COMS 4771 Machine Learning (Spring 2020)
Problem Set #4

Joseph High - jph2185@columbia.edu

April 25, 2020

Problem 1: Finding the value of a state under a policy

Proof.

First, using the law of total expectation and conditioning over $a \in \mathcal{A}$:

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \sum_a \mathbb{E}_\pi[G_t \mid S_t = s, a_t = a] \cdot \underbrace{P\left(a_t = a | s_t = s\right)}_{= \pi(a|s)}$$

$$= \sum_a \pi(a \mid s)\, \mathbb{E}_\pi[G_t \mid S_t = s, a_t = a]$$

Again, using the law of total expectation, but now conditioning over s':

$$= \sum_a \pi(a \mid s) \sum_{s'} \mathbb{E}_\pi[G_t \mid S_t = s, a_t = a, S_{t+1} = s'] \cdot \underbrace{P\left(S_{t+1} = s' \mid S_t = s, a_t = a\right)}_{= P(s'|s,a)}$$

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\, \mathbb{E}_\pi[G_t \mid S_t = s, a_t = a, S_{t+1} = s']$$

Substituting in the definition of G_t:

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\, \mathbb{E}_\pi\left[\mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k}\right] \,\middle|\, S_t = s, a_t = a, S_{t+1} = s'\right]$$

Pulling out the first summand, R_{t+1}, from the infinite series:

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\, \mathbb{E}_\pi\left[\mathbb{E}\left[R_{t+1} + \sum_{k=2}^{\infty} \gamma^{k-1} R_{t+k}\right] \,\middle|\, S_t = s, a_t = a, S_{t+1} = s'\right]$$

Re-indexing the infinite series and factoring out a γ :

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\, \mathbb{E}_\pi\left[\mathbb{E}\left[R_{t+1} + \gamma \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k+1}\right] \,\middle|\, S_t = s, a_t = a, S_{t+1} = s'\right]$$

Using the linearity property of expectation:

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a) \, \mathbb{E}_\pi \left[R_{t+1} + \gamma \, \mathbb{E} \left[\sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k+1} \right] \,\middle|\, S_t = s, a_t = a, S_{t+1} = s' \right]$$

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a) \, \mathbb{E}_\pi [R_{t+1} + \gamma G_{t+1} \mid S_t = s, a_t = a, S_{t+1} = s']$$

Again, using the linearity property of expectation:

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a) \Big[\mathbb{E}_\pi [R_{t+1} \mid S_t = s, a_t = a, S_{t+1} = s'] + \gamma \, \mathbb{E}_\pi [G_{t+1} \mid S_t = s, a_t = a, S_{t+1} = s'] \Big]$$

By the Markov property, $\mathbb{E}_\pi [G_{t+1} \mid S_t = s, a_t = a, S_{t+1} = s'] = \mathbb{E}_\pi [G_{t+1} \mid S_{t+1} = s']$. Applying this to the above:

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a) \Big[\underbrace{\mathbb{E}_\pi [R_{t+1} \mid S_t = s, a_t = a, S_{t+1} = s']}_{= R_a(s,s')} + \gamma \underbrace{\mathbb{E}_\pi [G_{t+1} \mid S_{t+1} = s']}_{= v_\pi(s')} \Big]$$

$$= \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a) \big[R_a(s, s') \; + \; \gamma v_\pi(s') \big]$$

Hence,

$$v_\pi(s) \; = \; \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a) \big[R_a(s, s') \; + \; \gamma v_\pi(s') \big]$$

□

Problem 2: Solving for a value function using linear algebra

In Problem 1 it was shown that

$$v_\pi(s) \;=\; \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a) \big[R_a(s, s') \;+\; \gamma v_\pi(s') \big]$$

Applying the assumption that all transitions are deterministic, i.e., $P(s' \mid s, a) = \mathbb{1}\{s' = next(s, a)\}$, we get

$$v_\pi(s) \;=\; \sum_a \pi(a \mid s) \sum_{s'} \mathbb{1}\{s' = next(s, a)\} \cdot \big[R_a(s, s') \;+\; \gamma v_\pi(s') \big]$$

That is, for each action a, given the current state s, there is exactly one subsequent state. Thus, for each action a, the second summation reduces to a single term:

$$\begin{aligned} v_\pi(s) \;&=\; \sum_a \pi(a \mid s) \big[R_a(s, s') \;+\; \gamma v_\pi(s') \big] \\ &=\; \sum_a \pi(a \mid s) R_a(s, s') \;+\; \gamma \sum_a \pi(a \mid s) v_\pi(s') \\ &=\; \mathbb{E}_\pi[R_a(s, s') \mid S_t = s] \;+\; \gamma\, \mathbb{E}_\pi[v_\pi(s') \mid S_t = s] \end{aligned}$$

Rearranging, we have

$$\begin{aligned} v_\pi(s) \;-\; \gamma\, \mathbb{E}_\pi[v_\pi(s') \mid S_t = s] \;&=\; \mathbb{E}_\pi[R_a(s, s') \mid S_t = s] \\ \implies \quad v_\pi(s) \;-\; \gamma \sum_{s'} P(s'|s) v_\pi(s') \;&=\; \mathbb{E}_\pi[R_a(s, s') \mid S_t = s] \\ \implies \quad v_\pi(s) \;-\; \gamma \sum_{s'} P(s'|s) v_\pi(s') \;&=\; \mathbb{E}_\pi[R_a(s, s') \mid S_t = s] \end{aligned}$$

where $P(s'|s)$ is the probability of transitioning from state s to the subsequent state s', and so $\sum_{s'} P(s'|s)$ is the sum of transition probabilities over all possible subsequent states.

The above can be expressed as a system of linear equations, where each equation is the value function at a particular current state s_i. All possible subsequent states are denoted by s_j (see below).

$$v_\pi(s_i) \;-\; \gamma \sum_{j=1}^{n} P(s_j|s_i) v_\pi(s_j) \;=\; \mathbb{E}_\pi[R_a(s, s') \mid S_t = s]$$

In matrix form, this is

$$\begin{bmatrix} v_\pi(s_1) \\ \vdots \\ v_\pi(s_n) \end{bmatrix} \;-\; \gamma \begin{bmatrix} p(s_1|s_1) & \cdots & p(s_n|s_1) \\ \vdots & \ddots & \vdots \\ p(s_1|s_n) & \cdots & p(s_n|s_n) \end{bmatrix} \begin{bmatrix} v_\pi(s_1) \\ \vdots \\ v_\pi(s_n) \end{bmatrix} \;=\; \begin{bmatrix} \mathbb{E}_\pi[R_a(s_1, s')] \\ \vdots \\ \mathbb{E}_\pi[R_a(s_n, s')] \end{bmatrix}$$

That is,

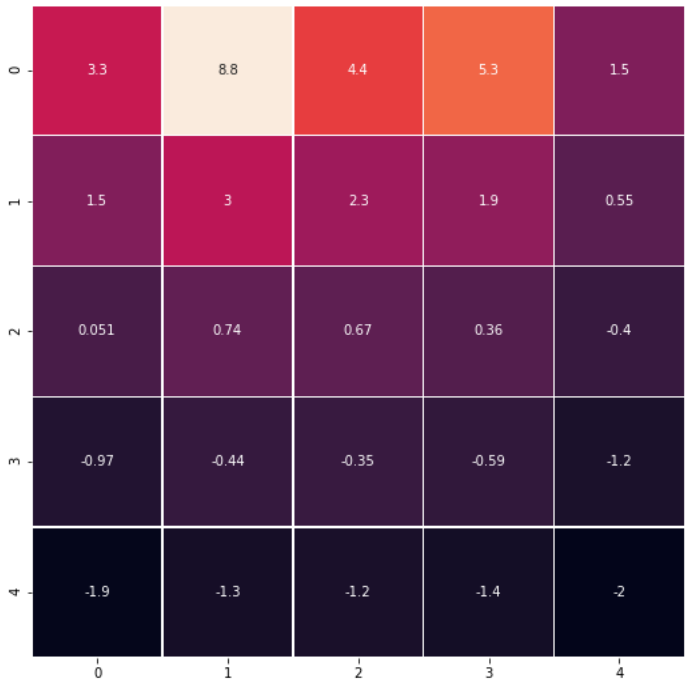
$$\begin{aligned} & \mathbf{v}_\pi \;-\; \gamma \mathbf{P} \mathbf{v}_\pi \;=\; \mathbf{R}_\pi \\ \Longrightarrow \qquad & (\mathbf{I} - \gamma \mathbf{P})\, \mathbf{v}_\pi \;=\; \mathbf{R}_\pi \end{aligned}$$

where $\mathbf{R}_\pi$ denotes the vector of expected returns on the right-hand side of the expression above.

Problem 3: Finding the value states "in the real world"

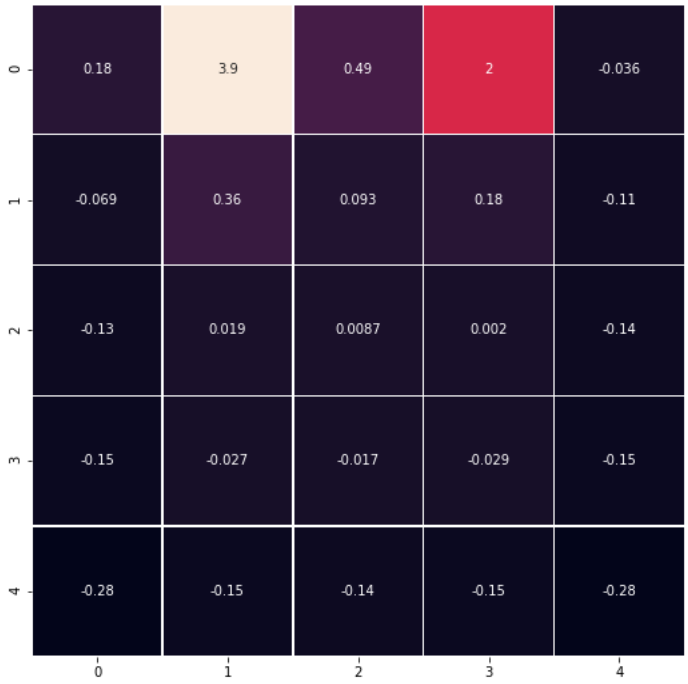
For uniform policy π and $\gamma = 0.9$, it can be seen that the north is favored. In particular, it appears that state A is the optimal state to be in. Note the negative values along the bottom of the grid. This is because of the increased probability of leaving the grid under this policy.

Figure 1: Grid for Uniform Policy π and $\gamma = 0.9$



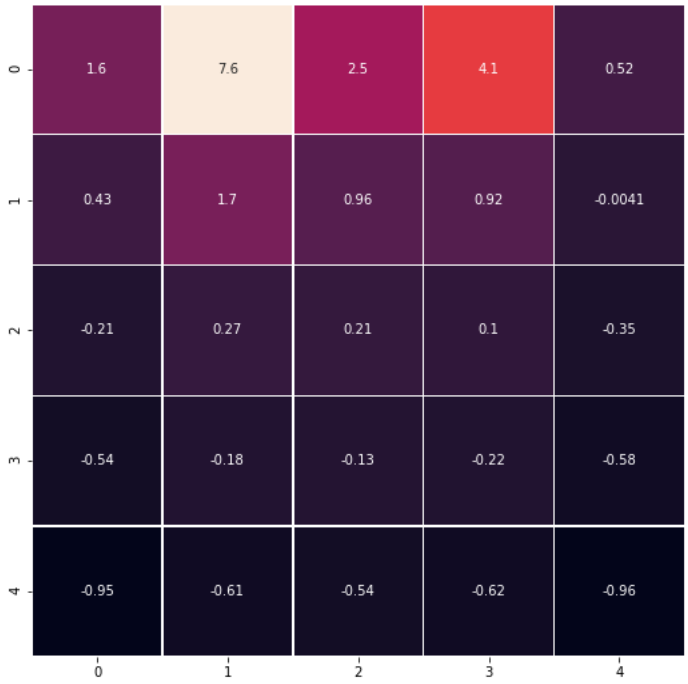
For non-uniform policy $\pi(north|s) = 0.7$ and 0.1, it can be seen that the north is favored here as well. Again, it appears that state A is the optimal state to be in. Note again, the negative values along the bottom of the grid.

Figure 2: Grid for $\pi(north|s) = 0.7$ and 0.1 for all other directions



For non-uniform policy $\pi(north|s) = \pi(south|s) = 0.4$ and $\pi(east|s) = \pi(west|s) = 0.1$, one will note that, again, the north is favored. Again, it appears that state A is the optimal state to be in. Note again, the negative values along the bottom of the grid.

Figure 3: Grid for $\pi(north|s) = \pi(south|s) = 0.4$ and $\pi(east|s) = \pi(west|s) = 0.1$



Problem 4: Finding an optimal value function

(a) Because a policy governs the actions taken and the associated probabilities, maximizing $v_\pi(s)$ for a given state s, over all policies π, is tantamount to choosing the action a that achieves the maximum value. Let a^* denote such an action. Then, $a^* = \arg\max_a\{v_\pi(s)\}$. Then, an optimal policy π^*, will always prescribe a^*. More specifically, for a given state s, an optimal policy will always distribute all of the weight/probability to a^*. Then, because $\sum_a \pi(a \mid s) = 1$, for a given s, we have that for an optimal policy π^*:

$$\pi^*(a \mid s) = \begin{cases} 1 & \text{if } a = \arg\max_a\{v_\pi(s)\} \\ 0 & \text{otherwise} \end{cases}$$

In other words, π^* is deterministic. The value function for an optimal policy, π^*, is such that

$$\begin{aligned} v_{\pi^*}(s) &= \sum_a \pi^*(a \mid s) \sum_{s'} P(s' \mid s, a)\big[R_a(s, s') \ + \ \gamma v_*(s')\big] \\ &= (\pi^*(a_1|s) + \cdots + \pi^*(a^*|s) + \cdots + \pi^*(a_k|s)) \sum_{s'} P(s' \mid s, a)\big[R_a(s, s') \ + \ \gamma v_*(s')\big] \\ &= (0 + \cdots + \ 1 \ + \cdots + 0) \sum_{s'} P(s' \mid s, a)\big[R_a(s, s') \ + \ \gamma v_*(s')\big] \\ &= \sum_{s'} P(s' \mid s, a)\big[R_a(s, s') \ + \ \gamma v_*(s')\big] \end{aligned}$$

In 4*b* it's argued that $v_{\pi^*}(s) = v_*(s)$. The optimal value function is then

$$\begin{aligned} v_*(s) &= \max_\pi\{v_\pi(s)\} \\ &= \max_\pi \left\{ \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\big[R_a(s, s') \ + \ \gamma v_\pi(s')\big] \right\} \end{aligned}$$

By the recursive definition of the value function, the optimal value at a state s must include the optimal value (discounted by γ) for each subsequent state, s'. It then follows that,

$$= \max_a \left\{ \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\big[R_a(s, s') \ + \ \gamma \max_\pi v_\pi(s')\big] \right\}$$

From the above argument that an optimal policy will always distribute all of the weight/probability to a^*:

$$= \max_a \left\{ \sum_{s'} P(s' \mid s, a)\big[R_a(s, s') \ + \ \gamma v_*(s')\big] \right\}$$

Therefore,

$$v_*(s) \;=\; \max_a \sum_{s'} P(s' \mid s, a)\big[R_a(s, s') \;+\; \gamma v_*(s')\big]$$

(b) An optimal policy can be found by recursively choosing the action that achieves the maximum value at each state s, and distributing all of the weight/probability to this action. That is, at each state s, choose the action $a^* = \arg\max_a v_\pi(s)$, and set $\pi(a^*|s) = 1$ and assign 0 to all other policies. Then at each state s, the value function for this optimal policy π^* is such that $v_\pi(s) \leq v_{\pi^*}(s)$ for all policies π and all $s \in \mathcal{S}$. However, this is also the relationship that $v_*(s)$ has with $v_\pi(s)$, for all policies π. Indeed, it follows from the definition of $v_*(s)$ being the optimal value function that $v_\pi(s) \leq v_*(s)$ for all policies π and all $s \in \mathcal{S}$. Therefore, it must be the case that $v_{\pi^*}(s) = v_*(s)$. Thus, the optimal policy is indeed optimal.

The optimal policy π^* is deterministic. Indeed, as discussed in part (a), it will always choose the action $a^* = \arg\max_a v_\pi(s)$. Therefore, all of the probability/weight will be allocated to a^* at each state, whereas all other actions will be assigned a probability of zero. This implies that the optimal policy is non-random. That is, the action it prescribes when in state s is a deterministic function of s.

(c) *Proof.* First, note that $q_\pi(s, a) \;=\; \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$. Now, starting from the definition of the value function, we have

$$\begin{aligned} v_\pi(s) \;&=\; \mathbb{E}_\pi[G_t \mid S_t = s] \\ &=\; \sum_a \underbrace{\mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]}_{=\, q_\pi(s,a)} \pi(a \mid s) \qquad \textit{by the law of total expectation} \\ &=\; \sum_a \pi(a \mid s) q_\pi(s, a) \end{aligned}$$

That is,

$$v_\pi(s) \;=\; \sum_a \pi(a \mid s) q_\pi(s, a)$$

In Problem 1, we found that

$$v_\pi(s) \;=\; \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\big[R_a(s, s') \;+\; \gamma v_\pi(s')\big]$$

Therefore,

$$\sum_a \pi(a \mid s) q_\pi(s, a) \;=\; \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a)\big[R_a(s, s') \;+\; \gamma v_\pi(s')\big]$$

$$\implies \quad q_\pi(s, a) \;=\; \sum_{s'} P(s' \mid s, a)\big[R_a(s, s') \;+\; \gamma v_\pi(s')\big]$$

Using this form of the q-function, we will maximize over all policies π:

$$\begin{aligned} q_*(s,a) \;=\; \max_\pi q_\pi(s,a) \;&=\; \max_\pi \sum_{s'} P(s' \mid s,a)\big[R_a(s,s') \;+\; \gamma v_\pi(s')\big] \\ &= \max_\pi \left\{ \sum_{s'} P(s' \mid s,a) R_a(s,s') \;+\; \gamma \sum_{s'} P(s' \mid s,a) v_\pi(s') \right\} \end{aligned}$$

Since we are maximizing over π, all terms independent of π can be moved outside of the max operator. While actions are dependent on π, the terms being moved outside of the max operator are conditioning on the action a. That is, a is given in those terms:

$$\begin{aligned} &= \sum_{s'} P(s' \mid s,a) R_a(s,s') \;+\; \gamma \sum_{s'} P(s' \mid s,a) \max_\pi v_\pi(s') \\ &= \sum_{s'} P(s' \mid s,a) R_a(s,s') \;+\; \gamma \sum_{s'} P(s' \mid s,a) v_*(s') \\ &= \sum_{s'} P(s' \mid s,a)\big[R_a(s,s') \;+\; \gamma v_*(s')\big] \\ \implies \quad q_*(s,a) \;&=\; \sum_{s'} P(s' \mid s,a)\big[R_a(s,s') \;+\; \gamma v_*(s')\big] \end{aligned}$$

Now, substituting in the expression for v_* from $4a$ into $q_*(s,a)$:

$$q_*(s,a) \;=\; \sum_{s'} P(s' \mid s,a)\big[R_a(s,s') + \gamma \max_{a'} \underbrace{\sum_{s''} P(s'' \mid s',a')[R_{a'}(s',s'') + \gamma v_*(s'')]}_{= q_*(s',a')}\big]$$

where s'' denotes the state subsequent to s'. Notice that the expression in the max operator satisfies the equation for q_* that was determined above. Hence, we have

$$q_*(s,a) \;=\; \sum_{s'} P(s' \mid s,a)\big[R_a(s,s') + \gamma \max_{a'} q_*(s',a')\big]$$

□

Problem 5: Finding the optimal policy using iterative methods

Proof.

$$
\begin{aligned}
\|V_t(s) - V_*(s)\|_\infty &= \max_{s\in\mathcal{S}} |V_t(s) - V_*(s)| \\
&= \max_{s\in\mathcal{S}} \Bigg| \max_a \sum_{s'} P(s'|s,a)\big[R_a(s,s') \;+\; \gamma V_{t-1}(s')\big] \\
&\qquad\qquad - \max_a \sum_{s'} P(s'|s,a)[R_a(s,s') \;+\; \gamma V_*(s')] \Bigg| \\
&\leq \max_{s\in\mathcal{S}} \Bigg| \max_a \Bigg\{ \sum_{s'} P(s'|s,a)\big[R_a(s,s') \;+\; \gamma V_{t-1}(s')\big] \\
&\qquad\qquad - \sum_{s'} P(s'|s,a)\big[R_a(s,s') \;+\; \gamma V_*(s')\big] \Bigg\} \Bigg| \\
&= \max_{s\in\mathcal{S}} \Bigg| \max_a \underbrace{\sum_{s'} P(s'|s,a)}_{=\,1} \big[\gamma V_{t-1}(s') \;-\; \gamma V_*(s')\big] \Bigg| \\
&= \gamma \max_{s\in\mathcal{S}} \big|V_{t-1}(s') \;-\; V_*(s')\big| \\
&\leq \gamma \max_{s\in\mathcal{S}} \big|V_{t-1}(s) \;-\; V_*(s)\big| \\
&= \gamma \|V_{t-1}(s) - V_*(s)\|_\infty
\end{aligned}
$$

Therefore,

$$\|V_t(s) - V_*(s)\|_\infty \;\leq\; \gamma \|V_{t-1}(s) - V_*(s)\|_\infty$$

In general, using the same argument as above, it can be shown that for all $k \in \{1, 2, \ldots, t-1\}$ and any $\gamma \in (0, 1)$

$$\|V_{t-k}(s) - V_*(s)\|_\infty \;\leq\; \gamma \|V_{t-k-1}(s) - V_*(s)\|_\infty$$

Then,

$$
\begin{aligned}
\|V_t(s) - V_*(s)\|_\infty &\leq \gamma \|V_{t-1}(s) - V_*(s)\|_\infty \\
&\leq \gamma^2 \|V_{t-2}(s) - V_*(s)\|_\infty \\
&\leq \gamma^3 \|V_{t-3}(s) - V_*(s)\|_\infty \\
&\;\;\vdots \qquad\qquad \vdots \\
&\leq \gamma^{t-1} \|V_1(s) - V_*(s)\|_\infty \\
&\leq \gamma^t \|V_0(s) - V_*(s)\|_\infty
\end{aligned}
$$

$$\implies \quad \|V_t(s) - V_*(s)\|_\infty \;\leq\; \gamma^t \|V_0(s) - V_*(s)\|_\infty \qquad \square$$

Now to show that the inequality above implies convergence.

Since $\gamma \in (0,1)$, $\gamma^t \longrightarrow 0$ as $t \longrightarrow \infty$

$\implies \quad \gamma^t \|V_0(s) - V_*(s)\|_\infty \longrightarrow 0$ as $t \longrightarrow \infty$

Then, the inequality $\|V_t(s) - V_*(s)\|_\infty \leq \gamma^t \|V_0(s) - V_*(s)\|_\infty$ implies that

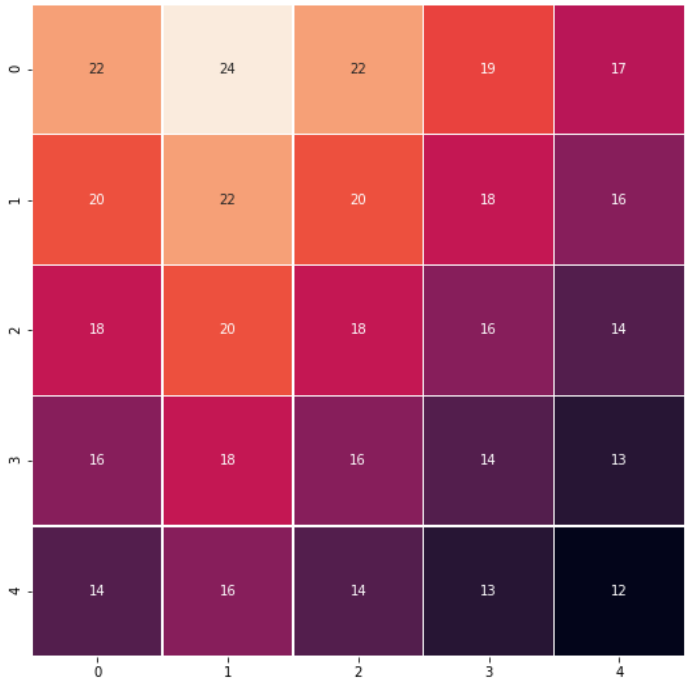
$\|V_t(s) - V_*(s)\|_\infty \longrightarrow 0$ as $t \longrightarrow \infty$

Therefore, the value function converges to the optimal value function as $t \longrightarrow \infty$.

Problem 6: Find the optimal value function for gridworld

The optimal value function was computed using the value iteration approach (see below). This optimal policy in this case is conceptually sound given that the greatest reward is in A.

Figure 4: Optimal Value Grid



Problem 7: A model-free approach

i) At every time step, the Q estimate changes. In particular, it increases or decreases depending on the magnitude of the reward associated with the action taken. For larger rewards, there is a larger change in the Q estimate, while for smaller rewards, there is a smaller change in the estimate of Q. However, by implementing an ϵ-greedy policy and for a reasonable value of α, the changes in the Q estimate between states will diminish. Indeed, if α is sufficiently small, your policy will learn on the most up-to-date information available. Therefore, for such an α and by acting ϵ-greedily with respect to the Q estimate, the policy will improve at every timestep. As the policy improves, your estimate of Q improves, and since the change in the estimates decrease at every timestep, the Q-estimate will converge to the optimal state action function for large values of t (Recall the result from Problem 5).

ii) Despite having started the homework early, I was not able to complete 7(ii) in time. However, I am actively working on this problem, so that I don't miss out on anything. Sorry to disappoint.

Extra Credit

i) From the definition of G_t, we have

$$G_t \;=\; \mathbb{E}\left[\sum_{n=1}^{\infty}\gamma^{n-1}R_{t+n}\right] \;=\; \mathbb{E}\left[R_{t+1} \;+\; \gamma\sum_{n=1}^{\infty}\gamma^{n-1}R_{t+n+1}\right]$$

$$=\; \mathbb{E}[R_{t+1}] \;+\; \gamma\,\mathbb{E}\left[\sum_{n=1}^{\infty}\gamma^{n-1}R_{t+n+1}\right]$$

$$=\; R_{t+1} + \gamma G_{t+1}$$

We also know that $V_\pi(S_t) = \mathbb{E}[G_t \mid S_t = s]$. Because, $V_\pi(S_t)$ is the expected value of G_t, conditional on a specific state, G_t can be estimated by $V_\pi(S_t)$ (biased estimate). Then,

$$G_t \;\approx\; R_{t+1} + \gamma V_\pi(S_{t+1})$$

The recursive rule then becomes

$$V_\pi(S_t) \longleftarrow V_\pi(S_t) + \alpha[R_{t+1} + \gamma V_\pi(S_{t+1}) - V_\pi(S_t)]$$

At two timesteps ahead, G_t can be expressed using following recursive argument:

$$\begin{aligned} G_t &= R_{t+1} + \gamma G_{t+1} \\ &= R_{t+1} + \gamma(R_{t+2} + \gamma G_{t+2}) \\ &= R_{t+1} + \gamma R_{t+2} + \gamma^2 G_{t+2} \end{aligned}$$

Now, since G_{t+2} can be estimated by $V_\pi(S_{t+2})$, the 2-steps-ahead estimate of G_t is

$$G_t \approx R_{t+1} + \gamma R_{t+2} + \gamma^2 V_\pi(S_{t+2})$$

and so, the recursive rule for two timesteps ahead becomes

$$V_\pi(S_t) \longleftarrow V_\pi(S_t) + \alpha[R_{t+1} + \gamma R_{t+2} + \gamma^2 V_\pi(S_{t+2}) - V_\pi(S_t)]$$

ii) We can write G_t in terms of the expected reward n-steps ahead using the same argument for two steps ahead in part (i):

$$\begin{aligned} G_t &= R_{t+1} + \gamma G_{t+1} \\ &= R_{t+1} + \gamma(R_{t+2} + \gamma G_{t+2}) \\ &= R_{t+1} + \gamma R_{t+2} + \gamma^2 G_{t+2} \\ &= R_{t+1} + \gamma R_{t+2} + \gamma^2(R_{t+3} + \gamma G_{t+3}) \\ &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 G_{t+3} \\ &\;\;\vdots \\ &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n G_{t+n} \\ &= \sum_{k=1}^{n} \gamma^{k-1} R_{t+k} + \gamma^n G_{t+n} \end{aligned}$$

By estimating G_{t+n} with $V_\pi(S_{t+n})$, the n-steps-ahead estimate of G_t is

$$G_t \approx \sum_{k=1}^{n} \gamma^{k-1} R_{t+k} + \gamma^n V_\pi(S_{t+n})$$

The recursive update rule for the n-steps-ahead estimate is then

$$V_\pi(S_t) \longleftarrow V_\pi(S_t) + \alpha \left[\sum_{k=1}^{n} \gamma^{k-1} R_{t+k} + \gamma^n V_\pi(S_{t+n}) - V_\pi(S_t) \right]$$

iii) Our n-steps-ahead estimate of G_t is

$$G_t \approx \sum_{k=1}^{n} \gamma^{k-1} R_{t+k} + \gamma^n V_\pi(S_{t+n})$$

The sum of discounted rewards, R_{t+k}, adds noise/variance to the estimate of G_t since each term is non-deterministic (or random), and because there is randomness in each transition from state-to-state since the action chosen is also random. Then, because G_t incurs noise from each additional transition (i.e., each additional $\gamma^k R_{t+k}$ term, $k \in \{2, ..., n\}$), the variance will increase as n increases.

$V_\pi(S_{t+n})$ introduces bias into the estimate of G_t since it is an estimate of the value function, v_π, at time step $t+n$. However, there is a discount factor γ^n on $V_\pi(S_{t+n})$ in the estimated expression for G_t, where $\gamma \in (0, 1)$. Then, as n increases, $\gamma^n V_\pi(S_{t+n})$ decreases, and so the bias in the estimate decreases.

iv) Set $a_n = (1-\lambda)\lambda^n$ for $0 < \lambda < 1$. Indeed, λ^n gives less weight to the $G_t^{(n)}$ for larger values of n, making the $G_t^{(n)}$ worth exponentially less for larger values of n. The $(1-\lambda)$ ensures that the weighted sum is a convex combination. Indeed,

$$\begin{aligned}(1-\lambda)\sum_{n=1}^{n} \lambda^n &= (1-\lambda)(1 + \lambda + \lambda^2 + \cdots + \lambda^n) \\ &= 1 - \lambda^n \longrightarrow 1 \text{ as } n \longrightarrow \infty\end{aligned}$$

That is,

$$(1-\lambda)\sum_{n=1}^{\infty} \lambda^n = 1$$

The recursive update rule is then,

$$V_\pi(S_t) \longleftarrow V_\pi(S_t) + \alpha \left[(1-\lambda) \sum_{n=1}^{\infty} \lambda^n G_t^{(n)} - V_\pi(S_t) \right]$$

v) This approach requires that all n returns be determined before updating the value function estimate. Therefore, the algorithm makes updates every n-steps, or at the end of each episode.

vi) Because $\lambda \in (0, 1)$, the $(1 - \lambda)\lambda^n$ weights on $G_t^{(n)}$ are exponentially smaller for smaller values of λ, decreasing the impact the noise has on each k-steps-ahead estimate coming from the rewards and transitions/actions. Similarly, the weights are exponentially larger for larger values of λ, increasing the impact the noise has on each estimate. Therefore, the larger the value of λ, the larger the variance. On the other hand, larger values of λ will result in smaller bias.