Ex 12-1

(a) ① Too many states/actions stored in memory.

② learning the utility value of each state individually would be too slow.

③ cannot represent continuous state space

④ impossible for unseen states (generalize failure)

(b) $$J(w) = \mathbb{E}_\pi\left[\left(U_\pi(s) - \hat{U}(s,w)\right)^2\right]$$

$$= \sum_{s\in S} \mu(s)\left[U_\pi(s) - \hat{U}(s,w)\right]^2$$

$\sum_{s\in S} \mu(s) = 1$, probability density function or weighting

Usually $\mu(s)$ = fraction of time spent

(c) $$\nabla J(w) = -2 \sum_{s\in S} \mu(s)\left[U_\pi(s) - \hat{U}(s,w)\right] \nabla_w \hat{U}(s,w)$$

(d) $$w \leftarrow w + \Delta w = w + \left(-\tfrac{1}{2}\alpha \nabla_w J(w)\right)$$

$$= w + \alpha \sum_{s\in S} \mu(s)\left[U_\pi(s) - \hat{U}(s,w)\right] \nabla_w \hat{U}(s,w)$$

SGD: $$w \leftarrow w + \alpha\left[U_\pi(s) - \hat{U}(s,w)\right] \nabla_w \hat{U}(s,w)$$

(e) $$\nabla_w \hat{U}(s,w) = \nabla_w\left[x(s)^T w\right] = x(s)$$

(f) MC: $\Delta w = \alpha\left(G_t - \hat{U}(s,w)\right) \nabla_w \hat{U}(s,w)$ $\quad G_t = \sum_{i=1}^{n} \gamma^{i-1} R_{t+i}$

TD(0): $\Delta w = \alpha\left(R_{t+1} + \gamma \hat{U}(s_{t+1},w) - \hat{U}(s,w)\right) \nabla_w \hat{U}(s,w)$

(g) $\nabla_w$ only computes $\hat{U}(s,w)$ estimate, not the target.

(b) only the last transition from 7 to it self decrease $w_8$ but not as much as other transitions increased it.

(c) diverge to infinity.

(d) on-policy. transition (1-6) → 7 only once. target policy $\pi$ remains in state 7.