EE 223: Stochastic Estimation and Control **Spring 2007**

Lecture 4 — January 25

Lecturer: Venkat Anantharam *Scribe: Yusik Kim*

For the finite horizon, fully observed problem where we have

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

and $x_0, w_0, w_1, ...w_{N-1}$ independent, our objective is to minimize

$$E[\sum_{k=0}^{N-1} g_k(x_k, u_k, w_k) + g_N(x_N)]$$

over all admissible strategies. To solve the problem, we define functions recursively from the end:

$$J_N(x_N) = g_N(x_N)$$
$$J_k(x_k) = \inf_{u \in \{\text{allowed controls at } x_k\}} E[g_k(x_k, u, w_k) + J_{k+1}(f_k(x_k, u, w_k))], k = 0, 1, ..., N-1$$

The main result is that if $\mu_k(x_k), k = 0, 1, .., N-1$ are any choice of minimizing functions in the respective definitions (of $J_k(x_k)$ at time k), then the Markov strategy $\pi = (\mu_0, \mu_1, ..., \mu_{N-1})$ is optimal. However in this theorem, there is no guarantee that minimizers actually exist. This point is illustrated in the following example:

Example: Non-existence of minimizers
Assume the following:

1. $N = 1$
2. State space at times 0 and 1 is $\{0\}$
3. Control space at time 0 when in state 0 is $\{1, 2, 3, ...\} := \mathbb{N}$
4. Terminal cost $g_1(0) = 0$
5. $g_0(0, u) = \frac{1}{u}, u \in \mathbb{N}$

Note that $P(X_1 = 0 | X_0 = 0, u_0 = u) = 1$. Here the procedure for finding an optimal strategy suggested by the theorem fails. To see this, $J_1(0) = 0$, $J_0(0) = \inf_{u \in \mathbb{N}}[\frac{1}{u} + 0] = 0$, but there is no minimizer.

We will spend the next few lectures exploring some examples of what the dynamic programming framework buys you.

Example: Secretary Problem
You are interviewing M secretaries and wish to pick the best (reward 1). If you pick any non-best secretary, you get reward 0. However, the interviews take place in sequence and you have to decide whether or not to pick the current interviewee irrevocably immediately after the interview. All relative orders of the quality are equally probable. Some thought reveals that this problem fits in our general stochastic control framework (finite horizon, fully observed), with a horizon of $N = M + 1$ as follows. (Since the problem involves rewards rather than costs, we will treat it as such, and where we had infimum in the definition of the optimal cost-to-go functions, we will now have a supremum in optimal reward-to-go functions, for which we will continue to use the notation $J_k(x_k)$, $k = 0, 1, \ldots, N$.) In the following we are visualizing secretaries as continuing to appear for interview even though a choice might already have been made.

The state space at time 0 is taken to be 0.

At any time $k = 1, \ldots, N-1$ the state space is taken to be the set

$$\{(\Delta, k, 1), (\Delta, k, 0), (k, 1), (k, 0)\} .$$

The state $(\Delta, k, 1)$ means that a decision was already made to pick one of the earlier secretaries, the current time is k (i.e. the k-th secretary has currently entered the interview process) and that the secretary that was picked earlier is the best one of all the secretaries so far.

The state $(\Delta, k, 0)$ means that a decision was already made to pick one of the earlier secretaries, the current time is k and the secretary that was picked earlier is not the best one of all the secretaries so far.

The state $(k, 1)$ means that a decision has not yet been made to pick one of the secretaries seen so far, the current time is k and the current secretary, (i.e. the k-th secretary) is the best one of all the secretaries so far.

The state $(k, 0)$ means that a decision has not yet been made to pick one of the secretaries seen so far, the current time is k and the current secretary is not the best one of all the secretaries so far.

At time N the state space is taken to be T. This represents a terminal state.

Now define independent random variables $w_0, ..., w_{N-2}, w_{N-1}$ where

$$\begin{aligned} P(w_k = 1) &= \frac{1}{k+1} \\ P(w_k = 0) &= \frac{k}{k+1}, \quad k = 0, 1, ..., N-2 \\ P(w_{N-1} = 0) &= 1. \end{aligned}$$

At each state at each time $k = 0, 1, \ldots, N-1$ there are two control actions available $u = 0$ denotes the choice to pick the current secretary. $u = 1$ denotes the choice to not pick the current secretary. In "Δ" states the choice of control action is irrelevant, but we may as well pretend there is a choice, for aesthetic reasons.

The state evolves as $x_{k+1} = f_k(x_k, u_k, w_k)$ for suitable f_k. These are described next.

At $k = 0$, the choice to "pick the current secretary" ($u = 0$) leads to the state $(\Delta, 1, 0)$ with probability 1. In effect the choice has been made to not engage in interviewing secretaries.

At $k = 0$ the choice $u = 1$ leads to $(1, 1)$ with probability 1.

(Note that both control choices result in deterministic transitions, so they can be expressed in terms of the deterministic variable w_0.)

For each $k = 1, \ldots, N-2$, from state $(\Delta, k, 0)$ both control actions $u = 0, 1$ lead deterministically to state $(\Delta, k+1, 0)$. Indeed, once it is the case that the secretary picked is not the best among those seen so far, the secretary can never again be best. Also, once one of the "Δ states" has been entered (once a secretary has been picked) the "Δ-states" cannot be left.

For each $k = 1, \ldots, N-2$, from state $(\Delta, k, 1)$ both control actions $u = 0, 1$ lead to state $(\Delta, k+1, 1)$ with probability $\frac{k}{k+1}$ and to state $(\Delta, k+1, 0)$ with probability $\frac{1}{k+1}$. Note that this transition can be described via a deterministic function, using the random variable w_k.

For each $k = 1, \ldots, N-2$, from state $(k, 1)$ the control action $u = 0$ leads to state $(\Delta, k+1, 1)$ with probability $\frac{k}{k+1}$ and to state $(\Delta, k+1, 0)$ with probability $\frac{1}{k+1}$. This transition can also be described via a deterministic function, using the random variable w_k.

For each $k = 1, \ldots, N-2$, from state $(k, 0)$ the control action $u = 0$ leads to the state $(\Delta, k+1, 0)$ with probability 1.

For each $k = 1, \ldots, N-2$, from both states $(k, 1)$ and $(k, 0)$ the control action $u = 1$ leads to state $(k+1, 1)$ with probability $\frac{1}{k+1}$ and to state $(k+1, 0)$ with probability $\frac{k}{k+1}$. This transition can also be described via a deterministic function, using the random variable w_k.

Finally, for $k = N-1$ both controls, from any state, lead to the terminal state T. These transitions, being deterministic, can be expressed in terms of the deterministic variable w_{N-1}.

We take the *reward* at the terminal state to be $g_N(T) = 0$. We also have that all $g_k(x_k, u_k, w_k) = 0$ except for 2 cases:

$g_{N-1}((N-1, 1), 0, 0) = 1$ (we have avoided choosing a secretary all the way down to the last secretary and this one turns out to be the best, hence the best overall, and we make the decision to pick this secretary, so we get a reward of 1)

$g_{N-1}((\Delta, N-1, 1), 0/1, 0) = 1$ (we picked a secretary before time $N-1$, but it turns

out at time $N-1$ that this secretary is the best of all so far, hence is the absolute best, so again we realize the reward of 1 from having picked this secretary)

We next find the optimal reward-to-go functions. We have

$$J_N(T) = 0 \ .$$

We observe that the optimal reward-to-go functions can be explicitly computed for all the "Δ" states since there is no control choice out of such states. For instance, we would have

$$J_{N-1}(\Delta, N-1, 1) = 1 \text{ and } J_{N-1}(\Delta, N-1, 0) = 0 \ .$$

We have

$$\begin{aligned} J_k(\Delta, k, 1) &= \frac{k}{k+1} J_{k+1}(\Delta, k+1, 1) \\ &= \frac{k}{k+1}\frac{k+1}{k+2}\cdots\frac{N-2}{N-1} J_{N-1}(\Delta, N-1, 1) \\ &= \frac{k}{M} \ , \end{aligned}$$

and

$$J_k(\Delta, k, 0) = 0 \ .$$

We may now turn to the more interesting issue of determining the optimal-cost-to-go (and the associated optimizers) in the non-Δ states. We see, for instance that

$$J_{N-1}(N-1, 1) = \max(1+0, 0+0) = 1 \ ,$$

where the first choice corresponds to the (smart) action of picking the current secretary and the second choice to the (dumb) action of rejecting this secretary. The optimal control in state $(N-1, 1)$ is clearly to pick the current secretary.

Likewise, we see that

$$J_{N-1}(N-1, 0) = \max(0+0, 0+0) = 0 \ ,$$

so either control is optimal in this state.

More generally, we have, for $k = 1, 2, \ldots, N-2$,

$$J_k(k, 1) = \max(\frac{k}{M}, \frac{1}{k+1} J_{k+1}(k+1, 1) + \frac{k}{k+1} J_{k+1}(k, 0))$$

and

$$\begin{aligned} J_k(k, 0) &= \max(0, \frac{1}{k+1} J_{k+1}(k+1, 1) + \frac{k}{k+1} J_{k+1}(k+1, 0)) \\ &= \frac{1}{k+1} J_{k+1}(k+1, 1) + \frac{k}{k+1} J_{k+1}(k+1, 0) \end{aligned}$$

while

$$J_0(0) = \max(0, J_1(1, 1)) = J_1(1, 1) \ .$$

We will analyze these equations next time to learn some nice properties of the optimal strategy in the secretary problem.