Chapter 4

State estimation

This chapter addresses the problem of state estimation for stochastic systems. First, linear systems will be considered. In the second part of the chapter, several approaches for state estimation of nonlinear stochastic systems will be discussed.

4.1 State space representation of stochastic systems

The *state variables* (or simply *the state*) of a dynamic system are the variables that one needs to know at a generic time t, to determine the evolution of the system at all future times $\tau > t$, provided that the input signal $u(\tau)$, $\tau > t$, is known. For deterministic discrete-time systems, this allows one to write the input-state-output (i/s/o) representation

$$\begin{cases} x(t+1) = f(x(t), u(t), t) \\ y(t) = h(x(t), u(t), t) \end{cases} \tag{4.1}$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^m$ is the input and $y(t) \in \mathbb{R}^p$ is the output. For linear time-invariant (LTI) systems, model (4.1) boils down

to

$$\begin{cases} x(t+1) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases}$$

where A, B, C, D are constant matrices of suitable dimensions.

When the involved signals are stochastic processes, the system becomes a *stochastic system* and the state can be interpreted in terms of an important class of stochastic processes.

Definition 4.1 (Markov process)**.** The s.p. $x(t)$ is a Markov process if, given a sequence of time instants $t_1 < t_2 < \cdots < t_k < t_{k+1}$, then the conditional pdf of $x(t)$ is such that

$$f_{\boldsymbol{x}}\left(x(t_{k+1})|x(t_k), x(t_{k-1}), \cdots, x(t_1)\right) = f_{\boldsymbol{x}}\left(x(t_{k+1})|x(t_k)\right)$$

The above definition states that for a Markov process the conditional pdf with respect to past observations is equal to the conditional pdf with respect to the most recent one. In other words, it is not necessary to keep track of all past observations, because the state vector $x(t_k)$ contains all the necessary information to compute the a posteriori pdf of the next state $x(t_{k+1})$.

Hereafter, we consider the following general class of systems:

$$\begin{cases} x(t+1) = f(x(t), u(t), w(t)) \\ y(t) = h(x(t), u(t)) + v(t) \end{cases} \tag{4.2}$$

where $w(t) \in \mathbb{R}^d$ and $v(t) \in \mathbb{R}^p$ are stochastic processes, on which we make the following assumption.

Assumption 4.1. *For system* (4.2)*, we assume that:*

i) $\mathbf{E}\left[\begin{pmatrix} w(t) \\ v(t) \end{pmatrix}\right] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$*;*

ii) $\mathbf{E}\left[\begin{pmatrix} w(t) \\ v(t) \end{pmatrix}\begin{pmatrix} w(t) \\ v(t) \end{pmatrix}^T\right] = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix}$ *where Q and R are the covariance matrices of $w(t)$ and $v(t)$, respectively;*

iii) $w(t)$ *and* $v(t)$ *are independent white processes;*

iv) $x(0)$ *is a random vector, independent from* $w(t)$ *and* $v(t)$*, with mean* m_0 *and covariance matrix* P_0*;*

v) $u(t)$ *is a deterministic (known) signal.*

The s.p. $w(t)$ is often referred to as *disturbance process* and models the stochastic component of the dynamic model $f(\cdot)$ (unmodeled dynamics, disturbances, etc.). The s.p. $v(t)$ is the so-called *measurement noise*, which represents the error of the sensor measuring the output function $h(\cdot)$. It is easy to see that under Assumption 4.1, the state $x(t)$ of model (4.2) is a Markov process.

For an LTI system, model (4.2) can be written as

$$\begin{cases} x(t+1) = Ax(t) + Bu(t) + Gw(t) \\ y(t) = Cx(t) + Du(t) + v(t) \end{cases} \tag{4.3}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $G \in \mathbb{R}^{n\times d}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$, and

$$\begin{pmatrix} w(t) \\ v(t) \end{pmatrix} \sim WP\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}\right) \tag{4.4}$$

with $Q \in \mathbb{R}^{d\times d}$, $R \in \mathbb{R}^{p\times p}$. Hereafter, it is assumed for simplicity $D = 0$, as it occurs in many real systems[1]. We will refer to the 6-tuple $\mathcal{M} = \{A, B, G, C, Q, R\}$ as the *model* of the stochastic system (4.3)-(4.4).

Let us now state a result concerning the mean and covariance functions of the processes $x(t)$ and $y(t)$, defined by system (4.3)-(4.4). Recall the notations $m_{\boldsymbol{x}}(t) = \mathbf{E}[x(t)]$ and $R_{\boldsymbol{x}}(t,s) = \mathbf{E}\left[(x(t) - m_{\boldsymbol{x}}(t))(x(s) - m_{\boldsymbol{x}}(s))^T\right]$ for the mean and covariance function of $x(t)$.

[1]The extension to the case $D \neq 0$ is trivial, as one can replace $y(t)$ with $y(t) - Du(t)$, being $u(t)$ known.

Theorem 4.1. *Consider system* (4.3) *and assume for simplicity* $D = 0$. *Then, under Assumption 4.1,one has*

$$\begin{aligned} m_{\boldsymbol{x}}(t) &= A^t m_0 + \sum_{i=0}^{t-1} A^{t-1-i} B u(i) && (4.5)\\ m_{\boldsymbol{y}}(t) &= C m_{\boldsymbol{x}}(t) && (4.6)\\ R_{\boldsymbol{x}}(t+\tau,t) &= A^\tau R_x(t,t) \triangleq A^\tau P(t) && (4.7)\\ R_{\boldsymbol{y}}(t+\tau,t) &= \begin{cases} CA^\tau P(t)C^T & \text{if } \tau > 0 \\ CP(t)C^T + R & \text{if } \tau = 0 \end{cases} && (4.8) \end{aligned}$$

where

$$P(t+1) = AP(t)A^T + GQG^T. \tag{4.9}$$

Proof

By taking the expected value of the first equation in (4.3), one gets

$$\mathbf{E}\left[x(t+1)\right] = A\mathbf{E}\left[x(t)\right] + Bu(t) + G\mathbf{E}\left[w(t)\right]$$

which results in

$$m_{\boldsymbol{x}}(t+1) = Am_{\boldsymbol{x}}(t) + Bu(t). \tag{4.10}$$

Then, (4.5) follows from the total response of the LTI deterministic system (4.10). Similarly, one obtains (4.6).

Let us define $\tilde{x}(t) = x(t) - m_{\boldsymbol{x}}(t)$. By using (4.10), one has

$$\begin{aligned} R_{\boldsymbol{x}}(t+\tau,t) =& \mathbf{E}\left[(x(t+\tau) - m_{\boldsymbol{x}}(t+\tau))(x(t) - m_{\boldsymbol{x}}(t))^T\right] = \\ =& \mathbf{E}\left[(Ax(t+\tau-1) + Gw(t+\tau-1) - Am_{\boldsymbol{x}}(t+\tau-1))\tilde{x}(t)^T\right] = \\ =& \mathbf{E}\left[(A\tilde{x}(t+\tau-1) + Gw(t+\tau-1))\tilde{x}(t)^T\right] = \\ =& A\mathbf{E}\left[\tilde{x}(t+\tau-1)\tilde{x}(t)^T\right] + G\underbrace{\mathbf{E}\left[w(t+\tau-1)\tilde{x}(t)^T\right]}_{0 \text{ for } \tau\geq 1} = \\ =& AR_x(t+\tau-1,t) \end{aligned}$$

where we have exploited the fact that $w(t)$ is a white process and hence $\tilde{x}(t)$, which depends on samples of w up to time $t-1$, is uncorrelated with

$w(t+\tau-1)$, $\forall\tau>0$. By iterating backwards, one gets

$$\begin{aligned} R_x(t+\tau,t) =& AR_x(t+\tau-1,t) = \\ =& A^2R_x(t+\tau-2,t) = \\ & \vdots \\ =& A^\tau R_x(t,t), \quad \text{for } \tau \geq 1. \end{aligned}$$

For $\tau = 0$ one has

$$\begin{aligned} P(t+1) = R_x(t+1,t+1) &= \mathbf{E}\left[\tilde{x}(t+1)\tilde{x}(t+1)^T\right] = \\ &= \mathbf{E}\left[(A\tilde{x}(t)+Gw(t))(A\tilde{x}(t)+Gw(t))^T\right] = \\ &= AP(t)A^T + GQG^T \end{aligned}$$

where once again we exploited $\mathbf{E}\left[\tilde{x}(t)w(t)^T\right] = 0$.

Finally, (4.8) can be proven in the same way, by exploiting (4.7) and the fact that $v(t)$ is uncorrelated with $w(t)$ and $x(0)$, and hence also with $\tilde{x}(t)$. □

Remark 4.1. It is worth stressing that all the results in Theorem 4.1 can be easily extended to the case of a linear time-varying system, in which the matrices of model $\mathcal{M}$ change with time. For example, if the first equation in (4.3) is $x(t+1) = A(t)x(t) + G(t)w(t)$ (we set $B = 0$ for simplicity), (4.5) becomes

$$m_{\boldsymbol{x}}(t) = A(t-1)A(t-2)\cdots A(0)m_0 = \prod_{i=0}^{t-1} A(i)\, m_0$$

The other equations can be amended in a similar way.

The matrix $P(t) = R_{\boldsymbol{x}}(t,t)$ is the autocovariance matrix of process $x(t)$ and equation (4.9), which describes its evolution in time, is called *recursive Lyapunov equation*. The fact that both $m_{\boldsymbol{x}}(t)$ and $P(t)$ are not constant in time implies that in general both $x(t)$ and $y(t)$ are not stationary process. The following result provides conditions under which they are asymptotically stationary.

Theorem 4.2. *Let λ_i $i = 1 \cdots n$, be the eigenvalues of A and assume that $|\lambda_i| < 1$, $\forall i$, (i.e., the system is asymptotically stable). Then:*

$$\lim_{t\to\infty} P(t) = \bar{P} \qquad \forall\, P(0) = P_0 > 0$$

where $\bar{P} = \bar{P}^T \geq 0$ is the unique solution of the Lyapunov equation

$$\bar{P} = A\bar{P}A^T + GQG^T.$$

Moreover, $x(t)$ and $y(t)$ are asymptotically stationary stochastic processes with zero mean and covariance matrices

$$\begin{aligned} R_{\boldsymbol{x}}(\tau) &= A^\tau \bar{P}\ , \qquad \tau \geq 0, \\ R_{\boldsymbol{y}}(\tau) &= \begin{cases} CA^\tau \bar{P}^T & \text{if } \tau > 0 \\ C\bar{P}C^T + R & \text{if } \tau = 0 \end{cases} . \end{aligned}$$

4.2 The state estimation problem

Consider the linear stochastic system

$$\begin{cases} x(t+1) = Ax(t) + Bu(t) + Gw(t) \\ y(t) = Cx(t) + v(t) \end{cases} \tag{4.11}$$

and let Assumption 4.1 hold. We are now ready to formulate the state estimation problem for system (4.11).

Problem 4.1. *(State estimation problem). At each time t, find an estimate of $x(t)$, based on the knowledge of the input sequence $\{u(0), u(1), \ldots, u(t-1)\}$ and the output measurements*

$$\mathcal{Y}^t = \{y(0), y(1), \cdots, y(t)\}.$$

Being both the data set $\mathcal{Y}^t$ and the quantity to be estimated $x(t)$ random variables, this is clearly a Bayesian estimation problem. Therefore, the MSE estimate $\hat{x}(t)$, minimizing the mean square error $\mathbf{E}\left[||x(t) - \hat{x}(t)||^2\right]$, is the conditional mean with respect to the data, i.e., $\hat{x}_{MSE}(t) = \mathbf{E}\left[x(t)|\mathcal{Y}^t\right]$.

In order to compute the MSE solution, it is necessary to know the joint pdf of the state $x(t)$ and the data $\mathcal{Y}^t$. If we restrict our attention to linear estimator, we can compute the linear MSE solution, which requires only knowledge of the mean and covariance functions of the involved processes. In fact, one has

$$\hat{x}_{LMSE}(t) = m_{\boldsymbol{x}}(t) + P_{x(t),Y^t}[P_{Y^t}]^{-1}(Y^t - m_{\boldsymbol{Y^t}}), \tag{4.12}$$

where P_{Y^t} is the covariance matrix of the data vector

$$Y^t = \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(t) \end{bmatrix},$$

while $P_{x(t),Y^t}$ is the cross-covariance matrix between $x(t)$ and Y^t. Although such quantities can be computed from model (4.11), by using Assumption 4.1, equation (4.12) cannot be employed in practice because the dimension of the involved covariance matrices grows as time passes. In fact, $P_{Y^t} \in \mathbb{R}^{(t+1)p\times(t+1)p}$ and for large values of t the computation of its inverse is practically infeasible.

In order to provide a computationally efficient approach to solve Problem 4.1, we aim at finding a *recursive solution* of the form

$$\hat{x}(t+1) = \Phi_t\, \hat{x}(t) + \Psi_t\, y(t+1)$$

where Φ_t and Ψ_t are suitable time-varying matrices which are used to compute a linear combination of the current estimate $\hat{x}(t)$ and the next measurement $y(t+1)$, providing the new estimate $\hat{x}(t+1)$, based on the data set $\mathcal{Y}^{t+1}$. The key idea is that the current estimate $\hat{x}(t)$ embeds all the information provided by the data up to time t, $\mathcal{Y}^t$. Clearly, the gains Φ_t and Ψ_t must be computed in order to minimize the mean square estimation error. The solution is based on a 2-step procedure and is known as the *Kalman Filter* (KF).

4.3 The Kalman Filter

Let us first introduce the notation that will be used in the construction of the LMSE state estimator. We denote by:

- $\hat{x}(t|t)$ the LMSE estimate of $x(t)$ based on Y^t;
- $\hat{x}(t+1|t)$ the LMSE estimate of $x(t+1)$ based on Y^t (LMSE 1-step ahead prediction);
- $P(t|t) = \mathbf{E}\left[(x(t) - \hat{x}(t|t))(x(t) - \hat{x}(t|t))^T\right] \in \mathbb{R}^{n\times n}$ the covariance matrix of the estimation error at time t;
- $P(t+1|t) = \mathbf{E}\left[(x(t+1) - \hat{x}(t+1|t))(x(t+1) - \hat{x}(t+1|t))^T\right] \in \mathbb{R}^{n\times n}$ the covariance matrix of the 1-step ahead prediction error at time t.

We aim at constructing the LMSE estimate of $x(t)$ through a two-step recursive procedure, also known as *prediction-correction algorithm*:

1. *Prediction*: Given $\hat{x}(t|t), P(t|t)$ and the model $\mathcal{M}$, compute $\hat{x}(t+1|t), P(t+1|t)$
2. *Correction*: Given $\hat{x}(t+1|t), P(t+1|t)$, the new measurement $y(t+1)$ and the model $\mathcal{M}$, compute $\hat{x}(t+1|t+1), P(t+1|t+1)$
3. Set $t \leftarrow t+1$ and go to step 1.

4.3.1 The prediction step

Given $\hat{x}(t|t), P(t|t)$ and the model

$$x(t+1) = Ax(t) + Bu(t) + Gw(t)$$

we want to compute the LMSE estimate $\hat{x}(t+1|t)$ of $x(t+1)$, based on data up to time t. The objective is to minimze $\mathbf{E}\left[||x(t+1) - \hat{x}(t+1|t)||^2\right]$ which is equivalent to minimize

$$\mathbf{E}\left[(x(t+1) - \hat{x}(t+1|t))(x(t+1) - \hat{x}(t+1|t))^T\right]$$

in the matricial sense. Let us introduce the notation for the estimation error $\tilde{x}(t|t) = x(t) - \hat{x}(t|t)$ and the prediction error $\tilde{x}(t+1|t) = x(t+1) - \hat{x}(t+1|t)$. Then one has

$$\begin{aligned} & \min_{\hat{x}} \mathbf{E}\left[(x(t+1) - \hat{x}(t+1|t))(x(t+1) - \hat{x}(t+1|t))^T\right] = \\ = & \min_{\hat{x}} \mathbf{E}\left[(Ax(t) + Bu(t) + Gw(t) - \hat{x}(t+1|t))(\cdots)^T\right] = \\ = & \min_{\hat{x}} \mathbf{E}\left[(A\hat{x}(t|t) + A\tilde{x}(t|t) + Bu(t) + Gw(t) - \hat{x}(t+1|t))(\cdots)^T\right] \end{aligned} \quad (4.13)$$

where the notation $(\cdots)^T$ is used to denote the transpose of the same term that appears on the left. Let

$$\begin{aligned} ① &= A\hat{x}(t|t) + Bu(t) - \hat{x}(t+1|t), \\ ② &= A\tilde{x}(t|t) + Gw(t). \end{aligned}$$

It is easy to see that $\mathbf{E}\left[①②^T\right] = 0$. In fact, a property of the LMSE estimate is that its estimation error is uncorrelated from the data, i.e. $\mathbf{E}\left[\tilde{x}(t|t)\, Y^t\right] = 0$. The term ① is a linear combination of the data up to time t, namely Y^t and $u(t)$. On the other hand, in ② both $\tilde{x}(t|t)$ and $w(t)$ are uncorrelated from Y^t. Hence, (4.13) becomes

$$\begin{aligned} \min_{\hat{x}(t+1|t)} \quad & \mathbf{E}\left[(A\hat{x}(t|t) + Bu(t) - \hat{x}(t+1|t))(\cdots)^T\right] \\ & + \mathbf{E}\left[(A\tilde{x}(t|t) + Gw(t))(\cdots)^T\right]. \end{aligned} \quad (4.14)$$

While the second term in (4.14) does not depend on the prediction $\hat{x}(t+1|t)$, the first term can be minimized by setting

$$\hat{x}(t+1|t) = A\hat{x}(t|t) + Bu(t) \quad (4.15)$$

which turns out to be the sought LMSE 1-step ahead state prediction. The corresponding error covariance is given by the second term in (4.14)

$$\begin{aligned} P(t+1|t) &= \mathbf{E}\left[(A\tilde{x}(t|t) + Gw(t))(\cdots)^T\right] = \\ &= \mathbf{E}\left[A\tilde{x}(t|t))\tilde{x}^T(t|t)A^T\right] + \mathbf{E}\left[Gw(t)w^T(t)G^T\right] = \\ &= A\mathbf{E}\left[\tilde{x}(t|t))\tilde{x}^T(t|t)\right]A^T + G\mathbf{E}\left[w(t)w^T(t)\right]G^T = \\ &= AP(t|t)A^T + GQG^T \end{aligned} \quad (4.16)$$

where we have exploited the fact that $\mathbf{E}\left[\tilde{x}(t|t)w(t)\right] = 0$, being $w(t)$ uncorrelated with both $x(t)$ and Y^t.

4.3.2 The correction step

Now assume that $\hat{x}(t+1|t)$, $P(t+1|t)$ and $y(t+1)$ are available. In order to derive the corrected estimate $\hat{x}(t+1|t+1)$, which incorporates also the new information provided by $y(t+1)$, we need to introduce a decoupling property of the LMSE estimate. Consider the expression of the LMSE estimate of a random variable $\boldsymbol{x}$ based on data $\boldsymbol{y}$

$$\hat{\boldsymbol{x}}_{LMSE} = m_{\boldsymbol{x}} + P_{\boldsymbol{xy}} P_{\boldsymbol{y}}^{-1}(\boldsymbol{y} - m_{\boldsymbol{y}}). \tag{4.17}$$

and assume that $\boldsymbol{y}$ is partitioned in two subvectors, i.e.

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}.$$

Then, (4.17) becomes

$$\hat{\boldsymbol{x}}_{\text{LMSE}} = m_{\boldsymbol{x}} + \begin{bmatrix} P_{\boldsymbol{xy_1}} & P_{\boldsymbol{xy_2}} \end{bmatrix} \begin{bmatrix} P_{\boldsymbol{y_1}} & P_{\boldsymbol{y_1 y_2}} \\ P_{\boldsymbol{y_2 y_1}} & P_{\boldsymbol{y_2}} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{y_1} - m_{\boldsymbol{y_1}} \\ \boldsymbol{y_2} - m_{\boldsymbol{y_2}} \end{bmatrix}$$

The computation of the matrix inverse simplifies if $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$ are uncorrelated. Indeed, if $P_{\boldsymbol{y_1 y_2}} = 0$ one has

$$\begin{aligned}
\hat{\boldsymbol{x}}_{\text{LMSE}} =& m_{\boldsymbol{x}} + \begin{bmatrix} P_{\boldsymbol{xy_1}} & P_{\boldsymbol{xy_2}} \end{bmatrix} \begin{bmatrix} P_{\boldsymbol{y_1}} & 0 \\ 0 & P_{\boldsymbol{y_2}} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{y}_1 - m_{\boldsymbol{y_1}} \\ \boldsymbol{y}_2 - m_{\boldsymbol{y_2}} \end{bmatrix} = \\
=& m_{\boldsymbol{x}} + \begin{bmatrix} P_{\boldsymbol{xy_1}} & P_{\boldsymbol{xy_2}} \end{bmatrix} \begin{bmatrix} P_{\boldsymbol{y_1}}^{-1} & 0 \\ 0 & P_{\boldsymbol{y_2}}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_1 - m_{\boldsymbol{y_1}} \\ \boldsymbol{y}_2 - m_{\boldsymbol{y_2}} \end{bmatrix} = \\
=& m_{\boldsymbol{x}} + P_{\boldsymbol{xy_1}} P_{\boldsymbol{y_1}}^{-1}(y_1 - m_{\boldsymbol{y_1}}) + P_{\boldsymbol{xy_2}} P_{\boldsymbol{y_2}}^{-1}(y_2 - m_{\boldsymbol{y_2}}) \\
=& \hat{\boldsymbol{x}}_{\text{LMSE}}|\boldsymbol{y_1} + P_{\boldsymbol{xy_2}} P_{\boldsymbol{y_2}}^{-1}(y_2 - m_{\boldsymbol{y_2}})
\end{aligned} \tag{4.18}$$

where $\hat{\boldsymbol{x}}_{\text{LMSE}}|\boldsymbol{y_1}$ is the LMSE estimate of $\boldsymbol{x}$ based on the observation of $\boldsymbol{y_1}$. This suggests that if one is able to decompose the available data in two

subsets that are uncorrelated with each other, it is possible to first compute the LMSE estimate based on the first data subset and then update it when the second data subset becomes available, by just adding the correction term $P_{\boldsymbol{xy_2}} P_{\boldsymbol{y_2}}^{-1}(y_2 - m_{\boldsymbol{y_2}})$.

Since in general it is not true that $y(t+1)$ is uncorrelated with Y^t (notice that if this occurs for every time t, y is an uncorrelated process), we want to find a new process that contains only the information provided by $y(t+1)$ which is not correlated with Y^t. This turns out to be the so called *innovation process* $e(t)$, which is defined as

$$\begin{aligned} e(t+1) &\triangleq y(t+1) - C\hat{x}(t+1|t) = \\ &= Cx(t+1) + v(t+1) - C\hat{x}(t+1|t) = \\ &= C\tilde{x}(t+1|t) + v(t+1). \end{aligned} \tag{4.19}$$

Proposition 4.1. *The innovation process $e(t)$ defined in (4.19) has the following properties.*

1. *$e(t+1)$ is a linear combination of the data $Y^{t+1} = \begin{bmatrix} Y^t \\ y(t+1) \end{bmatrix}$;*
2. *The process $e(t)$ is a sequence of uncorrelated random variables.*

Proof

1. Being $e(t+1) = y(t+1) - C\hat{x}(t+1|t)$, where $\hat{x}(t+1|t)$ is a linear combination of data Y^t, one has that $e(t+1)$ is a linear combination of Y^{t+1}.
2. Being $e(t+1) = C\tilde{x}(t+1|t) + v(t+1)$, we have that $\tilde{x}(t+1|t)$ is uncorrelated with Y^t because the LMSE estimation error is uncorrelated with the data used to compute the estimate, while $v(t+1)$ is uncorrelated with Y^t because it is white and independent from $w(t)$ and $x(0)$ (and hence also from $x(t)$). Hence, $e(t+1)$ is uncorrelated with Y^t and therefore also with $e(i)$, $i = 0, 1, \ldots, t$, which are linear combinations of Y^i. □

By exploiting Proposition 4.1, we compute the LMSE estimate of $x(t+1)$ based on

$$\begin{bmatrix} Y^t \\ e(t+1) \end{bmatrix} = \begin{bmatrix} I & 0 \\ * & 1 \end{bmatrix} \begin{bmatrix} Y^t \\ y(t+1) \end{bmatrix}$$

i.e., on a nonsingular linear transformation of the original dataset Y^{t+1}. By assuming $\boldsymbol{y_1} = Y^t$ and $\boldsymbol{y_2} = e(t+1)$, from (4.18) one gets

$$\hat{x}(t+1|t+1) = \hat{x}(t+1|t) + P_{x(t+1)e(t+1)} P^{-1}_{e(t+1)}(e(t+1) - m_{e(t+1)}) \quad (4.20)$$

where we need to compute the mean and covariance function of $e(t+1)$ and its cross covariance with $x(t+1)$. The mean is given by

$$\begin{aligned} m_{e(t+1)} &= \mathbf{E}\left[C\tilde{x}(t+1|t) + v(t+1)\right] = \\ &= C\mathbf{E}\left[\tilde{x}(t+1|t)\right] + \mathbf{E}\left[v(t+1)\right] = 0 \end{aligned} \quad (4.21)$$

in which the first term is zero because the LMSE estimate is unbiased. The cross-covariance between $x(t+1)$ and $e(t+1)$ is given by

$$\begin{aligned} &P_{x(t+1)e(t+1)} = \mathbf{E}\left[(x(t+1) - m_{x(t+1)})e^T(t+1)\right] = \\ &= \mathbf{E}\left[(x(t+1) - m_{x(t+1)})(C\tilde{x}(t+1|t) + v(t+1))^T\right] = \\ &= \mathbf{E}[\{\underbrace{\tilde{x}(t+1|t)}_{①} + \underbrace{\hat{x}(t+1|t) - m_{x(t+1)}}_{②}\}\{\underbrace{C\tilde{x}(t+1|t)}_{③} + \underbrace{v(t+1)}_{④}\}^T]. \end{aligned}$$

We claim that the only product which is not zero is the one between the terms ① and ③. In fact, $\tilde{x}(t+1|t)$ is uncorrelated with $\hat{x}(t+1|t)$, which is a linear combination of Y^t, while $v(t+1)$ is uncorrelated with both $\hat{x}(t+1|t)$ and $\tilde{x}(t+1|t)$, which depend on y and w up to time t. Therefore,

$$P_{x(t+1)e(t+1)} = \mathbf{E}\left[\tilde{x}(t+1|t)\tilde{x}^T(t+1|t)\right] C^T = P(t+1|t)C^T \quad (4.22)$$

Then, let us compute the covariance matrix of $e(t+1)$

$$\begin{aligned} P_{e(t+1)} &= \mathbf{E}\left[(C\tilde{x}(t+1|t) + v(t+1))(C\tilde{x}(t+1|t) + v(t+1))^T\right] = \\ &= C\mathbf{E}\left[(\tilde{x}(t+1|t)(\tilde{x}(t+1|t))^T\right] C^T + \mathbf{E}\left[v(t+1)v^T(t+1)\right] = \\ &= CP(t+1|t)C^T + R \end{aligned} \quad (4.23)$$

By substituting (4.19), (4.21), (4.22) and (4.23) into (4.20), one gets

$$\begin{aligned} \hat{x}(t+1|t+1) = \hat{x}(t+1|t) + P(t+1|t)C^T[CP(t+1|t)C^T + R]^{-1} \cdot \\ (y(t+1) - C\hat{x}(t+1|t)) \end{aligned}$$

which can be written in a shorter form as

$$\hat{x}(t+1|t+1) = \hat{x}(t+1|t) + K(t+1)(y(t+1) - C\hat{x}(t+1|t)) \tag{4.24}$$

where we have introduced the *Kalman gain*

$$K(t+1) = P(t+1|t)C^T[CP(t+1|t)C^T + R]^{-1}.$$

Finally, we need to compute the covariance of the estimation error at time $t+1$, which is given by

$$\begin{aligned}
&P(t+1|t+1) = \mathbf{E}\left[\tilde{x}(t+1|t)\tilde{x}^T(t+1|t)\right] = \\
&= \mathbf{E}\left[(x(t+1) - \hat{x}(t+1|t) - K(t+1)[C\tilde{x}(t+1|t) + v(t+1)])(\cdots)^T\right] = \\
&= \mathbf{E}\left[((I - K(t+1)C)\tilde{x}(t+1|t) - K(t+1)v(t+1))(\cdots)^T\right] \\
&= (I - K(t+1)C)P(t+1|t)(I - K(t+1)C)^T + K(t+1)RK(t+1)^T \\
&= P(t+1|t) - P(t+1|t)C^T[CP(t+1|t)C^T + R]^{-1}CP(t+1|t) \\
&= P(t+1|t)[I - C^T K(t+1)^T].
\end{aligned} \tag{4.25}$$

4.3.3 Initialization

In order to start the iteration of the Kalman filter, one needs to choose $\hat{x}(0|-1)$ and $P(0|-1)$, i.e., the state estimate and the corresponding covariance error *before* the first measurement $y(0)$ is processed. The natural choice are clearly the mean and covariance matrix of $x(0)$, if these quantities are known, i.e.

$$\begin{aligned}
\hat{x}(0|-1) &= m_0 \\
P(0|-1) &= P_0
\end{aligned}$$

Otherwise, if such quantities are not available, one can choose $\hat{x}(0|-1)$ as any vector which is compatible with the a priori information on $x(0)$, and set

$$P(0|-1) = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

with $\lambda_i > 0$, "big enough" so that the resulting confidence interval of $x_i(0)$ covers the initial uncertainty associated to the i-th element of the state vector.

4.4 Properties of the Kalman Filter

Summing up the derivation carried out in the previous section, the Kalman Filter algorithm is defined as follows.

Initialization: $\hat{x}(0|-1) = m_0,\ P(0|-1) = P_0$

For $t = 0, 1, 2, \cdots$

$$K(t) = P(t|t-1)C^T[CP(t|t-1)C^T + R]^{-1} \tag{4.26}$$

$$\hat{x}(t|t) = \hat{x}(t|t-1) + K(t)(y(t) - C\hat{x}(t|t-1)) \tag{4.27}$$

$$P(t|t) = P(t|t-1)[I - C^T K(t)^T] \tag{4.28}$$

$$\hat{x}(t+1|t) = A\hat{x}(t|t) + Bu(t) \tag{4.29}$$

$$P(t+1|t) = AP(t|t)A^T + GQG^T \tag{4.30}$$

end

It is worth stressing that the derivation of the Kalman filter does not change if the model $\mathcal{M}$ is time-varying. Hence, the above equation can be easily extended to the time-varying case by just setting $A = A(t)$, $B = B(t)$, $C = C(t)$, $G = G(t)$, $Q = Q(t)$, and $R = R(t)$.

Another interesting observation concerns the fact that the evolution of the matrices $P(t|t)$, $P(t+1|t)$ and $K(t)$ does not depend on the data set $\{u(t), y(t)\}$ which is processed. In fact, the input-output data affect only the estimates $\hat{x}(t|t)$, $\hat{x}(t+1|t)$. This means that the *quality* of the estimates (which is determined by the covariances of the estimation errors) depends only on the model $\mathcal{M}$ and not on the actual data realization. Moreover, the sequences $P(t|t)$, $P(t+1|t)$, $K(t)$, for $t = 0, 1, \ldots$, can be computed offline, *before* the filter is applied to a data stream, and they remain the same for all data sets. This is particularly useful in those applications in which the computational burden at each iteration is critical and must be kept as low as possible.

In the following, we analyze other useful properties of the KF algorithm.

4.4.1 The information matrix

Let us rewrite the correction equation for the covariance of the estimation error as

$$P(t|t) = P(t|t-1) - P(t|t-1)C^T[CP(t|t-1)C^T + R]^{-1}CP(t|t-1) \quad (4.31)$$

and define the *Information matrix*

$$I(t|t) \triangleq P(t|t)^{-1}.$$

The information matrix is the inverse of the covariance matrix, so the larger is $I(t|t)$, the smaller is the uncertainty associated to the estimate $\hat{x}(t|t)$ of $x(t)$ (i.e., the more accurate is the estimate). By recalling the Matrix Inversion Lemma

$$(\bar{A} - \bar{B}\bar{C}\bar{D})^{-1} = \bar{A}^{-1} + \bar{A}^{-1}\bar{B}(\bar{C}^{-1} - \bar{D}\bar{A}^{-1}\bar{B})^{-1}\bar{D}\bar{A}^{-1} \quad (4.32)$$

and applying it to (4.31) with $\bar{A} = P(t|t)$, $\bar{B} = P(t|t-1)C^T$, $\bar{C} = [CP(t|t-1)C^T + R]^{-1}$ and $\bar{D} = CP(t|t-1)$, one gets

$$I(t|t) = I(t|t-1) + C^T R^{-1} C$$

where $I(t|t-1) = P(t|t-1)^{-1}$ and $C^T R^{-1} C$ can be interpreted as the quantity of information provided by the new measurement $y(t)$. In the scalar case ($n = p = 1$), one has $y(t) = c\,x(t) + v(t)$, with $c \in \mathbb{R}$ and

$$C^T R^{-1} C = \frac{c^2}{\sigma_v^2}$$

which can be seen as a sort of signal-to-noise ratio, between the *signal* $c\,x(t)$ measured by the output sensor and the measurement noise $v(t)$.

4.4.2 The Kalman Filter as a dynamic system

It is easy to see that the KF is a dynamic system processing the input and output data to produce state estimates. By substituting (4.27) into (4.29), one gets

$$\begin{aligned}\hat{x}(t+1|t) &= A\{\hat{x}(t|t-1) + K(t)(y(t) - C\hat{x}(t|t-1))\} + Bu(t) = \\ &= (A - AK(t)C)\hat{x}(t|t-1) + AK(t)y(t) + Bu(t) \\ &= (A - AK(t)C)\,\hat{x}(t|t-1) + \begin{bmatrix} AK(t) & B \end{bmatrix} \begin{bmatrix} y(t) \\ u(t) \end{bmatrix} \end{aligned} \tag{4.33}$$

which is the equation of a linear time-varying system, whose state vector is $\hat{x}(t+1|t)$ and the input vector is $\begin{bmatrix} y(t) \\ u(t) \end{bmatrix}$. Notice that the KF is an inherently time-varying system, due to the fact that the Kalman gain $K(t)$ changes at every time instant. The corresponding equation of the prediction error turns out to be

$$\begin{aligned}\tilde{x}(t+1|t) &= x(t+1) - \hat{x}(t+1|t) \\ &= Ax(t) + Bu(t) + Gw(t) \\ &\quad -A\{\hat{x}(t|t-1) + K(t)(y(t) - C\hat{x}(t|t-1))\} - Bu(t) = \\ &= A\tilde{x}(t|t-1) + Gw(t) - AK(t)\{Cx(t) + v(t) - C\hat{x}(t|t-1)\} \\ &= (A - AK(t)C)\,\tilde{x}(t|t-1) + Gw(t) - AK(t)v(t)\end{aligned} \tag{4.34}$$

which leads to the evolution of the average prediction error

$$\mathbf{E}\left[\tilde{x}(t+1|t)\right] = (A - AK(t)C)\,\mathbf{E}\left[\tilde{x}(t|t-1)\right]. \tag{4.35}$$

For the covariance of the 1-step ahead prediction error, by substituing (4.28) into (4.30) one gets

$$\begin{aligned}P(t+1|t) &= AP(t|t-1)A^T + GQG^T \\ &\quad - AP(t|t-1)C^T[CP(t|t-1)C^T + R]^{-1}CP(t|t-1)A^T\end{aligned} \tag{4.36}$$

which can be seen as a dynamic system whose state is the matrix $P(t|t-1)$. Equation (4.36) is known as the *Discrete Riccati equation (DRE).*

4.4.3 The Kalman Filter as a state observer

Being the aim of the KF the computation of a state estimate, there is clearly a connection with the classical Luenberger state observer for deterministic systems. Consider the deterministic system

$$\begin{cases} x(t+1) = Ax(t) + Bu(t) \\ y(t) = Cx(t) \end{cases}$$

then, the Luenberger observer is given by

$$\hat{x}(t+1) = A\hat{x}(t) + Bu(t) + L(y(t) - C\hat{x}(t)). \tag{4.37}$$

If we introduce the state estimation error $\tilde{x}(t) = x(t) - \hat{x}(t)$, we obtain the error dynamics

$$\tilde{x}(t+1) = (A - LC)\tilde{x}(t) \tag{4.38}$$

If the pair (A, C) is detectable, it is always possible to find a matrix L such that $(A - LC)$ has all its eigenvalues inside the unit circle. This implies that system (4.38) is asymptotically stable and hence $\lim_{t\to\infty} \tilde{x}(t) = 0$ for every initial condition $\tilde{x}(0)$. Equation (4.37) can be rewritten as

$$\hat{x}(t+1) = (A - LC)\hat{x}(t) + \begin{bmatrix} L & B \end{bmatrix} \begin{bmatrix} y(t) \\ u(t) \end{bmatrix}$$

in which it is easy to recognize the same structure as in (4.33). Indeed, they turn out to be the same equation if we set $L = AK(t)$. A similar analogy can be observed between equations (4.38) and (4.35). This means that the KF can be seen as a time-varying state observer for the linear stochastic system (4.11).

4.4.4 The innovation process

Let us consider the innovation process $e(t) = y(t) - C\hat{x}(t|t-1)$. It is possible to reformulate the KF as a dynamic system that processes the signals $y(t)$

and $u(t)$, to generate $e(t)$ as the output. In fact, by recalling (4.33) one can write

$$\begin{cases} \hat{x}(t+1|t) = (A - AK(t)C)\hat{x}(t|t-1) + \begin{bmatrix} AK(t) & B \end{bmatrix} \begin{bmatrix} y(t) \\ u(t) \end{bmatrix} \\ e(t) = -C\hat{x}(t|t-1) + \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} y(t) \\ u(t) \end{bmatrix} \end{cases}$$

This can be seen as a *whitening filter*, i.e., a system which takes the (usually) correlated process $y(t)$ and transforms it into the sequence of uncorrelated random variables $e(t)$ (which is also a white process in the Gaussian case $e(t)$).

4.4.5 Extension to non white disturbance and noise processes

So far, we have supposed that the disturbance process $w(t)$ and the measurement noise $v(t)$ are white. This assumption can be relaxed, provided that a model for such processes is known. For example, assume that $w(t)$ is a stochastic process generated by the system

$$\begin{aligned} x_w(t+1) &= A_w x_w(t) + B_w \xi(t) \\ w(t) &= C_w x_w(t) + D_w \xi(t) \end{aligned}$$

where $\xi(t) \sim WP(0, Q_\xi)$ and it is assumed to be uncorrelated with the measurement noise $v(t) \sim WP(0, R)$. Notice that for an asymptotically stationary s.p., matrices A_w, B_w, C_w, D_w can be derived by computing a state space realization of the canonical spectral factor of process $w(t)$. Hence, we can write

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + Gw(t) \\ &= Ax(t) + Bu(t) + G(C_w x_w(t) + D_w \xi(t)). \end{aligned}$$

By defining the extended state vector

$$\bar{x}(t) = \begin{bmatrix} x(t) \\ x_w(t) \end{bmatrix}$$

one can write the system

$$\begin{bmatrix} x(t+1) \\ x_w(t+1) \end{bmatrix} = \begin{bmatrix} A & GC_w \\ 0 & A_w \end{bmatrix} \begin{bmatrix} x(t) \\ x_w(t) \end{bmatrix} + \begin{bmatrix} GD_w \\ B_w \end{bmatrix} \xi(t) + \begin{bmatrix} B \\ 0 \end{bmatrix} u(t) \tag{4.39}$$

$$y(t) = \begin{bmatrix} C & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ x_w(t) \end{bmatrix} + v(t) \tag{4.40}$$

Hence, one can apply the standard KF to system (4.39)-(4.40), thus obtaining an estimate $\hat{\bar{x}}(t|t) = \begin{bmatrix} \hat{x}(t|t) \\ \hat{x}_w(t|t) \end{bmatrix}$ of the extended state. Notice that, besides the desired estimate of the state of the original system, this contains also the estimate of the disturbance state $x_w(t)$ as a byproduct.

Similarly, assume that $v(t)$ is not white and it is generated by the system

$$x_v(t+1) = A_v x_v(t) + B_v \xi(t)$$
$$v(t) = C_v x_v(t) + D_v \xi(t)$$

where $\xi(t) \sim WP(0, Q_\xi)$ and is uncorrelated with the process disturbance $w(t) \sim WP(0, Q)$. By defining the extended state vector

$$\bar{x}(t) = \begin{bmatrix} x(t) \\ x_v(t) \end{bmatrix}$$

one obtains the extended system equations

$$\begin{bmatrix} x(t+1) \\ x_v(t+1) \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & A_v \end{bmatrix} \begin{bmatrix} x(t) \\ x_v(t) \end{bmatrix} + \begin{bmatrix} G & 0 \\ 0 & B_v \end{bmatrix} \begin{bmatrix} w(t) \\ \xi(t) \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} u(t) \tag{4.41}$$

$$y(t) = \begin{bmatrix} C & C_v \end{bmatrix} \begin{bmatrix} x(t) \\ x_v(t) \end{bmatrix} + D_v \, \xi(t) \tag{4.42}$$

where the extended process $\bar{w}(t) = \begin{bmatrix} w(t) \\ \xi(t) \end{bmatrix}$ is a white process with zero mean and covariance matrix $\begin{bmatrix} Q & 0 \\ 0 & Q_\xi \end{bmatrix}$.

Also in this case, one may think to apply the standard KF to system (4.41)-(4.42), thus obtaining an estimate $\hat{\bar{x}}(t|t) = \begin{bmatrix} \hat{x}(t|t) \\ \hat{x}_v(t|t) \end{bmatrix}$ of the extended state, which contains the desired state estimate, along with the estimate of the noise state $x_v(t)$. However, it should be noticed that it is not true that the extended process $\bar{w}(t)$ and the new output noise $D_v\xi(t)$ are uncorrelated. In fact, one has $\mathbf{E}\left[\bar{w}(t)(D_v\xi(t))^T\right] = [0 \quad Q_\xi D_v^T]$. Therefore, it is customary to use the version of the Kalman Filter in which item ii) in Assumption 4.1 is replaced by

$$\mathbf{E}\left[\begin{pmatrix} w(t) \\ v(t) \end{pmatrix}\begin{pmatrix} w(t) \\ v(t) \end{pmatrix}^T\right] = \begin{bmatrix} Q & N \\ N & R \end{bmatrix}$$

where $N \in \mathbb{R}^{d\times p}$ is the cross-covariance between $w(t)$ and $v(t)$. The equations of such a version of the Kalman Filter turn out to be as follows.

$$\hat{x}(t|t) = \hat{x}(t|t-1) + K(t)(y(t) - C\hat{x}(t|t-1)) \tag{4.43}$$

$$P(t|t) = P(t|t-1)[I - C^T K^T(t)] \tag{4.44}$$

$$K(t) = P(t|t-1)C^T[CP(t|t-1)C^T + R]^{-1} \tag{4.45}$$

$$\begin{aligned} \hat{x}(t+1|t) = {} & A\hat{x}(t|t) + Bu(t) \\ & + GN[CP(t|t-1)C^T + R]^{-1}(y(t) - C\hat{x}(t|t-1)) \end{aligned} \tag{4.46}$$

$$\begin{aligned} P(t+1|t) = {} & AP(t|t)A^T + GQG^T - GN[CP(t|t-1)C^T + R]^{-1}N^T G^T \\ & - AP(t|t-1)C^T[CP(t|t-1)C^T + R]^{-1}N^T G^T \\ & - GN[CP(t|t-1)C^T + R]^{-1}CP(t|t-1)A^T \end{aligned} \tag{4.47}$$

which in predictor form takes the more coincise form

$$\hat{x}(t+1|t) = A\hat{x}(t|t-1) + Bu(t) + K_c(t)(y(t) - C\hat{x}(t|t-1)) \tag{4.48}$$

$$P(t+1|t) = AP(t|t)A^T + GQG^T - K_c(t)[CP(t|t-1)C^T + R]K_c(t)^T \tag{4.49}$$

with

$$K_c(t) = (AP(t|t-1)C^T + GN)[CP(t|t-1)C^T + R]^{-1}. \tag{4.50}$$

Finally, the case in which both $w(t)$ and $v(t)$ are not white can be treated in the same way, by combining the two approaches outlined above.

4.5 Asymptotic behavior of the KF

A fundamental question on the performance of the Kalman Filter concerns its asymptotic behavior. In particular, we would like to answer the following questions

1. What is the asymptotic expected value of the estimation error, i,e.

$$\lim_{t\to+\infty} \mathbf{E}\left[\tilde{x}(t|t)\right] = \lim_{t\to+\infty} \mathbf{E}\left[x(t) - \hat{x}(t|t)\right] \tag{4.51}$$

   and how does it depend on the initialization of the filter? If the limit in (4.51) is equal to zero, this means that the KF is an asymptotically unbiased estimator.

2. What is the asymptotic behavior of the DRE (4.36)? Does the limit $\lim_{t\to+\infty} P(t+1|t)$ exist? How does it depend on the initial condition $P(0|-1)$?

3. Does the Kalman gain $K(t)$ converge to a constant matrix? And if this is the case, what are the properties of the asymptotic version of system (4.33)?

The next result provide an answer to the above questions.

Theorem 4.3. *Let A, B, C, G, Q, R be constant matrices. Define $H \in \mathbb{R}^{n\times d}$ such that $HH^T = GQG^T$ and assume that the pair (A, C) is detectable and the pair (A, H) is stabilizable. Then, the following results hold.*

1. $\lim_{t\to+\infty} \mathbf{E}\left[\tilde{x}(t|t)\right] = \lim_{t\to+\infty} \mathbf{E}\left[\tilde{x}(t+1|t)\right] = 0\ ,\ \forall \hat{x}(0|-1) \in \mathbb{R}^n$

2. $\lim_{t\to+\infty} P(t+1|t) = P_\infty < +\infty,\ \forall\, P(0|-1) > 0$ *where P_∞ is the unique positive semidefinite solution of the Algebraic Riccati Equation (ARE)*

$$P_\infty = AP_\infty A^T + GQG^T - AP_\infty C^T[CP_\infty C^T + R]^{-1}CP_\infty A^T \tag{4.52}$$

3. *Let*

$$K_\infty = \lim_{t\to\infty} K(t) = P_\infty C^T[CP_\infty C^T + R]^{-1}.$$

Then, the matrix $(A - AK_\infty C)$ *has all its eigenvalues inside the unit circle.*

The first item in Theorem 4.3 states that the state estimates provided by the KF are asymptotically unbiased. Even more important is the result in item 2: the covariance of the 1-step ahead prediction error always converges to the same constant matrix P_∞, whatever is the initial covariance matrix $P(0|-1)$ chosen to start the KF iterations. Notice that the same occurs for the covariance matrix of the state estimation error

$$\lim_{t\to+\infty} P(t|t) = P_\infty - P_\infty C^T[CP_\infty C^T + R]^{-1}CP_\infty = P_\infty(I - C^T K_\infty^T)$$

It is also worth pointing out that the above results hold under quite mild assumptions. The most important one concerns detectability of the pair (A, C): in fact, if the system is not detectable, it means that there is a subsystem which is not asymptotically stable and it is also unobservable. Hence, the covariance of the state estimation error for such subsystem will eventually grow to infinity. Recall that if a system is fully observable, it is also detectable.

The other assumption in Theorem 4.3, namely stabilizability of the pair (A, H), is essentially technical and it is necessary to guarantee that the ARE (4.52) has a unique positive semidefinite solution. Recall that if (A, H) is fully reachable, it is also stabilizable. This assumption can be further relaxed to exclude only unreachable eigenvalues of A with modulus exactly equal to 1: in such a case the ARE may have multiple positive semidefinite solutions, but they will be ordered (in matrical sense) and $P(t+1|t)$ will converge to the largest one.

Finally, the third item in Theorem 4.3 suggests that the time-varying evolution of the average prediction error in equation (4.35), converges to an asymptotically stable time-invariant system. This suggests that one may want to use the constant gain K_∞ right from the start, in place of the time-varying Kalman gain $K(t)$, as explained next.

4.5.1 The asymptotic Kalman Filter

Let us assume that instead of changing the Kalman gain $K(t)$ at every KF iteration, we want to use a constant gain. A natural choice is to use K_∞, because we know that it minimizes the asymptotic mean square estimation error. Then, the recursion (4.27)-(4.30) simplifies to

$$\hat{x}(t+1|t) = A\hat{x}(t|t) + Bu(t) \tag{4.53}$$

$$\hat{x}(t+1|t+1) = \hat{x}(t+1|t) + K_\infty(y(t+1) - C\hat{x}(t+1|t)) \tag{4.54}$$

which we will refer to as the *Asymptotic Kalman Filter*. Clearly, while (4.53)-(4.54) has the advantage of being a linear time-invariant system, not requiring the computation of the covariance matrices $P(t|t)$, $P(t+1|t)$ at every time step, it does not guarantee anymore to minimize the estimation MSE for every t, but only as time approaches infinity.

In order to understand which is the trade-off between using a time-invariant filter and the optimal KF, let us consider the following simple example involving a scalar state ($n = 1$)

$$\begin{cases} x(t+1) = a\,x(t) \\ y(t) = x(t) + v(t) \end{cases} \tag{4.55}$$

where $a \in \mathbb{R}$ and $v(t) \sim WP(0, r)$. The associated KF equations are

$$\begin{aligned} \hat{x}(t+1|t) &= a\,\hat{x}(t|t) \\ P(t+1|t) &= a^2 P(t|t) \\ \hat{x}(t+1|t+1) &= \hat{x}(t+1|t) + K(t+1)(y(t+1) - \hat{x}(t+1|t)) \\ P(t+1|t+1) &= P(t+1|t) - \frac{P^2(t+1|t)}{P(t+1|t)+r}. \end{aligned}$$

In prediction form, they become

$$\hat{x}(t+1|t) = a\hat{x}(t|t-1) + aK(t)(y(t) - \hat{x}(t|t-1)) \tag{4.56}$$

$$P(t+1|t) = a^2 P(t|t-1) - \frac{a^2 P^2(t|t-1)}{P(t|t-1)+r} = \frac{a^2 P(t|t-1) r}{P(t|t-1)+r} \tag{4.57}$$

The estimation error is given by

$$\begin{aligned}\tilde{x}(t+1|t) &= x(t+1) - \hat{x}(t+1|t) \\ &= ax(t) - a\hat{x}(t|t-1) - aK(t)(x(t) + v(t) - \hat{x}(t|t-1)) \\ &= a\tilde{x}(t|t-1) - aK(t)\tilde{x}(t|t-1) - aK(t)v(t) \\ &= a(1-K(t))\tilde{x}(t|t-1) - aK(t)v(t) \end{aligned} \tag{4.58}$$

where

$$K(t) = \frac{P(t|t-1)}{P(t|t-1)+r}.$$

Now, let us assume that we want to use an LTI filter with constant gain K. From (4.58), the corresponding estimation error equation will be

$$\tilde{x}(t+1) = a(1-K)\tilde{x}(t) - aKv(t).$$

By taking the expected value, being $\mathbf{E}[v(t)] = 0$, one gets

$$\mathbf{E}[\tilde{x}(t+1)] = a(1-K)\mathbf{E}[\tilde{x}(t)]$$

and hence the *bias error* will tend to zero whenever $|a(1-K)| < 1$. In particular, convergence will be faster as K approaches 1. On the other hand, if we consider the variance of the estimation error

$$P(t) = \mathbf{E}\left[(\tilde{x}(t) - m_{\tilde{\boldsymbol{x}}(\boldsymbol{t})})^2\right]$$

by exploiting the fact that $\tilde{x}(t|t)$ and $v(t)$ are uncorrelated, we get

$$\begin{aligned} P(t+1) &= \mathbf{E}\left[\left\{a(1-K)(\tilde{x}(t) - m_{\tilde{\boldsymbol{x}}(\boldsymbol{t})}) - aKv(t)\right\}^2\right] \\ &= a^2(1-K)^2\mathbf{E}\left[(\tilde{x}(t) - m_{\tilde{\boldsymbol{x}}(\boldsymbol{t})})^2\right] + a^2K^2\mathbf{E}\left[v^2(t)\right] \\ &= a^2(1-K)^2P(t) + a^2K^2r \end{aligned}$$

which is a first order dynamic system in the variable $P(t)$, forced by the constant input a^2K^2r. If $a^2(1-K)^2 < 1$ such a system is asymptotically stable and one has

$$\lim_{t\to\infty} P(t) = \frac{1}{1-a^2(1-K)^2}a^2K^2r.$$

Therefore, it is easy to see that the asymptotic value of the variance of the estimation error will tend to zero as K approaches zero. Summing up, we are faced to the typical bias-variance trade off: we need a "large" filter gain ($K \to 1$) to reduce the bias error as fast as possible, and a "small" gain ($K \to 0$) to reduce the asymptotic variance (uncertainty) associated to the estimate. This is the reason why the MSE filter is indeed a time-varying one: it employs a large gain during the transient, in order to rapidly reduce the bias error, and a small gain asymptotically, to reduce the variance of the estimation error.

Figures 4.1-4.3 report the results of a numerical test performed on system (4.55) with $a = 0.9$, $r = 0.04$, and $x(0) = 1$. Fig. 4.1 shows the evolution of the true state $x(t)$ and the noisy observation $y(t)$. Fig. 4.2 reports on top the estimates of two LTI filters with constant gain $K = 0.9$ and $K = 0.1$, respectively. It can be seen that the former quickly reduces the bias error but shows a remarkable uncertainty in the estimates; conversely, the latter has a small asymptotic variance of the estimate but it is quite slow in tracking the true state evolution. The bottom plot reports the estimate provided by the Kalman Filter, initialized with $\hat{x}(0|-1) = 0$ and $P(0|-1) = 1$. It can be observed that the KF succeeds in both reducing the initial bias error and keeping small the asymptotic error variance. This is clearly due to the time varying gain $K(t)$, shown in Fig. 4.3, along with the error variance $P(t|t)$.

The considered one-dimensional example is also helpful to provide an insight in the asymptotic results of Theorem 4.3. Since there is no process disturbance in system (4.55), one has $Q = 0$ and the ARE (4.52) reduces to

$$P_\infty = \frac{a^2 P_\infty r}{P_\infty + r}$$

which can be written as

$$P_\infty^2 + P_\infty r(1 - a^2) = 0$$

and therefore it has two solutions

$$\begin{cases} P_\infty = 0 \\ P_\infty = r(a^2 - 1) \end{cases} \tag{4.59}$$

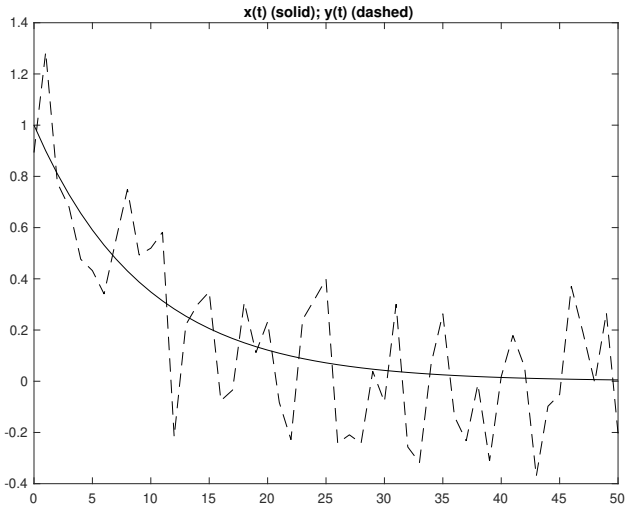

Figure 4.1: Evolution of system (4.55) with $a = 0.9$, $r = 0.04$, and $x(0) = 1$: $x(t)$ (solid); $y(t)$ (dashed).

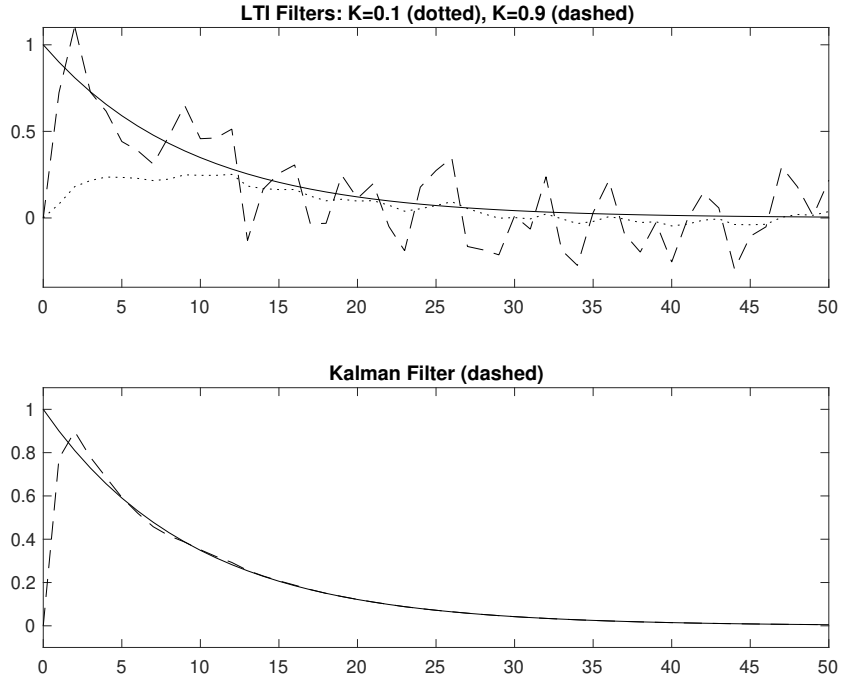

Figure 4.2: Top: performance of LTI filters with $K = 0.9$ (dashed) and $K = 0.1$ (dotted), with respect to the true state $x(t)$ (solid). Bottom: KF estimate $\hat{x}(t|t)$ (dashed) compared to true state $x(t)$ (solid).

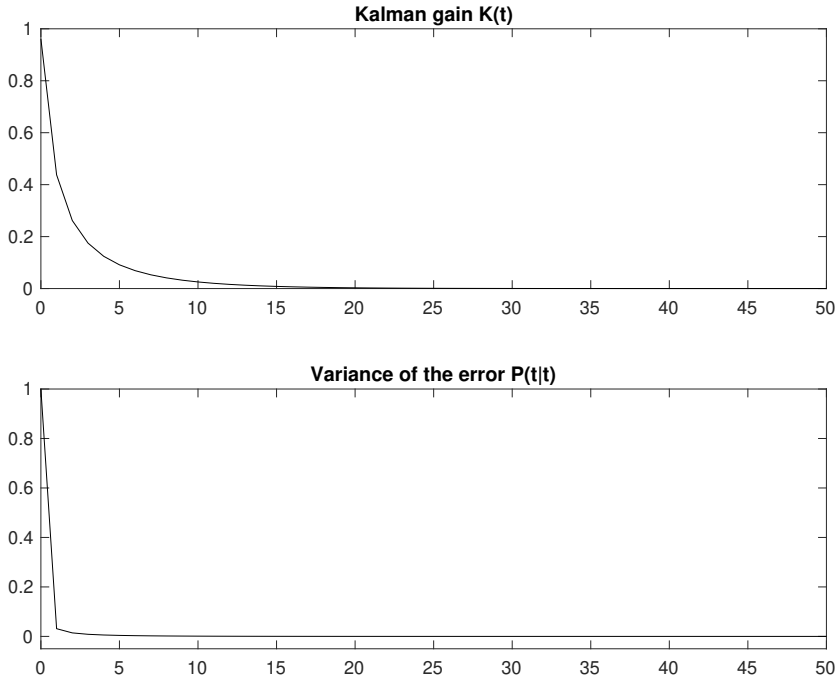

Figure 4.3: Top: Kalman gain $K(t)$. Bottom: variance of estimation error $P(t|t)$.

If $a^2 < 1$, then the only positive semidefinite solution is $P_\infty = 0$. This is consistent with the fact that, being $H = 0$, the pair $(A, H) = (a, 0)$ is not reachable but it is indeed stabilizable (the system is already asymptotically stable!). Hence, Theorem 4.3 guarantees that

$$\lim_{t\to\infty} P(t|t-1) = P_\infty = 0$$
$$\lim_{t\to\infty} K(t) = \frac{P_\infty}{P_\infty + r} = K_\infty = 0$$

This is precisely what can be obtained by studying the one-dimensional (nonlinear) system (4.57), which describes the behavior of the variance $P(t|t-1)$. Notice that in this example the asymptotic Kalman filter simply ignores the output data ($K_\infty = 0$) and it just waits that the state estimate converges to zero, like the true state does.

Conversely, if $a^2 > 1$, the pair $(a, 0)$ is not stabilizable and hence we cannot apply Theorem 4.3. Indeed, both solutions (4.59) are positive semidefi-

nite. Nevertheless, by analyzing again system (4.57), it can be shown that:

$$\lim_{t\to\infty} P(t|t-1) = r(a^2-1)$$
$$\lim_{t\to\infty} K(t) = \frac{P_\infty}{P_\infty + r} = \frac{r(a^2-1)}{r(a^2-1)+1} = \frac{a^2-1}{a^2}$$

i.e., the covariance converges to the largest solution of the ARE. We stress that the solution $K_\infty = 0$ this time is unacceptable because the system is unstable and ignoring the output data would lead to divergence of the error variance. Instead, $K_\infty = \frac{a^2-1}{a^2}$ leads to

$$\lim_{t\to\infty}(A - AK(t)C) = a - aK_\infty = a - a\frac{a^2-1}{a^2} = \frac{1}{a}$$

which is indeed inside the unit circle, as $a^2 > 1$ implies $\left|\frac{1}{a}\right| < 1$. This confirms that the asymptotic KF is an asymptotically stable LTI filter, guaranteeing that the expected value of the state estimation error converges to zero also for unstable systems.

It is also interesting to analyze the case in which the process disturbance is present. Let

$$x(t+1) = ax(t) + w(t)$$

with $w(t) \sim WP(0, q)$ and $h = \sqrt{q} \neq 0$, which implies that the pair (a, h) is always reachable and hence also stabilizable. The ARE becomes

$$P_\infty = a^2 P_\infty + q - \frac{a^2 P_\infty^2}{P_\infty + r} = \frac{a^2 P_\infty r}{P_\infty + r} + q$$

or equivalently

$$P_\infty^2 + P_\infty[r(1-a^2) - q] - rq = 0$$

which has always a unique positive solution $P_\infty > 0$, for every $a \in \mathbb{R}$, $q > 0$, $r > 0$. Hence, Theorem 4.3 guarantees that

$$\lim_{t\to\infty} P(t|t-1) = P_\infty.$$

Notice that in this case P_∞ cannot be equal to zero because $w(t)$ steadily injects uncertainty into the system.

4.6 Recursive system identification

Among the wide variety of applications of the KF, a popular one concerns the problem of online system identification. Consider the parametric system identification problem addressed in Chapter 3 and in particular the one concerning *linear regression models* (such as ARX, FIR, etc.). The problem aim is to find the parameter vector θ minimizing the 1-step ahead output prediction error, i.e.

$$\min_{\theta} \sum_{t=1}^{N} (y(t) - \hat{y}(t|t-1))^2 \tag{4.60}$$

where the 1-step ahead prediction is a linear function of the parameters, $\hat{y}(t|t-1) = \varphi^T(t)\theta$. As an example, for the ARX model

$$\begin{aligned} &y(t) + a_1 y(t-1) + a_2 y(t-2) + \cdots + a_{n_a} y(t-n_a) = \\ &= b_1 u(t-1) + b_2 u(t-2) + \cdots + b_{n_b} u(t-n_b) + e(t) \end{aligned}$$

with $e(t) \sim WP(0, \sigma_e^2)$, the regressor vector $\varphi(t)$ is a collection of observed past values of the input and output signals

$$\varphi(t) = \begin{bmatrix} -y(t-1) & \cdots & -y(t-n_a) & u(t-1) & \cdots & u(t-n_b) \end{bmatrix}^T$$

while the vector $\theta \in \mathbb{R}^d$, with $d = n_a + n_b$, contains the unknown parameters to be estimated

$$\theta = \begin{bmatrix} a_1 & a_2 & \cdots & a_{n_a} & b_1 & b_2 & \cdots & b_{n_b} \end{bmatrix}^T.$$

We already know that the unique solution of problem

$$\min_{\theta} \sum_{t=1}^{N} (y(t) - \varphi^T(t)\theta)^2$$

is given by the least squares estimate

$$\hat{\theta}_N^{LS} = \left(\sum_{t=1}^{N} \varphi(t)\varphi^T(t) \right)^{-1} \sum_{t=1}^{N} \varphi(t) y(t) \tag{4.61}$$

provided that the matrix $\sum_{t=1}^{N} \varphi(t)\varphi^T(t) \in \mathbb{R}^{d\times d}$ is nonsingular. Being the sum of N matrices of rank 1, a necessary condition is that $N \geq d$.

The computation of (4.61) is straightforward once we have collected the entire input-output dataset $\{u(1), y(1), \cdots u(N), y(N)\}$. However, in many practical applications one would like to estimate the parameter vector θ *on-line*, i.e., while the input-output data are being collected. A typical example is that of *fault detection*, in which the estimate of the parameter θ is continuously updated and compared to an ideal reference θ_0, in order to check if the system is significantly deviating from its nominal behavior. Another application is *adaptive control*, in which the identified model is used to update in real time the parameters of a feedback controller, which is designed according to the estimated parameter vector $\hat{\theta}$. This means that at each time t we aim at computing

$$\hat{\theta}_t = \arg\min_{\theta} \sum_{k=1}^{t} (y(k) - \varphi^T(k)\theta)^2.$$

Obviously, the solution is given by the least squares estimate

$$\hat{\theta}_t^{LS} = \left(\sum_{k=1}^{t} \varphi(k)\varphi^T(k) \right)^{-1} \sum_{k=1}^{t} \varphi(k)y(k) \tag{4.62}$$

but it would be clearly impractical to use equation (4.62) to compute the parameter estimates at each time t, as one needs to process a steadily increasing number of data as t grows.

As an alternative, we look for a *recursive solution* that updates the estimate at time t by using only the previous estimate at time $t-1$ and the new data pair $y(t), \varphi(t)$. Let

$$R(t) = \sum_{k=1}^{t} \varphi(k)\varphi^T(k) \qquad f(t) = \sum_{k=1}^{t} \varphi(k)y(k)$$

And rewrite the equation of the least squares estimate (4.62) as

$$\hat{\theta}_t = R(t)^{-1} f(t).$$

Now, it is easy to see that $R(t)$ and $f(t)$ can be updated in a recursive manner, by using just the current data pair $y(t), \varphi(t)$ at time t, according to

$$\begin{aligned} R(t) &= R(t-1) + \varphi(t)\varphi^T(t) \ , \\ f(t) &= f(t-1) + \varphi(t)y(t) \ . \end{aligned}$$

By exploiting the fact that $\hat{\theta}_{t-1} = R(t-1)^{-1}f(t-1)$ implies $f(t-1) = R(t-1)\hat{\theta}_{t-1}$, one has

$$\begin{aligned} \hat{\theta}_t &= R(t)^{-1}f(t) = R(t)^{-1}(f(t-1) + \varphi(t)y(t)) \\ &= R(t)^{-1}\left(R(t-1)\hat{\theta}_{t-1} + \varphi(t)y(t)\right) = \\ &= R(t)^{-1}\left[\left(R(t) - \varphi(t)\varphi^T(t)\right)\hat{\theta}_{t-1} + \varphi(t)y(t)\right] = \\ &= R(t)^{-1}R(t)\hat{\theta}_{t-1} + R(t)^{-1}\varphi(t)y(t) - R(t)^{-1}\varphi(t)\varphi^T(t)\hat{\theta}_{t-1} = \\ &= \hat{\theta}_{t-1} + R(t)^{-1}\varphi(t)\left[y(t) - \varphi^T(t)\hat{\theta}_{t-1}\right]. \end{aligned}$$

Hence, one can update the estimate $\hat{\theta}_t$ at each time t by using the following recursive algorithm

$$\begin{aligned} \hat{\theta}_t &= \hat{\theta}_{t-1} + R(t)^{-1}\varphi(t)\left[y(t) - \varphi^T(t)\hat{\theta}_{t-1}\right] \\ R(t+1) &= R(t) + \varphi(t+1)\varphi^T(t+1) \end{aligned} \tag{4.63}$$

Inverting the matrix $R(t) \in \mathbb{R}^{d\times d}$ requires $O(d^3)$ operations at each time step. In order to derive a more efficient algorithm, it is better to propagate the inverse $P(t) = R(t)^{-1}$. In fact, by exploiting the Matrix Inversion Lemma (4.32), one gets

$$\begin{aligned} P(t) &= \left[R(t-1) + \varphi(t)\varphi^T(t)\right]^{-1} = \\ &= P(t-1) - P(t-1)\varphi(t)\left[1 + \varphi^T(t)P(t-1)\varphi(t)\right]^{-1}\varphi^T(t)P(t-1) = \\ &= P(t-1) - \frac{P(t-1)\varphi(t)\varphi^T(t)P(t-1)}{1 + \varphi^T(t)P(t-1)\varphi(t)} \end{aligned}$$

Moreover,

$$\begin{aligned} R(t)^{-1}\varphi(t) &= P(t)\varphi(t) \\ &= P(t-1) - \frac{P(t-1)\varphi(t)\varphi^T(t)P(t-1)}{1+\varphi^T(t)P(t-1)\varphi(t)}\varphi(t) \\ &= \frac{P(t-1)\varphi(t)}{1+\varphi^T(t)P(t-1)\varphi(t)}. \end{aligned}$$

By substituting the above relationships into (4.63) one obtains the *Recursive Least Squares (RLS)* algorithm

$$\begin{aligned} \hat{\theta}_t &= \hat{\theta}_{t-1} + L(t)[y(t) - \varphi^T(t)\hat{\theta}_{t-1}] \\ L(t) &= \frac{P(t-1)\varphi(t)}{1+\varphi^T(t)P(t-1)\varphi(t)} \\ P(t) &= P(t-1) - \frac{P(t-1)\varphi(t)\varphi^T(t)P(t-1)}{1+\varphi^T(t)P(t-1)\varphi(t)} \end{aligned} \tag{4.64}$$

The RLS algorithm has computational complexity $O(d^2)$. It should be remarked that the first equation in (4.64) uses the output prediction error $y(t) - \hat{y}(t|t-1) = y(t) - \varphi^T(t)\hat{\theta}_{t-1}$ in order to update the parameter estimate from time $t-1$ to time t, by means of the RLS gain $L(t)$, which in turn depends on the matrix $P(t)$ and the regressor vector $\varphi(t)$.

It is interesting to highlight that the RLS algorithm can be seen as an application of the Kalman filter to a time-varying linear system. In fact, define the state vector $x(t)$ as the unknown parameter vector θ to be estimated. Being θ constant, for a linear regression model $y(t) = \varphi^T(t)\theta + e(t)$ one can write the state space model

$$\begin{aligned} x(t+1) &= x(t) \\ y(t) &= \varphi^T(t)x(t) + e(t). \end{aligned} \tag{4.65}$$

This clearly falls within the general framework of the LMSE state estimation problem (4.11), by setting $A = I$, $B = G = Q = 0$, $C = \varphi^T(t)$ and $v(t) = e(t) \sim WP(0, \sigma_e^2)$, i.e., $R = \sigma_e^2$. If we apply to system (4.65) the KF

algorithm, we get the recursions

$$\begin{aligned} \hat{x}(t) &= \hat{x}(t-1) + K(t)[y(t) - \varphi^T(t)\hat{x}(t-1)] \\ K(t) &= \frac{P(t-1)\varphi(t)}{\sigma_e^2 + \varphi^T(t)P(t-1)\varphi(t)} \\ P(t) &= P(t-1) - \frac{P(t-1)\varphi(t)\varphi^T(t)P(t-1)}{\sigma_e^2 + \varphi^T(t)P(t-1)\varphi(t)} \end{aligned} \tag{4.66}$$

where we used the shorthand notations $\hat{x}(t)$ and $P(t)$, in place of $\hat{x}(t|t) = \hat{x}(t|t-1)$ and $P(t|t) = P(t|t-1)$, respectively. It is easy to see that equations (4.66) coincide with those in (4.64), if we set $\sigma_e^2 = 1$. Notice that this can be done without loss of generality, as one can always scale the linear regression model by the standard deviation σ_e thus obtaining a new output equation

$$\tilde{y}(t) = \tilde{\varphi}^T(t)\theta + \tilde{e}(t) \tag{4.67}$$

where $\tilde{y}(t) = \frac{y(t)}{\sigma_e}$, $\tilde{\varphi}(t) = \frac{\varphi(t)}{\sigma_e}$, $\tilde{e}(t) = \frac{e(t)}{\sigma_e}$ and $\mathbf{E}\left[\tilde{e}^2(t)\right] = 1$. If we apply the RLS algorithm (4.64) to the linear regression model (4.67) , the resulting sequence of estimates $\hat{\theta}_t$ is exactly the same as the sequence $\hat{x}(t)$ returned by the recursion (4.66).

4.6.1 Initialization of the RLS algorithm

There are several possible ways to initialize the RLS recursion. If one uses the form (4.63), the matrices $R(t)$ must be full rank (i.e., $\text{rank}(R(t)) = d$), in order to be invertible. Hence, one can use an initial batch of data for $t = 1, 2, \ldots, t_0$ such that $\text{rank}(R(t_0)) = d$ (notice that it must be $t_0 \geq d$). Then, set

$$\hat{\theta}_{t_0} = R(t_0)^{-1} f(t_0)$$

and iterate the RLS algorithm from t_0 onwards (i.e., for $t = t_0+1, t_0+2, \cdots$). Notice that this ensures that the resulting estimates $\hat{\theta}_t$ are exactly equal to the batch solution (4.62), for all $t > t_0$.

On the other hand, if one uses the form (4.64), it is possible to start the recursion straight away at time $t = 1$ (i.e., when the first data pair $y(1), \varphi(1)$

is available). This can be done by choosing any $\hat{\theta}_0 \in \mathbb{R}^d$ and any matrix $P(0) = P_0 \in \mathbb{R}^{d\times d}$ such that $P_0 = P_0^T > 0$. In this case, the RLS algorithm will return at every time t the solution of the following optimization problem

$$\hat{\theta}_t = \arg\min_{\theta} \sum_{k=1}^{t} \left(y(t) - \varphi^T(t)\theta\right)^2 + (\theta - \hat{\theta}_0)P_0^{-1}(\theta - \hat{\theta}_0) \tag{4.68}$$

where it is apparent the influence of the initialization. Clearly, the larger is P_0, the smaller is such an influence. In any case, the effect of the initialization will decay asymptotically and the estimate resulting from (4.68) will converge to the least squares solution (4.62), as t grows.

4.6.2 Slowly time-varying parameters

The interpretation of the RLS algorithm as an application of the Kalman filter, allows one to devise some useful extensions of the recursive parameter estimation setting. Let us first consider the case of *slowly time-varying parameters*, i.e., parameters that may show a drift which is typically much slower than the dynamics of the output process $y(t)$. This can be modeled by the random walk

$$x(t+1) = x(t) + w(t)$$

where $w(t) \sim WP(0, Q)$. The smaller is Q, the slower is the time variation in the parameter vector θ (which is modeled by $x(t)$). By applying the KF recursions, one gets the algorithm

$$\begin{aligned} \hat{x}(t) &= \hat{x}(t-1) + K(t)[y(t) - \varphi^T(t)\hat{x}(t-1)] \\ K(t) &= \frac{P(t-1)\varphi(t)}{\sigma_e^2 + \varphi^T(t)P(t-1)\varphi(t)} \\ P(t) &= P(t-1) - \frac{P(t-1)\varphi(t)\varphi^T(t)P(t-1)}{\sigma_e^2 + \varphi^T(t)P(t-1)\varphi(t)} + Q \end{aligned} \tag{4.69}$$

where the only difference with the standard RLS recursion in (4.64) is just the presence of the matrix Q in the last equation.

4.6.3 Exponential data weighting

A possible problem of the RLS algorithm is that it may not be sufficiently reactive to *abrupt changes* (i.e., rapid variations in θ), once a large amount of data has been processed. The interpretation in terms of KF provides an explanation for such a phenomenon: the matrix $P(t)$ represents the covariance of the estimation error and therefore it will become smaller and smaller as long as new data are processed. This will lead in turn to a vary small gain $L(t)$ and thus the algorithm will be almost insensitive even to large output prediction errors $y(t) - \varphi^T(t)\hat{x}(t-1)$.

A key idea to overcome this problem is to weight more the last data with respect to the old ones. For example, this can be done by introducing an *exponential data weighting*, which corresponds to solving the optimization problem

$$\min_{\theta} \sum_{k=1}^{t} \lambda^{t-k} \left(y(t) - \varphi^T(t)\theta\right)^2 + \lambda^t(\theta - \hat{\theta}_0)P_0^{-1}(\theta - \hat{\theta}_0)$$

where $\lambda \in (0,1)$ is the *forgetting factor*, enforcing higher weights for more recent data. Typical values for λ are chosen in the range $[0.95, 0.999]$, with smaller values corresponding to more reactive algorithms. The resulting Exponentially Weighted RLS algorithm (EW-RLS) turns out to be

$$\begin{aligned}
\hat{\theta}_t &= \hat{\theta}_{t-1} + L(t)\left[y(t) - \varphi^T(t)\hat{\theta}_{t-1}\right] \\
L(t) &= \frac{P(t-1)\varphi(t)}{\lambda + \varphi^T(t)P(t-1)\varphi(t)} \\
P(t) &= \frac{1}{\lambda}\left[P(t-1) - \frac{P(t-1)\varphi(t)\varphi^T(t)P(t-1)}{1 + \varphi^T(t)P(t-1)\varphi(t)}\right]
\end{aligned}$$

The presence of λ in the second and third equation prevents $P(t)$ and $L(t)$ from vanishing to zero. On the other hand, a too small λ will lead to large values of the covariance $P(t)$ of the estimation error, which clearly affects the quality of the estimate (once again, a reduction in the bias error may result in an increase of the error variance).