Fast Gradient-Descent Method for TD learning with Linear Function Approximation

link

by: Richard S. Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, Eric Wiewiora

COMP 767 – Reinforcement Learning
March 10th

~Nicolas Angelard-Gontier - 260532513

Motivation

Gradient Descent for Linear Function Approximation:

- On-policy:
  - Monte-Carlo converges
  - TD, SARSA, Q-learning are only semi-gradient! (assume $\nabla_\theta target = 0$)
- Off-policy:
  - TD, SARSA, Q-learning convergence cannot be guaranteed!

- Off-policy training is useful for the exploration-exploitation tradeoff & for intra-option learning

Linear value-function approximation

- We have: $V_\theta(s) = \theta^T \Phi_s$
- Goal: $V_\theta(s) \simeq V(s)$
- Let's define the TD-error as: $\delta_t = \underbrace{r_{t+1} + \gamma \theta_t^T \Phi_{t+1}}_{\text{TD target}} - \theta_t^T \Phi_t$
- Usual update rule: $\theta_{t+1} = \theta_t + \frac{1}{2} \alpha_t \nabla_\theta (\delta_t)^2$

  $\theta_{t+1} = \theta_t + \alpha_t \delta_t \Phi_t$ (semi-gradient!)
- With the goal mentioned above, a natural objective function to minimize for θ is:

$$MSVE(\theta) = \sum_s d_s [V_\theta(s) - V(s)]^2 = \|V_\theta - V\|_D^2$$

The trick is...

- To use a different objective function:
- How closely the approximate value function satisfies the Bellman equation?
  - Note: the true value function does satisfy the Bellman eq: *V* = *T V* with *T* being the Bellman operator
- $MSBE(\theta) = \sum_s d_s [V_\theta(s) - TV_\theta(s)]^2 = \| V_\theta - TV_\theta \|_D^2$

  still not good enough: no convergence to the minimum of the MSBE
  because *T* follows state dynamics irrespectively of structure of function approximator, so TV_θ will never be representable as V_θ : $V_\theta \neq TV_\theta \forall \theta$

$$MSVE(\theta) = \sum_s d_s [V_\theta(s) - V(s)]^2 = \|V_\theta - V\|_D^2$$

$$MSBE(\theta) = \sum_s d_s [V_\theta(s) - TV_\theta(s)]^2 = \|V_\theta - TV_\theta\|_D^2$$

- Introduce the projection operator Π which maps any value function *v* to the nearest value function representable by our linear function approximator:

$$\Pi v = V_\theta \; where \; \theta = argmin_\theta \|V_\theta - v\|_D^2$$

- In a linear setup where $V_\theta = \Phi\theta$ the projection operator is independent of θ:

$$\Pi = \Phi(\Phi^T D \Phi)^{-1} \Phi^T D$$

TD fixpoint! :)

- Now our value function approximation satisfies the "projected bellman equation": $V_\theta = \Pi T V_\theta$

$$\cancel{MSVE(\theta)=\sum_s d_s[V_\theta(s)-V(s)]^2=\|V_\theta-V\|_D^2}$$

$$\cancel{MSBE(\theta)=\sum_s d_s[V_\theta(s)-TV_\theta(s)]^2=\|V_\theta-TV_\theta\|_D^2}$$

$$MSPBE(\theta)=\sum_s d_s[V_\theta(s)-\Pi TV_\theta(s)]^2=\|V_\theta-\Pi TV_\theta\|_D^2$$

$$=\ldots\textit{see paper}\ldots$$

$$=E[\delta\Phi]^T E[\Phi\Phi^T]^{-1}E[\delta\Phi]$$

Modifiable parameter: $w\in\Re^n$

$$w\simeq E[\Phi\Phi^T]^{-1}E[\delta\Phi]$$

- GTD2:

$$\frac{-1}{2}\nabla_\theta MSPBE(\theta)=E[(\Phi-\gamma\Phi')\Phi^T]E[\Phi\Phi^T]^{-1}E[\delta\Phi]$$

$$\simeq E[(\Phi-\gamma\Phi')\Phi^T]w$$

We have:

$\theta_{t+1}=\theta_t+\alpha_t(\Phi_t-\gamma\Phi_t')(\Phi_t^T w_t)$ and $w_{t+1}=w_t+\beta_t(\delta_t-\Phi_t^T w_t)\Phi_t$

$$MSPBE(\theta) = \sum_s d_s [V_\theta(s) - \Pi T V_\theta(s)]^2 = \|V_\theta - \Pi T V_\theta\|_D^2$$

$$= E[\delta \Phi]^T E[\Phi \Phi^T]^{-1} E[\delta \Phi]$$

Modifiable parameter: $w \in \Re^n$

$$w \simeq E[\Phi \Phi^T]^{-1} E[\delta \Phi]$$

- TDC: $\frac{-1}{2} \nabla_\theta MSPBE(\theta) = E[(\Phi - \gamma \Phi')\Phi^T] E[\Phi \Phi^T]^{-1} E[\delta \Phi]$

$$= (E[\Phi \Phi^T] - \gamma E[\Phi' \Phi^T]) E[\Phi \Phi^T]^{-1} E[\delta \Phi]$$

$$= E[\delta \Phi] - \gamma E[\Phi' \Phi^T] E[\Phi \Phi^T]^{-1} E[\delta \Phi]$$

$$\simeq E[\delta \Phi] - \gamma E[\Phi' \Phi^T] w$$

Same as conventional linear TD

Correction to follow MSPBE instead of MSVE

$$\theta_{t+1} = \theta_t + \alpha_t \boxed{\delta_t \Phi_t} - \alpha \boxed{\gamma \Phi_t'(\Phi_t^T w_t)} \quad \& \quad w_{t+1} = w_t + \beta_t(\delta_t - \Phi_t^T w_t)\Phi_t$$

Conclusion

- Both GTD2 and TDC converge (proved in paper)
- Time complexity O(n) (with $\theta \in \Re^n$)
- Memory complexity O(n)