Problem 1

(a) Show that $V_*(s) = \max_a Q_*(s,a)$ ——(1)

$$Q_*(s,a) = R_s^a + \gamma \sum_{s'} P_{ss'}^a V_*(s') \quad \text{——(2)}$$

Pf: For (1): Since $V^\pi(s) = \sum_a \pi(a|s) \cdot Q^\pi(s,a)$, then

$$V_*(s) = \max_\pi V^\pi(s) = \max_\pi \sum_a \pi(a|s) Q^\pi(s,a) \leq \max_\pi \sum_a \pi(a|s) Q_*(s,a)$$

$$\leq \max_a Q_*(s,a) \quad \text{——(1.1)}$$

Next, we prove this by contradiction:

Suppose $V_*(s) < \max_a Q_*(s,a)$ for some state s

Moreover, let π_* be a policy such that $Q^{\pi_*}(s,a) = \max_\pi Q^\pi(s,a) \equiv Q_*(s,a)$

Now, consider one-step policy improvement from π_* and derive an improved policy π':

$$\pi'(s) = \arg\max_a Q^{\pi_*}(s,a), \quad \forall s \quad \text{——(1.2)}$$

Then ..., we have: for any state s,

$$\max_a Q_*(s,a) > V_*(s) = \max_\pi V^\pi(s) \geq V^{\pi'}(s) \quad \text{——(1.3)}$$

$$= Q^{\pi'}(s, \pi'(s))$$

$$= R_s^{\pi'(s)} + \gamma \cdot \sum_{s'} P(s'|s, \pi'(s)) \cdot V^{\pi'}(s')$$

$$\geq R_s^{\pi'(s)} + \gamma \cdot \sum_{s'} P(s'|s, \pi'(s)) \cdot V^{\pi_*}(s')$$

$$= \max_a Q^{\pi_*}(s,a)$$

by (1.1)

by definition of $V_*(s)$

by the fact that π' is a deterministic policy

by Bellman expectation equation

by monotonic policy improvement

by (1.2)

this is a contradiction

(Cont.).

Therefore, we conclude that $V_*(s) = \max_a Q_*(s,a)$.

For (2):

Recall that $Q^\pi(s,a) = R_s^a + \gamma \cdot \sum_{s'} P_{ss'}^a V^\pi(s')$

$$Q_*(s,a) = \max_\pi Q^\pi(s,a)$$

$$= \max_\pi R_s^a + \gamma \cdot \sum_{s'} P_{ss'}^a V^\pi(s')$$

$$\leq R_s^a + \gamma \cdot \sum_{s'} P_{ss'}^a \cdot V_*(s') \quad \text{——————(1.4)}$$

Moreover, a by-product of (1.3) is that by defining a policy π' as:

$$\pi'(s) = \operatorname{argmax}_a Q_*(s,a), \forall s.$$

we have $V_*(s) = V^{\pi'}(s)$, for any s. ——————(1.5)

Therefore,

$$R_s^a + \gamma \cdot \sum_{s'} P_{ss'}^a \cdot V_*(s') \overset{\text{by (1.5)}}{=} R_s^a + \gamma \cdot \sum_{s'} P_{ss'}^a \cdot V^{\pi'}(s)$$

$$= Q^{\pi'}(s,a) \quad \text{(by the definition of } Q^\pi \text{ and } V^\pi\text{)}$$

$$\leq \max_\pi Q^\pi(s,a)$$

$$= Q_*(s,a) \quad \text{(by the definition of } Q_*(s,a)\text{)} \quad \text{——————(1.6)}$$

By (1.4) and (1.6), we conclude that $Q_*(s,a) = R_s^a + \gamma \cdot \sum_{s'} P_{ss'}^a \cdot V_*(s')$

□

(b). Consider the operator T^* as:

$$[T^*Q](s,a) := R_s^a + \gamma \sum_{s'} P_{ss'}^a \cdot \max_{a'} Q(s',a')$$

Show that T^* is a γ-contraction with L_∞-norm.

Pf: Consider two action-value functions Q and $\hat{Q}$:

$$\|T^*Q - T^*\hat{Q}\|_\infty$$

$$= \max_{(s,a)} \left| [T^*Q](s,a) - [T^*\hat{Q}](s,a) \right|$$

$$= \max_{(s,a)} \left| \left(R_s^a + \gamma \sum_{s'} P_{ss'}^a \cdot \max_{a'} Q(s',a') \right) - \left(R_s^a + \gamma \sum_{s'} P_{ss'}^a \cdot \max_{a''} \hat{Q}(s',a'') \right) \right|$$

$$= \max_{(s,a)} \gamma \left| \sum_{s'} P_{ss'}^a \cdot \left(\max_{a'} Q(s',a') - \max_{a''} \hat{Q}(s',a'') \right) \right|$$

$$\leq \max_{(s,a)} \gamma \cdot \sum_{s'} P_{ss'}^a \cdot \left| \max_{a'} Q(s',a') - \max_{a''} \hat{Q}(s',a'') \right|$$

$$\leq \max_{a'} \left| Q(s',a') - \hat{Q}(s',a'') \right|$$

$$\leq \max_{(s,a)} \gamma \sum_{s'} P_{ss'}^a \cdot \max_{a'} \left| Q(s',a') - \hat{Q}(s',a'') \right|$$

$$\leq \gamma \cdot \max_{(s',a')} \left| Q(s',a') - \hat{Q}(s',a'') \right| = \gamma \cdot \|Q - \hat{Q}\|_\infty .$$

Hence, T^* is indeed a γ-contraction operator with L_∞-norm. □

Problem 2

(a). (i) Show that $d_p(aU, aV) = |a| \cdot d_p(U,V)$, for any $a \in \mathbb{R}$.

We start by showing that $d_p(aU, aV) \le |a| \cdot d_p(U,V)$.

Let U and V be two random variables with marginal CDFs F_U, F_V

For ease of exposition, we assume that the density functions of U and V exist and are denoted by f_U and f_V.

(In fact, the discrete cases are more straightforward and omitted here)

For any joint density $f_{U,V}$ of U and V, for any $a \neq 0$, by Lemma 1 we know the joint density of $f_{aU,aV}$ can be derived as

$$f_{aU,aV}(u,v) = \frac{1}{a^2} f_{U,V}\left(\begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \end{bmatrix} \right)$$

$$= \frac{1}{a^2} f_{UV}\left(\frac{u}{a}, \frac{v}{a}\right)$$

Therefore, under the joint density $f_{U,V}$, we have

$$\|aU - aV\|_p^p = \int |u - v|^p \cdot f_{aU,aV}(u,v)\, du\, dv$$

$$= \int |u - v|^p \cdot \left(\frac{1}{a^2} f_{UV}\left(\frac{u}{a}, \frac{v}{a}\right)\right) du\, dv$$

$$= \int |u - v|^p \cdot f_{UV}\left(\frac{u}{a}, \frac{v}{a}\right) \cdot d\left(\frac{u}{a}\right) \cdot d\left(\frac{v}{a}\right)$$

$$= |a|^p \cdot \int |u' - v'|^p \cdot f_{UV}(u', v')\, du'\, dv'$$

(change of variables: let $u' = \frac{u}{a}$, $v' = \frac{v}{a}$)

$$= |a|^p \cdot \|U - V\|_p^p$$

This implies that $\|aU - aV\|_p = |a| \cdot \|U - V\|_p$.

> **Lemma 1:** Let X_1, X_2, Y_1, Y_2 be random variables that satisfy
> $Y_1 = aX_1 + bX_2$, $Y_2 = cX_1 + dX_2$.
> Let $\Sigma = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Then, we have
> $$f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{|\det(\Sigma)|} f_{X_1,X_2}\left(\Sigma^{-1}[y_1,y_2]^T\right)$$

(Cont.)

Therefore, under any joint density $f_{U,V}$, we have $\|aU-aV\|_p = |a|\cdot\|U-V\|_p$. ——(*)

Then, by the definition of Wasserstein metric,

$$d_p(aU,aV) = \inf_{\text{all possible joint of } aU,aV} \|aU-aV\|_p \leq \inf_{\text{all possible joint of } U,V} \|aU-aV\|_p$$

(infimum over a smaller set)

$$= \inf_{\text{all possible joint of } U,V} |a|\cdot\|U-V\|_p \quad \cdots\cdots \text{by (*)}$$

$$= |a|\cdot d_p(U,V). \quad \cdots\cdots \text{by the definition of } d_p(U,V).$$

Next, we can show $d_p(aU,aV) \geq |a|\cdot d_p(U,V)$ by reusing the above result.

Specifically, $d_p(U,V) = d_p\left(\frac{1}{a}(aU), \frac{1}{a}(aV)\right) \leq \left|\frac{1}{a}\right|\cdot d_p(aU,aV)$, for any $a\neq 0$.

Finally, it is straightforward to see that if $a=0$, then we have $d_p(aU,aV) = 0 = 0\cdot d_p(U,V)$

Hence, we conclude that $d_p(aU,aV) = |a|\cdot d_p(U,V)$, for any $a\in\mathbb{R}$.

□

(ii) Show that $d_p(A+U, A+V) \le d_p(U,V)$, where A is a random variable independent of U and V.

Similar to (i), we start by considering all possible joint distributions of U and V.

Under any joint distribution F_{UV} of U and V, since A is independent of U, V, then the joint distribution of $(A+U), (A+V)$ is well-defined.

Then, we can have $\|(A+U)-(A+V)\|_p = \|U-V\|_p$.

Therefore, we have

$$d_p(A+U, A+V) = \inf_{\substack{\text{all possible} \\ \text{joint of } (A+U) \text{ and } (A+V)}} \|(A+U)-(A+V)\|_p \le \inf_{\substack{\text{all possible} \\ \text{joint of } U \text{ and } V}} \|(A+U)-(A+V)\|_p$$

$$= \inf_{\substack{\text{all possible} \\ \text{joint of } U \text{ and } V}} \|U-V\|_p$$

$$= d_p(U,V).$$

□

(iii) Show that $d_p(QU, QV) \le q \cdot d_p(U,V)$, where $Q \sim \text{Bernoulli}(q)$ and is independent of U, V.

Similar to (ii), we shall consider all possible joint distributions of U and V.

Under any joint distribution F_{UV} of U and V, since Q is independent of U, V, then the joint distribution of QU, QV is also well-defined.

Then, it is easy to verify that

$$\| QU - QV \|_p = q \cdot \| U - V \|_p .$$

Therefre, we have

$$d_p(QU, QV) = \inf_{\substack{\text{all possible} \\ \text{joint of } QU \text{ and } QV}} \| QU - QV \|_p \le \inf_{\substack{\text{all possible} \\ \text{joint of } U,V}} \| QU - QV \|_p$$

$$= \inf_{\substack{\text{all possible} \\ \text{joint of } U,V}} q \cdot \| U - V \|_p$$

$$= q \cdot d_p(U,V).$$

□

(b) Consider $Z_1, Z_2 \in \mathcal{Z}$

By definition, we know

$$\bar{d}_p(B^\pi Z_1, B^\pi Z_2) = \sup_{s,a} d_p(B^\pi Z_1(s,a), B^\pi Z_2(s,a)).$$

Next, we have

$$d_p(B^\pi Z_1(s,a), B^\pi Z_2(s,a))$$

$$= d_p(r(s,a) + \gamma P^\pi Z_1(s,a), r(s,a) + \gamma P^\pi Z_2(s,a)) \quad \text{.... by the definition of } B^\pi$$

$$\leq d_p(\gamma P^\pi Z_1(s,a), \gamma P^\pi Z_2(s,a)) \quad \text{..... by Problem 2(a) and that } r(s,a) \text{ is independent from } \gamma P^\pi Z_1, \gamma P^\pi Z_2.$$

$$= \gamma \cdot d_p(P^\pi Z_1(s,a), P^\pi Z_2(s,a)) \quad \text{— (**)} \quad \text{.... by Problem 2(a)}$$

Consider a random experiment in which a state-action pair (s',a') is drawn randomly based on P^π given that the current state-action pair is (s,a).

For each $(\bar{s},\bar{a})$, define a random variable $A_{\bar{s},\bar{a}} = \begin{cases} 1, & \text{if } s'=\bar{s},\ a'=\bar{a} \\ 0, & \text{otherwise.} \end{cases}$

Then, by the Partition lemma, (*) can be further bounded as

Hence, $\{A_{\bar{s},\bar{a}}\}$ form a partition of the sample space

$$(**) \leq \gamma \sum_{\bar{s},\bar{a}} d_p\Big(\underbrace{A_{\bar{s},\bar{a}}(P^\pi Z_1(s,a))}_{A_{\bar{s},\bar{a}} \cdot Z_1(\bar{s},\bar{a})}, \underbrace{A_{\bar{s},\bar{a}}(P^\pi Z_2(s,a))}_{A_{\bar{s},\bar{a}} \cdot Z_2(\bar{s},\bar{a})}\Big)$$

$$\leq \gamma \cdot \sum_{\bar{s},\bar{a}} P(s'=\bar{s}, a'=\bar{a} \mid s,a) \cdot d_p(Z_1(\bar{s},\bar{a}), Z_2(\bar{s},\bar{a}))$$

$$\leq \gamma \cdot \sup_{\bar{s},\bar{a}} d_p(Z_1(\bar{s},\bar{a}), Z_2(\bar{s},\bar{a}))$$

□

(Cont.).

Therefore, by putting everything together, we have

$$\begin{aligned}
\overline{d}_p(B^\pi Z_1, B^\pi Z_2) &= \sup_{s,a} d_p(B^\pi Z_1(s,a), B^\pi Z_2(s,a)) \\
&\leq \sup_{s,a} \left(\gamma \cdot \sup_{\bar{s},\bar{a}} d_p(Z_1(\bar{s},\bar{a}), Z_2(\bar{s},\bar{a})) \right) \\
&= \gamma \cdot \sup_{s,a} d_p(Z_1(s,a), Z_2(s,a)) \\
&= \gamma \cdot \overline{d}_p(Z_1, Z_2).
\end{aligned}$$

□