Deep Q Learning: From Paper to Code

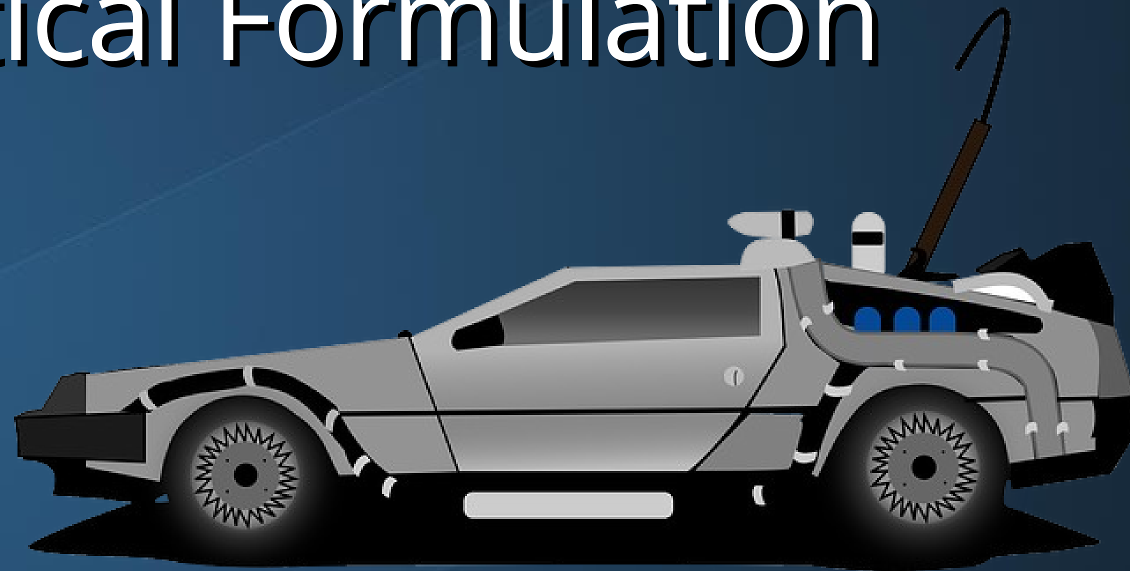
Markov Decision Processes

Last Time ...

- Interactions of agent and environment
- Agent learns and acts
- Environment is what is acted upon
- Rewards tell the agent what is good

Mathematical Formulation

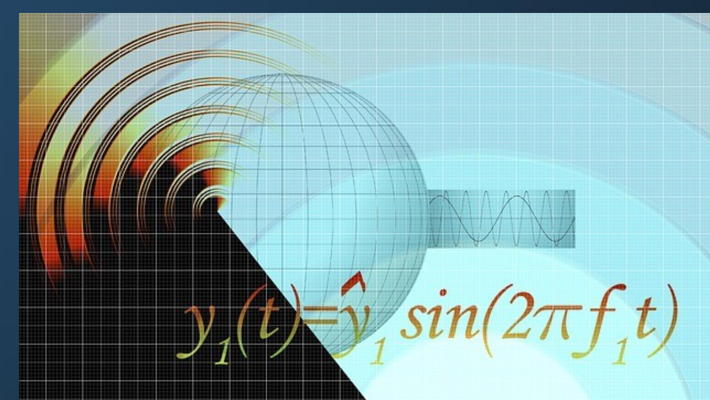
$$\left(S_1, A_1, R_1, S_2, A_2, R_2, \ldots\right)$$



Actions affect all future rewards

State depends only on previous state and action

<u>Markov</u> <u>Decision</u> <u>Process</u>



Mathematical abstraction

Probabilistic Transitions

Actions cause state transitions

$$p(s', r | a, s) \neq 1$$

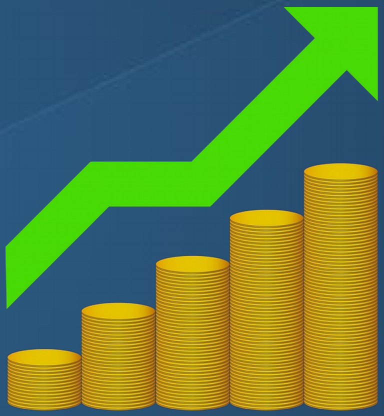
$$\sum_{s', r} p(s', r | a, s) = 1$$

Probabilities define our dynamics

$$r(s,a) = E[R_t | S_{t-1} = s, A_{t-1} = a] = \sum_{r \in R} r \sum_{s' \in S} p(s', r | s, a)$$

Expected reward → outcome * probability

Maximizing Rewards & Episodic Tasks



Series of rewards → expected return

Sum of rewards that follow current time

$$\boxed{G_t = R_{t+1} + R_{t+2} + R_{t+3}, \ldots, R_T}$$

Episode: discrete period of game play

Episodic Game Play

Terminal state is unique

$$\boxed{G_T = 0}$$

Ensures sum over rewards finite

Reward Discounting

Not all tasks are episodic!

$$\sum_{t=0}^{\infty} R_t \rightarrow \infty$$

Fix by discounting

Discount factor → γ

Reward Discounting

$$0 \leq \gamma \leq 1$$

$1 \rightarrow \gamma$ Far sighted

$0 \rightarrow \gamma$ Myopic

$$0.95 \leq \gamma \leq 0.99$$

$$\boxed{G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}}$$

Reward Discounting

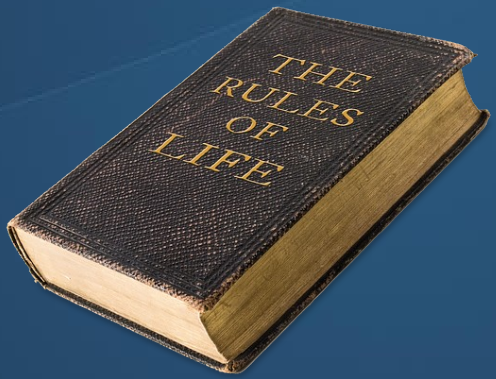
$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \ldots$$

$$G_t = R_{t+1} + \gamma \left(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \ldots \right)$$

$$\boxed{G_t = R_{t+1} + \gamma G_{t+1}}$$

But wait… how can we know future rewards?

The Policy

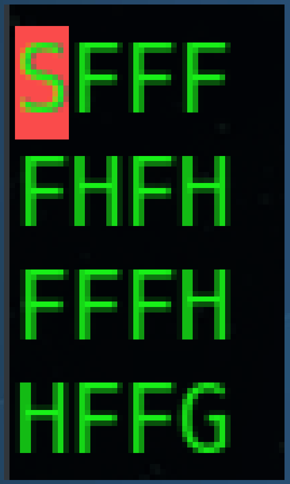


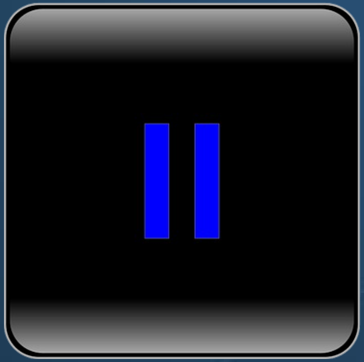
Mapping of states to actions

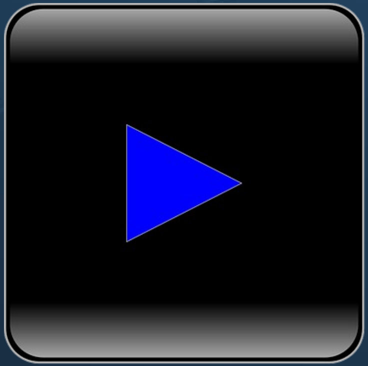
Can be probabilistic

Π

Next Exercise







- Frozen Lake environment
- Reasonable deterministic policy
- 1000 games
- Plot win % over trailing 10 games

Summary

- MDP determined by previous states and actions
- Governed by probability distribution
- Agent maximizes rewards over time
- Policy tells us how agent will act in some state

Up Next