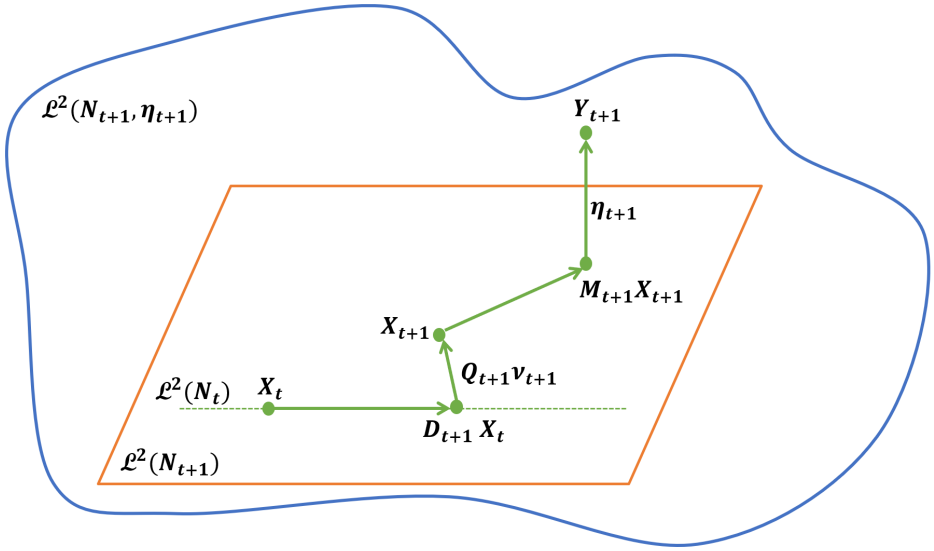
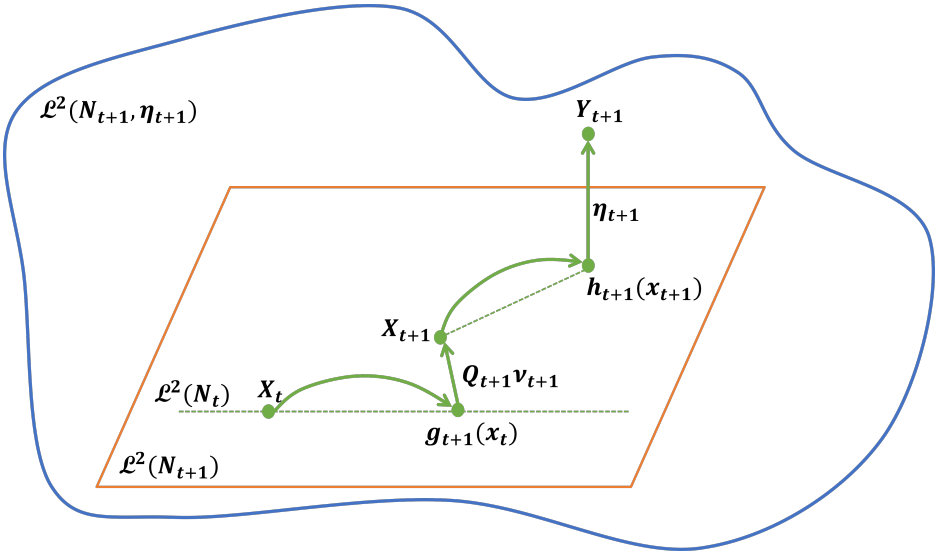
A tutorial on Kalman filtering.
The linear, extended and unscented filters.

Fernando Sansò, Chiara Pileggi, Ludovico Biagi

July 10, 2024





A tutorial on Kalman filtering. The linear, extended and unscented filters

Fernando Sansò:
GReD, Emeritus Professor of Politecnico di Milano; fernando.sanso060545@gmail.com

Chiara Pileggi:
Politecnico di milano, DICA; chiara.pileggi@polimi.it

Ludovico Biagi:
Politecnico di Milano, DICA; ludovico.biagi@polimi.it

GReD - c/o ComoNExT, via Cavour 2, 22074 Lomazzo (CO) Italy
https://www.g-red.eu/

ISBN: 9791221031799

Introduction

Kalman filtering is a tool to predict the time evolution of the state of a system, based on equations of the dynamics of the system and equations of observation describing the relation between the state vector X and observable quantities. In this tutorial the time in Kalman Filter models is discrete, with regular intervals, and all unknowns are finite dimensional vectors. This makes its mathematical basis much simpler than that required by continuous time models with functions as unknowns, i.e. systems with infinite degrees of freedom. When the dynamics of the system and the observation equations are linear the main problem of the filter, namely finding the best prediction of the state at time t based on the known observations up to the same time t, has an exact solution and, when appropriate hypotheses are set up on the noises of the system and of the observations, it can be fastly computed so that we can get the solution in Real Time. This is important, specially in navigation applications. When equations are non-linear though, an exact solution of the filter is typically missing, and one has to resort to approximations. A linearization approach leads to the so-called *Extended Kalman Filter* and has wide applications, yet sometimes the prediction process fails, diverging from the true solution. In addition it is known that the covariance of the prediction error can be significantly underestimated, so new methods have been studied to find approximate solutions that must have two characteristics: they must be better than linearized solutions, to avoid the above drawbacks, and they must be swiftly computable in order to provide results in almost real time.

For instance, an obvious approximation can be built making use of Montecarlo methods; this however requires a large number of samples extracted from prior distributions. This method is therefore hardly applicable to real time problems. In the last three decades yet a new method has advanced, called the *Unscented Kalman Filter*, that, though resembling the Montecarlo Method, in reality requires a considerably small number of computations, making it apt to solve real time problems. This tutorial strives to present the linear, extended and unscented filters with the due mathematical rigour, having in mind an application to kinematic positioning of a cell phone, combining terrestrial observations with GNSS positioning, which will be object of a final numerical example in the text.

1 An introduction to the stochastic approximation of random variables

Purpose of the section is to recall the solution of the following problem: given a vector random variable (X, Y), to find the "best" (in a suitable sense) approximation of X by some function of Y.

Results are mostly recalled without proofs, for which see [15, 13].

Definition 1. Our basic probability sample space is $R^n \otimes R^m$ with a given probability distribution,

$$(X, Y) \in R^n \otimes R^m \ , \quad P(dx, dy)$$

Hypothesis. We assume that $P(dx, dy)$ has finite second order statistics, i.e. there exist finite

$$\begin{vmatrix} \mu_X \\ \mu_Y \end{vmatrix} = E\left\{\begin{vmatrix} X \\ Y \end{vmatrix}\right\} \tag{1}$$

$$C_{\begin{vmatrix} X \\ Y \end{vmatrix}} = \begin{vmatrix} C_X & C_{XY} \\ C_{YX} & C_Y \end{vmatrix} = E\left\{\begin{vmatrix} X - \mu_X \\ Y - \mu_y \end{vmatrix} \left| X^T - \mu_X^T \ Y^T - \mu_Y^T \right|\right\} \tag{2}$$

Definition 2. $\mathcal{L}^2(X, Y)$ is the space of functions $g(X, Y)$ such that

$$E\{g^2(X, Y)\} < +\infty \ . \tag{3}$$

Remark 1. Since

$$\begin{aligned} E\{g^2(X, Y)\} &= \mu_g^2 + E\{(g(X, Y) - \mu_g)^2\} \\ \mu_g &= E\{g(X, Y)\} = \int_{R^n \otimes R^m} g(x, y) P(dx, dy) \end{aligned} \tag{4}$$

(3) is equivalent to: μ_g and $\sigma(g)$ are finite,

$$|\mu_g| < +\infty \ , \ \sigma^2(g) = E\{(g(x, y) - \mu_g)^2\} < +\infty \ . \tag{5}$$

Proposition 1. *$\mathcal{L}^2(X,Y)$ is a Hilbert space with scalar product*

$$g,h \in \mathcal{L}^2(X,Y)\ ,\ < g,h >= E\{g(X,Y)h(X;Y)\} \equiv \mu_g\mu_h + C_{gh}\ ; \tag{6}$$

namely $\mathcal{L}^2(X,Y)$ is complete under the norm

$$\parallel g \parallel^2_{\mathcal{L}^2} =< g,g >= E\{g^2(X,Y)\}\ . \tag{7}$$

Remark 2. The subspace of $\mathcal{L}^2(X,Y)$,

$$\mathcal{L}^2_0(X,Y) = \{g \in \mathcal{L}^2\ ;\ \mu_g = 0\} \tag{8}$$

is a closed subspace of $\mathcal{L}^2(X,Y)$, i.e. it is a Hilbert space itself. On $\mathcal{L}^2_0(X,Y)$ we have

$$\begin{cases} g \in \mathcal{L}^2_0(X,Y)\ ; \parallel g \parallel^2 = \sigma^2(g) = \int g^2 P(dx,dy) \\ g,h \in \mathcal{L}^2_0(X,Y)\ ; < g,h >= C_{gh} = \int ghP(dx,dy)\ . \end{cases} \tag{9}$$

Definition 3. The marginal distributions of (X) and (Y) are given by

$$\begin{array}{l} P(dx) \equiv \int_{(y)} P(dx\ dy) \\ P(dy) \equiv \int_{(x)} P(dx\ dy) \end{array} \tag{10}$$

Remark 3. We have clearly (by Fubini's theorem)

$$\begin{array}{l} E\{g(X)\} = \int_{(x)} g(x)P(dx) \equiv \int_{(x,y)} g(x)P(dx\ dy) \\ E\{h(Y)\} = \int_{(y)} h(y)P(dy) = \int_{(x,y)} h(y)P(dx\ dy)\ . \end{array} \tag{11}$$

Definition 4. We define $\mathcal{L}^2(X), \mathcal{L}^2(Y)$ as

$$\mathcal{L}^2(X) \equiv \{g(X) \in \mathcal{L}^2(X,Y)\} \tag{12}$$

$$\mathcal{L}^2(Y) \equiv \{h(Y) \in \mathcal{L}^2(X,Y)\} \tag{13}$$

It is easy to prove that $\mathcal{L}^2(X), \mathcal{L}^2(Y)$ are closed subspaces of $\mathcal{L}^2(X,Y)$. This means, e.g. in the case of $\mathcal{L}^2(X)$, that if $g_n(X) \in \mathcal{L}^2(X)$ and

$$g_n \underset{\mathcal{L}^2(X,Y)}{\longrightarrow} U \in \mathcal{L}^2(X,Y)\ , \tag{14}$$

then necessarily there is a $g(X)$ such that

$$U = g(X) \in \mathcal{L}^2(X,Y)\ . \tag{15}$$

Definition 5. Let $g = g(X,Y) \in \mathcal{L}^2(X,Y)$; we define the conditional mean of g, given Y, as

$$F_g(Y) \equiv E\{g(X,Y)|_Y\} \equiv Pr_{\mathcal{L}^2(Y)}[g(X,Y)] \tag{16}$$

where

$$Pr_{\mathcal{L}^2(Y)} = \text{ orthogonal projection on } \mathcal{L}^2(Y) \tag{17}$$

is the operator of orthogonal (in the sense of $\mathcal{L}^2(X,Y)$) projection on $\mathcal{L}^2(Y)$. Since $\mathcal{L}^2(Y)$ is a closed subspace of $\mathcal{L}^2(X,Y)$, the definition (16) is consistent.

Proposition 2. *The conditional mean enjoys the following properties*

i) $F_g(y)$ is linear in g, i.e.

$$F_{\lambda g_1 + \mu g_2}(Y) = \lambda F_{g_1}(Y) + \mu F_{g_2}(Y) \tag{18}$$

ii) $\forall g \in \mathcal{L}^2(Y)$, i.e. $g = g(Y)$

$$E\{g(Y)|_Y\} = F_g(Y) \equiv g(Y) \tag{19}$$

iii) the operator $g \to F_g$ is non-expansive, namely

$$\| F_g \|_{\mathcal{L}^2(X,Y)} \equiv E\{F_g(Y)^2\} \leq \| g \|^2_{\mathcal{L}^2(X,Y)} = E\{g^2(X,Y)\} \tag{20}$$

with equality iff $g \in \mathcal{L}^2(Y)$,

iv) $g - F_g \perp \mathcal{L}^2(Y)$, i.e.

$$< g - F_g, h > \equiv E\{[g(X,Y) - F_g(Y)]h(Y)\} \equiv 0 \quad \forall h \in \mathcal{L}^2(Y) \ , \tag{21}$$

v)

$$E_Y\{F_g(Y)\} = E_Y\{E\{g(X,Y)|_Y\}\} \equiv E_{(X,Y)}\{g(X,Y)\} \ , \tag{22}$$

vi)

$$F_g(y_0) = \int_{(x)} g(x,y_0) \frac{P(dx, dy(y_0))}{P(dy(y_0))} \tag{23}$$

where the conditional distribution of X, given Y, is

$$\frac{P(dx, dy(y_0))}{P(dy(y_0))} = \lim_{\substack{|\Delta y| \to 0 \\ y_0 \in \Delta y}} \frac{P(dx, \Delta y)}{P(\Delta y)} \tag{24}$$

the limit holding $P(dy)$ almost surely, i.e. for $dy \in S$, with $P(Y \in S) = 1$.

vii) assume X and Y are stochastically independent, i.e. equivalently

$$P(dx, dy) = P(dx)P(dy) \ , \tag{25}$$

then $\forall g(X), h(Y) \in \mathcal{L}^2(X,Y)$ we have

$$E\{g(X)h(Y)\} = E_X\{g(X)\}E_Y\{h(Y)\} \ ; \tag{26}$$

moreover, in the above hypothesis, the relation

$$E\{g(X)|_Y\} = E\{g(X)\} \ , \tag{27}$$

holds almost surely.

Proof. $i), ii), iii), iv)$ are standard properties of orthogonal projections. $v)$ comes on taking $h(Y) \equiv 1$ in (21). $vi)$ comes on writing explicitly (21) in the form

$$\forall h \in \mathcal{L}^2(Y) \ , \ \int_{(x,y)} h(y) \cdot g(x,y) P(dx, dy) \equiv \int_Y h(y) F_g(y) P(dy)$$

implying

$$\int_{(x)} g(x,y) P(dx, dy) = F_g(y) P(dy) \ . \tag{28}$$

When $P(dy) \neq 0$ (28) implies (23), (24).

Finally, question (26) is an obvious consequence of (28) and of the definition of the mean (see (4)). As for (27), we notice that (26) can be rewritten as

$$E\{g(X)h(Y)\} = E_Y\{\mu_g h(Y)\} \tag{29}$$

where $\mu_g = E_X\{g(X)\}$. On the other hand, using (19), we find

$$E\{g(X)h(Y)\} = E_Y\{h(Y)E\{g(X)|_Y\}\} \ ; \tag{30}$$

so we can write

$$E_Y\{h(Y)E\{g(X)|_Y\} = E_Y\{\mu_g h(Y)\}\} \tag{31}$$

that has to hold for every $h(Y) \in \mathcal{L}^2(Y)$.

Therefore (31) implies (27). □

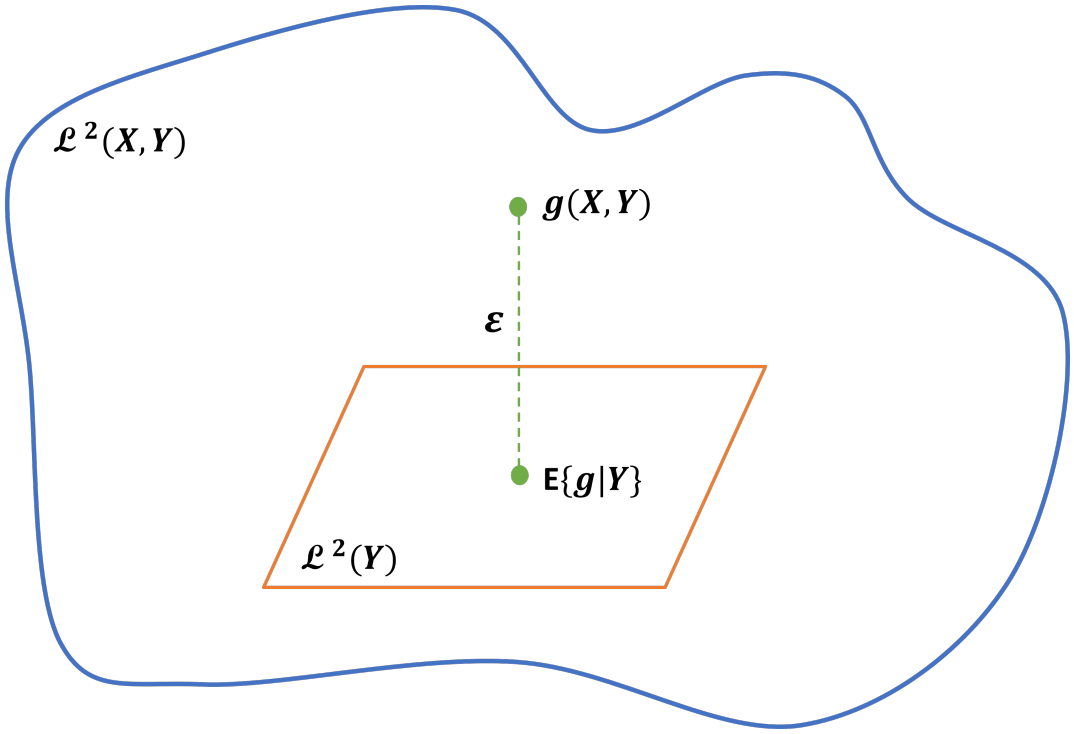


Figure 1: Geometry of Hilbert approximation of $g(X,Y)$ by projection of $\mathcal{L}^2(X,Y)$ onto $\mathcal{L}^2(Y)$.

Proposition 3. *$F_g(y) = E\{g(X,Y)|_Y\}$ is the "best" approximation of $g(X,Y)$ with a function of Y only, in the sense of minimising the distance between $g(X,Y)$ and $h(Y) \in \mathcal{L}^2(Y)$,*

$$F_g = \arg \min_{h \in \mathcal{L}^2(Y)} \| g(X,Y) - h(Y) \|^2_{\mathcal{L}^2(X,Y)} \ , \tag{32}$$

i.e. the variance of the "prediction error" for a general $h(Y)$, namely

$$\varepsilon(X,Y) = g(X,Y) - h(Y) \ . \tag{33}$$

Proof. This is a prominent property of the orthogonal projection on a closed subspace in the Hilbert space $\mathcal{L}^2(X,Y)$. □

Note: $F_g(Y)$ can be therefore considered as the solution of the minimum mean square prediction error (m.m.s.e.) principle (that we call the Wiener–Kolmogorov principle (WK)), applied to $g(X,Y)$. By using the property (22) we see that

$$\begin{aligned} \mu_F &= E_Y\{F_g(Y)\} = E_Y\{E\{g(X,Y)|_Y\}\} = \\ &= E\{g(X,Y)\} = \mu_g \end{aligned} \tag{34}$$

Therefore the prediction error can be expressed as

$$\varepsilon_F = g(X,Y) - F_g(Y) \equiv (g(X,Y) - \mu_g) - (F_g(Y) - \mu_F) \ . \tag{35}$$

Since ε_F is orthogonal to $F_g(Y) - \mu_F \in \mathcal{L}^2(Y)$ (see Fig. 1) we see by the Pythagorean rule that

$$\begin{aligned} \sigma^2(\varepsilon_F) &\equiv \sigma^2(g) - \sigma^2(F_g) \equiv \\ &\equiv E\{[g(X,Y) - \mu_g]^2\} - E\{[F_g(Y) - \mu_F]^2\} \end{aligned} \tag{36}$$

Example 1 (normal distribution). Assume that $P(dx\ dy)$ is normal, namely

$$P(dx\ dy) = G_{m+n}\left(\begin{vmatrix} X \\ Y \end{vmatrix}\ ;\ \begin{vmatrix} \mu_X \\ \mu_Y \end{vmatrix}\ ,\ \begin{vmatrix} C_X & C_{XY} \\ C_{YX} & C_Y \end{vmatrix}\right) dx\ dy \tag{37}$$

where the probability density $G_k(t;\mu_T, C_T)$ is given by

$$\begin{cases} G_k(t;\mu_T.C_T) = \frac{1}{2\pi^{k/2}\sqrt{\det(C_T)}} e^{-\frac{1}{2}Q} \\ Q = (t-\mu_T)^T C_T^{-1}(t-\mu_T)\ . \end{cases} \tag{38}$$

On using the formula of the blockwise inverse of $C\begin{vmatrix} X \\ Y \end{vmatrix}$,

$$\begin{vmatrix} C_X & C_{XY} \\ C_{YX} & C_Y \end{vmatrix}^{-1} = \begin{vmatrix} \Gamma^{-1} & -\Gamma^{-1}C_{XY}C_Y^{-1} \\ -C_Y^{-1}C_{YX}\Gamma^{-1} & C_Y^{-1} + C_Y^{-1}C_{YX}\Gamma^{-1}C_{XY}C_Y^{-1} \end{vmatrix} \tag{39}$$
$$(\Gamma = C_X - C_{XY}C_Y^{-1}C_{YX})\ ,$$

the quadratic form of the normal (38) can be decomposed as

$$\begin{cases} Q = (x-\mu_X|_Y)^T\Gamma^{-1}(x-\mu_X|_Y) + (y-\mu_Y)^T C_Y^{-1}(y-\mu_Y) \\ \mu_X|_Y = \mu_X + C_{XY}C_Y^{-1}(y-\mu_Y)\ . \end{cases} \tag{40}$$

Moreover, a similar formula for the determinants of a block-partitioned matrix gives

$$\det\begin{vmatrix} C_X & C_{XY} \\ C_{YX} & C_Y \end{vmatrix} = \det(C_Y)\det(\Gamma)\ . \tag{41}$$

Therefore we see that the probability distribution of (X, Y) is partitioned according to

$$P(dx, dy) = G_n(x;\mu_X|_Y, \Gamma)dx G_m(y;\mu_Y, C_Y)dy\ . \tag{42}$$

Since

$$P(dy) = \int_{(x)} P(dx\ dy) \equiv G_m(y;\mu_Y, C_Y)dy \tag{43}$$

from (42), (43) and (23) we see that, taken any $g = g(X)$ we have

$$F_g(Y) = \int_{(x)} g(x)G_n(x;\mu_X|_Y, \Gamma)dx\ , \tag{44}$$

which at once shows that the conditional distribution $P(dx|_Y)$ is given by

$$P(dx|_Y) = G_n(x;\mu_X|_Y, \Gamma)dx\ . \tag{45}$$

This in turn means that

$$F_X(Y) \equiv E\{X|_Y\} = \mu_X|_Y \equiv \mu_X + C_{XY}C_Y^{-1}(Y-\mu_X)\ , \tag{46}$$

namely the conditional mean of X given Y in a normal family is a linear, affine function of Y. In particular (46) is affine because by putting $Y = 0$ into it we get, in general

$$E\{X|_Y = 0\} = \mu_X|_Y = \mu_X - C_{XY}C_Y^{-1}\mu_Y \neq 0\ . \tag{47}$$

Furthermore, the covariance of the prediction error ε,

$$\varepsilon = X - [\mu_X + C_{XY}C_Y^{-1}(Y-\mu_Y)] = X - \mu_X|_Y\ , \tag{48}$$

is read out of (38), (39), (40) as

$$C_\varepsilon = \Gamma = C_X - C_{XY}C_Y^{-1}C_{YX}\ . \tag{49}$$

Remark 4. Strictly speaking, the Definition 5 of conditional mean has been given for a scalar variable g, yet it is clear that the definition can be extended to any vector function $g(X,Y)$ by applying (16) component-wise. In this way the notation $F_X(Y)$, for the vector X, is justified. Moreover, recalling the minimum mean square error (m.m.s.e.) property illustrated in Proposition 3, we could say that $F_X(Y)$ satisfies the WK principle

$$F_X(Y) = \arg \min_{h \in [\mathcal{L}^2(Y)]^n} E\{|X - h(Y)|^2\} \ , \tag{50}$$

where $h(Y)$ is a vector of n components, each of which has to belong to $\mathcal{L}^2(Y)$.

As for the prediction error

$$\varepsilon = X - F_X(Y) \ , \tag{51}$$

we can easily find its dispersion, i.e. the covariance matrix C_ε, by generalizing the orthogonality criterion used in (36).

We find then

$$C_\varepsilon = C_X - C_{F_X} \ . \tag{52}$$

Definition 6. We define the linear space

$$L^2(Y) \equiv \{h^T Y + \ell \ , \ \forall h \in R^m, \ \ell \in \mathcal{R}\} \ . \tag{53}$$

On account of the relation

$$\| \ h^T Y + \ell \ \|^2_{\mathcal{L}^2(X,Y)} = E\{(h^T Y + \ell)^2\} = (\ell + h^T \mu_Y)^2 + h^T C_Y h \tag{54}$$

it is clear that

$$L^2(Y) \subset \mathcal{L}^2(Y) \subset \mathcal{L}^2(X,Y) \ .$$

Furthermore, assuming that C_Y is strictly positive definite, as we do, the $L^2(Y)$ convergence of a sequence $\{h_k^T Y + \ell_k\}$ is equivalent to the ordinary, termwise convergence of $\{h_k^T, \ell_k\}$ in $R^m \otimes R$.

Therefore $L^2(Y)$ is a closed subspace of $\mathcal{L}^2(X,Y)$, so that the orthogonal projection theorem in $\mathcal{L}^2(X,Y)$ holds for this space. A similar definition can be adopted for $L^2(X)$ and even for

$$L^2(X,Y) \equiv \{k^T X + h^T Y + \ell \ ; \ k \in R^n \ , \ h \in R^m \ , \ \ell \in R\} \tag{55}$$

Proposition 4. *Given any $U \in \mathcal{L}^2(X,Y)$ there is only one $U|_Y \in L^2(Y)$, called the* linear regressor *of U on Y satisfying the WK principle, namely*

$$U|_Y = \arg \min_{h^T, \ell} E\{(U - h^T Y - \ell)^2\} \ , \tag{56}$$

given by

$$U|_Y = \mu_U + C_{UY} C_Y^{-1} (Y - \mu_Y) \ , \tag{57}$$

i.e.

$$\ell = \mu_U - C_{UY} C_Y^{-1} \mu_Y \ , \ h = C_Y^{-1} C_{YU} \ ; \tag{58}$$

the mean square (m.s.) prediction error of $U|_Y$ is given by

$$E\{(U - U|_Y)^2\} = \sigma_U^2 - C_{UY} C_Y^{-1} C_{YU} \ . \tag{59}$$

Proof. We compute

$$\begin{aligned} & E\{[U - (h^T Y + \ell)]^2\} = \\ & = E\{[(U - \mu_U) - h^T (Y - \mu_Y) \ - (\ell - \mu_U + h^T \mu_Y)]^2\} = \\ & = |\ell - \mu_U + h^T \mu_Y|^2 + \sigma_U^2 - 2 h^T C_{YU} + h^T C_Y h \ . \end{aligned} \tag{60}$$

Then finding the minimum of (60) with respect to h, ℓ is an ordinary l.s. problem, the solution of which yields (57), (58), (59). □

Proposition 5. *Proposition 4 generalizes to the approximation of any vector $V \in [\mathcal{L}^2(X,Y)]^p$, i.e. a random vector on R^p, with the linear regression formulas*

$$V|_Y = \mu_V + C_{VY}C_Y^{-1}(Y - \mu_Y) \tag{61}$$
$$\varepsilon = V - V|_Y = V - \mu_V - C_{VY}C_Y^{-1}(Y - \mu_Y) \tag{62}$$
$$C_\varepsilon = C_V - C_{VY}C_Y^{-1}C_{YV} \tag{63}$$

Remark 5. In particular, by choosing $V = X$ we find the linear regressor of X on Y

$$X|_Y = \mu_X + C_{XY}C_Y^{-1}(Y - \mu_Y) \tag{64}$$

with prediction error covariance

$$C_\varepsilon = C_X - C_{XY}C_Y^{-1}C_{YX} \ . \tag{65}$$

As we see, not by chance, these formulas are identical to the general regression formulas for normal variates (46), (49). This happens because it is intrinsic to the shape of normal distributions to have a regression function which is linear in Y. However, in the definition of the linear regressor no hypothesis is done on the distribution $P(dx, dy)$.

The large number of applications of (64), (65) is due to the fact that often a normal approximation of $P(dx, dy)$ gives good results. It is clear that a condition for the linear regressor to be "good" is exactly that the general regressor

$$E\{X|_Y\} = F_X(Y) \tag{66}$$

can be reasonably approximated by a linear function in the volume of (X, Y) where most of the probability is concentrated (see Appendix). The theory of Unscented Transforms is designed to improve on this linear approximation by a simple, numerically efficient approach.

Definition 7. As we have done in Remark 2 with respect to $\mathcal{L}^2(X,Y)$, we can define a subspace of $L^2(X,Y)$ (see (55)) as

$$L_0^2(X,Y) \equiv \{k^T(X - \mu_X) + h^T(Y - \mu_Y) \ ; \ k \in R^n, \ h \in R^m\} \tag{67}$$

of linear functions of (X, Y) that have zero average too. It is clear that $L_0^2(X,Y) = \mathcal{L}_0^2(X,Y) \cap L^2(X,Y)$, and therefore it is a closed subspace of both.

Moreover, also in this case we can generalize (67) to any random vector $V \in [L_0^2(X,Y)]^p$, i.e.

$$[L_0^2(X,Y)]^p \equiv \tag{68}$$
$$\equiv \{V = A(X - \mu_X) + B(Y - \mu_Y) \ ; \ A \in \mathcal{R}^p \otimes \mathcal{R}^n, \ B \in R^p \otimes R^m\} \ .$$

In particular, for (X, Y) with 0 average, the regression formula (64) reads

$$X|_Y = C_{XY}C_Y^{-1}Y \ , \tag{69}$$

while the covariance of the corresponding prediction error ε, is always given by (65). Finally we observe that often in mathematical literature $L_0^2(X,Y)$ is also called the span of $(X - \mu_X, Y - \mu_Y)$, i.e. the subspace generated by linear combinations of $X - \mu_X$ and $Y - \mu_Y$.

Remark 6. Assume that Y is split into two vectors $Y = (Y_1, Y_2), Y_1 \in R^{m_1}, Y_2 \in \mathcal{R}^{m_2}, m_1 + m_2 = m$, which are uncorrelated, namely

$$C_{Y_1Y_2} = E\{(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})^T\} = 0 \tag{70}$$

then clearly the covariance C_Y is block-diagonal, hence

$$C_Y^{-1} = \left| \begin{array}{cc} C_{Y_1}^{-1} & 0 \\ 0 & C_{Y_2}^{-1} \end{array} \right| \ ; \tag{71}$$

accordingly we see that the following nice orthogonal decomposition holds

$$\begin{aligned} X-\mu_X|_Y &= C_{XY}C_Y^{-1}(Y-\mu_Y) = C_{XY_1}C_{Y_1}^{-1}(Y_1-\mu_{Y_1}) + \\ &\quad + C_{XY_2}C_{Y_2}^{-1}(Y_2-\mu_{Y_2}) = \\ &= (X-\mu_X)|_{Y_1} + (X-\mu_X)|_{Y_2}\ . \end{aligned} \tag{72}$$

In particular, when X has zero average (72) becomes simply

$$X|_Y = X|_{Y_1} + X|_{Y_2} \tag{73}$$

a formula that will be useful in the sequel.

2 The Kalman Filter: linear theory

Definition 8 (Discrete time Dynamic System). We consider the time evolution of a system described by a state vector $\{\overline{X}_t\} \in R^n$. We consider the time t as a discrete variable

$$t = 0, 1, 2 \dots \tag{74}$$

The evolution law of the System is given by the linear equation

$$\overline{X}_{t+1} = D_{t+1}\overline{X}_t + u_{t+1} + Q_{t+1}\nu_{t+1}\ , \tag{75}$$

where $\{D_t\}$ is a sequence of dynamic matrices, u_{t+1} a sequence of deterministic inputs, $\{\nu_t\}$ a sequence of random noises, white in time, such that

$$E\{\nu_t\} = 0\ ,\ E\{\nu_t\nu_{t'}^T\} = \delta_{tt'}C_{\nu_t}\ . \tag{76}$$

The presence of the matrix Q_{t+1} is necessary in the dynamic model because many times we want to apply the noise only to some specific component of $\overline{X}_t$, so that in general the dimension of ν_t is smaller than that of $\overline{X}_t$. Since (75) is a difference equation, we need to specify initial conditions to identify unambiguously a solution.. To make it simple, we assume that

$$\overline{X}_0 = 0 \ \text{ with } P = 1 \tag{77}$$

so that we can claim that

$$C_{\overline{X}_0} = 0\ . \tag{78}$$

Indeed $\overline{X}_0$ could be any other constant vector, with (78) being always valid.

Definition 9 (Observations model). We assume that at each discrete time t we are able to perform an "observation" on the system, which is described by a linear vector function of $\{\overline{X}_t\}$ plus some observation noise, i.e.

$$\overline{Y}_t \in R^m\ ,\ \overline{Y}_t = M_t\overline{X}_t + \eta_t\ ,\ (t = 1, 2\dots) \tag{79}$$

where M_t is a constant (i.e. not random) matrix.

We assume that η_t is a noise, white in time, i.e.

$$E\{\eta_t\} = 0\ ,\ E\{\eta_t\eta_{t'}^T\} = \delta_{tt'}C_{\eta_t} \tag{80}$$

with C_{η_t} strictly positive. Indeed there is no need to add to the vector of observables $\overline{Y}_t$ a component that would be fully linearly dependent on the others. Moreover, we make the hypothesis that the observation noise $\{\eta_t\}$ is stochastically independent of the system noise $\{\nu_t\}$; this implies

$$t', t'' = 1 \dots t+1\ ,\ C_{\eta_{t'}\nu_{t''}} = E\{\eta_{t'}\nu_{t''}^T\} = 0 \tag{81}$$

Let us notice that we do not need to add a constant vector to the model (79) because this can always be included it into the definition of $\overline{Y}_t$.

The problem (first formulation)

Given the system (75), the initial condition (77) and the information coming from the observation vectors $\{\overline{Y}_{t'}\ ;\ t' = 1, 2 \dots t+1\}$, assuming that $\{\nu_{t'}\}, \{\eta_{t'}\}$ satisfy (76), (80), (81), to find the prediction of the sequence $\{\overline{X}_{t'}\ ;\ t' = 1 \dots t+1\}$ optimal in the WK sense.

This is what one of the authors has called Geodetic Navigation (see [16]) and can be solved equivalently either by a batch least squares algorithm, or by a two-stage algorithm, one flowing from the time $t' = 1$ to $t' = t+1$, called the Kalman Filter, the other flowing back, called the Kalman Smoother (see [7]).

Here we are interested in the Kalman Filter part of the process, because this is particularly suited to the realtime navigation problem, where the question is to get the best estimate of X_t for any time t, given all the observations up to this instant, without waiting for information coming in the future. As we are going to see, this problem has a nice sequential solution, easy to be numerically implemented.

The reduced problem (second formulation, the Kalman Filter)

We are looking for the best linear predictor of $\overline{X}_{t'}$ given $\overline{I}_{t'}^T = [\overline{Y}_1^T \dots \overline{Y}_{t'}^T], (t' = 1 \dots t+1)$ and the associated covariance matrix of the prediction error.

In abstract the solution of this problem is given by formula (64), (65), nevertheless it is convenient to reduce them to a recursive scheme which is the essence of the Kalman Filter.

Before doing that, it is useful to separate in all random variables the part of the mean from the stochastic residual, namely we put

$$\overline{X}_t = \mu_{\overline{X}_t} + X_t\ ,\ E\{X_t\} = 0 \tag{82}$$

$$\overline{Y}_t = \mu_{\overline{Y}_t} + Y_t\ ,\ E\{Y_t\} = 0 \tag{83}$$

Taking the average of (75), (79), we find

$$\mu_{\overline{X}_{t+1}} = D_{t+1}\mu_{\overline{X}_t} + u_{t+1} \tag{84}$$

$$\mu_{\overline{Y}_{t+1}} = M_{t+1}\mu_{\overline{X}_{t+1}} \tag{85}$$

It follows that the recursive equations (84), (85) can determine $\mu_{\overline{X}_{t+1}}$ at any time, starting from the known $\mu_{\overline{X}_0} = 0$, and subsequently $\mu_{\overline{Y}_0}$.

It is interesting to write explicitly the solution of (84). We introduce the "propagator" matrix from time τ to time $t+1$ as

$$\Phi_{t+1,\tau} = \prod_{k=\tau+1}^{t+1} D_k\ ,\ \Phi_{t+1,t+1} = I \tag{86}$$

and we find that

$$\mu_{\overline{X}_{t+1}} = \sum_{\tau=1}^{t+1} \Phi_{t+1,\tau} u_\tau\ , \tag{87}$$

as one can easily verify by substituting in (84).

In particular, if we use (87) with $t = 0, \tau = 1$ and consider that $\Phi_{11} = I$, we get

$$\mu_{\overline{X}_1} = u_1 \tag{88}$$

as it is already obvious from (84), because $\mu_{\overline{X}_0} = 0$. Due to the linearity of our equations the stochastic part of the system and of the observations, run according to the laws

$$X_{t+1} = D_{t+1}X_t + Q_{t+1}\nu_{t+1}\ , \tag{89}$$

$$Y_{t+1} = M_{t+1}X_{t+1} + \eta_{t+1}\ . \tag{90}$$

We underline that, since both $\overline{Y}_{t'}$ and $\mu_{\overline{Y}_{t'}}$ are known, the same holds for

$$Y_{t'} = \overline{Y}_{t'} - \mu_{\overline{Y}_{t'}}\ ,\ t' = 1 \dots t+1\ . \tag{91}$$

Remark 7. By applying (86), (87) to (89) we see that

$$X_{t+1} = \sum_{\tau=1}^{t+1} \Phi_{t+1,\tau} Q_\tau \nu_\tau \ : \tag{92}$$

then it is clear that, if we put

$$N_{t'}^T \equiv [\nu_1^T \nu_2^T \dots \nu_{t'}^T] \ , \tag{93}$$

we have

$$X_{t'} \in [L_0^2(N_{t'})]^n \ \ t' = 1 \dots t+1 \tag{94}$$

and in particular, due to (76),

$$X_{t'} = \sum_{\tau=1}^{t'} \Phi_{t',\tau} Q_\tau \nu_\tau \tag{95}$$

is an orthogonal decomposition of $X_{t'}$ on the orthogonal basis of $L_0^2(N_{t+1})$ given by the components of the noise N_{t+1}.

An easy consequence of (94) is that no $X_{t'}$ can depend on future noises $\nu_t (t > t')$, i.e. $X_{t'}$ and ν_t are uncorrelated or orthogonal,

$$E\{X_{t'} \nu_t^T\} = 0 \ , \ t > t' \ . \tag{96}$$

On the other hand, the $\eta_{t'}$ are orthogonal to one another(uncorrelated) and also orthogonal to $[L_0^2(N_{t+1})]^n$ due to (81). So (90) is an orthogonal decomposition of Y_{t+1} and, as we see,

$$Y_{t'} \in [L_0^2(N_{t'}, \eta_{t'})]^m \ . \tag{97}$$

To continue it is useful to define a number of new random variables

Definition 10. We define

i) $I_{t+1}^T \equiv [Y_i^T \dots Y_{t+1}^T]$
ii) $\widehat{X}_{t'} = X_{t'}|_{I_{t'}}$
iii) $\widehat{\varepsilon}_{t'} = X_{t'} - \widehat{X}_{t'}$, (prediction error of $\widehat{X}_{t'}$)
iv) $\widetilde{X}_{t'+1} = X_{t'+1}|_{I_{t'}}$
v) $\widetilde{Y}_{t'+1} = Y_{t'+1}|_{I_{t'}}$
vi) $\widetilde{\varepsilon}_{t'+1} = X_{t'+1} - \widetilde{X}_{t'+1}$(prediction error of $\widetilde{X}_{t'+1}$)
vii) $\delta Y_{t+1} = Y_{t+1} - Y_{t+1}|_{I_t} \equiv Y_{t+1} - \widetilde{Y}_{t+1}$

Remark 8. Let us recall that given two zero mean vectors X, Y, $X|_Y$ is just the orthogonal projection, componentwise, of X on $L_0^2(Y)$ so that $X|_Y \in [L_0^2(Y)]^n$ and its prediction error $\varepsilon = X - X|_Y$ is orthogonal (uncorrelated) to $[L_0^2(Y)]^n$. Therefore, from the above definition we can deduce that

$$\widehat{X}_{t'} \ , \ \widetilde{X}_{t'+1} \in [L_0^2(I_{t'})]^n \tag{98}$$
$$\widehat{\varepsilon}_{t'} \ , \ \widetilde{\varepsilon}_{t'+1} \perp [L_0^2(I_{t'})]^n \ ; \tag{99}$$

more precisely

$$\widehat{\varepsilon}_{t'} \in [L_0(N_{t'}, I_{t'})]^n \tag{100}$$
$$\widetilde{\varepsilon}_{t'+1} \in [L_0(N_{t'+1}, I_{t'})]^n \ . \tag{101}$$

Furthermore, we have

$$\delta Y_{t+1} \perp [L_0^2(I_t)]^m \ . \tag{102}$$

Finally, by (102) we can claim that

$$L_0^2(I_{t+1}) = L_0^2(I_t, Y_{t+1}) = L_0^2(I_t, \delta Y_t) \ . \tag{103}$$

Proposition 6 (Sequential solution of the Kalman Filter)**.** *Our target is to establish a sequential algorithm to compute* $\widehat{X}_{t+1}$ *and* $C_{\widehat{\varepsilon}_{t+1}}$.

The solution is given by formulas where, from $\widehat{X}_t, C_{\widehat{\varepsilon}_t}$ *one can derive* $\widehat{X}_{t+1}, C_{\widehat{\varepsilon}_{t+1}}$ *and the initial formulas of the sequence, giving* $\widehat{X}_1, C_{\widehat{\varepsilon}_1}$*:*

- $$\widehat{X}_{t_{t+1}} = \widetilde{X}_{t+1} + K_{t+1}M_{t+1}^T(M_{t+1}K_{t+1}M_{t+1}^T + C_{\eta_{t+1}})^{-1}\delta Y_{t+1} \ ; \tag{104}$$

where

$$\widetilde{X}_{t+1} = D_{t+1}\widehat{X}_t \tag{105}$$
$$\delta Y_{t+1} = Y_{t+1} - \widetilde{Y}_{t+1} = Y_{t+1} - M_{t+1}\widetilde{X}_{t+1} \tag{106}$$
$$K_{t+1} = D_{t+1}C_{\widehat{\varepsilon}_t}D_{t+1}^T + Q_{t+1}C_{\nu_{t+1}}Q_{t+1}^T \ , \tag{107}$$

and

- $$C_{\widehat{\varepsilon}_{t+1}} = K_{t+1} - K_{t+1}M_{t+1}^T C_{\delta Y_{t+1}}^{-1} M_{t+1}K_{t+1} \tag{108}$$

where

$$C_{\delta Y_{t+1}} = M_{t+1}K_{t+1}M_{t+1}^T + C_{\eta_{t+1}} \tag{109}$$

The initialization of the algorithm is given by

- $$\widehat{X}_1 = K_1 M_1^T C_{\eta_1}^{-1}\delta Y_1 \tag{110}$$

- $$C_{\widehat{\varepsilon}_1} = K_1 - K_1 M_i^T C_{Y_1}^{-1} M_1 K_1 \tag{111}$$

where $K_1 = Q_1 C_{\nu_1} Q_1^T$ *and* $C_{Y_1} = M_1 K_1 M_1^T + C_{\eta_1}$.

Proof. Before starting the proof, let us recall that all vectors here have zero mean, so that the concepts of zero correlation and orthogonality do coincide. We know that

$$\widehat{X}_{t+1} = X_{t+1}|_{I_{t+1}} \equiv X_{t+1}|_{I_t, \delta Y_{t+1}} \ .$$

Recalling that I_t and δY_{t+1} are orthogonal (see (102)) and that $X|_Y$ is the orthogonal projection of X on $L_0^2(Y)$, we can write

$$\begin{aligned} \widehat{X}_{t+1} &= X_{t+1}|_{I_t} + X_{t+1}|_{\delta Y_{t+1}} = \\ &= \widetilde{X}_{t+1} + L\delta Y_{t+1} \ , \end{aligned} \tag{112}$$

where, according to (69),

$$L = C_{X_{t+1}\delta Y_{t+1}} C_{\delta Y_{t+1}}^{-1} \ , \tag{113}$$

From the dynamic equation (89) we find

$$\widetilde{X}_{t+1} = X_{t+1}|_{I_t} = D_{t+1}X_t|_{I_t} + Q_{t+1}\nu_{t+1}|_{I_t} \tag{114}$$

On the other hand, ν_{t+1} is orthogonal to all $\eta_{t'}$ and to all past $X_{t'}$ and therefore to the whole I_t, so $\nu_{t+1}|_{I_t} \equiv 0$. Taking into account the definition (98), we see that (114) reads

$$\widetilde{X}_{t+1} = D_{t+1}\widehat{X}_t \ . \tag{115}$$

So (105) is proved and since $\widehat{X}_t$ is assumed to be known $\widetilde{X}_{t+1}$ is known too.

From (90) we find

$$\widetilde{Y}_{t+1} = Y_{t+1}|_{I_t} = M_{t+1}X_{t+1}|_{I_t} + \eta_{t+1}|_{I_t} \ . \tag{116}$$

But

$$X_{t+1}|_{I_t} = \widetilde{X}_{t+1} \in [L^2_0(I_t)]^m \ ; \tag{117}$$

moreover, η_{t+1} is orthogonal to all $Y_{t'}, t' \leq t$, by (97) and so it is orthogonal to the whole I_t, which yields

$$\eta_{t+1}|_{I_t} = 0 \ . \tag{118}$$

Therefore (115) reads

$$\widetilde{Y}_{t+1} = M_{t+1}\widetilde{X}_{t+1} \tag{119}$$

what proves (106). Since $\widetilde{X}_{t+1}$ is known, so are $\widetilde{Y}_{t+1}$ and $\delta Y_{t+1} = Y_{t+1} - \widetilde{Y}_{t+1}$.

Now, let us write

$$\begin{aligned} \delta Y_{t+1} &= Y_{t+1} - \widetilde{Y}_{t+1} = M_{t+1}(X_{t+1} - \widetilde{X}_{t+1}) + \eta_{t+1} = \\ &= M_{t+1}\widetilde{\varepsilon}_{t+1} + \eta_{t+1} \ . \end{aligned} \tag{120}$$

On the other hand, recalling (89) and (115),

$$\begin{aligned} \widetilde{\varepsilon}_{t+1} &\equiv D_{t+1}(X_t - \widehat{X}_t) + Q_{t+1}\nu_{t+1} = \\ &= D_{t+1}\widehat{\varepsilon}_t + Q_{t+1}\nu_{t+1} \ . \end{aligned} \tag{121}$$

From (100) we know that $\widehat{\varepsilon}_t$ is a linear function of N_t and I_t, so that (121) is an orthogonal decomposition because ν_{t+1} is orthogonal to both; therefore, according to the definition (107),

$$C_{\widetilde{\varepsilon}_{t+1}} = D_{t+1}C_{\widehat{\varepsilon}_t}D^T_{t+1} + Q_{t+1}C_{\nu_{t+1}}Q^T_{t+1} = K_{t+1} \ . \tag{122}$$

Since we assume to know $C_{\widehat{\varepsilon}_t}, K_{t+1}$ is a known matrix too. Moreover, since η_{t+1} is orthogonal to ν_{t+1} and to $\widehat{\varepsilon}_t$, also (120) is an orthogonal decomposition and we find

$$C_{\delta Y_{t+1}} = M_{t+1}K_{t+1}M^T_{t+1} + C_{\eta_{t+1}} \ . \tag{123}$$

Further on, using (120) and recalling that η_t is orthogonal to all $X_{t'} \in N_{t'}$, we get

$$C_{X_{t+1}\delta Y_{t+1}} = C_{X_{t+1}\widetilde{\varepsilon}_{t+1}}M^T_{t+1} \ . \tag{124}$$

But

$$X_{t+1} = \widetilde{X}_{t+1} + \widetilde{\varepsilon}_{t+1} \tag{125}$$

is by definition an orthogonal decomposition, so that $C_{\widetilde{X}_{t+1}\widetilde{\varepsilon}_{t+1}} = 0$ and

$$C_{X_{t+1}\widetilde{\varepsilon}_{t+1}} = C_{\widetilde{\varepsilon}_{t+1}} = K_{t+1} \tag{126}$$

Putting together (126), (124) and (123) into (113) we find

$$L_{t+1} = K_{t+1}M^T_{t+1}(M_{t+1}K_{t+1}M^T_{t+1} + C_{\eta_{t+1}})^{-1} \tag{127}$$

that inserted into (112) gives (104).

Next we use (104) in the definitions of $\widehat{\varepsilon}_{t+1}, \widetilde{\varepsilon}_{t+1}$ to see that

$$\widehat{\varepsilon}_{t+1} = X_{t+1} - \widehat{X}_{t+1} = \widetilde{\varepsilon}_{t+1} - L_{t+1}\delta Y_{t+1} \tag{128}$$

with L_{t+1} as in (127).

On the other hand, $\widehat{\varepsilon}_{t+1}$ is orthogonal to $(I_t, \delta Y_t)$, so rewriting (128) as

$$\begin{aligned} \widetilde{\varepsilon}_{t+1} &= \widehat{\varepsilon}_{t+1} + L\delta Y_{t+1} = \\ &= \widehat{\varepsilon}_{t+1} + K_{t+1}M^T_{t+1}C^{-1}_{\delta Y_{t+1}}\delta Y_{t+1} \end{aligned} \tag{129}$$

we arrive at the relation

$$C_{\widetilde{\varepsilon}_{t+1}} = C_{\widehat{\varepsilon}_{t+1}} + K_{t+1} M_{t+1}^T C_{\delta Y_{t+1}}^{-1} M_{t+1} K_{t+1} \ . \tag{130}$$

Inverting (130) and recalling (122) we get

$$C_{\widehat{\varepsilon}_{t+1}} = K_{t+1} - K_{t+1} M_{t+1}^T C_{\delta Y_{t+1}}^{-1} M_{t+1} K_{t+1} \tag{131}$$

and (108) is proved.

Finally (110) and (111) are nothing but (104), (108) when we take into account that

$$X_0 \equiv 0 \ , \ X_1 = Q_1 \nu_1 \ , \ \widetilde{X}_1 = 0 \ , \ \widetilde{Y}_1 = 0 \ , \ \delta Y_1 \equiv Y_1 \ .$$

The proof of Proposition 6 is complete. □

At the level of names we observe that often in Kalman Filter literature equations (105), (107) are called the *prediction part*, while (104), (108) are called the *update part* of the filter.

Remark 9. Since it is well-known that the optimal prediction of the random variable X in a linear model,

$$Y = AX + \nu \tag{132}$$

with the addition, in a Bayesian fashion, of some prior information on X, e.g.

$$X_0 = X + e \tag{133}$$

is equivalent to a least squares solution, where X is now considered a deterministic parameter estimated from (132), (133) for instance assuming to know C_ν and C_e and that ν and e are not correlated, it seems here interesting to explore in the present context the equivalence between the two points of view. In our case, in fact, we have an information, coming from the past of the series (X_t, Y_t), which is summarized by $\widetilde{X}_{t+1}$; we can write

$$\widetilde{X}_{t+1} = X_{t+1} - \widetilde{\varepsilon}_{t+1} \tag{134}$$

with

$$C_{\widetilde{\varepsilon}_{t+1}} = K_{t+1} \tag{135}$$

as given in (122).

To this we can add the observation equation

$$Y_{t+1} = M_{t+1} X_{t+1} + \eta_{t+1} \tag{136}$$

with a known $C_{\eta_{t+1}}$. Moreover, we have already clarified that η_{t+1} and $\widetilde{\varepsilon}_{t+1}$ are orthogonal to one another, while discussing equation (120).

The answer to the above question is supplied by the following proposition.

Proposition 7 (the Equivalence Lemma)**.** *The WK predictor of the random variable X_{t+1} given by (104) is the same as the least squares estimator with "observation equations" (134), (136) and covariance*

$$\begin{vmatrix} K_{t+1} & 0 \\ 0 & C_{\eta_{t+1}} \end{vmatrix} \tag{137}$$

so that the solution (104) can be understood in the light of either two interpretations.

Proof. We start by providing the l.s. solution of (134), (136), (137). We notice that the design matrix A is in this case

$$A = \begin{vmatrix} I \\ M_{t+1} \end{vmatrix} \tag{138}$$

so that the normal matrix S is

$$\begin{aligned} S_{t+1} &= [I\ M_{t+1}^T] \begin{bmatrix} K_{t+1}^{-1} & 0 \\ 0 & C_{\eta_{t+1}}^{-1} \end{bmatrix} \begin{bmatrix} I \\ M_{t+1} \end{bmatrix} = \\ &= (K_{t+1}^{-1} + M_{t+1}^T C_{\eta_{t+1}}^{-1} M_{t+1}) \end{aligned} \tag{139}$$

The l.s. solution of the problem is then

$$\widehat{x}_{ls} = S_{t+1}^{-1}(K_{t+1}^{-1}\widetilde{X}_{t+1} + M_{t+1}^T C_{\eta_{t+1}}^{-1} Y_{t+1}) \ . \tag{140}$$

Now we use the matrix identity

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \ , \tag{141}$$

valid for all invertible A, C and for B, D with dimensions suitable to make $A + BCD$ an invertible square matrix.

After introducing the convenient abbreviation (see (123))

$$C_{\eta_{t+1}} + M_{t+1}K_{t+1}M_{t+1}^T \equiv C_{\delta Y_{t+1}} \tag{142}$$

we get from (140),

$$\widehat{X}_{ls} = (K_{t+1} - K_{t+1}M_{t+1}^T C_{\delta Y_{t+1}}^{-1} M_{t+1}K_{t+1})^{-1}(K_{t+1}^{-1}\widetilde{X}_{t+1} + M_{t+1}^T C_{\eta_{t+1}}^{-1}) \ . \tag{143}$$

After multiplying the parentheses and using the identity (see (142))

$$C_{\delta Y_{t+1}}^{-1} M_{t+1}K_{t+1}M_{t+1}^T = I - C_{\delta Y_{t+1}}^{-1} C_{\eta_{t+1}} \ , \tag{144}$$

we find

$$\begin{aligned} \widehat{X}_{ls} &= \widetilde{X}_{t+1} - K_{t+1}M_{t+1}^T C_{\delta Y_{t+1}}^{-1} M_{t+1}\widetilde{X}_{t+1} + \\ &\quad (K_{t+1}M_{t+1}^T C_{\eta_{t+1}}^{-1} - K_{t+1}M_{t+1}^T C_{\eta_{t+1}}^{-1} + K_{t+1}M_{t+1}^T C_{\delta Y_{t+1}}^{-1}) Y_{t+1} \\ &= \widetilde{X}_{t+1} + K_{t+1}M_{t+1}^T C_{\delta Y_{t+1}}^{-1}(Y_{t+1} - M_{t+1}\widetilde{X}_{t+1}) \ . \end{aligned} \tag{145}$$

Since $Y_{t+1} - M_{t+1}\widetilde{X}_{t+1} = Y_{t+1} - \widetilde{Y}_{t+1} = \delta Y_{t+1}$, a quick comparison with (104) shows that in fact $\widehat{X}_{ls} = \widehat{X}_{t+1}$, namely the equivalence is established. A similar reasoning shows that the estimation error covariance matrix of the l.s. estimator

$$C_{\widehat{X}_{ls}} = S_{t+1}^{-1} \tag{146}$$

is equal to the prediction error covariance matrix $C_{\widehat{\varepsilon}_{t+1}}$, given by (108). □

Remark 10 (computability). Although equivalent on an analytical ground the two solutions (104) and (140) do not share the same numerical properties.

In fact (104) and (108) require the solution of systems of dimension m, while (140) and (146) are reduced to the solution of systems of dimension n. There are problems where n (state dimension) is much larger than m (observation vector dimension); for instance this happens in oceanography or in atmospheric dynamics (see [3], [4]). In this case formulas (104), (108) are recommended. In the case of GNSS–INS aided navigation for instance n is typically smaller than m so that the least squares formulas (140), (146) are numerically advantageous.

Remark 11 (non-zero average model). We would like to remove the limitation that we have put on the dynamic system by assuming that $X_0 = 0$ with $P = 1$. Let us first consider the case that we know that the trajectory of states starts from a known constant point $\overline{X}_0 \neq 0$. If we go back to the original dynamic model (75) and to the subsequent discussion it is obvious that the solution of the evolution equation can be written as

$$\overline{X}_t = X_{ht} + U_t + X_t \tag{147}$$

where

$$X_{ht} = \Phi_{t,0}\overline{X}_0, \ U_t = \sum_{\tau_1}^{t} \Phi_{t,\tau} u_\tau \ ; \tag{148}$$

here the suffix h stems for homogeneous. Accordingly, the best predictor $\widehat{\overline{X}}_t$ of X_t on the basis of the information contained in I_t is

$$\widehat{\overline{X}}_t = \Phi_{t,0}\overline{X}_0 + U_t + \widehat{X}_t \ , \tag{149}$$

where $\widehat{X}_t$ is the sequential solution of the Kalman Filter described in Proposition 6. It follows that

$$\widehat{\widetilde{\varepsilon}}_t = \overline{X}_t - \widehat{\overline{X}}_t = X_t - \widehat{X}_t = \widehat{\varepsilon}_t \tag{150}$$

so that the covariance of the prediction error is the same as that of the stochastic part

$$C_{\widehat{\widetilde{\varepsilon}}_t} = C_{\widehat{\varepsilon}_t} \tag{151}$$

given in Proposition 6.

However, it is not common to have a precise knowledge of the initial state vector $\overline{X}_0$, rather we often have a guess for the value of its entries and a guess for how large can be the error of such a knowledge.

We formalize the situation by assuming that we have an *observation equation* at time 0, namely

$$Y_0 = \overline{X}_0 + \eta_0 \ , \tag{152}$$

with

$$E\{\eta_0\} = 0 \ , \ C_{\eta_0} = \sigma_0^2 I \ ; \tag{153}$$

here the choice of the covariance is just dictated by a criterion of simplicity and the value of σ_0^2 should be large enough to accommodate for the vague information we have on $\overline{X}_0$.

With this prior information we can augment the observations vector by including Y_0 in the first position, namely by defining

$$\overline{I}_t = \begin{vmatrix} Y_0 \\ Y_1 \\ \dots \\ Y_t \end{vmatrix} \ . \tag{154}$$

Now we can reconsider the construction of the Kalman Filter starting from the first step, i.e. from

$$\overline{X}_1 = D_1\overline{X}_0 + u_1 + Q_1\nu_1 \ . \tag{155}$$

It is clear that we cannot simply proceed by conditioning (155) to $\overline{I}_0$, i.e. to Y_0, because $\overline{X}_0$, though unknown, is a constant and we would have

$$\overline{X}_1|_{\overline{I}_0} = D_1\overline{X}_0 + u_1 \tag{156}$$

which is of no use, precisely because we don't know $\overline{X}_0$. Since the only information on $\overline{X}_0$ we have i Y_0 itself, a common sense choice here is to call

$$\widetilde{\overline{X}}_1 = D_1 Y_0 + u_1 \tag{157}$$

which we can compute, and subsequently

$$\widetilde{\widetilde{\varepsilon}}_1 = X_1 - \widetilde{\overline{X}}_1 = -D_1\eta_0 + Q_1\nu_1 \ . \tag{158}$$

From here on we follow the logical development of the Kalman Filter, interpreted as a sequential least squares algorithm.

We collect the information we have on $\overline{X}_1$ in the two equations

$$\begin{cases} \widetilde{\overline{X}}_1 = \overline{X}_1 - \widetilde{\widetilde{\varepsilon}}_1 \\ Y_1 = M_1\overline{X}_1 + \eta_1 \end{cases} \tag{159}$$

where $\widetilde{\overline{X}}_1$ is given by (157) and its error $\widetilde{\overline{\varepsilon}}_1$ is given by (158), with covariance

$$C_{\widetilde{\overline{\varepsilon}}_1} = \sigma_0^2 D_1^T + Q_1 C_{\nu_1} Q_1^T = K_1 \ . \tag{160}$$

Since η_1 is independent if η_0 and η_1, and hence on $\widetilde{\overline{\varepsilon}}_1$ too (see (158)), the least squares solution of (159) is given by

$$\widehat{\overline{X}}_1 = (K_1^{-1} + M_1^T C_{\eta_1} M_1)^{-1} (K_1^{-1} \widetilde{\overline{X}}_1 + M_1^T C_{\eta_1}^{-1} Y_1) \tag{161}$$

with

$$\widehat{\overline{\varepsilon}}_1 = \overline{X}_1 - \widehat{\overline{X}_1} \ ; \ C_{\widehat{\overline{\varepsilon}}_1}^{-1} = (K_1^{-1} + M_1^T C_{\eta_1} M_1)^{-1} \ . \tag{162}$$

Running in revers order the proof of Proposition 7 we see that (161), (162) are equivalent to

$$\widehat{\overline{X}}_1 = \widetilde{\overline{X}}_1 + K_1 M_1^T (M_1 K_1 M_1^T + C_{\eta_1})^{-1} (Y_{t+1} - \widetilde{Y}_{t+1}) \ , \tag{163}$$

where we have put

$$\widetilde{Y}_1 = M_1 \widetilde{\overline{X}}_1 \ , \tag{164}$$

and

$$C_{\widehat{\overline{\varepsilon}}_1} = K_1 - K_1 M_1^T (M_1 K_1 M_1^T + C_{\eta_1})^{-1} M_1 K_1 \ . \tag{165}$$

A quick comparison with (104), (108) shows that (163), (164) follow the general form of the previously derived Kalman Filter, with the only difference that in the present case neither $\widehat{\overline{X}}_1$ nor $\widetilde{\overline{X}}_1$ are zero mean variates.

The same consideration can be done for the subsequent steps, so that the formulas of filter continue to be valid, only triggered by the initial solution (163), (165).

Yet it has to be stressed that the symbols $\widehat{\overline{X}}t, \widetilde{\overline{X}}_t$ used in this remark are not exactly the same as those defined in the previous part of this section. In particular, it is not true that

$$\widehat{\overline{X}}_t = \overline{X}_t|_{\overline{I}_t} \ . \tag{166}$$

In fact, in our derivation here, we cannot advocate the orthogonal decomposition (112) because in this case $\overline{I}_t$ and $Y_{t+1} - \widetilde{Y}_{t+1}$ are not anymore orthogonal. One can see that already in the first step, namely that

$$E\{\delta Y_1 Y_0^T\} \neq 0 \ . \tag{167}$$

In fact, recalling (164) and (157),

$$\begin{aligned} &E\{\delta Y_1 Y_0^T\} = E\{[M_1(D_1 X_0 + u_1 + Q_1 \nu_1) - M_1(D_1 Y_0 + u_1) + \\ &+\eta_1][X_0 + \eta_0]^T\} = -M_1 D_1 E\{\eta_0 \eta_0^T\} = -\sigma_0^2 M_1 D_1 \ . \end{aligned} \tag{168}$$

As a matter of fact, to derive optimal estimates one should use a batch least squares solution at any time t or, equivalently, the combined use of Kalman Filter and Smoother. Yet since such solutions are not easily implemented in recursive real-time computation, we will stop here the discussion just observing that the solution here proposed is not really optimal but good enough for our purposes.

Example 2. A trolley is moving along a rectilinear rail with constant velocity v and random fluctuations ν_t. At $t = 0$ the trolley is moving from the origin, $\overline{X}_0 = 0$, position known without error. Every second, $t = 1, 2 \dots T + 1$, a distance meter takes the position of the trolley from the origin, with a white noise measurement η_t. We want to set up a Kalman filter to get at every time $t + 1$ the best estimate of the position, based on all measurements at $t' = 1, 2 \dots t + 1$.

Initial condition:

$$\overline{X}_0 = 0 \ , \ \sigma^2(\overline{X}_0) = 0 \ ;$$

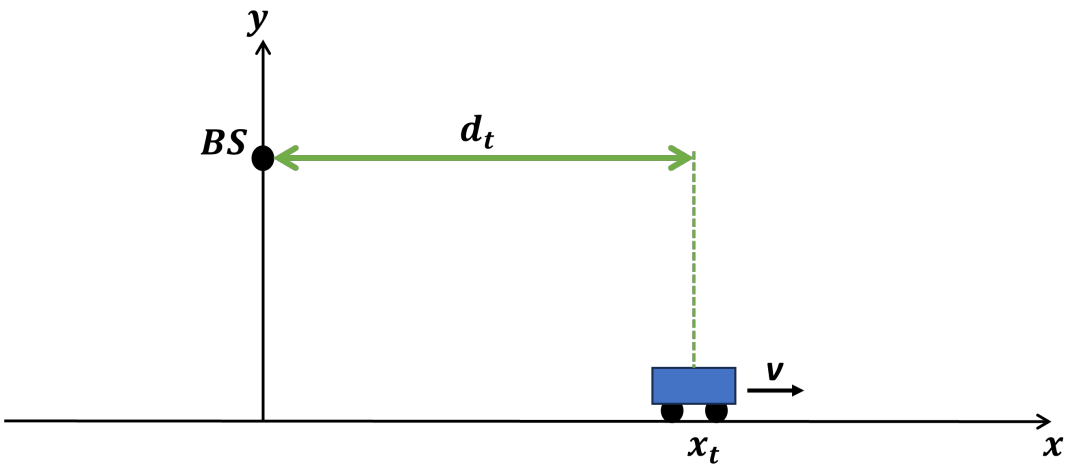


Figure 2: The measurement arrangement of Example 2.

dynamic model:

$$\overline{X}_{t+1} = \overline{X}_t + v + \nu_{t+1} \ , \ \sigma^2_{\nu_t} = \sigma^2_\nu \ ;$$

observational model:

$$\overline{Y}_{t+1} = \overline{X}_{t+1} + \eta_{t+1} \ , \ \sigma^2_{\eta_{t+1}} = \sigma^2_\eta \ .$$

With our notation we have

$$D_{t+1} = 1 \ , \ M_{t+1} = 1 \ .$$

Mean motion

$$\mu_{\overline{X}_{t+1}} = \mu_{\overline{X}_t} + v \rightarrow \mu_{\overline{X}_{t'}} = vt \ , \ t' = 0, 1 \ldots T+1 \ ,$$

mean of the observations

$$\mu_{\overline{Y}_{t+1}} = v(t+1) \ \ t = 0, \ldots T \ .$$

Stochastic part of the motion

$$\overline{X}_t = \mu_{X_t} + X_t \ ; \ X_{t+1} = X_t + \nu_{t+1}$$

and of the observations

$$\overline{Y}_{t+1} = \mu_{\overline{Y}_{t+1}} + Y_{t+1} \ ; \ Y_{t+1} = X_{t+1} + \eta_{t+1} \ .$$

Given

$$X_0, \sigma^2(X_0) = 0 \ , \ \widehat{X}_t, \sigma^2(\widehat{\varepsilon}_t) \ , \ \widehat{\varepsilon}_t = X_t - \widehat{X}_t \ ,$$

we have to find

$$\widehat{X}_{t+1}, \sigma^2(\widehat{\varepsilon}_{t+1}) \ ,$$

where $\widehat{\varepsilon}_{t+1} = X_{t+1} - \widehat{X}_{t+1}$. We have

$$\begin{aligned}
\widetilde{X}_{t+1} &= X_{t+1}|_{I_t} = \widehat{X}_t \ , \\
\widetilde{\varepsilon}_{t+1} &= X_{t+1} - \widetilde{X}_{t+1} = X_{t+1} - \widehat{X}_t = X_t + \nu_{t+1} - \widehat{X}_t = \widehat{\varepsilon}_t + \nu_{t+1} \ , \\
\widetilde{Y}_{t+1} &= \widetilde{X}_{t+1} = \widehat{X}_t \ , \\
\delta Y_{t+1} &= Y_{t+1} - \widetilde{Y}_{t+1} = Y_{t+1} - \widetilde{X}_{t+1} = Y_{t+1} - \widehat{X}_t = \\
&= \widetilde{\varepsilon}_{t+1} + \eta_{t+1} \\
\sigma^2_{\widetilde{\varepsilon}_{t+1}} &= \sigma^2_{\widehat{\varepsilon}_t} + \sigma^2_\nu \\
\sigma^2_{\delta Y_{t+1}} &= \sigma^2_{\widetilde{\varepsilon}_{t+1}} + \sigma^2_\eta = \sigma^2_{\widehat{\varepsilon}_t} + \sigma^2_\nu + \sigma^2_\eta \\
\sigma_{\widetilde{\varepsilon}_{t+1}\delta Y_{t+1}} &= \sigma^2_{\widetilde{\varepsilon}_{t+1}} \\
\widehat{X}_{t+1} &= \widetilde{X}_{t+1} + \frac{\sigma_{\widetilde{\varepsilon}_{t+1}\delta Y_{t+1}}}{\sigma^2_{\delta Y_{t+1}}} \cdot \delta Y_{t+1} = \\
&= \widehat{X}_t + \frac{\sigma_{\widetilde{\varepsilon}_{t+1}}}{\sigma^2_{\widetilde{\varepsilon}_{t+1}} + \sigma^2_\eta}(Y_{t+1} - \widehat{X}_t) \\
\sigma^2_{\widehat{\varepsilon}_{t+1}} &= \sigma^2_{\widetilde{\varepsilon}_{t+1}} - \frac{\sigma^4_{\widetilde{\varepsilon}_{t+1}}}{\sigma^2_{\widetilde{\varepsilon}_{t+1}} + \sigma^2_\eta} = \frac{\sigma^2_{\widetilde{\varepsilon}_{t+1}} \cdot \sigma^2_\eta}{\sigma^2_{\widetilde{\varepsilon}_{t+1}} + \sigma^2_\eta} \ .
\end{aligned}$$

It is interesting to remark that for $t \to +\infty$, $\sigma^2_{\widehat{\varepsilon}_t}$ has a limit $\sigma^2_{\widehat{\varepsilon}_\infty}$ given by

$$\sigma^2_{\widehat{\varepsilon}_\infty} = 1/2(\sqrt{\sigma^4_\nu + 4\sigma^2_\nu\sigma^2_\eta} - \sigma^2_\nu) \ .$$

On the contrary, without observations, we would have a variance of the stochastic part obeying

$$\sigma^2_{X_{t+1}} = \sigma^2_{X_t} + \sigma^2_\nu \ ,$$

so having an unbounded trend

$$\sigma^2_{X_t} = t\sigma^2_{X_0} \ .$$

3 Non-linear dynamic systems: the extended Kalman Filter

We generalize the model discussed in §2, by assuming that both dynamic equations and observation equations are non-linear, namely

$$X_{t+1} = g_{t+1}(x_t) + Q_{t+1}\nu_{t+1} \ , \tag{169}$$

$$Y_{t+1} = h_{t+1}(X_{t+1}) + \eta_{t+1} \ . \tag{170}$$

We make the same hypotheses (76), (80), (81) on ν_t, η_t and (77), (78) on the initial state X_0. We will also use the definitions

$$N^T_{t'} = \{\nu^T_1 \dots \nu^T_{t'}\} \ , \ I^T_{t'} \equiv \{Y^T_1 \dots Y^T_{t'}\} \ . \tag{171}$$

Yet from now on we will set the hypothesis that $\{\nu_t\}, \{\eta_t\}$ are normally distributed, so that for such noises the concept of orthogonality and stochastic independence do coincide; as an alterative we could always assume directly that they are stochastically independent, so that relations like $t' > t, E\{\nu_{t'}|_{N_t}\} = 0$ or $\forall t, t' E\{\eta_{t'}|_{N_t}\} = 0$ hold true.

In this section we first want to explore the differences between the non-linear and the linear model and then we will study how to properly use the linearization technique and the results of §2, to provide an approximate solution to our problem, which is always to find

$$\widehat{X}_{t+1} = E\{X_{t+1}|_{I_{t+1}}\} \tag{172}$$

and $C_{\widehat{\varepsilon}_{t+1}}$, where $\widehat{\varepsilon}_{t+1}$ is the prediction error

$$\widehat{\varepsilon}_{t+1} = X_{t+1} - \widehat{X}_{t+1} \ , \tag{173}$$

knowing $\widehat{X}_t$ and $C_{\widehat{\varepsilon}_t}$.

We first notice that (169) seems to be different from (75), not only because in (75) $\overline{X}_{t+1}$ depends linearly on $\overline{X}_t$, but also because in this equation we have a non-homogenous constant term, the deterministic input u_{t+1}. As we have seen, the sequence $\{u_{t'}\}$ essentially determines the evolution of the mean of $\{X_{t'}\}$ (see (87)), which can be separated from the zero mean stochastic part, as in (89), (90).

In the present context though, due to the non-linearity of (169), one cannot separate the evolution of the mean and of the stochastic residual. So in general in (169), (170) $\{X_{t'}\}, \{Y_{t'}\}$ cannot be considered as zero mean variables.

On the other hand, by writing the dependency of X_{t+1} from X_t as $g_{t+1}(X_t)$, namely a general non-linear function changing with time, we have a model that can account for the presence of any input, on condition that this is a known constant vector.

Example 3. To gain insight in the effects of the non-linearity in the dynamics of the system we just write the first two steps.

We have

$$\begin{aligned} &X_0 = 0 \; C_0 = 0 \\ &X_1 = g_1(0) + Q_1\nu_1 \\ &X_2 = g_2(X_1) + Q_2\nu_2 \equiv g_2[g_1(0) + Q_1\nu_1] + Q_2\nu_2 \; . \end{aligned}$$

As we see, in general

$$\mu_1 = E\{X_1\} = g_1(0) \neq 0 \; .$$

In addition, X_2 is not anymore a linear function of $(\nu_1^T, \nu_2^T) \equiv N_2^T$, so $X_2 \notin L^2(N_2)$. Nevertheless, if $g_2(X_1)$ has finite second moments, what we assume to be true, we will have

$$X_2 \in \mathcal{L}^2(N_2)$$

and also

$$\nu_2 \perp \mathcal{L}^2(N_1)$$

so that

$$\nu_2 \perp X_1 \; .$$

Generalizing the above example one realizes that, if we put the hypotheses

$$g_{t+1}(X_t) \in \mathcal{L}^2(N_{t'}) \quad \forall t' \geq t \tag{174}$$

we can describe the dynamics of the system by the following steps:

a) at time t the system undergoes a transformation sending $X_t \in \mathcal{L}^2(N_t)$ to $g_{t+1}(X_t) \in \mathcal{L}^2(N_t)$

b) at time $t+1$ the system accumulates the noise $Q_{t+1}\nu_{t+1}$ by moving in $\mathcal{L}^2(N_{t+1})$ orthogonally to $\mathcal{L}^2(N_t)$.

This means for instance that

$$E\{\nu_{t+1}|N_t\} = 0 \; . \tag{175}$$

Once the state X_{t+1} has been established at time $t+1$, an observation Y_{t+1} is done which is attained by moving X_{t+1} to $h(X_{t+1}) \in \mathcal{L}^2(N_{t+1})$ and adding a noise η_{t+1}.

The situation is illustrated in Fig. 2, where the motion along $\mathcal{L}^2(N_{t'})$ as opposed to the motion along $L^2(N_{t'})$ typical of linear systems is represented. Of course the space of the observations $Y_{t'}$ is different from that of the states so the last two steps of the path have to be understood only in a symbolic way.

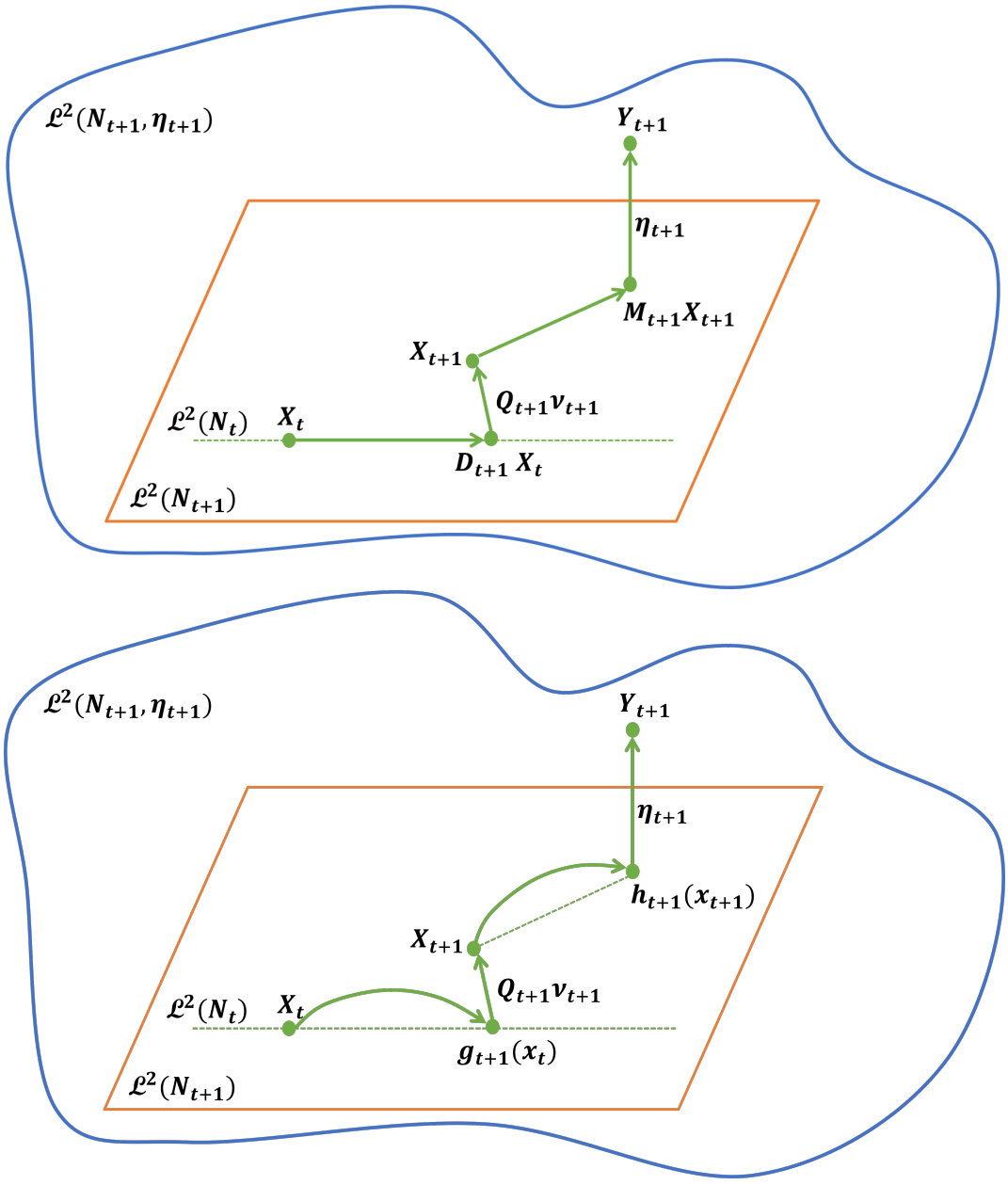


Figure 3: Symbolic representation of the formation of the observation Y_{t+1}. a) above, in a linear model, b) below, in a non-linear model, assuming that the dimension of X_t and Y_t is just 1.

Remark 12 (The linearization issue). As often when facing a non-linear problem, like producing the prediction (172) and the associated covariance matrix $C_{\widehat{\varepsilon}_{t+1}}$, a first approach is to linearize the equations, solve the linearized problem and then iterate with the hope of converging to the wanted solution. In this case yet we have one further problem. In fact, a first idea could be to try to linearize the predictor around the mean of $\{X_{t+1}\}$ and to compute such a mean iteratively by the approximate equation

$$\mu_{X_{t+1}} \cong g_{t+1}(\mu_{X_t}) \; ; \tag{176}$$

in this equation the approximation is in that we have put

$$E\{g_{t+1}(X_t)\} \cong g_{t+1}(\mu_{X_t}) \; . \tag{177}$$

This comes from a "linearization" of the dynamic equation (169), namely

$$\begin{aligned} X_{t+1} &= g_{t+1}(\mu_{X_t} + \delta X_t) + Q_{t+1}\nu_{t+1} \cong \\ &\cong g_{t+1}(\mu_{X_t}) + G_{t+1}(\mu_{X_t})\delta X_t + Q_{t+1}\nu_{t+1} \; , \end{aligned} \tag{178}$$

where

$$G_{t+1}(\mu_{X_t}) = \left.\frac{\partial g_{t+1}(X)}{\partial X}\right|_{X=\mu_{X_t}} \tag{179}$$

and $\delta X_t = X_t - \mu_{X_t}$ is the residual of X_t with zero mean. On taking the average of (178) one gets (177). The error in (177), when second derivatives of $g_{t+1}(X)$ are bounded, is a quadratic term in δX_t and so it is controlled by

$$E\{\delta X_t^2\} = Tr(C_{X_t}) \; . \tag{180}$$

Yet from (178) we see that

$$\begin{aligned} Tr(C_{X_{t+1}}) &= Tr(G_{t+1}C_{X_t}G_{t+1}^T) + Tr(Q_{t+1}C_{\nu_{t+1}}Q_{t+1}^T) \equiv \\ &\equiv Tr(C_{X_t}G_{t+1}^T G_{t+1}) + Tr(C_{\nu_{t+1}}Q_{t+1}^T Q_{t+1}) \; . \end{aligned} \tag{181}$$

Now let us assume that

$$\begin{cases} G_{t+1}^T G_{t+1} \geq cI \\ Q_{t+1}^T Q_{t+1} \geq qI \end{cases} \tag{182}$$

meaning that neither the dynamics nor the influence of the noise ν tend to zero with epochs t tending to infinity; then (182) gives

$$Tr(C_{X_{t+1}}) \geq cTr(C_{X_t}) + qTr(C_{\nu_{t+1}}) \; . \tag{183}$$

Further, assuming that $C_{\nu_{t+1}}$ is uniformly positive, i.e. the noise doesn't tend to faint, (183) tells that the sequence $Tr(C_{X_t})$ might not be bounded and it is as a matter of fact unbounded if $c > 1$. Just think of the most elementary example, namely $\delta X_t \in \mathcal{R}$ a random walk, i.e. a process satisfying

$$\delta X_{t+1} = \delta X_t + \nu_t \quad (E\{\nu_t^2\} = \sigma_\nu^2) \; . \tag{184}$$

In this case in fact

$$\sigma^2(\delta X_{t+1}) = \sigma^2(\delta X_t) + \sigma_\nu^2 \tag{185}$$

i.e.

$$\sigma^2(\delta X_t) = t \cdot c_\nu^2 \tag{186}$$

showing that δX_t has a stochastic divergence in time. All that says that choosing μ_{X_t} as linearization point is not a good idea and could lead, as in the example above, to inconsistent conclusions. A better idea then could be to choose for X_t our best estimate up to the time t, namely $\widehat{X}_t$. This however poses a different problem because $\widehat{X}_t$ is not fixed but a random variable. How to deal with this occupies the rest of this section and constitutes the heart of the Extended Kalman Filter approach.

The Extended Kalman Filter

It is our purpose to find

$$\widehat{X}_{t+1} = E\{X_{t+1}|_{I_{t+1}}\} \tag{187}$$

$$\begin{cases} C_{\widehat{\varepsilon}_{t+1}} = E\{\widehat{\varepsilon}_{t+1}\widehat{\varepsilon}_{t+1}^T\} \\ \widehat{\varepsilon}_{t+1} = X_{t+1} - \widehat{X}_{t+1} \ , \end{cases} \tag{188}$$

by applying at the best a linearization approach to the model (169), (170) and knowing

$$\widehat{X}_t = E\{X_t|_{I_t}\} \ , \ C_{\widehat{\varepsilon}_t} \ . \tag{189}$$

This program however is too ambitious for a general non-linear problem, because, due to the non-linear relations (146), (169), even assuming the noises normally distributed, the variates X_t, Y_t cannot be considered normally distributed too. As consequence, the concepts of orthogonality and stochastic independence do not coincide, and in particular the comfortable decomposition

$$E\{X_{t+1}|_{I_t,\delta Y_{t+1}}\} = E\{X_{t+1}|_{I_t}\} + E\{X_{t+1}|_{\delta Y_{t+1}}\} \ , \tag{190}$$

splitting the Kalman filter into the characteristic prediction-update steps, doesn't hold anymore.

So we have to restrict ourselves to the less optimal but still effective idea to use a linear regression to perform the filter, i.e. we transform (186), (187), (188) into the search of

$$\widehat{X}_{t+1} = X_{t+1}|_{I_{t+1}} \ , \tag{191}$$

$$C_{\widehat{\varepsilon}_{t+1}} = E\{\widehat{\varepsilon}_{t+1}\widehat{\varepsilon}_{t+1}^T\} \ , \tag{192}$$

$$\widehat{\varepsilon}_{t+1} = X_{t+1} - \widehat{X}_{t+1} \ , \tag{193}$$

given

$$\widehat{X}_t = X_t|_{I_t}, C_{\widehat{\varepsilon}_t} \ . \tag{194}$$

Notice that we are keeping the same symbol as in (186), (187), (188) although now $\widehat{X}_t, \widehat{\varepsilon}_t$ have a different meaning. In any way, we will continue to use the same notation as in the linear case and in particular we will put

$$\widetilde{X}_{t+1} = X_{t+1}|I_t \ , \ \widetilde{\varepsilon}_{t+1} = \widetilde{X}_{t+1} - X_t \ , \widetilde{Y}_{t+1} = Y_{t+1}|I_t \ , \ \delta Y_{t+1} = Y_{t+1} - \widetilde{Y}_{t+1} \ . \tag{195}$$

We underline that, since the operation $U|_{I_t}$ is a projection of U on $L^2(I_t)$, all orthogonality relations (e.g. $\widehat{\varepsilon}_{t+1} \perp L^2(I_{t+1})$, $\widetilde{\varepsilon}_{t+1} \perp L^2(I_t)$ etc.) continue to be valid.

Moreover, recalling the regression formula (57), we see that

$$\mu_{X_{t+1}} = \mu_{\widehat{X}_{t+1}} = \mu_{\widetilde{X}_{t+1}} \ , \tag{196}$$

namely $\widehat{\varepsilon}_{t+1}, \widetilde{\varepsilon}_{t+1}$ have zero average. The same holds indeed for the residuals $\widehat{\varepsilon}_{t'}, \widetilde{\varepsilon}_{t'}$ at every time.

Remark 13. It has to be well understood that despite the restriction of our predictors to $L^2(I_t)$, instead of $\mathcal{L}^2(I_t)$, the problem is still non-linear because (169), (170) continue to hold; only the relation between the predictor $\widehat{X}_{t+1}$ and I_{t+1} is now linear. So the development of the filter still has to be achieved by a suitable linearization.

To start the implementation of (191) we deal first with

$$X_{t+1}|_{I_t} \tag{197}$$

and apply a linearization approximation, namely

$$\begin{aligned} \widetilde{X}_{t+1} &= \{g(X_t) + Q_{t+1}\nu_{t+1}\}|_{I_t} \\ &= \{g(\widehat{X}_t + \widehat{\varepsilon}_{t+1})|_{I_t} = \\ &\cong \{g(\widehat{X}_t) + \widehat{G}_{t+1}\widehat{\varepsilon}_t\}|_{I_t} \ . \end{aligned} \tag{198}$$

In the last equation we have taken into account that $\nu_{t+1}|_{I_t} = 0$ and we have put

$$\widehat{G}_{t+1} = G_{t+1}(\widehat{X}_t) = \left.\frac{\partial g_{t+1}(X)}{\partial X}\right|_{X=\widehat{X}_t} \ . \tag{199}$$

In (198) the last step is just an approximation via linearization. Now it is important to realize that $g(\widehat{X}_t), \widehat{G}_{t+1}$ are known fixed quantities, according to (194), so that we can proceed from (194), getting

$$\widetilde{X}_{t+1} \cong g_{t+1}(\widehat{X}_t) + \widehat{G}_{t+1}\widehat{\varepsilon}_t|_{I_t} \equiv g_{t+1}(\widehat{X}_t) \ , \tag{200}$$

in fact by its very definition $\widehat{\varepsilon}_t = X_t - \widehat{X}_t$ is orthogonal to $L^2(I_t)$.

Relation (200) tells us that $\widetilde{X}_{t+1}$ is a known vector on the basis of the data (194).

Now let us use the orthogonal decomposition (73), that we apply to the zero average variable $\widetilde{\varepsilon}_{t+1}$, namely

$$(X_{t+1} - \widetilde{X}_{t+1})|_{I_{t+1}} = \widetilde{\varepsilon}_{t+1}|_{I_{t+1}} = \widetilde{\varepsilon}_{t+1}|_{I_t} + \widetilde{\varepsilon}_{t+1}|_{\delta Y_{t+1}} \ . \tag{201}$$

On the other hand $\widetilde{\varepsilon}_{t+1}|_{I_t} = 0$ by its very definition and also

$$\widetilde{X}_{t+1}|_{I_{t+1}} = \widetilde{X}_{t+1} \tag{202}$$

because $L^2(I_t)$ is a subspace of $L^2(I_{t+1})$. So (201) can be rewritten as

$$\widehat{X}_{t+1} = \widetilde{X}_{t+1} + \widetilde{\varepsilon}_{t+1}|_{\delta Y_{t+1}} \ . \tag{203}$$

This development is exact because $\widetilde{X}_{t+1} \in L^2(I_t)$, while δY_{t+1} is orthogonal to this space, as already observed.

On the other hand, by the regression formula (69) for zero mean random variables, we have

$$\widetilde{\varepsilon}_{t+1}|_{\delta Y_{t+1}} = C_{\widetilde{\varepsilon}_{t+1}\delta Y_{t+1}} C^{-1}_{\delta Y_{t+1}} \delta Y_{t+1} \ . \tag{204}$$

To proceed we need linearized formulas for $\widetilde{Y}_{t+1}$ and $\widetilde{\varepsilon}_{t+1}$.

We have

$$\begin{aligned} \widetilde{Y}_{t+1} &= Y_{t+1}|_{I_t} = \{h_{t+1}(\widetilde{X}_{t+1} + \widetilde{\varepsilon}_{t+1}) + \eta_{t+1}\}|_{I_t} = \\ &\cong \{h_{t+1}(\widetilde{X}_{t+1}) + \widetilde{H}_{t+1}\widetilde{\varepsilon}_{t+1}\}|_{I_t} \end{aligned} \tag{205}$$

because η_{t+1} is orthogonal to $L^2(I_t)$ and we have put

$$\widetilde{H}_{t+1} = \left.\frac{\partial h_{t+1}(X)}{\partial X}\right|_{X=\widetilde{X}_{t+1}} \ . \tag{206}$$

Once more, $h_{t+1}(\widetilde{X}_{t+1}), \widetilde{H}_{t+1}$ are fixed known quantities and $\widetilde{\varepsilon}_{t+1}|_{I_t} = 0$ because $\widetilde{\varepsilon}_{t+1}$ is orthogonal to $L^2(I_t)$; we have then

$$\widetilde{Y}_{t+1} \cong h_{t+1}(\widetilde{X}_{t+1}) \tag{207}$$

and

$$\begin{aligned} \delta Y_{t+1} &= Y_{t+1} - \widetilde{Y}_{t+1} \cong h(X_{t+1}) - h(\widetilde{X}_{t+1}) + \eta_{t+1} \\ &\cong \widetilde{H}_{t+1}\widetilde{\varepsilon}_{t+1} + \eta_{t+1} \ . \end{aligned} \tag{208}$$

Moreover, always developing to the first order, we can write

$$\begin{aligned} \widetilde{\varepsilon}_{t+1} &= X_{t+1} - \widetilde{X}_{t+1} = \\ &\cong g(\widehat{X}_t + \widehat{\varepsilon}_t) - g(\widehat{X}_t) + Q_{t+1}\nu_{t+1} \\ &\cong \widehat{G}_{t+1}\widehat{\varepsilon}_t + Q_{t+1}\nu_{t+1} \ . \end{aligned} \tag{209}$$

Noting that both (208), (209) are orthogonal decompositions, we find in sequence

$$C_{\widetilde{\varepsilon}_{t+1}} = K_{t+1} = \widehat{G}_{t+1} C_{\widehat{\varepsilon}_t} \widehat{G}_{t+1}^T + Q_{t+1} C_{\nu_{t+1}} Q_{t+1}^T \ , \tag{210}$$

$$C_{\delta Y_{t+1}} = \widetilde{H}_{t+1} K_{t+1} \widetilde{H}_{t+1}^T + C_{\eta_{t+1}} \ , \tag{211}$$

$$C_{\widetilde{\varepsilon}_{t+1}\delta Y_{t+1}} = K_{t+1} \widetilde{H}_{t+1}^T \ .$$

These relations used with (204), (200) in (191) give the Extended Kalman Filter predictor

$$\widehat{X}_{t+1} = g_{t+1}(\widehat{X}_t) + K_{t+1} \widetilde{H}_{t+1}^T [\widetilde{H}_{t+1} K_{t+1} \widetilde{H}_{t+1}^T + C_{\eta_{t+1}}]^{-1} \delta Y_{t+1} \tag{212}$$

where on right hand side we have only known quantities given as data of the problem As for the covariance of the prediction error $C_{\widehat{\varepsilon}_{t+1}}$, we proceed as in the linear case.

Let us write

$$\begin{aligned} \widehat{\varepsilon}_{t+1} &= X_{t+1} - \widehat{X}_{t+1} \\ &= X_{t+1} - \widetilde{X}_{t+1} - K_{t+1} \widetilde{H}_{t+1}^T C_{\delta Y_{t+1}}^{-1} = \\ &= \widetilde{\varepsilon}_{t+1} - K_{t+1} \widetilde{H}_{t+1}^T C_{\delta Y_{t+1}}^{-1} \delta Y_{t+1} \ ; \end{aligned} \tag{213}$$

we notice that $\delta Y_{t+1} \in L^2(I_{t+1})$ while $\widehat{\varepsilon}_{t+1}$ is by definition orthogonal to this space ($\widehat{X}_{t+1}$ is the orthogonal projection of X_{t+1} onto $L^2(I_{t+1})$), therefore solving for $\widetilde{\varepsilon}_{t+1}$ (213), we obtain

$$C_{\widetilde{\varepsilon}_{t+1}} = C_{\widehat{\varepsilon}_{t+1}} + K_{t+1} \widetilde{H}_{t+1}^T C_{\delta Y_{t+1}}^{-1} \widetilde{H}_{t+1} K_{t+1} \ . \tag{214}$$

Recalling that $C_{\widetilde{\varepsilon}_{t+1}} = K_{t+1}$, (214) yields

$$C_{\widehat{\varepsilon}_{t+1}} = K_{t+1} - K_{t+1} \widetilde{H}_{t+1}^T C_{\delta Y_{t+1}}^{-1} \widetilde{H}_{t+1} K_{t+1} \tag{215}$$

where K_{t+1} is given by (210) and $C_{\delta Y_{t+1}}$ by (211); the posed problem is therefore solved.

Remark 14. As underlined in literature (see [9]) the Extended Kalman Filter propagates the evolution of the state of the system by the non-linear dynamics $g_{t+1}(X)$ and the forward prediction equation

$$\widetilde{X}_{t+1} \cong g_{t+1}(\widehat{X}_t) \ ; \tag{216}$$

this at least takes into account the non-linear character of the motion. On the contrary the propagation of the covariance of the prediction error is purely linearized. The effect is that of a considerable distortion of the covariance, that is then influencing the following steps of the updating of the trajectory $\{X_t\}$ so much as to lead in some cases to an inconsistent estimate of it, diverging from the true path. It is for this reason that several methods have been invented to overcome the simple linearization procedure. Among them one particularly simple to implement and relatively light from the numerical point of view is the so-called *Unscented Kalman Filter*, that we present in the rest of the tutorial.

Example 4. The system is a point P moving in the plane, subject to a dynamic model with a constant velocity v in the direction of the x axis, plus some noise in both x and y directions. So the state of the system is described by two components collected in a vector:

$$\bar{X}_t = \begin{vmatrix} \bar{x}_t \\ \bar{y}_t \end{vmatrix}, \tag{217}$$

obeying the law:

$$\bar{X}_{t+1} = \begin{vmatrix} \bar{x}_{t+1} \\ \bar{y}_{t+1} \end{vmatrix} = \begin{vmatrix} \bar{x}_t \\ \bar{y}_t \end{vmatrix} + \begin{vmatrix} v \\ 0 \end{vmatrix} + \begin{vmatrix} \nu_{t+1} \\ \omega_{t+1} \end{vmatrix} = \bar{X}_t + \bar{V}_t + n_{t+1} \tag{218}$$

where

$$E\{n\} = 0, C_n = \sigma_0^2 I. \tag{219}$$

In what follows we shall assume $\sigma_0 = 0.1$ that means a noise of $10cm$. Moreover we will take also $v = 1$, considered as $1m/s$ corresponding to a walking velocity of $3.6km/h$. The motion starts at a point of the negative x axis $(-L, 0)$ without error. In this example we will take $L = 120m$, so that after two minutes walk the point is in the

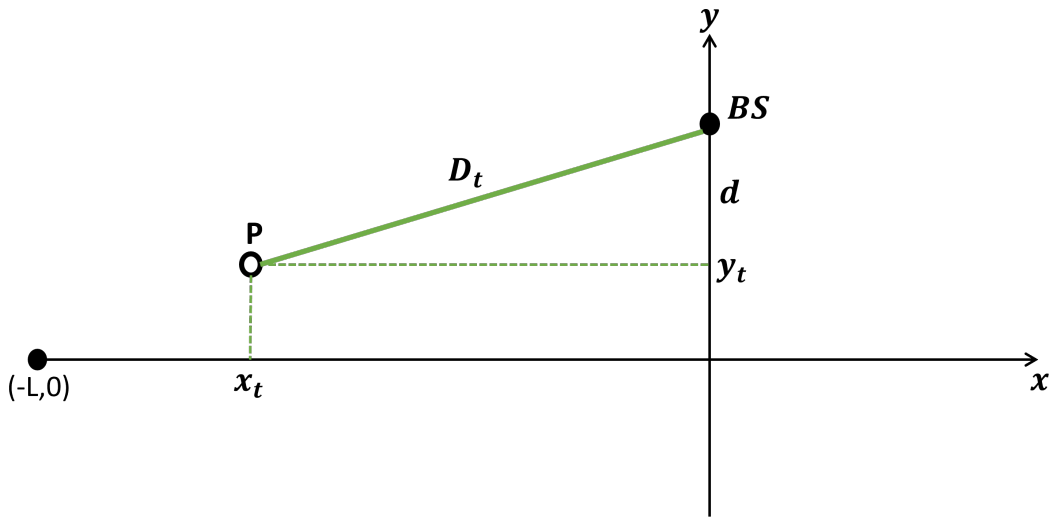


Figure 4: The 2D navigation of Example 4.

surroundings of the origin (see Fig. 4).
Since the dynamic of the system is linear, it is convenient to split the evolution into the deterministic part

$$E\{\bar{X}_t\} = \mu_{\bar{X}_t} = \bar{X}_0 + \bar{V}t = \begin{vmatrix} -L + vt \\ 0 \end{vmatrix}, \tag{220}$$

satisfying the equation

$$\mu_{\bar{X}_{t+1}} = \mu_{\bar{X}_t} + \bar{V}, \tag{221}$$

and the zero mean stochastic part

$$X_t = \bar{X}_t - \mu_{\bar{X}_t} = \begin{vmatrix} x_t \\ y_t \end{vmatrix}, \tag{222}$$

satisfying the equation

$$X_{t+1} = X_t + n_{t+1}. \tag{223}$$

Now we add a flow of observations namely distances at $1Hz$ rate from P and a Base Station BS placed on the y axis at coordinates $(0, d)$ (see Figure 4).
The observation equations have therefore the form

$$W_{t+1} = h(\bar{X}_{t+1}) + \eta_{t+1}, \tag{224}$$

with η_t a white noise in time,

$$E\{\eta_t\} = 0, \sigma^2(\eta_t) = \sigma_0^2, \tag{225}$$

and independent of n_t.
In this case

$$h(\bar{X}_{t+1}) = \sqrt{\bar{x}_{t+1}^2 + (d - \bar{y}_{t+1})^2} = \sqrt{(-L + v(t+1) + x_{t+1})^2 + (d - y_{t+1})^2} = D_{t+1}. \tag{226}$$

Our purpose is to retrieve the trajectory of P up to the time $t = 120s$, by applying the Extended Kalman Filter.
Following the reasoning of this section we find, from the dynamic equation,

$$Q_{t+1} = I, G_{t+1} = I \tag{227}$$

from (200)

$$\tilde{X}_{t+1} = \hat{X}_t \tag{228}$$

from (206)

$$H_{t+1} = \frac{1}{h(\hat{X}_t)} \begin{bmatrix} -L + v(t+1) + \hat{x}_t & \hat{y}_t - d \end{bmatrix} \tag{229}$$

from (207)

$$\tilde{W}_{t+1} = h(\hat{X}_t) = \sqrt{(-L + v(t+1) + \hat{x}_t)^2 + (\hat{y}_t - d)^2} \tag{230}$$

and

$$\delta W_{t+1} = W_{t+1} - \tilde{W}_{t+1}, \tag{231}$$

from (210) and (211)

$$K_{t+1} = C_{\hat{\epsilon}_t} + \sigma_0^2, \tag{232}$$

$$C_{\delta W_{t+1}} = \sigma^2(\delta W_{t+1}) = H_{t+1} K_{t+1} H_{t+1}^T + \sigma_0^2 \tag{233}$$

and finally from (212)

$$\hat{X}_{t+1} = \begin{vmatrix} \hat{x}_{t+1} \\ \hat{y}_{t+1} \end{vmatrix} + (C_{\epsilon_t} \hat{H}_{t+1}^T) \cdot \frac{\delta W_{t+1}}{H_{t+1} C_{\hat{\epsilon}_t} H_{t+1}^T + \sigma_0^2}. \tag{234}$$

Moreover from (214) we get the covariance propagation

$$C_{\hat{\epsilon}_{t+1}} = K_{t+1} - \frac{1}{\sigma^2(\delta W_{t+1})} (K_{t+1} H_{t+1}^T)(H_{t+1} K_{t+1}). \tag{235}$$

4 The Unscented Transformations: the one-dimensional case

The method has been invented (see [9]) to produce the propagation of mean and covariance through a non-linear transformation

$$Y = g(X) \ , \ X \in R^n \ , \ Y \in R^m \tag{236}$$

with an improvement with respect to the simple linear approximation, i.e.

$$\mu_Y \cong g(\mu_X) \tag{237}$$

$$\begin{cases} C_Y \cong G C_X G^T \\ G = \frac{\partial g}{\partial X}\Big|_{X=\mu_X} \end{cases} , \tag{238}$$

but without resorting to the use of second and higher order derivatives of $g(X)$, that becomes numerically heavy specially for real time applications that we have in mind.

We restrict ourselves to the hypothesis that X is a normal variate, with mean μ_X and covariance C_X, sending to Remark 13 a comment on the significance of such hypothesis.

Following the authors of [9], the "intuition" of the approach is in that "with a fixed number of parameters it should be easier to approximate a Gaussian distribution, than it is to approximate a non-linear function/transofrmation".

The trick devised is then, to use just a discrete distribution to approximate the normal distribution of X and then to apply the non-linear transformation to the discrete variable X' to obtain a discrete Y' that approximates Y.

We start by analyzing the problem of a one-dimensional variable with a transformation $g : R^1 \to R^1$, because in this case formulas are particularly transparent.

Assume X to be normal with mean μ_X and variance σ_X^2. Then with a discrete distribution on two argumental values only we can generate a variable X' with the same mean and variance. In fact, let us put

$$X' = \begin{cases} X'_1 = \mu_X + d\sigma_X & p_1 = p \\ X'_2 = \mu_X - d\sigma_X & p_2 = 1 - p \end{cases} \tag{239}$$

If we impose

$$\begin{cases} \mu_{X'} = \mu_X + pd\sigma_X - (1-p)d\sigma_X \equiv \mu_X \\ \sigma^2_{X'} = pd^2\sigma_X^2 + (1-p)d^2\sigma_X^2 \equiv \sigma_X^2 \end{cases} \tag{240}$$

we find

$$p = 1/2 \ , \ d = 1 \ . \tag{241}$$

So

$$X'_I \equiv \left\{ \begin{array}{ll} \mu_X + \sigma_X & p = 1/2 \\ \mu_X - \sigma_X & 1 - p = 1/2 \end{array} \right. \tag{242}$$

has the same mean and variance as X. Now, since X'_I is symmetric it is obvious that its third central moment, as all odd central moments, is zero

$$\overline{\mu}_3(X'_2) = E\{(X' - \mu_{X'_I})^3\} = 0 \ ; \tag{243}$$

this says in statistical terms that X'_I is not skewed, similarly to what happens to the normal variate X. As for the fourth central moment though, one has

$$\overline{\mu}_4(X'_I) = E\{(X' - \mu_{X'_I})^4\} = \sigma^4_{X'_I} \equiv \sigma^4_X \ . \tag{244}$$

This is in contrast with the case of a normal distribution which we know to have a kurtosis equal to 3, namely

$$\overline{\mu}_4(X) = 3\sigma^4_X \ . \tag{245}$$

If we wanted X and X' to have equal moments up to the fourth, still keeping a symmetric discrete distribution, we could put

$$X'_{II} \equiv \left\{ \begin{array}{lll} \mu_X - d\sigma_X & p & \\ \mu_X & q & (q + 2p = 1) \\ \mu_X + d\sigma_X & p \ , & \end{array} \right. \tag{246}$$

Due to its symmetry we obviously have at once

$$\mu_{X'_{II}} = \mu_X \ , \ \overline{\mu}_3(X'_{II}) = \overline{\mu}_3(X) = 0 \ .$$

On the other hand, imposing equality of variance and fourth central moments we get

$$\left\{ \begin{array}{l} \sigma^2_{X'_{II}} = 2d^2\sigma^2_X p \equiv \sigma^2_X \\ \overline{\mu}_4(X'_{II}) = 2d^4\sigma^4_X p = 3\sigma^4_X \end{array} \right. \tag{247}$$

that imply

$$d = \sqrt{3} \ , \ p = 1/6 \ , \ q = 2/3 \ . \tag{248}$$

The points $X'_{Ii} \equiv (\mu_X - \sigma_X, \mu_X + \sigma_X)$ in the first solution or $X'_{IIi} \equiv (\mu_X - \sqrt{3}\sigma_X, \mu_X, \mu_X + \sqrt{3}\sigma_X)$ in the second are denominated σ-points of the Gaussian distribution of X.

Let us see now what are the consequences of one or the other solution when we apply it to a non-linear transformation $R^1 \to R^1$, namely

$$Y = g(X) \ . \tag{249}$$

The idea is that if we can approximate the normal variate X, then we can use

$$Y' = g(X') \tag{250}$$

as an approximation of Y.

In particular we would like to know how good is the approximation

$$\mu_Y \cong \mu_{Y'} \tag{251}$$

and

$$\sigma^2_Y \cong \sigma^2_{Y'} \ , \tag{252}$$

comparing the first with the second choice of X', as described above.

We make the hypothesis that g is regular up to the sixth order derivatives, so that we can exploit Taylor's formula up to this order. After posing $X = \mu_X + \delta X$, we compute

$$Y = g(\mu_X + \delta X) = g(\mu_X) + g'\delta X + \frac{1}{2}g''\delta X^2 + \frac{1}{6}g'''\delta X^3 + \frac{1}{24}g^{iv}\delta X^4 + O_5 \ , \tag{253}$$

where all derivatives are evaluated at μ_X.

Recalling that

$$E\{\delta X\} = E\{\delta X^3\} = E\{\delta X^5\} = 0 \ ; \ E\{\delta X^2\} = \sigma_X^2 \ , \ E\{\delta X^4\} = 3\sigma_X^4 \ , \tag{254}$$

we find

$$\mu_Y = g(\mu_X) + \frac{1}{2}g''\sigma_X^2 + \frac{1}{8}g^{iv}\sigma_X^4 + O_6 \ . \tag{255}$$

This yields

$$Y - \mu_Y = g'\delta X + \frac{1}{2}g''(\delta X^2 - \sigma_X^2) + \frac{1}{6}g'''\delta X^3 + O_4 \tag{256}$$

and then

$$\begin{aligned} (Y - \mu_Y)^2 &= g'^2\delta X^2 + \frac{1}{4}g''^2(\delta X^2 - \sigma_X^2)^2 + g'g''\delta X(\delta X^2 - \sigma_X^2) + \\ &+ \frac{1}{3}g'g'''\delta X^4 + O_5 \ . \end{aligned} \tag{257}$$

Averaging and recalling (245) we have finally

$$\begin{aligned} \sigma_Y^2 &= g'^2\sigma_X^2 + \frac{1}{4}g''^2 \cdot 2\sigma_X^4 + g'g'''\sigma_X^4 - O_6 \\ &= g'^2\sigma_X^2 + (\frac{1}{2}g''^2 + g'g''')\sigma_X^4 + O_6 \ . \end{aligned} \tag{258}$$

Now we repeat the above computation for the two choices

$$X'_I \equiv \left\{ \begin{array}{ll} \mu_X + \sigma_X & p = 1/2 \\ \mu_X - \sigma_X & p = 1/2 \end{array} \right. \ ; \ X'_{II} = \left\{ \begin{array}{ll} \mu_X + \sqrt{3}\sigma_X & p = 1/6 \\ \mu_X & p = 2/3 \\ \mu_X - \sqrt{3}\sigma_X & p = 1/6 \end{array} \right. \tag{259}$$

giving rise to the corresponding

$$Y'_I = g(X'_I) \ ; \ Y'_{II} = g(X'_{II}) \ , \tag{260}$$

namely

$$Y'_I = \left\{ \begin{array}{ll} g(\mu_X + \sigma_X) & p = 1/2 \\ g(\mu_X - \sigma_X) & p = 1/2 \end{array} \right. \tag{261}$$

and

$$Y'_{II} = \left\{ \begin{array}{ll} g(\mu_X + \sqrt{3}\sigma_X) & p = 1/6 \\ g(\mu_X) & p = 2/3 \\ g(\mu_X - \sqrt{3}\sigma_X) & p = 1/6 \end{array} \right. \tag{262}$$

Developing both Y'_I, Y'_{II} in powers of σ_X up to order 4 and recalling that

$$\left\{ \begin{array}{l} \mu_{X'_I} = \mu_X \ ; \sigma^2_{X'_I} = \sigma^2_X \ ; \ \overline{\mu}_4(X'_I) = \sigma^4_X \\ \mu_{X'_{II}} = \mu_X \ ; \ \sigma^2_{X'_{II}} = \sigma^2_X \ ; \ \overline{\mu}_4(X'_{II}) = 3\sigma^4_X \end{array} \right. \tag{263}$$

we get

$$\begin{cases} \mu_{Y'_I} = g(\mu_X) + \frac{1}{2}g''\sigma_X^2 + \frac{1}{24}g^{iv}\sigma_X^4 + O_6 \\ \mu_{Y'_{II}} = g(\mu_X) + \frac{1}{2}g''\sigma_X^2 + \frac{1}{8}g^{iv}\sigma_X^4 + O_6 \end{cases} \tag{264}$$

and

$$\begin{cases} \sigma^2_{Y'_I} = g'^2\sigma_X^2 + \frac{1}{6}g'g'''\sigma_X^2 + O_6 \\ \sigma^2_{Y'_{II}} = g'^2\sigma_X^2 + (\frac{1}{2}g'' + g'g''')\sigma_X^4 + O_6 \ . \end{cases} \tag{265}$$

The conclusion is that

$$\mu_Y - \mu_{Y'_I} = \frac{1}{12}g^{iv}\sigma_X^4 + O_6 = O_4 \ , \tag{266}$$

$$\sigma_Y^2 - \sigma^2_{Y'_I} = \left(\frac{1}{2}g''^2 + \frac{5}{6}g'g'''\right)\sigma_X^4 + O_6 = O_4 \tag{267}$$

and

$$\mu_Y - \mu_{Y'_{II}} = O_6 \tag{268}$$

$$\sigma_Y^2 - \sigma^2_{Y'_{II}} = O_6 \ , \tag{269}$$

showing the improvement of the second choice with respect to the first.

Before concluding the example we would like to stress that $\mu_{Y'_I}, \mu_{Y'_{II}}, \sigma^2_{Y'_I}, \sigma^2_{Y'_{II}}$ are calculated by the ordinary statistical formulas

$$\mu_{Y'} = \sum_i Y'_i p_i \ , \tag{270}$$

$$\sigma^2_{Y'} = \sum_i (Y'_i - \mu_{Y'})^2 p_i \tag{271}$$

where argumental values Y'_i and probabilities p_i are given in (261), (262), depending on the choice taken.

Remark 15. One could argue that the hypothesis of normality of X is too restrictive. In reality, however, the procedure of the unscented transform aims at approximating the mean and variance (in the 1D case) of Y and thus, once X has a distribution around μ_X bulky and not too skewed, the approximation in practice works sufficiently well, even if X is not Gaussian. What is important is that in the interval of high probability of X the function g be smooth enough to be well approximated by a Taylor development up to the fourth order. This should be more than sufficient to get an improvement on the simple linear approximation. Different is the problem of the required smoothness of g to guarantee that the error estimates (266), (267), (268), (269) are realistic. It has been asserted (see [9]) that in principle the Unscented Transform works even with a discontinuous function for which no linearization can be applied.

This of course is true, however a warning on such a mater is certainly justified, as shown in next Example 5.

Example 5. In this example we examine the difference between the two choices of σ-points, X'_I and X'_{II}, with a very simple function

$$Y = g(X) = (a + X)^2 = a^2 + 2aX + X^2 \ .$$

We take for the X distribution, a standard Gaussian, so that $\mu_X = 0, \sigma_X = 1$ and the σ-points of X'_I and X'_{II} are respectively $\{1, -1\}, \{\sqrt{3}, 0, -\sqrt{3}\}$.

We have then

$$Y'_I = \begin{cases} (a+1)^2 & p = 1/2 \\ (a-1)^2 & p = 1/2 \end{cases} \qquad Y'_{II} = \begin{cases} (a+\sqrt{3})^2 & p = 1/6 \\ a^2 & p = 2/3 \\ (a-\sqrt{3})^2 & p = 1/6 \end{cases}$$

Computing means and covariances we get

$$\mu_{Y'_I} = a^2 + 1 \ , \ \sigma^2_{Y'_I} = 4a^2$$

$$\mu_{Y'_{II}} = a^2 + 1 \ , \ \sigma^2_{Y'_{II}} = 4a^2 + 2 \ .$$

Since by a direct calculus, applying the known information $\mu_X = 0$, $\sigma_X^2 = 1, \overline{\mu}_3(X) = 0, \overline{\mu}_4(X) = 3$, we have

$$\mu_Y = a^2 + 1 \ , \ \sigma_Y^2 = 4a^2 + 2 \ ,$$

we see that Y_I' can reproduce the mean but not the variance of Y, while Y_{II}' reproduces both. This result was indeed to be expected, since Y_{II}' can produce an approximation up to the fourth order, which is exactly what is necessary to compute σ_Y^2.

Example 6. In this example we want to illustrate the effects of a certain irregularity in $g(X)$. So we take

$$g(X) = \left\{ \begin{array}{ll} X & X \geq 0 \\ 0 & X < 0 \end{array} \right. ;$$

as distribution of X we choose again a standardized normal, so that $\mu_X = 0, \sigma_X^2 = 1$. Therefore the σ-points of X are again

$$X_I' = \left\{ \begin{array}{ll} 1 & p = 1/2 \\ -1 & p = 1/2 \end{array} \right. , \ X_{II}' = \left\{ \begin{array}{ll} \sqrt{3} & p = 1/6 \\ 0 & p = 2/3 \\ -\sqrt{3} & p = 1/6 \end{array} \right. ,$$

and the corresponding Y' variables are

$$Y_I' = \left\{ \begin{array}{ll} 1 & p = 1/2 \\ 0 & p = 1/2 \end{array} \right. \quad Y_{II}' = \left\{ \begin{array}{ll} \sqrt{3} & p = 1/6 \\ 0 & p = 5/6 \end{array} \right. ,$$

Then we have

$$\begin{aligned} \mu_{Y_I'} &= \ 0{,}50 \ \ , \ \sigma_{Y_I'}^2 = \ 0{,}25 \\ \mu_{Y_{II}'} &= \ 0{,}29 \ \ , \ \sigma_{Y_{II}'}^2 = \ 0{,}42 \ \ . \end{aligned}$$

These figures have to be compared with

$$\mu_Y = \frac{1}{\sqrt{2\pi}} \int_0^{+\infty} y e^{-\frac{y^2}{2}} dy = \frac{1}{\sqrt{2\pi}} \cong \ 0{,}40$$

and

$$\sigma_Y^2 = \frac{1}{\sqrt{2\pi}} \int_0^{+\infty} y^2 e^{-\frac{y^2}{2}} dy - \ 0{,}16 \ = \ 0{,}34 \ .$$

As we can see the differences are not huge, yet clearly not negligible. Moreover, in this case the second choice doesn't display an improvement on the first.

5 The Unscented Transform: general case

In this section we aim at repeating the reasoning of the preceding section for the general case

$$Y = g(X) \ ; \ X \in R^n \ , \ Y \in R^m \ . \tag{272}$$

As before, we assume X to be normally distributed with mean μ_X and covariance C_X. In order to find the σ-point of X, to build subsequently some discrete X', we need first to notice that the $n$dimensional standardization of X is done by the formulas

$$X = \mu_X + TZ \ , \ Z = T^{-1}(X - \mu_X) \tag{273}$$

where T is any square root of C_X, namely

$$TT^T = C_X \ . \tag{274}$$

Indeed many matrices T can be used to implement (274); among them we recall the symmetric square root of C_X, usually denoted $C_X^{1/2}$, and the factor T of the Cholesky decomposition, which is easy to compute.

One important remark is that the Z variable defined by (273) is first of all normal itself and furthermore

$$\mu_Z = 0 \ , \ C_Z = I_{(n)} \ . \tag{275}$$

We note too that, given the particular form of the normal density (see (38)), the components of $Z, Z_i (1 = 1 \dots n)$, are stochastically independent, so that the mean of products of functions of different components is just the product of the means of these functions. In particular, for the moments of third and fourth order of Z we have

$$\mu_3(Z) = E\{Z_i Z_j Z_k\} = 0 \quad \forall i, j, k \tag{276}$$

and

$$\mu_4(Z) = E\{Z_i Z_j Z_k Z_\ell\} = \begin{cases} 3 & i = j = k = \ell \\ & (i = j, k = \ell)(i \neq k) \\ 1 & (i = k, j = \ell)(i \neq j) \\ & (i = \ell, j = k)(i \neq j) \\ 0 & \text{all the other cases.} \end{cases} \tag{277}$$

We summarize the long description (277) into a unique formula, namely

$$\mu_4(Z) = E\{Z_i Z_j Z_k Z_\ell\} = \delta_{ij}\delta_{k\ell} + \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk} \ . \tag{278}$$

It is worth mentioning that for Z, as for all variates with zero mean, simple moments are the same as central moments. Endowed with (278) we can easily compute the moments of $\delta X = X - \mu_X$, i.e. the central moment of X. In fact, first for all components i, j, k

$$\overline{\mu}_3(X) = E\{\delta X_i \delta X_j \delta X_k\} = 0 \ ; \tag{279}$$

the same is true in general for any odd moment of $\delta X = X - \mu_X$ because this is an odd function of δX while the Gaussian probability density is an even function of δX and the integral of the product is zero.

Moreover, with the help of (273), we have

$$\overline{\mu}_4(X) = E\{\delta X_i \delta X_j \delta X_k \delta X_\ell\} = \tag{280}$$

$$= \sum_{q,r,s,t=1}^{n} T_{iq} T_{jr} T_{ks} T_{\ell t} (\delta_{qr}\delta_{st} + \delta_{qs}\delta_{rt} + \delta_{qt}\delta_{rs}) \ . \tag{281}$$

Let us make the computation of the first term:

$$\sum_{q,r,s,t=1}^{n} T_{iq} T_{jr} T_{ks} T_{\ell t} \delta_{qr}\delta_{st} = \sum_q T_{iq} T_{jq} \cdot \sum_s T_{ks} T_{\ell s} =$$
$$= C_{X,ij} C_{X,k\ell} \ , \tag{282}$$

where (274) has been used.

By cycling the indexes we arrive then at the formula

$$\begin{aligned} \overline{\mu}_{4,ijk\ell}(X) &= E\{\delta X_i, \delta X_j \delta X_k \delta X_\ell\} = \\ &= C_{X,ij} C_{X,k\ell} + C_{X,ik} C_{X,j\ell} + C_{X,i\ell} C_{X,jk} \ . \end{aligned} \tag{283}$$

Now we build the unscented variable X' by exploiting an analogy to the one-dimensional case, i.e. using two different choices of the σ-points in an effort to produce the identity of mean, covariance and central fourth moment tensor of X and X'.

Subsequently, we will analyze the effect of the two choices in the approximate transformation of X' to $Y' = g(X')$, compared with $g(X)$. The idea is to produce first σ-points of Z and then to transform them into σ-points of X.

First choice

The analogous to the choice (239) for an n-dimensional Z is

$$\arg(Z'_I) = \left\{ \begin{array}{l} Z'_i = du_i \\ Z'_{-i} = -du_i \end{array} \right. \quad 1 = 1, 2 \dots n \ , \tag{284}$$

with u_i any orthogonal basis of R^n. In this way we have $2n\sigma$-points symmetrically distributed.

Of course the basis $\{u_i\}$ is arbitrary, yet the easiest choice is to use the standard basis $\{e_i\}$, with unit vectors having components

$$\{e_i\}_j = e_{i,j} = \delta_{ij} \ , \tag{285}$$

namely 1 as i-th component and 0 all the others. Now we construct on such σ-points the distribution of a random variable Z' by assuming that it is uniform in all directions, so imitating the behaviour of a standard normal distribution. In this way we fix the probability attributed to each σ-point as

$$p_i = \frac{1}{2n} \ . \tag{286}$$

It is clear that, with such a symmetric distribution, we have

$$\mu_{Z'_I} = 0 \ , \ \overline{\mu}_3(Z'_I) = 0 \ ; \tag{287}$$

of course all higher odd moments are zero too. Now we fix d in such a way that the covariance of Z' is $I_{(n)}$, exactly as for Z. We have

$$\sum_{\substack{i=-n \\ i\neq 0}}^{n} Z'_i Z'^T_i p_i = \frac{1}{2n} d^2 \cdot 2 \sum_{i=1}^{n} e_i e_i^T = \frac{d^2}{n} I_{(n)} \ ; \tag{288}$$

here we have used the fact that $Z'_{-i} Z'^T_{-i} = Z'_i Z'^T_i$.

Therefore, if we put

$$d = \sqrt{n} \tag{289}$$

we have

$$C_{Z'_I} = I_{(n)} \ . \tag{290}$$

So we arrive at a discrete, symmetric variable having the required properties

$$Z'_I = \left\{ \begin{array}{l} Z'_i = \sqrt{n} e_i \\ Z'_{-i} = -\sqrt{n} e_i \end{array} \right. \quad p_i = \frac{1}{2n} \ , \tag{291}$$

$$\mu_{Z'_I} = 0 \ , \ C_{Z'_I} = I_{(n)} \ . \tag{292}$$

We have already observed that $\overline{\mu}_3(Z'_I) = 0$, as it is for Z, so we ask ourselves what is the situation with $\overline{\mu}_4(Z'_I)$. In analogy with the 1D case we have $\overline{\mu}_4(Z'_I) \neq \overline{\mu}_4(Z)$. In fact, recalling (285), we find

$$\begin{aligned} \overline{\mu}_4(Z'_I)_{qrst} &= E\{(Z')_q (Z')_r (Z')_s (Z')_t\} = \\ &= \frac{1}{2n} \sum_{\substack{i=-n \\ i\neq 0}}^{n} Z'_{i,q} Z'_{i,r} Z'_{i,s} Z'_{i,t} = n \sum_{i=1}^{n} e_{i,q} e_{i.r} e_{i.s} e_{i.t} = \\ &= n \sum_{i=1}^{n} \delta_{iq} \delta_{ir} \delta_{is} \delta_{it} = n \delta_{qr} \delta_{qs} \delta_{qt} \ . \end{aligned} \tag{293}$$

This is different from $\overline{\mu}_4(Z)$; in fact, it is enough to take $q = r \neq s = t$ in (278) and (293), to realize that

$$q \neq s \ , \ \ \overline{\mu}_4(Z)_{qqss} = 1 \neq \overline{\mu}_4(Z'_I)_{qqss} = 0 \ . \tag{294}$$

Indeed from the variable Z'_I one can construct the corresponding variable X' from

$$X'_I = \mu_X + TZ'_I \ , \tag{295}$$

namely a variable with a uniform distribution on

$$\arg(X'_I) = \left\{ \begin{array}{l} \mu_X + \sqrt{n}Te_i = \mu_X + \sqrt{n}T_i \\ \mu_X - \sqrt{n}Te_i = \mu_X - \sqrt{n}T_i \end{array} \right. \ i = 1, \dots, n \tag{296}$$

where T_i is the i-th column of the matrix T. Hence each σ-point $X'_{Ii} = \mu_X \pm \sqrt{n}T_i$ bears a probability $p_i = \frac{1}{2n}$, and we find

$$\left\{ \begin{array}{l} E\{X'_I\} = \mu_X \\ C_{X'_I} = TT^T = C_X \ , \end{array} \right. \tag{297}$$

as required to X'_I. In addition, by symmetry

$$\overline{\mu}_3(X'_I) = 0 \ . \tag{298}$$

On the contrary, for the fourth central moment of X'_I we have in general

$$\overline{\mu}_4(X'_I)_{qrst} \neq \overline{\mu}_4(X)_{qrst} \tag{299}$$

More precisely the inequality (299) becomes, in terms of the elements of T,

$$\begin{array}{l} n\sum_{i=1}^{n} T_{qi}T_{ri}T_{si}T_{ti} \neq (\sum_{i=1}^{n} T_{qi}T_{ri})(\sum_{i=1}^{n} T_{si}T_{ti}) + \\ (\sum_{i=1}^{n} T_{qi}T_{si})(\sum_{i=1}^{n} T_{ri}T_{ti}) + (\sum_{i=1}^{n} T_{qi}T_{ti})(\sum_{i=1}^{n} T_{ri}T_{si}) \ . \end{array} \tag{300}$$

That the inequality (300) is true in general can be verified by the elementary examples where T is diagonal, or its entries are only 1 or 0.

We pass now to study the performance of the choice X'_I, when we use it to produce the approximation

$$Y'_I = g(X'_I) = \left\{ \begin{array}{l} g(X'_i) \\ g(X'_{-i}) \end{array} \right. \ i = 1, \dots n \ , \ \ p_i = \frac{1}{2n} \tag{301}$$

to the variable Y. It is easy to predict that the error in $\mu_{Y'_I}$ and $C_{Y'_I}$ is O_4, yet we see this in detail.

Since we need only a Taylor development of $g(X)$ up to order 4, we prefer to use a simple index notation instead of one more synthetic. So we write

$$\begin{array}{rcl} Y & = & g(X) = g(\mu_X + \delta X) = g(\mu_X) + \sum_i \delta X_i g_i + \\ & & + \frac{1}{2}\sum_{ij} \delta X_i \delta X_j g_{ij} + \frac{1}{6}\sum_{ijk} \delta X_i \delta X_j \delta X_k g_{ijk} + \\ & & + \frac{1}{24}\sum_{ijk\ell} \delta X_i \delta X_j \delta X_k \delta X_\ell g_{ijk\ell} + O_5 \ ; \end{array} \tag{302}$$

here we have put

$$g_i = \frac{\partial g(\mu_X)}{\partial X_i} \ , \ g_{ij} = \frac{\sigma^2 g(\mu_X)}{\partial X_i \partial X_j} \ , \ g_{ijk} = \frac{\partial^3 g(\mu_X)}{\partial X_i \partial X_j \partial X_k} \ , \ g_{ijk\ell} = \frac{\partial^4 g(\mu_X)}{\partial X_i \partial X_j \partial X_k \partial X_\ell} \ . \tag{303}$$

Recalling that the mean of odd orders is zero, we get

$$\mu_Y = g(\mu_X) + \frac{1}{2}\sum_{ij} C_{Xij} g_{ij} + \frac{1}{24}\sum_{ijk\ell} \overline{\mu}_4(X)_{ijk\ell} g_{ijk\ell} + O_6 \ . \tag{304}$$

Therefore

$$Y-\mu_Y = \sum_i \delta X_i g_i + \frac{1}{2}\sum_{ij}(\delta X_i X_j - C_{Xij})g_{ij} + \tag{305}$$
$$+\frac{1}{6}\sum_{jk}\delta X_i \delta X_j \delta X_k g_{ijk} + O_4 \ ,$$

where we have stopped the Taylor formula at order 3 because we are going to use it in a quadratic form where terms O_4 multiplied by O_1 become at least O_5.

Then, returning to the maximum order 4, we can write

$$(Y-\mu_Y)(Y-\mu_Y)^T = \sum_{i,j}\delta X_i \delta X_j g_i g_j^T + \tag{306}$$
$$+\frac{1}{4}\sum_{ijk\ell}(\delta X_i \delta X_j - C_{Xij})(\delta X_k \delta X_\ell - C_{Xk\ell})g_{ij}g_{k\ell}^T +$$
$$+\frac{1}{6}\sum_{ijk\ell}\delta X_i \delta X_j \delta X_k \delta X_\ell (g_i g_{ik\ell}^T + g_{jk\ell}g_i^T) +$$
$$+ \text{ odd terms } + O_6 \ .$$

So, averaging we arrive at

$$C_Y = E\{(Y-\mu_Y)(Y-\mu_Y)^T\} = \sum_{ij} C_{Xij}g_i g_j^T \ + \tag{307}$$
$$+\frac{1}{4}\sum_{ijk\ell}\overline{\mu}_{Yijk\ell}(X)C_{Xij}C_{Xk\ell} - \frac{1}{4}(\sum C_{Xij}g_{ij})(\sum C_{Xk\ell}g_{k\ell})^T +$$
$$\frac{1}{6}\sum_{ijk\ell}\overline{\mu}_{Yijk\ell}(X)(g_i g_{jk\ell}^T + g_{jk\ell}g_i^T) + O_6 \ .$$

Of course, the same formulas (304) and (307) hold for both our original X and its discrete approximation X_I'. Therefore on account of the identities $\mu_X = \mu_{X_I'}, C_X = C_{X_I'}$ and of the inequality (299), between fourth order moments, we have obviously

$$\mu_Y - \mu_{Y_I'} = O_4 \ , \tag{308}$$
$$C_Y - C_{Y_I'} = O_4 \ . \tag{309}$$

There is no need to say that for the discrete approximation Y_I', given by (301), mean and covariance matrix have to be computed by the ordinary formulas

$$\mu_{Y_I'} = \frac{1}{2n}\sum_{\substack{i=-n \\ i\neq 0}}^{n} g(X_i) \ , \tag{310}$$

$$C_{Y_I'} = \frac{1}{2n}\sum_{\substack{i=-n \\ i\neq 0}}^{n}[g(X_i)-\mu_{Y_I'}][g(X_i)-\mu_{Y_I'}]^T \ . \tag{311}$$

How to improve the above estimates, taking into account the fourth order moments, will take us busy for the rest of the chapter. Before however, we have to discuss shortly a direction taken by the literature [8] which to our advise is not so elegant and well-settled as it is for the 1D case.

Remark 16. Since in the one-dimensional case the addition of the origin as argumental value of X' allows to impose the equality of the fourth order moment of X and X', the inventors of the Unscented Transform have thought to follow the same path in the multi-dimensional case. So they propose ([8], [9]) an X' with argumental values

$$1 = 1 \dots n \quad X' = \begin{cases} X_i = \mu_X + dT_i & p_i = p \\ X_0 = \mu_X & p_0 = 1 - 2np \\ T_{-i} = -T_i & p_{-i} = p \end{cases} \tag{312}$$

where the parameters to be adjusted are d and p.

Instead of p as an unknown parameter, S. Julier and J.K. Uhlmann prefer to use a parameter $\mathcal{K}$ related to p by the formula

$$p = \frac{1}{2(n+\mathcal{K})} \ ; \tag{313}$$

it is clear that since it has to be $2np < 1$, $\mathcal{K}$ so defined has to be positive. From (313) descends that

$$p_0 = \frac{\mathcal{K}}{n+\mathcal{K}} \ . \tag{314}$$

It is clear by symmetry that $\mu_X = \mu_{X'}$, so we have to impose only

$$\begin{aligned} C_{X'} &= \frac{d^2}{2(n+\mathcal{K})} \sum_{\substack{i=-n \\ i\neq 0}}^{n} T_i T_i^T = \\ &= \frac{d^2}{n+\mathcal{K}} \sum_{i=1}^{n} T_i T_i^T = C_X \ . \end{aligned} \tag{315}$$

Since $\sum_{i=1}^{n} T_i T_i^T = C_X$, we find

$$d = \sqrt{n+\mathcal{K}} \tag{316}$$

and $\mathcal{K}$ still remains undetermined. So in principle it could be used to reduce the errors in the fourth order moments of X', to make them closer to the fourth order moments of X'. On the other hand, if we construct the transformed σ points

$$Y_i' = g(X_i) \ , \ p_i = \frac{1}{2(n+\mathcal{K})} \ , \ i = \pm 1 \ldots \pm n \ , \ p_0 = \frac{\mathcal{K}}{n+\mathcal{K}} \tag{317}$$

we see that discrete mean and covariance become

$$\mu_{Y'} = \frac{1}{n+\mathcal{K}} \{\mathcal{K} g(\mu_X) + \frac{1}{2} \sum_{\substack{i=-n \\ i\neq 0}}^{n} g(X_i')\} \ , \tag{318}$$

$$\begin{aligned} C_{Y'} &= \frac{1}{n+\mathcal{K}} \{\mathcal{K}[g(\mu_X) - \mu_{Y'}][g(\mu_X) - \mu_{Y'}]^T + \\ &+ \frac{1}{2} \sum_{\substack{i=-n \\ i\neq 0}}^{n} [g(X_i') - \mu_{Y'}][g(X_i') - \mu_{Y'}]^T\} \ . \end{aligned} \tag{319}$$

If by any chance the optimal $\mathcal{K}$ is negative, what can happen, we see that (314) still retains its meaning, while $C_{Y'}$ might not be positive definite.

This leads us out of the theoretical frame of the present development and we think this is due to the particular choice of X', or of the underlying Z'

$$Z' = \begin{cases} \mu_X \pm \sqrt{n+\mathcal{K}} e_i & i = 1 \ldots n \quad p_i = \frac{1}{2(n+\mathcal{K})} \\ \mu_X & i = o \qquad p_0 \frac{\mathcal{K}}{n+\mathcal{K}} \end{cases} \ , \tag{320}$$

which has a substantial difference from Z. In fact, while the components of Z are stochastically independent, the components of Z', in (320), are not.

This is the reason why fourth order moments of X and X' cannot be the same, at least in general.

Some remedies are proposed in literature [8] trying to make the use of X', from (312), consistent. Yet we prefer another approach, that is in our opinion a more natural generalization of the 1D choices and works as easily at least for small values of n.

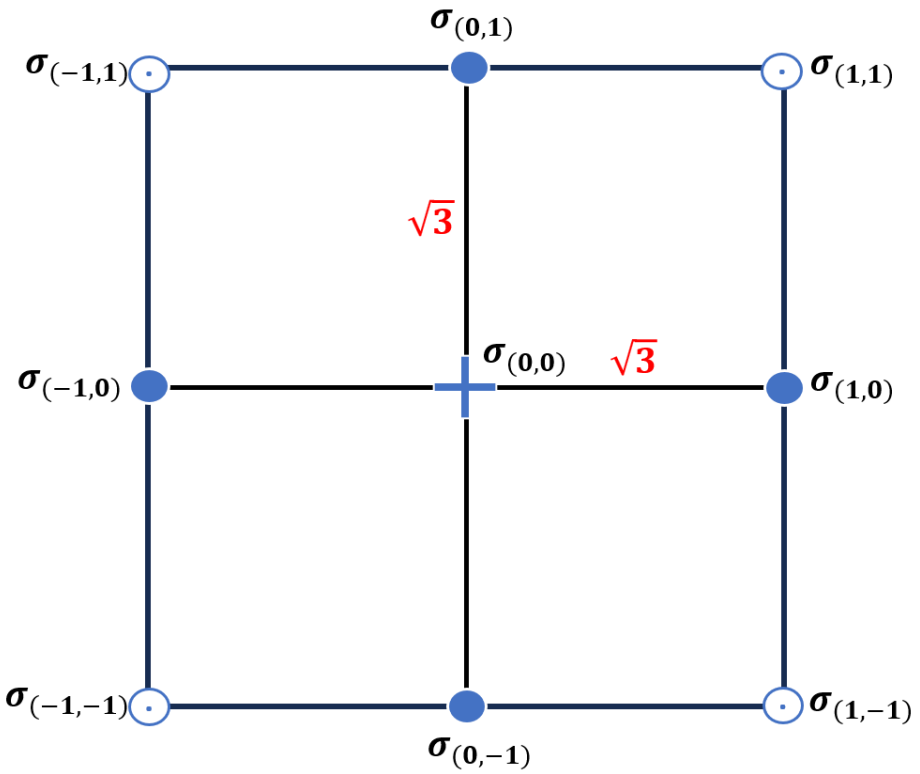


Figure 5: The grid (322) in 2D and the distribution of the sigma points:$\odot$ $p = \frac{1}{36}$, $\bullet$ $p = \frac{1}{9}$, $+$ $p = \frac{4}{9}$.

Second choice: the Normo-Kurtic Unscented Transform

We stipulate in this case that the σ-points of Z are given as follows: let's introduce a multi-index

$$\alpha = (\alpha_1, \alpha_2 \dots \alpha_n) \; ; \; \alpha_k = -1, 0, 1 \; , \tag{321}$$

thus we put

$$\arg Z'_{II} \equiv \{Z'_\alpha = \sqrt{3}\alpha \; , \; \forall \alpha\} \; . \tag{322}$$

It is clear that the grid (322) is just the Cartesian product of the grid $(-\sqrt{3}, 0, \sqrt{3})$, used in the 1D case, by itself along each coordinate axis of R^n. On such a grid it is possible to distribute probabilities so that the components of Z' become independent. It is in fact enough to put

$$P(Z_\alpha = \sqrt{3}\overline{\alpha}) = p_{\overline{\alpha}} = p_{\alpha_1} p_{\alpha_2} \dots p_{\alpha_n} \tag{323}$$

where

$$p_{\alpha_k} = \begin{cases} 1/6 & \alpha_k = 1 \\ 2/3 & \alpha_k = 0 \\ 1/6 & \alpha_k = -1 \; . \end{cases} \tag{324}$$

The situation is illustrated in Fig. 5 for a 2D case.

It is obvious that by construction the marginal distributions of Z'_{II}, i.e. the distributions of each component Z'_i of Z'_{II}, are all equal one to the other and correspond to

$$Z'_i = \begin{cases} 1 & p = 1/6 \\ 0 & p = 2/3 \\ -1 & p = 1/6 \end{cases} \; . \tag{325}$$

Accordingly we have

$$\forall i \ , \quad E\{Z_i'\} = 0 \ , \ E\{Z_i'^2\} = 1 \ , \ E\{Z_i'^4\} = 3 \ . \tag{326}$$

Moreover, always by construction, any two different components of Z' are independent, so

$$i \neq j \quad E\{Z_i' Z_j'\} = E\{Z_i'\} E\{Z_j'\} = 0 \ . \tag{327}$$

Clearly all third order moments are zero, as it is evident also for symmetry reasons. As for fourth order moments, also thanks to the stochastic independence, we have

$$i \neq j \neq k \neq \ell \ E\{Z_i' Z_j' Z_k' Z_\ell'\} = 0 \ ; \tag{328}$$

similarly in case that two components are equal but the others are different, one has

$$i \neq k \neq \ell \quad E\{Z_i'^2 Z_k' Z_\ell'\} = 0 \ . \tag{329}$$

On the contrary, when we have the components equal in couples, we get

$$E\{Z_j'^2 Z_k'^2\} = E\{Z_j'^2\} \cdot E\{Z_k'^2\} = 1 \ . \tag{330}$$

Likewise if we have a cube of one coordinate and a first order of another coordinate, we get

$$E\{Z_i^3 Z_k\} = 0 \tag{331}$$

Finally fourth powers of a single coordinate are already given in (326).
Putting together all relations from (326) to (331) we see that

$$\begin{cases} \mu_{Z_{II}'} = 0 \ , \ M_3(Z_{II}') = 0 \\ M_2(Z_{II}') = \{\delta_{ik}\} = I_{(n)} \\ M_4(Z_{II}')_{ijk\ell} = \delta_{ij}\delta_{d\ell} + \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk} \ , \end{cases} \tag{332}$$

exactly as we have for the first 4 moments of the standard normal Z. Based on these results, we can at once maintain that

$$X_{II}' = \mu_X + T Z_{II}' \ , \ X = \mu_X + T Z \ , \tag{333}$$

do have common moments up to the fourth. Even more, if we put

$$Y_{II}' = g(X_{II}') \ , \tag{334}$$

by going back to (304), (307), exploiting the equality of the 4 moments we have established the following proposition.

Proposition 8. *With the choice (334), (333), under the hypothesis that X is normal and g a vector differentiable up to the sixth order, we have*

$$\mu_Y - \mu_{Y_{II}'} = O_6 \ , \tag{335}$$

$$C_Y - C_{Y_{II}} = O_6 \ . \tag{336}$$

This is exactly the result we were looking for. Indeed the method is not devoid of critical points which we examine now.

Remark 17. One could object that Proposition 8 holds only under the hypothesis that X is a normal variate. Yet the argument has been already discussed in Remark 13, and here the same arguments hold. It is clear that for a non-Gaussian random variable X, (335), (336) can hold only approximately. In case of a symmetric variable X, we still have O_4 errors for mean and covariance and this can be achieved already with the simpler choice of X_I'. On the other hand, when X is seriously skewed, i.e. it has non-zero third moments, then only an O_3 error estimate can be deemed. In any event the purpose of improving on the simple linearized formulas, is achieved, with the advantage of avoiding computations of the derivatives of g up to the second order.

Remark 18. More serious is the objection that with the choice X'_{II} we have to determine 3^n σ-points, while for the choice X'_I this figure is only $2n$ or $2n+1$ if we add the origin). Yet, as anticipated, when the number of degrees of freedom of X is small, for instance $n = 3$ as the position of a point in space, the difference is certainly manageable from the computational point of view. On the other hand, when n is as large as 10 or so, the number of σ-points increases to a level that makes a simple Montecarlo Method competitive with the choice for the Unscented Transform.

A remedy to this drawback can come from a circumstance not infrequently met. Namely, assume that the random variable X can be split into two groups of variables

$$X = \left| \begin{array}{c} \xi \\ \eta \end{array} \right| , \tag{337}$$

such that in the range of variability of η the non-linear function

$$g(X) = g(\xi, \eta) , \tag{338}$$

can be linearized in η, with a second order error that we consider negligible. Then instead of (338) we can write

$$\begin{array}{rcll} g(X) & = & g(\mu_\xi + \delta\xi, \mu_\eta + \delta\xi) \cong & (339) \\ & = & g(\mu_\xi + \delta\xi, \mu_\eta) + g_\eta(\mu_\xi + \delta\xi, \mu_\eta)\delta\eta \ ; & \end{array}$$

if we further assume that g_η is slowly varying with $\delta\xi$ and considering that g_η is already multiplied buy the small residual $\delta\eta$, we can write a simplified model of (338) as

$$g(X) \cong g(\mu_\xi + \delta\xi, \mu_\eta) + g_\eta(\mu_\xi, \mu_\eta)\delta\eta \ . \tag{340}$$

Such a model mixes a non-linear dependence on ξ, which we assume to be necessary to keep, given the particular combination of the size of $\delta\xi$ and of the sensitivity of g with respect to this variable, and a linear dependence on η. So in synthesis we assume to have a model with the following form:

$$Y = f(\xi) + H\eta \tag{341}$$

and we want to propagate mean and covariance by knowing μ_ξ, μ_η and

$$C_X = \left| \begin{array}{cc} C_\xi & C_{\xi\eta} \\ C_{\eta\xi} & C_\eta \end{array} \right| . \tag{342}$$

As before it is convenient to reduce the problem to random variables that decouple (ξ, η), namely we define the block-Cholesky decomposition of C_X as

$$C_X = TT^T \ , \ T = \left| \begin{array}{cc} T_\xi & 0 \\ T_{\eta\xi} & T_\eta \end{array} \right| , \tag{343}$$

and we define two independent standard normal variates (ω, ζ) by

$$X = \left| \begin{array}{c} \xi \\ \eta \end{array} \right| \ = \ T \left| \begin{array}{c} \omega \\ \zeta \end{array} \right| \ = \ \left| \begin{array}{c} T_\xi\omega \\ T_{\eta\xi}\omega + T_\eta\zeta \end{array} \right| . \tag{344}$$

Substituting in (341) we get

$$\begin{array}{l} Y = F(\omega) + K\zeta \ , \\ (F(\omega) = f(\mu_\xi + T_\xi\omega) + HT_{\eta\xi}\omega \ , \ K = HT_\eta) \ . \end{array} \tag{345}$$

Now, since ω is stochastically independent of ζ, from (345) we can write

$$\left\{ \begin{array}{l} \mu_Y = \mu_F + K\mu_\zeta \\ C_Y = C_F + KC_\zeta K^T \end{array} \right. \ ; \tag{346}$$

in this way the problem has reduced to the estimate of μ_F, C_F for the non-linear function

$$U = F(\omega) \ . \tag{347}$$

If the dimension of ω is small, we can then apply the Normo-Kurtic Unscented Transform to this part of the problem, so finding the solution of (346) with what we could call a Partial Unscented Transform.

6 The Unscented Kalman Filter

We can return now to the non-linear problem, expressed in Chapter 3 by equations (187), (188), (189) and set up a solution, based on the idea of the Unscented Transform, outperforming the linearized approach of the Extended Kalman Filter.

Let us recall that the case of the Unscented Transform concept is to substitute any R.V. X with a discrete R.V. X' having argumental values $(X')_i$ with probabilities p_i such that X' and X have the same means and covariances, whatever is the chosen transform. When X is a normal variate, with a suitable choice of the Unscented Transform, the errors in the mean and covariance are O_6, but even in the worst case they are O_3, which is better than the O_2 errors of a linearized theory.

Here we will always assume normality of our random variables, that has a main consequence that two orthogonal zero mean variables $X - \mu_X, Y - \mu_Y$, i.e.

$$C_{XY} = E\{(X - \mu_X)(Y - \mu_Y)^T\} = 0 \ , \tag{348}$$

are also stochastically independent, so that, recalling (27),

$$E\{F(X)|_Y\} \equiv E\{F(X)\} \tag{349}$$

for almost all Y, with probability 1, and whatever $F(X) \in \mathcal{L}^2(X)$.

Furthermore, one has that the optimal approximation of X given Y is also the regressor of X on Y, namely (see Proposition 4 in Chapter 1)

$$\widehat{X} = E\{X|_Y\} = X|_Y = \mu_X + C_{XY}C_Y^{-1}(Y - \mu_Y) \tag{350}$$

with error variance (covariance)

$$\widehat{\varepsilon} = \widehat{X} - X \ , \ C_{\widehat{\varepsilon}} = C_{XY}C_Y^{-1}C_{YX} \ . \tag{351}$$

Alternatively we could say that we continue with the choice made in Chapter 2 to restrict ourselves to the search of regression predictors of our variables, what leads us directly to (350), (351).

With the above proviso, let us return to the problem of predicting $\widehat{X}_{t+1}$, that, recalling (203), we write as

$$\widehat{X}_{t+1} = \widetilde{X}_{t+1} + E\{X_{t+1} - \widetilde{X}_{t+1}|\delta Y_{t+1}\} \ . \tag{352}$$

As in the previous chapter

$$\delta Y_{t+1} = Y_{t+1} - \widetilde{Y}_{t+1} \tag{353}$$

and

$$\widetilde{Y}_{t+1} = Y_{t+1}|_{I_t} = E\{Y_{t+1}|_{I_t}\} \ , \tag{354}$$

that we will elaborate here below.

Furthermore

$$\begin{aligned} \widetilde{X}_{t+1} &= X_{t+1}|_{I_t} = E\{X_{t+1}|I_t\} = \\ &= E\{g_{t+1}(\widehat{X}_t + \widehat{\varepsilon}_t) + Q_{t+1}\nu_{t+1}|_{I_t}\} \end{aligned} \tag{355}$$

On the other hand $\nu_{t+1}|_{I_t} = E\{\nu_{t+1}|_{I_t}\} = 0$, while by conditioning $g_{t+1}(\widehat{X}_t + \widehat{\varepsilon}_t)$ to I_t we fix $\widehat{X}_t$ (which is in $\mathcal{L}^2(I_t)$) and let only $\widehat{\varepsilon}_t$ to vary. But $\widehat{\varepsilon}_t$ is orthogonal to $\mathcal{L}^2(I_t)$ and so independent of I_t, so that the conditioning on I_t does not act on $\widehat{\varepsilon}_t$ and the conditional mean in (355) becomes an unconditional mean

$$\widetilde{X}_{t+1} = E_{\widehat{\varepsilon}_t}\{g_{t+1}(\widehat{X}_t + \widehat{\varepsilon}_t)\} \ . \tag{356}$$

Formula (356) is the mean of a non-linear transformation of $\widehat{\varepsilon}_t$ and we know how to approximate it, by the theory of Chapter 5. In fact, since $E\{\widehat{\varepsilon}_t\} = 0$ and $C_{\widehat{\varepsilon}_t}$ is known we can build the σ-points of $\widehat{\varepsilon}_t$ substituting it with an unscented variable $\widehat{\varepsilon}'_t$, having the same 0 mean and the same covariance. So (356) becomes

$$\widetilde{X}_{t+1} \cong E'\{g_{t+1}(\widehat{X}_t + \widehat{\varepsilon}'_t)\} \tag{357}$$

where we have denoted with E' the average with respect to $\widehat{\varepsilon}'_t$.

Switching to the prediction of $E\{X_{t+1} - \widetilde{X}_{t+1}|_{\delta Y_{t+1}}\}$, we first observe that

$$E\{X_{t+1} - \widetilde{X}_{t+1}|_{\delta Y_{t+1}}\} = E\{\widetilde{\varepsilon}_{t+1}|_{\delta Y_{t+1}}\} \tag{358}$$

On the other hand, we know that

$$E\{\widetilde{\varepsilon}_{t+1}|_{\delta Y_{t+1}}\} = C_{\widetilde{\varepsilon}_{t+1}\delta Y_{t+1}} C^{-1}_{\delta Y_{t+1}} \delta Y_{t+1} \ , \tag{359}$$

according to the (approximate) normal distribution hypothesis of all the variates. So we need $\widetilde{Y}_{t+1}$, to compute δY_{t+1}, and $C_{\widetilde{\varepsilon}_{t+1}\delta Y_{t+1}}, C_{\delta Y_{t+1}}$. For this purpose we need to transform $\widetilde{\varepsilon}_{t+1}$ into $\widetilde{\varepsilon}'_{t+1}$, i.e., since $E\{\widetilde{\varepsilon}_{t+1}\} = 0$, we need $C_{\widetilde{\varepsilon}_{t+1}}$.

As $\widetilde{\varepsilon}_{t+1}$ is independent of I_t and $g_{t+1}(X_t)$ is independent of ν_{t+1}, recalling also (349), we get

$$\begin{aligned} C_{\widetilde{\varepsilon}_{t+1}} &= E\{(X_{t+1} - \widetilde{X}_{t+1})(X_{t+1} - \widetilde{X}_{t+1})^T\} = \\ &= E\{(X_{t+1} - \widetilde{X}_{t+1})(X_{t+1} - \widetilde{X}_{t+1})^T|_{I_t}\} = \\ &= E\{[g_{t+1}(\widehat{X}_t + \widehat{\varepsilon}_t) - \widetilde{X}_{t+1}][g_{t+1}(\widehat{X}_t + \widehat{\varepsilon}_t) - \widetilde{X}_{t+1}]^T\} + \\ &\quad + Q_{t+1} C_{\nu_{t+1}} Q^T_{t+1} \ ; \end{aligned} \tag{360}$$

let us underline that in the last line of (360) $\widehat{X}_t$ has to be considered as a constant, due to the conditioning on I_t, while the average E applies only to $\widehat{\varepsilon}_t$. Now we substitute $\widehat{\varepsilon}_t$ with the unscented $\widehat{\varepsilon}'_t$, that we have already constructed to compute $\widetilde{X}_{t+1}$, and E with E'. Therefore (360), with its last term, becomes a computable formula and $C_{\widetilde{\varepsilon}_{t+1}}$ is known, with the approximation pertaining to the chosen Unscented Transform.

For reasons that will become soon clear, it is convenient to compute already now the cross-variance $C_{\widetilde{\varepsilon}_{t+1}\widehat{\varepsilon}_t}$.

We have, by reasoning always conditionally to I_t,

$$\begin{aligned} C_{\widetilde{\varepsilon}_{t+1}\widehat{\varepsilon}_t} &= E\{\widetilde{\varepsilon}_{t+1}\widehat{\varepsilon}^T_t\} = E\{[g_{t+1}(\widehat{X}_t + \widehat{\varepsilon}_t) - \widetilde{X}_{t+1} + \nu_{t+1}]\widehat{\varepsilon}^T_t|_{I_t}\} = \\ &= E'\{[g_{t+1}(\widehat{X}_t + \widehat{\varepsilon}_t) - \widetilde{X}_{t+1}]\widehat{\varepsilon}^T_t\} \ , \end{aligned} \tag{361}$$

where E' means that the average has to be taken with respect to $\widehat{\varepsilon}_t$ only and $\widehat{X}_t, \widetilde{X}_{t+1}$ have to be considered as constant vectors. Indeed formula (361) can be implemented by using the σ-points of $\widehat{\varepsilon}$.

At this point we have the full covariance matrix

$$C_{\left|\begin{matrix}\widetilde{\varepsilon}_{t+1}\\ \widehat{\varepsilon}_t\end{matrix}\right|} = \left|\begin{matrix} C_{\widetilde{\varepsilon}_{t+1}} & C_{\widetilde{\varepsilon}_{t+1}\widehat{\varepsilon}_t} \\ C_{\widehat{\varepsilon}_t\widetilde{\varepsilon}_{t+1}} & C_{\widehat{\varepsilon}_t}\end{matrix}\right| \tag{362}$$

and so, also recalling that both variables have zero mean, we can define the joint σ-points of $\left|\begin{matrix}\widetilde{\varepsilon}_{t+1}\\ \widehat{\varepsilon}_t\end{matrix}\right|$ as well as its unscented counterpart $\left|\begin{matrix}\widetilde{\varepsilon}'_{t+1}\\ \widehat{\varepsilon}'_t\end{matrix}\right|$; one has to notice that these new $\widehat{\varepsilon}'_{t,i}$ do not coincide with those previously defined, yet if we reverse the order and build the block matrix $C_{\widehat{\varepsilon}_t}, C_{\widehat{\varepsilon}_t\widetilde{\varepsilon}_{t+1}}, C_{\widetilde{\varepsilon}_{t+1}}$ to compute the σ-points of $\left|\begin{matrix}\widehat{\varepsilon}_t\\ \widetilde{\varepsilon}_{t+1}\end{matrix}\right|$, we see that we can profitably use the Cholesky decomposition of $C_{\widehat{\varepsilon}_t}$ and the already computed σ-points of $\widehat{\varepsilon}_t$. At this point we can switch to compute $\widetilde{Y}_{t+1}$ by

$$\begin{aligned} \widetilde{Y}_{t+1} &= E\{h_{t+1}(X_{t+1})|_{I_t}\} + E\{\eta_{t+1}|_{I_t}\} = \\ &= E\{h_{t+1}(\widetilde{X}_{t+1} + \widetilde{\varepsilon}_{t+1})|_{I_t}\} = \\ &\cong E'\{h_{t+1}(\widetilde{X}_{t+1} + \widetilde{\varepsilon}'_{t+1})\} \ , \end{aligned} \tag{363}$$

where $\widetilde{X}_{t+1}$ is constant (known) and E' applies only to $\widetilde{\varepsilon}'_{t+1}$. Therefore $\widetilde{Y}_{t+1}$ can now be computed and so δY_{t+1} is known too.

Then we can write

$$\delta Y_{t+1} = h_{t+1}(X_{t+1}) - \widetilde{Y}_{t+1} + \eta_{t+1} \ . \tag{364}$$

where $\widetilde{Y}_{t+1} = E\{Y_{t+1}|_{I_t}\}$. Indeed $h_{t+1} - \widetilde{Y}_{t+1}$ and η_{t+1} are each independent of I_t; moreover, they are clearly independent of one another.

Therefore

$$C_{\delta Y_{t+1}} = C_{h_{t+1}-\widetilde{Y}_{t+1}} + C_{\eta_{t+1}} \ . \tag{365}$$

On the other side, applying again (349),

$$\begin{aligned} C_{h_{t+1}-\widetilde{Y}_{t+1}} &= E\{[h_{t+1}(X_{t+1}) - \widetilde{Y}_{t+1}][h_{t+1}(X_{t+1}) - \widetilde{Y}_{t+1}]^T\} = \\ &= E\{[h_{t+1}(X_{t+1}) - \widetilde{Y}_{t+1}][h_{t+1}(X_{t+1}) - \widetilde{Y}_{t+1}]^T|_{I_t}\} = \\ &= E\{[h_{t+1}(\widetilde{X}_{t+1} + \widetilde{\varepsilon}_{t+1}) - \widetilde{Y}_{t+1}][h_{t+1}(\widetilde{X}_{t+1} + \widetilde{\varepsilon}_{t+1}) - \widetilde{Y}_{t+1}|_{I_t}\} = \\ &\cong E'\{[h_{t+1}(\widetilde{X}_{t+1} + \widetilde{\varepsilon}'_{t+1}) - \widetilde{Y}_{t+1}][h_{t+1}(\widetilde{X}_{t+1} + \widetilde{\varepsilon}'_{t+1}) - \widetilde{Y}_{t+1}\} \ . \end{aligned} \tag{366}$$

Once more, under E', $\widetilde{X}_{t+1}$ and $\widetilde{Y}_{t+1}$ have to be considered constant due to conditioning. Then $C_{h_{t+1}-\widetilde{Y}_{t+1}}$ can be computed and $C_{\delta Y_{t+1}}$ too. Finally, since we have already constructed the Unscented, discrete variables $\widetilde{\varepsilon}'_{t+1}, \widehat{\varepsilon}'_t$ it becomes straightforward to compute

$$C_{\widetilde{\varepsilon}_{t+1}\delta Y_{t+1}} = E\{\widetilde{\varepsilon}_{t+1}\delta Y_{t+1}^T\} \ . \tag{367}$$

In fact, since $\widetilde{\varepsilon}_{t+1}, \delta Y_{t+1}$ are independent of I_t,

$$\begin{aligned} E\{\widetilde{\varepsilon}_{t+1}\delta Y_{t+1}^T\} &= E\{[g_{t+1}(\widehat{X}_t + \widehat{\varepsilon}_t) - \widetilde{X}_{t+1} + \nu_{t+1}][h_{t+1}(\widetilde{X}_{t+1} + \widetilde{\varepsilon}_{t+1}) + \\ &\quad -\widetilde{Y}_{t+1} + \eta_{t+1}]^T|_{I_t}\} = \\ &= E\{[g(\widehat{X}_t + \widehat{\varepsilon}_t) - \widetilde{X}_{t+1}][h_{t+1}(\widetilde{X}_{t+1} + \widetilde{\varepsilon}_{t+1}) - \widetilde{Y}_{t+1}]^T\} \ , \end{aligned} \tag{368}$$

and this average can be computed with the joint unscented variable $\begin{vmatrix} \widetilde{\varepsilon}'_{t+1} \\ \widehat{\varepsilon}'_t \end{vmatrix}$. As before, $\widehat{X}_t, \widetilde{X}_{t+1}$ have to be taken as constant in (368).

Wrapping up into (352), (359), we have all the elements to compute

$$\widehat{X}_{t+1} = \widetilde{X}_{t+1} + C_{\widetilde{\varepsilon}_{t+1}\delta Y_{t+1}} C^{-1}_{\delta Y_{t+1}} \delta Y_{t+1} \ . \tag{369}$$

To conclude our step from time t to time $t+1$ we need the covariance matrix $C_{\widehat{\varepsilon}_{t+1}}$. Here we reason as we did in Chapter 4; from (369) we have

$$\widehat{\varepsilon}_{t+1} = \widetilde{\varepsilon}_{t+1} - C_{\widetilde{\varepsilon}_{t+1}\delta Y_{t+1}} C^{-1}_{\delta Y_{t+1}} \delta Y_{t+1} \ , \tag{370}$$

or, setting $L_{t+1} = C_{\widetilde{\varepsilon}_{t+1}\delta Y_{t+1}} C^{-1}_{\delta Y_{t+1}}$,

$$\widetilde{\varepsilon}_{t+1} = \widehat{\varepsilon}_{t+1} + L_{t+1}\delta Y_{t+1} \ . \tag{371}$$

But $\delta Y_{t+1} \in \mathcal{L}^2(N_{t+1}, I_{t+1})$ while $\widehat{\varepsilon}_{t+1}$ is orthogonal to this space, so that

$$\begin{aligned} C_{\widetilde{\varepsilon}_{t+1}} &= C_{\widehat{\varepsilon}_{t+1}} + L_{t+1} C_{\delta y_{t+1}} L^T_{t+1} = \\ &= C_{\widehat{\varepsilon}_{t+1}} + C_{\widetilde{\varepsilon}_{t+1}\delta Y_{t+1}} C^{-1}_{\delta Y_{t+1}} C_{\delta Y_{t+1}\widetilde{\varepsilon}_{t+1}} \ , \end{aligned} \tag{372}$$

which gives

$$C_{\widehat{\varepsilon}_{t+1}} = C_{\widetilde{\varepsilon}_{t+1}} - C_{\widetilde{\varepsilon}_{t+1}\delta Y_{t+1}} C^{-1}_{\delta Y_{t+1}} C_{\delta Y_{t+1}\widetilde{\varepsilon}_{t+1}} \ , \tag{373}$$

where all matrices have already been computed and are therefore known. The step of transition from t to $t+1$ of the Unscented Kalman Filter is completed. To summarize the procedure we represent it in Fig. 6.

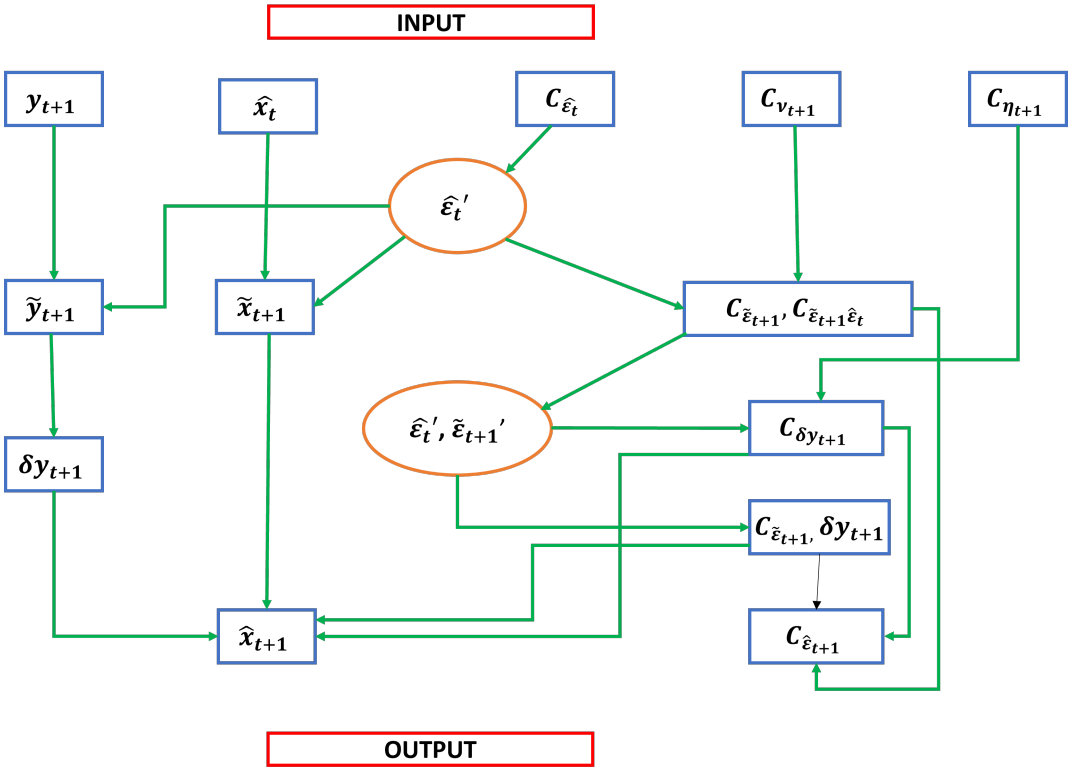


Figure 6: Scheme of the Unscented Kalman Filter; rectangles: computed quantities; ellipses: Discrete Random Variables.

Remark 19. As already stated in Proposition 7 of Chapter 2, an equivalent interpretation of Kalman Filtering can be given in terms of a least squares solution of the updating, based on the predicted state $\widetilde{X}_{t+1}$ and its error covariance matrix $C_{\widetilde{\varepsilon}_{t+1}}$, and on the innovation information δY_{t+1} with its error covariance matrix $C_{\eta_{t+1}}$. In Chapter 2 the context was that of the linear Kalman Filter, yet the same remark holds in the framework of non-linear filtering, although the least squares method is difficult to apply here for a general non-linear model of the dynamics. In fact, let us recall that by definition

$$\widetilde{X}_{t+1} = E\{g_{t+1}(\widehat{X}_t + \widehat{\varepsilon}_t)|_{I_t}\} \tag{374}$$

where the conditioning to I_t makes $\widehat{X}_t$ a constant vector, while $\widehat{\varepsilon}_t$ is a zero mean normal variate with covariance $C_{\widehat{\varepsilon}_t}$.

So for a general vector function $g_{t+1}(\cdot)$ it is difficult to compute (374), if not impossible in exact form. Nevertheless, if the dynamic model is linear, as it happens when it represents just a "smoothing" of the trajectory $\{X_t\}$, since $E\{\widehat{\varepsilon}_t\} = 0$, one can put

$$\widetilde{X}_{t+1} = D_{t+1}\widehat{X}_t \ , \tag{375}$$

i.e. $\widetilde{X}_{t+1}$ is computable and the least squares approach can be followed without any further ado even if the observation equations

$$Y_{t+1} = h_{t+1}(X_t) + \eta_{t+1} \tag{376}$$

have a non-linear form.

Example 7. A trolley is moving along a rectilinear rail with constant velocity and random fluctuations, as illustrated in Figure (7). At time $t = 0$ the trolley is moving from the origin, $X_0 = -120m$, position known without error. The trolley is moving with constant velocity of $1m/s$ along x axis, with the addition of a white noise of $\pm 0.1m$ (Figure 8).

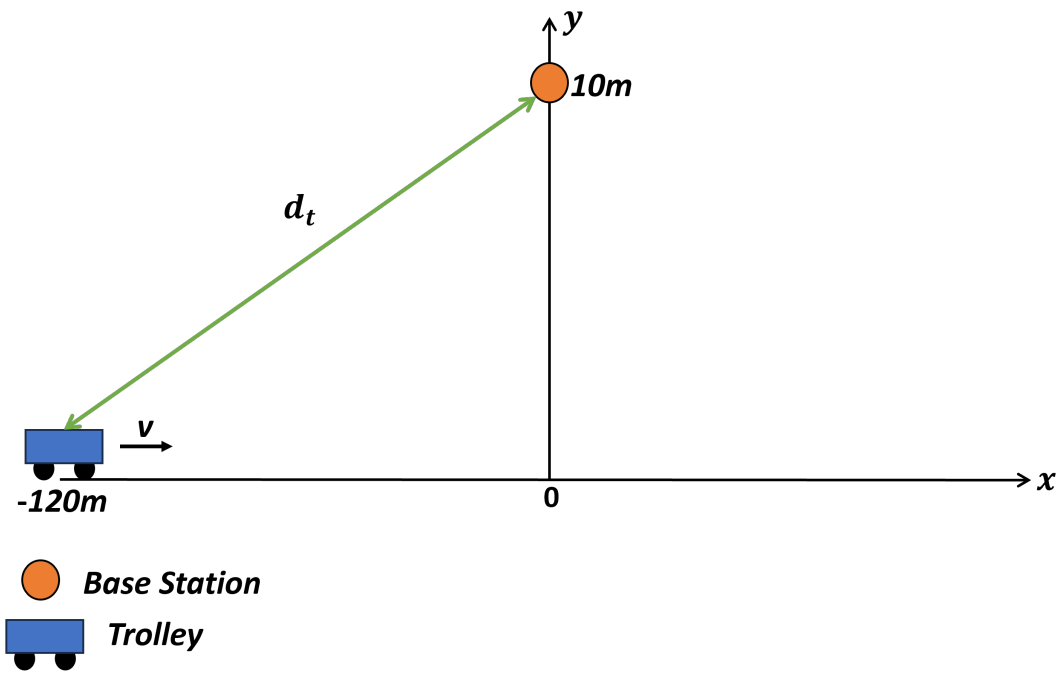


Figure 7: Example schema.

The objective is to set up a Kalman Filter to get at every time the best estimate of the position, based on all the previous measurements. Given the presence of a station at $[0, 10]m$, our Kalman Filter's observations will be the distances between the trolley and the station at each epoch: this is analogous to Time of Arrival (ToA) measurements. To make the scenario more practical , these observations are intentionally imbued with noise: the measurement noise level is equal to $\pm 0.1m$, mirroring real-world conditions. In this example, the model is represented by the equation

$$X_{t+1} = D_{t+1}X_t + u_{t+1} + Q_{t+1}v_{t+1} \tag{377}$$

while the observations are expressed through the equation

$$Y_{t+1} = h(X_{t+1}) + \nu_{t+1}; \tag{378}$$

where:

$$X = \left|x\right|, D = \left[1\right], u = \left[1m/s \cdot 1s\right], Q = \left[1\right], h = \sqrt{x^2 + y_{BS}^2}. \tag{379}$$

The problem is linear in the dynamic model but not linear in the observation equations. Note that the model of the example is very simple, just to allow an elementary check. Note also that the velocity is kept as a known constant while, typically, in more realistic applications is itself part of the state vector to be estimated. Commencing from these noise-affected distances, we estimated the position estimation of the trolley by employing the three different Kalman algorithms described in the text:

- Extended Kalman Filter (EKF),
- Unscented Kalman Filter according to Julier and Uhlmann (2004)(JU-UKF),
- Normo-Kurtic Unscented Kalman Filter as discussed in this paper (NK-UKF).

In this specific example, JU-UKF generates three σ points as follows:

$$\begin{aligned} &\chi_0 = \hat{\mathbf{x}}_{t-1} \\ &\chi_1 = \hat{\mathbf{x}}_{t-1} + \sqrt{2+\lambda} \cdot T_1 \quad \chi_2 = \hat{\mathbf{x}}_{t-1} - \sqrt{2+\lambda} \cdot T_1 \end{aligned} \tag{380}$$

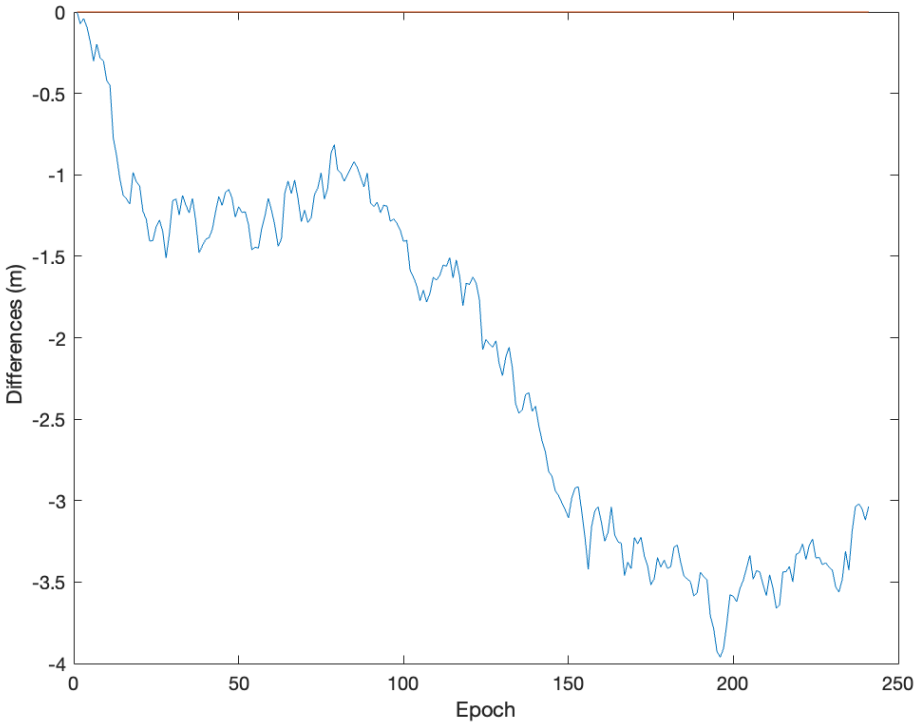


Figure 8: Trajectory: differences (random walk) between the simulated and the nominal trajectory.

λ is a scaling parameter with $\lambda \in \mathbb{R}$ computed according to Julier and Uhlmann and T_i represents the i-th column of matrix T which is derived from a Cholesky decomposition applied to the covariance matrix. In our case:

$$T = T_1 = \left[0.1\right]. \tag{381}$$

Similarly, NK-UKF generates three σ points as follows:

$$\begin{aligned} &\chi_0 = \hat{\mathbf{x}}_{t-1} \\ &\chi_1 = \hat{\mathbf{x}}_{t-1} + \sqrt{3} \cdot T_1 \quad \chi_2 = \hat{\mathbf{x}}_{t-1} - \sqrt{3} \cdot T_1 \end{aligned} \tag{382}$$

In the first case the model's standard deviation v and the standard deviation of the observations ν are set to $0.1m$ each. We conducted a comparative analysis of the three filters to assess their accuracy.

Error	**EKF**	**JU-UKF**	**NK-UKF**
Mean (m)	-0.02	-0.02	-0.04
Standard deviation (m)	0.11	0.10	0.12
Max (m)	0.53	0.54	0.55

Table 1: Example 7. Error statistics using EKF, JU-UKF and NK-UKF.

Table 1 displays the statistics for the error in x-component for all the employed algorithms. In Figure 9, you can observe the corresponding plots of the errors for the three filters.

In conclusion, as it was easy to predict, at the level of this elementary example the results of the three methods are very similar. As expected, the maximum errors are present around $x = 0$, where the observation equations are ill conditioned with respect to the unknown. In particular, no advantage is visible due to the use of the Normo Kurtic unscented filter. More tests will be needed to investigate more complex scenarios.

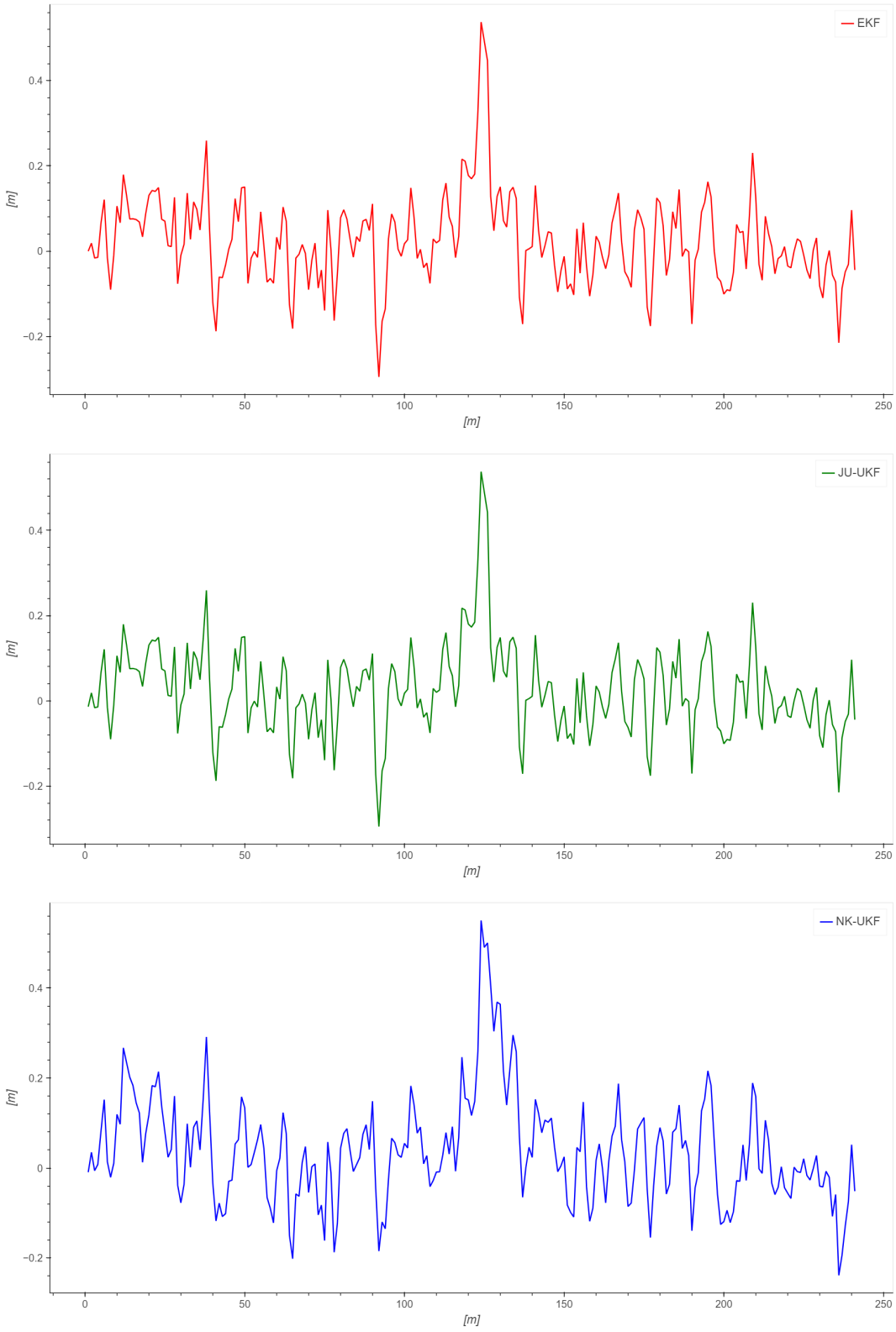


Figure 9: Example 7. X Errors for EKF (above) JU-UKF (middle) and NK-UKF (bottom) filters.

7 Appendix

In the text we have often used the hypothesis of a normal approximation of the distribution of (X, Y) with the purpose of using the approximate relation

$$E\{X|_Y\} \cong X|_Y = \mu_X + C_{XY}C_Y^{-1}(Y - \mu_Y) \ . \tag{383}$$

In reality the quality of the approximation in (383) depends on how much "linear" is

$$\overline{X}(Y) = E\{X|_Y\} \tag{384}$$

at least in a zone around μ_Y with very high probability. To quantify such error is relatively easy. Let us call

$$X(Y) = X|_Y \ , \tag{385}$$

then we can set up an index for the linearization error, with the formula

$$L\mathcal{E}^2 = \underset{Y}{E}\{|\overline{X}(Y) - X_L(Y)|^2\} \tag{386}$$

To compute (386) we consider the residual errors of the approximations of X with $\overline{X}(Y)$ and with $X_L(Y)$,

$$\varepsilon = X - \overline{X}(Y) \ , \ \varepsilon_L = X - X_L(Y) \tag{387}$$

and their mean squared values

$$E\{|\varepsilon|^2\} = Tr(C_\varepsilon) \tag{388}$$

$$E\{|\varepsilon_L|^2\} = Tr(C_{\varepsilon_L}) = Tr(C_{XY}C_Y^{-1}C_{YX}) \ . \tag{389}$$

Now we observe that

$$\begin{aligned} Tr(C_{\varepsilon_L}) &= E\{|X - X_L(Y)|^2\} = E\{|X - \overline{X}(Y) + \overline{X}(Y) - X_L(Y)|^2\} = \\ &= E\{|X - \overline{X}(Y)|^2\} + E\{|\overline{X}(Y) - X_L(Y)|^2\} + \\ &\quad +2E\{[\overline{X}(Y) - X_L(Y)]^T[X - \overline{X}(Y)]\} \\ &= Tr(C_\varepsilon) + L\mathcal{E}^2 \ , \end{aligned} \tag{390}$$

because, considering that $\overline{X}(Y) - X_L(Y)$ is function of Y only,

$$\begin{aligned} &E\{[\overline{X}(Y) - X_L(Y)]^T[X - \overline{X}(Y)]\} = \\ &= \underset{Y}{E}\{[\overline{X}(Y) - X_L(Y)]^T E\{X - \overline{X}(Y)|_Y\}\} = 0 \ . \end{aligned} \tag{391}$$

Therefore (390) gives

$$L\mathcal{E}^2 = Tr(C_{\varepsilon_L}) - Tr(C_\varepsilon) \ . \tag{392}$$

An elementary example will help appreciating the above discussion.

Example 8. Let $(X, Y) \in R^2$ have a p.d.f. defined by

$$\begin{aligned} &f_{XY}(x, y) = f_X|_Y(x|_y) f_Y(y) \\ &f_Y(y) = 1/2\chi_{[-1,1]}(y) = \begin{cases} 1/2 & |y| < 1 \\ 0 & |y| > 1 \end{cases} \\ &f_{X|_Y}(x|_y) \sim \mathcal{N}(\overline{X}(y), 1) \\ &\overline{X}(y) = y + \ 0.1 \ y^2 \end{aligned}$$

so that $\overline{X}(y)$ is "almost" linear where $f_Y(y) \neq 0$.

An elementary computation gives

$$\mu_Y = 0$$
$$\mu_X = \underset{Y}{E}\{\overline{X}(Y)\} = \frac{0.1}{3}$$
$$\sigma_{XY} = E\{XY\} = \underset{Y}{E}\{Y\overline{X}(Y)\} = \frac{1}{3}\ ,\ \sigma^2(Y) = \frac{1}{3}$$

so that the linear regressor of X on Y is

$$X_L(Y) = \frac{0.1}{3} + \frac{\sigma_{XY}}{\sigma_Y^2} Y$$

Therefore

$$\overline{X}(Y) - X_L(Y) = \ 0.1\ \left(Y^2 - \frac{1}{3}\right)$$

and

$$L\mathcal{E}^2 = \ 0.01\ \underset{Y}{E}\left\{Y^4 - \frac{2}{3}Y^2 + \frac{1}{9}\right\} = \ 0.009\ \ ,$$

namely $L\mathcal{E} = \ 0.094$, which is small compared to

$$\sigma_{\varepsilon_L} = \frac{\sigma_{XY}}{\sigma_Y} = \ 0.577\ .$$

References

[1] Pierre Del Moral (2013) Mean Field Simulation for Monte Carlo Integration. Monograph on statistics and applied probability. Chapman and Hall/CRC.

[2] A. Dermanis, R. Rummel (2000) Data analysis methods in geodesy. In: Geomatic Methods for the Analysis of Data in the Earth Sciences LNES. Springer Verlag, Berlin Heidelberg.

[3] G. Evensen (2007) Data assimilation: the Ensamble Kalman Filter. Springer Verlag, Berlin.

[4] M. Ghil, P. Rizzoli Malanotte (1991) Data assimilation in Meteorology and Oceanography. Adv. in Geohpysics v33. Academic Press Inc.

[5] Mohinder S. Grewal, Angus P. Andrews (2001) Kalman Filtering: Theory and Practice with MATLAB, 2nd Ed. Wiley and Sons Inc., New York

[6] F. Gustafsson, G. Hendeby (2012) Some Relations Between Extended and Unscented Kalman Filters. Signal Processing, IEEE Transactions, v60, N2.

[7] C. Jekeli (2001) Inertial Navigation System with Geodetic Applications. W. de Gruyter, Berlin-Boston.

[8] Simon J. Julier (2002) The Scaled Unscented Transformation. Proc. of the American Control Conference, Anchorage, AK May 8-10, 2022.

[9] Simon J. Julier, Jeffrey K. Uhlmann (1996) A General Method for Approximating Nonlinear Transformation of Probability Distribution. Technical Report, University of Oxford, Oxford.

[10] T. Kailath (1981) Lectures on Wiener and Kalman Filtering. CISM, 140. Springer Verlag, New York

[11] R.E. Kalman (1961) A new approach to linear filtering and prediction theory. In: Trans. ASME Ser D, J Basic Eng. **83** pp. 95–107

[12] Steven M. Kay (1993) Fundamentals of Statistical Signal Processing. Vol. 1 Estimation Theory. Prentice-Hall, Englewood Cliffs, NJ.

[13] J. Lamperti (1977) Stochastic processes: a survey of mathematical theory. In: Applied Math. Sciences. Springer Verlag, New York

[14] Peter S. Maybach (1979) Stochastic Models Estimation and Control. Mathematics in Science and Engineering. VIGI-1. Academic press, New York, San Francisco, London.

[15] A. Papoulis, S. U. Pillai (2002) Probability, Random Variables and Stochastic Processes. McGraw-Hill

[16] Sansò F. (2006) Navigazione Geodetica e Posizionamento Cinematico. Polipress, Politecnico di Milano.

[17] Eric A. Wan and Rudolph van der Merwe (2000) The Unscented Kalman Filter for Nonlinear Estimation. Proc. of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium, 4 October 2000. 153-158

[18] Wikipedia: Ensamble Kalman Filter.

[19] Zhe Chen (2003) Bayesian Filtering: From Kalman Filters to Particle Filters, and Beyond. Statistics: A Journal of Theoretical and Applied Statistics 182(1)