中文答案与解析

强化学习期中模拟卷 A

使用方式

本卷是根据你提供的考试通知、复习照片和课件编写的原创模拟题，不是老师发布的真题。先用英文试卷限时 120 分钟，再看本册。题号 A1-E1 与英文卷一一对应。

部分	分值	建议用时
选择题 A1-A10	20	15 分钟
简答题 B1-B5	25	30 分钟
MDP 计算 C1-C3	25	30 分钟
学习算法计算 D1-D2	20	25 分钟
最优策略计算 E1	10	15 分钟
检查	合计 100 分	另留 5 分钟

本次出题重点

照片圈出的 RL-4、RL-5、RL-6、RL-7 内容构成计算题主体；RL-8 的函数近似、RL-9 的 DQN、RL-10 的 Double DQN / 优先经验回放 / Dueling DQN、RL-11 的随机策略与策略梯度进入选择和简答题。

RL-2、RL-3 用于基础概念。RL-12 虽未列在照片中，但你提供了课件且通知写明考前所授全部内容，因此安排了 A10 一道 2 分的学习与规划概念题。是否确已讲完 RL-12，应以课堂进度为准。

先记住计算的顺序

① 看奖励在哪一步收到。② 看下一状态是否终止。③ 确认目标：完整回报、按策略平均、实际下一动作，还是最大动作价值。④ 再代入更新。

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

DP 的同步迭代用上一轮整张表。在线 TD 按时间顺序使用已经更新过的数。题目指定独立计算时，必须重置到题目初值。

评分约定

选择题每题 2 分，不倒扣。简答和计算按列出的要点给分；同义解释、等价公式均可。后续步骤若只继承前面的算术错误，可保留正确方法的分数，避免重复扣同一个错误。

本册列出的页码均是原 PDF 从第 1 页开始的物理页码。照片个别手写范围不够清楚时，按相邻主题一并核对；范围与题目对照见最后一页。

A1-A5：基础、回报与策略评估

答案速查

题号	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
答案	B	C	A	D	B	C	A	D	B	C

A1：B。Markov 不等于“确定性”

给定当前状态和动作后，更早的历史不再提供预测下一状态及奖励所需的额外信息。转移仍可随机。状态可以压缩历史，不需要把完整历史逐项复制。奖励也可以依赖动作。

得分：选 B 得 2 分。对应：RL-3 状态与 Markov 性；RL-4 的 MDP 定义。

A2：C。第一个奖励不乘 γ

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

A 的 5 是未折扣求和。常见错误是把第一个奖励也乘 γ，或忘记第三个奖励的折扣次数为 2。

得分：选 C 得 2 分。对应：RL-4，第 18-25 页。

A3：A。按策略概率平均

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

固定策略 π 下，下一步会按 π 选动作，所以先取各动作价值，再按选中概率加权。最大值属于贪心选择或最优性关系，不能把任意策略的平均直接替换为 max。C 没有概率权重，D 则取了最差动作。

得分：选 A 得 2 分。对应：RL-4，第 42、48-52 页。

A4：D。k 是整轮更新编号

同步更新中，V_{k+1} 的所有位置都从 V_k 读取。B 描述的是原地更新，与本题明确规定的同步更新不同。k 既不是回合编号，也不是环境时间步；含义要看算法语境。

得分：选 D 得 2 分。对应：RL-5，第 8-14 页。

A5：B。探索也可能抽中最优动作

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

其余三个动作各 0.05，总和为 1。0.80 只算了利用分支；0.20 是进入随机探索分支的概率。

得分：选 B 得 2 分。对应：RL-7，第 16-18 页。

A6-A10：控制与深度强化学习

A6：C。实际下一动作与最大值

SARSA 的目标是 r+γQ(s',a')，其中 a' 是行为策略实际选出的下一动作。Q-learning 的目标是 r+γ max Q(s',a')，行为仍可 ε-greedy 探索，因此是异策略控制。二者都不需要已知转移概率模型。

得分：选 C 得 2 分。对应：RL-7，第 34-36、55-61 页。

A7：A。共享参数带来泛化能力

当状态很多或连续时，逐状态存表代价大，且不能有效利用状态间相似性。函数近似用参数 w 表示价值，可从见过的状态推广到未见过的状态。近似会有误差，也不保证所有训练方法收敛。

得分：选 A 得 2 分。对应：RL-8，第 9-11 页。

A8：D。先减去原始优势的平均值

$$\bar{A} = \frac{2-1+0}{3} = \frac{1}{3}$$

$$Q(s,a) = V(s) + A(s,a) - \bar{A}$$

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

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

A 漏减均值。B 使用的是减最大优势的另一种聚合方式，不符合题目指定的 mean-centered 版本。C 忽略了状态价值分支。

得分：选 D 得 2 分。对应：RL-10a，第 32-39 页。

A9：B。随机策略的适用理由

石头剪刀布中，固定可预测的动作容易被对手针对。均匀随机策略可以避免这种可利用性。不能推出“所有完全可观测的有限折扣 MDP 都必须用随机最优策略”；标准条件下存在确定性的平稳最优策略。

得分：选 B 得 2 分。对应：RL-11，第 12-19 页。

A10：C。模型产生模拟经验

Dyna 把真实交互、模型学习和利用模型规划结合起来。模型预测下一状态及奖励，让智能体在两次真实交互之间额外做价值更新。模型并非凭空已知；若模型有误差，也可能影响学习结果。

得分：选 C 得 2 分。对应：RL-12，学习与规划、Dyna 相关内容。此题为照片以外的范围补充。

B1-B2：三种评估方法与 GLIE

B1：DP、MC、TD(0) [5 分]

方法	需要环境模型	使用样本	自举
标准 DP	是	否，求完整期望	是
MC	否	是，完整回报	否
TD(0)	否	是，一步转移	是

自举（bootstrapping）指更新目标中使用另一个当前估计值。DP 和 TD 的目标含 V(s')；标准完整回合 MC 使用实际回报 G，不靠 V(s') 来构成目标。

对固定策略，完整 MC 回报 G_t 是 $v_\pi(S_t)$ 的无偏样本，但沿途多次随机转移和动作通常带来较大方差。实际 TD 目标 $r+\gamma V(s')$ 使用近似 V，通常会带来偏差，但方差往往较低。

若把 V 换成真实 v_π，一步 TD 目标在给定当前状态下同样无偏。“TD 永远有偏”和“方差在任何例子里都更小”都说得过强。这里的比较针对目标估计，并非宣称所有有限步估计器都无偏。

评分：每行三项判断全部正确各 1 分，共 3 分；MC 的偏差/方差说明 1 分；TD 的自举偏差和通常较低方差 1 分。

英文关键表达：DP uses a full model expectation; MC samples complete returns without bootstrapping; TD samples and bootstraps.

B2：GLIE [5 分]

① Infinite exploration：在无限训练的理论极限下，所有相关状态-动作对被访问无穷多次。

② Greedy in the limit：策略最终趋于对当前动作价值估计贪心，非贪心动作的总概率趋于 0。

固定 $\varepsilon=0.1$ 时，即使访问充分，也始终保留非贪心动作的固定概率，不能满足第②条。只有 $\varepsilon_k \to 0$ 也不够，因为可能减得过快，或有些状态很难到达，从而破坏第①条。

MC 是“用完整回报更新价值”的方法；ε-greedy 是“如何选择动作”的规则；GLIE 是“探索策略随训练变化时应满足的性质”。MC 配上 ε-greedy 不自动等于 GLIE。

评分：两条定义各 1 分；固定 ε 的解释 1 分；仅 $\varepsilon \to 0$ 不足的理由 1 分；三者角色区分 1 分。

对应：RL-6，第 21-28、40-43 页；RL-7，第 16-24 页。GLIE 为 ε-greedy 的相邻补充内容。

B3-B4：DQN 及其改进

B3：DQN 与 Double DQN [5 分]

经验回放把历史转移存入缓冲区并抽样训练，减少按时间连续取样带来的相关性，也能重复利用数据。目标网络在一段时间内保持相对固定，让回归目标变化较慢；它不是永远不更新。

非终止转移的目标分别为：

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

$$a^* = \arg\max_{a'} Q(s', a'; \theta)$$

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

Double DQN 用在线网络选动作，用目标网络评价该动作，减轻“在有噪声估计上取最大值”造成的过估计。它并不保证每一个样本的误差都更小，也不保证消除一切偏差。

评分：回放目的 1 分；目标网络目的 1 分；两个目标各 1 分；过估计及选择/评价解耦 1 分。

B4：优先回放与 Dueling [5 分]

比例优先回放通常设 $p_i = |\delta_i| + \eta$，其中 $\eta > 0$ 是小偏移量；采样概率随 p_i 增大而增大，例如 $P(i) = p_i^\kappa / \Sigma_j p_j^\kappa$，$\kappa \geq 0$ 控制优先程度。

较大 TD 误差往往表示这条经验还没有学好。η 防止零误差经验完全失去抽样机会。非均匀采样改变了训练数据分布，因此引入重要性采样权重来校正相对于均匀缓冲区目标的偏差。

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

β=1 对应完整的重要性采样校正；实际可逐渐增加 β，并把权重归一化。这里 κ 是采样优先指数，β 是校正指数，η 是优先偏移，不要与学习率 α 或 ε-greedy 的 ε 混淆。

Dueling 网络分为状态价值 V(s) 和动作优势 A(s,a) 两条输出流。均值中心化的聚合式为：

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

直接写 Q=V+A 会产生分解不唯一的问题：V 加常数、A 减同一常数，Q 不变。中心化约束使有效优势均值为零；原始优势输出整体平移仍会被抵消。这里 V 是聚合分支，不必等同于任意行为策略下的 v_π。

评分：误差决定优先级 1 分；正偏移目的 1 分；重要性权重目的 1 分；两条流 1 分；正确聚合式 1 分。

对应：RL-9，第 30-32 页；RL-10a，第 3-11、14-27、32-39 页。

B5：REINFORCE、基线与 Actor-Critic

无折扣回合任务的更新 [1 分]

题目明确为 γ=1 的 episodic task，避免不同折扣目标记法的差别。用梯度上升增加预期回报：

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

基线作为乘数中的控制变量处理，策略更新不对该基线反向求导。若基线另有可训练参数，可单独更新。

为什么不改变期望梯度 [1 分]

固定状态 s，b(s) 不依赖本次抽到的动作，因此可移到动作求和外：

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

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

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

因此减掉它不改变期望策略梯度。若基线任意依赖动作，上面的推导一般不成立。

基线的目的 [1 分]

选择合适基线可降低梯度估计的方差，使学习更稳定。常用 b(s)=V(s)。此时 G_t-V(S_t) 表示本次结果比该状态的一般水平好多少。不能说“所有基线都必然降低方差”。

Actor 和 Critic [1 分]

Actor 是策略 π_θ，决定如何选动作；Critic 估计 V、Q 或优势，对动作效果提供学习信号。Critic 学习评价，Actor 据此改进策略。

用一步 TD 误差估计优势 [1 分]

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

本题 γ=1。若下一状态终止，$V_w(S_{t+1})$=0。在 V_w=v_π 精确成立时，给定 (S_t,A_t)，TD 误差的条件期望等于 $A_\pi(S_t,A_t)$；用近似 critic 时是近似优势信号。

英文关键表达：The actor updates the policy; the critic estimates values. An action-independent baseline leaves the expected gradient unchanged and can reduce variance.

对应：RL-11，第 41-58 页。终点不自举，策略参数用梯度上升。

C1：MRP 的回报、方程与迭代

(a) 样本回报 [2 分]

从第 0 步看，剩余奖励依次是 1、1、2；从第 1 步看，剩余奖励只有 1、2：

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

每个结果 1 分。注意到达终点前收到的奖励 2 仍然计入回报。

(b) 贝尔曼方程 [2 分]

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

每条方程 1 分。状态 A 的下一状态不确定，要按转移概率平均；不能把其中一个后继值直接拿来代替期望。

(c) 解联立方程 [2 分]

把第二式代入第一式：

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

每个值 1 分。单次样本 G_0=3.08 不必等于期望值 v(A)，因为样本只是众多可能轨迹之一。

(d) 同步迭代 [2 分]

第 1 轮所有后继值都为 0，所以 V_1=(1,2)。

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

因此 V_2=(2.2,2.2)。每轮完整向量 1 分。

易错点：计算 V_2(B) 时必须用 V_1(A)=1，而不能用刚算出的 V_2(A)=2.2。对应：RL-4，第 18-25、44-56 页；RL-5，第 8-14 页。

C2：策略诱导模型与策略改进

(a) 按策略加权转移和奖励 [2 分]

状态顺序为 (X,Y,T)。X 以 0.5 留在 X、0.5 去 Y；Y 选 retry 的概率为 0.25，再乘其去 X 的概率 0.5，得到 0.125。

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

$r^\pi(X)=0.5\times0+0.5\times2=1$；$r^\pi(Y)=0.25\times0+0.75\times4=3$。矩阵 1 分，奖励向量 1 分。

(b) 评估原策略 [2 分]

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

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

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

所以 $v_\pi(X)\approx2.3830$，$v_\pi(Y)\approx3.1489$。方程 1 分，解 1 分。

(c) 动作价值先固定当前动作 [2 分]

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

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

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

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

每个值 0.5 分。这里已经指定当前动作，不能再乘当前动作被 π 选中的概率；后续才跟随 π。

(d) 贪心改进并重新评估 [2 分]

X 选 go，因为 168/47>56/47；Y 选 finish，因为 4>28/47。所以 π'=(go,finish)。

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

策略 1 分，新价值 1 分。原来的 $q_\pi(X,go)\approx3.5745$ 不是新策略价值 4，因为新策略在后续 Y 也改了动作。

对应：RL-4，第 38-56 页；RL-5，第 28-36 页。

C3：格子世界逐轮更新

(a) 规则 [1 分]

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

f(s,a) 为执行动作后的格子；越界时 f(s,a)=s。两个终点始终为 0，不继续按 -1 更新。γ=1 已代入。

(b) 第 1 轮 [2 分]

所有非终点后继价值初始都为 0，所以每个非终点是 -1。

V1：第 1 行	0	-1	-1
第 2 行	-1	-1	-1
第 3 行	-1	-1	0

非终点值 1 分，正确保持终点为 0 得 1 分。

(c) 第 2 轮 [2 分]

邻接一个终点的格子，四个后继价值是一个 0 和三个 -1，因此 -1+(-3)/4=-1.75。其余非终点四个后继均为 -1，因此为 -2。

V2：第 1 行	0	-1.75	-2
第 2 行	-1.75	-2	-1.75
第 3 行	-2	-1.75	0

四个邻终点格子 1 分，其余格子及终点 1 分。

(d) 第 3 轮的两个位置 [2 分]

(1,2) 按上、下、左、右的后继 V_2 值为 -1.75、-2、0、-2：

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

(2,2) 的四个后继 V_2 均为 -1.75：

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

每个结果及其后继代入 1 分。

(e) 用 V2 做一步贪心选择 [2 分]

在 (1,2)，按上、下、左、右分别得到：

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

最大值是 -1，所以选 Left。四个动作值 1 分，动作 1 分。这里比较最大值；负数中 -1 大于 -3。

对应：RL-5，第 8-14、28-36 页。V_k 是经过 k 轮评估的估计；不能把 k 当成当前位置走过的步数。

D1：同一回合，三种更新方式

(a) 从后往前算回报 [2 分]

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

G_0 正确 1 分，G_1 和 G_2 均正确 1 分。

(b) 首次访问 MC，固定步长 [2 分]

A 在 t=0 首次出现，所以只用 G_0；B 在 t=1 首次出现，用 G_1。A 在 t=2 再出现不触发本部分的第二次更新。

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

每个值 1 分。

(c) 每次访问 MC，只有本回合样本的均值 [2 分]

A 贡献两份回报：5.23 和 3；B 贡献一份：4.7。

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

每个值 1 分。本小题题目指定“无历史样本、直接求均值”，不能继续沿用 (b) 的固定 α 更新。

(d) 在线 TD(0)，立即保留更新 [3 分]

第一条 A→B，奖励 1，用初始 V(B)=2：

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

第二条 B→A，奖励 2，必须用刚更新的 V(A)=1.9：

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

第三条 A→T，奖励 3，终点后继价值为 0：

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

最终 (V(A),V(B))=(2.45,2.855)。三条更新各 1 分。

(e) 为什么结果不同 [1 分]

MC 的目标是完整实际回报，TD 的目标包含当前的下一状态估计；此外本题各方法的访问计数和更新方式不同。经过有限的一条轨迹，估计不必相同。不能把“一次结果不同”说成算法必然收敛到不同的真实值。

对应：RL-6，第 21-28、40-43 页及 MC 的首次/每次访问概念。

D2：四种目标与终止处理

(a) ε-greedy [2 分]

利用分支概率为 0.7，全部给 a1；探索分支在三个动作中均分，每个得到 0.3/3=0.1。

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

最优动作概率 1 分，其余概率及总和正确 1 分。

(b) SARSA：使用实际选中的 a2 [2 分]

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

目标 1 分，更新 1 分。虽然 a1 价值最大，实际下一步选的是 a2，SARSA 就把 a2 的后续价值放进目标。

(c) Q-learning：使用最大值 4 [2 分]

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

目标 1 分，更新 1 分。此小题重新从旧值 2 出发，不能从 SARSA 的 1.95 接着更新。

(d) 在线网络与目标网络分别做什么 [3 分]

标准 DQN 在目标网络的 (2,5,3) 上取最大值，选择/评价的是 a2：

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

Double DQN 先在在线网络的 (4,1,3) 上选 a1，再取目标网络对 a1 的评价 2：

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

两个目标各 1 分；明确 DQN 用 a2、Double 用在线选出的 a1 再由目标网络评价，共 1 分。argmax 给的是动作编号，max 给的是数值。

(e) 下一状态终止 [1 分]

所有这些目标都不再加下一状态价值：target=1。表格更新为 2+0.5(1-2)=1.5。目标与更新各 0.5 分。

对应：RL-7，第 16-18、34-36、55-61 页；RL-9，第 30-32 页；RL-10a，第 3-11 页。

E1：完整策略迭代与价值迭代

(a) 初始策略 π0=(a,d) [2 分]

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

每个值 1 分。S 永远选 a 虽然不终止，但回报 1+0.5+0.25+...=2 收敛。

(b) 固定当前动作，后续遵循 π0 [3 分]

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

所以 $\pi_1(S)=a$，$\pi_1(U)=c$。四个动作值各 0.5 分，改进策略 1 分。

(c) 重新评估，并检查是否还会变化 [3 分]

新策略下 S 仍不断选择 a，故 $v_{\pi 1}(S)=2$；U 直接选择 c 到终点，故 $v_{\pi 1}(U)=3$。

状态	第一动作价值	第二动作价值	贪心动作
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

再次贪心改进仍为 (a,c)，策略稳定。它满足贝尔曼最优性方程，因此 v*=(2,3)，π*=(a,c)。本题两个状态的最大值均严格大于另一动作，所以最优平稳动作选择唯一。

评分：新价值 1 分；动作比较 1 分；稳定与最优结论 1 分。

(d) 价值迭代：每轮直接取 max [2 分]

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

轮次	S 的两个候选值	U 的两个候选值	完整向量 (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)

正确最优性更新 0.5 分，三个向量各 0.5 分。第 2 轮在 S 出现临时平局，不代表最终最优策略也平局；用粗略 V_k 做贪心与用 v* 做贪心是两回事。

对应：RL-4，第 57-68 页；RL-5，第 28-36、48-53 页。

按错题决定回看哪里

主题与原 PDF 页码	本卷题号
RL-2 / RL-3：基础、状态、Markov 性	A1
RL-4：18-25；38-56；57-68	A2、A3、C1、C2、E1
RL-5：8-14 策略评估；28-36 策略改进；48-53 价值迭代	A4、C1-C3、E1
RL-6：21-28 TD 与 MC 比较；40-43 MC/DP/TD；另复习首次与每次访问	B1、D1
RL-7：16-18 ε-greedy；34-36 SARSA；55-61 Q-learning；另复习 23-24 GLIE	A5、A6、B2、D2
RL-8：9-11 函数近似的目的	A7
RL-9：30-32 DQN	B3、D2(d)
RL-10a：3-11 Double；14-27 优先回放；32-39 Dueling	A8、B3、B4、D2(d)
RL-11：12-19 随机策略；41-58 REINFORCE、基线、Actor-Critic	A9、B5
RL-12：学习与规划、Dyna（照片以外的补充）	A10

照片中 RL-4 的中段末页手写不够清晰，本卷把相邻的 53-56 页贝尔曼式与矩阵式一并纳入；RL-11 的“41-53-58”按 41-58 的连续主题理解。具体范围仍以老师的正式说明为准。本卷分配用于训练，不代表各章实际考试占比。

判错后，检查以下六件事

1. 折扣位置：第一个奖励不折扣；以后依次乘 γ、γ^2……。

2. 概率位置：固定动作算 Q 时，不再乘当前动作概率；算 V 时才对动作按 π 加权。环境随机转移仍需按 P 加权。

3. 更新顺序：同步 DP 用旧整表，在线 TD 用当前已更新的估计。

4. 目标来源：MC 用 G；TD 用 $r+\gamma V'$；SARSA 用实际 Q'；Q-learning 用 max Q'。

5. 深度 Q：Double 由在线网络选动作、目标网络评价；Dueling 是 V 与 A 两条输出流；PER 改的是回放抽样。

6. 极限与有限步：$\varepsilon\to 0$ 不自动等于 GLIE；一轮贪心改进不一定已最优；V_2 的平局不代表 v^* 也平局。

模拟考后的记录

总分：____ / 100。最需要重做的三题：_____、_____、_____。

建议先不看答案重做错题，必须写出“目标值→误差→新估计”或“贝尔曼方程→求解→动作比较”，再核对解析。