Exercise 1 (Bellman equations and value iteration)**.** Consider an MDP with finite sets of states $\mathcal{S}$ and actions $\mathcal{A}$, transition probabilities $P(s' \mid s, a)$, rewards $r(s, a)$ with $|r(s, a)| \le R$ for any (s, a), and a discount factor $\gamma \in [0, 1)$. For a deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ and a state s, the value of π is

$$V^\pi(s) = \mathbb{E}\Big[\sum_{t \ge 0} \gamma^t \, r(S_t, A_t) \,\Big|\, S_0 = s \Big] \ , \tag{1}$$

where $A_t = \pi(S_t)$ and $S_{t+1} \sim P(\cdot \mid S_t, A_t)$. For $V \in \mathbb{R}^{\mathcal{S}}$, let $\|V\|_\infty = \max_s |V(s)|$, and define the Bellman operators T^π and T by

$$\begin{aligned} (T^\pi V)(s) &= r\big(s, \pi(s)\big) + \gamma \sum_{s'} P\big(s' \mid s, \pi(s)\big)\, V(s') \ , \\ (TV)(s) &= \max_a \Big(r(s, a) + \gamma \sum_{s'} P(s' \mid s, a)\, V(s') \Big) \ . \end{aligned} \tag{2}$$

1. Show that $\|V^\pi\|_\infty \le R/(1-\gamma)$ for any policy π.
2. Show that $V^\pi = T^\pi V^\pi$ (the Bellman equation of π).
   Hint: separate the term $t = 0$ of (1), and condition the rest on S_1: by the Markov property, given $S_1 = s'$, the process $(S_{t+1})_{t \ge 0}$ is the MDP started from s'.
3. Let P_π be the matrix with entries $P_\pi(s, s') = P(s' \mid s, \pi(s))$ and r_π the vector with entries $r(s, \pi(s))$. Show that $\|P_\pi V\|_\infty \le \|V\|_\infty$ for any V, deduce that $I - \gamma P_\pi$ is invertible, and that $V^\pi = (I - \gamma P_\pi)^{-1} r_\pi$.
4. Show that $|\max_a x_a - \max_a y_a| \le \max_a |x_a - y_a|$ for any real numbers $(x_a)_{a \in \mathcal{A}}$ and $(y_a)_{a \in \mathcal{A}}$.
5. Show that $\|TV - TW\|_\infty \le \gamma \|V - W\|_\infty$ for any V, W: T is a *contraction*. Show the same for T^π.
6. Deduce that T has at most one fixed point. Value iteration computes $V_0 = 0$ and $V_{k+1} = TV_k$; admitting that T has a fixed point $V^\star$, show that
$$\|V_k - V^\star\|_\infty \le \gamma^k \, \frac{R}{1-\gamma} \ .$$
7. Using $\log(1/\gamma) \ge 1 - \gamma$, deduce that $k \ge \frac{1}{1-\gamma} \log \frac{R}{\varepsilon(1-\gamma)}$ iterations are enough for an error at most $\varepsilon > 0$. Give this number for $\gamma = 0.99$, $R = 1$ and $\varepsilon = 0.01$.
8. Let $\pi^\star$ be greedy with respect to $V^\star$: $\pi^\star(s)$ maximizes $r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V^\star(s')$. Show that $T^{\pi^\star} V^\star = V^\star$, and deduce that $V^{\pi^\star} = V^\star$.
9. Show that $T^\pi V \le TV$ and that $V \le W$ implies $T^\pi V \le T^\pi W$, for any π and any V, W (componentwise inequalities). Deduce that $V^\pi \le V^\star$ for any policy π: the greedy policy $\pi^\star$ is optimal.
   Hint: show by induction that $(T^\pi)^k V^\star \le V^\star$, and let $k \to \infty$.

Exercise 2 (Q-learning by hand)**.** An MDP has three states $1, 2, 3$ and two actions, a (advance) and b (back). The transitions are deterministic:

- in state 1, action a leads to state 2 with reward 0, and action b leads to state 1 with reward 1;
- in state 2, action a leads to state 3 with reward 0, and action b leads to state 1 with reward 0;
- in state 3, both actions lead to state 1 with reward 4.

Q-learning keeps a table $Q(s, x)$, initialized at 0, and after each transition (s, x, r, s') applies

$$Q(s, x) \leftarrow Q(s, x) + \alpha\Big(r + \gamma \max_y Q(s', y) - Q(s, x) \Big) \ . \tag{3}$$

In this exercise, $\alpha = \gamma = 1/2$.

1. The agent observes the transitions $(1, a, 0, 2)$, $(2, a, 0, 3)$, $(3, a, 4, 1)$, $(1, a, 0, 2)$, $(2, a, 0, 3)$, in this order. Give the table Q after each transition.
2. Explain why $Q(1, a)$ is still 0, and how many passes through the cycle are needed before $Q(1, a)$ becomes positive.
3. For $\gamma = 1/2$, admit that the optimal values are $V^\star(1) = 2$, $V^\star(2) = 5/2$ and $V^\star(3) = 5$. Compute $Q^\star(1, a)$ and $Q^\star(1, b)$ with $Q^\star(s, x) = r(s, x) + \gamma V^\star(s'(s, x))$, where $s'(s, x)$ is the state reached from s with action x, and give the optimal action in state 1.
4. In a general MDP, with $Q^\star(s, x) = r(s, x) + \gamma \sum_{s'} P(s' \mid s, x) \max_y Q^\star(s', y)$, show that if $Q = Q^\star$, the target $r + \gamma \max_y Q(s', y)$ of (3) has expectation $Q^\star(s, x)$ given (s, x), where $s' \sim P(\cdot \mid s, x)$. Deduce that $Q^\star$ is a fixed point of the update (3) on average.