Q1 ①

action & sequence : [1 0 0]

R : [-2 -8 9] ⇒ future return : [-1 1 9]

Model O/P

[[0.59 0.41] P : 0.41 0.47 0.67

[0.47 0.53] $\log_e P$: -0.892 -0.755 -0.4

[0.67 0.33]]

$\log P * R = 0.812 - 0.765 - 3.6$

Total loss = -($\sum$)

= [-3.463] = 3.463

② Action [0 0 0 2]

r [-8 -2 5 6]

P : 0.79 0.03 0.52 0.56

$\log P$: -0.24 -3.51 -0.65 -0.58

→ ①

	V_0	V_1	V_2	V_3
	-3.53	-1.4	-4.95	-9.52

delta = $r + \delta\, u(S_{t+1}) - V(S_t)$

$\delta_0 = -8 + (-1.4) + 3.53 = -5.87$

$\delta_1 = -2 + (-4.95) + 1.4 = -5.55$ δ = [-5.87 -5.55 0.43]

$\delta_2 = 5 + (-9.52) + 4.95 = 0.43$

Actor loss [-1.41 -19.48 0.27]

$g = (g * \delta)$

Date / /

Subject

critic loss = [34.46 30.8 0.18]

σ^2

overall loss = [33.05 11.32 0.45]

↓

actor loss + critic loss

Subject _____ Date / /

③

a_0	a_1	a_2	a_3	a_4	r_0	r_1	r_2	r_3	r_4
3	2	3	4	3	5	-1	-3	4	8

$T[0]$	$T[1]$	$T(2)$	$T(3)$	$T(4)$	V_0	V_1	V_2	V_3	V_4
5.86	-3.98	-7.89	-7	-2.89	-3.88	-2.17	9.26	-4.72	-1.73

$\gamma = 0.97, N = 0.99, \epsilon = 0.2$

1 Advantage:

$$G = r_t + \gamma V(S_{t+1}) - V(S_t)$$

$$G_0 = 5 + 0.97 * -2.17 + 3.88 = 6.78$$
$$G_1 = -1 + 0.97 * -9.26 + 2.17 = -7.81$$
$$G_2 = -3 + 0.97 * -4.72 + 9.26 = 1.682$$
$$G_3 = 4 + 0.97 * -1.73 + 4.72 = 10.4$$

$$A = \begin{bmatrix} 1 & \gamma N & \gamma^2 N^2 & \gamma^3 N^3 \\ 0 & 1 & \gamma N & \gamma^2 N^2 \\ 0 & 0 & 1 & \gamma N \\ 1 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} G_0 \to 6.78 \\ G_1 \to -7.81 \\ G_2 \to 1.682 \\ G_3 \to 10.4 \end{bmatrix}$$

$$A_1 = 6.78 + (0.97 * 0.99 * -7.81) + (0.97^2 * 0.99^2 * 10.6) + (0.97^3 * 0.99^3 * 10.4) = 10.04$$

$$A_2 = -7.81 + (0.97 * 0.99 * 1.682) + (0.97^2 * 0.99^2 * 10.4) = 3.4$$

$$A_3 = 1.682 + (0.97 * 0.99 * 10.4) = 11.67$$

$A_4 = 10.4$	A_0	A_1	A_2	A_3
	10.04	3.4	11.67	10.4

② actor loss

π_{old} : 0.39 0.07 0.06 0.18 0.04

π_{new} 0.17 0.14 0.17 036 0.24

Ration $\frac{new}{old}$

rt :

R_0	r_1	r_2	r_3	r_4
0.436	2	7.83	2	6

$\epsilon = 0.2$ → $1-\epsilon = 0.8$

→ $1+\epsilon = 1.2$

clipped rt :

fro	r_1	r_2	r_3	r_4
0.8	1.2	1.2	1.2	1.2

rt At : 0 1 2 3 4

4.38 6.8 33.03 20.8 0

rt clipped At : 0 1 2 3 4

6.03 4.03 14 12.48 0

min (clipped - unclipped)

0 1 2 3 4

4.38 4.08 14 12.48 0

actor loss = mean = 6.982

③ critic loss

Return: R ++ Vold

0	1	2	3	4
4.18	-0.58	3.78	3.4	-2.89

Return - Vnew

0	1	2	3	4
8.06	1.59	8.04	8.12	-1.16

Square(Return . Vnew)

0	1	2	3	4
64.96	2.53	170.04	65.93	1.35

critical loss = Vloan = 60.96