Reinforcement Learning: Tutorial 8 (revision)
(week from 23. 3. 2015)

This sheet contains a selection of exam questions from previous years. Please check also questions of earlier tutorials.

MABs

- Explain the ε-greedy action selection method with respect to the multi-arm bandit (MAB) problem.

- What do you understand by the term ‘optimistic initialisation’, and what is its effect on the learning process in the MAB problem?

- Explain the concept of regret as a measure of online algorithms. Does the ε-greedy method achieve zero regret? Why?

MDPs

- Explain what is mixing time, with reference to MDPs

- How can you characterise the asymptotic behaviour of an MDP?

- How does a semi-Markov Decision Process (SMDP) differ from an MDP? Explain, using expressions for the reward, state transition law and cost criterion and an example of a problem domain where this formulation is beneficial.

Value functions and policies

- What is the purpose of an exponentially weighted average of a one-dimensional time series?

- Explain, with suitable expressions for value function updating, the policy improvement step in a general policy iteration method for dynamic programming. Why is it needed?

- Explain, in conceptual terms and with an example, the difference between first-visit and every-visit computations of returns for the Monte Carlo policy evaluation method. In statistical terms, what precisely is the every-visit method computing?

- What are the forward and backward interpretations of the TD(λ) algorithm? How do they differ, and how do they aid analysis and/or implementations?
  What is the difference between the accumulating traces and replacing traces versions of TD algorithms? Using a suitable example MDP, explain why one of these might be preferred over the other.

- What are the advantages and disadvantages of TD(0), Monte Carlo and TD(λ) learning, and when would you prefer to use each?

Reinforcement learning (general)

- Reinforcement learning systems do not need to be “taught” by knowledgeable “teachers”; they learn from their own experience. But teachers of various types can still be helpful. Describe three different ways in which a teacher might facilitate learning. For each, give a specific example scenario and explain what makes learning more efficient.

- Explain, with suitable diagrams and equations, how you might use the actor-critic architecture to solve an infinite horizon control problem. What type of reward would you use? In this setting, what are the benefits of separating the actor and critic components?

- Describe (using symbols and pseudocode) the SARSA and Q-learning algorithms. Give a description both of the difference in optimal policy and of the difference in computational steps. In some applications, it is empirically observed, although not theoretically justified, that SARSA converges faster than Q-learning. Describe possible reasons for this effect.

- Explain why using afterstates can be useful.

Applications

- What are the generalisation issues that one faces in applying reinforcement learning to real-world problems? Illustrate your answer with an example problem of your choice.

- Consider the problem faced by a new generation of smart-phones in learning to use appropriate prompts, from a pre-defined set, with the goal of achieving user satisfaction in a bank of user-interface tasks. The user only gives feedback in terms of satisfaction with the overall interaction, without any additional critique. Describe the structure of an on-policy and off-policy solution to the learning problem. What are the pros and cons of each approach?

- A computer manufacturing company can be in one of two states. In state *hot*, it has a successful product that sells well. In state *cold*, the product sells poorly. In state *hot*, the company can advertise its product and receive an immediate reward of 4 units, and the transition probabilities are $P_{adv}(hot,hot)=0.8$ and $P_{adv}(hot,cold)=0.2$. If it does not advertise, the immediate reward of 6 units and the transition probabilities are $P_{notadv}(hot,hot)=0.5$ and $P_{notadv}(hot,cold)=0.5$. While in state cold, the company can do research to improve its product, in which case the immediate reward is -5 and the transition probabilities are $P_{res}(cold,hot)=0.7$ and $P_{res}(cold,cold)=0.3$. If in state cold, and no research is carried out, immediate reward is -3 and transition probabilities are $P_{notres}(cold,hot)=0.4$ and $P_{notres}(cold,cold)=0.6$.
  Consider the infinite horizon discounted version of this decision problem. Give a graphical depiction of this decision process in terms of its transition structure.
  Give the complete Bellman optimality equations for this problem - enumerating the function for different states and actions. What is the value iteration backup for the state hot?
  Assuming a direct experience of the form: hot, adv, 4, cold, and a table lookup value function, give the explicit TD(0) backup for the state hot.

- Consider the problem of synthesizing a strategy for profitable trading in an electronic market - the state space consists of cash positions and holdings of stocks, the action space consists of the amounts of sale/purchase in these stocks (by transacting in cash). To make matters concrete, imagine you are dealing in k stocks, and you define profitability/performance over a horizon of M discrete time steps. Using explicit symbols for all variables and appropriate equations, explain how you could pose this problem of learning a trading strategy as an on-policy learning problem. Explain why (with at least one specific example scenario) off-policy methods might yield very poor quality strategies in this domain.

- The problem of learning for legged locomotion, modelled in terms of multilinked pendulum systems, can be posed and solved using the actor-critic architecture. Explain, with block diagrams and/or equations, how you would do this.

- Consider the game of Tic-Tac-Toe (or any board game with a finite number of moves). How do you define the state space and state value function for such a game? One could define an afterstate in terms of board positions after the agent has made its move. How does this impact on the information required to calculate a value function, and what might be the advantage of such a scheme - explain with a simple example scenario.

- Suppose you have the task of getting a mine-detection system to plot a safe path through a minefield. The mines are reasonably easily detectable and you have a few small, relatively cheap (and therefore, to some extent, disposable) robots that you can use to aid you. Discuss how you might use reinforcement learning to find a safe path.

- Your company is employed by a market gardener growing pumpkins. The gardener wishes to control the amount of water supplied by the water sprinklers so that the pumpkins grow well. Information such as soil humidity, air temperature, weather forecast, opening of sprinkler valves, etc. is available. How would you set this up as a reinforcement learning problem? Explain particularly why training data is quite hard to get and how you would deal with this fact.

Function approximation for RL

- Explain the process of policy evaluation with function approximation, by giving expressions for the Mean Squared Error performance measure and a gradient descent parameter estimation iteration. What are the ‘training data’, which act as input to the function approximator - how are they generated, and what is the key requirement on the data?

- What is a feature vector and how may one be used in the representation of a state? Why would one choose to use a feature-vector representation of a state?

- There are a number of counter-examples highlighting the difficulty of off-policy learning with function approximation. What is the essential source of difficulty for these methods? Explain using suitable equations.

POMDP

- What is a belief state in the context of Partially Observable Markov Decision Processes (POMDP)? How does one modify the Bellman equation for the optimal value function to accommodate this concept?

- Explain, with equations and/or figures, the value iteration process for a POMDP (assuming access to observations before making actions).

- Explain, in outline form, the MC-POMDP approximation procedure.

- In problem domains such as spatial navigation, two major weakness of the general (PO)MDP formulation are that (i) there is no natural way to describe metrics in the environment and (ii) (PO)MDPs conflate properties of the environment with properties of the agent. Using concrete application examples, explain why these are serious limitations, and how they impede efficient learning.

- What do you understand by the term “particle filter”? Explain - with equations - how this can be used to define an approximate solution method for POMDPs.

A few general remarks

As you may expect, most of the questions of this year's exam will not be found on this sheet. Further- more, because the content of the course has changed somewhat over the years, some of the questions would not be posed in the same way today. In some cases questions aimed at combinations of concepts such that the classification I have tried above is not always accurate. All question that can occur in the exam should be answerable based on the Lecture slides and the Sutton & Barto book.

For a successful exam please keep in mind:

- Make sure that you address all sub-questions. In many cases, if there is a maximum of k marks, your answer should contain k facts.
- There is no problem with brief answers, but you will loose marks if you don't explain your answers. Give a reason even it might seem rather trivial, which is not meant to imply that non-trivial reasons are not preferred.
- If something is hard to explain, there is never a problem to add a figure.