Challenge Question Solution

Recall in 6.1, we compared TD target and MC target, where the MC target G_t is the actual return following time t, and the TD target being $R_{t+1} + \gamma V(S_{t+1})$, in which R_{t+1} is the observed reward at time t and $V(S_{t+1})$ is our value estimate of S_{t+1}.

Given $S_t = s$, and S_{t+1}, we can rewrite G_t as $R_{t+1} + \gamma G_{t+1}$.

Therefore,

$$E[G_t \mid S_t = s, S_{t+1}]$$

$$= E[R_{t+1} + \gamma G_{t+1} \mid S_t = s, S_{t+1}]$$

Since reward R_{t+1} is observed and γ is a constant discount rate,

$$= R_{t+1} + \gamma E[G_{t+1} \mid S_{t+1}] \quad \text{under the markov assumption}$$

Recall the definition of the value function in 3.5, $v_\pi(s) = E_\pi[G_t \mid S_t = s]$

$$= R_{t+1} + \gamma v_\pi(S_{t+1})$$

Because we assume in the question that we have found a V that exactly equals v_π, we have

$$E[G_t \mid S_t = s, S_{t+1}] = R_{t+1} + \gamma V(S_{t+1}) \quad \text{(TD target)}$$

Provided the variance decomposition equation of G_t, we have

$$\text{Var}(G_t \mid S_{t=s}) = E[\text{Var}(G_t \mid S_t = s, S_{t+1})] + \text{Var}(E[G_t \mid S_t = s, S_{t+1}] \mid S_t = s)$$

MC Target

$$= \underbrace{E[\text{Var}(G_t \mid S_{t=s}, S_{t+1})]}_{\geq 0} + \text{Var}(\underbrace{R_{t+1} + \gamma V(S_{t+1})}_{\text{TD target}} \mid S_t = s)$$

Since the variance of G_t is always greater or equal to 0, so does its expectation. Hence, we've shown that the Monte Carlo target has at least as high of variance as the TD target, whose variance is just the second term in our final result.