Cliffwalking with Eligibility Trace

2017- March -24

Di Wu

ID：260562997

Outline

- Eligibility traces
- Accumulating, Dutch, and Replacing Traces
- Simulation:
  - Simulation settings
  - Simulation results for different cases

Eligibility traces

- Another way of combing Monte Carlo and Temporal Difference methods
- λ return is now: $$G_t^{\lambda} \doteq (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{(n)}$$
- Depending on choice of λ. It can be MC (λ= 1) and TD (λ= 0).
- It can be intepreted as average of n-step returns

Eligibility traces

- According to the incrementing strategy, there are mainly three eligibility traces.
- Accumulating trace [1]:

$$\mathbf{e}_0 \doteq \mathbf{0},$$
$$\mathbf{e}_t \doteq \nabla \hat{v}(S_t,\boldsymbol{\theta}_t) + \gamma\lambda\mathbf{e}_{t-1}$$

Eligibility traces

- Replacing trace[1]:

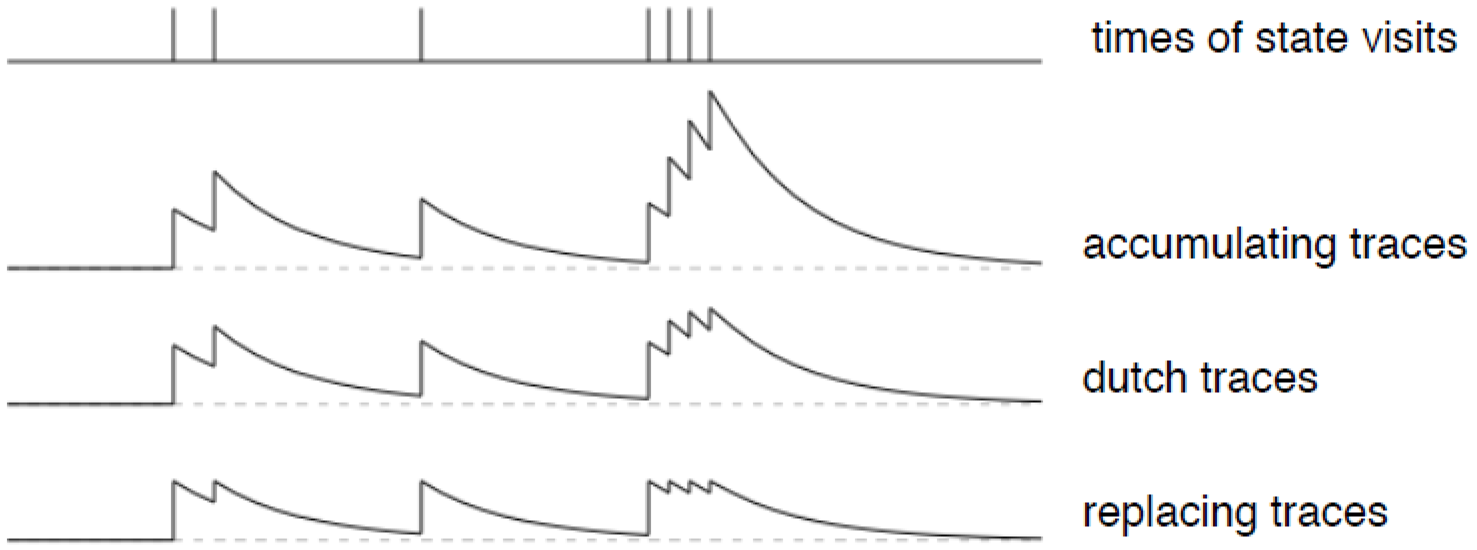
$$e_{i,t} \doteq \begin{cases} 1 & \text{if } \phi_{i,t} = 1 \\ \gamma\lambda e_{i,t-1} & \text{otherwise.} \end{cases}$$

  - Suitable for task with binary features.
- Dutch trace[2][3]:

$$\mathbf{e}_t \doteq \gamma\lambda\mathbf{e}_{t-1} + \left(1 - \alpha\gamma\lambda\mathbf{e}_{t-1}^\top\phi_t\right)\phi_t.$$

Eligibility traces

- Three traces are different in incrementing



This graph is from Sutton's RL course slide

Pseudo code for Sarsa(λ)

- Sarsa(λ):

Initialize $Q(s,a)$ arbitrarily and $e(s,a) = 0$, for all s, a
Repeat (for each episode):
    Initialize s, a
    Repeat (for each step of episode):
        Take action a, observe r, s'
        Choose a' from s' using policy derived from Q (e.g., ε-greedy)
        $\delta \leftarrow r + \gamma Q(s',a') - Q(s,a)$
        $e(s,a) \leftarrow e(s,a) + 1$
        For all s, a:
            $Q(s,a) \leftarrow Q(s,a) + \alpha \delta e(s,a)$
            $e(s,a) \leftarrow \gamma \lambda e(s,a)$
        $s \leftarrow s'$; $a \leftarrow a'$
    until s is terminal

This is from Sutton's RL book website

Pseudo code for Q(λ)

- Q(λ):

Initialize $Q(s,a)$ arbitrarily and $e(s,a) = 0$, for all s, a
Repeat (for each episode):
    Initialize s, a
    Repeat (for each step of episode):
        Take action a, observe r, s'
        Choose a' from s' using policy derived from Q (e.g., ε-greedy)
        $a^* \leftarrow \arg\max_b Q(s', b)$ (if a' ties for the max, then $a^* \leftarrow a'$)
        $\delta \leftarrow r + \gamma Q(s', a^*) - Q(s, a)$
        $e(s,a) \leftarrow e(s,a) + 1$
        For all s, a:
            $Q(s,a) \leftarrow Q(s,a) + \alpha \delta e(s,a)$
            If $a' = a^*$, then $e(s,a) \leftarrow \gamma \lambda e(s,a)$
                        else $e(s,a) \leftarrow 0$
        $s \leftarrow s'$; $a \leftarrow a'$
    until s is terminal

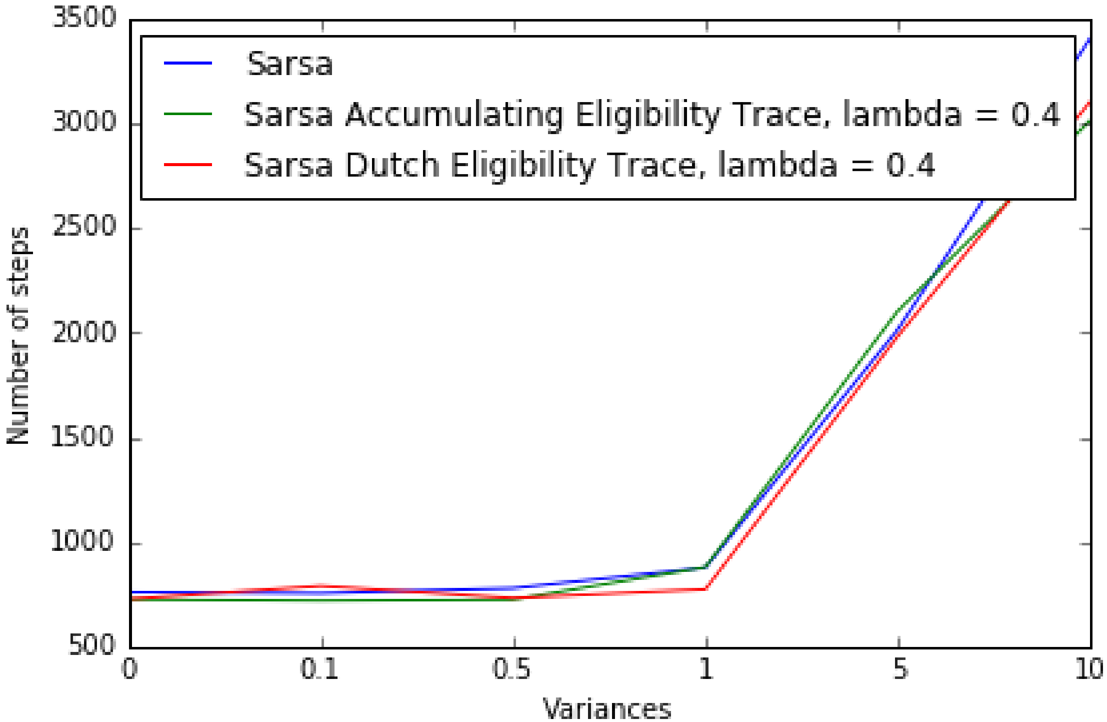
This is from Sutton's RL book website

Simulation results description

- I use the Weiwei Zhang's implementation (Sarsa) on CliffWalking as baselines.
- I implemented the following algorithms:
  - Sarsa with Dutch trace
  - Sarsa with Accumulating trace
  - Q Learning with Accumulating trace
  - Q Learning with Dutch trace

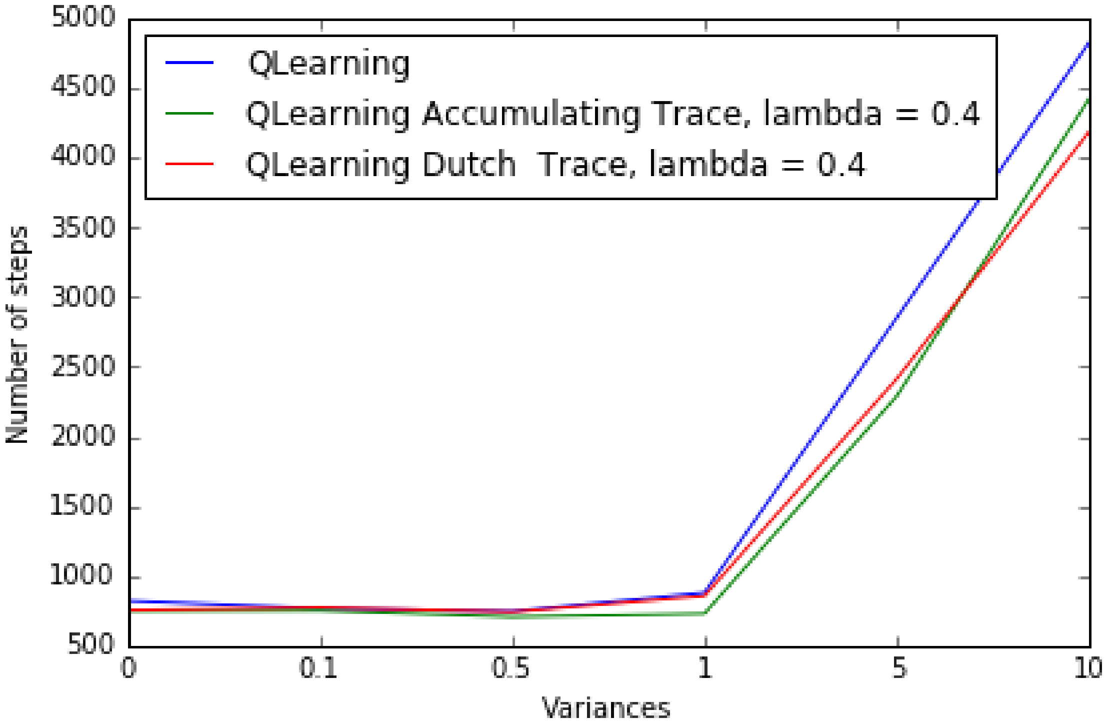
Simulation Results

- Simulation results for Sarsa with dutch trace and accumulating trace



Simulation Results

- Simulation results for Q Learning with dutch trace and accumulating trace



Reference

- [1] Milan, Stephanie, et al. "The impact of physical maltreatment history on the adolescent mother–infant relationship: Mediating and moderating effects during the transition to early parenthood." *Journal of Abnormal Child Psychology* 32.3 (2004): 249-261.

- [2] van Seijen, Harm, and Richard S. Sutton. "True Online TD (lambda)." ICML. Vol. 14. 2014.
- [3] Van Seijen, Harm, et al. "True online temporal-difference learning." Journal of Machine Learning Research 17.145 (2016): 1-40.

Thanks!