Reinforcement Learning with Long Short-Term Memory

Written by Bram Bakker (NIPS 2001)

Presented by Tom Bosc

Introduction

- POMDP: Partially Observable MDP
- Instead of states, the agent get an *observation*
- Equivalence between belief MDP and POMDP: the state space of a Belief MDP is on the probability simplex of the states of the POMDP
- Agent needs to estimate the state it's in. Need for memory.
- Relies on several techniques/tricks:
  - Advantage learning
  - Eligibility traces
  - Guided exploration
- In depth experiments: Nielsen 2006: Solving POMDP with RL and extended LSTM

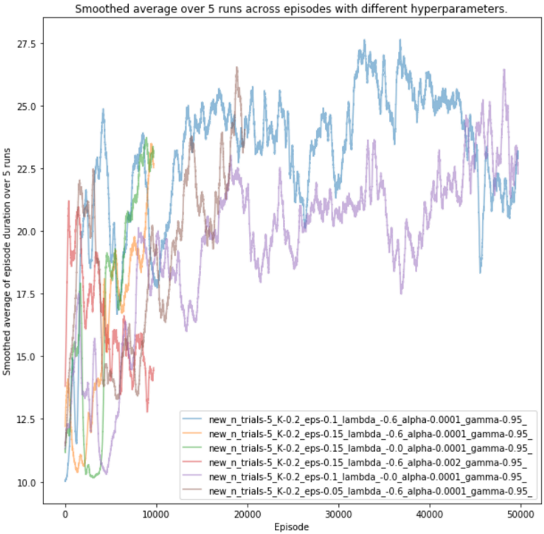
Advantage learning

- Motivation: fine discretization in time leads to slow learning
- $E^{TD}(t) = V_{s_t} + \frac{r_t + \gamma V(s_{t+1}) - V(s_t)}{K} - A(s_t, a_t)$ where $V_{s_t} = max_a A(s_t, a)$ and $K \in [0, 1]$
- Retrieve Q-learning with $K = 1$

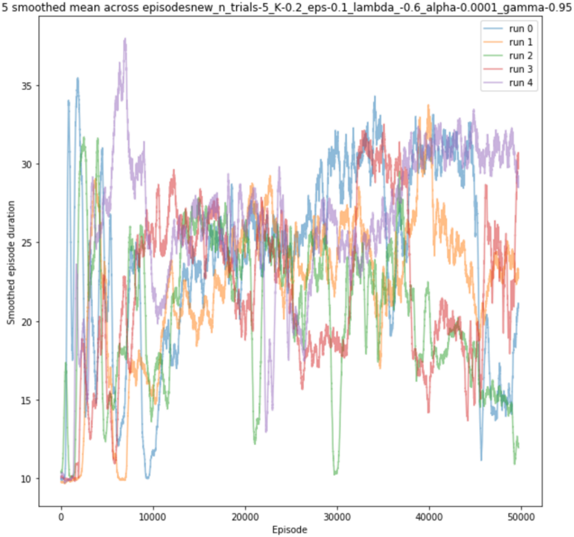
Guided exploration

- Idea: on states with poor value estimates, the agent will explore more.
- Predict TD error with another function approximator $y_v(t)$.
- Target $y_v^d = |E^{TD}(t)| \; + \beta y_v(t+1)$ where β is a discount factor (notice bootstrap here)
- Then, temperature of Boltzmann distribution is $\tau = C * y_v(t)$.
- Low temperature either when TD error is low or when future TD error is low.

Experiments on cart pole



Experiments on cart pole



Conclusion

- RNNs can deal with POMDP.
- Online (no BPTT), no experience replay (although code help)
- Hard to tune.
- According to Nielsen 2006 (on a maze problem): K matters, low λ is better, guided exploration is useful (but not on cart pole according to Bakker 2001)