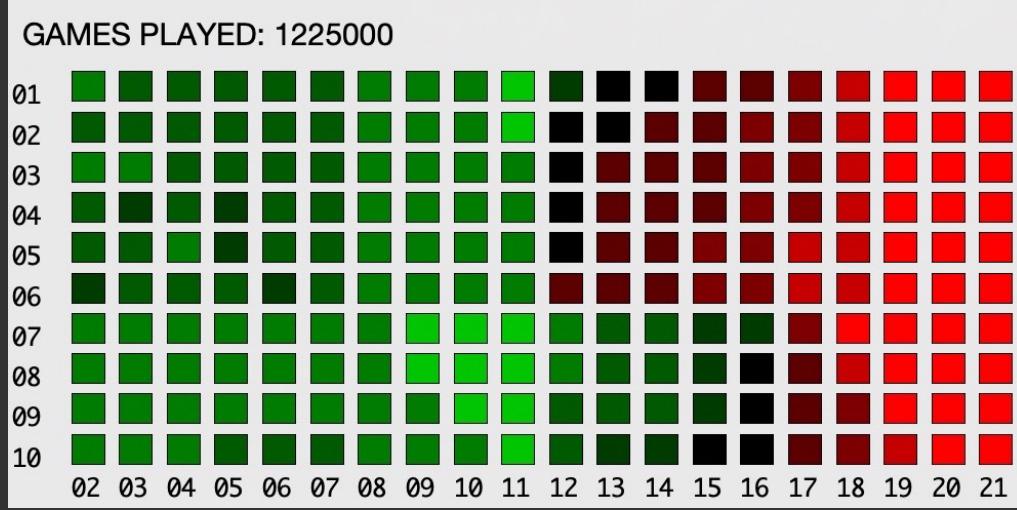
dealer's
up card



player's card total

aces count

as one



aces count
as one or eleven

