Announcements

• Homework 2 due 1 minute ago. Homework 3 released tonight.

• Project proposals will be reviewed and feedback provided this week.

• Chapter 9 reading is updated on the course website (less to read now).
Project Literature Review

• The next phase of your project is a literature review.

• An essential element to any research project.
  • Minimum expectation is that your survey cites at least 10 relevant references.
  • For each surveyed source, *briefly describe* (1-2 sentences), say why it is relevant to your project, and then say how it is different from your project.

  • Survey should be submitted as a pdf on Canvas.

  • The survey is also a secondary check-in on project direction.

  • See [https://pages.cs.wisc.edu/~jphanna/teaching/2022fall_cs839/project.html](https://pages.cs.wisc.edu/~jphanna/teaching/2022fall_cs839/project.html) for more details.
Today

- Finish RTDP
- Planning at decision-time:
  - Heuristic Search
  - Roll-out Algorithms
  - MCTS
Trajectory Sampling

- Uniform sampling of states can be inefficient.
- It may be more effective to focus value back-ups on states that the agent will visit often.
- How to know what states the agent will visit?
  - Initialize the agent in a start state and follow the current policy from there.
  - Simulate entire trajectories within the model or real world. Back-up the values for these states.
Real-time Dynamic Programming

• Key Idea: perform a value-iteration update on each state as it is visited.

• For n episodes:
  
  • Start in initial state, $S_0$.

  • Repeat $A_t \sim \pi(A = a \mid S_t)$, $S', R \sim \text{Model}(S_t, A_t)$ where $\pi$ is $\epsilon$-greedy.

  • At each step, $t$, apply the value iteration update to $Q(S_t, A_t)$:

    $$Q(s, a) \leftarrow \sum_{s', r} p(s', r \mid s, a) \left[ r + \gamma \max_{a'} Q(s', a') \right]$$
RTDP Example
Planning at Decision Time

- So far we considered using planning to improve the value-function and speed-up policy iteration.

- Now we consider using planning to immediately compute an action for a given state.

\[
\pi(s) \leftarrow \arg \max_a \sum_{s', r} p(s', r | s, a)[r + \gamma v_\pi(s')]
\]

\[
\pi(s) \leftarrow \arg \max_a \sum_{s', r} p(s', r | s, a)[r + \gamma \max_{a'} \sum_{s'', r'} p(s'', r' | s'', a')[r' + \gamma v_\pi(s'')]]
\]

- Slow deliberation before making a decision.

- Contrasts with immediate decision-making of model-free methods.
Heuristic Search

• Motivation: model is perfect and action-value function is imperfect.

• Focus memory and computation on immediate relevant state and next decision.

• Deeper search generally leads to a better action choice at expense of more computation.
Rollout Algorithms

- **Rollout**: following a policy until termination, i.e., rolling out the policy.

- Monte Carlo learning at decision-time; improve upon the roll-out policy.

- Rolling out the policy only requires a sample model.

- Computation time is a limiting factor. Roll-out algorithms computation affected by:
  
  - Speed to sample from model.
  
  - Speed to execute rollout policy.

\[
\begin{align*}
Q(s, a) &= 1/3 \\
Q(s, a) &= -1/3
\end{align*}
\]
Monte Carlo Tree Search

Iteration #1

- $Q(s,a) = 1, \quad n=1$
- Rollout Policy
- $G = 1$

Iteration #2

- $Q(s,a) = 1, \quad n=1$
- $Q(s,a) = -1, \quad n=1$
- Rollout Policy
- $G = -1$

Iteration #3

- $Q(s,a) = 1, \quad n=1$
- $Q(s,a) = -1, \quad n=1$
- $Q(s,a) = -1, \quad n=1$
- $G = -10$

Josiah Hanna, University of Wisconsin — Madison
Paul’s Presentation

Slides
Part I Summary

- Functions (policies, value functions, and models) have been represented as look-up tables.
- We have seen 4 types of algorithms:
  - Dynamic programming methods.
  - Model-free Monte Carlo methods.
  - Model-free temporal difference learning methods.
  - Model-based learning and planning methods.
- All algorithms we have seen are instances of generalized policy iteration:
  \[ \pi_0 \rightarrow q_0 \rightarrow \cdots \rightarrow \pi_k \rightarrow q_k \rightarrow \pi_{k+1} \rightarrow \cdots \rightarrow q_\star \rightarrow \pi_\star \]
Part I Summary

- Much intuition and understanding carries forward as we move into Part II.
  - Returns and values defined similarly.
  - On-policy and off-policy methods.
  - Exploration vs. Exploitation trade-off.

- Looking ahead:
  - The learning agent has limited capacity to model $V_\pi(s)$ for all $s$.
  - The learning agent may never visit the same state twice.
Summary

• Models can be used to produce simulated experience to learn from.
  • Makes better use of finite data.
  • Distribution models permit expected updates such as those made by RTDP.
  • Sample models are generally easier to acquire and can be used with sample updates.

• Models can be used to compute a better decision than acting greedily w.r.t. an inaccurate action-value function.
  • Requires planning at decision time.
  • Computation is the bottleneck for making better decisions.
Action Items

• Homework 3 released tonight.

• Begin literature review.

• Begin reading Chapter 9.