ELE8088: Control & Estimation Theory **QUB, 2021**

Handout 14: Kalman Filter

Lecturer: Pantelis Sopasakis *Date:* _____

Topics: Gauss-Markov model ∘ Kalman Filter ∘ KF is BLUE ∘ Bayesian Interpretation of the KF ∘ Maximum likelihood estimation ∘ Maximum *a posteriori* estimation ∘ KF is a recursive MAPE ∘ Forward Dynamic Programming and KF.

Last update: May 29, 2022 at 16:45:37

14.1 Gauss-Markov Model

Consider the linear dynamical system (without an input)

$$x_{t+1} = A_t x_t + G_t w_t, \tag{14.1a}$$
$$y_t = C_t x_t + v_t, \tag{14.1b}$$

where $x_t \in \mathbb{R}^{n_x}$ is the system state, $y_t \in \mathbb{R}^{n_y}$ is the *output*, $w_t \in \mathbb{R}^{n_w}$ is a noise term acting on the system dynamics known as *process noise*, and $v_t \in \mathbb{R}^{n_v}$ is a measurement noise term.

Assumptions: (i) $\mathbb{E}[w_t] = 0$ and $\mathbb{E}[v_t] = 0$ for all $t \in \mathbb{N}$, (ii) x_0, $(w_t)_t$ and $(v_t)_t$ are mutually independent random variables[1], (iii) w_t and v_t are normally distributed and $\mathbb{E}[w_t w_t^\intercal] = Q_t$, $\mathbb{E}[v_t v_t^\intercal] = R_t$. Lastly, x_0 is a random variable and (iv) $x_0 \sim \mathcal{N}(\tilde{x}_0, P_0)$.

As an example, consider the system

$$x_{t+1} = \left[\begin{smallmatrix} 0.5 & 0.3 \\ -0.2 & 0.5 \end{smallmatrix}\right] x_t + w_t, \tag{14.2}$$

where $w_t \sim \mathcal{N}\left(\left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0.10 & 0.05 \\ 0.05 & 0.15 \end{smallmatrix}\right]\right)$, and the initial condition:

$$x_0 \sim \mathcal{N}\left(\left[\begin{smallmatrix} 5 \\ -1 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0.9 & 0.4 \\ 0.4 & 0.3 \end{smallmatrix}\right]\right). \tag{14.3}$$

The evolution of the system states — which are random variables — starting from the above initial condition is illustrated in Figure 14.1.

For the system in Equation (14.1), define $\tilde{x}_t = \mathbb{E}[x_t]$; then,

$$\tilde{x}_{t+1} = \mathbb{E}[x_{t+1}] = \mathbb{E}[A_t x_t + G_t w_t] = A_t \mathbb{E}[x_t] = A_t \tilde{x}_t. \tag{14.4}$$

Define $P_t = \mathrm{Var}[x_t]$. Then,

$$\begin{aligned} P_{t+1} &= \mathbb{E}\Big[(x_{t+1} - \tilde{x}_{t+1})(x_{t+1} - \tilde{x}_{t+1})^\intercal\Big] \\ &= \mathbb{E}\Big[(A_t x_t + G_t w_t - A_t \tilde{x}_t)(A_t x_t + G_t w_t - A_t \tilde{x}_t)^\intercal\Big] \\ &= \mathbb{E}\Big[(A_t(x_t - \tilde{x}_t) + G_t w_t)(A_t(x_t - \tilde{x}_t) + G_t w_t)^\intercal\Big] \\ &= \mathbb{E}\Big[A_t(x_t - \tilde{x}_t)(x_t - \tilde{x}_t)^\intercal A_t^\intercal + 2A_t(x_t - \tilde{x}_t) w_t^\intercal G_t^\intercal + G_t w_t w_t^\intercal G_t^\intercal\Big] \\ &= A_t P_t A_t^\intercal + G_t Q_t G_t^\intercal. \end{aligned} \tag{14.5}$$

[1] therefore, $\mathbb{E}[w_t w_l^\intercal] = 0$ for $t \neq l$, and $\mathbb{E}[v_t v_l^\intercal] = 0$ for $t \neq l$

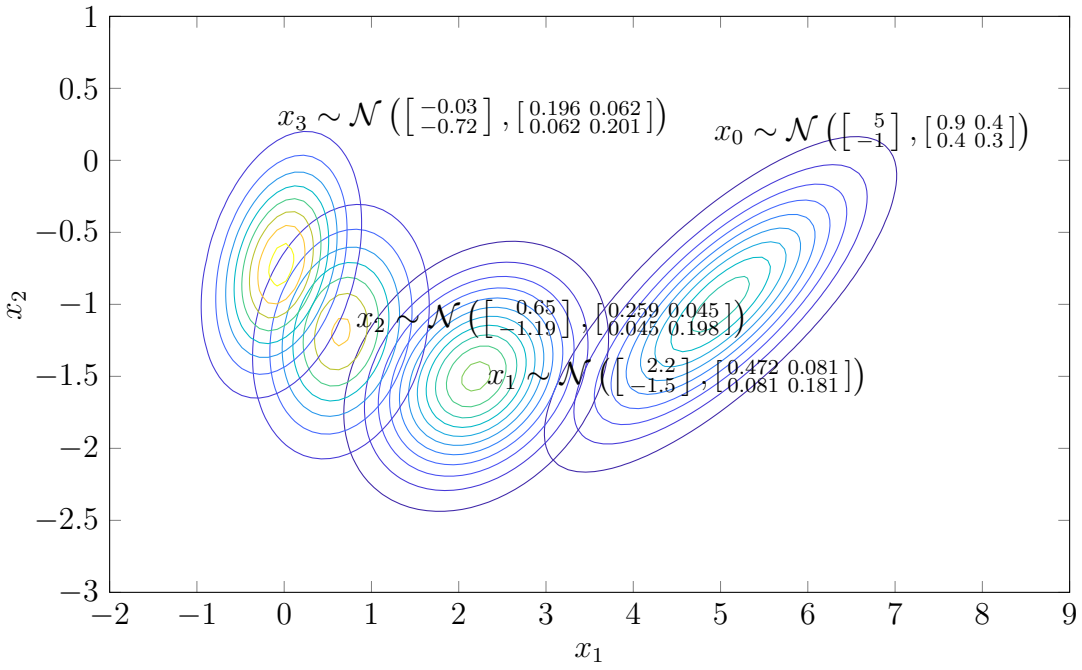

Figure 14.1: Illustration of the Gauss-Markov model

Exercise 1 (🔥🔥). For simplicity, assume that $A_t = A$, $B_t = B$, $G_t = G$ and $Q_t = Q$ for all $t \in \mathbb{N}$. Determine the covariance matrix $P_{t,l} := \mathrm{Cov}(x_t, x_l)$ for $t, l \in \mathbb{N}$, as a function of P_0, A, G, Q, t and l.

Exercise 2 (🔥🔥🔥). Show that the random process $(x_t)_t$ is a Markov process, i.e.,

$$p_{x_{t+1}|x_0,x_1,\ldots,x_t}(x_{t+1} \mid x_0, x_1, \ldots, x_t) = p_{x_{t+1}|x_t}(x_{t+1} \mid x_t), \tag{14.6}$$

but $(y_t)_t$ is not necessarily Markovian[2]. More generally, for $0 \leq t_0 < t_1 < \ldots < t_k \leq t$, show that

$$p_{x_{t+1}|x_{t_0},x_{t_1},\ldots,x_{t_k}}(x_{t+1} \mid x_{t_0}, x_{t_1}, \ldots, x_{t_k}) = p_{x_{t+1}|x_{t_k}}(x_{t+1} \mid x_{t_k}). \tag{14.7}$$

Exercise 3 (🔥🔥). Consider the following dynamical system

$$x_{t+1} = \begin{bmatrix}1 & 0\\0.1 & 1\end{bmatrix} x_t + \begin{bmatrix}1\\0\end{bmatrix} w_t, \tag{14.8a}$$

$$y_1 = \begin{bmatrix}1 & 1\end{bmatrix} x_t + v_t, \tag{14.8b}$$

where $w_t \sim \mathcal{N}(0, 1)$, $v_t \sim \mathcal{N}(0, 5)$ and $x_0 \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}20 & 5\\5 & 20\end{bmatrix}\right)$. Write a Python or MATLAB script to determine $\tilde{x}_{20}$, $P_{20} = \mathbb{E}[(x_{20} - \tilde{x}_{20})(x_{20} - \tilde{x}_{20})^\intercal]$ and $\mathrm{Cov}(x_{20}, x_{50})$.

[2]It suffices to show that $\mathbb{E}[y_2 \mid y_1]$ is not the same as $\mathbb{E}[y_2 \mid y_0, y_1]$.

Exercise 4 (☕☕). [Important] Consider a dynamical system of the form $x_{t+1} = Ax_t + w_t$, where w_t is a time-uncorrelated random variable with zero mean (not necessarily normally distributed) and the pdf of w_t is a function p_{w_t}. Let us denote the conditional density function of x_{t+1} given x_t by $p_{x_{t+1}|x_t}$; then show that[3]

$$p_{x_{t+1}|x_t}(x_{t+1} \mid x_t) = p_{w_t}(x_{t+1} - Ax_t). \tag{14.9}$$

[3]Hint: to show that two pfds are equal, you may show that the corresponding cdfs are equal (why?) — the latter is easier.

14.2 Kalman Filter

The random variables x_0 and y_0 are jointly normal with $\mathbb{E}[x_0] = \tilde{x}_0$, $\mathbb{E}[y_0] = \mathbb{E}[C_0 x_0 + v_0] = C_0\tilde{x}_0$. The variance of x_0 is $\mathrm{Var}[x_0] = P_0$. The variance of y_0 is

$$\mathrm{Var}[y_0] = \mathrm{Var}[C_0 x_0 + v_0] = C_0 P_0 C_0^\intercal + R_0. \tag{14.10}$$

The covariance of x_0 with y_0 is

$$\begin{aligned}
\mathrm{Cov}(x_0, y_0) &= \mathbb{E}[(x_0 - \tilde{x}_0)(y_0 - \tilde{y}_0)^\intercal], \qquad \text{where } \tilde{y}_0 = \mathbb{E}[y_0] \\
&= \mathbb{E}[(x_0 - \tilde{x}_0)(C_0 x_0 + v_0 - C_0\tilde{x}_0)^\intercal] \\
&= \mathbb{E}[(x_0 - \tilde{x}_0)(C_0(x_0 - \tilde{x}_0) + v_0)^\intercal] \\
&= \mathbb{E}[(x_0 - \tilde{x}_0)(x_0 - \tilde{x}_0)^\intercal C_0^\intercal + (x_0 - \tilde{x}_0)v_0^\intercal] = P_0 C_0^\intercal.
\end{aligned} \tag{14.11}$$

Therefore,

$$\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \tilde{x}_0 \\ C_0\tilde{x}_0 \end{bmatrix}, \begin{bmatrix} P_0 & P_0 C_0^\intercal \\ C_0 P_0 & C_0 P_0 C_0^\intercal + R_0 \end{bmatrix} \right). \tag{14.12}$$

Suppose we measure y_0. What is x_0 given y_0? Since (x_0, y_0) is jointly normally distributed as in Equation (14.12), $x_0 \mid y_0$ is normally distributed. We may define the estimator $\hat{x}_{0|0} := \mathbb{E}[x_0 \mid y_0]$, which is[4]

$$\hat{x}_{0|0} = \tilde{x}_0 + P_0 C_0^\intercal (C_0 P_0 C_0^\intercal + R_0)^{-1}(y_0 - C_0\tilde{x}_0), \tag{14.13}$$

and the estimator variance, $\Sigma_{0|0} := \mathrm{Var}[x_0 \mid y_0]$, which is[5]

$$\Sigma_{0|0} = P_0 - P_0 C_0^\intercal (C_0 P_0 C_0^\intercal + R_0)^{-1} C_0 P_0 \tag{14.14}$$

Having observed y_0 at $t = 0$ we want to estimate x_1; we compute $\hat{x}_{1|0} := \mathbb{E}[x_1 \mid y_0]$ which is

$$\hat{x}_{1|0} = A_0 \hat{x}_{0|0}. \tag{14.15}$$

The estimator variance, $\Sigma_{1|0} = \mathrm{Var}[x_1 \mid y_0]$, is

$$\Sigma_{1|0} = A_0 \Sigma_{0|0} A_0^\intercal + G_0 Q_0 G_0^\intercal. \tag{14.16}$$

[4] See Handout 11, Theorem 11.1 (conditioning of multivariate normals)

[5] See Handout 11, Section 11.3.1 (multivariate normal distributions).

The output at $t = 1$ given the observation of y_0 is expected to be

$$\hat{y}_{1|0} = \mathbb{E}[y_1 \mid y_0] = C_1 \hat{x}_{1|0}, \tag{14.17}$$

and its (conditional) variance is

$$\mathrm{Var}[y_1 \mid y_0] = C_1 \Sigma_{1|0} C_1^\intercal + R_1, \tag{14.18}$$

(can you see why?) and the covariance between x_1 and y_1, conditional on y_0, is

$$\mathrm{Cov}(x_1, y_1 \mid y_0) := \mathbb{E}[(x_1 - \tilde{x}_1)(y_1 - \tilde{y}_1)^\intercal \mid y_0] = \Sigma_{1|0} C_1^\intercal. \tag{14.19}$$

Overall,

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \Bigg| \, y_0 \sim \mathcal{N}\left(\begin{bmatrix} \hat{x}_{1|0} \\ C_1 \hat{x}_{1|0} \end{bmatrix}, \begin{bmatrix} \Sigma_{1|0} & \Sigma_{1|0} C_1^\intercal \\ C_1 \Sigma_{1|0} & C_1 \Sigma_{1|0} C_1^\intercal + R_1 \end{bmatrix} \right). \tag{14.20}$$

Once we obtain a measurement y_1,

$$\hat{x}_{1|1} = \mathbb{E}[x_1 \mid y_0, y_1] = \hat{x}_{1|0} + \Sigma_{1|0} C_1^\intercal (C_1 \Sigma_{1|0} C_1^\intercal + R_1)^{-1} (y_1 - C_1 \hat{x}_{1|0}), \tag{14.21a}$$

$$\Sigma_{1|1} = \mathrm{Var}[x_1 \mid y_0, y_1] = \Sigma_{1|0} - \Sigma_{1|0} C_1^\intercal (C_1 \Sigma_{1|0} C_1^\intercal + R_1)^{-1} C_1 \Sigma_{1|0}. \tag{14.21b}$$

Kalman Filter Equations:

$$\text{Measurement update} \left[\begin{array}{l} \hat{x}_{t|t} = \hat{x}_{t|t-1} + \Sigma_{t|t-1} C_t^\intercal (C_t \Sigma_{t|t-1} C_t^\intercal + R_t)^{-1} (y_t - C_t \hat{x}_{t|t-1}) \\ \Sigma_{t|t} = \Sigma_{t|t-1} - \Sigma_{t|t-1} C_t^\intercal (C_t \Sigma_{t|t-1} C_t^\intercal + R_t)^{-1} C_t \Sigma_{t|t-1} \end{array} \right.$$

$$\text{Time update} \left[\begin{array}{l} \hat{x}_{t+1|t} = A_t \hat{x}_{t|t} \\ \Sigma_{t+1|t} = A_t \Sigma_{t|t} A_t^\intercal + G_t Q_t G_t^\intercal \end{array} \right.$$

$$\text{Initial conditions} \left[\begin{array}{l} \hat{x}_{0|-1} = \tilde{x}_0 \\ \Sigma_{0|-1} = P_0 \end{array} \right.$$

Note that we have defined: (i) $\hat{x}_{t|t} := \mathbb{E}[x_t \mid y_0, y_1, \ldots, y_t]$, (ii) $\hat{x}_{t+1|t} := \mathbb{E}[x_{t+1} \mid y_0, y_1, \ldots, y_t]$, (iii) $\Sigma_{t|t} := \mathrm{Var}[x_t \mid y_0, y_1, \ldots, y_t]$, and (iv) $\Sigma_{t+1|t} := \mathrm{Var}[x_{t+1} \mid y_0, y_1, \ldots, y_t]$.

Exercise 5 (☕☕☕). Derive the Kalman filter equations for the dynamical system

$$\begin{aligned} x_{t+1} &= A_t x_t + B_t u_t + G_t w_t, \\ y_t &= C_t x_t + D_t u_t + v_t, \end{aligned}$$

where u_t is a control signal that can be observed and w_t, v_t and x_0 are as before.

Exercise 6 (☕). Implement the Kalman filter for the following dynamical system

$$x_{t+1} = x_t + w_t, \tag{14.22a}$$

$$y_t = \begin{bmatrix} 1 & -2 \end{bmatrix} x_t + v_t, \tag{14.22b}$$

where $x_t \in \mathbb{R}^2$, $y_t \in \mathbb{R}$, $w_t \sim \mathcal{N}(0, I)$, $v_t \sim \mathcal{N}(0, 5)$, $x_0 \sim \mathcal{N}(5, 100)$.

Remarks: The covariance matrices are updated according to

$$\Sigma_{t+1|t} = A_t \Sigma_{t|t-1} A_t^\intercal + A_t \Sigma_{t|t-1} C_t^\intercal (C_t \Sigma_{t|t-1} C_t^\intercal + R_t)^{-1} C_t \Sigma_{t|t-1} A_t^\intercal + G_t Q_t G_t^\intercal, \tag{14.23}$$

which is a *Riccati recursion*! We know that the Riccati recursion — under certain assumptions[6] — converges to a steady-state matrix Σ_∞, i.e., $\Sigma_{t+1|t} \to \infty$ as $t \to \infty$. The covariance matrices can be computed without the need to obtain any system data (independent of y_t). The state estimates are essentially conditional expectations As such, the Kalman filter is the "best" we can achieve (minimum conditional variance estimator).

[6] **Exercise 7 (☕).** What are these assumptions?

14.3 Application: GPS positioning system

14.3.1 Problem Statement and Kalman Filter

The initial position, x_0, of a vehicle is inexactly known: $x_0 \sim \mathcal{N}(0, 100)$. The velocity of the vehicle, u_t, is approximately $10\,\mathrm{m/s}$ in the following sense $u_t \sim \mathcal{N}(10, 8)$. The position of the vehicle can be measured using a GPS system every $h = 0.05\,\mathrm{s}$ using a sensor with additive noise $v_t \sim \mathcal{N}(0, 15)$. The dynamics of the position of the vehicle is

$$x_{t+1} = x_t + hu_t, \tag{14.24a}$$

$$y_t = x_t + v_t. \tag{14.24b}$$

However, u_t is not a zero-mean random variable!

The velocity can be written as $u_t = \bar{u}_t + w_t$, where $\bar{u}_t = 10\,\mathrm{m/s}$ and $w_t \sim \mathcal{N}(0, 8)$. Since $\bar{u}_t$ is constant, $\bar{u}_{t+1} = \bar{u}_t$, so

$$\begin{bmatrix} x_{t+1} \\ \bar{u}_{t+1} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_t \\ \bar{u}_t \end{bmatrix}}_{\boldsymbol{x}_t} + \underbrace{\begin{bmatrix} h \\ 0 \end{bmatrix}}_{G} w_t. \tag{14.25}$$

The state of the system is $\boldsymbol{x}_t = (x_t, \bar{u}_t)$ with

$$\boldsymbol{x}_0 \sim \mathcal{N}\left(\left[\begin{smallmatrix} 0 \\ 10 \end{smallmatrix}\right], \left[\begin{smallmatrix} 100 & 0 \\ 0 & 0 \end{smallmatrix}\right]\right).$$

The output of the system is $y_t = [1\ 0]\,\boldsymbol{x}_t + v_t$. Overall, the system is in the form of Equation (14.1), where

$$A = \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix}, G = \begin{bmatrix} h \\ 0 \end{bmatrix}, C = \begin{bmatrix} 1 & 0 \end{bmatrix}, P_0 = \begin{bmatrix} 100 & 0 \\ 0 & 0 \end{bmatrix}, Q = 8, R = 15,$$

and $\bar{\boldsymbol{x}}_0 = \left[\begin{smallmatrix} \tilde{x}_0 \\ \tilde{\bar{u}}_0 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 0 \\ 10 \end{smallmatrix}\right]$.

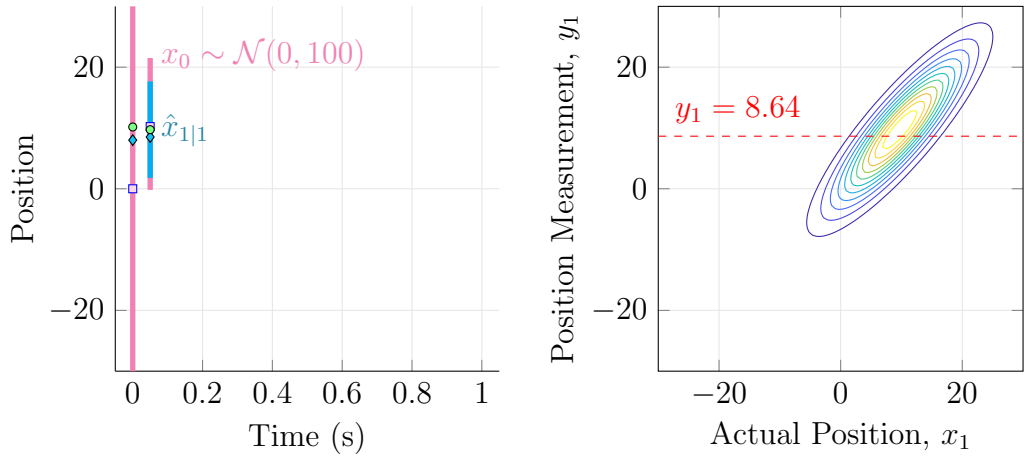

Figure 14.2: (Left) Estimated positions, $\hat{x}_{t|t-1}$ (□), $\hat{x}_{t|t}$ (♦) and the unknown true position (●). The vertical bars show the $\pm 3\sigma^2$-intervals around the estimated positions. (Right) Joint distribution of (x_1, y_1) and a measurement $y_1 = 8.64$.

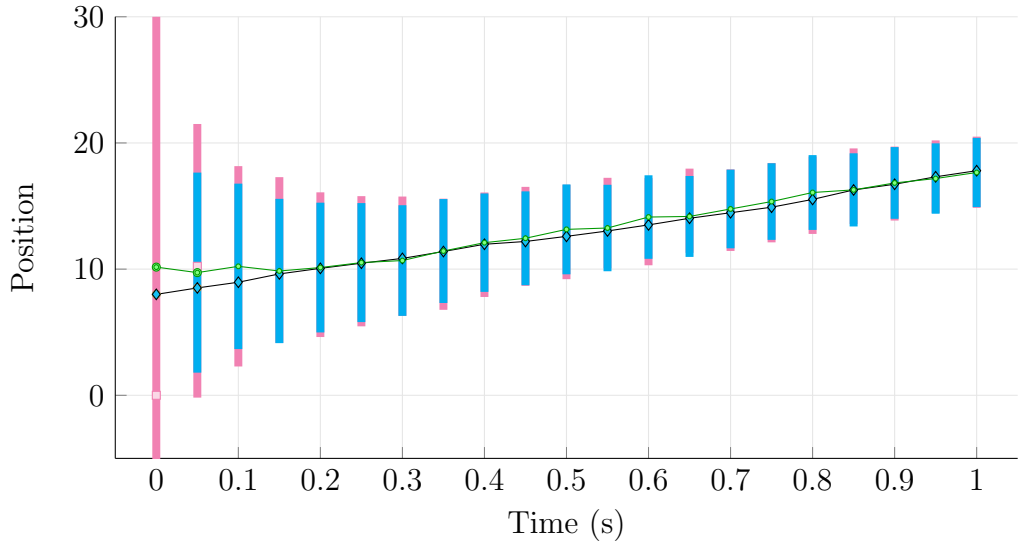

Figure 14.3: Estimated positions, $\hat{x}_{t|t-1}$ (□), $\hat{x}_{t|t}$ (♦) and the unknown true position (●). We observe that one new data becomes available, the variance of the estimated position decreases. We also see that $\Sigma_{t+1|t}$ converges to a finite variance and that in this example the estimation error is very small.

14.3.2 Intermittent measurements

Suppose we obtain measurements at $t = 0, 1, \ldots, t_1$, but then the connection to the GPS breaks so we have no measurements from $t_1 + 1$ to $t_2 - 1$. At time t_2 the connection is recovered. In that case, we can still apply the Kalman filter by applying only time update steps when we do not have measurements. This is illustrated in Figure 14.4.

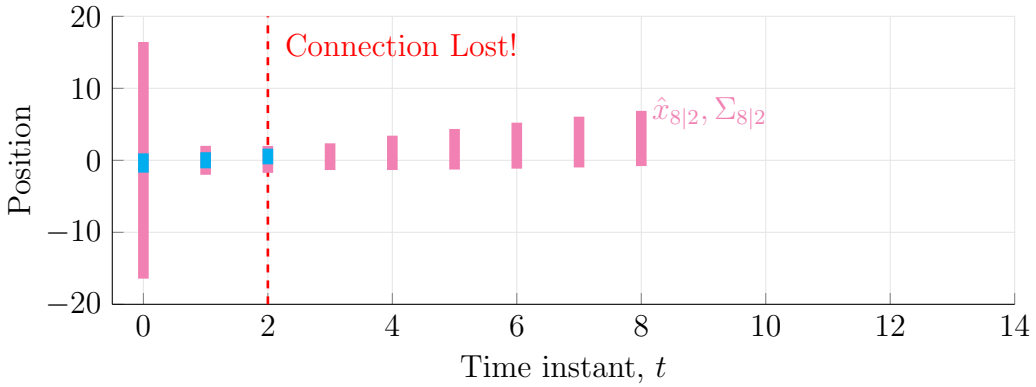

Figure 14.4: The connection is lost at $t_1 = 2$ and there are no measurements from time 3 to time 7; then, at $t_2 = 8$, the connection to the GPS is recovered (not shown here). Meanwhile, we compute the estimates $\hat{x}_{3|2}, \ldots, \hat{x}_{8|2}$ and variances $\Sigma_{3|2}, \ldots, \Sigma_{8|2}$.

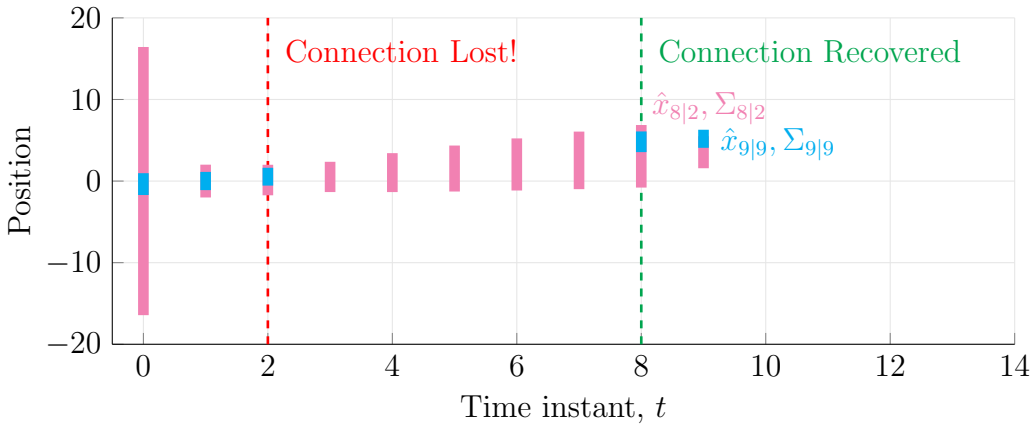

Figure 14.5: The connection is recovered at time $t_2 = 8$; subsequently, we can keep interleaving time and measurement update steps.

14.4 The KF is BLUE

In this section we will drop the normality assumptions. The Kalman filter is a linear (affine[7]) estimator. By combining the measurement and time updates of the Kalman filter, we can see that

$$\boxed{\hat{x}_{t+1|t} = \underbrace{A_t\hat{x}_{t|t-1}}_{\text{system dynamics}} + K_t \underbrace{(y_t - C_t\hat{x}_{t|t-1})}_{\text{prediction error}},} \tag{14.26}$$

where $K_t = A_t\Sigma_{t|t-1}C_t^\intercal(C_t\Sigma_{t|t-1}C_t^\intercal + R_t)^{-1}$. The Kalman filter is an *affine filter*. It is in fact the *"best"* affine filter (will explain). Recall that, by definition, $\hat{x}_{t+1|t} = \mathbb{E}[x_{t+1} \mid y_0, \ldots, y_t]$, therefore, the KF is *unbiased*, i.e.,

$$\mathbb{E}[\hat{x}_{t+1|t} - x_{t+1}] = \mathbb{E}[\mathbb{E}[x_{t+1} \mid y_0, \ldots, y_t] - x_{t+1}] = 0. \tag{14.27}$$

The conditional expectation of a random variable X given $Y = y$ minimises the following function[8]

$$f(z; y) = \mathbb{E}\left[\|X - z\|^2 \mid Y = y\right]. \tag{14.28}$$

This means that

$$f\left(\mathbb{E}[X \mid Y = y]; y\right) \leq f(z; y), \tag{14.29}$$

for all estimators $z(y)$. Moreover, define the function $F(z; y) = \mathbb{E}[(X - z)(X - z)^\intercal \mid Y = y]$. Then,

$$\boxed{F\left(\mathbb{E}[X \mid Y = y]; y\right) \preccurlyeq F(z; y),} \tag{14.30}$$

However, $\mathbb{E}[X \mid Y = y]$ can be difficult to determine (without the normality assumption); in general, it is a *nonlinear* function of y.

Question: What is the best *linear* (affine) estimator we can construct?

Suppose that X and Y are jointly distributed. Then the *best* (minimum variance) estimator of X given $Y = y$ is $\mathbb{E}[X \mid Y = y]$. What is the best *affine* estimator?

Problem statement: Without assuming that X and Y are (jointly) normally distributed, suppose we are looking for estimators of the form $\widehat{X}(y) = Ay + b$, i.e., *affine* estimators. We

[7] Often the Kalman Filter is said to be "linear," but it is in fact an "affine" one.

[8] See Handout 12, Minimum Variance Estimation Theorem.

seek to determine $A = A^\star$ and $b = b^\star$ so that $\widehat{X}(y) = A^\star y + b^\star$ is the *best affine estimator*, i.e., the best estimator among all affine ones, i.e.,

$$\mathbb{E}\left[\|X - A^\star y + b^\star\|^2\right] \leq \mathbb{E}\left[\|X - Ay + b\|^2\right], \tag{14.31}$$

for any A and b, and its conditional counterpart

$$\mathbb{E}\left[\|X - (A^\star y + b^\star)\|^2 \mid Y = y\right] \leq \mathbb{E}\left[\|X - (Ay + b)\|^2 \mid Y = y\right],$$

also holds for any A and b.

Theorem 14.1 (KF is BLUE) *Suppose that (X, Y) are jointly distributed with means $\mathbb{E}[X] = m_x$, $\mathbb{E}[Y] = m_y$ and covariance matrix*

$$\mathrm{Var}\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}, \tag{14.32}$$

with $\Sigma_{yy} \succ 0$. The best affine estimator of X given Y is $\widehat{X}(Y) = A^\star Y + b^\star$ with

$$A^\star = \Sigma_{xy}\Sigma_{yy}^{-1}, \text{ and } b^\star = m_x - A^\star m_y. \tag{14.33}$$

In particular, $\mathbb{E}\left[\|X - \widehat{X}(Y)\|^2\right] \leq \mathbb{E}\left[\|X - (AY + b)\|^2\right]$, for any parameters A and b.

Remarks: The estimator can be written as $\widehat{X}(Y) = m_x + A^\star(Y - m_y)$. The Kalman filter is the best affine filter in the sense that it minimises the *mean square error*, $\mathbb{E}[\|X - \widehat{X}\|^2]$. Theorem 14.1 does not require that X or Y be Gaussian. We can prove a similar result for the covariance matrix $\mathbb{E}[(X - \widehat{X})(X - \widehat{X})^\intercal]$ (guess what...). Later we will show that KF is the *best affine* filter.

Before the Proof. We will use the following observations: for an n-dimensional random variable Z:

$$\mathbb{E}[\|Z\|^2] = \mathbb{E}[Z^\intercal Z] = \mathbb{E}[\mathrm{trace}(ZZ^\intercal)] = \mathrm{trace}\,\mathbb{E}[ZZ^\intercal]. \tag{14.34}$$

Secondly, $\mathrm{Var}[Z] = \mathbb{E}[ZZ^\top] - \mathbb{E}[Z]\mathbb{E}[Z]^\intercal$, so $\mathbb{E}[ZZ^\top] = \mathrm{Var}[Z] + \mathbb{E}[Z]\mathbb{E}[Z]^\intercal$, therefore,

$$\mathbb{E}[\|Z\|^2] = \mathrm{trace}\,\mathbb{E}[ZZ^\top] = \mathrm{trace}\,\mathrm{Var}[Z] + \mathrm{trace}\left(\mathbb{E}[Z]\mathbb{E}[Z]^\intercal\right). \tag{14.35}$$

Note also that $\mathbb{E}[X - AY - b] = m_x - Am_y - b$.

Proof: It is

$$\begin{aligned}\mathbb{E}[\|X - AY - b\|^2] &= \operatorname{trace}\operatorname{Var}[X - AY - b] + \operatorname{trace}(\mathbb{E}[X - AY - b](\,\cdot\,)^\intercal) \\ &= \operatorname{trace}\operatorname{Var}[X - AY - b] + \|m_x - Am_y - b\|^2, \end{aligned} \tag{14.36}$$

where

$$\begin{aligned}\operatorname{Var}(X - AY - b) &= \mathbb{E}[(X - m_x - A(Y - m_y))(\,\cdot\,)^\intercal] \\ &= \mathbb{E}[(X - m_x)(X - m_x)^\intercal] + \mathbb{E}[A(Y - m_y)(Y - m_y)^\intercal A^\intercal] \\ &\qquad - A(Y - m_y)\mathbb{E}[X - m_x]^\intercal - \mathbb{E}[X - m_x](Y - m_y)^\intercal A^\intercal \\ &= \Sigma_{xx} + A\Sigma_{yy}A^\intercal - A\Sigma_{yx} - \Sigma_{xy}A^\intercal, \end{aligned} \tag{14.37}$$

therefore,

$$\mathbb{E}[\|X - AY - b\|^2] = \operatorname{trace}\left[\Sigma_{xx} + A\Sigma_{yy}A^\intercal - A\Sigma_{yx} - \Sigma_{xy}A^\intercal\right] + \|m_x - Am_y - b\|^2. \tag{14.38}$$

Now observe that

$$(A - \Sigma_{xy}\Sigma_{yy}^{-1})\Sigma_{yy}(A - \Sigma_{xy}\Sigma_{yy}^{-1})^\intercal = A\Sigma_{yy}A^\intercal - \Sigma_{xy}A^\intercal - A\Sigma_{yx} + \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}. \tag{14.39}$$

The mean square error can be written as

$$\begin{aligned}\mathbb{E}[\|X - AY - b\|^2] = &\underbrace{\operatorname{trace}[\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}]}_{\text{independent of } A \text{ and } b} \\ &+ \operatorname{trace}[(A - \Sigma_{xy}\Sigma_{yy}^{-1})\Sigma_{yy}(A - \Sigma_{xy}\Sigma_{yy}^{-1})^\intercal] + \|m_x - Am_y - b\|^2. \end{aligned} \tag{14.40}$$

All terms are nonnegative. The term $\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$ is the Schur complement of Σ, so it is positive semidefinite. The first term is independent of A and b. The second term can be made 0 by taking $A = \Sigma_{xy}\Sigma_{yy}^{-1}$ and the third term vanishes if we take $b = m_x - Am_y$. ∎

Exercise 8 (☕). Suppose that (X, Y) are jointly distributed with means $\mathbb{E}[X] = m_x$, $\mathbb{E}[Y] = m_y$ and covariance matrix

$$\operatorname{Var}\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}, \tag{14.41}$$

with $\Sigma_{yy} \succ 0$. Showthat the best affine estimator of X given Y is $\widehat{X}(Y) = A^\star Y + b^\star$ with

$$A^\star = \Sigma_{xy}\Sigma_{yy}^{-1}, \text{ and } b^\star = m_x - A^\star m_y, \tag{14.42}$$

in the sense that

$$\mathbb{E}\left[(X-\widehat{X}(y))(X-\widehat{X}(y))^\intercal\right] \preccurlyeq \mathbb{E}\left[(X-Ay-b)(X-Ay-b)^\intercal\right], \tag{14.43}$$

for any parameters A and b.

Hint: follow the steps of the proof of the theorem we just stated, omitting the trace.

Exercise 9 [Estimator bias and variance] **(☕).** Show that the best linear estimator, $\widehat{X}(Y) = A^\star Y + b^\star$, with

$$A^\star = \Sigma_{xy}\Sigma_{yy}^{-1},$$
$$b^\star = m_x - A^\star m_y.$$

is *unbiased*, i.e., $\mathbb{E}[X-\widehat{X}(Y)] = 0$ and its variance (the variance of the estimator error, $X-\widehat{X}(Y)$) is

$$\mathrm{Var}[X-\widehat{X}(Y)] = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}.$$

Assumptions: x_0, $(w_t)_t$ and $(v_t)_t$ are mutually independent random variables (not necessarily Gaussian) with $\mathbb{E}[w_t]=0$, $\mathbb{E}[v_t]=0$, $\mathbb{E}[x_0]=\tilde{x}_0$, and $\mathrm{Var}[w_t]=Q_t$, $\mathrm{Var}[v_t]=R_t$, $\mathrm{Var}[x_0]=P_0$. Then (x_0,y_0) is jointly distributed with mean

$$\mathbb{E}\begin{bmatrix}x_0\\y_0\end{bmatrix} = \begin{bmatrix}\tilde{x}_0\\C_0\tilde{x}_0\end{bmatrix}, \tag{14.44}$$

and variance-covariance matrix

$$\mathrm{Var}\begin{bmatrix}x_0\\y_0\end{bmatrix} = \begin{bmatrix}P_0 & P_0C_0^\intercal\\C_0P_0 & C_0P_0C_0^\intercal+R_0\end{bmatrix}. \tag{14.45}$$

By the BAE Theorem, the best affine estimator of x_0 given y_0 is

$$\hat{x}_0(y_0) = \tilde{x}_0 + P_0C_0^\intercal(C_0P_0C_0^\intercal+R_0)^{-1}(y_0-C_0\tilde{x}_0). \tag{14.46}$$

and the error covariance is

$$\mathrm{Var}[x_0-\hat{x}_0(y_0)] = P_0 - P_0C_0^\intercal(C_0P_0C_0^\intercal+R_0)^{-1}C_0P_0. \tag{14.47}$$

These are the same formulas as in the Kalman filter!

We can easily show recursively that the Kalman filter is the Best (minimum variance, minimal covariance matrix) Linear (actually affine) Unbiased Estimator (BLUE). "Best linear" means that it is the best among all linear estimators — however, **there may be some nonlinear estimator that leads to a lower variance**. Without the normality assumption, the Kalman filter is not a minimum variance estimator.

14.5 Bayesian Interpretation

We will give an alternative interpretation and derivation of the Kalman filter equations. We will first describe the maximum a posteriori (MAP) Bayesian estimation methodology. This will be the basis for the Moving Horizon Estimation methodology that is applicable to nonlinear and constrained systems.

14.5.1 Interlude: Maximum Likelihood Estimation

Suppose a random variable Y has a pdf $p_Y(\cdot\,;\theta)$, that depends parametrically on a (scalar or vector) θ. Note that θ is treated as an unknown, but *deterministic* (not random) parameter and *we do not assume* that we have some prior information about it. For example, if $Y \sim \mathcal{N}(\mu, \sigma^2)$, the parameter vector of the distribution of Y is $\theta = (\mu, \sigma^2)$. Having obtained some independent samples, $\boldsymbol{y}_N = (y_i)_{i=1}^N$, from this distribution, the objective is to estimate θ.

Let us give a few examples. A coin has probability p to land heads and $1-p$ to land tails. We toss it N times. How can we estimate p from our observations? The number of customers entering a store every day is known to follow the Poisson distribution with parameter λ. We observe and record the total number of customers on N different days. How can be estimate λ? The wealth of people in society can be modelled by the Pareto distribution with parameters α and x_m. We randomly select a sample of N people and record their wealth. How can we estimate α and x_m? What is the *most likely* value of these parameters given the observed data? In general, a parameter $\theta \in \Theta$ needs to be estimated from independent samples y_i, $i = 1, \ldots, N$ that follow the same distribution with pdf $p(\,\cdot\,;\theta)$

Suppose that the random variables $\boldsymbol{y}_N = (y_i)_{i=1}^N$ are samples generated by a probability distribution with density $p(\,\cdot\,;\theta)$, there $\theta \in \Theta$ is a parameter from a parameter space Θ. Define the *likelihood function* of θ given $\boldsymbol{y}_N$ as

$$L(\theta; \boldsymbol{y}_N) = p(\boldsymbol{y}_N; \theta). \tag{14.48}$$

The *maximum likelihood estimate* of θ is

$$\hat{\theta}_{\text{mle}} \in \arg\max_{\theta \in \Theta} L(\theta; \boldsymbol{y}_N). \tag{14.49}$$

It is often convenient to work with the log-likelihood function

$$\ell(\theta; \boldsymbol{y}_N) = \log L(\theta; \boldsymbol{y}_N), \tag{14.50}$$

with the convention $\log 0 = -\infty$. In terms of the the log-likelihood function,

$$\hat{\theta}_{\text{mle}} \in \arg\max_{\theta \in \Theta} \ell(\theta; \boldsymbol{y}_N). \tag{14.51}$$

If $y_1, \ldots, y_N$ are independent, then

$$L(\theta; \boldsymbol{y}_N) = p(\boldsymbol{y}_N; \theta) = \prod_{i=1}^{N} p(y_i; \theta) \Rightarrow \ell(\theta, \boldsymbol{y}_N) = \sum_{i=1}^{N} \log p(y_i; \theta). \tag{14.52}$$

Example. Suppose that the samples $y_1, \ldots, y_N$ are independent and follow the univariate normal distribution with an unknown mean μ_0 and unknown variance σ_0^2. The parameter we want to estimate is $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times [0, \infty)$.

The maximum likelihood estimate is

$$\hat{\theta} \in \arg\max_{\theta \in \Theta} \ell(\theta; y_1, \ldots, y_N), \tag{14.53}$$

where

$$\begin{aligned} \ell(\theta; y_1, \ldots, y_N) &= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2} \right] \\ &= -\frac{N}{2} \log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(y_i - \mu)^2}{2\sigma^2}. \end{aligned} \tag{14.54}$$

In order to determine the maximum, we differentiate with respect to μ and σ^2

$$\frac{\partial}{\partial \mu} \ell(\theta; y_1, \ldots, y_N) = \frac{1}{\sigma^2} \left[\sum_{i=1}^{N} y_i - N\mu \right], \tag{14.55a}$$

$$\frac{\partial}{\partial \sigma^2} \ell(\theta; y_1, \ldots, y_N) = \frac{1}{2\sigma^2} \left(\sum_{i=1}^{N} \frac{(y_i - \mu)^2}{\sigma^2} - N \right). \tag{14.55b}$$

By setting the derivatives equal to zero and solving for μ and σ^2 we find that

$$\hat{\mu} = \tfrac{1}{N} \sum_{i=1}^{N} y_i, \tag{14.56a}$$

$$\hat{\sigma}^2 = \tfrac{1}{N} \sum_{i=1}^{N} (y_i - \hat{\mu})^2. \tag{14.56b}$$

Note that the estimate $\hat{\mu}$ (of μ_0) is unbiased

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} y_i\right] = \mu_0, \tag{14.57}$$

But the estimate $\hat{\sigma}^2$ (of σ_0^2) is not unbiased since[9]

$$\mathbb{E}[\hat{\sigma}^2] = \frac{N-1}{N}\sigma_0^2, \tag{14.58}$$

nevertheless, $\hat{\sigma}^2$ is *asymptotically unbiased* since $\mathbb{E}[\hat{\sigma}^2] \to \sigma_0^2$ as $N \to \infty$.

Exercise 10 (☕). (i) Show that the maximum value of the likelihood given in Equation (14.54) is

$$\ell(\hat{\theta}) = -\frac{N}{2}(1 + \log(2\pi\hat{\sigma}^2)). \tag{14.59}$$

(ii) Show that the Hessian of ℓ at $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$ is negative semidefinite, therefore, indeed the maximum of ℓ is attained at $\hat{\theta}$.

Exercise 11 (☕). Show that for the MAP estimates $\hat{\mu}$ and $\hat{\sigma}^2$ given in the previous slide, we have

$$\mathbb{E}[\hat{\mu}] = \mu_0, \tag{14.60a}$$

$$\mathbb{E}[\hat{\sigma}^2] = \frac{N-1}{N}\sigma_0^2. \tag{14.60b}$$

Exercise 12 [Exponential distribution] **(☕).** The exponential distribution with parameter $\lambda > 0$ is

$$p(x;\lambda) = \lambda e^{-\lambda x} 1_{[0,\infty)}(x), \tag{14.61}$$

where $1_{[0,\infty)}(x) = 1$ for $x \geq 0$ and $1_{[0,\infty)}(x) = 0$ otherwise. Determine the max likelihood estimate of λ given a set of independent samples $x_1, \ldots, x_N$. Spoilers: you should find that

$$\hat{\lambda} = \frac{N}{\sum_{i=1}^{N} x_i}. \tag{14.62}$$

Show that this estimator is not unbiased.

[9] See Exercise 11

14.5.2 Maximum a posteriori Estimation

In maximum likelihood estimation (MLE) we treated θ as non-random parameter. In maximum a posteriori estimation (MAPE) we treat θ as a random variable for which there is some known prior, $p(\theta)$.

The *posterior distribution* of θ, given some observation x, becomes[10]

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)} = \frac{p(x \mid \theta)p(\theta)}{\int_\Theta p(x,\theta)\mathrm{d}\theta} = \frac{p(x \mid \theta)p(\theta)}{\int_\Theta p(x \mid \theta)p(\theta)\mathrm{d}\theta}. \tag{14.63}$$

Note that the denominator does not depend on θ, so we can write

$$p(\theta \mid x) \propto p(x \mid \theta)p(\theta), \tag{14.64}$$

where the symbol "$\propto$" means "proportional to"[11].

The *maximum a posteriori estimate* consists in estimating θ by maximising the posterior distribution, $p(\theta \mid x)$, that is

$$\boxed{\hat{\theta}_{\text{map}}(x) \in \arg\max_{\theta\in\Theta} p(\theta \mid x) = \arg\max_{\theta\in\Theta} p(x \mid \theta)p(\theta).} \tag{14.65}$$

Example. Suppose that $x_1, \ldots, x_N$ are independent samples that follow the univariate normal distribution, $\mathcal{N}(\mu, \sigma^2)$, with some *known* variance σ^2 and unknown mean μ.

We will treat μ as a random variable. We assume that $\mu \sim \mathcal{N}(\mu_0, \sigma_\mu^2)$. Then,

$$\begin{aligned}
\hat{\mu}_{\text{map}} &\in \arg\max_\mu p(\mu \mid x_1, \ldots, x_N) = \arg\max_\mu p(x_1, \ldots, x_N \mid \mu)p(\mu) \\
&= \arg\max_\mu \underbrace{\frac{1}{\sqrt{2\pi\sigma_\mu^2}} \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_\mu^2}\right]}_{p(\mu)} \underbrace{\prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]}_{p(x_1,\ldots,x_N\mid\mu)} \\
&= \arg\max_\mu -\frac{(\mu-\mu_0)^2}{2\sigma_\mu^2} - \sum_{i=1}^N \frac{(x_i-\mu)^2}{2\sigma^2}.
\end{aligned}$$

After some algebraic manipulations we may find that

$$\hat{\mu}_{\text{map}} = \frac{\sigma^2\mu_0 + \sigma_\mu^2 \sum_{i=1}^N x_i}{N\sigma_\mu^2 + \sigma^2}.$$

[10] By Bayes rule: $p(\theta \mid x)p(x) = p(x \mid \theta)p(\theta)$

[11] This means that $p(\theta \mid x) = C(x)p(x \mid \theta)p(\theta)$, where the constant $C(x)$, which does not depend on θ, is $C(x) = p(x)^{-1}$

Another Example:[12] Let X be a real-valued continuous random variable with pdf

$$p_X(x) = 6x(1-x)1_{[0,1]}(x), \tag{14.66}$$

where $1_{[0,1]}(x) = 1$ for $x \in [0,1]$ and $1_{[0,1]}(x) = 0$ otherwise. Suppose that Y is another real-valued random variable and $Y \mid X = x \sim \text{Geometric}(x)$, i.e.,

$$p_{Y|X}(y \mid x) = x(1-x)^{y-1}. \tag{14.67}$$

The objectives are to (i) determine a MAP estimate of X given $Y = 2$, and (ii) determine the MAP estimate of X given $Y = y$ [the second one is left to you as an exercise].

The posterior distribution of X given $Y = y$ is

$$\begin{aligned} p_{X|Y}(x \mid y) &\propto p_{Y|X}(y \mid x)p_X(x) \\ &= x(1-x)^{y-1} \cdot 6x(1-x)1_{[0,1]}(x) \\ &\propto x^2(1-x)^y 1_{[0,1]}(x), \end{aligned} \tag{14.68}$$

and its logarithm is

$$\log p_{X|Y}(x \mid y) = 2\log(x) + y\log(1-x) + \text{const},$$

defined for $x \in [0,1]$[13]. The MAP estimate of X given $Y = 2$ is determined by

$$\hat{x}_{\text{MAP}}(2) \in \arg\max_{x \in [0,1]} 2\log(x) + 2\log(1-x) = \tfrac{1}{2}.$$

Estimation from noisy measurements: Suppose $X \sim \mathcal{N}(0, \sigma_X^2)$, $N \sim \mathcal{N}(0, \sigma_N^2)$ are independent and $Y = X + N$. We measure Y and need to estimate X.

Minimum variance estimation: We have shown that

$$\hat{x}_{\text{mve}}(y) := \mathbb{E}[X \mid Y = y] = \frac{\sigma_X^2 y}{\sigma_X^2 + \sigma_N^2}.$$

Maximum likelihood estimation. It is $(X \mid Y = y) = y - N \sim (y, \sigma_N^2)$ and

$$\begin{aligned} \hat{x}_{\text{mle}}(y) &\in \arg\max p_{X|Y}(x \mid y) \\ &= \arg\max \tfrac{1}{\sigma_N\sqrt{2\pi}} \exp\left(-\frac{(x-y)^2}{2\sigma_N^2}\right) = y. \end{aligned}$$

[12] Credit: This is based on Problem 1 in https://www.probabilitycourse.com/chapter9/9_2_0_ch_probs.php.

[13] Alternatively, we can write $\log p_{X|Y}(x \mid y) = 2\log(x) + y\log(1-x) + \text{const} - \delta_{[0,1]}(x)$, where $\delta_{(0,1)}$ is the *indicator function* of $(0,1)$, i.e., $\delta_{(0,1)}(x) = 0$ for $x \in (0,1)$ and $\delta_{(0,1)}(x) = \infty$ for $x \notin (0,1)$

Maximum a posteriori estimation. The MAP estimate is

$$\hat{x}_{\text{mape}}(y) \in \arg\max_x p_{Y|X}(y \mid x) p_X(x) = \arg\max_x -\left[\frac{(y-x)^2}{2\sigma_N^2} + \frac{x^2}{2\sigma_X^2}\right] = \frac{\sigma_X^2 y}{\sigma_X^2 + \sigma_N^2}.$$

Overall,

Method	Estimate, $\hat{x}(y)$	Notes
MVE	$\hat{x}_{\text{mve}}(y) = \frac{\sigma_X^2 y}{\sigma_X^2+\sigma_N^2}$	Unbiased, minimum variance
MLE	$\hat{x}_{\text{mle}}(y) = y$	Ignores distr of X (prior)
MAPE	$\hat{x}_{\text{map}}(y) = \frac{\sigma_X^2 y}{\sigma_X^2+\sigma_N^2}$	Same as MVE (in this case)

14.5.3 KF is a recursive MAP estimator

Consider the dynamical system

$$x_{t+1} = Ax_t + w_t, \tag{14.69a}$$
$$y_t = Cx_t + v_t, \tag{14.69b}$$

where $w_t \sim \mathcal{N}(0, Q)$ is time-independent, $v_t \sim \mathcal{N}(0, R)$ is time-independent, w_t is independent of v_t and $x_0 \sim \mathcal{N}(\bar{x}_0, P_0)$ is independent of w_0 and v_0[14].

$$p_{w_t}(w) \propto \exp\left[-\tfrac{1}{2}\|w\|^2_{Q^{-1}}\right], \tag{14.70a}$$
$$p_{v_t}(v) \propto \exp\left[-\tfrac{1}{2}\|v\|^2_{R^{-1}}\right], \tag{14.70b}$$
$$p_{x_0}(x_0) \propto \exp\left[-\tfrac{1}{2}\|x_0 - \bar{x}_0\|^2_{P_0^{-1}}\right]. \tag{14.70c}$$

Given a set of measurements $y_0, y_1, \ldots, y_{N-1}$ we need to estimate $x_0, x_1, \ldots, x_N$.

In Exercises we will state a number of important results that will be essential to develop the Bayesian interpretation of the Kalman filter. There are good candidates for the final exam.

Exercise 13 (☕☕). Let $N \in \mathbb{N}$. Use the properties of the conditional density function to show that

$$p(y_0, y_1, \ldots, y_N) = p(y_0) \prod_{t=1}^{N} p(y_t \mid y_{0:t-1}), \tag{14.71}$$

[14] Recall the notation $\|w\|^2_{Q^{-1}} := w^\intercal Q^{-1} w$.

where $y_{0:t} = (y_0, \dots, y_t)$. How can we determine $p(y_N)$ if we have $p(y_0, \dots, y_N)$?

Exercise 14 (☕☕). Use the properties of the conditional density function and the Markovianity of $(x_t)_t$ to show that for any $N \in \mathbb{N}$!

$$p(x_0, x_1, \dots, x_N) = p(x_0) \prod_{t=0}^{N-1} p_{w_t}(x_{t+1} - Ax_t). \tag{14.72}$$

How can we determine $p(x_N)$ if we have $p(x_0, \dots, x_N)$?

Exercise 15 (☕☕). Show that for all $N \in \mathbb{N}$!

$$p(x_0, \dots, x_N, w_0, \dots, w_{N-1}) = \begin{cases} 0, & \text{if not } x_{t+1} = Ax_t + w_t, \\ p_{x_0}(x_0) \prod_{t=0}^{N-1} p_{w_t}(w_t), & \text{otherwise} \end{cases}$$

Exercise 16 (☕☕). Use Bayes rule to show that !

$$p(x_0, \dots, x_N \mid y_0, \dots, y_{N-1}) \propto p_{x_0}(x_0) \prod_{t=0}^{N-1} p_{v_t}(y_t - Cx_t) p_{w_t}(x_{t+1} - Ax_t). \tag{14.73}$$

where $N \in \mathbb{N}$. How can we determine $p(x_0 \mid y_0, \dots, y_{N-1})$ in terms of $p_{x_0}(x_0)$, $p_{v_t}(y_t - Cx_t)$, and $p_{w_t}(x_{t+1} - Ax_t)$ for $t \in \mathbb{N}_{[0,N-1]}$?

From Exercise 16 the log-likelihood is (omitting the constant term)

$$\begin{aligned} &\log p(x_0, \dots, x_N \mid y_0, \dots, y_{N-1}) \\ &\quad = \log p_{x_0}(x_0) + \sum_{t=1}^{N-1} \log p_{v_t}(y_t - Cx_t) + \log p_{w_t}(x_{t+1} - Ax_t) \\ &\quad = -\tfrac{1}{2}\|x_0 - \bar{x}_0\|^2_{P_0^{-1}} + \sum_{t=0}^{N-1} -\|y_t - Cx_t\|^2_{R^{-1}} - \|x_{t+1} - Ax_t\|^2_{Q^{-1}}. \end{aligned} \tag{14.74}$$

Using the result of this exercise, we have the MAP estimate

$$\begin{aligned}(\hat{x}_t)_{t=0}^{N-1} &= \operatorname*{arg\,max}_{x_0,\dots,x_{N-1}} p(x_0,\dots,x_{N-1} \mid y_0,\dots,y_N) \\ &= \operatorname*{arg\,max}_{x_0,\dots,x_{N-1}} \log p(x_0,\dots,x_{N-1} \mid y_0,\dots,y_N) \\ &= \operatorname*{arg\,max}_{x_0,\dots,x_{N-1}} -\tfrac{1}{2}\|x_0-\bar{x}_0\|^2_{P_0^{-1}} + \sum_{t=0}^{N-1} -\|y_t - Cx_t\|^2_{R^{-1}} - \|x_{t+1}-Ax_t\|^2_{Q^{-1}} \\ &= \operatorname*{arg\,min}_{x_0,\dots,x_{N-1}} \tfrac{1}{2}\|x_0-\bar{x}_0\|^2_{P_0^{-1}} + \sum_{t=0}^{N-1} \|y_t - Cx_t\|^2_{R^{-1}} + \|x_{t+1}-Ax_t\|^2_{Q^{-1}}. \end{aligned} \tag{14.75}$$

In fact we can write it as

$$(\hat{x}_t)_{t=0}^{N-1} = \operatorname*{arg\,min}_{\substack{x_0,\dots,x_N,\\ w_0,\dots,w_{N-1},\\ v_0,\dots,v_{N-1},\\ x_{t+1}=Ax_t+w_t, t\in\mathbb{N}_{[0,N-1]}\\ y_t = Cx_t+v_t, t\in\mathbb{N}_{[0,N]}}} \tfrac{1}{2}\|x_0-\bar{x}_0\|^2_{P_0^{-1}} + \sum_{t=0}^{N-1} \tfrac{1}{2}\|v_t\|^2_{R^{-1}} + \tfrac{1}{2}\|w_t\|^2_{Q^{-1}}. \tag{14.76}$$

We need to solve the problem

$$\operatorname*{minimise}_{\substack{x_0,\dots,x_N,\\ w_0,\dots,w_{N-1},\\ v_0,\dots,v_{N-1}}} \tfrac{1}{2}\|x_0-\bar{x}_0\|^2_{P_0^{-1}} + \sum_{t=0}^{N-1} \tfrac{1}{2}\|v_t\|^2_{R^{-1}} + \tfrac{1}{2}\|w_t\|^2_{Q^{-1}}, \tag{14.77a}$$

$$\text{subject to: } x_{t+1} = Ax_t + w_t, t \in \mathbb{N}_{[0,N-1]}, \tag{14.77b}$$

$$y_t = Cx_t + v_t, t \in \mathbb{N}_{[0,N]}, \tag{14.77c}$$

which is akin to a linear-quadratic optimal control problem. Key difference: *arrival cost* instead of *terminal cost* and the initial condition is not fixed. The problem can be written as follows:

$$\operatorname*{minimise}_{\substack{x_0,\dots,x_N,\\ w_0,\dots,w_{N-1}}} \underbrace{\tfrac{1}{2}\|x_0-\bar{x}_0\|^2_{P_0^{-1}}}_{\ell_{x_0}(x_0)} + \sum_{t=0}^{N-1} \underbrace{\tfrac{1}{2}\|y_t - Cx_t\|^2_{R^{-1}} + \tfrac{1}{2}\|w_t\|^2_{Q^{-1}}}_{\ell_t(x_t,w_t)}, \tag{14.78a}$$

$$\text{subject to: } x_{t+1} = Ax_t + w_t, t \in \mathbb{N}_{[0,N-1]}. \tag{14.78b}$$

This is known as the *full information estimation* problem.

! **Exercise 17 (☕☕).** An alternative way to state the estimation problem is to use the result of Exercise 15 and consider the posterior distribution of $x_0,\dots,x_N,w_0,\dots,w_{N-1}$

given the measurements $y_0, \ldots, y_{N-1}$. Define $x_{0:N} = (x_0, \ldots, x_N)$ and, likewise, $w_{0:N-1} = (w_0, \ldots, w_{N-1})$. Show that

$$p(x_{0:N}, w_{0:N-1} \mid y_{0:N-1}) \propto p(x_{0:N}, w_{0:N-1}) p(y_{0:N-1} \mid x_{0:N}, w_{0:N-1}), \tag{14.79a}$$

where

$$p(y_{0:N-1} \mid x_{0:N}, w_{0:N-1}) = \prod_{t=0}^{N-1} p_{y_t|x_t,w_t}(y_t \mid x_t, w_t) = \prod_{t=0}^{N-1} p_{v_t}(y_t - Cx_t). \tag{14.79b}$$

Derive the full information state estimation problem (for the estimation of $x_{0:N}$ and $w_{0:N-1}$ given $y_{0:N-1}$) using Equation (14.79). Use the convention $\log 0 = -\infty$.

14.5.4 Forward DP Solution

The estimation problem becomes

$$\underset{x_0,\ldots,x_{N-1}}{\text{minimise}}\ \ell_{x_0}(x_0) + \sum_{t=0}^{N-1} \ell_t(x_t, w_t), \tag{14.80a}$$

$$\text{subject to: } x_{t+1} = Ax_t + w_t, t \in \mathbb{N}_{[0,N-1]}. \tag{14.80b}$$

We apply the DP procedure in a *forward* fashion:

$$\begin{aligned}
\widehat{V}_N^\star &= \min_{\substack{x_0,\ldots,x_N \\ w_0,\ldots,w_{N-1}}} \left\{ \ell_{x_0}(x_0) + \sum_{t=0}^{N-1} \ell_t(x_t, w_t) \;\middle|\; \begin{array}{l} x_{t+1} = Ax_t + w_t, \\ t \in \mathbb{N}_{[0,N-1]} \end{array} \right\} \\
&= \min_{\substack{x_1,\ldots,x_N \\ w_1,\ldots,w_{N-1}}} \left\{ \underbrace{\min_{x_0,w_0} \{\ell_{x_0}(x_0) + \ell_0(x_0, w_0) \mid x_1 = Ax_0 + w_0\}}_{V_1^\star(x_1)} \right. \\
&\qquad \left. + \sum_{t=1}^{N-1} \ell_t(x_t, w_t) \;\middle|\; \begin{array}{l} x_{t+1} = Ax_t + w_t, \\ t \in \mathbb{N}_{[1,N-1]} \end{array} \right\} \\
&= \min_{\substack{x_1,\ldots,x_N \\ w_1,\ldots,w_{N-1}}} \left\{ V_1^\star(x_1) + \sum_{t=1}^{N-1} \ell_t(x_t, w_t) \;\middle|\; \begin{array}{l} x_{t+1} = Ax_t + w_t, \\ t \in \mathbb{N}_{[1,N-1]} \end{array} \right\}.
\end{aligned} \tag{14.81}$$

The *forward* dynamic programming procedure can be written as

$$V_0^\star(x_0) = \ell_{x_0}(x_0), \tag{14.82a}$$

$$V_{t+1}^\star(x_{t+1}) = \min_{x_t, w_t} \left\{ V_t^\star(x_t) + \ell_t(x_t, w_t) \mid x_{t+1} = Ax_t + w_t \right\}, \tag{14.82b}$$

$$\widehat{V}_N^\star = \min_{x_N} V_N^\star(x_N). \tag{14.82c}$$

We shall prove that the solution of this problem yields the Kalman filter!

Remark. The MAP estimation approach can be also applied when (i) the involved random variables are not normally distributed, (ii) the dynamics is nonlinear, (iii) the system is constrained (e.g., we know that $x_t \in \mathcal{X}$), (iv) the disturbances are bounded (e.g., $w_t \in \mathcal{W}$, $v_t \in \mathcal{V}$).

14.5.5 The Kalman Filter as a Forward DP*

To show that Equations (14.82) yield the Kalman Filter, we need two lemmata. Firstly, Woodbury's matrix inversion lemma which we state without a proof (see also Handout 4).

Lemma 14.2 (Woodbury matrix identity) *The following holds*

$$(S + UCV)^{-1} = S^{-1} - S^{-1}U(C^{-1} + VS^{-1}U)^{-1}VS^{-1}. \tag{14.83}$$

Secondly, we state and prove the following lemma.

Lemma 14.3 (Factorisation of sum of quadratics) *Let* $Q_1 \in \mathbb{S}^n_{++}$, $Q_2 \in \mathbb{S}^m_{++}$, $F \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$. *Define*

$$q(x) = \tfrac{1}{2}\|x - \bar{x}\|^2_{Q_1^{-1}} + \tfrac{1}{2}\|Fx - b\|^2_{Q_2^{-1}}. \tag{14.84}$$

Then for all $x \in \mathbb{R}^n$,

$$q(x) = \tfrac{1}{2}\|x - x^\star\|^2_{W^{-1}} + c, \tag{14.85}$$

where

$$x^\star = \bar{x} + Q_1 F^\intercal S^{-1}(b - F\bar{x}), \tag{14.86a}$$
$$W = Q_1 - Q_1 F^\intercal S^{-1} F Q_1, \tag{14.86b}$$
$$c = \tfrac{1}{2}\|F\bar{x} - b\|^2_{S^{-1}}, \tag{14.86c}$$

and where $S = Q_2 + FQ_1F^\intercal$. *Moreover,*

$$x^\star = \arg\min_x q(x), \quad \textit{and} \quad \min_x q(x) = c. \tag{14.87}$$

Before we give the proof, recall the following very useful identity which holds for $Q \in \mathbb{S}^n_+$

$$\nabla\left(\tfrac{1}{2}\|Ax - b\|^2_Q\right) = A^\intercal Q(Ax - b). \tag{14.88}$$

Proof: Since q is a convex quadratic function, it has a minimiser, $x^\star$, which satisfies $\nabla q(x^\star) = 0$. The idea is that we can use Taylor's expansion to write q as follows

$$\begin{aligned} q(x) &= q(x^\star) + \nabla q(x^\star)^\intercal (x - x^\star) + \tfrac{1}{2}\|x - x^\star\|^2_{\nabla^2 q(x^\star)} \\ &= q(x^\star) + \tfrac{1}{2}\|x - x^\star\|^2_{\nabla^2 q(x^\star)}, \end{aligned} \tag{14.89}$$

It suffices to determine $x^\star$, $q(x^\star)$ and $\nabla^2 q(x^\star)$. Let us first determine $x^\star$: we have $\nabla q(x) = Q_1^{-1}(x - \bar{x}) + F^\intercal Q_2^{-1}(Fx - b)$, and we need to solve $\nabla q(x^\star) = 0$, that is

$$Q_1^{-1}(x^\star - \bar{x}) + F^\intercal Q_2^{-1}(Fx^\star - b) = 0, \tag{14.90}$$

from which

$$\begin{aligned} x^\star &= (Q_1^{-1} + F^\intercal Q_2^{-1} F)^{-1}(Q_1^{-1}\bar{x} + F^\intercal Q_2^{-1} b) \\ &= \bar{x} + (Q_1^{-1} + F^\intercal Q_2^{-1} F)^{-1}(Q_1^{-1}\bar{x} + F^\intercal Q_2^{-1} b - (Q_1^{-1} + F^\intercal Q_2^{-1} F)\bar{x}) \\ &= \bar{x} + (Q_1^{-1} + F^\intercal Q_2^{-1} F)^{-1} F^\intercal Q_2^{-1}(b - F\bar{x}) \end{aligned}$$

and now we apply Woodbury's matrix identity (Lemma 14.2) to the term $(Q_1^{-1} + F^\intercal Q_2^{-1} F)^{-1}$

$$\begin{aligned} &= \bar{x} + (Q_1 - Q_1 F^\intercal (Q_2 + FQ_1F^\intercal)^{-1} FQ_1) F^\intercal Q_2^{-1}(b - F\bar{x}) \\ &= \bar{x} + Q_1 F^\intercal \left[I - \underbrace{(Q_2 + FQ_1F^\intercal)}_{S}{}^{-1} \underbrace{FQ_1F^\intercal}_{S - Q_2} \right] Q_2^{-1}(b - F\bar{x}) \\ &= \bar{x} + Q_1 F^\intercal S^{-1}(b - F\bar{x}), \end{aligned} \tag{14.91}$$

(see Equation (14.86a)). Next, let us determine $q(x^\star)$

$$\begin{aligned}
q(x^\star) &= \tfrac{1}{2}\|x^\star - \bar{x}\|^2_{Q_1^{-1}} + \tfrac{1}{2}\|Fx^\star - b\|^2_{Q_2^{-1}} \\
&= \tfrac{1}{2}\left\|Q_1F^\intercal S^{-1}(b - F\bar{x})\right\|^2_{Q_1^{-1}} + \tfrac{1}{2}\left\|F\bar{x} - b + FQ_1F^\intercal S^{-1}(b - F\bar{x})\right\|^2_{Q_2^{-1}} \\
&= \tfrac{1}{2}\left\|Q_1F^\intercal S^{-1}(b - F\bar{x})\right\|^2_{Q_1^{-1}} + \tfrac{1}{2}\left\|\left(I - FQ_1F^\intercal S^{-1}\right)(b - F\bar{x})\right\|^2_{Q_2^{-1}} \\
&= \tfrac{1}{2}\left\|Q_1F^\intercal S^{-1}(b - F\bar{x})\right\|^2_{Q_1^{-1}} + \tfrac{1}{2}\left\|\left(I - (S - Q_2)S^{-1}\right)(b - F\bar{x})\right\|^2_{Q_2^{-1}} \\
&= \tfrac{1}{2}\left\|Q_1F^\intercal S^{-1}(b - F\bar{x})\right\|^2_{Q_1^{-1}} + \tfrac{1}{2}\left\|Q_2S^{-1}(b - F\bar{x})\right\|^2_{Q_2^{-1}} \\
&= \tfrac{1}{2}(b - F\bar{x})^\intercal S^{-1}FQ_1F^\intercal S^{-1}(b - F\bar{x})\tfrac{1}{2}(b - F\bar{x})^\intercal S^{-1}Q_2S^{-1}(b - F\bar{x}) \\
&= \tfrac{1}{2}(b - F\bar{x})^\intercal S^{-1}(\underbrace{FQ_1F^\intercal + Q_2}_{S})S^{-1}(b - F\bar{x}) \\
&= \tfrac{1}{2}(b - F\bar{x})^\intercal S^{-1}(b - F\bar{x}) = c. \qquad (14.92)
\end{aligned}$$

Lastly, it is easy to see that

$$\nabla^2 q(x) = Q_1^{-1} + F^\intercal Q_2^{-1}F = W^{-1}, \qquad (14.93)$$

where the second equality is due to Lemma 14.2. This completes the proof. ∎

We shall now apply Lemma 14.3 to the case of the Kalman filter.

Proposition 14.4 (MAPE ≡ KF) *Suppose that*

$$\ell_{x_0}(x_0) = \tfrac{1}{2}\|x_0 - \bar{x}_0\|^2_{P_0^{-1}}, \qquad (14.94a)$$

$$\ell_t(x_t, w_t) = \tfrac{1}{2}\|y_t - Cx_t\|^2_{R^{-1}} + \tfrac{1}{2}\|w_t\|^2_{Q^{-1}}. \qquad (14.94b)$$

Then, the procedure of Equation (14.82) yields the KF equations.

Indeed, the first step in the forward DP is

$$\begin{aligned}
V_1^\star(x_1) &= \min_{x_0, w_0}\left\{V_0^\star(x_0) + \ell_0(x_0, w_0) \mid x_1 = Ax_0 + w_0\right\} \\
&= \min_{x_0, w_0}\left\{\tfrac{1}{2}\|x_0 - \bar{x}_0\|^2_{P_0^{-1}} + \tfrac{1}{2}\|y_0 - Cx_0\|^2_{R^{-1}} + \tfrac{1}{2}\|w_0\|^2_{Q^{-1}} \mid x_1 = Ax_0 + w_0\right\} \\
&= \min_{x_0}\left\{\tfrac{1}{2}\|x_0 - \bar{x}_0\|^2_{P_0^{-1}} + \tfrac{1}{2}\|y_0 - Cx_0\|^2_{R^{-1}} + \tfrac{1}{2}\|x_1 - Ax_0\|^2_{Q^{-1}}\right\}. \qquad (14.95)
\end{aligned}$$

We can use Lemma 14.3 to write the sum of the first two terms as follows:

$$\tfrac{1}{2}\|x_0 - \bar{x}_0\|^2_{P_0^{-1}} + \tfrac{1}{2}\|y_0 - Cx_0\|^2_{R^{-1}} = \tfrac{1}{2}\|x_0 - x_0^\star\|^2_{W_0^{-1}} + \tfrac{1}{2}\|y_0 - Cx_0^\star\|^2_{S_0^{-1}}, \tag{14.96}$$

where

$$S_0 = R + CP_0C^\intercal, \tag{14.97a}$$
$$W_0 = P_0 - P_0C^\intercal S_0^{-1}CP_0, \tag{14.97b}$$
$$x_0^\star = \bar{x}_0 + P_0C^\intercal S_0^\intercal(y_0 - C\bar{x}_0). \tag{14.97c}$$

Note that $W_0 = \Sigma_{0|0}$ and $x_0^\star = \hat{x}_{0|0}$ (see Equations (14.13) and (14.14)). Next, we have

$$V_1^\star(x_1) = \underbrace{\tfrac{1}{2}\|y_0 - Cx_0^\star\|^2_{S_0^{-1}}}_{\text{constant}} + \min_{x_0}\left\{\tfrac{1}{2}\|x_0 - x_0^\star\|^2_{W_0^{-1}} + \tfrac{1}{2}\|x_1 - Ax_0\|^2_{Q^{-1}}\right\}. \tag{14.98}$$

From Equation (14.87) we have

$$V_1^\star(x_1) = \text{constant} + \tfrac{1}{2}\|x_1 - A\hat{x}_{0|0}\|^2_{\overline{S}_0^{-1}}, \tag{14.99}$$

where $\hat{x}_{1|0} = A\hat{x}_{0|0}$ and

$$\overline{S}_0 = Q + AW_0A^\intercal = \Sigma_{1|0}, \tag{14.100}$$

(compare this with Equation (14.16)). Note that $V_1^\star$ has the same form as $V_1^\star$, so the same procedure can be repeated. This will yield the same iterates as in Section 14.2.

14.6 References

1. B.D.O. Anderson and J.B. Moore, Optimal Filtering, Dover Books on Electrical Engineering, 2005

2. D. Simon, Optimal State Estimation: Kalman, H_∞, and Nonlinear Approaches, Wiley-Interscience, 2006