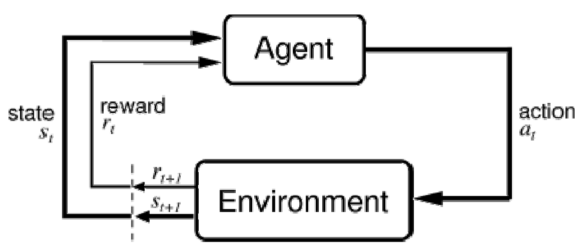
CS 287: Advanced Robotics
Fall 2009

Lecture 9: Reinforcement Learning 1: Bandits

Pieter Abbeel
UC Berkeley EECS

Reinforcement Learning



- Model: Markov decision process (S, A, T, R, γ)
  - Goal: Find π that maximizes expected sum of rewards
- T and R might be unknown

[Drawing from Sutton and Barto, Reinforcement Learning: An Introduction, 1998]

Exploration vs. exploitation

- = classical dilemma in reinforcement learning
- A conceptual solution: Bayesian approach:
  - State space = { x : x = probability distribution over T, R}
    - For known initial state --- tree of sufficient statistics could suffice
  - Transition model: describes transitions in new state space
  - Reward = standard reward
- Today: one particular setting in which the Bayesian solution is in fact computationally practical

Multi-armed bandits

- Slot machines
- Clinical trials
- Advertising
- Merchandising

Multi-armed bandits

Consider slot machines $H_1, H_2, \ldots, H_n$.

Slot machine i has pay-off $= \begin{cases} 0, & \text{with probability } 1-\theta_i \\ 1, & \text{with probability } \theta_i \end{cases}$

where θ_i is unknown.

Now the objective to maximize is:

$$E[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)] \text{ (where } s_t \text{ is unchanged).}$$

Information state

$$\begin{pmatrix} \text{\# of successes on } H_1 \\ \text{\# of failures on } H_1 \\ \text{\# of successes on } H_2 \\ \text{\# of failures on } H_2 \\ \vdots \\ \text{\# of successes on } H_n \\ \text{\# of failures on } H_n \end{pmatrix}$$

Semi-MDP

- Transition model:

$$P(s_{t_{k+1}} = s', t_{k+1} = t_k + \Delta | s_k = s, a_k = a)$$

- Objective:

$$E[\sum_{k=0}^{\infty} \gamma^{t_k} R(s_{t_{k+1}}, \Delta_k, s_{t_k})].$$

- Bellman update:

$$V(s) = \max_a \sum_{s',\Delta} P(s', \Delta | s, a)[R(s, \Delta, a, s') + \gamma^{\Delta} V(s')]$$

Optimal stopping

- A specialized version of the Semi-MDP is is the "Optimal stopping problem". At each of the times, we have two choices:
  - 1. continue
  - 2. stop and accumulate reward g for current time and for all future times.
- The optimal stopping problem has the following Bellman update:

$$V(s) = \max\{\sum_{s',\Delta} P(s', \Delta | s)[R(s, \Delta, s') + \gamma^{\Delta} V(s')], \frac{g}{1-\gamma})$$

Optimal stopping

- Optimal stopping Bellman update:

$$V(s) = \max\{\sum_{s',\Delta} P(s',\Delta|s)[R(s,\Delta,s') + \gamma^{\Delta} V(s')], \frac{g}{1-\gamma})$$

- Hence, for fixed g, we can find the value of each state in the optimal stopping problem by dynamic programming
- However, we are interested in g*(s) for all s:

$$g^*(s) = \min\{g| \frac{g}{1-\gamma} \geq \max_{\tau} \mathbb{E}_{\tau}[\sum_{k=0}^{\tau-1} \gamma^{t_k} R(s_{t_k}, \Delta_k, s_{tk+1}) + \sum_{t=\tau}^{\infty} \gamma^t g]$$

- Note: τ is a random variable, which denotes the stopping time. It is the policy in this setting.
- Any stopping policy can be represented as a set of states in which we decided to stop. The random variable τ takes on the value = time when we first visit a state in the stopping set.

Optimal stopping

- One approach:
  - Solve the optimal stopping problem for many values of g, and for each state keep track of the smallest value of g which causes stopping

Reward rate

- Reward rate

$$\sum_{t=0}^{\Delta_{t_k}-1} \gamma^t r(s_{t_k}, \Delta_{t_k}, s_{t_{k+1}}) = R(s_{t_k}, \Delta_k, s_{t_{k+1}})$$

- Expected reward rate

$$\bar{r}(s) = \mathbb{E}_{\Delta,s'}[r(s,\Delta,s')] = \sum_{\Delta,s'} P(s',\Delta|s) r(s,\Delta,s')$$

Basic idea to find g*

Now consider

$$s^* - \arg\max_s \bar{r}(s).$$

Of all states, the state s^* would require the highest payoff to be willing to stop. Namely,

$$g^*(s^*) = \bar{r}(s^*)$$

This means that when $g < g^*(s^*)$, the optimal stopping policy will choose to continue at s^*.
Note that for $s \neq s^*$, $g^*(s) < g^*(s^*)$.

To compute $g^*(s)$ for the other states, we consider a new semi-MDP which differs from the existing one only in that we always continue when at state s^*. This is equivalent to letting the new state space $\bar{\mathcal{S}} = \mathcal{S} \backslash \{s^*\}$.

Finding the optimal stopping costs

$while |\bar{\mathcal{S}}| > 0:$

$\mathcal{S} \leftarrow \mathcal{S} \backslash \{s^*\}$

Adjust the transition model and the reward function accordingly—namely, assuming we always continue when visiting state s^*.

Compute reward rates$r$ by solving: $\sum_{t=0}^{\Delta-1} \gamma^t \bar{r}(s, \Delta, s') = \bar{R}(s, \Delta, s')$

Compute expected reward rates $\bar{r}(s) = \mathbb{E}_{\Delta, s'} r(s, \Delta, s')$.

$s^* \leftarrow \arg\max_s \bar{r}(s)$

$g^*[s^*] \leftarrow \bar{r}(s^*)$

Solving the multi-armed bandit

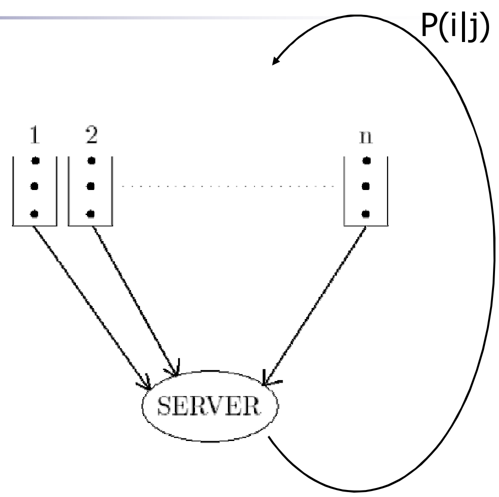
- 1. Find the optimal stopping cost $g^*(s_t^{(i)})$ for each bandit's current state
- 2. When asked to act, pull an arm i such that

$$i \in \arg\max_i g^{*(i)}(s_t^{(i)})$$

Key requirements

- Reward at time t only depends on state of M_i at time t
- When pulling M_i, only state of M_i changes
- Note: M_i need not "just" be a bandit; we just need to be able to compute its optimal stopping cost

Example: cashier's nightmare



P(i|j): probability of joining queue i after being served in queue j
c_i : cost of a customer being in queue i

Further readings

- Gittins, J.C., D.M. Jones. 1974. A dynamic allocation index for the sequential design of experiments.*["Gittins indices"]*
- Different family of approaches: regret-based
  - Lai and Robbins, 1985
  - Auer +al, UCB algorithm (1998)

  Type of result: after n plays, the regret is bounded by an expression O(log n)

  After n plays the regret is defined by:

  $$n\mu^* - \sum_j \mu_j \mathrm{E}[T_j(n)] \text{where } \mu^* = \max_j \mu_j$$

  Deterministic policy: UCB1.
  Initialization: Play each machine once.
  Loop:

  – Play machine j that maximizes $\bar{x}_j + \sqrt{\frac{2 \ln n}{n_j}}$, where $\bar{x}_j$ is the average reward obtained from machine j, n_j is the number of times machine j has been played so far, and n is the overall number of plays done so far.

- Loosening and strengthening assumptions, e.g.,
  - Guha, S., K. Munagala. 2007. Approximation algorithms for budgeted learning problems. *STOC '07.*
  - Various Robert Kleinberg publications
  - "contextual bandit" setting