Multi-step RL: Unifying Algorithm

Kirill Bobyrev

November 6, 2017

Plan

Results

Introducing algorithm $Q(\sigma)$ with following properties

- Model-free
- Can be trained both off- and on-policy
- Uses *n*-step backup rule
- Generalizes a huge number of existing algorithms while subsuming them as special cases
- Performs better given a reasonable choice of hyperparameter σ

Monte Carlo methods

- Sample many episodes
- MC every-visit backup: $V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$
- Does not need environment model

TD methods

- Combines Monte Carlo and Dynamic Programming
- Does not need environment model
- Uses bootstrapping for updates
- Sample many steps instead of methods
- One-step TD backup: $V(S_t) \leftarrow V(S_t) + \alpha[R_t - \gamma V(S_t)]$

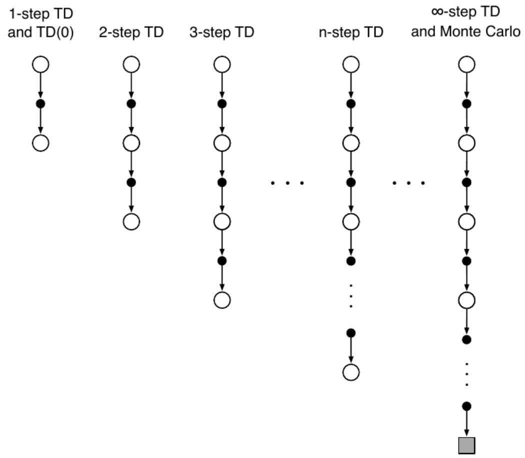
From 1-step to *n*-step

Define multi-step return for TD:
$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n V_{t+n-1}(S_{t+n})$
Using this multi-step return:

- Monte Carlo backup uses $G_{t:T}$
- One-step TD backup uses $G_{t:t+1}$

n-step TD: $V_{t+n}(S_t) \doteq V_{t+n-1}(S_t) + \alpha[G_{t:t+n} - V_{t+n-1}(S_t)]$
$Q(\sigma)$ is based on *n*-step Sarsa and *n*-step Tree Backup

Backups: From one-step TD to MC

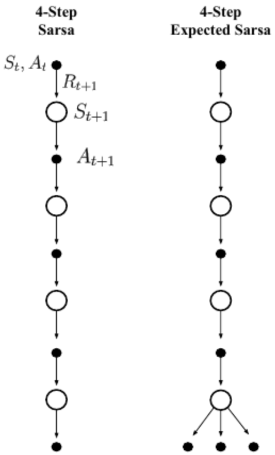


n-step Sarsa

$$\delta_t^S = R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)$$

$$\delta_t^{ES} \doteq R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q_t(S_{t+1}, a) - Q_{t-1}(S_t, A_t)$$

n-step Sarsa and Expected Sarsa backup mechanisms



Tree Backup

Tree Backup is the multi-step generalization of Expected Sarsa backup

$$G_{t:t+1} \doteq R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q_t(S_{t+1}, a) = \delta_t^{ES} + Q_{t-1}(S_{t+1}, a)$$

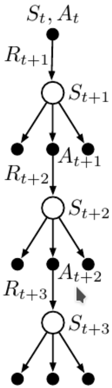
Hence *n*-step return of Tree Backup is a sum of TD errors:

$$G_{t:t+n} \doteq Q_{t-1}(S_t, A_t) + \sum_{k=t}^{\min t+n-1, T-1} \delta_k^{ES} \prod_{i=t+1}^{k} \gamma\pi(A_i|S_i)$$

Taking update rule from *n*-step Sarsa:

$$Q_{t+n}(S_t, A_t) \doteq Q_{t+n-1}(S_t, A_t) + \alpha[G_{t:t+n} - Q_{t+n-1}(S_t, A_t)]$$

Tree Backup backup mechanism



the 3-step
tree backup

Sarsa and Tree Backup off-policy learning

Sarsa | Tree Backup

Relation to other algorithms

Two families of multi-step algorithms:

- Algorithms that backup their actions and samples (Sarsa and Expected Sarsa)
- Algorithms that consider an expectation over all actions in their backup (Expected Sarsa and Tree Backup)

New parameter $\sigma \in [0, 1]$, which controls the degree of sampling at each step of the backup, unifies both families

Details

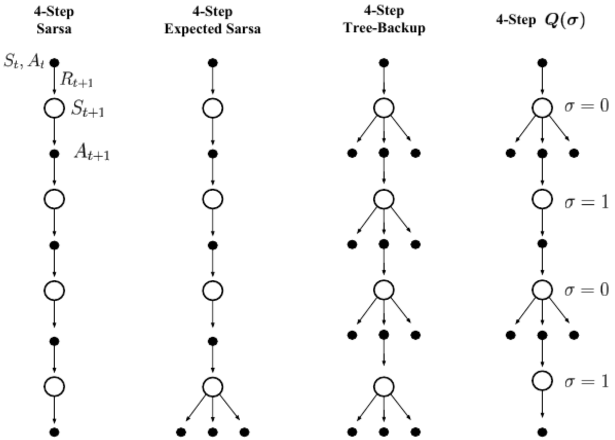
Error modification:

$$\begin{aligned}\delta_t^\sigma &= \sigma_{t+1}\delta_t^S + (1-\sigma_{t+1})\delta_t^{ES} \\ &= R_{t+1} + \gamma[\sigma_{t+1}Q_t(S_{t+1}, A_{t+1}) + (1-\sigma_{t+1})V_{t+1}] - Q_{t-1}(S_t, A_t)\end{aligned}$$

Resulting return:

$$G_t^{(n)} = Q_{t-1}(S_t, A_t) + \sum_{k=t}^{\min t+n-1, T-1} \delta_k^\sigma \prod_{i=t+1}^{k} \gamma[(1-\sigma_i)\pi(A_i|S) + \sigma_i]$$

Backup comparisons



$Q(\sigma)$ off-policy learning

$Q(\sigma)$ importance sampling for off-policy learning combines the off-policy learning ideas for base algorithms:

$$\rho_{t+1}^{t+n} = \prod_{k=t+1}^{\min t+n-1, T-1} (\sigma_k \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} + 1 - \sigma_k)$$

σ choosing strategies

- Constant $\sigma = C$
- Altering $\sigma(t) = 1, 0, 1, 0, \ldots = [t \mod 2 = 0]$
- Random $\sigma(t) \sim \mathcal{U}[0, 1]$
- Decreasing or increasing over t (between 0 and 1)

Algorithm description

Initialize $S_0 \neq S_T$; select A_0 according to $\pi(.|S_0)$
Store S_0, A_0 and $Q(S_0, A_0)$
for $t \leftarrow 0, T + n - 1$ **do**
    if $t < T$ **then**
        Take action A_t, get R, observe and store S_{t+1}
        if S_{t+1} is S_T **then**
            Store $\delta_t^\sigma = R - Q(S_t, A_t)$
        else
            Select and store A_{t+1} according to $\pi(.|S_{t+1}$
            Store $Q(S_{t+1}, A_{t+1})$, σ_{t+1}, $\pi(A_{t+1}|S_{t+1})$
            Calculate and store δ_t^σ
        end if
    end if
    if $t \geq n$ **then**
        Calculate and store $G_t^{(n)}$
        Perform backup: $Q(S_t, A_t) \leftarrow S(_t, A_t) + \alpha[G_t^{(n)} - Q(S_t, A_t)]$
    end if
end for

Stochastic Windy Gridworld Environment

o	o	o	o	o	o	o	o	o	o
o	o	o	o	o	o	o	o	o	o
o	o	o	o	o	o	o	o	o	o
S	o	o	o	o	o	o	G	o	o
o	o	o	o	o	o	o	o	o	o
o	o	o	o	o	o	o	o	o	o
o	o	o	o	o	o	o	o	o	o
0	0	0	1	1	1	2	2	1	0

- Tabular navigation environment, agent is moved by upward "wind" by x cells specified below each corresponding column at the end of its turn
- Environment gives reward of -1 after each step
- Agent returns to the closest valid state upon exiting the world
- Stochastic modification: agent ends up in one of 8 adjacent states with $p = 0.1$

Comparing Sarsa, Tree Backup, $Q(0.5)$ and dynamic σ

Synopsis

- *n*-step algorithms are derived from MC and one-step TD methods
- $Q(\sigma)$ unifies *n*-step Sarsa and Tree-backup
- $Q(\sigma)|_{\sigma=0}$ is Tree Backup
- $Q(\sigma)|_{\sigma=1}$ is *n*-step Sarsa

References

Kristopher De Asis, J. Fernando Hernandez-Garcia, G. Zacharias Holland, Richard S. Sutton.
Multi-step Reinforcement Learning: A Unifying Algorithm.
arXiv, 3 Mar 2017.

Richard S. Sutton, Andrew G. Barto.
Reinforcement Learning: An Introduction.
MIT Press, Cambridge, MA, 19 Jun 2017 Draft.

Materials

Presentation, code and other materials are available in the GitHub repository