facebook
Artificial Intelligence Research

How to solve an MDP: Dynamic Programming

Matteo Pirotta
Facebook AI Research (on leave from Inria Lille)

Outline

How to solve *exactly* an MDP

Dynamic Programming

Bellman Equations

Value Iteration

Policy Iteration

The Optimization Problem

$$\max_{\pi} \; V^{\pi}(s_0) =$$

$$= \max_{\pi} \; \mathbb{E}\big[r(s_0, d_0(a_0|s_0)) + \gamma r(s_1, d_1(a_1|s_0, s_1)) + \gamma^2 r(s_2, d_2(a_2|s_0, s_1, s_2)) + \dots\big]$$

$$= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | a_t \sim d_t(\cdot|h_t)\right]$$

Plan to Simplify the Optimization Problem

1. Reduce the search space
   - i. History-based $\Rightarrow$ Markov decision rules
   - ii. Non-stationary $\Rightarrow$ Stationary policies
   - $\Rightarrow$ Focus on *stationary policies with Markov decision rules*
2. Leverage Markov property of the MDP to "simplify" the value function
3. Stochastic $\Rightarrow$ Deterministic decision rules

$\Rightarrow$ Focus on stationary policies with *deterministic* Markov decision rules

From History-Based to Markov Policies

Theorem (Bertsekas [2007])

Consider an MDP with $|A| < \infty$ and an initial distribution β over states such that $\left|\{s \in S : \beta(s) > 0\}\right| < \infty$. For any policy π, let

$$p_t^{\pi}(s,a) = \mathbb{P}\left[S_t = s, A_t = a\right]; \qquad p_t^{\pi}(s) = \mathbb{P}\left[S_t = s\right].$$

Then for any history-based policy π there exists a Markov policy $\overline{\pi}$ such that

$$p_t^{\overline{\pi}}(s,a) = p_t^{\pi}(s,a); \qquad p_t^{\overline{\pi}}(s) = p_t^{\pi}(s)$$

for any $s \in S$, $a \in A$ and $t \in \mathbb{N}^+$.

Markov policies are as “expressive” as history-based policies

Proof: From History-Based to Markov Policies

For any $\pi = (d_0, d_1, \ldots)$ with d_t a randomized history-dependent decision rule, let $\overline{\pi} = (\overline{d}_0, \overline{d}_1, \ldots)$ be a randomized Markov policy such that

$$\overline{d}_t(a|s) = \frac{p_t^{\pi}(s,a)}{p_t^{\pi}(s)}$$

Base case. For any s, $p_0^{\overline{\pi}}(s) = p_0^{\pi}(s)$ by definition. And

$$p_0^{\overline{\pi}}(s,a) = p_0^{\overline{\pi}}(s)\overline{d}_0(a|s) = p_0^{\overline{\pi}}(s)\frac{p_0^{\pi}(s,a)}{p_0^{\pi}(s)} = p_0^{\overline{\pi}}(s)\frac{p_0^{\pi}(s,a)}{p_0^{\overline{\pi}}(s)} = p_0^{\pi}(s,a)$$

Proof: From History-Based to Markov Policies

Induction. For any s and some $t > 0$, $p_t^{\overline{\pi}}(s) = p_t^{\pi}(s)$ and $p_t^{\overline{\pi}}(s,a) = p_t^{\pi}(s,a)$ by inductive assumption. Then

$$
\begin{aligned}
p_{t+1}^{\overline{\pi}}(s_{t+1}) &= \sum_{s_t,a_t} p_t^{\overline{\pi}}(s_t,a_t) p(s_{t+1}|s_t,a_t) \\
&= \sum_{s_t,a_t} p_t^{\overline{\pi}}(s_t) \overline{d}_t(a_t|s_t) p(s_{t+1}|s_t,a_t) \\
&= \sum_{s_t,a_t} p_t^{\overline{\pi}}(s_t) \frac{p_t^{\pi}(s_t,a_t)}{p_t^{\pi}(s_t)} p(s_{t+1}|s_t,a_t) \\
&= \sum_{s_t,a_t} p_t^{\overline{\pi}}(s_t) \frac{p_t^{\pi}(s_t,a_t)}{p_t^{\overline{\pi}}(s_t)} p(s_{t+1}|s_t,a_t) \\
&= \sum_{s_t,a_t} p_t^{\pi}(s_t,a_t) p(s_{t+1}|s_t,a_t) \\
&= p_{t+1}^{\pi}(s_{t+1})
\end{aligned}
$$

Similar for $p_{t+1}^{\overline{\pi}}(s_{t+1},a_{t+1}) = p_{t+1}^{\pi}(s_{t+1},a_{t+1})$.

The Discounted Occupancy Measure

$$
\begin{aligned}
V^{\pi}(s) &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, d_t(s_t))\right] \\
&= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[r(s_t, d_t(s_t))\right] \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s,a} \mathbb{P}\left[S_t = s, A_t = a\right] r(s,a) \\
&= \frac{1}{1-\gamma} \sum_{s,a} (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}\left[S_t = s, A_t = a\right] r(s,a) \\
&= \frac{1}{1-\gamma} \sum_{s,a} \rho_{\gamma}^{\pi}(s,a) r(s,a)
\end{aligned}
$$

From Non-Stationary to Stationary Policies

Theorem (?)

Consider an MDP with $|A| < \infty$ *and an initial distribution* β *over states such that* $\left|\{s \in S : \beta(s) > 0\}\right| < \infty$.
Then for any non-stationary policy π *there exists a stationary policy* $\overline{\pi}$ *such that*

$$\rho_\gamma^{\overline{\pi}}(s,a) = \rho_\gamma^{\pi}(s,a); \qquad \rho_\gamma^{\overline{\pi}}(s) = \rho_\gamma^{\pi}(s)$$

for any $s \in S$, $a \in A$ *and* $t \in \mathbb{N}^+$.

- Stationary policies are as "expressive" as non-stationary policies
- Stationary policies can "generate" any value function

Proof: From Non-Stationary to Stationary Policies

State discounted occupancy measure for stationary policy $\overline{\pi}$ (with Markov decision rules)

$$
\begin{aligned}
\rho_\gamma^{\overline{\pi}}(s) &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}\left[S_t = s\right] \\
&= (1-\gamma)\beta(s) + (1-\gamma)\sum_{t=1}^{\infty}\gamma^t \mathbb{P}\left[S_t = s\right] \\
&= (1-\gamma)\beta(s) + (1-\gamma)\gamma\sum_{t=1}^{\infty}\gamma^{t-1}\sum_{s'}\sum_{a}\mathbb{P}\left[S_{t-1} = s', A_{t-1} = a\right]p(s|s',a) \\
&= (1-\gamma)\beta(s) + \gamma\sum_{s'}(1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1}\mathbb{P}\left[S_{t-1} = s'\right]\sum_{a}\overline{\pi}(a|s')p(s|s',a) \\
&= (1-\gamma)\beta(s) + \gamma\sum_{s'}(1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1}\mathbb{P}\left[S_{t-1} = s'\right]p^{\overline{\pi}}(s|s') \\
&= (1-\gamma)\beta(s) + \gamma\sum_{s'}\rho_\gamma^{\overline{\pi}}(s')p^{\overline{\pi}}(s|s')
\end{aligned}
$$

Proof: From Non-Stationary to Stationary Policies

Moving to *matrix* formulation

$$[\boldsymbol{\rho}_\gamma^{\overline{\pi}}]_s = \rho_\gamma^{\overline{\pi}}(s)$$
$$[P^{\overline{\pi}}]_{s,s'} = p^{\overline{\pi}}(s'|s)$$

$$\rho_\gamma^{\overline{\pi}}(s) = (1-\gamma)\beta(s) + \gamma \sum_{s'} \rho_\gamma^{\overline{\pi}}(s') p^{\overline{\pi}}(s|s')$$
$$\Rightarrow \boldsymbol{\rho}_\gamma^{\overline{\pi}} = (1-\gamma)\beta(s) + \gamma P^{\overline{\pi}} \boldsymbol{\rho}_\gamma^{\overline{\pi}}$$
$$\Rightarrow \boldsymbol{\rho}_\gamma^{\overline{\pi}} = (1-\gamma)\beta(s)\left(I - \gamma P^{\overline{\pi}}\right)^{-1}$$

Proof: From Non-Stationary to Stationary Policies

Moving to *matrix* formulation

$$[\boldsymbol{\rho}_\gamma^{\overline{\pi}}]_s = \rho_\gamma^{\overline{\pi}}(s)$$
$$[P^{\overline{\pi}}]_{s,s'} = p^{\overline{\pi}}(s'|s)$$

$$\rho_\gamma^{\overline{\pi}}(s) = (1-\gamma)\beta(s) + \gamma \sum_{s'} \rho_\gamma^{\overline{\pi}}(s') p^{\overline{\pi}}(s|s')$$
$$\Rightarrow \boldsymbol{\rho}_\gamma^{\overline{\pi}} = (1-\gamma)\beta(s) + \gamma P^{\overline{\pi}} \boldsymbol{\rho}_\gamma^{\overline{\pi}}$$
$$\Rightarrow \boldsymbol{\rho}_\gamma^{\overline{\pi}} = (1-\gamma)\beta(s)\big(I - \gamma P^{\overline{\pi}}\big)^{-1}$$

The eigenvalues of a stochastic matrix P^π all belongs to $[0,1]$. As a consequence, $-\gamma \notin span(P^\pi)$ thus $I - \gamma P^\pi$ is invertible.

Proof: From Non-Stationary to Stationary Policies

For any non-stationary policy π define a stationary policy $\overline{\pi}$

$$\overline{\pi}(a|s') = \frac{\rho_\gamma^\pi(s,a)}{\rho_\gamma^\pi(s)}$$

$$\begin{aligned}
\rho_\gamma^\pi(s) &= (1-\gamma)\beta(s) + \gamma \sum_{s'} \sum_{a} (1-\gamma) \sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{P}\left[S_{t-1} = s', A_{t-1} = a\right] p(s|s',a) \\
&= (1-\gamma)\beta(s) + \gamma \sum_{s'} \sum_{a} \rho_\gamma^\pi(s',a) p(s|s',a) \\
&= (1-\gamma)\beta(s) + \gamma \sum_{s'} \sum_{a} \overline{\pi}(a|s') \rho_\gamma^\pi(s') p(s|s',a) \\
&= (1-\gamma)\beta(s) + \gamma \sum_{s'} \rho_\gamma^\pi(s') \sum_{a} \overline{\pi}(a|s') p(s|s',a) \\
&= (1-\gamma)\beta(s) + \gamma \sum_{s'} \rho_\gamma^\pi(s') p^{\overline{\pi}}(s|s')
\end{aligned}$$

Proof: From Non-Stationary to Stationary Policies

Moving to the *matrix* formulation

$$\rho_\gamma^{\pi}(s) = (1-\gamma)\beta(s) + \gamma \sum_{s'} \rho_\gamma^{\pi}(s') p^{\overline{\pi}}(s|s')$$

$$\Rightarrow \boldsymbol{\rho}_\gamma^{\pi} = (1-\gamma)\beta(s)\big(I - \gamma P^{\overline{\pi}}\big)^{-1}$$

$$\Rightarrow \boldsymbol{\rho}_\gamma^{\pi} = \boldsymbol{\rho}_\gamma^{\overline{\pi}}$$

The Optimization Problem

$$\max_{\pi} \; V^{\pi}(x_0) =$$

$$= \max_{\pi} \; \mathbb{E}\big[r(s_0, d_0(a_0|s_0)) + \gamma r(s_1, d_1(a_1|s_0, s_1)) + \gamma^2 r(s_2, d_2(a_2|s_0, s_1, s_2)) + \ldots\big]$$

$$= \max_{\pi \in \Pi^{MRS}} \; \mathbb{E}\big[r(s_0, \pi(a_0, s_0)) + \gamma r(s_1, \pi(a_1|s_1)) + \gamma^2 r(s_2, \pi(a_2|s_2)) + \ldots\big]$$

The Optimization Problem

$$\max_{\pi} \; V^{\pi}(x_0) =$$

$$= \max_{\pi} \; \mathbb{E}\big[r(s_0, d_0(a_0|s_0)) + \gamma r(s_1, d_1(a_1|s_0, s_1)) + \gamma^2 r(s_2, d_2(a_2|s_0, s_1, s_2)) + \dots\big]$$

$$= \max_{\pi \in \Pi^{MRS}} \; \mathbb{E}\big[r(s_0, \pi(a_0, s_0)) + \gamma r(s_1, \pi(a_1|s_1)) + \gamma^2 r(s_2, \pi(a_2|s_2)) + \dots\big]$$

Even if we restrict to deterministic policies we still have $|A|^{|S|}$ policies to check

Outline

The Bellman Equation

Theorem

For any stationary deterministic policy $\pi = (d, d, \dots)$, at any state $s \in S$, the state value function satisfies the Bellman equation:

$$V^\pi(s) = r(s, \pi(s)) + \gamma \sum_y p(y|s, \pi(s)) V^\pi(y).$$

Proof: The Bellman Equation

For any stationary policy $\pi = (d, d, \dots)$,

$$V^\pi(s) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)) \,|\, s_0 = s; \pi\Big]$$ *[value function]*

$$= r(s, \pi(s)) + \mathbb{E}\Big[\sum_{t=1}^{\infty} \gamma^t r(s_t, \pi(s_t)) \,|\, s_0 = s; \pi\Big]$$

$$= r(s, \pi(s))$$ *[Markov property]*

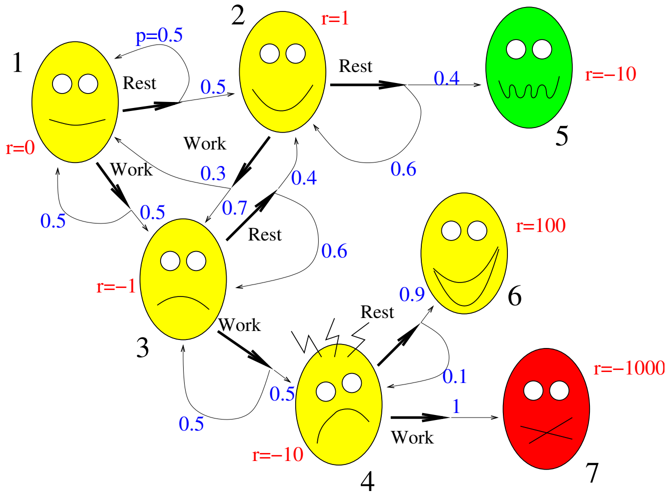
$$+ \gamma \sum_{s'} \mathbb{P}(s_1 = s' \,|\, s_0 = s; \pi(s_0)) \mathbb{E}\Big[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, \pi(s_t)) \,|\, s_1 = s'; \pi\Big]$$

$$= r(s, \pi(s))$$ *[MDP and change of "time"]*

$$+ \gamma \sum_{s'} p(s'|s, \pi(s)) \mathbb{E}\Big[\sum_{t'=0}^{\infty} \gamma^{t'} r(s_{t'}, \pi(s_{t'})) \,|\, s_{0'} = s'; \pi\Big]$$

$$= r(s, \pi(s)) + \gamma \sum_{s'} p(s'|s, \pi(s)) V^\pi(s')$$ *[value function]*

The student dilemma

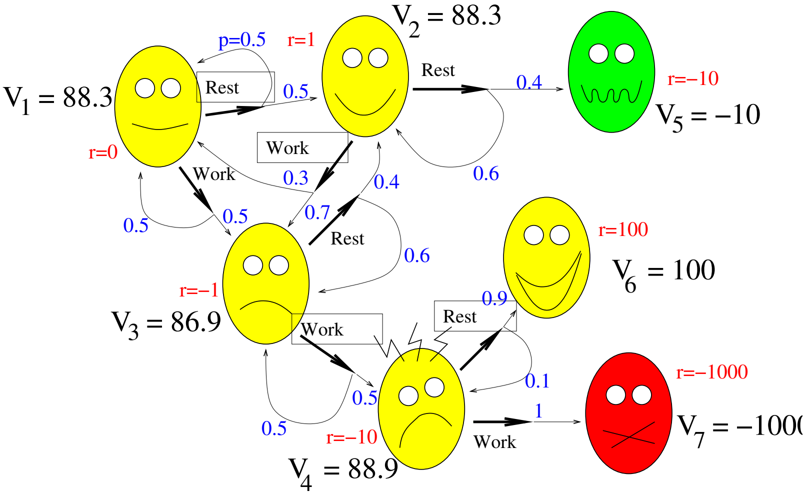


The student dilemma

- *Model*: all the transitions are Markov, states x_5, x_6, x_7 are terminal.
- *Setting*: infinite horizon with terminal states.
- *Objective*: find the policy that maximizes the expected sum of rewards before achieving a terminal state.

Notice: not a discounted infinite horizon setting! But the Bellman equations hold unchanged.

The student dilemma



The student dilemma

Computing V_4:

$$V_6 = 100$$
$$V_4 = -10 + (0.9V_6 + 0.1V_4)$$

$$\Rightarrow V_4 = \frac{-10 + 0.9V_6}{0.9} = 88.8$$

The student dilemma

Computing V_3: *no need* to consider all possible trajectories

$$V_4 = 88.8$$
$$V_3 = -1 + (0.5V_4 + 0.5V_3)$$

$$\Rightarrow V_3 = \frac{-1 + 0.5V_4}{0.5} = 86.8$$

The student dilemma

Computing V_3: *no need* to consider all possible trajectories

$$V_4 = 88.8$$
$$V_3 = -1 + (0.5V_4 + 0.5V_3)$$

$$\Rightarrow V_3 = \frac{-1 + 0.5V_4}{0.5} = 86.8$$

and so on for the rest...

Bellman Equation: a System of Equations

The Bellman equation

$$V^{\pi}(s) = r(s, \pi(s)) + \gamma \sum_{y} p(y|s, \pi(s)) V^{\pi}(y).$$

is a **linear** system of equations with $S = |\mathcal{S}|$ unknowns and S linear constraints.

Matrix notation

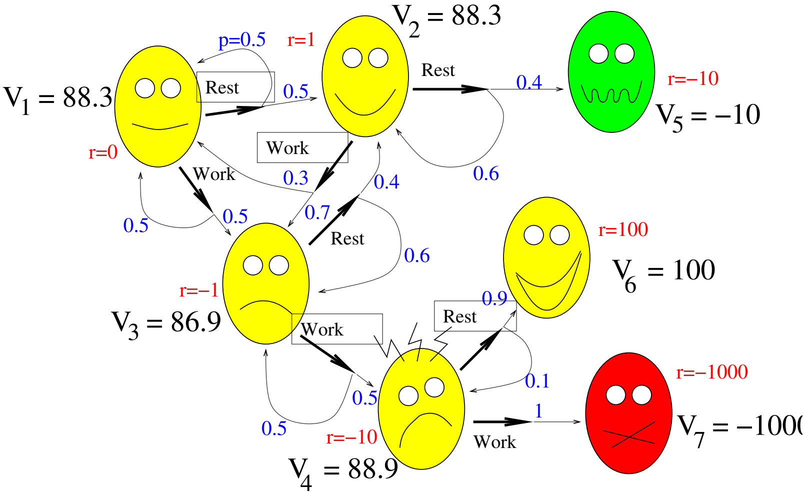
$$V^{\pi} \in \mathbb{R}^{S}, \quad r^{\pi} \in \mathbb{R}^{S}, \quad P^{\pi} \in \mathbb{R}^{S \times S}$$

then

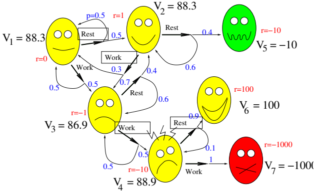
$$\begin{aligned} & V^{\pi} = r^{\pi} + \gamma P^{\pi} V^{\pi} \\ \implies \quad & V^{\pi} = (I - \gamma P^{\pi})^{-1} r^{\pi} \end{aligned}$$

👉 V^{π} can be compute inverting a $S \times S$ matrix ($O(S^3)$ time)

The student dilemma



The student dilemma



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

System of equations

$$\begin{cases} V_1 &= 0 + 0.5V_1 + 0.5V_2 \\ V_2 &= 1 + 0.3V_1 + 0.7V_3 \\ V_3 &= -1 + 0.5V_4 + 0.5V_3 \\ V_4 &= -10 + 0.9V_6 + 0.1V_4 \\ V_5 &= -10 \\ V_6 &= 100 \\ V_7 &= -1000 \end{cases} \Rightarrow$$

$$(V, R \in \mathbb{R}^7,\ P \in \mathbb{R}^{7\times 7})$$

$$V = R + PV$$

$$\Downarrow$$

$$V = (I - P)^{-1} R$$

Outline

The Optimal Bellman Equation

Bellman's Principle of Optimality Bellman [1957]:

> *"An optimal policy has the property that, whatever the initial state and the initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision."*

The Optimal Bellman Equation

Bellman's Principle of Optimality Bellman [1957]:

> *"An optimal policy has the property that, whatever the initial state and the initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision."*

$$V^\star = \max_{\pi \in \Pi^{\mathsf{MRS}}} V^\pi = \max_{\pi \in \Pi^{\mathsf{MRS}}} \{ r^\pi + \gamma P^\pi V^\pi \}$$

There always exists an optimal policy that is deterministic!

The Optimal Bellman Equation

Theorem

The optimal value function $V^\star$ (i.e., $V^\star = \max_\pi V^\pi$) is the solution to the optimal Bellman equation:

$$V^\star(s) = \max_{a \in A} \Big[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^\star(s') \Big].$$

and any optimal policy is such that

$$\pi^\star(a|s) \geq 0 \Leftrightarrow a \in \arg\max_{a' \in A} \Big[r(s,a') + \gamma \sum_{s'} p(s'|s,a) V^\star(s') \Big].$$

The Optimal Bellman Equation

Theorem

The optimal value function $V^\star$ (i.e., $V^\star = \max_\pi V^\pi$) is the solution to the optimal Bellman equation:

$$V^\star(s) = \max_{a \in A} \Big[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^\star(s') \Big].$$

and any optimal policy is such that

$$\pi^\star(a|s) \geq 0 \Leftrightarrow a \in \arg\max_{a' \in A} \Big[r(s,a') + \gamma \sum_{s'} p(s'|s,a) V^\star(s') \Big].$$

👍 There is always a deterministic policy

Proof: The Optimal Bellman Equation

For any policy $\pi = (a, \pi')$ (possibly non-stationary),

$$
\begin{aligned}
V^\star(x) &= \max_\pi \mathbb{E}\Big[\sum_{t \geq 0} \gamma^t r(x_t, \pi(x_t)) \,|\, x_0 = x; \pi\Big] \\
&= \max_{(a,\pi')} \Big[r(x,a) + \gamma \sum_y p(y|x,a) V^{\pi'}(y) \Big] \\
&= \max_a \Big[r(x,a) + \gamma \sum_y p(y|x,a) \max_{\pi'} V^{\pi'}(y) \Big] \\
&= \max_a \Big[r(x,a) + \gamma \sum_y p(y|x,a) V^\star(y) \Big].
\end{aligned}
$$

Proof: The Optimal Bellman Equation

We have

$$\max_{\pi'} \sum_y p(y|x,a) V^{\pi'}(y) \leq \sum_y p(y|x,a) \max_{\pi'} V^{\pi'}(y)$$

But, let $\overline{\pi}(y) = \arg\max_{\pi'} V^{\pi'}(y)$

$$\sum_y p(y|x,a) \max_{\pi'} V^{\pi'}(y) = \sum_y p(y|x,a) V^{\overline{\pi}}(y) \leq \max_{\pi'} \sum_y p(y|x,a) V^{\pi'}(y)$$

■

System of Equations

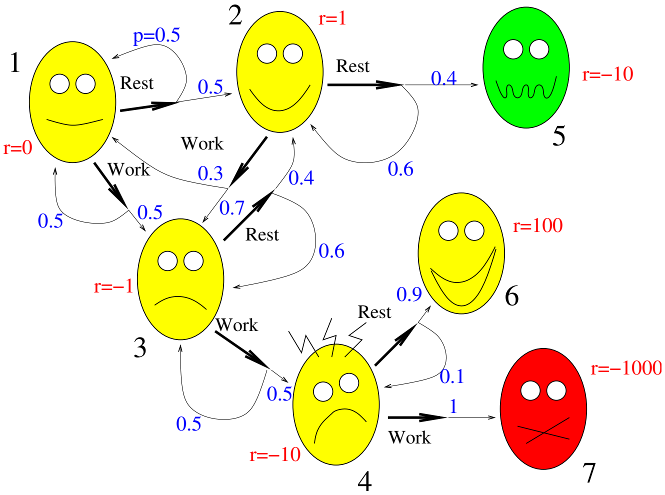
The optimal Bellman equation

$$V^\star(s) = \max_{a \in A} \Big[r(s,a) + \gamma \sum_{s'} p(y|s,a) V^\star(s') \Big].$$

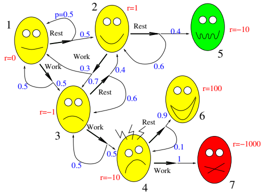
is a **non-linear** system of equations with N unknowns and N non-linear constraints (i.e., the $\max$ operator).

$\implies$ no simple matrix inversion technique ...

The Student Dilemma



The Student Dilemma



$$V^\star(x) = \max_{a \in A} \Big[r(x,a) + \gamma \sum_y p(y|x,a) V^\star(y) \Big]$$

System of equations

$$\begin{cases}
V_1 &= \max\{0 + 0.5V_1 + 0.5V_2;\ \ 0 + 0.5V_1 + 0.5V_3\} \\
V_2 &= \max\{1 + 0.4V_5 + 0.6V_2;\ \ 1 + 0.3V_1 + 0.7V_3\} \\
V_3 &= \max\{-1 + 0.4V_2 + 0.6V_3;\ \ -1 + 0.5V_4 + 0.5V_3\} \\
V_4 &= \max\{-10 + 0.9V_6 + 0.1V_4;\ \ -10 + V_7\} \\
V_5 &= -10 \\
V_6 &= 100 \\
V_7 &= -1000
\end{cases}$$

$\Rightarrow$ too complicated, we need to find an alternative solution.

The Bellman Operators

Notation. w.l.o.g. a discrete state space $|S| = N$ and $V^\pi \in \mathbb{R}^N$.

Definition

For any $W \in \mathbb{R}^N$, the *Bellman operator* $\mathcal{T}^\pi : \mathbb{R}^N \to \mathbb{R}^N$ is

$$\mathcal{T}^\pi W(s) = r(s, \pi(s)) + \gamma \sum_{s'} p(s'|s, \pi(s)) W(s'),$$

and the *optimal Bellman operator* (or dynamic programming operator) is

$$\mathcal{T} W(s) = \max_{a \in A} \big[r(s, a) + \gamma \sum_{s'} p(s'|s, a) W(s) \big].$$

Banach Fixed-Point Theorem

Review this theorem

The Bellman Operators

Proposition

Properties of the Bellman operators

1. *Monotonicity*: *for any* $W_1, W_2 \in \mathbb{R}^N$, *if* $W_1 \leq W_2$ *component-wise, then*

$$\begin{aligned} \mathcal{T}^\pi W_1 &\leq \mathcal{T}^\pi W_2, \\ \mathcal{T} W_1 &\leq \mathcal{T} W_2. \end{aligned}$$

The Bellman Operators

Proposition

Properties of the Bellman operators

1. *Monotonicity*: *for any* $W_1, W_2 \in \mathbb{R}^N$, *if* $W_1 \leq W_2$ *component-wise, then*

$$\begin{aligned} \mathcal{T}^\pi W_1 &\leq \mathcal{T}^\pi W_2, \\ \mathcal{T} W_1 &\leq \mathcal{T} W_2. \end{aligned}$$

2. *Additivity*: *for any scalar* $c \in \mathbb{R}$,

$$\begin{aligned} \mathcal{T}^\pi (W + cI_N) &= \mathcal{T}^\pi W + \gamma c I_N, \\ \mathcal{T} (W + cI_N) &= \mathcal{T} W + \gamma c I_N, \end{aligned}$$

The Bellman Operators

Proposition

3. *Contraction in L_∞-norm*: for any $W_1, W_2 \in \mathbb{R}^N$

$$\begin{aligned} \|\mathcal{T}^\pi W_1 - \mathcal{T}^\pi W_2\|_\infty &\leq \gamma \|W_1 - W_2\|_\infty, \\ \|\mathcal{T} W_1 - \mathcal{T} W_2\|_\infty &\leq \gamma \|W_1 - W_2\|_\infty. \end{aligned}$$

The Bellman Operators

Proposition

3. *Contraction in L_∞-norm: for any $W_1, W_2 \in \mathbb{R}^N$*

$$\begin{aligned} ||\mathcal{T}^\pi W_1 - \mathcal{T}^\pi W_2||_\infty &\leq \gamma ||W_1 - W_2||_\infty, \\ ||\mathcal{T} W_1 - \mathcal{T} W_2||_\infty &\leq \gamma ||W_1 - W_2||_\infty. \end{aligned}$$

4. *Fixed point: For any policy π*

V^π *is the unique fixed point of* $\mathcal{T}^\pi$ $\quad (V^\pi = \mathcal{T}^\pi V^\pi)$

$V^\star$ *is the unique fixed point of* $\mathcal{T}$ $\quad (V^\star = \mathcal{T} V^\star)$

The Bellman Operators

Proposition

3 *Contraction in L_∞-norm*: *for any $W_1, W_2 \in \mathbb{R}^N$*

$$\begin{aligned} ||\mathcal{T}^\pi W_1 - \mathcal{T}^\pi W_2||_\infty &\leq \gamma ||W_1 - W_2||_\infty, \\ ||\mathcal{T} W_1 - \mathcal{T} W_2||_\infty &\leq \gamma ||W_1 - W_2||_\infty. \end{aligned}$$

4 *Fixed point*: *For any policy π*

V^π *is the unique fixed point of* $\mathcal{T}^\pi$ $(V^\pi = \mathcal{T}^\pi V^\pi)$

$V^\star$ *is the unique fixed point of* $\mathcal{T}$ $(V^\star = \mathcal{T} V^\star)$

For any $W \in \mathbb{R}^N$ and any stationary policy π

$$\begin{aligned} \lim_{k\to\infty} (\mathcal{T}^\pi)^k W &= V^\pi \\ \lim_{k\to\infty} (\mathcal{T})^k W &= V^\star. \end{aligned}$$

Proof: Contraction of the Bellman Operator

For any $s \in S$

$$
\begin{aligned}
&|\mathcal{T}W_1(s) - \mathcal{T}W_2(s)| \\
&\quad = \left| \max_a \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) W_1(s') \right] - \max_{a'} \left[r(s,a') + \gamma \sum_{s'} p(s'|s,a') W_2(s') \right] \right| \\
&\quad \le \max_a \left| \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) W_1(s') \right] - \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) W_2(s') \right] \right| \\
&\quad = \gamma \max_a \sum_{s'} p(s'|s,a) |W_1(s') - W_2(s')| \\
&\quad \le \gamma \|W_1 - W_2\|_\infty \max_a \sum_{s'} p(s'|s,a) = \gamma \|W_1 - W_2\|_\infty,
\end{aligned}
$$

■

Same proof applies for $\mathcal{T}^\pi$

State-Action Value Function

Definition

In discounted infinite horizon problems, for any policy π, the *state-action value function* (or Q-function) $Q^\pi : S \times A \mapsto \mathbb{R}$ is

$$Q^\pi(s, a) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a, a_t = \pi(s_t), \forall t \geq 1\Big],$$

The optimal Q-function is

$$Q^\star(s, a) = \max_\pi Q^\pi(s, a),$$

Greedy Policy

The *greedy* policy with respect to a value $V \in \mathbb{R}^S$, is defined as

$$\pi(s) \in \arg\max_{a \in \mathcal{A}} \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V(s') \right]$$

The *greedy* policy with respect to a value $Q \in \mathbb{R}^{S \times A}$, is defined as

$$\pi(s) \in \arg\max_{a \in \mathcal{A}} Q(s,a)$$

from Bellman optimality equations

$$\pi^\star = \text{greedy}(V^\star) \quad \text{or} \quad \pi^\star = \text{greedy}(Q^\star)$$

State-Action Value Function Operators*

$$\mathcal{T}^{\pi} Q(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) Q(s,\pi(s))$$

$$\mathcal{T} Q(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_{a'} Q(s,a')$$

*Abuse of notation for the operators

State-Action and State Value Function

$$Q^\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^\pi(s')$$

$$V^\pi(s) = Q^\pi(s, \pi(s))$$

$$Q^\star(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^\star(s')$$

$$V^\star(s) = Q^\star(s, \pi^\star(s)) = \max_{a \in A} Q^\star(s,a)$$

Dynamic Programming

Bellman Equations

Value Iteration

Policy Iteration

Value Iteration: the Idea

1. Let V_0 be *any* vector in R^N

Value Iteration: the Idea

1. Let V_0 be *any* vector in R^N
2. At each iteration $k = 1, 2, \ldots, K$

Value Iteration: the Idea

1. Let V_0 be *any* vector in R^N
2. At each iteration $k = 1, 2, \dots, K$
   - Compute $V_{k+1} = \mathcal{T} V_k$

Value Iteration: the Idea

1. Let V_0 be *any* vector in R^N
2. At each iteration $k = 1, 2, \dots, K$
   - Compute $V_{k+1} = \mathcal{T} V_k$
3. Return the *greedy* policy

$$\pi_K(s) \in \arg\max_{a \in A} \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V_K(s') \right].$$

Value Iteration: the Guarantees

Theorem

Let $V_0 \in \mathbb{R}^N$ be an arbitrary function, then the sequence of functions $\{V_k\}_k$ generated by value iteration ***converges*** *to the optimal value function $V^\star$.*

Furthermore, let $\varepsilon > 0$ and $\max_{s,a} |r(s,a)| \leq r_{\max} < \infty$, then after ***at most***

$$K = \frac{\log(r_{\max}/\varepsilon)}{\log(1/\gamma)}$$

iterations $\|V_K - V^\star\|_\infty \leq \varepsilon$.

Proof: Value Iteration

- From the *fixed point* property of $\mathcal{T}$ and $V_k = \mathcal{T} V_{k-1}$

$$\lim_{k\to\infty} V_k = \lim_{k\to\infty} \mathcal{T}^k V_0 = V^\star$$

Proof: Value Iteration

- From the *fixed point* property of $\mathcal{T}$ and $V_k = \mathcal{T} V_{k-1}$

$$\lim_{k\to\infty} V_k = \lim_{k\to\infty} \mathcal{T}^k V_0 = V^\star$$

- From the *contraction* property of $\mathcal{T}$

$$\begin{aligned}
&\|V^\star - V_{k+1}\|_\infty \\
&\quad = \|\mathcal{T} V^\star - \mathcal{T} V_k\|_\infty && \textit{[value iteration and Bellman eq.]} \\
&\quad \leq \gamma \|V_k - V^\star\|_\infty && \textit{[contraction]} \\
&\quad \leq \gamma^{k+1} \|V^\star - V_0\|_\infty && \textit{[recursion.]} \\
&\quad \to 0
\end{aligned}$$

Proof: Value Iteration

- From the *fixed point* property of $\mathcal{T}$ and $V_k = \mathcal{T} V_{k-1}$

$$\lim_{k\to\infty} V_k = \lim_{k\to\infty} \mathcal{T}^k V_0 = V^\star$$

- From the *contraction* property of $\mathcal{T}$

$$\begin{aligned}
&\|V^\star - V_{k+1}\|_\infty \\
&\quad = \|\mathcal{T}V^\star - \mathcal{T}V_k\|_\infty && \textit{[value iteration and Bellman eq.]} \\
&\quad \le \gamma\|V_k - V^\star\|_\infty && \textit{[contraction]} \\
&\quad \le \gamma^{k+1}\|V^\star - V_0\|_\infty && \textit{[recursion.]} \\
&\quad \to 0
\end{aligned}$$

- *Convergence rate*. Let $\varepsilon > 0$ and $||r||_\infty \le r_{\max}$, then after *at most*

$$\|V^\star - V_{k+1}\|_\infty \le \gamma^{k+1}\|V^\star - V_0\|_\infty < \varepsilon \;\;\Rightarrow\;\; K \ge \frac{\log(r_{\max}/\varepsilon)}{\log(1/\gamma)}$$

Value Iteration: the Guarantees

Corollary

Let V_K the function computed after K iterations by value iteration, then the greedy policy

$$\pi_K(s) \in \arg\max_{a \in A} \left[r(x,a) + \gamma \sum_y p(y|s,a) V_K(y) \right]$$

is such that

$$\underbrace{\|V^\star - V^{\pi_K}\|_\infty}_{\text{performance loss}} \leq \frac{2\gamma}{1-\gamma} \underbrace{\|V^\star - V_K\|_\infty}_{\text{approx. error}} .$$

Furthermore, there exists $\epsilon > 0$ such that if $\|V_K - V^\star\|_\infty \leq \epsilon$, then π_K is optimal.

Proof: Performance Loss

$$
\begin{aligned}
\|V^\star - V^{\pi_k}\|_\infty &\le \|\mathcal{T}V^\star - \mathcal{T}^{\pi_k}V_k\|_\infty + \|\mathcal{T}^{\pi_k}V_k - \mathcal{T}^{\pi_k}V^{\pi_k}\|_\infty \\
&\le \|\mathcal{T}V^\star - \mathcal{T}V_k\|_\infty + \gamma\|V - V^{\pi_k}\|_\infty \\
&\le \gamma\|V^\star - V_k\|_\infty + \gamma(\|V_k - V^\star\|_\infty + \|V^\star - V^{\pi_k}\|_\infty) \\
&\le \frac{2\gamma}{1-\gamma}\|V^\star - V_k\|_\infty.
\end{aligned}
$$

■

Value Iteration: the Idea

Termination condition

$$span(V_k - V_{k-1}) := \max_s |V_k(s) - V_{k-1}(s)| - \min_s |V_k(s) - V_{k-1}(s)| \leq \varepsilon$$

Performance guarantees

$$\underbrace{\|V^\star - V^\pi\|_\infty}_{\textit{performance loss}} \leq \frac{\gamma}{1-\gamma}\varepsilon$$

Value Iteration

Input: $\mathcal{S}$, $\mathcal{A}$, r, p, ϵ

Set $V_0(s) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $k = 0$

repeat

  for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**

    Compute

$$V_{k+1}(s) = \mathcal{T} V_k(s) = \max_a \left\{ r(s,a) + \gamma \sum_{s'} p(s'|s,a) V_k(s') \right\}$$

  end

  $k = k + 1$

until $\|V_{k+1} - V_k\|_\infty < \epsilon$

return greedy policy $\pi_\epsilon(s) = \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s'} p(s'|s,a) V_k(s') \right\}$

Value Iteration: the Complexity

Time complexity

- Each iteration takes $O(S^2 A)$ operations

$$V_{k+1}(s) = \mathcal{T} V_k(s) = \max_{a \in A} \Big[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V_k(s') \Big]$$

- The computation of the greedy policy takes $O(S^2 A)$ operations

$$\pi_K(s) \in \arg\max_{a \in A} \Big[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V_K(s') \Big]$$

- Total time complexity $O(KS^2 A)$

Space complexity

- Storing the MDP: dynamics $O(S^2 A)$ and reward $O(SA)$.
- Storing the value function and the optimal policy $O(S)$.

Value Iteration: Extensions and Implementations

Asynchronous VI.

1. Let V_0 be any vector in R^N
2. At each iteration $k = 1, 2, \dots, K$
   - **Choose a state s_k**
   - Compute $V_{k+1}(s_k) = \mathcal{T} V_k(s_k)$
3. Return the greedy policy

$$\pi_K(s) \in \arg\max_{a \in A} \Big[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V_K(s') \Big].$$

Comparison

- Reduced time complexity to $O(SA)$
- Using round-robin, number of iterations increased by at most $O(KS)$ but much smaller in practice if states are properly *prioritized*
- Convergence guarantees if no *starvation*

Value Iteration: Extensions and Implementations

Q-iteration

Input: $\mathcal{S}$, $\mathcal{A}$, r, p, ϵ
Set $Q_0(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$

repeat
  for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
    Compute

$$Q_{k+1}(s,a) = \mathcal{T} Q_k(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q_k(s',a') \right]$$

  end
  $k = k+1$
until $\|Q_{k+1} - Q_k\|_{\infty,\infty} < \epsilon$
return greedy policy $\pi_\epsilon(s) = \arg\max_{a \in \mathcal{A}} Q_k(s,a)$

Comparison

- Increased space and time complexity to $O(SA)$ and $O(S^2A^2)$
- Reduced time complexity to compute the greedy policy $O(SA)$
- *Bonus:* computing the greedy policy from the Q-function **does not require** the MDP

Dynamic Programming

Bellman Equations

Value Iteration

Policy Iteration

Policy Iteration: the Idea

1. Let π_0 be *any* stationary policy

Policy Iteration: the Idea

1. Let π_0 be *any* stationary policy
2. At each iteration $k = 1, 2, \ldots, K$

Policy Iteration: the Idea

1. Let π_0 be *any* stationary policy
2. At each iteration $k = 1, 2, \ldots, K$
   - *Policy evaluation* given π_k, compute V^{π_k}.

Policy Iteration: the Idea

1. Let π_0 be *any* stationary policy
2. At each iteration $k = 1, 2, \dots, K$
   - *Policy evaluation* given π_k, compute V^{π_k}.
   - *Policy improvement*: compute the *greedy* policy

$$\pi_{k+1}(s) \in \arg\max_{a \in A} \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^{\pi_k}(s') \right].$$

Policy Iteration: the Idea

1. Let π_0 be *any* stationary policy
2. At each iteration $k = 1, 2, \ldots, K$
   - *Policy evaluation* given π_k, compute V^{π_k}.
   - *Policy improvement*: compute the *greedy* policy

$$\pi_{k+1}(s) \in \arg\max_{a \in A} \Big[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^{\pi_k}(s') \Big].$$

3. Stop if $V^{\pi_k} = V^{\pi_{k-1}}$

Policy Iteration: the Idea

1. Let π_0 be *any* stationary policy
2. At each iteration $k = 1, 2, \dots, K$
   - *Policy evaluation* given π_k, compute V^{π_k}.
   - *Policy improvement*: compute the *greedy* policy

$$\pi_{k+1}(s) \in \arg\max_{a \in A} \Big[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^{\pi_k}(s') \Big].$$

3. Stop if $V^{\pi_k} = V^{\pi_{k-1}}$
4. Return the last policy π_K

Policy Iteration: the Guarantees

Proposition

The policy iteration algorithm generates a sequences of policies with non-decreasing performance

$$V^{\pi_{k+1}} \geq V^{\pi_k},$$

and it converges to $\pi^\star$ in a finite number of iterations.

Proof: Policy Iteration

From the definition of the Bellman operators and the greedy policy π_{k+1}

$$V^{\pi_k} = \mathcal{T}^{\pi_k} V^{\pi_k} \leq \mathcal{T} V^{\pi_k} = \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \tag{1}$$

Proof: Policy Iteration

From the definition of the Bellman operators and the greedy policy π_{k+1}

$$V^{\pi_k} = \mathcal{T}^{\pi_k} V^{\pi_k} \leq \mathcal{T} V^{\pi_k} = \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \tag{1}$$

and from the monotonicity property of $\mathcal{T}^{\pi_{k+1}}$, it follows that

$$\begin{aligned} V^{\pi_k} &\leq \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \\ \mathcal{T}^{\pi_{k+1}} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^2 V^{\pi_k}, \\ &\dots \\ (\mathcal{T}^{\pi_{k+1}})^{n-1} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k}, \\ &\dots \end{aligned}$$

Proof: Policy Iteration

From the definition of the Bellman operators and the greedy policy π_{k+1}

$$V^{\pi_k} = \mathcal{T}^{\pi_k} V^{\pi_k} \leq \mathcal{T} V^{\pi_k} = \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \tag{1}$$

and from the monotonicity property of $\mathcal{T}^{\pi_{k+1}}$, it follows that

$$\begin{aligned} V^{\pi_k} &\leq \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \\ \mathcal{T}^{\pi_{k+1}} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^2 V^{\pi_k}, \\ &\dots \\ (\mathcal{T}^{\pi_{k+1}})^{n-1} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k}, \\ &\dots \end{aligned}$$

Joining all the inequalities in the chain we obtain

$$V^{\pi_k} \leq \lim_{n\to\infty} (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k} = V^{\pi_{k+1}}.$$

Proof: Policy Iteration

From the definition of the Bellman operators and the greedy policy π_{k+1}

$$V^{\pi_k} = \mathcal{T}^{\pi_k} V^{\pi_k} \leq \mathcal{T} V^{\pi_k} = \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \tag{1}$$

and from the monotonicity property of $\mathcal{T}^{\pi_{k+1}}$, it follows that

$$\begin{aligned} V^{\pi_k} &\leq \mathcal{T}^{\pi_{k+1}} V^{\pi_k}, \\ \mathcal{T}^{\pi_{k+1}} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^2 V^{\pi_k}, \\ &\dots \\ (\mathcal{T}^{\pi_{k+1}})^{n-1} V^{\pi_k} &\leq (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k}, \\ &\dots \end{aligned}$$

Joining all the inequalities in the chain we obtain

$$V^{\pi_k} \leq \lim_{n \to \infty} (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k} = V^{\pi_{k+1}}.$$

Then $(V^{\pi_k})_k$ is a non-decreasing sequence.

Policy Iteration: the Guarantees

Since a finite MDP admits a finite number of policies, then the termination condition is eventually met for a specific k.
Thus eq. 1 holds with an equality and we obtain

$$V^{\pi_k} = \mathcal{T} V^{\pi_k}$$

and $V^{\pi_k} = V^\star$ which implies that π_k is an optimal policy. ■

Policy Iteration: Complexity

Notation. For any policy π the reward *vector* is $r^\pi(x) = r(x, \pi(x))$ and the transition *matrix* is $[P^\pi]_{x,y} = p(y|x, \pi(x))$

Policy Iteration: Complexity

Policy Evaluation Step

- *Direct computation.* For any policy π compute

$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi.$$

Complexity: $O(S^3)$.

Policy Iteration: Complexity

Policy Evaluation Step

- *Direct computation.* For any policy π compute

$$V^{\pi} = (I - \gamma P^{\pi})^{-1} r^{\pi}.$$

Complexity: $O(S^3)$.

- *Iterative policy evaluation.* For any policy π

$$\lim_{n \to \infty} \mathcal{T}^{\pi} V_0 = V^{\pi}.$$

Complexity: An ε-approximation of V^{π} requires $O\Big(S^2 \frac{\log(1/\epsilon)}{\log(1/\gamma)}\Big)$ steps.

Policy Iteration: Complexity

Policy Evaluation Step

- *Direct computation.* For any policy π compute

$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi.$$

  Complexity: $O(S^3)$.

- *Iterative policy evaluation.* For any policy π

$$\lim_{n\to\infty} \mathcal{T}^\pi V_0 = V^\pi.$$

  Complexity: An ε-approximation of V^π requires $O\Big(S^2 \frac{\log(1/\epsilon)}{\log(1/\gamma)}\Big)$ steps.

- *Monte-Carlo simulation.* In each state s, simulate n trajectories $((s_t^i)_{t\geq 0,})_{1\leq i\leq n}$ following policy π and compute

$$\hat{V}^\pi(s) \simeq \frac{1}{n}\sum_{i=1}^{n}\sum_{t\geq 0}\gamma^t r(s_t^i, \pi(s_t^i)).$$

  Complexity: In each state, the approximation error is $O\Big(\frac{r_{\max}}{1-\gamma}\sqrt{\frac{1}{n}}\Big)$

Policy Iteration: Complexity

Policy Improvement Step

- If the policy is *evaluated with* V, then complexity $O(SA)$

Policy Iteration: Complexity

Policy Improvement Step

- If the policy is *evaluated with V*, then complexity $O(SA)$
- If the policy is *evaluated with Q*, then complexity $O(A)$

$$\pi_{k+1}(s) \in \arg\max_{a \in A} Q^{\pi_k}(s, a),$$

Number of Iterations

- At most $O\left(\frac{SA}{1-\gamma} \log(\frac{1}{1-\gamma})\right)$
- Other results exist that do not depend on γ

Comparison between Value and Policy Iteration

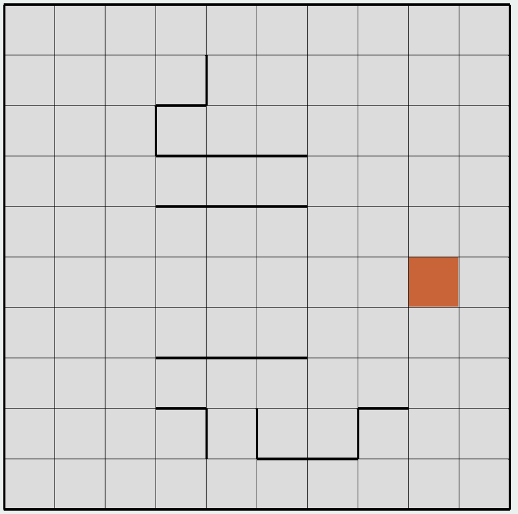
Value Iteration

- *Pros:* each iteration is very *computationally efficient*.
- *Cons:* convergence is only *asymptotic*.

Policy Iteration

- *Pros:* converge in a *finite* number of iterations (often small in practice).
- *Cons:* each iteration requires a full *policy evaluation* and it might be expensive.

The Grid-World Problem



Other Algorithms

- Linear programming
- Modified Policy Iteration
- λ-Policy Iteration
- Primal-dual formulations

Outline

Markov Decision Process

[Puterman, 1994]

A **finite-horizon** Markov decision process (MDP) is a tuple $M = \langle \mathcal{S}, \mathcal{A}, r_h, p_h, H \rangle$

- State space $\mathcal{S}$
- Action space $\mathcal{A}$
- Horizon H
- Transition distribution $p_h(\cdot|s,a) \in \Delta(\mathcal{S})$, $h = 1, \dots, H$
- Reward distribution with expectation $r_h(s,a) \in [0,1]$, $h = 1, \dots, H$

An agent acts according to a *time-variant policy*

$$\pi_h : \mathcal{S} \to \mathcal{A} \qquad h = 1, \dots, H$$

Value Functions and Optimality

Value functions

$$
\begin{aligned}
Q_h^{\pi}(s,a) &= r_h(s,a) + \mathbb{E}\left[\sum_{l=h+1}^{H} r_l(s_l, \pi_l(s_l))\right] \\
V_h^{\pi}(s) &= Q_h^{\pi}(s, \pi_h(s))
\end{aligned}
$$

Optimality

$$
\begin{aligned}
Q_h^{\star}(s,a) &= \sup_{\pi} Q_h^{\pi}(s,a) \\
\pi_h^{\star}(s) &= \arg\max_{a \in \mathcal{A}} Q_h^{\star}(s,a)
\end{aligned}
$$

Value Functions and Optimality

Value functions

$$Q^{\pi}_h(s,a) = r_h(s,a) + \mathbb{E}\Big[\sum_{l=h+1}^{H} r_l(s_l, \pi_l(s_l))\Big]$$
$$V^{\pi}_h(s) = Q^{\pi}_h(s, \pi_h(s))$$

Optimality

$$Q^{\star}_h(s,a) = \sup_{\pi} Q^{\pi}_h(s,a)$$
$$\pi^{\star}_h(s) = \arg\max_{a\in\mathcal{A}} Q^{\star}_h(s,a)$$

Remark: given $r_h(s,a) \in [0,1]$, then $Q_h(s,a), V_h(s) \in [0, H-(h-1)]$

Bellman Equations

Policy Bellman equation

$$
\begin{aligned}
Q_h^\pi(s,a) &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[Q_{h+1}^\pi(s', \pi_{h+1}(s'))\right] \\
&= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[V_{h+1}^\pi(s')\right]
\end{aligned}
$$

Optimal Bellman equation

$$
\begin{aligned}
Q_h^\star(s,a) &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[\max_{a' \in \mathcal{A}} Q_{h+1}^\star(s', a')\right] \\
&= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[V_{h+1}^\star(s')\right]
\end{aligned}
$$

Value Iteration (aka Backward Induction)

Input: $\mathcal{S}$, $\mathcal{A}$, r_h, p_h
Set $Q^\star_{H+1}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$

for $h = H, \dots, 1$ **do**
  for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
    Compute

$$\begin{aligned} Q^\star_h(s,a) &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[\max_{a' \in \mathcal{A}} Q^\star_{h+1}(s',a')\right] \\ &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[V^\star_{h+1}(s')\right] \end{aligned}$$

  end
end
return $\pi^\star_h(s) = \arg\max_{a \in \mathcal{A}} Q^\star_h(s,a)$

☞ the algorithm always converges in H steps to the unique optimal solution

Outline

Stochastic Shortest Path Problem

[Bertsekas, 2007]

Let $\overline{s}$ be a *terminal state*. Then

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\tau_\pi(s)} r(s_t, a_t) | s_0 = s \right]$$

with $\tau_\pi(s) = \inf\{\tau \in \mathbb{N} : s_{t+1} = \overline{s} | s_1 = s, \pi\}$ (*hitting time*)

With the convention:

- $p(\overline{s}|\overline{s}, a) = 1$, $r(\overline{s}, a) \geq 0$ for nay a
- $r(s, a) < 0$ for all $s \in \mathcal{S} \setminus \{\overline{s}\}$ and a
- $|r(s, a)| \leq r_{\max}$, for any (s, a)

since r is bounded we can restrict out attention to **stationary deterministic policies**

Properties

- It features two possibly conflicting objectives
  - quickly reaching the terminal state
  - while minimizing the costs along the way
- Policies may never reach the terminal state
- The number of summands may differ from one trajectory to another

Proper Policies

A stationary policy π is *proper* if $\exists n \in \mathbb{N}$ such that $\forall s \in \mathcal{S}$ the probability of reaching the terminal state $\overline{s}$ aftern n steps is strictly positive:

$$\rho_\pi = \max_s \mathbb{P}(s_n \neq \overline{s} | s_0 = s, \pi) < 1$$

👉 i.e., $\overline{s}$ is reached with probability 1 from any state $\mathcal{S}$

Proper Policies

A stationary policy π is *proper* if $\exists n \in \mathbb{N}$ such that $\forall s \in \mathcal{S}$ the probability of reaching the terminal state $\overline{s}$ aftern n steps is strictly positive:

$$\rho_\pi = \max_s \mathbb{P}(s_n \neq \overline{s} | s_0 = s, \pi) < 1$$

i.e., $\overline{s}$ is reached with probability 1 from any state $\mathcal{S}$

Properties:

- π proper policy $\Rightarrow V^\pi$ is bounded i.e., $\|V^\pi\|_\infty < \infty$
- π non-proper policy $\Rightarrow \exists s \in \mathcal{S} : V^\pi(s) = -\infty$

Bellman Operator

[Bertsekas, 2007]

Bellman Operator:

$$\mathcal{T}W(s) = \max_{a \in \mathcal{A}} \left(r(s,a) + \sum_{s'} p(s'|s,a) W(s') \right)$$

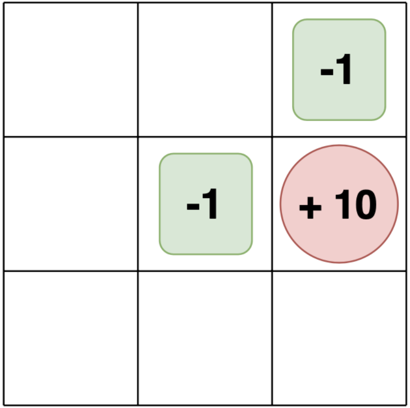
Under certain properties [?]

$\mathcal{T}$ is a contraction in a (∞, μ)-norm $\|\cdot\|_{\infty,\mu}$, i.e., exists μ and $\beta < 1$ such that

$$\|\mathcal{T}v - \mathcal{T}u\|_{\infty,\mu} \leq \beta \|\mathcal{T}v - \mathcal{T}u\|_{\infty,\mu}$$

$\implies$ value iteration converges

Exercise: Value Iteration



Try value iteration (on paper) on this very simple grid world with a single terminal state in red to see that it can converge.

Outline

Classification

If an MDP M is

- *ergodic* then it is possible to go from any state to any other state under *any* deterministic stationary policy

$$\forall s, s', \forall \pi : \mathcal{S} \to \mathcal{A}, \exists t < \infty, \text{ s.t. } \mathbb{P}_\pi^M\big(s_t = s' | s_0 = s\big) > 0$$

- *communicating* then it is possible to go from any state to any other state under *a specific* deterministic stationary policy

$$\forall s, s', \exists \pi : \mathcal{S} \to \mathcal{A}, \exists t < \infty, \text{ s.t. } \mathbb{P}_\pi^M\big(s_t = s' | s_0 = s\big) > 0$$

☞ A communicating MDP has *finite diameter*

$$D_M = \max_{s,s' \in \mathcal{S}} \min_{\pi : \mathcal{S} \to \mathcal{A}} \mathbb{E}\big[T_\pi^M(s, s')\big]$$

Classification

If an MDP M is

- *ergodic* then it is possible to go from any state to any other state under *any* deterministic stationary policy

$$\forall s, s', \forall \pi : \mathcal{S} \to \mathcal{A}, \exists t < \infty, \text{ s.t. } \mathbb{P}_\pi^M\big(s_t = s' | s_0 = s\big) > 0$$

- *communicating* then it is possible to go from any state to any other state under *a specific* deterministic stationary policy

$$\forall s, s', \exists \pi : \mathcal{S} \to \mathcal{A}, \exists t < \infty, \text{ s.t. } \mathbb{P}_\pi^M\big(s_t = s' | s_0 = s\big) > 0$$

☞ A communicating MDP has *finite diameter*

$$D_M = \max_{s,s' \in \mathcal{S}} \underbrace{\min_{\pi:\mathcal{S}\to\mathcal{A}} \mathbb{E}\big[T_\pi^M(s,s')\big]}_{\text{shortest path}}$$

Gain and Bias

Gain of a deterministic stationary policy π

$$g_M^\pi(s) = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} r(s_t, a_t) \middle| s_0 = s, a_t = \pi(s_t)\right]$$

Bias of a deterministic stationary policy π

$$h_M^\pi(s) := C\text{-}\lim_{T\to\infty} \mathbb{E}\left[\sum_{t=1}^{T} \big(r(s_t, a_t) - g_M^\pi(s_t)\big) \middle| s_0 = s, a_t = \pi(s_t)\right]$$

Span of the bias function

$$\mathrm{sp}\big(h_M^\pi\big) = \max_s h_M^\pi(s) - \min_s h_M^\pi(s)$$

Bellman operators

$$= \sum_{s'} p(s'|s,a)h(s')$$

Bellman operator $L_M^a : \mathbb{R}^S \to \mathbb{R}^S$

$$L_M^a h(s) = r(s,a) + \underbrace{p(\cdot|s,a)^\mathsf{T} h}$$

Optimal Bellman operator $L_M^\star : \mathbb{R}^S \to \mathbb{R}^S$

$$L_M^\star h(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + p(\cdot|s,a)^\mathsf{T} h \right\}$$

Optimality gap of action a at s

$$\delta_M^\star(s,a) \;=\; L_M^\star h_M^\star(s) - L_M^a h_M^\star(s)$$

a.k.a. advantage function

Optimality

Optimal policy and *optimal gain*

$$\pi^\star_M \in \arg\max_\pi g^\pi_M(s) \qquad g^\star_M = g^{\pi^\star}_M(s) \;\; \forall s \in \mathcal{S}$$

Optimality equation

$$h^\star_M(s) + g^\star_M = L^\star_M h^\star_M(s)$$

Greedy policy w.r.t. $h^\star_M$ is optimal

$$\pi^\star_M(s) \in \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + p(\cdot|s,a)^\mathsf{T} h^\star_M \right\}$$

Set of optimal actions in state s

$$\Pi^\star_M(s) = \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + p(\cdot|s,a)^\mathsf{T} h^\star_M \right\}$$

Optimality

deterministic stationary

Optimal policy and *optimal gain*

$$\pi^\star_M \in \arg\max_{\pi} g^\pi_M(s) \qquad g^\star_M = g^{\pi^\star}_M(s) \quad \forall s \in \mathcal{S}$$

Optimality equation

$$h^\star_M(s) + g^\star_M = L^\star_M h^\star_M(s)$$

Greedy policy w.r.t. $h^\star_M$ is optimal

$$\pi^\star_M(s) \in \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + p(\cdot|s,a)^\mathsf{T} h^\star_M \right\}$$

Set of optimal actions in state s

$$\Pi^\star_M(s) = \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + p(\cdot|s,a)^\mathsf{T} h^\star_M \right\}$$

Optimality

deterministic stationary

constant gain*

Optimal policy and *optimal gain*

$$\pi^\star_M \in \arg\max_\pi g^\pi_M(s) \qquad g^\star_M = g^{\pi^\star}_M(s) \quad \forall s \in \mathcal{S}$$

Optimality equation

$$h^\star_M(s) + g^\star_M = L^\star_M h^\star_M(s)$$

Greedy policy w.r.t. $h^\star_M$ is optimal

$$\pi^\star_M(s) \in \arg\max_{a\in\mathcal{A}} \left\{ r(s,a) + p(\cdot|s,a)^\mathsf{T} h^\star_M \right\}$$

Set of optimal actions in state s

$$\Pi^\star_M(s) = \arg\max_{a\in\mathcal{A}} \left\{ r(s,a) + p(\cdot|s,a)^\mathsf{T} h^\star_M \right\}$$

*In communicating MDPs

Summary

- Bellman equations and Bellman operators (and their properties)
- Value iteration (algorithm, guarantees, and complexity)
- Policy iteration (algorithm, guarantees, and complexity)

Bibliography

R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, N.J., 1957.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. II*. Athena Scientific, 3rd edition, 2007.

M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, Etats-Unis, 1994.

Thank you!

facebook
Artificial Intelligence Research