Graduate Artificial Intelligence
CS 640
Value and Policy Iteration

Jeffrey Considine

jconsidi@bu.edu

Plan for Today

- Policies
- Value iteration methods
- Policy iteration methods

Last Time

Re: specifying optimal values for Markov decision processes,

- What's the best possible expected value for this state?
- Specified recursively in terms of best possible expected values for other states.

Still open: how do we do this maximization?

What is a Policy?

A policy is a function mapping states to actions.

- A deterministic policy returns a single action.

- A probabilistic policy returns a probability distribution of actions.

Usually denoted as or with subscripts for context…

Probabilistic vs Deterministic Policies

Deterministic policy advantages

- Usually sufficient for optimal performance.
  - Exceptions with simultaneous adversarial choices.
- Much smaller to represent.
  - actions for states

Probabilistic policy advantages

- Can always represent deterministic policies.
  - All probability on one action.
- Often convenient for numerical optimizations.

Representing Policies for Finite MDPs

- Represent a deterministic policy as a table of actions.

- Represent a probabilistic policy as an matrix for states and actions.

Evaluating Policies for MDPs (take 1)

Previously,

For a specific policy (optimal or not),

Evaluating Policies for MDPs (take 2)

Rewrite

to

Evaluating Policies for MDPs (take 3a)

Rewrite

with

Evaluating Policies for MDPs (take 3b)

Rewrite

to

Rewriting with and yields a Markov reward process.

And we already know how to solve for their values.

What is an Optimal Policy?

So, what is ?

A policy is optimal if and only if

Optimal policies always exist. Is that surprising?

Any Questions?

???

Numerical Problems of Analytical Solution

Previously for Markov reward processes, we derived this solution.

- The interpretation of is the expected number of -discounted visits to each state.
- If , then at least one state has an infinite number of visits, and the matrix inversion will fail.
- This breaks the calculation even if those states have zero rewards.

Computational Cost of Analytical Solution

If , then what is the computational cost of the analytical solution?

???

Value Iteration

1. Set .
2. For i = 0,1, 2, …
    - For all states ,
        - Set

What is ?

How do we implement that expectation?

Value Iteration as Fixed Horizon Evaluations

Prove: For integer , is the best possible value achieved within steps.

- Base case: is the best possible value achieved within 0 steps.
  - 0 steps so 0 value.
- Inductive hypothesis: Assume that is the best possible value achieved within steps
- Induction step: Then is the best possible value achieved within steps.
  - This is true by construction.

Bounding Distance Values

How close is a given to ?

How much discounted reward is remaining after steps?

How to Act given (Near) Optimal Values?

For a given , pick the value maximizing action .

Any Questions?

???

Policy Iteration

- Policy iteration methods iterate on policies instead of value function.
- Similar intuitions but often converging in fewer iterations.

- Basic idea:
  - Start with any policy .
  - Calculate .
  - Calculate the best assuming will be used for one step and followed by using .

Naïve Implementations of Policy Iteration

1. Initialize to any policy.
2. For i = 0,1, 2, …
   a) Compute using value iteration.
   b) Set

Ignoring cost of 2a, why might this be better than value iteration?

Incremental Value Updates

1. is always a lower bound on since is achievable.
2. is achieved through specifically.
3. is no worse than .
4. is no worse than .
5. We can speedup computation of by initializing with .

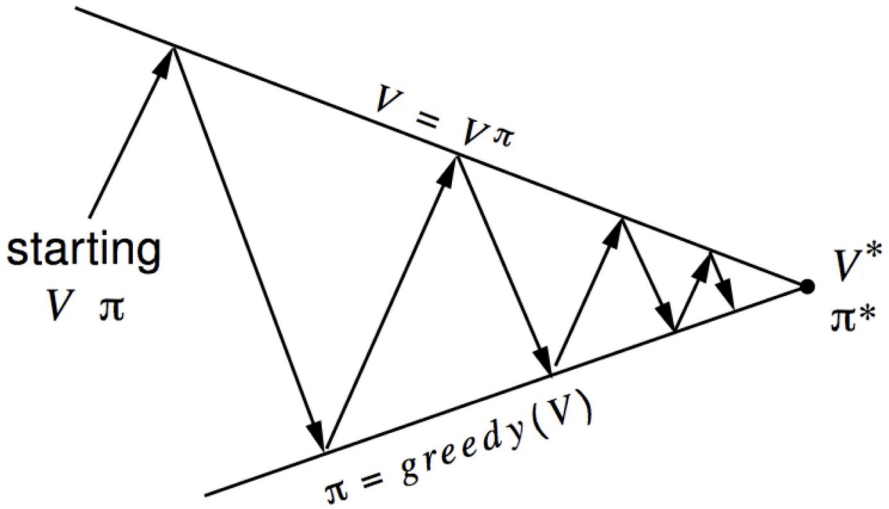
Policy Iteration with Warm Start Values

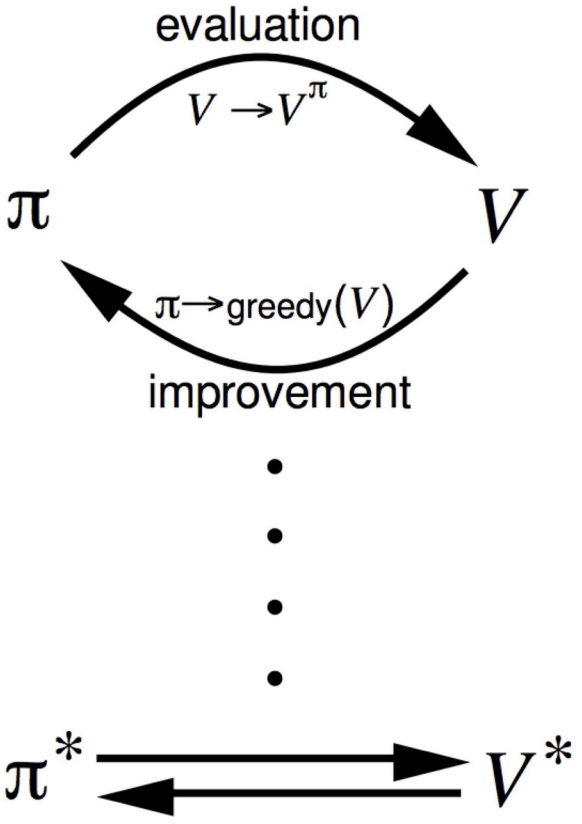
1. Initialize to any policy.
2. Set using value iteration.
3. For i = 0, 1, 2, …
    a) Set
    b) Compute starting from .

Interleaved Policy and Value Updates

1. Initialize to any policy.
2. Set using value iteration.
3. For i = 0, 1, 2, …
    a) Set
    b) Set

Convergence of Policy Iteration





Policy evaluation Estimate v_π

Iterative policy evaluation

Policy improvement Generate $\pi' \geq \pi$

Greedy policy improvement

igure by David Silver.

ttps://davidstarsilver.wordpress.com/wp-content/uploads/2025/04/lecture-3-planning-by-dynamic-programming-.

Any Questions?

???