Q.1 a) in online Q learning, all $Q(s,a)$ pairs are parameterized by same weight parameters θ. So upod updating θ, will change $Q(s,a)$ of all (s,a) pairs and won't retain the old value.

In watkins Q learning, all (s,a) pairs have their seperate $Q(s,a)$ values. So updating value of one (s,a) pair doesn't affect the Q value of another pair (s',a').

b). ~~Again, the update will suffer~~

Since the replay buffer here is static. The stochastic batch upate will make the approximator Q_ϕ to only learn Q values corresponding to the replay buffer we have. As Q_ϕ keeps updating, we need to collect more replay samples using current Q function to help the agent learn better and converge to the optimal state-value function.

Since after each update, policy changes, one need to collect samples from updated policy not old one.

Also, if target networks are not used, the network will further face convergence issues.

Q.3(d) $\nabla_\theta J(\theta) = E_{\pi_\theta}\left[\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t)\, \psi_t\right]$

$\psi_t = G_{t:\infty} - b(s_t)$

$\nabla_\theta J(\theta) = E_{\pi_\theta}\left[\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t)\, (G_{t:\infty} - b(s_t))\right]$

writing in terms of trajectories τ

$\nabla_\theta J(\theta) = \underset{\tau\sim\pi_\theta(\tau)}{E}\left[\nabla_\theta \log \pi_\theta(\tau)\,(G_{t:\infty}^{(\tau)} - b)\right]$

Variance (Var) $= \underset{\tau\sim\pi_\theta(\tau)}{E}\left[(\nabla_\theta \log \pi_\theta(\tau)\,(G_{t:\infty}^{(\tau)} - b))^2\right]$

$- \underset{\tau\sim\pi_\theta(\tau)}{E}\left[\nabla_\theta \log \pi_\theta(\tau)\,(G_{t:\infty}^{(\tau)} - b)\right]^2$

baselines are unbiased so it is just $\underset{\tau\sim\pi_\theta(\tau)}{E}[\nabla_\theta \log \pi_\theta(\tau)\, G_{t:\infty}]$

$Var = \underset{\tau\sim\pi_\theta(\tau)}{E}\left[(\nabla_\theta \log \pi_\theta(\tau)\,(G_{t:\infty}^{(\tau)} - b))^2\right] - \underset{\tau\sim\pi_\theta(\tau)}{E}\left[\nabla_\theta \log \pi_\theta(\tau)\, G_{t:\infty}\right]$

$\frac{d(Var)}{db} = -2\underset{\tau\sim\pi_\theta(\tau)}{E}\left[\nabla_\theta \log \pi_\theta^2(\tau)\, G_{t:\infty}^{(\tau)}\right] + 2b \underset{\tau\sim\pi_\theta(\tau)}{E}\left[\nabla_\theta \log \pi_\theta^2(\tau)\right]$

$\frac{d(Var)}{db} = 0 \Rightarrow b = \dfrac{\underset{\tau\sim\pi_\theta(\tau)}{E}\left[\nabla_\theta \log^2 \pi_\theta(\tau)\, G_{t:\infty}^{(\tau)}\right]}{\underset{\tau\sim\pi_\theta(\tau)}{E}\left[\nabla_\theta \log \pi_\theta^2(\tau)\right]}$