Function Approximators for Reinforcement Learning

Keith L. Downing

The Norwegian University of Science and Technology (NTNU)
Trondheim, Norway
keithd@idi.ntnu.no

January 26, 2020
Large RL Search Spaces:

State space = cross product of all variable states.

<table>
<thead>
<tr>
<th>Robotics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Sensor readings</td>
</tr>
<tr>
<td>2 Location, orientation, velocity, acceleration</td>
</tr>
<tr>
<td>3 Battery level</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Texas Hold ’Em Poker</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 A player’s hole cards, of which there are (\binom{52}{2}) possibilities.</td>
</tr>
<tr>
<td>2 Number of raises in the current betting round.</td>
</tr>
<tr>
<td>3 Number of players still active in the current betting round.</td>
</tr>
<tr>
<td>4 Money in the pot (preferably discretized into k different bins)</td>
</tr>
<tr>
<td>5 The (discretized and binned) strength rating of the player’s hand - based on her hole cards plus the mutual cards (flop, turn, river).</td>
</tr>
<tr>
<td>6 Number of active players known to bluff frequently</td>
</tr>
</tbody>
</table>
Explosion of possible states → tables are impossible to build.

Neural net = a function approximator (funcapp) from states to evaluations (critic) or states to actions (actor).

The funcapp must handle every possible state as input.

Key generality assumption: For the vast majority of cases, similar states map to similar values.
Learning with Tables -vs- Funcapps

- Using tables for state values in TD(0), update the table entry $V(s)$ for state $s$ via:

$$V(s) \leftarrow V(s) + \alpha [r + \gamma V(s') - V(s)]$$

where $\delta = r + \gamma V(s') - V(s)$; episode = ... $s \rightarrow s' \rightarrow s''$ ...

- $\delta$ = error in the current estimate of $V(s)$.
- Clearly, this update ONLY affects $V(s)$.
- With a Function Approximator, there is no entry just for $V(s)$. There are model parameters $(\theta)$, where $|\theta| << |S|$.
- Based on $\delta$, how can we modify $\theta$ to get a better approximation of $V(s)$...but without ruining our estimates for $V(s') \forall s' \neq s$?
- If the funcapp is an NN, how do we modify the weights when given $\delta$ to reduce our overall error ($\forall s \in S$)?
- This has become a supervised learning problem, with each training case = features $(s) +$ target $(r + \gamma V(s'))$.
Stochastic Gradient Descent (SGD) in RL Funcapps

Loss Function $L(\theta)$

$$L(\theta) = \frac{1}{2} \sum_{s \in MB} (V^*(s) - V_\theta(s))^2$$

where

- $MB$ = a minibatch of states
- $V^*(s)$ = actual value of $s$ = the target
- $V_\theta(s)$ = estimate of $V^*(s)$ = output of NN with params $\theta$

- Using a minibatch of 1 case ($s$), $\forall w \in \theta$, compute:

$$\nabla_w L(\theta) = \frac{\partial L(\theta)}{\partial w} = [V^*(s) - V_\theta(s)] \nabla_w (-V_\theta(s))$$

- The weight update (where $\alpha$ = learning rate):

$$\Delta w = -\alpha \nabla_w L(\theta) = \alpha [V^*(s) - V_\theta(s)] \nabla_w V_\theta(s)$$
What is the target, $V^*(s)$?

Assume $s = s_t =$ state visited at step $t$ of an episode.

- **1-step TD: TD(0)** \( \{ \gamma = \text{discount} \} \)
  \[
  V^*(s_t) = r_{t+1} + \gamma V_\theta(s_{t+1})
  \]

- **N-Step TD**
  \[
  V^*(s_t) = G_t = r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{n-1} r_{t+n} + \gamma^n V_\theta(S_{t+n})
  \]

- **Monte Carlo** \( \{ T = \text{final step of the episode} \} \)
  \[
  V^*(s_t) = G_t = r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{T-t-1} r_T
  \]

For TD methods, bootstrapping (red terms) makes $V^*(S_t)$ a function of $\theta$, since $V_\theta(S_{t+n})$ is a function of $\theta$. Hence, \( \frac{\partial L(\theta)}{\partial w} \) is more complex; this is **Semi-Gradient Descent**...but it still works pretty well.
NN = funcapp $V_\theta$ that maps states to evaluations.

1. Compute the symbolic gradients: $\forall w : \nabla_w V_\theta$
2. Do one episode of RL search.
   Along the way (TD) or at episode end (MC), collect cases: $(s_t, V^*(s_t))$.
3. $\forall (s, V^*(s)) \in$ cases: 
   \{Train $V_\theta$ on each case\}
   - $\forall w \in \theta$:
     - Instantiate the symbolic gradients with s and $V_\theta(s)$, yielding numeric gradients: $\forall w : \nabla_w V_\theta(s)$
     - Update w:
       $$\triangle w \leftarrow \alpha [V^*(s) - V_\theta(s)] \nabla_w V_\theta(s)$$
4. Go to 2

- Gradients such as $\nabla_w V_\theta$ are standard backprop calculations.
- For 1-step TD, $\triangle w \leftarrow \alpha \delta \nabla_w V_\theta(s)$
Semi-Gradient Descent for TD(\(\lambda\))

- \(V^*(s_t)\) is the same as for 1-step TD, regardless of \(\lambda\).
  \[
  V^*(s_t) = r_{t+1} + \gamma V_\theta(s_{t+1})
  \]

- TD Error is then:
  \[
  \delta = V^*(s_t) - V_\theta(s_t) = r_{t+1} + \gamma V_\theta(s_{t+1}) - V_\theta(s_t)
  \]

- Each weight, \(w_j\) has an eligibility \(e_j\), updated via:
  \[
  e_j \leftarrow \gamma \lambda e_j + \nabla_{w_j} V_\theta(s_t)
  \]

  where \(\gamma =\) discount, \(\lambda =\) eligibility decay factor.

- Weight updates (with \(\alpha =\) learning rate) are:
  \[
  w_j \leftarrow w_j + \alpha \delta e_j
  \]
Eligibility Traces in Backprop Nets

Eligibilities decay on each step and are incremented when $\nabla_{w_j} V_\theta(s_t) \neq 0$

$$e_j \leftarrow \gamma \lambda e_j + \nabla_{w_j} V_\theta(s_t)$$
$$w_j \leftarrow w_j + \alpha \delta e_j$$

$\frac{dV(S)}{dw} = \frac{d(Sum)}{dw} \times \frac{d(B)}{d(Sum)} \times ...$

$= A \times \frac{d(B)}{d(Sum)} \times ...$ when abs(A) is non-zero

=> w does not always contribute to V(S)

=> w's eligibility trace is not always incremented.
As an RL system runs through an episode,
the NN processes a different state $s_t$ at each time $t$.
Each $s_t$ produces a different $\nabla_{w_j} V_\theta(s_t)$ for each weight, $w_j$, yielding differential weight updates.
If the TD error ($\delta_t$) for $s_t$ is high, then any weight ($w^*$) with a strong contribution to $V_\theta(s_t)$ will get a large modification.
Such weights can continue to receive large modifications on state $s_{t'}$ (for $t' > t$) as long as $\delta_{t'}$ remains high...even if $\nabla_{w^*} V_\theta(s_{t'})$ is low.
This effect gradually wears off with eligibility decay.
It can only be re-strengthened by another high gradient for $w^* : \nabla_{w^*} V_\theta(s_{t'})$. 
TD-Gammon (Tesauro, 1995)

\[ V(S_t) = r^* + V(S_{t+1}^*) - V(S_t) \]

\[ \delta_t = r^* + V(S_{t+1}^*) - V(S_t) \]

The state of the board is denoted by \( S_t \).

The diagram illustrates the process of learning in TD-Gammon, where the agent learns to choose actions to maximize the expected cumulative reward. The state of the board at time \( t \) is denoted by \( S_t \), and the action taken at time \( t \) is denoted by \( a_t \). The next state is \( S_{t+1} \), and the reward is \( r_t \). The value function \( V(S_t) \) is approximated using function approximators, and the difference between the estimated value and the actual reward is used for learning.

Keith L. Downing

Function Approximators for Reinforcement Learning
General Observations

**ANN as Value function**

Many search spaces are too large to yield accurate $V(s)$ entries via standard RL exploration. ANNs are an adaptive approximator of $V(s)$ that allow systems to bootstrap to effective policies.

**RL to Supervised Learning via TD**

1. In TD, each move leads to a comparison of expectations to slightly more informed expectations, and any difference between the two is a **surprise** that leads to learning.

2. If an ANN (or any system) can predict future states (based on current states and actions), then

3. when the future arrives, reality serves as the **correct answer**, which, when compared to the prediction, provides an error term, which can guide learning.

4. This can be done at every timestep.

5. So **prediction + comparison to reality** turns RL into supervised learning.