Today

Policy Iteration

- Example
- Properties and Convergence

Policy Iteration

Policy iteration is an alternative to value iteration for solving infinite horizon MDPs

Basic Idea:

Step 1: *Choose an initial policy*

Step 2: *Evaluate the policy*

Step 3: *Find a better policy*

Step 4: *If the new policy is optimal then stop. Otherwise go to step 2.*

Policy Iteration

Algorithm (Policy Iteration):

1. Set $n = 0$ and select an arbitrary decision rule $d_0 \in D$.

2. ***Policy Evaluation***: *Obtain* v^n *by solving* $(I - \lambda P_{d_n})v = r_{d_n}$

3. ***Policy Improvement***: *Choose* d_{n+1} *to satisfy*:

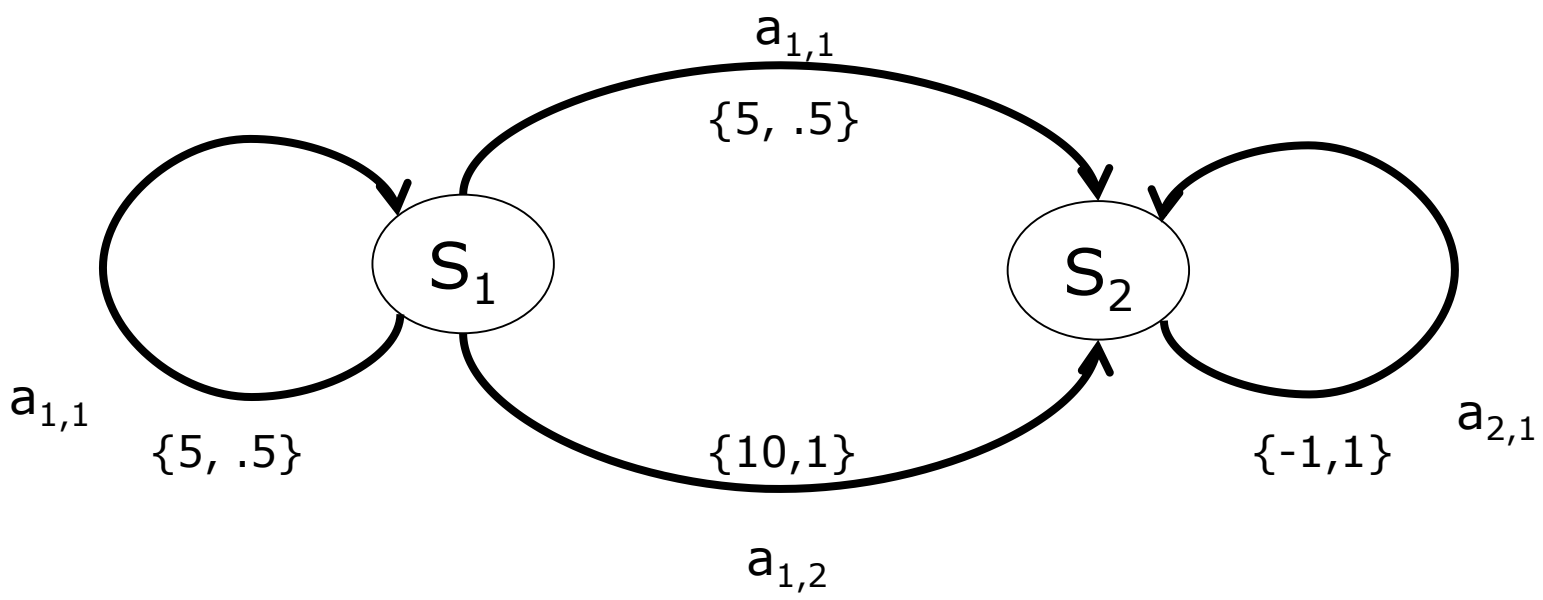
$$d_{n+1} \in argmax_{d \in D}\{r_d + \lambda P_d v^n\},$$

   setting $d_{n+1} = d_n$ if possible.

4. If $d_{n+1} = d_n$, stop and set $d^* = d_n$. Otherwise $n = n + 1$ and return to step 2.

Example Revisited: 2 State MDP

In state S_1 actions $a_{1,1}$ and $a_{1,2}$ are available; in state S_2 only $a_{2,1}$ is available. Rewards and transition probabilities are defined below as {r,p}



Exercise: Use **policy iteration** to find an optimal policy. Assume λ = 0.95.

Policy Iteration – Why does it work?

When there are a finite number of policies **policy iteration** generates a finite sequence of decision rules $\{d_n\}$, and value functions $\{v_n\}$.

Proposition (6.4.1 Puterman): Let v^n and v^{n+1} be successive values generated by the policy iteration algorithm. Then $v^{n+1} \geq v^n$.

Proof: Completed in class

Policy Iteration - Convergence

The policy iteration algorithm converges finitely to an optimal solution (assuming a finite set of states and actions)

Theorem (6.4.2 Puterman): The policy iteration algorithm terminates in a finite number of iterations, with a solution of the optimality equations and an optimal policy, d^*.

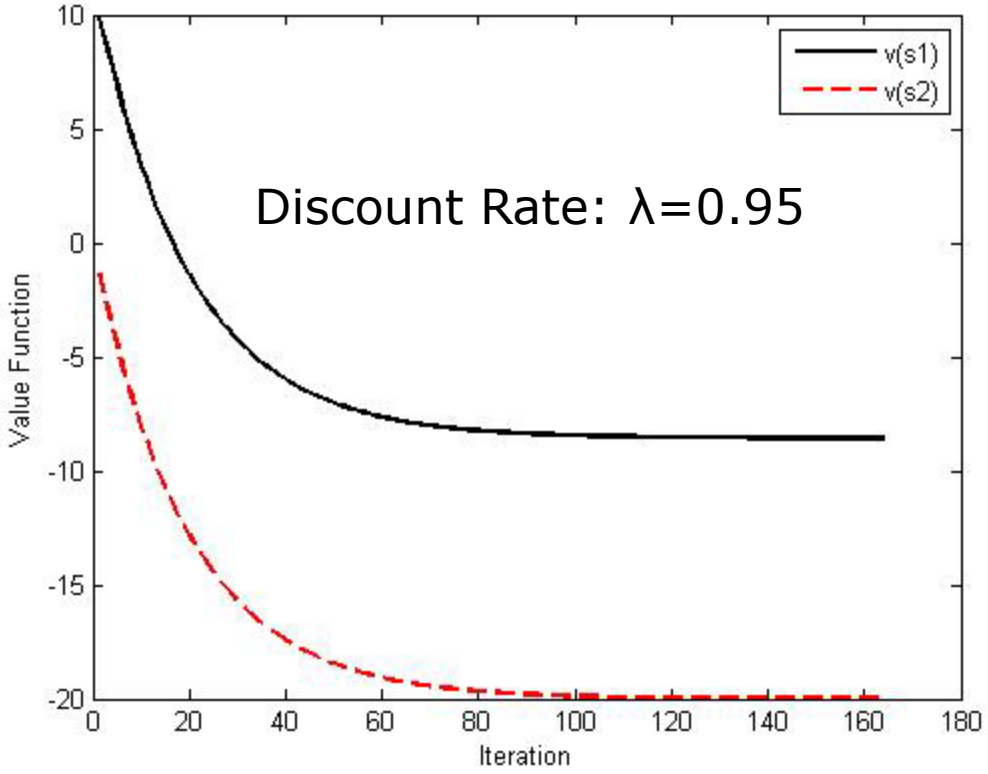
Proof: Completed in class

Restatement of this theorem in words:

- If an improvement to a policy is possible then policy iteration will find it
- If no improvement is possible then the algorithm terminates an optimal solution to $v = Lv$

Recall: Value Iteration From Last Class

Recall that it took 162 iterations of value iteration to converge to a solution within 0.01 of optimal on the 2-state example



Policy iteration found the (provably optimal) solution in 2 iterations.

Value vs Policy Iteration

- Value iteration generates ϵ- optimal solutions; Policy iteration is exact.

- Policy iteration solves a system of linear equations at each iteration.

- There are several variants of value and policy iteration that can accelerate convergence:
  - Gauss-Seidel Value Iteration
  - Modified Policy Iteration

- Modifications to value iteration can be incorporated into the policy evaluation step of policy iteration

For Next Class

Before next class look at the following article which illustrates an application of infinite horizon dynamic programming:

- Alagoz, O., Maillart, L., Schaefer, A., Roberts, M., 2004, The Optimal Timing of Living-Donor Liver Transplant, Operations Research, 50(1), 1420-1430

There will be a short quiz on the article at the start of next class