BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

WORK INTEGRATED LEARNING PROGRAMME

Course Handout

Part A: Content Design

Course Title	Deep Reinforcement Learning
Course No(s) **Credit Units**	4
Credit Model	2 - 0.5 - 1.5. 1 unit for class room hours, 0.5 unit for Tutorial, 1.5 units for Student preparation. 1 unit = 32 hours
Content Authors	S.P. Vimal
Version	1.0

Date 17, May, 2023

Course Objectives

C01: Understand

a. the conceptual, mathematical foundations of deep reinforcement learning b. various classic & state-of-the-art Deep Reinforcement Learning algorithms

C02: Implement and Evaluate the deep reinforcement learning solutions to various problems like planning, control, and decision making in various domains

C03: Provide conceptual, mathematical, and practical exposure on DRL

c. to understand the recent developments in deep reinforcement learning and

d. to enable modeling new problems as DRL problems.

Text Book(s)

T1	Reinforcement Learning: An Introduction, Richard S. Sutton and Andrew G. Barto, Second Ed. , MIT Press
T2	Foundations of Deep Reinforcement Learning: Theory and Practice in Python (Addison-Wesley Data & Analytics Series) 1st Edition by Laura Graesser and Wah Loon Keng

Reference Book(s) & other resources

R1	Deep Reinforcement Learning, Aske Plaat, Springer Nature (2022), [Available from: https://arxiv.org/pdf/2201.02135.pdf]

R2	[Links to related research articles inserted as hyperlinks against some topics]

Content Structure

1. Introduction: Introducing RL
    1.1. Introduction to Reinforcement Learning (RL); Examples; Elements of Reinforcement Learning (Policy, Reward, Value, Model of the environment) & their characteristics; Example: RL for Tic-Tac-Toe; Historical Background;
    1.2. Multi-armed Bandit Problem - Motivation and Problem Statement; Incremental solution to the stationary & non-stationary MAB problems; Exploration vs. Exploitation tradeoff; Bandit Gradient Algorithm as Stochastic Gradient Ascent; Associative Search
2. MDP: Framework
    2.1. (Finite) Markov Decision Processes: Modelling Agent-Environment interaction using MDP; Examples; Discussion on Goals ,
    2.2. Rewards & Returns; Policy and Value Functions;
    2.3. Bellman Equation for value functions;
    2.4. Optimal Policy and Optimal Value functions;
3. Approaches to Solving Reinforcement Problems
    3.1. Dynamic Programming Solution (Policy Iteration; Value Iteration; Generalized policy iteration; Efficiency of Dynamic Programming)
    3.2. Monte Carlo (MC) Methods (MC prediction, MC control, incremental MC.)

[Mid-Semester Exam]

3.3. Temporal-Difference (TD) Learning
3.4. Discussion on Other Classic Approaches that combine 3.1, 3.2,3.3 4. Discuss the classification of (Deep) Reinforcement Learning Approaches, Algorithms, and Applications
    4.1. Model-Based vs. Model Free;
    4.2. Value-based vs. Policy-Based;
    4.3. On-Policy vs. Off-Policy;
    4.4. Deep Learning as a Function Approximator and Review of Related Literature
5. Value-Based DRL Methods
    5.1. Function approximation; Feature Construction for Linear Methods (Tile Coding, Asymmetric Tile Coding);
    5.2. Linear function approximation; Semi-Gradient TD methods; Off-policy function approximation TD divergence;
    5.3. Deep Q Network; Double DQN; Rainbow

6. Policy Gradients Methods
    6.1. Policy Gradient Methods, Policy Gradient Theorem,
    6.2. REINFORCE algorithm, REINFORCE with baseline algorithm, 6.3. Actor-Critic methods, REINFORCE algorithm for continuing problems (problems without episode boundaries)

7. Model-Based Deep RL
    7.1. Upper-Confidence-Bound Action Selection,

7.2. Monte-Carlo tree search,
7.3. AlphaGo Zero, MuZero, PlaNet

8. Imitation Learning
   8.1. Introduction to Imitation Learning;
   8.2. Imitation Learning Via Supervised Learning, Behavior Cloning, Inverse Reinforcement Learning,
   8.3. GAIL; Dataset augmentation, DAGGER;
   8.4. Applications in autonomous Driving, Game Playing, and Robotics;

9. (Optional Content) Multi-Agent RL
   9.1. Understanding multi-agent environment;
   9.2. Cooperative vs Competitive agents, centralized vs. decentralized RL ; 9.3. Proximity Primal Optimization (Surrogate Objective Function, Clipping) ; Multi-agent PPO

10. (Optional Content) Special Topics
   10.1. Discussion at a high level on a few selected topics from - Safety In Reinforcement Learning: Constrained RL, Safe Exploration, Adversarial Training, Corrigibility, Distributional Shift, Human-in-the-Loop, Formal methods in Safe RL, Offline/Batch Reinforcement Learning

11. Course Summary

Learning Outcomes

After successfully completing this course, the students will be able to

LO-1: understand the fundamental concepts of reinforcement learning (RL), and algorithms and apply them to solving problems, including control, decision-making, and planning.

LO-2: Implement DRL algorithms, and handle challenges in training due to stability and convergence

LO-3: evaluate the performance of DRL algorithms, including metrics such as sample efficiency, robustness, and generalization.

LO-4: understand the challenges and opportunities of applying DRL to real-world problems & model real-life problems

Part B: Learning Plan

Academic Term	
Course Title	Deep Reinforcement Learning
Course No	
Lead Instructor	

Detailed Plan of Contact Sessions

Session #	# of Hrs	Topics	Ref

1	2	Introduction to Reinforcement Learning (RL); Examples; Elements of Reinforcement Learning (Policy, Reward, Value, Model of the environment) & their characteristics; Example: RL for Tic-Tac-Toe; Historical Background;	[TB] Ch-1
2-3	3	Multi-armed Bandit Problem - Motivation and Problem Statement; Incremental solution to the stationary & non-stationary MAB problems; Exploration vs. Exploitation tradeoff; Bandit Gradient Algorithm as Stochastic Gradient Ascent; Associative Search	[TB] Ch-2
3-4	2	(Finite) Markov Decision Processes: Modelling Agent-Environment interaction using MDP; Examples; Discussion on Goals , Rewards & Returns; Policy and Value Functions; Bellman Equation for value functions; Optimal Policy and Optimal Value functions;	[TB] Ch-3
4-6	3	Introduction to Dynamic Programming; Policy Iteration; Value Iteration; Generalized policy iteration; Efficiency of Dynamic Programming	[TB] Ch-4
7-8	2	MC prediction, MC control, incremental MC.	[TB] Ch-5
9-10	3	TD Prediction, On-policy temporal difference model-free (SARSA), Off-policy temporal difference model-free (Q-Learning), Expected SARSA, N-step TD policy evaluation, N-step Sarsa, Off-policy n-step Sarsa, Tree backup algorithm	[TB] Ch-6,7
11	1	Discussion on the classification of (Deep) Reinforcement Learning Approaches, Algorithms, Applications: Model-Based vs. Model Free; Value based vs. Policy-Based; On-Policy vs Off-Policy;	Notes
Mid-Semester Exam			

12	3	Function approximation; Feature Construction for Linear Methods (Tile Coding, Asymmetric Tile Coding); Linear function approximation; Semi-Gradient TD methods; Off-policy function approximation TD divergence; Deep Q Network; Double DQN; Rainbow	[TB-2] Ch-9, [DQN] [DDQN] [Rainbow]
13	3	Policy Gradient Methods, Policy Gradient Theorem, REINFORCE algorithm, REINFORCE with baseline algorithm, Actor-Critic methods, REINFORCE algorithm for continuing problems (problems without episode boundaries)	[TB-2] Ch-13,
14-15	2	Upper-Confidence-Bound Action Selection, Monte-Carlo tree search, AlphaGo Zero, MuZero, PlaNet	[Alphazero] [AlphaGo Zer o] [MuZero] [PlaNet]

15-16	3	Imitation Learning: Introduction, Imitation Learning; Imitation Learning Via Supervised Learning, Behavior Cloning, Inverse Reinforment Learning, GAIL; Dataset augmentation , DAGGER; Applications in autonomous Driving, Game Playing, Robotics;	[DeepMimic] [BAIL] [ACM-SUR-IL]
[optional]	3	Understanding multi-agent environment; cooperative vs competitive agents, centralized vs. decentralized RL ; Proximity Primal Optimization (Surrogate Objective Function, Clipping) ; Multi-agent PPO	[MARL]
[optional]	1	Discussion on few selected advanced topics; Course Summary.	[NAS -1] [NAS-2] [SafeRL-Sur] [OfflineRL]

[RL - Reinforcement Learning, DRL: Deep Reinforcement Learning]

Detailed Plan for Lab work

Lab No.	Lab Objective Lab Sheet Access URL Session Reference
1	Bandit Gradient Algorithm NA
2-4	Implementing Dynamic programming, Monte carlo, SARSA, Q-Learning; NA
3	Deep Q Network NA
4	REINFORCE NA

5 Imitation Learning NA

Evaluation Scheme:

Legend: EC = Evaluation Component; AN = After Noon Session; FN = Fore Noon Session

No	Name	Type	Duration Weight Schedule Remarks
	EC-1(a) Quizzes	Online	5 % TBA Two quizzes of 5% will be conducted, and the score of highest will be taken towards grading; Strictly no make-up;
	Assignments	Take Home	25% TBA

EC-2	Mid-Semester Test	Closed Book	30% TBA
EC-3	Comprehensive Exam	Open Book	40% TBA

Note:
Syllabus for Mid-Semester Test (Closed Book): Topics in Session Nos. 1 to 8
Syllabus for Comprehensive Exam (Open Book): All topics (Session Nos. 1 to 16)

Important links and information:

Elearn portal: https://elearn.bits-pilani.ac.in or Canvas
Students are expected to visit the Elearn portal on a regular basis and stay up to date with the latest announcements and deadlines.

Contact sessions: Students should attend the online lectures as per the schedule provided on the Elearn portal.

Evaluation Guidelines:

1 EC-1 consists of two Quizzes. Students will attempt them through the course pages on the Elearn portal. Announcements will be made on the portal, in a timely manner.

2 EC-2 consists of either one or two Assignments. Students will attempt them through the course pages on the Elearn portal. Announcements will be made on the portal, in a timely manner.

3 For Closed Book tests: No books or reference material of any kind will be permitted. 4 For Open Book exams: Use of books and any printed / written reference material (filed or bound) is permitted. However, loose sheets of paper will not be allowed. Use of calculators is permitted in all exams. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.

5 If a student is unable to appear for the Regular Test/Exam due to genuine exigencies, the student should follow the procedure to apply for the Make-Up Test/Exam which will be made available on the Elearn portal. The Make-Up Test/Exam will be conducted only at selected exam centres on the dates to be announced later.

Plagiarism Policy:

All submissions for graded components must be the result of your original effort. It is strictly prohibited to copy and paste verbatim from any sources, whether online or from your peers. The use of unauthorized sources or materials, as well as collusion or unauthorized collaboration to gain an unfair advantage, is also strictly prohibited. Please note that we will not distinguish between the person sharing their resources and the one receiving them for plagiarism, and the consequences will apply to both parties equally.

In cases where suspicious circumstances arise, such as identical verbatim answers or a significant overlap of unreasonable similarities in a set of submissions, will be investigated, and severe punishments will be imposed on all those found guilty of plagiarism.