Reinforcement Learning Midterm - Form A

Answer key, worked solutions, and marking guide

This booklet answers the original Form A practice paper. Question numbers A1-E1 match the English question booklet. All numerical results and marking allocations follow the Chinese solution booklet.

Section	Questions	Points
I. Single-choice	A1-A10	20
II. Short-answer	B1-B5	25
III. Calculation - MDP	C1-C3	25
IV. Calculation - Learning algorithms	D1-D2	20
V. Calculation - Optimal policy	E1	10
Total	21 numbered questions	100

Single-choice answer key

Q.	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Key	B	C	A	D	B	C	A	D	B	C

Conventions used throughout

R_{t+1} is received on the transition after action A_t. A terminal state has value zero, but the reward received on entering it still counts. Exploration in ε-greedy is uniform over all available actions, including greedy actions.

Synchronous DP uses the previous complete value vector. Online TD immediately retains each update for subsequent transitions. Independent methods restart from the initial estimates specified in the question.

$$\text{new estimate} = \text{old estimate} + \alpha(\text{target} - \text{old estimate})$$

How to use this booklet

Try the English paper under the 120-minute limit before checking the answers. The short-answer solutions are model responses: equivalent explanations and mathematically equivalent formulas are acceptable. Extra explanatory notes help with revision and need not all be reproduced in an exam.

For calculations, show the target or Bellman equation, substitute the numbers, and give the result. Where a later result inherits an earlier arithmetic error, award valid method marks rather than repeatedly penalizing the same error.

Scope and references

The paper emphasizes the photographed review topics. A10 adds a two-point Dyna question from the supplied RL-12 lecture, which was not listed in the photograph. Lecture references use physical PDF page numbers, counted from page 1. The final page maps questions to lecture topics.

This is an original practice paper and solution guide, not an official examination or a prediction of the actual questions.

A1-A5: Foundations and evaluation

A1. B - The Markov property [2 points]

Given the current state and action, the conditional distribution of the next state and reward does not require earlier history. Transitions may still be random. A state can summarize relevant history without storing every past observation. Rewards may depend on the action.

Reference: RL-3, states and the Markov property; RL-4, the MDP definition.

A2. C - Discounted return [2 points]

$$G_0 = 2 + 0.5(-1) + 0.5^2(4) = 2 - 0.5 + 1 = 2.5$$

The first reward is not discounted. The third reward is multiplied by γ^2. Option A gives the undiscounted sum, 5.

Reference: RL-4, pp. 18-25.

A3. A - Average over the policy [2 points]

$$v_\pi(s) = \sum_a \pi(a \mid s) q_\pi(s, a)$$

Under a fixed policy, actions are selected with probabilities $\pi(a|s)$, so the state value is the probability-weighted average of their action values. A maximum describes a greedy choice and cannot replace this average for an arbitrary policy. An unweighted sum also omits the required probabilities.

Reference: RL-4, pp. 42 and 48-52.

A4. D - Use the previous complete sweep [2 points]

In synchronous policy evaluation, every entry of V_{k+1} is computed using V_k. Using values already updated during the same sweep would be an in-place update, contrary to the convention in this question.

Here k counts evaluation sweeps. It is not an episode number or an environment time step; an index must be interpreted in its algorithmic context.

Reference: RL-5, pp. 8-14.

A5. B - Exploration can also select the greedy action [2 points]

$$P(\text{greedy}) = (1 - \varepsilon) + \frac{\varepsilon}{4} = 0.8 + 0.05 = 0.85$$

Each other action has probability 0.05, and the four probabilities sum to 1. The value 0.80 counts only the exploitation branch; 0.20 is the probability of entering the random-exploration branch.

Reference: RL-7, pp. 16-18. Award 2 points for the correct option in each question.

A6-A10: Control and deep RL

A6. C - Actual next action versus a maximum [2 points]

SARSA uses the target r+γQ(s',a'), where a' is the next action actually selected by the behavior policy. Q-learning uses r+γ max Q(s',a') in its target while the behavior policy may continue exploring. Neither method requires a known transition model.

Reference: RL-7, pp. 34-36 and 55-61.

A7. A - Parameter sharing and generalization [2 points]

A parameterized value function avoids storing an independent entry for every state and can generalize from visited states to unseen states. This is useful in large or continuous state spaces. Approximation may introduce error and does not guarantee convergence of every learning algorithm.

Reference: RL-8, pp. 9-11.

A8. D - Subtract the mean raw advantage [2 points]

$$\bar{A} = \tfrac{2-1+0}{3} = \tfrac{1}{3}, \qquad Q(s,a) = V(s) + A(s,a) - \bar{A}$$

$$Q(s,\cdot) = \left(5+2-\tfrac{1}{3},\ 5-1-\tfrac{1}{3},\ 5-\tfrac{1}{3}\right)$$

$$= \left(\tfrac{20}{3}, \tfrac{11}{3}, \tfrac{14}{3}\right)$$

Option A omits mean subtraction. Option B subtracts the maximum advantage, a different aggregation rule from the mean-centered rule specified. Option C omits the value stream.

Reference: RL-10a, pp. 32-39.

A9. B - Avoid predictable behavior [2 points]

In rock-paper-scissors, an adaptive opponent can exploit a predictable deterministic action. Uniform random play avoids that particular exploitability. This does not imply that stochastic policies are strictly better in every fully observed finite discounted MDP; under standard assumptions, such an MDP has a deterministic stationary optimal policy.

Reference: RL-11, pp. 12-19.

A10. C - Generate simulated experience [2 points]

Dyna combines real interaction, model learning, and planning. A learned model predicts rewards and successor states, allowing extra value updates from simulated transitions between real interactions. The model must be learned from data, and model errors can affect the resulting policy.

Reference: RL-12, integrating learning and planning / Dyna. This is the supplementary topic outside the photographed list. Award 2 points for each correct option.

B1-B2: Evaluation methods and GLIE

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

Method	Model required?	Samples?	Bootstraps?
Standard DP	Yes	No: full expectation	Yes
MC	No	Yes: complete return	No
TD(0)	No	Yes: one transition	Yes

Bootstrapping means using another current value estimate in the update target. DP and TD use estimated successor values. Standard complete-episode MC uses the observed return instead.

For a fixed policy, the complete return G_t is an unbiased sample of $v_\pi(S_t)$, but it usually has relatively high variance because it depends on many random future events. A practical TD target $r+\gamma V(s')$ uses an approximate value estimate, so it may be biased, but its variance is often lower.

If V equals the true v_π, the one-step TD target is also unbiased conditional on the current state. The usual variance comparison is not a universal ordering for every possible example. These statements concern targets, not the unbiasedness of every finite-step estimator.

Marking: 1 point for each fully correct table row (3 total); 1 for the MC bias/variance explanation; 1 for the TD explanation.

B2. GLIE and ε-greedy [5 points]

GLIE means Greedy in the Limit with Infinite Exploration. It requires both of the following:

1. Infinite exploration: every relevant state-action pair is visited infinitely often in an indefinitely continued learning process.

2. Greediness in the limit: the policy approaches a greedy policy with respect to its current action-value estimates; the total probability of nongreedy actions approaches zero.

With fixed $\varepsilon=0.1$, nongreedy actions retain positive probability whenever such actions exist, so the policy does not generally become greedy in the limit. Conversely, $\varepsilon_k \to 0$ alone does not guarantee sufficient exploration: decay may be too fast, or some states may remain insufficiently visited.

MC specifies how values are updated using complete returns. ε-greedy specifies how actions are chosen. GLIE specifies asymptotic exploration and greediness requirements for a sequence of policies. Using ε-greedy within MC control does not automatically establish GLIE.

Marking: 1 point for each GLIE condition; 1 for fixed ε; 1 for why decay alone is insufficient; 1 for distinguishing MC, ε-greedy, and GLIE.

References: RL-6, pp. 21-28 and 40-43; RL-7, pp. 16-24. GLIE is adjacent review material to ε-greedy.

B3-B4: DQN and its improvements

B3. DQN and Double DQN [5 points]

Experience replay stores previous transitions and samples them for training. This reduces correlations caused by training on consecutive transitions and allows data reuse. A target network changes more slowly than the online network, providing a more stable regression target; it is updated periodically rather than frozen forever.

For a nonterminal transition, the targets are:

$$y_{\text{DQN}} = r + \gamma \max_{a'} Q(s', a'; \theta^-)$$

$$a^* = \arg\max_{a'} Q(s', a'; \theta), \qquad y_{\text{Double}} = r + \gamma Q(s', a^*; \theta^-)$$

Double DQN uses the online network to select an action and the target network to evaluate that action. Separating these roles reduces overestimation associated with maximizing noisy value estimates. It does not guarantee a smaller error on every sample or eliminate all bias.

Marking: replay purpose 1; target-network purpose 1; each target formula 1; overestimation / selection-evaluation separation 1.

B4. Prioritized replay and dueling networks [5 points]

Proportional prioritized replay typically defines a positive priority using the absolute TD error:

$$p_i = |\delta_i| + \eta, \qquad P(i) = \frac{p_i^\kappa}{\sum_j p_j^\kappa}, \quad \eta > 0,\ \kappa \geq 0$$

A larger TD error often indicates an experience that has not yet been learned well. The positive offset η prevents zero-error transitions from losing all chance of being sampled. Nonuniform replay changes the sampling distribution, so importance-sampling weights correct bias relative to a uniform-buffer objective:

$$w_i \propto \left(\frac{1}{NP(i)}\right)^\beta$$

β=1 gives full importance-sampling correction before any common rescaling. In practice, β may be increased gradually and weights normalized. κ controls prioritization; β controls correction. These are distinct from the learning rate and ε-greedy exploration.

A dueling network has a state-value stream V(s) and an action-advantage stream A(s,a). Mean-centered aggregation is:

$$Q(s, a) = V(s) + A(s, a) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s, a')$$

Without centering, adding a constant to V and subtracting it from every A leaves Q unchanged. Centering gives the effective advantages zero mean; a common shift of the raw advantage outputs still cancels. The value stream is not necessarily the value of an arbitrary behavior policy.

Marking: priority from TD error 1; positive offset 1; importance-weight purpose 1; two streams 1; correct aggregation 1.

References: RL-9, pp. 30-32; RL-10a, pp. 3-11, 14-27, and 32-39.

B5: Policy gradients and actor-critic

REINFORCE ascent update [1 point]

The question specifies an undiscounted episodic task, so γ=1. With an action-independent state baseline, one update is:

$$\theta \leftarrow \theta + \alpha[G_t - b(S_t)]\nabla_\theta \log \pi_\theta(A_t \mid S_t)$$

The baseline is treated as a control variate in this actor update, not differentiated through the multiplying return term. A parameterized baseline can be trained separately.

Why the baseline leaves the expectation unchanged [1 point]

At a fixed state s, b(s) is independent of the sampled action. Therefore:

$$\sum_a \pi_\theta(a \mid s) b(s) \nabla_\theta \log \pi_\theta(a \mid s)$$

$$= b(s) \sum_a \nabla_\theta \pi_\theta(a \mid s)$$

$$= b(s) \nabla_\theta \sum_a \pi_\theta(a \mid s) = b(s) \nabla_\theta 1 = 0$$

Subtracting this term does not change the expected policy gradient. An arbitrary action-dependent baseline does not generally satisfy this argument.

Purpose of the baseline [1 point]

A suitable baseline reduces the variance of the gradient estimate and can make learning more stable. A common choice is b(s)=V(s). Then G_t-V(S_t) measures how much better or worse this outcome was than the usual value of the state. An arbitrary baseline is not guaranteed to reduce variance.

Actor and critic [1 point]

The actor is the policy π_θ and determines how actions are selected. The critic estimates values or advantages and supplies an evaluation signal that guides policy improvement. The critic learns an evaluation; the actor improves the policy using that evaluation.

One-step TD error as an advantage estimate [1 point]

$$\delta_t = R_{t+1} + \gamma V_w(S_{t+1}) - V_w(S_t)$$

Here γ=1, and the successor value is zero for a terminal state. If V_w equals the true v_π, the conditional expectation of δ_t given (S_t,A_t) equals A_π(S_t,A_t). With an approximate critic, it provides an approximate advantage signal.

Reference: RL-11, pp. 41-58. Use gradient ascent for the policy objective, and do not bootstrap beyond a terminal state.

C1: Returns and an MRP

(a) Sample returns [2 points]

At time 0, the remaining rewards are 1, 1, and 2. At time 1, only 1 and 2 remain:

$$G_0 = 1 + 0.8(1) + 0.8^2(2) = 3.08$$
$$G_1 = 1 + 0.8(2) = 2.6$$

Award 1 point for each return. The reward 2 received on entering T still counts, even though the terminal state's continuation value is zero.

(b) Bellman expectation equations [2 points]

$$v(A) = 1 + 0.8[0.5v(A) + 0.5v(B)]$$
$$v(B) = 2 + 0.8[0.25v(A) + 0.75(0)] = 2 + 0.2v(A)$$

Award 1 point per equation. Average over all possible successor states using their transition probabilities.

(c) Exact solution [2 points]

Substitute the equation for v(B) into the equation for v(A):

$$v(A) = 1 + 0.4v(A) + 0.4[2 + 0.2v(A)]$$
$$v(A) = 1.8 + 0.48v(A) \quad \Rightarrow \quad 0.52v(A) = 1.8$$
$$v(A) = \tfrac{45}{13} \approx 3.4615, \qquad v(B) = \tfrac{35}{13} \approx 2.6923$$

Award 1 point per value. The single sampled return 3.08 need not equal v(A): one trajectory is only one realization of the random process, whereas v(A) is an expectation over possible trajectories.

(d) Synchronous iterative evaluation [2 points]

All successor values in V_0 are zero, so the first sweep gives V_1=(1,2).

$$V_2(A) = 1 + 0.8[0.5(1) + 0.5(2)] = 2.2$$
$$V_2(B) = 2 + 0.8[0.25(1)] = 2.2$$

Thus V_2=(2.2,2.2), in state order (A,B). Award 1 point for each complete vector.

Common mistake

When calculating V_2(B), use V_1(A)=1. Using the newly computed value V_2(A)=2.2 would violate the synchronous-update convention.

References: RL-4, pp. 18-25 and 44-56; RL-5, pp. 8-14.

C2: Induced model and policy improvement

(a) Policy-induced transition and reward model [2 points]

Use state order (X,Y,T). From Y, the probability of reaching X is the probability of selecting retry multiplied by its probability of reaching X: 0.25×0.5=0.125.

P^{π} row	X	Y	T	r^{π}
X	0.5	0.5	0	1
Y	0.125	0	0.875	3
T	0	0	1	0

The rewards are $r^{\pi}(X)=0.5\times0+0.5\times2=1$ and $r^{\pi}(Y)=0.25\times0+0.75\times4=3$. Award 1 point for the matrix and 1 for the reward vector.

(b) Evaluate the original policy [2 points]

$$v_X = 1 + 0.25v_X + 0.25v_Y, \qquad v_Y = 3 + 0.0625v_X$$

$$v_X = 1.75 + 0.265625v_X \quad \Rightarrow \quad v_X = \frac{112}{47}$$

$$v_Y = 3 + \frac{1}{16}\frac{112}{47} = \frac{148}{47}$$

Therefore $v_{\pi}(X)\approx2.3830$ and $v_{\pi}(Y)\approx3.1489$. Award 1 point for the equations and 1 for the solution.

(c) Fix the current action before evaluating it [2 points]

$$q_{\pi}(X, \text{wait}) = 0 + 0.5v_X = \frac{56}{47} \approx 1.1915$$

$$q_{\pi}(X, \text{go}) = 2 + 0.5v_Y = \frac{168}{47} \approx 3.5745$$

$$q_{\pi}(Y, \text{retry}) = 0 + 0.5[0.5v_X] = \frac{28}{47} \approx 0.5957$$

$$q_{\pi}(Y, \text{finish}) = 4$$

Award 0.5 point per action value. Since the current action is already specified, do not multiply by its probability under π. The policy determines subsequent actions.

(d) Improve and then re-evaluate [2 points]

Choose go at X because 168/47>56/47. Choose finish at Y because 4>28/47. Thus π'=(go,finish).

$$v_{\pi'}(Y) = 4, \qquad v_{\pi'}(X) = 2 + 0.5(4) = 4$$

Award 1 point for the policy and 1 for its values. The original $q_{\pi}(X,\text{go})\approx3.5745$ is not the new policy value 4: the new policy also changes the later decision at Y.

References: RL-4, pp. 38-56; RL-5, pp. 28-36.

C3: Gridworld evaluation, sweep by sweep

(a) Update rule [1 point]

$$V_{k+1}(s) = \frac{1}{4} \sum_{a \in \{U, D, L, R\}} [-1 + V_k(f(s, a))]$$

The function f(s,a) gives the successor cell; if the move would leave the grid, f(s,a)=s. Both terminal values stay at zero. The formula applies only to nonterminal cells, with γ=1 already substituted.

(b) First sweep [2 points]

Every successor estimate starts at zero, so all nonterminal values become -1:

V1, row 1	0	-1	-1
Row 2	-1	-1	-1
Row 3	-1	-1	0

Award 1 point for the nonterminal values and 1 for keeping both terminal values at zero.

(c) Second sweep [2 points]

A cell adjacent to a terminal has one successor value 0 and three successor values -1. Its update is -1+(-3)/4=-1.75. The other nonterminal cells have four successor values -1, giving -2.

V2, row 1	0	-1.75	-2
Row 2	-1.75	-2	-1.75
Row 3	-2	-1.75	0

Award 1 point for the four cells adjacent to terminals and 1 for the remaining cells.

(d) Two entries of the third sweep [2 points]

At (1,2), successor values in Up, Down, Left, Right order are -1.75, -2, 0, -2:

$$V_3(1, 2) = -1 + \frac{-1.75 - 2 + 0 - 2}{4} = -2.4375$$

At (2,2), all four successor values are -1.75:

$$V_3(2, 2) = -1 + \frac{4(-1.75)}{4} = -2.75$$

Award 1 point for each correct calculation with its successor values.

(e) Greedy action using V2 [2 points]

At (1,2), the one-step action values in Up, Down, Left, Right order are:

$$(-1 - 1.75,\ -1 - 2,\ -1 + 0,\ -1 - 2) = (-2.75, -3, -1, -3)$$

Choose Left, whose action value -1 is the largest. Award 1 point for the four values and 1 for the action. Remember that -1 is greater than -3.

Reference: RL-5, pp. 8-14 and 28-36. Here k counts complete evaluation sweeps, not environment steps.

D1: MC prediction and online TD(0)

(a) Compute returns backward [2 points]

$$G_2 = 3, \qquad G_1 = 2 + 0.9(3) = 4.7$$
$$G_0 = 1 + 0.9(4.7) = 5.23$$

Award 1 point for G_0 and 1 for both G_1 and G_2.

(b) First-visit MC with a constant step size [2 points]

A first appears at t=0, so use G_0. B first appears at t=1, so use G_1. The second occurrence of A at t=2 does not trigger another update in this part.

$$V(A) = 1 + 0.5(5.23 - 1) = 3.115$$
$$V(B) = 2 + 0.5(4.7 - 2) = 3.35$$

Award 1 point per value.

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

There are two returns for A, 5.23 and 3, and one return for B, 4.7:

$$V_{\text{every}}(A) = \frac{5.23 + 3}{2} = 4.115, \qquad V_{\text{every}}(B) = 4.7$$

Award 1 point per estimate. This part explicitly requests arithmetic means from this episode alone, with no previous samples. Do not reuse the constant-step update from part (b).

(d) Online TD(0) in chronological order [3 points]

First transition A→B, reward 1: use the initial V(B)=2.

$$V(A) \leftarrow 1 + 0.5[1 + 0.9(2) - 1] = 1.9$$

Second transition B→A, reward 2: use the already updated V(A)=1.9.

$$V(B) \leftarrow 2 + 0.5[2 + 0.9(1.9) - 2] = 2.855$$

Third transition A→T, reward 3: the terminal continuation value is zero.

$$V(A) \leftarrow 1.9 + 0.5[3 - 1.9] = 2.45$$

The final estimates are (V(A),V(B))=(2.45,2.855). Award 1 point for each transition update.

(e) Why the estimates differ [1 point]

MC targets complete observed returns; TD targets include the current successor-state estimate. The parts also use different visit-counting and averaging conventions. Estimates need not agree after a single finite episode. Different intermediate results do not establish that the methods must converge to different true values.

Reference: RL-6, pp. 21-28 and 40-43, plus first-visit and every-visit MC.

D2: Four targets and terminal handling

(a) ε-greedy probabilities [2 points]

The exploitation branch has probability 0.7 and chooses a1. The exploration branch assigns 0.3/3=0.1 to each action:

$$P(a_1) = 0.8, \qquad P(a_2) = 0.1, \qquad P(a_3) = 0.1$$

Award 1 point for the greedy-action probability and 1 for the other probabilities and correct total.

(b) SARSA uses the selected action a2 [2 points]

$$y_{\text{SARSA}} = 1 + 0.9Q(s', a_2) = 1 + 0.9(1) = 1.9$$
$$Q_{\text{new}}(s, a) = 2 + 0.5(1.9 - 2) = 1.95$$

Award 1 point for the target and 1 for the updated value. Although a1 has the largest estimate, the behavior policy actually selected a2, so SARSA uses Q(s',a2).

(c) Q-learning uses the maximum next value [2 points]

$$y_{\text{QL}} = 1 + 0.9\max(4, 1, 3) = 4.6$$
$$Q_{\text{new}}(s, a) = 2 + 0.5(4.6 - 2) = 3.3$$

Award 1 point for the target and 1 for the update. Restart from Q(s,a)=2; do not continue from SARSA's result 1.95.

(d) Standard DQN versus Double DQN [3 points]

Standard DQN maximizes the target-network outputs (2,5,3). The maximum is 5 at a2:

$$y_{\text{DQN}} = 1 + 0.9(5) = 5.5$$

Double DQN selects a1 using the online outputs (4,1,3), then evaluates a1 using the target-network value 2:

$$a^* = a_1, \qquad y_{\text{Double}} = 1 + 0.9(2) = 2.8$$

Award 1 point per target and 1 for correctly identifying the evaluated actions and selection/evaluation roles. Argmax returns an action index; max returns a value.

(e) Terminal successor [1 point]

All these targets omit the successor-value term when the transition terminates. The target is 1, and the tabular update is:

$$Q_{\text{new}}(s, a) = 2 + 0.5(1 - 2) = 1.5$$

Award 0.5 point for the target and 0.5 for the updated value.

References: RL-7, pp. 16-18, 34-36, and 55-61; RL-9, pp. 30-32; RL-10a, pp. 3-11.

E1: Finding and verifying an optimal policy

(a) Evaluate π0=(a,d) [2 points]

$$v_{\pi_0}(S) = 1 + 0.5v_{\pi_0}(S) \quad \Rightarrow \quad v_{\pi_0}(S) = 2$$
$$v_{\pi_0}(U) = 1 + 0.5v_{\pi_0}(S) = 2$$

Award 1 point per value. Selecting a forever at S does not terminate, but the discounted return 1+0.5+0.25+… converges to 2.

(b) Improve the policy [3 points]

Fix the current action, then follow π0 afterward:

$$q_{\pi_0}(S, a) = 1 + 0.5(2) = 2, \quad q_{\pi_0}(S, b) = 0 + 0.5(2) = 1$$
$$q_{\pi_0}(U, c) = 3, \quad q_{\pi_0}(U, d) = 1 + 0.5(2) = 2$$

Thus $\pi_1(S)=a$ and $\pi_1(U)=c$. Award 0.5 point per action value and 1 point for the improved policy.

(c) Re-evaluate and verify optimality [3 points]

Under π1, S still repeatedly selects a, so $v_{\pi 1}(S)=2$. U selects c and terminates, so $v_{\pi 1}(U)=3$. Recompute the one-step action values using these new continuation values:

State	First action value	Second action value	Greedy action
S	a: 1+0.5×2=2	b: 0+0.5×3=1.5	a
U	c: 3	d: 1+0.5×2=2	c

The policy remains (a,c), so policy improvement is stable. These values satisfy the Bellman optimality equations. Therefore v*=(2,3) and π*=(a,c). Each state's maximum is strict, so its optimal stationary action choice is unique.

Marking: new policy values 1; action comparisons 1; stability and optimality conclusion 1.

(d) Value iteration [2 points]

$$V_{k+1}(S) = \max\{1 + 0.5V_k(S),\ 0.5V_k(U)\}$$
$$V_{k+1}(U) = \max\{3,\ 1 + 0.5V_k(S)\}$$

Sweep	Candidates at S	Candidates at U	Vector (S,U)
V1	1, 0	3, 1	(1,3)
V2	1.5, 1.5	3, 1.5	(1.5,3)
V3	1.75, 1.5	3, 1.75	(1.75,3)

Award 0.5 point for correct optimality backups and 0.5 for each vector. The temporary tie at S in sweep 2 does not imply a tie under v*. A greedy choice based on an approximate V_k can differ from one based on v*.

References: RL-4, pp. 57-68; RL-5, pp. 28-36 and 48-53.

Use your errors to guide revision

Lecture topics and physical PDF pages	Questions
RL-2 / RL-3: foundations, states, Markov property	A1
RL-4: pp. 18-25, 38-56, 57-68	A2, A3, C1, C2, E1
RL-5: pp. 8-14 evaluation; 28-36 improvement; 48-53 value iteration	A4, C1-C3, E1
RL-6: pp. 21-28 MC vs. TD; 40-43 MC/DP/TD; first-/every-visit MC	B1, D1
RL-7: pp. 16-18 ε-greedy; 34-36 SARSA; 55-61 Q-learning; 23-24 GLIE	A5, A6, B2, D2
RL-8: pp. 9-11 value function approximation	A7
RL-9: pp. 30-32 DQN	B3, D2(d)
RL-10a: pp. 3-11 Double DQN; 14-27 prioritized replay; 32-39 dueling	A8, B3, B4, D2(d)
RL-11: pp. 12-19 stochastic policies; 41-58 REINFORCE, baselines, actor-critic	A9, B5
RL-12: learning and planning / Dyna, supplementary to the photo	A10

The handwritten endpoint of the middle RL-4 range was unclear, so the adjacent Bellman and matrix material on pp. 53-56 was included. The RL-11 notation "41-53-58" was interpreted as the continuous topic range pp. 41-58. Follow the instructor's official scope if it differs.

Six checks before submitting an answer

1. Discount positions: the first reward has coefficient 1, then γ, then γ^2.

2. Probabilities: computing Q fixes the current action; computing V averages over actions using π. Random state transitions still require weighting by P.

3. Update order: synchronous DP uses the old complete vector; online TD uses estimates already updated earlier in the trajectory.

4. Targets: MC uses G; TD uses $r+\gamma V'$; SARSA uses the selected next Q; Q-learning uses the maximum next Q.

5. Deep-Q variants: Double separates selection and evaluation; dueling separates value and advantage streams; prioritized replay changes sampling.

6. Limits: $\varepsilon \to 0$ alone does not establish GLIE; one greedy improvement need not be optimal; a tie under V_2 need not persist under v^*.

Practice record

Score: _____ / 100. Three questions to redo: _____, _____, _____.

Redo them without the solution first. Write either "target → error → updated estimate" or "Bellman equations → solution → action comparison" before checking your work.