

# Lecture 3: Planning by Dynamic Programming

David Silver

# Outline

- 1 Introduction
- 2 Policy Evaluation
- 3 Policy Iteration
- 4 Value Iteration
- 5 Extensions to Dynamic Programming
- 6 Contraction Mapping

# What is Dynamic Programming?

Dynamic **sequential or temporal component** to the problem  
Programming optimising a “program”, i.e. a policy

- c.f. linear programming
- A method for solving complex problems
- By breaking them down into subproblems
  - Solve the subproblems
  - Combine solutions to subproblems

# Requirements for Dynamic Programming

Dynamic Programming is a very general solution method for problems which **have two properties:**

- Optimal substructure
  - *Principle of optimality* applies
  - Optimal solution can be decomposed into subproblems
- **Overlapping subproblems**
  - Subproblems recur many times
  - Solutions can be **cached and reused**
- Markov decision processes satisfy both properties
  - Bellman equation gives **recursive decomposition**
  - Value function stores and **reuses solutions**

# Planning by Dynamic Programming

- Dynamic programming assumes full knowledge of the MDP
- It is used for *planning* in an MDP
- **For prediction:**
  - Input: MDP  $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$  and policy  $\pi$
  - or: MRP  $\langle \mathcal{S}, \mathcal{P}^\pi, \mathcal{R}^\pi, \gamma \rangle$
  - Output: value function  $v_\pi$
- **Or for control:**
  - Input: MDP  $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$
  - Output: optimal value function  $v_*$
  - and: optimal policy  $\pi_*$

## Other Applications of Dynamic Programming

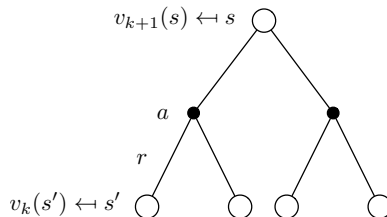
Dynamic programming is used to solve many other problems, e.g.

- Scheduling algorithms
- String algorithms (e.g. sequence alignment)
- Graph algorithms (e.g. shortest path algorithms)
- Graphical models (e.g. Viterbi algorithm)
- Bioinformatics (e.g. lattice models)

# Iterative Policy Evaluation

- Problem: evaluate a given policy  $\pi$
- Solution: iterative application of Bellman expectation backup
- $v_1 \rightarrow v_2 \rightarrow \dots \rightarrow v_\pi$
- Using *synchronous* backups,
  - At each iteration  $k + 1$
  - For all states  $s \in \mathcal{S}$
  - Update  $v_{k+1}(s)$  from  $v_k(s')$
  - where  $s'$  is a successor state of  $s$
- We will discuss *asynchronous* backups later
- Convergence to  $v_\pi$  will be proven at the end of the lecture

# Iterative Policy Evaluation (2)



$$v_{k+1}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left( \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_k(s') \right)$$

$$\mathbf{v}^{k+1} = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi \mathbf{v}^k$$



# Evaluating a Random Policy in the Small Gridworld



actions

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	

$r = -1$   
on all transitions

- Undiscounted episodic MDP ( $\gamma = 1$ )
- Nonterminal states 1, ..., 14
- One terminal state (shown twice as shaded squares)
- Actions leading out of the grid leave state unchanged
- Reward is  $-1$  until the terminal state is reached
- Agent follows uniform random policy

$$\pi(n|\cdot) = \pi(e|\cdot) = \pi(s|\cdot) = \pi(w|\cdot) = 0.25$$

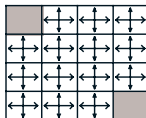
# Iterative Policy Evaluation in Small Gridworld

$v_k$  for the  
Random Policy

$k = 0$

0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0

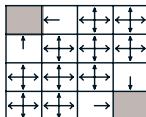
Greedy Policy  
w.r.t.  $v_k$



← random  
policy

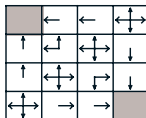
$k = 1$

0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0



$k = 2$

0.0	-1.7	-2.0	-2.0
-1.7	-2.0	-2.0	-2.0
-2.0	-2.0	-2.0	-1.7
-2.0	-2.0	-1.7	0.0





# How to Improve a Policy

- Given a policy  $\pi$ 
  - **Evaluate** the policy  $\pi$

$$v_{\pi}(s) = \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \dots | S_t = s]$$

- **Improve** the policy by acting greedily with respect to  $v_{\pi}$

$$\pi' = \text{greedy}(v_{\pi})$$

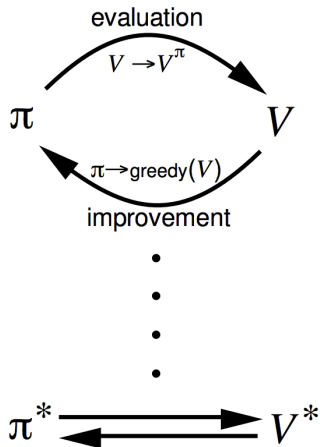
- In Small Gridworld improved policy was optimal,  $\pi' = \pi^*$
- In general, need more iterations of improvement / evaluation
- But this process of **policy iteration** always converges to  $\pi^*$

# Policy Iteration



**Policy evaluation** Estimate  $v_\pi$   
Iterative policy evaluation

**Policy improvement** Generate  $\pi' \geq \pi$   
Greedy policy improvement

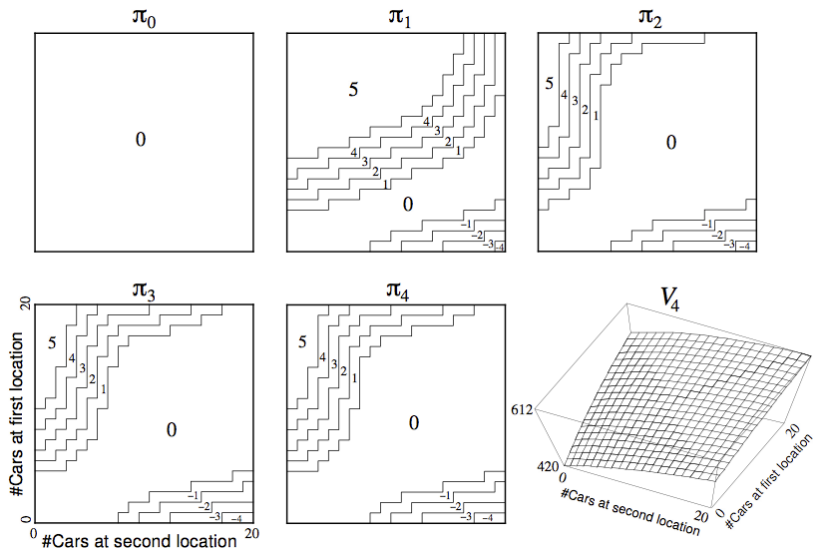


# Jack's Car Rental



- States: Two locations, maximum of 20 cars at each
- Actions: Move up to 5 cars between locations overnight
- Reward: \$10 for each car rented (must be available)
- Transitions: Cars returned and requested randomly
  - Poisson distribution,  $n$  returns/requests with prob  $\frac{\lambda^n}{n!} e^{-\lambda}$
  - 1st location: average requests = 3, average returns = 3
  - 2nd location: average requests = 4, average returns = 2

# Policy Iteration in Jack's Car Rental



# Policy Improvement

- Consider a deterministic policy,  $a = \pi(s)$
- We can *improve* the policy by acting greedily

$$\pi'(s) = \operatorname{argmax}_{a \in \mathcal{A}} q_{\pi}(s, a)$$

- This improves the value from any state  $s$  over one step,

$$q_{\pi}(s, \pi'(s)) = \max_{a \in \mathcal{A}} q_{\pi}(s, a) \geq q_{\pi}(s, \pi(s)) = v_{\pi}(s)$$

- It therefore improves the value function,  $v_{\pi'}(s) \geq v_{\pi}(s)$

$$\begin{aligned} v_{\pi}(s) &\leq q_{\pi}(s, \pi'(s)) = \mathbb{E}_{\pi'} [R_{t+1} + \gamma v_{\pi}(S_{t+1}) \mid S_t = s] \\ &\leq \mathbb{E}_{\pi'} [R_{t+1} + \gamma q_{\pi}(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s] \\ &\leq \mathbb{E}_{\pi'} [R_{t+1} + \gamma R_{t+2} + \gamma^2 q_{\pi}(S_{t+2}, \pi'(S_{t+2})) \mid S_t = s] \\ &\leq \mathbb{E}_{\pi'} [R_{t+1} + \gamma R_{t+2} + \dots \mid S_t = s] = v_{\pi'}(s) \end{aligned}$$



## Policy Improvement (2)

- If improvements stop,

$$q_{\pi}(s, \pi'(s)) = \max_{a \in \mathcal{A}} q_{\pi}(s, a) = q_{\pi}(s, \pi(s)) = v_{\pi}(s)$$

- Then the Bellman optimality equation has been satisfied

$$v_{\pi}(s) = \max_{a \in \mathcal{A}} q_{\pi}(s, a)$$

- Therefore  $v_{\pi}(s) = v_{*}(s)$  for all  $s \in \mathcal{S}$
- so  $\pi$  is an optimal policy

# Modified Policy Iteration

- Does policy evaluation need to converge to  $v_\pi$ ?
- Or should we introduce a stopping condition
  - e.g.  $\epsilon$ -convergence of value function
- Or simply stop after  $k$  iterations of iterative policy evaluation?
- For example, in the small gridworld  $k = 3$  was sufficient to achieve optimal policy
- Why not update policy every iteration? i.e. stop after  $k = 1$ 
  - This is equivalent to *value iteration* (next section)

# Generalised Policy Iteration

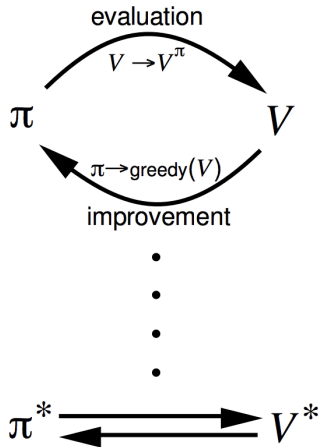


**Policy evaluation** Estimate  $v_\pi$

**Any** policy evaluation algorithm

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

**Any** policy improvement algorithm



# Principle of Optimality

Any optimal policy can be subdivided into two components:

- An optimal first action  $A_*$
- Followed by an optimal policy from successor state  $S'$

## Theorem (Principle of Optimality)

*A policy  $\pi(a|s)$  achieves the optimal value from state  $s$ ,  $v_\pi(s) = v_*(s)$ , if and only if*

- *For any state  $s'$  reachable from  $s$*
- *$\pi$  achieves the optimal value from state  $s'$ ,  $v_\pi(s') = v_*(s')$*

# Deterministic Value Iteration

- If we know the solution to subproblems  $v_*(s')$
- Then solution  $v_*(s)$  can be found by one-step lookahead

$$v_*(s) \leftarrow \max_{a \in \mathcal{A}} \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_*(s')$$

- The idea of value iteration is to apply these updates iteratively
- Intuition: start with final rewards and work backwards
- Still works with **loopy**, stochastic MDPs

# Example: Shortest Path

g			

Problem

0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

 $V_1$ 

0	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	-1

 $V_2$ 

0	-1	-2	-2
-1	-2	-2	-2
-2	-2	-2	-2
-2	-2	-2	-2

 $V_3$ 

0	-1	-2	-3
-1	-2	-3	-3
-2	-3	-3	-3
-3	-3	-3	-3

 $V_4$ 

0	-1	-2	-3
-1	-2	-3	-4
-2	-3	-4	-4
-3	-4	-4	-4

 $V_5$ 

0	-1	-2	-3
-1	-2	-3	-4
-2	-3	-4	-5
-3	-4	-5	-5

 $V_6$ 

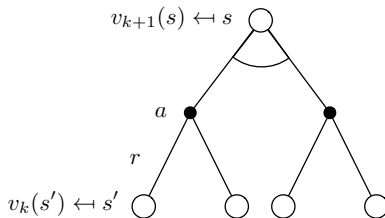
0	-1	-2	-3
-1	-2	-3	-4
-2	-3	-4	-5
-3	-4	-5	-6

 $V_7$

# Value Iteration

- Problem: find optimal policy  $\pi$
- Solution: iterative application of Bellman optimality backup
- $v_1 \rightarrow v_2 \rightarrow \dots \rightarrow v_*$
- Using synchronous backups
  - At each iteration  $k + 1$
  - For all states  $s \in \mathcal{S}$
  - Update  $v_{k+1}(s)$  from  $v_k(s')$
- Convergence to  $v_*$  will be proven later
- Unlike policy iteration, there is no explicit policy
- Intermediate value functions may not correspond to any policy

# Value Iteration (2)



$$v_{k+1}(s) = \max_{a \in \mathcal{A}} \left( \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_k(s') \right)$$

$$\mathbf{v}_{k+1} = \max_{a \in \mathcal{A}} \mathbf{R}^a + \gamma \mathbf{P}^a \mathbf{v}_k$$



## Example of Value Iteration in Practice

<http://www.cs.ubc.ca/~poole/demos/mdp/vi.html>

# Synchronous Dynamic Programming Algorithms

Problem	Bellman Equation	Algorithm
Prediction	Bellman Expectation Equation	Iterative Policy Evaluation
Control	Bellman Expectation Equation + Greedy Policy Improvement	Policy Iteration
Control	Bellman Optimality Equation	Value Iteration

- Algorithms are based on state-value function  $v_{\pi}(s)$  or  $v_{*}(s)$
- Complexity  $O(mn^2)$  per iteration, for  $m$  actions and  $n$  states
- Could also apply to action-value function  $q_{\pi}(s, a)$  or  $q_{*}(s, a)$
- Complexity  $O(m^2n^2)$  per iteration

# Asynchronous Dynamic Programming

- DP methods described so far used *synchronous* backups
- i.e. all states are backed up in parallel
- *Asynchronous DP* backs up states individually, in any order
- For each selected state, apply the appropriate backup
- Can significantly reduce computation
- Guaranteed to converge if all states continue to be selected

# Asynchronous Dynamic Programming

Three simple ideas for asynchronous dynamic programming:

- *In-place* dynamic programming
- *Prioritised sweeping*
- *Real-time* dynamic programming

# In-Place Dynamic Programming

- Synchronous value iteration stores two copies of value function  
for all  $s$  in  $\mathcal{S}$

$$v_{new}(s) \leftarrow \max_{a \in \mathcal{A}} \left( \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_{old}(s') \right)$$

$$v_{old} \leftarrow v_{new}$$

- In-place value iteration only stores one copy of value function  
for all  $s$  in  $\mathcal{S}$

$$v(s) \leftarrow \max_{a \in \mathcal{A}} \left( \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v(s') \right)$$

# Prioritised Sweeping

- Use magnitude of Bellman error to guide state selection, e.g.

$$\left| \max_{a \in \mathcal{A}} \left( \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v(s') \right) - v(s) \right|$$

- Backup the state with the largest remaining Bellman error
- Update Bellman error of affected states after each backup
- Requires knowledge of reverse dynamics (predecessor states)
- Can be implemented efficiently by maintaining a priority queue

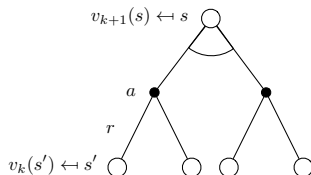
# Real-Time Dynamic Programming

- Idea: only states that are relevant to agent
- Use agent's experience to guide the selection of states
- After each time-step  $S_t, A_t, R_{t+1}$
- Backup the state  $S_t$

$$v(S_t) \leftarrow \max_{a \in \mathcal{A}} \left( \mathcal{R}_{S_t}^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{S_t s'}^a v(s') \right)$$

# Full-Width Backups

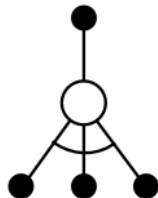
- DP uses *full-width* backups
- For each backup (sync or async)
  - Every successor state and action is considered
  - Using knowledge of the MDP transitions and reward function
- DP is effective for medium-sized problems (millions of states)
- For large problems DP suffers Bellman's *curse of dimensionality*
  - Number of states  $n = |\mathcal{S}|$  grows exponentially with number of state variables
- Even one backup can be too expensive





# Sample Backups

- In subsequent lectures we will consider *sample backups*
- Using sample rewards and sample transitions  
 $\langle S, A, R, S' \rangle$
- Instead of reward function  $\mathcal{R}$  and transition dynamics  $\mathcal{P}$
- Advantages:
  - Model-free: no advance knowledge of MDP required
  - Breaks the curse of dimensionality through sampling
  - Cost of backup is constant, independent of  $n = |\mathcal{S}|$



# Approximate Dynamic Programming

- Approximate the value function
- Using a *function approximator*  $\hat{v}(s, \mathbf{w})$
- Apply dynamic programming to  $\hat{v}(\cdot, \mathbf{w})$
- e.g. Fitted Value Iteration repeats at each iteration  $k$ ,
  - Sample states  $\tilde{\mathcal{S}} \subseteq \mathcal{S}$
  - For each state  $s \in \tilde{\mathcal{S}}$ , estimate target value using Bellman optimality equation,

$$\tilde{v}_k(s) = \max_{a \in \mathcal{A}} \left( \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a \hat{v}(s', \mathbf{w}_k) \right)$$

- Train next value function  $\hat{v}(\cdot, \mathbf{w}_{k+1})$  using targets  $\{\langle s, \tilde{v}_k(s) \rangle\}$

## Some Technical Questions

- How do we know that value iteration converges to  $v_*$ ?
- Or that iterative policy evaluation converges to  $v_\pi$ ?
- And therefore that policy iteration converges to  $v_*$ ?
- Is the solution unique?
- How fast do these algorithms converge?
- These questions are resolved by *contraction mapping theorem*

# Value Function Space

- Consider the vector space  $\mathcal{V}$  over value functions
- There are  $|\mathcal{S}|$  dimensions
- Each point in this space fully specifies a value function  $v(s)$
- What does a Bellman backup do to points in this space?
- We will show that it brings value functions *closer*
- And therefore the backups must converge on a unique solution

## Value Function $\infty$ -Norm

- We will measure distance between state-value functions  $u$  and  $v$  by the  $\infty$ -norm
- i.e. the largest difference between state values,

$$\|u - v\|_{\infty} = \max_{s \in \mathcal{S}} |u(s) - v(s)|$$

# Bellman Expectation Backup is a Contraction

- Define the *Bellman expectation backup operator*  $T^\pi$ ,

$$T^\pi(v) = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi v$$

- This operator is a  $\gamma$ -contraction, i.e. it makes value functions closer by at least  $\gamma$ ,

$$\begin{aligned} \|T^\pi(u) - T^\pi(v)\|_\infty &= \|(\mathcal{R}^\pi + \gamma \mathcal{P}^\pi u) - (\mathcal{R}^\pi + \gamma \mathcal{P}^\pi v)\|_\infty \\ &= \|\gamma \mathcal{P}^\pi(u - v)\|_\infty \\ &\leq \|\gamma \mathcal{P}^\pi\| \|u - v\|_\infty \\ &\leq \gamma \|u - v\|_\infty \end{aligned}$$

# Contraction Mapping Theorem

## Theorem (Contraction Mapping Theorem)

*For any metric space  $\mathcal{V}$  that is complete (i.e. closed) under an operator  $T(v)$ , where  $T$  is a  $\gamma$ -contraction,*

- *$T$  converges to a unique fixed point*
- *At a linear convergence rate of  $\gamma$*

# Convergence of Iter. Policy Evaluation and Policy Iteration

- The Bellman expectation operator  $T^\pi$  has a unique fixed point
- $v_\pi$  is a fixed point of  $T^\pi$  (by Bellman expectation equation)
- By contraction mapping theorem
- Iterative policy evaluation converges on  $v_\pi$
- Policy iteration converges on  $v_*$



# Bellman Optimality Backup is a Contraction

- Define the *Bellman optimality backup operator*  $T^*$ ,

$$T^*(v) = \max_{a \in \mathcal{A}} \mathcal{R}^a + \gamma \mathcal{P}^a v$$

- This operator is a  $\gamma$ -contraction, i.e. it makes value functions closer by at least  $\gamma$  (similar to previous proof)

$$\|T^*(u) - T^*(v)\|_\infty \leq \gamma \|u - v\|_\infty$$

# Convergence of Value Iteration

- The Bellman optimality operator  $T^*$  has a unique fixed point
- $v_*$  is a fixed point of  $T^*$  (by Bellman optimality equation)
- By contraction mapping theorem
- Value iteration converges on  $v_*$