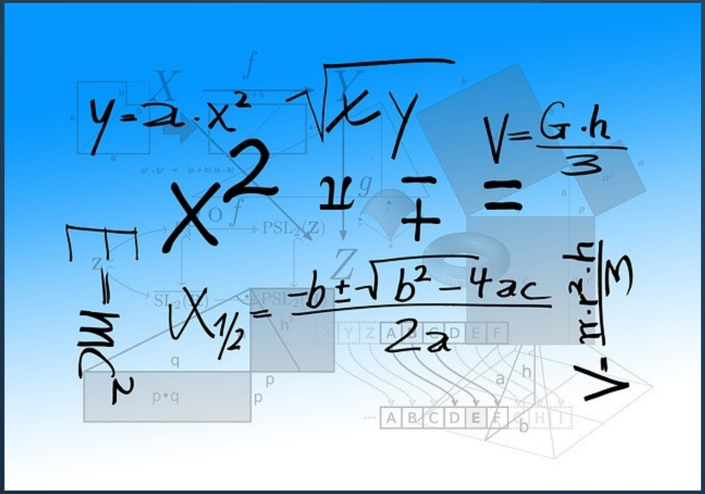
Deep Q Learning: From Paper to Code

The Explore-Exploit Dilemma

Last Time

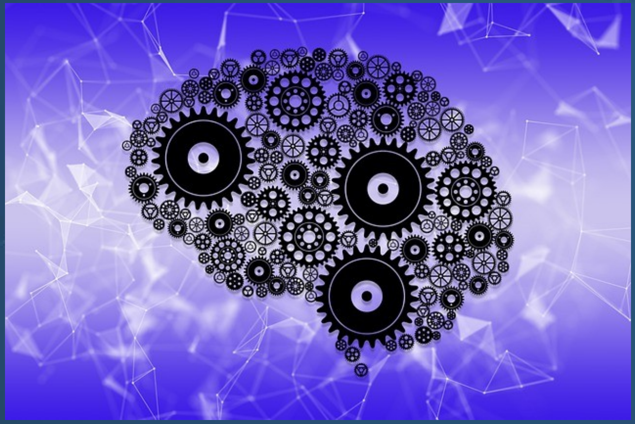
Model based vs. model free

Model free learning → trial & error



Model based → solve equations

Learning vs. Maximizing Rewards



How to learn & max rewards?

Opportunity cost of greed

Explore-Exploit

Best known action → greed

Sub optimal action → exploration

How to balance the two is a dilemma

Quick Example

- Penalty of -1 for each step
- Reward of 0 for winning
- Goal is to minimize negative reward

Escape in as few moves as possible

Quick Example

Have to start with estimate

$$v_{\pi}(s) < 0 \; \forall \; s \in S$$

Set initial estimate to 0 and greedy policy

How does the optimism play out?

Get Used to Disappointment

Disappointment

Exploration

Optimistic

Initial

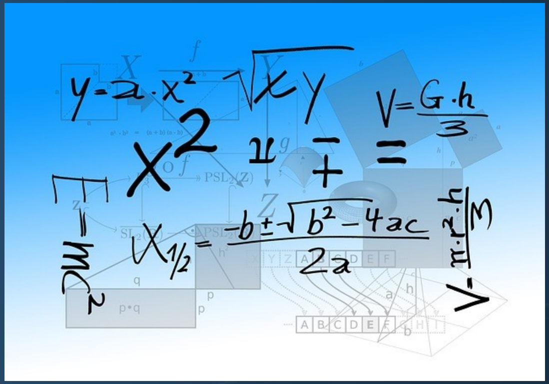
Values

Back to hope

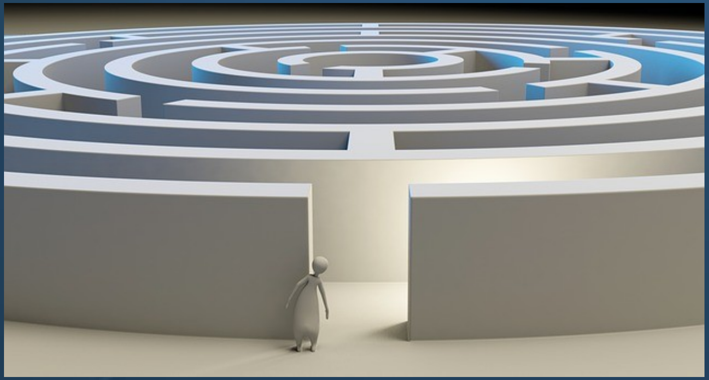
Success

Epsilon Greedy

Parameter for action selection



Random number generator



Explore entirety of state space

Decrease epsilon over time

Epsilon must stay finite

Summary

- Never certain estimates are accurate
- Number of solutions – use epsilon greedy
- Some moves explore, others greed

Up Next