Eligibility Traces for Options

Ayush Jain

April 20th, 2017

Motivation

- Problem of Temporal Abstraction
  - Connect high and low level behavior with minimum changes to RL framework.
  - How ??? - Options
- Intra-Option Learning
  - Take advantage of each fragment of experience
  - Incremental, step-by-step updates
- Eligibility Traces
  - Interpolation between TD(0) and MC
  - Implement λ-Return, control bias-variance tradeoff
- Off-Policy Evaluation
  - Enable agent to use experience to learn about many different policies, each belonging to a different macro-action

Options framework

A Markov option o : ($I \in S$, $\pi: S \times A \rightarrow [0,1]$, $\beta: S \rightarrow [0,1]$)

Hierarchical policy over options – $\mu: S \times A \rightarrow [0,1]$

- An initializable option is selected with probability $\mu(o|s)$
- Option’s internal policy is followed to select actions
- Option terminates with $\beta(s)$, new option is selected again with μ

Intra-Option Learning

- Take advantage of each fragment of experience
- SMDP learning: option executed to termination keeping track of rewards, update applied only to the option taken
- Intra-Option learning: after each primitive action, update every option that could have taken that action, based on reward observed and bootstrapping from next state’s value

Off-Policy Evaluation

Per-Decision Importance Sampling Approach

- Weigh updates with a factor correcting trajectory probability; or simply product of importance sampling ratios for 0 – t
- Behavior policy b needs to be known, high variance if π and b are too different.

Tree Backup

- Combine value estimates for actions with their probabilities under the target policy
- New target is formed using old estimates of values for actions not taken and new estimate of value for the action taken, iterated over many steps
- Behavior b can be unknown, cuts traces quickly

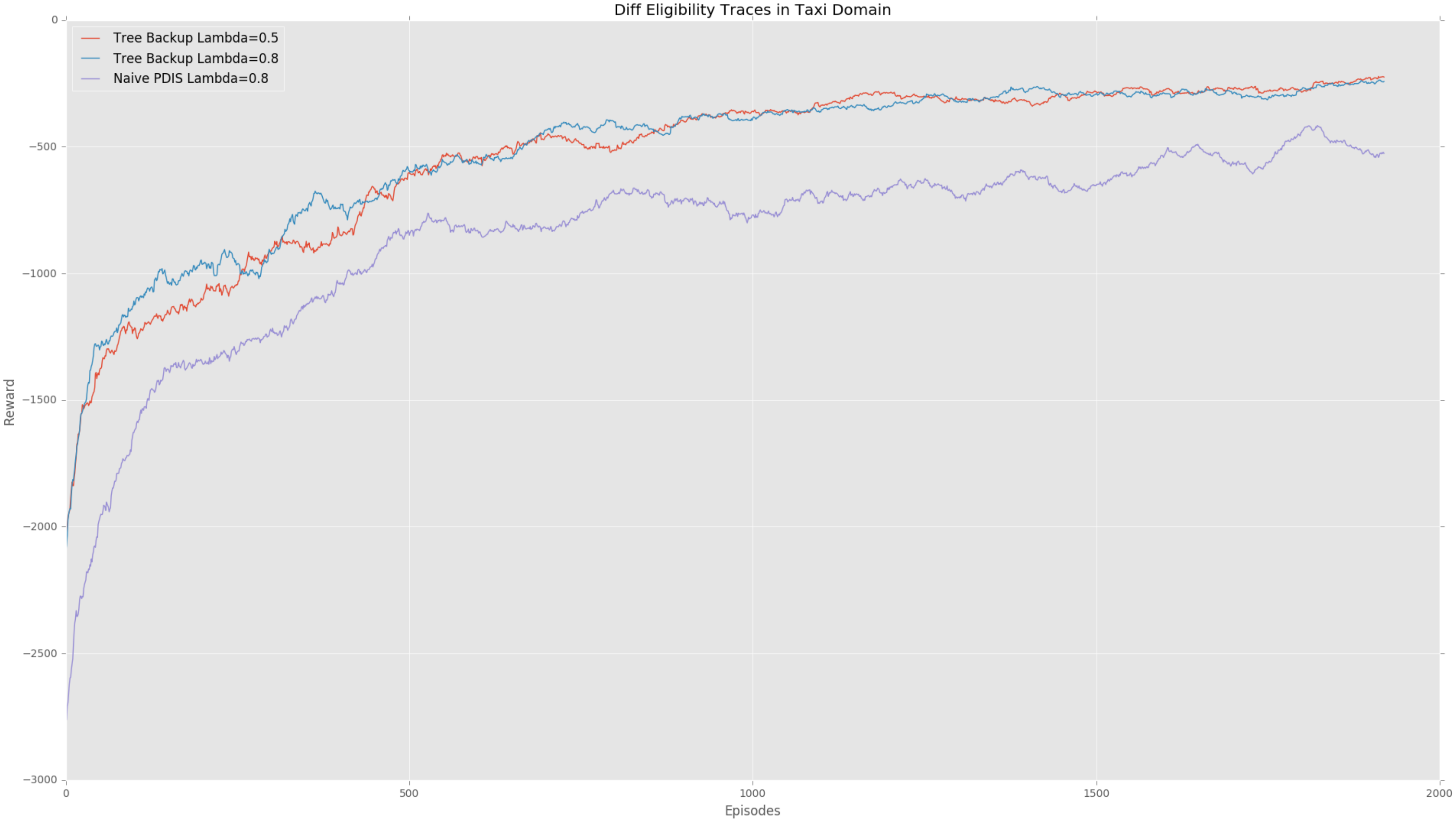
Recognizers

- Function c: S x A→[0,1] indicates to what extend an action is recognized in a state. Recognizer with a behavior policy defines target policy.
- Π: c(s,a)xb(s,a)/μ, where recognition probability μ = Σ c(s,a)xb(s,a)

Traces for Options

On blackboard

Results


Diff Eligibility Traces in Taxi Domain
Tree Backup Lambda=0.5
Tree Backup Lambda=0.8
Naive PDIS Lambda=0.8
Reward
0
−500
−1000
−1500
−2000
−2500
−3000
Episodes
0
500
1000
1500
2000