Chapter 1

Probability reminds

1.1 The Bayes Rule

Let A and B be two events. Then:

$$p(A|B) = p(B|A)\frac{p(A)}{p(B)} \tag{1.1}$$

This property simply follows from the definition of the conditioned probability: $p(A|B) = \frac{p(A\cap B)}{p(B)}$. Then we have $p(A \cap B) = p(A|B)p(B)$ and also $p(A \cap B) = p(B \cap A) = p(B|A)p(A)$, that is $p(A|B)p(B) = p(B|A)p(A)$, from which (1.1) follows.

If there is a third conditioning event C, we may write

$$p(A|B, C) = p(B|A, C)\frac{p(A|C)}{p(B|C)}.$$

In fact,

$$p(A|B, C) = p(A|B \cap C) = \frac{p(A \cap B \cap C)}{p(B \cap C)} = \frac{p(B \cap A \cap C)}{p(B \cap C)} = p(B|A \cap C)\frac{p(A \cap C)}{p(B \cap C)}$$

$$= p(B|A \cap C)\frac{p(A|C)p(C)}{p(B|C)p(C)} = p(B|A, C)\frac{p(A|C)}{p(B|C)}.$$

1.2 Total Probability Theorem

Let A be an event and B_i, $i = 1, 2, \ldots, n$ be a partition of the sample space Ω, i.e. $B_i \cap B_j = \emptyset$ and $\bigcup_{i=1}^{n} B_i = \Omega$. Then

$$p(A) = \sum_{i=1}^{n} p(A|B_i)p(B_i) \tag{1.2}$$

This property follows form the Bayes rule and from the fact that $p(A) = \sum_{i=1}^{n} p(A \cap B_i)$.

1.3 Independence

Two events A and B (or two random variables) are said *independent* if $p(A|B) = p(A)$ (that is the probability that A occurs is not modified by the fact that event B occurred or less). From the definition of conditioned probability it follows the well known property that if A and B are independent

$$p(A \cap B) = p(A|B)p(B) \underbrace{=}_{indip.} p(A)p(B),$$

that is the probability of the simultaneous occurrence of two independent events is the product of their probabilities.

1.4 Expectation, variance, covariance and correlation

We deal only with continuous random variables (the definition for the discrete case is completely equivalent). Given a (continuous) random variable with probability density function (pdf) $p(x)$, the *expectation* of x is the quantity

$$m_x = E[x] = \int_{-\infty}^{\infty} xp(x)dx$$

and is also called the *mean* of x. It is also possible to compute the expectation of any function $f(x)$ of the random variable x by $E[f(x)] = \int_{-\infty}^{\infty} f(x)p(x)dx$. In particular, the following quantity

$$\sigma_x^2 = E[(x - m_x)^2]$$

is the *variance* of the random variable x and tells about the uncertainty of x ($\sigma_x^2 = 0$ is the variance of a deterministic quantity). The quantity σ_x is called the *standard deviation* of x.

Given two random variables x and y, the *covariance* of x and y is defined as:

$$\sigma_{xy} = E[(x - m_x)(y - m_y)]$$

and is a measure of the connection between the two variables. It is often used a normalized version of the covariance which is the *correlation*

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Observing that the covariance of two random variables x and y can be seen as an inner product between them and that the standard deviation of a random variable x can be seen as the norm of x induced by this inner product, the Cauchy-Schwarz inequality implies:

$$1 \leq \rho_{xy} \leq 1,$$

that is, the correlation between x and y ranges from -1 (x and y negatively correlated) to 1 (x and y positively correlated). If $\rho_{xy} = 0$, x and y are said *uncorrelated*. It is important to observe that two independent random variables are always uncorrelated. In fact we have:

$$\sigma_{xy} = E[(x - m_x)(y - m_y)] = \int\int (x - m_x)(y - m_y)p(x, y)dxdy$$

$$\underbrace{=}_{\text{if independent}} \int (x - m_x)p(x)dx \int (y - m_y)p(y)dy = 0$$

So, **independence implies uncorrelation**, while the vice versa is in general not true: it holds, as mentioned afterwards, when x and y are jointly Gaussian. An example of uncorrelated but dependent random variables will be sketched next.

The extension to random vectors is immediate. Let $x \in \Re^n$ be a random vector with pdf $p(x) = p(x_1, x_2, \ldots, x_n)$. Then:

$$m_x = E[x] = \int \ldots \int xp(x)dx_1 \ldots dx_n = \int \ldots \int \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} p(x_1, \ldots, x_n)dx_1 \ldots dx_n$$

$$= \begin{bmatrix} \int \dots \int x_1 p(x_1, x_2, \dots, x_n) dx_1 \dots dx_n \\ \vdots \\ \int \dots \int x_n p(x_1, x_2, \dots, x_n) dx_1 \dots dx_n \end{bmatrix}$$

$$= \begin{bmatrix} \int dx_1 x_1 \int dx_2 \dots \int dx_n p(x_1, x_2, \dots, x_n) \\ \vdots \\ \int dx_n x_n \int dx_1 \dots \int dx_{n-1} p(x_1, x_2, \dots, x_n) \end{bmatrix} = \begin{bmatrix} \int x_1 p(x_1) dx_1 \\ \vdots \\ \int x_n p(x_n) dx_n \end{bmatrix} = \begin{bmatrix} E[x_1] \\ \vdots \\ E[x_n] \end{bmatrix}$$

since, as mentioned in Section 1.2 for the discrete case, $\int dx_2 \dots \int dx_n p(x_1, \dots, x_n) = p(x_1)$ (and similarly for the other terms). In place of the variance of a vector x we introduce the *covariance matrix*

$$\Sigma_x = E[(x - m_x)(x - m_x)']$$

which is a symmetric positive semidefinite $n \times n$ matrix. It is positive semidefinite because, for any constant vector a, $a'\Sigma_x a$ is the variance of the random variable $a'x$, which can not be negative.

1.5 Gaussian Random Vectors

A (scalar) Gaussian random variable x takes values on $(-\infty, \infty)$ according to the following probability density function:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}} \tag{1.3}$$

where m is the expected value of x:

$$m = E[x] = \int_{-\infty}^{\infty} x\, p(x)\, dx$$

and σ^2 is the variance of x:

$$\sigma^2 = E[(x-m)^2] = \int_{-\infty}^{\infty} (x-m)^2\, p(x)\, dx = E[x^2] - m^2.$$

It is often used the notation $x \sim \mathcal{N}(m, \sigma^2)$ to indicate that x is a Gaussian random variable with mean m and variance σ^2. We will also use the notation $\mathcal{N}(\xi; m, \sigma^2)$ to indicate that the Gaussian pdf $\mathcal{N}(m, \sigma^2)$ given in (1.3) is evaluated at $x = \xi$.

Definition 1 *A vector $x \in \Re^n$, $x = [x_1, x_2, \dots, x_n]'$, is a* **Gaussian random vector** *if any linear combination of its variables is a Gaussian random variable. That is, for any $a \in \Re^n$, $y = a' \cdot x \sim \mathcal{N}(m_y, \sigma_y^2)$ for proper m_y and σ_y^2.*

Clearly, given a Gaussian random vector $x = [x_1, x_2, \dots, x_n]'$, any x_i is a Gaussian random variable: it is enough to take in Definition 1 a such that $a_i = 1$ and $a_j = 0$ for all $j \neq i$. The vice versa is not true in general, i.e., if $x_1 \sim \mathcal{N}(m_1, \sigma_1^2)$ and $x_2 \sim \mathcal{N}(m_2, \sigma_2^2)$ are Gaussian, the vector $x = [x_1, x_2]'$ is not necessarily a Gaussian random vector, as the following example shows.

Example 1 *Let $x_1 \sim \mathcal{N}(0, \sigma^2)$ and define*

$$x_2 = \begin{cases} x_1 & \text{if } |x_1| \leq 1 \\ -x_1 & \text{if } |x_1| > 1 \end{cases}$$

It is possible to see that also x_2 is a Gaussian random variable, in particular $x_2 \sim \mathcal{N}(0, \sigma^2)$. However, $x = [x_1, x_2]'$ is not a Gaussian random vector: just take the linear combination $y = x_1 + x_2$ to see that

$$y = \begin{cases} 2x_1 & \text{if } |x_1| \leq 1 \\ 0 & \text{if } |x_1| > 1 \end{cases}$$

is clearly not a Gaussian random variable (in particular $|y| \leq 2$ does not range in $(-\infty, \infty)$).

However, if $x_1 \sim \mathcal{N}(m_1, \sigma_1^2)$ and $x_2 \sim \mathcal{N}(m_2, \sigma_2^2)$ are Gaussian and *independent*, then $x = [x_1, x_2]'$ is a Gaussian random vector. This can be proven using the characteristic function[1]. In fact, if x_1 and x_2 are two independent Gaussian random variables, any linear combination of them is the sum of two properly scaled independent Gaussian random variables. Since the characteristic function of the sum of two independent random variables is the product of their characteristic functions (this holds in general also for non Gaussian random variables and depends on the fact that the pdf of the sum of two independent random variables is the convolution of their pdf), using the expression of the characteristing function of a Gaussian random variable, we get the desired property that the linear combination of independent Gaussian random variables is Gaussian. We omit the details of this proof.

Using the characteristic function, it is also possible to show the following important result, which proof is not difficult but is omitted for brevity.

Theorem 1 *Let $x = [x_1, x_2, \ldots, x_n]'$ be a* **Gaussian random vector**. *Then*

$$p(x) = \frac{1}{\sqrt{det(2\pi \cdot \Sigma)}} e^{-\frac{1}{2}(x-m)'\Sigma^{-1}(x-m)} \tag{1.4}$$

where $m = E[x] \in \Re^n$ is the expected value of x and $\Sigma = E[(x-m)(x-m)'] \in \Re^{n \times n}$ is the covariance matrix *of x.*

Similarly to the scalar case, it is often used the notation $x \sim \mathcal{N}(m, \Sigma)$ to indicate that x is a Gaussian random vector with mean m and covariance matrix Σ. The elements of Σ are such that $\Sigma_{ii} = \sigma_i^2 = E[(x_i - m_i)^2]$ is the variance of x_i and $\Sigma_{ij} = \sigma_{ij} = E[(x_i - m_i)(x_j - m_j)]$ is the covariance of x_i and x_j. As mentioned above, if $\sigma_{ij} = 0$, x_i and x_j are uncorrelated, and it is easy to verify from the pdf expression given in Theorem 1 that two jointly Gaussian random variables are independent if and only if they are uncorrelated[2]. In fact we already know that independence always implies uncorrelation. But, if two joint Gaussian variables x and y are uncorrelated, (x, y) is a Gaussian random vector with a diagonal covariance matrix Σ. It is then possible to factorize the pdf, i.e. $p(x, y) = p(x)p(y)$ which proves the independence of x and y.

1.5.1 Properties of Gaussian random vectors

We present some important properties of Gaussian random vectors, useful for the derivation of the Kalman Filter.

1. **A linear transformation of a Gaussian random vector gives a Gaussian random vector.**

   Let $x = [x_1, \ldots, x_n]' \sim \mathcal{N}(m, \Sigma)$ be a Gaussian random vector and consider $y = Ax$, where A is a $q \times n$ matrix. Then also $y \in \Re^q$ is a Gaussian random vector, with $y \sim \mathcal{N}(A \cdot m, A\Sigma A')$.

   In fact, if x is Gaussian, for all $a \in \Re^n$, $a'x$ is a Gaussian random variable. Also y is Gaussian if for all $b \in \Re^q$, $b'y$ is a Gaussian random variable. Now, $b'y = b'Ax = a'x$ (with $a = A'b$) is a Gaussian random variable.

   Since

   $$E[y] = E[Ax] = A \cdot E[x] = A \cdot m$$

[1] The characteristic function of a random variable x with density $p(x)$ is defined as $F_x(jv) = E[e^{jvx}]$, where j is the imaginary unit. Actually, the characteristic function is a sort of Fourier Transform for probability density functions. This definition can be naturally extended also to random vectors x by taking $F_x(jv) = E[e^{jv'x}] = E[e^{j(v_1x_1 + \ldots + v_nx_n)}]$.

[2] Notice that this is not true for two generic Gaussian random variables: in particular it is possible to show that for a proper value of σ the two variables in Example 1 are uncorrelated with $E[(x_1 - m_1)(x_2 - m_2)] = E[x_1x_2] = \int_{|x_1|<1} x_1^2 p(x_1)dx_1 - \int_{|x_1|>1} x_1^2 p(x_1)dx_1 = 0$ but are clearly not independent. This is because they are Gaussian but not *jointly* Gaussian.

and

$$E[(y-m_y)(y-m_y)'] = E[A(x-m)(x-m)'A'] = AE[(x-m)(x-m)']A' = A\Sigma A',$$

it follows that $y \sim \mathcal{N}(A \cdot m, A\Sigma A')$.

2. **The sum of two independent Gaussian random vectors is a Gaussian random vector.**

   Let $x_1 \sim \mathcal{N}(m_1, \Sigma_1)$ and $x_2 \sim \mathcal{N}(m_2, \Sigma_2)$ be two independent Gaussian random vectors. Then $z = x_1 + x_2$ is a Gaussian random vector $z \sim \mathcal{N}(m_1 + m_2, \Sigma_1 + \Sigma_2)$.

   Now, as mentioned, the sum of two independent Gaussian random variables is a Gaussian random variable. Its mean is the sum of the means and its variance is the sum of the variances (mean and variance of the sum of two independent random variables is the sum of the means and respectively of the variances also in the case of non Gaussian random variables, as it is easy to verify). This implies that the vector $w = [x_1', x_2']'$ is a Gaussian vector. In fact, for any a_1, $y_1 = a_1' \cdot x_1$ is a Gaussian random variable and for any a_2, $y_2 = a_2' \cdot x_2$ is a Gaussian random variable. Clearly y_1 and y_2 are independent for any a_1 and a_2, hence it is a Gaussian variable also the sum $y_1 + y_2$. But then, for any $c \in \Re^{2n}$, letting $c = [a_1' \ a_2']'$, it follows that $c'w = a_1'x_1 + a_2'x_2 = y_1 + y_2$ is a Gaussian random variable. So w is a Gaussian random vector.

   Consider now the linear transformation

$$z = Aw = [I_n \quad I_n] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

   where I_n is the identity matrix of dimension n. Clearly $z = x_1 + x_2$. According to property 1, z is also a Gaussian random vector with

$$m_z = A \cdot m_w = [I_n \quad I_n] \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = m_1 + m_2$$

   and

$$\Sigma_z = A\Sigma_w A' = [I_n \quad I_n] \begin{bmatrix} \Sigma_1 & 0_n \\ 0_n & \Sigma_2 \end{bmatrix} \begin{bmatrix} I_n \\ I_n \end{bmatrix} = \Sigma_1 + \Sigma_2.$$

3. **The affine combination of independent Gaussian random vectors is a Gaussian random vector**.

   Combining the two previous properties, one gets the following property. Let $x \sim \mathcal{N}(m_x, \Sigma_x)$ and $y \sim \mathcal{N}(m_y, \Sigma_y)$ be two independent Gaussian random vectors and b a constant vector. Then

$$z = Ax + By + b$$

   is also a Gaussian random vector with mean

$$m_z = Am_x + Bm_y + b$$

   and covariance matrix

$$\Sigma_z = A\Sigma_x A' + B\Sigma_y B'.$$

Chapter 2

Estimators

Let $x \in \Re^n$ be a vector of unknown to be estimated and let $Y = [y_1, y_2, \ldots, y_t]'$ be a vector of (e.g. scalar) measurements correlated with x under a proper model[1] in such a way that it is well defined the conditioned probability of x given Y

$$p(x|Y).$$

Define, accordingly,

$$E[x|Y] = \int_{-\infty}^{\infty} xp(x|Y)dx$$

the expectation of x given Y.

An estimator of x given Y is a quantity $\hat{x}$ which depends *deterministically* on Y and tries to guess the value of x, according to some criterion. We introduce some possible estimators.

2.1 Least Square estimator

A LSQ estimator of x given Y is the quantity $\hat{x}_{LSQ}$ minimizing the expected square error, i.e.

$$\hat{x}_{LSQ} = \arg\min_{\hat{x}} E[(\hat{x} - x)'(\hat{x} - x)|Y]$$

The following facts hold, as shown subsequently:

1. The LSQ estimator is the conditioned expected value of x, i.e.

$$\hat{x}_{LSQ} = \bar{x}_Y = E[x|Y].$$

2. If

$$\hat{P} = E[(\hat{x} - x)(\hat{x} - x)'|Y]$$

is the covariance matrix associated with a given estimator $\hat{x}$, then

- the covariance matrix $\hat{P}_{LSQ}$ of $\hat{x}_{LSQ}$ is the covariance matrix of x given Y (i.e. $\hat{P}_{LSQ} = P_Y$, where $P_Y = E[(x - \bar{x}_Y)(x - \bar{x}_Y)'|Y]$) and is such that $\hat{P} - \hat{P}_{LSQ}$ is positive semidefinite (i.e. the LSQ estimator is a minimum variance estimator, in the sense that in the scalar case it has the minimum variance while in the vectorial case it is the estimator $\hat{x}$ minimizing $w'\hat{P}w$ for any w, in such a way that, if we want to estimate a linear combination $w'x$ of x, the estimator with minimum variance is $w'\hat{x}_{LSQ}$);
- the expected square error is the trace of $\hat{P}$, hence $\hat{x}_{LSQ}$, by definition, minimizes $tr(\hat{P})$.

[1] You may think for example that $y_i = C_i x + n_i$ where n_i is some noise, $i = 1, 2, \ldots$ and that in the absence of noises it is possible to uniquely determine x from Y. However this is not essential in the definitions given in this section.

Let's show the first fact, that $\hat{x}_{LSQ} = E[x|Y]$ (denoted for brevity $\bar{x}_Y$). Take any estimator $\hat{x} = \bar{x}_Y + v_Y$, where v_Y is a deterministic vector function of Y. Then we have, for the expected square error:

$$E[(\hat{x}-x)'(\hat{x}-x)|Y] = E[(\bar{x}_Y + v_Y - x)'(\bar{x}_Y + v_Y - x)|Y] = E[(\bar{x}_Y - x)'(\bar{x}_Y - x)|Y]$$

$$+2E[v_Y'(\bar{x}_Y - x)|Y] + v_Y'v_Y = trace(P_Y) + 2v_Y'E[(\bar{x}_Y - x)|Y] + \|v_Y\|^2 = trace(P_Y) + \|v_Y\|^2$$

Clearly, since P_Y does not depend on $\hat{x}$, this error is minimized by taking $v_Y = 0$, i.e. $\hat{x}_{LSQ} = \bar{x}_Y$. This also implies that $P_{LSQ} = P_Y$.

Similarly, for any estimator $\hat{x} = \bar{x}_Y + v_Y$, we have:

$$\hat{P} = E[(\hat{x}-x)(\hat{x}-x)'|Y] = E[(\bar{x}_Y + v_Y - x)(\bar{x}_Y + v_Y - x)'|Y] = E[(\bar{x}_Y - x)(\bar{x}_Y - x)'|Y]$$

$$+2E[v_Y(\bar{x}_Y - x)'|Y] + v_Y v_Y' = P_Y + 2v_Y E[(\bar{x}_Y - x)'|Y] + v_Y v_Y' = P_Y + v_Y v_Y' = P_Y + V = \hat{P}_{LSQ} + V$$

for some positive semidefinite matrix $V = v_Y v_Y'$. This shows the second fact, that $\hat{P} - \hat{P}_{LSQ}$ is positive semidefinite. The third fact follows, as mentioned, from the definition of $\hat{x}_{LSQ}$.

2.2 Maximum Likelihood estimator

A ML estimator $\hat{x}_{ML}$ of x given Y is the quantity x maximizing the probability of obtaining the measurements Y, i.e.

$$\hat{x}_{ML} = \arg\max_x p(Y|x)$$

2.3 Maximum A Posteriori estimator

A MAP estimator $\hat{x}_{MAP}$ of x given Y is the most likely x corresponding to the measurements Y, i.e.

$$\hat{x}_{MAP} = \arg\max_x p(x|Y)$$

This estimator differs from the ML estimator because the ML ignores the *a priori* probability density of x. To see this, let $p(x)$ be the a priori probability of x (i.e. the pdf describing the possible values of x before taking the measurements; if $p(x)$ is uniform, that is no prior information is available on possible values of x, it is $\hat{x}_{ML} \equiv \hat{x}_{MAP}$). Then, applying the Bayes rule, we have:

$$\hat{x}_{MAP} = \arg\max_x p(x|Y) \overbrace{=}^{Bayes} \arg\max_x p(Y|x)\frac{p(x)}{p(Y)} = \arg\max_x p(Y|x)p(x)$$

where the last equality depends on the fact that $p(Y)$ is independent of x. So:

$$\begin{aligned} \hat{x}_{MAP} &= \arg\max_x p(Y|x)p(x) \\ \hat{x}_{ML} &= \arg\max_x p(Y|x) \end{aligned}$$

In robotics applications, it is important to take into account the prior knowledge of x, i.e. $p(x)$. For this reason, a MAP estimator will be usually more appealing w.r.t. a ML estimator.

Chapter 3

The Kalman Filter

Consider the discrete time linear system:

$$x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1} \tag{3.1}$$
$$y_k = Cx_k + v_k \tag{3.2}$$

where $x_k \in \Re^n$, $y_k \in \Re^q$, $u_k \in \Re^p$ and the vectors w_k, v_k are two Gaussian white sequences[1] of noise, independent one each other and independent of x_0, which is also assumed Gaussian. In particular:

$$x_0 \sim \mathcal{N}(m_0, P_0)$$
$$w_k \sim \mathcal{N}(0, Q_k) \quad v_k \sim \mathcal{N}(0, R_k)$$

where R_k is assumed strictly positive definite (this is usually of practical interest, since it corresponds to the fact that it is not possible to find a linear combination of the measurements y_k which is not noisy). We have assumed that the noise w_{k-1} in (3.1) is not pre-multiplied by any matrix. This is w.l.o.g. since, if we have $x_k = Ax_{k-1} + Bu_{k-1} + B_w w_{k-1}$, it is possible to define a noise $\bar{w}_k = B_w w_k$ which is still a white Gaussian sequence and has a covariance matrix given by $\bar{Q}_k = B_w Q_k B_w'$. Also v_k in (3.2) could be pre-multiplied by a matrix C_v: the only assumption in this case is that the resulting covariance matrix $\bar{R}_k = C_v R_k C_v'$ is strictly positive definite. Finally, the case the noises are not zero mean can be easily handled with minor modifications (just subsume the known expected value of the noise w_{k-1} in the constant term Bu_{k-1} and add a constant vector to the equation in (3.2)).

Under these assumptions (i.e.: *linearity of the dynamics, Gaussianity of the noise and independence of the sequences*) we have the following strong property: **the probability density function of x_k conditioned on past measurements (and controls) is Gaussian**. More in detail, let

$$y^k = \{y_1, y_2, \ldots, y_k\} \qquad \text{and} \qquad u^{k-1} = \{u_0, u_1, \ldots, u_{k-1}\}$$

denote the sequence of measurements and control inputs available at time k. Then, the *a priori* posterior $p(x_k|u^{k-1}, y^{k-1})$ (i.e. considering all previous measurements and controls but not the last available measurement y_k) and the *a posteriori* posterior $p(x_k|u^{k-1}, y^k)$ (i.e. considering all available measurements and controls, i.e. also the last measurement y_k) are *Gaussian* densities! We will adopt the following notation:

$$p(x_k|u^{k-1}, y^{k-1}) = \mathcal{N}(m_k^-, P_k^-) \tag{3.3}$$
$$p(x_k|u^{k-1}, y^k) = \mathcal{N}(m_k, P_k) \tag{3.4}$$

This fact, which will be shown in this section, has the following strong implications:

[1] A sequence is said *Gaussian* if any set of its elements has a jointly Gaussian pdf. It is said *white* if each element ν_k in the sequence is 0-mean and is uncorrelated with the others, i.e., $E[\nu_k \nu_h'] = S_k \delta(k-h)$, for some positive semidefinite matrix S_k (where $\delta(k-h) = 1$ if $h = k$ and is 0 otherwise). Since in the Gaussian case uncorrelation implies independence, a white Gaussian sequence is also an *i.i.d.* (independent and identically distributed) Gaussian sequence, where i.i.d. means that all the elements in the sequence have the same pdf and are independent one each other.

- Since a Gaussian density is completely described through its mean and covariance matrix, an algorithm which is able to compute the updated means m_k^- and m_k and covariance matrices P_k^- and P_k *provides the complete description of the possible values the state* x_k *can take.* The algorithm which performs this service is the *Kalman Filter.*
- Since the Gaussian density is unimodal, with the maximum coincident with the expected value, according to what mentioned in Chapter 2, the *Kalman filter provides through the updated value of the means* m_k^- *and* m_k *both the LSQ and the MAP estimates of the state* x_k *of the system.* But, also, through the covariance matrices P_k^- and P_k, it gives a measurement of the quality of the estimate.

3.1 Proof of Gaussianity

In this section we prove (3.3)-(3.4) and provide the recursive equations to compute m_k^-, m_k, P_k^- and P_k, which are indeed the equations of the Kalman Filter.

3.1.1 The prediction step of the Kalman Filter

The prior density $p(x_k|u^{k-1}, y^{k-1})$ which uses the last control u_{k-1} but not the last measurement y_k is a *prediction* on the next value x_k of the state. We show (3.3) and provide the equations of the prediction step of the Kalman Filter, namely:

$$\begin{aligned} m_k^- &= Am_{k-1} + Bu_{k-1} && (3.5)\\ P_k^- &= AP_{k-1}A' + Q_{k-1} && (3.6) \end{aligned}$$

This can be shown in a very straightforward manner by considering the results reported in Chapter 1. In particular, by induction on the fact that $x_0 \sim \mathcal{N}(m_0, P_0)$ is a Gaussian random vector, by assuming that $p(x_{k-1}|u^{k-2}, y^{k-1}) = \mathcal{N}(m_{k-1}, P_{k-1})$ is Gaussian, i.e. that x_{k-1} given all available measurements y^{k-1} and controls u^{k-2} is a Gaussian vector, we show that this holds also for $p(x_k|u^{k-1}, y^{k-1})$, i.e. that x_k given all available controls u^{k-1} and past measurements y^{k-1}, is still a Gaussian vector. Now

$$x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1}$$

is an affine combination of independent Gaussian vectors (x_{k-1} and w_{k-1}). Hence, according to Chapter 1, x_k is also a Gaussian vector with:

$$\begin{aligned} m_k^- = E[x_k|u^{k-1}, y^{k-1}] &= E[Ax_{k-1} + Bu_{k-1} + w_{k-1}|u^{k-1}, y^{k-1}] = AE[x_{k-1}|u^{k-1}, y^{k-1}] + Bu_{k-1} \\ &= AE[x_{k-1}|u^{k-2}, y^{k-1}] + Bu_{k-1} = Am_{k-1} + Bu_{k-1}. \end{aligned} \quad (3.7)$$

Similarly, omitting for brevity the conditioning information $\{u^{k-1}, y^{k-1}\}$ from the expectations and using the independence of w_{k-1} w.r.t. x_{k-1} (which indeed depends on w^{k-2}):

$$\begin{aligned} P_k^- &= E[(x_k - m_k^-)(x_k - m_k^-)'] = E[(A(x_{k-1} - m_{k-1}) + w_{k-1})(A(x_{k-1} - m_{k-1}) + w_{k-1})'] \\ &= AE[(x_{k-1} - m_{k-1})(x_{k-1} - m_{k-1})']A' + E[w_{k-1}w_{k-1}'] = AP_{k-1}A' + Q_{k-1}. \end{aligned} \quad (3.8)$$

3.1.2 The correction step

Exploiting also the last available measurement y_k, we *correct* the prior estimation m_k^- on x_k as follows.

$$p(x_k|u^{k-1}, y^k) = p(x_k|y_k, u^{k-1}, y^{k-1}) \underbrace{=}_{Bayes} p(y_k|x_k, u^{k-1}, y^{k-1}) \frac{p(x_k|u^{k-1}, y^{k-1})}{p(y_k|u^{k-1}, y^{k-1})}$$

$$= \eta \;\; p(y_k|x_k, u^{k-1}, y^{k-1}) p(x_k|u^{k-1}, y^{k-1}) \tag{3.9}$$

where the last equality depends on the fact that $p(y_k|u^{k-1}, y^{k-1})$ is a constant equal for all x_k and is written as a normalizing factor η. Now,

$$p(y_k|x_k, u^{k-1}, y^{k-1}) = p(y_k|x_k)$$

since, from $y_k = Cx_k + v_k$, if we know x_k, y_k is independent of the past controls and measurements (also in view of the hypothesis on the sequence v_k). Now, we have just shown that

$$p(x_k|u^{k-1}, y^{k-1}) = \mathcal{N}(m_k^-, P_k^-)$$

and, in addition,

$$p(y_k|x_k) = p_{v_k}(y_k - Cx_k)$$

where $p_{v_k}(s) = \mathcal{N}(s; 0, R_k) = \frac{1}{det(2\pi R_k)^{1/2}} e^{-\frac{1}{2} s' R_k s}$ is the pdf of the noise v_k. So, (3.9) can be written as follows:

$$p(x_k|u^{k-1}, y^k) = \eta \;\; p(y_k|x_k) p(x_k|u^{k-1}, y^{k-1}) = \eta \;\; \mathcal{N}(y_k - Cx_k; 0, R_k) \mathcal{N}(x_k; m_k^-, P_k^-). \tag{3.10}$$

This is the product between two Gaussians, where the argument of the two exponentials sums and gives (omitting the factor $-1/2$):

$$argExp = (x_k - m_k^-)'(P_k^-)^{-1}(x_k - m_k^-) + (y_k - Cx_k)' R_k^{-1} (y_k - Cx_k)$$

Omitting for simplicity the sub k and the superscript $-$, the previous equation can be written:

$$argExp = (x - m)' P^{-1} (x - m) + (y - Cx)' R^{-1} (y - Cx)$$

Expanding the products, we get:

$$argExp = x'(P^{-1} + C'R^{-1}C)x - 2(m'P^{-1} + y'R^{-1}C)x + G \tag{3.11}$$

where G is a quantity independent of x which can be subsumed in the normalization term η in (3.9). To show that $p(x_k|u^{k-1}, y^k)$ in (3.10) is a Gaussian density, we have to show that $argExp$ is a quadratic form of x, i.e., that there exist a symmetric and positive definite matrix S and a vector a such that

$$x'(P^{-1} + C'R^{-1}C)x - 2(m'P^{-1} + y'R^{-1}C)x = (x - a)' S^{-1} (x - a) + H \tag{3.12}$$

where H, again, is some constant quantity independent of x. Expanding the products, we obtain:

$$(x - a)' S^{-1} (x - a) = x' S^{-1} x - 2a' S^{-1} x + a' S^{-1} a \tag{3.13}$$

Combining (3.12) and (3.13), we can conclude on the Gaussianity of $p(x_k|u^{k-1}, y^k)$ if it is possible to determine S and a (which will correspond respectively to P_k and m_k) such that:

$$S^{-1} = P^{-1} + C'R^{-1}C \tag{3.14}$$

$$a'S^{-1} = m'P^{-1} + y'R^{-1}C \tag{3.15}$$

We exploit the following linear algebra result, where A, B, C and D are matrices (with proper dimensions) with A and C invertible:

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1} DA^{-1} \tag{3.16}$$

Applying this formula to (3.14) we see that S exists and is given by:

$$S = (\underbrace{P^{-1}}_{A} + \underbrace{C'}_{B} \underbrace{R^{-1}}_{C} \underbrace{C}_{D})^{-1} = P - PC'(CPC' + R)^{-1} CP = [I - PC'(CPC' + R)^{-1} C]P. \tag{3.17}$$

If S exists, also a is well defined and this concludes the proof that $p(x_k|u^{k-1}, y^k)$ is a Gaussian pdf. About the computation of $m_k = a$ and of $P_k = S$ we proceed as follows.

Reintroducing the sub k and the superscript $-$ in (3.17) we obtain:

$$P_k = [I - P_k^- C'(CP_k^- C' + R_k)^{-1} C] P_k^-$$

Defining the *Kalman gain*:

$$K_k = P_k^- C'(CP_k^- C' + R_k)^{-1}, \tag{3.18}$$

we finally have:

$$P_k = [I - K_k C] P_k^- \tag{3.19}$$

As for a (i.e. m_k), we have from (3.15):

$$S^{-1} a = C' R^{-1} y + P^{-1} m$$

i.e.

$$a = S(C'R^{-1}y + P^{-1}m) = S[C'R^{-1}(Cm + y - Cm) + P^{-1}m] = S\underbrace{[C'R^{-1}C + P^{-1}]}_{S^{-1}} m + SC'R^{-1}(y - Cm)$$

$$= m + L(y - Cm) \tag{3.20}$$

where $L = SC'R^{-1}$. Using the expression of S given in (3.17):

$$L = [I - PC'(CPC' + R)^{-1}C]PC'R^{-1} = PC'R^{-1} - PC'(CPC' + R)^{-1}CPC'R^{-1}$$

$$= PC'[R^{-1} - (CPC' + R)^{-1}CPC'R^{-1}]$$

Using again the formula reported in (3.16):

$$L = PC'[\underbrace{R^{-1}}_{A} \underbrace{-}_{B} \underbrace{(CPC' + R)^{-1}}_{C} \underbrace{CPC'R^{-1}}_{D}] = PC'(R + R(-CPC'R^{-1}R + CPC' + R)^{-1}CPC'R^{-1}R)^{-1}$$

$$= PC'(R + RR^{-1}CPC'R^{-1}R)^{-1} = PC'(R + CPC')^{-1}$$

Reintroducing also here the sub k and the superscript $-$, we obtain:

$$L = P_k^- C'(R_k + CP_k^- C')^{-1}$$

which is exactly the Kalman Gain (see (3.18))! So from (3.20), where a is the mean m_k of $p(x_k|u^{k-1}, y^k)$ and $m = m_k^-$, we finally have:

$$m_k = m_k^- + K_k \underbrace{(y_k - Cm_k^-)}_{innovation}, \tag{3.21}$$

The term denoted as *innovation* in (3.21) plays a crucial role: it provides a *correction* to the predicted value m_k^- according to the difference between the actual measurement y_k and the expected measurement Cm_k^-. The optimal weight to assign to the correction w.r.t. the predicted value m_k^- is given by the Kalman Gain. In conclusion, remembering that the estimation provided by the Kalman Filter is the mean of the distributions, i.e. $\hat{x}_k^- = m_k^-$ and $\hat{x}_k = m_k$, the equations (3.7), (3.8), (3.18), (3.21) and (3.19) define the Kalman Filter and are summarized for convenience in the following algorithm.

Algorithm 1 The Kalman Filter (KF)

Assume $x_0 \sim \mathcal{N}(m_0, P_0)$ and initialize the filter by $\hat{x}_0 = m_0$. At each time $k \geq 1$ we have the following recursive equations:

$$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_{k-1} \tag{3.22}$$

$$P_k^- = AP_{k-1}A' + Q_{k-1} \tag{3.23}$$

$$K_k = P_k^- C'(CP_k^- C' + R_k)^{-1} \tag{3.24}$$

$$\hat{x}_k = \hat{x}_k^- + K_k(y_k - Cm_k^-) \tag{3.25}$$

$$P_k = [I - K_k C] P_k^- \tag{3.26}$$

3.2 Observations

These observations are taken from the textbook "Dalla Mora, Germani, Manes: Introduzione alla teoria dell'identificazione dei sistemi".

- As mentioned, the Kalman filter provides the LSQ and the MAP estimation of the state at all times k and also, through the covariance matrix, the quality of the estimates.
- The Kalman Gain, which involves the main burden of computations, can be computed off line, being independent of the measurements.
- The Kalman filter is optimal at all times k if it starts with the correct estimate and the correct covariance matrix. However, as discussed shortly below, under proper conditions, the KF is asymptotically optimal under any initialization.
- Even if the system is stationary (i.e. A, B, C and D are constant and $R_k = R$, $Q_k = Q$) the KF is not stationary (K_k remains time dependent). However, under proper conditions, the covariance matrices P_k and the Kalman gain K_k converge to a steady state value $\bar{P}$ and $\bar{K}$ and the (stationary) KF (i.e. the one which uses the stationary Kalman gain) is asymptotically optimal.

Consider now a stationary system. Using (3.16), (3.23), (3.24) and (3.26), we have:

$$P_{k+1} = (I - K_k)(AP_kA' + Q) = [\underbrace{I}_{A} \underbrace{-(AP_kA' + Q)C'}_{B} \underbrace{(C(AP_kA' + Q)C' + R)^{-1}}_{C} \underbrace{C}_{D}](AP_kA' + Q)$$

$$= \left[I + (AP_kA' + Q)C'(-C(AP_kA' + Q)C' + C(AP_kA' + Q)C' + R)^{-1}C\right]^{-1}(AP_kA' + Q)$$

$$= \left[I + (AP_kA' + Q)C'R^{-1}C\right]^{-1}(AP_kA' + Q)$$

Assuming P_k reaches a steady state value $\bar{P}$, the previous equation becomes (at steady state):

$$\bar{P} = \left[I + (A\bar{P}A' + Q)C'R^{-1}C\right]^{-1}(A\bar{P}A' + Q) \tag{3.27}$$

which is known as the Algebraic Riccati Equation (ARE).

Now, if the couple (A, C) is detectable, then, for any positive semidefinite P_0, $P_k \to \bar{P}$, which is a finite positive semidefinite matrix solution to the ARE (3.27) ($\bar{P}$ is in general not unique and depends on P_0).

Let $\bar{P}$ be a positive semidefinite matrix solution to the ARE (3.27). If (A, Q) is stabilizable, $\bar{P}$ is the unique positive semidefinite solution to the ARE (3.27) and is such that the dynamic matrix of the Kalman filter has all the eigenvalues with modulus less than one.

The following theorem summarizes these properties (see the cited text for its proof).

Theorem 2 *If the couple (A, C) is detectable and the couple (A, Q) is stabilizable then:*

1. *$\exists!$ $\bar{P} \geq 0$ solution of the ARE (3.27) which is such that:*

$$\lim_{k\to\infty} P_k = \bar{P} \text{ for all } P_0$$

2. *The KF is stable, that is the dynamic matrix of the filter has all the eigenvalues with modulus less than one (and so the filter is independent at steady state of the initial estimate $\hat{x}_0$).*
3. *The KF with the steady state Kalman Gain is asymptotically optimal.*