Aalto University
School of Science

Policy iteration method for solving Markov decision processes

Einari Tuukkanen
Presentation *13*
06.11.2020

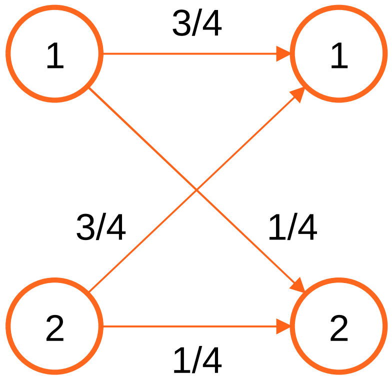
MS-E2191 Graduate Seminar on Operations Research
Fall 2020

In this presentation...

- Quick recap of Markov decision process (MDP) and value iteration (VI)
- Introducing policy iteration (PI) with an example
- Pros and cons of PI
- PI improvements
- References
- Homework

Markov decision process & value iteration

- Process in state *i*
- States $s \in \boldsymbol{S}$
- Actions $a \in \boldsymbol{A}$
- Cost (or reward) **R(**s, a, s’**)**
- Transition probabilities **P(**s, a, s’**)**



Value Iteration method

$$V_{i+1}(s) = \max_{a \in A} \sum_{s' \in S} P(s,a,s')[R(s,a,s') + \gamma V_i(s')]$$

Policy iteration method

Step 1: Initialization

Step 2: Policy evaluation

Step 3: Policy improvement

MDP – Example from presentation 5

- Actions have desired outcome with **P=0.7**, discount factor **ɣ=0.5**
- All other transitions equally likely
- Actions: **N**(orth), **E**(east), **S**(outh), **W**(est), States **S**: 1, 2, 3, 4, 5, 6, 7, 8, 9
- Objective to maximize reward

States

7	8	9
4	5	6
1	2	3

Rewards

0	0	100
0	0	0
0	0	0

Policy μ^0

↑	↑	↑
↑	↑	↑
↑	↑	↑

MDP – Modified example

- Option to stay in place with certainty
- If not staying in place, always move to a neighbouring state
- Reward gained from transition to **S=9** regardless of the action chosen

- Valid states **S**: 5, 8, 9
- Other states (and outside of map) have value 0 and policy to always stay put
- E.g. **S=5, A=N**: P(8)=0.7 (ignoring obstacles)
- E.g. **S=8, A=S**: P(5)=0.7, P(9)=0.1

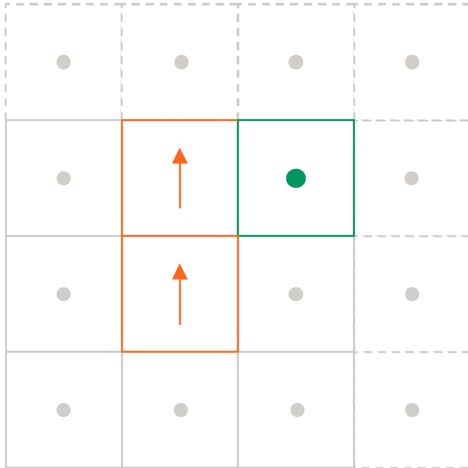
States

7	8	9
4	5	6
1	2	3

Rewards

0	0	100
0	0	0
0	0	0

Policy μ^0



Policy iteration steps

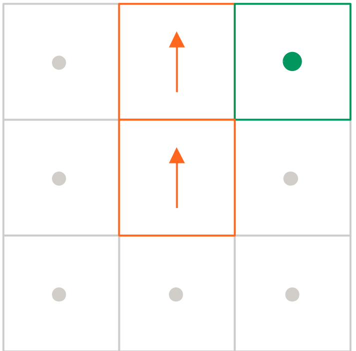
Step 1: Initialization

Guess an initial stationary policy μ^0

Step 2: Policy evaluation

Step 3: Policy improvement

Example policy μ^0



Step 2: Policy evaluation

(BAD) EXAMPLE

Path	Prob.	Utility
↑	0.7	0
→	0.1	0
←	0.1	0
↓	0.1	0

Path	Prob.	Utility
↑ →	0.07	50
↑ ←	0.07	0
↑ ↑	0.49	0
...	...	...

Naïve method:

$$V_\mu(5) \approx 0.07 \cdot 50 = 3.5$$

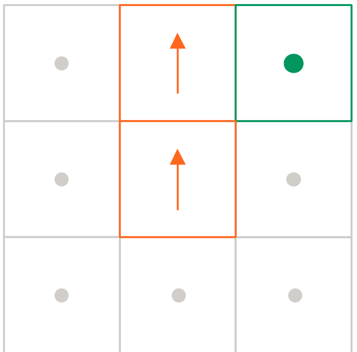
States

7	8	9
4	5	6
1	2	3

Rewards

0	0	100
0	0	0
0	0	0

Policy μ^0



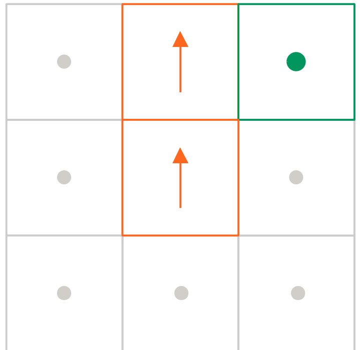
Step 2: Policy evaluation

EXAMPLE

$$V_\mu(s) = \sum_{s' \in S} P(s, \mu(s), s')\left[R(s, \mu(s), s') + \gamma V_\mu(s')\right], \forall s \in S$$

on this slide notate $v_s = V_{\mu^0}(s)$

Policy μ^0



Consider values at states 5, 8 and 9

$\mathbf{v_5}$ = 0.7 * (0 + 0.5 * $\mathbf{v_8}$) + 0.3 * (0 + 0.5 * 0)

$\mathbf{v_8}$ = 0.1 * (100 + 0.5 * $\mathbf{v_9}$) + 0.1 * (0 + 0.5 * $\mathbf{v_5}$) + 0.8 * (0 + 0.5 * 0)

$\mathbf{v_9}$ = 1 * (100 + 0.5 * $\mathbf{v_9}$)

Solve the linear system

$\mathbf{v_9}$ = 200

$\mathbf{v_8}$ = 10 + 0.05 * 200 + 0.05 * 0.35 * $\mathbf{v_8}$ ⇔ $\mathbf{v_8}$ = 8000/393 = 20.356...

$\mathbf{v_5}$ = 0.7 * 0.5 * 8000/393 = 2800/393 = 7.124...

Policy Iteration steps

Step 1: Initialization

Guess an initial stationary policy μ^0

Step 2: Policy evaluation

$$V_\mu(s) = \sum_{s' \in S} P(s, \mu(s), s')\left[R(s, \mu(s), s') + \gamma V_\mu(s')\right], \forall s \in S$$

Step 3: Policy improvement

Step 3: Policy improvement

- We can always find at least equally good policy
- Roll-out policy
  - For each state, choose the maximizing action and assume current policy elsewhere
- Finite number of states and actions

➔ Eventually terminates with an optimal policy

Step 3: Policy improvement

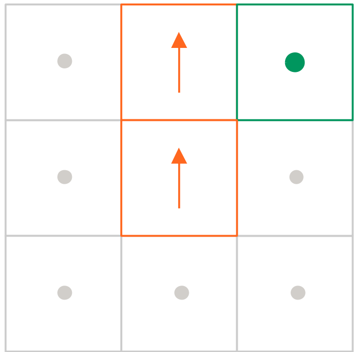
EXAMPLE

$$\mu^1(s) = \arg\max_{a \in A} \sum_{s' \in S} P(s,a,s')\left[R(s,a,s') + \gamma V_{\mu^0}(s')\right], \forall s \in S$$

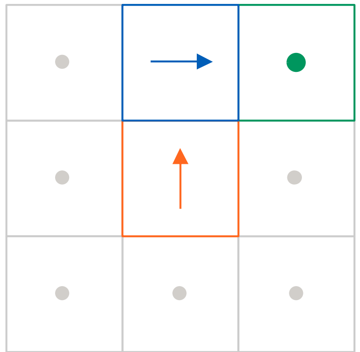
Remember the solution from the last step

$v_9 = 200$, $v_8 \approx 20.35$, $v_5 \approx 7.12$

Baseline policy μ^0



Improved policy μ^1



$\mu^1(5) = \arg\max$
- ↑ : **7.12**
- → : $0.1 \cdot 0.5 \cdot 20.35 \approx 1.01$
- ↓ : $0.1 \cdot 0.5 \cdot 20.35 \approx 1.01$
- ← : $0.1 \cdot 0.5 \cdot 20.35 \approx 1.01$

$\mu^1(8) = \arg\max$
- ↑ : 20.35
- → : **$0.7\ (100 + 0.5 \cdot 200) + 0.1 \cdot 0.5 \cdot 7.124 \approx 140.35$**
- ↓ : $0.7 \cdot 0.5 \cdot 7.124 + 0.1 \cdot (100 + 0.5 \cdot 200) \approx 22.49$
- ← : $0.1 \cdot 0.5 \cdot 7.124 + 0.1 \cdot (100 + 0.5 \cdot 200) \approx 20.35$

Policy iteration steps

Step 1: Initialization

Guess an initial stationary policy μ^0

Step 2: Policy evaluation

$V_\mu(s) = \sum_{s' \in S} P(s, \mu(s), s')\left[R(s, \mu(s), s') + \gamma V_\mu(s')\right], \; \forall s \in S$

Step 3: Policy improvement

$\mu^{k+1}(s) = \arg\max_{a \in A} \sum_{s' \in S} P(s, a, s')\left[R(s, a, s') + \gamma V_{\mu^k}(s')\right], \; \forall s \in S,$

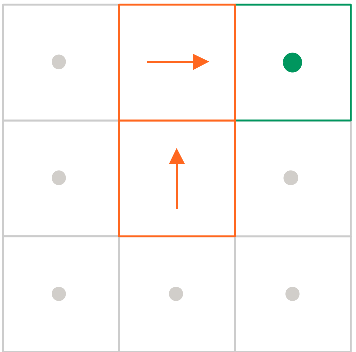
repeat steps 2 and 3 until μ is unchanged

Iter 2, step 2: Policy evaluation
EXAMPLE

$$V_\mu(s) = \sum_{s' \in S} P(s, \mu(s), s')\left[R(s, \mu(s), s') + \gamma V_\mu(s')\right], \forall s \in S$$

on this slide notate $v_s = V_{\mu^1}(s)$

Current policy μ^1



Consider values at states 5, 8 and 9

v$_5$ = 0.7 * (0 + 0.5 * **v**$_8$) + 0.3 * (0 + 0.5 * 0)

v$_8$ = **0.7** * (100 + 0.5 * **v**$_9$) + 0.1 * (0 + 0.5 * **v**$_5$) + 0.2 * (0 + 0.5 * 0)

v$_9$ = 1 * (100 + 0.5 * **v**$_9$)

Solve the linear system

v$_9$ = 200

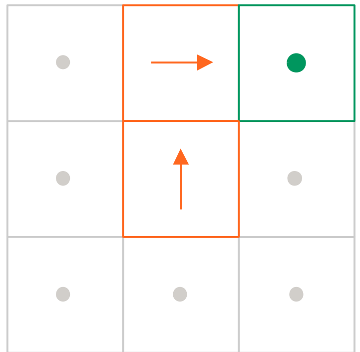
v$_8$ = 56000/393 = **142.493**...

v$_5$ = 19600/393 = **49.872**...

Iter 2, step 3: Policy improvement

EXAMPLE

Current policy μ^1



$$\mu^2(s) = \arg\max_{a \in A} \sum_{s' \in S} P(s,a,s')\left[R(s,a,s') + \gamma V_{\mu^1}(s')\right],\ \forall s \in S$$

Remember the solution from the last step
$v_9 = 200$, $v_8 \approx 142.49$, $v_5 \approx 49.87$

The policy does not change
➔ We have reached the optimal policy

$\mu^2(5) = \arg\max$
- ↑ : **49.87**
- → : $0.1 \cdot 0.5 \cdot 142.49 \approx 7.12$
- ↓ : $0.1 \cdot 0.5 \cdot 142.49 \approx 7.12$
- ← : $0.1 \cdot 0.5 \cdot 142.49 \approx 7.12$

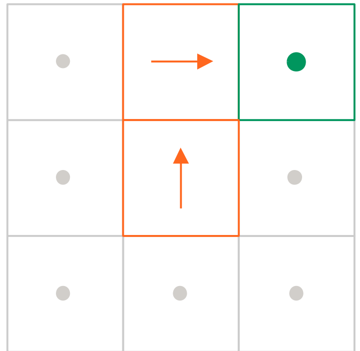
$\mu^2(8) = \arg\max$
- ↑ : $0.1 \cdot (100 + 0.5 \cdot 200) + 0.1 \cdot 0.5 * 49.87 \approx 22.49$
- → : **142.49**
- ↓ : $0.7 \cdot 0.5 \cdot 49.87 + 0.1 \cdot (100 + 0.5 \cdot 200) \approx 37.45$
- ← : $0.1 \cdot 0.5 \cdot 49.87 + 0.1 \cdot (100 + 0.5 \cdot 200) \approx 22.49$

Iter 2, step 3: Policy improvement

EXAMPLE

$$\mu^2(s) = \arg\max_{a \in A} \sum_{s' \in S} P(s,a,s')[R(s,a,s') + \gamma V_{\mu^1}(s')], \forall s \in S$$

Current policy μ^1



Remember the solution from the last step
$v_9 = 200$, $v_8 \approx 142.49$, $v_5 \approx 49.87$

The policy does not change
➔ We have reached the optimal policy

Optimal policy

$\mu^*(5)$ = ”North”, $\mu^*(8)$ = ”East”

Optimal values

$V^*(5) \approx 49.87$, $V^*(8) \approx 142.49$,
$V^*(9) = 200$

Aalto University
School of Science

Pros and cons

Pros

- Finite-time convergence to the optimal policy
- Typically terminates (or gets close to optimal) in remarkably few iterations

Cons

- Possibly requires solving of large linear systems

→ Poor performance when number of states is high

On each iteration of PI

- $card(S)$ linear equations
- $card(S)$ unknowns
- $O(card(S)^3)$ solution

Iteration of VI only $O(card(S) \cdot card(A))$

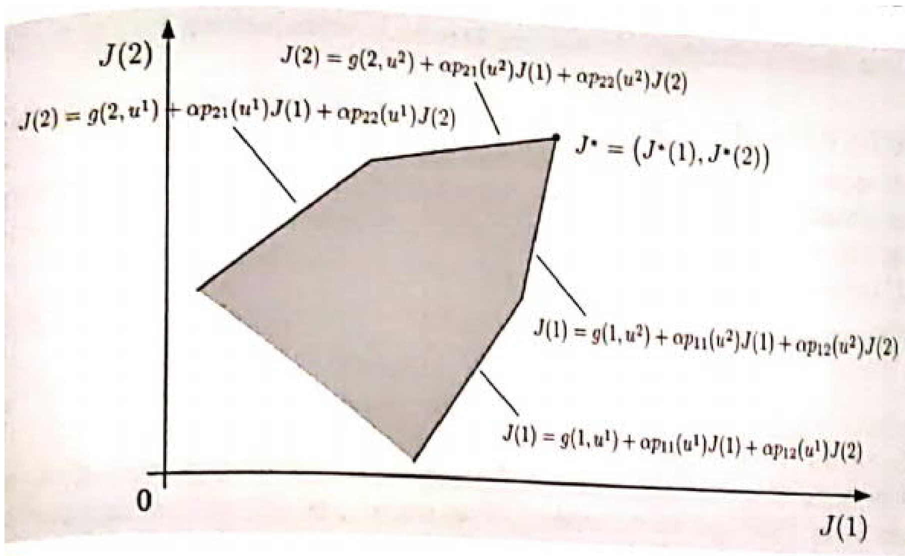
Complexity reference: 10 Lecture 23: Markov Decision Processes Policy Iteration

Improving PI method

- Optimistic policy iteration
  - In evaluation step, solve the equation system (approximately) using VI
- Linear programming methods

Linear programming methods in PI

- Aims directly for an optimal policy
- To find out optimal $V^*(1), \dots, V^*(n)$ solve the following problem in $z_1, \dots, z_n$



$$\max \sum_{s \in S} z_s$$

s.t. $z_s \leq \sum_{s' \in S} P(s,a,s')[R(s,a,s') + \gamma V(s')], \forall s \in S, a \in A(s)$

Aalto University
School of Science

References

Bertsekas, D. P. (2012). Dynamic programming and optimal control (Vol. 2, 4th ed.) Approximate Dynamic Programming. Belmont, MA: Athena scientific.

Howard, R. A. (1960). Dynamic programming and markov processes. John Wiley & Sons

https://ocw.mit.edu/courses/aeronautics-and-astronautics/16-410-principles-of-autonomy-and-decision-making-fall-2010/lecture-notes/MIT16_410F10_lec23.pdf, 5.11.2020

Homework

Consider the following problem

$S = \{1, 2\}, \qquad A = \{a_1, a_2\}$

$$P(s, a_1, s') = \begin{pmatrix} p_{11}(a_1), p_{12}(a_1) \\ p_{21}(a_1), p_{22}(a_1) \end{pmatrix} = \begin{pmatrix} 3/4, 1/4 \\ 3/4, 1/4 \end{pmatrix}$$

$$P(s, a_2, s') = \begin{pmatrix} p_{11}(a_2), p_{12}(a_2) \\ p_{21}(a_2), p_{22}(a_2) \end{pmatrix} = \begin{pmatrix} 1/4, 3/4 \\ 1/4, 3/4 \end{pmatrix}$$

Discount factor $\gamma = 0.9$

Baseline policy $\mu^0(1) = a_1, \ \ \mu^0(2) = a_2$

Transition costs $g(s, a)$		
States / Actions	**s=1**	**s=2**
a=a$_1$	2	1
a=a$_2$	0.5	3

E.g. $g(2, a_1) = 1$

Homework

Find the **minimizing** optimal policy and cost for the problem. Report the optimal actions $\mu^*(s)$ and values $V^*(s)$ in each state.

DL: 13.11. 9:00, einari.tuukkanen@aalto.fi