Safe and efficient off-policy reinforcement learning

Rehas Sachdeva, Sreeja Kamishetty, Sairam Kolla

International Institute of Information Technology, Hyderabad

25/11/2017

Overview of the Presentation

- Recap.
- Proofs of γ - contraction property of RQ operator, convergence of Retrace(λ) algorithm
- Implementation of Retrace λ algorithm on famous Cart Pole Balancing Problem.
- Results.

Recap

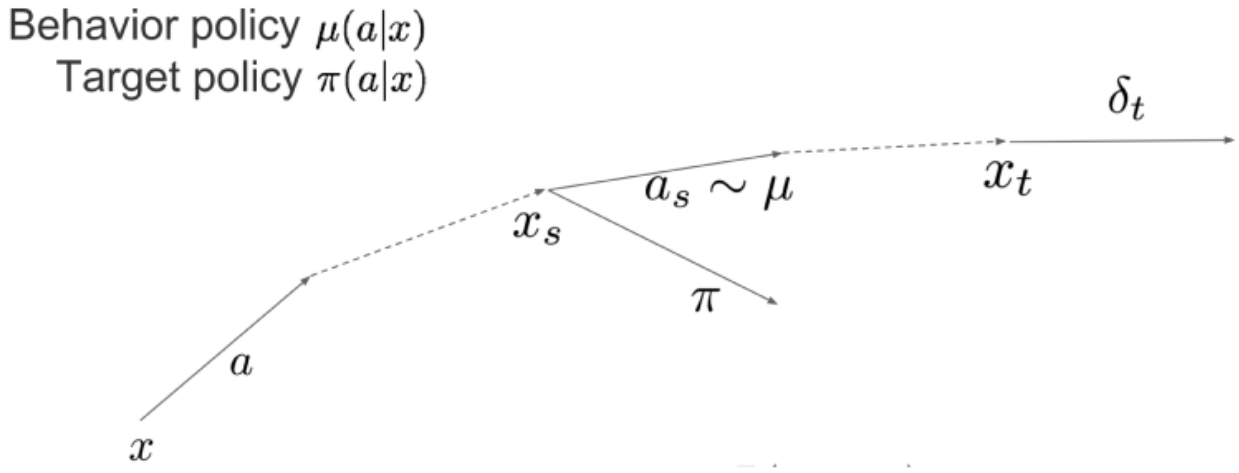
- In our first evaluation we have discussed about different Off-policy algorithms, introduced Retrace(λ) and discussed why Retrace(λ) is better than others.

General Off-Policy return based algorithm can be represented by:

$$\Delta Q(x,a) = \sum_{t \geq 0} \gamma^t \Big(\prod_{1 \leq s \leq t} c_s \Big) \Big(\underbrace{r_t + \gamma \mathbb{E}_\pi Q(x_{t+1}, \cdot) - Q(x_t, a_t)}_{\delta_t} \Big)$$

Algorithm:	Trace coefficient:	Problem:
IS	$c_s = \frac{\pi(a_s \vert x_s)}{\mu(a_s \vert x_s)}$	high variance
$Q^\pi(\lambda)$	$c_s = \lambda$	not safe (off-policy)
$TB(\lambda)$	$c_s = \lambda \pi(a_s \vert x_s)$	not efficient (on-policy)

Retrace(λ)



From the presentation by the authors: `https://ewrl.files.wordpress.com/2016/12/munos.pdf`

- $\Delta Q(x,a) = \gamma^t(\prod_{1\leq s\leq t} \lambda \min(1, \frac{\pi(a_s|x_s)}{\mu(a_s|x_s)}))\delta_t$

\+ Variance is bounded

\+ Convergent for any π and μ

\+ Uses full returns when on-policy

– Doesn't work if μ is unknown or non-Markov ($\leftrightarrow$ Tree-Backup)

Properties of Retrace(λ) and Comparisons of different algorithms

- it has low variance.
- it safely uses samples collected from any behavior policy, whatever the degree of "off-policyness".
- it is efficient as it makes the best use of samples collected from near on-policy behavior policies.

Algorithm:	Trace coefficient:	Problem:
IS	$c_s = \frac{\pi(a_s\|x_s)}{\mu(a_s\|x_s)}$	high variance
$Q^\pi(\lambda)$	$c_s = \lambda$	not safe (off-policy)
$TB(\lambda)$	$c_s = \lambda\pi(a_s\|x_s)$	not efficient (on-policy)
$Retrace(\lambda)$	$c_s = \lambda \min\left(1, \frac{\pi(a_s\|x_s)}{\mu(a_s\|x_s)}\right)$	none!

Notations

These are basic notations used in the proof.

- state $x \in \mathcal{X}$
- action $a \in \mathcal{A}$
- discount factor $\gamma \in [0, 1]$
- immediate reward $r \in \mathbb{R}$
- policies $\pi, \mu : \mathcal{X} \times \mathcal{A} \mapsto [0, 1]$
- value function
  $Q^{\pi}(x, a) := \mathbb{E}_{\pi}[r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots | x_0 = x, a_0 = a]$
- optimal value function $Q^* := \max_{\pi} Q^{\pi}$
- $\mathbb{E}_{\pi} Q(x, \cdot) := \sum_a \pi(a|x) Q(x, a)$

Theorem 1

Statement: The operator R defined by the general return-based off-policy algorithms has unique fixed point Q^π. Furthermore, if for each $a_s \in A$ and each history F_s we have $c_s = c_s(a_s, F_s) \in [0, \frac{\pi(a_s|x_s}{\mu(a_s|x_s)}]$, then for any Q-function Q.

$$||RQ - Q^\pi|| \leq \gamma ||Q - Q^\pi||$$

Theorem 1 Proof

$$\mathcal{R}Q(x,a) = Q(x,a) + \mathbb{E}_\mu\Big[\sum_{t\geq 0}\gamma^t(c_1\dots c_t)\big(r_t + \gamma\mathbb{E}_\pi Q(x_{t+1},\cdot) - Q(x_t,a_t)\big)\Big]$$
$$= \mathbb{E}_\mu\Big[\sum_{t\geq 0}\gamma^t(c_1\dots c_t)\big(r_t + \gamma\big[\mathbb{E}_\pi Q(x_{t+1},\cdot) - c_{t+1}Q(x_{t+1},a_{t+1})\big]\big)\Big]$$

Thus

$$(\mathcal{R}Q_1 - \mathcal{R}Q_2)(x,a) = \mathbb{E}_\mu\Big[\sum_{t\geq 0}\gamma^{t+1}(c_1\dots c_t)\Big(\mathbb{E}_\pi(Q_1-Q_2)(x_{t+1},\cdot) - c_{t+1}(Q_1-Q_2)(x_{t+1},a_{t+1})\Big)\Big]$$
$$= \mathbb{E}_\mu\Big[\sum_{t\geq 0}\gamma^{t+1}(c_1\dots c_t)\sum_a\big(\pi(a|x_{t+1}) - \mu(a|x_{t+1})c_{t+1}(a)\big)(Q_1-Q_2)(x_{t+1},a)\Big]$$

- which is a linear combination weighted by non-negative coefficients.

Theorem 1 Contd...

$$
\begin{aligned}
\text{Sum of the coeff.} &= \mathbb{E}_\mu\Big[\sum_{t\geq 0}\gamma^{t+1}(c_1\dots c_t)\sum_a\big(\pi(a|x_{t+1})-\mu(a|x_{t+1})c_{t+1}(a)\big)\Big] \\
&= \mathbb{E}_\mu\Big[\sum_{t\geq 0}\gamma^{t+1}(c_1\dots c_t)(1-c_{t+1})\Big] \\
&= \gamma-(1-\gamma)\mathbb{E}_\mu\Big[\sum_{t\geq 1}\gamma^{t}(c_1\dots c_t)\Big] \\
&\in [0,\gamma]
\end{aligned}
$$

Thus $\|\mathcal{R}Q_1-\mathcal{R}Q_2\|_\infty \leq \gamma\|Q_1-Q_2\|_\infty$

- Therefore, we have proved that R is gamma-contraction mapping.

Theorem 2

Statement: Consider an arbitrary sequence of behavior policies (μ_k) (which may depend on (Q, k)) and a sequence of target policies (π_k) that are increasingly greedy w.r.t. the sequence (Q, k) and $Q_{k+1} = \mathcal{R}_k Q_k$ where the return operator R_k is defined by general off-policy algorithm for π_k and μ_k and a Markovian $c_s = c(a_s, x_s) \in [0, \frac{\pi_k(a_s|x_s)}{\mu(a_s|x_s)}]$. Assume the target policies π_k are ε_k-away from the greedy policies w.r.t. Q_k , in the sense that $\mathcal{T}^{\pi_k} Q_k \geq \mathcal{T} Q_k - \varepsilon_k ||Q_k|| e$, where e is the vector with 1-components. Further suppose that $\mathcal{T}^{\pi_0} Q_0 \geq Q_0$. Then for any k $\geq$ 0.

$$||Q_{k+1} - Q^*|| \leq \gamma ||Q_k - Q^*|| + \varepsilon_k ||Q_k||$$

In consequence, if $\varepsilon_k \to 0$, then $Q_k \to Q^*$

Theorem 2 Proof

Define the (sub)-probability transition operator,
$(P^{c\mu}Q)(x,a) := \sum_{x'}\sum_{a'} p(x'|x,a)\mu(a'|x')c(a',x')Q(x',a')$

Since,

$$\mathcal{R}Q(x,a) := Q(x,a) + \mathbb{E}[\sum_{t\geq 0}\gamma^t(\prod_{s=1}^{t} c_s)(r_t + \gamma\mathbb{E}_\pi Q(x_{t+1},.) - Q(x_t,a_t))]$$

The Retrace(λ) operator then writes,

$$\begin{aligned}\mathcal{R}_k Q &= Q + \sum_{t\geq 0}\gamma^t(P^{c\mu_k})^t(\mathcal{T}^{\pi_k}Q - Q) \\ &= Q + (\mathrm{I} - \gamma P^{c\mu_k})^{-1}(\mathcal{T}^{\pi_k}Q - Q)\end{aligned} \tag{1}$$

Proof contd..

Upper bound on $Q_{k+1} - Q*$ since $Q_{k+1} = \mathcal{R}_k Q_k$, we have

$$\begin{aligned}
Q_{k+1} - Q^* &= Q_k - Q^* + (\mathrm{I} - \gamma P^{c\mu_k})^{-1}[\mathcal{T}^{\pi_k} Q_k - Q_k] \\
&= (\mathrm{I} - \gamma P^{c\mu_k})^{-1}[\mathcal{T}^{\pi_k} Q_k - Q_k + (\mathrm{I} - \gamma P^{c\mu_k})(Q_k - Q)] \\
&= (\mathrm{I} - \gamma P^{c\mu_k})^{-1}[\mathcal{T}^{\pi_k} Q_k - Q^* - \gamma P^{c\mu_k}(Q_k - Q^*)] \\
&= (\mathrm{I} - \gamma P^{c\mu_k})^{-1}[\mathcal{T}^{\pi_k} Q_k - \mathcal{T} Q^* - \gamma P^{c\mu_k}(Q_k - Q^*)] \\
&\leq (\mathrm{I} - \gamma P^{c\mu_k})^{-1}[\gamma P^{\pi_k}(Q_k - Q*) - \gamma P^{c\mu_k}(Q_k - Q*)] \\
&= \gamma(\mathrm{I} - \gamma P^{c\mu_k})^{-1}[P^{\pi_k} - P^{c\mu_k}](Q_k - Q*) \\
&= A_k(Q_k - Q*)
\end{aligned} \tag{2}$$

where $A_k := \gamma(\mathrm{I} - \gamma P^{c\mu_k})^{-1}[P^{\pi_k} - P^{c\mu_k}]$. Also A_k has only non zero elements and they sum up to at most γ. Hence the LHS is upper bounded as,

$$Q_{k+1} - Q* \leq \gamma ||Q_k - Q * ||e \tag{3}$$

Proof contd..

Lower bound on $Q_{k+1} - Q*$ we have,

$$\begin{aligned} Q_{k+1} &= Q_k + (\mathrm{I} - \gamma P^{c\mu_k})^{-1}[\mathcal{T}^{\pi_k} Q_k - Q_k] \\ &= Q_k + \sum_{i \geq 0} \gamma^i (P^{c\mu_k})^i (\mathcal{T}^{\pi_k} Q_k - Q_k) \\ &= \mathcal{T}^{\pi_k} Q_k + \sum_{i \geq 1} \gamma^i (P^{c\mu_k})^i (\mathcal{T}^{\pi_k} Q_k - Q_k) \\ &= \mathcal{T}^{\pi_k} Q_k + \gamma P^{c\mu_k} (\mathrm{I} - \gamma P^{c\mu_k})^{-1} (\mathcal{T}^{\pi_k} Q_k - Q_k) \end{aligned} \tag{4}$$

From the definition of ϵ_k we have,

$$\begin{aligned} \mathcal{T}^{\pi_k} Q_k &\geq \mathcal{T} Q_k - \varepsilon_k ||Q_k|| \\ &\geq \mathcal{T}^{\pi^*} Q_k - \varepsilon_k ||Q_k|| \end{aligned} \tag{5}$$

Proof Contd..

Thus

$$\begin{aligned} Q_{k+1} - Q* &= Q_{k+1} - \mathcal{T}^{\pi_k} Q_k + \mathcal{T}^{\pi_k} Q_k - \mathcal{T}^{\pi^*} Q_k + \mathcal{T}^{\pi^*} Q_k - \mathcal{T}^{\pi^*} Q^* \\ &\geq Q_{k+1} - \mathcal{T}^{\pi_k} Q_k + \gamma P^{\pi^*}(Q_k - Q^*) - \varepsilon_k ||Q_k|| e \end{aligned} \tag{6}$$

Using results from (4) and (6) we get,

$$\begin{aligned} Q_{k+1} - Q* \geq \gamma P^{c\mu_k} (\mathrm{I} - \gamma P^{c\mu_k})^{-1} (\mathcal{T}^{\pi_k} Q_k - Q_k) \\ + \gamma P^{\pi^*}(Q_k - Q^*) - \varepsilon_k ||Q_k|| e \end{aligned} \tag{7}$$

Proof contd..

Lower bound on $\mathcal{T}^{\pi_k} Q_k - Q_k$ By hypothesis, (π_k) is increasingly greedy w.r.t (Q_k), thus

$$
\begin{aligned}
\mathcal{T}^{\pi_{k+1}} Q_{k+1} - Q_{k+1} &\geq \mathcal{T}^{\pi_k} Q_{k+1} - Q_{k+1} \\
&= \mathcal{T}^{\pi_k} \mathcal{R}_k Q_k - \mathcal{R}_k Q_k \\
&= r + (\gamma P^{\pi_k} - \mathrm{I}) \mathcal{R}_k Q_k \\
&= r + (\gamma P^{\pi_k} - \mathrm{I})[Q_k + (\mathrm{I} - \gamma P^{c\mu_k})^{-1} (\mathcal{T}^{\pi_k} Q_k - Q_k)] \\
&= \mathcal{T}^{\pi_k} Q_k - Q_k \\
&\quad + (\gamma P^{\pi_k} - \mathrm{I})(\mathrm{I} - \gamma P^{c\mu_k})^{-1} (\mathcal{T}^{\pi_k} Q_k - Q_k) \\
&= \gamma [P^{\pi_k} - P^{c\mu_k}](\mathrm{I} - \gamma P^{c\mu_k})^{-1} (\mathcal{T}^{\pi_k} Q_k - Q_k) \\
&= B_k (\mathcal{T}^{\pi_k} Q_k - Q_k)
\end{aligned}
\tag{8}
$$

where $B_k := \gamma [P^{\pi_k} - P^{c\mu_k}](\mathrm{I} - \gamma P^{c\mu_k})^{-1}$

Proof Contd..

Since B_k has non negative elements, the following inequality holds

$$\mathcal{T}^{\pi_k} Q_k - Q_k \geq B_{k-1} B_{k-2} ... B_0 (\mathcal{T}^{\pi_0} Q_0 - Q_0)$$

Since $\mathcal{T}^{\pi_0} Q_0 - Q_0 \geq 0$, equation (7) implies that,

$$Q_{k+1} - Q* \geq \gamma P^{\pi^*} (Q_k - Q^*) - \varepsilon_k ||Q_k|| e$$

Combining this with (3) we deduce,

$$||Q_{k+1} - Q*|| \leq \gamma ||Q_k - Q^*|| - \varepsilon_k ||Q_k||$$

Now assume that $\epsilon \Rightarrow 0$. We first deduce that Q_k is bounded. Indeed as soon as $\epsilon < (1-\gamma)/2$, we have

$$\begin{aligned} ||Q_{k+1}|| &\leq ||Q_k|| + \gamma ||Q_k - Q^*|| + \frac{1-\gamma}{2} ||Q_k|| \\ &\leq (1+\gamma) ||Q^*|| + \frac{1+\gamma}{2} ||Q_k|| \end{aligned}$$

Proof Contd..

Thus

$$\lim_{k \to \infty} ||Q_k||_\infty \leq \frac{1+\gamma}{1-(1+\gamma)/2} ||Q^*||$$

Since Q_k is bounded, we deduce that $\lim_{k \to \infty} ||Q_k||_\infty = Q^*$

Algorithm Implementation on Cart Pole Balancing Problem

- **Cart Pole Balancing Problem:** A pole is attached by an un-actuated joint to a cart, which moves along a frictionless track. The system is controlled by applying a force of $+1$ or -1 to the cart. The pendulum starts upright, and the goal is to prevent it from falling over.
- A reward of $+1$ is provided for every timestep that the pole remains upright. The episode ends when the pole is more than 15 degrees from vertical, or the cart moves more than 2.4 units from the center.

Results

- We define goal average episode length to be 195. We say that goal is reached when the average episode length of last 100 episodes becomes $\geq$ goal average episode length. Algorithms are compared on the basis of time to reach the goal.

	episodes	λ	γ	ϵ
Retrace(λ)	325	1	0.95	0.5
Tree-backup(λ)	432	1	1	0.5
Q learning	325	-	1	0.5