1. Robust Value Iteration

A robust MDP (RMDP) is an extension of the regular MDP to the case where the transition probability themselves are uncertain, and are only known to lie in some set, called the uncertainty set.

Formally, for each state $s \in S$ and action $a \in A$, the transition probability $p_{s,a}(s') \text{ B } p(s' \mid s,a)$ belongs to a *finite* set $\text{P}_{s,a}$, i.e., each element in $\text{P}_{s,a}$ is a probability measure over S. We let P denote the set of all MDPs such that their transitions probabilities $p_{s,a}(s')$ for each s and a belong to $\text{P}_{s,a}$. The goal is to optimize the *worst-case* expected discounted return

$$\max_{\pi} \min_{P \in \text{P}} \text{E}^{P}\left[\sum_{t=0}^{\infty} \gamma^{t} r\left(s_t, \pi(s_t)\right)\right].$$

a. Define a 'value function' for a fixed policy $V^{\pi}(s)$ and an optimal value function $V^{*}(s)$ for this setting.

b. The robust Bellman operator is defined as follows:

$$T^{r}V(s) = \max_{a} \min_{p_{s,a} \in \text{P}_{s,a}} \left\{ r(s,a) + \gamma \sum_{y} p_{s,a}(y) V(y) \right\}$$

Show that T^{r} is a $\gamma-$contraction. Propose a corresponding value iteration algorithm, and show that it converges.

It can be shown (not here) that the fixed point of T^{r} is indeed the optimal value function.

Solution

a. Similarly to regular MDPs, we define

$$V^{\pi}(s) = \min_{P \in \text{P}} \text{E}^{P}\left[\sum_{t=0}^{\infty} \gamma^{t} r\left(s_t, \pi(s_t)\right) \middle| s_0 = s\right]$$

$$V^{*}(s) = \max_{\pi} \min_{P \in \text{P}} \text{E}^{P}\left[\sum_{t=0}^{\infty} \gamma^{t} r\left(s_t, \pi(s_t)\right) \middle| s_0 = s\right]$$

b. We need to show that $\left|T^{r}V(s) - T^{r}\bar{V}(s)\right| \le \gamma \left|V(s) - \bar{V}(s)\right|$.

Choose some s. Consider first the case where $T^{r}V(s) \ge T^{r}\bar{V}(s)$. We have that

$$
\begin{aligned}
T^r V(s) - T^r \bar{V}(s) &= \max_a \min_{p_{s,a} \in P_{s,a}} \left\{ r(s,a) + \gamma \sum_y p_{s,a}(y) V(y) \right\} \\
&\quad - \max_a \min_{p_{s,a} \in P_{s,a}} \left\{ r(s,a) + \gamma \sum_y p_{s,a}(y) \bar{V}(y) \right\} \\
&\le \min_{p_{s,a^*} \in P_{s,a^*}} \left\{ r(s,a^*) + \gamma \sum_y p_{s,a^*}(y) V(y) \right\} \\
&\quad - \min_{p_{s,a^*} \in P_{s,a^*}} \left\{ r(s,a^*) + \gamma \sum_y p_{s,a^*}(y) \bar{V}(y) \right\} \\
&= \gamma \left(\min_{p_{s,a^*} \in P_{s,a^*}} \sum_y p_{s,a^*}(y) V(y) - \min_{p_{s,a^*} \in P_{s,a^*}} \sum_y p_{s,a^*}(y) \bar{V}(y) \right) \\
&\le \gamma \left(\sum_y p^*(y) V(y) - \sum_y p^*(y) \bar{V}(y) \right) \\
&\le \gamma \left| \sum_y p^*(y) V(y) - \sum_y p^*(y) \bar{V}(y) \right| \\
&\le \gamma \left\| V(y) - \bar{V}(y) \right\|_\infty
\end{aligned}
$$

Where $a^* = \arg\max_a \min_{p_{s,a} \in P_{s,a}} \left\{ r(s,a) + \gamma \sum_y p_{s,a}(y) V(y) \right\}$ and $p^* = \arg\min_{p_{s,a^*} \in P_{s,a^*}} \sum_y p_{s,a^*}(y) \bar{V}(y)$. A similar result holds for the case $T^r V(s) \le T^r \bar{V}(s)$.

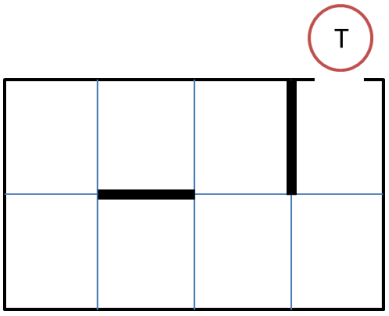
A value iteration algorithm proceeds as follows:

1. Initialize some arbitrary $V_0(s)$ for all $s \in S$.
2. Repeat: $V_{k+1} = T^r V_k$.

Convergence follows from the contraction property, and the fixed-point theorem.

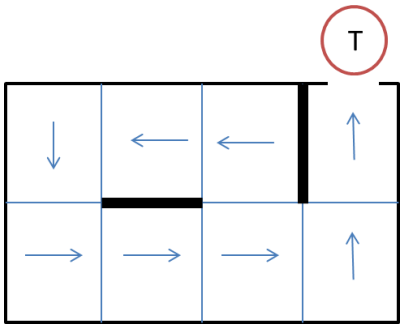
2. Policy iteration vs. value iteration

Consider the following deterministic grid-world domain. The reward for each nonterminal state is -1, and there is no discounting ($\gamma = 1$). The actions are moving north\south\east\west, and moving into a wall is not allowed.



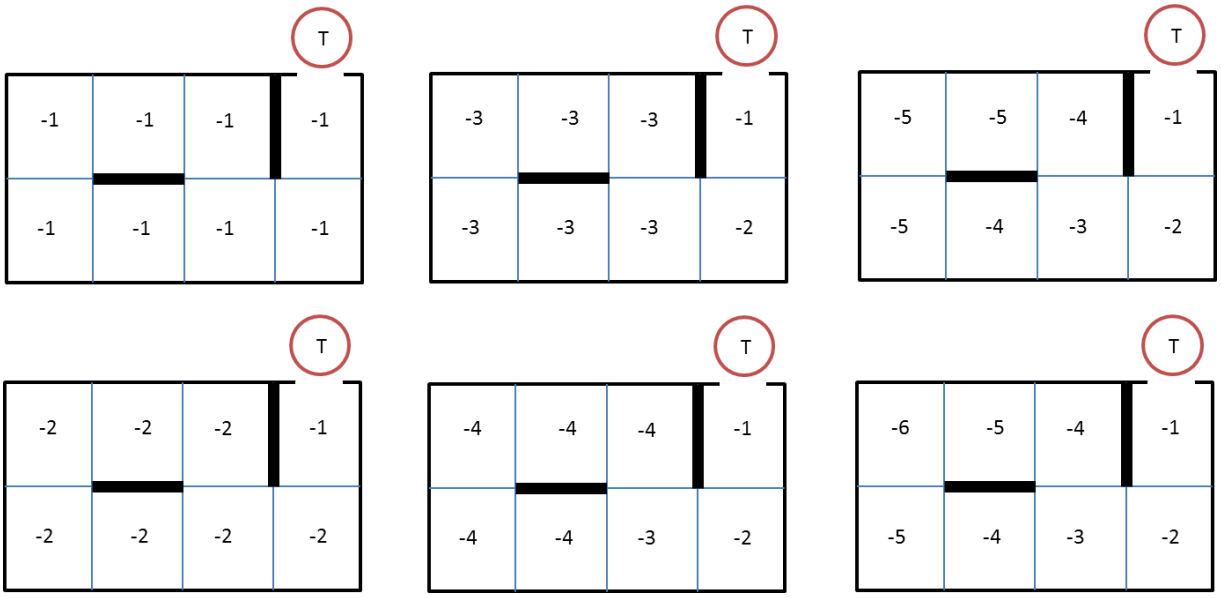
a. Compute value iteration for this domain, starting from $V_0(s) = 0$ for all states. After how many iterations is the greedy policy optimal?

b. Compute policy iteration for this domain, starting from the following policy



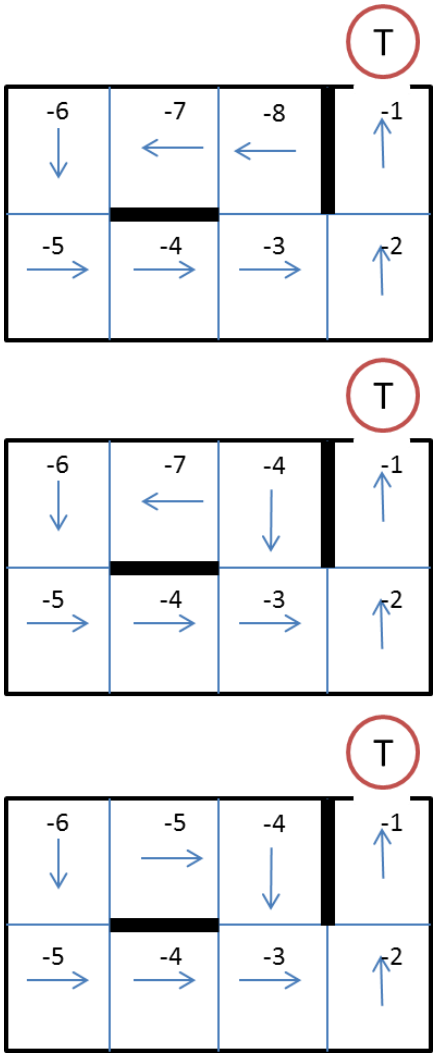
Solution:

a.



After the 5th iteration the greedy policy is optimal.

b.



3. Monotonic Improvement of The Greedy Policy

Let $\bar{\pi}$ be the greedy policy w.r.t a value function of some policy π, $v^{\pi} \in \mathbb{R}^{N}$,

$$\bar{\pi} \in \arg\max_{a \in A} r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) v^{\pi}(s').$$

a. Prove the following Lemma: $v^{\pi} \leq v^{\bar{\pi}}$, and equality holds if and only if π is the optimal policy.

Solution:

First, $T^{\bar{\pi}} v^{\pi} = T v^{\pi}$ due to the construction of the greedy policy,

$$T^{\bar{\pi}} v^{\pi} = r(s, \bar{\pi}(s)) + \gamma \sum_{s' \in S} p(s' \mid s, \bar{\pi}(s)) v^{\pi}(s')$$

$$= \max_{a \in A} r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) v^{\pi}(s') = T v^{\pi}.$$

We thus have that,

$$v^{\pi} = T^{\pi}v^{\pi} \leq Tv^{\pi} = T^{\overline{\pi}}v^{\pi}$$

Since $T^{\overline{\pi}}$ is monotone operator we can apply this relation iteratively and get,

$$v^{\pi} \leq T^{\overline{\pi}}v^{\pi} \leq (T^{\overline{\pi}})^2 v^{\pi} \leq \cdots \leq \lim_{n\to\infty} (T^{\overline{\pi}})^n \, v^{\pi} = v^{\overline{\pi}}.$$

The equality part is left as an exercise.