Gambler’s problem

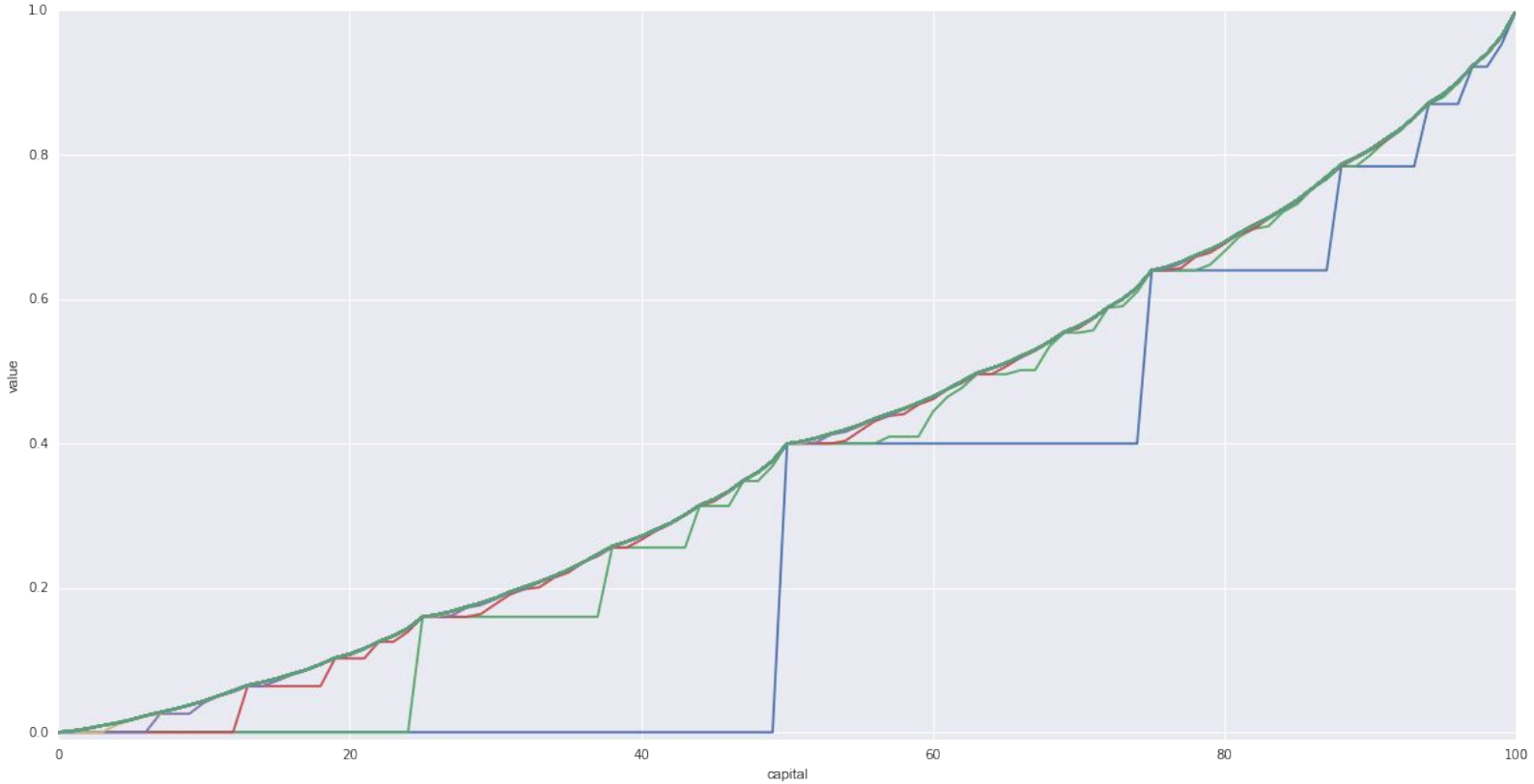
Reinforcement Learning - Discussion

Ashish Bora

Problem Description

- A gambler can bet integer dollars per timestep
- Wins amount equal to bet with probability 0.4
- Loses the bet with probability 0.6
- Game over if gambler runs out of money
- Gambler wins on reaching 100

Solution -- value iteration



Why monotonic?

- Intuition?

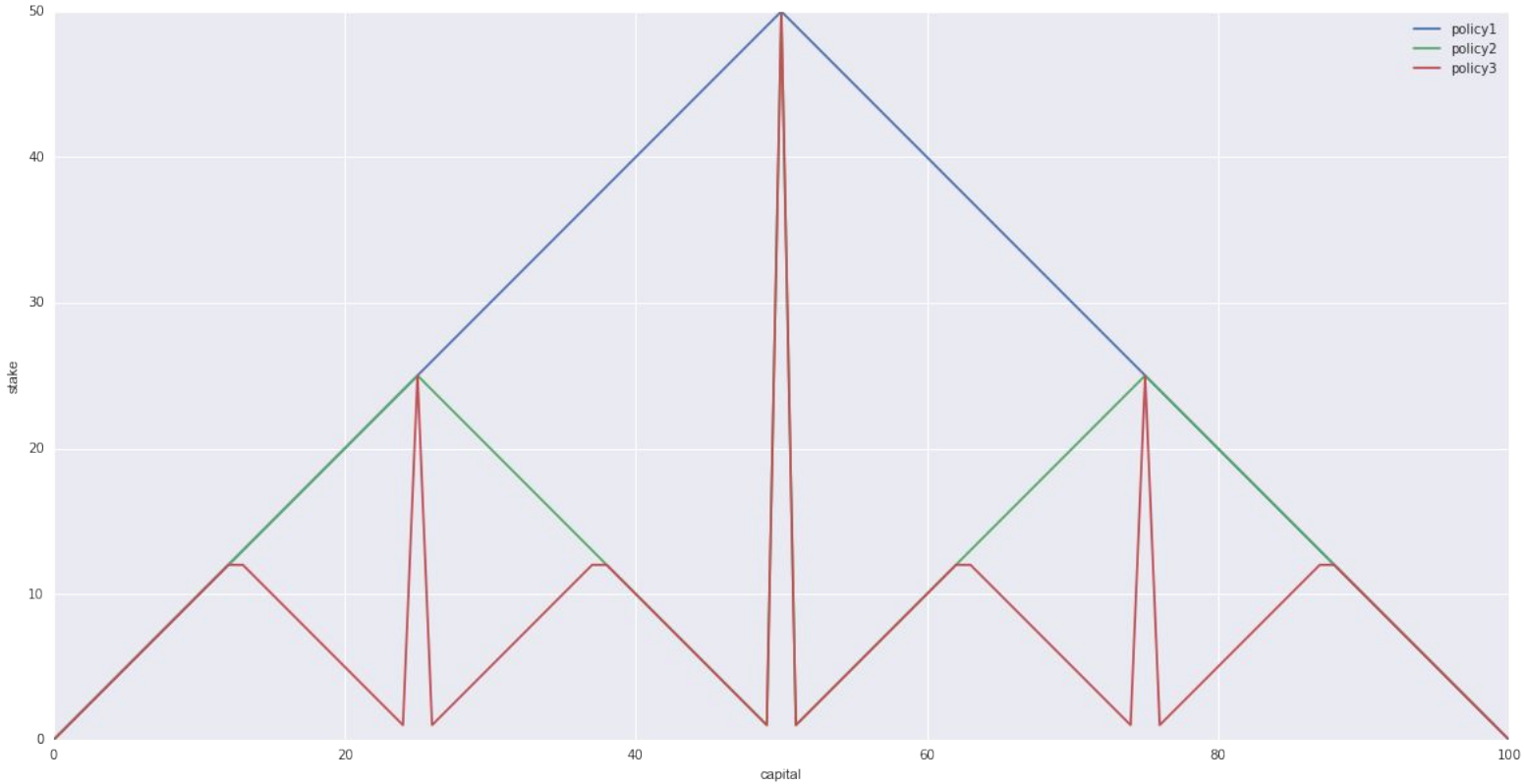
Why monotonic?

- Intuition?
- Proof?

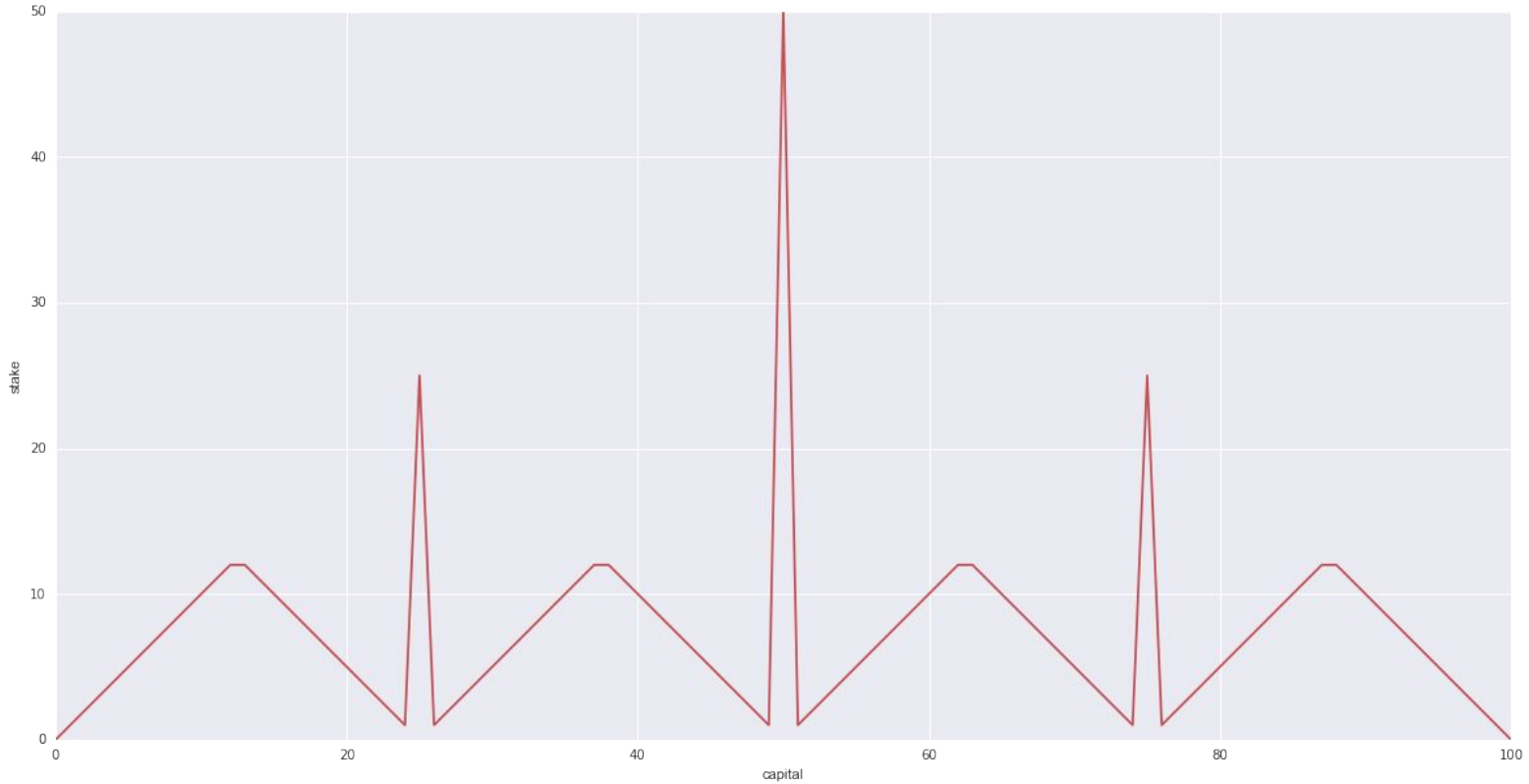
Why monotonic?

- Intuition?
- Proof?
  - Starting at captial x+1, with probability 1, we can either win or go to x.
  - If current capital is y, repeatedly bet min((y - x), y, 100 - y)
  - P(win | x) <= P(win | x+1)

Solution -- optimal policies



Why is the spiky graph optimal?



Why is the spiky graph optimal?

- Hypothesis?

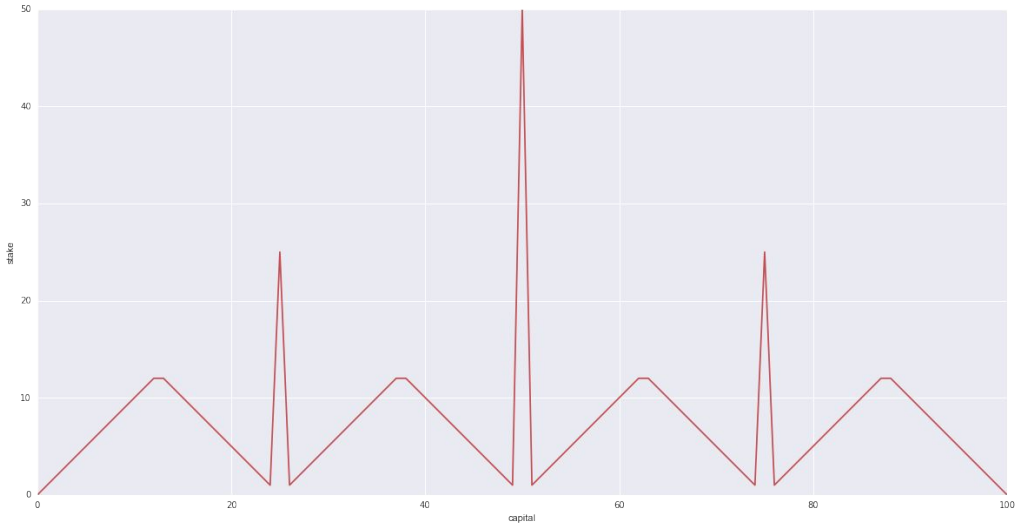
Why is the spiky graph optimal?

- Hypothesis?
  - The agent has set some goalposts and is trying to bet such that in worst case, it still is at the goalpost it just passed.
  - The goalposts are set according to number of hops needed to reach 100

Tests for this hypothesis?

Tests for this hypothesis? - one idea

Agent takes low risk actions to stay ahead of goalpost.

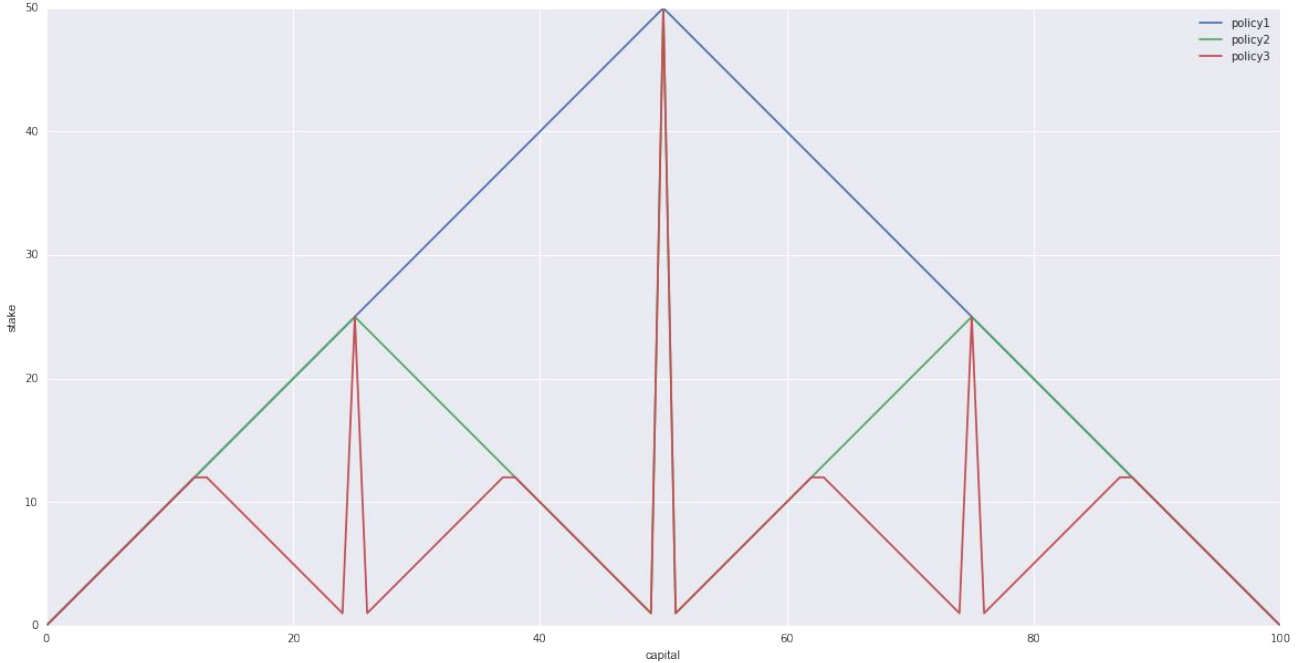


Can we use this?

Tests for this hypothesis? - one idea

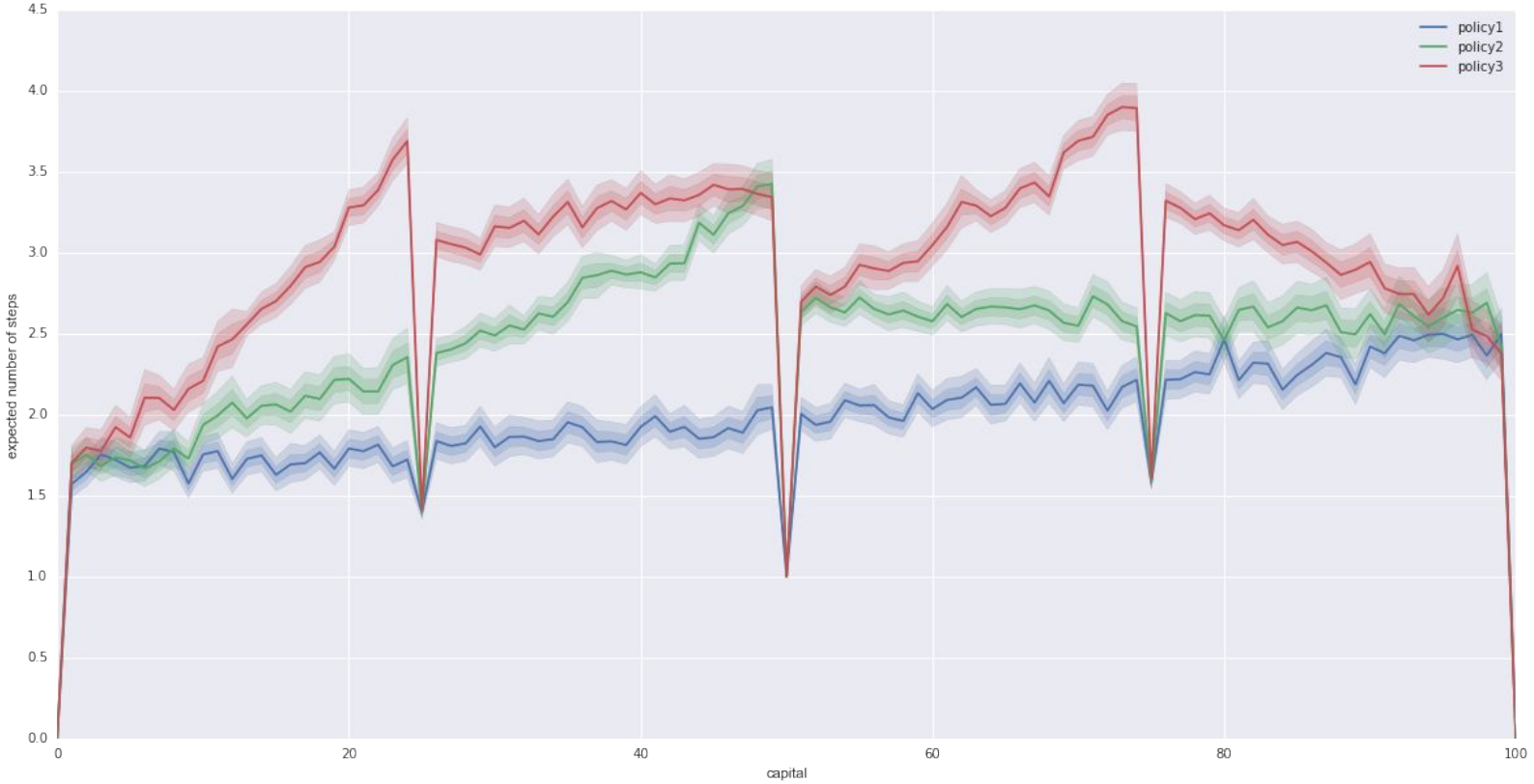
- Low risk (smaller bets) will lead to higher number of steps. Check of this is true
- Experiment
  - For each policy, for each starting point, estimate mean number of steps till end of episode

Optimal policies (reminder)



- Can you say something about values at 25, 50, and 75?
- Can you guess the full outcome?

Test for hypothesis -- Result (n = 500)

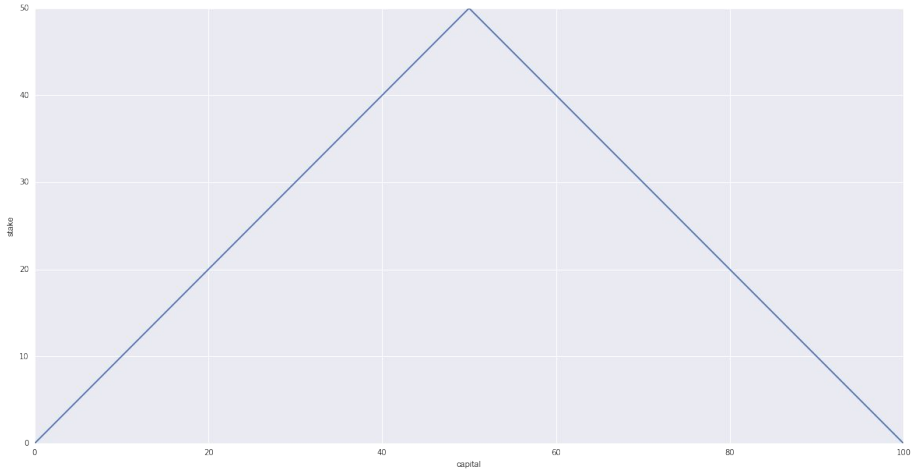


Another test -- ε discouragement

- Disincentivize more hops
- Give a small negative reward at each step
- Experiment
  - At each step, agent receives negative ε
  - If at 100, receives reward +1
- Can you guess how the following will look like for different values of ε?
  - value function
  - optimal policies
  - expected number of steps per episode for each policy

ε discouragement -- Results

- For small ε, value function remains almost same
- All spiky extrema destroyed(!), even for ε = 1e-10
- Envelope policy remains optimal



ε discouragement -- Moral?

- Some optimal policies may be destroyed by very small changes to reward function
- Analogy -- Maze problem
- Thoughts?