Theory Question 1 - Linear convergence of Policy Iteration

Problem 1. (a) (solution from Peng)

Proof.

$$\begin{aligned}
V^{\pi_t}(s) &= \sum_a \pi(a \mid s) \sum_{s',r} p\left(s', r \mid s, a\right)\left[r+\gamma V^{\pi_t}\left(s'\right)\right], \quad \text{for all } s \in \mathcal{S} \ldots\ldots \text{(Bellman Equation)} \\
&= \sum_{s',r} p\left(s', r \mid s, a\right)\left[r+\gamma V^{\pi_t}\left(s'\right)\right] \ldots\ldots (\ \pi(a \mid s)=1 \text{ because of deterministic policy}) \\
&= \sum_{s',r} p\left(s', r \mid s, a\right) r+\sum_{s',r} p\left(s', r \mid s, a\right) \gamma V^{\pi_t}\left(s'\right) \\
&<= \sum_{s',r} p\left(s', r \mid s, a\right) \max_a r+\sum_{s',r} p\left(s', r \mid s, a\right) \max_a \gamma V^{\pi_t}\left(s'\right) \\
&\ldots\ldots (r <= \max_a r, V^{\pi_t}(s') <= \max_a V^{\pi_t}\left(s'\right), \quad \text{for all } a \in \mathcal{A}) \\
&= \max_a r+\sum_{s'} p\left(s' \mid s, a\right) \max_a \gamma V^{\pi_t}\left(s'\right) \ldots\ldots (\sum_{s',r} p\left(s', r \mid s, a\right)=1) \\
&= \max_a r+\max_a \gamma \sum_{s'} p\left(s' \mid s, a\right) V^{\pi_t}\left(s'\right) \\
&= \max_a \left[r(s, a)+\gamma \mathbb{E}_{s' \mid s, a}\left[V^\pi\left(s'\right)\right]\right] \ldots\ldots \text{(Definition of Expectation)} \\
&= B V^{\pi_t}(s)
\end{aligned}$$

■

1(a) (solution from Paul)

$$\begin{aligned}
B V^{\pi_t}(x) &= \max_a \left[r(x, a)+\gamma \mathbb{E}_{x' \mid x, a}\left[V^{\pi_t}(x')\right]\right] \\
&\geq r(x, \pi_t(x))+\gamma \mathbb{E}_{x' \mid x, \pi_t(x)}\left[V^{\pi_t}(x')\right] \\
&= V^{\pi_t}(x)
\end{aligned}$$

First, let's define some notation. In this context, we have a state space $\mathcal{S}$ and an action space $\mathcal{A}$. We also have a discount factor $\gamma \in [0,1)$ and a reward function $r: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. We're interested in finding a policy $\pi: \mathcal{S} \rightarrow \mathcal{A}$ that maximizes the expected discounted sum of rewards starting from each state.

Now, let's focus on the Bellman operator. This is defined as:

$$B V^\pi(x)=\max_a \left[r(x, a)+\gamma \mathbb{E}_{x' \mid x, a}\left[V^\pi(x')\right]\right]$$

Here, $V^\pi(x)$ is the value function for policy π at state x, which represents the expected discounted sum of rewards starting from x and following policy π. The Bellman operator essentially says that the value function for policy π at state x should be equal to the maximum expected sum of rewards that can be obtained by taking any action a in state x and then following policy π from the resulting state x'.

Now, let's move on to the proof. We want to show that the value function for policy π_{t+1} is greater than or equal to the value function for policy π_t at all states. In other words:

$$V^{\pi_{t+1}}(x) \geq V^{\pi_t}(x) \forall x \in \mathcal{S}$$

We can use the Bellman operator to help us with this proof. First, note that for any policy π, we have:

$$BV^{\pi}(x) \geq V^{\pi}(x) \forall x \in \mathcal{S}$$

This is because the Bellman operator takes the maximum over all actions, so it must be greater than or equal to the expected sum of rewards for any single action.

Problem 1. (b) (solution from Peng)

Proof.

$$\begin{aligned} V^{\pi_{t+1}}(s) &= \max_a \left[r(s,a) + \gamma \mathbb{E}_{s'|s,a}\left[V^{\pi_{t+1}}\left(s'\right)\right]\right] \\ &>= \max_a \left[r(s,a) + \gamma \mathbb{E}_{s'|s,a}\left[V^{\pi_t}\left(s'\right)\right]\right] \ldots\ldots (V^{\pi_{t+1}} >= V^{\pi_t} \qquad \text{due to greedy policy }) \\ &= BV^{\pi_t}(s) \end{aligned}$$

■

1(b) (solution from Paul)

Algorithm 1 Fixed Point Iteration

```
Initialize $U_0 = V^{\pi_t}$.
for $i = 1$ to $T$ do
    $U_i = r(x, \pi_{t+1}(x)) + \mathbb{E}_{x'|x,\pi_{t+1}(x)}[U_{i-1}]$
end for
```

From fixed point iteration

$$U_1(x) = BV^{\pi_t} \geq V^{\pi_t}(x) = U_0(x)$$

By induction, U_i is monotonically increasing:

$$\begin{aligned} U_i(x) &= r(x, \pi_{t+1}(x)) + E_{x'|x,\pi_{t+1}(x)}[U_{i-1}(x')] \\ &\geq r(x, \pi_{t+1}(x)) + E_{x'|x,\pi_{t+1}(x)}[U_{i-2}(x')] \\ &= U_{i-1}(x) \end{aligned}$$

Therefore

$$V^{\pi_{t+1}} = U_{i\to\infty} \geq U_1 = BV^{\pi_t}$$

The proof is showing that the value function $V^{\pi_{t+1}}$ obtained from the policy π_{t+1} using the fixed point iteration algorithm is greater than or equal to the value function BV^{π_t} obtained from the policy π_t using the Bellman operator.

First, the algorithm initializes U_0 to be the value function of the policy π_t, denoted as V^{π_t}. Then, the algorithm iteratively updates U_i by taking the maximum reward obtained by taking action $\pi_{t+1}(x)$ and the expected value of U_{i-1} obtained by transitioning to the next state x', which is the Bellman update for the value function.

Next, the proof uses induction to show that U_i is monotonically increasing. The base case is given by $U_1(x) = BV^{\pi_t}(x) \geq V^{\pi_t}(x) = U_0(x)$, which is true because the Bellman operator always increases the value function.

Then, assuming that $U_{i-1}(x) \leq U_i(x)$ for all x, the proof shows that $U_i(x) \leq U_{i+1}(x)$ for all x. This is true because the Bellman update ensures that $U_i(x)$ can only increase as the algorithm iterates.

Finally, the proof concludes that $V^{\pi_{t+1}}$, the value function obtained from the fixed point iteration algorithm, is greater than or equal to U_1, which is greater than or equal to BV^{π_t}.

Therefore, $V^{\pi_{t+1}} \geq BV^{\pi_t}$, which means that the value function obtained from the new policy is at least as good as the value function obtained from the old policy. This is a key result in policy iteration, as it ensures that each iteration of the algorithm improves the policy until convergence.

Problem 1. (c) (solution from Peng)

Proof. (a) Prove contract mapping: $|BV_1 - BV_2|_\infty \leq \gamma \|V_1 - V_2\|_\infty$

$$\begin{aligned}
|BV_1(s) - BV_2(s)| = &| \max_a \sum_{s'} p\left(s' \mid s, a\right)\left(r\left(s, a\right) + \gamma V_1\left(s'\right)\right) \\
&- \max_a \sum_{s'} p\left(s' \mid s, a\right)\left(r\left(s, a\right) + \gamma V_2\left(s'\right)\right) | \\
\leq &\max_a \sum_{s'} p\left(s' \mid s, a\right)\left|r\left(s, a\right) + \gamma V_1\left(s'\right) - r\left(s, a\right) - \gamma V_2\left(s'\right)\right| \\
\leq &\gamma \max_a \sum_{s'} p\left(s' \mid s, a\right)\left|V_1\left(s'\right) - V_2\left(s'\right)\right| \\
\leq &\gamma \ \max_{s'} |V_1(s') - V_2(s')| \\
= &\gamma \|V_1 - V_2\|_\infty
\end{aligned}$$

Since the above inequality hold for each s, thus, from $\max_s |BV_1(s) - BV_2(s)| \leq \gamma \|U(s) - V(s)\|_\infty$ we can get $|BV_1 - BV_2|_\infty \leq \gamma \|V_1 - V_2\|_\infty$.

(b) Prove $V^* = BV^*$
$BV^* = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s'|s,a}\left[V^{\pi^*}\left(s'\right)\right]\right] = V^*$

(c) Prove Linear convergence:

$$\begin{aligned}
\frac{\|V^{\pi_{t+1}} - V^*\|_\infty}{\|V^{\pi_t} - V^*\|_\infty} &<= \frac{\|BV^{\pi_t} - V^*\|_\infty}{\|V^{\pi_t} - V^*\|_\infty} \ldots\ldots (V^{\pi_{t+1}} >= BV^{\pi_t} \Rightarrow \|V^{\pi_{t+1}} - V^*\| <= \|BV^{\pi_t} - V^*\|) \\
&= \frac{\|BV^{\pi_t} - BV^*\|_\infty}{\|V^{\pi_t} - V^*\|_\infty} \ldots\ldots (V^* = BV^*) \\
&<= \gamma \frac{\|V^{\pi_t} - V^*\|_\infty}{\|V^{\pi_t} - V^*\|_\infty} \ldots\ldots (\text{ Contract mapping}) \\
&= \gamma \ldots\ldots (\text{ Definition of Linear convergence})
\end{aligned}$$

Thus, $\left\|V^{\pi_t} - V^*\right\|_\infty <= \gamma \left\|V^{\pi_{t+1}} - V^*\right\|_\infty <= ... <= \gamma^t \left\|V^{\pi_0} - V^*\right\|_\infty = 0$ In the end, we get $\lim_{t\to\infty} V_t = V^*$

■

1(c) (solution from Paul)
From above, and $V^*(x) \geq V^\pi(x)$ for any policy π:

$$||V^{\pi_{t+1}}(x) - V^*||_\infty \leq ||BV^{\pi_t}(x) - V^*||_\infty$$

Because $BV^* = V^*$ and the contractive property of B we proceed by induction:

$$\begin{aligned} ||V^{\pi_{t+1}}(x) - V^*||_\infty &\leq ||BV^{\pi_t}(x) - BV^*||_\infty \\ &\leq ||B(V^{\pi_t}(x) - V^*)||_\infty \\ &\leq \gamma||V^{\pi_t}(x) - V^*||_\infty \\ &\leq \gamma^t||V^{\pi_1}(x) - V^*||_\infty \end{aligned}$$

This expression decays exponentially as t goes to ∞.

The proof is related to the convergence rate of the policy iteration algorithm, and it shows that the difference between the value function of the policy at time $t+1$, $V^{\pi_{t+1}}$, and the optimal value function, V^*, decreases exponentially with the number of iterations t.

The proof begins by using the fact that $V^x \geq V^\pi(x)$ for any policy π. This means that the optimal value function V^* is always greater than or equal to the value function of any policy π. Using this inequality, the proof then applies the infinity-norm ($|\cdot|_\infty$) to both sides of the Bellman equation for $V^{\pi_{t+1}}$, which gives:

$$||V^{\pi_{t+1}}(x) - V^*||_\infty \leq ||BV^{\pi_t}(x) - V^*||_\infty$$

where $BV^{\pi_t}(x)$ is the Bellman backup of V^{π_t}, which is computed by applying the Bellman operator B to V^{π_t}.

The proof then proceeds by using the contractive property of B. This property means that the distance between the Bellman backups of two value functions is always less than or equal to the distance between the value functions themselves. Specifically, we have:

$$||BV^{\pi_t}(x) - BV^*||_\infty \leq ||V^{\pi_t}(x) - V^*||_\infty$$

Using these inequalities and the induction hypothesis, the proof then shows that:

$$\begin{aligned} ||V^{\pi_{t+1}}(x) - V^*||_\infty &\leq ||BV^{\pi_t}(x) - BV^*||_\infty \\ &\leq ||B(V^{\pi_t}(x) - V^*)||_\infty \\ &\leq \gamma||V^{\pi_t}(x) - V^*||_\infty \\ &\leq \gamma^t||V^{\pi_1}(x) - V^*||_\infty \end{aligned}$$

where γ is the discount factor and $V^{\pi_1}(x)$ is the value function of an arbitrary policy π_1. This expression decays exponentially as t goes to infinity, which means that the difference between $V^{\pi_{t+1}}$ and V^* approaches zero very quickly as the number of iterations increases.

Theory Question 2 - Policy Iteration vs Value Iteration

The aim of both these iterative processes is to get to an optimal policy for a given Markov Decision Process (MDP). The policy iteration uses the Bellman expectation equation and a greedy policy improvement to iteratively improve its policy. It reaches the optimal policy only once the estimates of state-value function converge to $\lim_{k\to\infty} v_k$ or at least until the difference between the old and new state-value function estimates becomes very small ($\Delta < \theta$). This can mean several iterations where the optimal policy has already been achieved, but the estimates of the state-value function (output of the Bellman expectation equation) are still converging.

The value iteration on the other hand does not further iterate once the optimal policy has been achieved. It works by turning the Bellman optimality equation into an update rule. The iterations here are truncated to where the Bellman optimality equation for all state action pairs is met and we have an optimal policy.

The main advantage of the value iteration over the policy iteration is that it can converge faster if the state space is large. In problems with more limited state spaces (like the Gridworld), the policy iteration converges with fewer steps, as calculating a good approximation of the action value is easy. Another advantage of the value iteration is that it only relies on the Bellman optimality equation rather than requiring both the Bellman expectation equation and a greedy policy improvement as the policy iteration does. This alteration between the policy evaluation and the policy improvement requires an additional loop, but tends to make the training process of the policy iteration more stable (i.e. less oscillation). Both these algorithms can be applied when we know the MDP to finite environments, where actions, states and rewards are finite too. Common applications are certain boardgames, video games, as well as control and navigation of robots that can be modelled as Markov Decision Processes.