Markov Decision Processes with Ordinal Rewards: Reference Point-Based Preferences

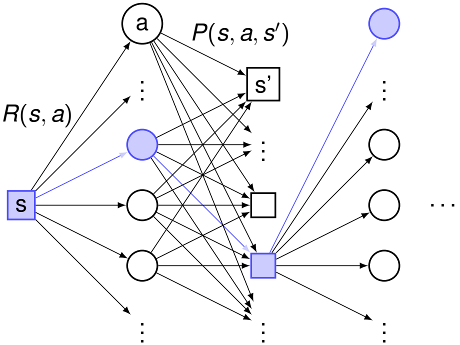
Paul Weng

LIP6, UPMC, Paris

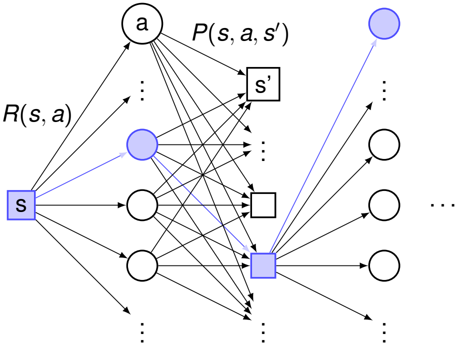
14/06/2011

21st International Conference on Automated Planning and Scheduling
Freiburg, Germany

Sequential Decision Making under Uncertainty



Sequential Decision Making under Uncertainty



MDP

- S set of states
- A set of actions
- $P : S \times A \times S \rightarrow [0, 1]$
- $R : S \times A \rightarrow \mathbb{R}$

- history γ
- $\succsim$ over policies π

Value Functions and Solution Methods

Value functions

- $v_t^\pi(s) = R(s, \pi(s)) + \beta \sum_{s' \in S} P(s, \pi(s), s') v_{t-1}^\pi(s')$

Value Functions and Solution Methods

Value functions

- $v_t^\pi(s) = R(s, \pi(s)) + \beta \sum_{s' \in S} P(s, \pi(s), s') v_{t-1}^\pi(s')$
- $\pi \succsim \pi' \Leftrightarrow \forall s, v^\pi(s) \geq v^{\pi'}(s)$

Value Functions and Solution Methods

Value functions

- $v_t^\pi(s) = R(s, \pi(s)) + \beta \sum_{s' \in S} P(s, \pi(s), s') v_{t-1}^\pi(s')$
- $\pi \succsim \pi' \Leftrightarrow \forall s, v^\pi(s) \geq v^{\pi'}(s)$
- $v^*(s) = \max_{a \in A} R(s, a) + \beta \sum_{s' \in S} P(s, a, s') v^*(s')$

Value Functions and Solution Methods

Value functions

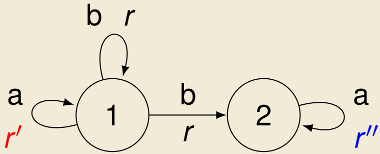
- $v_t^\pi(s) = R(s, \pi(s)) + \beta \sum_{s' \in S} P(s, \pi(s), s') v_{t-1}^\pi(s')$
- $\pi \succsim \pi' \Leftrightarrow \forall s, v^\pi(s) \geq v^{\pi'}(s)$
- $v^*(s) = \max_{a \in A} R(s, a) + \beta \sum_{s' \in S} P(s, a, s') v^*(s')$

Family of solution methods

- Value iteration
- Policy iteration
- Linear Programming

Optimal Policies Depend on the Reward Function...

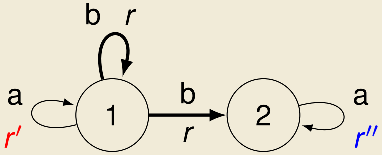
Example with $\beta = 0.5$



- $r \succ r' \succ r''$

Optimal Policies Depend on the Reward Function...

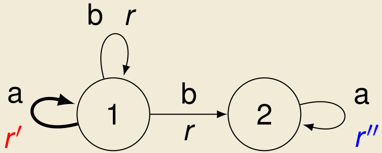
Example with $\beta = 0.5$



- $r \succ r' \succ r''$
- $2 \succ 1 \succ 0$

Optimal Policies Depend on the Reward Function...

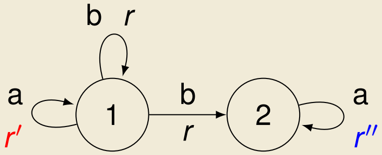
Example with $\beta = 0.5$



- $r \succ r' \succ r''$
- $2 \succ 1 \succ 0$
- $10 \succ 9 \succ 0$

Optimal Policies Depend on the Reward Function...

Example with $\beta = 0.5$



- $r \succ r' \succ r''$
- $2 \succ 1 \succ 0$
- $10 \succ 9 \succ 0$

. . . Except for One Simple Case

Proposition

If $R(s, a) \in \{0, r\}$, changing r does not impact optimal policies.

Difficulty of Defining the Reward Function

When is it easy to define numeric rewards?

- rewards = money, length, duration...
- ex: stochastic shortest path problem

Difficulty of Defining the Reward Function

When is it easy to define numeric rewards?

- rewards = money, length, duration...
- ex: stochastic shortest path problem

When is it difficult?

- values not known precisely or of qualitative nature
- ex: video games where reward represents utility

Difficulty of Defining the Reward Function

When is it easy to define numeric rewards?

- rewards = money, length, duration...
- ex: stochastic shortest path problem

When is it difficult?

- values not known precisely or of qualitative nature
- ex: video games where reward represents utility

Ordinal Reward MDP (OMDP)

- $R : S \times A \to E$
- $E = \{r_1 > r_2 \ldots > r_n\}$

Towards Preference over vectors

Histories

- γ yields a sequence of ordinal rewards $r_1, \ldots, r_n$
- Idea: count the number of each reward yielded by γ
- γ valued by $(N_1^\beta(\gamma), ..., N_n^\beta(\gamma))$

Towards Preference over vectors

Histories

- γ yields a sequence of ordinal rewards $r_1, \ldots, r_n$
- Idea: count the number of each reward yielded by γ
- γ valued by $(N_1^\beta(\gamma), ..., N_n^\beta(\gamma))$

H. preference over histories = preference over vectors

Towards Preference over vectors

Histories

- γ yields a sequence of ordinal rewards $r_1, \ldots, r_n$
- Idea: count the number of each reward yielded by γ
- γ valued by $(N_1^\beta(\gamma), ..., N_n^\beta(\gamma))$

H. preference over histories = preference over vectors

Policies in a state

- application of π in a state yields a probability distribution over histories
- π valued by the expectation of vectors $(N_1^\beta(\gamma), ..., N_n^\beta(\gamma))$

Assumptions for a Numeric Reward Functions

Axioms

A1. $\succsim$ is a complete preorder on $\mathbb{R}^n_+$

Assumptions for a Numeric Reward Functions

Axioms

A1. $\succsim$ is a complete preorder on $\mathbb{R}^n_+$

A2. $N \succsim N' \Leftrightarrow \forall i = 1, \dots, n, N + e_i \succsim N' + e_i$

Assumptions for a Numeric Reward Functions

Axioms

A1. $\succsim$ is a complete preorder on $\mathbb{R}^n_+$

A2. $N \succsim N' \Leftrightarrow \forall i = 1, \ldots, n, N + e_i \succsim N' + e_i$

A3. $N \succ N' \Rightarrow \exists n \in \mathbb{N}, nN + M \succsim nN' + M'$

Assumptions for a Numeric Reward Functions

Axioms

A1. $\succsim$ is a complete preorder on $\mathbb{R}^n_+$

A2. $N \succsim N' \Leftrightarrow \forall i = 1, \ldots, n, N + e_i \succsim N' + e_i$

A3. $N \succ N' \Rightarrow \exists n \in \mathbb{N}, nN + M \succsim nN' + M'$

Theorem

The two following propositions are equivalent:

(i) $\succsim$ *satisfies Axioms A1, A2 and A3.*

(ii) there exists a function $u : E \to \mathbb{R}$ *such that* $\forall N, N' \in \mathbb{R}^n$*:*

$$N \succsim N' \Leftrightarrow \sum_{k=1}^{n} N_k u(e_k) \geq \sum_{k=1}^{n} N'_k u(e_k)$$

Assumptions for Reference Point-Based Preferences

Additional Axioms

A4. $e_1 \succsim e_2 \succsim \ldots \succsim e_n$

Assumptions for Reference Point-Based Preferences

Additional Axioms

A4. $e_1 \succsim e_2 \succsim \ldots \succsim e_n$

A5. $N \sim N + e_{k_0}$

Assumptions for Reference Point-Based Preferences

Additional Axioms

A4. $e_1 \succsim e_2 \succsim \ldots \succsim e_n$

A5. $N \sim N + e_{k_0}$

Corollary

The two following propositions are equivalent:

(i) $\succsim$ satisfies Axioms A1 to A5.

(ii) there exists a reference point $\tilde{N} \in \mathbb{R}^n_+$ such that $\forall N, N' \in \mathbb{R}^n$:

$$N \succsim N' \Leftrightarrow \phi_{\tilde{N}}(N) \geq \phi_{\tilde{N}}(N')$$

where $\phi_{\tilde{N}}(N) = \sum_{k=1}^{k_0-1} N_k \sum_{j=k}^{k_0-1} \tilde{N}_j - \sum_{k=k_0+1}^{n} N_k \sum_{j=k_0+1}^{k} \tilde{N}_j$

Interpretation of $\phi_{\tilde{N}}$ (1/2)

Positive Feedbacks ($k_0 = n$)

- $\phi_{\tilde{N}}(N) = \sum_{k=1}^{n-1} N_k \sum_{j=k}^{n-1} \tilde{N}_j$
- $\phi_{\tilde{N}}(N)$: number of times a reward selected in N is better than one selected in $\tilde{N}$

Interpretation of $\phi_{\tilde{N}}$ (1/2)

Positive Feedbacks ($k_0 = n$)

- $\phi_{\tilde{N}}(N) = \sum_{k=1}^{n-1} N_k \sum_{j=k}^{n-1} \tilde{N}_j$
- $\phi_{\tilde{N}}(N)$: number of times a reward selected in N is better than one selected in $\tilde{N}$

Example ($n = 3$)

$N = (1, 0, 2)$ $N' = (0, 2, 1)$ $\tilde{N} = (1, 2, 0)$

$\phi_{\tilde{N}}(N) = 1 \times (1 + 2) + 0 = 3$ $\phi_{\tilde{N}}(N') = 0 + 2 \times 2 = 4$

Interpretation of $\phi_{\tilde{N}}$ (1/2)

Positive Feedbacks ($k_0 = n$)

- $\phi_{\tilde{N}}(N) = \sum_{k=1}^{n-1} N_k \sum_{j=k}^{n-1} \tilde{N}_j \quad \phi'_{\tilde{N}}(N) = \frac{\phi_{\tilde{N}}(N)}{\sum_{k=1}^{n} N_k \sum_{k=1}^{n} \tilde{N}_k}$
- $\phi_{\tilde{N}}(N)$: number of times a reward selected in N is better than one selected in $\tilde{N}$
- $\phi'_{\tilde{N}}(N)$: probability that a reward drawn from N is better than one drawn in $\tilde{N}$

Example ($n = 3$)

$N = (1, 0, 2) \quad N' = (0, 2, 1) \quad \tilde{N} = (1, 2, 0)$

$\phi_{\tilde{N}}(N) = 1 \times (1 + 2) + 0 = 3 \quad \phi_{\tilde{N}}(N') = 0 + 2 \times 2 = 4$

Interpretation of $\phi_{\tilde{N}}$ (2/2)

Negative Feedbacks ($k_0 = 1$)

- $\phi_{\tilde{N}}(N) = -\sum_{k=2}^{n} N_k \sum_{j=2}^{k} \tilde{N}_j \quad \phi'_{\tilde{N}}(N) = 1 + \frac{\phi_{\tilde{N}}(N)}{\sum_{k=1}^{n} N_k \sum_{k=1}^{n} \tilde{N}_k}$

Positive and Negative Feedbacks ($1 < k_0 < n$)

$$\phi_{\tilde{N}}(N) = \sum_{k=1}^{k_0-1} N_k \sum_{j=k}^{k_0-1} \tilde{N}_j - \sum_{k=k_0+1}^{n} N_k \sum_{j=k_0+1}^{k} \tilde{N}_j$$

Vade Mecum

How to Use Reference Point-Based Preference OMDPs

- define an OMDP
- pick a reference point
- determine vector $\tilde{N}$ and compute associated rewards

$$u^{\tilde{N}}(r_k) \begin{cases} = 0 & \text{if } k = k_0 \\ = \sum_{j=k}^{k_0-1} \tilde{N}_j & \text{if } k < k_0 \\ = -\sum_{j=k_0+1}^{k} \tilde{N}_j & \text{if } k > k_0 \end{cases}$$

- solve with any standard method

Choosing a Reference Point

- step of the qualitative scale E
- probability distribution over E
- history
- policy

Reference-Point Based Preferences in Standard MDPs: One-Shot Decision

Principle

- compute an optimal policy π^* of (S, A, P, R)

Reference-Point Based Preferences in Standard MDPs: One-Shot Decision

Principle

- compute an optimal policy π^* of (S, A, P, R)
- compute $\tilde{N}$ and then $R^{\tilde{N}}$

Reference-Point Based Preferences in Standard MDPs: One-Shot Decision

Principle

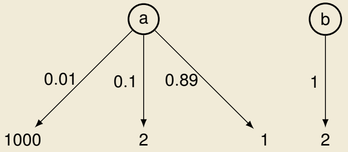
- compute an optimal policy π^* of (S, A, P, R)
- compute $\tilde{N}$ and then $R^{\tilde{N}}$
- compute an optimal policy π^{**} of $(S, A, P, R^{\tilde{N}})$

Reference-Point Based Preferences in Standard MDPs: One-Shot Decision

Principle

- compute an optimal policy π^* of (S, A, P, R)
- compute $\tilde{N}$ and then $R^{\tilde{N}}$
- compute an optimal policy π^{**} of $(S, A, P, R^{\tilde{N}})$

Example

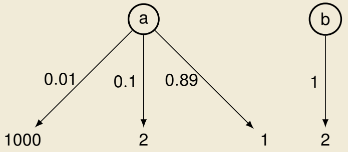


Reference-Point Based Preferences in Standard MDPs: One-Shot Decision

Principle

- compute an optimal policy π^* of (S, A, P, R)
- compute $\tilde{N}$ and then $R^{\tilde{N}}$
- compute an optimal policy π^{**} of $(S, A, P, R^{\tilde{N}})$

Example



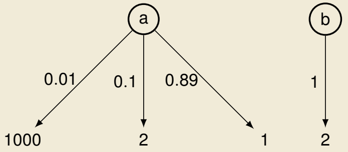
$$V^a = 10 + 0.2 + 0.89 = 11.09$$

Reference-Point Based Preferences in Standard MDPs: One-Shot Decision

Principle

- compute an optimal policy π^* of (S, A, P, R)
- compute $\tilde{N}$ and then $R^{\tilde{N}}$
- compute an optimal policy π^{**} of $(S, A, P, R^{\tilde{N}})$

Example



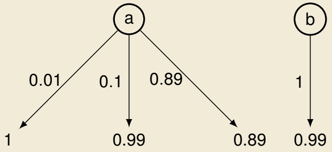
$$V^a = 10 + 0.2 + 0.89 = 11.09$$
$$V^b = 2$$

Reference-Point Based Preferences in Standard MDPs: One-Shot Decision

Principle

- compute an optimal policy π^* of (S, A, P, R)
- compute $\tilde{N}$ and then $R^{\tilde{N}}$
- compute an optimal policy π^{**} of $(S, A, P, R^{\tilde{N}})$

Example



$V^a = 10 + 0.2 + 0.89 = 11.09$
$V^b = 2$

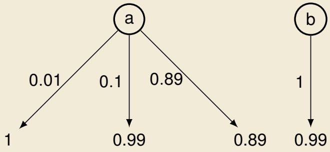
$\tilde{N} = (0.01, 0.1, 0.89)$

Reference-Point Based Preferences in Standard MDPs: One-Shot Decision

Principle

- compute an optimal policy π^* of (S, A, P, R)
- compute $\tilde{N}$ and then $R^{\tilde{N}}$
- compute an optimal policy π^{**} of $(S, A, P, R^{\tilde{N}})$

Example



$V^a = 10 + 0.2 + 0.89 = 11.09$
$V^b = 2$

$\tilde{N} = (0.01, 0.1, 0.89)$
$V^a = 0.9011$
$V^b = 0.99$

Conclusion and Future Work

- how to define a semantically justified reward function

- experimental evaluation
- relax some of the axioms
- more qualitative preference relations over vectors