B9140 Dynamic Programming and Reinforcement Learning Fall 2017

Prof. Daniel Russo September 11, 2017

Homework 1, Due in class Monday September 18

When formulating a problem and/or providing a dynamic programming recursion, make sure to clearly define the state space, action space, cost function, and state dynamics. When characterizing an optimal policy, make sure to clearly define the nature of the state that it takes as input and the action that it produces as output.

1 Deterministic Costs

In class, we formulated a problem where the cost incurred in period k, $g_k(x_k, u_k, w_k)$, is a function not only of the state x_k and control u_k but of the random disturbance w_k. Consider a modified MDP with the same transition dynamics

$$x_{k+1} = f_k(x_k, u_k, w_k) \quad k \in \{0, 1, \dots, N-1\}$$

but where costs incurred at stage k are a deterministic function of $\tilde{g}_k(x_k, u_k)$ of the state and control, defined by

$$\tilde{g}_k(x, u) = \mathbb{E}[g_k(x, u, w_k)] \quad \forall x \in \mathcal{X}_k,\ u \in U_k(x).$$

Show that the optimal cost–to–go function $J^*(x_0)$ and the optimal policy is the same for both problems. (*You may assume for simplicity that there is a unique optimal policy for the problem with random costs* $g_k(x_k, u_k, w_k)$.)

Solution

We will argue that every step of the DP algorithm is the same for each problem.
Precisely, the DP algorithm with random stage costs produces a sequence of cost-to-go functions $(J_0^*, \dots J_N^*)$ and an optimal policy $(\mu_0, \dots, \mu_N)$. The DP algorithm for the problem with deterministic stage costs produces a sequence of cost-to-go functions we will denote by $(\tilde{J}_0, \dots \tilde{J}_N)$ and an optimal policy we will denote by $(\tilde{\mu}_0, \dots, \tilde{\mu}_N)$. We'll show $J_k^* = \tilde{J}_k^*$ and $\mu_k = \tilde{\mu}_k$ for every k.

Working backward through time we have that for all $x \in \mathcal{X}_n$

$$J_N^*(x) = \min_{u \in U_N(x)} \mathbb{E}[g_N(x, u, w)] = \min_{u \in U_N(x)} \tilde{g}_N(x, u) = \tilde{J}_N^*(x)$$

and

$$\mu_N(x) = \arg\min_{u \in U(x)} \mathbb{E}[g_N(x, u, w)] = \arg\min_{u \in U(x)} \tilde{g}_N(x, u) = \tilde{\mu}_N(x)$$

Proceeding by induction, assume $J^*_{k+1} = \tilde{J}^*_{k+1}$. Then for all $x \in \mathcal{X}_k$

$$\begin{aligned} J^*_k(x) &= \min_{u \in U_k(x)} \mathbb{E}[g_k(x,u,w) + J^*_{k+1}(f_k(x,u,w))] \\ &= \min_{u \in U_k(x)} \tilde{g}_k(x,u) + \mathbb{E}[\tilde{J}^*_{k+1}(f_k(x,u,w))] = \tilde{J}^*_k(x). \end{aligned}$$

Therefore $J^*_k = \tilde{J}^*_k$ for $k = 0, 1, \ldots, N$. Similarly, we conclude

$$\begin{aligned} \mu_k(x) &= \arg\min_{u \in U_k(x)} \mathbb{E}[g_k(x,u,w) + J^*_{k+1}(f_k(x,u,w))] \\ &= \arg\min_{u \in U_k(x)} \tilde{g}_k(x,u) + \mathbb{E}[\tilde{J}^*_{k+1}(f_k(x,u,w))] = \tilde{\mu}_k(x). \end{aligned}$$

2 Optimal Sequential Search

Consider the problem of actively searching for the location of an unknown target $z^* \in [0,1]$. At each time k, we query a location $u_k \in [0,1]$ and are told whether z^* is smaller or larger than u_k. (We observe $\mathbf{1}\{z^* > u_k\}$) Based on these observations, we can construct increasingly refined intervals $[a_k, b_k] \subseteq [a_{k-1}, b_{k-1}] \subseteq ... \subseteq [0,1]$ that are guaranteed to contain z^*. In particular, $[a_1, b_1] = [0, u_0]$ if we observe that $z^* \leq u_0$ and is $[u_0, 1]$ otherwise.

We will use dynamic programming to study how to sequentially acquire information about z^* in an optimal manner. Assume the location of the target z^* is drawn uniformly at random from $[0,1]$. The objective is to sequentially choose the querry points $u_0, u_2, ...u_{N-1}$ to minimize $\mathbb{E}\left[\log(b_N - a_N)\right]$.

a) Formulate this problem as a finite horizon Markov decision process.

b) Solve for the optimal policy $\mu^*_{N-1}(a_{N-1}, b_{N-1})$ and cost-to-go function J^*_{N-1} at stage $N-1$. **Hint:** it is easier to work with the variable $p_k \equiv (u_k - a_k)/(b_k - a_k) \in [0,1]$

c) Prove that a myopic policy is optimal. That is, show $\mu^*_k = \mu^*_{N-1}$ for all k.

Solution

a) The state space $\mathcal{X} = \{[a,b] \in [0,1]^2 : a \leq b\}$ is the set of intervals in [0,1] and the set of controls is $\mathcal{U}([a_k, b_k]) = [a_k, b_k]$. The transition dynamics are defined by

$$[a_{k+1}, b_{k+1}] = \begin{cases} [u_k, b_k] & w/\ Prob.\ (b_k - u_k)/(b_k - a_k) \\ [a_k, u_k] & otherwise \end{cases}$$

b)

$$\begin{aligned} J_{N-1}([a,b]) &= \min_{p \in [0,1]} \{p \log(p|b-a|) + (1-p)\log((1-p)|b-a|)\} \\ &= \log(b-a) + \min_{p \in [0,1]} \{p\log(p) + (1-p)\log(1-p)\} = \log(b-a) + \log(1/2) \end{aligned}$$

The optimal choice is $p_{N-1} = 1/2$, or equivalently, the optimal policy is $\mu^*_{N-1}([a,b]) = (a+b)/2$. This can be seen by setting the derivative of $\{p\log(p) + (1-p)\log(1-p)\}$ to zero:

$$\frac{p}{p} - \frac{1-p}{1-p} + \log(p) - \log(1-p) = 0 \iff p = 1-p \iff p = 1/2.$$

c) We show by induction that for all k, $J_k([a_k, b_k]) = \log(b_k - a_k) + c_k$ where c_k is a constant. Since c_k doesn't affect our choice of action, it follows immediately that a myopic policy is optimal.

Base case: $J_N([a,b]) \overset{Def}{=} \log(b-a) \implies c_N = 0$

Induction step: Suppose the claim holds for period $k+1$. Then proceeding as in part (b)

$$\begin{aligned} J_k([a,b]) &= \log(b-a) + \underbrace{\min_{p\in[0,1]} \{p\log(p) + (1-p)\log(1-p)\} + c_{k+1}}_{c_k} \\ &= \log(b-a) + c_k \end{aligned}$$

Comments for those familiar with information theory

Knowledge of information theory is not needed, or necessarily even helpful for solving this problem, but it does provide an interesting interpretation of this result.

Based on our first N observations, we can update the prior distribution of z^* to a posterior distribution, which is $\mathrm{Uniform}(\mathrm{a_N}, \mathrm{b_N})$. The terminal cost $\mathbb{E}\left[\log(b_N - a_N)\right]$ is the differential entropy of the posterior distribution of z^*, a common measure of our uncertainty about z^*. This result shows a myopic policy, which always samples a point that leads to the largest expected reduction in the entropy of z^* in *this period*, is in fact optimal for the multi-period problem. This result can be generalized to situations where z^* is drawn from any continuous distribution on [0,1]. More can be learned from the recent papers (Bisection Search with Noisy Responses R. Waeber, P.I. Frazier & S.G. Henderson, 2013) and (Twenty Questions with Noise: Bayes Optimal Policies for Entropy Loss B. Jedynak, P.I. Frazier & R. Sznitman, 2012).

3 Optimal Stopping

This problem asks you to solve problem 3.19 of Bertsekas Vol. 1. The solution from the textbook author is provided in a separate document. The solutions are from an earlier version of the text, and therefore label this as problem 4.19.