Temporal-Difference Networks

Sutton and Tanner, 2004

Presented by Tom Bosc

Introduction

From Sutton's 14 principles:

- (7) all world knowledge can be well thought of as predictions of experience.
- (9) temporal-difference learning is not just for rewards, but for learning about everything, for all world knowledge. any moment-by-moment signal (e.g., a sensation or a state variable) can substitute for the reward in a temporal-difference error
- Recall: Usual TD learning:
  $V(s) \leftarrow V(s) + \alpha[R + \gamma V(s') - V(s)]$
- We should be able to **compose** predictions. Complex predictions can/should be based on other predictions.
- Can we bootstrap using other predictors than the one we are trying to update?

TD networks

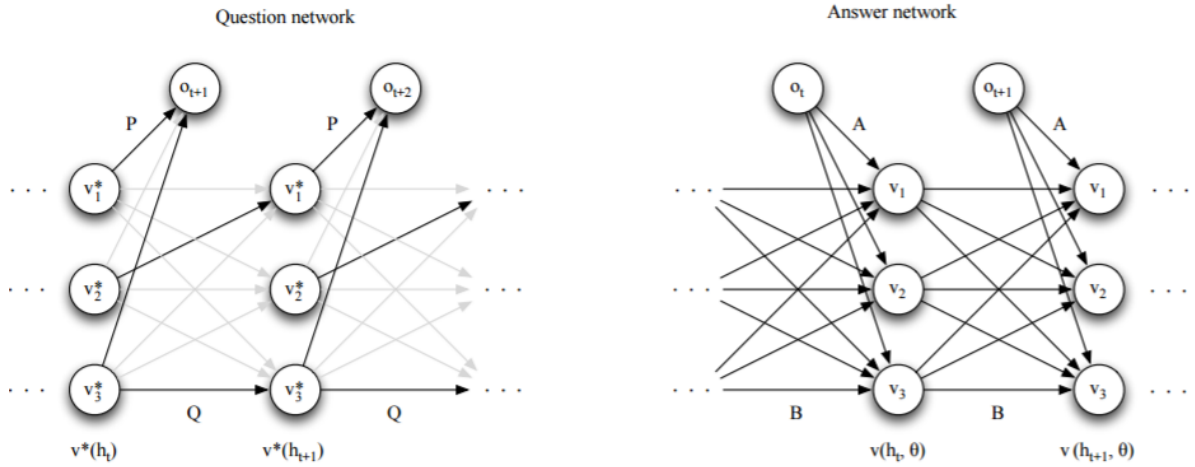

Figure 1: Question and answer networks (from Silver 2012)

- Prediction y_t, observation o_t, action a_t.
- **Answer** networks u with param W_t define the update $y_t = u(y_{t-1}, a_{t-1}, o_t, W_t)$
- **Question** networks define the targets $z_t = z(o_{t+1}, \tilde{y}_{t+1})$. $\tilde{y}_{t+1}$ is the prediction using old weights W_{t-1}.
- Order of computation: $y_t, a_t, o_{t+1}, x_{t+1}, \tilde{y}_{t+1}, z_t, W_{t+1}, y_{t+1}$
- Targets can be conditional on an action or even an option.

Example 3: Non-Markov

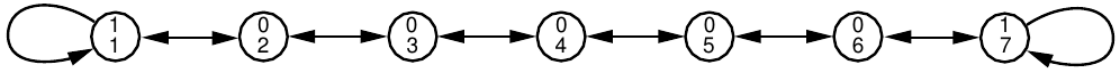

Figure 2: Random-walk MDP (Sutton et Tanner 2004)

- Observation: $o_t = 1$ when in state 1 or 7, $o_t = 0$ elsewhere.
- Prediction task: $z_t = [o_{t+2} \cdot o_{t+3} \cdot o_{t+43}]$
- $y_t = \sigma(W_t x_t)$ where features $x_t = [a_{t-1} \cdot o_t \cdot y_{t-1}]$

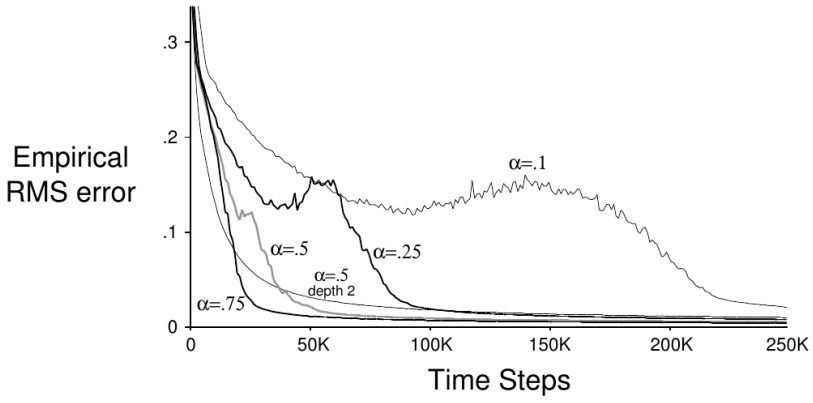

Figure 3: Results (Sutton et Tanner 2004)

Conclusion

- It is possible to use various prediction as features and bootstrap
- Especially useful in POMDP
- Related to General Value Functions (Sutton et al. 2011), Predictive State Representations (Littman et al., 2001)
- "Finally, we note that adding nodes to a question network produces new predictions and thus may be a way to address the discovery problem for predictive representations."

Bibliography

- fourteen declarative principles of experience-oriented intelligence, Sutton (online, not published)
- Temporal-Differences Networks, Sutton, Tanner, 2004
- Predictive Representations of States, Littman et al., 2001
- Gradient Temporal Difference Networks, Silver, 2012
- Horde: A Scalable Real-time Architecture for Learning Knowledge from Unsupervised Sensorimotor Interaction, Sutton et al., 2011