Important: Answer all 6 parts, 1(a) - 1(f). Each part is worth the same. Mark your answers with a tick on the separate multiple-choice answer sheet. **Do not** mark your answers on this question sheet.

Question 1
Consider the pseudo-code for the following algorithm:

initialise V randomly;
for *each episode* **do**
  initialise s randomly;
  for *each step of episode* **do**
    choose action a from policy π at s;
    take action a;
    observe immediate reward r, and successor state s';
    $\text{tmp} \rightarrow r + \gamma V(s') - V(s)$;
    $V(s) \leftarrow r + \alpha \text{tmp}$;
    $s \rightarrow s'$;
  end
end

Based on the content of our course identify precisely the class of algorithm to which the pseudocode above belongs:

A Greedy TD estimation
B TD value function estimation with exploring starts
C MC policy iteration
D α-greedy policy evaluation
E TD value iteration
F Generalised policy iteration
G MC value function estimation with exploring starts
H ϵ-greedy MC value function estimation
I MDP policy evaluation
J SARSA policy evaluation
K ϵ-greedy value function estimation
L MC policy estimation with exploring starts
M None of the above

Question 2 Consider an undiscounted ($\gamma = 1$) 3 state MDP, with states House and School being transient and Holidays being an absorbing state. The following two state-reward traces have been observed:
τ_1 = House +3 House +2 School -4 House +4 School -3 Holidays
τ_2 = House -2 House +3 School -3 Holidays
For the above two traces, perform Every-Visit MC estimation of the value function at V(House) and select the closest value from the choices below.

A 2.75	F 1.5	K 0.5	P -2.25	U -1.25
B 2.5	G 1.0	L 0.25	Q -2	V -1
C 2.25	H 1.25	M 0	R -1.75	W -0.75
D 2	I 1	N -2.75	S -1.5	X -0.5
E 1.75	J 0.75	O -2.5	T -1.0	Y -0.25

Question 3 For the same MDP as in Question 1(b), perform First-Visit MC estimation of the value function at $V(\text{House})$, and select the closest value from the choices below.

A 2.75	F 1.5	K 0.5	P -2.25	U -1.25
B 2.5	G 1.0	L 0.25	Q -2	V -1
C 2.25	H 1.25	M 0	R -1.75	W -0.75
D 2	I 1	N -2.75	S -1.5	X -0.5
E 1.75	J 0.75	O -2.5	T -1.0	Y -0.25

Question 4
Consider the pseudo-code for Deep Q-learning:

```
1  Initialise Q randomly
2  Initialise 𝒟 to empty
3  Initialise 𝒳 to empty
4  for each episode do
5  |  S ←S_init
6  |  for each step of episode do
7  |  |  Choose A from S using policy derived from Q
8  |  |  Take A, observe R and S'
9  |  |  Compute x = R + γ max_a Q(S', a) − Q(S, A)
10 |  |  Store |x| in 𝒳
11 |  |  Store transition (S, A, R, S') in 𝒟
12 |  end
13 |  Get mini-batch ℬ ⊂ 𝒟 by sampling according to 𝒳
14 |  Update Q using ℬ
15 end
```

Which extension to Deep Q-learning is most accurately represented by lines 9, 10, and 13? Select your answer from the options below.

A Policy gradients
B Action sampling for continuous actions
C Double Q-learning
D Waiting for a minimum replay buffer size before training
E Model-based deep reinforcement learning
F Using a target network
G Cross entropy method
H Multi-step TD learning
I Prioritised experience replay
J Model predictive control
K Clipping of the TD error

Question 5 In each iteration of Monte Carlo Tree Search, there are four distinct algorithms. Each of the four lines below describes a different one of these four algorithms, and represents a unique component of that algorithm:

1 Modify the values of a set of nodes.

2 Take a path through the tree by considering the value of each node.

3 Take a path through the tree by taking random actions.

4 Create a new child node if possible.

What is the correct order of these four algorithms, in one iteration of Monte Carlo Tree Search? Select your answer from the options below. For example, selecting [G] indicates you believe that each iteration should start with line **2**, then do line **1**, then do line **3**, and finally do line **4**.

[A] **1**, **2**, **3**, **4**	[F] **1**, **4**, **3**, **2**	[K] **2**, **4**, **1**, **3**	[P] **3**, **2**, **4**, **1**	[U] **4**, **2**, **1**, **3**
[B] **1**, **2**, **4**, **3**	[G] **2**, **1**, **3**, **4**	[L] **2**, **4**, **3**, **1**	[Q] **3**, **4**, **1**, **2**	[V] **4**, **2**, **3**, **1**
[C] **1**, **3**, **2**, **4**	[H] **2**, **1**, **4**, **3**	[M] **3**, **1**, **2**, **4**	[R] **3**, **4**, **2**, **1**	[W] **4**, **3**, **1**, **2**
[D] **1**, **3**, **4**, **2**	[I] **2**, **3**, **1**, **4**	[N] **3**, **1**, **4**, **2**	[S] **4**, **1**, **2**, **3**	[X] **4**, **3**, **2**, **1**
[E] **1**, **4**, **2**, **3**	[J] **2**, **3**, **4**, **1**	[O] **3**, **2**, **1**, **4**	[T] **4**, **1**, **3**, **2**	

Question 6
Consider the following equation for computing the policy gradient, with on-policy learning and Monte-Carlo sampling of trajectories $\tau^{(1)} \dots \tau^{(M)}$:

$$\nabla_\theta J(\theta) \approx \frac{1}{M} \sum_{i=1}^{i=M} \sum_{t=0}^{t=H} Z R(\tau^{(i)})$$

What term does Z represent? Select your answer from the options below. In this context, $Q_\theta(S, A)$ is an approximation of the Q-value, $\pi_\theta(A|S)$ is the policy's action distribution, $V_\theta(S)$ is an approximation of the state-value, and N is the number of transitions in each mini-batch sampled from the experience replay buffer.

[A] $\sum_{j=1}^{j=N} \log Q_\theta(s_j^{(i)}, a_j^{(i)})$
[B] $\sum_{j=1}^{j=N} \log \pi_\theta(a_j^{(i)}|s_j^{(i)})$
[C] $\sum_{j=1}^{j=N} \log V_\theta(s_j^{(i)})$
[D] $\nabla_\theta \sum_{j=1}^{j=N} \log Q_\theta(s_j^{(i)}, a_j^{(i)})$
[E] $\nabla_\theta \sum_{j=1}^{j=N} \log \pi_\theta(a_j^{(i)}|s_j^{(i)})$
[F] $\nabla_\theta \sum_{j=1}^{j=N} \log V_\theta(s_j^{(i)})$
[G] $\nabla_\theta \log Q_\theta(s_t^{(i)}, a_t^{(i)})$
[H] $\nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})$
[I] $\nabla_\theta \log V_\theta(s_t^{(i)})$
[J] $\log Q_\theta(s_t^{(i)}, a_t^{(i)})$
[K] $\log \pi_\theta(a_t^{(i)}|s_t^{(i)})$
[L] $\log V_\theta(s_t^{(i)})$

CO424 exam: Answer sheet for multiple-choice questions.

. . . . . .

Write your first name, last name, CID

Answers must be given exclusively on this sheet, by ticking one box per question. Answers given on the other sheets will be ignored. All questions have only one answer. If you need to make corrections, use the provided correction stickers to cover up answers you want to change.

Question 1: A B C D E F G H I J K L M

Question 2: A B C D E F G H I J K L M N
O P Q R S T U V W X Y

Question 3: A B C D E F G H I J K L M N
O P Q R S T U V W X Y

Question 4: A B C D E F G H I J K

Question 5: A B C D E F G H I J K L M N
O P Q R S T U V W X

Question 6: A B C D E F G H I J K L

Answers:
1 B
2 M
3 M
4 I
5 L
6 H