Reinforcement Learning in Drug Dosing

Adam E. Gaweda, Mehmet K. Muezzinoglu, George R Aronoff, Alfred A. Jacobs, Jacek M. Zurada, Michael E. Brier
University of Louisville and the Department of Veterans Affairs, Louisville, KY

Outline

• Reinforcement Learning
  – Q-learning
• Drug Dosing as Markovian Decision Process
  – Anemia Management
• Experimental Results
  – Comparison with Anemia Management Protocol
• Conclusions

Reinforcement Learning

• Rooted in psychology (animal learning), systematic trial-and-error process to solve staged decision problems.
• A methodology that evaluates the long-term effects of interactions of an agent with its environment toward a goal.
• Inputs are state transitions along a learning episode, corresponding actions and incurred immediate rewards.
• Outcome is a value function that ranks states (or state/action pairs) by their contributions to the goal.
• Strict relation to control theory – Simulation-based (approximate) DP.

Markovian Decision Process (MDP)

• Problem of estimating long-term benefits by combining short-term observations.
• Value of a state (or state/action pair) is a translation of long term benefit into computational domain.
• $V(s)$ is usually unavailable due to process uncertainty, i.e. insufficient information about state transitions and rewards.
• In such cases, one needs to estimate $V(s)$, by simulation.
• Manipulate the simulation to maximize the information gain.

Policy

• Maps the state space to actions (control law).
• We assume that it is an algebraic map in this work.
• Can be optimized by policy iteration methods: Start from an initial policy and improve it along a (trial-and-error) learning process.
• GAs perform search directly on the policy space guided by scalar evaluations of entire policies.
• Policy iteration methods under RL are different: They seek to evaluate the states making use of policies.
• On-policy and off-policy methods:
  – On-policy: Policy being improved is demonstrated in simulation.
  – Off-policy: Subject policy is improved due to the observed effects of other policy.

Q-Learning (1)

• An off-policy RL control method.
• Maintains the Q-table along the learning process.

$$Q(s, a) = E \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$$

• Seeks an approximate solution to Bellman equation:

$$Q(s, a) = s(s, a) + \gamma \max_{a'} Q(s', a')$$

• At each observation $s \rightarrow s'$ and $a \rightarrow a'$

$$Q(s, a) = (1 - \alpha) Q(s, a) + \alpha [r(s, a) + \gamma \max_{a'} Q(s', a')]$$
Q-Learning (2)

- Having updated the state value estimate, the policy is improved:
  \[ p(s) = \max_a Q(s,a) \]

- The following simulation step is performed by picking the new action
  \[ a = \begin{cases} 
  \text{arbitrary action} & \text{with probability } \epsilon \\
  \tilde{a} & \text{with probability } 1 - \epsilon
  \end{cases} \]

- Exploration is critical in learning.
- Taking the best action is critical for control during learning.
- State-dependent could be a solution.

Drug Dosing as MDP

- *Drug dose - action*
- *Physiologic response, toxic effects - state*
- *Reward = f( target, state, action )*

Example - Anemia Management

- **Erythropoietin (EPO)**
- **Hemoglobin (HgB)**
- *Reward = \[ \begin{cases} 
  1 & \text{if } 11 \text{ g/dL} \leq \text{HgB} \leq 12 \text{ g/dL} \\
  -1 & \text{if } \text{HgB} > 12 \text{ g/dL} \text{ or } \text{HgB} < 11 \text{ g/dL}
  \end{cases} \]*

Modeling Dose – Response

- *Classification*
- *Prediction*

Q-Learning for Optimal Drug Dosing

- We first picked a finite set of representative HgB levels and the best-guess policy.
- We allowed exploration (nonzero ) only when HgB=11.5.
- Note that the method would learn how to behave only at the representative states.
- To extend the skeleton policy to [8,15], we designed an RBF *policy* network.
Drug Dosing Simulation

Simulation Example
Normal Responder

Simulation Example
Poor Responder

Results - RL vs. AMP

<table>
<thead>
<tr>
<th></th>
<th>HgB Level (g/dL)</th>
<th>HgB Std Dev (g/dL)</th>
<th>Mean Total EPO (1,000 U)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q-learning</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal Responders</td>
<td>11.59</td>
<td>0.29</td>
<td>589.3 (344.5, 834.0)</td>
</tr>
<tr>
<td>Poor Responders</td>
<td>11.16</td>
<td>0.74</td>
<td>1,145.3 (926.6, 1,363.9)</td>
</tr>
<tr>
<td>AMP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal Responders</td>
<td>11.66</td>
<td>0.32</td>
<td>610.6 (356.9, 864.2)</td>
</tr>
<tr>
<td>Poor Responders</td>
<td>11.51</td>
<td>0.67</td>
<td>1,075.4 (942.5, 1,208.3)</td>
</tr>
</tbody>
</table>

Conclusions

- An approach to pharmacotherapy by mimicking goal-oriented learning and adaptation as performed by humans.
- In its current form it does as good as AMP, a well-established dynamic policy.
- Proposed scheme is only a first step and includes many issues that needs more attention: state space quantization, information gain due to exploration, learning speed.
- What if the patient simulation is stochastic and/or non-stationary?

Research Direction

Can RL handle drug dosing problem? Yes, proved in the previous work.
Can RL do as good as an MD? Yes, proved here.
Can RL handle model/measurement uncertainties? We know it can… Will be applied to our problem.
How could learning speed be improved? Guided exploration may be the answer.