Analysis of Tree backup algorithm : Tabular case, Linear Function approximation and gradient correction

Ahmed Touati

April 21, 2017

Abstract

We provide a new and old analysis of tree backup algorithm in tabular case as well as with linear function value approximation. We derive a new algorithm based on Tree backup returns and with gradient correction in order to get convergence guarantees provided certain conditions.

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:

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

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 (Sutton 1998, appendix).
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 converges, it converges 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 described in the following section.

Note:An interesting link that we could established (and maybe is already known) is that the projected operator is a contraction with respect to the norm $||.||_{D^\mu}$ then the matrix A is negative definite. In fact, it was established in [4] that the projected operator is a contraction implies that the operator $f = D^\mu(I-R)$ is strongly monotone. $f(x) = D^\mu(x - R(x)) = (I-\lambda\gamma P^{\mu\pi})^{-1}(T^\pi - I)(x) = -Kx + (I-\lambda\gamma P^{\mu\pi})^{-1}r$. As f is a linear function, then f is monotone if and only if the symmetric matrix $K + K^T$ is definite negative which is equivalent to key matrix K is definite negative (Sutton 1998, appendix). (Recall that f is monotone operator if $\langle f(x) - f(y), x - y\rangle \geq 0 \forall x, y \in \mathcal{X}$)

3 Gradient Tree Backup

3.1 Derivation of algorithm

We derive here our new algorithm. Our objective function that we could minimize is the Mean Square Projected Error.

$$\begin{aligned}
\mathbf{MSPBE}(\theta) &= \frac{1}{2}||\Pi^\mu R(\Phi\theta) - \Phi\theta||_{D^\mu} \\
&= \frac{1}{2}||\Pi^\mu (R(\Phi\theta) - \Phi\theta)||_{D^\mu} \\
&= \frac{1}{2}[\Pi^\mu (R(\Phi\theta) - \Phi\theta)]^T D^\mu [\Pi^\mu (R(\Phi\theta) - \Phi\theta)] \\
&= \frac{1}{2}[\Phi^T D^\mu (R(\Phi\theta) - \Phi\theta)]^T (\Phi^T D^\mu \Phi)^{-1} \Phi^T D^\mu [\Phi(\Phi^T D^\mu \Phi)^{-1} \Phi^T D^\mu (R(\Phi\theta) - \Phi\theta)] \\
&= \frac{1}{2}[\Phi^T D^\mu (R(\Phi\theta) - \Phi\theta)]^T (\Phi^T D^\mu \Phi)^{-1} [\Phi^T D^\mu (R(\Phi\theta) - \Phi\theta)] \\
&= \frac{1}{2}||\Phi^T D^\mu (R(\Phi\theta) - \Phi\theta)||_{M^{-1}} \\
&= \frac{1}{2}||\Phi^T D^\mu [(I - \lambda\gamma P^{\mu\pi})^{-1}(T^\pi - \lambda\gamma P^{\mu\pi})\Phi\theta - \Phi\theta]||_{M^{-1}} \\
&= \frac{1}{2}||\Phi^T D^\mu (I - \lambda\gamma P^{\mu\pi})^{-1}(\gamma P^\pi - I)\Phi\theta + \Phi^T D^\mu (I - \lambda\gamma P^{\mu\pi})^{-1} r]||_{M^{-1}} \\
&= \frac{1}{2}||A\theta + b||_{M^{-1}}
\end{aligned}$$

Where $M = \Phi^T D^\mu \Phi = \mathbb{E}_\mu[\Phi\Phi^T]$, A and b are defined in the previous section.
We could derive our updates from computing gradients of the above expression as it is done in [4], but then we will obtain a gradient that is a product of two expectations. So, this double sampling makes not straightforward to obtain an unbiased estimator of the gradient. [4] used two times scale stochastic approximations. The main drawback of this method is that the derived algorithm is not true stochastic gradient methods with respect to their original objective.
Instead, [5] suggested to solve to the problem with a principled way by converting the original minimizing problem into primal-dual saddle point.
Here, we choose to follow the approach suggested by [5].
Let's recall the definition of the convex conjugate of a real-valued function f:

$$f^*(y) = \sup_{x \in X}(\langle y, x\rangle - f(x))$$

If f is convex, we have $f^{**} = f$
If $f(x) = \frac{1}{2}||x||_{M^{-1}}$, $f^*(x) = \frac{1}{2}||x||_M$. Note that by going to the convex conjugate, we don't need to invert the matrix M.
Let's go back to original minimization problem:

$$\begin{aligned}
\min_\theta \mathbf{MSPBE}(\theta) &\Leftrightarrow \min_\theta \frac{1}{2}||A\theta + b||^2_{M^{-1}} \\
&\Leftrightarrow \min_\theta \max_\omega(\langle A\theta + b, \omega\rangle - \frac{1}{2}||\omega||^2_M)
\end{aligned}$$

We apply now the gradient updates for saddle-point problem (ascent in ω and descent in θ)

$$\begin{aligned}
\omega_{k+1} &= \omega_k + \alpha_k(A\theta_k + b - M\omega_k) \\
\theta_{k+1} &= \theta_k - \alpha_k(A^T\omega_k)
\end{aligned}$$

As the quantities A, b and M are defined by expectation, we could derive stochastic updates of the former gradient updates by drawing samples for A, b and providing then an unbiased estimates of gradients.

As we are considering multi-step backups, the forward view is not efficient, so, we have to derive an eligibility traces for our estimates. Let's e the eligibility traces vector having the same number of components as θ. Then, our estimates becomes:

$$\begin{aligned}
e_k &= \lambda\gamma\pi(x_k, a_k)e_{k-1} + \phi(x_k, a_k) \\
\hat{A}_k &= e_k(\gamma\mathbb{E}_\pi[\phi(x_{k+1}, .)] - \phi(x_k, a_k)])^T \\
\hat{b}_k &= r(x_k, a_k)e_k \\
\hat{M}_k &= \Phi(x_k, a_k)\Phi(x_k, a_k)^T
\end{aligned}$$

And the parameters updates becomes:

$$\begin{aligned}
\omega_k &= \omega_{k-1} + \alpha_k(\hat{A}_k\theta_{k-1} + \hat{b}_k - \hat{M}_k\omega_{k-1}) \\
&= \omega_{k-1} + \alpha_k[\delta_k e_k - w_k^T\phi(x_k, a_k)\phi(x_k, a_k)] \\
\theta_k &= \theta_{k-1} - \alpha_k(\hat{A}^T\omega_{k-1}) \\
&= \theta_{k-1} - \alpha_k(w_{k-1}^T e_k(\gamma\mathbb{E}[\phi(x_{k+1}, .)] - \phi(x_k, a_k)]))
\end{aligned}$$

Let's show that $\mathbb{E}_\mu[\hat{A}_k] = A$. Let's Δ_t denotes $[\gamma\mathbb{E}_\pi\phi(x_{t+1}, .)^T - \phi(x_t, a_t)^T]$

$$\begin{aligned}
A &= \mathbb{E}_\mu[\sum_{t=0}^{\infty}(\lambda\gamma)^t(\prod_{i=1}^{t}\pi_i)\phi(x_0, a_0)\Delta_t] \\
&= \mathbb{E}_\mu[\sum_{t=k}^{\infty}(\lambda\gamma)^{t-k}(\prod_{i=k+1}^{t}\pi_i)\phi(x_k, a_k)\Delta_t] \\
&= \mathbb{E}_\mu[\phi(x_k a_k)\Delta_k + \sum_{t=k+1}^{\infty}(\lambda\gamma)^{t-k}(\prod_{i=k+1}^{t}\pi_i)\phi(x_k, a_k)\Delta_t] \\
&= \mathbb{E}_\mu[\phi(x_k a_k)\Delta_k + \sum_{t=k}^{\infty}(\lambda\gamma)^{t-k+1}(\prod_{i=k+1}^{t+1}\pi_i)\phi(x_k, a_k)\Delta_{t+1}]] \\
&= \mathbb{E}_\mu[\phi(x_k a_k)\Delta_k + \lambda\gamma\pi(x_{k+1}, a_{k+1})\phi(x_k a_k)\Delta_{k+1} + \sum_{t=k+1}^{\infty}(\lambda\gamma)^{t-k+1}(\prod_{i=k+1}^{t}\pi_i)\phi(x_k, a_k)\Delta_{t+1}]] \\
&= \mathbb{E}_\mu[\phi(x_k a_k)\Delta_k + \lambda\gamma\pi(x_k, a_k)\phi(x_{k-1}a_{k-1})\Delta_k + \sum_{t=k+1}^{\infty}(\lambda\gamma)^{t-k+1}(\prod_{i=k+1}^{t}\pi_i)\phi(x_k, a_k)\Delta_{t+1}]] \\
&= \mathbb{E}_\mu[\Delta_k(\phi(x_k a_k) + \lambda\gamma\pi(x_k, a_k)\phi(x_{k-1}a_{k-1}) + (\lambda\gamma)^2\pi(x_k, a_k)\pi(x_{k-1}, a_{k-1})\phi(x_{k-2}a_{k-2}) + ...)] \\
&= \mathbb{E}_\mu[\Delta_k e_k] = \mathbb{E}_\mu[\hat{A}]
\end{aligned}$$

we have used in the 6th line the fact that $\mathbb{E}_\mu[\pi(x_{k+1}, a_{k+1})\phi(x_k a_k)] = \mathbb{E}_\mu[\pi(x_k, a_k)\phi(x_{k-1}a_{k-1})]$ because the expectation is over the stationary distribution.

Similarly, we could show that $\mathbb{E}_\mu[\hat{b}_k] = b$. And, we have obviously that $\mathbb{E}_\mu[\hat{M}_k] = M$. Then our updates corresponds to true stochastic gradient descent of the primal-dual saddle point problem. So, that we could provide convergence rate analysis using this framework instead of limiting ourself to asymptotic convergence when we use two-time scale stochastic approximation.

Algorithm 1 Gradient Tree-backup with eligibility traces

```
procedure (target policy $\pi$, behavior policy $\mu$)
    Initialize $\theta_0$ and $\omega_0$
    set $e_0 = 0$
    for k = 1 ... do
        Observe $x_k, a_k, r_k, x_{k+1}$ according to $\mu$
        Update traces
        $e_k = \lambda\gamma\pi(x_k, a_k)e_{k-1} + \phi(x_k, a_k)$
        Update parameters
        $\delta_k = r_t + \gamma\mathbb{E}_\pi[\theta_{k-1}^T\phi(x_{k+1}, .)] - \theta_{k-1}^T\phi(x_k, a_k)$
        $\omega_k = \omega_{k-1} + \alpha_k[\delta_k e_k - w_{k-1}^T\phi(x_k, a_k)\phi(x_k, a_k)]$
        $\theta_k = \theta_{k-1} - \alpha_k(w_{k-1}^T e_k(\gamma\mathbb{E}_\pi[\phi(x_{k+1}, .)] - \phi(x_k, a_k)]))$
```

3.2 Convergence analysis

To show convergence rate, we will proceed similarly as in [5]. We will use the following proposition which is proven in the section 3 of [6] but with some slightly modified conditions for simplification

Proposition:
We consider a differentiable convex-concave function f defined on $\varTheta \times \varOmega$, where $\varTheta$ and $\varOmega$ are two bounded closed convex sets whose diameters are upper bounded by $D > 0$.
we assume that we have an increasing sequence of σ-fields $\{F_t\}$ such that, θ_0, ω_0 are F_0 measurable and such that for $t \geq 1$,

$$\theta_t = \Pi_\varTheta(\theta_{t-1} - \gamma_t g_t^\theta)) \tag{1}$$

$$\omega_t = \Pi_Y(\omega_{t-1} + \gamma_t g_t^\omega)) \tag{2}$$

$$\text{output} : \bar{\theta}_T = \frac{\sum_{t=0}^T \gamma_t\theta_t}{\sum_{t=0}^T \gamma_t}, \bar{\omega}_T = \frac{\sum_{t=0}^T \gamma_t\omega_t}{\sum_{t=0}^T \gamma_t}$$

where

- $\Pi_\varTheta$ and $\Pi_\varOmega$ are orthogonal projection respectively on $\varTheta$ and $\varOmega$.
- $\mathbb{E}(g_t^\theta|F_{t-1}) = \nabla_x f(\theta_{t-1}, \omega_{t-1})$ and $\mathbb{E}(g_t^\omega|F_{t-1}) = \nabla_y f(\theta_{t-1}, \omega_{t-1})$
- Its exists $G \geq 0$ such that, $\mathbb{E}(||g_t^\theta||^2) \leq G^2$ and $\mathbb{E}(||g_t^\omega||^2) \leq G^2$

(θ^*, ω^*) a saddle point of f i.e $\forall(\theta', \omega') \in \varTheta \times \varOmega, f(\theta^*, \omega') \leq f(\theta^*, \omega^*) \leq f(\theta', \omega^*)$ Then, if $\gamma_t = O(\frac{1}{\sqrt{T}})$ then, $(\bar{\theta}_T, \bar{\omega}_T)$ convergences to (θ^*, ω^*) with $O(1/\sqrt{t})$ rate.
In our case,

- $f(\theta, \omega) = \langle A\theta + b, \omega\rangle - \frac{1}{2}||\omega||_M^2$.
- $g_t^\omega = \hat{A}_t\theta + \hat{b}_t - \hat{M}_t\omega$
- $g_t^\theta = \hat{A}_t^T\omega$

In order to be able to apply the statement of the proposition above we need to some assumptions.:

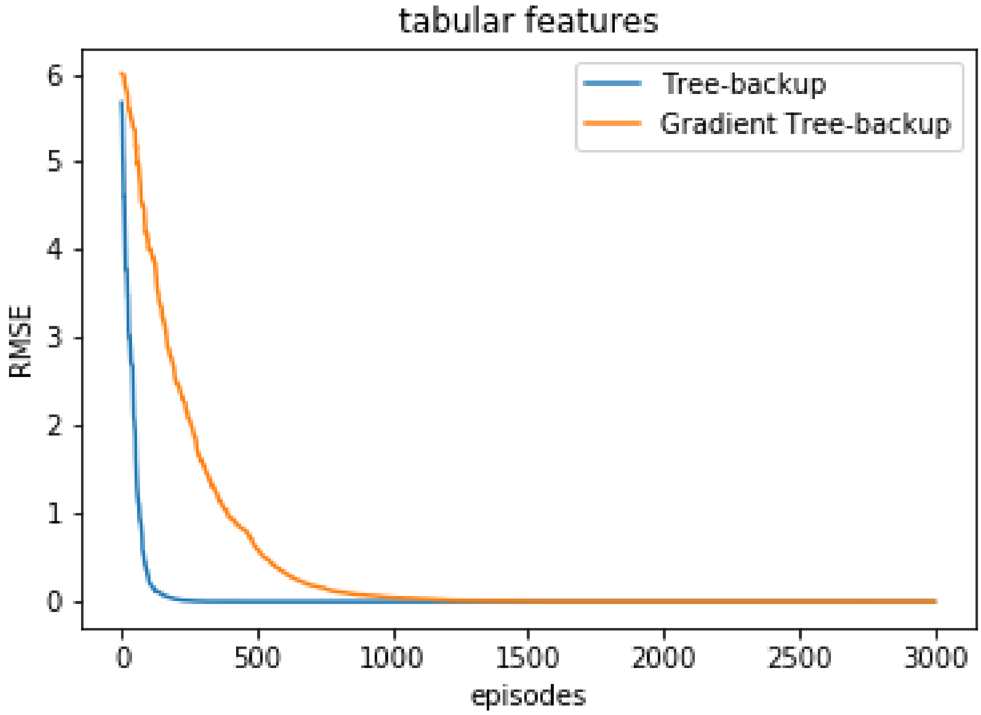
- **Assumption 1**: The feasible sets Θ and Ω are bounded closed convex sets
- **Assumption 2**: The features vectors and the rewards are uniformly bounded.
- **Assumption 3**: The matrices A and M are non-singular

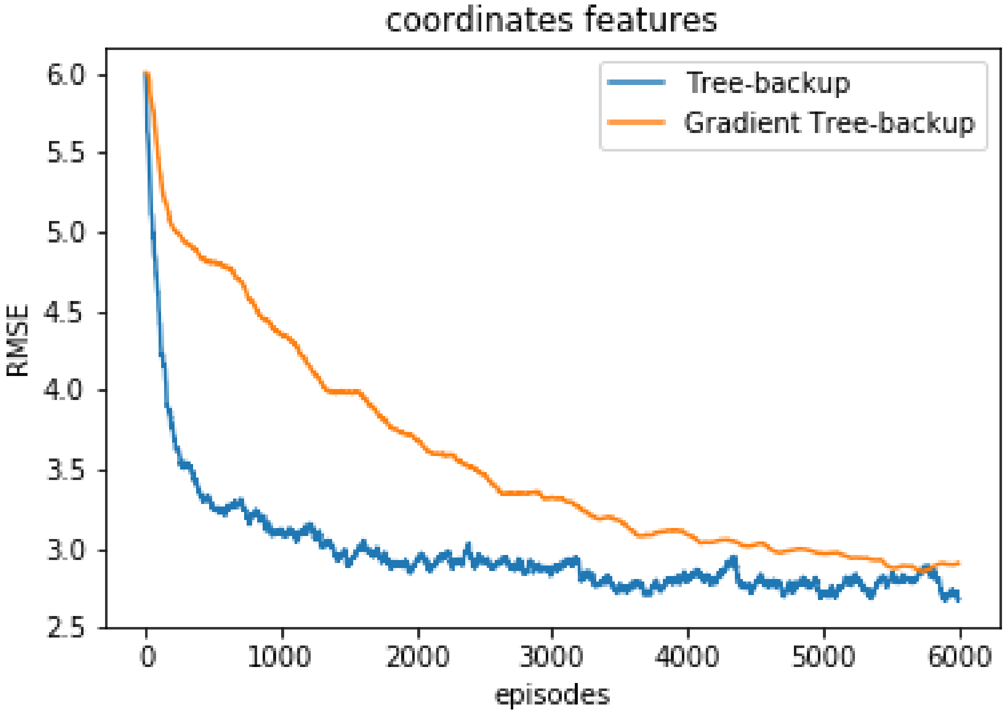
Using the assumptions 1 and 3, we could show the estimated gradients have bounded variance than we could reach $O(1/\sqrt{t})$ convergence rate.
The assumptions 2 allows us to state that the our primal-dual problem has unique saddle point which is $(\theta^*, \omega^*) = (-A^{-1}b, 0)$.

3.3 Preliminary Experimental results

Here we present some preliminary experimental results. We consider a simple gridworld with two terminal states as test environment. we consider tow setting:

- Tabular features: actions and states are one hot encoded. In this case: running Tree backup with LFA is equivalent to running tabular tree-backup. So, it is expected that Tree backup would be faster that Gradient based Tree backup.
- Coordinates features: 4-dimensional features. Each state is represented by its coordinates in the grid and the action is represented by its direction. for example, the (state, action) = (5, up) is represented by (0, 1, 0, -1).





4 Future work

Our preliminary experiments shows that Gradient Tree-backup seems to converge to the desired solution but it converges slowly comparing to standard tree-backup when it converges. So more experiments have to be done in order to conclude.
As we use saddle point formulation to derive our algorithm, accelerated methods are available to improve the convergence rate, we cite particularly the extra-gradient method. Or, we could add l2 regularization term to our objective function so our function would be strongly concave-convex and we could then use proximal method (see [7]).

[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

[6] A. NEMIROVSKI, A. JUDITSKY, G. LAN, and A. SHAPIRO (2009) *Robust Stochastic approximation approach to stochastic programming.*Society for Industrial and Applied Mathematics

[7] P Balamurugan and Francis Bach (2016) *Stochastic Variance Reduction Methods for Saddle-Point Problems.* Neural Information Processing Systems