EE 223: Stochastic Systems: Estimation and Control	**Fall 2006**
Lecture 11 — February 20	
Lecturer: Venkat Anantharam	*Scribe: Saurabh Amin*

We continue our discussion on partially observed dynamic programming from last time. This lecture will revisit the material of Lecture 10 before making further progress.

11.1 Finite Horizon Partially Observed DP Problem

Consider the stochastic system model starting from initial state x_0 in the state space form

$$\begin{aligned} x_{k+1} &= f_k(x_k, u_k, w_k) \;\; k = 0, \ldots, N-1 \\ y_0 &= h_0(x_0, v_0) \\ y_k &= h_k(x_k, u_{k-1}, v_k) \;\; k = 1, \ldots, N \end{aligned}$$

with $x_0, w_0, \ldots, w_{N-1}, v_0, \ldots, v_N$ mutually independent random variables.
A strategy Ψ is comprised of functions $u_k = \nu_k(y_0, \ldots, y_k)$, $k = 0, \ldots, N-1$. The aim is to minimize the following cost function

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

with $U_k^{\Psi} = \nu_k(Y_0^{\Psi}, \ldots, Y_k^{\Psi})$. Notice that the essential difference with respect to the fully observed case is that the controller has to be a function of only the past observations. We convert the partially observed problem to a fully observed one with a new definition of state. Recall the notation from Lecture 10:

$$\begin{aligned} \eta_k &\triangleq (y_0, \ldots, y_k, u_0, \ldots, u_{k-1}) \\ I_k^{\Psi} &\triangleq (Y_0^{\Psi}, \ldots, Y_k^{\Psi}, U_0^{\Psi}, \ldots, U_{k-1}^{\Psi}) \end{aligned}$$

We want to find a new state (also known as *information state*) that satisfies the following:

- The information state at time k is a function of information available upto time k,
- The information state at time $k+1$ can be determined from the information state at time k, control applied at time k, and observation obtained at time $k+1$.

We saw last time that these properties are satisfied by the *conditional law* of X_k^{Ψ} given I_k^{Ψ}. We denote the conditional law by λ_k^{Ψ}. The conditional law is a probability distribution

valued random variable taking values in the space of all probability distributions on the state space $\mathcal{X}_k$ at time k and is a (deterministic) function of I_k^Ψ. That is, for all $A \subseteq \mathcal{X}_k$

$$\lambda_k^\Psi(A) = P(X_k^\Psi \in A | I_k^\Psi) = E[1(X_k^\Psi \in A | I_k^\Psi] = \int_A p^\Psi(x_k | I_k^\Psi) dx_k \ ,$$

where the last of these expressions uses a density type notation. For example, if $\mathcal{X}_k$ is a finite set of cardinality 5, then λ_k^Ψ lies in the unit simplex in $\mathbf{R}^5$. To understand the difference between conditional law and conditional expectation, let us recall the definition of the latter.

Remark 11.1. *(Conditional Expectation) Given random variables X and $Y_1, \ldots, Y_m$, the conditional expectation $E[X|Y_1, \ldots, Y_m]$ is the function of $Y_1, \ldots, Y_m$ which is the best estimate of X given $Y_1, \ldots, Y_m$ in the least squares sense. This may be characterized by the requirement that $[X - E[X|Y_1, \ldots, Y_m]]$ is orthogonal to all functions of the conditioning variables, i.e., $E[(X - E[X|Y_1, \ldots, Y_m])g(Y_1, \ldots, Y_m)] = 0$ for all $g(Y_1, \ldots, Y_m)$.*

One can view the conditional law of X as the function of $Y_1, \ldots, Y_m$ that best estimates the *distribution of* X (and not just X) given $Y_1, \ldots, Y_m$. Therefore, the conditional law is much more informative than the conditional expectation, in that it will enable us to simultaneously estimate all functions of X given $Y_1, \ldots, Y_m$.

In Lecture 10, using a density type notation we defined functions $p_{k|k}(x_k|\eta_k)$ and $p_{k+1|k}(x_{k+1}|\eta_k, u_k)$, starting with $p_{0|-1}(x_0) = p(x_0)$, the initial law, that could be recursively calculated as

$$p_{k+1|k+1}(x_{k+1}|\eta_{k+1}) = \frac{p(y_{k+1}|x_{k+1}, u_k) p_{k+1|k}(x_{k+1}|\eta_k, u_k)}{\int p(y_{k+1}|x_{k+1}, u_k) p_{k+1|k}(x_{k+1}|\eta_k, u_k) dx_{k+1}}$$

$$p_{k+1|k}(x_{k+1}|\eta_k, u_k) = \int p(x_{k+1}|x_k, u_k) p_{k|k}(x_k|\eta_k) dx_k$$

Here, $p(y_{k+1}|x_{k+1}, u_k)$ comes from the observation equation $y_{k+1} = h_{k+1}(x_{k+1}, u_k, v_{k+1})$ and $p(x_{k+1}|x_k, u_k)$ comes from the state equation $x_{k+1} = f_k(x_k, u_k, w_k)$. Also recall that these functions do not depend on strategy Ψ for $k = 0, \ldots, N-1$.

Under strategy Ψ, we saw that we could compute the conditional law λ_k^Ψ as begin given by

$$x_k \mapsto p_{k|k}(x_k | I_k^\Psi)$$

in density type notation. Moving further, we want to see that the original cost functions can be expressed as functions of the information state.

First observe that since w_k is independent of (I_k^Ψ, X_0^Ψ), we can work with $\tilde{g}_k(x_k, u_k) \triangleq E_{w_k}[g_k(x_k, u_k, w_k)]$, to write $E[g_k(X_k^\Psi, U_k^\Psi, w_k)] = E[\tilde{g}_k(X_k^\Psi, U_k^\Psi)]$. This is because

$$E[g_k(X_k^\Psi, U_k^\Psi, w_k)] = E[E[g_k(X_k^\Psi, U_k^\Psi, w_k) \mid X_k^\Psi, U_k^\Psi]] = E[\tilde{g}_k(X_k^\Psi, U_k^\Psi)] \ ,$$

where the second step comes from the independent of w_k from (X_k^Ψ, U_k^Ψ), which is a consequence of the independence of w_k from (I_k^Ψ, X_0^Ψ).

We now claim that

$$E[\tilde{g}_k(X_k^\Psi, U_k^\Psi)] = E[G_k(\lambda_k^\Psi, U_k^\Psi)]$$

for some function G_k of laws on the state space at time k and controls at time k. This follows from the fact that U_k^Ψ is a function of I_k^Ψ and so $E[\tilde{g}_k(X_k^\Psi, U_k^\Psi)]$ can be expressed in terms of the conditional law of X_k^Ψ given I_k^Ψ, namely λ_k^Ψ. Indeed

$$E[\tilde{g}_k(X_k^\Psi, U_k^\Psi)] = E[E[\tilde{g}_k(X_k^\Psi, U_k^\Psi)|I_k^\Psi]] = E[\int \tilde{g}_k(x, U_k^\Psi)\lambda_k^\Psi(dx)] \ ,$$

where the quantity in the expectation in the RHS term depends on I_k^Ψ and is hence random. Thus, what we call $G_k(\lambda, u)$ for λ a probability distribution on $\mathcal{X}_k$ and u a control would be

$$G_k(\lambda, u) = \int \tilde{g}_k(x, u)\lambda(dx) \ .$$

Thus we arrive at the following conclusion: Our control problem has been reformulated as a fully observed control problem for a new dynamical system whose states are conditional laws of the original state (also called the information state). There is only one gap: we need to argue that the evolution of the new state can be treated as being driven at each step by a random variable that is independent from time to time and such that these variables are independent of the initial condition. We will accept for the moment that this true; this gap will be closed in one of the later lectures. So, by the theory of fully observed DP, there exists an optimal control that depends only on the current information state.

11.2 Finite State Partially Observed Markov Decision Process

Suppose the state space at each time is $\mathcal{X} = \{1, \ldots, d\}$. Controls are drawn from a finite set $\mathcal{U}$ at each time. The observations take values in a finite set$\mathcal{Y}$ at each time. For each $u \in \mathcal{U}$, we have a transition probability matrix $\mathbf{P}(u) \triangleq [p_{ij}(u)]$. The observation at time k has the conditional probability described by $q(y_k|x_k, u_{k-1})$, where $q(y \mid j, u)$ gives a probability distribution on $\mathcal{Y}$ for each $(j, u) \in \mathcal{X} \times \mathcal{U}$. The information state at time k can be determined in terms of a row vector of size d.

$$\pi_{k|k}(\eta_k) = [p_{k|k}(1|\eta_k), \ldots, p_{k|k}(d|\eta_k)] \ ,$$

whose dynamics are given by

$$\pi_{k+1|k+1}(\eta_{k+1}) \propto \pi_{k|k}(\eta_k)\mathbf{P}(u_k)D(u_k, y_{k+1}) \ ,$$

where $D(u_k, y_{k+1})$ denotes the $d \times d$ diagonal matrix with entry $q(y_{k+1} \mid i, u_k)$ in the (i,i)-th location. The starting condition $\pi_{0|0}(\eta_0)$ can be computed from $\pi_{0|-1}$, which is the initial distribution of the state. We can compute these quantities offline. Note that the quantity on the RHS of these equations denote the unnormalized $\pi_{k+1|k+1}(\eta_{k+1})$ and these evolve according to a linear equation. The information state at time k for any strategy Ψ will be $\pi_{k|k}(I_k^\Psi)$. The optimal control at time k, as part of the optimal strategy Ψ, will be a function only of the information state $\pi_{k|k}(I_k^\Psi)$.

11.3 Sequential Probability Ratio Test

We first recall the static binary hypothesis testing problem in a Bayesian setting. The decision maker makes an observation from a finite set $\mathcal{Y}$. Let α be the prior probability that hypothesis H_0 is true. $(p_0(y), y \in \mathcal{Y})$ is the distribution of the observation given that hypothesis H_0 is true. $(p_1(y), y \in \mathcal{Y})$ is the distribution of the observation given that hypothesis H_1 is true. The problem is to decide which hypothesis is true based on the observation.

If the decision maker makes the correct decision, zero cost is incurred. If H_0 (H_1) holds and H_1 (H_0) is decided, cost L_1 (L_0) is incurred. A decision rule is a map $d : \mathcal{Y} \mapsto \{0,1\}$. The optimal decision rule can be obtained by minimizing the total expected cost

$$\alpha L_1 \sum_{y:d(y)=1} p_0(y) + (1-\alpha) L_0 \sum_{y:d(y)=0} p_1(y)$$

For any given $y \in \mathcal{Y}$, the decision maker can either take the hit $\alpha L_1 p_0(y)$ or $(1-\alpha)L_0 p_1(y)$. The decision rule will be optimal if and only if the smaller hit is chosen (if there is a tie it doesn't matter what decision is made). That is, an optimal decision rule is

$$d^*(y) = 1 \Leftrightarrow \frac{(1-\alpha)L_0 p_1(y)}{\alpha L_1 p_0(y)} \geq 1$$

This is a threshold rule based on the *likelihood ratio* $\frac{p_1(y)}{p_0(y)}$: if the likelihood ratio is big enough (at least $\frac{\alpha L_1}{(1-\alpha)L_0}$) the decision maker should decide H_1 on observing y.

In the next lecture, we will consider the sequential hypothesis testing problem in which instead of deciding based on one sample, the decision maker has the option of getting new samples, at a cost C, up to a total of M samples. All samples are independent over time conditioned on the true hypothesis, and the underlying hypothesis does not change over time.