Analysis of Tree backup algorithm

Tabular case, Linear Function approximation and gradient Tree backup

Ahmed Touati

Reinforcement Learning class project

Motivation and contributions

- Tree backup is an off-policy multi-step temporal difference learning proposed by Doina and al (2000).
- Tree-backup corrects the discrepancy between target/behavior policy by scaling returns by target policy probabilities.
- No importance sampling ratio.
- Good empirical performance.

My contribution is mainly theoretical understanding:

- Tabular Case: new convergence proof than the proof showed in the original article
- Linear Function Approximation: divergence issues understanding.
- Derivation of new algorithm: Gradient Tree backup.
- Derivation Eligibility traces of the new algorithm.
- Convergence rate proof.

Tabular case

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

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

  where $\pi_i = \pi(x_i, a_i)$
- The λ return extension considers exponentially weighted sums of n-steps returns:

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

- We could show that:

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

  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:

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

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

Convergence Tabular case

- Convergence result could be obtained by applying general results of Robbins-Monro stochastic approximation methods for solving $Q = RQ$. where R is the tree-backup operator defined by:

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

where:

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

- $R = I + (I - \lambda\gamma P^{\mu\pi})^{-1}(T^\pi - I) = (I - \lambda\gamma P^{\mu\pi})^{-1}(T^\pi - \lambda\gamma P^{\mu\pi})$
- the mapping R is γ maximum norm contraction.

Linear Function Approximation

- let $Q(x,a) = \theta^T \phi(x,a)$. The tree-backup with VFA is then:

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

where $\delta_t^{\pi} = r_t + \gamma \mathbb{E}_{\pi}\theta^T\phi(x_{t+1}, .) - \theta^T\phi(x_t, a_t)$

- let's rearrange the update: $\theta_{k+1} = \theta_k + \alpha_k(A_k\theta_k + b_k)$ where

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

$$b_k = \sum_{t=0}^{\infty}(\lambda\gamma)^t(\prod_{i=1}^{t}\pi_i)r_t\phi(x,a)$$

- Make expectation over trajectories generated by μ

$$A = \mathbb{E}_{\mu}[A_k] = \Phi^T D^{\mu}(I - \lambda\gamma P^{\mu\pi})^{-1}(\gamma P^{\pi} - I)\Phi$$

$$b = \mathbb{E}_{\mu}[b_k] = \Phi^T D^{\mu}(I - \lambda\gamma P^{\mu\pi})^{-1} r$$

- Unfortunately, the matrix A is not necessarily definite negative. (In particular, in the case of $\lambda = 0$, $A = \Phi^T D^{\mu}(\gamma P^{\pi} - I)\Phi$ which is the matrix we obtain for off-policy temporal difference learning TD(0))

Gradient Tree backup: motivation

- When FVA algorithm converges, it converges to $\theta^* = -A^{-1}b$. We could 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:

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

- we could prove that $\text{MSPBE}(\theta) = \frac{1}{2}||A\theta + b||^2_{M^{-1}}$ where $||.||_{M^{-1}}$ is the Euclidian norm weighted by the inverse of the matrix $M = \Phi^T D^\mu \Phi = \mathbb{E}_\mu[\Phi\Phi^T]$

Gradient Tree backup: Derivation

- We could derive our updates from computing gradients of the above expression, but then we will obtain a gradient that is a product of two expectations $\Rightarrow$ double sampling $\Rightarrow$ not true stochastic gradient methods !!
- Instead, we cast our problem into saddle-point problem using Fenchel duality.
- the convex conjugate of a real-valued function f:

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

  If f is convex, we have $f^{**} = f$
  If $f(x) = \frac{1}{2}||x||^2_{M^{-1}}$, $f^*(x) = \frac{1}{2}||x||^2_M$
-
  $$\begin{aligned} \min_\theta \mathrm{MSPBE}(\theta) &\Leftrightarrow \min_\theta \frac{1}{2}||A\theta + b||^2_{M^{-1}} \\ &\Leftrightarrow \min_\theta \max_\omega(< A\theta + b, \omega > -\frac{1}{2}||\omega||^2_M) \end{aligned}$$

Gradient tree backup

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

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

- Let's e the eligibility traces vector having the same number of components as θ. Then, our estimates becomes:

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

Gradient Tree backup algorithm

Algorithm 1 Gradient Tree-backup with eligibility traces

procedure (target policy π, behavior policy μ)
  Initialize θ_0 and ω_0
  set $e_0 = 0$
  for k = 1 ... **do**
    Observe x_k, a_k, r_k, x_{k+1} according to μ
    Update traces
    $e_k = \lambda\gamma\pi(x_k, a_k)e_{k-1} + \phi(x_k, a_k)$
    Update parameters
    $\delta_k = r_t + \gamma\mathbb{E}_\pi[\theta_{k-1}^T\phi(x_{k+1}, .)] - \theta_{k-1}^T\phi(x_k, a_k)$
    $\omega_k = \omega_{k-1} + \alpha_k[\delta_k e_k - w_{k-1}^T\phi(x_k, a_k)\phi(x_k, a_k)]$
    $\theta_k = \theta_{k-1} - \alpha_k(w_{k-1}^T e_k(\gamma\mathbb{E}_\pi[\phi(x_{k+1}, .)] - \phi(x_k, a_k)]))$

Convergence rate analysis

Proposition: We consider a differentiable convex-concave function f defined on $X \times Y$, where X and Y 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, x_0, y_0 are F_0 measurable and such that for $t \geq 1$,

$$x_t = \Pi_X(x_{t-1} - \gamma_t g_t^x)) \tag{1}$$

$$y_t = \Pi_Y(y_{t-1} + \gamma_t g_t^y)) \tag{2}$$

$$\text{output} : \bar{x}_T = \frac{\sum_{t=0}^T \gamma_t x_t}{\sum_{t=0}^T \gamma_t}, \bar{y}_T = \frac{\sum_{t=0}^T \gamma_t y_t}{\sum_{t=0}^T \gamma_t}$$

where

- Π_X and Π_Y are orthogonal projection respectively on X and Y.
- $\mathbb{E}(g_t^x | F_{t-1}) = \nabla_x f(x_{t-1}, y_{t-1})$ and $\mathbb{E}(g_t^y | F_{t-1}) = \nabla_y f(x_{t-1}, y_{t-1})$
- Its exists $G \geq 0$ such that, $\mathbb{E}(||g_t^x||^2) \leq G^2$ and $\mathbb{E}(||g_t^y||^2) \leq G^2$

(x^*, y^*) a saddle point of f i.e $\forall (x', y') \in X \times Y, f(x^*, y') \leq f(x^*, y^*) \leq f(x', y^*)$
Then, $(\bar{x}_T, \bar{y}_T)$ convergences to (x^*, y^*) with $O(1/\sqrt{t})$ rate.

Questions?