CS7643: Deep Learning
Fall 2020
HW5 Solutions

Ruize Yang

November 5, 2020

Problem 1a

1.(a). When stare at S_1 and always choose stay, it will always in S_1. The reward of each step is -2.

$$\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t) = -2 \sum_{t=0}^{\infty} \gamma^t = -\frac{2}{1-\gamma}.$$

Problem 1b

1.(b) If $\gamma \geq 0.1367$, $(a_1, a_2) = (go, go)$. If $\gamma < 0.1367$, $(a_1, a_2) = (stay, go)$ or $(stay, stay)$.

There are four possible policy:

① $(stay, stay) \Rightarrow S_1$ ② $(go, go) \Rightarrow S_1 \rightleftarrows S_2$

③ $(stay, go) \Rightarrow S_1$ ④ $(go, stay) \Rightarrow S_1 \rightarrow S_2$

The reward of ①、③ is $-\frac{2}{1-\gamma}$, and is higher than that of ④, because $S_1 \rightarrow S_2$ is -3, and stay in S_1 is -2.

The reward of ② is $-3+5\gamma$, because at the second step, the agent takes "go" from S_2 and ends the episode.

$-3+5\gamma \geq -\frac{2}{1-\gamma} \Rightarrow \gamma \geq 0.1367$., $\Rightarrow$ $0.1367 \leq \gamma < 1$, $(a_1, a_2) = (go, go)$

$0 \leq \gamma < 0.1367$, $(a_1, a_2) = (stay, stay)$

or $(a_1, a_2) = (stay, go)$.

Problem 1c

1 (c). $\gamma = 1$.

$V_1(S_1) = \max\{r(S_1, stay) + V_0(S_1),\ r(S_1, go) + V_0(S_2)\}$

$= \max\{-2+0,\ -3+0\} = -2$

$V_1(S_2) = \max\{r(S_2, stay) + V_0(S_2),\ r(S_2, go)\}$

$= \max\{-2+0,\ 5\} = 5$

$V_1 = [-2, 5]$.

$V_2(S_1) = \max\{r(S_1, stay) + V_1(S_1),\ r(S_1, go) + V_{\cancel{0}1}(S_2)\}$

$= \max\{-2-2,\ -3+5\} = 2$.

$V_2(S_2) = \max\{r(S_2, stay) + V_1(S_2),\ r(S_2, go)\}$

$= \max\{-2+5,\ 5\} = 5$

$V_2 = [2, 5]$.

$V_3(S_1) = \max\{r(S_1, stay) + V_2(S_1),\ r(S_1, go) + V_2(S_2)\}$

$= \max\{-2+2,\ -3+5\} = 2$.

$V_3(S_2) = \max\{r(S_2, stay) + V_2(S_2),\ r(S_2, go)\}$

$= \max\{-2+5,\ 5\} = 5$.

$V_3 = [2, 5]$.

V_2, V_3 are both optimal, but V^3 is better, because it shows V converges.

Problem 2a

2 (a). V^1: $\| V^1 - V^* \| = \max\{ |-2-2|, |5-5| \} = 4.$

V^2: $\| V^2 - V^* \| = \max\{ |2-2|, |5-5| \} = 0$

V^3: $\| V^3 - V^* \| = \max\{ |2-2|, |5-5| \} = 0.$

The error decreases monotonically

Problem 2b

2(b). For state s:

$$T(V)(s) = \max_a \sum_{s'} p(s' \mid s,a)[r(s,a) + \gamma V(s')]$$

$$T(V')(s) = \max_a \sum_{s'} P(s' \mid s,a)[r(s,a) + \gamma V'(s')]$$

$$\|T(V)(s) - T(V')(s)\| \leq \max_a \Big| \sum_{s'} p(s' \mid s,a)[r(s,a)+\gamma V(s')] - \sum_{s'} p(s' \mid s,a)[r(s,a)+\gamma V'(s')] \Big|$$

$$= \max_a \Big| \sum_{s'} P(s' \mid s,a)[\gamma V(s') - \gamma V'(s')] \Big|$$

$$= \gamma \max_a \sum_{s'} P(s' \mid s,a) \left| V(s') - V'(s') \right|.$$

$$\|T(V) - T(V')\|_\infty = \gamma \max_{a,s} \sum_{s'} p(s' \mid s,a) \left| V(s') - V'(s') \right|$$

$$\leq \gamma \max_{a,s} \sum_{s'} P(s' \mid s,a) \|V - V'\|_\infty$$

$$= \gamma \|V - V'\|_\infty \max_{a,s} \sum_{s'} P(s' \mid s,a)$$

$$= \gamma \|V - V'\|_\infty$$

$$\Rightarrow \|T(V) - T(V')\|_\infty \leq \gamma \|V - V'\|_\infty, \quad \forall\, V, V'.$$

Problem 2c

2(c). $\|V^{n+1} - V^*\|_\infty = \|V^{n+1} - T(V^{n+1}) + T(V^{n+1}) - V^*\|_\infty$

$$\le \|V^{n+1} - T(V^{n+1})\|_\infty + \|T(V^{n+1}) - V^*\|_\infty$$

$$= \|T(V^n) - T(V^{n+1})\|_\infty + \|T(V^{n+1}) - T(V^*)\|_\infty$$

$$\le \gamma \|V^n - V^{n+1}\|_\infty + \gamma \|V^{n+1} - V^*\|_\infty.$$

$$(1-\gamma)\|V^{n+1} - V^*\|_\infty \le \gamma \|V^n - V^{n+1}\|_\infty$$

$$\|V^{n+1} - V^*\|_\infty \le \frac{\gamma}{1-\gamma} \|V^n - V^{n+1}\|_\infty.$$

As $\{V^n\}$ is a cauchy sequence, then $\forall \epsilon, \exists n.$ s.t. $\|V^n - V^{n+1}\| < \epsilon.$

$$\Rightarrow \|V^{n+1} - V^*\|_\infty \le \frac{\gamma}{1-\gamma} \epsilon.$$

Problem 2d

2(d) Define the sequence $\{x_n\}$ as $x_n = T(x_{n-1})$,

$\|T(x_{n+1}) - T(x_n)\|_\infty \le \alpha \| x_{n+1} - x_n\|_\infty = \alpha\|T(x_n) - T(x_{n-1})\|_\infty \le \alpha^n \| x_1 - x_0\|_\infty$

$0 \le \alpha < 1$, $\|x_1 - x_0\|_\infty$ is a constant.

For $m > n$, $\|T(x_m) - T(x_n)\|_\infty \le \|T(x_m) - T(x_{m-1})\|_\infty + \cdots + \|T(x_{n+1}) - T(x_n)\|_\infty$

$$
\begin{aligned}
&\le \alpha^{m-1}\|x_1 - x_0\|_\infty + \cdots + \alpha^n \|x_1 - x_0\|_\infty \\
&= \alpha^n \|x_1 - x_0\|_\infty \sum_{k=0}^{m-n-1} \alpha^k \\
&\le \alpha^n \|x_1 - x_0\|_\infty \sum_{k=0}^{\infty} \alpha^k \\
&= \alpha^n \|x_1 - x_0\|_\infty \frac{1}{1-\alpha}
\end{aligned}
$$

For any ϵ, $\exists\, n$. s.t. $\alpha^n \|x_1 - x_0\| \frac{1}{1-\alpha} < \epsilon$, $\{x_n\}$ is cauchy sequence.

So $\{x_n\}$ will converge, let it converges to x^*.

$x^* = \lim_{n\to\infty} x_n = \lim_{n\to\infty} T(x_{n-1}) = T(\lim_{n\to\infty} x_{n-1}) = T(x^*)$

x^* is a fixed point of $\{x_n\}$.

Assume there is another fixed point x', $T(x') = x'$.

$\|x^* - x'\|_\infty = \|T(x^*) - T(x')\|_\infty \le \alpha \|x^* - x'\|_\infty$.

$\alpha < 1 \Rightarrow \|x^* - x'\|_\infty = 0 \Rightarrow x^* = x'$.

So the fixed point x^* exists and is unique.

Problem 3a

Problem 3b

Problem 3c

Problem 3d

Problem 4a

4 (a). $\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)]$

$$\Rightarrow \frac{1}{N}\sum_{i=1}^{N}(R(\tau_i) - b)\nabla_\theta \log \pi_\theta(\tau_i)$$

$$= \frac{1}{N}\sum_{i=1}^{N} R(\tau_i)\nabla_\theta \log \pi_\theta(\tau_i) - \frac{1}{N}\sum_{i=1}^{N} b\nabla_\theta \log \pi_\theta(\tau_i)$$

$$\frac{1}{N}\sum_{i=1}^{N} b\nabla_\theta \log \pi_\theta(\tau_i) = \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(\tau_i)\cdot b]$$

$$= \mathbb{E}_{s_{0:t}, a_{0:t-1}}[\mathbb{E}_{s_{t+1:T}, a_{t:T-1}}[\nabla_\theta \log \pi_\theta(\tau_i)\cdot b]]$$

$$= \mathbb{E}_{s_{0:t}, a_{0:t-1}}[b\cdot \mathbb{E}_{s_{t+1:T}, a_{t:T-1}}[\nabla_\theta \log \pi_\theta(\tau_i)]]$$

$$= \mathbb{E}_{s_{0:t}, a_{0:t-1}}[b\cdot \mathbb{E}_{a_t}[\nabla_\theta \log \pi_\theta(\tau_i)]]$$

$$= \mathbb{E}_{s_{0:t}, a_{0:t-1}}[b\cdot \int \frac{\nabla_\theta \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)}\pi_\theta(a_t|s_t)da_t$$

$$= \mathbb{E}_{s_{0:t}, a_{0:t-1}}[b\cdot \nabla_\theta \int \pi_\theta(a_t|s_t)\,da_t]$$

$$= \mathbb{E}_{s_{0:t}, a_{0:t-1}}[b\cdot \nabla_\theta 1]$$

$$= \mathbb{E}_{s_{0:t}, a_{0:t-1}}[b\cdot 0]$$

$$= 0$$

So $\frac{1}{N}\sum_{i=1}^{N}(R(\tau_i)-b)\nabla_\theta \log \pi_\theta(\tau_i) = \frac{1}{N}\sum_{i=1}^{N} R(\tau_i)\nabla_\theta \log \pi_\theta(\tau_i)$,
doesn't change the gradient estimate.

Problem 4b

4(b). $x = R(\tau)\nabla_\theta \log \pi_\theta(\tau)$, $x' = (R(\tau)-b)\nabla_\theta \log \pi_\theta(\tau)$

$Var(x') = \mathbb{E}([(R(\tau)-b)\nabla_\theta \log \pi_\theta(\tau)]^2) - [\mathbb{E}((R(\tau)-b)\nabla_\theta \log \pi_\theta(\tau))]^2$

~~The second term~~

$Var(x) = \mathbb{E}([R(\tau)\nabla_\theta \log \pi_\theta(\tau)]^2) - [\mathbb{E}(R(\tau)\nabla_\theta \log \pi_\theta(\tau))]^2$

The second term in $Var(x)$ and $Var(x')$ are the same, only the first term is different.

$$\mathbb{E}([(R(\tau)-b)\nabla_\theta \log \pi_\theta(\tau)]^2) \cancel{= \mathbb{E}([R(\tau)-b]^2)\,\mathbb{E}([\nabla_\theta \log \pi_\theta(\tau)]^2)} \le \mathbb{E}([R(\tau)\nabla_\theta \log \pi_\theta(\tau)]^2)$$

$$\cancel{\le \mathbb{E}(R(\tau)^2)\,\mathbb{E}([\nabla_\theta \log \pi_\theta(\tau)]^2)}$$

$\Rightarrow Var(x') \le Var(x)$

$$\frac{\partial Var(x')}{\partial b} = \frac{\partial}{\partial b}\left[\mathbb{E}([(R(\tau)-b)\nabla_\theta \log \pi_\theta(\tau)]^2) - [\mathbb{E}(R(\tau)-b)\nabla_\theta \log \pi_\theta(\tau))]^2\right]$$

$$= \frac{\partial}{\partial b}\mathbb{E}([(R(\tau)-b)\nabla_\theta \log \pi_\theta(\tau)]^2)$$

$$= \frac{\partial}{\partial b}\mathbb{E}([R(\tau)-b]^2[\nabla_\theta \log \pi_\theta(\tau)]^2)$$

$$= \frac{\partial}{\partial b}\mathbb{E}([R^2(\tau)-2bR(\tau)+b^2][\nabla_\theta \log \pi_\theta(\tau)]^2)$$

$$= \frac{\partial}{\partial b}\mathbb{E}(R^2(\tau)[\nabla_\theta \log \pi_\theta(\tau)]^2) - 2b\,\mathbb{E}(R(\tau)[\nabla_\theta \log \pi_\theta(\tau)]^2 + b^2\,\mathbb{E}([\nabla_\theta \log \pi_\theta(\tau)$$

$$= \frac{\partial}{\partial b}[-2b\,\mathbb{E}(R(\tau)[\nabla_\theta \log \pi_\theta(\tau)]^2) + b^2\,\mathbb{E}([\nabla_\theta \log \pi_\theta(\tau)]^2)]$$

$$= -2\,\mathbb{E}(R(\tau)[\nabla_\theta \log \pi_\theta(\tau)]^2) + 2b\,\mathbb{E}([\nabla_\theta \log \pi_\theta(\tau)]^2)$$

Set $\frac{\partial Var(x')}{\partial b}$ to 0.

$$\cancel{\Rightarrow \mathbb{E}(R(\tau))\,\mathbb{E}(\nabla_\theta \log \pi_\theta(\tau))^2 = 2b\,\mathbb{E}(\nabla_\theta \log \pi_\theta(\tau))^2 \Rightarrow b = R(\tau)}$$

$$-2\,\mathbb{E}(R(\tau)[\nabla_\theta \log \pi_\theta(\tau)]^2) + 2b\,\mathbb{E}([\nabla_\theta \log \pi_\theta(\tau)]^2) = 0$$

$$\Rightarrow b = \frac{\mathbb{E}[R(\tau)[\nabla_\theta \log \pi_\theta(\tau)]^2)}{\mathbb{E}([\nabla_\theta \log \pi_\theta(\tau)]^2)}$$

Problem 5

The authors show an algorithm to formulate RL as supervised learning problem, called Upside-Down RL. They use the desired reward and desired horizon as input command, the model then learns the actions to achieve that reward in that amount of steps. The training is split into Training and Gathering which update each other. They test the algorithm on several tasks, which shows Upside-Down RL can solve RL problems and has better performance when the reward is sparse.
To make such a input command, need to check each possible reward value and the actions to get the reward. This is fine with small number of possible rewards, e.g. the reward is given at the end of episode and have a few known value. This is is the case of LunarLanderSparse, where UDRL outperforms other algorithm. But a task with continuously valued rewards observed at arbitrary time steps would be difficult for UDRL.