CS885 Reinforcement Learning
Lecture 10: June 1, 2018

Bayesian RL

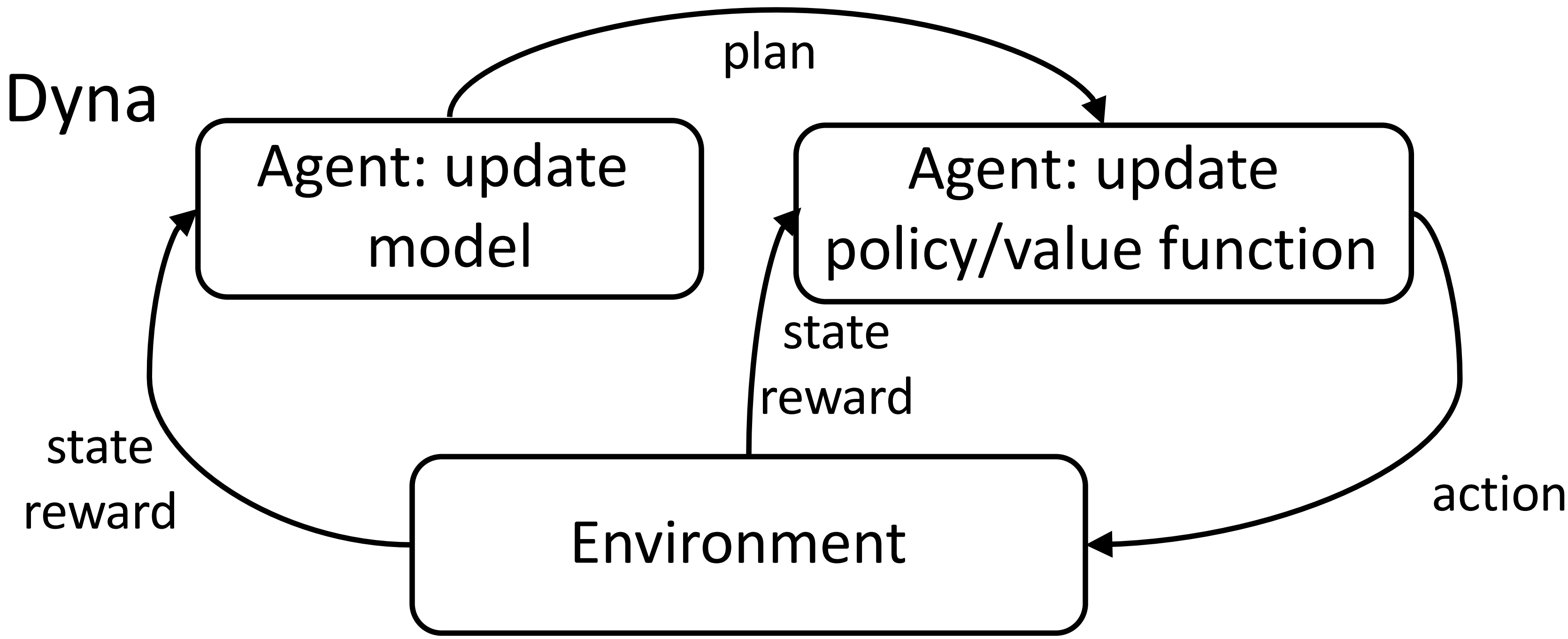
Reading: Michael O'Gordon Duff's PhD Thesis (2002)

Outline

- Model-based Bayesian RL
  - Value iteration with belief model
  - Thompson sampling in Bayesian RL
  - PILCO: model-based Bayesian actor critic

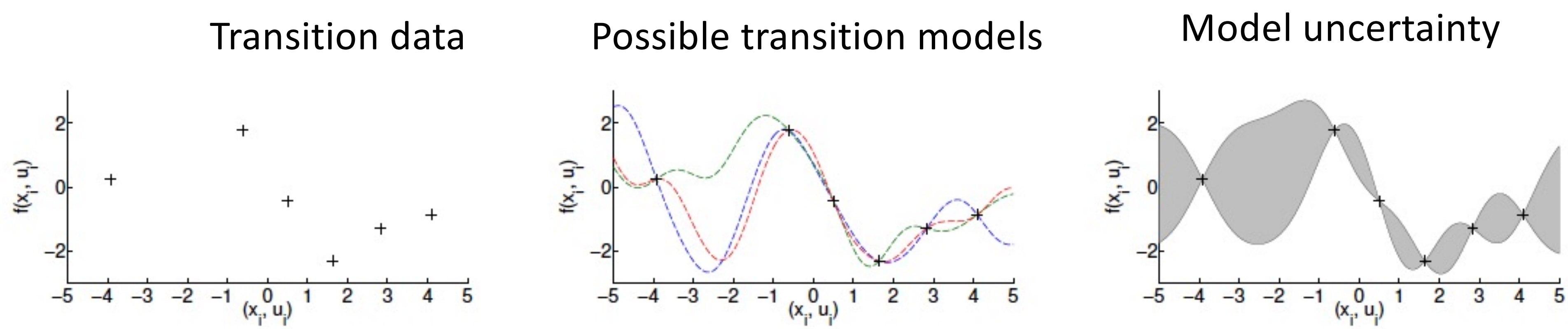
Model-free vs model-based RL

- **Model-free RL:** unbiased direct learning, needs many interactions with environment
- **Model-based:** biased indirect learning via a model, if bias is not too important then less data needed



Biased model

- Problem:
  - Model learned from finite amount of data
  - Model is necessarily imperfect
  - There is a risk that planning will overfit the model inaccuracies and produce a bad policy
- Solution: represent uncertainty in model



Bayesian RL

- Explicit representation of uncertainty
- Benefits
  - Balance exploration/exploitation tradeoff
  - Mitigate model bias
  - Reduce data needs
- Drawback
  - Complex computation

Traditional RL

- Reinforcement Learning
  - States: $\boldsymbol{s} \in \boldsymbol{S}$
  - Actions: $\mathbf{a} \in \boldsymbol{A}$
  - Rewards: $\mathbf{r} \in \mathbb{R}$
  - Unknown model: $\mathbf{Pr}(\boldsymbol{r}, \boldsymbol{s}' | \boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\theta})$

- Goal: find policy $\pi: S \to A$

  and/or value function $Q: S \times A \to \mathbb{R}$

Bayesian RL

- Idea: augment state with distribution about unknown parameters
  - Information states: $(\boldsymbol{s}, \boldsymbol{b}) \in \boldsymbol{S} \times \boldsymbol{B}$
    - Physical states: $\boldsymbol{s} \in \boldsymbol{S}$
    - Belief states: $\boldsymbol{b} \in \boldsymbol{B}$ where $b(\theta) = \Pr(\theta)$
  - Actions: $\mathbf{a} \in \boldsymbol{A}$
  - Rewards: $\mathbf{r} \in \mathbb{R}$
  - Known model: $\mathbf{Pr}(\boldsymbol{r}, \boldsymbol{s}', \boldsymbol{b}' | \boldsymbol{s}, \boldsymbol{b}, \boldsymbol{a})$
- Goal: find policy $\pi: S \times B \to A$

  and/or value function $Q: S \times B \times A \to \mathbb{R}$

Model in Bayesian RL

- Claim: the model in Bayesian RL is known!

$$\Pr(r, s', b' | s, b, a) = \underbrace{\Pr(r, s' | s, b, a)}_{\text{Physical model}} \underbrace{\Pr(b' | r, s', s, b, a)}_{\text{belief model}}$$

- Idea: integrate out unknown θ

$$\Pr(r, s' | s, b, a) = \int_{\theta} \Pr(r, s' | s, a, \theta)\, b(\theta)\, d\theta$$

- Idea: b' is the posterior belief

$$\Pr(b' | r, s', s, b, a) = \begin{cases} 1 & \text{if } b'(\theta) = b^{s,a,s',r} = b(\theta | s, a, s', r) \\ 0 & \text{otherwise.} \end{cases}$$

Maze Example

3	r	r	r	+1
2	u		u	-1
1	u	l	l	l
	1	2	3	4

$\gamma = 1$

Reward is -0.04 for non-terminal states

Transition model (when ignoring boundaries):

$$\Pr(i',j'|i,j,right,\theta) = \begin{cases} \theta & i' = i+1 \text{ and } j' = j \\ \frac{1-\theta}{2} & i' = i \text{ and } (j' = j+1 \; or \; j' = j-1) \\ 0 & \text{otherwise} \end{cases}.$$

$$\Pr(i',j'|i,j,up,\theta) = \begin{cases} \theta & i' = i \text{ and } j' = j+1 \\ \frac{1-\theta}{2} & (i' = i+1 \; or \; i' = i-1) \text{ and } j' = j. \\ 0 & \text{otherwise} \end{cases}.$$

(similarly for the other actions)

Belief state

- Let's model our uncertainty with respect to θ by a Beta distribution

$$b(\theta) = k\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Belief update: Bayes theorem

$$\begin{aligned} b'(\theta) &= b^{s,a,s'}(\theta) \\ &= b(\theta|s,a,s') \\ &\propto b(\theta)Pr(s'|s,a,\theta) \end{aligned}$$

Example belief update

- Prior

$$b(\theta) = Beta(\theta; \alpha, \beta) = k\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Posterior for $i, j, up \rightarrow i', j'$ where $i' = i$ and $j' = j + 1$
- Belief update: Bayes theorem

$$\begin{aligned} b'(\theta) = b^{s,a,s'}(\theta) = b(\theta|s,a,s') &= b(\theta|i,j,up,i',j') \\ &\propto b(\theta)Pr(i',j'|i,j,up,\theta) \\ &= k\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta \\ &= k\theta^{\alpha}(1-\theta)^{\beta-1} \propto Beta(\theta; \alpha+1, \beta) \end{aligned}$$

Physical Model

- Consider $s = (i, j), a = right, s' = (i', j')$
  where $i' = i$ and $j' = j - 1$

- Predictive distribution

$$\Pr(s'|s,b,a) = \int_\theta \Pr(s'|s,a,\theta)\, b(\theta)\, d\theta$$
$$= \int_\theta \Pr(i',j'|i,j,right,\theta)\, Beta(\theta;\alpha,\beta) d\theta$$
$$= \int_\theta \frac{(1-\theta)}{2} k\theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = \frac{\beta}{2}$$

Planning

- Since the model is known, treat Bayesian RL as an MDP
- Benefits:
  - Solve RL problem by planning (e.g., value/policy iteration)
  - Optimal exploration/exploitation tradeoff
- Drawback:
  - Complex computation
- Bellman's Equation:

$$V^*(s,b) = \max_a E[r|s,b,a] + \gamma \sum_{s'} \Pr(s'|s,a,b)\, V^*(s', b^{s,a,s'}) \;\; \forall s$$

where $E[r|s,b,a] = \int_\theta b(\theta) \int_r pdf(r|s,a,\theta)\, r\, dr\, d\theta$

Value Iteration

- Traditional MDP

valueIteration(MDP)

$V_0^*(s) \leftarrow \max_a E[r|s,a] \;\; \forall s$

For $t = 1$ to h do

$\qquad V_t^*(s) \leftarrow \max_a E[r|s,a] + \gamma \sum_{s'} \Pr(s'|s,a) \, V_{t-1}^*(s') \;\; \forall s$

Return V^*

- Information state MDP

valueIteration(BayesianRL)

$V_0^*(s,b) \leftarrow \max_a E[r|s,b,a] \;\; \forall s$

For $t = 1$ to h do

$\qquad V_t^*(s,b) \leftarrow \max_a E[r|s,b,a] + \gamma \sum_{s'} \Pr(s'|s,a,b) \, V_{t-1}^*(s', b^{s,a,s'}) \;\; \forall s$

Return V^*

Exploration/exploitation tradeoff

- Dilemma:
  - ~~Maximize immediate rewards (exploitation)?~~
  - ~~Or, maximize information gain (exploration)?~~
- **Wrong question!**

- Single objective: max expected total rewards
  - $V^{\pi}(s,b) = \sum_t \gamma^t E[r_t | s_t, b_t]$
  - Optimal policy π^*: $V^{\pi^*}(s,b) \geq V^{\pi}(w,b)$ for all *s,b*
    - **Optimal exploration/exploitation tradeoff (given prior knowledge)**

Bayesian RL

- Two phases:
  - Offline planning (without the environment)

Find π^* and/or V^*
by policy/value iteration or any other algorithm

  - Online execution (with the environment)

Initialize $s_0, b_0, n \leftarrow 0$
Repeat
    Execute policy $a_n \leftarrow \pi(s_n, b_n)$
    receive s_{n+1} and r_n from the environment
    Belief update: $b_{n+1}(\theta) = b_n^{s_n,a_n,r_n,s_{n+1}}(\theta) = b_n(\theta|s_n, a_n, r_n', s_{n+1}')$
    $n \leftarrow n + 1$

Challenges in Bayesian RL

- Offline planning is notoriously difficult
  - Use function approximators (e.g., Gaussian process or neural net) for model, V^* and π^*
  - Continuous belief space
  - Problem: a good plan should implicitly account for all possible environments, which is intractable

- Alternative: online partial planning
  - Thompson sampling
  - PILCO (Model-based Bayesian Actor Critic)

Thompson Sampling in Bayesian RL

- Idea: Sample models θ_i at each step and plan for the corresponding MDP_{θ_i}'s

ThompsonSamplingInBayesianRL(s,b)
Repeat
    Sample $\theta_1, \dots, \theta_k \sim \Pr(\theta)$
    $Q^*_{\theta_i} \leftarrow solve\left(MDP_{\theta_i}\right) \; \forall i$
    $\hat{Q}(s,a) \leftarrow \frac{1}{k}\sum_{i=1}^{k} Q^*_{\theta_i}(s,a) \;\; \forall a$
    $a^* \leftarrow \mathrm{argmax_a}\; \hat{Q}(s,a)$
    Execute a^* and receive r, s'
    $b(\theta) \leftarrow b(\theta)\Pr(r, s' | s, a^*, \theta)$
    $s \leftarrow s'$

Model-based Bayesian Actor Critic

- PILCO: Deisenroth, Rasmussen (2011)
  - $b(\theta)$: Gaussian Process transition model
- Deep PILCO: Gal, McCallister, Rasmussen (2016)
  - $b(\theta)$: Bayesian neural network transition model

PILCO(s, b, π)
    Repeat
        Repeat
            Critic: $V_b^\pi \leftarrow policyEvaluation(b, \pi)$
            Actor: $\pi \leftarrow \pi + \alpha\, \partial V_b^\pi / \partial \pi$
        $a \leftarrow \pi(s, b)$
        Execute a and receive r, s'
        $b \leftarrow b^{s,a,r,s'}$ and $s \leftarrow s'$

Unprecedented Data Efficiency

Table 1. PILCO's data efficiency scales to high dimensions.

	cart-pole	cart-double-pole	unicycle
state space	$\mathbb{R}^4$	$\mathbb{R}^6$	$\mathbb{R}^{12}$
# trials	≤ 10	20–30	≈ 20
experience	$\approx 20\,\mathrm{s}$	$\approx 60\,\mathrm{s}$–$90\,\mathrm{s}$	$\approx 20\,\mathrm{s}$–$30\,\mathrm{s}$
parameter space	$\mathbb{R}^{305}$	$\mathbb{R}^{1816}$	$\mathbb{R}^{28}$

Cartpole problem