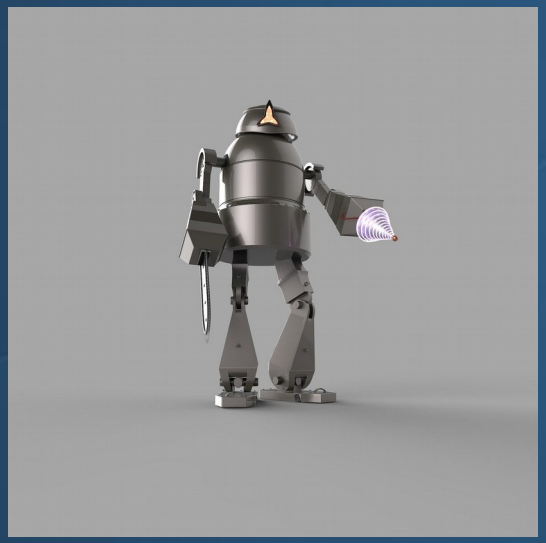
Actor Critic Methods: From Paper to Code

Monte Carlo Prediction Problem

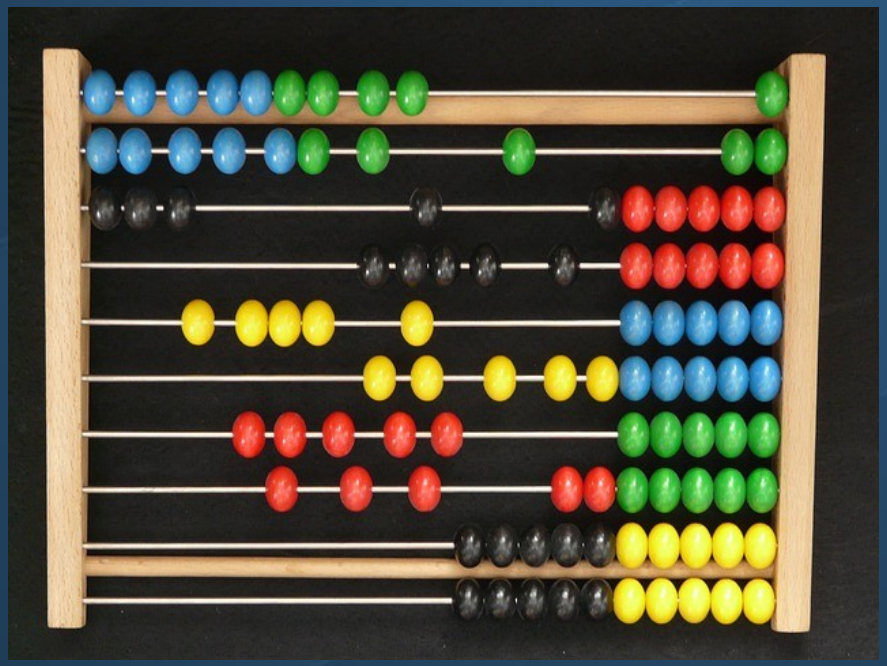
Monte Carlo (MC) Methods



Model free algorithms

Average received rewards

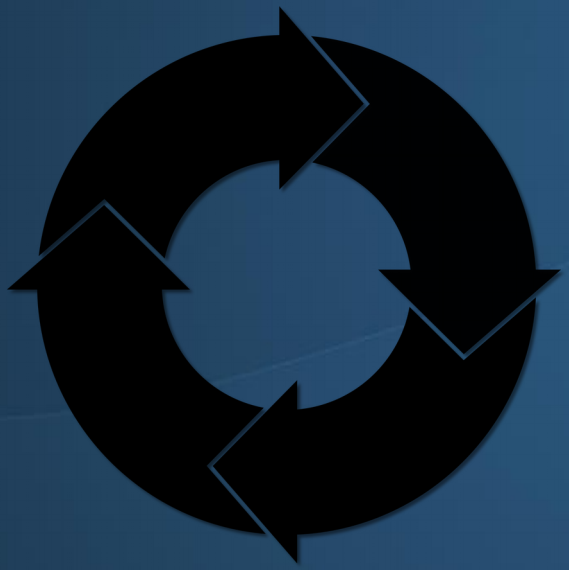
Prediction & Control Problems



Calculate v_π → prediction

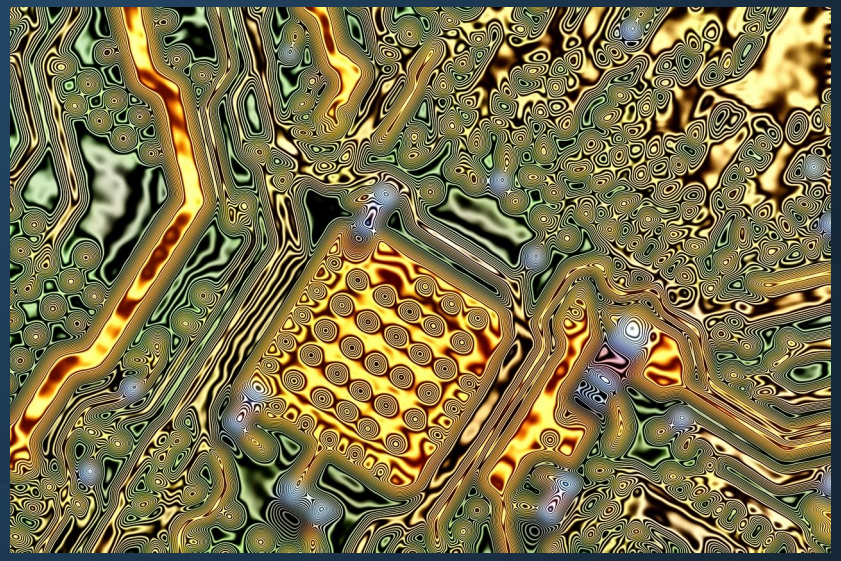
Improve π → control

Generalized Policy Iteration



Calculate v_π → make greedy → repeat

Can do in series or parallel



First vs. Every Visit MC

- Tracking rewards received after visiting states
- Rewards received after first visit → First visit MC
- Rewards received after every visit → Every visit MC

Blackjack Overview

- Player vs. Dealer; first to 21 wins, > 21 is a loss (bust)
- Ace worth 1 or 11, other face cards worth 10
  - Ace that doesn’t cause a bust is called usable
- One dealer card is showing
- Inifinte deck with replacement → no counting
- State space is a 3 tuple:
  - Player sum (4-21), dealer showing card (ace – 10), boolean for a usable ace
- Reward +1 for winning, 0 for draw, -1 for loss
- Policy: draw new card (hit) if player total < 20, else stick

Algorithm Overview

Initialize the policy to be evaluated

Initialize the value function arbitrarily

Initialize list of returns for all states in the state space

Repeat for large number of episodes:

- Generate episode using policy
- For each state s in the agent's memory:
  - Calculate the return that followed first visit to s
  - Append return G to list of returns
  - Calculate the average of the returns for state s

500,000 games → print value of state (21, 2, True)

Split into agent class and main function

Conclusion

- Use experience to estimate value of policy
- Iterate estimation and improvement
- Found value of policy with first visit MC prediction