Indian Institute Of Technology Patna
Department of Computer Science and Engineering
CS-6131 Reinforcement Learning
End Semester Examination (2024)

27th November 2024

Full Marks: 100
Duration: 3 hours

Instruction

Make reasonable assumptions as and whenever necessary. You can answer the questions in any sequence. **However, answers to all the parts of any particular question should appear together**. Markings will be based on the correctness and soundness of the answers. Proper indentation and appropriate comments (if necessary) are mandatory.

stochastic policy
fails on large state space
maxima in ~~[illegible]~~ optimi
slow convergence

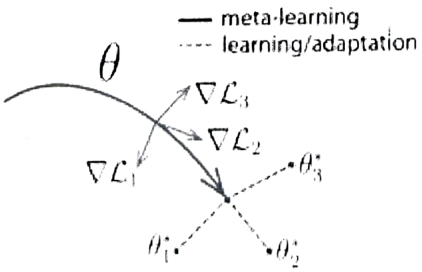
1.
   a. "The Optimal Policy from the value-based method is Deterministic." Justify your answer. (**6 Marks**)
   b. (i) Valuc-bascd rcinforccment learning (RL) algorithms, such as Q-learning, often face several challenges. Discuss the key pitfalls associated with value-based methods in reinforcement learning. (**5 Marks**) (ii) How does overestimation bias manifest in Q-learning, and what is its potential impact on the learned policy? Provide a detailed explanation and solution. (**5 Marks**)
   c. Derive the objective function of Policy Gradient. (**4 Marks**)
   d. Write down the steps of the REINFORCE algorithm. What is the intuitive reasoning behind the high variance in REINFORCE? (**6 Marks**)
   e. Does adding a baseline introduce bias in REINFORCE? How can we combine baseline and causality? (**4 Marks**)
   f. Justify "The Reinforce Estimators known to have high variance" (**3 Marks**)
   g. Write short notes on variance reduction methods. When can we combine two variance reduction methods together? (**4 +2 =6 Marks**)

2. a. Elucidate the goal of the actor-critic algorithm by incorporating a value baseline function expressed through mathematical derivations (**5 Marks**) b. Discuss the optimal neural network configuration choice for critic design (**3 Marks**) c. Describe the Actor-Critic Algorithm. What is an advantage function? (**3+2=5 Marks**) d. What is the difference between Asynchronous Advantage Actor-Critic (A3C) and Synchronous Advantage Actor-Critic (A2C)? (**3 Marks**) e. Discuss the advantages of using an Actor-Critic algorithm compared to pure policy-based and value-based methods. (**3 Marks**) f. "Correlation Problem can not be solved in Policy Gradient Algorithm but can be solved with Actor-Critic." Justify your answer (**5 Marks**)

5 2. (a)What are the limitations of Instruction Fine-Tuning? What is RLHF? With an example, describe its advantage over the traditional automated reward function. (**5 Marks**) b. Explain with a diagram how modern RLHF works. (**3 Marks**) c. In what specific scenario could Reinforcement Learning from Human Feedback (RLHF) be applied to enhance the performance of ChatGPT, particularly in optimizing its ability to generate contextually accurate and user-aligned responses? (**5 Marks**) 13

3. a.Explain the concept of prompt engineering and discuss its significance in fine-tuning the performance of large language models (LLMs) (**3 Marks**). b. Explain the concept of "options" in the context of HRL. How are options defined, and how do they influence the learning process? (**3 Marks**) c. You are tasked with developing a prompt to generate a detailed scientific explanation ~~about Machine~~ Translation. Describe the structure of the prompt you would use for the following scenarios: (i) Zero-shot Prompting (ii) Few-shot Prompting (**5 Marks**). d. Imagine you are working on a project that involves training a generative AI model to generate synthetic medical data for research purposes. The data must resemble real patient data, but it must adhere to strict privacy regulations like HIPAA (Health Insurance Portability and Accountability Act). The generative model should maintain statistical consistency with real data while ensuring that no personally identifiable information (PII) is ever revealed. (i) How would you design the architecture of a generative model to generate synthetic medical data while ensuring privacy and compliance with regulations like HIPAA? (ii) Discuss the potential risks and challenges involved in using generative AI for this purpose and how they can be mitigated. (**6 Marks**) 17

4. a.Explain the key components of the Model-Agnostic Meta-Learning (MAML) algorithm. What is the significance of second-order derivatives in MAML and how does it enable the fast adaptation of models to new tasks? (**5 Marks**) b.In the context of meta-learning, explain the following diagram: (**2 Marks**)


— meta-learning
---- learning/adaptation
θ
$\nabla\mathcal{L}_3$
$\nabla\mathcal{L}_2$
$\nabla\mathcal{L}_1$
θ_3^*
θ_1^*
θ_2^*