Problem 1:

(a). There are only three possible scenarios for $\hat{\nabla} V$:

Scenario #1: $S_0 = s,\ a_1 = a,\ S_1 = \text{terminal}$ (this scenario happens with prob. $\frac{1}{10}$)

It is easy to verify that
$$\frac{\partial \log \pi_\theta(a|s)}{\partial \theta_a} = 1 - \pi_\theta(a|s)$$
$$\frac{\partial \log \pi_\theta(a|s)}{\partial \theta_b} = -\pi_\theta(b|s)$$
$$\frac{\partial \log \pi_\theta(a|s)}{\partial \theta_c} = -\pi_\theta(c|s)$$

Therefore, we have
$$\hat{\nabla} V = \begin{bmatrix} 1-\pi_\theta(a|s) \\ -\pi_\theta(b|s) \\ -\pi_\theta(c|s) \end{bmatrix} \times r(s,a) = 100 \times \begin{bmatrix} \frac{9}{10} \\ -\frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}$$

Scenario #2: $S_0 = s,\ a_1 = b,\ S_1 = \text{terminal}$ (this scenario happens with prob. $\frac{5}{10}$)

Similar to scenario #1, we have
$$\hat{\nabla} V = \begin{bmatrix} -\pi_\theta(a|s) \\ 1-\pi_\theta(b|s) \\ -\pi_\theta(c|s) \end{bmatrix} \times r(s,b) = 98 \times \begin{bmatrix} -\frac{1}{10} \\ \frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}$$

Scenario #3: $S_0 = s,\ a_1 = c,\ S_1 = \text{terminal}$ (this scenario happens with prob. $\frac{4}{10}$)
$$\hat{\nabla} V = \begin{bmatrix} -\pi_\theta(a|s) \\ -\pi_\theta(b|s) \\ 1-\pi_\theta(c|s) \end{bmatrix} \times r(s,c) = 95 \times \begin{bmatrix} -\frac{1}{10} \\ -\frac{5}{10} \\ \frac{6}{10} \end{bmatrix}$$

(Conti).

Therefore, we know

$$E[\hat{\nabla} V] = \frac{1}{10} \times \left(100 \times \begin{bmatrix} \frac{9}{10} \\ -\frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}\right) + \frac{5}{10} \times \left(98 \times \begin{bmatrix} -\frac{1}{10} \\ \frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}\right) + \frac{4}{10} \times \left(95 \times \begin{bmatrix} -\frac{1}{10} \\ -\frac{5}{10} \\ \frac{6}{10} \end{bmatrix}\right)$$

$$= \begin{bmatrix} \frac{30}{100} \\ \frac{50}{100} \\ \frac{-80}{100} \end{bmatrix} = \begin{bmatrix} \frac{3}{10} \\ \frac{5}{10} \\ -\frac{8}{10} \end{bmatrix}$$ ← This is actually the true policy gradient $\nabla_\theta V^{\pi_\theta}$

By definition, we have

$$Cov(\hat{\nabla} V, \hat{\nabla} V) = E[\hat{\nabla} V \hat{\nabla} V^T] - (E[\hat{\nabla} V]) \cdot (E[\hat{\nabla} V])^T = \begin{bmatrix} 894.03 & -509.75 & -384.28 \\ -509.75 & 2352.75 & -1843 \\ -384.28 & -1843 & 2227.28 \end{bmatrix}$$

- $$E[\hat{\nabla} V \hat{\nabla} V^T] = \frac{1}{10} \times \left(100 \times \begin{bmatrix} \frac{9}{10} \\ -\frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}\right) \cdot \left(100 \times \begin{bmatrix} \frac{9}{10} \\ -\frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}\right)^T$$

$$+ \frac{5}{10} \times \left(98 \times \begin{bmatrix} -\frac{1}{10} \\ \frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}\right) \cdot \left(98 \times \begin{bmatrix} -\frac{1}{10} \\ \frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}\right)^T$$

$$+ \frac{4}{10} \times \left(95 \times \begin{bmatrix} -\frac{1}{10} \\ -\frac{5}{10} \\ \frac{6}{10} \end{bmatrix}\right) \cdot \left(95 \times \begin{bmatrix} -\frac{1}{10} \\ -\frac{5}{10} \\ \frac{6}{10} \end{bmatrix}\right)^T$$

$$= \begin{bmatrix} 894.12 & -509.6 & -384.52 \\ -509.6 & 2353 & -1843.4 \\ -384.52 & -1843.4 & 2227.92 \end{bmatrix}.$$

- $$(E[\hat{\nabla} V])(E[\hat{\nabla} V])^T = \begin{bmatrix} 0.09 & 0.15 & -0.24 \\ 0.15 & 0.25 & -0.40 \\ -0.24 & -0.40 & 0.64 \end{bmatrix}$$

□

(b). $V^{\pi_\theta}(s) = \pi_\theta(a|s)\cdot r(s,a) + \pi_\theta(b|s)\cdot r(s,b) + \pi_\theta(c|s)\cdot r(s,c)$

$$= \frac{1}{10}\times 100 + \frac{5}{10}\times 98 + \frac{4}{10}\times 95$$

$$= 97$$

Then, we can get the estimated policy gradient with baseline for the three scenarios:

Scenario#1: $$\tilde{\nabla} V = \begin{bmatrix} 1-\pi_\theta(a|s) \\ -\pi_\theta(b|s) \\ -\pi_\theta(c|s) \end{bmatrix} \times \left(r(s,a) - V^{\pi_\theta}(s) \right) = 3\times \begin{bmatrix} \frac{9}{10} \\ -\frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}$$

Scenario#2: $$\tilde{\nabla} V = \begin{bmatrix} -\pi_\theta(a|s) \\ 1-\pi_\theta(b|s) \\ -\pi_\theta(c|s) \end{bmatrix} \times \left(r(s,b) - V^{\pi_\theta}(s) \right) = 1\times \begin{bmatrix} -\frac{1}{10} \\ \frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}$$

Scenario#3: $$\tilde{\nabla} V = \begin{bmatrix} -\pi_\theta(a|s) \\ -\pi_\theta(b|s) \\ 1-\pi_\theta(c|s) \end{bmatrix} \times \left(r(s,c) - V^{\pi_\theta}(s) \right) = -2\times \begin{bmatrix} -\frac{1}{10} \\ -\frac{5}{10} \\ \frac{6}{10} \end{bmatrix}$$

Then, similar to Problem 1(a), we have

$$E[\tilde{\nabla} V] = \frac{1}{10}\times\left(3\times\begin{bmatrix} \frac{9}{10} \\ -\frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}\right) + \frac{5}{10}\times\left(1\times\begin{bmatrix} -\frac{1}{10} \\ \frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}\right) + \frac{4}{10}\times\left(-2\times\begin{bmatrix} -\frac{1}{10} \\ -\frac{5}{10} \\ \frac{6}{10} \end{bmatrix}\right)$$

$$= \begin{bmatrix} \frac{3}{10} \\ \frac{5}{10} \\ -\frac{8}{10} \end{bmatrix}$$

(Conti.)

$$\mathrm{Cov}(\tilde{\nabla}V, \tilde{\nabla}V) = E[\tilde{\nabla}V \tilde{\nabla}V^T] - (E[\tilde{\nabla}V])\cdot(E[\tilde{\nabla}V])^T = \begin{bmatrix} 0.66 & -0.50 & -0.16 \\ -0.50 & 0.50 & 0 \\ -0.16 & 0 & 0.16 \end{bmatrix}.$$

P.4

- $$E[\tilde{\nabla}V \tilde{\nabla}V^T] = \frac{1}{10} \times \left(3 \times \begin{bmatrix} \frac{9}{10} \\ -\frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}\right)\cdot\left(3 \times \begin{bmatrix} \frac{9}{10} \\ -\frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}\right)^T$$

$$+ \frac{5}{10} \times \left(1 \times \begin{bmatrix} -\frac{1}{10} \\ \frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}\right)\cdot\left(1 \times \begin{bmatrix} -\frac{1}{10} \\ \frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}\right)^T$$

$$+ \frac{4}{10} \times \left(-2 \times \begin{bmatrix} -\frac{1}{10} \\ -\frac{5}{10} \\ \frac{6}{10} \end{bmatrix}\right)\cdot\left(-2 \times \begin{bmatrix} -\frac{1}{10} \\ -\frac{5}{10} \\ \frac{6}{10} \end{bmatrix}\right)^T$$

$$= \begin{bmatrix} 0.75 & -0.35 & -0.40 \\ -0.35 & 0.75 & -0.40 \\ -0.40 & -0.40 & 0.80 \end{bmatrix}$$

- $$(E[\tilde{\nabla}V])\cdot(E[\tilde{\nabla}V])^T = \begin{bmatrix} 0.09 & 0.15 & -0.24 \\ 0.15 & 0.25 & -0.40 \\ -0.24 & -0.40 & 0.64 \end{bmatrix}$$

$\square$

(c) Let $B(s)$ be the baseline function.

Then, we can write down the estimated policy gradient (denoted by ∇V_B) for the three scenarios:

Scenario #1: $S_0 = s,\ a_1 = a,\ S_1 = \text{terminal}$.

$$\nabla V_B = \begin{bmatrix} 1-\pi_\theta(a|s) \\ -\pi_\theta(b|s) \\ -\pi_\theta(c|s) \end{bmatrix} \times \left(r(s,a) - B(s) \right) = (100 - B(s)) \cdot \begin{bmatrix} \frac{9}{10} \\ -\frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}$$

Scenario #2: $S_0 = s,\ a_1 = b,\ S_1 = \text{terminal}$

$$\nabla V_B = \begin{bmatrix} -\pi_\theta(a|s) \\ 1-\pi_\theta(b|s) \\ -\pi_\theta(c|s) \end{bmatrix} \times \left(r(s,b) - B(s) \right) = (98 - B(s)) \cdot \begin{bmatrix} -\frac{1}{10} \\ \frac{5}{10} \\ -\frac{4}{10} \end{bmatrix}$$

Scenario #3: $S_0 = s,\ a_1 = c,\ S_1 = \text{terminal}$

$$\nabla V_B = \begin{bmatrix} -\pi_\theta(a|s) \\ -\pi_\theta(b|s) \\ 1-\pi_\theta(c|s) \end{bmatrix} \times \left(r(s,c) - B(s) \right) = (95 - B(s)) \cdot \begin{bmatrix} -\frac{1}{10} \\ -\frac{5}{10} \\ \frac{6}{10} \end{bmatrix}$$

It is easy to verify that $E[\nabla V_B] = \begin{bmatrix} \frac{3}{10} \\ \frac{5}{10} \\ \frac{-8}{10} \end{bmatrix}$, which does not depend on $B(s)$.

Moreover, $E\left[(\nabla V_B)\cdot(\nabla V_B)^T\right] = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}$

Let's evaluate C_{11}, C_{22}, C_{33} separately in the next page.

(Cont.)

$$C_{11} = \frac{1}{10} \times \left[\left(100 - B(s)\right) \cdot \left(\frac{9}{10}\right) \right]^2 + \frac{5}{10} \times \left[\left(98 - B(s)\right) \cdot \left(-\frac{1}{10}\right) \right]^2 + \frac{4}{10} \times \left[\left(95 - B(s)\right) \cdot \left(-\frac{1}{10}\right) \right]^2$$

$$= \frac{1}{1000} \left[90 \cdot B(s)^2 - 17940 \cdot B(s) + \text{some constant} \right]$$

$$C_{22} = \frac{1}{10} \times \left[\left(100 - B(s)\right) \cdot \left(-\frac{5}{10}\right) \right]^2 + \frac{5}{10} \times \left[\left(98 - B(s)\right) \cdot \left(\frac{5}{10}\right) \right]^2 + \frac{4}{10} \times \left[\left(95 - B(s)\right) \cdot \left(-\frac{5}{10}\right) \right]^2$$

$$= \frac{1}{1000} \left[250 \cdot B(s)^2 - 48500 \cdot B(s) + \text{some constant} \right]$$

$$C_{33} = \frac{1}{10} \times \left[\left(100 - B(s)\right) \cdot \left(-\frac{4}{10}\right) \right]^2 + \frac{5}{10} \times \left[\left(98 - B(s)\right) \cdot \left(-\frac{4}{10}\right) \right]^2 + \frac{4}{10} \times \left[\left(95 - B(s)\right) \cdot \left(\frac{6}{10}\right) \right]^2$$

$$= \frac{1}{1000} \left[240 \, B(s)^2 - 46240 \cdot B(s) + \text{some constant} \right]$$

$$\mathrm{Tr}\left(\mathrm{Cov}(\nabla V_B, \nabla V_B) \right) = C_{11} + C_{22} + C_{33} + \left(\text{some constant from } (E[\nabla V_B])(E[\nabla V_B])^T \right)$$

$$= \frac{1}{1000} \left[580 \cdot B(s)^2 - 112680 \cdot B(s) + \text{some constant} \right]$$

Then, it is easy to verify that $\mathrm{Tr}\left(\mathrm{Cov}(\nabla V_B, \nabla V_B) \right)$ is minimized at

$$B(s) = \frac{112680}{2 \times 580} \approx 97.14 .$$

(Remark: One can see that $V^{\pi_\theta}(s)$ is a near-optimal baseline in this problem). $\square$.

Problem 2

(a) Show that $\overbrace{E_{\tau \sim p_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t,a_t)\right]}^{LHS} = \overbrace{\frac{1}{1-\gamma} E_{s \sim d_\mu^{\pi_\theta}} E_{a \sim \pi_\theta(\cdot|s)}\left[f(s,a)\right]}^{RHS}$.

pf:

$$RHS = \frac{1}{1-\gamma}\sum_s \underbrace{d_\mu^{\pi_\theta}(s)}\cdot \sum_a \pi_\theta(a|s)\cdot f(s,a)$$

by the definition of $d_\mu^{\pi_\theta}(s)$

$$= \cancel{\frac{1}{1-\gamma}}\sum_s \left(\sum_{s_0} \mu(s_0)\cdot \cancel{(1-\gamma)}\cdot \sum_{t=0}^{\infty} \gamma^t \cdot P(s_t=s|s_0,\pi_\theta)\right)\cdot \sum_a \pi_\theta(a|s)\cdot f(s,a)$$

$$= \sum_{s_0} \mu(s_0)\cdot \left[\sum_s \sum_a \pi_\theta(a|s)\cdot \sum_{t=0}^{\infty} \gamma^t \cdot P(s_t=s|s_0,\pi_\theta)\cdot f(s,a)\right]$$

$$= \sum_{s_0} \mu(s_0)\cdot \left[\sum_s \sum_a \left(\sum_{t=0}^{\infty} \gamma^t \cdot \underbrace{P(s_t=s, a_t=a|s_0,\pi_\theta)}_{\substack{\| \\ \sum_{\substack{\tau=\{s_0,a_0,\cdots\} \\ \text{with } s_t=s,\ a_t=a}} P(\tau|s_0,\pi_\theta)}}\cdot f(s,a)\right)\right]$$

Let's reorganize this by figuring out how much each trajectory τ contributes

$\Rightarrow$ Each τ shall contribute $P(\tau|s_0,\pi_\theta)\cdot \sum_{t=0}^{\infty}\gamma^t f(s_t,a_t)$

$$= \sum_{s_0} \mu(s_0)\cdot \left[\sum_{\tau=\{s_0,a_0,\cdots\}} P(\tau|s_0,\pi_\theta)\left(\sum_{t=0}^{\infty} \gamma^t f(s_t,a_t)\right)\right]$$

$$= \sum_{\text{all possible } \tau} P(\tau|\pi_\theta)\left(\sum_{t=0}^{\infty} \gamma^t f(s_t,a_t)\right)$$

$$= E_{\tau \sim p_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t,a_t)\right] \equiv LHS.$$

$\square$

(b). For episodic environments, show that

$$\nabla_\theta V^{\pi_\theta}(\mu) = E_{\tau \sim p_\mu^{\pi_\theta}}\left[\sum_{t=0}^{T-1} \gamma^t A^{\pi_\theta}(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)\right].$$

Pf: The major difference between episodic and continuing environments is the existence of a "terminal state".

Let S_* be the terminal state of an episodic environment.

Once the agent reaches state S_*, it will stay at S_* forever (and hence the episode ends).

Moreover, $Q^\pi(s_*,a)=0$, $V^\pi(s_*)=0$, $A^\pi(s_*,a)=0$, for all a and all π

Let T be the episode length of a trajectory τ,

Then, we have

by the policy gradient expression (P5) in Lec 8

$$\nabla_\theta V^{\pi_\theta}(\mu) = E_{\tau \sim p_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi_\theta}(s_t,a_t)\cdot\nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

$$= E_{\tau \sim p_\mu^{\pi_\theta}}\left[\sum_{t=0}^{T-1} \gamma^t A^{\pi_\theta}(s_t,a_t)\cdot\nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

$\square$