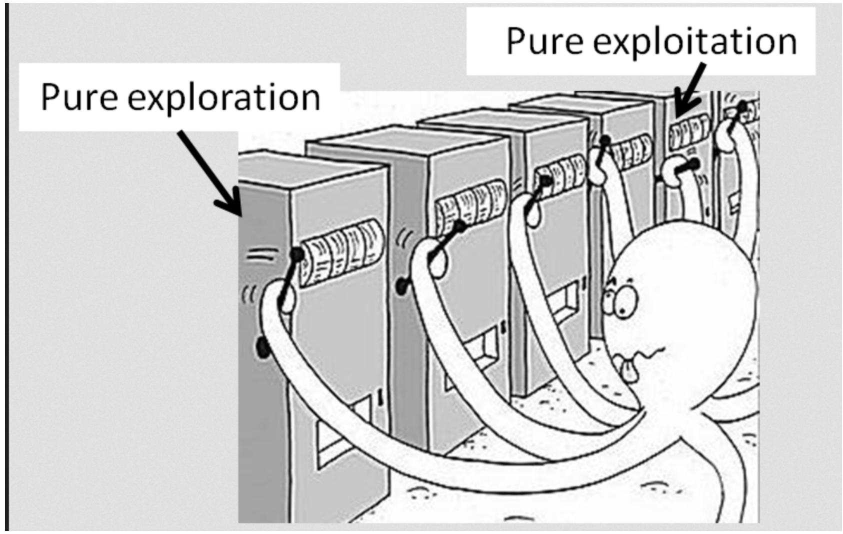
Bayesian Reinforcement Learning

Inference on Transition Dynamics

Haochen Wu
12/18/2019



Introduction

- Background
  - Markov Decision Process <S,A,T,R>
- Motivation
  - Exploration-Exploitation Dilemma
  - Planning under unknown system dynamics
- Objectives
  - Parameterize the transition dynamics for inference
  - Maintain the belief on transition probability distribution
  - Find the policy that naturally optimizes the expected total reward given exploration-exploitation dilemma

Problem Formulation

- Bayesian Reinforcement Learning is defined as a tuple $< \overline{S}, A, \overline{T}, R >$
  - $\overline{S}: S \times B$ is defined by augmenting the physical world state s and belief state $b(\theta)$ on transition dynamics
  - S, A, R are the same as the regular MDP
  - $\overline{T}: S \times B \times A \times S \times B \rightarrow [0,1]$, transition probability from augmented state (s, b) to (s', b') under action a
    - $\overline{T}(s, b, a, s', b') = \Pr(s', b'|s, b, a) = \Pr(s'|s, b, a)\Pr(b'|s, b, a, s')$
    - The first term could be easily determined under given belief over the transition model parameters
    - In the second term, b' is the posterior distribution over transition parameters given prior b and data (s, a, s)
- Dirichlet Distribution – Conjugate Prior
  - Let $\theta_{s,a}^{s'}$ denote each transition parameter, satisfying $\sum_{s' \in S} \theta_{s,a}^{s'} = 1$, which is the standard $|S| - 1$ simplex.
  - Transition dynamics for each state-action pair could be modeled as a Dirichlet Distribution denoted as $Dir(\alpha)$
  - α is a vector of $|S|$ real numbers. $b(\theta_{s,a}) \sim Dir(\alpha)$
- Belief Monitoring using Bayes Rule – $b'(\theta_{s,a}) \propto \Pr(s'|s, a, \theta_{s,a})\, b(\theta_{s,a}|s, a, s')$
  - $Dir(\alpha'_{s,a}) \propto \theta_{s,a}^{s'} Dir(\alpha_{s,a}) = \theta_{s,a}^{s'} \prod_{s'' \in S} {\theta_{s,a}^{s''}}^{\alpha_{s,a}^{s''} - 1}$
  - Given data (s, a, s), the belief update on transition parameters becomes $\alpha_{s,a}^{s'} = \alpha_{s,a}^{s'} + 1$

Methods

- Offline approach is computationally intractable
  - Belief space is infinite and continuous.
  - Parameterizing the distribution still has a large number of states.
- Proposed algorithm
  - Thompson Sampling – sampling from the posterior belief
    - choose the action that maximizes the expected reward over the samples
  - Treating Thompson Sampling as a trial of playing a game
  - Experience Replay with Thompson Sampling
    - Eliminates ϵ-greedy exploration-exploitation dilemma
    - Keeps a memory of past behaviors and randomly samples previous transitions
    - Speeds up the credit propagation and adapt to recent experiences

Algorithm 2 Bayesian RL with Experience Replay

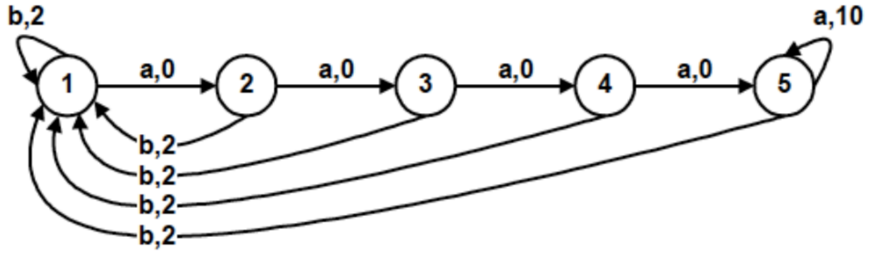
Initialize replay memory $\mathcal{D}$ with capacity N
Initialize prior $\{\alpha_{a,s}\}$ for transition parameters
while max episode is not reached **do**
    $s = s_0$
    while max iteration is not reached **do**
        Sample $\theta^k_{a,s}$ from $\alpha_{a,s}, \forall a \in A, s \in S$
        for $i = 1 : k$ **do**
            $Q_i(s,a) = solveMDP(\theta^i)$
        end for
        $\overline{Q}(s,a) = \frac{1}{k}\sum_i Q_i(s,a)$
        $a^* = argmax_a \overline{Q}(s,a)$
        Execute a^* and receive data (s,a,s',r)
        Store (s,a,s',r) into $\mathcal{D}$
        Randomly select a minibatch (s,a,s',r) from $\mathcal{D}$
        $\alpha^{s'}_{s,a} = \alpha^{s'}_{s,a} + 1$
        $s = s'$
    end while
end while

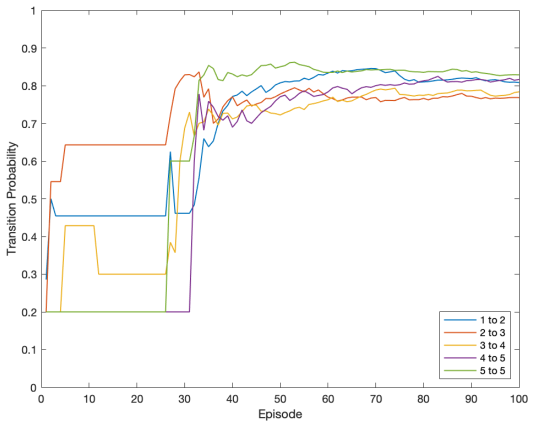
Results – On Chain Problem

True transition probability:
Action a: 0.8 move to right, 0.2 go to state 1
Action b: 0.2 move to right, 0.8 go to state 1

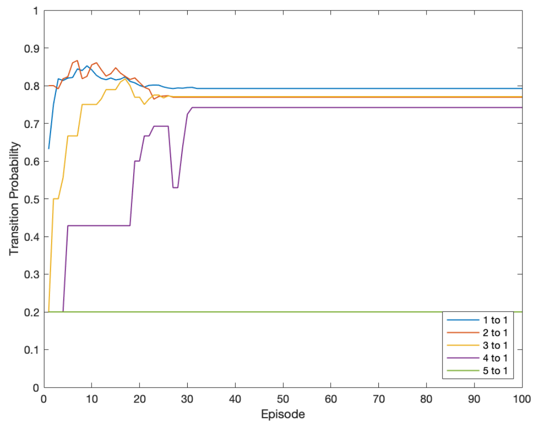
Reward:
Action a: 10 at state 5
Action b: 2 at each state

- Chain Problem
  - The agent would get comfortable with taking action b
  - Bayesian RL would help the agent step out of its comfort zone
- Expected Regret [1]
  - The lose amount of not taking an optimal action in each step
  - Equivalent but more convenient measure of expected total reward
- Posteriors on Transition Dynamics



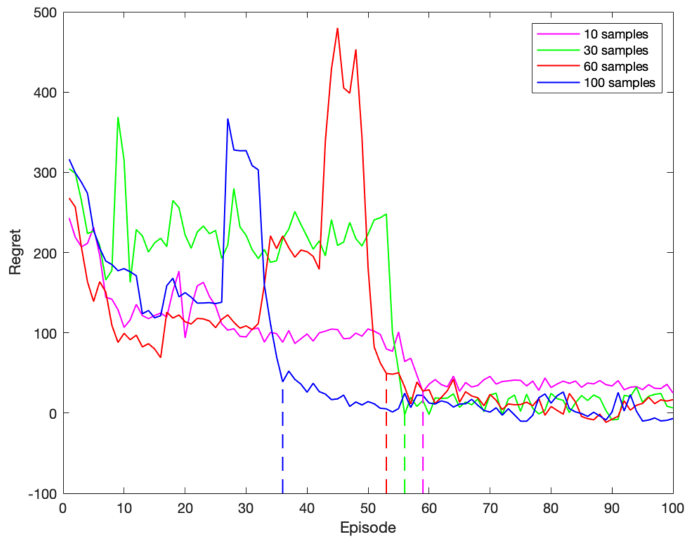


Action a, move to right



Action b, set to initial

n_s	Ep_r	$Score_b$	$Score_a$
10	59	0.1356	0.1463
30	56	3.9107	15.2500
60	53	2.0000	18.5319
100	36	3.0278	18.1250



[1] Agrawal and Goyal, ICAIS, 2013

Conclusions

- Model the problem as Bayesian Reinforcement learning
  - Learning the parameters of system transition dynamics
  - Solving the exploration-exploitation dilemma
  - Find the optimal action under uncertain transition to maximize the expected total rewards
- Experience Replay with Thompson Sampling
  - An online algorithm to interact with the environment while monitoring the belief over transition parameters
  - Q-Value is approximated by Thompson Sampling
  - The algorithm could be extended to inferring the transition dynamics and learning the Q-Value in parallel by representing the Q-Value as Q-Network