Reinforcement Learning for Healthcare

Taylor W. Killian
University of Toronto / Vector Institute

W’20 CSC 2541 - Lecture 7
20 February 2020
The basic unit of any RL or decision-making problem is the transition between observations after applying some action.

* This lecture does not provide an in-depth introduction to RL. For a better and more thorough introduction, please review Sutton and Barto [2017] or David Silver's RL lectures (on youtube).
Repeated interactions with an environment introduce a series of observations that can be ordered to accomplish some task.

With a specified task, each interaction with the environment can be defined to provide some auxiliary signal reflecting the utility (cost) of subsequent actions.
Foundations

\[ M = \{S, A, T, R, \gamma\} \]

Policies \( \pi \) are optimized by maximizing the “expected future reward” based on some discounted horizon of the reward gained in subsequent interactions with the environment.

\[
V^\pi(s_t) = \mathbb{E} \left[ \sum_{i=t}^{T} \gamma^{i-1} r_i \right]
\]

\[
V^\pi(s_t) = \max_a Q^\pi(s_t, a_t)
\]

\[
Q^\pi(s_t, a_t) = R(s_t, a_t) + \gamma \max_a Q^\pi(s_{t+1}, a)
\]
Foundations

\[ \mathcal{M} = \{S, A, T, R, \gamma\} \]

**Policy Iteration**

**Policy Gradients**

\[
V^\pi(s_t) = \mathbb{E} \left[ \sum_{i=t}^{T} \gamma^{i-1} r_i \right]
\]

**Value Iteration**

**Q-Learning**

\[
Q^\pi(s_t, a_t) = R(s_t, a_t) + \gamma \max_{a} Q^\pi(s_{t+1}, a)
\]

Based on the choice of policy optimization algorithm, you will utilize one or the other value function representation.
What’s Learnable in RL?

\[ \mathcal{M} = \{ S, A, T, R, \gamma \} \]

\[ \pi(a_t | s_t; \psi) \]

Beyond parameters for the policy, depending on your particular research question and the format of your data, you may be able to learn or infer other elements of the MDP.
Open Questions in RL

- Sample complexity
- Exploration v. Exploitation
- Representation Learning
- IRL vs. interactive feed-forward design
- Safety / Causality
  (i.e. is a developing policy guaranteed to “do no harm”)
Recent exciting demonstrations of RL successes have showed the flexibility and representational power when learning complex behaviors from sequential observations and guiding utility (cost) signal.

Recent exciting demonstrations of RL successes have showed the flexibility and representational power when learning complex behaviors from sequential observations and guiding utility (cost) signal.

The practice of medicine is inherently a sequential decision making problem:

- Clinicians, with their best understanding of a patient’s status, propose a treatment.
- The patient’s status may or may not change as a result of the prescribed treatment.
- Eventual outcomes are a noisy measure of the affect the prior treatment decisions had on the patient.

Can we utilize the Reinforcement Learning framework to describe/explain and augment clinician decision making?
Limitations of RL in Healthcare

Learning optimal treatment policies from observational data—*an offline and off-policy RL task*—is complicated by:

1. the inability to explore, and
2. a shrinking volume of training observations as top strategies are discovered

These two limitations severely complicate the ability to develop proactive RL algorithms/policies that suggest *what to do*
Limitations of RL in Healthcare

Beyond the inability to explore and diminishing training support there are other significant challenges to using RL algorithms for healthcare:

1. Unclear objectives
   a. What is a stable and clinically relevant reward?
   b. What motivates the clinician when there are competing priorities?

2. Biased measurements and noisy partial observations
   a. Oftentimes routine tests and measurements are missing which may indicate the clinician’s belief about the patient’s condition

3. Clinical practice varies widely between doctors and institutions.
   a. There is no clear understanding of what the best “expert” policy is to learn from.
   b. Sets of observations may differ between institutions
Limitations of RL in Healthcare (RL4H)

Beyond the inability to explore and diminishing training support there are other significant challenges to using RL algorithms for healthcare:

1. Unclear objectives
   a. What is a stable and clinically relevant reward?

2. Biased measurements and noisy partial observations
   a. Oftentimes routine tests and measurements are missing which may indicate the clinician's belief about the patient's condition

3. Clinical practice varies widely between doctors and institutions.
   a. There is no clear understanding of what the best “expert” policy is to learn from.
   b. Sets of observations may differ between institutions.

In order to make RL in Healthcare feasible, we need to fundamentally rethink how we utilize standard RL forms and frameworks to develop sequentially relevant insights into clinical decision making.
RL within Healthcare is not Totally New

Clinical Data Based Optimal STI Strategies for HIV

• Shortreed, et al [2011] provide one of the first rigorous studies of how RL may be leveraged in a healthcare setting.

• They perform empirical investigations of how Q-learning may be used in observational settings, finally applying their insights to the treatment of Schizophrenia.
Raghu, et al [2017] implement Deep Q-learning methodologies to develop treatment strategies for septic patients in Intensive Care Units, comparing with how learned strategies differ from actual doctor decisions.

Prasad, et al [2017] utilize GP regression for missing data imputation and then leverage FQI and NNs to develop policies for extubation of patients who are on mechanical ventilation.

Focusing on Value Functions

\[ V^\pi(s_t) = \mathbb{E} \left[ \sum_{i=t}^{T} \gamma^{i-1} r_i \right] \]

\[ V^\pi(s_t) = \max_a Q^\pi(s_t, a_t) \]

\[ Q^\pi(s_t, a_t) = R(s_t, a_t) + \gamma \max_a Q^\pi(s_{t+1}, a) \]

Traditionally, RL seeks to maximize the value function by approximating the effect of subsequent actions from the current observation (the Q-function) in developing a policy. The max operator excludes potentially valuable signal from the outcomes of actions that aren’t locally optimal. Without exploration, the outcomes following these actions aren’t incorporated into the value estimate for corresponding states.
Summary

- The practice of medicine is inherently a sequential decision making problem.
- While there are some complications with utilizing RL in observational settings, there is great promise with the framework being able to better describe the sequential nature of the patient-clinician decision problem. We cannot blindly implement SOTA RL approaches and expect the same kind of “superhuman” results → We need to be vigilant and thoughtful about how we develop RL research in healthcare settings.

Largely, open problems in RL map neatly into open problems in ML4H. There’s a lot of exciting developments to come in the near future!

Taylor’s Opinion