CSCE 790-003, Spring 2022
Assignment 1

Username: [zhongs@email.sc.edu]
Name: [Shan Zhong]

By turning in this assignment, I agree by the honor code of USC Columbia.

**Submission.** You need to submit the following files to Blackboard:

- A pdf file named as assignment1_⟨*username*⟩.pdf, where you replace ⟨*username*⟩ with your email username. This pdf file contains your anwsers to the written problems, including problems 1, 2, 3, and 4(c). You can either edit the assignment1.tex file to fill in your answers and submit the pdf file generated from the edited tex file, or scan and submit your hand-written answers.
- A zip file named as assignment1_⟨*username*⟩.zip, where you replace ⟨*username*⟩ with your email username. This zip file contains a single file specified in problem 4.

# 1 Bellman Optimality Operator Is a Contraction [25pt]

Recall the definition of the Bellman optimality operator $T^* : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$, where $|S|$ is the number of states in state space $S$:

$$(T^*V)(s) := \max_a R(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s')$$

for each $s \in S$. Suppose $\gamma < 1$. Prove that $T^*$ is a $\gamma$-contraction in the infinity norm $\|\cdot\|_\infty$, i.e. for any $V, V' \in \mathbb{R}^{|S|}$

$$\|T^*V' - T^*V\|_\infty \leq \gamma \|V' - V\|_\infty$$

where $\|V\|_\infty := \max_{s \in S} |V(s)|$.

Hint: Since it is the infinity norm, we need to show the inequality holds for each $s \in S$, i.e. $T^*V'(s) - T^*V(s) \leq \gamma\|V' - V\|_\infty$. For any $s \in S$. Let

$$a_s^* = \arg\max_a R(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s')$$
$$a_s'^* = \arg\max_a R(s,a) + \gamma \sum_{s'} P(s'|s,a)V'(s')$$

be the actions selected by the $T^*$ operator for $V$ and $V'$, respectively.

If $(T^*V')(s) \geq (T^*V)(s)$, try to show that the following holds:

$$\begin{aligned}
|(T^*V')(s) - (T^*V)(s)| =& (T^*V')(s) - (T^*V)(s) \qquad \text{(Since we assumed } (T^*V')(s) \geq (T^*V)(s)\text{)} \\
=& R(s, a_s'^*) + \gamma \sum_{s'} P(s'|s, a_s'^*)V'(s') - (R(s, a_s^*) + \gamma \sum_{s'} P(s'|s, a_s^*)V(s')) \\
\leq& \gamma\|V' - V\|_\infty
\end{aligned}$$

What if $(T^*V')(s) < (T^*V)(s)$ instead?

First we have:

$$
\begin{aligned}
(T^*V)(s) :=& \max_{a\in A}\big(R(s,a)\big) + \gamma\sum_{s'\in S} P(s'|s,a)V(s') \\
(T^*V')(s) - (T^*V)(s) =& \big[\max_{a\in A}\big(R(s,a)\big) + \gamma\sum_{s'\in S} P(s'|s,a)V(s')\big] - \\
& \big[\max_{a\in A}\big(R(s,a)\big) - \gamma\sum_{s'\in S} P(s'|s,a)V'(s')\big] \\
=& \gamma\sum_{s'\in S} P(s'|s,a)V(s') - \gamma\sum_{s'\in S} P(s'|s,a)V'(s') \\
=& \gamma\sum_{s'\in S} P(s'|s,a)\big(V(s') - V'(s')\big)
\end{aligned}
$$

Then:

$$
\begin{aligned}
\forall s\in S, P(s'|s,a)\geq 0, \big(V'(s)-V(s)\big) \leq& \max_{s\in S}\big|V'(s)-V(s)\big| \\
\forall s\in S, P(s'|s,a)\big(V'(s)-V(s)\big) \leq& P(s'|s,a)\max_{s\in S}\big|V'(s)-V(s)\big| \\
\sum_{s\in S}\big[P(s'|s,a)\big(V'(s)-V(s)\big)\big] \leq& \sum_{s\in S}\big[P(s'|s,a)\max_{s\in S}\big|V'(s)-V(s)\big|\big]
\end{aligned}
$$

Also we have $\sum_{s\in S} P(s'|s,a) = 1, 0 < \gamma < 1$:

$$
\begin{aligned}
(T^*V')(s) - (T^*V)(s) =& \gamma\sum_{s'\in S} P(s'|s,a)\big(V(s') - V'(s')\big) \\
\leq& \big(\sum_{s\in S} P(s'|s,a)\big)\gamma\max_{s\in S}\big|(T^*V')(s) - (T^*V)(s)\big| \\
=& \gamma\max_{s\in S}\big|V'(s) - V(s)\big| \\
=& \gamma\|V' - V\|_\infty
\end{aligned}
$$

As choose of $s$ is arbitrary,

$$
\|T^*V' - T^*V\|_\infty = \max_{s\in S}\big|(T^*V')(s) - (T^*V)(s)\big| \leq \gamma\|V' - V\|_\infty
$$

## 2 Performance Difference Lemma and Policy Improvement Theorem [25pt]

(a) Let's first prove a useful lemma, called Performance Difference Lemma: For any policies $\pi, \pi'$, and any state $s \in S$,

$$V^{\pi'}(s) - V^{\pi}(s) = \mathbb{E}_{s' \sim d_s^{\pi'}}[A^{\pi}(s', \pi')]$$

where $d_s^{\pi'}$ is the discounted occupancy induced by policy $\pi'$ by starting from state $s$, i.e.

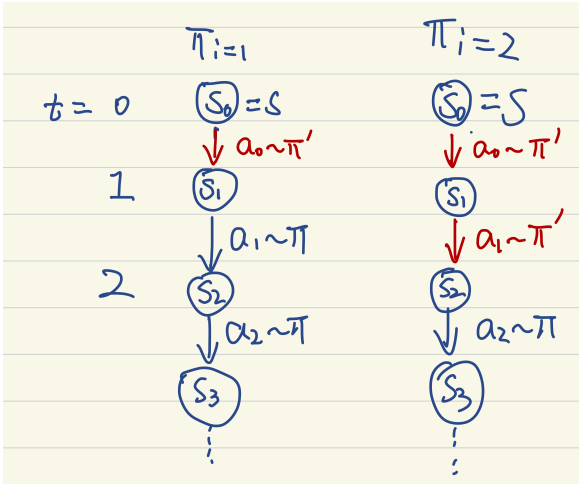
$$d_s^{\pi'}(s') = \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s' | s_0 = s, \pi')$$

and $A^{\pi}(s', \pi')$ is defined as

$$A^{\pi}(s', \pi') := \mathbb{E}_{a' \sim \pi'(s')}[Q^{\pi}(s', a')] - V^{\pi}(s').$$

$A^{\pi}$ is referred to as the advantage function of $\pi$.

Hint: To prove this lemma, consider a sequence of (possibly non-stationary) policies $\{\pi_i\}_{i \geq 0}$, where $\pi_0 = \pi$, $\pi_\infty = \pi'$. For any intermediate $i$, $\pi_i$ is the non-stationary policy that follows $\pi'$ for the first $i$ time steps (i.e. time steps $t$ such that $0 \geq t < i$) and then switches to $\pi$ for time steps $t \geq i$, as shown in the figure below comparing of $\pi_{i=1}$ and $\pi_{i=2}$.




Now we can rewrite the LHS of the statement as:

$$V^{\pi'}(s) - V^{\pi}(s) = \sum_{i=0}^{\infty} (V^{\pi_{i+1}}(s) - V^{\pi_i}(s))$$

For each term $(V^{\pi_{i+1}}(s) - V^{\pi_i}(s))$ on the RHS, observe that $\pi_{i+1}$ and $\pi_i$ are both identical to $\pi'$ for the first $i$ time steps, which induces the same state distribution at time step $i$, $\Pr(s_i | s_0 = s, \pi')$. They are also both identical to $\pi$ starting from state $s_{i+1}$ at time step $i+1$; so conditioned on $(s_i = s, a_i = a)$, the expected total reward for the remainder of the trajectory is $\gamma^i Q^{\pi}(s, a)$. Therefore, we have

$$V^{\pi_{i+1}}(s) - V^{\pi_i}(s) = \gamma^i \sum_{s'} \Pr(s_i = s' | s_0 = s, \pi')(\mathbb{E}_{a_i \sim \pi'(s')}[Q^{\pi}(s', a_i)] - \mathbb{E}_{a_i \sim \pi(s')}[Q^{\pi}(s', a_i)]).$$

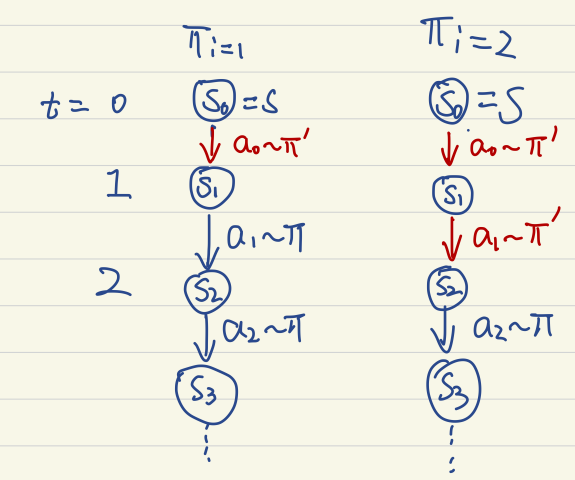
because the difference between $V^{\pi_{i+1}} - V^{\pi_i}$ only starts from $s_i$ at time step $i$, where $\pi_{i+1}$ and $\pi_i$ choose action $a_i$ according $\pi'$ and $\pi$, respectively.

As we have:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, Q^\pi(s,a) = \mathbb{E}[G_t | s_t = s, a_t = a, \pi]$$

$$\begin{aligned}
V^\pi(s) &= \mathbb{E}[G_t | s_t = s, \pi] \\
&= \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, \pi] = \sum_{k=0}^{\infty} \gamma^k \, \mathbb{E}[r_{t+k+1} | s_t = s, \pi] \\
&= \sum_{k=0}^{\infty} \gamma^k \sum_{s_t \in S} \sum_{a_t \in A} \mathbb{E}[r_{t+k+1} | s_t = s, a_t = a, \pi] \Pr(a_t = a | s_t = s, \pi) \Pr(s_t = s | s_0, \pi) \\
&= \sum_{k=0}^{\infty} \gamma^k \sum_{s_t \in S} (\mathbb{E}_{a_t \sim \pi(s_t)}[Q^\pi(s,a)]) \Pr(s_t = s | s_0, \pi)
\end{aligned}$$

Consider a sequence of (possibly non-stationary) policies $\{\pi_i\}_{i \geq 0}$, where $\pi_0 = \pi$, $\pi_\infty = \pi'$. For any intermediate $i$, $\pi_i$ is the non-stationary policy that follows $\pi'$ for the first $i$ time steps (i.e. time steps $t$ such that $0 \geq t < i$) and then switches to $\pi$ for time steps $t \geq i$. Then we have for any policies $\pi, \pi'$, and any state $s \in S$:




For each term $(V^{\pi_{i+1}}(s) - V^{\pi_i}(s))$, $\pi_{i+1}$ and $\pi_i$ are both identical to $\pi'$ for the first $i$ time steps, which induces the same state distribution at time step $i$, $\Pr(s_i | s_0 = s, \pi')$. They are also both identical to $\pi$ starting from state $s_{i+1}$ at time step $i+1$; so conditioned on $(s_i = s, a_i = a)$,

the expected total reward for the remainder of the trajectory is $\gamma^i Q^\pi(s,a)$.

$$
\begin{aligned}
V^{\pi'}(s) - V^{\pi}(s) =& V^{\pi\infty}(s) - V^{\pi_0}(s) = \sum_{i=0}^{\infty}(V^{\pi_{i+1}}(s) - V^{\pi_i}(s)) \\
&= \sum_{i=0}^{\infty}\Big(\sum_{t=0}^{\infty}\gamma^t \sum_{s_t\in S}(\mathbb{E}_{a_t\sim\pi_{i+1}(s_t)}[Q^\pi(s',a)])\Pr(s_t=s'|s_0,\pi_{i+1}) - \\
&\qquad \sum_{t=0}^{\infty}\gamma^t \sum_{s_t\in S}(\mathbb{E}_{a_t\sim\pi_{i}(s_t)}[Q^\pi(s',a)])\Pr(s_t=s'|s_0,\pi_{i})\Big) \\
&= \sum_{i=0}^{\infty}\Big(\sum_{t=0}^{i}\gamma^t \sum_{s_t\in S}(\mathbb{E}_{a_t\sim\pi'(s_t)}[Q^\pi(s',a)])\Pr(s_t=s'|s_0,\pi') + \\
&\qquad \sum_{t=i+1}^{\infty}\gamma^t \sum_{s_t\in S}(\mathbb{E}_{a_t\sim\pi(s_t)}[Q^\pi(s',a)])\Pr(s_t=s'|s_0,\pi) - \\
&\qquad \sum_{t=0}^{i-1}\gamma^t \sum_{s_t\in S}(\mathbb{E}_{a_t\sim\pi'(s_t)}[Q^\pi(s',a)])\Pr(s_t=s'|s_0,\pi')\Big) - \\
&\qquad \sum_{t=i}^{\infty}\gamma^t \sum_{s_t\in S}(\mathbb{E}_{a_t\sim\pi(s_t)}[Q^\pi(s',a)])\Pr(s_t=s'|s_0,\pi)\Big) \\
&= \sum_{i=0}^{\infty}\Big(\gamma^i \sum_{s_i\in S}(\mathbb{E}_{a_i\sim\pi'(s_i)}[Q^\pi(s',a)])\Pr(s_i=s'|s_0,\pi') - \\
&\qquad \gamma^i \sum_{s_i\in S}(\mathbb{E}_{a_i\sim\pi(s_i)}[Q^\pi(s',a)])\Pr(s_i=s'|s_0,\pi')\Big) \\
&= \sum_{i=0}^{\infty}\gamma^i \sum_{s_i\in S}\Pr(s_i=s'|s_0,\pi')\big(\mathbb{E}_{a_i\sim\pi'(s_i)}[Q^\pi(s',a)] - \mathbb{E}_{a_i\sim\pi(s_i)}[Q^\pi(s',a)]\big) \\
&= \sum_{s_i\in S}\sum_{i=0}^{\infty}\gamma^i \Pr(s_i=s'|s_0,\pi')\big(\mathbb{E}_{a_i\sim\pi'(s_i)}[Q^\pi(s',a)] - V^\pi(s')\big)
\end{aligned}
$$

Also as:

$$
\begin{aligned}
d_s^{\pi'}(s') &= \sum_{t=0}^{\infty}\gamma^t \Pr(s_t=s'|s_0=s,\pi') \\
A^\pi(s',\pi') &:= \mathbb{E}_{a'\sim\pi'(s')}[Q^\pi(s',a')] - V^\pi(s')
\end{aligned}
$$

$$
\begin{aligned}
V^{\pi'}(s) - V^{\pi}(s) &= \sum_{s_i\in S} d_s^{\pi'}(s')\big(A^\pi(s',\pi')\big) \\
&= \mathbb{E}_{s'\sim d_s^{\pi'}}[A^\pi(s',\pi')]
\end{aligned}
$$

(b) Using the performance difference lemma, prove the policy improvement theorem, i.e. prove $V^{\pi_{k+1}}(s) \geq V^{\pi_k}(s)$, where $\pi_{k+1}$ and $\pi_k$ are consecutive policies in policy iteration.

$$\begin{aligned} V^{\pi_{k+1}}(s) - V^{\pi_k}(s) &= \gamma^i \sum_{s'} \Pr(s_i = s' | s_0 = s, \pi')(\mathbb{E}_{a_i \sim \pi'(s')}[Q^{\pi}(s', a_i)] - V^{\pi}(s')) \\ &= d_s^{\pi'}(s')[A^{\pi}(s', \pi')] \end{aligned}$$

As $A^{\pi}(s', \pi')$ is defined as the advantage function that alwasy $\geq 0$, $d_s^{\pi'}(s') \geq 0$ as well, we have $V^{\pi_{k+1}}(s) \geq V^{\pi_k}(s)$

## 3 Bounding the Performance of Greedy Policy [25pt]

(a) Consider the sequence of iterates, $V_0, V_1, ..., V_k, ...$, in value iteration, where $V_k = T^* V_{k-1}$. Suppose $\gamma < 1$. Since $T^*$ is a contraction, $\|V_k - V_{k-1}\|_\infty$ decreases as $k$ increases, $\|V_k - V_{k-1}\|_\infty = \|T^* V_{k-1} - T^* V_{k-2}\|_\infty \leq \gamma \|V_{k-1} - V_{k-2}\|_\infty$. Suppose $\|V_k - V_{k-1}\|_\infty \leq \epsilon$ for some large enough $k$. Show that $\|V^* - V_k\|_\infty \leq \frac{\epsilon}{1-\gamma}$. In words, if the iterates are close, then they are close to the optimal state-value $V^*$.

Hint: Pick some integer $n \geq 1$, we have

$$\begin{aligned}\|V^* - V_k\|_\infty =& \|V^* - V_{k+n} + V_{k+n} - V_{k+n-1} + ... + V_{k+1} - V_k\|_\infty \\ \leq& \|V^* - V_{k+n}\|_\infty + \sum_{i=1}^{n} \|V_{k+i} - V_{k+i-1}\|_\infty \qquad \text{(triangle inequality)}\end{aligned}$$

We have:

$$\begin{aligned}\|V^* - V_k\|_\infty =& \|V^* - V_{k+n} + V_{k+n} - V_{k+n-1} + ... + V_{k+1} - V_k\|_\infty \\ \leq& \|V^* - V_{k+n}\|_\infty + \sum_{i=1}^{n} \|V_{k+i} - V_{k+i-1}\|_\infty \qquad \text{(triangle inequality)}\end{aligned}$$

As $\|V_k - V_{k-1}\|_\infty \leq \epsilon$ for some large enough $k$, also we have $\|V_k - V_{k-1}\|_\infty = \|T^* V_{k-1} - T^* V_{k-2}\|_\infty \leq \gamma \|V_{k-1} - V_{k-2}\|_\infty$

$$\begin{aligned}\|V_{k+i} - V_{k+i-1}\|_\infty &\leq \gamma \|V_{k+i-1} - V_{k+i-2}\|_\infty \leq \gamma^2 \|V_{k+i-2} - V_{k+i-3}\|_\infty \cdots \\ &\leq \gamma^i \|V_{k+i-i} - V_{k+i-i-1}\|_\infty = \gamma^i \|V_k - V_{k-1}\|_\infty \leq \gamma^i \epsilon\end{aligned}$$

$$\begin{aligned}\|V^* - V_k\|_\infty &\leq \|V^* - V_{k+n}\|_\infty + \sum_{i=1}^{n} \|V_{k+i} - V_{k+i-1}\|_\infty \\ &\leq \max_{s \in S} |V^*(s) - V_{k+n}(s)| + \sum_{i=1}^{n} \gamma^i \epsilon\end{aligned}$$

Also:

$$\begin{aligned}V^*(s) &= \mathbb{E}[\sum_{n=0}^{\infty} \gamma^k r_{t+n+1} | s_t = s, \pi] = \sum_{n=0}^{\infty} \mathbb{E}[\gamma^k r_{t+n+1} | s_t = s, \pi] \\ &= \sum_{n=0}^{k-1} \mathbb{E}[\gamma^k r_{t+n+1} | s_t = s, \pi] + \sum_{n=k}^{\infty} \mathbb{E}[\gamma^k r_{t+n+1} | s_t = s, \pi] \\ &= C + \sum_{i=1}^{\infty} \|V_{k+i} - V_{k+i-1}\|_\infty = C + \frac{\epsilon}{1-\gamma}\end{aligned}$$

For some constant $C$, and from Cauchy's property, for some large $k$:

$$\sum_{n=k}^{\infty} \mathbb{E}[\gamma^k r_{t+n+1} | s_t = s, \pi] \leq \frac{\epsilon}{1-\gamma}$$

Thus $V^*(s)$ defined as a sum of series converges, we assumed there exist one unique $V^*(s)$ such that $V^*(s) \geq V(s)$, $\forall V \in \mathbb{V}$, and we can always construct a sequence of $V_n$ such that

$$\lim_{n\to\infty} V_n(s) = V^*(s)$$

Thus for that particular sequence of $V_n$:

$$\begin{aligned}\lim_{n\to\infty} \|V^* - V_k\|_\infty &\leq \lim_{n\to\infty} \big[\max_{s\in S} |V^*(s) - V_{k+n}(s)| + \sum_{i=1}^{n} \gamma^i \epsilon\big] \\ &= \lim_{n\to\infty} \big[0 + \epsilon\gamma^i\big] \\ &= \frac{\epsilon}{1-\gamma}\end{aligned}$$

(b) Suppose $\|V - V^*\|_\infty \leq \epsilon$ for some $V \in \mathbb{R}^{|S|}$. Let $\pi$ be the greedy policy with respect to $V$, $\pi(s) = \arg\max_a R(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s')$. Show that $\|V^* - V^\pi\|_\infty \leq \frac{2\gamma\epsilon}{1-\gamma}$.
Hint: Using the fact that $T^\pi V = T^* V$ since $\pi$ is greedy with respect to $V$, write $\|V^* - V^\pi\|_\infty = \|V^* - T^*V + T^\pi V - V^\pi\|_\infty$ and apply the triangle inequality.

$$\begin{aligned}T^\pi V(s) &= \mathbb{E}[r(s, a = \pi(s)) + \gamma V(s'|s, \pi(s))] \\ &= R(s, \pi(s)) + \gamma \sum_{s'} P^\pi_{s,s'} V(s') \\ &= \arg\max_a R(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \\ &= T^* V(s)\end{aligned}$$

For some choose of $V$ such that $\|V - V^*\|_\infty \leq \epsilon$:

$$\begin{aligned}\|V^* - V^\pi\|_\infty &= \|V^* - T^*V + T^\pi V - V^\pi\|_\infty \\ &= \max_{s\in S} |V^*(s) - T^*V(s) + T^\pi V(s) - V^\pi(s)| \\ &\leq \max_{s\in S} |T^*V(s) - V^*(s)| + \max_{s\in S} |T^\pi V(s) - V^\pi(s)|\end{aligned}$$

As we have $V^*$ and $T^*$ to be the optimal state value and Bellman operator, we have $T^*V^* = V^*$ that iterates with stationary property. Also for $V^\pi$ and $T^\pi$ to be the $\pi$ optimal state value and $\pi$ optimal operator which works the same as Bellman operator, we have $T^\pi V^\pi = V^\pi$ holds. As well from the first problem that $\|T^*V' - T^*V\|_\infty \leq \gamma \|V' - V\|_\infty$ for any $V, V' \in \mathbb{R}^{|S|}$, thus:

$$\begin{aligned}\|V^* - V^\pi\|_\infty &\leq \max_{s\in S} |T^*V(s) - T^*V^*(s)| + \max_{s\in S} |T^\pi V(s) - T^\pi V^\pi(s)| \\ &= \gamma \max_{s\in S} |V(s) - V^*(s)| + \gamma \max_{s\in S} |V(s) - V^\pi(s)|\end{aligned}$$

Similarily as (a) did, we can prove that $\max_{s\in S}|V(s)-V^*(s)| \leq \frac{\epsilon}{1-\gamma}$ and $\max_{s\in S}|V(s)-V^\pi(s)| \leq \frac{\epsilon}{1-\gamma}$ for our choose of $V$, thus

$$\begin{aligned}\|V^*-V^\pi\|_\infty \leq & \gamma \max_{s\in S}|V(s)-V^*(s)| + \gamma \max_{s\in S}|V(s)-V^\pi(s)| \\ = & \gamma\frac{\epsilon}{1-\gamma} + \gamma\frac{\epsilon}{1-\gamma} \\ = & \frac{2\gamma\epsilon}{1-\gamma}\end{aligned}$$

# 4 Frozen Lake MDP [25pt]

Implement value iteration and policy iteration for the Frozen Lake environment from OpenAI Gym. We have provided custom versions of this environment in the starter code in folder `assignment1_coding`.

Make sure you use Python 3 and have installed the dependencies in `requirements.txt`.

This problem is credited to Emma Brunskill.

(a) **(coding)** Read through `vi_and_pi.py` and implement `policy_evaluation`, `policy_improvement` and `policy_iteration`. The stopping tolerance (defined as $\max_s |V_{old}(s) - V_{new}(s)|$) is tol $= 10^{-3}$. Use $\gamma = 0.9$. Return the optimal value function and the optimal policy.

(b) **(coding)** Implement `value_iteration` in `vi_and_pi.py`. The stopping tolerance is tol $= 10^{-3}$. Use $\gamma = 0.9$. Return the optimal value function and the optimal policy.

(c) **(written)** Run both methods on the Deterministic-4x4-FrozenLake-v0 and Stochastic-4x4-FrozenLake-v0 environments. In the second environment, the dynamics of the world are stochastic. How does stochasticity affect the number of iterations required, and the resulting policy?

(a) policy iteration: value function: [0.59 0.656 0.729 0.656 0.531 0. 0.81 0. 0.478 0. 0.9 0. 0. 0. 1. 0. ] policy: [2 2 1 0 3 3 1 3 3 2 1 3 3 2 2 3]

(b) value iteration: value function: [0.59 0.656 0.729 0.656 0.656 0. 0.81 0. 0.729 0.81 0.9 0. 0. 0.9 1. 0. ] policy: [1 2 1 0 1 0 1 0 2 1 1 0 0 2 2 0] (c) The value iteration method converges faster, as policy evaluation takes time to simulate data. When changing to Stochastic iteration method, it takes much more time to converge for both methods, even can not converge. Reducing the tolerance could lead to converage but with chances that can not find the Goal point.