Fast Gradient-Descent Methods for Temporal Difference Learning with Linear Function Approximation

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S.,Silver, D., Szepesvari, Cs., and Wiewiora, E.

Greta Laage

March 10, 2017

Objective

Finding an off-policy temporal difference algorithm effective on large applications with linear function approximation

Objective

Finding an off-policy temporal difference algorithm effective on large applications with linear function approximation

Framework

- TD methods based on gradient descent and linear function approximation
- Most popular methods (TD(λ),Q-learning,Sarsa) not true gradient-descent methods. Norrower and less robust conditions to convergence. Convergence not guaranteed with off-policy training
- Existing methods: non-gradient-descent approaches, second-order methods $O(n^2)$(LSTD),...
- But $TD(\lambda)$ is $O(n)$

Linear Function Approximation

$$V_\theta(s) = \theta^T \phi_s \simeq V(s) = \mathbb{E}\{\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_o = s\}$$

Paper settings

- first state of each transition is chosen iid according to a arbitrary distribution d that may be unrelated to P (off policy learning)
- probability over (s_k, s'_k, r_k)
- $\phi_k = \phi_{s_k}$ and $\phi'_k = \phi_{s'_k}$
- TD error: $\delta_k = r_k + \gamma \theta_k^T \phi'_k - \theta_k^T \phi'_k$

Conventional linear TD algorithm: one-step TD learning

One independant update to θ for each state transition and associated reward

$$\theta_{k+1} \leftarrow \theta_k + \alpha_k \delta_k \phi_k$$

Objective function: MSPBE

Usually, $MSBE(\theta) = \|V_\theta - TV_\theta\|_D^2$ for GD algorithm. But :

- Most TD algorithms (TD, LSTD, GTD) do not converge to the minimum of $MSBE$
- TV_θ not be representable as V_θ

Projection Π

Projects v to the nearest value function representable by the approximator

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

Mean squared projected Bellman error

$$MSPBE(\theta) = \|V_\theta - \Pi TV_\theta\|_D^2 = \mathbb{E}[\delta\phi]^T\mathbb{E}[\phi\phi^T]\mathbb{E}[\delta\phi]$$

Algorithms converge to the fixpoint $V_\theta = \Pi TV_\theta$

GTD Algorithm

Algorithm updates

Norm of the expected TD update *NEU* where $u_k = \mathbb{E}[\delta\phi]$

$$NEU(\theta) = \mathbb{E}[\delta\phi]^T\mathbb{E}[\delta\phi] \qquad -\frac{1}{2}\nabla NEU(\theta) = \mathbb{E}[(\phi - \gamma\phi')\phi^T]\mathbb{E}[\delta\phi]$$

$$\theta_{k+1} = \theta_k + \alpha_k(\phi_k - \gamma\phi'_k)(\phi_k^T u_k)$$
$$u_{k+1} = u_k + \beta_k(\delta_k\phi_k - u_k)$$

Convergence analysis

Convergence to the TD solution with probability on some conditions

Proof: recursive stochastic algorithms $x_{k+1} = x_k + \alpha_k(h(x_k) + M_{k+1})$
+ ordinary differential equation approach $\dot{x} = h(x(t))$
Theorem 2.2 of Borkar & Meyn (2000): x_k *converges to the unique global asymptotically stable equilibrium*

GTD2 Algorithm

Algorithm updates

$$-\frac{1}{2}\nabla MSPBE(\theta) \approx \mathbb{E}[(\phi - \gamma\phi')\phi^T]\omega, \quad \omega = \mathbb{E}[\phi\phi^T]^{-1}\mathbb{E}[\delta\phi]$$

$$\theta_{k+1} = \theta_k + \alpha_k(\phi_k - \gamma\phi'_k)(\phi_k^T\omega_k)$$
$$\omega_{k+1} = \omega_k + \beta_k(\delta_k - \phi_k^T\omega_k)\phi_k$$

Theorem1 Convergence analysis

- $\beta_k = \eta\alpha_k$, $\eta > 0$, $\beta_k \in (0, 1]$
- $\sum \alpha_k = \infty$ and $\sum \alpha_k^2 < \infty$
- $\mathbb{E}[\phi_k(\phi_k - \gamma\phi'_k)^T]$ and $\mathbb{E}[\phi_k\phi_k^T]$ non singular
- (ϕ_k, r_k, ϕ'_k) iid sequence with uniformly bounded second moments

θ_k converges with probability one to the fixpoint $V_\theta = \Pi T V_\theta$

Proof: ordinary differential approach and theorem 2.2 of Borkar & Meyn(2000)

TDC Algorithm

Algorithm updates

$$-\frac{1}{2}\nabla MSPBE(\theta) \approx \mathbb{E}[\delta\phi] - \gamma\mathbb{E}[\phi'\phi^T]\omega, \quad \omega = \mathbb{E}[\phi\phi^T]^{-1}\mathbb{E}[\delta\phi]$$

$$\theta_{k+1} = \theta_k + \alpha_k\delta_k\phi_k - \alpha\gamma\phi'_k(\phi_k^T\omega_k)$$
$$\omega_{k+1} = \omega_k + \beta_k(\delta_k - \phi_k^T\omega_k)\phi_k$$

Theorem2 Convergence analysis

- $\alpha_k, \beta_k > 0$, $\frac{\alpha_k}{\beta_k} \underset{k\infty}{\rightarrow} 0$
- $\sum \alpha_k = \infty$ and $\sum \alpha_k^2 < \infty$
- $\sum \beta_k = \infty$ and $\sum \beta_k^2 < \infty$
- $\mathbb{E}[\phi_k(\phi_k - \gamma\phi'_k)^T]$ and $\mathbb{E}[\phi_k\phi_k^T]$ non singular
- (ϕ_k, r_k, ϕ'_k) iid sequence with uniformly bounded second moments

θ_k converges with probability one to the fixpoint $V_\theta = \Pi T V_\theta$

Proof: Two timescale stochastic approximation recursion, Borkar(1997). Faster recursion on ω_k and slower recursion on θ_k

Random Walk

Figure 1: 5 states Random Walk - Sutton&Barto

	Tabular features	Inverted features	Dependent features
	Tabular features	Inverted features	Dependent features
A	$(1,0,0,0,0)$	$(0,\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2})$	$(1,0,0)$
B	$(0,1,0,0,0)$	$(\frac{1}{2},0,\frac{1}{2},\frac{1}{2},\frac{1}{2})$	$\frac{1}{\sqrt{(2)}},\frac{1}{\sqrt{(2)}},0)$
C	$(0,0,1,0,0)$	$(\frac{1}{2},\frac{1}{2},0,\frac{1}{2},\frac{1}{2})$	$(\frac{1}{\sqrt{(3)}},\frac{1}{\sqrt{(3)}},\frac{1}{\sqrt{(3)}})$
D	$(0,0,0,1,0)$	$(\frac{1}{2},\frac{1}{2},\frac{1}{2},0,\frac{1}{2})$	$(0,\frac{1}{\sqrt{(2)}},\frac{1}{\sqrt{(2)}})$
E	$(0,0,0,0,1)$	$(\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},0)$	$(0,0,1)$

Table 1: Different features

Baird's counterexample

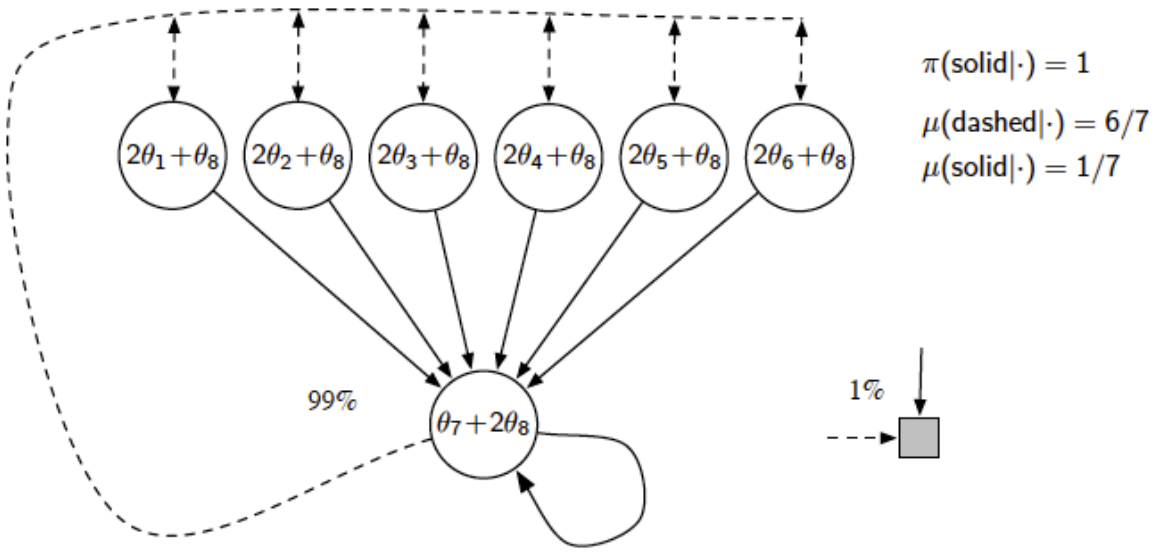

Figure 2: Baird counterexample - Sutton&Barto

$$V_\theta = \Phi\theta, \Phi = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 \end{pmatrix}$$