Multi-Step Average-Reward Prediction via Differential TD(λ)

Abhishek Naik Richard S. Sutton
abhishek.naik@ualberta.ca rsutton@ualberta.ca
UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

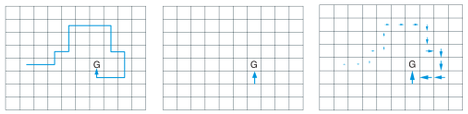
- Algorithms for **multi-step** temporal credit assignment work in **average-reward** reinforcement learning.
- Notably, the two ideas of eligibility traces and maximizing the average reward respectively explain certain neurophysiological and behavioral data.

Ideas

Maximize Average Reward

Maximize the *rate* of reward
instead of
the (discounted) *sum* of rewards.

Perform Multi-step Credit Assignment



Trajectory One-step **Multi-step**

Problem Setting

- Continuing
- On-policy
- Prediction

$$\ldots \; S_t \; A_t \; R_{t+1} \; S_{t+1} \; A_{t+1} \; R_{t+2} \; \ldots$$

$$r(\pi) \doteq \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[R_t \,|\, S_0, A_{0:t-1} \sim \pi]$$

$$v_\pi(s) \doteq \sum_{t=1}^{\infty} \mathbb{E}[R_t - r(\pi) \,|\, S_0 = s, A_{0:t-1} \sim \pi] \quad \forall s$$

Bellman equation

$$v_\pi(s) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s', r \mid s, a) \left[r - r(\pi) + v_\pi(s') \right]$$

Goal: estimate v_π and $r(\pi)$

Learnable parameters: $\mathbf{w} \in \mathbb{R}^d, \bar{R} \in \mathbb{R}$
$d \ll |S|$

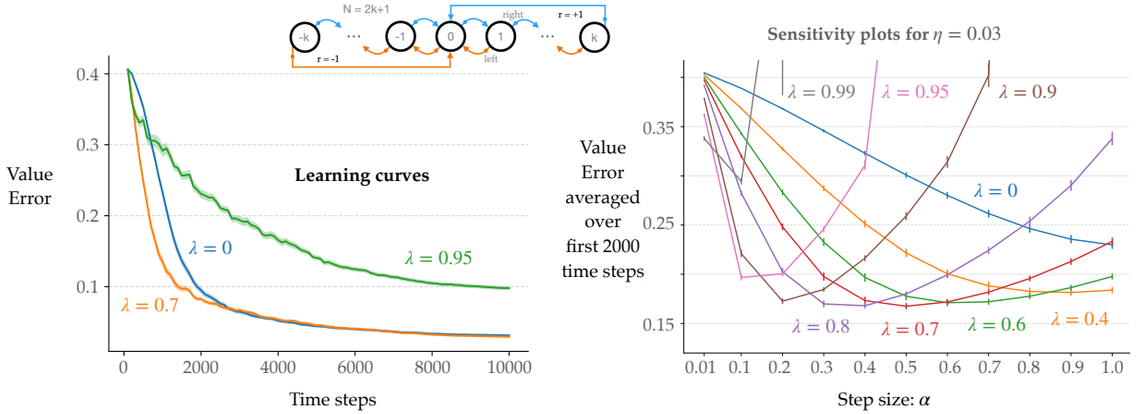
Algorithm and Convergence Result

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t$$

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \delta_t \mathbf{z}_t$$
$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha \delta_t$$
$$\mathbf{z}_{t+1} \doteq \lambda \mathbf{z}_t + \mathbf{x}_{t+1}$$

Under mild technical conditions,
linear on-policy Differential TD(λ) converges w.p.1
for $\lambda \in [0,1)$ to the continuing TD fixed point.

Preliminary Empirical Results



- On-policy Differential TD(λ) converges to the fixed point predicted by the theory.
- An intermediate value of λ results in the most sample efficiency.

Further Comments

- Differential TD(λ) is off-policy ready.
- Biological evidence of the two ideas gives further credence to their presence in the theory of intelligence.

Check out the paper