A2

1. DQN

1.1

(a)

1. By approximating the Q-value using deep learning, DQN can efficiently handle large state spaces.
2. Since Q-learning with VFA can be prone to divergence, DQN implements the Experience Replay method, which breaks the correlation between samples and enhances learning efficiency.
3. DQN also differs from Q-learning by employing a seperate target network. When computing target Q-value for gradient descent, DQN uses fixed Q network to handle learning stability.

(b)

I think that the component that contributes most to performance gain is D, the replay memory. Neural Networks assume that the training data is independent and identically distributed (i.i.d.). However, in reinforcement learning (or MDP setting), samples are collected sequentially and are therefore highly correlated. The replay memory D solves this problem by randomly sampling mini-batches from its buffer, breaking temporal correlations. Making the training data closer to the i.i.d. assumption is critical for stabilizing the training of the neural network and thus the agent's learning process.

(c)

Since simple Atari games like Pong have short playtime, the target network will remain fixed with its initial random values for the entire training process. It will lead to a meaningless process, because the main Q-network will try to learn a target based on incorrect random values. Consequently, the Agent will not be able to find an optimal policy.

2. Policy Gradient

2.1 ~ 2.6 : Coding

2.7

(a)

Since we can express return G_t by recursive relationship,

$G_t = r_t + \gamma G_{t+1},\ G_t = r_t + \gamma r_{t+1} + \ldots + \gamma^{T-t} r_T$

we can use a backward pass starting form the end of the episode (T).

(b)

$L_t^{CLIP} = \min(r_t(\theta)A_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)$

1. $1 - \epsilon \leq r_t(\theta) \leq 1 + \epsilon$ (Non clipped case)

   $L_t^{CLIP} \propto r_t(\theta)A_t$

   $\nabla_\theta L \propto \nabla_\theta r_t(\theta) \cdot A_t$

   - Since $\nabla_\theta r_t(\theta) \neq 0$, the only case gradient of the loss function has zero value is $A_t = 0$

   In this case, the action has average value and the objection function of this sample will be zero. Since there is no error (this action is perfectly on average) we don't have to change distribution of this action.

2. Suppose that our objective functions has clipped.

   $L_t^{CLIP} \propto (1 - \epsilon \text{ or } 1 + \epsilon)A_t$

   This implies that our new policy has changed too much from original policy $(r_t(\theta) \geq 1 + \epsilon)$ and advantage function A_t has a positive value (which makes $(1 + \epsilon)A_t$ smaller than $r_t(\theta)A_t$), vice versa.

   $\nabla_\theta L \propto 0$

   Since it exceeded the trusted region, PPO doesn't updates this action.

(c)

Since REINFORCE is an on-policy algorithm, it doesn't use the same batch of samples for multiple updates. Therefore, we can simply compute log-probability using the current policy.

However, PPO imporves sample efficiency by performing off-policy updates. It updates its policy for several epochs using the same batch of samples. Since our policy updates for every epoch, we need to cache the log-probability of the original policy that sampled these actions. If the log-probability had not been collected during the rollout, we can pre-cache it by passing samples to our

original policy before applying PPO updates. Or, we could store and maintain our original policy and use it whenever we need its log-probability.

3. Distributions induced by a policy

(a)

$$\rho^{\pi}(\tau) = \prod_{t=0}^{\infty} P(s_{t+1}|s_t, a_t) \cdot \pi(a_t|s_t)$$

(b)

$p^{\pi}(s_t = s) = \sum_{s'} p^{\pi}(s_t = s') \sum_a \pi(a|s') P(s|s', a)$

(c) Prove the following identity :

$\mathbb{E}_{\tau \sim \rho^{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)\right] = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi}}\left[\mathbb{E}_{a \sim \pi(s)}[f(s, a)]\right]$

$\mathbb{E}_{\tau \sim \rho^{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)\right] = \sum_{t=0}^{\infty} \gamma^t \left[\mathbb{E}_{\tau \sim \rho^{\pi}}[f(s_t, a_t)]\right]$

$= \sum_{t=0}^{\infty} \gamma^t \Big[\sum_{s \in S} \sum_{a \in A} p^{\pi}(s_t = s) \pi(a|s) f(s, a)\Big]$

$= \sum_{s \in S} \sum_{a \in A} \Big[\underbrace{\Big(\sum_{t=0}^{\infty} \gamma^t p^{\pi}(s_t = s)\Big)}_{\frac{d^{\pi}(s)}{1-\gamma}} \pi(a|s) f(s, a)\Big]$

$= \frac{1}{1-\gamma} \sum_{s \in S} \sum_{a \in A} d^{\pi}(s) \pi(a|s) f(s, a)$

$= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi}}\left[\mathbb{E}_{a \sim \pi(s)}[f(s, a)]\right]$

(d) Prove that the following statement holds for all policies

$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{(1-\gamma)} \mathbb{E}_{s \sim d^{\pi}}\left[\mathbb{E}_{a \sim \pi(s)}\left[A^{\pi'}(s, a)\right]\right]$

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

$V^{\pi}(s) - V^{\pi'}(s) = \mathbb{E}_{a \sim \pi(s)}\left[r(s, a) + \gamma \mathbb{E}_{s'|s,a} V^{\pi}(s')\right] - V^{\pi'}(s)$

$$= \mathbb{E}_{a\sim\pi(s)}\Big[Q^{\pi'}(s,a) - V^{\pi'}(s)\Big] + \mathbb{E}_{a\sim\pi(s)}\Big[r(s,a) + \gamma\mathbb{E}_{s'|s,a}V^{\pi}(s') - Q^{\pi'}(s,a)\Big]$$

- $Q^{\pi'}(s,a) = r(s,a) + \gamma\mathbb{E}_{s'|s,a}V^{\pi'}(s')$

$$= \mathbb{E}_{a\sim\pi(s)}\Big[A^{\pi'}(s,a)\Big] + \mathbb{E}_{a\sim\pi(s)}\Big[\gamma\mathbb{E}_{s'|s,a}\Big[V^{\pi}(s') - V^{\pi'}(s')\Big]\Big]$$

$$= \mathbb{E}_{a\sim\pi(s)}\Big[A^{\pi'}(s,a)\Big] + \gamma\mathbb{E}_{(a\sim\pi(s),s'|s,a)}\Big[V^{\pi}(s') - V^{\pi'}(s')\Big] = \Delta(s)$$

Then

$$\Delta(s) = \mathbb{E}_{a\sim\pi(s)}\Big[A^{\pi'}(s,a)\Big] + \gamma\mathbb{E}_{(a\sim\pi(s),s'|s,a)}\Big[\Delta(s')\Big]$$

Set $s = s_0$

$$\Delta(s_0) = \mathbb{E}_{a\sim\pi(s_0)}\Big[A^{\pi'}(s_0,a)\Big] + \gamma\mathbb{E}_{(a\sim\pi(s_0),s_1|s_0,a)}\Big[\mathbb{E}_{a\sim\pi(s_1)}\Big[A^{\pi'}(s_1,a)\Big] + \gamma\mathbb{E}_{(a\sim\pi(s_1),s_2|s_1,a)}\Big[\Delta(s_2)\Big]\Big]$$

$$\vdots$$

$$= \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t A^{\pi'}(s_t,a_t)|s_0\Big]$$

- Expected value of [discounted sum of advantage function of π'] following π, starting from s_0

$$= \sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{\tau\sim\pi}\Big[A^{\pi'}(s_t,a_t)\Big]$$

$$= \sum_{t=0}^{\infty}\gamma^t\Big[\textstyle\sum_s p^{\pi}(s_t=s)\mathbb{E}_{a\sim\pi(s)}[A^{\pi'}(s,a)]\Big]$$

$$= \textstyle\sum_s\sum_{t=0}^{\infty}\gamma^t p^{\pi}(s_t=s)\mathbb{E}_{a\sim\pi(s)}\Big[A^{\pi'}(s,a)\Big]$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi},\, a\sim\pi(s)}\Big[A^{\pi'}(s,a)\Big]$$