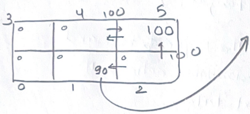
The basic premise is that you teach the algorithm to take certain actions based on prior experience by rewarding or punishing actions.

actions that lead to less reward are shunned / looked down upon.

very bad!

So, how it does that



If this is not punishing actions, I don't know what is.

Exploration

from State 2 and 4, goes to state 5, very rewarding. (100)

Exploitation

Just Use whatever learned. (Record is kept track in q-table), use that stoff.

Now, from state 2, say take left action (to state 1) get immediate reward, 0,

0 + 0.9% of biggest reward available from any action in state 1

0 + 0.9% of 100

90