ECE183DA (Winter 2022)

Design of Robotic Systems I

Prof. Ankur Mehta ⟨mehtank@ucla.edu⟩

Problem set 8 | Reinforcement learning

Key takeaways

After this lecture, you should understand:

- What assumptions / constraints on problem statements we have been making when deriving our earlier collection of algorithms, and what problems we might face that don't hold to those specifications.
- How we can generate and use datasets as a substitute for explicit knowledge of givens within a problem statement.
- Where we can apply function approximation within our algorithms to fill in the gaps in our dataset, and the corresponding adjustments necessary on those algorithms.
- How value iteration and policy iteration can be therefore modified into Q-learning and policy gradient approaches to machine learning.

Assignment

8(a). What is the expected limit

$$\lim_{T\to\infty} E[\rho(T)]$$

for an ε-greedy strategy with uniform random exploration on a standard multi-armed bandit? Can you explain in one English sentence a simple modification to this strategy that reduces this regret with negligible additional computational overhead?

8(b). In the original Q-learning algorithm, a single iterated update step computed a gradient descent on θ to drive Q_θ towards Q^*. To improve convergence properties, this algorithm was modified to split the update into two steps:

(i) Hold $\hat{Q}$ constant and iteratively update θ to minimize loss $\mathcal{L}(\theta) = \|r + \gamma \max_{a'} \hat{Q}(s', a') - Q_\theta(s, a)\|$;

(ii) After many updates of (i), set $\hat{Q}$ to Q_θ.

What process is being executed by each of these two steps? That is, describe conceptually in two to three English sentences each (without using math):

- what function is being approximated by θ in (i), and
- what is the assignment in (ii) meant to accomplish?

8(c). When doing policy gradient approaches, we approximate the expectation over trajectories via sampling. In two to three English sentences, explain why we might want to sample trajectories from different starting states, even when we know the exact initial state of our system.

8(d). Would you be willing to let us use your correct responses as (anonymized) examples for the class?