# 9

# Constrained Estimation

## 9.1 Overview

This chapter introduces the reader to the issues involved in constrained estimation. We adopt a stochastic framework and model the underlying system via a set of stochastic difference equations in which the noise has a known probability density function. This leads to a stochastic interpretation of the resulting estimators. Alternatively, one can interpret the resulting optimisation problems in a purely deterministic framework.

We begin with fixed horizon *constrained* linear estimation problems. We will see that the resulting optimisation problems are *similar* to the problems that arise in constrained control. Indeed, they only differ by virtue of the boundary conditions imposed. In the next chapter we will show that the connection is actually deeper than similarity. Indeed, we will show that, for the linear constrained case, the problems are formally dual to each other. We then consider rather general nonlinear estimation problems. Finally, the moving horizon implementation of these estimators is discussed and illustrated by examples.

Potential applications of the ideas presented here include any estimation problem where the variables are known, a priori, to satisfy various constraints. Examples are:

(i) State estimation problems in physical systems where constraints are known to apply, for example, in a distillation column where the liquid levels in the trays are known to lie between two levels (empty and full).
(ii) More general state estimation problems in process control where key variables (for example, disturbances) are known to lie in certain regions.
(iii) Channel equalisation problems in digital communication systems where the transmitted signal is known to belong to a finite alphabet (say $\pm 1$).
(iv) Estimation problems with general distributions where the distribution can be approximated in different regions by different Gaussian distributions.

## 9.2 Simple Linear Regression

To motivate the more general results to follow, let us first consider a simple linear regression problem:

$$\begin{aligned} x_{k+1} &= x_k = x_0 \quad \text{for } k = 0, \ldots, N-1, \\ y_k &= Cx_k + v_k \quad \text{for } k = 1, \ldots, N, \end{aligned} \tag{9.1}$$

where $x_k \in \mathbb{R}^n$ and where $\{y_k\}$ is a given sequence of scalar observations. Say that $\{v_k\}$ is an i.i.d. sequence having a distribution $p_v(v_k)$ obtained by truncating on the interval $[-b, b]$ a Gaussian distribution with zero mean and variance $\sigma^2$, that is,

$$p_v(v_k) = \begin{cases} \dfrac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{v_k^2}{2\sigma^2}\right\}}{\int_{-b}^{b} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\alpha^2}{2\sigma^2}\right\} d\alpha} & \text{if } |v_k| \le b, \\ 0 & \text{otherwise.} \end{cases} \tag{9.2}$$

Also, assume that $x_0$ has a Gaussian distribution: $N(\mu_0, P_0)$ with $P_0 > 0$.

In the sequel, we will need to refer frequently to *conditional probability density functions*. These take the general form of the probability density for a random variable $a$ evaluated at $\hat{a}$ (say), given that another random variable $b$ takes the specific value $\hat{b}$. We will express this density as $p_{a|b}(a = \hat{a}|b = \hat{b})$. Often we will simplify the notation to $p_{a|b}(\hat{a}|\hat{b})$.

Let $\mathbf{y}_N = \begin{bmatrix} y_1 & \ldots & y_N \end{bmatrix}^{\mathrm{T}}$ and let $\mathbf{y}_N^d = \begin{bmatrix} y_1^d & \ldots & y_N^d \end{bmatrix}^{\mathrm{T}}$ denote the given observations. Then, using Bayes' rule and the independence assumption, the joint probability density function for the data $\mathbf{y}_N^d$ and initial state estimate $\hat{x}_0$ can be obtained as follows:

$$\begin{aligned} p_{y_1,x_0}(y_1^d, \hat{x}_0) &= p_{y_1|x_0}(y_1^d|\hat{x}_0)\, p_{x_0}(\hat{x}_0), \\ p_{y_2,y_1,x_0}(y_2^d, y_1^d, \hat{x}_0) &= p_{y_2|y_1,x_0}(y_2^d|y_1^d, \hat{x}_0)\, p_{y_1,x_0}(y_1^d, \hat{x}_0) \\ &= p_{y_2|y_1,x_0}(y_2^d|y_1^d, \hat{x}_0)\, p_{y_1|x_0}(y_1^d|\hat{x}_0)\, p_{x_0}(\hat{x}_0) \\ &= p_{y_2|x_0}(y_2^d|\hat{x}_0)\, p_{y_1|x_0}(y_1^d|\hat{x}_0)\, p_{x_0}(\hat{x}_0), \\ &\vdots \\ p_{\mathbf{y}_N,x_0}(\mathbf{y}_N^d, \hat{x}_0) &= p_{x_0}(\hat{x}_0) \prod_{k=1}^{N} p_{y_k|x_0}(y_k^d|\hat{x}_0). \end{aligned} \tag{9.3}$$

Also note from (9.1) and (9.2) that

$$p_{y_k|x_0}(y_k^d|\hat{x}_0) = p_v(y_k^d - C\hat{x}_0)$$

$$= \begin{cases} \dfrac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y_k^d - C\hat{x}_0)^2}{2\sigma^2}\right\}}{\int_{-b}^{b} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\alpha^2}{2\sigma^2}\right\} d\alpha} & \text{if } |y_k^d - C\hat{x}_0| \le b, \\ 0 & \text{otherwise.} \end{cases}$$

Then, using the above in (9.3), we finally obtain

$$p_{\mathbf{y}_N,x_0}(\mathbf{y}_N^d,\hat{x}_0)=\begin{cases} f_1(\mathbf{y}_N^d,\hat{x}_0) & \text{if } |y_k^d - C\hat{x}_0| \le b, \quad k=1,\dots,N,\\ 0 & \text{otherwise,}\end{cases} \tag{9.4}$$

where

$$f_1(\mathbf{y}_N^d,\hat{x}_0) \triangleq \beta \exp\left\{-(\hat{x}_0-\mu_0)^{\mathrm{T}}\frac{P_0^{-1}}{2}(\hat{x}_0-\mu_0)\right\} \times \prod_{k=1}^{N}\frac{\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{\frac{-(y_k^d-C\hat{x}_0)^2}{2\sigma^2}\right\}}{\int_{-b}^{b}\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{\frac{-\alpha^2}{2\sigma^2}\right\}d\alpha}, \tag{9.5}$$

where $\beta \triangleq (2\pi)^{-\frac{n}{2}}(\det P_0)^{-\frac{1}{2}}$.

The estimation problem is as follows: Given $\mathbf{y}_N^d$, make some statement about the value of $x_0$. Based on $p_{\mathbf{y}_N,x_0}(\mathbf{y}_N^d,\hat{x}_0)$ we can express the *a posteriori* distribution of $x_0$ given $\mathbf{y}_N$ as follows:

$$p_{x_0|\mathbf{y}_N}(\hat{x}_0|\mathbf{y}_N^d)=\frac{p_{\mathbf{y}_N,x_0}(\mathbf{y}_N^d,\hat{x}_0)}{p_{\mathbf{y}_N}(\mathbf{y}_N^d)}, \tag{9.6}$$

where $p_{\mathbf{y}_N}(\mathbf{y}_N^d)$ is independent of $x_0$ and satisfies

$$p_{\mathbf{y}_N}(\mathbf{y}_N^d)=\int_{\mathbb{R}^n} p_{\mathbf{y}_N,x_0}(\mathbf{y}_N^d,\alpha)d\alpha. \tag{9.7}$$

The a posteriori distribution $p_{x_0|\mathbf{y}_N}(\hat{x}_0|\mathbf{y}_N^d)$ summarises "what we know about $x_0$ given the observations $\mathbf{y}_N^d$." If we require a specific estimate, then we can obtain this from $p_{x_0|\mathbf{y}_N}(\hat{x}_0|\mathbf{y}_N^d)$. Possible estimates are:

(i) Conditional mean

$$\hat{x}_0^{[1]} \triangleq \mathbf{E}\{x_0|\mathbf{y}_N^d\}=\int_{\mathbb{R}^n}\alpha\; p_{x_0|\mathbf{y}_N}(\alpha|\mathbf{y}_N^d)\; d\alpha. \tag{9.8}$$

(ii) A posteriori most probable

$$\hat{x}_0^{[2]} \triangleq \arg\max_{\hat{x}_0}\; p_{x_0|\mathbf{y}_N}(\hat{x}_0|\mathbf{y}_N^d)=\arg\max_{\hat{x}_0}\; p_{\mathbf{y}_N,x_0}(\mathbf{y}_N^d,\hat{x}_0). \tag{9.9}$$

Note that, in general, $\hat{x}_0^{[1]} \neq \hat{x}_0^{[2]}$. A simple two-state case is illustrated in Figure 9.1. In the *unconstrained* Gaussian case we have that the conditional mean coincides with the a posteriori most probable estimate (denoted $\hat{x}_0$ in the figure). However, in the presence of constraints, the a posteriori probability density is nonzero only in a restricted region illustrated by the shaded area[1] in

[1] For simplicity, all the truncated distributions are illustrated in this chapter without scaling.

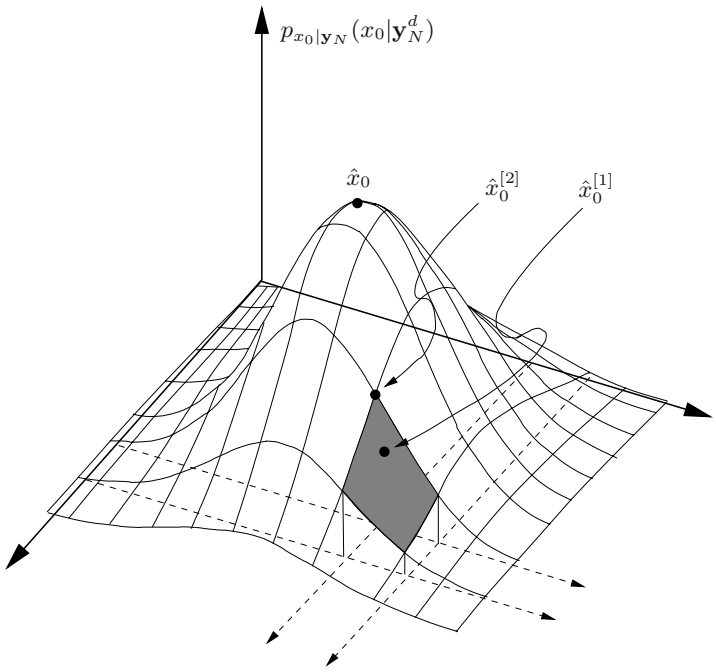


**Figure 9.1.** Illustration of the conditional mean and the a posteriori most probable estimate. (The points shown should actually be on the $x_0$-plane.)

Figure 9.1. In this case, we see that the conditional mean $\hat{x}_0^{[1]}$ will, in general, differ from the a posteriori most probable $\hat{x}_0^{[2]}$.

In the sequel, we will mainly focus on the a posteriori most probable estimate since this is found via a *constrained optimisation procedure* which is similar to the optimal control problems addressed earlier.

Returning to our special case of simple linear regression, we see from (9.9), that the a posteriori most probable estimate is obtained by maximising (9.4)–(9.5). In turn, this is equivalent to minimising $-\ln p_{\mathbf{y}_N,x_0}(\mathbf{y}_N^d,\hat{x}_0)$ where

$$-\ln p_{\mathbf{y}_N,x_0}(\mathbf{y}_N^d,\hat{x}_0)=\sum_{k=1}^{N}\frac{1}{2\sigma^2}\hat{v}_k^2+\frac{1}{2}(\hat{x}_0-\mu_0)^{\mathrm{T}}P_0^{-1}(\hat{x}_0-\mu_0)+\ \text{constant},$$

subject to the constraints

$$\begin{aligned}&\hat{v}_k=y_k^d-C\hat{x}_0\quad\text{for } k=1,\dots,N,\\&\hat{v}_k\in[-b,b]\quad\text{for } k=1,\dots,N.\end{aligned}$$

We recognise this as a standard constrained quadratic optimisation problem in the variable $\hat{x}_0$.

## 9.3 Linear State Estimation with Constraints

Here we generalise the ideas presented in Section 9.2 to the following linear Markov model:

$$\begin{aligned} x_{k+1} &= Ax_k + Bw_k, \\ y_k &= Cx_k + v_k, \end{aligned} \tag{9.10}$$

where $x_k \in \mathbb{R}^n$, $w_k \in \mathbb{R}^m$, $y_k \in \mathbb{R}^r$ and $v_k \in \mathbb{R}^r$. Suppose that $\{w_k\}$, $\{v_k\}$, $x_0$ are i.i.d. sequences having truncated Gaussian distributions, that is,

$$p_w(w_k) = \begin{cases} \dfrac{\beta_w \exp\left\{-\frac{1}{2} w_k^{\mathrm{T}} Q^{-1} w_k\right\}}{\beta_w \int_{\Omega_1} \exp\left\{-\frac{1}{2}\nu^{\mathrm{T}} Q^{-1}\nu\right\} d\nu} & \text{for } w_k \in \Omega_1, \\ 0 & \text{otherwise,} \end{cases} \tag{9.11}$$

$$p_v(v_k) = \begin{cases} \dfrac{\beta_v \exp\left\{-\frac{1}{2} v_k^{\mathrm{T}} R^{-1} v_k\right\}}{\beta_v \int_{\Omega_2} \exp\left\{-\frac{1}{2}\nu^{\mathrm{T}} R^{-1}\nu\right\} d\nu} & \text{for } v_k \in \Omega_2, \\ 0 & \text{otherwise,} \end{cases} \tag{9.12}$$

$$p_{x_0}(x_0) = \begin{cases} \dfrac{\beta_{x_0} \exp\left\{-\frac{1}{2}(x_0-\mu_0)^{\mathrm{T}} P_0^{-1}(x_0-\mu_0)\right\}}{\beta_{x_0} \int_{\Omega_3} \exp\left\{-\frac{1}{2}(\nu-\mu_0)^{\mathrm{T}} P_0^{-1}(\nu-\mu_0)\right\} d\nu} & \text{for } x_0 \in \Omega_3, \\ 0 & \text{otherwise,} \end{cases} \tag{9.13}$$

where $Q > 0$, $R > 0$, $P_0 > 0$, $\beta_w \triangleq (2\pi)^{-\frac{m}{2}} (\det Q)^{-\frac{1}{2}}$, $\beta_v \triangleq (2\pi)^{-\frac{r}{2}} (\det R)^{-\frac{1}{2}}$, $\beta_{x_0} \triangleq (2\pi)^{-\frac{n}{2}} (\det P_0)^{-\frac{1}{2}}$, $\Omega_1 \subset \mathbb{R}^m$, $\Omega_2 \subset \mathbb{R}^r$ and $\Omega_3 \subset \mathbb{R}^n$.

We define

$$\mathbf{y}_N = \left[y_1^{\mathrm{T}} \ldots y_N^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{9.14}$$

$$\mathbf{y}_N^d = \left[y_1^{d^{\mathrm{T}}} \ldots y_N^{d\ \mathrm{T}}\right]^{\mathrm{T}}, \tag{9.15}$$

$$\mathbf{x}_N = \left[x_0^{\mathrm{T}} \ldots x_N^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{9.16}$$

$$\hat{\mathbf{x}}_N = \left[\hat{x}_0^{\mathrm{T}} \ldots \hat{x}_N^{\mathrm{T}}\right]^{\mathrm{T}}. \tag{9.17}$$

From Bayes' rule and the Markovian structure of (9.10) we have that

$$\begin{aligned} p_{x_{k+1},\ldots,x_0}(\hat{x}_{k+1}, \hat{x}_k, \hat{x}_{k-1}, \ldots, \hat{x}_0) &= p_{x_{k+1}|x_k,\ldots,x_0}(\hat{x}_{k+1}|\hat{x}_k, \hat{x}_{k-1}, \ldots, \hat{x}_0) \\ &\quad \times p_{x_k,\ldots,x_0}(\hat{x}_k, \hat{x}_{k-1}, \ldots, \hat{x}_0) \\ &= p_{x_{k+1}|x_k}(\hat{x}_{k+1}|\hat{x}_k) \\ &\quad \times p_{x_k,\ldots,x_0}(\hat{x}_k, \hat{x}_{k-1}, \ldots, \hat{x}_0), \end{aligned}$$

and also

$$\begin{aligned} p_{y_k,x_k,\ldots,x_0}(\hat{y}_k^d,\hat{x}_k,\hat{x}_{k-1},\ldots,\hat{x}_0) &= p_{y_k|x_k,\ldots,x_0}(y_k^d|\hat{x}_k,\hat{x}_{k-1},\ldots,\hat{x}_0) \\ &\quad \times p_{x_k,\ldots,x_0}(\hat{x}_k,\hat{x}_{k-1},\ldots,\hat{x}_0) \\ &= p_{y_k|x_k}(y_k^d|\hat{x}_k) \\ &\quad \times p_{x_k,\ldots,x_0}(\hat{x}_k,\hat{x}_{k-1},\ldots,\hat{x}_0). \end{aligned}$$

It then follows that the joint probability density function for $\mathbf{y}_N$ and $\mathbf{x}_N$ defined in (9.14) and (9.16), respectively, is given by

$$\begin{aligned} p_{\mathbf{y}_N,\mathbf{x}_N}(\mathbf{y}_N = \mathbf{y}_N^d, \mathbf{x}_N = \hat{\mathbf{x}}_N) = p_{x_0}(x_0 = \hat{x}_0) \prod_{k=1}^{N} \Big[ & p_{y_k|x_k}(y_k = y_k^d | x_k = \hat{x}_k) \\ & \times \; p_{x_k|x_{k-1}}(x_k = \hat{x}_k | x_{k-1} = \hat{x}_{k-1}) \Big]. \end{aligned} \tag{9.18}$$

We next develop an explicit expression for the joint density function in (9.18). We begin with the nonsingular case when $w_k \in \mathbb{R}^n$ $(m = n)$ and $B$ is nonsingular in (9.10).

**Lemma 9.3.1** *For the model described in* (9.10) *to* (9.17)*, and subject to* $w_k \in \mathbb{R}^n$ $(m = n)$ *and* $B$ *nonsingular, the joint probability density function* (9.18) *for* $\mathbf{y}_N$ *and* $\mathbf{x}_N$ *satisfies*

$$\begin{aligned} p_{\mathbf{y}_N,\mathbf{x}_N}(\mathbf{y}_N = \mathbf{y}_N^d, \mathbf{x}_N = \hat{\mathbf{x}}_N) = \text{constant} \times & \exp\left\{ -\frac{1}{2} \sum_{k=0}^{N-1} \hat{w}_k^{\mathrm{T}} Q^{-1} \hat{w}_k \right\} \\ & \times \exp\left\{ -\frac{1}{2} \sum_{k=1}^{N} \hat{v}_k^{\mathrm{T}} R^{-1} \hat{v}_k \right\} \\ & \times \exp\left\{ -\frac{1}{2} (\hat{x}_0 - \mu_0)^{\mathrm{T}} P_0^{-1} (\hat{x}_0 - \mu_0) \right\}, \end{aligned} \tag{9.19}$$

*whenever*

$$\begin{aligned} \hat{w}_k &\in \Omega_1 \quad \textit{for } k = 0,\ldots,N-1, \\ \hat{v}_k &\in \Omega_2 \quad \textit{for } k = 1,\ldots,N, \\ \hat{x}_0 &\in \Omega_3, \end{aligned}$$

*where*

$$\begin{aligned} \hat{x}_{k+1} &= A\hat{x}_k + B\hat{w}_k \quad \textit{for } k = 0,\ldots,N-1, \\ \hat{v}_k &= y_k^d - C\hat{x}_k \quad \textit{for } k = 1,\ldots,N. \end{aligned}$$

*Proof.* From (9.10), (9.11) and (9.12), we have, using the rule of transformation of probability density functions:

$$\begin{aligned} p_{x_{k+1}|x_k}(x_{k+1} = \hat{x}_{k+1}|x_k = \hat{x}_k) &= \text{constant} \times p_w(\hat{w}_k) \\ &= \text{constant} \times \exp\left\{-\frac{1}{2}\hat{w}_k^{\mathrm{T}} Q^{-1} \hat{w}_k\right\}, \end{aligned}$$

whenever $\hat{w}_k \in \Omega_1$ and satisfies $\hat{x}_{k+1} = A\hat{x}_k + B\hat{w}_k$. Also,

$$\begin{aligned} p_{y_k|x_k}(y_k = y_k^d|x_k = \hat{x}_k) &= \text{constant} \times p_v(\hat{v}_k) \\ &= \text{constant} \times \exp\left\{-\frac{1}{2}\hat{v}_k^{\mathrm{T}} R^{-1} \hat{v}_k\right\}, \end{aligned}$$

whenever $\hat{v}_k \in \Omega_2$ and satisfies $y_k^d = C\hat{x}_k + \hat{v}_k$. Finally, using (9.13), and substituting all expressions into (9.18), the result follows. □

**Remark 9.3.1.** In the general case, when $w_k \in \mathbb{R}^m$ with $m < n$ in (9.10), the linear equality $\hat{x}_{k+1} - A\hat{x}_k = B\hat{w}_k$, implies that $\hat{x}_{k+1} - A\hat{x}_k$ can only take values in the range space of $B$. Hence, we need to account for the fact that $\hat{x}_{k+1} - A\hat{x}_k$ has a singular distribution[2] in $\mathbb{R}^n$. We can easily deal with this situation by introducing a linear transformation in the state space as follows.

Assume that $B$ has full column rank. Let $T_1$ be a basis for the range space of $B$ (which, in particular, could be chosen equal to $B$) and choose any $T_2$ such that $T = [T_1 \;\; T_2]$ is nonsingular. We partition $T^{-1}$ as follows:

$$T^{-1} = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix},$$

where $S_1$ is an $m \times n$ matrix. Then $T^{-1}T = I_n$ implies

$$S_1 T_1 = I_m, \quad S_2 T_1 = 0_{(n-m)\times m}.$$

Hence, since $B = T_1 \bar{B}_1$ for some nonsingular $m \times m$ matrix $\bar{B}_1$, we have, using the above equations, that

$$T^{-1}B = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} T_1 \bar{B}_1 = \begin{bmatrix} \bar{B}_1 \\ 0 \end{bmatrix}. \tag{9.20}$$

Partition $\bar{x}_{k+1}$ as

$$\bar{x}_{k+1} \triangleq T^{-1} x_{k+1}. \tag{9.21}$$

Then, from (9.10), $\bar{x}_{k+1}$ satisfies

$$\bar{x}_{k+1} = \bar{A}x_k + \bar{B}w_k, \tag{9.22}$$

[2] A singular distribution is a distribution in $\mathbb{R}^n$ which is concentrated in a lower dimensional subspace, that is, the probability associated with any set not intersecting the subspace is zero (Anderson 1958).

where

$$\bar{A} \triangleq T^{-1}A \triangleq \begin{bmatrix} \bar{A}_1 \\ \bar{A}_2 \end{bmatrix}, \qquad \bar{B} \triangleq T^{-1}B = \begin{bmatrix} \bar{B}_1 \\ 0 \end{bmatrix}, \tag{9.23}$$

using (9.20). Let

$$\bar{x}_{k+1} \triangleq \begin{bmatrix} \bar{x}'_{k+1} \\ \bar{x}''_{k+1} \end{bmatrix},$$

where $\bar{x}'_{k+1} \in \mathbb{R}^m$. Then, from (9.22)–(9.23), we can write

$$\begin{bmatrix} \bar{x}'_{k+1} \\ \bar{x}''_{k+1} \end{bmatrix} = \begin{bmatrix} \bar{A}_1 \\ \bar{A}_2 \end{bmatrix} x_k + \begin{bmatrix} \bar{B}_1 \\ 0 \end{bmatrix} \; w_k. \tag{9.24}$$

Hence, using the rule of transformation of probability density functions, we have, from (9.21) and (9.24), that

$$\begin{aligned} p_{x_{k+1}|x_k}(\hat{x}_{k+1}|\hat{x}_k) &= \text{constant} \times p_{\bar{x}_{k+1}|x_k}(\hat{\bar{x}}_{k+1}|\hat{x}_k) \\ &= \text{constant} \times p_{\bar{x}'_{k+1}|x_k}(\hat{\bar{x}}'_{k+1}|\hat{x}_k) \times \delta_{n-m}[\hat{\bar{x}}''_{k+1} - \bar{A}_2\hat{x}_k] \\ &= \text{constant} \times p_w(\hat{w}_k) \times \delta_{n-m}[\hat{\bar{x}}''_{k+1} - \bar{A}_2\hat{x}_k], \end{aligned}$$

whenever $\hat{w}_k \in \Omega_1$ and satisfies $\hat{\bar{x}}'_{k+1} = \bar{A}_1\hat{x}_k + \bar{B}_1\hat{w}_k$. In the above equations, $\delta_{n-m}[\,\cdot\,]$ is the Dirac delta function defined on $\mathbb{R}^{n-m}$, that is, $\delta_{n-m}[\eta] = \delta(\eta_1) \times \cdots \times \delta(\eta_{n-m})$, where $\eta = \begin{bmatrix} \eta_1 & \cdots & \eta_{n-m} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{n-m}$.

We can thus write

$$p_{x_{k+1}|x_k}(\hat{x}_{k+1}|\hat{x}_k) = \text{constant} \times p_w(\hat{w}_k) \times \delta_{n-m}[\hat{\bar{x}}''_{k+1} - \bar{A}_2\hat{x}_k], \tag{9.25}$$

where $\hat{x}_{k+1}$ is restricted to those values reachable from $\hat{w}_k$, that is, such that $\hat{x}_{k+1} = A\hat{x}_k + B\hat{w}_k$ for some $\hat{w}_k \in \Omega_1$. We thus see that $p_{x_{k+1}|x_k}(\,\cdot \mid \cdot\,)$ has a density function in $\mathbb{R}^n$ corresponding to those values of $x_{k+1}$ that are reachable from $w_k$.

When defining the joint a posteriori most probable [JAPMP] estimator below, we will maximise the envelope of the delta function in (9.25). For notational convenience, we define this envelope as

$$p'_{x_{k+1}|x_k}(\hat{x}_{k+1}|\hat{x}_k) \triangleq \text{constant} \times p_w(\hat{w}_k)$$

whenever $\hat{w}_k \in \Omega_1$ and satisfies $\hat{x}_{k+1} = A\hat{x}_k + B\hat{w}_k$. Hence, in the sequel, probability densities $p$ corresponding to singular distributions should be interpreted as the envelope $p'$ defined above. ∘

The general estimation problem is: Given the observations $\mathbf{y}_N^d = [{y_1^d}^{\mathrm{T}} \ldots {y_N^d}^{\mathrm{T}}]^{\mathrm{T}}$, make some statement about the states $\mathbf{x}_N = [x_0^{\mathrm{T}} \ldots x_N^{\mathrm{T}}]^{\mathrm{T}}$. From the joint probability density function (9.19), we can express the *a posteriori distribution* of $\mathbf{x}_N$ given $\mathbf{y}_N$ as follows:

$$p_{\mathbf{x}_N|\mathbf{y}_N}(\hat{\mathbf{x}}_N|\mathbf{y}_N^d) = \frac{p_{\mathbf{y}_N,\mathbf{x}_N}(\mathbf{y}_N^d, \hat{\mathbf{x}}_N)}{p_{\mathbf{y}_N}(\mathbf{y}_N^d)}, \tag{9.26}$$

where $p_{\mathbf{y}_N}(\mathbf{y}_N^d)$ is a data dependent term which does not depend on $\mathbf{x}_N$.

The a posteriori distribution $p_{\mathbf{x}_N|\mathbf{y}_N}(\hat{\mathbf{x}}_N|\mathbf{y}_N^d)$ summarises "what we know about $\mathbf{x}_N$ given the observations $\mathbf{y}_N^d$." As foreshadowed in Remark 9.3.1, our aim is to find the *joint a posteriori most probable* [JAPMP] state estimates $\hat{\mathbf{x}}_N = [\hat{x}_0^{\mathrm{T}} \dots \hat{x}_N^{\mathrm{T}}]^{\mathrm{T}}$ given the observations $\hat{\mathbf{y}}_N^d$; that is,

$$\hat{\mathbf{x}}_N^* \triangleq \arg\max_{\hat{\mathbf{x}}_N} \; p_{\mathbf{x}_N|\mathbf{y}_N}(\hat{\mathbf{x}}_N|\mathbf{y}_N^d). \tag{9.27}$$

Note that (9.27) is equivalent to maximising the joint probability density function, since, as noticed in (9.26), both functions are related by a term that does not depend on $\mathbf{x}_N$. Thus, the joint maximum a posteriori estimate is given by

$$\begin{aligned}\hat{\mathbf{x}}_N^* &\triangleq \arg\max_{\hat{\mathbf{x}}_N} \; p_{\mathbf{x}_N|\mathbf{y}_N}(\hat{\mathbf{x}}_N|\mathbf{y}_N^d) \\ &= \arg\max_{\hat{\mathbf{x}}_N} \; p_{\mathbf{y}_N,\mathbf{x}_N}(\mathbf{y}_N^d, \hat{\mathbf{x}}_N) \\ &= \arg\min_{\hat{\mathbf{x}}_N} \; -\ln p_{\mathbf{y}_N,\mathbf{x}_N}(\mathbf{y}_N^d, \hat{\mathbf{x}}_N).\end{aligned} \tag{9.28}$$

The preceding discussion leads, upon substitution of (9.19) into (9.28), to the following optimisation problem.

**Estimation Problem**

Given the observations $\{y_1^d, \dots, y_N^d\}$ and the knowledge of $\mu_0$ (the mean value of $x_0$), solve:

$$\mathcal{P}_e: \qquad V_N^{\mathrm{OPT}}(\mu_0, \{y_k^d\}) \triangleq \min V_N(\{\hat{x}_k\}, \{\hat{v}_k\}, \{\hat{w}_k\}), \tag{9.29}$$

subject to:

$$\hat{x}_{k+1} = A\hat{x}_k + B\hat{w}_k \quad \text{for } k = 0, \dots, N-1, \tag{9.30}$$

$$\hat{v}_k = y_k^d - C\hat{x}_k \quad \text{for } k = 1, \dots, N, \tag{9.31}$$

$$\hat{w}_k \in \Omega_1 \quad \text{for } k = 0, \dots, N-1, \tag{9.32}$$

$$\hat{v}_k \in \Omega_2 \quad \text{for } k = 1, \dots, N, \tag{9.33}$$

$$\hat{x}_0 \in \Omega_3, \tag{9.34}$$

where

$$\begin{aligned}V_N(\{\hat{x}_k\}, \{\hat{v}_k\}, \{\hat{w}_k\}) \triangleq \frac{1}{2}(\hat{x}_0 - \mu_0)^{\mathrm{T}} P_0^{-1}(\hat{x}_0 - \mu_0) \\ + \frac{1}{2}\sum_{k=0}^{N-1} \hat{w}_k^{\mathrm{T}} Q^{-1} \hat{w}_k + \frac{1}{2}\sum_{k=1}^{N} \hat{v}_k^{\mathrm{T}} R^{-1} \hat{v}_k.\end{aligned} \tag{9.35}$$

We see that the above problem is very *similar* to the constrained linear quadratic optimal control problems discussed earlier (see, for example, (5.49)–(5.54) in Chapter 5) save that they have different boundary conditions and initial and terminal state weightings. The two problems are compared in Table 9.1.

| | Constrained control | Constrained estimation |
|---|---|---|
| Model | $x_{k+1} = Ax_k + Bu_k$ | $\hat{x}_{k+1} = A\hat{x}_k + B\hat{w}_k$ |
| Initial condition | $x_0$ (given) | $\hat{x}_0 \in \Omega_3$ |
| Initial state weighting | $\frac{1}{2}x_0^{\mathrm{T}}Qx_0$ (given) | $\frac{1}{2}(\hat{x}_0 - \mu_0)^{\mathrm{T}}P_0^{-1}(\hat{x}_0 - \mu_0)$, $\mu_0$ given |
| Terminal state weighting | $\frac{1}{2}x_N^{\mathrm{T}}Px_N$ | $\frac{1}{2}(y_N^d - C\hat{x}_N)^{\mathrm{T}}R^{-1}(y_N^d - C\hat{x}_N)$, $y_N^d$ given |

**Table 9.1.** Comparison between the optimisation problems corresponding to constrained control and constrained estimation.

## 9.4 Extensions to Other Constraints and Distributions

The development in Section 9.3 was based on an assumption of truncated Gaussian noise. This result is interesting in its raw form but becomes a powerful tool when utilised as a basic building block to solve more general problems. Several alternatives are discussed below indicating how the core ideas of Section 9.3 can be used in more general problems.

### 9.4.1 Nonzero-mean Truncated Gaussian Noise

It is very straightforward to add a nonzero mean assumption to the truncated Gaussian noise assumption. The appropriate changes to (9.11) and (9.12) are

$$p_w(w_k) = \begin{cases} \dfrac{\beta_w \exp\left\{-\frac{1}{2}(w_k - \mu_w)^{\mathrm{T}}Q^{-1}(w_k - \mu_w)\right\}}{\beta_w \int_{\Omega_1} \exp\left\{-\frac{1}{2}(\nu - \mu_w)^{\mathrm{T}}Q^{-1}(\nu - \mu_w)\right\} d\nu} & \text{for } w_k \in \Omega_1, \\ 0 & \text{otherwise,} \end{cases}$$

$$p_v(v_k) = \begin{cases} \dfrac{\beta_v \exp\left\{-\frac{1}{2}(v_k - \mu_v)^{\mathrm{T}}R^{-1}(v_k - \mu_v)\right\}}{\beta_v \int_{\Omega_2} \exp\left\{-\frac{1}{2}(\nu - \mu_v)^{\mathrm{T}}R^{-1}(\nu - \mu_v)\right\} d\nu} & \text{for } v_k \in \Omega_2, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mu_w$ and $\mu_v$ are the "prior" means, that is, the means of the Gaussian distributions before truncation.

The corresponding change in the objective function (9.35) is

$$\begin{aligned} V_N(\{\hat{x}_k\},\{\hat{v}_k\},\{\hat{w}_k\}) \triangleq \frac{1}{2}(\hat{x}_0-\mu_0)^{\mathrm{T}}P_0^{-1}(\hat{x}_0-\mu_0) \\ + \frac{1}{2}\sum_{k=0}^{N-1}(\hat{w}_k-\mu_w)^{\mathrm{T}}Q^{-1}(\hat{w}_k-\mu_w) \\ + \frac{1}{2}\sum_{k=1}^{N}(\hat{v}_k-\mu_v)^{\mathrm{T}}R^{-1}(\hat{v}_k-\mu_v). \end{aligned}$$

The use of a nonzero mean for the underlying distribution allows one, for example, to build new zero-mean distributions such as the one illustrated in Figure 9.2.

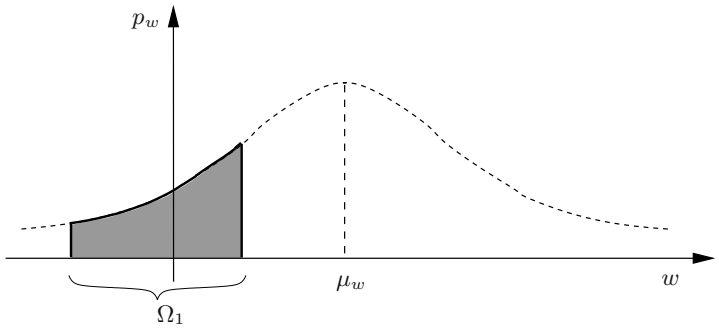


**Figure 9.2.** Zero-mean distribution formed by truncating a nonzero-mean Gaussian distribution.

### 9.4.2 Combinations of Truncated Gaussian Noise

A further embellishment is to have different truncated Gaussian distributions in different regions. For example, we could have

$$p_w(w_k) = \frac{\beta_{w_i}\exp\left\{-\frac{1}{2}(w_k-\mu_i)^{\mathrm{T}}Q_i^{-1}(w_k-\mu_i)\right\}}{\sum_{i=1}^{L}\beta_{w_i}\int_{\Omega_i}\exp\left\{-\frac{1}{2}(\nu-\mu_i)^{\mathrm{T}}Q_i^{-1}(\nu-\mu_i)\right\}d\nu},$$

for $w_k \in \Omega_i$, $i = 1,\ldots,L$, and zero otherwise, where $\Omega_i \subset \mathbb{R}^m$ are convex sets that have an empty intersection pairwise. A simple example is shown in Figure 9.3.

The associated optimisation problem can be solved by partitioning the problem into constrained sub-problems, each of which is convex in a convex

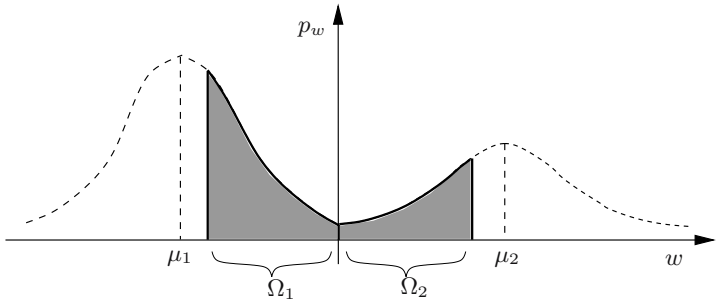


**Figure 9.3.** Combination of two nonzero-mean truncated Gaussian distributions.

region. One then simply chooses the global optimum as the minimum of the individual sub-problems. This idea was described in general terms in Section 2.7 of Chapter 2.

Thus, say that we have a scalar disturbance $\{w_k\}$ and an $N$-step optimisation horizon. Also, say that the distribution of $\{w_k\}$ is divided into $L$ nonoverlapping regions, each containing a different truncated Gaussian distribution. Then one needs to solve $L^N$ separate QP problems. As an illustration, with $L = 2$ (as in Figure 9.3) and $N = 5$, then one needs to solve $2^5 = 32$ QP problems.

**Remark 9.4.1.** Actually, the above idea is an interesting precursor to ideas that will be presented in Chapter 13 when we treat finite alphabet estimation problems. The latter case can be thought of as the limiting version of the idea presented above in which each region contains a *point mass distribution*. In this case, the optimisation problem requires $L^N$ objective function evaluations rather than $L^N$ QP problems. ∘

### 9.4.3 Multiconvex Approximations of Arbitrary Distributions

A further generalisation of these ideas is to use a staircase approximation to an arbitrary distribution. Thus, consider the smooth, but otherwise arbitrary, distribution in Figure 9.4, together with a staircase approximation.

In each region, the probability density function is approximated by a uniform distribution, that is,

$$p_w(w_k) \approx c_i \quad \text{for} \quad w_k \in \Omega_i,$$

where $c_i > 0$ is a constant. In this case, we have

$$\ln p_w(w_k) \approx \ln c_i \quad \text{for} \quad w_k \in \Omega_i.$$

The objective function (9.35) splits into $L^N$ (where $L$ is the number of regions $\Omega_i$ in the staircase approximation) convex functions $V_N^i$, each of them

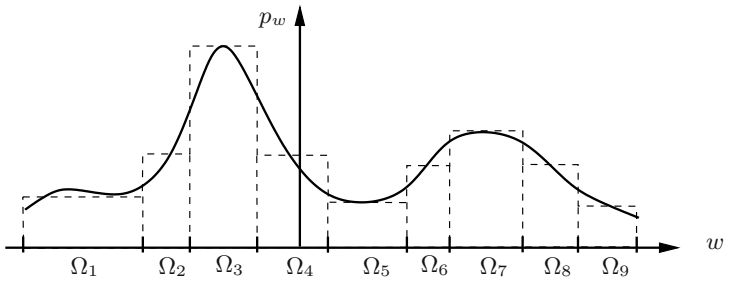


**Figure 9.4.** Arbitrary distribution and staircase approximation.

having the form:[3]

$$V_N^i(\{\hat{x}_k\},\{\hat{v}_k\},\{\hat{w}_k\}) \triangleq \frac{1}{2}(\hat{x}_0-\mu_0)^{\mathrm{T}}P_0^{-1}(\hat{x}_0-\mu_0) - \sum_{k=0}^{N-1}\ln\ \ell_k + \frac{1}{2}\sum_{k=1}^{N}\hat{v}_k^{\mathrm{T}}R^{-1}\hat{v}_k,$$

where $\ell_k \in \{c_1,\ldots,c_L\}$, $k=0,\ldots,N-1$.

The global solution is computed as the minimum of the $L^N$ convex optimisation sub-problems. (Note that the term $-\sum_{k=0}^{N-1}\ln\ \ell_k$ is constant for each sub-problem and, hence, it does not affect each minimiser. However, these terms must be included in the evaluation of each sub-problem when computing the global optimum.)

### 9.4.4 Discussion

We have seen above that one can treat very general estimation problems by combining convex optimisation with constraints. Note that the juxtaposition of *constraints* and regional *convexity* is the key idea to solving these problems.

## 9.5 Dynamic Programming

As for constrained control problems, we can utilise dynamic programming to solve the constrained estimation problem. Here it is most convenient to use *forward dynamic programming* whereas previously we used *reverse dynamic programming* (see Section 3.4 in Chapter 3).

We return to the problem of constrained estimation described in (9.29)–(9.35). We note that the objective function from time 0 to $k$ (that is, (9.35) for $N=k$) is a function of the initial state estimate $\hat{x}_0$, the choice of the input

[3] Notice that we assume that $v_k$ has a Gaussian distribution, but the idea is readily extended to arbitrary distributions for $v_k$, also.

noise sequence $\hat{w}_0, \ldots, \hat{w}_{k-1}$, and the given data $\mu_0, y_1^d, \ldots y_k^d$. If, for given $\hat{x}_0$, we optimise with respect to $\hat{w}_0, \ldots, \hat{w}_{k-1}$, then the resulting partial value function (at time $k$) is a function of $\hat{x}_0$ and $\mu_0, y_1^d, \ldots y_k^d$. For the purposes of the dynamic programming argument it is actually more convenient to make the partial value function a function of $\hat{x}_k$ and $\mu_0, y_1^d, \ldots y_k^d$. This is possible since (9.30) allows us to express $\hat{x}_0$ as a function of $\hat{x}_k$ (together with the given sequence $\hat{w}_0, \ldots, \hat{w}_{k-1}$) provided $A$ is nonsingular. Thus, assuming that $A$ is nonsingular, the partial value function at time $k$ is

$$V_k^{\mathrm{OPT}}(\hat{x}_k, \mu_0, y_1^d, \ldots, y_k^d) \triangleq \min_{\hat{w}_0,\ldots,\hat{w}_{k-1}} \Big\{ \frac{1}{2}(\hat{x}_0 - \mu_0)^{\mathrm{T}} P_0^{-1}(\hat{x}_0 - \mu_0) + \frac{1}{2}\sum_{j=0}^{k-1} \hat{w}_j^{\mathrm{T}} Q^{-1} \hat{w}_j + \frac{1}{2}\sum_{j=1}^{k} (y_j^d - C\hat{x}_j)^{\mathrm{T}} R^{-1}(y_j^d - C\hat{x}_j) \Big\},$$

subject to:

$$\hat{x}_j = A^{-1}(\hat{x}_{j+1} - B\hat{w}_j) \quad \text{for } j = 0, \ldots, k-1, \tag{9.36}$$

$$\hat{w}_j \in \Omega_1 \quad \text{for } j = 0, \ldots, k-1, \tag{9.37}$$

$$y_j^d - C\hat{x}_j \in \Omega_2 \quad \text{for } j = 1, \ldots, k, \tag{9.38}$$

$$\hat{x}_0 \in \Omega_3. \tag{9.39}$$

Then, the forward dynamic programming algorithm proceeds as follows. We start with the partial value function at time 0, which, for $\hat{x}_0 \in \Omega_3$, is defined as

$$V_0^{\mathrm{OPT}}(\hat{x}_0, \mu_0) \triangleq \frac{1}{2}(\hat{x}_0 - \mu_0)^{\mathrm{T}} P_0^{-1}(\hat{x}_0 - \mu_0). \tag{9.40}$$

Next, for $\hat{x}_1 \in \mathbb{R}^n$ such that $y_1^d - C\hat{x}_1 \in \Omega_2$, the partial value function at time 1 is computed as

$$V_1^{\mathrm{OPT}}(\hat{x}_1, \mu_0, y_1^d) = \min_{\hat{w}_0} \Big\{ V_0^{\mathrm{OPT}}(A^{-1}\hat{x}_1 - A^{-1}B\hat{w}_0, \mu_0) + \frac{1}{2}\hat{w}_0^{\mathrm{T}} Q^{-1}\hat{w}_0 + \frac{1}{2}(y_1^d - C\hat{x}_1)^{\mathrm{T}} R^{-1}(y_1^d - C\hat{x}_1) \Big\}, \tag{9.41}$$

subject to:

$$\hat{w}_0 \in \Omega_1, \tag{9.42}$$

$$A^{-1}\hat{x}_1 - A^{-1}B\hat{w}_0 \in \Omega_3. \tag{9.43}$$

Finally, for $k \geq 1$, and $\hat{x}_{k+1} \in \mathbb{R}^n$ such that $y_{k+1}^d - C\hat{x}_{k+1} \in \Omega_2$,

$$V_{k+1}^{\mathrm{OPT}}(\hat{x}_{k+1}, \mu_0, y_1^d, \ldots, y_{k+1}^d)$$
$$= \min_{\hat{w}_k} \Big\{ V_k^{\mathrm{OPT}}(A^{-1}\hat{x}_{k+1} - A^{-1}B\hat{w}_k, \mu_0, y_1^d, \ldots, y_k^d) + \frac{1}{2}\hat{w}_k^{\mathrm{T}} Q^{-1}\hat{w}_k$$
$$+ \frac{1}{2}(y_{k+1}^d - C\hat{x}_{k+1})^{\mathrm{T}} R^{-1}(y_{k+1}^d - C\hat{x}_{k+1}) \Big\}, \tag{9.44}$$

subject to:

$$\hat{w}_k \in \Omega_1, \tag{9.45}$$
$$y_k^d - C(A^{-1}\hat{x}_{k+1} - A^{-1}B\hat{w}_k) \in \Omega_2. \tag{9.46}$$

In the absence of constraints, the above dynamic programming algorithm leads to the well-known Kalman filter. This is explained in the next section.

## 9.6 Linear Gaussian Unconstrained Problems

For the case of linear Gaussian unconstrained problems, the dynamic programming algorithm of Section 9.5 can be solved explicitly. As expected, the optimal estimator in this case is the Kalman filter, as we show in the following results.

**Lemma 9.6.1 ( Dynamic Programming for Linear Gaussian Estimation)** *Assume that $A$ is nonsingular.*[4] *In the absence of constraints (that is, $\Omega_1 = \mathbb{R}^m$ in (9.37), $\Omega_2 = \mathbb{R}^r$ in (9.38) and $\Omega_3 = \mathbb{R}^n$ in (9.39)), the dynamic programming problem specified in (9.40)–(9.46) has the solution*

$$V_k^{\mathrm{OPT}}(\hat{x}_k, \mu_0, y_1^d, \ldots, y_k^d) = \frac{1}{2}(\hat{x}_k - \hat{x}_{k|k})^{\mathrm{T}} P_{k|k}^{-1}(\hat{x}_k - \hat{x}_{k|k}) + \text{ constant}, \tag{9.47}$$

*where $\hat{x}_{k|k}$ is a function of $\mu_0, y_1^d, \ldots, y_k^d$ defined via the following recursion:*

$$\hat{x}_{0|0} = \mu_0, \tag{9.48}$$
$$P_{0|0} = P_0, \tag{9.49}$$

*and, for $j = 0, \ldots, k-1$,*

$$\hat{x}_{j+1|j} = A\hat{x}_{j|j}, \tag{9.50}$$
$$\hat{x}_{j+1|j+1} = \hat{x}_{j+1|j} + P_{j+1|j}C^{\mathrm{T}}(R + CP_{j+1|j}C^{\mathrm{T}})^{-1}(y_{j+1}^d - C\hat{x}_{j+1|j}), \tag{9.51}$$
$$P_{j+1|j} = AP_{j|j}A^{\mathrm{T}} + BQB^{\mathrm{T}}, \tag{9.52}$$
$$P_{j+1|j+1} = P_{j+1|j} - P_{j+1|j}C^{\mathrm{T}}(R + CP_{j+1|j}C^{\mathrm{T}})^{-1}CP_{j+1|j}. \tag{9.53}$$

*Proof.* We use induction, and assume that $V_k^{\mathrm{OPT}}(\hat{x}_k, \mu_0, y_1^d, \ldots, y_k^d)$ is a quadratic function of $\hat{x}_k$ of the form

[4] Here we assume $A$ nonsingular, but the result holds for any matrix $A$.

$$V_k^{\text{OPT}}(\hat{x}_k, \mu_0, y_1^d, \ldots, y_k^d) = \frac{1}{2}(\hat{x}_k - \hat{x}_{k|k})^{\text{T}} P_{k|k}^{-1}(\hat{x}_k - \hat{x}_{k|k}) + \text{constant}, \quad (9.54)$$

where $\hat{x}_{k|k}$ is the function of $\mu_0, y_1^d, \ldots, y_k^d$ defined via (9.48)–(9.53) for $j = 0, \ldots, k-1$. We note from (9.40) and (9.48)–(9.49) that the induction hypothesis holds for $k = 0$.

We next assume that (9.54) holds for $k$ and show, by performing the minimisation (9.44), that the results holds for $k+1$. (Note that, in this case, we should obtain that equations (9.50)–(9.53) apply for $j = k$.)

**Step 1:** As the first step towards performing the minimisation in (9.44), we begin by adding the term $\frac{1}{2}\hat{w}_k^{\text{T}} Q^{-1} \hat{w}_k$ to (9.54) and substituting $\hat{x}_k = A^{-1}\hat{x}_{k+1} - A^{-1}B\hat{w}_k$, and then minimise with respect to $\hat{w}_k$. We denote the resulting value function by $W_{k+1}^{\text{OPT}}(\hat{x}_{k+1}, \mu_0, y_1^d, \ldots, y_k^d)$, that is:

$$\begin{aligned} W_{k+1}^{\text{OPT}}(\hat{x}_{k+1}, \mu_0, y_1^d, \ldots, y_k^d) = \min_{\hat{w}_k} \Big\{ & \frac{1}{2}(A^{-1}\hat{x}_{k+1} - A^{-1}B\hat{w}_k - \hat{x}_{k|k})^{\text{T}} P_{k|k}^{-1} \\ & (A^{-1}\hat{x}_{k+1} - A^{-1}B\hat{w}_k - \hat{x}_{k|k}) \\ & + \frac{1}{2}\hat{w}_k^{\text{T}} Q^{-1} \hat{w}_k \Big\} + \text{ constant}. \end{aligned} \quad (9.55)$$

Differentiating the argument of the min in (9.55) with respect to $\hat{w}_k$ and equating to zero gives

$$B^{\text{T}} A^{-T} P_{k|k}^{-1} \left(\hat{x}_{k|k} - A^{-1}\hat{x}_{k+1} + A^{-1}B\hat{w}_k\right) + Q^{-1}\hat{w}_k = 0,$$

or

$$\begin{aligned} \hat{w}_k &= -\left(B^{\text{T}} A^{-T} P_{k|k}^{-1} A^{-1} B + Q^{-1}\right)^{-1} B^{\text{T}} A^{-T} P_{k|k}^{-1} \left(\hat{x}_{k|k} - A^{-1}\hat{x}_{k+1}\right) \\ &\triangleq -(\Gamma + \Theta)^{-1} B^{\text{T}} A^{-T} P_{k|k}^{-1}\ \alpha, \end{aligned} \quad (9.56)$$

where we have used the definitions

$$\Gamma \triangleq B^{\text{T}} A^{-T} P_{k|k}^{-1} A^{-1} B, \quad (9.57)$$

$$\Theta \triangleq Q^{-1}, \quad (9.58)$$

$$\alpha \triangleq \left(\hat{x}_{k|k} - A^{-1}\hat{x}_{k+1}\right). \quad (9.59)$$

Back-substituting (9.56) into (9.55), we obtain

$$
\begin{aligned}
W_{k+1}^{\mathrm{OPT}}(\hat{x}_{k+1}, \mu_0, y_1^d, \ldots, y_k^d) = & \frac{1}{2}\left[\alpha - A^{-1}B(\Gamma+\Theta)^{-1}B^{\mathrm{T}}A^{-T}P_{k|k}^{-1}\alpha\right]^{\mathrm{T}} P_{k|k}^{-1} \\
& \left[\alpha - A^{-1}B(\Gamma+\Theta)^{-1}B^{\mathrm{T}}A^{-T}P_{k|k}^{-1}\alpha\right] \\
& + \frac{1}{2}\alpha^{\mathrm{T}}P_{k|k}^{-1}A^{-1}B(\Gamma+\Theta)^{-1}\Theta \\
& (\Gamma+\Theta)^{-1}B^{\mathrm{T}}A^{-T}P_{k|k}^{-1}\alpha + \text{ constant} \qquad (9.60) \\
\triangleq & \frac{1}{2}\alpha^{\mathrm{T}}S\alpha + \text{ constant}, \qquad (9.61)
\end{aligned}
$$

where

$$
\begin{aligned}
S = & \left[I - A^{-1}B(\Gamma+\Theta)^{-1}B^{\mathrm{T}}A^{-T}P_{k|k}^{-1}\right]^{\mathrm{T}} P_{k|k}^{-1} \\
& \left[I - A^{-1}B(\Gamma+\Theta)^{-1}B^{\mathrm{T}}A^{-T}P_{k|k}^{-1}\right] \\
& + P_{k|k}^{-1}A^{-1}B(\Gamma+\Theta)^{-1}\Theta(\Gamma+\Theta)^{-1}B^{\mathrm{T}}A^{-T}P_{k|k}^{-1} \\
= & \ P_{k|k}^{-1} - 2P_{k|k}^{-1}A^{-1}B(\Gamma+\Theta)^{-1}B^{\mathrm{T}}A^{-T}P_{k|k}^{-1} \\
& + P_{k|k}^{-1}A^{-1}B(\Gamma+\Theta)^{-1}\left\{B^{\mathrm{T}}A^{-T}P_{k|k}^{-1}A^{-1}B\right\}(\Gamma+\Theta)^{-1}B^{\mathrm{T}}A^{-T}P_{k|k}^{-1} \qquad (9.62) \\
& + P_{k|k}^{-1}A^{-1}B(\Gamma+\Theta)^{-1}\Theta(\Gamma+\Theta)^{-1}B^{\mathrm{T}}A^{-T}P_{k|k}^{-1}.
\end{aligned}
$$

We note that the term in the { } in (9.62) is equal to $\Gamma$ defined in (9.57). Hence, the last three terms above can be combined to give

$$
S = P_{k|k}^{-1} - P_{k|k}^{-1}A^{-1}B(\Gamma+\Theta)^{-1}B^{\mathrm{T}}A^{-T}P_{k|k}^{-1}. \qquad (9.63)
$$

Substituting (9.63) and (9.57)–(9.59) in (9.61), we have

$$
\begin{aligned}
W_{k+1}^{\mathrm{OPT}}(\hat{x}_{k+1}, \mu_0, y_1^d, \ldots, y_k^d) = & \frac{1}{2}\left(\hat{x}_{k+1} - A\hat{x}_{k|k}\right)^{\mathrm{T}} A^{-T}\left\{P_{k|k}^{-1} - P_{k|k}^{-1}A^{-1}B\right. \\
& \left.\left[B^{\mathrm{T}}A^{-T}P_{k|k}^{-1}A^{-1}B + Q^{-1}\right]^{-1} B^{\mathrm{T}}A^{-T}P_{k|k}^{-1}\right\} \\
& A^{-1}(\hat{x}_{k+1} - A\hat{x}_{k|k}) + \text{ constant}, \\
\triangleq & \frac{1}{2}\left(\hat{x}_{k+1} - \hat{x}_{k+1|k}\right)^{\mathrm{T}}(P_{k+1|k})^{-1}\left(\hat{x}_{k+1} - \hat{x}_{k+1|k}\right) \\
& + \text{ constant},
\end{aligned}
$$

where we have used

$$
\begin{aligned}
& \hat{x}_{k+1|k} \triangleq A\hat{x}_{k|k}, \qquad \text{(which gives (9.50) for } j = k\text{)}, \\
& P_{k+1|k} \triangleq A\left\{P_{k|k}^{-1} - P_{k|k}^{-1}A^{-1}B\left[B^{\mathrm{T}}A^{-T}P_{k|k}^{-1}A^{-1}B + Q^{-1}\right]^{-1}\right. \\
& \qquad \left. B^{\mathrm{T}}A^{-T}P_{k|k}^{-1}\right\}^{-1} A^{\mathrm{T}}. \qquad (9.64)
\end{aligned}
$$

We also we note from (9.64) that

$$P_{k+1|k} = \Big\{ A^{-T} P_{k|k}^{-1} A^{-1} - A^{-T} P_{k|k}^{-1} A^{-1} B \\ \Big[ B^{\mathrm{T}} A^{-T} P_{k|k}^{-1} A^{-1} B + Q^{-1} \Big]^{-1} B^{\mathrm{T}} A^{-T} P_{k|k}^{-1} A^{-1} \Big\}^{-1}.$$

Using the matrix inversion lemma, we have

$$P_{k+1|k} = A P_{k|k} A^{\mathrm{T}} + BQB^{\mathrm{T}},$$

as in (9.52) for $j = k$. Thus, summarising step 1, we have shown that

$$W_{k+1}^{\mathrm{OPT}}(\hat{x}_{k+1}, \mu_0, y_1^d, \ldots, y_k^d) = \frac{1}{2} \left( \hat{x}_{k+1} - \hat{x}_{k+1|k} \right)^{\mathrm{T}} P_{k+1|k}^{-1} \left( \hat{x}_{k+1} - \hat{x}_{k+1|k} \right) \\ + \text{ constant}, \tag{9.65}$$

where $\hat{x}_{k+1|k}$ and $P_{k+1|k}$ satisfy (9.50) and (9.52), respectively, for $j = k$.

**Step 2:** We next add the term $\frac{1}{2}(y_{k+1}^d - C\hat{x}_{k+1})^{\mathrm{T}} R^{-1} (y_{k+1}^d - C\hat{x}_{k+1})$ to (9.65) to obtain

$$V_{k+1}^{\mathrm{OPT}}(\hat{x}_{k+1}, \mu_0, y_1^d, \ldots, y_{k+1}^d) = \frac{1}{2} (\hat{x}_{k+1} - \hat{x}_{k+1|k})^{\mathrm{T}} P_{k+1|k}^{-1} (\hat{x}_{k+1} - \hat{x}_{k+1|k}) \\ + \frac{1}{2} (y_{k+1}^d - C\hat{x}_{k+1})^{\mathrm{T}} R^{-1} (y_{k+1}^d - C\hat{x}_{k+1}) \\ + \text{ constant}. \tag{9.66}$$

We want to write (9.66) as a perfect square, that is,

$$V_{k+1}^{\mathrm{OPT}}(\hat{x}_{k+1}, \mu_0, y_1^d, \ldots, y_{k+1}^d) = \frac{1}{2} (\hat{x}_{k+1} - \hat{x}_{k+1|k+1})^{\mathrm{T}} P_{k+1|k+1}^{-1} \\ (\hat{x}_{k+1} - \hat{x}_{k+1|k+1}) + \text{ constant}. \tag{9.67}$$

To find the expression for $\hat{x}_{k+1|k+1}$ used in (9.67), we note that $\hat{x}_{k+1|k+1}$ is the minimum of $V_{k+1}^{\mathrm{OPT}}(\hat{x}_{k+1}, \mu_0, y_1^d, \ldots, y_{k+1}^d)$. Hence, to obtain $\hat{x}_{k+1|k+1}$, we differentiate (9.66) with respect to $\hat{x}_{k+1}$, evaluate at $\hat{x}_{k+1} = \hat{x}_{k+1|k+1}$ and set the result to zero, that is,

$$P_{k+1|k}^{-1} (\hat{x}_{k+1|k+1} - \hat{x}_{k+1|k}) - C^{\mathrm{T}} R^{-1} (y_{k+1}^d - C\hat{x}_{k+1|k+1}) = 0.$$

Adding and subtracting $C^{\mathrm{T}} R^{-1} C \hat{x}_{k+1|k}$, and rearranging, we have

$$(P_{k+1|k}^{-1} + C^{\mathrm{T}} R^{-1} C) \hat{x}_{k+1|k+1} = (P_{k+1|k}^{-1} + C^{\mathrm{T}} R^{-1} C) \hat{x}_{k+1|k} \\ + C^{\mathrm{T}} R^{-1} (y_{k+1}^d - C\hat{x}_{k+1|k}).$$

From the above expression we obtain

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

which gives (9.51) for $j = k$.

Similarly, to find the expression for $P_{k+1|k+1}$ used in (9.67), we differentiate (9.66) twice with respect to $\hat{x}_{k+1}$. This gives

$$P_{k+1|k+1}^{-1} \triangleq P_{k+1|k}^{-1} + C^{\mathrm{T}}R^{-1}C,$$

which, using the matrix inversion lemma, gives (9.53) for $j = k$.

Thus, we have established (9.67) and induction completes the proof. □

We can use the characterisation of the partial value functions given in Lemma 9.6.1 to derive the optimal estimator where we optimise with respect to both $\{\hat{w}_0, \ldots, \hat{w}_{k-1}\}$ and $\hat{x}_0$ (or, equivalently, $\hat{x}_k$). In particular, we have the following important result.

**Theorem 9.6.2 (Kalman Filter)** *The optimal estimate $\hat{x}_k$ for $x_k$ given the data $\mu_0$, $y_1^d$, $\ldots y_k^d$, satisfies*

$$\hat{x}_k = \hat{x}_{k|k},$$

*where $\hat{x}_{k|k}$ satisfies the recursions* (9.48) *to* (9.53)*.*

*Proof.* The optimal choice $\hat{x}_k = \hat{x}_{k|k}$ follows immediately by minimising (9.47) with respect to $\hat{x}_k$ since $\hat{x}_k$ is unconstrained here. □

**Remark 9.6.1 (Optimal Smoother).** Actually, the minimisation of (9.47) with respect to $\hat{x}_k$ yields optimal estimates of all states $x_0, \ldots, x_k$ given data up to time $k$. These are called optimal *smoothed* estimates, and will be denoted by $\hat{x}_{j|k}$ for $j = 0, \ldots, k$. They can be computed simply by running $\hat{x}_{k-1} = A^{-1}\hat{x}_k - A^{-1}B\hat{w}_{k-1}$ backwards starting from $\hat{x}_k = \hat{x}_{k|k}$ and using $\hat{w}_{k-1}, \hat{w}_{k-2}, \ldots, \hat{w}_0$ as in (9.56). Defining $\hat{w}_{j|k} \triangleq \hat{w}_j$, for $j = 0, \ldots, k-1$, the *optimal smoother* is then given by the recursion

$$\hat{x}_{j|k} = A^{-1}\hat{x}_{j+1|k} - A^{-1}B\hat{w}_{j|k} \quad \text{for } j = 0, \ldots, k-1,$$

where

$$\hat{w}_{j|k} = -\left(B^{\mathrm{T}}A^{-T}P_{j|j}^{-1}A^{-1}B + Q^{-1}\right)^{-1} B^{\mathrm{T}}A^{-T}P_{j|j}^{-1}\left(\hat{x}_{j|j} - A^{-1}\hat{x}_{j+1|k}\right),$$

and $\hat{x}_{j|j}$ and $P_{j|j}$ are given by (9.50)–(9.53). ∘

## 9.7 Nonlinear Problems

The above circle of ideas can be extended to nonlinear and/or non-Gaussian problems. Consider the following nonlinear Markov model:

$$x_{k+1} = f(x_k, w_k), \tag{9.68}$$

$$y_k = h(x_k) + v_k, \tag{9.69}$$

where $f$ and $h$ are continuously differentiable functions of their arguments, and $\partial f/\partial w_k$ is nonsingular. In (9.68)–(9.69), $\{w_k\}$ and $\{v_k\}$ are i.i.d. sequences having probability density functions that satisfy

$$p_w(w_k) = \begin{cases} p_1(w_k) & \text{for } w_k \in \Omega_1, \\ 0 & \text{otherwise,} \end{cases}$$

and such that $-\ln p_1(w_k) = \ell_1(w_k)$; and

$$p_v(v_k) = \begin{cases} p_2(v_k) & \text{for } v_k \in \Omega_2, \\ 0 & \text{otherwise,} \end{cases}$$

and such that $-\ln p_2(v_k) = \ell_2(v_k)$. Also, we assume

$$p_{x_0}(x_0) = \begin{cases} p_3(x_0) & \text{for } x_0 \in \Omega_3, \\ 0 & \text{otherwise,} \end{cases}$$

and $-\ln p_3(x_0) = \ell_3(x_0)$.

Using the rule of transformation of probability density functions for the model (9.68)–(9.69) we have:

$$\begin{aligned} p_{y_k|x_k}(y_k = y_k^d | x_k = \hat{x}_k) &= p_v(v_k = y_k^d - h(\hat{x}_k)), \\ p_{x_{k+1}|x_k}(x_{k+1} = \hat{x}_{k+1} | x_k = \hat{x}_k) &= p_w(w_k = \hat{w}_k) \left| \det \frac{\partial x_{k+1}}{\partial w_k}\bigg|_{\hat{x}_k, \hat{w}_k} \right|^{-1} \\ &= p_w(w_k = \hat{w}_k) \left| \det \frac{\partial f(\hat{x}_k, \hat{w}_k)}{\partial w_k} \right|^{-1}, \end{aligned}$$

for all $\hat{w}_k \in \Omega_1$ such that $\hat{x}_{k+1} = f(\hat{x}_k, \hat{w}_k)$.

Then, using the vector definitions in (9.14)–(9.17), the negative logarithm of the joint probability density function for states and outputs satisfies

$$\begin{aligned} -\ln p_{\mathbf{y}_N, \mathbf{x}_N}(\mathbf{y}_N = \mathbf{y}_N^d, \mathbf{x}_N = \hat{\mathbf{x}}_N) = \ell_3(\hat{x}_0) &+ \sum_{k=1}^{N} \ell_2(y_k^d - h(\hat{x}_k)) \\ &+ \sum_{k=0}^{N-1} \left[ \ell_1(\hat{w}_k) + \ln \left| \det \frac{\partial f(\hat{x}_k, \hat{w}_k)}{\partial w_k} \right| \right], \end{aligned} \tag{9.70}$$

subject to the constraints

$$\hat{x}_{k+1} = f(\hat{x}_k, \hat{w}_k) \quad \text{for } k = 0, \ldots, N-1, \tag{9.71}$$

$$\hat{w}_k \in \Omega_1 \quad \text{for } k = 0, \ldots, N-1, \tag{9.72}$$

$$y_k^d - h(\hat{x}_k) \in \Omega_2 \quad \text{for } k = 1, \ldots, N, \tag{9.73}$$

$$\hat{x}_0 \in \Omega_3. \tag{9.74}$$

Hence, we can find the JAPMP estimate (9.27) by minimising (9.70) subject to (9.71)–(9.74) (see (9.28)).

## 9.8 Relationship to Chapman–Kolmogorov Equation

We next relate the above ideas to the Chapman–Kolmogorov[5] equation for recursive nonlinear filtering. The latter equation allows one to *recursively* compute $p_{x_k|y_k,\ldots,y_1}(x_k|y_k, y_{k-1}, \ldots, y_1)$. Specifically, using the Markovian structure of (9.68), (9.69), we have, from Bayes' rule:

**Time Update[6] (Chapman-Kolmogorov Equation)**

$$p_{x_k|y_{k-1},\ldots,y_1}(x_k|y_{k-1}, \ldots, y_1)$$

$$= \int_{\mathbb{R}^n} p_{x_k,x_{k-1}|y_{k-1},\ldots,y_1}(x_k, x_{k-1}|y_{k-1}, \ldots, y_1)dx_{k-1} \tag{9.75}$$

$$= \int_{\mathbb{R}^n} p_{x_k|x_{k-1},y_{k-1},\ldots,y_1}(x_k|x_{k-1}, y_{k-1}, \ldots, y_1)$$

$$\times\, p_{x_{k-1}|y_{k-1},\ldots,y_1}(x_{k-1}|y_{k-1}, \ldots, y_1)dx_{k-1} \tag{9.76}$$

$$= \int_{\mathbb{R}^n} p_{x_k|x_{k-1}}(x_k|x_{k-1}) p_{x_{k-1}|y_{k-1},\ldots,y_1}(x_{k-1}|y_{k-1}, \ldots, y_1)dx_{k-1}, \quad k \geq 1. \tag{9.77}$$

**Observation Update[7]**

$$p_{x_k|y_k,\ldots,y_1}(x_k|y_k, \ldots, y_1)$$

$$= \frac{p_{y_k|x_k,y_{k-1},\ldots,y_1}(y_k|x_k, y_{k-1}, \ldots, y_1)\; p_{x_k|y_{k-1},\ldots,y_1}(x_k|y_{k-1}, \ldots, y_1)}{p_{y_k|y_{k-1},\ldots,y_1}(y_k|y_{k-1}, \ldots, y_1)} \tag{9.78}$$

$$= \frac{p_{y_k|x_k}(y_k|x_k)\; p_{x_k|y_{k-1},\ldots,y_1}(x_k|y_{k-1}, \ldots, y_1)}{p_{y_k|y_{k-1},\ldots,y_1}(y_k|y_{k-1}, \ldots, y_1)}, \quad k \geq 0, \tag{9.79}$$

[5] Sometimes misspelled in Australia as Kolmogoroo.

[6] In passing from (9.75) to (9.76) we use Bayes' rule, and from (9.76) to (9.77) we use the Markovian property of (9.68).

[7] Equality (9.78) follows from Bayes' rule, and, in passing from (9.78) to (9.79) we use the Markovian property of (9.69).

where

$$p_{y_k|y_{k-1},\ldots,y_1}(y_k|y_{k-1},\ldots,y_1) = \int_{\mathbb{R}^n} p_{y_k|x_k}(y_k|x_k) p_{x_k|y_{k-1},\ldots,y_1}(x_k|y_{k-1},\ldots,y_1) dx_k. \quad (9.80)$$

Notice that $p_{x_k|x_{k-1}}$ and $p_{y_k|x_k}$, needed in the evaluation of equations (9.77) and (9.79)–(9.80) are given in Section 9.7 above.

Given $p_{x_k|y_k,\ldots,y_1}(x_k|y_k,\ldots,y_1)$, one can then compute various estimates, for example:

(i) Conditional mean

$$\hat{x}_k^{[1]} = \int_{\mathbb{R}^n} x_k \; p_{x_k|y_k,\ldots,y_1}(x_k|y_k,\ldots,y_1) dx_k. \quad (9.81)$$

(ii) A posteriori most probable

$$\hat{x}_k^{[3]} = \arg\max_{x_k} \; p_{x_k|y_k,\ldots,y_1}(x_k|y_k,\ldots,y_1). \quad (9.82)$$

Thus the Chapman–Kolmogorov equation (9.77) and the observation update equation (9.80) offer more flexibility than the optimisation approach presented in Section 9.3 (for the linear constrained case) and Section 9.7 (for the nonlinear constrained case) since they describe the entire conditional distribution of $x_k$ given the (past) data $y_1,\ldots,y_k$. Given this distribution, one can then compute various estimates, for example, those given in (9.81) and (9.82). On the other hand, the Chapman–Kolmogorov equation is, in general, difficult to solve and require various approximations to be used, for example, those used in particle filtering (see, for example, Doucet, de Freitas and Gordon 2001). By way of contrast, the optimisation approach of Sections 9.3 and 9.7 can be solved via optimal control methods.

Finally, we note that the following two estimates are not, in general, equal:

(i) Joint a posteriori most probable [JAPMP]

$$\left[\hat{x}_0^{[2]},\ldots,\hat{x}_N^{[2]}\right] \triangleq \arg\max_{x_0,\ldots,x_N} \; p_{x_0,\ldots,x_N|y_1,\ldots,y_N}(x_0,\ldots,x_N|y_1,\ldots,y_N). \quad (9.83)$$

(ii) A posteriori most probable [APMP]

$$\hat{x}_N^{[3]} \triangleq \arg\max_{x_N} \; p_{x_N|y_1,\ldots,y_N}(x_N|y_1,\ldots,y_N) \quad (9.84)$$

$$= \arg\max_{x_N} \int_{\mathbb{R}^n\times\cdots\times\mathbb{R}^n} p_{x_0,\ldots,x_N|y_1,\ldots,y_N}(x_0,\ldots,x_N|y_1,\ldots,y_N) dx_0\ldots dx_{N-1}. \quad (9.85)$$

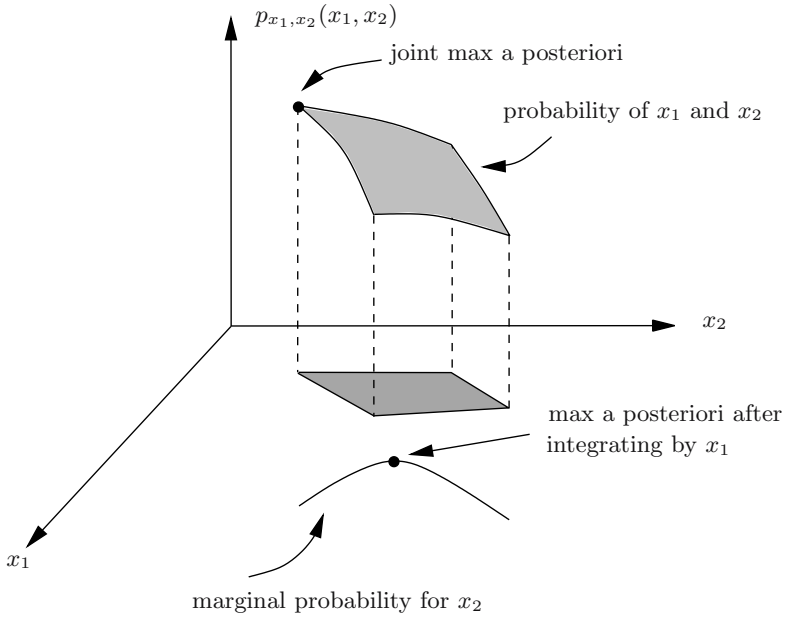


**Figure 9.5.** Difference between joint a posteriori maximum probability and a posteriori maximum probability.

This is illustrated in Figure 9.5.

However, if we use the conditional mean (9.81) as an estimate then we get the same answer whether we use the joint distribution for $\{x_0, \ldots, x_N\}$ or the marginal distribution for $x_N$. This follows because

$$\begin{aligned}
\hat{x}_N^{[1]} &= \int_{\mathbb{R}^n} x_N \; p_{x_N|y_N,\ldots,y_1}(x_N|y_N,\ldots,y_1)dx_N \\
&= \int_{\mathbb{R}^n} x_N \left[ \int_{\mathbb{R}^n \times \cdots \times \mathbb{R}^n} p_{x_N,\ldots,x_0|y_N,\ldots,y_1}(x_N,\ldots,x_0|y_N,\ldots,y_1) \; dx_{N-1}\ldots dx_0 \right] dx_N \\
&= \int_{\mathbb{R}^n \times \cdots \times \mathbb{R}^n} x_N \; p_{x_N,\ldots,x_0|y_N,\ldots,y_1}(x_N,\ldots,x_0|y_N,\ldots,y_1)dx_N \ldots dx_0,
\end{aligned}$$

since

$$\begin{aligned}
&p_{x_N|y_N,\ldots,y_1}(x_N|y_N,\ldots,y_1) \\
&\qquad = \int_{\mathbb{R}^n \times \cdots \times \mathbb{R}^n} p_{x_N,\ldots,x_0|y_N,\ldots,y_1}(x_N,\ldots,x_0|y_N,\ldots,y_1)dx_{N-1}\ldots dx_0.
\end{aligned}$$

## 9.9 Moving Horizon Estimation

As with control, we can readily convert the fixed horizon estimators discussed above into *moving horizon estimators* [MHE]. An issue to be addressed in this context is whether or not the situation allows *data-smoothing*; that is, whether one can collect data beyond the time at which the state estimate is required.

In some applications, for example, control, one requires that the estimate apply to the most recent state; that is, it is not possible to collect data beyond the point where the state estimate is defined. In other applications, for example, telecommunications, one can tolerate a delay between the last time at which the data are collected and the time at which the estimate is defined. In the latter situation we say that a *smoothed* state estimate is required.

To cover both of the above scenarios, we let $i$ denote the "time" at which the estimate is required. We also fix integers $L_1 \geq 0$ and $L_2 \geq 0$ and suppose for the moment that

$$x_{i-L_1} \sim N(z_{i-L_1}, P_{i-L_1}), \tag{9.86}$$

where $z_{i-L_1}$ is a given a priori estimate for $x_{i-L_1}$ having a Gaussian distribution. The matrix $P_{i-L_1}^{-1}$ reflects the degree of belief in this a priori estimate. We will treat the data in blocks of length $N = L_1 + L_2$. We assume that the estimate of $x_i$ can be based on data collected between $i - L_1$ and $i + L_2 - 1$. We then formulate the fixed horizon optimisation problem as in (9.29)–(9.35) over the interval $[i - L_1, i + L_2 - 1]$. That is, the corresponding sequences are indexed by $k = i - L_1, i - L_1 + 1, \ldots, i + L_2 - 1$. This yields the required estimate (or smoother for $L_2 > 1$, see Remark 9.6.1) of $x_i$.

The next question is how to turn this into a moving horizon procedure. The idea is to store the final state estimate $\hat{x}_{i+L_2-1}$ obtained from the above fixed horizon optimisation together with some measure of our degree of belief in this estimate, which we denote $P_{i+L_2-1}^{-1}$. The pair $(\hat{x}_{i+L_2-1}, P_{i+L_2-1})$ will be used to initialise a fixed horizon optimisation problem $L_1 + L_2$ steps ahead (that is, they will take the role of $(z_{i-L_1}, P_{i-L_1})$ in (9.86)).

We use again a Gaussian approximation when we return to this estimate. Of course, due to the constraints, we appreciate that the a posteriori distribution of the state will not be Gaussian. However, a Gaussian approach is justified on the following grounds:

(i) The "initial state information" is of diminishing importance as the block length $N$ increases.
(ii) Making a Gaussian approximation greatly simplifies the problems.
(iii) We can, at least, be compatible with the unconstrained case by determining $P_{i-L_1}$ from ordinary linear estimation theory.

Finally, the MHE is organised as illustrated in Figure 9.6. (Note that we need storage for $L_1 + L_2$ past state estimates to initialise subsequent blocks.)

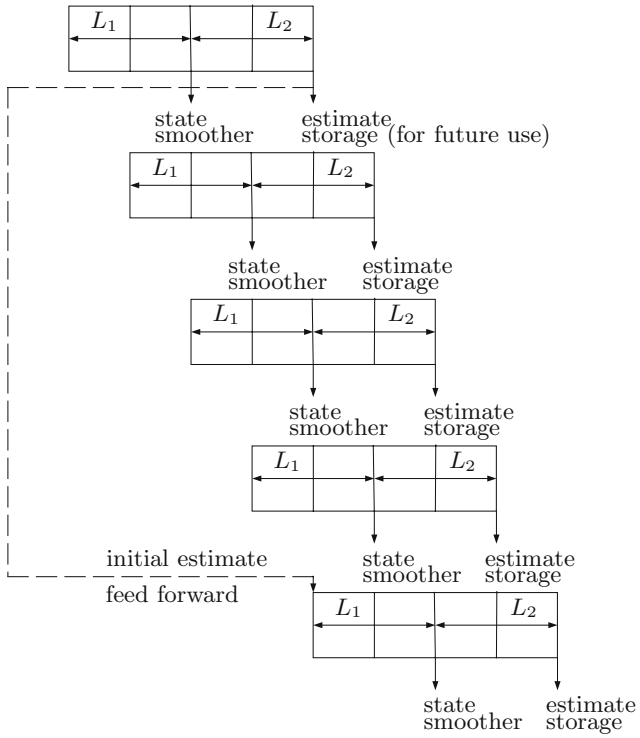


**Figure 9.6.** Graphical representation of MHE.

We next illustrate the idea of constrained estimation by three simple examples.

**Example 9.9.1.** Consider the same model as used in Example 1.3.1 of Chapter 1, which we repeat here for convenience:

$$y_k = w_k - 1.7w_{k-1} + 0.72w_{k-2} + v_k. \tag{9.87}$$

Rather than a binary signal, we here consider that the input noise $w_k$ has a truncated Gaussian distribution. We assume that the measurement noise $v_k$ has a Gaussian distribution. The details are:

- input noise variance prior to truncation: $Q = 1$;
- input noise mean prior to truncation: $\mu_w = 0$;
- measurement noise variance: $R = 0.2$;
- truncation interval: $w_k \in [-1, 1]$;
- input noise variance after truncation: $\approx 0.293$;
- input noise mean after truncation: 0.

Two estimators were compared, namely the MHE using $N = L_1 + 1 = 2$, $L_2 = 1$, incorporating the constraint $|w_k| \leq 1$, and a standard linear Kalman filter based on $R = 0.2$ and the true input variance of 0.293. The initial estimates as in (9.86) were selected as follows: $z_{i-N}$ is stored and propagated as in Figure 9.6; $P_{i-N}$ is set equal to the corresponding value for the Kalman filter. The results are shown in Figure 9.7. Some observations from this figure are:

(i) The linear Kalman filter performs quite well in this example. (This is not surprising since it is, after all, the best linear unbiased estimator.)
(ii) The estimates provided by the linear Kalman filter occasionally lie outside the range $\pm 1$. (Again, this is not surprising since this estimator is unconstrained.)
(iii) The MHE is slightly better but the result is marginal. (Again, this is not surprising in view of observation (i).) ○

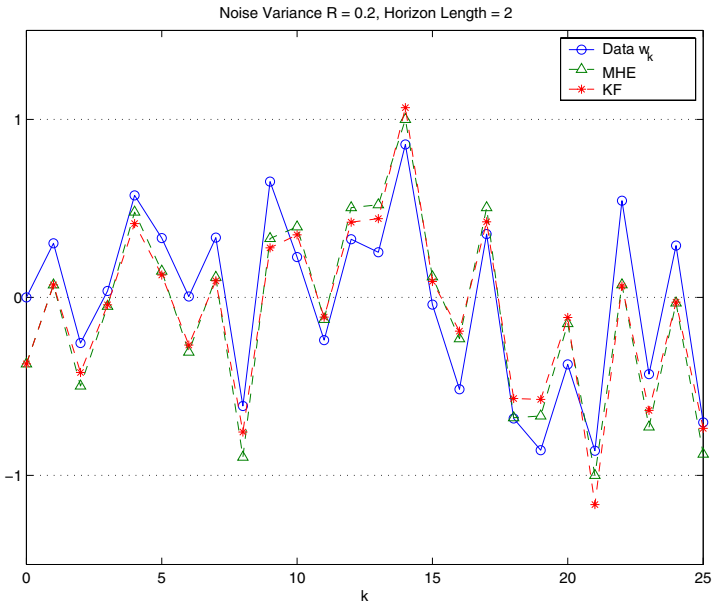


**Figure 9.7.** Comparison of MHE and Kalman filter with correct variance: data (circle-solid line), estimate provided by the MHE (triangle-dashed line) and estimate provided by the Kalman filter (star-dashed line).

**Example 9.9.2.** Here we consider the same model (9.87) as in Example 9.9.1, save that we change the input to a nonzero-mean truncated Gaussian distribution as illustrated in Figure 9.2. The details are:

- input noise variance prior to truncation: $Q = 1$;
- input noise mean prior to truncation: $\mu_w = 1.5$;
- truncation interval: $w_k \in [-1.5, 0.5]$;
- input noise variance after truncation: $\approx 0.175$;
- input noise mean after truncation: $\approx 0$.

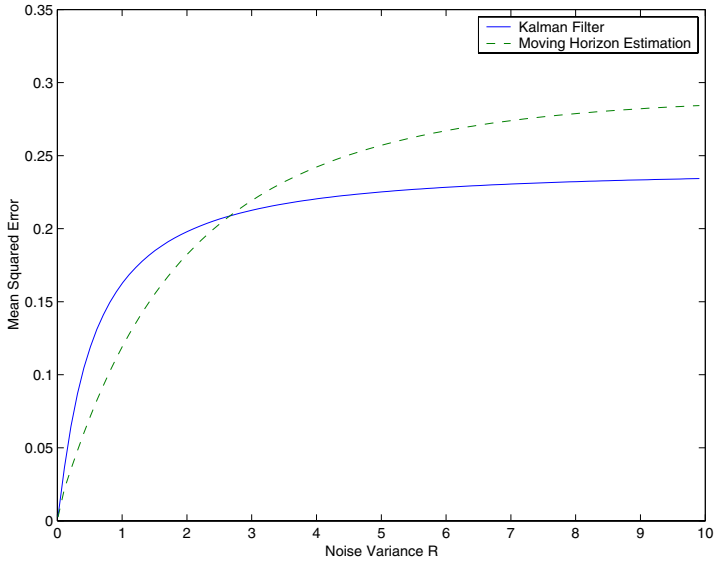


**Figure 9.8.** Comparison of mean square estimation error achieved by the MHE (dashed line) and the Kalman filter with correct variance (solid line).

Two estimators were compared, namely, MHE with $N = L_1 + 1 = 5$, $L_2 = 1$, and using the given constraints; and a standard linear Kalman filter based on the true variance. Figure 9.8 compares the mean square estimation errors for a range of measurement noise variances $R$.

It can be seen from Figure 9.8 that the MHE outperforms the Kalman filter save in the presence of large measurement noise. This result is in good accord with intuition since, for large measurement noise, the observations are basically ignored. This means that the Kalman filter gives the a priori mean, which is zero, whereas the MHE gives $w_k = 0.5$ since this corresponds to the point where the a priori probability is maximal.

○

**Example 9.9.3.** Here we consider the same channel model (9.87) as in Examples 9.9.1 and 9.9.2, save that now the input $w_k$ is distributed as the combination of two nonoverlapping, nonzero-mean truncated Gaussian distributions as in Figure 9.9. The distribution can be described by two regions: the "left region" is a Gaussian distribution $N(-1.5, 0.1)$ truncated between $[-1, 0]$, and

the "right region" is a Gaussian distribution $N(1.5, 0.1)$ truncated between $[0, 1]$. The resulting distribution has mean $\approx 0$ and variance $\approx 0.872$.

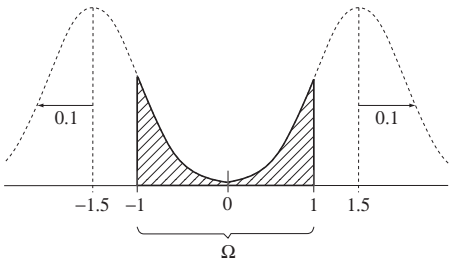


**Figure 9.9.** Combining the tails of two truncated Gaussian distributions.

We will compare the performance of the Kalman filter and the MHE for the above problem. The Kalman filter assumes a Gaussian approximation of the distribution, with zero mean and variance 0.872. For the MHE, we consider no smoothing, that is, $L_2 = 1$. The initial weighting $P$ is set equal to the value of the steady state error covariance of the Kalman filter, and the initial estimate is forwarded as in Figure 9.6. To find the optimal input sequence $\{\hat{w}_0, \ldots, \hat{w}_{N-1}\}$ the estimator solves, at each step, $2^N$ separate QP problems (see Section 9.4.2). The global optimum is the minimum of the individual sub-problems.

In Figure 9.10, we compare the Kalman filter estimates with those of the MHE for different measurement noise variances and different horizon lengths. In Figure 9.10 (a), incorporating mixed distributions with the MHE method and horizon 1 gives estimates that are closer to the boundary. On the other hand, the unconstrained Kalman filter exceeds the limits and tends to estimate near the zero mean. In Figure 9.10 (b) we see that the MHE performs more poorly as more measurement noise is introduced, since, in this case, the MHE tends to give the point where the a priori probability is maximal. By increasing the horizon length to 2 (see Figure 9.10 (c)), the estimator uses more data, resulting in better estimates. However, the number of sub-problems also increases. In Figure 9.10 (d), the horizon was increased to 4, showing a slight improvement in performance.

It should be observed that, since the distribution of the data points $w_k$ is close to the boundary, and with additive measurement noise, the MHE will give estimates that are close to the boundary. In the limiting case, when the distribution approaches a point mass distribution, the estimation problem will resemble that of the finite alphabet estimation problem, which is discussed in Chapter 13.

○

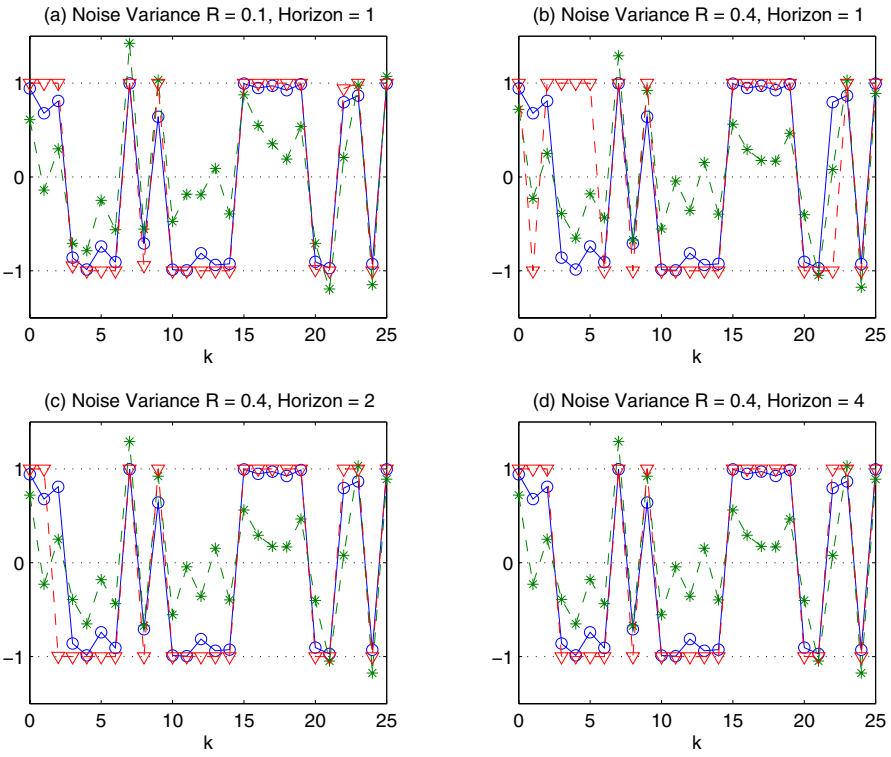


**Figure 9.10.** Data $w_k$ (circle-solid line), Kalman filter estimates (star-dashed line), MHE estimates (triangle-dashed line) for different measurement noise variances $R$ and different horizons $N$.

## 9.10 Further Reading

For complete list of references cited, see References section at the end of book.

### General

A useful introduction to estimation is given in Jazwinski (1970). The original derivation of the discrete Kalman filter used the concept of orthogonal projection (Kalman 1960a). The variational approach to estimation was first taken by Bryson and Frazier (1963). The solution of the continuous least square problem via dynamic programming was first given by Cox (1964).

### Section 9.4.3

The idea of utilising constraints in the context of approximating arbitrary distributions appears in Robertson and Lee (2002).

**Section 9.9**

Early work on moving horizon estimation appears in Michalska and Mayne (1995). See also Rao, Rawlings and Lee (2001) and Rao, Rawlings and Mayne (2003).