MIDTERM PRACTICE PAPER - FORM A

120 minutes / 100 points

Original practice questions based on the supplied lectures and photographed review priorities. This is not an official examination or a prediction of the actual questions.

Examination conditions

The announced examination is on 13 November 2026, 09:30-11:30. For this practice paper, allow exactly two hours.

Closed-book, except for two double-sided sheets of notes. A standalone calculator with calculation functions only is permitted. Computers, tablets, phones, and other electronic devices are not permitted.

Section	Questions	Points	Suggested time
I. Single-choice	A1-A10: 10 x 2	20	15 minutes
II. Short-answer	B1-B5: 5 x 5	25	30 minutes
III. Calculation - MDP	C1-C3: 8 + 8 + 9	25	30 minutes
IV. Calculation - Learning algorithms	D1-D2: 10 + 10	20	25 minutes
V. Calculation - Optimal policy	E1: 10	10	15 minutes
Checking time			5 minutes

Instructions and conventions

Answer all 21 numbered questions. Each single-choice question has exactly one correct answer. There is no penalty for an incorrect answer. Show your working in Sections III-V; method marks are available.

Use the reward-on-transition convention: R_{t+1} is received after action A_t. All terminal states have value zero and no future rewards. A reward on entry into a terminal state is still counted.

For every ε-greedy question, exploration is uniform over ALL available actions, including greedy actions. If several actions are tied for the maximum, split the exploitation probability equally among them.

Unless stated otherwise, use synchronous DP updates: every value in sweep k+1 is computed from the complete vector in sweep k. Different questions and independently specified algorithms start from their own stated initial values.

Give exact fractions or at least four decimal places when appropriate. Write short answers in English. The separate answer booklet provides Chinese explanations and marking guidance.

Name: _____ Student ID: _____

Single-choice questions: A1-A5

10 questions x 2 points = 20 points. Select exactly one option for each question.

A1. The Markov property [2 points]

Which statement correctly describes a Markov state in an MDP?

A. The next state must be deterministic.

B. Given the current state and action, the next-state and reward distribution does not require the earlier history.

C. The current state must contain a copy of the entire history.

D. Rewards must be independent of the action.

A2. Return and discounting [2 points]

An episode produces rewards 2, -1, and 4, then terminates. If γ = 0.5, what is the return at the start?

A. 5

B. 1.5

C. 2.5

D. 3

A3. State value and action value [2 points]

For a fixed policy π, which expression equals $v_\pi(s)$?

A. $\sum_a \pi(a \mid s) q_\pi(s,a)$ B. $\max_a q_\pi(s,a)$

C. $\sum_a q_\pi(s,a)$ D. $\min_a q_\pi(s,a)$

A4. What does one sweep mean? [2 points]

In synchronous iterative policy evaluation, which values are used to compute $V_{k+1}(s)$?

A. The exact optimal values at all successor states.

B. Newly updated values from earlier states in the same sweep.

C. The return from one complete sampled episode only.

D. Values from the previous complete sweep, V_k.

A5. ε-greedy action probability [2 points]

There are four actions, a unique greedy action, and ε = 0.2. What is the probability of selecting the greedy action?

A. 0.80 B. 0.85 C. 0.20 D. 0.95

Single-choice questions: A6-A10

A6. On-policy and off-policy control [2 points]

With an exploratory behavior policy, which statement is correct?

A. SARSA always uses the largest Q value at the next state.

B. Q-learning requires its behavior policy to be greedy at every step.

C. SARSA uses the action actually selected next; Q-learning uses a greedy next-action value in its target.

D. Both methods require the transition probability model.

A7. Value function approximation [2 points]

Why replace a lookup table with a parameterized value function in a large state space?

A. To share parameters and generalize across states while avoiding a separate stored value for every state.

B. To guarantee exact values after one update.

C. To remove the need for any experience or rewards.

D. To ensure all possible training algorithms converge.

A8. Dueling network aggregation [2 points]

A dueling network uses mean-centered advantages. Its outputs are $V(s) = 5$ and raw advantages $(2, -1, 0)$. What are its three Q values?

A. $(7, 4, 5)$

B. $(5, 2, 3)$

C. $(2, -1, 0)$

D. $(20/3, 11/3, 14/3)$

A9. Stochastic policies [2 points]

Which example gives a valid reason to learn a stochastic policy directly?

A. A stochastic policy is strictly better in every fully observed finite discounted MDP.

B. In rock-paper-scissors against an adaptive opponent, a predictable deterministic action can be exploited.

C. A stochastic policy makes all rewards deterministic.

D. A stochastic policy requires no optimization objective.

A10. Learning and planning [2 points]

In Dyna-style learning, what is the role of a learned environment model?

A. It permanently replaces all real interaction before any data are collected.

B. It stores only the greedy action and cannot predict a next state.

C. It produces simulated transitions for additional value updates between real interactions.

D. It guarantees that approximation errors can never affect the learned policy.

Answer record: A1 ___ A2 ___ A3 ___ A4 ___ A5 ___ A6 ___ A7 ___ A8 ___ A9 ___ A10 ___

Short-answer questions: B1-B3

5 questions x 5 points = 25 points. Brief explanations are sufficient.

B1. DP, MC, and TD(0) [5 points]

For standard policy evaluation, compare DP, Monte Carlo, and TD(0) in terms of (i) requiring an environment model, (ii) sampling, and (iii) bootstrapping. Then explain the usual bias-variance difference between a complete MC return and a TD target using an approximate value function.

Method	Model required?	Samples?	Bootstraps?
DP			
MC			
TD(0)			

B2. GLIE and ε-greedy [5 points]

State the two conditions defining GLIE. Explain why a fixed $\varepsilon = 0.1$ does not satisfy GLIE, and why ε_k tending to zero alone is insufficient. Finally, distinguish the roles of MC, ε-greedy, and GLIE in MC control.

B3. DQN and Double DQN [5 points]

Explain one purpose of experience replay and one purpose of a target network. Write the nonterminal target for standard DQN and for Double DQN, using online parameters θ and target parameters θ^-. State the estimation problem that Double DQN is designed to reduce.

Short-answer questions: B4-B5

B4. Prioritized replay and dueling networks [5 points]

In proportional prioritized experience replay, explain how the magnitude of the TD error affects sampling, why a small positive priority offset is useful, and why importance-sampling weights are introduced. For a dueling network, name its two output streams and write the mean-centered formula that combines them into Q values.

B5. Policy gradients and actor-critic [5 points]

Consider an undiscounted episodic task. Write a REINFORCE ascent update with an action-independent state baseline b(s). Explain why subtracting this baseline does not change the expected policy gradient, and give its main purpose. Identify the roles of the actor and critic. Finally, write a one-step TD error that can serve as an advantage estimate.

C1. Returns and a Markov reward process

8 points / 4 subquestions

Consider an MRP with nonterminal states A and B and terminal state T. Let $\gamma = 0.8$. The immediate reward is deterministic given the current state and is received on the transition. The transition model is:

Current state	Immediate reward	Next-state probabilities
A	1	P(A)=0.5, P(B)=0.5
B	2	P(A)=0.25, P(T)=0.75
T	0	Absorbing; V(T)=0

(a) Sample returns [2 points]

One sampled episode is A → A → B → T, with rewards 1, 1, 2. Compute G_0 and G_1.

(b) Bellman equations [2 points]

Write the two Bellman expectation equations for v(A) and v(B).

(c) Exact values [2 points]

Solve the equations from (b) for v(A) and v(B).

(d) Iterative evaluation [2 points]

Starting from $V_0(A)=V_0(B)=0$, compute the complete vectors V_1 and V_2 using synchronous updates. List the entries in the order (A, B).

C2. From an MDP to a policy-induced MRP

8 points / 4 subquestions

An MDP has states X, Y, and terminal state T, with $\gamma = 0.5$. Rewards are received on transitions. All transitions not shown have probability zero.

State	Action	Reward	Next state
X	wait	0	X with probability 1
X	go	2	Y with probability 1
Y	retry	0	X with probability 0.5; T with probability 0.5
Y	finish	4	T with probability 1

The policy chooses wait and go with equal probability at X. At Y, it chooses retry with probability 0.25 and finish with probability 0.75. Use state order (X, Y, T), with an absorbing zero-reward terminal row.

(a) Induced model [2 points]

Write the 3 x 3 matrix P^{π} and the reward vector r^{π}.

(b) Policy values [2 points]

Write and solve the Bellman equations for $v_{\pi}(X)$ and $v_{\pi}(Y)$.

(c) Action values [2 points]

Compute all four nonterminal action values $q_{\pi}(X,\text{wait})$, $q_{\pi}(X,\text{go})$, $q_{\pi}(Y,\text{retry})$, and $q_{\pi}(Y,\text{finish})$.

(d) Policy improvement [2 points]

Give a deterministic policy π' greedy with respect to q_{π}. Evaluate $v_{\pi'}(X)$ and $v_{\pi'}(Y)$.

C3. Gridworld policy evaluation

9 points / 5 subquestions

Consider the 3 x 3 grid below. Coordinates are (row, column). The top-left and bottom-right cells are terminal. The policy selects Up, Down, Left, and Right with probability 1/4 each. Actions are otherwise deterministic.

T: (1,1)	(1,2)	(1,3)
(2,1)	(2,2)	(2,3)
(3,1)	(3,2)	T: (3,3)

An attempted move outside the grid leaves the agent in the same cell. Every action from a nonterminal cell gives reward -1, including a move into a terminal cell. After termination there are no further rewards. Let $\gamma = 1$, and initialize every value to zero.

(a) Update rule [1 point]

Write the synchronous policy-evaluation update for a nonterminal cell, including how an out-of-grid action is handled.

(b) First sweep [2 points]

Compute the entire 3 x 3 matrix V_1.

(c) Second sweep [2 points]

Compute the entire 3 x 3 matrix V_2.

(d) Third sweep [2 points]

Compute $V_3(1,2)$ and $V_3(2,2)$. Show the four successor values used in each calculation.

(e) One-step greedy choice [2 points]

Using V_2 as the continuation value, compute the four action values at (1,2) and identify the greedy action.

D1. MC prediction versus online TD(0)

10 points / 5 subquestions

A fixed policy generates the episode below. T is terminal; the numbers over the arrows are rewards. Let γ = 0.9.

$$S_0 = A \xrightarrow{1} S_1 = B \xrightarrow{2} S_2 = A \xrightarrow{3} S_3 = T$$

For parts (b) and (d), independently start from V(A)=1, V(B)=2 and V(T)=0, with α = 0.5. Do not carry an updated value from one part into another.

(a) Returns [2 points]

Compute G_0, G_1, and G_2.

(b) First-visit MC [2 points]

Process this episode using first-visit MC and the constant-step update V(s) ← V(s)+α[G-V(s)]. Update each state only from its first occurrence in the episode. Find the final V(A) and V(B).

(c) Every-visit sample mean [2 points]

In a separate estimator, use only the returns from this episode and take their arithmetic mean for each state. There are no previous samples. Compute the every-visit MC estimates of A and B; do not use α in this part.

(d) Online TD(0) [3 points]

Starting again from the stated initial values, process all three transitions in chronological order. Immediately keep each update for use in later transitions. Show each updated value and the final pair (V(A), V(B)).

(e) Interpretation [1 point]

Why can the MC and TD estimates differ after the same episode?

D2. SARSA, Q-learning, and deep Q targets

10 points / 5 subquestions

Observe a nonterminal transition (s,a,r=1,s'). Initially Q(s,a)=2, γ=0.9 and α=0.5. The next-state action values are:

Action at s'	a1	a2	a3
Current Q / online-network output	4	1	3

The behavior policy uses ε-greedy with ε=0.3. The next action actually selected is a2. Treat the updates in (b) and (c) independently, each starting from Q(s,a)=2.

(a) Behavior probabilities [2 points]

Compute the probability of each action at s'.

(b) SARSA [2 points]

Compute the SARSA target and the new Q(s,a).

(c) Q-learning [2 points]

Compute the Q-learning target and the new Q(s,a).

(d) DQN and Double DQN targets [3 points]

For a separate deep-network calculation, retain the same r and γ. The online-network outputs at s' are (4,1,3), and the target-network outputs are (2,5,3), in action order (a1,a2,a3). Compute the standard DQN target and the Double DQN target. State which action each target evaluates. No parameter-gradient update is required.

(e) Terminal transition [1 point]

Instead suppose s' is terminal, with the same reward r=1. Starting again from Q(s,a)=2, give the target and the tabular updated value. Explain what happens to the next-state term.

E1. Finding and checking an optimal policy

10 points / 4 subquestions

Consider the following deterministic MDP with γ=0.5, nonterminal states S and U, and terminal state T. Rewards are received on transitions.

State	Action	Reward	Next state
S	a	1	S
S	b	0	U
U	c	3	T
U	d	1	S

Start policy iteration with $\pi_0(S)=a$ and $\pi_0(U)=d$.

(a) Evaluate the initial policy [2 points]

Find $v_{\pi 0}(S)$ and $v_{\pi 0}(U)$. A trajectory need not terminate for its discounted value to be well defined.

(b) Improve the policy [3 points]

Compute all four action values under π_0 and give a greedy policy π_1.

(c) Verify optimality [3 points]

Evaluate π_1. Then compare both actions at each state using $v_{\pi 1}$. Decide whether another policy improvement changes the policy. State v^* and an optimal policy π^*.

(d) Value iteration [2 points]

Independently start value iteration from $V_0(S)=V_0(U)=0$. Compute V_1, V_2, and V_3 in state order (S,U), using synchronous optimality backups.

END OF PAPER. Check that you have answered all 21 numbered questions.