order of questions are different for different students. Hope you will get the context.

(1 A). Policy gradient rule allows us to sample efficiently because

$$\nabla J(\theta) = \sum \underbrace{d_{\pi}(\theta)} \sum \nabla \pi_{\theta}(s,a)\, Q(s,a)$$

↳ If it was $\underline{\nabla d_{\pi}(\theta)}$, then we couldn't have sampled!

(2 A). Baseline is used to reduce the variance as discussed in class

(3 A). (a). Semigradient TD(0) does converge! we have nice proofs.

(b). Gradient MC will converge but to an approximated solution!

(4 A).

$$\theta_1 = \theta_0 + \alpha\, G\, \nabla \log \pi(s,a).$$

$$= \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 0.5 \cdot (1) \cdot \Big[\underbrace{\phi(1,a) - \sum_{x} \pi(1,x)\, \phi(1,x).}_{\text{this is the gradient.}} \Big]$$

Note $\pi_0(1,a) = \dfrac{e^0}{e^0 + e^0} = 1/2$

similarly $\pi_0(1,b) = \dfrac{e^0}{e^0 + e^0} = 1/2$

$$\theta_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 0.5 \cdot 1 \cdot \left[\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix} \right]$$

$$= \begin{bmatrix} 1/4 \\ -1/4 \end{bmatrix}.$$

(5 A).

$$v^{\pi}(1) = 0 + 0.5\, v^{\pi}(2). \qquad v^{\pi}(2) = 1 + 0.5\, v^{\pi}(1).$$

$\Rightarrow$ $v^{\pi}(2) = 4/3$ & $v^{\pi}(1) = 2/3$.

$\therefore \quad \theta_1 = 2/3$ and $\theta_1 + \theta_2 = 4/3$

$\Rightarrow \theta_2 = 2/3.$

NOTE: Correct answer is not given in options. My mistake. Will add extra mark for those who attempted this question. Apologies for inconvinience.