Planning and Optimal Control

B.1 Markov Decision Processes

Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)
Institute for Computer Science
University of Hildesheim, Germany

Syllabus

		A. Models for Sequential Data
Tue. 22.10.	(1)	1. Markov Models
Tue. 29.10.	(2)	2. Hidden Markov Models
Tue. 5.11.	(3)	3. State Space Models
Tue. 12.11.	(4)	3b. (ctd.)
		B. Models for Sequential Decisions
Tue. 19.11.	(5)	1. Markov Decision Processes
Tue. 26.11.	(6)	1b. (ctd.)
Tue. 3.12.	(7)	2. Introduction to Reinforcement Learning
Tue. 10.12.	(8)	3. Monte Carlo and Temporal Difference Methods
Tue. 17.12.	(9)	4. Q Learning
Tue. 24.12.	—	— *Christmas Break* —
Tue. 7.1.	(10)	5. Policy Gradient Methods
Tue. 14.1.	(11)	tba
Tue. 21.1.	(12)	tba
Tue. 28.1.	(13)	8. Reinforcement Learning for Games
Tue. 4.2.	(14)	Q&A

Outline

Outline

Markov Decision Process (MDP)

An MDP (S, A, T, p, r) is a controlled stochastic Markov processes:

- finite set S called **states**,
- finite set A called **actions** (aka **controls**, **decisions**),
- set $T \subseteq \mathbb{N}$ called **time steps**,
- function $p : S \times A \rightarrow \Delta(S)$ called **state transition probability** and
  - usually written $p(s_{t+1} \mid s_t, a_t)$
  - often represented by stochastic transition matrices P_a, $a \in A$
- function $r : S \times A \rightarrow \mathbb{R}$ called **reward**.
  - often represented by vectors $r_a \in \mathbb{R}^S$, $a \in A$

Note: $\Delta(S) := \{p : S \rightarrow \mathbb{R}_0^+ \mid \sum_{s \in S} p(s) = 1\}$ probability functions over S.

Example: Find a way out of a Labyrinth

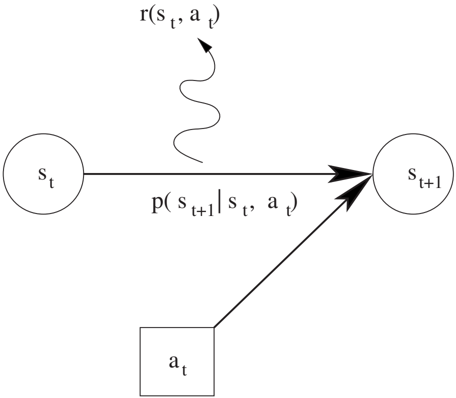
- $S := \{(x, y) \mid x, y \in \{1, 2, 3, 4, 5\}\}$
  $\setminus\{(2, 2), (2, 3), (2, 4), \quad (4, 2), (4, 3), (4, 4), (4, 5), \quad (3, 2)\}$
  walkable tiles,
  $s_0 := (1, 1)$ start location

- $A := \{(+1, 0), (-1, 0), (0, +1), (0, -1)\}$ movement
  right/left/up/down

- $p(s + a \mid s, a) := 1$, if $(s + a) \in S$
  else $p(s \mid s, a) := 1$.

- $r(s, a) = 1$ if $s = (5, 5)$ (exit),
  $r(s, a) = 0$ for all other s.

MDPs



[source: ?, p.5]

Markov Property

Markov property:

$$p(s_{t+1} \mid s_{0:t}, a_{0:t}) = p(s_{t+1} \mid s_t, a_t)$$

Action Policies

A **policy** (aka **strategy**):

$$\pi : (S \times A)^* \times S \to \Delta(A)$$

- $\pi(h, s)$ chooses a probabilistic action a if in state s with history $h = ((s_0, a_0), (s_1, a_1), \ldots, (s_{t-1}, a_{t-1}))$
- **Markov policy**: does not depend on history:

$$\pi(h, s) = \pi(h', s) \quad \forall h, h'$$

  - but may depend on time (non-stationary)
  - then just write as $\pi : T \times S \to \Delta(A)$
- **deterministic policy**: chosen action is deterministic:

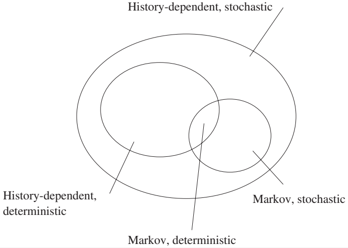
$$\forall h, s \; \exists a : \pi(h, s)(a) = 1$$

  - then just write as $\pi : (S \times A)^* \times S \to A$
- deterministic Markov policy: choose next action in each state
  - $\pi : T \times S \to A$

Action Policies / Policy Spaces

$$\Pi^{\text{MDS}} := \{\pi : S \rightarrow A\}$$
$$\Pi^{\text{MAS}} := \{\pi : S \rightarrow \Delta(A)\}$$
$$\Pi^{\text{MD}} := \{\pi : T \times S \rightarrow A\}$$
$$\Pi^{\text{MA}} := \{\pi : T \times S \rightarrow \Delta(A)\}$$
$$\Pi^{\text{HD}} := \{\pi : (S \times A)^* \times S \rightarrow A\}$$
$$\Pi^{\text{HA}} := \{\pi : (S \times A)^* \times S \rightarrow \Delta(A)\}$$

M=Markov vs. H=history-dependent
D=deterministic vs. A=stochastic
S=stationary vs. .=non-stationary



Stochastic State / Action / Reward Processes for a Policy

For an MDP (p, r),
a start state $s_0 \in S$ and
a policy π,

let

$$
\begin{array}{lll}
s_0 & a_0 \sim \pi(s_0) & r_0 := r(s_0, a_0) \\
s_1 \sim p(s_0, a_0), & a_1 \sim \pi(((s_0, a_0)), s_1) & r_1 := r(s_1, a_1) \\
\vdots & \vdots & \vdots \\
s_{t+1} \sim p(s_t, a_t), & a_{t+1} \sim \pi(((s_0, a_0), \ldots, (s_t, a_t)), s_{t+1}), & r_{t+1} := r(s_{t+1}, a_{t+1})
\end{array}
$$

describing three stochastic processes:

- the stochastic process s_t of states visited,
- the stochastic process a_t of actions taken and
- the stochastic process r_t of rewards gained

by policy π starting in s_0 for MDP (p, r).

Example: Walk on a Line

$$
\begin{aligned}
S &:= \{-10, -9, -8, \ldots, -1, 0, 1, 2, \ldots, 10\} \\
s_0 &:= 0 \\
A &:= \{+1, -1\} \\
p(s' \mid s, a) &:= \begin{cases} 1, & \text{if } s' = s + a, (s+a) \in S \quad \text{valid move} \\ 1, & \text{if } s' = s \qquad , (s+a) \notin S \quad \text{invalid move} \\ 0, & \text{else} \end{cases} \\
r(s, a) &:= \begin{cases} 1, & \text{if } s = 9, a = +1 \\ 0, & \text{else} \end{cases}
\end{aligned}
$$

Example: Walk on a Line / Go Always Left

$$\pi^L(s) := -1, \quad \text{go always left}$$

s_t	0	-1	-2	...	-8	-9	-10	-10	-10	...
a_t	-1	-1	-1	...	-1	-1	-1	-1	-1	...
r_t	0	0	0	...	0	0	0	0	0	...

- deterministic state/action/reward sequence
- total reward $\sum_{t\in\mathbb{N}} r_t = 0$

Example: Walk on a Line / Go Always Right

$$\pi^R(s) := +1, \quad \text{go always right}$$

$$\begin{array}{c|cccccccccc} s_t & 0 & 1 & 2 & \ldots & 8 & 9 & 10 & 10 & 10 & \ldots \\ a_t & +1 & +1 & +1 & \ldots & +1 & +1 & +1 & +1 & +1 & \ldots \\ r_t & 0 & 0 & 0 & \ldots & 0 & 1 & 0 & 0 & 0 & \ldots \end{array}$$

- deterministic state/action/reward sequence
- total reward: 1

Example: Walk on a Line II

a distribution of MDPs:

$$S := \{-10, -9, -8, \ldots, -1, 0, 1, 2, \ldots, 10\}$$
$$s_0 := 0$$
$$A := \{+1, -1\}$$
$$p(s' \mid s, a) := \begin{cases} 1, & \text{if } s' = s + a, (s+a) \in S \quad \text{valid move} \\ 1, & \text{if } s' = s \qquad , (s+a) \notin S \quad \text{invalid move} \\ 0, & \text{else} \end{cases}$$

every second MDP:

$$r(s, a) := \begin{cases} 1, & \text{if } s = 9, a = +1 \\ 0, & \text{else} \end{cases}$$

every other second MDP:

$$r(s, a) := \begin{cases} 1, & \text{if } s = -9, a = -1 \\ 0, & \text{else} \end{cases}$$

Example: Walk on a Line II / Go Always Left

$$\pi^L(s) := -1, \quad \text{go always left}$$

every second MDP:

s_t	0	−1	−2	...	−8	−9	−10	−10	−10	...
a_t	−1	−1	−1	...	−1	−1	−1	−1	−1	...
r_t	0	0	0	...	0	0	0	0	0	...

every other second MDP:

s_t	0	−1	−2	...	−8	−9	−10	−10	−10	...
a_t	−1	−1	−1	...	−1	−1	−1	−1	−1	...
r_t	0	0	0	...	0	1	0	0	0	...

- deterministic state/action sequence, stochastic reward sequence
- total expected reward: $\mathbb{E}(\sum_{t\in\mathbb{N}} r_t) = 0.5$

Example: Walk on a Line II/ Go Always Right

$$\pi^R(s) := +1, \quad \text{go always right}$$

every second MDP:

$$\begin{array}{c|cccccccccc}
s_t & 0 & 1 & 2 & \ldots & 8 & 9 & 10 & 10 & 10 & \ldots \\
a_t & +1 & +1 & +1 & \ldots & +1 & +1 & +1 & +1 & +1 & \ldots \\
r_t & 0 & 0 & 0 & \ldots & 0 & 1 & 0 & 0 & 0 & \ldots
\end{array}$$

every other second MDP:

$$\begin{array}{c|cccccccccc}
s_t & 0 & 1 & 2 & \ldots & 8 & 9 & 10 & 10 & 10 & \ldots \\
a_t & +1 & +1 & +1 & \ldots & +1 & +1 & +1 & +1 & +1 & \ldots \\
r_t & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots
\end{array}$$

- deterministic state/action sequence, stochastic reward sequence
- total expected reward: 0.5

Example: Walk on a Line II/ Go Left then Right

$$\pi^{LR}(t, s) := \begin{cases} -1, & \text{if } t < 10 \\ +1, & \text{else} \end{cases}$$

every second MDP:

s_t	0	−1	...	−8	−9	−10	−9	...	8	9	10	10	...
a_t	−1	−1	...	−1	−1	+1	+1	...	+1	+1	+1	+1	...
r_t	0	0	...	0	0	0	0	...	0	1	0	0	...

every other second MDP:

s_t	0	−1	...	−8	−9	−10	−9	...	8	9	10	10	...
a_t	−1	−1	...	−1	−1	+1	+1	...	+1	+1	+1	+1	...
r_t	0	0	...	0	1	0	0	...	0	0	0	0	...

- deterministic state/action sequence, stochastic reward sequence
- total expected reward: 1

Value Function and Markov Decision Problem

- to evaluate the quality of a policy,
  the reward process usually is aggregated by a
  scalar performance criterion.
  - e.g., expected sum, expected average, expected discounted sum

- each policy π is then described by a
  reward value for each initial state s_0,
  called **value function**:

$$V^\pi : S \rightarrow \mathbb{R}$$

$$\text{e.g., } V^\pi(s) := E\Big(\sum_{t=0}^{\infty} r_t \;\Big|\; \begin{array}{l} s_0 := s, \\ a_t \sim \pi(s_t), \\ r_t := r(s_t, a_t) \\ s_{t+1} \sim p(s_{t+1} \mid s_t, a_t), \end{array} \Big)$$

- **Markov Decision Problem**: find the optimal policy π^* with

$$V^{\pi^*}(s) \geq V^\pi(s) \quad \forall s \in S, \; \forall \pi \in \Pi$$

Markov Decision Problem

Given an MDP (p, r) and
a value criterion $V : \mathbb{R}^* \to \mathbb{R}$ that aggregates rewards

find a policy

$$\pi^* : S \to A$$

s.t. the expected value is maximial, i.e.,

$$V^{\pi^*}(s) \geq V^{\pi}(s), \quad \forall s \in S, \pi \in \Pi$$

$$\text{with } V^{\pi}(s) := \mathbb{E}(V((r_t)_{t \in \mathbb{N}}) \mid \begin{array}{l} s_0 := s, \\ a_t := \pi(s_t), \\ r_t = r(s_t, a_t), \\ s_{t+1} \sim p(s_{t+1} \mid s_t, a_t) \end{array})$$

Outline

Value Function for the Finite Criterion

$$V_N^\pi(s) := \mathbb{E}(\sum_{t=0}^{N-1} r_t \mid s_0 = s) = \mathbb{E}(r_0 + r_1 + r_2 + \ldots + r_{N-1} \mid s_0 = s)$$

- assumes that the process has to finish within finite horizon of N steps

Value Function for the Discounted Criterion

$$V_\gamma^\pi(s) := \mathbb{E}(\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s) = \mathbb{E}(r_0 + \gamma r_1 + \gamma^2 r_2 \ldots + \gamma^t r_t + \ldots \mid s_0 = s)$$

- infinite horizon
- assumes that future rewards are discounted by factor $\gamma \in (0, 1)$, e.g., $\gamma := 1/(1 + \text{inflation rate})$ for monetary rewards

Value Function for the Total Reward Criterion

$$V^\pi(s) := \mathbb{E}(\sum_{t=0}^{\infty} r_t \mid s_0 = s) = \mathbb{E}(r_0 + r_1 + r_2 \ldots + r_t + \ldots \mid s_0 = s)$$

- assumes that rewards can be summed infinitely, e.g.,
  - because they shrink quickly enough (like discounting enforces)
  - because they eventually become 0 (as a goal has been reached)
    - finite, but unknown horizon; optimal stopping
  - etc.

Value Function for the Average Criterion

$$V^{\pi}_{\text{avg}}(s) := \lim_{N \to \infty} \frac{1}{N} \mathbb{E}(\sum_{t=0}^{N-1} r_t \mid s_0 = s)$$
$$= \lim_{N \to \infty} \frac{1}{N} \mathbb{E}(r_0 + r_1 + r_2 + \ldots + r_{N-1} \mid s_0 = s)$$

- measures average reward per step
  - in a potentially infinite horizon

Performance Criteria

finite criterion with length N:

$$\mathbb{E}(\sum_{t=0}^{N-1} r_t \mid s_0) = \mathbb{E}(r_0 + r_1 + r_2 + \ldots + r_{N-1} \mid s_0)$$

discounted criterion with factor $\gamma \in (0, 1)$:

$$\mathbb{E}(\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0) = \mathbb{E}(r_0 + \gamma r_1 + \gamma^2 r_2 \ldots + \gamma^t r_t + \ldots \mid s_0)$$

total reward criterion:

$$\mathbb{E}(\sum_{t=0}^{\infty} r_t \mid s_0) = \mathbb{E}(r_0 + r_1 + r_2 \ldots + r_t + \ldots \mid s_0)$$

average criterion:

$$\lim_{N\to\infty} \frac{1}{N} \mathbb{E}(\sum_{t=0}^{N-1} r_t \mid s_0) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}(r_0 + r_1 + r_2 + \ldots + r_{N-1} \mid s_0)$$

Performance Criteria / Properties

1. performance criteria are additive in r_t
2. performance criteria are expectations over the policy-specific reward process

$\rightsquigarrow$ Bellman optimality principle:
all sub-policies of an optimal policy are optimal sub-policies.

Outline

Equivalence of Stochastic Markov Policies and History-dependent Policies

For

- any MDP (p, r),
- any value criterion V
  (either finite, discounted, total reward or average criterion), and
- any stochastic **history-dependent** policy π

there exists an equivalent

(generally non-stationary) stochastic **Markov** policy π',

i.e., with the same value function:

$$V^{\pi'}(s) = V^{\pi}(s), \quad \forall s \in S$$

Equivalence of … / Proof

Denote marginals as

$$P^\pi(a_t = a \mid s_t = s', s_0 = s) := \frac{\sum_{s_{1:t-1} \in S^{t+1}, a_{0:t-1} \in A^t} p^\pi(a_t = a \mid s_t = s', s_{1:t-1}, a_{0:t-1}, s_0 = s)}{\sum_{s_{1:t-1} \in S^{t+1}, a_{0:t-1} \in A^t, a' \in A} p^\pi(a_t = a' \mid s_t = s', s_{1:t-1}, a_{0:t-1}, s_0 = s)}$$

Define (a generally non-stationary) π' via

$$\pi'(a_t = a \mid s_t = s') := P^\pi(a_t = a \mid s_t = s', s_0 = s)$$

and show

$$P^{\pi'}(s_t = s', a_t = a \mid s_0 = s) = P^\pi(s_t = s', a_t = a \mid s_0 = s)$$

via induction over t:

- $t = 0$: clear.

Equivalence of . . . / Proof

show

$$P^{\pi'}(s_t = s', a_t = a \mid s_0 = s) = P^{\pi}(s_t = s', a_t = a \mid s_0 = s)$$

via induction over t:

- $t > 0$:

$$\begin{aligned}
&P^{\pi}(s_t = s' \mid s_0 = s) \\
&= \sum_{\tilde{s}\in S, \tilde{a}\in A} P^{\pi}(s_{t-1} = \tilde{s}, a_{t-1} = \tilde{a} \mid s_0 = s)\ p(s' \mid \tilde{s}, \tilde{a}) \\
&\overset{\text{ind.ass.}}{=} \sum_{\tilde{s}\in S, \tilde{a}\in A} P^{\pi'}(s_{t-1} = \tilde{s}, a_{t-1} = \tilde{a} \mid s_0 = s)\ p(s' \mid \tilde{s}, \tilde{a}) \\
&= P^{\pi'}(s_t = s' \mid s_0 = s)
\end{aligned}$$

Equivalence of . . . / Proof

$$
\begin{aligned}
&P^{\pi'}(s_t = s', a_t = a \mid s_0 = s) \\
&= P^{\pi'}(a_t = a \mid s_t = s')\, P^{\pi'}(s_t = s' \mid s_0 = s) \\
&= P^{\pi}(a_t = a \mid s_t = s', s_0 = s)\, P^{\pi}(s_t = s' \mid s_0 = s) \\
&= P^{\pi}(s_t = s', a_t = a \mid s_0 = s)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}(r(s_t, a_t) \mid s_0 = s, \pi) &= \sum_{s' \in S, a \in A} r(s', a)\; P^{\pi}(s_t = s', a_t = a \mid s_0 = s) \\
&= \sum_{s' \in S, a \in A} r(s', a)\; P^{\pi'}(s_t = s', a_t = a \mid s_0 = s) \\
&= \mathbb{E}(r(s_t, a_t) \mid s_0 = s, \pi')
\end{aligned}
$$

and thus

$$V^{\pi}(s) = V^{\pi'}(s)$$

Markov State Process

Although an MDP is Markov,
the stochastic state process s_t by a policy not necessary will be Markov.

For an

- MDP (p, r) and
- a **Markov** policy π,

the stochastic state process s_t is Markov with transition matrix

$$P_{\pi,s,s'} := P^\pi(s_{t+1} = s' \mid s_t = s) = \sum_{a \in A} \pi(s, a)\, p(s' \mid s, a)$$

Proof:

$$\begin{aligned} P^\pi(s_{t+1} \mid s_{0:t}) = & \sum_{a \in A} P^\pi(a_t = a \mid s_{0:t})\, P^\pi(s_{t+1} \mid s_{0:t}, a_t = a) \\ = & \sum_{a \in A} \pi(s_t, a)\, p(s_{t+1} \mid s_t, a) \\ = & P^\pi(s_{t+1} \mid s_t) \end{aligned}$$

Valued Markov Processes

Such a Markov state process together with its rewards

$$r_\pi(s) := \sum_{a \in A} \pi(s,a)\, r(s,a)$$

also is called **Valued Markov Process**

Outline

Optimal Policy for the Finite Criterion

optimal value for the last n steps:

$$V_n^*(s) := \max_{a_{N-n},\, a_{N-n+1},\ldots,a_{N-1}} \mathbb{E}(r_{N-n} + r_{N-n+1} + \ldots + r_{N-1} \mid s_{N-n} = s)$$

$$V_1^*(s) := \max_{a_{N-1}} \mathbb{E}(r_{N-1} \mid s_{N-1} = s) = \max_a r_{N-1}(s, a)$$

Optimal Policy for the Finite Criterion

optimal value for the last n steps:

$$V_n^*(s) := \max_{a_{N-n},\, a_{N-n+1},\ldots,a_{N-1}} \mathbb{E}(r_{N-n} + r_{N-n+1} + \ldots + r_{N-1} \mid s_{N-n} = s)$$

$$V_1^*(s) := \max_{a_{N-1}} \mathbb{E}(r_{N-1} \mid s_{N-1} = s) = \max_a r_{N-1}(s, a)$$

$$\begin{aligned}
V_2^*(s) &:= \max_{a_{N-2},a_{N-1}} \mathbb{E}(r_{N-2} + r_{N-1} \mid s_{N-2} = s) \\
&= \max_{a_{N-2}} r_{N-2}(s, a_{N-2}) + \max_{a_{N-1}} \mathbb{E}(r_{N-1} \mid s_{N-2} = s) \\
&= \max_{a_{N-2}} r_{N-2}(s, a_{N-2}) + \sum_{s'} p(s' \mid s, a_{N-2}) \max_{a_{N-1}} \mathbb{E}(r_{N-1} \mid s_{N-1} = s') \\
&= \max_a r_{N-2}(s, a) + \sum_{s'} p(s' \mid s, a) V_1^*(s')
\end{aligned}$$

Optimal Policy for the Finite Criterion

optimal value for the last n steps:

$$V_n^*(s) := \max_{a_{N-n}, a_{N-n+1}, \ldots, a_{N-1}} \mathbb{E}(r_{N-n} + r_{N-n+1} + \ldots + r_{N-1} \mid s_{N-n} = s)$$

$$V_1^*(s) := \max_{a_{N-1}} \mathbb{E}(r_{N-1} \mid s_{N-1} = s) = \max_a r_{N-1}(s, a)$$

$$\begin{aligned} V_2^*(s) &:= \max_{a_{N-2}, a_{N-1}} \mathbb{E}(r_{N-2} + r_{N-1} \mid s_{N-2} = s) \\ &= \max_{a_{N-2}} r_{N-2}(s, a_{N-2}) + \max_{a_{N-1}} \mathbb{E}(r_{N-1} \mid s_{N-2} = s) \\ &= \max_{a_{N-2}} r_{N-2}(s, a_{N-2}) + \sum_{s'} p(s' \mid s, a_{N-2}) \max_{a_{N-1}} \mathbb{E}(r_{N-1} \mid s_{N-1} = s') \\ &= \max_a r_{N-2}(s, a) + \sum_{s'} p(s' \mid s, a) V_1^*(s') \\ &\vdots \end{aligned}$$

$$V_{n+1}^*(s) = \max_a r_{N-1-n}(s, a) + \sum_{s'} p(s' \mid s, a) V_n^*(s')$$

Optimal Policy for the Finite Criterion

The optimal value functions $V^*_{1:N}$ (for remaining steps $n = 1 : N$) are the unique solutions of the set of equations

$$V^*_{n+1}(s) = \max_{a \in A} \left(r_{N-1-n}(s,a) + \sum_{s' \in S} p_{N-1-n}(s' \mid s,a) V^*_n(s') \right), \quad s \in S, \; n = 0 : N\text{–}1$$

$$V^*_0(s) := 0$$

from which an optimal (generally non-stationary) policy $\pi^*_{1:N}$ can be computed via

$$\pi^*_t(s) \in \arg\max_{a \in A} \left(r_t(s,a) + \sum_{s' \in S} p_t(s' \mid s,a)\, V^*_{N-1-t}(s') \right), \quad s \in S, \; t = 0 : N\text{–}1$$

Optimal Policy for the Finite Criterion / Proof

$$V^*_{n+1}(s) = \max_{a \in A} \left(r_{N-1-n}(s,a) + \sum_{s' \in S} p_{N-1-n}(s' \mid s,a) V^*_n(s') \right), \quad s \in S, \; n = 0 : N{-}1$$

For $n = 0$: just optimize reward of last step:

$$V^*_1(s) = \max_{a \in A} r_{N-1}(s,a), \quad s \in S$$

For $n > 0$:

- optimize sum of reward $r_{N-1-n}(s,a)$ of current step and
- reward $V^*_n(s')$ of future n steps from follow-up state s'
  - weighted by how likely an action will bring us to follow-up state s'

Optimal Policy for the Finite Criterion / Idea

- the optimal policy for the finite criterion can be computed recursively
  - backwards in time: $\pi_{N-1}, \pi_{N-2}, \ldots, \pi_0$
  - along with the value functions of remaining steps $V_1, V_2, \ldots, V_N$
- it can be chosen deterministic and Markov
  - but in general, not stationary

Find Optimal Policy for Finite Criterion

opt-policy-finite(p, r, S, A, N):
  for $s \in S : V_0(s) := 0$
  for $n := 0 : N-1$:
    $t := N - 1 - n$
    for $s \in S$:
      choose $a^* \in \arg\max_{a \in A} (r_t(s,a) + \sum_{s' \in S} p_t(s' \mid s, a)\, V^*_n(s'))$
      $\pi^*_t(s) := a^*$
      $V^*_{n+1}(s) := r_t(s, a^*) + \sum_{s' \in S} p_t(s' \mid s, a^*) V^*_n(s')$
  return V^*, π^*

Outline

L_π operator

Given

- an MDP (p, r),
- a discount factor $\gamma \in (0, 1)$ and
- a stationary Markov policy π,

define the L_π **operator** on value functions:

$$L_\pi V := r_\pi + \gamma P_\pi V, \quad V \in \mathbb{R}^S$$

Note: P_π and r_π are state transition matrix and rewards of the valued Markov process by π.

Value Function of the Discounted Criterion

Given

- an MDP (p, r),
- a discount factor $\gamma \in (0, 1)$ and
- a stationary Markov policy π,

then the value function V_γ^π is the only fixpoint of L_π

$$V = L_\pi V$$

and equivalently

$$V_\gamma^\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

Value Function of the Discounted Criterion / Proof

Stochastic matrix P_π has all eigenvalues ≤ 1,
$\rightsquigarrow \gamma P_\pi$ with $\gamma \in (0,1)$ has all eigenvalues < 1
$\rightsquigarrow I - \gamma P_\pi$ is invertible.

$$(I - \gamma P_\pi)^{-1} r_\pi = \sum_{k=0}^{\infty} \gamma^k P_\pi^k r_\pi$$

Remember, if $(I-A)^{-1}$ exists, then

$$(I-A)^{-1} = \sum_{k=0}^{\infty} A^k$$

simply as

$$(I-A)\sum_{k=0}^{\infty} A^k = \sum_{k=0}^{\infty} A^k - \sum_{k=1}^{\infty} A^k = I$$

Value Function of the Discounted Criterion / Proof

On the other hand,

$$
\begin{aligned}
V^\pi_\gamma(s) &= \mathbb{E}(\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s) \\
&= \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}(r(s_t, a_t) \mid s_0 = s) \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s' \in S, a \in A} P^\pi(s_t = s', a_t = a \mid s_0 = s) \, r(s', a) \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s' \in S, a \in A} \pi(s', a) \, P^\pi(s_t = s' \mid s_0 = s) \, r(s', a) \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s' \in S} P^\pi(s_t = s' \mid s_0 = s) \, r_\pi(s') \\
&= (\sum_{t=0}^{\infty} \gamma^t P^t_\pi r_\pi)(s)
\end{aligned}
$$

Bellman Equation

Given

- an MDP (p, r) and
- a discount factor $\gamma \in (0, 1)$,

define the **dynamic programming operator** L on value functions:

$$LV := \max_{\pi \in \Pi^{\text{MAS}}} L_\pi V = \max_{\pi \in \Pi^{\text{MAS}}} (r_\pi + \gamma P_\pi V), \quad V \in \mathbb{R}^S$$

Theorem (Bellman equation)

The optimal value functions V^*_γ are the only fixpoints of L:

$$LV = V$$

or equivalently

$$V(s) = \max_{a \in A} \left(r(s, a) + \gamma \sum_{s' \in S} p(s' \mid s, a)\, V(s') \right)$$

Bellman equation / Proof (Overview)

In 5 steps:

1. stationary Markov policies maximizing the one step expected reward yield the same value, wether they are deterministic or stochastic:

$$\max_{\pi \in \Pi^{\text{MDS}}} (r_\pi + \gamma P_\pi V) = \max_{\pi \in \Pi^{\text{MAS}}} (r_\pi + \gamma P_\pi V)$$

2. value functions being shrunken by L, upper bound the optimal value function:

$$LV \leq V \Rightarrow V_\gamma^* \leq V$$

3. value functions being inflated by L, lower bound the optimal value function:

$$LV \geq V \Rightarrow V \leq V_\gamma^*$$

4. thus, any fixpoint of L is an optimal value function.

5. L has fixpoints (because it is a contraction)

Find Optimal Policy for Discounted Criterion / LP

- There are several algorithms to find optimal policies for the discounted criterion.
- The 3 most important:
  1. via a linear program (LP)
  2. **value iteration**
  3. **policy iteration**
- idea of 1. via a linear program:
  - optimize over all value functions being upper bounds of V^*_γ
    - can be encoded via constraints $V \geq LV$ (see proof step 2 of Bellman equation)
  - within upper bounds, optimal policies minimize $\sum_{s \in S} V(s)$

$$\min_{V \in \mathbb{R}^S} \sum_{s \in S} V(s)$$

$$\text{s.t.} \quad V(s) \geq r(s,a) + \gamma \sum p(s' \mid s,a)\, V(s') \quad \forall s \in S, a \in A$$

Find Optimal Policy for Discounted Criterion / LP

opt-policy-discounted-lp(p, r, S, A, γ):

$$V^*_\gamma = \text{argmin-solve-lp}_{V \in \mathbb{R}^S} \sum_{s \in S} V(s)$$
$$\text{s.t.} \quad V(s) \geq r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s, a)\, V(s') \quad \forall s \in S, a \in A$$

for $s \in S$:

choose $\pi^*(s) \in \arg\max_{a \in A} (r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s, a)\, V^*_\gamma(s'))$

return V^*_γ, π^*

Find Opt. Policy for Discounted Criterion / Value Iteration

- idea: iterate fixpoint equation for the optimal value function:

$$V^{(n+1)} := LV^{(n)}$$

- works from any initialization $V^{(0)}$
- stop once $||V^{(n+1)} - V^{(n)}|| < \epsilon$
  for some prescribed threshold ϵ
- variants:
  - use already computed $V^{(n+1)}(s)$ to compute $V^{(n+1)}(s')$
    (instead of $V^{(n)}(s)$; called Gauss-Seidel)
  - reestimate $V(s)$ in random order of s
    (called asynchronous dynamic programming)
  - reestimate $V(s)$ proportional to their last change
    (also: prune some states s)

Find Opt. Policy for Discounted Criterion / Value Iteration

opt-policy-discounted-value-iteration$(p, r, S, A, \gamma, \epsilon)$:
  initialize $V^{(0)}$ arbitrarily
  $n := 0$
  repeat
    $n := n + 1$
    for $s \in S$:
      $V^{(n)}(s) := \max_{a \in A}(r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s, a)\, V^{(n-1)}(s'))$
  until $||V^{(n)} - V^{(n-1)}|| < \epsilon$
  $V^*_\gamma := V^{(n)}$
  for $s \in S$:
    choose $\pi^*(s) \in \arg\max_{a \in A}(r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s, a)\, V^*_\gamma(s'))$
  return V^*_γ, π^*

Find Opt. Policy for Discounted Criterion / Policy Iteration

One step look-ahead policy improvement:
Let $\pi \in \Pi^{MAS}$. Then the one step look-ahead policy π'

$$\pi' \in \arg\max_{\pi' \in \Pi^{MAS}} (r_{\pi'} + \gamma P_{\pi'} V_\gamma^\pi)$$

has a value function $V_\gamma^{\pi'}$ that upper bounds / improves π:

$$V_\gamma^{\pi'} \geq V_\gamma^\pi$$

without improvement only if π was already optimal ($V_\gamma^{\pi'} = V_\gamma^\pi$ iff $\pi = \pi^*$).

Find Opt. Policy for Discounted Criterion / Policy Iteration

opt-policy-discounted-policy-iteration(p, r, S, A, γ):
  initialize $\pi^{(0)}$ arbitrarily
  $n := 0$
  repeat
    $V^{(n)} := (I - \gamma P_{\pi^{(n)}})^{-1} r_{\pi^{(n)}}$
    for $s \in S$:
      choose $\pi^{(n+1)}(s) \in \arg\max_{a \in A}(r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s, a) V^{(n)}(s'))$
    $n := n + 1$
  until $\pi^{(n)} = \pi^{(n-1)}$
  return $V^{(n-1)}, \pi^{(n)}$

Outline

Value Functions for Total Reward

- for total reward, value functions are limits.
- for some MDPs these limits may not exist.
  - example:

$$S := \{1,2\}, \quad P := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad A := \{1\}, \quad r_1 := \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$\pi := \{(1,1),(2,1)\}$$

$$r_t = 1, -1, 1, -1, 1, -1, \ldots \text{ whose sum does not converge}$$

- specific conditions on the MDP are required for these limits to exist
  - positive MDPs
  - negative MDPs

Positive MDPs

An MDP (p, r) is called **positive**, if

i) for all states there exists an action with non-negative reward and

ii) for all policies the **positive value function**

$$V_+^\pi(s) := \mathbb{E}(\sum_{t=0}^{\infty} \max(0, r_t) \mid s_0 = s)$$

is finite for all states.

Optimal Policies for Positive MDPs under Total Reward

Let operators L_π and L be defined as before (for $\gamma := 1$).
Given a positive MDP (p, r),

i) V^π is the minimum solution of $V = L_\pi V$ in $(\mathbb{R}_0^+)^S$
(for all policies $\pi \in \Pi^{\text{MDS}}$).

ii) V^* is the minimum solution of $V = LV$ in $(\mathbb{R}_0^+)^S$.

iii) $\pi \in \Pi^{\text{HA}}$ optimal iff V^π is a fixpoint of $V = LV$.

iv) if $\pi \in \arg\max_{\pi \in \Pi^{\text{MA}}}(r_\pi + P_\pi V^*)$
and $\lim_{N\to\infty} P_\pi^N V^*(s) = 0$ for all states s,
then π is optimal.

Find an Optimal Policy for Positive MDPs under Total Reward

- value iteration:
  - converges monotonously to V^* if $0 \leq V_0 \leq V^*$
    - e.g., $V_0 := 0$ will do.
- policy iteration:
  - ensure that its value function stays in $(\mathbb{R}_0^+)^S$
  - force $V^{(n)}(s) := 0$ for all recurrent states s in Markov chain $P_{\pi^{(n)}}$

Find Opt. Policy for Total Reward Criterion, Positive MDP / Policy Iteration

opt-policy-total-pos-policy-iteration(p, r, S, A):
  initialize $\pi^{(0)}$ s.t. $r_{\pi^{(0)}} \geq 0$
  $n := 0$
  repeat
    $V^{(n)} :=$ minimum solution of
      $V^{(n)}(s) = r(s, \pi^{(n)}(s)) + \sum_{s' \in S} p(s' \mid s, \pi^{(n)}(s)) V^{(n)}(s')$
    for $s \in S$:
      choose $\pi^{(n+1)}(s) \in \arg\max_{a \in A}(r(s, a) + \sum_{s' \in S} p(s' \mid s, a) V^{(n)}(s'))$
        (choose $\pi^{(n+1)}(s) = \pi^{(n)}(s)$ if it is still among maximal actions)
    $n := n + 1$
  until $\pi^{(n)} = \pi^{(n-1)}$
  return $V^{(n-1)}, \pi^{(n)}$

Negative MDPs

An MDP (p, r) is called **negative**, if

i) all rewards are negative and

ii) there exists a policy with value function having finite values for all states.

Summary

- **Markov Decision Processes** (**MDPs**) describe Markov processes that
  - can be controlled/manipulated by **actions**/**decisions**
  - yield **rewards** depending on current state and action.
- A **policy** describes which action to choose in which situation.
  - **Markov policy**: depends only on current state, not on history.
  - **stationary**: does not depend on current time.
  - **deterministic policy**: choose a single action, not stochastic.
- An MDP, a start state and a policy define three **stochastic processes for states, actions and rewards**.
- A **performance criterion** describes how to aggregate a stochastic reward process to a scalar **value**.
  - sum, sum of first N, discounted sum, average
  - expectation
  - called **total reward**, **finite**, **discounted**, **average** criterion.

Summary (2/3)

- The **value function** of a policy gives the value for a policy for each start state.
- The **Markov Decision Problem**, **to find the optimal policy** for an MDP, is formalized as finding a policy with maximal value function (for all states).
- Optimal policies for these four criteria always can be chosen Markov.
  - no need for history-dependent policies.
  - but they are non-stationary in general.
- The state process of an MDP under a Markov policy is Markov
  - together with the reward process called **Valued Markov Process**.

Summary (3/3)

- Criteria and algorithms for optimal policies differ depending on the criterion.
- For the finite criterion, an optimal policy can be computed through a simple recursive scheme backwards in time.
  - optimal policy can be chosen deterministic
  - but will in general be non-stationary.
- For the discounted criterion,
  - optimal policies are the fixpoints of the **dynamic programming operator** L (**Bellman equation**).
    - choose best policy according to one step look ahead and value function of the input policy.
  - via linear programming: find policy with maximal sum of values respecting Bellman equations.
  - **value iteration**: iterate dynamic programming operator on the value function.
  - **policy iteration**: iterate one step look ahead improvement of current policy.

Further Readings

- Markov decision processes:
  - Frederick Garica, Emmanuel Rachelson (2010): *Markov Decision Processes*, ch. 1 in **?**.

References