Multi-agent RL Cooperative Case

Ron Parr

CSCI2951-F

A taxonomy of multiagent scenarios

- Do all agents share common goals (shared reward function)?
  - This is the fully cooperative case
  - Generally considered one of the easier cases
- Do agents have adversarial goals (zero sum)?
  - This is the fully adversarial case
  - Middle ground in terms of difficulty
- Do agents have varying objectives that may partially overlap?
  - Fully general case
  - Requires full machinery of game theory, and can be messy

Cooperative Case

- Why is this easier?
- Can view agents as one big agent with large action space - cross product of individual agent action spaces
- Challenges/issues:
  - Action space can grow exponentially with number of agents
  - Picking the maximizing action can itself be an expensive process
  - Picking overall maximal action requires perfect knowledge of joint state and/or fully communication between agents at every decision

Dealing with large action spaces

- Q-values
  - May need to store many Q-functions and/or use a complex function approximation architecture (single network with both state and actions as input)
  - Iterating over all actions can be expensive
- Policy function (e.g. for REINFORCE or other policy search methods)
  - Requires a distribution over joint action space
  - Many parameters to optimize (slow, local optima)
  - Not so easily combined with variance reduction methods which often implicitly involve a max over all actions to store, e.g., a value function

Decentralization

- What if agents don't (or can't) communicate, so they are forced to pick their actions in isolation
- Challenges:
  - Unless problem is deterministic and agents have full environment models (so that they can simulate what other agents are doing in the heads), other agent's state and actions are not known
  - If other agent policies are known (but state is not known), then problem is a POMDP from the perspective of each agent
  - At training time, if actions of other agents are not known and other agents are also training, then problem becomes non-stationary

Centralized Training/Decentralized Acting

- Assume:
  - Each agent maintains a local set of Q-functions and/or policy functions
  - Individual agents see part of the state, but not full state, not other agent actions
  - "Bird's Eye" learning algorithm sees all agent observations, actions
- Training:
  - Centralized learner sends updates/gradient signals to individual agents to guide them towards maximizing global reward
- Implementation:
  - Can do policy gradient with a joint policy function that is product of individual agent policy functions
  - Can do Q-function approximation where global Q-value is sum of local Q-values

Finding optimal **decentralized** policies

- Given a model, we can formalize the problem of finding an optimal, decentralized policy for all agents
- This is called a decentralized POMDP (DEC-POMDP)
- Even harder than for a single agent POMDP because multiple agents must take into account uncertainty over each other agent's observations and actions
- DEC-POMDP formulations are typically EXP-hard (guaranteed exponential time), and optimal solutions are policy only for 1's of states
- Some practical success using policy search methods

Another approach: Learned communication

- Suppose agents are allowed to send signals to (a subset of) the other agents
- Augment action space to include sending signals (shooting a flare, making a sound, etc.)
- Augment observable state of other agents to see these signals

- Challenges:
  - Is this realistic?
  - Agents must discover their own language
  - Can encounter many local optima
  - Can require tons of training

Cooperative Case Summary

- Good news:
  - Really just a big action space
- Bad news:
  - Big action spaces are hard
  - “Big action space” view isn’t quite right w/o full communication
  - Removing full communication assumption introduces partial observability
  - Handling partial observability optimally is intractable
- Current approaches:
  - Learn some sort of fully or partially decentralized policy
  - Trade off optimality for tractability
  - Not many compelling theoretical results in this area