1. Model-Based Learning

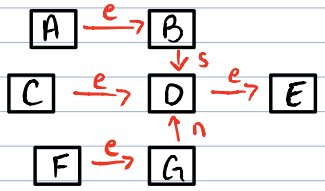
	State	Action	New State	Reward
Episode 1:	A	east	B	-1
	B	south	D	-2
	D	east	E	10

	State	Action	New State	Reward
Episode 2:	F	east	G	-1
	G	north	D	-2
	D	east	E	-10

	State	Action	New State	Reward
Episode 3:	C	east	D	-1
	D	east	B	-1
	B	south	D	-2
	D	east	G	-1
	G	north	D	-1
	D	east	E	10

S	a	S'	Count	reward
A	e	B	1	-1
B	s	D	11	-2, -2
D	e	E	111	10, 10, 10
D	e	B	1	-1
D	e	G	1	-1
F	e	G	1	-1
G	n	D	11	-2, -1
C	e	D	1	-1

A.

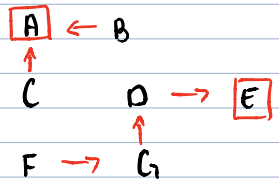


B.

T(A, east, B) = 1 b/c 1 out of 1 times A moves east to B
T(B, south, D) = 1 b/c 2 out of 2 times B moves south to D
T(D, east, E) = 3/5 or 0.60 b/c 3 out of 5 times D moves east to E
these 2 are included for visuals
X T(D, east, B) = 1/5 or 0.20 b/c 1 out of 5 times D moves east to B
X T(D, east, G) = 1/5 or 0.20 b/c 1 out of 5 times D moves east to G
T(F, east, G) = 1 b/c 1 out of 1 times F moves east to G
T(G, north, D) = 1 b/c 2 out of 2 times G moves north to D
T(C, east, D) = 1 b/c 1 out of 1 times C moves east to D.

R (A, east, B) = -1
R (B, south, D) = -2
R (D, east, E) = 10
R (D, east, B) = -1
R (D, east, G) = -1
R (F, east, G) = -1
R (G, north, D) = -2 or -1, averages to -1.5
R (C, east, D) = -1

2. Model-Free Passive Reinforcement Learning



$\gamma = 0.9$

Episode 1:

State	Action	New State	Reward
F	east	C	-3
C	north	A	-1
A	exit	x	10

Episode 2:

State	Action	New State	Reward
B	west	D	-2
D	east	E	-3
E	exit	x	5

Episode 3:

State	Action	New State	Reward
F	east	G	-1
G	north	D	-2
D	east	E	-5
E	exit	x	7

Episode 4:

State	Action	New State	Reward
G	north	D	-1
D	east	B	-2
B	west	A	-4
A	exit	x	8

■ = with discount → $\gamma V(s')$

In A:
Episode 1: $0.9^0 \cdot 10 = 10$
Episode 4: $0.9^0 \cdot 8 = 8$
$10 + 8 = 18/2 = \boxed{9}$

In B:
Episode 2: $(0.9^0 \cdot -2) + (0.9^1 \cdot -3) + (0.9^2 \cdot 5) = -0.65$
Episode 4: $(0.9^0 \cdot -4) + (0.9^1 \cdot 8) = 3.2$
$-0.65 + 3.2 = 2.55/2 = \boxed{1.275}$

In C:

Episode 1: $(0.9^0 \cdot -1) + (0.9^1 \cdot 10) = 8$

$\boxed{8}$

In D:

Episode 2: $(0.9^0 \cdot -3) + (0.9^1 \cdot 5) = 1.5$

Episode 3: $(0.9^0 \cdot -5) + (0.9^1 \cdot 7) = 1.3$

Episode 4: $(0.9^0 \cdot -2) + (0.9^1 \cdot -4) + (0.9^2 \cdot 8) = 0.88$

$(1.5 + 1.3 + 0.88)/3 = \boxed{1.23}$

In E:

Episode 2: $0.9^0 \cdot 5 = 5$

Episode 3: $0.9^0 \cdot 7 = 7$

$5 + 7 = 12/2 = \boxed{6}$

In F:

Episode 1: $(0.9^0 \cdot -3) + (0.9^1 \cdot -1) + (0.9^2 \cdot 10) = 4.2$

Episode 3: $(0.9^0 \cdot -1) + (0.9^1 \cdot -2) + (0.9^2 \cdot -5) + (0.9^3 \cdot 7) = -6.84$

$4.2 - 6.84/2 = \boxed{-1.32}$

In G:

Episode 3: $(0.9^0 \cdot -2) + (0.9^1 \cdot -5) + (0.9^2 \cdot 7) = -0.83$

Episode 1: $(0.9^0 \cdot -1) + (0.9^1 \cdot -2) + (0.9^2 \cdot -4) + (0.9^3 \cdot 8) = -6.03$

$-0.83 - 6.03/2 = \boxed{-3.43}$

3. Approximate Q-Learning

a.

$$\hat{Q}(s,a) = w_e f_e + w_m f_m + w_x f_x$$

$$\hat{Q}(s, \text{east}) = 6 \cdot 0.7 + -3 \cdot 0.1 + -5 \cdot 0.4$$

$$1.2 + -0.3 + -2$$

$$\boxed{1.9}$$

b. $Q(s,\text{east}) = 6 \cdot \text{Exit} - 3 \cdot \text{Monster} - 5 \cdot \text{bomb}$

$1.2 - 0.3 - 2$

1.9

$Q(s, \text{east}) = 1.9$

$$r + \gamma \max_{a'} Q(s', a') = -1{,}000 + 0$$

$$\text{difference} = -1{,}000 - 1.9$$

$$= \boxed{-1{,}001.9}$$

c.

$$w_{\text{Exit}} \leftarrow 6 + (\alpha [-1{,}001.9] \cdot 0.7) \quad \alpha = 0.2$$

$6 - 140.266 = -134.266$

$$w_{\text{Monster}} \leftarrow -3 + (\alpha [-1{,}001.9] \cdot 0.1)$$

$-3 - 20.038 = -23.038$

$$w_{\text{Bomb}} \leftarrow -5 + (\alpha [-1{,}001.9] \cdot 0.4)$$

$-5 - 80.152 = -85.152$

$$Q(s, \text{east}) = -134.266 f_e - 23.038 f_m - 85.152 f_x$$