Extending Gradient-based TD and TDC to Nonlinear Cases

COMP 767 – Reinforcement Learning

Mathieu Nassif

McGill University

March 24, 2017

Convergent Temporal-Difference Learning with Arbitrary Smooth Function Approximation

By Hamid R. Maei, Csaba Szepesvári, Shalabh Bhatnagar, Doina Precup, David Silver and Richard S. Sutton

Context

- Value Function v
- Value Function Approximation v_θ
- Objective: Find value of parameter θ

Context

- Value Function v
- Value Function Approximation v_θ
- Objective: Find value of parameter θ

- Finite MDP
- Error function: mean-square projected Bellman error

Context

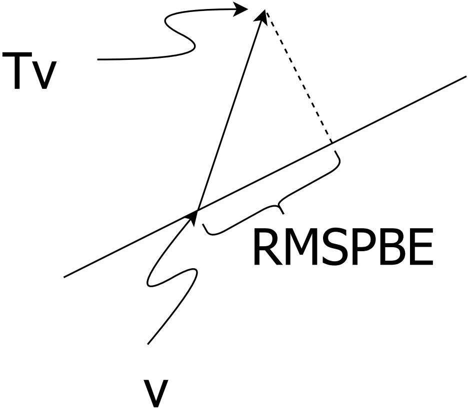
- Value Function v
- Value Function Approximation v_θ
- Objective: Find value of parameter θ

- Finite MDP
- Error function: mean-square projected Bellman error

- Linear Case: $v_\theta(s) = \theta^T \phi(s)$
- $\phi : \mathcal{S} \to \mathbb{R}^n$

Current Approaches

Minimize mean-square projected Bellman error



Implementation

Based on $w \approx \mathbb{E}[\phi\phi^T]^{-1}\mathbb{E}[\delta\phi]$

$\triangleright$ $w_{k+1} = w_k + \beta_k(\delta_k - \phi_k^T w_k)\phi_k$

Gradient-based TD (GTD2)

$\triangleright$ $\theta_{k+1} = \theta_k + \alpha_k(\phi_k - \gamma\phi'_k)(\phi_k^T w_k)$

TD with corrections (TDC)

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

Implementation

Based on $w \approx \mathbb{E}[\phi\phi^T]^{-1}\mathbb{E}[\delta\phi]$

$\triangleright\ w_{k+1} = w_k + \beta_k(\delta_k - \phi_k^T w_k)\phi_k$

Gradient-based TD (GTD2)

$\triangleright\ \theta_{k+1} = \theta_k + \alpha_k(\phi_k - \gamma\phi'_k)(\phi_k^T w_k)$

TD with corrections (TDC)

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

- Converge almost surely in the linear case
- Each step executes in $O(n)$

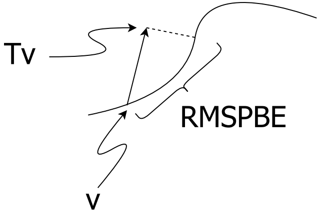
Nonlinear Cases

What if the approximation function is not linear? Is there a way to adapt the preceding algorithms?

Problem

What can be the objective function?

Naive projected Bellman error

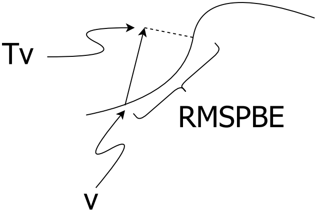


- Computationally hard

Problem

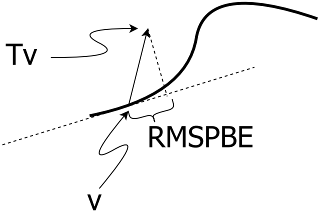
What can be the objective function?

Naive projected Bellman error



- Computationally hard

Alternative function



- Much easier

Implementation I

Based on the expression of the MSPBE as

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

(similar to the linear case: $MSPBE(\theta) = \mathbb{E}[\delta\phi]^T \mathbb{E}[\phi\phi^T]^{-1} \mathbb{E}[\delta\phi]$)
We substitute $\nabla v_\theta(s) = \phi$.

We introduce

$$h_k = (\delta_k - \phi_k^T w_k) \nabla^2 v_{\theta_k}(s_k) w_k \tag{1}$$

Weights:

$$w_{k+1} = w_k + \beta_k (\delta_k - \phi_k^T w_k) \phi_k \tag{2}$$

Difference is in the ϕ. Otherwise, it is the same expression.

Implementation II

GTD2:

$$\theta_{k+1} = \Gamma\left(\theta_k + \alpha_k\left((\theta_k - \gamma\theta'_k)(\theta_k^T w_k) - h_k\right)\right) \tag{3}$$

TDC:

$$\theta_{k+1} = \Gamma\left(\theta_k + \alpha_k\left(\delta_k\phi_k - \gamma\phi'_k(\phi_k^T w_k) - h_k\right)\right) \tag{4}$$

Where Γ is a projection into a compact set with a set boundary. This is used to ensure divergence cannot happen at the firsts stages. In practice, it is often unused.

Implementation III

These expressions result from expanding the gradient of the MSPBE.

Convergence for GTD2

- We focus on the proof of convergence of GTD2, for brevity. The proof of convergence for TDC is similar.
- Due to time/space constraints, some details of the proof will be left apart. We will focus on the intuitive ideas where the technical details could hide them.

Conditions

Throughout the proof, we assume the following.

- $v_\theta(s)$ is at least three times continuously differentiable with respect to θ, for any s where $d(s) > 0$.
- The sequences $\{\alpha_k\}$ and $\{\beta_k\}$, $k \in \mathbb{N} \cup \{0\}$, contain only positive elements and respect $\sum \alpha_k = \infty$ and $\sum \alpha_k^2 < \infty$ (similarly for the β_k). Also, $\lim_{k\to\infty} \frac{\alpha_k}{\beta_k} = 0$.
- All matrices for which we take the inverse are non-singular.

- The notation and context is the same as the one seen in class.

Gradient of the Objective Function

Let $J(\theta)$ represents the mean-square projected Bellman error. Then,

$$\frac{1}{2}[\nabla J(\theta)]_i = -(\partial_i \mathbb{E}[\delta\phi])^T \mathbb{E}[\phi\phi^T]^{-1}\mathbb{E}[\delta\phi] - \frac{1}{2}\mathbb{E}[\delta\phi]^T \partial_i(\mathbb{E}[\phi\phi^T]^{-1})\mathbb{E}[\delta\phi]$$
$$= -\mathbb{E}[\partial_i(\delta\phi)]^T w + \frac{1}{2} w^T \mathbb{E}[\partial_i(\phi\phi^T)]w$$
$$= -\mathbb{E}[(\partial_i\delta)\phi^T w] - \mathbb{E}[\delta(\partial_i\phi^T)w] + \mathbb{E}[\phi^T w(\partial_i\phi^T)w]$$

First line is applying the gradient, second line is using the definition of w and changing the order of expectation and derivative, and third line is using the identity $\frac{1}{2}w^T\partial_i(\phi\phi^T)w = \phi^T w(\partial_i\phi^T)w$.

Finally, using $\nabla\delta = \gamma\phi' - \phi$ and $\nabla\phi^T = \nabla^2 v_\theta(s)$

$$\frac{1}{2}[\nabla J(\theta)]_i = -\mathbb{E}[(\gamma\phi' - \phi)\phi^T w] - \mathbb{E}[(\delta - \phi^T w)\nabla^2 v_\theta(s) w]$$

We rewrite the first term as $-\mathbb{E}[\delta\phi] - \gamma\mathbb{E}[\phi'\phi^T w]$ and the last term as $h(\theta, w)$.

Proof of Convergence I

We show a (partial) proof of convergence of GTD2, omitting some details to focus on high-level ideas.
The proof of convergence is done in 4 steps.

1. Rewrite the equations 2 and 3 as

$$w_{k+1} = w_k + \beta_k(f(\theta_k, w_k) + M_{k+1})$$

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

with

$$f(\theta_k, w_k) =$$

Proof of Convergence II

2. We can show that there exist a compact set $B \subset \mathbb{R}^{2n}$ such that
   1. Functions f and g are Lipschitz continuous over B, because $v_\theta(s)$ is three times continuously differentiable,
   2. $(M_k, \mathcal{G}_k)$ and $(N_k, \mathcal{G}_k)$ are martingale difference sequences (a softer condition than i.i.d. sequence), where $\mathcal{G}_k$ is the sigma field generated by θ_i, w_i, r_i, s_i, $0 \leqslant i \leqslant k$ and s'_j, $0 \leqslant j < k$, by definition and $\mathbb{E}[M_{k+1}|\mathcal{G}_k] = \mathbb{E}[N_{k+1}|\mathcal{G}_k] = 0$.
   3. Given a starting point in B, the sequences $\{(w_k(\theta), \theta)\}$ and $\{(w, \theta_k)\}$ stay in B almost surely. This follows using convergence, and because we work in a compact set.

The last condition shows that the set will be used to contain (almost surely) the values of the iteration.

Proof of Convergence III

3. Given the operator $\hat{\Gamma}$, such that $\hat{\Gamma} v(\theta)$ is $v(\theta)$ if θ is in the interior of the compact set C, and its projection to the tangent space at θ otherwise.

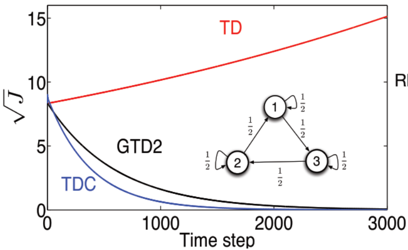
   Using a similar method (and intuition) as for the linear case, we show that the sequence of θ_k converges almost surely to the set of asymptotically stable equilibria of $\dot{\theta} = \hat{\Gamma} g(\theta, w_\theta)$, where w_θ is the equilibrium point of $\dot{w} = \mathbb{E}[\delta_\theta \phi_\theta] - \mathbb{E}[\phi_\theta \phi_\theta^T] w_\theta$.

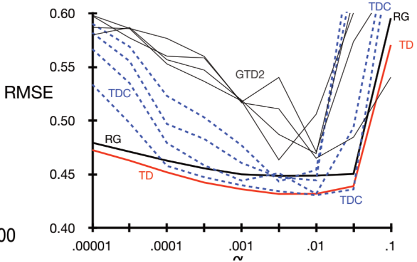
   $w_\theta = \mathbb{E}[\phi_\theta \phi_\theta^T]^{-1} \mathbb{E}[\delta_\theta \phi_\theta]$

Proof of Convergence IV

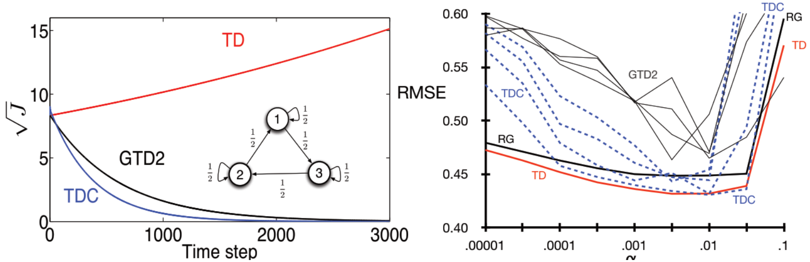
4. Finally, using the expression of the gradient of the objective function, we see that $g(\theta, w_\theta) = -\frac{1}{2}\nabla J(\theta)$. Thus, the iterations converge.

A Quick Note on Empirical Results





A Quick Note on Empirical Results



Nonlinear methods converge, whereas traditional TD diverges!
TDC performs almost as well as TD, but GTD2 is slightly worse.