Analysis of Tree backup algorithm

Ahmed Touati
University of Montreal
ahmed.touati@umontreal.ca

Abstract

We provide a new and old analysis of tree backup algorithm in tabular case as well as with linear function value approximation. The divergence issue that we show here motivates us to derive a new algorithm that we will develop in details in our class project.

1 Tabular tree backup

1.1 Definition

Tree-backup algorithm $TB(\lambda)$ is an off-policy multi-step temporal difference learning where samples generated by a behavior policy are used to learn a target policy. Tree-backup corrects the discrepancy between target/behavior policy by scaling returns by target policy probabilities.

The n-steps tree-backup return is defined by:

$$TB^{(n)} = \sum_{t=0}^{n} \gamma^t (\prod_{i=1}^{t} \pi_i)(r_t + \gamma \mathbb{E}_{\pi}^{a \neq a_{t+1}} Q(x_{t+1}, .)) + (\prod_{i=1}^{n+1} \pi_i) \gamma^{n+1} Q(x_{n+1}, a_{n+1})$$

where $\pi_i = \pi(x_i, a_i)$
The case n=0 recovers the expected SARSA return.
The λ return extension considers exponentially weighted sums of n-steps returns:

$$TB^{\lambda} = (1-\lambda) \sum_{n=0}^{\infty} \lambda^n \, TB^{(n+1)}$$

Let's rewrite the λ return:

$$\begin{aligned}
TB^{\lambda} &= (1-\lambda) \sum_{n=0}^{\infty} \lambda^n [\sum_{t=0}^{n} \gamma^t (\prod_{i=1}^{t} \pi_i)(r_t + \gamma \mathbb{E}_{\pi}^{a \neq a_{t+1}} Q(x_{t+1}, .)) + (\prod_{i=1}^{n+1} \pi_i) \gamma^{n+1} Q(x_{n+1}, a_{n+1})] \\
&= (1-\lambda) \sum_{t=0}^{\infty} \gamma^t (\prod_{i=1}^{t} \pi_i)(r_t + \gamma \mathbb{E}_{\pi}^{a \neq a_{t+1}} Q(x_{t+1}, .)) \sum_{n=t}^{\infty} \lambda^n + \sum_{n=0}^{\infty} (\prod_{i=1}^{n+1} \pi_i) \gamma^{n+1} Q(x_{n+1}, a_{n+1})(\lambda^n - \lambda^{n+1}) \\
&= \sum_{t=0}^{\infty} (\lambda\gamma)^t (\prod_{i=1}^{t} \pi_i)(r_t + \gamma \mathbb{E}_{\pi} Q(x_{t+1}, .) - \gamma Q(x_{t+1}, a_{t+1})) + \sum_{n=0}^{\infty} (\prod_{i=1}^{n+1} \pi_i) \gamma^{n+1} Q(x_{n+1}, a_{n+1})(\lambda^n - \lambda^{n+1}) \\
&= \sum_{t=0}^{\infty} (\lambda\gamma)^t (\prod_{i=1}^{t} \pi_i)(r_t + \gamma \mathbb{E}_{\pi} Q(x_{t+1}, .)) - \sum_{n=0}^{\infty} (\prod_{i=1}^{n+1} \pi_i)(\lambda\gamma)^{n+1} Q(x_{n+1}, a_{n+1}) \\
&= Q(x_0, a_0) + \sum_{t=0}^{\infty} (\lambda\gamma)^t (\prod_{i=1}^{t} \pi_i)(r_t + \gamma \mathbb{E}_{\pi} Q(x_{t+1}, .) - Q(x_t, a_t)) \\
&= Q(x_0, a_0) + \sum_{t=0}^{\infty} (\lambda\gamma)^t (\prod_{i=1}^{t} \pi_i) \delta_t^{\pi}
\end{aligned}$$

where $\delta_t^\pi = r_t + \gamma \mathbb{E}_\pi Q(x_{t+1}, .) - Q(x_t, a_t)$ The off-line update of tree back-up algorithm is then:

$$Q_{t+1}(x,a) = Q_t(x,a) + \alpha_t \sum_{t=0}^{\infty} (\lambda\gamma)^t (\prod_{i=1}^{t} \pi_i) \delta_t^\pi$$

where $x_1, a_1, r_1, ..., x_1, a_t, r_t, ...$ is trajectory generated by the policy μ

1.2 Convergence result

Convergence result could be obtained by applying general results of Robbins-Monro stochastic approximation methods for solving $Q = RQ$, when the mapping R is weighted maximum norm contraction (Prop 4.4 in [3]). Let's rewrite tree-backup update:

$$Q_{k+1}(x,a) = (1-\alpha_k) Q_k(x,a) + (1-\alpha_k)(RQ_k(x,a) + w_k(x,a))$$

where R is the tree-backup operator defined by:

$$\begin{aligned}
(RQ)(x,a) &= Q(x,a) + \mathbb{E}_\mu[\sum_{t=0}^{\infty} (\lambda\gamma)^t (\prod_{i=1}^{t} \pi_i) \delta_t^\pi] \\
&= Q(x,a) + \sum_{t=0}^{\infty} (\lambda\gamma)^t \mathbb{E}_{x_{1:t+1}, a_{1:t+1}}[(\prod_{i=1}^{t} \pi_i) \delta_t^\pi] \\
&= Q(x,a) + \sum_{t=0}^{\infty} (\lambda\gamma)^t \mathbb{E}_{x_{1:t}, a_{1:t}}[(\prod_{i=1}^{t} \pi_i)(r_t + \gamma \mathbb{E}_{x_{t+1}}[\mathbb{E}_\pi Q(x_{t+1}, .)|F_t] - Q(x_t, a_t))] \\
&= Q(x,a) + \sum_{t=0}^{\infty} (\lambda\gamma)^t \mathbb{E}_{x_{1:t}, a_{1:t}}[(\prod_{i=1}^{t} \pi_i)(r_t + \gamma \sum_{x' \in X} \sum_{a' \in A} p(x'|x_t, a_t)\pi(a'|x') Q(x', a') - Q(x_t, a_t))] \\
&= Q(x,a) + \sum_{t=0}^{\infty} (\lambda\gamma)^t \mathbb{E}_{x_{1:t}, a_{1:t}}[(\prod_{i=1}^{t} \pi_i)(r_t + \gamma P^\pi Q(x_t, a_t) - Q(x_t, a_t))] \\
&= Q(x,a) + \sum_{t=0}^{\infty} (\lambda\gamma)^t \mathbb{E}_{x_{1:t}, a_{1:t}}[(\prod_{i=1}^{t} \pi_i)(T^\pi Q(x_t, a_t) - Q(x_t, a_t))] \\
&= Q(x,a) + \sum_{t=0}^{\infty} (\lambda\gamma)^t \mathbb{E}_{x_{1:t-1}, a_{1:t-1}}[(\prod_{i=1}^{t-1} \pi_i) \sum_{x' \in X} \sum_{a' \in A} p(x'|x_{t-1}, a_{t-1})\pi(a'|x')\mu(a'|x') \\
&(T^\pi Q(x', a') - Q(x', a'))] \\
&= Q(x,a) + \sum_{t=0}^{\infty} (\lambda\gamma)^t \mathbb{E}_{x_{1:t-1}, a_{1:t-1}}[(\prod_{i=1}^{t-1} \pi_i) P^{\mu\pi}(T^\pi - I) Q(x_{t-1}, a_{t-1})] \\
&= Q(x,a) + \sum_{t=0}^{\infty} (\lambda\gamma)^t (P^{\mu\pi})^t (T^\pi - I) Q(x,a) \\
&= Q(x,a) + (I - \lambda\gamma P^{\mu\pi})^{-1}(T^\pi - I) Q(x,a)
\end{aligned}$$

where:

$$\begin{aligned}
P^\pi Q(x,a) &= \sum_{x' \in X} \sum_{a' \in A} p(x'|x,a)\pi(a'|x') Q(x', a') \\
P^{\pi\mu} Q(x,a) &= \sum_{x' \in X} \sum_{a' \in A} p(x'|x,a)\pi(a'|x')\mu(a'|x') Q(x', a') \\
T^\pi &= r + \gamma P^\pi
\end{aligned}$$

We obtain then

$$R = I + (I - \lambda\gamma P^{\mu\pi})^{-1}(T^\pi - I) = (I - \lambda\gamma P^{\mu\pi})^{-1}(T^\pi - \lambda\gamma P^{\mu\pi})$$

The noise term is defined by:

$$w_k(x,a) = Q_k(x,a) + \sum_{t=0}^{\infty}(\lambda\gamma)^t(\prod_{i=1}^{t}\pi_i)\delta_t^{\pi} - RQ_k(x,a)$$

In particular, we have $\mathbb{E}[w_k|F_k] = 0$ Q^{π} is fix point of the operator R. It lasts to show that R is a contraction with respect to the maximum norm $||||_{\infty}$.

$$\begin{aligned}
RQ - Q^{\pi} &= Q - Q^{\pi} + (I - \lambda\gamma P^{\mu\pi})^{-1}(\gamma P^{\pi} - I)(Q - Q^{\mu}) \\
&= (I - \lambda\gamma P^{\mu\pi})^{-1}(I - \lambda\gamma P^{\mu\pi} + P^{\pi} - I)(Q - Q^{\mu}) \\
&= \gamma(I - \lambda\gamma P^{\mu\pi})^{-1}(P^{\pi} - \lambda P^{\mu\pi})(Q - Q^{\pi})
\end{aligned}$$

So for all $(x,a) \in X, A$, we have: All the entries of the matrix $(I - \lambda\gamma P^{\mu\pi})^{-1}$ is non-negative, the entries of the matrix $P^{\pi} - \lambda P^{\mu\pi}$ are non-negative too as $p(x'|x,a)\pi(a'|x')(1 - \lambda\mu(a'|x')) \geq 0$. let $\mathbf{1}$ the vector whose all entries are equal to one. In particular, as we P^{π} is stochastic matrix, we have $P^{\pi}\mathbf{1} = \mathbf{1}$

$$\begin{aligned}
|RQ(x,a) - Q^{\pi}(x,a)| &= |\gamma(I - \lambda\gamma P^{\mu\pi})^{-1}(P^{\pi} - \lambda P^{\mu\pi})(Q(x',a') - Q^{\pi}(x',a'))| \\
&\leq \gamma(I - \lambda\gamma P^{\mu\pi})^{-1}(P^{\pi} - \lambda P^{\mu\pi})\mathbf{1}(x,a)||Q - Q^{\pi}||_{\infty} \\
&= \gamma(I - \lambda\gamma P^{\mu\pi})^{-1}(\mathbf{1} - \lambda P^{\mu\pi}\mathbf{1})(x,a)||Q - Q^{\pi}||_{\infty} \\
&= (\gamma\sum_{t\geq 0}(\gamma\lambda)^t(P^{\mu\pi})^t\mathbf{1} - \sum_{t\geq 0}(\gamma\lambda)^{t+1}(P^{\mu\pi})^{t+1}\mathbf{1})(x,a)||Q - Q^{\pi}||_{\infty} \\
&= [(1-\gamma)(\sum_{t\geq 0}(\gamma\lambda)^t(P^{\mu\pi})^t\mathbf{1})(x,a) + 1]||Q - Q^{\pi}||_{\infty} \\
&= [(\gamma-1)(1 + \sum_{t\geq 1}(\gamma\lambda)^t(P^{\mu\pi})^t\mathbf{1})(x,a)) + 1]||Q - Q^{\pi}||_{\infty} \\
&\leq [(\gamma-1) + 1]||Q - Q^{\pi}||_{\infty} \\
&= \gamma||Q - Q^{\pi}||_{\infty}
\end{aligned}$$

We conclude that $||RQ - Q^{\pi}||_{\infty} \leq \gamma||Q - Q^{\pi}||_{\infty}$ and the operator R is then γ pseudo-contraction around Q^{π} with respect to the maximum norm, we could then apply the Prop 4.4 in [3] and conclude that Q_t converges to R-fixed point Q^{π} with probability one.

2 Tree backup with linear Value Function approximation

We tackle in this section the following question:
Could we extend tabular Tree backup algorithm mechanistically to the linear Value function approximation setting?

2.1 Definition

As in the tabular case, we describe here the tree backup with VFA.
let $Q(x,a) = \theta^T\phi(x,a)$. The n-steps return:

$$TB^{(n)} = \sum_{t=0}^{n}\gamma^t(\prod_{i=1}^{t}\pi_i)(r_t + \gamma\mathbb{E}_{\pi}^{a\neq a_{t+1}}\theta^T\phi(x_{t+1},.)) + (\prod_{i=1}^{n+1}\pi_i)\gamma^{n+1}\theta^T\phi(x_{n+1},a_{n+1})$$

The λ-return is:

$$TB^{\lambda} = \theta^T\phi(x_0,a_0) + \sum_{t=0}^{\infty}(\lambda\gamma)^t(\prod_{i=1}^{t}\pi_i)\delta_t^{\pi}$$

where $\delta_t^{\pi} = r_t + \gamma\mathbb{E}_{\pi}\theta^T\phi(x_{t+1},.) - \theta^T\phi(x_t,a_t)$
The tree-backup with VFA is then:

$$\theta_{k+1} = \theta_k + \alpha_k(\sum_{t=0}^{\infty}(\lambda\gamma)^t(\prod_{i=1}^{t}\pi_i)\delta_t^{\pi})\phi(x,a)$$

2.2 Convergence understanding ??

In this section, we will analyze the convergence or divergence of the algorithm in the framework of the ODE (Ordinary differential equations) approach which is the main tool used in the convergence proofs for FVA algorithms. We consider in particular the Prop 4.8 in [3] which considers the Markov process defined by

$$\theta_{k+1} = \theta_k + \alpha_k(A(X_k)\theta + b(X_k))$$

where X takes values in a set X and A maps every $X \in X$ to a square matrix $A(X)$, b maps every $X \in X$ to a vector and α is a non-negative scalar stepsize. The Prop 4.8 states that under some conditions, the sequence θ_k converges to the unique solution of θ^* the system $A\theta * +b =$, where $A = \mathbb{E}[A(X_k)]$ and $b = \mathbb{E}[b(X_k)]$ where the expectation is with respect to the stationary distribution induced by the ergodic Markov chain X_k.
One of the crucial condition is the matrix A is negative definite.

Let's find the matrix A that corresponds to tree backup

$$\begin{aligned}
\theta_{k+1} &= \theta_k + \alpha_k(\sum_{t=0}^{\infty}(\lambda\gamma)^t(\prod_{i=1}^{t}\pi_i)[r_t + \gamma\mathbb{E}_\pi\theta^T\phi(x_{t+1},.) - \theta^T\phi(x_t,a_t)])\phi(x,a) \\
&= \theta_k + \alpha_k(\sum_{t=0}^{\infty}(\lambda\gamma)^t(\prod_{i=1}^{t}\pi_i)\phi(x,a)[\gamma\mathbb{E}_\pi\phi(x_{t+1},.)^T - \phi(x_t,a_t)^T]\theta_k + \sum_{t=0}^{\infty}(\lambda\gamma)^t(\prod_{i=1}^{t}\pi_i)r_t\phi(x,a)) \\
&= \theta_k + \alpha_k(A_k\theta + b_k)
\end{aligned}$$

where

$$\begin{aligned}
A_k &= \sum_{t=0}^{\infty}(\lambda\gamma)^t(\prod_{i=1}^{t}\pi_i)\phi(x,a)[\gamma\mathbb{E}_\pi\phi(x_{t+1},.)^T - \phi(x_t,a_t)^T] \\
b_k &= \sum_{t=0}^{\infty}(\lambda\gamma)^t(\prod_{i=1}^{t}\pi_i)r_t\phi(x,a)
\end{aligned}$$

Notice here that k is used to index trajectories whereas k indexed transition of the Markov chain in the Prop 4.8 in [3] but convergence results still applies in our case. (see also Prop 6.6 in [3]).

let's then compute the matrix $A = \mathbb{E}[A_k]$ where expectation is with respect the trajectories generated by the behavior policies μ. Let d be stationary distribution induced by μ. Using similar derivation as in the first section, we get:

$$\begin{aligned}
\mathbb{E}[\sum_{t=0}^{\infty}(\lambda\gamma)^t(\prod_{i=1}^{t}\pi_i)Q(x,a)[\gamma\mathbb{E}_\pi Q(x_{t+1},.) - Q(x_t,a_t)]] &= \mathbb{E}[Q(x,a)(I - \lambda\gamma P^{\mu\pi})^{-1}(\gamma P^\pi - I)Q(x,a)] \\
&= \sum_{x,a} d(x,a)Q(x,a)(I - \lambda\gamma P^{\mu\pi})^{-1}(\gamma P^\pi - I)Q(x,a) \\
&= Q^T D^\mu(I - \lambda\gamma P^{\mu\pi})^{-1}(\gamma P^\pi - I)Q
\end{aligned}$$

where D^μ is a diagonal matrix whose diagonal entries are the stationary probabilities
Now, if we consider $Q = \Phi\theta$ (Φ is a matrix whose rows are $\phi(x,a)$), we have $Q(x,a) = \theta^T\phi(x,a)$.
Then

$$\theta^T\mathbb{E}[\sum_{t=0}^{\infty}(\lambda\gamma)^t(\prod_{i=1}^{t}\pi_i)\phi(x,a)[\gamma\mathbb{E}_\pi\phi(x_{t+1},.) - \phi(x_t,a_t)]^T]\theta = \theta^T\Phi^T D^\mu(I - \lambda\gamma P^{\mu\pi})^{-1}(\gamma P^\pi - I)\Phi\theta$$

Since the vector θ is arbitrary, it follows that:

$$A = \Phi^T D^\mu(I - \lambda\gamma P^{\mu\pi})^{-1}(\gamma P^\pi - I)\Phi$$

Similarly, we could show that:

$$b = \Phi^T D^\mu(I - \lambda\gamma P^{\mu\pi})^{-1}r$$

If we assume that Φ is full rank, The matrix is negative definite if and only if the key matrix $K = D^{\mu}(I - \lambda\gamma P^{\mu\pi})^{-1}(I - \gamma P^{\pi})$ is negative definite.
Unfortunately, for arbitrary target/behavior policies, the matrix K is not necessarily negative positive.
In particular, in the case of $\lambda = 0$, $K = D^{\mu}(\gamma P^{\pi} - I)$ which is basically the matrix we obtain for off-policy temporal difference learning TD(0). So in this case, the matrix may have positive eigenvalues. (See Example 6.7 in [3]).

When the algorithm convergences, it convergences to $\theta^* = -A^{-1}b$. In [4], it was shown also that θ^* is the fixed point of the projected operator

$$\Phi\theta^* = \Pi^{\mu} R(\Phi\theta^*)$$

where $\Pi^{\mu} = \Phi(\Phi^T D^{\mu}\Phi)^{-1}\Phi^T D^{\mu}$ is the projection onto the space $S = \{\Phi\theta | \theta \in \mathbb{R}^d\}$ with respect to the weighted Euclidean norm $||.||_{D^{\mu}}$. So, Other way to estimate θ^* is by minimizing the Mean Squared Projected Error (MSPBE) given as follows:

$$\mathbf{MSPBE}(\theta) = \frac{1}{2}||\Pi^{\mu} R(\Phi\theta) - \Phi\theta||_{D^{\mu}}$$

Which gives the new algorithm that we will derive in our project.

[1] Doina Precup, Richard Sutton & Sanjoy Dasgupta (2000) Eligibility traces for off-policy evaluation, *International Conference in Machine Learning*.

[2] Tsitsiklis, J. N., and Van Roy, B. (1997) An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control* 42:674–690.

[3] Dimitry P. Bertsekas and John N. Tsitsiklis (1996) *Neuro-Dynamic Programming*.

[4] Richard S. Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvari and Eric Wiewiora (2009a)*A convergent O(n) algorithm for off-policy temporal-difference learning with linear function approximation*. In Advances in Neural Information Processing Systems 21, pp. 1609–1616. MIT Press

[5] Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, Marek Petrik (2015) *Finite-Sample Analysis of Proximal Gradient TD Algorithms*.Journal of Machine Learning Research (JMLR), 13:3041-3074, 2012