Convergent Temporal-Difference Learning with Arbitrary Smooth Function Approximation

H. R. Maei, C. Szepesvari, S. Bhatnagar, D. Precup, D. Silver, R. S. Sutton

24 mars 2017

Motivations

TD algorithms are great but potentially **unstable** in the **off-policy** case, or when the function approximation is **non-linear**.

Gradient methods (TDC, GTD...) help in the off-policy case.

The problem in the non-linear case remains $\Rightarrow$ **non-linear extensions of GTD and TDC**

Convergent as long as the function approximation is **smooth** enough.

The objective function

Linear TDC, GTD :

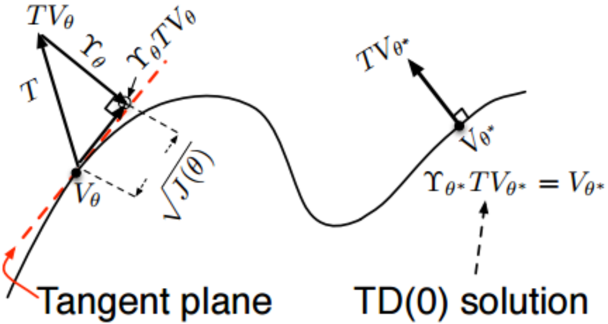
$$J(\theta) = MSPBE(\theta) = \|\Pi(TV_\theta - V_\theta)\|_D$$

where Π projects onto $\mathcal{M} = \{V_\theta | \theta \in \mathbb{R}^n\} = \{\Phi\theta | \theta \in \mathbb{R}^n\}$ (linear).

Non-linear TDC, GTD : $\mathcal{M}$ is a non-linear manifold... $\Rightarrow$ projection onto the tangent space of $\mathcal{M}$ at θ : $T\mathcal{M}_\theta = \{\Phi_\theta\theta | \theta \in \mathbb{R}^n\}$ where $(\Phi_\theta)_{s,i} = \frac{\partial}{\partial\theta_i} V_\theta(s)$:

$$J(\theta) = \|\Pi_\theta(TV_\theta - V_\theta)\|$$

Visually...



Υ_θ : projection on the tangent plane
$\Pi_\theta(TV_\theta - V_\theta) = \Upsilon_\theta TV_\theta - V_\theta$

Derivations

Let $\Phi_\theta \equiv \nabla V_\theta(s)$

$$J(\theta) = \mathbb{E}[\delta\Phi_\theta]^T \mathbb{E}[\Phi_\theta\Phi_\theta^T]\mathbb{E}[\delta\Phi_\theta]$$

Gradient of $J(\theta)$

$$\begin{aligned} -\frac{1}{2}\nabla J(\theta) &= -\mathbb{E}[(\gamma\Phi'_\theta - \Phi_\theta)\Phi_\theta^T\omega] + h(\theta,\omega) && \text{(GTD)} \\ &= -\mathbb{E}[\delta\Phi_\theta] - \gamma\mathbb{E}[\Phi'_\theta\Phi_\theta^T\omega] + h(\theta,\omega) && \text{(TDC)} \end{aligned}$$

where $\omega = \mathbb{E}[\Phi_\theta\Phi_\theta^T]^{-1}\mathbb{E}[\delta\Phi_\theta]$ and $h(\theta,\omega) = -\mathbb{E}[\nabla^2 V_\theta(s)\omega]$

In comparison with the linear case, the only difference is the presence of a second order term $h(\theta,\omega)$.

Non-linear GTD/TDC updates

Non-linear GTD

$$\theta_{k+1} = \theta_k + \alpha_k \left[(\Phi_k - \gamma \Phi'_k)(\Phi_k^T w_k) - h_k \right]$$

Non-linear TDC

$$\theta_{k+1} = \theta_k + \alpha_k \left[\delta_k \Phi_k - \gamma \Phi'_k (\Phi_k^T w_k) - h_k \right]$$

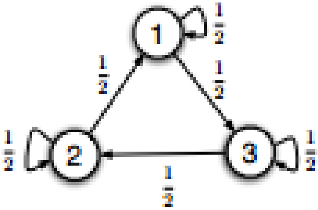
Same approximation trick for ω (and h) as in the linear case :

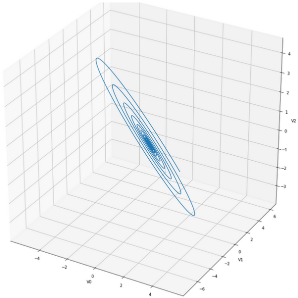
$$\omega_{k+1} = \omega_k + \beta_k (\delta_k - \Phi_k^T \omega_k) \Phi_k$$
$$h_k = (\delta_k - \Phi_k^T w_k) \nabla^2 V_{\theta_k}(s_k) \omega_k$$

The spiral counterexample

From cite

$$V_\theta(s) = (a[s]\cos(\lambda\theta - b[s]\sin(\lambda\theta)))\exp(\epsilon\theta)$$





The spiral counterexample