# Dynamic Programming

IV125, Adam Ivora

March 8, 2021

## 1 Dynamic programming

Dynamic programming is a collection of algorithms that can be used to compute optimal policies given a perfect model of the environment as a Markov Decision Process.

We will discuss four of the algorithms:

- Policy evaluation (prediction) - allows us to compute the state-value function $v_\pi(s)$ for an arbitrary policy $\pi$,
- Policy improvement - allows us to find a better policy $\pi'$ given a value function $v_\pi$
- Policy iteration - allows us to find an optimal policy using policy evaluation and policy improvement $\pi_*$
- Value iteration - allows us to find an optimal policy without policy evaluation$\pi_*$

## 2 Environment

- $\mathcal{S}$ - a finite set of states
- $\mathcal{S}^+ = \mathcal{S} \cup \{\bot\}$ - a set of states plus the terminal state if it is needed
- $\mathcal{R}$ - a finite set of possible rewards
- $\mathcal{A} : \mathcal{S} \to \mathcal{P}(A)$ - a mapping from states to a finite subset of possible actions
- $p(s', r \mid s, a)$ - are dynamics of the environment such that $s \in \mathcal{S}, a \in \mathcal{A}(s), r \in \mathcal{R}, s' \in \mathcal{S}^+$

### 2.1 Stochastic policy

The following sections assume a stochastic policy $\pi$ such that:

- $\pi : (A \times \mathcal{S}) \to [0, 1]$,
- $\pi(a \mid s)$ is a way to denote $\pi(a, s)$,
- domain of $\pi$ is $D(\pi) = \{(a, s) \mid a \in \mathcal{A}(s), s \in \mathcal{S}\}$,
- $\forall (a, s) \in D(\pi) : \pi(a \mid s) \geq 0$,
- $\forall s \in \mathcal{S} : \sum_{a \in \mathcal{A}(s)} \pi(a \mid s) = 1$.

## 3 Bellman expectation equation

For a fixed stochastic policy $\pi$, discount factor $\gamma \in [0, 1)$ (with the possibility that $\gamma = 1$ if all episodes terminate) it holds that $\forall s \in \mathcal{S}$:

$$v_\pi(s) = \sum_{a \in \mathcal{A}(s)} \pi(a \mid s) \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a)[r + \gamma \cdot v_\pi(s')],$$

and $v_\pi(\bot) = 0$, if $\bot$ is defined.

## 3.1 Analytic solution

If we fully know the dynamics $p(s', r \mid s, a)$, then the Bellman expectation equation is a system of $|\mathcal{S}|$ linear equations with $\mathcal{S}$ unknowns, which are $v_\pi(s)$ for each $s \in \mathcal{S}$.

$$\begin{aligned} v_\pi(s) &= \sum_{a \in \mathcal{A}(s)} \pi(a \mid s) \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a)[r + \gamma \cdot v_\pi(s')] \\ &= \sum_{a \in \mathcal{A}(s)} \pi(a \mid s) \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a) \cdot r + \gamma \cdot \sum_{a \in \mathcal{A}(s)} \pi(a \mid s) \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a) \cdot v_\pi(s') \end{aligned}$$

We will introduce the Bellman operator to show how the solution can be computed exactly.

## 3.2 Bellman operator

Denote by $V$ a vector space of all state-value functions $\boldsymbol{v} = (v(s_1), v(s_2), \ldots, v(s_{|\mathcal{S}|}))$ ($\dim V = |\mathcal{S}|$).
Then Bellman operator $B_\pi : V \to V$ is defined by:

$$B_\pi(\boldsymbol{v}) = R_\pi + \gamma \cdot P_\pi \cdot \boldsymbol{v},$$

where $R_\pi$ is a vector of expected rewards for each state $(R_\pi(s_1), R_\pi(s_2), \ldots, R_\pi(s_{|\mathcal{S}|}))$ and

$$R_\pi(s) = \sum_{a \in \mathcal{A}(s)} \pi(a \mid s) \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a) \cdot r,$$

and where $P_\pi$ is an $|\mathcal{S}| \times |\mathcal{S}|$ (stochastic) matrix which elements are

$$P_\pi(s, s') = \sum_{a \in \mathcal{A}(s)} \pi(a \mid s) \sum_{r \in R} p(s', r \mid s, a)$$

Let's show the mapping has a fixed point $B_\pi(\boldsymbol{v_\pi}) = \boldsymbol{v_\pi}$ for a simple concrete case where $\mathcal{S} = \{s_1, s_2, s_3\}$:

$$
\begin{aligned}
B_\pi(\boldsymbol{v}) &= R_\pi + \gamma \cdot P_\pi \cdot \boldsymbol{v} \\
B_\pi \begin{pmatrix} v(s_1) \\ v(s_2) \\ v(s_3) \end{pmatrix} &= \begin{pmatrix} R_\pi(s_1) \\ R_\pi(s_2) \\ R_\pi(s_3) \end{pmatrix} + \gamma \cdot \begin{pmatrix} P_\pi(s_1, s_1) & P_\pi(s_1, s_2) & P_\pi(s_1, s_3) \\ P_\pi(s_2, s_1) & P_\pi(s_2, s_2) & P_\pi(s_2, s_3) \\ P_\pi(s_3, s_1) & P_\pi(s_3, s_2) & P_\pi(s_3, s_3) \end{pmatrix} \cdot \begin{pmatrix} v(s_1) \\ v(s_2) \\ v(s_3) \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
B_\pi(v(s_1)) &= R_\pi(s_1) + \gamma \cdot (P_\pi(s_1, s_1) \cdot v(s_1) + P_\pi(s_1, s_2) \cdot v(s_2) + P_\pi(s_1, s_3) \cdot v(s_3)) \\
B_\pi(v(s_1)) &= \sum_{a \in \mathcal{A}(s_1)} \pi(a \mid s_1) \sum_{\substack{s' \in \mathcal{S} \\ r \in \mathcal{R}}} p(s', r \mid s_1, a) \cdot r \\
&\quad + \gamma \sum_{a \in \mathcal{A}(s_1)} \pi(a \mid s_1) \sum_{r \in R} p(s_1, r \mid s_1, a) \cdot v(s_1) \\
&\quad + \gamma \sum_{a \in \mathcal{A}(s_1)} \pi(a \mid s_1) \sum_{r \in R} p(s_2, r \mid s_1, a) \cdot v(s_2) \\
&\quad + \gamma \sum_{a \in \mathcal{A}(s_1)} \pi(a \mid s_1) \sum_{r \in R} p(s_3, r \mid s_1, a) \cdot v(s_3) \\
&= \sum_{a \in \mathcal{A}(s_1)} \pi(a \mid s_1) \left[ \sum_{\substack{s' \in \mathcal{S} \\ r \in \mathcal{R}}} (p(s', r \mid s_1, a) \cdot r) + \sum_{s' \in \mathcal{S}} \sum_{r \in R} \gamma \cdot p(s', r \mid s_1, a) \cdot v(s') \right] \\
&= \sum_{a \in \mathcal{A}(s_1)} \pi(a \mid s_1) \left[ \sum_{\substack{s' \in \mathcal{S} \\ r \in \mathcal{R}}} p(s', r \mid s_1, a) \cdot r + \gamma \cdot p(s', r \mid s_1, a) \cdot v(s') \right] \\
&= \sum_{a \in \mathcal{A}(s_1)} \pi(a \mid s_1) \sum_{\substack{s' \in \mathcal{S} \\ r \in \mathcal{R}}} p(s', r \mid s_1, a) \cdot [r + \gamma \cdot v(s')] \\
&\implies B_\pi(v_\pi(s)) = \sum_{a \in \mathcal{A}(s)} \pi(a \mid s) \sum_{\substack{s' \in \mathcal{S} \\ r \in \mathcal{R}}} p(s', r \mid s, a) \cdot [r + \gamma \cdot v_\pi(s')] = v_\pi(s)
\end{aligned}
$$

We can calculate $v_\pi$ analytically:

$$
\begin{aligned}
R_\pi + \gamma P_\pi \boldsymbol{v_\pi} &= \boldsymbol{v_\pi} \\
\boldsymbol{v_\pi} - \gamma P_\pi \boldsymbol{v_\pi} &= R_\pi \\
\boldsymbol{v_\pi}(1 - \gamma P_\pi) &= R_\pi \\
\boldsymbol{v_\pi} &= R_\pi (1 - \gamma P_\pi)^{-1}
\end{aligned}
$$

# 4 Policy evaluation

## 4.1 Synchronous iterative policy evaluation

For a fixed policy $\pi$, $\forall s \in \mathcal{S}$:

$$
\begin{aligned}
v_0(s) &= \begin{cases} \text{arbitrary} & s \in \mathcal{S}, \\ 0 & s = \bot. \end{cases} \\
v_{k+1}(s) &= \sum_{a \in \mathcal{A}(s)} \pi(a \mid s) \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a)[r + \gamma \cdot v_k(s')]
\end{aligned}
$$

and $v_k(\bot) = 0$ for all $k$, if present. Then

$$
\lim_{k \to \infty} v_k = v_\pi.
$$

*Proof.* The case of $\gamma \in (0, 1)$ is proven here, however the result can be extended to the case when $\gamma = 1$ if all episodes terminate.

It should be easy to see that synchronous iterative policy evaluation is equivalent to iterative application of the Bellman operator $B_\pi : V \to V$. For all $k > 0$,

$$v_k(s) = B_\pi^k(v_0(s)) = B_\pi(B_\pi(B_\pi(\dots B_\pi(v_0(s)))))$$

We consider the metric space $(V, d)$, where $V$ is the vector space over the value function vectors and $d$ is a metric induced by an $L_\infty$ norm:

$$\forall \boldsymbol{v} \in V : ||\boldsymbol{v}||_\infty = \max_{s \in \mathcal{S}} |v(s)|$$
$$\forall \boldsymbol{v_1}, \boldsymbol{v_2} \in V : d(\boldsymbol{v_1}, \boldsymbol{v_2}) = ||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty = \max_{s \in S} |v_1(s) - v_2(s)|$$

The operator $B_\pi$ is a $\gamma$-contraction which means that:

$$\forall \boldsymbol{v_1}, \boldsymbol{v_2} \in V : d(B_\pi(\boldsymbol{v_1}), B_\pi(\boldsymbol{v_2})) \leq \gamma \cdot d(\boldsymbol{v_1}, \boldsymbol{v_2})$$

We show that it is true:

$$\begin{aligned}
||B_\pi(\boldsymbol{v_1}) - B_\pi(\boldsymbol{v_2})||_\infty &= ||(R_\pi + \gamma \cdot P_\pi \cdot \boldsymbol{v_1}) - (R_\pi + \gamma \cdot P_\pi \cdot \boldsymbol{v_2})||_\infty \\
&= ||\gamma \cdot P_\pi(\boldsymbol{v_1} - \boldsymbol{v_2})||_\infty \\
&= ||\gamma \cdot P_\pi \begin{pmatrix} v_1(s_1) - v_2(s_1) \\ v_1(s_2) - v_2(s_1) \\ \vdots \\ v_1(s_{|\mathcal{S}|}) - v_2(s_{|\mathcal{S}|})) \end{pmatrix} ||_\infty \\
&\leq ||\gamma \cdot P_\pi \begin{pmatrix} ||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty \\ ||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty \\ \vdots \\ ||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty \end{pmatrix} ||_\infty \\
&= ||\gamma \cdot P_\pi \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} (||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty)||_\infty \\
&= ||\gamma \cdot \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} (||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty)||_\infty \\
&= \gamma \cdot || \begin{pmatrix} ||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty \\ ||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty \\ \vdots \\ ||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty \end{pmatrix} ||_\infty \\
&= \gamma \cdot ||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty
\end{aligned}$$

$$||B_\pi(\boldsymbol{v_1}) - B_\pi(\boldsymbol{v_2})||_\infty \leq \gamma \cdot ||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty,$$

as $P_\pi \cdot (1,1,\ldots,1)^\top = (1,1,\ldots,1)^\top$:

$$P_\pi \cdot (1,1,\ldots,1)^\top = (1,1,\ldots,1)^\top$$

$$\begin{pmatrix} \sum_{s' \in \mathcal{S}} P_\pi(s_1, s') \\ \sum_{s' \in \mathcal{S}} P_\pi(s_2, s') \\ \vdots \\ \sum_{s' \in \mathcal{S}} P_\pi(s_{|\mathcal{S}|}, s') \end{pmatrix} = \begin{pmatrix} \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}(s_1)} \pi(a \mid s_1) \sum_{r \in R} p(s', r \mid s_1, a) \\ \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}(s_2)} \pi(a \mid s_2) \sum_{r \in R} p(s', r \mid s_2, a) \\ \vdots \\ \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}(s_{|\mathcal{S}|})} \pi(a \mid s_{|\mathcal{S}|}) \sum_{r \in R} p(s', r \mid s_{|\mathcal{S}|}, a) \end{pmatrix}$$

$$\forall i \in \{1,\ldots,|\mathcal{S}|\} : \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}(s_i)} \pi(a \mid s_i) \sum_{r \in R} p(s', r \mid s_i, a) = \sum_{\substack{s' \in \mathcal{S} \\ a \in \mathcal{A}(s_i) \\ r \in \mathcal{R}}} \pi(a \mid s_i) \cdot p(s', r \mid s_i, a)$$

$$= p(s' \in \mathcal{S}, r \in \mathcal{R} \mid s_i, a \in \mathcal{A}(s_i)) = 1$$

Now that we know $||B_\pi(\boldsymbol{v_1}) - B_\pi(\boldsymbol{v_2})||_\infty \leq \gamma \cdot ||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty$ ($B_\pi$ is a $\gamma$-contraction), we can use the fact to find the fixed point and show that it is unique (by using the Banach Contraction Principle).

Assumptions of BCP: a contraction mapping $B_\pi$ on a complete metric space $(V, d)$ with contraction constant $\gamma < 1$.
Fix $\boldsymbol{v_0} \in V$ to an arbitrary value.

Define a sequence $\{v_k\}$ in $V$ by

$$\boldsymbol{v_{k+1}} = B_\pi(\boldsymbol{v_k}) = B_\pi^{k+1}(\boldsymbol{v_0}), k \geq 0.$$

Because $B_\pi$ is a $\gamma$-contraction, we have:

$$d(\boldsymbol{v_k}, \boldsymbol{v_{k+1}}) = d(B_\pi(\boldsymbol{v_{k-1}}), B_\pi(\boldsymbol{v_k})) \leq \gamma \cdot d(\boldsymbol{v_{k-1}}, \boldsymbol{v_k})$$
$$d(\boldsymbol{v_k}, \boldsymbol{v_{k+1}}) \leq \gamma^k \cdot d(\boldsymbol{v_0}, \boldsymbol{v_1})$$

For any $m, n$ such that $m > n$ it means

$$d(\boldsymbol{v_n}, \boldsymbol{v_m}) \leq \sum_{i=n}^{m-1} d(\boldsymbol{v_i}, \boldsymbol{v_{i+1}})$$
$$\leq \sum_{i=n}^{m-1} \gamma^i \cdot d(\boldsymbol{v_0}, \boldsymbol{v_1}) \leq \frac{\gamma^n}{1-\gamma} \cdot d(\boldsymbol{v_0}, \boldsymbol{v_1}).$$

**Definition 4.1.** A sequence $\{a_n\}$ is said to be a Cauchy sequence iff for any $\epsilon > 0$ there exists $N$ such that $d(a_n, a_m) < \epsilon$ for all $m, n \geq N$.

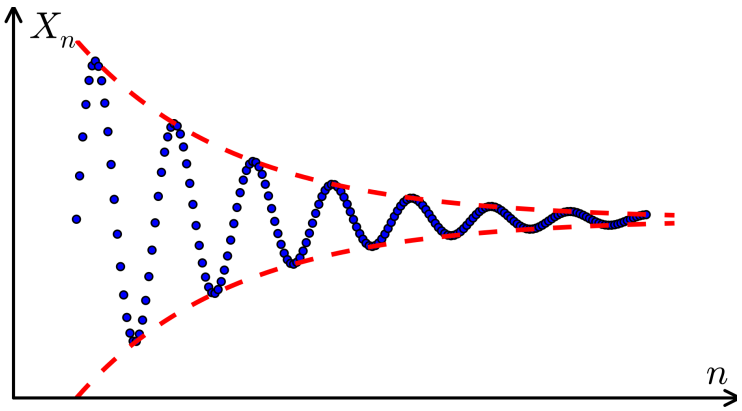


Figure 1: A Cauchy sequence

**Theorem 4.2** (Cauchy Criterion)**.** *In a complete metric space, a sequence is Cauchy iff it converges.*

We can find $N$ for any $\epsilon > 0$ such that $d(a_n, a_m) < \epsilon$ for all $m, n \geq N$:

$$
\begin{aligned}
d(\boldsymbol{v_n}, \boldsymbol{v_m}) &\leq \frac{\gamma^n}{1-\gamma} \cdot d(\boldsymbol{v_0}, \boldsymbol{v_1}) < \epsilon \\
\gamma^n &< \epsilon \cdot \frac{1-\gamma}{d(\boldsymbol{v_0}, \boldsymbol{v_1})} \\
n &> \log_\gamma \left( \epsilon \cdot \frac{1-\gamma}{d(\boldsymbol{v_0}, \boldsymbol{v_1})} \right) \\
N &= \left\lceil \log_\gamma \left( \epsilon \cdot \frac{1-\gamma}{d(\boldsymbol{v_0}, \boldsymbol{v_1})} \right) \right\rceil \\
&\implies d(\boldsymbol{v_n}, \boldsymbol{v_m}) \leq \frac{\gamma^N}{1-\gamma} \cdot d(\boldsymbol{v_0}, \boldsymbol{v_1}) < \epsilon.
\end{aligned}
$$

Thus, $\{v_k\}$ is a Cauchy sequence. Because $\{v_k\}$ is a Cauchy sequence, it satisfies the Cauchy criterion and converges.

That means there exists a convergence point $\boldsymbol{x}$:

$$
\begin{aligned}
\boldsymbol{x} &= \lim_{k \to \infty} \boldsymbol{v_k} = \lim_{k \to \infty} \boldsymbol{v_{k-1}} = B_\pi(\boldsymbol{x}) \\
B_\pi \left( \lim_{k \to \infty} \boldsymbol{v_k} \right) &= \lim_{k \to \infty} \boldsymbol{v_k}
\end{aligned}
$$

the limit of the iterative application of $B_\pi$ on $\boldsymbol{v_0}$ always converges to a fixed point $\boldsymbol{x}$ such that $B_\pi(\boldsymbol{x}) = \boldsymbol{x}$. But we already know one fixed point of the mapping, it is the solution $\boldsymbol{v}_\pi = v_\pi(s)\ \forall s \in \mathcal{S}$ to the Bellman expectation equation, which is the state-value function for an arbitrary policy $\pi$.

In other words,

$$
\forall \boldsymbol{v_0} \in V : \lim_{k \to \infty} B_\pi^k(\boldsymbol{v_0}) = \boldsymbol{v}_\pi,
$$

where $\boldsymbol{v}_\pi \in V$ is the state-value function (vector of dimension $|\mathcal{S}|$) of an arbitrary policy $\pi$.

The last thing to show is that the fixed point is unique. Let $\boldsymbol{x}, \boldsymbol{y}$ be fixed points of $B_\pi$, then

$$
\begin{aligned}
d(\boldsymbol{x}, \boldsymbol{y}) = d(B_\pi(\boldsymbol{x}), B_\pi(\boldsymbol{y})) &\leq \gamma \cdot d(\boldsymbol{x}, \boldsymbol{y}) \\
d(\boldsymbol{x}, \boldsymbol{y}) &\leq \gamma \cdot d(\boldsymbol{x}, \boldsymbol{y}) \\
(1-\gamma) \cdot d(\boldsymbol{x}, \boldsymbol{y}) &\leq 0
\end{aligned}
$$

$(1-\gamma) > 0$, thus $d(\boldsymbol{x}, \boldsymbol{y}) = 0$ (as the distance must be non-negative) and $\boldsymbol{x} = \boldsymbol{y}$. □

## 4.2 Asynchronous iterative policy evaluation

It is not needed to keep the values $v_k(s)$ during the iterative policy evaluation sweep. We can change the values of $v(s)$ in-place.

**Iterative Policy Evaluation, for estimating $V \approx v_\pi$**

Input $\pi$, the policy to be evaluated
Algorithm parameter: threshold $\epsilon > 0$ determining accuracy of estimation
Initialize $V(s)$ for all $s \in \mathcal{S}$ arbitrarily, $V(\perp) = 0$
**repeat**
  $\Delta = 0$;
  **foreach** $s \in \mathcal{S}$ **do**
    $v = V(s)$;
    $V(s) = \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)[r + \gamma \cdot V(s')]$;
    **if** $|v - V(s)| > \Delta$ **then**
      $\Delta = |v - V(s)|$;
    **end if**
  **end foreach**
**until** $\Delta < \epsilon$;

It can be proven that also this form of policy evaluation converges to $v_\pi$.

## 4.3 Gridworld

We consider a simple case of gridworld:

| | | | |
|---|---|---|---|
| $\perp$ | 1 | 2 | 3 |
| 4 | 5 | 6 | 7 |
| 8 | 9 | 10 | 11 |
| 12 | 13 | 14 | $\perp$ |

which can be represented as a finite MDP:

- $\mathcal{S} = \{1, 2, \ldots, 14\}$
- $\mathcal{S}^+ = \mathcal{S} \cup \{\perp\}$
- $\mathcal{R} = \{-1\}$
- $\forall s \in \mathcal{S} : \mathcal{A}(s) = \{\texttt{left}, \texttt{up}, \texttt{right}, \texttt{down}\}$
- $p(s', r \mid s, a)$
  state transition: when on the edge, stay on the current cell; otherwise, move to the direction specified by the action
  reward: -1 always (except the terminal state)
- $\gamma = 1$

## 4.4 Calculation of $v_\pi$

We will calculate the value function $v_\pi$ of the equiprobable random policy:

$$\forall s, a \in \mathcal{A}(s) : \pi(a \mid s) = \frac{1}{4}$$

a) analytically

$$\begin{aligned}
v_\pi(\perp) &= 0 \\
v_\pi(1) &= \frac{1}{4} \cdot p(\perp, -1 \mid 1, \texttt{left}) \cdot (-1 + v_\pi(\perp)) \\
&\quad + \frac{1}{4} \cdot p(1, -1 \mid 1, \texttt{up}) \cdot (-1 + v_\pi(1)) \\
&\quad + \frac{1}{4} \cdot p(2, -1 \mid 1, \texttt{right}) \cdot (-1 + v_\pi(2)) \\
&\quad + \frac{1}{4} \cdot p(5, -1 \mid 1, \texttt{down}) \cdot (-1 + v_\pi(5)) \\
v_\pi(1) &= -1 + \frac{1}{4} \cdot (v_\pi(1) + v_\pi(2) + v_\pi(5)) \\
&\implies 3v_\pi(1) - v_\pi(2) - v_\pi(5) = -4
\end{aligned}$$

$$\begin{aligned}
v_\pi(2) &= -1 + \frac{1}{4} \cdot (v_\pi(1) + v_\pi(2) + v_\pi(3) + v_\pi(6)) \\
&\implies -v_\pi(1) + 3v_\pi(2) - v_\pi(3) - v_\pi(6) = -4 \\
&\vdots
\end{aligned}$$

The system of equations can be represented by an $|\mathcal{S}| \times |\mathcal{S}|$ matrix:

$$\begin{pmatrix} 3 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & -1 & 3 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 3 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 3 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 3 \end{pmatrix} \cdot \boldsymbol{v_\pi} = \begin{pmatrix} -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \\ -4 \end{pmatrix}$$

We can solve the system using any method for solving linear equations and get

$$\boldsymbol{v_\pi} = \begin{pmatrix} -14 & -20 & -22 & -14 & -18 & -20 & -20 & -20 & -20 & -18 & -14 & -22 & -20 & -14 \end{pmatrix}^\top,$$

which are the exact values of the state-value function $v_\pi$ for the states $s \in \mathcal{S}$.

b) (synchronous) policy evaluation

| | | | |
|---|---|---|---|
| $\perp$ | 1 | 2 | 3 |
| 4 | 5 | 6 | 7 |
| 8 | 9 | 10 | 11 |
| 12 | 13 | 14 | $\perp$ |

We use the iterative policy evaluation update rule with $v_0(s) = 0$ for all $s \in \mathcal{S}^+$.

$k = 1$:

$$\begin{aligned} v_1(1) &= \frac{1}{4} \cdot p(\perp, -1 \mid 1, \texttt{left}) \cdot (-1 + v_0(\perp)) \\ &\quad + \frac{1}{4} \cdot p(1, -1 \mid 1, \texttt{up}) \cdot (-1 + v_0(1)) \\ &\quad + \frac{1}{4} \cdot p(2, -1 \mid 1, \texttt{right}) \cdot (-1 + v_0(2)) \\ &\quad + \frac{1}{4} \cdot p(5, -1 \mid 1, \texttt{down}) \cdot (-1 + v_0(5)) \\ &= \frac{1}{4} \cdot (-4 + 4 \cdot 0) = -1 \\ \forall s \in \mathcal{S} : v_1(s) &= -1 \end{aligned}$$

$k = 2$ :

$$\begin{aligned} v_2(1) &= \frac{1}{4} \cdot p(\perp, -1 \mid 1, \texttt{left}) \cdot (-1 + v_1(\perp)) \\ &\quad + \frac{1}{4} \cdot p(1, -1 \mid 1, \texttt{up}) \cdot (-1 + v_1(1)) \\ &\quad + \frac{1}{4} \cdot p(2, -1 \mid 1, \texttt{right}) \cdot (-1 + v_1(2)) \\ &\quad + \frac{1}{4} \cdot p(5, -1 \mid 1, \texttt{down}) \cdot (-1 + v_1(5)) \\ &= \frac{1}{4} \cdot (-4 + 0 - 1 - 1 - 1) = -\frac{7}{4} = -1.75 \\ &\vdots \end{aligned}$$

$k = 3$ :

$$
\begin{aligned}
v_3(1) &= \frac{1}{4} \cdot p(\perp, -1 \mid 1, \texttt{left}) \cdot (-1 + v_2(\perp)) \\
&+ \frac{1}{4} \cdot p(1, -1 \mid 1, \texttt{up}) \cdot (-1 + v_2(1)) \\
&+ \frac{1}{4} \cdot p(2, -1 \mid 1, \texttt{right}) \cdot (-1 + v_2(2)) \\
&+ \frac{1}{4} \cdot p(5, -1 \mid 1, \texttt{down}) \cdot (-1 + v_2(5)) \\
&= \frac{1}{4} \cdot (-4 + 0 - \frac{7}{4} - 2 - 2) = -\frac{39}{16} = -2.4375 \\
&\vdots
\end{aligned}
$$

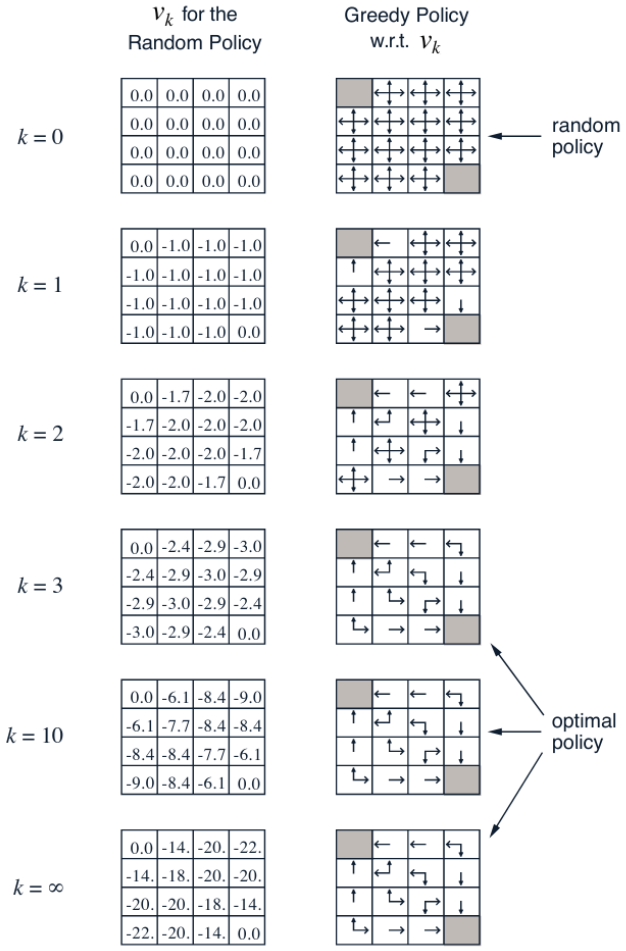


Figure 2: Convergence of iterative policy evaluation on a small gridworld.

# 5 Policy improvement

The following explanation is only for the case of deterministic policies $\pi$ such that:

- $\mathcal{A} : \mathcal{S} \to 2^A$
- $\pi : \mathcal{S} \to A$
- $\forall s \in \mathcal{S} : \pi(s) \in \mathcal{A}(s)$

The ideas should "easily" extend to stochastic policies.

## 5.1 Bellman equation for deterministic policies

As we know, the Bellman expectation equation for stochastic policies is

$$v_\pi^{\text{stoch}}(s) = \sum_{a \in \mathcal{A}(s)} \pi(a \mid s) \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a)[r + \gamma \cdot v_\pi(s')]$$

The Bellman expectation equation for deterministic policies is

$$v_\pi(s) = \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, \pi(s))[r + \gamma \cdot v_\pi(s')],$$

and

$$q_\pi(s, a) = \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a)[r + \gamma \cdot v_\pi(s')].$$

## 5.2 Bellman operator for deterministic policies

The Bellman operator for deterministic policies can be defined using the same matrices as for the stochastic case.

Denote by $V$ a vector space of all state-value functions $\boldsymbol{v} = (v(s_1), v(s_2), \ldots, v(s_{|\mathcal{S}|}))$ ($\dim V = |\mathcal{S}|$).
Then Bellman operator $B_\pi : V \to V$ is defined by:

$$B_\pi(\boldsymbol{v}) = R_\pi + \gamma \cdot P_\pi \cdot \boldsymbol{v},$$

where $R_\pi$ is a vector of expected rewards for each state $(R_\pi(s_1), R_\pi(s_2), \ldots, R_\pi(s_{|\mathcal{S}|}))$ and

$$R_\pi(s) = \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, \pi(s)) \cdot r,$$

and where $P_\pi$ is an $|\mathcal{S}| \times |\mathcal{S}|$ (stochastic) matrix which elements are

$$P_\pi(s, s') = \sum_{r \in R} p(s', r \mid s, \pi(s)).$$

## 5.3 Policy improvement theorem

Let $\pi, \pi' \in A^{\mathcal{S}}$ be deterministic policies such that

$$\forall s \in \mathcal{S} : q_\pi(s, \pi'(s)) \geq v_\pi(s).$$

Then the policy $\pi'$ is at least as good as $\pi$:

$$\forall s \in \mathcal{S} : v_{\pi'}(s) \geq v_\pi(s).$$

If there is strict inequality for at least one $s \in \mathcal{S}$ in the first equation, there is a strict inequality for at least one $s \in \mathcal{S}$ in the second equation.

*Proof.* (Original, from the book, without explanation.)

$$\begin{aligned}
v_\pi(s) &\leq q_\pi(s, \pi'(s)) \\
&= \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = \pi'(s)] \\
&= \mathbb{E}_{\pi'}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s] \\
&\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s] \\
&= \mathbb{E}_{\pi'}[R_{t+1} + \gamma\, \mathbb{E}'_\pi[R_{t+2} + \gamma v_\pi(S_{t+2}) \mid S_{t+1}, A_{t+1} = \pi'(S_{t+1})] \mid S_t = s] \\
&= \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(S_{t+2})) \mid S_t = s] \\
&\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 v_\pi(S_{t+3})) \mid S_t = s] \\
&\vdots \\
&\leq [R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \mid S_t = s] \\
&= v'_\pi(s).
\end{aligned}$$

□

**Definition 5.1** (Definitions and facts.)**.** We know that for arbitrary state-value function $\boldsymbol{v} \in V$ from the vector space $V$ of all state-value functions it holds that

$$\lim_{k\to\infty} B_\pi^k(\boldsymbol{v}) = \boldsymbol{v}_\pi,$$

where $\boldsymbol{v}_\pi \in V$ is the state-value function of the policy $\pi$.

We define a binary partial order relation $\geq$ between two state-value functions $\boldsymbol{v}, \boldsymbol{v'}$ from $V$ by

$$\boldsymbol{v'} \geq \boldsymbol{v} \iff (\forall s \in \mathcal{S})(v'(s) \geq v(s))$$

and a binary relation $>$ by

$$\boldsymbol{v'} > \boldsymbol{v} \iff (\forall s \in \mathcal{S})(v'(s) \geq v(s)) \wedge (\exists s' \in \mathcal{S})(v'(s') > v(s'))$$

The Bellman operator $B_\pi$ preserves $\geq$:

$$(\forall s \in \mathcal{S})(v(s) \geq v'(s)) \implies \left(\sum_{\substack{s'\in\mathcal{S}^+\\ r\in\mathcal{R}}} p(s', r \mid s, \pi(s))[r + \gamma \cdot v(s')]\right) \geq \left(\sum_{\substack{s'\in\mathcal{S}^+\\ r\in\mathcal{R}}} p(s', r \mid s, \pi(s))[r + \gamma \cdot v'(s')]\right)$$
$$\boldsymbol{v} \geq \boldsymbol{v'} \implies B_\pi(\boldsymbol{v}) \geq B_\pi(\boldsymbol{v}).$$

For all policies $\pi, \pi' \in A^{\mathcal{S}}$ and states $s \in \mathcal{S}$:

$$(B_{\pi'}(\boldsymbol{v}_\pi))(s) = \sum_{\substack{s'\in\mathcal{S}^+\\ r\in\mathcal{R}}} p(s', r \mid s, \pi'(s))[r + \gamma \cdot v_\pi(s')]$$
$$(B_{\pi'}(\boldsymbol{v}_\pi))(s) = q_\pi(s, \pi'(s))$$

*Proof.* (Policy improvement theorem, using Bellman operator.) We use the fixed point of the Bellman operator to prove the policy improvement theorem. There is an assumption that the discount factor $\gamma < 1$.

a) Let $\pi, \pi' \in A^{\mathcal{S}}$ be any pair of deterministic policies such that, for all $s \in \mathcal{S}$, $q_\pi(s, \pi'(s)) \geq v_\pi(s)$:

$$\begin{aligned}
(\forall s \in \mathcal{S})(q_\pi(s, \pi'(s)) &\geq v_\pi(s))\\
B_{\pi'}(\boldsymbol{v}_\pi) &\geq \boldsymbol{v}_\pi\\
B_{\pi'}(B_{\pi'}(\boldsymbol{v}_\pi)) &\geq B_{\pi'}(\boldsymbol{v}_\pi)\\
B^3_{\pi'}(\boldsymbol{v}_\pi)) &\geq B^2_{\pi'}(\boldsymbol{v}_\pi)\\
&\vdots\\
B^{k+1}_{\pi'}(\boldsymbol{v}_\pi)) &\geq B^k_{\pi'}(\boldsymbol{v}_\pi)\\
&\vdots
\end{aligned}$$

$$\begin{aligned}
\implies v_\pi \leq B_{\pi'}(\boldsymbol{v}_\pi)) \leq B^2_{\pi'}(\boldsymbol{v}_\pi) \leq \cdots \leq \lim_{k\to\infty} B^k_{\pi'}(\boldsymbol{v}_\pi) &= \boldsymbol{v}_{\pi'}\\
\boldsymbol{v}_{\pi'} &\geq \boldsymbol{v}_\pi\\
(\forall s \in \mathcal{S})(v_{\pi'}(s) &\geq v_\pi(s))\\
(\forall s \in \mathcal{S})(q_\pi(s, \pi'(s)) \geq v_\pi(s) &\implies v_{\pi'}(s) \geq v_\pi(s)).
\end{aligned}$$

b) Let $\pi, \pi' \in A^{\mathcal{S}}$ be any pair of deterministic policies such that, for all $s \in \mathcal{S}$, $q_\pi(s, \pi'(s)) \geq v_\pi(s)$ and

that there exists $s' \in \mathcal{S}$ such that $q_\pi(s', \pi'(s')) > v_\pi(s')$:

$$
\begin{aligned}
(\forall s \in \mathcal{S})(q_\pi(s, \pi'(s)) \geq v_\pi(s)) \wedge (\exists s' \in \mathcal{S})(q_\pi(s', \pi'(s')) > v_\pi(s')) \\
B_{\pi'}(\boldsymbol{v}_\pi) > \boldsymbol{v}_\pi \\
B_{\pi'}(B_{\pi'}(\boldsymbol{v}_\pi)) \geq B_{\pi'}(\boldsymbol{v}_\pi) \\
B^3_{\pi'}(\boldsymbol{v}_\pi)) \geq B^2_{\pi'}(\boldsymbol{v}_\pi) \\
\vdots \\
B^{k+1}_{\pi'}(\boldsymbol{v}_\pi)) \geq B^k_{\pi'}(\boldsymbol{v}_\pi) \\
\vdots \\
\implies v_\pi < B_{\pi'}(\boldsymbol{v}_\pi)) \leq B^2_{\pi'}(\boldsymbol{v}_\pi) \leq \cdots \leq \lim_{k\to\infty} B^k_{\pi'}(\boldsymbol{v}_\pi) = \boldsymbol{v}_{\pi'} \\
(\forall s \in \mathcal{S})(v'(s) \geq v(s)) \wedge (\exists s' \in \mathcal{S})(v'(s') > v(s'))
\end{aligned}
$$

$$
\begin{aligned}
(\forall s \in \mathcal{S})(q_\pi(s, \pi'(s)) \geq v_\pi(s)) \wedge (\exists s' \in \mathcal{S})(q_\pi(s', \pi'(s')) > v_\pi(s')) \\
\implies (\forall s \in \mathcal{S})(v'(s) \geq v(s)) \wedge (\exists s' \in \mathcal{S})(v'(s') > v(s'))
\end{aligned}
$$

The policy improvement theorem is hence proved. □

## 5.4 Bellman optimality equation

For a discount factor $\gamma \in [0, 1)$ (with the possibility that $\gamma = 1$ if all episodes terminate) it holds that $\forall s \in \mathcal{S}$:

$$v^*(s) = \max_{a \in \mathcal{A}(s)} \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a)[r + \gamma \cdot v^*(s')],$$

and $v^*(\bot) = 0$, if $\bot$ is defined.

The Bellman optimality equation is a set of $|\mathcal{S}|$ **non-linear** equations with $|\mathcal{S}|$ unknowns.

To prove that $v^*$ is unique, we will use a variant of the Bellman operator.

## 5.5 Bellman backup operator for deterministic policies

Denote by $V$ a vector space of all state-value functions $\boldsymbol{v} = (v(s_1), v(s_2), \ldots, v(s_{|\mathcal{S}|}))$ ($\dim V = |\mathcal{S}|$).
Then the Bellman backup operator $B^* : V \to V$ is defined as:

$$
\begin{aligned}
B^*(\boldsymbol{v}) &= B^*((v(s_1), v(s_2), \ldots, v(s_{|\mathcal{S}|}))^\top), \\
B^*(v(s)) &= \max_{a \in \mathcal{A}(s)} \left( R(s, a) + \gamma \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a) \cdot v(s') \right),
\end{aligned}
$$

where $R : (\mathcal{S} \times A) \to \mathbb{R}$ is a function defined for $s \in \mathcal{S}, a \in \mathcal{A}(s)$:

$$R(s, a) = \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a) \cdot r.$$

### 5.5.1 Fixed point

The iterative application of $B^*$ converges to a unique fixed point $v^*$:

$$\lim_{k\to\infty} (B^*)^k = v^*.$$

*Proof.* Similar to the proof of convergence of policy evaluation. The assumptions are the same (discount factor $\gamma \in [0, 1)$).

We will prove that the Bellman backup operator is a $\gamma$-contraction in the metric space induced by the $L_\infty$ norm and thus converges to a unique fixed point by the Banach contraction principle. As the fixed point is the Bellman optimality equation, the iterative application of the Bellman backup operator (value iteration) converges to the optimal value function $v^*$.

We consider the metric space $(V, d)$, where $V$ is the vector space over the value function vectors and $d$ is a metric induced by an $L_\infty$ norm:

$$\forall \boldsymbol{v} \in V : ||\boldsymbol{v}||_\infty = \max_{s \in \mathcal{S}} |v(s)|$$
$$\forall \boldsymbol{v_1}, \boldsymbol{v_2} \in V : d(\boldsymbol{v_1}, \boldsymbol{v_2}) = ||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty = \max_{s \in S} |v_1(s) - v_2(s)|$$

The operator $B^*$ is a $\gamma$-contraction which means that:

$$\forall \boldsymbol{v_1}, \boldsymbol{v_2} \in V : d(B^*(\boldsymbol{v_1}), B^*(\boldsymbol{v_2})) \leq \gamma \cdot d(\boldsymbol{v_1}, \boldsymbol{v_2})$$

We show that it is true.

**Lemma 5.2.** *The maximal absolute difference of two functions is greater than or equal to the absolute difference of maxima of them.*

$$\max_x |f(x) - g(x)| \geq |\max_x f(x) - \max_x g(x)|$$

*Proof.*

$$\begin{aligned} f(x) &\leq |f(x) - g(x)| + g(x) \\ \max_x f(x) &\leq \max_x (|f(x) - g(x)| + g(x)) \\ &\leq \max_x |f(x) - g(x)| + \max_x g(x) \\ \max_x f(x) - \max_x g(x) &\leq \max_x |f(x) - g(x)|. \end{aligned}$$

Similarly (we swap $f$ and $g$),

$$\max_x g(x) - \max_x f(x) \leq \max_x |g(x) - f(x)| = \max_x |f(x) - g(x)|,$$

thus

$$\max_x |f(x) - g(x)| \geq |\max_x f(x) - \max_x g(x)|.$$

□

**Theorem 5.3.** *The Bellman backup operator $B^*$ is a $\gamma$-contraction.*

*Proof.* By definition of $L_\infty$ norm,

$$||B^*(\boldsymbol{v_1}) - B^*(\boldsymbol{v_2})||_\infty = \max_{s \in \mathcal{S}} |B^*(v_1(s)) - B^*(v_2(s))|.$$

For every $s \in \mathcal{S}$,

$$\begin{aligned} |B^*(v_1(s)) - B^*(v_2(s))| &= \left| \max_{a \in \mathcal{A}(s)} \left( R(s,a) + \gamma \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a) \cdot v_1(s') \right) - \max_{a \in \mathcal{A}(s)} \left( R(s,a) + \gamma \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a) \cdot v_2(s') \right) \right| \\ &\leq \max_{a \in \mathcal{A}(s)} \left| R(s,a) + \gamma \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a) \cdot v_1(s') - R(s,a) - \gamma \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a) \cdot v_2(s') \right| \\ &= \gamma \max_{a \in \mathcal{A}(s)} \left| \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a) \cdot (v_1(s') - v_2(s')) \right|. \end{aligned}$$

We used the maximal absolute difference lemma on the second line.

$$
\begin{aligned}
||B^*(\boldsymbol{v_1}) - B^*(\boldsymbol{v_2})||_\infty &\leq \max_{s \in S} \left( \gamma \max_{a \in \mathcal{A}(s)} \left| \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a) \cdot (v_1(s') - v_2(s')) \right| \right) \\
&= \gamma \max_{s \in S} \max_{a \in \mathcal{A}(s)} \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a) \, |v_1(s') - v_2(s')| \\
&\leq \gamma \max_{s \in S} \max_{a \in \mathcal{A}(s)} \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a) ||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty \\
&= \gamma \cdot ||\boldsymbol{v_1} - \boldsymbol{v_2}||_\infty
\end{aligned}
$$

□

Therefore, $B^*$ is a $\gamma$-contraction. By the Banach contraction principle, we deduce that the iterative application of the Bellman backup operator converges to a single fixed point for arbitrary initial state-value function $\boldsymbol{v} \in V$:

$$\lim_{k \to \infty} (B^*)^k(\boldsymbol{v}) = \boldsymbol{v^*}$$

But we know the fixed point is the solution $v^*$ to the Bellman optimality equation. That means the iterative application of the Bellman backup operator converges to an optimal state-value function $v^*$ and that the Bellman optimality equation has a unique solution. □

## 5.6 Greedy policy

We can directly use the policy improvement theorem to construct a policy $\pi \in A^{\mathcal{S}}$ which meets the conditions of the theorem.

For all $s \in \mathcal{S}$:

$$
\begin{aligned}
\pi'(s) &= \arg\max_{a \in \mathcal{A}(s)} q_\pi(s, a) \\
&= \arg\max_{a \in \mathcal{A}(s)} \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a)[r + \gamma v_\pi(s')]
\end{aligned}
$$

We show that the greedy policy meets the conditions of the theorem. For all $s \in \mathcal{S}$,

$$q_\pi(s, \pi'(s)) = \max_{a \in \mathcal{A}(s)} q_\pi(s, a) \geq q_\pi(s, \pi(s)) = v_\pi(s)$$

The greedy policy meets the conditions of the policy improvement theorem, so it is at least as good as the original policy $\pi$.

Moreover, unless the policy $\pi$ is already optimal, it is better than the original policy $\pi$. Assume that the greedy policy $\pi'$ has the same state-value function as $\pi$. Then

$$
\begin{aligned}
\boldsymbol{v}_\pi &= \boldsymbol{v}_{\pi'} \\
\forall s \in \mathcal{S} : v_{\pi'}(s) &= \max_a \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a)[r + \gamma \cdot v'_\pi(s')] \\
v_{\pi'} &= v_*,
\end{aligned}
$$

by the Bellman optimality equation.

The process of making a new policy $\pi'(s)$ that improves on an original policy, by making it greedy with respect to the value function of the original policy $\pi$, is called policy improvement.

# 6 Policy iteration

Policy evaluation can be used to obtain a state-value function $v_\pi$, policy improvement can be used to obtain a better policy from $v_\pi$. When alternating policy evaluation and policy improvement steps, the resulting policy and state-value function converge to the optimal $\pi_*$ and $v_*$ (a finite MDP has only a finite number of policies and policy improvement always improves the policy unless it is already the optimal one):

$$\pi_0 \xrightarrow{E} v_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} v_{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} \cdots \xrightarrow{I} \pi_* \xrightarrow{E} v_*$$

**Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi^*$**

Algorithm parameter: threshold $\epsilon > 0$ determining accuracy of policy evaluation

1. Initialization
   Initialize $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$, $V(\bot) = 0$.

2. Policy Evaluation
   **repeat**
   - $\Delta = 0$;
   - **foreach** $s \in \mathcal{S}$ **do**
     - $v = V(s)$;
     - $V(s) = \sum_{s',r} p(s', r \mid s, \pi(s))[r + \gamma \cdot V(s')]$;
     - **if** $|v - V(s)| > \Delta$ **then**
       - $\Delta = |v - V(s)|$;
     - **end if**
   - **end foreach**

   **until** $\Delta < \epsilon$;

3. Policy Improvement
   *policy-stable* = *true*;
   **foreach** $s \in \mathcal{S}$ **do**
   - *old-action* = $\pi(s)$;
   - $\pi(s) = \arg\max_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma \cdot V(s')]$;
   - **if** *old-action* $\neq \pi(s)$ **then**
     - *policy-stable* = *false*;
   - **end if**

   **end foreach**
   **if** *policy-stable* **then**
   - return $V \approx v^*, \pi \approx \pi^*$

   **end if**
   go to 2.

# 7 Value iteration

Policy iteration converges to the optimal policy in the limit, but policy evaluation is expensive (it converges in the limit). Value iteration is a way to find an optimal policy without policy evaluation nor policy improvement steps. It uses the Bellman optimality equation as an update rule.

### 7.1 Synchronous value iteration

Value iteration (for deterministic policies) is defined as:

$$v_0(s) = \begin{cases} \text{arbitrary} & s \in \mathcal{S}, \\ 0 & s = \bot. \end{cases}$$

$$v_{k+1}(s) = \max_{a \in \mathcal{A}(s)} \sum_{\substack{s' \in \mathcal{S}^+ \\ r \in \mathcal{R}}} p(s', r \mid s, a) \left[ r + \gamma \cdot v_k(s') \right]$$

and $v_k(\bot) = 0$ for all $k$, if present.

### 7.2 Convergence

*Proof.* The iterative application of the Bellman backup operator is equivalent to value iteration. As it converges to a single fixed point $v^*$, which is the state-value function of the optimal policy, also the value iteration algorithm converges to the state-value function of the optimal policy. □

### 7.3 Asynchronous value iteration

As with policy evaluation, also value iteration converges in its asynchronous form. The algorithm iteratively approximates the optimal state-value function and when the updates are reasonably small, it returns the greedy policy based on the calculated state-value function. In the limit, the state-value function and thus the greedy policy are optimal.

**Value Iteration, for estimating $\pi \approx \pi^*$**

Algorithm parameter: threshold $\epsilon > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(\bot) = 0$
**repeat**
  $\Delta = 0$;
  **foreach** $s \in \mathcal{S}$ **do**
    $v = V(s)$;
    $V(s) = \max_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma V(s')]$;
    $\Delta = \max(\Delta, |v - V(s)|)$;
  **end foreach**
**until** $\Delta < \epsilon$;
Output a deterministic policy, $\pi \approx \pi^*$, such that
$\pi(s) = \arg\max_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma V(s')]$

## Links and sources

- `http://incompleteideas.net/book/the-book.html` - Reinforcement Learning Chapter 4 - Sutton, Barto
- `https://cs.stanford.edu/people/karpathy/reinforcejs/gridworld_dp.html` - Original Gridworld Demo
- `https://www.fi.muni.cz/~xivora/reinforcejs/gridworld_dp.html` - RL Book Gridworld Demo
- `https://www.andrew.cmu.edu/course/10-703/slides/lecture4_valuePolicyDP-9-10-2018.pdf` - policy evaluation convergence proof
- `https://www.springer.com/gp/book/9783319015859` - Banach contraction principle proof
- `https://www.cse.iitb.ac.in/~shivaram/resources/ijcai-2017-tutorial-policyiteration/tapi.pdf` - policy improvement theorem proof
- `https://www.cs.cmu.edu/~arielpro/15381f16/c_slides/781f16-11.pdf` - Bellman backup operator is a contraction proof
- `http://www.cs.cmu.edu/afs/cs/academic/class/15780-s16/www/slides/mdps.pdf` - value iteration convergence proof (not used in this text)
- `https://inst.eecs.berkeley.edu/~cs294-40/fa08/scribes/lecture3.pdf` - asynchronous value iteration convergence proof sketch (for an interested reader)