Convergent Temporal-Difference Learning with Arbitrary Smooth Function Approximation

Hugo Berard

COMP-767 Reinforcement Learning, McGill University

TD with Linear Approximation

Let's define the error function:

$$MSBE = ||V_\theta - TV_\theta||_D^2$$

We can project TV_θ on the linear space $\mathcal{M} = \{V_\theta = \Phi\theta | \theta \in \mathbb{R}^n\}$:

$$MSPBE = ||V_\theta - \Pi TV_\theta||_D^2$$

where $\Pi = \Phi(\Phi^T D\Phi)^{-1}\Phi^T D$ is the projection operator.

We can rewrite the error:

$$MSPBE = \mathbb{E}[\delta\phi]^T \mathbb{E}[\phi\phi^T]^{-1}\mathbb{E}[\delta\phi]$$

where $\mathbb{E}[\phi\phi^T] = \sum_s d_s\phi_s\phi_s^T = \Phi^T D\Phi$ and
$\mathbb{E}[\delta\phi^T] = \sum_s d_s\phi_s(R_s + \gamma\sum_s' P_{ss'}V_\theta(s') - V_\theta(s)) = \Phi^T D(TV_\theta - V_\theta)$

TD with Linear Approximation

If we take the gradient with respect to the parameters we can derive two algorithms:

GTD2:

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

where $\omega = \mathbb{E}[\phi\phi^T]^{-1}\mathbb{E}[\delta\phi]$

If we approximate ω with a linear predictor, we get the update rules:
$\theta_{k+1} = \theta_k + \alpha_k(\phi_k - \gamma\phi'_k)(\phi_k^T\omega_k)$ and $\omega_{k+1} = \omega_k + \beta_k(\delta_k - \phi_k^T\omega_k)\phi_k$

TDC:

$$-\frac{1}{2}\nabla MSPBE = \mathbb{E}[\delta\phi] - \gamma\mathbb{E}[\phi'\phi^T]\omega$$

If we approximate ω with a linear predictor, we get the update rules:
$\theta_{k+1} = \theta_k + \alpha_k(\delta_k\phi_k - \gamma\phi'_k(\phi_k^T\omega_k))$

Nonlinear Temporal Difference Learning

Let's define the tangent space $T\mathcal{M}_\theta$ as the plane orthogonal to the normal of $\mathcal{M}$ at θ and that passes through the origin.
$T\mathcal{M}_\theta = \{\Phi_\theta a | a \in \mathbb{R}^n\}$ where $(\Phi_\theta)_{s,i} = \frac{\partial}{\partial \theta_i} V_\theta(s)$.

The projection operator Π_θ on $T\mathcal{M}_\theta$ is thus similar to the linear case:

$$\Pi_\theta = \Phi_\theta(\Phi_\theta^T D \Phi_\theta)^{-1}\Phi_\theta^T D$$

and the objective functions becomes:

$$MSPBE = ||\Pi_\theta(V_\theta - TV_\theta)||_D^2$$

Similarly to the linear case we can rewrite the error:

$$MSPBE = \mathbb{E}[\delta \nabla V_\theta(s)]^T \mathbb{E}[\nabla V_\theta(s) \nabla V_\theta(s)]^{-1} \mathbb{E}[\delta \nabla V_\theta(s)]$$

Nonlinear Temporal Difference Learning

As in the linear case we can derive two gradient update:

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

with $h(\theta,\omega) = -\mathbb{E}[(\delta - \phi^T\omega)\nabla^2 V_\theta(s)\omega]$

and the updates becomes:

- GTD2: $\theta_{k+1} = \Gamma(\theta_k + \alpha_k((\phi_k - \gamma\phi'_k)(\phi_k^T\omega_k) - h_k))$

- TDC: $\theta_{k+1} = \Gamma(\theta_k + \alpha_k(\delta_k\phi_k - \gamma\phi'_k(\phi_k^T\omega_k) - h_k))$

and $h_{k+1} = (\delta_k - \phi_k^T\omega_k)\nabla^2 V_{\theta_k}(s_k)\omega_k$

where Γ is a projection on a compact set, and is necessary for convergence proof

Convergence proof

Let's rewrite the updates:

$$\omega_{k+1} = \omega_k + \beta_k(f(\theta_k, \omega_k) + M_{k+1})$$

$$\theta_{k+1} = \theta_k + \alpha_k(g(\theta_k, \omega_k) + N_{k+1}$$

with $f(\theta_k, \omega_k) = \mathbb{E}[\delta_k \phi_k | \theta_k] - \mathbb{E}[\phi_k \phi_k^T | \theta_k]\omega_k$,
$M_{k+1} = (\delta_k - \phi_k^T \omega_k)\phi_k - f(\theta_k, \omega_k)$,
$g(\theta_k, \omega_k) = \mathbb{E}[(phi_k - \gamma\phi_k')\phi_k^T \omega_k - h_k | \theta_k, \omega_k]$, and
$N_{k+1} = ((\phi_k - \gamma\phi_k')\phi_k^T \omega_k - h_k) - g(\theta_k, \omega_k)$

Convergence proof

We need to show that:

a) f and g are Lipshchitz continuous over a compact set $\mathcal{B}$.

b) $\mathbb{E}[M_{k+1}|\theta_k, \omega_k] = 0$ and $\mathbb{E}[N_{k+1}|\theta_k, \omega_k] = 0$

c) $(\omega_k(\theta), \theta)$ almost surely stays in $\mathcal{B}$ for any initial $(\omega_0(\theta), \theta) \in \mathcal{B}$

d) (ω, θ_k) almost surely stays in $\mathcal{B}$ for any initial $(\omega, \theta_0) \in \mathcal{B}$

From these conditions it follows that θ_k converges almost surely to the equilibria: $\dot{\theta} = \hat{\Gamma}(-\frac{1}{2}\nabla MSPBE)(\theta)$, where $\hat{\Gamma}v(\theta)$ is the projection to the of v on the tangent space of the compact set $\mathcal{C}$ at θ.

Convergence proof

a) is satisfied if V_θ is 3 times differentiable.

b) is straight forward from the definition of M_{k+1} and N_{k+1}

c) $w_k(\theta)$ converges to ω_θ which stays bounded if θ comes from a bounded set.

d) $\theta_k \in \mathcal{C}$ and $\mathcal{C}$ is a compact set thus θ_k is bounded.