Introduction to Control with Learning	March 28th, 2023

Lecture 3

Lecturer: Oron Sabag *Scribe: os*

In this lecture,

1. State-space models
2. The asymptotic observer
3. Reminder on least squares problem
4. Stochastic least squares
5. The Kalman filter
6. The innovations process (next lecture)

1 State-space models as mappings

Consider a discrete-time linear time-invariant (LTI) system represented by the difference equation:

$$y_k + a_1 y_{k-1} + a_2 y_{k-2} + \cdots + a_n y_{k-n} = b_0 u_k + b_1 u_{k-1} + b_2 u_{k-2} + \cdots + b_{n-1} u_{k-n+1}. \tag{1}$$

The inputs are the u_i while the outputs are the y_i.

It may seem that the output y_k only depends on $(u_{k-n+1}, \ldots, u_k)$ but, indeed, it depends on all inputs $(u_0, \ldots, u_k)$ via the dependence on past outputs.

Simple examples:

1. **Moving average**: if $a_1 = a_2 = \cdots = a_n = 0$.
2. **Auto-regressive**: if $b_1 = b_2 = \cdots = b_n = 0$. (Special case: Fibonacci sequence)
3. **ARMA**: auto-regressive moving average. Sometimes, used as ARMA(i, j) to include the order.

Taking the Z-transform of both sides (assuming zero initial conditions), we get:

$$Y(z) + a_1 z^{-1} Y(z) + a_2 z^{-2} Y(z) + \cdots + a_n z^{-n} Y(z) = b_0 U(z) + b_1 z^{-1} U(z) + b_2 z^{-2} U(z) + \cdots + b_{n-1} z^{-n+1} U(z \tag{2}$$

Factoring out $Y(z)$ and $U(z)$:

$$Y(z)\left(1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-n}\right) = U(z)\left(b_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_{n-1} z^{-n+1}\right). \tag{3}$$

Thus, the transfer function $H(z)$ is:

$$H(z) = \frac{Y(z)}{U(z)} = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_{n-1} z^{-n+1}}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-n}}. \tag{4}$$

The zeros of numerator are called the *zeros* of the transfer function $H(z)$. The zeros of denumerator are called the *poles* of the transfer function $H(z)$.

Controllable Canonical Form

For the discrete-time transfer function:

$$H(z) = \frac{b_0 z^{n-1} + b_1 z^{n-2} + \cdots + b_{n-1}}{z^n + a_1 z^{n-1} + \cdots + a_n}, \tag{5}$$

the state-space representation in controllable canonical form is:

$$A_c = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix}, \quad B_c = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad C_c = \begin{bmatrix} b_{n-1} & b_{n-2} & \cdots & b_1 & b_0 \end{bmatrix}, \quad D_c = 0 \tag{6}$$

Observable Canonical Form

For the discrete-time transfer function:

$$H(z) = \frac{b_0 z^{n-1} + b_1 z^{n-2} + \cdots + b_{n-1}}{z^n + a_1 z^{n-1} + \cdots + a_n}, \tag{7}$$

the state-space representation in observable canonical form is:

$$A_o = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & 0 \\ -a_2 & 0 & 1 & \cdots & 0 \\ -a_3 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_n & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad B_o = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad C_o = \begin{bmatrix} b_0 & b_1 & \cdots & b_{n-2} & b_{n-1} \end{bmatrix}, \quad D_o = 0 \tag{8}$$

Jordan Canonical Form

For the Jordan canonical form, consider a discrete-time system $x_{k+1} = Ax_k + Bu_k$ and $y_k = Cx_k + Du_k$. The Jordan canonical form involves transforming A into its Jordan form J, where $A = PJP^{-1}$ and J is a block diagonal matrix consisting of Jordan blocks. Each Jordan block corresponds to an eigenvalue of A.

The Jordan canonical form is:

$$J = \begin{bmatrix} J_1 & 0 & \cdots & 0 \\ 0 & J_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_p \end{bmatrix}, \tag{9}$$

where each J_i is a Jordan block.

If J_i is a Jordan block corresponding to the eigenvalue λ_i, it has the form:

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 \\ 0 & 0 & \lambda_i & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 1 \\ 0 & 0 & 0 & \cdots & \lambda_i \end{bmatrix}. \tag{10}$$

The state-space representation in Jordan canonical form is:

$$A_J = J, \quad B_J = P^{-1}B, \quad C_J = CP, \quad D_J = D. \tag{11}$$

2 The inverted pendelum

Linearized State-Space Model. We define the state vector as

$$x = \begin{bmatrix} x \\ \dot{x} \\ \theta \\ \dot{\theta} \end{bmatrix} \quad \text{and input } u = \text{force applied to the cart.} \tag{12}$$

Linearizing around the upright equilibrium ($\theta = 0$), the continuous-time dynamics take the form:

$$\dot{x} = Ax + Bu, \tag{13}$$

$$y = Cx. \tag{14}$$

The system matrices are:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -\dfrac{d}{M} & \dfrac{mg}{M} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -\dfrac{d}{ML} & -\dfrac{(M+m)g}{ML} & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \dfrac{1}{M} \\ 0 \\ \dfrac{1}{ML} \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}. \tag{15}$$

Interpretation:

- x: horizontal position of the cart.
- $\dot{x}$: cart velocity.
- θ: angle of the pendulum from vertical (upright at $\theta = 0$).
- $\dot{\theta}$: angular velocity of the pendulum.
- M: mass of the cart, m: mass of the pendulum, L: length to pendulum center of mass.
- g: gravitational acceleration, d: friction/damping coefficient on the cart.
- A: encodes system dynamics including gravity and coupling between cart and pendulum.
- B: describes how the applied force affects both cart and pendulum.
- C: output is the cart position.

In our path to solve the IP on a cart problem, we will now learn how to choose the controller when it does not have the full state vector. For example, here it is only a function of the position. This involves sort of "estimation", i.e., given a measurement, how to find a good estimate of the state.

3 The asymptotic observer

The setting is given by a linear system

$$\begin{aligned} x_{i+1} &= Ax_i + Bu_i, \quad x_0 \\ y_i &= Cx_i \end{aligned} \tag{16}$$

where:

1. x_i is the states sequence
2. u_i is the control sequence (which is available to us)
3. y_i is the measurements process that we observe.

The objective is to estimate the states. If the initial state is available to us, we can easily compute

$$\hat{x}_{i+1} = A\hat{x}_i + Bu_i, \quad \hat{x}_0 = x_0. \tag{17}$$

and obtain zero estimation error.

The question is how to estimate the states when we do not have the initial state? One solution is to use (17) with an initial guess of x_0. In particular, assume that $\hat{x}_0$ is our best guess for the initial state x_0 (in practice, we may take it to be the zeros vector).

Then, we can evaluate (17) with the initial condition $\hat{x}_0$. We examine the error evolution of the estimates in (17) with $\tilde{x}_0 = x_0 - \hat{x}_0 \neq 0$. Denote the error as

$$\begin{aligned}\tilde{x}_{i+1} &\triangleq x_{i+1} - \hat{x}_{i+1} \\ &= A\tilde{x}_i \\ &= A^{i-1}\tilde{x}_0 \end{aligned} \tag{18}$$

It becomes clear that the knowledge of the initial state is crucial. If $\tilde{x}_0 \neq 0$, then any unstable mode of A (i.e., an eigenvalue outside the unit circle) will lead to diverging error as $i \to \infty$ (assume that $\tilde{x}_0 \notin \text{null}(A)$).

The above scheme is called open-loop since we are not utilizing the knowledge of the measurements to improve our state estimates. A natural closed-loop linear estimate (that will be shown to be optimal) is

$$\hat{x}_{i+1} = \underbrace{A\hat{x}_i + Bu_i}_{prediction} + \underbrace{K(y_i - C\hat{x}_i)}_{update}, \quad \hat{x}_0. \tag{19}$$

We will later show how to find the constant K explicitly, but the focus here is to show the different behavior of the error when we have a closed loop

$$\begin{aligned}\tilde{x}_{i+1} &\triangleq x_{i+1} - \hat{x}_{i+1} \\ &= Ax_i + Bu_i - (A\hat{x}_i + Bu_i + K(y_i - C\hat{x}_i)) \\ &= (A - KC)\tilde{x}_i \\ &= (A - KC)^{i-1}\tilde{x}_0. \end{aligned} \tag{20}$$

The difference between (18) and (20) is fundamental and is the base for our class today: if we properly design the matrix K such that $\rho(A - KC) < 1$, then we will be able to achieve a vanishing error no matter how good is our initial guess $\hat{x}_0$.

Definition 1 (Observability)**.** A discrete-time system is said to be **observable** if, for any initial state x_0, the state x_0 can be determined from the output $y_1, y_2, \ldots, y_k$ over a finite time interval $[0, k]$.

Lemma 1. *The discrete-time system is observable if and only if the observability matrix $\mathcal{O}$ defined by:*

$$\mathcal{O} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} \tag{21}$$

has full rank, i.e.,

$$rank(\mathcal{O}) = n. \tag{22}$$

Proof is omitted since it is similar to the notion of controllability in Lecture 2.

4 Least squares - recap

Given is a set of linear equation

$$y = Hx, \tag{23}$$

where $y \in \mathbb{R}^n$ is the measurements vector and $H \in \mathbb{R}^{n \times m}$.

The vector x can be thought of as a state to be estimated from y. There are several regimes for this problem, and we are interested in the case where $\nexists x : y = Hx$. That is, the system is over-determined ($n \geq m$).

The least squares solution is a solution that minimizes

$$J(\hat{x}) = \|y - H\hat{x}\|^2 = (y - H\hat{x})^T(y - H\hat{x}). \tag{24}$$

The following result characterizes an optimal estimator.

Lemma 2. *A solution $\hat{x}$ is optimal, i.e., $J(\hat{x}) \leq J(x)$ for all x if the normal equations*

$$H^T H\hat{x} = H^T y \tag{25}$$

are satisfied. The optimal objective is

$$J(\hat{x}) = \|y\|^2 - \|H\hat{x}\|^2. \tag{26}$$

The proof is omitted and simply follows by computing the derivative (the objective is convex due to the positive semi definite property of the hessian.

The take-out from this simple exercise is the concept of projection. The optimal estimate satisfies the orthogonality principle $\|y\|^2 = \|y - H\hat{x}\|^2 + \|H\hat{x}\|^2$. This concept will also generalize to our more involved problem of Kalman filtering. **to add figure**

If H has a full rank, the solution is given by

$$\hat{x} = (H^T H)^{-1} H^T y. \tag{27}$$

Otherwise, we can show that any two optimal solutions must satisfy $H(\hat{x}_1 - \hat{x}_2) = 0$.

5 Stochastic least squares

In stochastic least squares, we have a similar setting but with random variables. Given are two random vectors X, Y with a probability density function $f_{X,Y}$. The objective is to construct an estimator for X given Y, denoted by $\hat{X}(Y)$. The error $\Delta = X - \hat{X}(Y)$ is a random variable on its own so we study the mean squared error (MSE):

$$MSE = \mathbb{E}[(X - \hat{X}(Y))^T(X - \hat{X}(Y))]. \tag{28}$$

It turns out that if MSE is the criteria, we can find the optimal estimator.

Lemma 3. *The estimator that minimizes* (28) *is the conditional expectation*

$$\hat{X} = \mathbb{E}[X|Y]. \tag{29}$$

In other words, for any $g : \mathcal{Y} \to \mathcal{X}$*, we have*

$$\mathbb{E}[(X - g(Y))^T(X - g(Y))] \geq \mathbb{E}[(X - \mathbb{E}[X|Y])^T(X - \mathbb{E}[X|Y])]. \tag{30}$$

Proof. The proof is a consequence of the law of total expectation

$$\begin{aligned}
&\mathbb{E}[(X - g(Y))^T(X - g(Y))] \\
&= \mathbb{E}[(X - \mathbb{E}[X|Y])^T(X - \mathbb{E}[X|Y])] + \mathbb{E}[(g(Y) - \mathbb{E}[X|Y])^T(g(Y) - \mathbb{E}[X|Y])] \\
&+ \mathbb{E}[(X - \mathbb{E}[X|Y])^T(g(Y) - \mathbb{E}[X|Y])] + \mathbb{E}[(g(Y) - \mathbb{E}[X|Y])^T(X - \mathbb{E}[X|Y])] \\
&\overset{(a)}{=} \mathbb{E}[(X - \mathbb{E}[X|Y])^T(X - \mathbb{E}[X|Y])] + \mathbb{E}[(g(Y) - \mathbb{E}[X|Y])^T(g(Y) - \mathbb{E}[X|Y])] \\
&\overset{(b)}{\geq} \mathbb{E}[(X - \mathbb{E}[X|Y])^T(X - \mathbb{E}[X|Y])].
\end{aligned} \tag{31}$$

where (a) follows from the law of total expectation as

$$\mathbb{E}_Y[\mathbb{E}_{X|Y}[(X - \mathbb{E}[X|Y])^T(g(Y) - \mathbb{E}[X|Y])]] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[(X - \mathbb{E}[X|Y])^T](g(Y) - \mathbb{E}[X|Y])] = 0,$$

and (b) follows from the non-negativity of the second term. □

The estimator is a function of the measurement Y. In order to compute the (mostly non-trivial) conditional expected value, we should know the density function $f_{X,Y}$. In the following, we restrict estimators to be linear which in turn simplifies the estimator implementation.

A linear estimator takes the form $\hat{X} = KY$, where K is a matrix to be optimized. The error covariance matrix of a linear estimator K is defined as

$$P(K) = \mathbb{E}[(X - KY)(X - KY)^T]. \tag{32}$$

Note that $P(K)$ is the error covariance matrix. Thus, we say K_0 is a linear least mean square estimator (LMMSE) if

$$P(K) \succeq P(K_0) \tag{33}$$

for any K. Alternatively, we can write (33) as

$$a^T P(K) a \geq a^T P(K_0) a, \tag{34}$$

for all K and vectors a. Informally, the LMMSE guarantees that the error is minimized in all directions.

Fact: The LLMSE is optimal if (X, Y) **are jointly Gaussian.**

The following result is in analogy with the deterministic least squares problem.

Theorem 1. *Any LLMSE K_0 satisfies the normal equations*

$$K_0 R_Y = R_{XY}, \tag{35}$$

where $R_Y = \mathbb{E}[YY^T]$ is the covariance of the measurements, and R_{XY} is the covariance between X and Y.

If R_Y is invertible, we obtain the well-known estimator

$$\hat{X} = R_{XY} R_Y^{-1} Y$$

with the estimation error covariance

$$P(K_0) = R_X - R_{XY} R_Y^{-1} R_{YX}. \tag{36}$$

Note that to implement the linear estimator, all we need is the first and second moments of the variables X, Y. This is in contrast to the estimator (29) that requires the knowledge of the density.

Proof. Sketch of proof: take derivative

$$\frac{\partial\, \mathbb{E}[KYY^T K^T]}{\partial K} = 2KR_Y$$

and

$$\frac{\partial\, \mathbb{E}[XY^T]K^T}{\partial K} = R_{XY},$$

and verify that it is a convex function. □

We can gain some intuition on the solution by writing the covariance matrix using an UDL factorization (see Lemma 4 in the Appendix)

$$\begin{pmatrix} R_X & R_{XY} \\ R_{YX} & R_Y \end{pmatrix} = \begin{pmatrix} I & R_{XY}R_Y^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} R_X - R_{XY}R_Y^{-1}R_{YX} & 0 \\ 0 & R_Y \end{pmatrix} \begin{pmatrix} I & 0 \\ R_Y^{-1}R_{XY} & I \end{pmatrix} \tag{37}$$

Since the first element of the diagonal matrix is the optimal error variance, we can write:

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} I & R_{XY}R_Y^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \tilde{X} \\ Y \end{pmatrix}. \tag{38}$$

This can be viewed as a Gram-Schmidt process to project the variable X onto the linear space spanned by Y with the inner product $< X, Y >= \mathbb{E}[XY^T]$. Thus, we will sometimes use this to write the estimator as a standard projection:

$$\hat{X} =< X, Y >< Y, Y >^{-1} Y =< X, Y > \|Y\|^{-2} Y. \tag{39}$$

6 Kalman filter

The model is given by a state-space

$$\begin{aligned} x_{i+1} &= Fx_i + Gw_i, \\ y_i &= Hx_i + v_i, \end{aligned} \tag{40}$$

where w_i, v_i are random variables to be defined shortly.

Definition 2 (Estimator)**.** An estimator is a sequence of mappings

$$\pi_i = \mathcal{Y}^i \to \mathcal{X}, \quad i \geq 0 \tag{41}$$

such that $\hat{x}_{i|i} = \pi_i(y^i)$. The corresponding error of an estimator is

$$x_i - \hat{x}_{i|i}. \tag{42}$$

Definition 3 (Predictor)**.** A predictor is a sequence of mappings

$$\pi_i = \mathcal{Y}^{i-1} \to \mathcal{X}, \quad i \geq 0 \tag{43}$$

such that $\hat{x}_{i|i-1} = \pi_i(y^{i-1})$. The corresponding error of a predictor is

$$x_i - \hat{x}_{i|i-1}. \tag{44}$$

The above definition can be unified with the estimator $\hat{x}_{i|j}$ that estimates x_i using the measurements $y_1, \ldots, y_j$. We will focus on linear predictors as similar derivations can be used to solve the general case.

The random variables (w_i, v_i) are zero mean and are not correlated among different times. Formally, the following covariance matrix describes the model:

$$\mathbb{E}[\begin{pmatrix} w_i \\ v_i \\ x_0 \end{pmatrix} \begin{pmatrix} w_j^T & v_j^T & x_0^T & 1 \end{pmatrix}] = \begin{pmatrix} \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix} \delta_{ij} & 0 & 0 \\ 0 & \Pi_0 & 0 \end{pmatrix}, \tag{45}$$

where $\begin{pmatrix} Q & S \\ S^T & R \end{pmatrix}$ and Π_0 are positive semidefinite matrices and δ_{ij} equals 1 if $i = j$ and is zero otherwise.

Note that w_i is uncorrelated as a process over time but its coordinates at a fixed time can be correlated via Q.

6.1 Proposed predictor

In a similar manner to the closed-loop estimator in (19), we can suggest the predictor

$$\hat{x}_{i+1|i} = F\hat{x}_{i|i-1} + K_i(y_i - H\hat{x}_{i|i-1}), \tag{46}$$

where K_i is an arbitrary matrix that can be optimized.

Theorem 2. *Consider the model in* (40) *and the predictor suggested in* (46). *If the error covariance matrix at time i is $P_i \triangleq \mathbb{E}[(x_i - \hat{x}_i)(x_i - \hat{x}_i)^T]$, the matrix K_i that minimizes P_{i+1} is given by*

$$K_{p,i} \triangleq (FP_iH^T + GS)R_{e,i}^{-1}, \tag{47}$$

with $R_{e,i} = HP_iH^T + R$.

Note that Theorem 2 does not show the optimality of the predictor among all linear mappings, it is just among those who use the innovation $y_i - H\hat{x}_{i|i-1}$.

Proof. The corresponding error is

$$\begin{aligned}\tilde{x}_{i+1} &= x_{i+1} - \hat{x}_{i+1|i} \\ &= (F - K_iH)\tilde{x}_i + \begin{pmatrix} G & -K_i \end{pmatrix} \begin{pmatrix} w_i \\ v_i \end{pmatrix}.\end{aligned} \tag{48}$$

The error covariance is

$$\begin{aligned}P_{i+1} &= E[\tilde{x}_{i+1}^T \tilde{x}_{i+1}] \\ &\overset{(a)}{=} (F - K_iH)P_i(F - K_iH)^T + \begin{pmatrix} G & -K_i \end{pmatrix} \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix} \begin{pmatrix} G^T \\ -K_i^T \end{pmatrix} \\ &= \begin{pmatrix} I & K_i \end{pmatrix} \begin{pmatrix} FP_iF^T + GQG^T & -(FP_iH^T + GS) \\ -(FP_iH^T + GS)^T & HP_iH^T + R \end{pmatrix} \begin{pmatrix} I \\ K_i^T \end{pmatrix}\end{aligned} \tag{49}$$

where (a) follows from (48) and the fact that (w_i, v_i) is uncorrelated with $\tilde{x}_i$.

We now use the LDU factorization of the middle matrix

$$\begin{pmatrix} FP_iF^T + GQG^T & -(FP_iH^T + GS) \\ -(FP_iH^T + GS)^T & HP_iH^T + R \end{pmatrix} \tag{50}$$

$$= \begin{pmatrix} I & -(FP_iH^T + GS)R_{e,i}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \Delta & 0 \\ 0 & R_{e,i} \end{pmatrix} \begin{pmatrix} I & 0 \\ -R_{e,i}^{-1}(FP_iH^T + GS)^T & I \end{pmatrix} \tag{51}$$

with

$$\begin{aligned}\Delta &= FP_iF^T + GQG^T - (FP_iH^T + GS)R_{e,i}^{-1}(FP_iH^T + GS)^T \\ &= FP_iF^T + GQG^T - K_{p,i}R_{e,i}K_{p,i}^T\end{aligned} \tag{52}$$

Combining (49) and (50) allows to us write the covariance as

$$P_{i+1} = \Delta + (K_i - K_{p,i})R_{e,i}(K_i - K_{p,i})^T \tag{53}$$

and is minimized (in a matrix sense) with $K_i = K_{p,i}$. □

In the next lecture, we will show the optimality of the suggested estimator in 19 and some implementation aspects of the Kalman filter.

A LDU factorization

The LDU factorization is very useful in the context of linear estimation.

Lemma 4. *If A and D are square matrices, and D is invertible, we can write*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \Delta_D & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & 0 \\ D^{-1}C & I \end{pmatrix}, \tag{54}$$

where $\Delta_D = A - BD^{-1}C$ is the Schur complement of the matrix.

If A is invertible, we have

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & \Delta_A \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix}, \tag{55}$$

where $\Delta_A = D - CA^{-1}B$ is the Schur complement of the matrix.

Proof is straightforward.