On Convergence of Emphatic Temporal-Difference Learning

Huizhen Yu
Reinforcement Learning and Artificial Intelligence Laboratory
Department of Computing Science, University of Alberta

28th Annual Conference on Learning Theory
Paris, France, July 6, 2015
Background: Off-Policy TD Learning

Off-policy policy evaluation problem

- MDP: finite state/action spaces, state-dependent discount factors
- Approximate policy evaluation with linear function approximation
- Evaluate multiple \textit{target policies} from one exploratory \textit{behavior policy}
- \textit{Standard TD algorithms need not converge} (Baird 1995; Tsitsiklis & Van Roy 1997)
- Least squares or gradient based methods are more complex

\textbf{Emphatic TD learning} (Sutton, Mahmood & White 2015)

- Aim to solve a projected multistep Bellman equation (like TD)
- \textit{Employ a novel weighting scheme} weight each time step \textit{non-uniformly} accommodate users’ interest in learning values at specific states yet \textit{maintain a salient stability property} (which TD lacks)
Emphatic TD Algorithms

Inputs: \( \{(S_t, A_t, R_t)\} \) from the behavior policy, interest function \( i: \mathcal{S} \rightarrow \mathbb{R}^+ \)

Outputs: \( \theta_t \in \mathbb{R}^n \), parameters of approximate value functions

**ETD(\( \lambda \))**:

\[
\theta_{t+1} = \theta_t + \alpha_t e_t \cdot \rho_t \left( R_t + \gamma_{t+1} \phi(S_{t+1})^\top \theta_t - \phi(S_t)^\top \theta_t \right)
\]

where

\[
e_t = \lambda_t \gamma_t \rho_{t-1} e_{t-1} + M_t \phi(S_t)
\] (eligibility trace)

\[
F_t = \gamma_t \rho_{t-1} F_{t-1} + i(S_t)
\] (follow-on trace)

\[
M_t = \lambda_t i(S_t) + (1 - \lambda_t) F_t
\] (emphasis for \( S_t \))

**ELSTD(\( \lambda \))**:\n
\[
\theta_t = -C_t^{-1} b_t
\] where

\[
C_{t+1} = (1 - \alpha_t) C_t + \alpha_t e_t \cdot \rho_t (\gamma_{t+1} \phi(S_{t+1})^\top - \phi(S_t)^\top)
\]

\[
b_{t+1} = (1 - \alpha_t) b_t + \alpha_t e_t \cdot \rho_t R_t
\]

*Emphasis weights reflect the occupation frequencies of the target policy rather than the behavior policy!*
Our Results: Stability and Convergence

“Mean ODE” for ETD(\(\lambda\)): \(\dot{\theta} = C\theta + b\)

- \(C \preceq 0\) always, thanks to the emphatic weighting scheme (Sutton et al., 2015).
- If \(C \prec 0\) (negative definite), the desired solution \(\theta^*\) with \(C\theta^* + b = 0\) is globally asymptotically stable.

We prove:

**Theorem (Stability property of \(C\)).**

\(C \prec 0\) iff the set of feature vectors of emphasized states, \(\{\phi(s) | s \in \mathcal{S}, \bar{M}_{ss} > 0\}\), contains \(n\) linearly independent vectors.

**Sufficient condition for \(C \prec 0\):**

\(\{\phi(s) | s \in \mathcal{S}, i(s) > 0\}\) contains \(n\) linearly independent vectors.

Can be satisfied easily **without model knowledge**
Our Results: Stability and Convergence

Main conditions in our analysis:
an ergodicity condition on the behavior policy, negative definiteness of $C$, standard diminishing stepsize condition $\sum_t \alpha_t = \infty, \sum_t \alpha_t^2 < \infty$

We prove:

**Theorem (Convergence of ELSTD($\lambda$)).**
For any given initial $(e_0, F_0, C_0, b_0)$, $\{(C_t, b_t)\}$ converges in $L^1$:

\[
\lim_{t \to \infty} \mathbb{E}[\|C_t - C\|] = 0, \quad \lim_{t \to \infty} \mathbb{E}[\|b_t - b\|] = 0,
\]

and hence $\theta_t = -C_t^{-1}b_t \to \theta^*$ in probability. If, in addition, $\alpha_t = 1/(t+1)$, then

\[
C_t \overset{a.s.}{\to} C, \quad b_t \overset{a.s.}{\to} b, \quad \theta_t \overset{a.s.}{\to} \theta^*.
\]

**Theorem (Convergence of ETD($\lambda$)).**
With $\alpha_t = O(1/t)$ and $\frac{\alpha_t - \alpha_{t+1}}{\alpha_t} = O(1/t)$, for any given initial $(e_0, F_0, \theta_0)$, $\theta_t \overset{a.s.}{\to} \theta^*$.

Proofs use: properties of trace iterates and the Markov chain $\{(S_t, A_t, e_t, F_t)\}$ including ergodicity; a line of analysis from (Yu 2012) for off-policy TD/LSTD; a convergence theorem for SA algorithms based on the “mean ODE” approach (Kushner & Yin 2003); relation between ETD and its constrained variant.

