An Emphatic Approach to the problem of Off-policy Temporal-Difference Learning

Sutton, R. S., Mahmood A. R., White M.

Greta Laage

March 16, 2017

Abstract

Off-policy TD learning with function approximation

- Emphasizing or de-emphasizing updates on different time steps
- Certain ways lead to stability under off-policy training
- One learned parameter vector and one step-size parameter

Gradient TD methods: TDC, $GTD(\lambda)$, $GQ(\lambda)$

- Model-free Gradient-TD methods with updates in $O(n)$
- Stability under off-policy training

Specificities of $ETD(\lambda)$

- State-dependent discounting
- Bootstrapping functions
- Varying interests for states

Off-Policy TD(0)

Off policy settings

Data from a continuing finite MDP

Behavior policy $\mu \neq \pi$ *target policy* $\boldsymbol{d}_\mu = \mathbb{P}[\underset{t \to \infty}{S_t = s}]$

Assumption of coverage: $\pi(a|s) > 0 \Rightarrow \mu(a|s) > 0$

Off Policy $TD(0)$

$$\theta_{t+1} = \theta_t + \rho_t \alpha \left(R_{t+1} + \gamma \theta_t^T \phi_{t+1} - \theta_t^T \phi_t \right) \phi_t$$

$$\theta_{t+1} = \theta_t + \alpha \left(\rho_t R_{t+1} \phi_t - \rho_t \phi_t (\phi_t - \gamma \phi_{t+1})^T \theta_t \right)$$

$\boldsymbol{A}$ matrix

$$\boldsymbol{A} = \boldsymbol{\Phi}^T \boldsymbol{D}_\mu (\boldsymbol{I} - \gamma \boldsymbol{P_\pi}) \boldsymbol{\Phi}$$

- Columns sum may be negative
- A not positive definite
- Divergence of the parameter is likely

Instability of Off-Policy $TD(0)$

Example of divergence

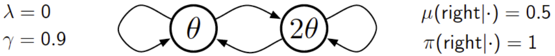


Figure 1: $\theta \rightarrow 2\theta$ example without a terminal state.

- $A = -0.2 < 0 =>$ Expected update and algorithm are not stable
- Only 2 transitions to the right that create updates and occur equally often
- From 1 to 2: $\theta + 16\alpha$ and from 2 to 2: $\theta - 8\alpha =>$ divergence

Emphatic $TD(\lambda)$

Emphasizing or de-emphasizing updates on different time steps

- Varies emphasis so as to reweight the distribution of linear $TD(\lambda)$ updates
- Goal : Creating a weighting equivalent to the *followon distribution*

Emphatic $TD(\lambda)$

Emphasizing or de-emphasizing updates on different time steps

- Varies emphasis so as to reweight the distribution of linear $TD(\lambda)$ updates
- Goal : Creating a weighting equivalent to the *followon distribution*

followon distribution: weights states according to their number of occurences before termination if the agent follows the target policy.

Stability: expected update over the distribution is a contraction (positive definite matrix). Prerequisite for full convergence of the stochastic algorithm.

Emphatic $TD(0)$

Off policy issue

μ may take the process to $d_\mu \neq d_\pi$ while the states might be similar because of FA.

Emphatic $TD(0)$

Off policy issue

μ may take the process to $d_\mu \neq d_\pi$ while the states might be similar because of FA.

Emphatic approach: New contemplated excursion from the current state at every time step:

- Excursion begin in a state sampled from d_μ following π
- Sequence of states and actions would exist
- Product of importance sampling ratios since the beginning of the excursion

Update at t emphasized proportional to a new scalar F_t, corrects for the state distribution.

Emphatic $TD(0)$

Off policy issue

μ may take the process to $d_\mu \neq d_\pi$ while the states might be similar because of FA.

Emphatic approach: New contemplated excursion from the current state at every time step:

- Excursion begin in a state sampled from d_μ following π
- Sequence of states and actions would exist
- Product of importance sampling ratios since the beginning of the excursion

Update at t emphasized proportional to a new scalar F_t, corrects for the state distribution.

$f(s) = d_\mu(s)\mathbb{E}_\mu[\underset{t\to\infty}{F_t} | S_t = s]$ is the *followon trace*. It is the expected number of time steps that would be spent in each state during an excursion starting from d_μ.

Emphatic $TD(0)$

Emphasis

$$F_t = \gamma \rho_{t-1} F_{t-1} + 1$$

Algorithm update

$$\begin{aligned}\theta_{t+1} &= \theta_t + \alpha \rho_t F_t \left(R_{t+1} + \gamma \theta_t^T \phi_{t+1} - \theta_t^T \phi_t \right) \phi_t \\ &= \theta_t + \alpha \left(\rho_t F_t \phi_t R_{t+1} - \rho_t F_t \phi_t (\phi_t - \gamma \phi_{t+1})^T \theta_t \right)\end{aligned}$$

$\boldsymbol{A}$ matrix

$$\boldsymbol{A} = \boldsymbol{\Phi}^T \boldsymbol{F} (\boldsymbol{I} - \gamma \mathbf{\Phi}_\pi) \boldsymbol{\Phi}$$

Diagonal of $\boldsymbol{F}$: $f(s) = d_\mu(s) \mathbb{E}_\mu [\underset{t \to \infty}{F_t} | S_t = s]$

$\boldsymbol{A}$ positive definite => algorithm stable

Emphatic $TD(0)$

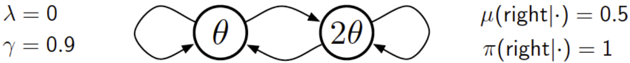


Figure 1: $\theta \rightarrow 2\theta$ example without a terminal state.

- $\boldsymbol{f(s)}$: $d_\mu +$ where to 1 step $+$ where to 2 steps ...
- $f(1) = d_\mu(1) = 0.5$: only in 1 if you start there
- $f(2) = 0.5 + 0.9 + 0.9^2 \ldots$: $\gamma = 0.9$, $\rho = 2$ and $\mu(right|.) = 0.5$

$\boldsymbol{F}$ emphasizes the second state which would occur more often under π compared to μ

General case of emphatic TD

Discount, interest and bootstrapping functions

$$\text{Discount function: } \gamma : \mathcal{S} \rightarrow [0,1] \text{ such that } \prod_{k=1}^{\infty} \gamma(S_{t+k} = 0) w.p.1$$

$$\text{Interest function: } i : \mathcal{S} \rightarrow [0, \infty[$$

$$\text{Bootstrapping function: } \lambda : \mathcal{S} \rightarrow [0,1]$$

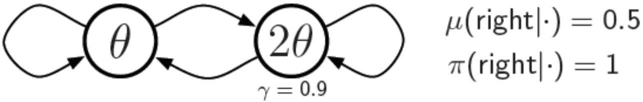
Specify a different degree of bootstrapping $1 - \lambda(s)$ for each state

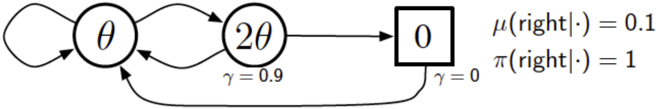
Objective function

$$MSVE(\theta) = \sum_{s \in \mathcal{S}} d_{\mu}(s) i(s) \left(v_{\pi} - \theta^T \phi(s) \right)^2$$

Empirical Example 1

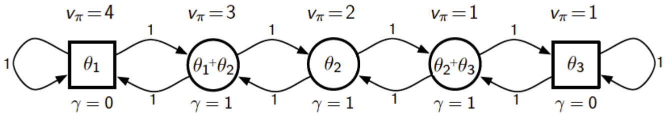
$$\lambda = 0$$
$$i(S_0) = 1, i(S_1) = 0$$





Empirical Example 2

$\lambda = 0$
$i(s) = 1 \quad \forall s$



$\mu(\text{left}|\cdot) = 2/3$
$\pi(\text{right}|\cdot) = 1$