Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division

Digital Learning Handout

Part A: Content Design

Course Title	Deep Reinforcement Learning
Course No(s)	AIML* ZG512
Credit Units	4
Credit Model	1.25 - 1.5 - 1.25 1 unit for classroom hours, 0.5 unit for Tutorial, 1.5 units for Student preparation. 1 unit = 32 hours
Course Author	Vimal S P
Version No:	3.0
Date:	21/10/2025

Course Description

Introduction and applications. Markov decision processes(MDP), Tabular MDP planning, Tabular RL policy evaluation, Q-learning, model-based RL, deep RL with function approximation, policy search, policy gradient, fast learning, applications in game playing, imitation learning, RL for neural architecture search, batch RL

Course Objectives

No	Course Objective
CO1	Understand the conceptual and mathematical foundations of Reinforcement Learning and its deep learning extensions
CO2	Explain and analyse classical and state-of-the-art Deep Reinforcement Learning algorithms for value-based and policy-based learning.
CO3	Design, implement, and evaluate DRL solutions for planning, control, and sequential decision-making tasks using standard frameworks and environments.
CO4	Identify, model, and investigate new problems as DRL tasks, recognising current research trends and limitations

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

Learning Outcomes: Students will be able to

LO1	Understand the fundamental concepts of reinforcement learning (RL) and algorithms and apply them to solving problems, including control, decision-making, and planning
LO2	Implement DRL algorithms, and handle challenges in training due to stability and convergence
LO3	Evaluate the performance of DRL algorithms, including metrics such as sample efficiency, robustness, and generalization
LO4	Understand the challenges and opportunities of applying DRL to real-world problems & model real-life problems

Modular Content Structure

1. Introduction: Introducing RL
    - 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;
    - Multi-armed Bandit Problem - Motivation and Problem Statement; Incremental solution to the stationary & non-stationary MAB problems; Exploration vs. Exploitation trade off; Bandit Gradient Algorithm as Stochastic Gradient Ascent; Associative Search
2. MDP: Framework
    - (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;
3. Approaches to Solving Reinforcement Problems
    - Dynamic Programming Solution (DP) (Policy Iteration; Value Iteration; Generalized policy iteration; Efficiency of Dynamic Programming)
    - Monte Carlo (MC) Methods (MC prediction, MC control, incremental MC.)
    - Temporal-Difference (TD) Learning
    - Discussion on Other Classic Approaches that combines DP, MC, TD
4. Discuss the classification of DRL Approaches, Algorithms, and Applications
    - Model-Based vs. Model Free;
    - Value-based vs. Policy-Based;
    - On-Policy vs. Off-Policy;
    - Deep Learning as a Function Approximator and Review of Related Literature

5. Value-Based DRL Methods
    - 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; Duelling networks, Prioritized Replay, Rainbow
6. Policy Gradients Methods
    - Policy Gradient Methods, Policy Gradient Theorem,
    - REINFORCE algorithm, REINFORCE with baseline algorithm,
    - Actor-Critic methods (A2C, A3C), REINFORCE algorithm for continuing problems (problems without episode boundaries)
    - Proximal Policy Optimization
    - Entropy regularization
    - Continuous Control Algorithms - Deterministic Policy Gradient, Soft SAC, DDPG, TD3
7. Model-Based Deep RL
    - Upper-Confidence-Bound Action Selection,
    - Monte-Carlo tree search,
    - AlphaGo Zero, MuZero, PlaNet
8. Imitation Learning
    - Introduction to Imitation Learning;
    - Imitation Learning Via Supervised Learning, Behaviour Cloning, Inverse Reinforcement Learning
    - GAIL; Dataset augmentation, DAGGER;
    - Applications in autonomous Driving, Game Playing, and Robotics;
9. (Optional Content) Multi-Agent RL
    - Understanding multi-agent environment;
    - Cooperative vs Competitive agents, centralized vs. decentralized RL;
    - Proximity Primal Optimization (Surrogate Objective Function, Clipping); Multi-agent PPO
10. (Optional Content) Special Topics
    - 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

Part B: Learning Plan

Contact Session	List of Topic Title	Sub-Topics	Reference
1	**Introduction: Introducing RL**	• 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	T1 Chapter-1
2	**Introduction: Multi-armed Bandit Problem**	• Multi-armed Bandit Problem - Motivation and Problem Statement • Incremental solution to the stationary & non-stationary MAB problems • Exploration vs. Exploitation trade-off • Bandit Gradient Algorithm as Stochastic Gradient Ascent • Associative Search	T1 Chapter-2
3	**MDP: Framework**	• 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	T1 Chapter-3
4	**Approaches to Solving Reinforcement Problems : Dynamic Programming**	• Introduction to Dynamic Programming • Policy Iteration • Value Iteration • Generalized policy iteration • Efficiency of Dynamic Programming	T1 Chapter-4
5	**Approaches to Solving Reinforcement Problems : Monte Carlo Method - I**	• Monte Carlo Methods I (On-policy) • Monte Carlo prediction (first-visit, every-visit) • Monte Carlo control (exploring starts, ε-soft policies)	T1 Chapter-5
6	**Approaches to Solving Reinforcement Problems : Monte Carlo Method - II**	• Monte Carlo Methods II (Off-policy) • Importance sampling (ordinary & weighted) • Off-policy prediction & control • Link between MC and TD	T1 Chapter-5

7	**Approaches to Solving Reinforcement Problems : Temporal Difference Learning I**	• Temporal Difference Learning • TD(0), SARSA, Q-Learning, Expected SARSA	T1 Chapter-6, 7
8	**Temporal Difference Learning II & Discuss the classification of DRL Approaches, Algorithms, and Applications**	• Temporal Difference Learning - n-step returns, TD(λ) • Taxonomy: 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	T1 Chapter-6, 7 , Notes
9	**Value-Based DRL Methods**	• 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; TD Gammon	T1 Chapter-9, 16
10	**Value-Based DRL Methods**	• Deep Q learning -DQN architecture, target networks, replay buffer • Instability and fixes	T2 Chapter-4, DQN Research Paper
11	**Value-Based DRL Methods**	• Extensions of DQN - Double DQN, Duelling networks • Prioritized replay • Rainbow	T2 Chapter-5
12	**Policy Gradients Methods**	• Policy Gradient Theorem • REINFORCE algorithm, REINFORCE with baseline algorithm • Actor-Critic methods • REINFORCE algorithm for continuing problems (problems without episode boundaries)	T2 Chapter-13
13	**Policy Gradients Methods**	• Advanced Policy Optimization • A2C/A3C • Proximal Policy Optimization (PPO) • Entropy regularization	T2 Chapter-6, 7, Notes
14	**Policy Gradients Methods**	• Continuous Control Algorithms • Deterministic Policy Gradient	Notes

		• DDPG, TD3 • Soft Actor–Critic (SAC)	
15	**Model-Based Deep RL**	• Model-Based & Imitation Learning • AlphaGo Zero, MuZero, PlaNet • Dreamer (latent world models)	Research Paper: [AlphaGoZero] [AlphaGo] [MuZero] ,[PlaNet]
16	**Imitation Learning**	• Imitation learning • Behavior Cloning, DAGGER, GAIL • Introduction to Multi-Agent RL ; Safe RL	[Deep Mimic] [BAIL] [ACM-SUR-IL]

Experiential Learning Components:

1. Lab work: 6
2. Project work: 0
3. Case Study: 0
4. Simulation: 0
5. Work Integrated Learning Assignment- 2 Assignments
6. Design work/ Field work: 0

Objective of Experiential Learning Component:
Learners will implement traditional reinforcement learning and deep learning techniques.

Scope of Experiential Learning Component:
Programming language – Python
Tools and libraries: Open AI Gym toolkit, Pytorch, Tensorflow, Keras, Numpy

Lab Infrastructure:
Google Colab

List of Experiments:

Lab No	Lab Objective	Session Reference
1	Implementing Bandit gradient Algorithm	2
2	Implementing Dynamic programming	3
3	Implementing Q-Learning	7
4	Implementing DQN & DDQN	10,11
5	Implementing Policy Gradient algorithms – REINFORCE, Actor Critic	12,13
6	Implementing Imitation learning algorithms	16

Evaluation Scheme:

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

Evaluation Component	Name (Quiz, Lab, Project, Mid-term exam, End semester exam, etc.)	Type (Open book, Closed book, Online, etc.)	Weight	Duration	Day, Date, Session, Time
EC – 1*	Quiz I	Online	5%	1 day	To be announced
	Lab Assignment I	Online	10-15%	10 days	To be announced
	Lab Assignment II	Online	10-15 %	10 days	To be announced
EC - 2	Mid-Semester Test	Closed Book	30%	2 hours	To be announced
EC - 3	Comprehensive Exam	Open Book	40%	2 ½ Hours	To be announced

EC1* (30%): Quiz: 5 %, Lab Assignment/Assignment: 25%
Syllabus for Mid-Semester Test (Closed Book): Topics in Contact session: 1 to 8
Syllabus for Comprehensive Exam (Open Book): All topics

Important Links and Information:

eLearn Portal: https://elearn.bits-pilani.ac.in
Students must visit the eLearn portal regularly and stay updated 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 either one or two Assignments and Quizzes. Two quizzes of 5% may be conducted, and the score of highest may be taken towards grading. Students will attempt them through the course pages on the eLearn portal. Announcements will be made on the portal in a timely manner.
2. For Closed Book tests: No books or reference material of any kind will be permitted.
3. For Open Book exams: "open book" means text/ reference books (publisher copy only) and does not include any other learning material. No other learning material will be permitted during the open book examinations. For Detailed Guidelines refer to the attached document.
   EC3 Guidelines
4. 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.

It shall be the responsibility of the individual student to be regular in maintaining the self-study schedule as given in the course handout, attend the online lectures, and take all the prescribed evaluation components such as Assignments/Quizzes, Mid-Semester Tests and Comprehensive Exams according to the evaluation scheme provided in the handout.

****