Midterm Review

CMPS 4660/6660: Reinforcement Learning

Midterm

- When: Oct 15 (R) 12:25-1:35 pm
- Where: Zoom meeting
  - camera on during the entire exam period
  - your exam will not be graded if you do not join the Zoom meeting
- Open-book and open-notes
  - You are NOT allowed to communicate with each other or search solutions online
- Office hours: W 10-11 am

Topics Covered

- Intro to RL
- Markov Decision Processes
- Dynamic Programming
- Model-Free Prediction
- Model-Free Control

Intro to RL

- Sequential decision making in uncertain environment
- Learning vs. Planning
- Exploration vs. Exploitation
- Goals and Rewards: Rewards Hypothesis
- Environment state vs. agent state: fully vs. partially observable environment

Markov Decision Processes

- Definition of MDP
  - Five elements $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$
  - Different ways of representing transition probabilities
  - Connections with Markov Chains and Markov Reward Processes
- Policy
  - history dependent vs. stationary policies
  - stochastic vs. deterministic policies
- Return
  - Episodic vs. Continuing Tasks
  - Why discount in continuing tasks?
  - "MDP with a terminal state" not required

Markov Decision Processes

- State-value and action-value functions
  - Connection between the two
- Prediction
  - Bellman Equations for state-value and action-value functions
  - Proof required for graduate students
- Control
  - Optimal Policy and Optimal Value Functions
  - Bellman Optimality Equations for state-value and action-value functions
  - Proof not required

Dynamic Programming

- Banach's fixed point theorem
  - Convergence in norm, contraction mappings, fixed point
  - Iterative convergence and uniqueness
  - Proof required for graduate students

- Policy Evaluation
  - Properties of Bellman operator T^{π}
  - Iterative policy evaluation
  - for state-value and action-value functions

Dynamic Programming

- Policy Optimization
  - Properties of Bellman optimality operator T^*
  - From optimal value to optimal policy (proof required for graduate students)
- Value Iteration
  - Algorithm and convergence result
  - Synchronous vs. Asynchronous VI
- Policy Iteration
  - Policy improvement theorem (proof required for graduate students)
  - Generalized Policy Iteration
- Linear Programming method for MDP
- POMDP not required

Mode-Free Prediction

- Model-free vs. model-based approaches
  - Model?
  - Model-free??

- Monte-Carlo Method
  - Algorithm
  - Incremental view: step-size
  - Convergence property
  - Stochastic Approximation not required

Mode-Free Prediction

- TD(0)
  - TD target, TD error
  - TD(0) vs. MC: bootstrapping, bias/variance tradeoff
  - Batch MC and TD
- n-step TD
- TD(λ)
  - forward-view vs. backward view
  - online vs. offline updates
  - TD(1) vs. MC
- TD vs. DP

Mode-Free Control

- Generalized Policy Iteration for model-free control
  - Why use action-value function?
  - Why is exploration necessary?
  - ϵ-greedy policy improvement

- On-policy Monte-Carol Control
  - Convergence: Greedy in the limit with infinite exploration (GLIE)