EE402 - Discrete Time Systems **Fall 2021**

Lecture 18

Lecturer: Elif Sarıtaş

Preliminaries

Multivariate Normal Distribution

A multivariate normal distribution for $x \in \mathbb{R}^{n_x}$

$$\begin{aligned} p(x) =& \mathcal{N}(x;\mu,\Sigma) \\ =& \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left\{ -\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu) \right\} \end{aligned}$$

where

$$\begin{aligned} \mu =& \mathrm{E}\left[x\right] \\ \Sigma =& \mathrm{E}\left[(x-\mu)(x-\mu)^{\mathrm{T}}\right] \end{aligned}$$

Remember that any covariance matrix is

i. a square matrix (in this case $\Sigma \in \mathbb{R}^{n_x \times n_x}$)

ii. symmetric, i.e., $\Sigma = \Sigma^{\mathrm{T}}$

iii. positive semi-definite, i.e., $a^{\mathrm{T}}\Sigma a \geq 0, \ \ \forall a$

Marginal Distribution: Given that $x \sim \mathcal{N}(x;\mu,\Sigma)$ where $x = \left[\begin{array}{cc} x_a & x_b \end{array}\right]^{\mathrm{T}}$, $\mu = \left[\begin{array}{cc} \mu_a & \mu_b \end{array}\right]^{\mathrm{T}}$, and $\Sigma = \left[\begin{array}{cc} \Sigma_a & \Sigma_c \\ \Sigma_c & \Sigma_b \end{array}\right]$, then

$$\begin{aligned} p(x_a) =& \mathcal{N}(x_a;\mu_a,\Sigma_a) \\ p(x_b) =& \mathcal{N}(x_b;\mu_b,\Sigma_b) \end{aligned}$$

Conditional Distribution: Given that $x \sim \mathcal{N}(x;\mu,\Sigma)$ where $x = \left[\begin{array}{cc} x_a & x_b \end{array}\right]^{\mathrm{T}}$, $\mu = \left[\begin{array}{cc} \mu_a & \mu_b \end{array}\right]^{\mathrm{T}}$, and $\Sigma = \left[\begin{array}{cc} \Sigma_a & \Sigma_c \\ \Sigma_c & \Sigma_b \end{array}\right]$, then

$$p(x_a \mid x_b) = \mathcal{N}(x_a;\mu_{a|b},\Sigma_{a|b})$$

where

$$\begin{aligned} \mu_{a|b} =& \mu_a + \Sigma_c\Sigma_b^{-1}(x_b - \mu_b) \\ \Sigma_{a|b} =& \Sigma_a - \Sigma_c\Sigma_b^{-1}\Sigma_c^{\mathrm{T}} \end{aligned}$$

Linear Combinations: Linear combinations of jointly Gaussian densities are also Gaussian, which is one of the main reasons for Gaussian densities being this popular. Given two jointly normal random variables, $(x, y) \sim \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix}; \begin{bmatrix} \mu_x & \\ \mu_y & , \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_y & \Sigma_{yx} \end{bmatrix}\right)$, then

$$\begin{aligned} z =& Ax + By + c \\ z \sim& \mathcal{N}(z; A\mu_x + B\mu_y + c, A\Sigma_x A^{\mathrm{T}} + B\Sigma_y B^{\mathrm{T}} + A\Sigma_{xy} B^{\mathrm{T}} + B\Sigma_{yx} A^{\mathrm{T}}) \end{aligned}$$

When x and y are independent, $\Sigma_{xy} = 0$, then covariance matrix of the z becomes

$$z \sim \mathcal{N}(z; A\mu_x + B\mu_y + c, A\Sigma_x A^{\mathrm{T}} + B\Sigma_y B^{\mathrm{T}})$$

Markov Property

A process, x, is called is Markov, if

$$p(x_k \mid x_{k-1}, x_{k-2}, \cdots, x_0) = p(x_k \mid x_{k-1}) \;\; \forall k.$$

That is the transition between the states depend only on the previous state, which is a summary of the past.

Estimators

An estimator is an algorithm to find one or more unknown values using observations. These unknowns may be

- stochastic or deterministic,
- static or dynamic,
- finite dimensional or infinite dimensional.

There are three inference problems in the Bayesian framework.

Prediction: In this problem, the estimate of an n-step ahead future state is computed by considering the measurements up to time k. That is we deal with the following distribution

$$p(x_{k+n} \mid Y_k), \;\; \text{where} \;\; n = 1, 2, \cdots$$

Filtering: In this case, we estimate the state at time k using the measurements up to and including the time k. The distribution we consider is

$$p(x_k \mid Y_k),$$

Smoothing: This time all of the measurements in an interval is used to estimate the state at a time instant in that interval. Hence, the distribution of interest is now

$$p(x_k \mid Y_T), \quad \text{where} \;\; k < T, \cdots$$

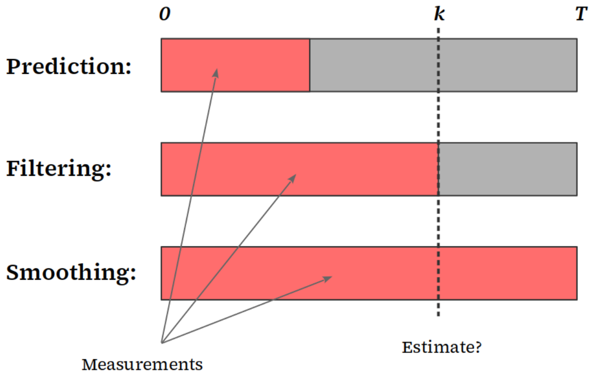

Figure 18.1: Inference problems (Courtesy of Murat Kumru)

Minimum Mean Square Error

Mean square error (MSE) is a measure that demonstrates how much estimates deviate from their true values. It is used as an performance indicator for estimators, and it is defined as

$$\mathrm{MSE}_\theta = \mathrm{E}\left[(\theta - \hat{\theta})^{\mathrm{T}}(\theta - \hat{\theta})\right]$$

where $\theta \sim p(\theta)$, and $\hat{\theta}$ is the estimate.Minimum mean square error is the estimator that minimizes MSE.

Discrete-Time Kalman Filter

Problem Definition

We consider a linear-Gaussian system, in which the models are linear and the noise characteristics are Gaussian.

$$\begin{aligned} x_{k+1} &= Gx_k + Hu_k + w_k \\ y_k &= Cx_k + v_k \end{aligned}$$

where

- $x_k \in \mathbb{R}^{n_x}$ is the state,
- $y_k \in \mathbb{R}^{n_x}$ is the measurement,
- $w_k \in \mathbb{R}^{n_x}$ is the process noise with $w_k \sim \mathcal{N}(w_k; 0, Q)$,
- $v_k \in \mathbb{R}^{n_x}$ is the measurement noise with $v_k \sim \mathcal{N}(v_k; 0, R)$,
- u_k is either deterministic or known, it may a function of known variables.
- We denote all of the measurements collected from the start up to and including the time k as $Y_k = \{y_0, y_1, \cdots, y_k\}$.

The process noise captures the uncertainties regarding the system dynamics, whereas the measurements noise depends on the noise characteristics of the sensors. The probabilistic model for this system is written as

$$\begin{aligned} p(x_{k+1} \mid x_k) =& \mathcal{N}(x_{k+1}; Gx_k + Hu_k, Q), \\ p(y_k \mid x_k) =& \mathcal{N}(y_k; Cx_k, R). \end{aligned}$$

This system can be modelled using block diagrams as follows.

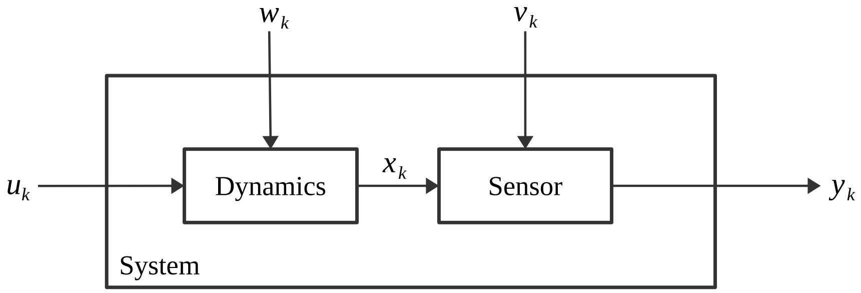

Figure 18.2: Block diagram of the system

The initial state is distributed according to

$$x_0 \sim \mathcal{N}(x_0; \hat{x}_{0|0}, P_{0|0}).$$

We assume that x_0, w_0, w_1, $\cdots$ and v_0, v_1, $\cdots$ are independent.

Goal: Given the measurements y_0, y_1, $\cdots$, y_k, find an estimate of x_k such that the loss function, $\mathrm{E}\left[(x_k - \hat{x}_k)^{\mathrm{T}}(x_k - \hat{x}_k)\right]$, is minimized.

Notation

We will use the following notations throughout the derivation, the first one of the densities below is called the predicted density, whereas the latter is the posterior density.

$$\begin{aligned} p(x_{k+1} \mid Y_k) =& \mathcal{N}(x_{k+1}; \hat{x}_{k+1|k}, P_{k+1|k}) \\ p(x_k \mid Y_k) =& \mathcal{N}(x_k; \hat{x}_{k|k}, P_{k|k}) \end{aligned}$$

Remark:

If we can find a linear estimator, the above equations will hold; as the linear combinations of independent Gaussians are also Gaussian.

Derivation

Let us first consider the prediction problem, as it is quite simple given the motion model. The prediction problem aims to find $\mathrm{E}\left[x_{k+1} \mid Y_k\right]$.

$$\begin{aligned} \hat{x}_{k+1|k} =& \mathrm{E}\left[x_{k+1} \mid Y_k\right] \\ =& \mathrm{E}\left[Gx_k + Hu_k + w_k \mid Y_k\right] \\ =& \mathrm{E}\left[Gx_k \mid Y_k\right] + \mathrm{E}\left[Hu_k \mid Y_k\right] + \mathrm{E}\left[w_k \mid Y_k\right] \\ =& G\mathrm{E}\left[x_k \mid Y_k\right] + Hu_k \\ =& G\hat{x}_{k|k} + Hu_k \end{aligned}$$

Note that the input, u_k, is deterministic and $\mathrm{E}\left[w_k \mid Y_k\right] = \mathrm{E}\left[w_k\right] = 0$.

The predicted error covariance is defined and computed as

$$\begin{aligned} P_{k+1|k} =& \mathrm{E}\left[(x_{k+1} - \hat{x}_{k+1|k})(x_{k+1} - \hat{x}_{k+1|k})^{\mathrm{T}} \mid Y_k\right] \\ =& \mathrm{E}\left[(Gx_k + Hu_k + w_k - G\hat{x}_{k|k} - Hu_k)(Gx_k + Hu_k + w_k - G\hat{x}_{k|k} - Hu_k)^{\mathrm{T}} \mid Y_k\right] \\ =& \mathrm{E}\left[\left(G(x_k - \hat{x}_{k|k}) + w_k\right)\left(G(x_k - \hat{x}_{k|k}) + w_k\right)^{\mathrm{T}} \mid Y_k\right] \\ =& G\mathrm{E}\left[(x_k - \hat{x}_{k|k})(x_k - \hat{x}_{k|k})^{\mathrm{T}} \mid Y_k\right] G^{\mathrm{T}} + \mathrm{E}\left[w_k {w_k}^{\mathrm{T}} \mid Y_k\right] \\ =& GP_{k|k}G^{\mathrm{T}} + Q \end{aligned}$$

where we use the fact that $\mathrm{E}\left[w_k {w_k}^{\mathrm{T}} \mid Y_k\right] = Q$.

Remark:

Both the predicted state and the predicted error covariance estimates depend on the filtered estimates on the previous time step.

Now in order to find the filter estimates at time $k+1$, the idea is to use the predicted estimates and incorporate the new measurement arriving at time $k+1$ into the procedure. Thus, we start this part of the derivation by making the guess that the filtered state estimate is a linear combination of the predicted state and the additional information obtained from the new measurement. That is

$$\hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + K_{k+1}(y_{k+1} - \hat{y}_{k+1|k})$$

where $\hat{y}_{k+1|k}$ is the predicted measurement,

$$\hat{y}_{k+1|k} = C\hat{x}_{k+1|k},$$

the predicted measurement error is named as innovation, and K is an unknown matrix named as Kalman gain. We wish to determine the Kalman gain such that the loss function is minimized. Then,

$$\begin{aligned}
\hat{x}_{k+1|k+1} =& \hat{x}_{k+1|k} + K_{k+1}(Cx_{k+1} + v_{k+1} - C\hat{x}_{k+1|k}) \\
P_{k+1|k+1} =& \mathrm{E}\left[(x_{k+1} - \hat{x}_{k+1|k+1})(x_{k+1} - \hat{x}_{k+1|k+1})^{\mathrm{T}}\right] \\
=& \mathrm{E}\Big[\big(x_{k+1} - \hat{x}_{k+1|k} - K_{k+1}(Cx_{k+1} + v_{k+1} - C\hat{x}_{k+1|k})\big) \\
& \quad \big(x_{k+1} - \hat{x}_{k+1|k} - K_{k+1}(Cx_{k+1} + v_{k+1} - C\hat{x}_{k+1|k})\big)^{\mathrm{T}}\Big] \\
=& \mathrm{E}\Big[\big((I - K_{k+1}C)(x_{k+1} - \hat{x}_{k+1|k}) - K_{k+1}v_{k+1}\big)\big((I - K_{k+1}C)(x_{k+1} - \hat{x}_{k+1|k}) - K_{k+1}v_{k+1}\big)^{\mathrm{T}}\Big] \\
=& (I - K_{k+1}C)\mathrm{E}\left[(x_{k+1} - \hat{x}_{k+1|k})(x_{k+1} - \hat{x}_{k+1|k})^{\mathrm{T}}\right](I - K_{k+1}C)^{\mathrm{T}} + K_{k+1}\left[v_{k+1}{v_{k+1}}^{\mathrm{T}}\right]{K_{k+1}}^{\mathrm{T}} \\
& - 2\mathrm{E}\left[K_{k+1}v_{k+1}(x_{k+1} - \hat{x}_{k+1|k})^{\mathrm{T}}(I - K_{k+1}C)^{\mathrm{T}}\right] \\
=& (I - K_{k+1}C)P_{k+1|k}(I - K_{k+1}C)^{\mathrm{T}} + K_{k+1}R{K_{k+1}}^{\mathrm{T}} \\
=& P_{k+1|k} + K_{k+1}CP_{k+1|k}C^{\mathrm{T}}{K_{k+1}}^{\mathrm{T}} - K_{k+1}CP_{k+1|k} - P_{k+1|k}C^{\mathrm{T}}{K_{k+1}}^{\mathrm{T}} + K_{k+1}R{K_{k+1}}^{\mathrm{T}}
\end{aligned}$$

Remember that our aim is to find the Kalman gain that minimizes the loss function, i.e.,

$$K_{k+1} = \underset{K}{\mathrm{argmin}}\mathrm{E}\left[(x_{k+1} - \hat{x}_{k+1|k+1})^{\mathrm{T}}(x_{k+1} - \hat{x}_{k+1|k+1})\right]$$

We observe that loss function is equivalent to the trace of the filtered error covariance.

$$\mathrm{E}\left[(x_{k+1} - \hat{x}_{k+1|k+1})^{\mathrm{T}}(x_{k+1} - \hat{x}_{k+1|k+1})\right] = \mathrm{tr}\left[P_{k+1|k+1}\right]$$

Therefore, instead of minimizing the loss function we can minimize the trace of the filtered error covariance. For that purpose, we take the derivative of it with respect to the Kalman gain and set it to zero.

$$\begin{aligned}
\frac{\partial \mathrm{tr}\left[P_{k+1|k+1}\right]}{\partial K_{k+1}} =& 2K_{k+1}(CP_{k+1|k}C^{\mathrm{T}} + R) - 2P^{\mathrm{T}}{}_{k+1|k}C^{\mathrm{T}} \\
=& 0 \\
K_{k+1} =& P_{k+1|k}C^{\mathrm{T}}(CP_{k+1|k}C^{\mathrm{T}} + R)^{-1} \\
=& P_{k+1|k}C^{\mathrm{T}}S^{-1}_{k+1|k}
\end{aligned}$$

We substitute the gain into the update equations

$$\begin{aligned}
\hat{x}_{k+1|k+1} =& \hat{x}_{k+1|k} + P_{k+1|k}C^{\mathrm{T}}S^{-1}_{k+1|k}{}_{k+1}(y_{k+1} - \hat{y}_{k+1|k}) \\
P_{k+1|k+1} =& P_{k+1|k} - P_{k+1|k}C^{\mathrm{T}}(CP_{k+1|k}C^{\mathrm{T}} + R)^{-1}CP_{k+1|k} \\
=& (I - K_{k+1}C)P_{k+1|k}
\end{aligned}$$

Remark:

Both the filtered state and the filtered error covariance estimates depend on the predicted estimates on the previous time step.

Properties of the Kalman Filter

1. The Kalman filter is a recursive algorithm, which is an implication of the model obeying the Markov property. The recursion visits the prediction and measurement update stages one after another till the end.

Kalman Filter Recursion:

Given the initial distribution $p(x_0) = \mathcal{N}(x_0; \hat{x}_{0|0}, P_{0|0})$

Repeat for each k:

- Prediction Update

$$\hat{x}_{k+1|k} = G\hat{x}_{k|k} + Hu_k \tag{18.1}$$

$$P_{k+1|k} = GP_{k|k}G^{\mathrm{T}} + Q \tag{18.2}$$

- Measurement Update

$$\hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + K_{k+1}(y_{k+1} - \hat{y}_{k+1|k}) \tag{18.3}$$

$$P_{k+1|k+1} = P_{k+1|k} - K_{k+1}S_{k+1|k}{K^{\mathrm{T}}}_{k+1} \tag{18.4}$$

where

$$\hat{y}_{k+1|k} = Cx_{k+1|k} \tag{18.5}$$

$$S_{k+1|k} = CP_{k+1|k}C^{\mathrm{T}} + R \tag{18.6}$$

$$K_{k+1} = P_{k+1|k}C^{\mathrm{T}}S^{-1}_{k+1|k} \tag{18.7}$$

2. The Kalman gain and the covariance update equations do not depend on the measurements, hence they can be calculated offline once the model is given.

3. The sufficient statistics to describe a Gaussian distribution are the mean and the covariance. Using Kalman filter, we compute this statistics, thus we make not only point estimations but also produce probability desity function estimates.

4. The Kalman filter is an optimal observer, you can see that (18.3) has the same form of a Luenberger's observer.

5. When we believe that we know the system dynamics quite surely, we keep the process noise covariance matrix, Q, small. Then, the resulting Kalman gain becomes also small that the effect of measurements on the estimates are kept small. Similarly, we can infer that when Q is high, the measurements are incorporated into the estimates to a bigger extent, as we have more trust on the measurements than the system model.

6. When we think that our sensor is highly noisy, we set the measurement noise covariance, R, to a high value, as a result the Kalman gain gets smaller. Therefore, estimates mostly depend on the predictions. In the case of a smaller R, the measurements affect the estimates more due to having a larger Kalman gain.

Example: 2-D random walk with scalar measurement. Consider the following system

$$\begin{aligned} x_{k+1} &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x_k + w_k \\ y_k &= \begin{bmatrix} 1 & 1 \end{bmatrix} + v_k \end{aligned}$$

where $w_k \sim \mathcal{N}(w_k; 0, Q)$, $v_k \sim \mathcal{N}(v_k; 0, 0.4)$, and $x_{0|0} \sim \mathcal{N}(x_0; 0, \mathrm{I}_{2\times 2})$ with $Q = 0.1 \cdot \mathrm{I}_{2\times 2}$. Given the measurements $y_1 = 1$ and $y_2 = -1.5$, compute the filtered mean and covariance estimates for the first two time instants.

Solution:

- When $k = 1$:

  Prediction Update:

$$\begin{aligned} x_{1|0} =& G x_{0|0} \\ =& \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ =& \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ P_{1|0} =& G P_{0|0} G^{\mathrm{T}} + Q \\ =& \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \\ =& \begin{bmatrix} 1.1 & 0 \\ 0 & 1.1 \end{bmatrix} \end{aligned}$$

  Measurement Update:

$$\begin{aligned} S_{1|0} =& C P_{1|0} C^{\mathrm{T}} + R \\ =& \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1.1 & 0 \\ 0 & 1.1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 0.4 \\ =& 2.5 \\ K_1 =& P_{1|0} C^{\mathrm{T}} S_{1|0}^{-1} \\ =& \frac{1}{2.5} \begin{bmatrix} 1.1 & 0 \\ 0 & 1.1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ =& \begin{bmatrix} 0.44 \\ 0.44 \end{bmatrix} \\ \hat{y}_1 =& C x_{1|0} \\ =& \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ =& 0 \end{aligned}$$

$$\begin{aligned}
x_{1|1} =& x_{1|0} + K_1(y_1 - \hat{y}_1) \\
=& \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.44 \\ 0.44 \end{bmatrix} (1 - 0) \\
=& \begin{bmatrix} 0.44 \\ 0.44 \end{bmatrix} \\
P_{1|1} =& P_{1|0} - K_1 S_{1|0} K_1^{\mathrm{T}} \\
=& \begin{bmatrix} 1.1 & 0 \\ 0 & 1.1 \end{bmatrix} - 2.5 \begin{bmatrix} 0.44 \\ 0.44 \end{bmatrix} \begin{bmatrix} 0.44 & 0.44 \end{bmatrix} \\
=& \begin{bmatrix} 0.62 & -0.48 \\ -0.48 & 0.62 \end{bmatrix}
\end{aligned}$$

- When $k = 2$:

  Prediction Update:

$$\begin{aligned}
x_{2|1} =& G x_{1|1} \\
=& \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.44 \\ 0.44 \end{bmatrix} \\
=& \begin{bmatrix} 0.44 \\ 0.44 \end{bmatrix} \\
P_{2|1} =& G P_{1|1} G^{\mathrm{T}} + Q \\
=& \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.62 & -0.48 \\ -0.48 & 0.62 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \\
=& \begin{bmatrix} 0.72 & -0.48 \\ -0.48 & 0.72 \end{bmatrix}
\end{aligned}$$

  Measurement Update:

$$\begin{aligned}
S_{2|1} =& C P_{2|1} C^{\mathrm{T}} + R \\
=& \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 0.72 & -0.48 \\ -0.48 & 0.72 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 0.4 \\
=& 0.88 \\
K_2 =& P_{2|1} C^{\mathrm{T}} S_{2|1}^{-1} \\
=& \frac{1}{0.88} \begin{bmatrix} 0.72 & -0.48 \\ -0.48 & 0.72 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\
=& \begin{bmatrix} 0.27 \\ 0.27 \end{bmatrix} \\
\hat{y}_2 =& C x_{2|1} \\
=& \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 0.44 \\ 0.44 \end{bmatrix} \\
=& 0.88
\end{aligned}$$

$$\begin{aligned}
x_{2|2} =& x_{2|1} + K_2(y_2 - \hat{y}_2) \\
=& \begin{bmatrix} 0.44 \\ 0.44 \end{bmatrix} + \begin{bmatrix} 0.27 \\ 0.27 \end{bmatrix} (-1.5 - 0.88) \\
=& \begin{bmatrix} 0.2 \\ 0.2 \end{bmatrix} \\
P_{2|2} =& P_{2|1} - K_2 S_{2|1} K_2^{\mathrm{T}} \\
=& \begin{bmatrix} 0.72 & -0.48 \\ -0.48 & 0.72 \end{bmatrix} - 0.88 \begin{bmatrix} 0.27 \\ 0.27 \end{bmatrix} \begin{bmatrix} 0.27 & 0.27 \end{bmatrix} \\
=& \begin{bmatrix} 0.66 & -0.54 \\ -0.54 & 0.66 \end{bmatrix}
\end{aligned}$$