B9120-001 Dynamic Programming Spring 2020

Prof. Daniel Russo

Homework Assignment 5 Solutions

Problem 1.8: The Restart Problem

Here we show that the index $m(x_0)$ of a state x_0 can be calculated by solving an infinite horizon discounted cost problem.

The restart problem: At state x_k at time k there are two options: Continue, which brings reward $\alpha^k R(x_k)$ and moves the state to $x_{k+1} = f(x_k, w)$, or (2) restart the project, which moves the state to x_0, brings reward $\alpha^k R(x_0)$ and then moves the project to state $f(x_0, w)$. The optimal cost-to-go fucntion satisfies Bellman's equation

$$J(x) = \max\left\{R(x_0) + \alpha E\left[J(f(x_0, w))\right], R(x) + \alpha E\left[J(f(x, w))\right]\right\} \tag{1}$$

The bandit problem: Now consider a multi-armed bandit problem with a single project. When this project is in state x_k and is worked on, it generates reward $\alpha^k R(x_k)$ and transitions to a new state $x_{k+1} = f(x_k, w)$.

Let the retirement reward be

$$M = m(x_0) \stackrel{\text{Def}}{=} \min\{\bar{M} : \tilde{J}\left(x_0, \bar{M}\right) = \bar{M}\}$$

where $\tilde{J}(x, \bar{M})$ denote the optimal reward attainable when the initial state is x and the retirement reward is $\bar{M}$.

Using $J(x) = \tilde{J}\left(x, M\right)$, the optimal cost to go for this problem satisfies Bellman's equation

$$J(x) = \max\left\{M, R(x) + \alpha E[J(f(x, w))]\right\}$$

Using that $M = R(x_0) + \alpha E[J(f(x_0, w))$, we conclude

$$J(x) = \max\left\{R(x_0) + \alpha E[J(f(x_0, w)), R(x) + \alpha E[J(f(x, w))]\right\},$$

which corresponds to (1).

Coin Tossing

We can treat this problem as a multi-armed bandit problem and apply the Gittins index theorem.

The state of coin i is described by a pair (α_i, β_i) where α_i is the number of heads observed so far, and β_i is number of tails observed so far. The state space of coin i is

$$\mathcal{X}_i = \{(\alpha, \beta) \in \{0, 1, 2, ...100\}^2 : \alpha + \beta \leq 100\}$$

and the state space of the MDP is $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times ... \times \mathcal{X}_{10}$.

That this problem can be modeled as a multi-armed bandit problem relies on the fact that (i) only the state of the coin that is tossed evolves and (ii) the probability of state transitions on that coin don't depend on the states of other coins.

Rewards functions and Transition Probabilities:

Given we have observed α heads and β tails from a coin, we can associate with that coin a posterior probability the coin has bias $2/3$ of

$$\pi(\alpha, \beta) := \mathbb{P}\left(\text{the coin's bias is } \frac{2}{3} \middle| \ \alpha, \beta\right)$$

which can be computed in explicitly by applying Bayes rule.

Now, let's use this to define state transition probabilities. Let

$$P\left((\alpha, \beta), (\alpha', \beta')\right)$$

denote the probability a coin transitions from state (α, β) to another state (α', β') when it's flipped. Clearly, we assign zero probabilities to all but two possible transitions (we observe heads or tails). We can write

$$p\left((\alpha, \beta), (\alpha + 1, \beta)\right) = \frac{2}{3}\pi(\alpha, \beta) + \frac{1}{3}\left(1 - \pi(\alpha, \beta)\right) = 1 - p\left((\alpha, \beta), (\alpha, \beta + 1)\right) \qquad \text{if } \alpha + \beta < 100$$

When a coin has been flipped 100 times, we say that it if selected, it stays in the same state and always gives a payoff of 0. (Essentially it cannot be flipped any more)

We associate a reward of 1 with a heads and a reward of 0 with tails, so

$$R\left((\alpha, \beta), (\alpha + 1, \beta)\right) = 1 \qquad R\left((\alpha, \beta), (\alpha, \beta + 1)\right) = 0.$$

Gittin's Index Solution: We now have a well defined multi-armed problem. An optimal policy is to play the coin with the highest gittin's index. In particular, when in the state is $x = (x_1, ..., x_{10})$, where $x_i = (\alpha_i, \beta_i)$ denotes the state of coin i, we flip a coin j where

$$\underbrace{m(x_j)}_{\substack{\text{Gittins index} \\ \text{of state } x_k}} = \max_i m(x_i).$$