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

Richard S. Sutton, A. Rupam Mahmood, Martha White

COMP767 - Reinforcement Learning

Claudio Sole

École polytechnique de Montréal

March 31, 2017

Introduction

- Setting:
  - off-policy $\rightarrow$ learn about v_π while behaving according to a *behaviour policy* μ
  - linear function approximation: $v_\pi(S_t) \approx \theta_t^T \phi(S_t)$
- Goal:
  Create a weighting equivalent to the **followon distribution**, which weights states according to how often they would occur prior to termination by discounting if the target policy was followed.
- Focus:
  Prove stability of the resulting algorithm, the *emphatic TD(λ)*: the expected update is a contraction involving a positive definite matrix

On-policy stability of TD(0) (1)

- Rewrite the TD(0) update to highlight stability issue:

$$
\begin{aligned}
\theta_{t+1} &= \theta_t + \alpha\Big(R_{t+1} + \gamma\theta_t^T\phi(S_{t+1}) - \theta_t^T\phi(S_t)\Big)\phi(S_t) \\
&= \theta_t + \alpha\Big(\underbrace{R_{t+1}\phi(S_t)}_{b_t\in\mathbb{R}^n} - \underbrace{\phi(S_t)(\phi(S_t) - \gamma\phi(S_{t+1})}_{A_t\in\mathbb{R}^{n\times n}}\theta_t\Big) \\
&= (I - \alpha A_t)\theta_t + \alpha b_t
\end{aligned} \tag{1}
$$

- Only A_t multiplies θ and is therefore critical for convergence

Idea

Suppose A_t diagonal. If A_t has some negative values, then the corresponding elements of $(I - \alpha A_t)$ will be greater then one thus increasing θ_t, leading to divergence. In general, θ_t will be reduced toward zero whenever A_t is positive definite.

On-policy stability of TD(0) (2)

Stability: Defined $A = \lim_{t\to\infty} \mathbb{E}[A_t]$ and $b = \lim_{t\to\infty} \mathbb{E}[b_t]$, a stochastic algorithm in the form of (1) is stable if the corresponding deterministic algorithm

$$\theta_{t+1} = \theta_t + \alpha(b - A\theta_t)$$

converges to unique fixed point independent form θ_0

$$\begin{aligned} A &= \lim_{t\to\infty} \mathbb{E}[A_t] \\ &= \Phi^T \underbrace{D_\pi(I - \gamma P_\pi)}_{key\ \ matrix} \Phi \end{aligned}$$

Following Sutton(1988) and Varga(1962), to assure the key matrix is positive definite we want to show that all his columns sum to a nonnegative number

On-policy stability of TD(0) (3)

To compute the columns sums:

$$\begin{aligned} 1^T D_\pi(I - \gamma P_\pi) &= d_\pi^T(I - \gamma P_\pi) \\ &= d_\pi^T - \gamma d_\pi^T P_\pi \\ &= d_\pi^T - \gamma d_\pi^T \\ &= (1 - \gamma) d_\pi \end{aligned}$$

all components of which are positive.

Instability of Off-policy TD(0)

- Importance sampling ratios $\rho_t = \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$
- Off-policy TD(0) update:

$$
\begin{aligned}
\theta_{t+1} &= \theta_t + \alpha\rho_t\Big(R_{t+1} + \gamma\theta_t^T\phi(S_{t+1}) - \theta_t^T\phi(S_t)\Big)\phi(S_t) \\
&= \theta_t + \alpha\Big(\underbrace{\rho_t R_{t+1}\phi_t}_{b_t} - \underbrace{\rho_t\phi_t(\phi_t - \gamma\phi_{t+1})^T}_{A_t}\theta_t\Big)
\end{aligned}
$$

- And thus the matrix A becomes

$$A = \lim_{t\to\infty}\mathbb{E}[A_t] = \Phi^T D_\mu(I - \gamma P_\pi)\Phi$$

Instability of Off-policy TD(0): Example

- r = 0 for every transition

$\lambda = 0$
$\gamma = 0.9$

$\mu(\text{right}|\cdot) = 0.5$
$\pi(\text{right}|\cdot) = 1$

- Columns of the key matrix may sum to negative number

$$\mathbf{D}_\mu(\mathbf{I} - \gamma\mathbf{P}_\pi) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \times \begin{bmatrix} 1 & -0.9 \\ 0 & 0.1 \end{bmatrix} = \begin{bmatrix} 0.5 & -0.45 \\ 0 & 0.05 \end{bmatrix}.$$

- Effect on updates ($\theta_0 = 10$):

$$\begin{aligned} \theta_{t+1} &= \theta_t + \rho_t\alpha\left(R_{t+1} + \gamma\boldsymbol{\theta}_t^\top\phi_{t+1} - \boldsymbol{\theta}_t^\top\phi_t\right)\phi_t \\ &= 10 + 2\cdot 0.1\,(0 + 0.9\cdot 10\cdot 2 - 10\cdot 1)\,1 \\ &= 10 + 1.6, \end{aligned}$$

$$\begin{aligned} \theta_{t+1} &= \theta_t + \rho_t\alpha\left(R_{t+1} + \gamma\boldsymbol{\theta}_t^\top\phi_{t+1} - \boldsymbol{\theta}_t^\top\phi_t\right)\phi_t \\ &= 10 + 2\cdot 0.1\,(0 + 0.9\cdot 10\cdot 2 - 10\cdot 2)\,2 \\ &= 10 - 0.8. \end{aligned}$$

Since these updates happen with same frequency ($d_\mu = [0.5, 0.5]$) a divergence occurs.

Off-policy stability of emphatic TD(0) (1)

- Followon trace:

$$F_t = \gamma \rho_{t-1} F_{t-1} + 1 \ \forall t > 0, F_0 = 1$$

- So the TD(0) update becomes:

$$\begin{aligned}\theta_{t+1} &= \theta_t + \alpha F_t \rho_t \Big(R_{t+1}\phi_t + \gamma \theta_t^T \phi_{t+1} - \theta_t^T \phi_t \Big) \phi_t \\ &= \theta_t + \alpha \Big(\underbrace{F_t \rho_t R_{t+1} \phi_t}_{b_t} - \underbrace{F_t \rho_t \phi_t (\phi_t - \gamma \phi_{t+1})^T}_{A_t} \theta_t \Big)\end{aligned}$$

- The expected A matrix thus becomes:

$$A = \lim_{t\to\infty} \mathbb{E}[A_t] = \Phi^T F (I - \gamma P_\pi) \Phi$$

where F is the diagonal matrix with diagonal elements $f(s) = d_\mu \lim_{t\to\infty} \mathbb{E}_\mu[F_t | S_t = s]$ components of the vector f given by

$$f = \Big(I - \gamma P_\pi^T \Big)^{-1} d_\mu$$

Off-policy stability of emphatic TD(0) (2)

- f is the expected number of steps that would be spent in each state during an excursion starting from the behaviour distribution d_μ and following π
- in the $\theta \to 2\theta$ example, the F matrix is

$$F = \begin{bmatrix} 0.5 & 0 \\ 0 & 9.5 \end{bmatrix}$$

$f(1) = d_\mu(1) = 0.5$

$f(2) = d_\mu(2) = 0.5 + 0.9 + 0.9^2 + \ldots = 9.5$

thus giving much more importance to the lower row

- what's the effect of F in the key matrix?

off-policy TD(0)

$$D_\mu(I - \gamma P_\pi) = \begin{bmatrix} 0.5 & -0.45 \\ 0 & 0.05 \end{bmatrix}$$

off-policy emphatic TD(0)

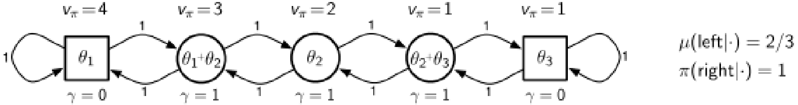
$$F(I - \gamma P_\pi) = \begin{bmatrix} 0.5 & -0.45 \\ 0 & 0.95 \end{bmatrix}$$

Generalizations I

- **discount function** $\gamma : S \to [0,1]$ such that $\prod_{k=1}^{\infty} \gamma(S_t + k) = 0$ w.p. $1\ \forall t$. Allows soft termination:

$$G_t = R_{t+1} + \gamma(S_{t+1}) T_{t+2} + \gamma(S_{t+1})\gamma(S_{t+2}) R_{t+3} + \ldots$$

  thus, if $\gamma(S_k) = 0$, the rewards accumulation is fully terminated at step $k > t$



Generalizations II

- **interest function** $i : S \to [0, \infty)$: explicitly specify states for which we want accurate estimates of value. The MSVE objective function thus becomes

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

- **Bootstrapping function** $\lambda : S \to [0, 1]$

References

[1] Sutton, R. S.,*Learning to predict by the methods of temporal diffrences*,Machine Learning 3:9–44, erratum p. 377., (1988)
[2] Sutton, Richard S., A. Rupam Mahmood, and Martha White., *"An emphatic approach to the problem of off-policy temporal-difference learning."*,The Journal of Machine Learning Research 17,(2015)
[3] Varga, R. S.,*Matrix Iterative Analysis*, Englewood Cliffs, NJ: Prentice-Hall,(1962)