

Answer any THREE questions. Each question is worth 20 marks. Use separate answer books for PART A and PART B. **Gatsby PhD students only:** answer *either* TWO questions from PART A and ONE question from PART B; *or* ONE question from PART A and TWO questions from PART B.

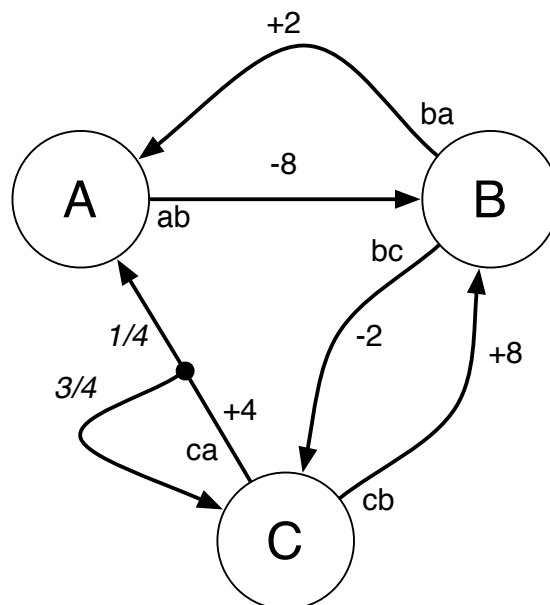
Marks for each part of each question are indicated in square brackets

Calculators are NOT permitted

Part A: Kernel Methods

Part B: Reinforcement Learning

1. Consider the following Markov Decision Process (MDP) with discount factor $\gamma = 0.5$. Upper case letters A, B, C represent states; arcs represent state transitions; lower case letters ab, ba, bc, ca, cb represent actions; signed integers represent rewards; and fractions represent transition probabilities.



- Define the *state-value function* $V^\pi(s)$ for a discounted MDP [1 marks]
- Write down the *Bellman expectation equation* for state-value functions [2 marks]

- Consider the uniform random policy $\pi_1(s,a)$ that takes all actions from state s with equal probability. Starting with an initial value function of $V_1(A) = V_1(B) = V_1(C) = 2$, apply one synchronous iteration of iterative policy evaluation (i.e. one backup for each state) to compute a new value function $V_2(s)$ [3 marks]
- Apply one iteration of greedy policy improvement to compute a new, deterministic policy $\pi_2(s)$ [2 marks]
- Consider a deterministic policy $\pi(s)$. Prove that if a new policy π' is greedy with respect to V^π then it must be better than or equal to π , i.e. $V^{\pi'}(s) \geq V^\pi(s)$ for all s ; and that if $V^{\pi'}(s) = V^\pi(s)$ for all s then π' must be an optimal policy. [5 marks]
- Define the *optimal state-value function* $V^*(s)$ for an MDP [1 marks]
- Write down the *Bellman optimality equation* for state-value functions [2 marks]
- Starting with an initial value function of $V_1(A) = V_1(B) = V_1(C) = 2$, apply one synchronous iteration of value iteration (i.e. one backup for each state) to compute a new value function $V_2(s)$. [3 marks]
- Is your new value function $V_2(s)$ optimal? Justify your answer. [1 marks]

[Total 20 marks]

2. Consider an undiscounted Markov Reward Process with two states A and B . The transition matrix and reward function are unknown, but you have observed two sample episodes:

$A + 3 \rightarrow A + 2 \rightarrow B - 4 \rightarrow A + 4 \rightarrow B - 3 \rightarrow \text{terminate}$

$B - 2 \rightarrow A + 3 \rightarrow B - 3 \rightarrow \text{terminate}$

In the above episodes, sample state transitions and sample rewards are shown at each step, e.g. $A + 3 \rightarrow A$ indicates a transition from state A to state A , with a reward of $+3$.

- Using first-visit Monte-Carlo evaluation, estimate the state-value function $V(A), V(B)$
[2 marks]
- Using every-visit Monte-Carlo evaluation, estimate the state-value function $V(A), V(B)$
[2 marks]
- Draw a diagram of the Markov Reward Process that best explains these two episodes (i.e. the model that maximises the likelihood of the data - although it is not necessary to prove this fact). Show rewards and transition probabilities on your diagram.
[4 marks]
- Define the Bellman equation for a Markov reward process
[2 marks]
- Solve the Bellman equation to give the true state-value function $V(A), V(B)$. Hint: solve the Bellman equations directly, rather than iteratively.
- What value function would batch TD(0) find, i.e. if TD(0) was applied repeatedly to these two episodes?
[2 marks]
- What value function would batch TD(1) find, using accumulating eligibility traces?
[2 marks]
- What value function would LSTD(0) find?
[2 marks]

[Total 20 marks]

3. A rat is involved in an experiment. It experiences one episode. At the first step it hears a bell. At the second step it sees a light. At the third step it both hears a bell and sees a light. It then receives some food, worth +1 reward, and the episode terminates on the fourth step. All other rewards were zero. The experiment is undiscounted.

- Represent the rat's state s by a vector of two binary features, $bell(s) \in \{0, 1\}$ and $light(s) \in \{0, 1\}$. Write down the sequence of feature vectors corresponding to this episode.

[3 marks]

- Approximate the state-value function by a linear combination of these features with two parameters: $b \cdot bell(s) + l \cdot light(s)$. If $b = 2$ and $l = -2$ then write down the sequence of approximate values corresponding to this episode.

[3 marks]

- Define the λ -return v_t^λ

[1 marks]

- Write down the sequence of λ -returns v_t^λ corresponding to this episode, for $\lambda = 0.5$ and $b = 2, l = -2$

[3 marks]

- Using the forward-view TD(λ) algorithm and your linear function approximator, what are the sequence of updates to weight b ? What is the total update to weight b ? Use $\lambda = 0.5, \gamma = 1, \alpha = 0.5$ and start with $b = 2, l = -2$

[3 marks]

- Define the TD(λ) *accumulating eligibility trace* \mathbf{e}_t when using linear value function approximation

[1 marks]

- Write down the sequence of eligibility traces \mathbf{e}_t corresponding to the bell, using $\lambda = 0.5, \gamma = 1$

[3 marks]

- Using the backward-view TD(λ) algorithm and your linear function approximator, what are the sequence of updates to weight b ? (Use offline updates, i.e. do not actually change your weights, just accumulate your updates). What is the total update to weight b ? Use $\lambda = 0.5, \gamma = 1, \alpha = 0.5$ and start with $b = 2, l = -2$

[3 marks]

$$\Delta b_1 = \alpha \delta_1 e_1 = 0.5(0 + -2 - 2)1 = -2$$

$$\Delta b_2 = \alpha \delta_2 e_2 = 0.5(0 + 0 - -2)1/2 = 1/2$$

$$\Delta b_3 = \alpha \delta_3 e_3 = 0.5(1 + 0 - 0)5/4 = 5/8$$

$$\sum \Delta b = (-2 + 1/2 + 5/8) = -7/8$$

[Total 20 marks]

END OF PAPER