(1A).

(a): $d^{\pi_\theta}(s) = \sum_{k=0}^{\infty} Pr\{s_0 \to s, k, \pi_\theta\}$.

$Q^{\pi_\theta}(s,a)$ = total rewards when action a is taken in state s and $\pi_\theta$ is followed subsequently.

(b). Both are fine as:

$\max_a Q^{\pi}(s,a)$ & $\sum_b \frac{Q^{\pi_\theta}(s,b)}{|A(s)|}$ do not depend on given action $\underline{a}$.

(2A). (a): $0.5 \cdot 0.5 \cdot 0.2 \left(\frac{1}{0.5 \cdot 0.5}\right)(1) + 0.5 \cdot 0.5 \cdot 0.8 \cdot 0.5 \cdot 0.5 \cdot 0.2 \left(\frac{1}{(0.5)^4}\right) + \cdots$

$= 0.2 \left[\sum_{t=0}^{\infty} (0.8)^t\right] = 1.$

(b). Variance $= 0.2 \sum_{t=0}^{\infty} (0.8)^t \, 2^{2t+2} = \infty$

(c). Weighted I.S works well.

Reason: Consider trajectories $\tau_1, \tau_2, \ldots \tau_n$ and the objective is to estimate $V^{\pi}(s_1)$. Let $I \subset \{1, 2, \ldots\}$ & $G_I = 1$.

using weighted I.S: $V^{\pi}(s_1) = \frac{\sum_{i \in I} \rho_i}{\sum_{i \in I} \rho_i} = 1$. exactly $V^{\pi}(s_1)$.

using ordinary I.S: $V^{\pi}(s_1) = \frac{\sum_{i \in I} \rho_i}{n}$, need not be 1.

(3A). (a). In Monte Carlo, Value function is updated after entire trajectory is generated. Whereas, in TD(0), update is done sample-by-sample.

As a result, MC has large variance, but less bias. If there is a very large availability of data (trajectories), one can prefer MC over TD(0).

(b). For MC, $V(B) = 1$ & $V(A) = 0$. [Took first visit MC. Every visit MC is also fine!].

$TD(0)$: $V(B) = (0.5)(0) + 0.5(1) = 0.5$

$V(B) = (0.5)(0.5) + 0.5(1) = 0.75$

$V(B) = (0.5)(0.75) + 0.5(1) = 0.875$

Note: Took $\gamma = 1$. Any other value of $\gamma$ is also fine!

$V(A) = 0.5(0) + 0.5[0 + 0.875]$
$= 0.4375$

$V(B) = (0.5)(0.875) + 0.5(0+0) = 0.4375$

(4A) Refer textbook.

(5A). (a). Let 'w' be the parameter.

For the sample $(s_t, a_t, r_t, s_{t+1})$ we have:

$$w_{t+1} \doteq w_t + \alpha[r_t + \gamma w_t^T x(s_{t+1}) - w_t^T x(s_t)]\, x(s_t).$$

(b). First, let us compute stationary distribution:

$$[\pi_1 \quad \pi_2] \begin{bmatrix} 0.7 & 0.3 \\ 1 & 0 \end{bmatrix} = [\pi_1 \quad \pi_2]$$

$0.3\pi_1 = \pi_2$ & $\pi_1 + \pi_2 = 1$

$\Rightarrow 1.3\pi_1 = 1 \Rightarrow \pi_1 = \frac{10}{13}$ & $\pi_2 = \frac{3}{13}$.

Now, $b = \frac{10}{13}\left[0.7\times5 + 0.3\times3\right]\begin{bmatrix}1\\0\end{bmatrix} +$

$\frac{3}{13}[3]\begin{bmatrix}1\\1\end{bmatrix}$

$= \begin{bmatrix}53/13\\9/13\end{bmatrix}$

&

$A = \frac{10}{13}\left[0.3\left[\begin{bmatrix}1\\0\end{bmatrix}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.5\\0.5\end{bmatrix}\right)^T\right] + 0.7\left[\begin{bmatrix}1\\0\end{bmatrix}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0.5\\0\end{bmatrix}\right)^T\right]\right] +$

$\frac{3}{13}\left[\begin{bmatrix}1\\1\end{bmatrix}\left(\begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix}0.5\\0\end{bmatrix}\right)^T\right]$

$= \frac{10}{13}\left[0.3\begin{bmatrix}0.5 & -0.5\\0 & 0\end{bmatrix} + 0.7\begin{bmatrix}0.5 & 0\\0 & 0\end{bmatrix}\right] +$

$\frac{3}{13}\begin{bmatrix}0.5 & 1\\0.5 & 1\end{bmatrix}$

$= \frac{10}{13}\begin{bmatrix}0.5 & -0.15\\0 & 0\end{bmatrix} + \frac{3}{13}\begin{bmatrix}0.5 & 1\\0.5 & 1\end{bmatrix}$

$= \begin{bmatrix}0.5 & 1.5/13\\1.5/13 & 3/13\end{bmatrix}$

$\therefore$ The point of convergence is $A^{-1}b$.

$\approx \begin{bmatrix}2.26 & -1.12\\-1.12 & 4.89\end{bmatrix}\begin{bmatrix}53/13\\9/13\end{bmatrix}$

$$\simeq \begin{bmatrix} 8.434 \\ -1.184 \end{bmatrix}$$