Least Squares TD

Michael Noseworthy

March 31, 2017

Review: TD(0)

- Recall the Linear TD update:

$$\theta_{t+1} = \theta_t + \alpha(R_{t+1} + \gamma\theta_t^T \phi_{t+1} - \theta_t^T \phi_t)\phi_t$$

- Rewrite this as:

$$\theta_{t+1} = \theta_t + \alpha\left(R_{t+1}\phi_t - \phi_t(\phi_t - \gamma\phi_{t+1})^T\theta_t\right)$$
$$\mathbb{E}\left[\theta_{t+1}|\theta_t\right] = \theta_t + \alpha\left(b - A\theta_t\right)$$

$$b = \mathbb{E}\left[R_{t+1}\phi_t\right] \;\; A = \mathbb{E}\left[\phi_t(\phi_t - \gamma\phi_{t+1})^T\right]$$

- At convergence, we get:

$$\theta_{TD} = A^{-1}b$$

LSTD(0)

- **Idea:** Instead let's estimate θ_{TD} directly.
- Keep estimates of A and b:

$$\hat{A}_t = \frac{1}{t}\sum_{k=0}^{t} \phi_t(\phi_t - \gamma\phi_{t+1})^T$$

$$\hat{b}_t = \frac{1}{t}\sum_{k=0}^{t} \phi_t R_{t+1}$$

- Whenever we want to estimate θ:

$$\theta_t = \hat{A}^{-1}\hat{b}$$

Recursive Implementation

- As presented, we require a matrix inverse! $O(n^3)$
- Instead, let's directly estimate the inverse
  - Sherman-Morrison formula

$$\hat{A}_t^{-1} = \hat{A}_{t-1}^{-1} - \frac{\hat{A}_{t-1}^{-1}\phi_t(\phi_t - \gamma\phi_{t+1})^T\hat{A}_{t-1}^{-1}}{1 + (\phi_t - \gamma\phi_{t+1})^T\hat{A}_{t-1}^{-1}\phi_t}$$

- New initialization parameter:

$$\hat{A}_{-1}^{-1} = \epsilon^{-1}\mathbf{I}$$

Model-Based RL

- If we knew P and R we could directly calculate the value function.
- Introduce sufficient statistics for the model:
  - n: Vector for the number of times a state has been visited ($N = diag(n)$)
  - C_{ij}: Transition counts from state i to state j
  - s: Vector for the sum of rewards for each state

$$v = N^{-1}s - N^{-1}Cv$$
$$v = (N - C)^{-1}s$$

Model-Based RL

- Consider the tabular case: $\phi_t = [\ldots 1 \ldots]^T$
- $\hat{b}$ from LSTD(0) is the same as s

$$\hat{b}_t = \sum_{k=0}^{t} \phi_t R_{t+1}$$

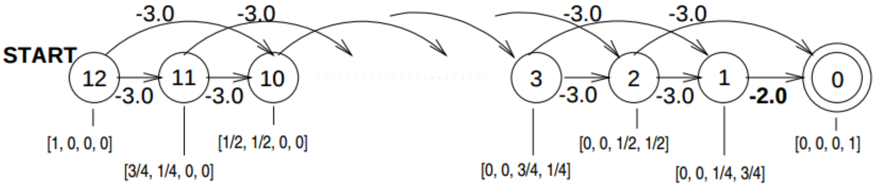
- $\hat{A}$ is the same as $N - C$:

$$\hat{A}_t = \sum_{k=0}^{t} \phi_t (\phi_t - \phi_{t+1})^T$$

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & -1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

- So $v = (N - C)^{-1} s = \hat{A}^{-1} \hat{b}$ for the tabular case.

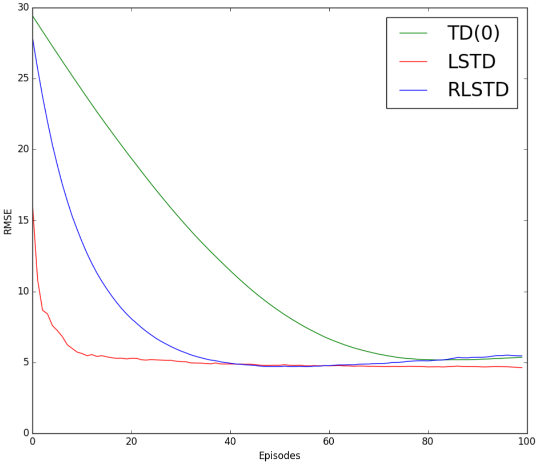
Experiments: Boyan Chain



$$\beta^* = (-24, -16, -8, 0)$$

- Compare algorithms:
  - TD(0)
  - LSTD(0)
  - RLSTD(0)

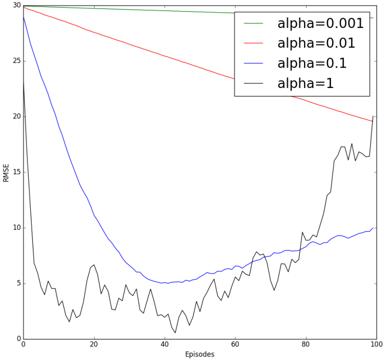
Results



- LSTD is statistically more efficient.
- But computationally more expensive!

Results

- TD is sensitive to α



- RLSTD is sensitive to ϵ