EE 223: Stochastic Estimation and Control	**Spring 2007**
Lecture 3 — January 23	
Lecturer: Venkat Anantharam	*Scribe: Nikhil Shetty*

3.1 Outline

In this lecture, we argue that the optimal strategy in the finite horizon fully observed stochastic control problem we have been considering so far is a Markov strategy. In addition, we define the optimal cost-to-go functions associated to the problem and sketch a proof of the theorem that says that optimal Markov strategies must be defined in terms of the minimizers in the equations defining these optimal cost-to-go functions. We do not concern ourselves with the question of existence of optimal strategies, rather, we discuss their nature, assuming they exist.

3.2 Assumptions and Definitions

The evolution equation of our dynamical system is:

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

where x_k is the state of the system at time k, u_k is the input applied at time k and w_k is a random variable. The functions f_k determine the system transition at time k. x_0 may be fixed or random.
At time k the controller has access to the state trajectory up to time k, i.e. $(x_0, \ldots, x_k)$. The aim of the controller is to choose the controls causally to minimize the objective function, written informally as:

$$E[\sum_{k=0}^{N-1} g_k(X_k, U_k, w_k) + g_N(X_N)]$$

where $g_k(x_k, u_k, w_k)$, $0 \leq k \leq N-1$, and $g_N(x_N)$ are some prescribed cost functions.

Formally, a strategy $\Psi = (\nu_0, \nu_1, \ldots, \nu_{N-1})$ is comprised of functions $\nu_k(x_0, x_1, \ldots, x_k)$ which take values in the allowed set of controls at state x_k for $0 \leq k \leq N-1$. Our aim is to minimize the following objective function over all strategies Ψ,

$$E[\sum_{k=0}^{N-1} g_k(X_k^\Psi, \nu_k(X_0^\Psi, X_1^\Psi, \ldots, X_k^\Psi), w_k) + g_N(X_N^\Psi)]$$

where $X_{k+1}^\Psi = f_k(X_k^\Psi, \nu_k(X_0^\Psi, X_1^\Psi, \ldots, X_k^\Psi), w_k) \quad k = 0, \ldots, N-1$.

We have $X_0 = X_0^\Psi$ for all Ψ. The $w_i's$ are assumed to be independent of each other and of X_0. The ν_k could also be randomized but we do not consider that here.

Markov strategy: A Markov strategy $\pi = (\mu_0, \ldots, \mu_{N-1})$ is prescribed by functions $\mu_k(x_k)$ taking values in the set of allowed controls at state x_k, $\quad k = 0, \ldots, N-1$..

Define functions $J_0, J_1, \ldots, J_N$ by backwards recursion as follows with $J_N(x_N) = g_N(x_N)$ and

$$J_k(x_k) = \overset{\inf}{u}\ E[g_k(x_k, u, w_k) + J_{k+1}(f_k(x_k, u, w_k))]$$

where in the infimum u ranges over the allowed control choices at state x_k. These are called the *optimal cost-to-go* functions. Note that these equations are defining *functions*. J_k is a function on the set of possible states at time k, defined in terms of the function J_{k+1}, which has already been defined in the backwards recursion procedure.

3.3 Theorems

1. Let $\mu_k(x_k)$ be any minimizer in the equation defining $J_k(x_k)$. (Note that this sentence is talking about a *function* $\mu_k(x_k)$, i.e. we want the control $\mu_k(x_k)$ to minimize in the expression defining $J_k(x_k)$ for *each* x_k.) Then consider the Markov strategy $\pi = (\mu_0, \ldots, \mu_{N-1})$. This is an optimal strategy.

2. For any Markov strategy $\pi = (\mu_0, \ldots, \mu_{N-1})$ that is optimal, $\mu_k(X_k^\pi)$ must be a minimizer in the definition of $J_k(X_k^\pi)$, almost surely, for each $k = 0, \ldots, N-1$.

3.3.1 Cost-to-go function

Given any strategy $\Psi = (\nu_0, \nu_1, \ldots, \nu_{N-1})$, we define for each $k = 0, \ldots, N-1$, the cost-to-go function of the strategy at time k, denoted Γ_k^Ψ. This is a random variable depending on $(X_0^\Psi, X_1^\Psi, \ldots, X_k^\Psi)$. It is defined via:

$$\Gamma_N^\Psi = g_N(X_N^\Psi) = E[g_N(X_N^\Psi) | X_0^\Psi, X_1^\Psi, \ldots, X_k^\Psi] \ ,$$

and

$$\Gamma_k^\Psi = E[\sum_{l=k}^{N-1} g_l(X_l^\Psi, \nu_l(X_0^\Psi, X_1^\Psi, \ldots, X_l^\Psi), w_l) + g_N(X_N^\Psi) | X_0^\Psi, X_1^\Psi, \ldots, X_k^\Psi]$$

The expected total cost of the strategy Ψ is thus seen to be given by $E[\Gamma_0^\Psi]$. (Note that Γ_0^Ψ is a random variable depending upon X_0.)

3.3.2 Comparison Principle

The main theorem is actually a consequence of the following "comparison principle" which says

Let $V_k(x_k), \quad k = 0, \ldots, N$, be any functions satisfying $V_N(x_N) \leq g_N(x_N)$ and

$$V_k(x_k) \leq \overset{\inf}{u} E[g_k(x_k, u, w_k) + V_{k+1}(f_k(x_k, u, w_k))]$$

where the infimum is taken over u ranging over the allowed control values at state x_k. Then, for any strategy Ψ, for every $k = 0, \ldots, N$, $V_k(X_k^\Psi) \leq \Gamma_k^\Psi$, almost surely.

Proof (by backward induction)

By definition of $V_N(x_N)$,
$V_N(X_N^\Psi) \leq g_N(X_N^\Psi) = \Gamma_N^\Psi$
Thus, the claim is true at time N. Assuming it is true at time $k+1$, we need to show that it is true at time k. By the definition of $V_k(x_k)$,

$$V_k(X_k^\Psi) \leq E[g_k(X_k^\Psi, \nu_k(X_0^\Psi, X_1^\Psi, \ldots, X_k^\Psi), w_k) +$$
$$V_{k+1}(f_k(X_k^\Psi, \nu_k(X_0^\Psi, X_1^\Psi, \ldots, X_k^\Psi), w_k)) | X_0^\Psi, X_1^\Psi, \ldots, X_k^\Psi]$$

Understanding this equation requires some thought. On the right hand side of this equation, we have an expression conditioned on $(X_0^\Psi, X_1^\Psi, \ldots, X_k^\Psi)$. The conditioning determines both the state X_k^Ψ and the control $\nu_k(X_0^\Psi, X_1^\Psi, \ldots, X_k^\Psi)$ that appear in the expression that is being conditioned. Now, appealing to the independence of w_k from the random variables that are being conditioned on, and appealing to the definition of $V_k(x_k)$ (taking X_k^Ψ for x_k and noting that this definition involves an infimum over *all* allowed controls at this state) the truth of this equation becomes apparent.

Also, $V_{k+1}(f_k(X_k^\Psi, \nu_k(X_0^\Psi, X_1^\Psi, \ldots, X_k^\Psi), w_k)) \leq \Gamma_{k+1}^\Psi$ by the inductive hypothesis.

Hence, $V_k(X_k^\Psi) \leq E[g_k(X_k^\Psi, \nu_k(X_0^\Psi, X_1^\Psi, \ldots, X_k^\Psi), w_k) + E[E[\sum_{l=k+1}^{N-1} g_l(X_l^\Psi, \nu_l(X_0^\Psi, X_1^\Psi, \ldots, X_l^\Psi), w_l) + g_N(X_N^\Psi) | X_0^\Psi, X_1^\Psi, \ldots, X_{k+1}^\Psi] | X_0^\Psi, X_1^\Psi, \ldots, X_k^\Psi] = \Gamma_k^\Psi$

3.3.3 Proof of Part 1 of theorem

We observe that $J_k(x_k)$ for $\quad k = 0, \ldots, N$ satisfy the inequalities that apply to the $V_k(x_k)$ for $\quad k = 0, \ldots, N$ in the comparison principle (in fact with equality). So $J_k(X_k^\Psi) \leq \Gamma_k^\Psi$ almost surely for all Ψ and all $k = 0, \ldots, N$. In particular, $J_0(X_0) = J_0(X_k^\Psi) \leq \Gamma_0^\Psi$ almost surely for all Ψ. Taking expectations in this equation we get $E[J_0(X_0)] \leq E[\Gamma_0^\Psi]$ for all Ψ. Thus we have found a *universal* lower bound on the expected total cost achievable by *any*

strategy. Thus, any strategy whose expected total cost equals this universal lower bound would be an optimal strategy.

To prove part 1 of the main theorem, let $\mu_0(x_0), \ldots, \mu_{N-1}(x_{N-1})$ be minimizers in the definition of $J_k(x_k)$ and let $\pi = (\mu_0(x_0), \ldots, \mu_{N-1}(x_{N-1}))$ be the corresponding Markov strategy. If you trace through the proof of comparison principle (replacing Ψ by π and each V_k by J_k), all inequalities become equalities. Let us do this explicitly. At time N we have $J_N(X_N^\pi) = g_N(X_N^\pi) = \Gamma_N^\pi$, where both equalities are by definition. Assuming it is true that $J_{k+1}(X_{k+1}^\pi) = \Gamma_{k+1}^\pi$ almost surely. we will show that $J_k(X_k^\pi) = \Gamma_k^\pi$ almost surely. Because $\mu_k(X_k^\pi)$ is a minimizer in the definition of $J_k(X_k^\pi)$ we have

$$J_k(X_k^\pi) = E[g_k(X_k^\pi, \mu_k(X_k^\pi), w_k) + J_{k+1}(f_k(X_k^\pi, \mu_k(X_k^\pi), w_k))|X_0^\pi, X_1^\pi, \ldots, X_k^\pi]$$

Since $X_{k+1}^\pi = f_k(X_k^\pi, \mu_k(X_k^\pi), w_k)$ and we have the inductive hypothesis, this equation can be written as

$$J_k(X_k^\pi) = E[g_k(X_k^\pi, \mu_k(X_k^\pi), w_k) + \Gamma_{k+1}^\pi|X_0^\pi, X_1^\pi, \ldots, X_k^\pi] \ .$$

However, the right hand side is just the definition of Γ_k^π, which completes the proof. (Note that in the case of Markov strategies the cost-to-go at any time is just a function of the current state, and this is also the case for the specific π that is being considered here.)

In particular, $\Gamma_0^\pi = J_0(X_0^\pi) = J_0(X_0)$ almost surely, and taking expectations we get $E[\Gamma_0^\pi] = E[J_0(X_0)]$, i.e. this Markov strategy π achieves the universal lower bound on the total expected cost, so it is an optimal strategy.

Please note that in this proof the existence of π was *hypothesized* and not proved. Indeed, such π may not exist, unless suitable technical assumptions are made. There is a rich literature full of existence theorems under various sorts of conditions.

3.3.4 Sketch of proof of Part 2 of theorem

Suppose $\pi = (\mu_0(x_0), \ldots, \mu_{N-1}(x_{N-1}))$ is an optimal Markov strategy. We prove by backward induction that $\mu_k(X_k^\pi)$ is a minimizer in the definition of $J_k(X_k^\pi)$ (almost surely).
Consider $k = N-1$. Suppose there is some function $\mu_{N-1}^{'}(x_{N-1})$ different from the function $\mu_{N-1}(x_{N-1})$ such that

$$E[g_{N-1}(X_{N-1}^\pi, \mu_{N-1}(X_{N-1}^\pi), w_{N-1}) + J_N(f_{N-1}(X_{N-1}^\pi, \mu_{N-1}(X_{N-1}^\pi), w_{N-1}))|X_{N-1}^\pi] \geq$$

$$E[g_{N-1}(X_{N-1}^\pi, \mu_{N-1}^{'}(X_{N-1}^\pi), w_{N-1}) + J_N(f_{N-1}(X_{N-1}^\pi, \mu_{N-1}^{'}(X_{N-1}^\pi), w_{N-1}))|X_{N-1}^\pi]$$

and that this inequality holds strictly with positive probability. (If $\mu_{N-1}(X_{N-1}^\pi)$ is not almost surely a minimizer in the definition of $J_{N-1}(X_{N-1}^\pi)$ then such $\mu_{N-1}^{'}(x_{N-1})$ must exist.) Now

consider the Markov strategy $\pi' = (\mu_0, \ldots, \mu_{N-2}, \mu'_{N-1})$. We will argue that the expected total cost achieved by this strategy is *strictly* better than that achieved by the strategy π, thus arriving at a contradiction. Since $J_N(x_N) = g_N(x_N)$ by definition, the preceding inequality can also be written as

$$E[g_{N-1}(X^\pi_{N-1}, \mu_{N-1}(X^\pi_{N-1}), w_{N-1}) + g_N(X^\pi_N)|X^\pi_{N-1}] \geq$$

$$E[g_{N-1}(X^\pi_{N-1}, \mu'_{N-1}(X^\pi_{N-1}), w_{N-1}) + g_N(X^{\pi'}_N)|X^\pi_{N-1}]$$

with the inequality being strict with positive probability. Both strategies result in exactly the same state evolution up to and including time $N-1$, i.e. $X^\pi_k = X^{\pi'}_k$ almost surely for $k = 0, \ldots, N-1$. Thus the difference between the total expected cost of the strategy π and that of the strategy π' is precisely the difference between the expectation of the left hand side and the expectation of the right hand side of the preceding inequality, and this is strictly positive, which is the sought for contradiction.

For the cases $k < N-1$, we will assume, for simplicity, that minimizers exist in the definitions for $J_l(x_l)$ for $l = k+1, \ldots, N-1$ (this is only to avoid the need to go through an approximation argument). Suppose there is some function $\mu'_k(x_k)$ different from the function $\mu_k(x_k)$ such that

$$E[g_k(X^\pi_k, \mu_k(X^\pi_k), w_k) + J_{k+1}(f_k(X^\pi_k, \mu_k(X^\pi_k), w_k))|X^\pi_k] \geq$$

$$E[g_k(X^\pi_k, \mu'_k(X^\pi_k), w_k) + J_{k+1}(f_k(X^\pi_k, \mu'_k(X^\pi_k), w_k))|X^\pi_k]$$

and that this inequality holds strictly with positive probability. (If $\mu_k(X^\pi_k)$ is not almost surely a minimizer in the definition of $J_k(X^\pi_k)$ then such $\mu'_k(x_k)$ must exist.) Now consider the Markov strategy $\pi' = (\mu_0, \ldots, \mu_{k-1}, \mu'_k, \tilde{\mu}_{k+1}, \ldots, \tilde{\mu}_{N-1})$, where $\tilde{\mu}_l(x_l)$, $l = k+1, \ldots, N-1$ are respectively minimizers in the definitions for $J_l(x_l)$ for $l = k+1, \ldots, N-1$, which are assumed to exist. We will argue that the expected total cost achieved by this strategy is *strictly* better than that achieved by the strategy π, thus arriving at a contradiction.

We consider as an intermediate the Markov strategy $\tilde{\pi} = (\mu_0, \ldots, \mu_{k-1}, \mu_k, \tilde{\mu}_{k+1}, \ldots, \tilde{\mu}_{N-1})$. We have $X^\pi_l = X^{\tilde{\pi}}_l$ for $l = 0, \ldots, k+1$. Considering the restricted optimization problem over the interval of time $k+1 \leq l \leq n$ (i.e. with the original one step and terminal costs, but thinking of the state at time $k+1$ as an initial condition) and appealing to the assumption that $\tilde{\mu}_l(x_l)$, $l = k+1, \ldots, N-1$ are respectively minimizers in the definitions for $J_l(x_l)$ for $l = k+1, \ldots, N-1$, and the first part of the theorem we get $\Gamma^{\tilde{\pi}}_{k+1} = J_{k+1}(X^{\tilde{\pi}}_{k+1})$. From this we learn that the Markov strategy $\tilde{\pi}$ is also an optimal strategy for the original problem. Thus, if we can show that the expected total cost achieved by the Markov strategy π' is strictly better than that achieved by the Markov strategy $\tilde{\pi}$, we would arrive at a contradiction.

Since the Markov strategy $\pi^{'}$ uses the minimizers $\tilde{\mu}_l(x_l)$, $l = k+1, \ldots, N-1$ from time $k+1$ onwards, we have $\Gamma^{\pi^{'}}_{k+1} = J_{k+1}(X^{\pi^{'}}_{k+1})$ (again, this comes from considering the restricted optimization problem over the interval of time $k+1 \leq l \leq n$). Now, the presumed inequality with which we started the discussion can be written as

$$E[g_k(X^{\pi}_k, \mu_k(X^{\pi}_k), w_k) + J_{k+1}(X^{\pi}_{k+1})|X^{\pi}_k] \geq$$

$$E[g_k(X^{\pi}_k, \mu^{'}_k(X^{\pi}_k), w_k) + J_{k+1}(X^{\pi^{'}}_{k+1})|X^{\pi}_k] \ ,$$

with the inequality holding strictly with positive probability. This can be further written as

$$E[g_k(X^{\tilde{\pi}}_k, \mu_k(X^{\tilde{\pi}}_k), w_k) + J_{k+1}(X^{\tilde{\pi}}_{k+1})|X^{\tilde{\pi}}_k] \geq$$

$$E[g_k(X^{\tilde{\pi}}_k, \mu^{'}_k(X^{\tilde{\pi}}_k), w_k) + J_{k+1}(X^{\pi^{'}}_{k+1})|X^{\tilde{\pi}}_k] \ ,$$

with the inequality holding strictly with positive probability, and again as

$$E[g_k(X^{\tilde{\pi}}_k, \mu_k(X^{\tilde{\pi}}_k), w_k) + \Gamma^{\tilde{\pi}}_{k+1}|X^{\tilde{\pi}}_k] \geq$$

$$E[g_k(X^{\tilde{\pi}}_k, \mu^{'}_k(X^{\tilde{\pi}}_k), w_k) + \Gamma^{\pi^{'}}_{k+1}|X^{\tilde{\pi}}_k] \ ,$$

with the inequality holding strictly with positive probability. Since $X^{\tilde{\pi}}_l = X^{\pi^{'}}_l$ for $l = 0, \ldots, k$, throwing in the costs incurred before time k and taking expectations in this inequality shows that the total expected cost of $\tilde{\pi}$ is strictly bigger than that for $\pi^{'}$, which is the sought for contradiction.