Multi-Agent Systems

Introduction to Reinforcement Learning:

Model-based Prediction and Control

Eric Pauwels (CWI & VU)

November 25, 2021

Reading

- Sutton & Barto: chapters 3 & 4

Outline

Reinforcement Learning: Markov Decision Process (MDP)

Optimal Policy and Bellman Optimality Equations

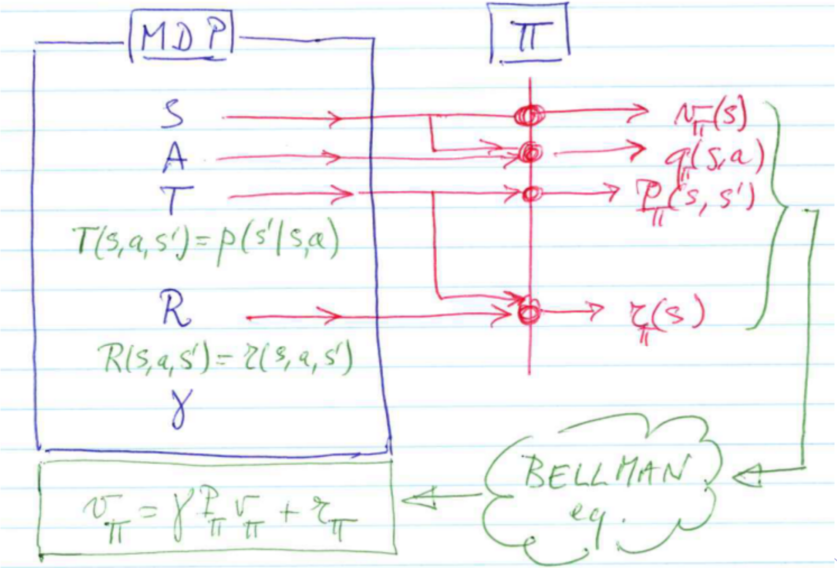
Taxonomy of RL problems

Model-based Prediction and Control

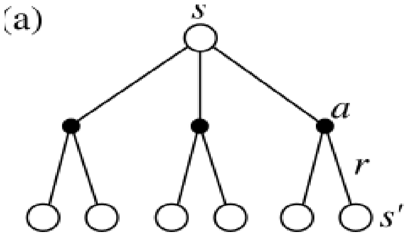
Markov decision processes (MDP)

- **Markov decision processes (MDP)** provide a **formal model** for a **sequential decision problem**;
- A **finite MDP** $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ consists of:
  - **Discrete time** $t = 0, 1, 2, \ldots$
  - A discrete set of **states** $s \in S$
  - A discrete set of **actions** $a \in A(s)$ for each s
  - A **transition function** $p(s'|s, a)$: probability of transitioning to state s' when taking action a at state s
  - A **reward function** $r(s, a, s') = E[r|s, a, s']$: expected reward when taking action a at state s and transitioning to s'
  - A planning horizon H or **discount factor** γ;
    - How important are future rewards?
    - shortsighted $\quad 0 \leftarrow \gamma \rightarrow 1 \quad$ farsighted

MDP – Policy – Bellman



Backup diagrams for Bellman equation (1)

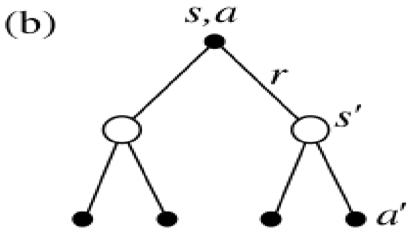


$$v_\pi(s) = \sum_a \pi(a \mid s) q_\pi(s, a) \qquad \text{(weighted mean over } a\text{)}$$

$$q_\pi(s, a) = \sum_{s'} p(s' \mid s, a)\Big[r(s, a, s') + \gamma v_\pi(s')\Big]$$

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a)\Big[r(s, a, s') + \gamma v_\pi(s')\Big]$$

Backup diagrams for Bellman equation (2)



$$q_\pi(s,a) = \sum_{s'} p(s' \mid s,a)\Big[r(s,a,s') + \gamma v_\pi(s')\Big]$$

$$q_\pi(s,a) = \sum_{s'} p(s' \mid s,a)\Big[r(s,a,s') + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s',a')\Big]$$

Bellman equation: Summary

- The definition of v_π can be rewritten recursively by making use of the transition model, yielding the **Bellman equation**:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a)\Big[r(s, a, s') + \gamma v_\pi(s')\Big]$$

- This is a set of **linear equations,** one for each state, the solution of which defines the value of π
- A similar recursive definition holds for Q-values:

$$q_\pi(s, a) = \sum_{s'} p(s' \mid s, a)\Big[r(s, a, s') + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s', a')\Big]$$

Matrix form of Bellman equation

$$\mathbf{v}_\pi = \gamma P_\pi \mathbf{v}_\pi + \mathbf{r}_\pi \qquad \text{or again} \qquad (I - \gamma P_\pi)\mathbf{v}_\pi = \mathbf{r}_\pi$$

where

- square matrix $P(s, s')$ is **transition probability** $s \to s'$ (under policy π):

$$P_\pi(s, s') := \sum_a \pi(a \mid s) p(s' \mid s, a)$$

- $\mathbf{r}_\pi(s)$ is **expected (immediate) reward** in s (under policy π):

$$\mathbf{r}_\pi(s) = \sum_a \pi(a \mid s) R(s, a) \quad \text{where } R(s, a) = \sum_{s'} p(s' \mid s, a) r(s, a, s')$$

Solving the Bellman equation

$$\mathbf{v}_\pi = \gamma P_\pi \mathbf{v}_\pi + \mathbf{r}_\pi \qquad \text{or again} \qquad (I - \gamma P_\pi)\mathbf{v}_\pi = \mathbf{r}_\pi$$

- Notice that $\mathbf{v} = (I - \gamma P)^{-1}\mathbf{r} = (I + \gamma P + \gamma^2 P^2 + \ldots)\,\mathbf{r}$
- **Alternatively:** solve iteratively (fix-point!)

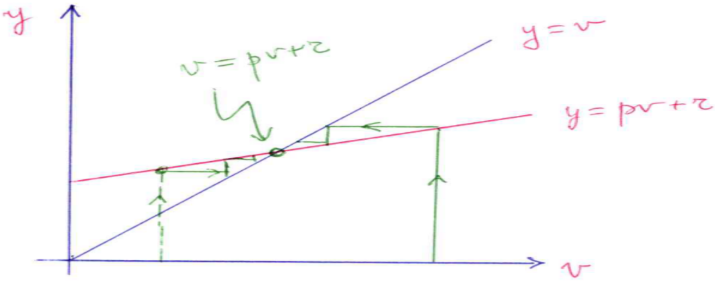
Solving Bellman eqs: Iteration to Fix-Point

Matrix form of **Bellman equation** corresponds to fix-point:

$$\mathbf{v} = \gamma P \mathbf{v} + \mathbf{r}$$

Iterative solution: (Dynamic Progamming DP) update rule:

$$\mathbf{v}^{k+1} = \gamma P \mathbf{v}^{k} + \mathbf{r}$$



Outline

Reinforcement Learning: Markov Decision Process (MDP)

Optimal Policy and Bellman Optimality Equations

Taxonomy of RL problems

Model-based Prediction and Control

Optimal value functions

- Value functions define a partial ordering over policies:

$$\pi \succ \pi' \iff v_\pi(s) \geq v_{\pi'}(s),\ \forall s \in S$$

- There can be multiple optimal policies but they all share the same **optimal state-value function**:

$$v^*(s) = \max_\pi v_\pi(s), \quad \forall s \in S$$

- They also share the same **optimal action-value function**:

$$q^*(s,a) = \max_\pi q_\pi(s,a), \quad \forall s \in S, a \in A$$

Backup Diagram for Bellman Optimality Equations

Optimize over actions!

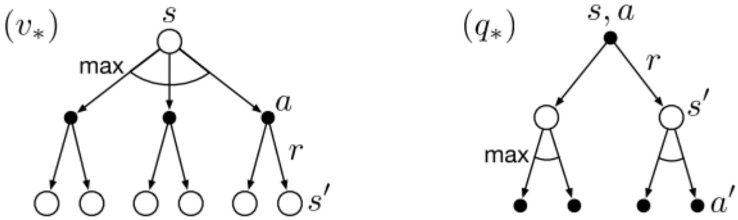


Figure 3.5: Backup diagrams for v_* and q_*

Bellman optimality equation in matrix form

$$
\begin{aligned}
v^*(s) &= \max_{a \in A} \sum_{s'} p(s' \mid s, a)\Big[r(s, a, s') + \gamma v^*(s')\Big] \\
&= \max_{a} \left(R(s, a) + \gamma \sum_{s'} \underbrace{p(s' \mid s, a)}_{T_a(s,s')} v^*(s') \right) \\
&= \max_{a} \left(R(s, a) + \gamma \sum_{s'} T_a(s, s') v^*(s') \right)
\end{aligned}
$$

or in matrix notation:

$$
\mathbf{v}^* = \max_a \left(R_a + \gamma T_a \mathbf{v}^* \right)
$$

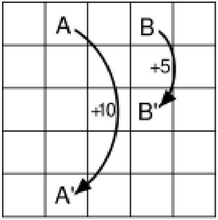
q^* versus v^*

$$V^*(s) = \max_a Q^*(s, a)$$

Why optimal value functions are useful

An optimal policy is **greedy** with respect to v^* or q^*:

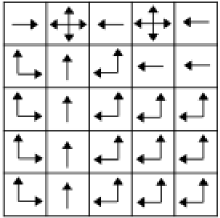
$$\pi^*(s) \in \arg\max_a q^*(s,a) = \arg\max_a \left[\sum_{s'} p(s' \mid a, s)(r(s,a,s') + \gamma v^*(s')) \right]$$



a) gridworld

22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

b) V^*



c) π^*

Outline

Reinforcement Learning: Markov Decision Process (MDP)

Optimal Policy and Bellman Optimality Equations

Taxonomy of RL problems

Model-based Prediction and Control

Model-based vs model-free

- **Model-based:** the MDP $= (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ is completely specified;
  - Solve the Bellman (optimality) equations
  - Suffices to focus on state value function $v(s)$;
- **Model-free:** only **direct experience**, i.e. sample paths (states, actions and rewards) are given. Put differently, only experience-based information is given!
  - Focus on state-action value function $q(s, a)$
  - Random search but Bellman equations allow to propagate values!

Taxonomy of RL problems

	Prediction *Estimation:* *Given π, what is v?*	**(Optimal) Control** *Optimisation:* *What is optimal π?*
model-based (MDP given)	Policy evaluation using Dyn. Programming (DP)	Policy improvement (+ Policy evaluation) = Policy iteration
model-free (MDP unknown)	Monte Carlo (MC) Temporal Diff^{ing} (TD) = "impatient MC" *bootstrapping!*	Generalized Policy Iteration *"simultaneous"*

Outline

Reinforcement Learning: Markov Decision Process (MDP)

Optimal Policy and Bellman Optimality Equations

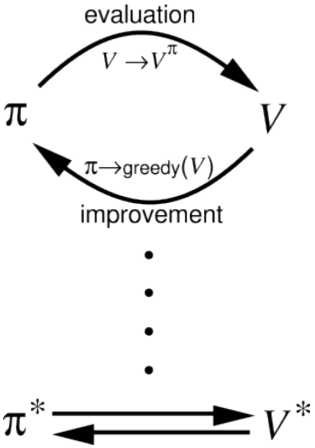
Taxonomy of RL problems

Model-based Prediction and Control

Model-based Prediction and Control

- **Dynamic Programming (DP):** Collection of algorithms that can be used to **compute optimal policy** given a **completely specified model** for the environment (MDP);
- **Policy evaluation:** given a policy π compute value functions $v_\pi(s)$ and $q_\pi(s,a)$;
- **Policy improvement:** given a policy π and corresponding value function v_π, can we find a better policy π' such that $v_{\pi'} \geq v_\pi$?
- **Policy iteration:** iteratively alternate between policy evaluation and improvement to find an optimal policy.

Policy evaluation, improvement and iteration



Policy **evaluation** (1)

- Rather than estimating value of each state independently, use Bellman equation to exploit the relationship between states
- Initial value function v_0 is chosen arbitrarily
- Policy evaluation = evaluate value function under the policy update rule:

$$v_{k+1}(s) \leftarrow \sum_{a} \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \left[r(s, a, s') + \gamma v_k(s') \right]$$

- Apply to every state in each **sweep** of the state space
- Repeat over many sweeps
- Converges to the fixed point $v^k = v_\pi$

Policy **evaluation (2)**: Algorithm

Input π, the policy to be evaluated;

Initialize $v(s) = 0$, for all $s \in \mathcal{S}$

Repeat:

$\Delta \leftarrow 0$;

for each $s \in S$: # single sweep over all states

$v \leftarrow v(s)$

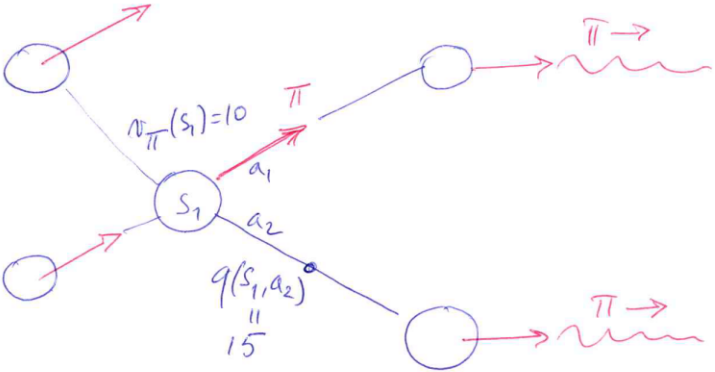
$v(s) \leftarrow \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a)(r(s, a, s') + \gamma v(s'))$

$\Delta \leftarrow \max(\Delta, |v - v(s)|)$

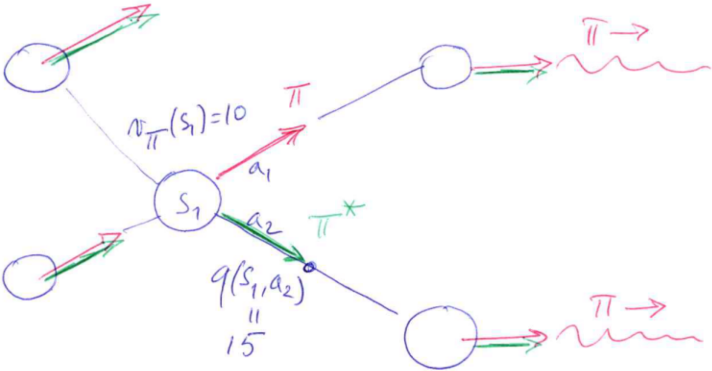
until $\Delta \leq$ small positive number;

Output: $v \approx v_\pi$

Policy **improvement**



Policy **improvement**



Policy **improvement** (1)

- Policy evaluation yields v_π, the true value of π
- Use this to incrementally improve the policy by considering whether for some state s there is a better action $a \neq \pi(s)$
- Is **choosing** a **in** s and **then using** π better than using π, i.e.,

$$q_\pi(s,a) = \sum_{s'} p(s' \mid s, a)\Big[r(s,a,s') + \gamma v_\pi(s')\Big] > v_\pi(s)?$$

- If so, then the **policy improvement theorem** tells us that changing π to take a in s will increase its value:

$$\forall s \in S, q_\pi(s, \pi'(s)) \geq v_\pi(s) \quad \Rightarrow \quad \forall s \in S, v_{\pi'}(s) \geq v_\pi(s)$$

- In our case, $\pi = \pi'$ except that $\pi'(s) = a \neq \pi(s)$

Policy **improvement** (2)

- Applying to all states yields the **greedy** policy w.r.t. v_π:

$$\pi'(s) \leftarrow \arg\max_a q_\pi(s, a)$$

$$v_{\pi'}(s) = \max_a \sum_{s'} p(s' \mid s, a)\Big[r(s, a, s') + \gamma v_\pi(s')\Big]$$

- If $\pi = \pi'$, then $v_\pi = v_{\pi'}$ and for all $s \in S$:

$$v_{\pi'}(s) = \max_{a \in A} \sum_{s'} p(s' \mid s, a)\Big[r(s, a, s') + \gamma v'_\pi(s')\Big]$$

- This is equivalent to the Bellman optimality equation, implying that $v_\pi = v_{\pi'} = v^*$ and $\pi = \pi' = \pi^*$

Policy **iteration** (1)

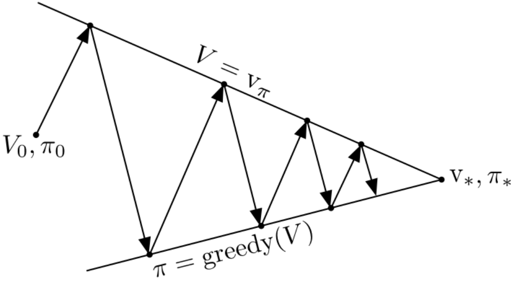
Policy iteration = policy evaluation + policy improvement

$$\pi_0 \xrightarrow{\mathrm{E}} V^{\pi_0} \xrightarrow{\mathrm{I}} \pi_1 \xrightarrow{\mathrm{E}} V^{\pi_1} \xrightarrow{\mathrm{I}} \pi_2 \xrightarrow{\mathrm{E}} \cdots \xrightarrow{\mathrm{I}} \pi^* \xrightarrow{\mathrm{E}} V^*,$$

- Policy improvement makes result of policy evaluation obsolete
- Return to policy evaluation to compute $v_{\pi'}$
- Converges to the fixed point $v_\pi = v^*$

Policy **iteration** (2): geometric analogy

A geometric metaphor for convergence of GPI:



Compare to **EM-algorithm** in ML.

Policy iteration (3): Algorithm (for deterministic policy)

1. Initialization
   $V(s) \in \Re$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Repeat
   - $\Delta \leftarrow 0$
   - For each $s \in \mathcal{S}$:
     - $v \leftarrow V(s)$
     - $V(s) \leftarrow \sum_{s'} \mathcal{P}_{ss'}^{\pi(s)} \left[\mathcal{R}_{ss'}^{\pi(s)} + \gamma V(s') \right]$
     - $\Delta \leftarrow \max(\Delta, |v - V(s)|)$

   until $\Delta < \theta$ (a small positive number)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   - $b \leftarrow \pi(s)$
   - $\pi(s) \leftarrow \arg\max_a \sum_{s'} \mathcal{P}_{ss'}^{a} \left[\mathcal{R}_{ss'}^{a} + \gamma V(s') \right]$
   - If $b \neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*

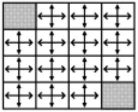
   If *policy-stable*, then stop; else go to 2

Stopping policy evaluation early

V_k for the Random Policy | Greedy Policy w.r.t. V_k

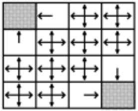
$k = 0$

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



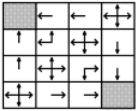
$k = 1$

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



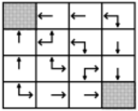
$k = 2$

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



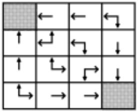
$k = 3$

0.0	-2.4	-2.9	-3.0
-2.4	-2.9	-3.0	-2.9
-2.9	-3.0	-2.9	-2.4
-3.0	-2.9	-2.4	0.0



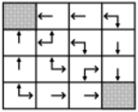
$k = 10$

0.0	-6.1	-8.4	-9.0
-6.1	-7.7	-8.4	-8.4
-8.4	-8.4	-7.7	-6.1
-9.0	-8.4	-6.1	0.0



$k = \infty$

0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0



Value iteration

- Compute optimal v^* first (iteratively), then derive optimal policy π^*
- Function iteration:

$$v_{k+1}(s) \leftarrow \max_a q_{k+1}(s,a),$$
$$q_{k+1}(s,a) \leftarrow \sum_{s'} p(s' \mid s,a)\Big[r(s,a,s') + \gamma v_k(s')\Big]$$

- Turns Bellman optimality equation into an update rule:

$$v_{k+1}(s) \leftarrow \max_a \sum_{s'} p(s' \mid s,a)\Big[r(s,a,s') + \gamma v_k(s')\Big]$$

Efficiency of dynamic programming

- An MDP has $|A|^{|S|}$ deterministic policies
- But the worst-case computational complexity of dynamic programming is polynomial in $|S|$ and $|A|$
- MDP planning can also be done with **linear programming**, which has better worst-case guarantees, but is impractical for large MDPs
- In very large MDPs, where even doing one sweep is infeasible, **asynchronous dynamic programming** must be used
- Convergence in the limit is guaranteed as long as every state is backed up infinitely often

Summary of terminology

- **Value iteration** algorithms search for optimal value function v^* from which policy is deduced:

$$v_1 \longrightarrow v_2 \longrightarrow \ldots \longrightarrow v^* \longrightarrow \pi^*$$

- **Policy iteration** algorithms evaluate the policy π by computing the (corresponding) value function v_π and uses v_π to improve the policy: va

$$\pi_1 \longrightarrow v_1 \longrightarrow \pi_2 \longrightarrow v_2 \longrightarrow \ldots \longrightarrow \pi^*$$

- **Policy search** algorithms use optimisation techniques to directly search for an optimal policy:

$$\pi_1 \longrightarrow \pi_2 \longrightarrow \ldots \longrightarrow \pi^*$$