5.2 Policy Evaluation by Monte Carlo Simulation

Thursday, February 11, 2021 2:33 PM

We wish to calculate the cost-to-go vector J^{μ} by simulation.

Notation is eased off, p_{ij} and $g(i,j)$ in place of $p_{ij}(\mu(i))$ and $g(i,\mu(i),j)$.

Consider the mth time a given state i_0 is encountered with $(i_0, i_1, \dots i_N)$ being the remainder of trajectory with $i_N = 0$.

Let $c(i_0, m)$ be the cumulative cost,

$$c(i_0,m) = g(i_0,i_1) + \dots g(i_{N-1}, i_N).$$

For all states, $J^{\mu}(i) = E[c(i,m)]$.

$$\Rightarrow J(i) = \frac{1}{K}\sum_{m=1}^{K} c(i,m).$$

We can iteratively calculate sample means,

$$J(i) = J(i) + \gamma_m\left[c(i,m) - J(i)\right], \quad m = 1,2,\dots$$

where $\gamma_m = \frac{1}{m}$ starting with $J(i) = 0$.

Thus, for each $k = 0,1,\dots N-1$,

$$J(i_k) = J(i_k) + \gamma(i_k)\left[g(i_k,i_{k+1}) + g(i_{k+1},i_{k+2}) + \dots g(i_{N-1},i_N) - J(i_k)\right].$$

<u>The Every-Visit Method</u>.

Consider state i which is encountered infinitely many times in the long run and is updated every time upon its visit. Since $K \to \infty$, $K_i \to \infty$,

The sample mean of all available cost samples $c(i,m,k)$ is given, asymptotically, by

$$\lim_{K\to\infty} \frac{\sum_{\{k|n_k\ge 1\}}\sum_{m=1}^{n_k} c(i,m,k)}{\sum_{\{k|n_k \ge 1\}} n_k} = \lim_{K_i\to\infty} \frac{\frac{1}{K_i}\sum_{\{k|n_k\ge1\}}\sum_{m=1}^{n_k} c(i,m,k)}{\frac{1}{K_i}\sum_{\{k|n_k\ge 1\}} n_k}.$$

$$= \frac{E\left[\sum_{m=1}^{n_k} c(i,m,k) \mid n_k \ge 1\right]}{E[n_k \mid n_k \ge 1]}.$$

Note that $E[c(i,m,k) \mid n_k \ge m] = J^{\mu}(i)$ is a consequence of Markov's property. Using Wald's Identity,

$$\frac{E\left[\sum_{m=1}^{n_k} c(i,m,k) \mid n_k \ge 1\right]}{E[n_k \mid n_k \ge 1]} = E\left[c(i,1,k) \mid n_k \ge 1\right] = J^{\mu}(i).$$

<u>The First-Visit Method</u>.

Every-visit method results in a biased estimator when the number of samples are finite.

Alternatively, we can use only the cost sample $c(i,1,k)$ corresponding to the first visit to state i.

This yields,
$$\frac{\sum_{\{k|n_k\ge 1\}} c(i,1,k)}{K_i}.$$

<u>5.2.2 Q-Factors and Policy Iteration</u>.

$$Q^{\mu}(i,u) = \sum_{j=0}^{n} p_{ij}(u)\left(g(i,u,j) + J^{\mu}(j)\right).$$

"Expected cost of starting in state 'i', using control 'u' and following policy 'μ' for subsequent stages."

Policy improvement can be executed as follows,

$$\bar{\mu}(i) = \arg\min_{u\in U(i)} Q^{\mu}(i,u). \quad , i = 1,2,\dots n$$

x ———————— x ———————— x