UC Berkeley
Department of Electrical Engineering and Computer Sciences

Electrical Engineering 126: Probability and Random Processes

Discussion 13
Fall 2017

---

1. **Dynamic Programming**

   A decision problem is characterized by (state, action, noise) (x_k, u_k, w_k) for each positive integer k. The dynamics is governed by: $x_{k+1} = f_k(x_k, u_k, w_k)$, where f_k is some bounded function. Consider N (a positive integer) to be the horizon and the controller wants to minimize a cost function $g_k(x_k, u_k, w_k)$, which is additive over the discrete time step. The terminal cost $g_N(x_N)$ is given. Formulate the problem as a dynamic program and find the total expected cost. Define a policy sequence $\mu = (\mu_0, \ldots, \mu_{N-1})$ as a mapping from state space to action space, i.e., $u_k = \mu_k(x_k)$. Find optimal policy as a function of total expected cost. Also state the dynamic programming update equations for the k-th iteration, using the principle of optimality.

   Solution:

   The total cost is $\mathbb{E}[g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, u_k, w_k)]$. If x_0 is the start state, and $J_\pi(x_0)$ denotes the cost starting from state x_0 of using policy π, then $J_\pi(x_0) = \mathbb{E}[g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), w_k)]$. The optimal cost starting from state x_0 is $J_{\pi^*}(x_0) = \min_{\pi \in \Pi} J_\pi(x_0)$.

   Principle of Optimality: Let $\pi^* = (\mu_0^*, \ldots, \mu_{N-1}^*)$ be the optimal policy. At time $i \in \{1, \ldots, N-1\}$, we are at state x_i. We want to minimize the cost-to-go from that state, $J_{\pi^*}(x_i) = \mathbb{E}[g_N(x_N) + \sum_{k=i}^{N-1} g_k(x_k, \mu_k(x_k), w_k)]$. The same policy is optimal for the subproblem from iterations i through N: $(\mu_i^*, \mu_{i+1}^*, \ldots, \mu_{N-1}^*)$.

   Principle of Optimality in DP Algorithms: Starting at x_0, let $J^*(x_0)$ denote the optimal cost. Let $J^*(x_k)$ be the optimal cost-to-go from state x_k. Then, for $k = 0, 1, \ldots, N-1$,

   $$J_k^*(x_k) = \min_{u_k} \mathbb{E}\big[g_k(x_k, u_k, w_k) + J_{k+1}^*\big(f_k(x_k, u_k, w_k)\big)\big].$$

   Note that $f_k(x_k, u_k, w_k) = x_{k+1}$, the next state.

   Remark: The equation says to minimize over the action space over the expected current cost and the cost-to-go from the next state, which is intuitive.

   Connection to Discounted Cost Problem: Suppose that the objective function is now $\mathbb{E}[g_N(x_N) + \sum_{k=1}^{N-1} \alpha^k g_k(x_k, u_k, w_k)]$, where $\alpha \in (0, 1)$ is the discount factor. Then, the DP iteration will be

   $$J_k^*(x_k) = \min_{u_k} \mathbb{E}\big[g_k(x_k, u_k, w_k) + \alpha J_{k+1}^*\big(f_k(x_k, u_k, w_k)\big)\big].$$

2. **Infinite-Horizon Discounted Cost MDP**

Consider a relay placement problem, where a deployment agent starts walking from state 0 on a line. He stops at regular intervals (consider the step length to be $\delta > 0$), and decides whether to place a relay there or not. At each step, he measures the power required to maintain a reasonable quality link. The goal of the agent is to minimize a linear combination of power cost and relay cost. Assume that the process restarts every time the agent deploys a relay. Also assume the length of the line is geometric with parameter θ. Formulate the problem as an infinite horizon MDP, with state space (r, γ) where r and γ are the respective distance and power from the previously placed relay. Using the Bellman equation, find the optimal policy structure. Mention a way to compute the optimal policy.

Solution:

For each positive integer i, let $\Gamma^{(i,i-1)}$ be the power (a random variable) between the ith and $(i-1)$th relay. Also, assume that ξ is the cost of each relay and there are N relays placed. N is a random variable since the line ends at random with probability θ at each step (geometric).

Objective: $\min_{\pi \in \Pi} \mathbb{E}_\pi[\sum_{i=1}^{N+1} \Gamma^{(i,i-1)} + \xi N]$.

For infinite-horizon problems we use the **Bellman equations**, which is just the **principle of optimality equation**.

- Here, the state space consists of tuples (r, γ) and the action space is $\{\text{place}, \text{do not place}\}$.
- The process *restarts* every time a relay is *placed*.

Equations: 0 is the start state. $J^*(r, \gamma) = \min\{C_\text{p}(r, \gamma), C_\text{np}(r, \gamma)\}$, where

$$C_\text{p}(r, \gamma) = \underbrace{\xi}_{\text{relay cost}} + \underbrace{\gamma}_{\text{power cost}} + \underbrace{J^*(0)}_{\text{restarts, so the next state is 0}}$$

and

$$C_\text{np}(r, \gamma) = \underbrace{\theta\,\mathbb{E}[\Gamma_{r+1}]}_{\substack{\theta \text{ probability that the line ends} \\ \mathbb{E}[\Gamma_{r+1}] \text{ is the terminal cost}}} + \underbrace{(1-\theta)}_{\text{line does not end}} \cdot \underbrace{\mathbb{E}[J^*(r+1, \Gamma_{r+1})]}_{\text{cost-to-go from next state}}.$$

The total cost is $J^*(0) = \theta\,\mathbb{E}[\Gamma_1] + (1-\theta)\,\mathbb{E}[J^*(1, \Gamma_1)]$, since at 0 we do not place the relay. So, $J^*(0) = C_\text{np}(0)$.

Optimal Policy: At state (r, γ), the agent places the relay if

$$C_\text{p}(r, \gamma) \le C_\text{np}(r, \gamma),$$

i.e., if $\xi + \gamma + J^*(0) \le \theta\,\mathbb{E}[\Gamma_{r+1}] + (1-\theta)\,\mathbb{E}[J^*(r+1, \Gamma_{r+1})]$. Define γ_th to be the value of γ which solves $\gamma = \theta\,\mathbb{E}[\Gamma_{r+1}] + (1-\theta)\,\mathbb{E}[J^*(r+1, \Gamma_{r+1})] - \xi - J^*(0)$. Then, the policy is to place a relay if $\gamma \le \gamma_\text{th}$, so the optimal policy is a **threshold policy**. This policy can be computed via **value iteration**.