Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Informatik
Computer Science

Probabilistic Foundations of Artificial Intelligence

Probabilistic Planning

Prof. Andreas Krause
Learning and Adaptive Systems (las.ethz.ch)

New topic: Probabilistic planning

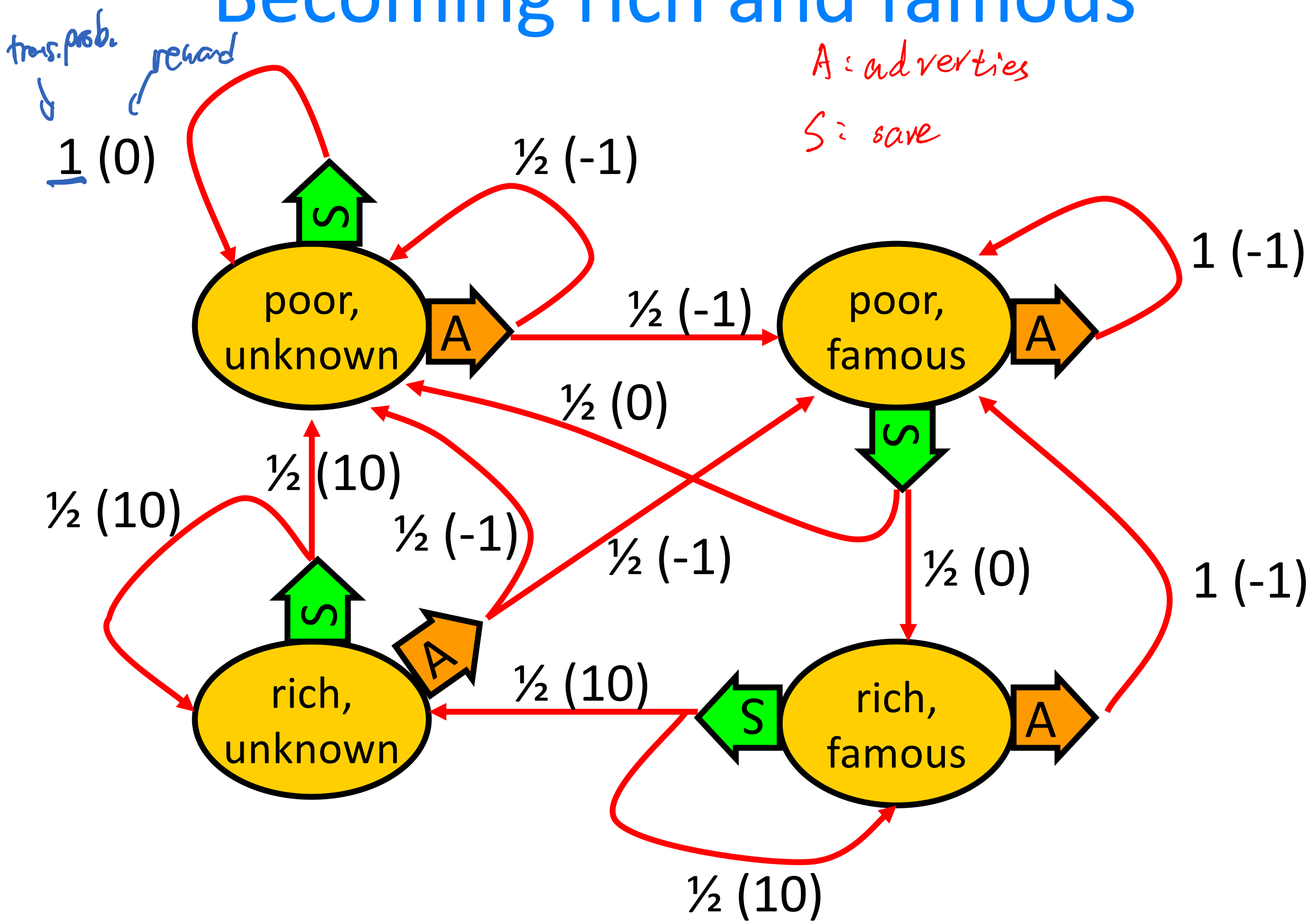
- So far: Probabilistic inference in dynamical models
  - E.g.: Tracking a robot based on noisy measurements

- Next: How should we control the robot to accomplish some goal / perform some task?

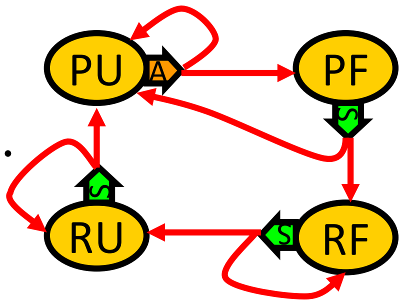
Markov Decision Processes

- An MDP is specified by
  - A set of states $X = \{1,\ldots, n\}$ …
  - A set of actions $A = \{1,\ldots, m\}$
  - Transition probabilities
    $P(x' \mid x, a) = \text{Prob}(\text{Next state} = x' \mid \text{Action } a \text{ in state } x)$
  - A reward function $r(x, a)$
    Reward can be random with mean $r(x, a)$;
    Reward may depend on x only or (x, a, x') as well.
- For now assume r and P are known!
- Want to choose actions to maximize reward

Becoming rich and famous



Planning in MDPs



- Deterministic policy $\pi: X \rightarrow A$
- Induces a **Markov chain**: $X_0, X_1, \ldots, X_t, \ldots$ with transition probabilities

$$\overset{\pi}{P}(X_{t+1}=x' \mid X_t=x) = P(x' \mid x, \pi(x))$$

- Expected value $J(\pi) = E[\quad r(X_0,\pi(X_0))$

$$+ \gamma\, r(X_1,\pi(X_1))$$
$$+ \gamma^2\, r(X_2,\pi(X_2))$$
$$+ \ldots \quad]$$

Computing the value of a policy

For a fixed policy define value function

$$\underline{V^\pi(x)} = J(\pi \mid X_0 = x) = \underline{\mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^{(t)} r(X_t, \pi(X_t)) \mid X_0 = x\Big]}$$

Recursion:

$$V^\pi(x) = \mathbb{E}\Big[\gamma^0 r(X_0, \pi(X_0)) + \sum_{t=1}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x\Big]$$

lin. of. exp.

$$= \mathbb{E}[r(X_0, \pi(X_0)) \mid X_0 = x] + \mathbb{E}\Big[\sum_{t=1}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x\Big]$$

index shift

$$= r(x, \pi(x)) + \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^{t+1} r(X_{t+1}, \pi(X_{t+1})) \mid X_0 = x\Big]$$

$$= r(x, \pi(x)) + \gamma\, \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^{t} r(X_{t+1}, \pi(X_{t+1})) \mid X_0 = x\Big] \leftarrow \mathbb{E}_{X_1}\big[\mathbb{E}_{X_{2,3,\dots}}\{\sum r \dots\}\big]$$

iter. expect.

$$= r(x, \pi(x)) + \gamma \sum_{x'} \underbrace{P(X_1 = x' \mid X_0 = x, \pi(x))}_{\mathbb{E}_{X_1}} \underbrace{\mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^{t} r(X_{t+1}, \pi(X_{t+1})) \mid X_1 = x'\Big]}_{V^\pi(x')}$$

stationarity

$$= r(x, \pi(x)) + \gamma \sum_{x'} P(x' \mid x, \pi(x))\, V^\pi(x')$$

Solving for the value of a policy

$$V^\pi(x) = r(x, \pi(x)) + \gamma \sum_{x'} P(x' \mid x, \pi(x)) V^\pi(x')$$

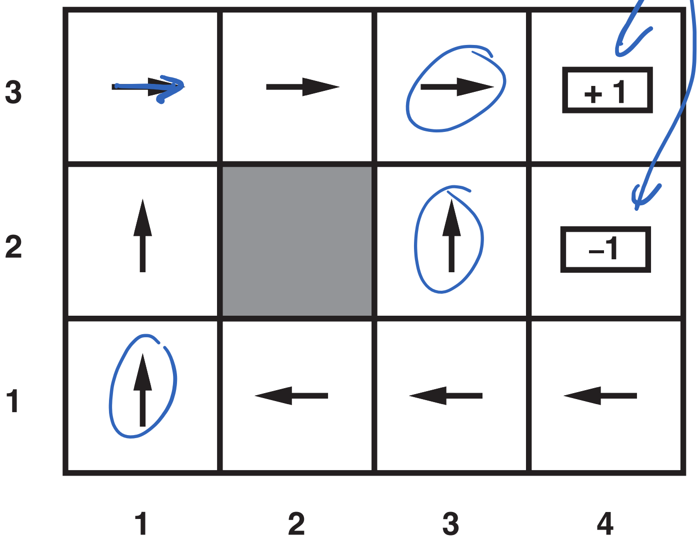
$$V^\pi = \begin{pmatrix} V^\pi(1) \\ \vdots \\ V^\pi(n) \end{pmatrix} \quad r^\pi = \begin{pmatrix} r(1,\pi(1)) \\ \vdots \\ r(n,\pi(n)) \end{pmatrix} \quad T^\pi = \begin{pmatrix} P(1 \mid 1,\pi(1)) & \cdots & P(n \mid 1,\pi(1)) \\ \vdots & & \vdots \\ P(1 \mid n,\pi(n)) & \cdots & P(n \mid n,\pi(n)) \end{pmatrix}$$

$$V^\pi = r^\pi + \gamma T^\pi V^\pi \Rightarrow V^\pi = \underbrace{(I - \gamma T^\pi)^{-1}}_{\text{Sol. exists if } \gamma < 1} r^\pi$$

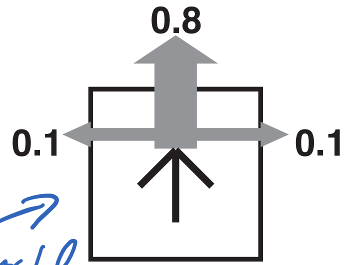
➔ **Can compute V^π exactly by solving linear system! ☺**

Value function illustration

absorbing (stay here)



3	0.812	0.868	0.918	+ 1
2	0.762		0.660	–1
1	0.705	0.655	0.611	0.388
	1	2	3	4



How can we find the optimal policy?

motion model

A simple algorithm

- For every policy π compute $J(\pi)$ $= \sum_x P(X_0 = x) V(x)$
- Pick $\pi^* = \operatorname{argmax} J(\pi)$

Is this a good idea?

policies is $O(|A|^{|X|})$

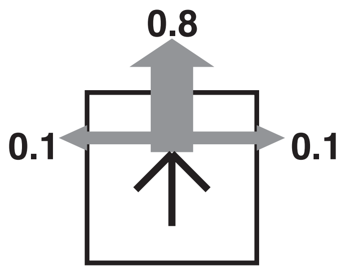
Suppose I give you the values

- Suppose you know V, and start in state x.
- Which action would you choose?

$$a^* \in \arg\max_a \; r(x,a) + \gamma \sum_{x'} P(x' \mid x,a)\, V(x')$$

$$\| \quad \pi(x)$$

3	0.812	0.868	0.918	+ 1
2	0.762		0.660	–1
1	0.705	0.655	0.611	0.388
	1	2	3	4



Value functions and policies

Every value function induces a policy

Value function V^π

$V^\pi(x) = r(x,\pi(x)) + \gamma \sum_{x'} P(x' | x,\pi(x))\, V^\pi(x')$

Greedy policy w.r.t. *V*

$\pi_V(x) = \text{argmax}_a\, r(x,a) + \gamma \sum_{x'} P(x' \mid x,a)\, V(x')$

Every policy induces a value function

Theorem (Bellman):

Policy optimal ⇔ greedy w.r.t. its induced value function!

$$V^*(x) = \max_a \left[r(x,a) + \gamma \sum_{x'} P(x' \mid x, a)\, V^*(x') \right]$$

Policy iteration

- Start with an arbitrary (e.g., random) policy π
- Until converged do:

  Compute value function $V^{\pi}(x)$

  Compute greedy policy π_G w.r.t. V^{π}

  Set $\pi \leftarrow \pi_G$

- Guaranteed to
  - Monotonically improve $V^{\pi_{t+1}}(x) \geq V^{\pi_t}(x) \quad \forall x, t$
  - Converge to an optimal policy π^* in $O^*(n^2\, m \,/\, (1\text{-}\gamma))$ iterations! [Ye '10]

Alternative approach

- Recall (Bellman): For the optimal policy π^* it holds

$$\boxed{V^*(x) = \max_a r(x,a) + \gamma \sum_{x'} P(x' \mid x, a)\, V^*(x')}$$

- Compute V^* using fixed point / dynamic programming:

$V_t(x)$ = Max. expected reward when starting in state x and world ends in t time steps

$$V_0(x) = \max_a r(x,a)$$

$$V_1(x) = \max_a r(x,a) + \gamma \sum_{x'} P(x' \mid x,a)\, V_0(x')$$

$$V_{t+1}(x) = \max_a r(x,a) + \gamma \sum_{x'} P(x' \mid x,a)\, V_t(x')$$

Value iteration

- Initialize $V_0(x) = \max_a r(x, a)$
- For $t = 1$ to ∞

  For each x, a, let

  $$Q_t(x,a) = r(x,a) + \gamma \sum_{x'} P(x' \mid x,a) V_{t-1}(x')$$

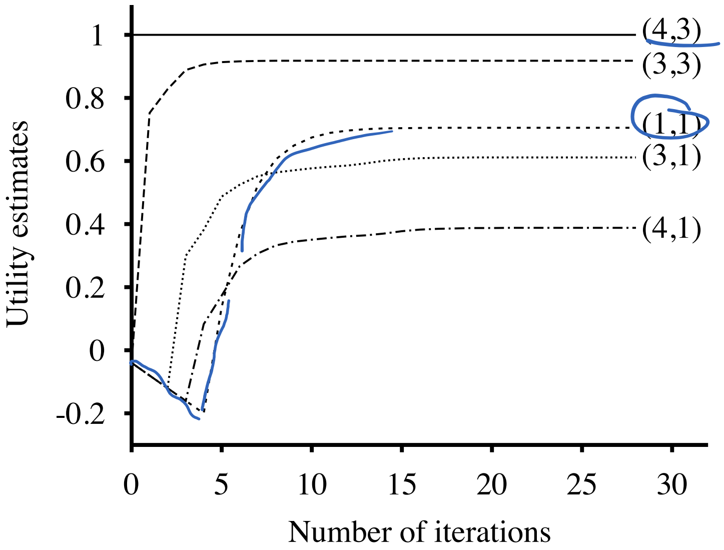
  For each x let $V_t(x) = \max_a Q_t(x,a)$

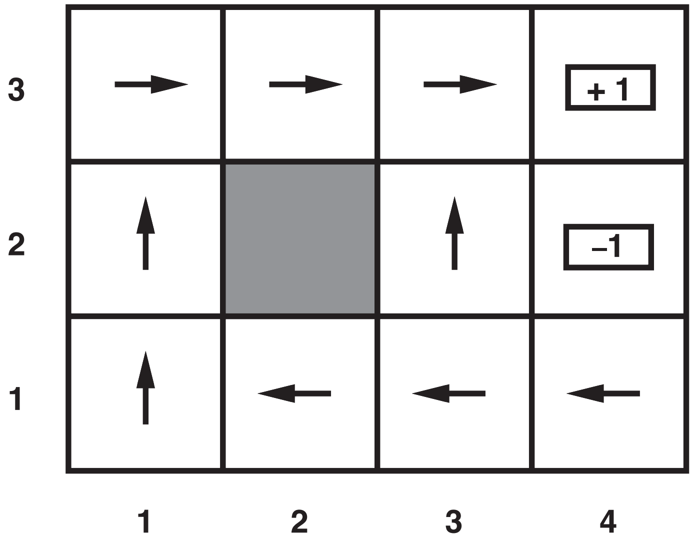
  Break if $||V_t - V_{t-1}||_\infty = \max_x |V_t(x) - V_{t-1}(x)| \leq \varepsilon$

- Then choose greedy policy w.r.t. V_t

- **Guaranteed to converge to ε-optimal policy!**

Value iteration





Convergence of Value Iteration

- Main ingredient of proof: Bellman update is a contraction

$B: \mathbb{R}^n \to \mathbb{R}^n \quad B: V \mapsto BV$

$(BV)(x) = \max_a r(x,a) + \gamma \sum_{x'} P(x' \mid x,a) V(x')$

Thm: $\forall V, V' \in \mathbb{R}^n \quad \|BV - BV'\|_\infty \le \gamma \cdot \|V - V'\|_\infty$

$\underbrace{\|B^t V_0 - \underbrace{B^t V^*}_{V^*}\|}_{\le \varepsilon} \le \gamma^t \underbrace{\|V_0 - V^*\|}_{\Delta}$

- A contraction has two important properties:
  - Existence of a unique fixed points:

$\exists!\ V^*: \ \underline{BV^* = V^*}$

  - Convergence to the fixed point:

$$\lim_{t \to \infty} B^t V_0 := \lim_{t \to \infty} \underbrace{B(B(B \ldots (B(V)) \ldots)}_{t\text{-times}} = V^*$$

Acknowledgments

- Slides based on material accompanying the textbook "AI: A Modern Approach" (3rd edition) by S. Russell and P. Norvig, as well as material by C. Guestrin and A.W.Moore