Mathematical Foundations of Reinforcement Learning

Homework 1

Due Date*: March 27, 2017

Problem 1. You suddenly realize to ward the end of the semester that you have three courses that have assigned a term project instead of a final exam. You quickly estimate howw much each one will take to get 100 points (equivalent to an A+) on the project. You then guess that if you invest t hours in a project, which you estimated would need T hours to get 100 points, then for $t < T$ your score will be

$$R = 100\sqrt{t/T} \tag{1}$$

That is there are declining marginal returns to putting more work into a project. So, if a project is projected to take 40 hours and you only invest 10, you estimate that your score will be 50 points (100 times the square root of 10 over 40) You decide that you cannot spend more than a total of 30 hours on the projects, and you want to choose a value of t for each project that is a multiple of 5 hours. You also feel that you need to spend at least 5 hours on each project (that is , you cannot completely ignore a project). The time you estimate to get full score on each of the four projects is given by

Project	Completion time T
1	20
2	15
3	10

You decide to solve the problem as a dynamic program.

(a) What is the state variable and decision epoch for this problem?

(b) What is your reward function?

(c) Write out the problem as an optimization problem.

(d) Set up the optimality equations.

(e) Solve the optimality equations to find the right time investment strategy.

Problem 2. You have to send a set of questionnaires to each of N population segments. The size of each population segment is given by ω_i. You have a budget of B questionnaires to allocate among the population segments. If you send x_i questionnaires to segment i, you will have a sampling error proportional to

$$f(x_i) = \frac{1}{\sqrt{x_i}} \tag{2}$$

You want to minimize the weighted sum of sampling errors, given by

$$F(x) = \sum_{i=1}^{N} \omega_i f(x_i) \tag{3}$$

You wish to find the allocation x that minimizes $F(x)$ subject to the budget constraint $\sum_{i=1}^{N} x_i \leq B$. Set up the optimality equations to solve this problem as a dynamic program (needless to say, we are only interested in integer solutions).

Problem 3. An oil company will order tankers to fill a group of large storage tanks. One full tanker is required to fill an entire storage tank. Orders are placed at the beginning of each four-week accounting period but do not arrive until the end of the accounting period. During this period the company may be able to sell $0, 1$, or 2 tanks of oil to one of the regional chemical companies (orders are conveniently made in units of storage tanks). The probability of a demand of 0, 1, or 2 is $0.40, 0.40$, and 0.20, respectively.
A tank of oil costs \$1.6 million (M) to purchase and sells for \$2M. It costs \$0.020M to store a tank of oil during each period (oil ordered in period t, which cannot be sold until period $t + 1$, is not charged to any holding cost in period t). Storage is only charged on oil that is in the tank at the beginning of the period and remains unsold during the period. It is possible to order more oil than can be stored. For example, the company may have two full storage tanks, order three more, and then only sell one. This means that at the end of the period, they will have four tanks of oil. Whenever they have more than two tanks of oil, the company must sell the oil directly from the ship for a price of \$0.70M. There is no penalty for unsatisfied demand.
An order placed in time period t must be paid for in time period t even though the order does not arrive until $t + 1$. The company uses an interest rate of 20 percent per accounting period (i.e., a discount factor of 0.80).

(a) Give an expression for the one-period reward function $r(s, d)$ for being in state s and making decision d. Compute the reward function for all possible states (0,1,2) and all possible decisions (0,1,2)

(b) Find the one-step probability transition matrix when your action is to order one or two tanks of oil. The transition matrix when you order zero is given by

From-to	0	1	2
0	1	0	0
1	0.6	0.4	0
2	0.2	0.4	0.4

(c) Write out the general form of the optimality equations and solve this problem in steady state.

(d) Solve the optimality equations using the value iteration algorithm, starting with $V(s) = 0$ for $s = 0, 1,$ and 2. You may use a programming environment, but the problem can be solved in a spreadsheet. Run the algorithm for 20 iterations. Plot $V^n(s)$ for $s = 0, 1, 2$, and give the optimal action for each state at each iteration.

(e) Give a bound on the value function after each iteration.

Problem 4. Every day a salesman visits N customers in order to sell the R identical items he has in his van. Each customer is visited exactly once, and each customer buys zero or one item. Upon arrival at a customer location, the salesman quotes one of the prices $0 < p_1 \leq p_2 \leq ... \leq p_m$. Given that the quoted price is p_i, a customer buys an item with probability r_i. Naturally r_i is decreasing in i. The salesman is interested in maximizing the total expected revenue for the day. Show that if $r_i p_i$ is increasing in i, then it is always optimal to quote the highest price p_m.

Problem 5. You are trying to find the best parking space to use that minimizes the time needed to get to your restaurant. There are 50 parking spaces, and you see spaces 1,2,...50 in order. As you approach each parking space, you see whether it is full or empty. We assume, somewhat heroically, that the probability that each space is occupied follows an independent Bernoulli process, which is to say that each space will be occupied with probability p but will be free with probability $1 - p$, and that each outcome is independent of the other.
It takes 2 seconds to drive past each parking space and it takes 8 seconds to walk past. That is, if we park in space n, it will require $8(50 - n$ seconds to walk to the restaurant. Furthermore it would have taken you $2n$ seconds to get to this space. If you get to the last space without finding an opening, then you will have to drive into a special lot down the block, adding 30 seconds to your trip.
We want to find an optimal strategy for accepting or rejecting a parking space.

(a) Give the sets of state and action spaces and the set of decision epochs.

(b) Give the expected reward function for each time period and the expected terminal reward function.

(c) Give a formal statement of the objective function.

(d) Give the optimality equations for solving this problem.

(e) You have just looked at space 45, which was empty. There are five more spaces remaining (46 through 50). What should you do? Using $p = 0.6$, find the optimal policy by solving your optimality equations for parking spaces 46 through 50.

(f) Five the optimal value of the objective function in part (e) corresponding to your optimal solution.

Problem 6. You have invested R_0 dollars in a stock market that evolves according to the equation

$$R_t = \gamma R_{t-1} + \epsilon_t \tag{4}$$

where ϵ_t is a discrete, positive random variable that is independent and identically distributed and where $0 < \gamma < 1$. If you sell the stock at the end of period t, it will earn a risk-less return r until time T, which means it will evolve according to

$$R_t = (1 + r)R_{t-1} \tag{5}$$

You have to sell the stock, all on the same day, some time before T.

(a) Write a dynamic programming recursion to solve the problem.

(b) Show that there exists a point in the time τ such that it is optimal to sell for $t \geq \tau$, and optimal to hold for $t < \tau$.

(c) How does your answer to (b) change if you are allowed to sell only a portion of the assets in a given period? That is, if you have R_t dollars in your account, you are allowed to sell $a_t \leq R_t$ at time t.