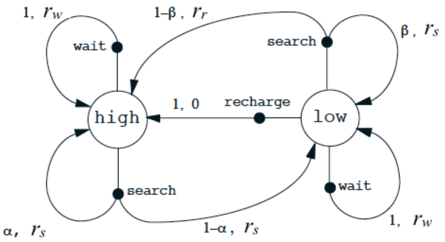
2 Markov Decision Processes

Plan: Understand the interplay between environment, observations, agents, actions, and rewards. We assume that the observation emitted by the environment is acquired by the agent in full. We also assume we have access to the underlying probability distributions.

2.1 The recycling robot

Problem setting At the beginning of each journey, the battery level of a recycling robot can be either *high* or *low*. If the level is *high*, the robot can either *search* or *wait*, if *low*, the robot can also *recharge*. Depending on the battery level and the action chosen, the new battery level is random: we only know the probability with which the battery level either changes or remains the same. The reward $r(k)$ received at the k-th journey, instead, is deterministic: it depends on the battery level, the action chosen and the new battery level (see Figure 1). The *cumulative reward* is the sum of the rewards in n journeys as in

$$\sum_{k=0}^{n} r(k)$$



s	a	s'	probability	reward
high	search	high	α	r_s
high	search	low	$1-\alpha$	r_s
low	search	high	$1-\beta$	$r_r < 0$
low	search	low	β	r_s
high	wait	high	1	r_w
high	wait	low	0	r_w
low	wait	high	0	r_w
low	wait	low	1	r_w
low	recharge	high	1	0
low	recharge	low	0	0

Figure 1: Diagram and table for the recycling robot: for each selected action with a certain battery level, the arrow from the filled circle to the new battery level displays the probability of reaching that level and the corresponding reward.

Question Which action should the robot take given the battery level and what is the maximum expected cumulative reward that can be achieved?

Fix some values for r_w, r_s, and r_r and adopt any deterministic policy. What is the expected cumulative reward starting from either battery level? Even the solution to this much simpler problem seems quite complicated.

2.2 Basics

We denote random variables with capital letters, their values in lower case. Matrix and vectors are in boldface.

Definition 2.1. *Markov Decision Process* (MDP)
A *Markov Decision Process* $\mathcal{M}$ is a 4-tuple $\mathcal{M} = \{\mathtt{S}, \mathtt{A}, \mathtt{R}, \mathbf{T}^a\}$ where

1. $\mathtt{S}$, a (finite) set of states,

2. $\mathtt{A}$, a (finite) set of actions,

3. $\mathtt{R}$ a set of rewards, and

4. $\mathbf{T}^a$ a transition matrix with $a \in \mathtt{A}$

Observation 2.1. *Markov property*
A Markov Decision Process is memoryless: if we know the state at a certain time t, we do not gain any information by knowing the process previous history. This is equivalent to write

$$\Pr(S_{t+1}|S_1 \ldots S_t) = \Pr(S_{t+1}|S_t)$$

An MDP is fully characterised by two probability distributions.

Definition 2.2. *Policy*
A *policy*

$$\pi(A_t|S_t) = \Pr(A_t|S_t)$$

is the probability distribution of the action A_t conditioned to the environment being in the state S_t, where $0 \le \pi(a, s) \le 1$ for all $a \in \mathtt{A}$ and $s \in \mathtt{S}$ and $\sum_{a \in \mathtt{A}} \pi(a, s) = 1$ for all $s \in \mathtt{S}$.

Observation 2.2. *Deterministic policy*
If for each state s the action a executed is always the same, then we write $a = a(s)$ and π reduces to a Kroneker's delta

$$\pi(a|s) = \delta_{aa(s)}$$

Definition 2.3. *Joint probability distribution*
The second ingredient is the *joint probability* with which the environment is in the state S_{t+1} and emits the reward R_{t+1} at time $t+1$, conditioned to having observed the state S_t and executed the action A_t at time t

$$\mathrm{P}_J(S_{t+1}, R_{t+1}|S_t, A_t) = \Pr(S_{t+1}, R_{t+1}|S_t, A_t)$$

We now use P_J to determine the transition probabilities between states.

Definition 2.4. *Transition matrix*
The $|\mathtt{S}| \times |\mathtt{S}|$ transition matrix $\mathbf{T}^a$, whose element $T^a_{ss'}$ gives the transition probability from state s to state s' under the action a, is obtained by summing over all possible rewards, or

$$T^a_{ss'} = \sum_{r' \in \mathtt{R}} \mathrm{P}_J(s', r'|s, a) = \sum_{r' \in \mathtt{R}} \Pr(s', r'|s, a) = \Pr(s'|s, a)$$

Observation 2.3. *Stochastic property*
The matrix $\mathbf{T}^a$ is stochastic, since $\forall a \in \mathtt{A}$

$$\sum_{s' \in \mathtt{S}} T^a_{ss'} = 1 \quad \forall s \in \mathtt{S}$$

with $0 \le T^a_{ss'} \le 1$ for all s and $s' \in \mathtt{S}$.

Since

$$\mathrm{P}_J(s', r'|s, a) = \Pr(r'|s, a, s')\Pr(s'|s, a) = \Pr(r'|s, a, s')T^a_{ss'}$$

we can viewed the joint probability P_J as the product between the probability of receiving the reward r' conditioned to $S_t = s$, $A_t = a$, and $S_{t+1} = s'$ with the transition probability from state s to state s' under the action a.

Back to the multi-armed bandit In this case the *environment* is always in one and the same **state**, while the *agent* picks one of finitely many **actions** and obtains a random **reward**. We can simplify the notation and agree that

- a policy $\pi(A_t)$ is a probability distribution on which arm to choose
- the joint probability reduces to the probability distribution of the rewards conditioned the chosen arm, $\Pr(R_{t+1}|A_t)$.

Back to the recycling robot We have $\mathtt{S} = \{low, high\}$, and $\mathtt{A} = \{s, w, r\}$ with s for *search*, w for *wait*, and r for *recharge*. With respect to the general setting we presented, the rewards are deterministic: $\mathtt{R} = \{\mathtt{0}, r_w, r_s, r_r\}$ with $r_r < 0$. With two states, three actions, and four rewards we have

$$|\mathtt{S}| \times |\mathtt{A}| \times \mathtt{S}| \times |\mathtt{R}| = 2 \times 3 \times 2 \times 4 = 48$$

probabilities, only eight of which strictly positive,

$$\begin{array}{ll}
\mathrm{P}_J(low, r_s|\, low, s) = \beta & \mathrm{P}_J(high, r_r|\, low, s) = 1 - \beta \\
\mathrm{P}_J(low, r_w|\, low, w) = 1 & \mathrm{P}_J(high, 0|\, low, r) = 1 \\
\mathrm{P}_J(low, r_s|\, high, s) = 1 - \alpha & \mathrm{P}_J(high, r_s|\, high, s) = \alpha \\
\mathrm{P}_J(high, r_w|\, high, w) = 1 & \mathrm{P}_J(high, 0|\, high, r) = 1
\end{array}$$

Since the reward $\mathcal{R}(s, a, s')$ is deterministic (see Figure 1), the transition matrices

$$\mathbf{T}^{\mathtt{s}} = \begin{pmatrix} \beta & 1-\beta \\ 1-\alpha & \alpha \end{pmatrix}, \quad \mathbf{T}^{\mathtt{w}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{T}^{\mathtt{r}} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$$

are simply obtained by the formula

$$T^a_{ss'} = \mathrm{P}_J(s', r'|s, a)$$

2.3 Prediction problem

Main ingredients

In the most general case, the immediate reward R_{t+1} can be viewed as a random variable which takes on the value r' according to the probability $\Pr(r'|s, a, s')$. For the multi-armed bandit there is no state to consider and the probability is only conditioned to the arm a, or $\Pr(r'|a)$. For the recycling robot, the reward for the triplet s, a, s' is fixed and the probability reduces to a Kroneker delta.

Definition 2.5. *Expected immediate reward*
We can write the expected immediate reward as a function of $S_t = s$, $A_t = a$, and $S_{t+1} = s'$ as

$$\mathcal{R}(s, a, s') = \sum_{r' \in \mathtt{R}} \mathrm{P}_J(s', r'|s, a) r' = \sum_{r' \in \mathtt{R}} \Pr(r'|s, a, s') T^a_{ss'} r' \qquad (1)$$

If we sum over all possible actions at time t and states at time $t+1$ we can obtain the expected immediate reward as a function of $S_t = s$ for a fixed policy π or

$$\mathcal{R}^\pi(s) = \sum_{a \in \mathtt{A}} \sum_{s' \in \mathtt{S}} \pi(a|s) \mathcal{R}(s, a, s') \qquad (2)$$

Definition 2.6. *Discounted reward*

$$\sum_{k=0}^{+\infty} \gamma^k R_{t+1+k} \quad \text{with} \quad 0 \leq \gamma < 1$$

For $\gamma = 0$ it reduces to the immediate reward (myopic view). It can also include the case in which the environment ends up in a terminal state after a finite number of steps.

In the case of *finite horizon* it is customary to set $\gamma = 1$ and write for the total reward

$$\sum_{k=0}^{K} R_{t+1+k}$$

Since

$$\sum_{k=0}^{+\infty} \gamma^k = \frac{1}{1-\gamma}$$

the quantity $1/(1-\gamma)$ can be used to estimate the number of steps after which the contribution to the discounted reward becomes less relevant.

We are particularly interested in evaluating the expected discounted reward which, for a fixed policy π, can be obtained starting from the state $S_t = s$.

Definition 2.7. *Value function*
For a fixed policy π with the environment in the state $S_t = s$ at time t we define the value of each state $s \in \mathrm{S}$ as

$$\mathcal{V}^\pi(s) = \mathbb{E}\left[\sum_{k=0}^{+\infty} \gamma^k R_{t+1+k}\right] = \sum_{a \in \mathtt{A}} \sum_{s' \in \mathtt{S}} \sum_{r' \in \mathtt{R}} \pi(a|s) \mathrm{P}_J(s', r'|s, a) \left(r' + \gamma r'' + \gamma^2 r''' + \ldots\right)$$

The value function for the state s is the expected discounted reward starting from the state s while adopting the policy π. We now show that $\mathcal{V}^\pi(s)$ can be computed recursively.

First term For all $s \in \mathtt{S}$, using Equation2 (1) and (2), we simply have

$$\mathbb{E}[R_{t+1}] = \sum_{a \in \mathtt{A}} \sum_{s' \in \mathtt{S}} \sum_{r' \in \mathtt{R}} \pi(a|s) \mathrm{P}_J(s', r'|s, a) r' = \sum_{a \in \mathtt{A}} \sum_{s' \in \mathtt{S}} \pi(a|s) \mathcal{R}(s, a, s') = \mathcal{R}^\pi(s)$$

The other terms What about the expected value of $\mathbb{E}[R_{t+1+k}]$ for $k > 0$? We start with $\mathbb{E}[R_{t+2}]$. Note that

$$\mathbb{E}[R_{t+2}] = \sum_{a \in \mathtt{A}} \sum_{s'' \in \mathtt{S}} \sum_{r'' \in \mathtt{R}} \pi(a|s) \mathrm{Pr}(s'', r''|s, a) r'' \neq \mathcal{R}^\pi(s')$$

because s and a refer to the state the environment is in at time t and not $t + 1$!

Using twice the law of total probability and the Markov property, we have that

$$\begin{aligned} \mathrm{Pr}(s'', r''|s, a) &= \sum_{s' \in \mathtt{S}} \sum_{a' \in \mathtt{A}} \mathrm{Pr}(s'', r''|s', a', s, a) \mathrm{Pr}(a'|s', s, a) \mathrm{Pr}(s'|s, a) \\ &= \sum_{s' \in \mathtt{S}} \sum_{a' \in \mathtt{A}} \mathrm{P}_J(s'', r''|s', a') \pi(a'|s') T^a_{ss'} \end{aligned}$$

and, hence,

$$\begin{aligned}
\mathbb{E}[R_{t+2}] &= \sum_{a \in \mathtt{A}} \sum_{s'' \in \mathtt{S}} \sum_{r'' \in \mathtt{R}} \pi(a|s) \mathrm{Pr}(s'', r''|s, a) r'' \\
&= \sum_{a \in \mathtt{A}} \sum_{s'' \in \mathtt{S}} \sum_{r'' \in \mathtt{R}} \sum_{a' \in \mathtt{A}} \sum_{s' \in \mathtt{S}} \pi(a|s) \mathrm{P}_J(s'', r''|s', a') \pi(a'|s') T^a_{ss'} r'' \\
&= \sum_{a \in \mathtt{A}} \pi(a|s) \sum_{s' \in \mathtt{S}} T^a_{ss'} \sum_{a' \in \mathtt{A}} \sum_{s'' \in \mathtt{S}} \pi(a'|s') \sum_{r'' \in \mathtt{R}} \mathrm{P}_J(s'', r''|s', a') r'' \\
&= \sum_{a \in \mathtt{A}} \pi(a|s) \sum_{s' \in \mathtt{S}} T^a_{ss'} \sum_{a' \in \mathtt{A}} \sum_{s'' \in \mathtt{S}} \pi(a'|s') \mathcal{R}(s', a', s'') \\
&= \sum_{a \in \mathtt{A}} \pi(a|s) \sum_{s' \in \mathtt{S}} T^a_{ss'} \mathcal{R}^\pi(s') = \sum_{s' \in \mathtt{S}} T^\pi_{ss'} \mathcal{R}^\pi(s')
\end{aligned}$$

where the $|\mathtt{S}| \times |\mathtt{S}|$ matrix $\mathbf{T}^\pi$ defined as

$$T^\pi_{ss'} = \sum_{a \in \mathtt{A}} \pi(a|s) T^a_{ss'}$$

is stochastic as well since

$$\sum_{s' \in \mathtt{S}} T^\pi_{ss'} = \sum_{s' \in \mathtt{S}} \sum_{a \in \mathtt{A}} \pi(a|s) T^a_{ss'} = \sum_{a \in \mathtt{A}} \pi(a|s) \sum_{s' \in \mathtt{S}} T^a_{ss'} = \sum_{a \in \mathtt{A}} \pi(a|s) \cdot 1 = 1$$

Repeating the same reasoning for $\mathbb{E}[R_{t+3}]$ we obtain

$$\mathbb{E}[R_{t+3}] = \sum_{s' \in \mathtt{S}} T^\pi_{ss'} \sum_{s'' \in \mathtt{S}} T^\pi_{s's''} \mathcal{R}^\pi(s'')$$

Observation 2.4. *Bellman linear equation*
For all $s \in \mathtt{S}$ we thus have that the value function satisfies the linear recursive equation

$$\begin{aligned}
\mathcal{V}^\pi(s) &= \mathcal{R}^\pi(s) + \gamma \sum_{s' \in \mathtt{S}} T^\pi_{ss'} \mathcal{R}^\pi(s') + \gamma^2 \sum_{s' \in \mathtt{S}} T^\pi_{ss'} \sum_{s'' \in \mathtt{S}} T^\pi_{s's''} \mathcal{R}^\pi(s'') + \dots \\
&= \mathcal{R}^\pi(s) + \gamma \sum_{s' \in \mathtt{S}} T^\pi_{ss'} \left(\mathcal{R}^\pi(s') + \gamma \sum_{s'' \in \mathtt{S}} T^\pi_{s's''} \mathcal{R}^\pi(s'') + \dots \right) \\
&= \mathcal{R}^\pi(s) + \gamma \sum_{s' \in \mathtt{S}} T^\pi_{ss'} \mathcal{V}^\pi(s') \qquad (3)
\end{aligned}$$

In words ...

Policy evaluation

Through a linear system of equations If $\mathcal{V}^\pi(s)$ and $\mathcal{R}^\pi(s)$ are seen as the s-th component of the $|\mathtt{S}|$-dimensional vectors $\mathbf{V}^\pi$ and $\mathbf{R}^\pi$, we can rewrite Equation (3) as

$$\mathbf{V}^\pi = \mathbf{R}^\pi + \gamma \mathbf{T}^\pi \mathbf{V}^\pi$$

the solution of which is

$$\mathbf{V}^\pi = (\mathbf{I} - \gamma \mathbf{T}^\pi)^{-1} \mathbf{R}^\pi$$

Existence and uniqueness Since $\mathbf{T}^\pi$ is stochastic, all of its eigenvalues are no greater than 1 in module. Therefore, since $0 \leq \gamma < 1$, the module of the smallest eigenvalue of $\mathbf{I} - \gamma \mathbf{T}^\pi$ is strictly larger than 0 and $\mathbf{I} - \gamma \mathbf{T}^\pi$ is always invertible.

Through iterative policy evaluation An alternative procedure to compute $\mathcal{V}^{\pi}$ can be obtained iteratively. Starting off with an arbitrary value for $\mathcal{V}_0^{\pi}$, from $k = 0$ on we write

$$\mathcal{V}_{k+1}^{\pi}(s) = \sum_{a \in \mathtt{A}} \sum_{s' \in \mathtt{S}} \pi(a|s) \left(\mathcal{R}(s, a, s') + \gamma T_{ss'}^{a} \mathcal{V}_k^{\pi}(s') \right) \tag{4}$$

Theorem 2.1. *A fixed point theorem*

$$\lim_{k \to +\infty} \mathcal{V}_k^{\pi}(s) = \mathcal{V}^{\pi}(s) \tag{5}$$

Proof
Indeed, we have

$$\begin{aligned} \left|\mathcal{V}_{k+1}^{\pi}(s) - \mathcal{V}^{\pi}(s)\right| &= \gamma \left| \sum_{a \in \mathtt{A}} \pi(a, s) \sum_{s' \in \mathtt{S}} T_{ss'}^{a} \left(\mathcal{V}_k^{\pi}(s') - \mathcal{V}^{\pi}(s') \right) \right| \\ &\leq \gamma \sum_{a \in \mathtt{A}} \pi(a, s) \sum_{s' \in \mathtt{S}} T_{ss'}^{a} \left| \mathcal{V}_k^{\pi}(s') - \mathcal{V}^{\pi}(s') \right| \\ &\leq \gamma e_k \end{aligned}$$

where

$$e_k = \max_s |\mathcal{V}_k^{\pi}(s) - \mathcal{V}^{\pi}(s)|$$

Since

$$e_{k+1} \leq \gamma e_k$$

we find that

$$\lim_{k \to +\infty} e_k = 0$$

■

The recycling robot, once again The recycling robot example the value function can be computed in closed form. We pick a policy $\pi = 1/3$ for each of the three actions and independently of the battery level. This clearly includes the silly case in which we recharge a battery without reason but simplifies the math. Not a bright policy, but it does the job.

For the stochastic matrix $\mathbf{T}^{\pi}$ we have

$$\mathbf{T}^{\pi} = \frac{1}{3} \left(\mathbf{T}^{s} + \mathbf{T}^{w} + \mathbf{T}^{r} \right) = \frac{1}{3} \begin{pmatrix} 1 + \beta & 2 - \beta \\ 1 - \alpha & 2 + \alpha \end{pmatrix}$$

Therefore, we find

$$\mathbf{I} - \gamma \mathbf{T}^{\pi} = \frac{1}{3} \begin{pmatrix} 3 - \gamma(1 + \beta) & -\gamma(2 - \beta) \\ -\gamma(1 - \alpha) & 3 - \gamma(2 + \alpha) \end{pmatrix}$$

and, through simple but lengthy algebra, we obtain

$$(\mathbf{I} - \gamma \mathbf{T}^{\pi})^{-1} = \frac{1}{(1 - \gamma)(3 - \gamma(\alpha + \beta))} \begin{pmatrix} 3 - \gamma(2 + \alpha) & \gamma(2 - \beta) \\ \gamma(1 - \alpha) & 3 - \gamma(1 + \beta) \end{pmatrix}$$

Since the only non zero expected immediate rewards are

$$\begin{array}{lll}
\mathcal{R}(low,s,low)=\beta r_s & \mathcal{R}(low,s,high)=(1-\beta)r_r & \mathcal{R}(low,w,low)=r_w \\
\mathcal{R}(high,s,low)=(1-\alpha)r_s & \mathcal{R}(high,s,high)=\alpha r_s &
\end{array}$$

for the expected immediate reward we find

$$\mathbf{R}^\pi = \frac{1}{3}\left(\begin{array}{c} (1-\beta)r_r+\beta r_s+r_w \\ r_s \end{array}\right)$$

We now plug in some numbers and check the results. If we set $\alpha=\beta=\gamma=1/2$, $r_s=2$, $r_w=0$ and $r_r=-4$ we obtain

$$(\mathbf{I}-\gamma\mathbf{T}^\pi)^{-1} = \frac{1}{5}\left(\begin{array}{cc} 7 & 3 \\ 1 & 9 \end{array}\right)$$

and since

$$\begin{array}{lll}
\mathcal{R}(low,s,low)=1 & \mathcal{R}(low,s,high)=-2 & \mathcal{R}(low,w,low)=0 \\
\mathcal{R}(high,s,low)=1 & \mathcal{R}(high,s,high)=1 &
\end{array}$$

we have

$$\mathbf{R}^\pi = \frac{1}{3}\left(\begin{array}{c} -1 \\ 2 \end{array}\right) \simeq \left(\begin{array}{c} -0.33 \\ 0.66 \end{array}\right)$$

Straightforward matrix-vector multiplication gives

$$\mathbf{V}^\pi = (\mathbf{I}-\gamma\mathbf{T}^\pi)^{-1}\,\mathbf{R}^\pi = \frac{1}{15}\left(\begin{array}{cc} 7 & 3 \\ 1 & 9 \end{array}\right)\left(\begin{array}{c} -1 \\ 2 \end{array}\right) = \frac{1}{15}\left(\begin{array}{c} -1 \\ 17 \end{array}\right) \simeq \left(\begin{array}{c} -0.07 \\ 1.13 \end{array}\right)$$

Equation (4) thus gives

$$\begin{array}{rcl}
\mathcal{V}^\pi_{k+1}(low) & = & \frac{1}{3}\left[\left(\frac{1}{2}\mathcal{R}(low,s,low)+\frac{1}{2}\mathcal{R}(low,s,high)\right)+\frac{1}{2}\left(\frac{1}{2}\mathcal{V}^\pi_k(low)+\frac{1}{2}\mathcal{V}^\pi_k(high)\right)\right. \\
 & + & \left(\frac{1}{2}\mathcal{R}(low,w,low)+\frac{1}{2}\mathcal{R}(low,w,high)\right)+\frac{1}{2}\left(1\cdot\mathcal{V}^\pi_k(low)+0\cdot\mathcal{V}^\pi_k(high)\right) \\
 & + & \left.\left(\frac{1}{2}\mathcal{R}(low,r,low)+\frac{1}{2}\mathcal{R}(low,r,high)\right)+\frac{1}{2}\left(0\cdot\mathcal{V}^\pi_k(low)+1\cdot\mathcal{V}^\pi_k(high)\right)\right]
\end{array}$$

and

$$\begin{array}{rcl}
\mathcal{V}^\pi_{k+1}(high) & = & \frac{1}{3}\left[\left(\frac{1}{2}\mathcal{R}(high,s,low)+\frac{1}{2}\mathcal{R}(high,s,high)\right)+\frac{1}{2}\left(\frac{1}{2}\mathcal{V}^\pi_k(low)+\frac{1}{2}\mathcal{V}^\pi_k(high)\right)\right. \\
 & + & \left(\frac{1}{2}\mathcal{R}(high,w,low)+\frac{1}{2}\mathcal{R}(high,w,high)\right)+\frac{1}{2}\left(0\cdot\mathcal{V}^\pi_k(low)+1\cdot\mathcal{V}^\pi_k(high)\right) \\
 & + & \left.\left(\frac{1}{2}\mathcal{R}(high,r,low)+\frac{1}{2}\mathcal{R}(high,r,high)\right)\frac{1}{2}\left(0\cdot\mathcal{V}^\pi_k(low)+1\cdot\mathcal{V}^\pi_k(high)\right)\right]
\end{array}$$

Upon substitution we finally obtain

$$\mathcal{V}^\pi_{k+1}(low) = -\frac{1}{3}+\frac{1}{4}\mathcal{V}^\pi_k(low)+\frac{1}{4}\mathcal{V}^\pi_k(high)$$

and

$$\mathcal{V}^\pi_{k+1}(high) = \frac{2}{3}+\frac{1}{12}\mathcal{V}^\pi_k(low)+\frac{5}{12}\mathcal{V}^\pi_k(high)$$

Setting $\mathcal{V}^\pi_0(low)=\mathcal{V}^\pi_0(high)=0$, the iterative policy evaluation converges (albeit quite slowly). Here are the first few iterations for our example

$$\mathcal{V}_1^\pi(low) = -0.33 \;\rightarrow\; \mathcal{V}_2^\pi(low) = -0.25 \;\rightarrow\; \mathcal{V}_3^\pi(low) = -0.17 \;\rightarrow \dots \quad \mathcal{V}^\pi(low) \simeq -0.07$$

$$\mathcal{V}_1^\pi(high) = 0.67 \;\rightarrow\; \mathcal{V}_2^\pi(high) = 0.92 \;\rightarrow\; \mathcal{V}_3^\pi(high) = 1.03 \;\rightarrow \dots \quad \mathcal{V}^\pi(high) \simeq 1.13$$

2.4 Control problem

Policy iteration

If

$$\mathcal{V}^*(s) = \max_\pi \mathcal{V}^\pi(s) \;\; \forall s \in \mathtt{S}$$

we clearly have

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

and we can state a fundamental result.

Theorem 2.2. *Bellman optimality equation*
The optimal value function is the unique solution to the equation

$$\mathcal{V}^*(s) = \max_{a \in \mathtt{A}} \sum_{s' \in \mathtt{S}} \left(\mathcal{R}(s,a,s') + \gamma T^a_{ss'} \mathcal{V}^*(s')\right)$$

Furthermore, an optimal policy is deterministic with $a(s)$ given by

$$a(s) = \arg\max_{a \in \mathtt{A}} \sum_{s' \in \mathtt{S}} \left(\mathcal{R}(s,a,s') + \gamma T^a_{ss'} \mathcal{V}^*(s')\right)$$

Proof
We first prove the existence of an optimal value function satisfying the Bellman optimality equation through a deterministic policy. Then, its uniqueness.

Existence We proceed by constructing an optimal deterministic policy solution of the Bellman optimality equation. The procedure we describe is known as **policy iteration**.

We start off with an arbitrary policy π_0 and compute $\mathcal{V}^{\pi_0}(s)$ as the fixed point of Equation (5). Then, we set $n = 0$ and repeat indefintely

Policy improvement: *greedily* select a deterministic policy $\pi(a|s) = \delta_{aa(s)}$ according to

$$a(s) = \arg\max_{a \in \mathtt{A}} \sum_{s' \in \mathtt{S}} \left(\mathcal{R}(s,a,s') + \gamma T^a_{ss'} \mathcal{V}^{\pi_0}(s')\right)$$

Value function estimation: compute $\mathcal{V}^\pi(s)$ as the fixed point of Equation (5) (or by solving the system of linear equations)

Policy update: $\pi_0 \leftarrow \pi$ and $n \leftarrow n + 1$

At each iteration we obtain

$$\mathcal{V}^\pi(s) \geq \mathcal{V}^{\pi_0}(s) \quad \forall s \in \mathtt{S}$$

Indeed, if

$$\mathbb{E}_0[\cdot] = \sum_{a \in \mathtt{A}} \sum_{s' \in \mathtt{S}} \sum_{r' \in \mathtt{R}} \pi_0(a|s) P_J(s', r'|s, a)(\cdot)$$

we have

$$\begin{aligned}
\mathcal{V}^{\pi_0}(s) &= \mathbb{E}_0[R_{t+1} + \gamma V^{\pi_0}(s')] \leq \mathbb{E}[R_{t+1} + \gamma V^{\pi_0}(s')] \\
&= \mathcal{R}^{\pi}(s) + \gamma \mathbb{E}[V^{\pi_0}(s')] \\
&= \mathcal{R}^{\pi}(s) + \gamma \mathbb{E}[\mathbb{E}_0[R_{t+2} + \gamma V^{\pi_0}(s'')]] \\
&\leq \mathcal{R}^{\pi}(s) + \gamma \mathbb{E}[R_{t+2} + \gamma V^{\pi_0}(s'')] \\
&= \mathcal{R}^{\pi}(s) + \gamma \mathcal{R}^{\pi}(s') + \gamma^2 \mathbb{E}[\mathbb{E}_0[R_{t+3} + \gamma V^{\pi_0}(s''')]] \\
&\leq \mathcal{R}^{\pi}(s) + \gamma \mathcal{R}^{\pi}(s') + \gamma^2 \mathbb{E}[R_{t+3} + \gamma V^{\pi_0}(s''')] \\
&= \mathcal{R}^{\pi}(s) + \gamma \mathcal{R}^{\pi}(s') + \gamma^2 \mathcal{R}^{\pi}(s'') + \gamma^3 \mathbb{E}[\mathbb{E}_0[R_{t+4} + \gamma V^{\pi_0}(s'''')]] \\
&= \ldots \\
&\leq \mathcal{V}^{\pi}(s)
\end{aligned}$$

where the inequality signs are due to the replacement of $\mathbb{E}_0[\cdot]$ with $\mathbb{E}[\cdot]$ and we repeatedly use the recursive formula for the value function $\mathcal{V}^{\pi_0}$.

■

Since at each iteration the value function does not decrease $\forall s \in \mathtt{S}$ and the number of states is finite, we have that for $n \to \infty$ the policy π must converge to some policy π^{∞} for which, by construction,

$$\mathcal{V}^{\pi^{\infty}}(s) \geq \mathcal{V}^{\pi}(s) \quad \forall s \in \mathtt{S}$$

Thus, the existence of an optimal value function (and of an optimal policy) is established.

Uniqueness We proceed by contradiction. We assume we have two different optimal solutions

$$V_1(s) = \max_{a \in \mathtt{A}} \sum_{s' \in \mathtt{S}} \left(\mathcal{R}(s,a,s') + \gamma T^a_{ss'} V_1(s') \right)$$

and

$$V_2(s) = \max_{a \in \mathtt{A}} \sum_{s' \in \mathtt{S}} \left(\mathcal{R}(s,a,s') + \gamma T^a_{ss'} V_2(s') \right)$$

Without loss of generality, since $V_1(s) \neq V_2(s)$ for some $s \in \mathtt{S}$, we write

$$V_2(s) = V_1(s) + e(s)$$

with s^* the state for which

$$s^* = \arg\max_{s \in \mathtt{S}} e(s) > 0$$

In particular, we have that for the state s^*

$$V_2(s^*) = V_1(s^*) + e(s^*) \tag{6}$$

However, since $\gamma < 1$,

$$\gamma \max_{a \in \mathtt{A}} \sum_{s'} T^a_{s^*s'} e(s') \leq \gamma e(s^*) < e(s^*)$$

from the fact that

$$\begin{aligned}
V_2(s^*) &= \max_{a \in \mathtt{A}} \sum_{s' \in \mathtt{S}} \left(\mathcal{R}(s^*,a,s') + \gamma T^a_{s^*s'} V_1(s') + \gamma T^a_{s^*s'} e(s') \right) \\
&\leq \underbrace{\max_{a \in \mathtt{A}} \sum_{s' \in \mathtt{S}} \left(\mathcal{R}(s^*,a,s') + \gamma T^a_{s^*s'} V_1(s') \right)}_{V_1(s^*)} + \gamma \max_{a \in \mathtt{A}} \sum_{s' \in \mathtt{S}} T^a_{s^*s'} e(s')
\end{aligned}$$

it follows that

$$V_2(s^*) < V_1(s^*) + e(s^*)$$

This inequality contradicts Equation (6).

■

Value iteration

In general, policy iteration converges in a relatively small number of steps. However, each step might require each of which requires the evaluation of a policy. Leveraging on the fixed point structure of the value function we can design an iterative algorithm which at each step reduces the maximal difference between $V_k(s)$ and $V_{k+1}(s)$ by at least a factor γ.

The recycling robot, one final time

We now follow step by step the described procedure to determine the optimal policy and estimate the optimal value function for the recycling robot (for the same parameter values). Starting off with the uniform policy, which we now call π_0 for the value function we obtained

$$\mathcal{V}^{\pi_0}(L) = -\frac{1}{15} \simeq -0.07 \quad \text{and} \quad \mathcal{V}^{\pi_0}(H) = -\frac{17}{15} \simeq= 1.13$$

We now apply the policy improvement step. In vector notation for each pair $(L, \cdot)$ we have

$$\begin{array}{lcl} (L,s) & \rightarrow & -1 + \frac{1}{2}\begin{pmatrix}\frac{1}{2} & \frac{1}{2}\end{pmatrix}^\top \begin{pmatrix} -0.07 \\ 1.13 \end{pmatrix} \simeq -0.74 \\ (L,w) & \rightarrow & 0 + \frac{1}{2}(1 \;\; 0)^\top \begin{pmatrix} -0.07 \\ 1.13 \end{pmatrix} \simeq -0.04 \\ (L,r) & \rightarrow & 0 + \frac{1}{2}(0 \;\; 1)^\top \begin{pmatrix} -0.07 \\ 1.13 \end{pmatrix} \simeq 0.57 \end{array}$$

from which we see that the action to choose, if the battery level is *low*, is *recharge*.

Similarly for each pair $(H, \cdot)$ we have

$$\begin{array}{lcl} (H,s) & \rightarrow & 2 + \frac{1}{2}\begin{pmatrix}\frac{1}{2} & \frac{1}{2}\end{pmatrix}^\top \begin{pmatrix} -0.07 \\ 1.13 \end{pmatrix} \simeq 2.30 \\ (H,w) & \rightarrow & 0 + \frac{1}{2}(0 \;\; 1)^\top \begin{pmatrix} -0.07 \\ 1.13 \end{pmatrix} \simeq 0.57 \\ (H,r) & \rightarrow & 0 + \frac{1}{2}(0 \;\; 1)^\top \begin{pmatrix} -0.07 \\ 1.13 \end{pmatrix} \simeq 0.57 \end{array}$$

from which we see that the action to choose, if the battery level is *high*, is *search*.

We now evaluate this policy double checking that for each state is an improvement with respect to π_0.

For the transition matrix $\mathbf{T}^\pi$ we pick the first row of $\mathbf{T}^r$ (because if the battery level is *low* the robot should *recharge*) and the last row of $\mathbf{T}^s$ (because if the battery level is *high* the robot should *search*) and find

$$\mathbf{T}^\pi = \frac{1}{2}\begin{pmatrix} 0 & 2 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{I} - \frac{1}{2}\mathbf{T}^\pi = \frac{1}{4}\begin{pmatrix} 4 & -1 \\ -1 & 3 \end{pmatrix}, \quad \text{and} \quad \left(\mathbf{I} - \frac{1}{2}\mathbf{T}^\pi\right)^{-1} = \frac{1}{5}\begin{pmatrix} 6 & 4 \\ 2 & 8 \end{pmatrix}$$

Since for the expected immediate reward for this deterministic policy we simply have

$$\mathbf{R}^{\pi} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$$

the evaluation gives

$$\begin{pmatrix} \mathcal{V}^{\pi}(L) \\ \mathcal{V}^{\pi}(H) \end{pmatrix} = \frac{1}{5} \begin{pmatrix} 6 & 4 \\ 2 & 8 \end{pmatrix} \begin{pmatrix} 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 1.6 \\ 3.2 \end{pmatrix}$$

This new policy is clearly better than the initial, and it is immediate to verify that there is no room for improvement.

3 Monte Carlo

4 Temporal Difference Learning

Let $Z_1, \dots, Z_n$ be *i.i.d* random variables with finite variance and μ as expected value. For each $n = 1, 2, \dots$ we recursively define the sequence $\mu_1, \cdots \mu_n$ as

$$\mu_n = \mu_{n-1} - \alpha_n(\mu_{n-1} - z_n)$$

with $\alpha_n > 0$ for all n. If $\alpha_n = 1/n$, μ_n is the empirical average or

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} z_i$$

From the law of large numbers we know that for $n \to \infty$ $\mathbb{E}[(\mu_n - \mu)^2] \to 0$. If $\alpha_n \neq 1/n$ under what conditions we expect that $\mathbb{E}[(\mu_n - \mu)^2] \to 0$ for $n \to \infty$? The question is not trivial since for $\alpha_n \neq 1/n$ μ_n is not the empirical average.

We start by considering

$$\mathbb{E}[(\mu_n - \mu)^2] = \mathbb{E}[(\mu_{n-1} - \alpha_n(\mu_{n-1} - z_n) - \mu)^2]$$

and noticing that the expectation is taken with respect to joint probability distribution of $Z_1, \dots, Z_n$ (since $Z_1, \dots, Z_n$ are independent the joint probability reduces to the product of the n marginal distributions). Using the linearity of the expectation of a sum of random variables we have

$$\mathbb{E}[(\mu_n - \mu)^2] = \mathbb{E}[(\mu_{n-1} - \mu)^2)] - 2\alpha_n\mathbb{E}[(\mu_{n-1} - z_n)(\mu_{n-1} - \mu)] + \alpha_n^2\mathbb{E}[(\mu_{n-1} - z_n)^2]$$

We now focus on the second and third term, since the first does not depend on Z_n. In particular,

$$\mathbb{E}[(\mu_{n-1} - z_n)(\mu_{n-1} - \mu)] = \mathbb{E}[(\mu_{n-1} - \mu)^2]$$

and

$$\mathbb{E}[(\mu_{n-1} - z_n)^2] \leq S^2$$

Taking the telescopic sum we finally obtain

$$\mathbb{E}[(\mu_{n+1} - \mu)^2] = \mathbb{E}[(\mu_1 - \mu)^2)] - 2\sum_{i=1}^{n}\alpha_n\mathbb{E}[(\mu_n - \mu)^2] + \sum_{i=1}^{n}\alpha_n^2\mathbb{E}[(\mu_n - z_n)^2] \geq 0$$

and thus

$$2\sum_{i=1}^{n}\alpha_n\mathbb{E}[(\mu_n - \mu)^2] \leq \mathbb{E}[(\mu_1 - \mu)^2)] + S^2\sum_{i=1}^{n}\alpha_n^2$$

Therefore, if $\sum_{i=1}^{n}\alpha_n^2$ converges and $\sum_{i=1}^{n}\alpha_n$ diverges, then $\mathbb{E}[(\mu_n - \mu)^2] \to 0$ which means that, with high probability, $\mu_n \to \mu$. We consider the simple one dimensional case. The result we describe lies at the basis of all the reinforcement learning algorithms we are going to discuss.

Observation 4.1. *Tower law*
We first remind an important result from probability theory. If X and Y are random variables,

$$\mathbb{E}_X[X] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]]$$

In words, the expected value of a random variable X can always be written as the expected value of the conditional expected value of X given Y. □

We aim at finding the root ξ of a function $f : (a, b) \to \mathbb{R}$, or $f(\xi) = 0$. For each point $x \in (a, b)$, we have access to a noisy sample of f at x, or

$$\phi(x) = f(x) + \epsilon$$

and we also assume that the noise ϵ is unbiased so that $\mathbb{E}[\epsilon] = 0$ or $\mathbb{E}[\phi(x)|x] = f(x)$ and the variance $Var(\phi(x)) = \mathbb{E}[\phi^2(x)] \leq G^2$ for some finite real number G.

We now study how to schedule the values of the strictly positive constants $\alpha_1, \alpha_2, \dots$ so that the sequence

$$x_{n+1} = x_n - \alpha_n \phi(x_n)$$

converges in probability to ξ.

Insert figure with $f(x) < 0$ for $x < \xi$ and $f(x) > 0$ for $x > \xi$. Comment about the meaning of the minus sign. If $\phi(x_n)$ and $f(x_n)$ have the same sign (very likely if x_n is far away from ξ) and α_n is not too large, x_{n+1} will lie closer to ξ with respect to x_n. Indeed, if $\xi < x_n$ then $\xi \leq x_{n+1} < x_n$ since $\alpha_n \phi(x_n) > 0$. Conversely, if $\xi > x_n$ then $\xi \geq x_{n+1} > x_n$ since $\alpha_n \phi(x_n) < 0$.

Under what assumptions on the α_n the expected value of the sequence converges to ξ?

We assume that f', the derivative of f, is not flat. This means that $\forall x \in (a, b)$

$$f'(x) \geq F > 0$$

We now study the convergence in expectation of the stochastic sequence $x_1, x_2, \dots$ to ξ. We start off with the quantity

$$\mathbb{E}[(x_{n+1} - \xi)^2]$$

where the expectation is taken with respect to all possible random sequences $x_1, \dots, x_{n+1}$, that is with respect to the joint probability distribution of the random variables $X_1, \dots, X_{n+1}$. We start by rewriting $\mathbb{E}[(x_{n+1} - \xi)^2]$ as

$$\mathbb{E}[(x_{n+1} - \xi)^2] = \mathbb{E}[(x_n - \xi)^2] - 2\alpha_n \mathbb{E}[(x_n - \xi)\phi(x_n)] + \alpha_n^2 \mathbb{E}[\phi^2(x_n)]$$

and we focus on the second and third term. Applying the *tower law* to the second term we obtain

$$\mathbb{E}[(x_n - \xi)\phi(x_n)] = \mathbb{E}[\mathbb{E}_{X_n}[(x_n - \xi)\phi(x_n)|x_n]] = \mathbb{E}[(x_n - \xi)f(x_n)]$$

where the outer expected value is taken with respect to the joint probability distribution of $X_1, \dots, X_n$.

We now multiply and divide by $(x_n - \xi)$ and obtain

$$\mathbb{E}\left[(x_n - \xi)^2 \frac{f(x_n)}{x_n - \xi}\right]$$

Since for all x_n

$$\frac{f(x_n)}{x_n - \xi} > F$$

we finally find

$$\mathbb{E}[(x_n - \xi)\phi(x_n)] \geq F\,\mathbb{E}\left[(x_n - \xi)^2\right]$$

For the hypothesis of finite variance

$$\mathbb{E}[\phi^2(x_n)] \geq G^2$$

If we now form the telescopic sum we obtain

$$\mathbb{E}[(x_{n+1} - \xi)^2] \leq \mathbb{E}[(x_0 - \xi)^2] - 2F \sum_{i=1}^{n} \alpha_i \mathbb{E}\left[(x_i - \xi)^2\right] + G^2 \sum_{i=1}^{n} \alpha_i^2$$

which, since $\mathbb{E}[(x_{n+1} - \xi)^2] \geq 0$, gives

$$\sum_{i=1}^{n} \alpha_i \mathbb{E}\left[(x_i - \xi)^2\right] \leq \frac{\mathbb{E}[(x_0 - \xi)^2]}{2F} + \frac{G^2}{2F} \sum_{i=1}^{n} \alpha_i^2$$

We now see that if $\sum_{i=1}^{n} \alpha_i^2$ converges to a finite value the right hand side is finite. If in addition $\sum_{i=1}^{n} \alpha_i$ diverges, for $i \to \infty$ it must be $\mathbb{E}\left[(x_i - \xi)^2\right] \to 0$.