EEET 4071 Advanced Control (2021)

Supplementary Lecture D: Deriving the Kalman filter

Dr. Yee Wei Law ⟨yeewei.law@unisa.edu.au⟩

The purpose of this supplementary lecture is to provide a Bayesian perspective to the Kalman filter and a derivation of the algorithm. The derivation should shed some useful insight into the working of the algorithm.

1 State estimation as a Bayesian filtering problem

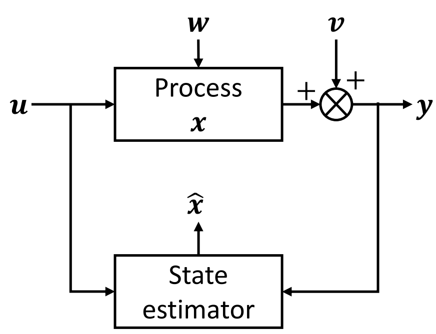

Figure 1: State estimation in the presence of process noise $\boldsymbol{w}$ and measurement noise $\boldsymbol{v}$.

State estimation (see Figure 1) is an area where control meets signal processing, machine learning, and other statistically based disciplines. In fact, state estimation can be viewed as a *Bayesian filtering* problem. The Bayesian perspective of state estimation is especially useful because it highlights the stochastic nature of the problem. The adjective "Bayesian" refers to one of the most fundamental theorems in statistics: Bayes' theorem. Given two *random variables* X and Y, representing "cause" and "effect" respectively, Bayes' theorem states

$$p_X(x|Y=y) = \frac{p_Y(y|X=x)p_X(x)}{p_Y(y)}, \tag{1}$$

or in a more user-friendly way,

$$p(\text{cause}|\text{effect}) = \frac{p(\text{effect}|\text{cause})p(\text{cause})}{p(\text{effect})},$$

where

- p_X denotes the probability density function (pdf) of random variable X;
- $p_X(x) \stackrel{\text{def}}{=} p_X(X=x)$ is called the *prior pdf* (note X is a random variable, x is a value);
- $p_Y(y|X=x)$ is called the *likelihood*;
- $p_X(x|Y=y)$ is called the *posterior pdf*;
- $p_Y(y)$ is called the *normalization constant*.

Detail: Prior and posterior

In the literature, it is common to use "posterior" and "prior" interchangeably with their Latin equivalents, which are "a posteriori" and "a priori" respectively.

Essentially, Bayes' theorem solves a *statistical inverse problem* (i.e., the statistical version of $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$). For state estimation, we usually deal with vectors, in which case Eq. (1) is written as

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}, \tag{2}$$

omitting the subscripts and the symbols for random vectors, i.e., $p(\boldsymbol{x}|\boldsymbol{y})$ actually means $p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{Y} = \boldsymbol{y})$.

To apply Bayes' theorem to state estimation, suppose we take $\boldsymbol{x}$ as a particular system state value, and $\boldsymbol{y}$ as a particular observed output value, as shown in Figure 1. Then,

- The *prior probability* is the probability of having state $\boldsymbol{x}$.
- The *likelihood* is the pdf of the observed output if the state is $\boldsymbol{x}$.
- The *posterior probability* gives us the probability of having state $\boldsymbol{x}$ when the observed output is $\boldsymbol{y}$.

Eq. (2) allows us to calculate the pdf of the current state given the current observed output, *if* the current observed output depends only on the current state, *and* the current state is independent of all preceding states and inputs — this might not reflect reality.

Let us extend Eq. (2) to take into account the *entire history of system input and output*, we can write

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k}, \boldsymbol{u}_{1:k}) = \frac{p(\boldsymbol{y}_{1:k}|\boldsymbol{x}_k, \boldsymbol{u}_{1:k})p(\boldsymbol{x}_k, \boldsymbol{u}_{1:k})}{p(\boldsymbol{y}_{1:k}, \boldsymbol{u}_{1:k})}, \tag{3}$$

where the subscript "$1:k$" means "the first to the kth time interval". Removing the condition $\boldsymbol{u}_{1:k}$ which is common to both sides of the equation above, we get

$$p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k}) = \frac{p(\boldsymbol{y}_{1:k}|\boldsymbol{x}_k)p(\boldsymbol{x}_k)}{p(\boldsymbol{y}_{1:k})}. \tag{4}$$

Note that $\boldsymbol{y}_{1:k}$ is a matrix whose number of columns increases with time. At some point, the matrix will become so large that the calculation of the conditional probabilities become intractable. Consequently, some assumptions need to be made to simplify the calculation of the posterior pdf [Sär13, p. 10].

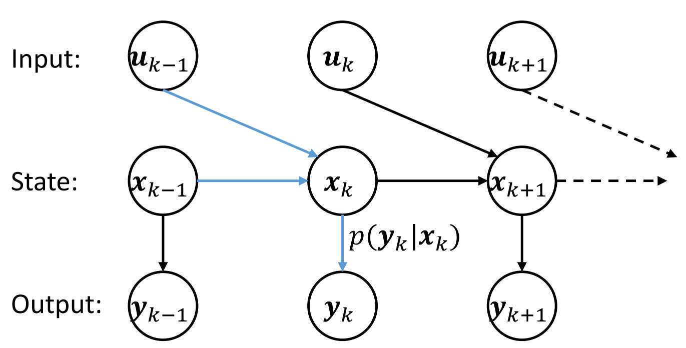

Figure 2: Assumptions for simplifying the calculation of $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})$.

For the assumptions, consider the discrete-time linear state equation: $\boldsymbol{x}_k = \mathbf{F}_{k-1}\boldsymbol{x}_{k-1} + \mathbf{G}_{k-1}\boldsymbol{u}_{k-1}$. This equation suggests

Assumption 1

The current state depends only on the previous state and previous input, i.e., the system is Markovian.

Next, consider the discrete-time linear output equation: $\boldsymbol{y}_k = \mathbf{H}_k \boldsymbol{x}_k$. This equation suggests

Assumption 2

The current output depends only on the current state.

The assumptions are illustrated in Figure 2, Applying the assumptions to Eq. (4), we get

$$\begin{aligned}
p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k}) &= \frac{p(\boldsymbol{y}_{1:k}|\boldsymbol{x}_k)p(\boldsymbol{x}_k)}{p(\boldsymbol{y}_{1:k})} \\
&= \frac{p(\boldsymbol{y}_k, \boldsymbol{y}_{1:(k-1)}|\boldsymbol{x}_k)p(\boldsymbol{x}_k)}{p(\boldsymbol{y}_k, \boldsymbol{y}_{1:(k-1)})} \qquad \text{(splitting } \boldsymbol{y}_{1:k} \text{ into } \boldsymbol{y}_k, \boldsymbol{y}_{1:(k-1)}\text{)} \\
&= \frac{p(\boldsymbol{y}_k|\boldsymbol{y}_{1:(k-1)}, \boldsymbol{x}_k)p(\boldsymbol{y}_{1:(k-1)}|\boldsymbol{x}_k)p(\boldsymbol{x}_k)}{p(\boldsymbol{y}_k|\boldsymbol{y}_{1:(k-1)})p(\boldsymbol{y}_{1:(k-1)})} \\
&= \frac{p(\boldsymbol{y}_k|\boldsymbol{y}_{1:(k-1)}, \boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{y}_{1:(k-1)})p(\boldsymbol{y}_{1:(k-1)})p(\boldsymbol{x}_k)}{p(\boldsymbol{y}_k|\boldsymbol{y}_{1:(k-1)})p(\boldsymbol{x}_k)p(\boldsymbol{y}_{1:(k-1)})} \qquad \text{(Bayes' theorem)} \\
&= \frac{p(\boldsymbol{y}_k|\boldsymbol{y}_{1:(k-1)}, \boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{y}_{1:(k-1)})}{p(\boldsymbol{y}_k|\boldsymbol{y}_{1:(k-1)})} \qquad \text{(cancelling out common factors)} \\
&= \frac{p(\boldsymbol{y}_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{y}_{1:(k-1)})}{\int p(\boldsymbol{y}_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{y}_{1:(k-1)})\,\mathrm{d}\boldsymbol{x}_k} \qquad (\because \text{Assumption 2}).
\end{aligned} \tag{5}$$

Above, the conditional probability $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:(k-1)})$ is given by the *Chapman-Kolmogorov equation* [RAG04, p. 5]:

$$\begin{aligned}
p(\boldsymbol{x}_k|\boldsymbol{y}_{1:(k-1)}) &= \int p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{u}_{k-1})p(\boldsymbol{x}_{k-1}, \boldsymbol{u}_{k-1}|\boldsymbol{y}_{1:(k-1)})\,\mathrm{d}\boldsymbol{x}_{k-1} \qquad (\because \text{Assumption 1}) \\
&= \int p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{u}_{k-1})p(\boldsymbol{x}_{k-1}|\boldsymbol{y}_{1:(k-1)})\,\mathrm{d}\boldsymbol{x}_{k-1}.
\end{aligned} \tag{6}$$

Eqs. (5)–(6) are called the *Bayesian filtering equations* [Sär13, Theorem 4.1]. The equations show that $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})$ can be calculated from $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:(k-1)})$, which in turn can be calculated from $p(\boldsymbol{x}_{k-1}|\boldsymbol{y}_{1:(k-1)})$, which in turn can be calculated from $p(\boldsymbol{x}_{k-1}|\boldsymbol{y}_{1:(k-2)})$, and so on and so forth until the calculation reaches $p(\boldsymbol{x}_0|\boldsymbol{y}_{-1}) = p(\boldsymbol{x}_0)$, since $\boldsymbol{y}_{-1}$ does not exist. Thus, there is a recursive relationship between the posterior pdf $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})$ and the prior pdf $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:(k-1)})$, that forms the basis of Bayesian (and hence Kalman) filtering. Once the posterior pdf is obtained, there are several statistically meaningful ways of extracting an estimate of $\boldsymbol{x}_k$, denoted $\hat{\boldsymbol{x}}_k$ from the posterior pdf:

- The *conditional mean estimate* [HL64] is defined as

$$\hat{\boldsymbol{x}}_{k|k}^{\mathrm{CM}} \stackrel{\mathrm{def}}{=} \mathsf{E}\left\{\boldsymbol{x}_k|\boldsymbol{y}_{1:k}\right\} = \int \boldsymbol{x}_k p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})\,\mathrm{d}\boldsymbol{x}_k, \tag{7}$$

  where E is the expectation operator, and the subscript "$k|k$" means the "the kth state given k observed outputs".

- The *maximum a posteriori estimate* (MAP estimate) is defined as

$$\hat{\boldsymbol{x}}_{k|k}^{\mathrm{MAP}} \stackrel{\mathrm{def}}{=} \arg\max_{\boldsymbol{x}_k} p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k}). \tag{8}$$

$\hat{\boldsymbol{x}}_{k|k}^{\mathrm{MAP}}$ corresponds to the *mode* of $\boldsymbol{x}_k$ (see Figure 3). MAP estimation is problematic when $p(\boldsymbol{x}_k|\boldsymbol{y}_{1:k})$ has multiple local maxima.

- The *minimax estimate* is defined as

$$\hat{\boldsymbol{x}}_{k|k}^{\mathrm{MM}} \stackrel{\mathrm{def}}{=} \arg\min_{\hat{\boldsymbol{x}}_{k|k}} \max \|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_{k|k}\|. \tag{9}$$

$\hat{\boldsymbol{x}}_{k|k}^{\mathrm{MM}}$ corresponds to the *median* of $\boldsymbol{x}_k$ (see Figure 3).

- The *maximum likelihood estimate* is equivalent to the MAP estimate, except that the prior distribution is assumed to be a uniform distribution.

- Other estimates include the minimum conditional inaccuracy, minimum conditional Kullback-Leibler divergence, and minimum free energy estimates [Che03, p. 10].

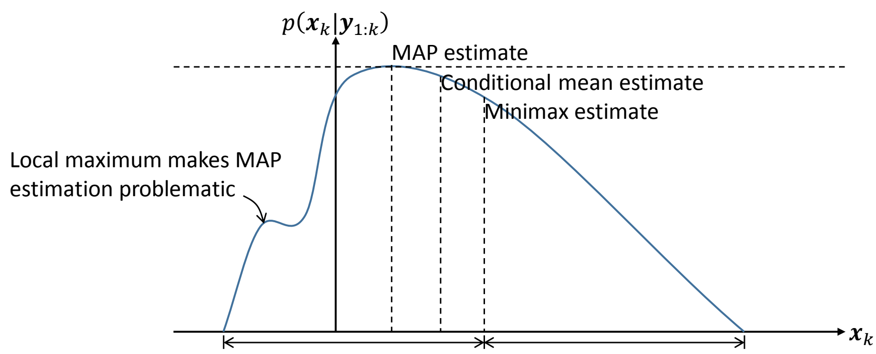

Figure 3: Different estimates based on the posterior pdf.

Among the different kinds of estimates above, the conditional mean estimate is *optimal* in the *linear Gaussian model*, in terms of the variance of the estimation error (equivalently, mean squared error).

Linear Gaussian model [Sär13, p. 37]

- The system is linear and stochastic.
- The conditional probabilities $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \boldsymbol{u}_{k-1})$ and $p(\boldsymbol{y}_k|\boldsymbol{x}_k)$ are Gaussian, or equivalently, the initial state $\boldsymbol{x}_0$, process noise $\boldsymbol{w}$, and measurement noise $\boldsymbol{v}$ (see Figure 1) follow some Gaussian distributions. Note the affine combination of Gaussian random vectors is also Gaussian.

In other words, in the linear Gaussian model, the conditional mean estimate is also the the *minimum variance estimate* or *minimum mean squared error estimate* (MMSE estimate):

$$\hat{\boldsymbol{x}}_{k|k}^{\mathrm{CM}} = \hat{\boldsymbol{x}}_{k|k}^{\mathrm{MMSE}} \stackrel{\mathrm{def}}{=} \arg\min_{\hat{\boldsymbol{x}}_{k|k}} \mathsf{E}\left\{\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_{k|k}\|^2|\boldsymbol{y}_{1:k}\right\}. \tag{10}$$

This model is exactly the basis of the Kalman filter, which provides MMSE estimates of system states. Before we discuss the model in detail, let us revise some essential statistical concepts.

2 Deriving the Kalman filter

The Kalman filter algorithm provided in Lecture 7 is repeated below:

Kalman filter

Initialization ($k = 0$):

$$\hat{\boldsymbol{x}}_0^+ = \mathsf{E}\{\boldsymbol{x}_0\}, \quad \mathbf{P}_0^+ = \mathsf{E}\left\{(\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0^+)(\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0^+)^\top\right\}. \tag{11}$$

Time update / prediction update equations for propagating *a priori* estimates ($k = 1, 2, \ldots$):

$$\hat{\boldsymbol{x}}_k^- = \mathbf{F}_{k-1}\hat{\boldsymbol{x}}_{k-1}^+ + \mathbf{G}_{k-1}\boldsymbol{u}_{k-1}, \tag{12}$$

$$\mathbf{P}_k^- = \mathbf{F}_{k-1}\mathbf{P}_{k-1}^+\mathbf{F}_{k-1}^\top + \mathbf{Q}_{k-1}. \tag{13}$$

Kalman gain update:

$$\mathbf{K}_k = \mathbf{P}_k^-\mathbf{H}_k^\top(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^\top + \mathbf{R}_k)^{-1}. \tag{14}$$

Measurement update equations for updating *a posteriori* estimates:

$$\hat{\boldsymbol{x}}_k^+ = \hat{\boldsymbol{x}}_k^- + \mathbf{K}_k(\boldsymbol{y}_k - \mathbf{H}_k\hat{\boldsymbol{x}}_k^-) = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\hat{\boldsymbol{x}}_k^- + \mathbf{K}_k\boldsymbol{y}_k, \tag{15}$$

Joseph stabilized version: $\mathbf{P}_k^+ = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^-(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)^\top + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^\top$. (16)

Standard form: $\mathbf{P}_k^+ = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^-$. (17)

Information filter form: $\mathbf{P}_k^+ = [(\mathbf{P}_k^-)^{-1} + \mathbf{H}_k^\top\mathbf{R}_k^{-1}\mathbf{H}_k]^{-1}$. (18)

In the linear Gaussian model,

- $\boldsymbol{x}_k$ is Gaussian.
- Since $\boldsymbol{x}_k$ and $\boldsymbol{v}_k$ are Gaussian, $\boldsymbol{y}_k = \mathbf{H}\boldsymbol{x}_k + \boldsymbol{v}_k$ is Gaussian.
- In fact, $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ are jointly Gaussian [AM79, p. 18].

Consequently, the conditional mean estimator, which is a linear estimator [AM79, Theorem 2.1], is also an MMSE estimator [AM79, Theorem 2.2]. Furthermore, the MMSE estimator is unbiased [AM79, Theorem 2.3], i.e., the estimation error is zero on average.

Detail: Linear filter

While MMSE estimators are generally nonlinear in terms of the observations, it is important to realize that the Kalman filter is a linear (more accurately, *affine*) filter. In general, for any jointly distributed random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, the conditional mean estimator $\mathsf{E}\{\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}\}$ can be expressed as some affine function of $\boldsymbol{y}$:

$$\mathbf{A}\boldsymbol{y} + \boldsymbol{b},$$

for some matrix $\mathbf{A}$ and vector $\boldsymbol{b}$, which are functions of the means of $\boldsymbol{X}$ and $\boldsymbol{Y}$, and covariances between $\boldsymbol{X}$ and $\boldsymbol{Y}$ [AM79, Theorem 2.1].

The facts we have collected so far culminate in the following paramount theorem for the derivation of the Kalman filter:

Theorem 1: Projection theorem (aka orthogonality principle) [AM79, Theorem 2.5]

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are jointly Gaussian, and $\hat{\boldsymbol{x}} = \mathsf{E}\{\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}\}$ is a linear MMSE estimate, then the error $(\boldsymbol{X} - \hat{\boldsymbol{x}})$ and observation $\boldsymbol{y}$ are mutually orthogonal, i.e.,

$$\mathsf{E}\left\{(\boldsymbol{X} - \hat{\boldsymbol{x}})\boldsymbol{y}^\top\right\} = \mathbf{0}. \tag{19}$$

For deriving the Kalman filter, this is the strategy we are following:

1. We start by defining the posterior estimate $\hat{\boldsymbol{x}}_k^+$ as a linear MMSE estimate, and expressing it as a *linear combination* of the prior estimate $\hat{\boldsymbol{x}}_k^-$ and the new measurement $\boldsymbol{y}_k$:

$$\hat{\boldsymbol{x}}_k^+ = \mathbf{L}_k \hat{\boldsymbol{x}}_k^- + \mathbf{K}_k \boldsymbol{y}_k, \tag{20}$$

where the weights $\mathbf{L}_k$ and $\mathbf{K}_k$ are to be determined.

Quiz 1

Why is the input $\boldsymbol{u}_k$ not included in Eq. (20)?

2. We then apply the orthogonality principle to determine $\mathbf{L}_k$ and $\mathbf{K}_k$. The orthogonality principle suggests

$$\mathsf{E}\left\{\boldsymbol{e}_k^+ \boldsymbol{y}_i^\top\right\} = \mathbf{0}, \qquad \forall i = 1, 2, \ldots, k-1, \tag{21}$$

where $\boldsymbol{e}_k^+ \stackrel{\text{def}}{=} \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k^+$. In other words,

$$\begin{aligned}
\mathsf{E}\left\{\boldsymbol{e}_k^+ \boldsymbol{y}_i^\top\right\} &= \mathsf{E}\left\{\left(\boldsymbol{x}_k - \mathbf{L}\hat{\boldsymbol{x}}_k^- - \mathbf{K}_k \mathbf{H}_k \boldsymbol{x}_k - \mathbf{K}_k \boldsymbol{v}_k\right) \boldsymbol{y}_i^\top\right\} \\
&= \mathsf{E}\left\{\left(\mathbf{I} - \mathbf{K}_k \mathbf{H}_k - \mathbf{L}_k\right) \boldsymbol{x}_k \boldsymbol{y}_i^\top + \mathbf{L}_k \left(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k^-\right) \boldsymbol{y}_i^\top\right\} \\
&= \left(\mathbf{I} - \mathbf{K}_k \mathbf{H}_k - \mathbf{L}_k\right) \mathsf{E}\left\{\boldsymbol{x}_k \boldsymbol{y}_i^\top\right\} = \mathbf{0}.
\end{aligned}$$

For the preceding equation to hold for arbitrary values of $\boldsymbol{x}_k$ and $\boldsymbol{y}_i$,

$$\mathbf{I} - \mathbf{K}_k \mathbf{H}_k - \mathbf{L}_k = \mathbf{0} \implies \mathbf{L}_k = \mathbf{I} - \mathbf{K}_k \mathbf{H}_k. \tag{22}$$

Substituting (22) into (20) gives us the time-update equation for $\hat{\boldsymbol{x}}_k^+$:

$$\hat{\boldsymbol{x}}_k^+ = \left(\mathbf{I} - \mathbf{K}_k \mathbf{H}_k\right) \hat{\boldsymbol{x}}_k^- + \mathbf{K}_k \boldsymbol{y}_k = \hat{\boldsymbol{x}}_k^- + \mathbf{K}_k \left(\boldsymbol{y}_k - \mathbf{H}_k \hat{\boldsymbol{x}}_k^-\right), \tag{23}$$

where $\mathbf{K}_k$ is called the **Kalman gain**; $(\boldsymbol{y}_k - \mathbf{H}_k \hat{\boldsymbol{x}}_k^-)$ is called an **innovation**. We still need to derive an expression for $\mathbf{K}_k$.

Earlier when we invoked the orthogonality principle:

$$\mathsf{E}\left\{\boldsymbol{e}_k^+ \boldsymbol{y}_i^\top\right\} = \mathbf{0}, \qquad \forall i = 1, 2, \ldots, k-1, \tag{24}$$

we did not consider the case $i = k$, which we do now:

$$\mathsf{E}\left\{\boldsymbol{e}_k^+ \boldsymbol{y}_k^\top\right\} = \mathbf{0} \qquad \text{and furthermore} \qquad \mathsf{E}\left\{\boldsymbol{e}_k^+ \hat{\boldsymbol{y}}_k^\top\right\} \approx \mathbf{0}, \tag{25}$$

where $\hat{\boldsymbol{y}}_k \stackrel{\text{def}}{=} \mathbf{H}_k \hat{\boldsymbol{x}}_k^-$ is the predicted value of $\boldsymbol{y}_k$ based on $\hat{\boldsymbol{x}}_k^-$. Changing $\approx$ in Eq. (25) to equality gives us

$$\mathsf{E}\left\{\boldsymbol{e}_k^+ \hat{\boldsymbol{y}}_k^\top\right\} = \mathsf{E}\left\{\boldsymbol{e}_k^+ \left(\mathbf{H}_k \hat{\boldsymbol{x}}_k^-\right)^\top\right\} = \mathbf{0}. \tag{26}$$

Therefore,

$$\mathsf{E}\left\{\boldsymbol{e}_k^+ \boldsymbol{y}_k^\top\right\} - \mathsf{E}\left\{\boldsymbol{e}_k^+ \hat{\boldsymbol{y}}_k^\top\right\} = \mathsf{E}\left\{\boldsymbol{e}_k^+ \left(\mathbf{H}_k \boldsymbol{e}_k^- + \boldsymbol{v}_k\right)^\top\right\} = \mathbf{0}, \tag{27}$$

where $\boldsymbol{e}_k^- \stackrel{\text{def}}{=} \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k^-$. We can simplify the above expression further by expressing $\boldsymbol{e}_k^+$ in terms of $\boldsymbol{e}_k^-$:

$$\begin{aligned}
\boldsymbol{e}_k^+ &= \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k^- - \mathbf{K}_k \boldsymbol{y}_k + \mathbf{K}_k \mathbf{H}_k \hat{\boldsymbol{x}}_k^- \\
&= \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k^- - \mathbf{K}_k \mathbf{H}_k \boldsymbol{x}_k - \mathbf{K}_k \boldsymbol{v}_k + \mathbf{K}_k \mathbf{H}_k \hat{\boldsymbol{x}}_k^- \\
&= \left(\mathbf{I} - \mathbf{K}_k \mathbf{H}_k\right) \boldsymbol{e}_k^- - \mathbf{K}_k \boldsymbol{v}_k.
\end{aligned} \tag{28}$$

Substituting (28) into (27) gives us

$$\begin{aligned}
\mathsf{E}\left\{\boldsymbol{e}_k^+(\mathbf{H}_k\boldsymbol{e}_k^- + \boldsymbol{v}_k)^\top\right\} &= \mathsf{E}\left\{\left[(\mathbf{I}-\mathbf{K}_k\mathbf{H}_k)\boldsymbol{e}_k^- - \mathbf{K}_k\boldsymbol{v}_k\right]\left[(\boldsymbol{e}_k^-)^\top\mathbf{H}_k^\top + \boldsymbol{v}_k^\top\right]\right\} \\
&= \mathsf{E}\left\{(\mathbf{I}-\mathbf{K}_k\mathbf{H}_k)\boldsymbol{e}_k^-(\boldsymbol{e}_k^-)^\top\mathbf{H}_k^\top - \mathbf{K}_k\boldsymbol{v}_k\boldsymbol{v}_k^\top\right\} \\
&= (\mathbf{I}-\mathbf{K}_k\mathbf{H}_k)\,\mathsf{E}\left\{\boldsymbol{e}_k^-(\boldsymbol{e}_k^-)^\top\right\}\mathbf{H}_k^\top - \mathbf{K}_k\,\mathsf{E}\left\{\boldsymbol{v}_k\boldsymbol{v}_k^\top\right\} \\
&= (\mathbf{I}-\mathbf{K}_k\mathbf{H}_k)\,\mathsf{E}\left\{\boldsymbol{e}_k^-(\boldsymbol{e}_k^-)^\top\right\}\mathbf{H}_k^\top - \mathbf{K}_k\mathbf{R}_k \\
&= \mathbf{0}.
\end{aligned} \tag{29}$$

At this point, we identify $\mathsf{E}\left\{\boldsymbol{e}_k^-(\boldsymbol{e}_k^-)^\top\right\}$ as the prior error covariance, hence

$$(\mathbf{I}-\mathbf{K}_k\mathbf{H}_k)\,\mathbf{P}_k^-\mathbf{H}_k^\top - \mathbf{K}_k\mathbf{R}_k = \mathbf{P}_k^-\mathbf{H}_k^\top - \mathbf{K}_k(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^\top + \mathbf{R}_k) = \mathbf{0}. \tag{30}$$

$$\therefore \mathbf{K}_k = \mathbf{P}_k^-\mathbf{H}_k^\top(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^\top + \mathbf{R}_k)^{-1}. \tag{31}$$

3. Having identified the gains $\mathbf{L}_k$ and $\mathbf{K}_k$ in Eq. (20), we can now relate the posterior state estimate $\hat{\boldsymbol{x}}_k^+$ to the prior state estimate $\hat{\boldsymbol{x}}_k^-$, and on top of that, the posterior error covariance $\mathbf{P}_k^+$ to the prior error covariance $\mathbf{P}_k^-$:

$$\mathbf{P}_k^+ = \mathsf{E}\left\{\boldsymbol{e}_k^+(\boldsymbol{e}_k^+)^\top\right\} = \mathsf{E}\left\{\left[(\mathbf{I}-\mathbf{K}_k\mathbf{H}_k)\boldsymbol{e}_k^- - \mathbf{K}_k\boldsymbol{v}_k\right]\left[(\boldsymbol{e}_k^-)^\top(\mathbf{I}-\mathbf{K}_k\mathbf{H}_k)^\top - \boldsymbol{v}_k^\top\mathbf{K}_k^\top\right]\right\}. \tag{32}$$

$$\therefore \mathbf{P}_k^+ = (\mathbf{I}-\mathbf{K}_k\mathbf{H}_k)\,\mathbf{P}_k^-\,(\mathbf{I}-\mathbf{K}_k\mathbf{H}_k)^\top + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^\top. \tag{33}$$

The above expression for $\mathbf{P}_k^+$ is called the **Joseph stabilized version** or **Joseph form** of the covariance update equation, and it is more numerically robust than Eqs. (17) and (18). We now have all the expressions necessary for computing posterior estimates from prior estimates.

4. We now look at computing prior estimates from posterior estimates, i.e., $\hat{\boldsymbol{x}}_k^-$ from $\hat{\boldsymbol{x}}_{k-1}^+$; $\mathbf{P}_k^-$ from $\mathbf{P}_{k-1}^+$. Bear in mind for this, $\boldsymbol{y}_k$ is not yet available. We simply apply the plant model to obtain

$$\hat{\boldsymbol{x}}_k^- = \mathbf{F}_{k-1}\hat{\boldsymbol{x}}_{k-1}^+ + \mathbf{G}_{k-1}\boldsymbol{u}_{k-1}. \tag{34}$$

To relate $\mathbf{P}_k^-$ to $\mathbf{P}_{k-1}^+$, we observe

$$\begin{aligned}
\boldsymbol{e}_k^- &= \boldsymbol{x}_k - \mathbf{F}_{k-1}\hat{\boldsymbol{x}}_{k-1}^+ - \mathbf{G}_{k-1}\boldsymbol{u}_{k-1} \\
&= \mathbf{F}_{k-1}\boldsymbol{x}_{k-1} + \mathbf{G}_{k-1}\boldsymbol{u}_{k-1} + \boldsymbol{w}_{k-1} - \mathbf{F}_{k-1}\hat{\boldsymbol{x}}_{k-1}^+ - \mathbf{G}_{k-1}\boldsymbol{u}_{k-1} \\
&= \mathbf{F}_{k-1}\boldsymbol{e}_{k-1}^+ + \boldsymbol{w}_{k-1}.
\end{aligned} \tag{35}$$

Consequently,

$$\mathbf{P}_k^- = \mathsf{E}\left\{\boldsymbol{e}_k^-(\boldsymbol{e}_k^-)^\top\right\} = \mathbf{F}_{k-1}\mathbf{P}_{k-1}^+\mathbf{F}_{k-1}^\top + \mathbf{Q}_{k-1}. \tag{36}$$

This concludes our derivation of the Kalman filter. References providing alternative derivatives include [AM79, Sects. 3.1, 5.4], [IN01, Sect. 2.5.1], [Lin07, Sect. 4.5], [Hay09, Sect. 14.3], [Gu12, Sect. 5.1.2].

References

[AM79] B. D. Anderson and J. B. Moore, *Optimal Filtering*, Prentice-Hall, Inc., 1979.

[Che03] Z. Chen, *Bayesian Filtering: From Kalman Filters to Particle Filters, and Beyond*, Tech. report, McMaster University, 2003.

[Gu12] G. Gu, *Discrete-Time Linear Systems: Theory and Design with Applications*, Springer, 2012.

[Hay09] S. S. Haykin, *Neural networks and learning machines*, 3rd ed., Pearson Education, Inc., 2009.

[HL64] Y. Ho and R. Lee, A Bayesian approach to problems in stochastic estimation and control, *IEEE Transactions on Automatic Control* **9** no. 4 (1964), 333–339. https://doi.org/10.1109/TAC.1964.1105763.

[IN01] E. Ikonen and K. Najim, *Advanced process identification and control*, CRC Press, 2001.

[Lin07] F. Lin, *Robust Control Design: An Optimal Control Approach*, John Wiley & Sons Ltd., 2007.

[RAG04] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman filter: Particle filters for tracking applications*, **685**, Artech house Boston, 2004.

[Sär13] S. Särkkä, *Bayesian Filtering and Smoothing*, Cambridge University Press, 2013.