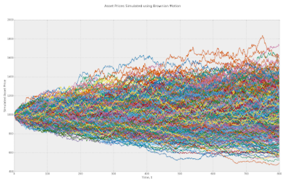


LINMA1731 Stochastic processes

Simon Desmidt

Année académique 2023-2024 - Q2

UCLouvain

Contents

Probability - Reminders

1.1 Probability and event spaces

- A sample space W is a nonempty collection of points called outcomes.
- An event space $\mathcal{F}$ of a sample space W is a nonempty collection of subsets of W, called events, that is closed under complementation and countable union, and with $W \in \mathcal{F}$.
- A measurable space $(W, \mathcal{F})$ is a pair consisting of a sample space W and an event space $\mathcal{F}$ of W.
- A probability measure P on a measurable space $(W, \mathcal{F})$ assigns a number $P(E)$ to every event E, such that P obeys the following rules, called axioms of probability :
  - $P(E) \geq 0$ for all $E \in \mathcal{F}$.
  - $P(W) = 1$.
  - $P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$, for all countable collection $\{E_i\}$ of pairwise disjoint events.
- A probability space is a triple $W, \mathcal{F}, P)$ consisting of a sample space W, an event space $\mathcal{F}$, and a probability measure P.
- Two events E_1 and E_2 are called independent if $P(E_1 \cap E_2) = P(E_1)P(E_2)$.
- The conditional probability of event F given event G is $P(F|G) = P(F \cap G)/P(G)$, assuming that $P(G) > 0$.
- Given a probability space $(W, \mathcal{F}, P)$, a real-valued random variable (rv) is a function $X : W \to \mathbb{R}$ with the property that if $A \in \mathcal{B}(\mathbb{R})$, then also $X^{-1}(A) \in \mathcal{F}$. This means that the function is bijective.
- The cumulative distribution function (cdf) of X is defined by $F_X(x) = P(\{w : X(w) \leq x\}), \forall x \in \mathbb{R}$. The cdf can be either discrete, continuous or mixed.
- The probability mass function (pmf) or distribution of a discrete random variable X defined on $(W, \mathcal{F}, P)$ is the set fucntion P_X defined by $P_X(A) = P(\{w : X(w) \in A\}), \forall A \in \mathcal{B}(\mathbb{R})$.
- For a continuous rv, the function T_X is called the probability density function (pdf) of X. If moreover F_X is differentiable at x, then $T_X(x) = \frac{dF_X}{dx}(x)$.

- A random vector is a vector-valued function $X : W \to \mathbb{R}^n$, with each of the n components being a rv.
- Two rv's X and Y are independent if $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for all x, y, or equivalently, if the pdf's exist : $T_{X,Y}(x,y) = T_X(x)T_Y(y)$ for all x, y.
- The expectation of a rv X with pdf T_X is defined by the following integral, if it exists : $\mathbb{E}(X) = \int xT_X(x)dx$.
- The expectation of an $\mathbb{R}^n$-valued random vector X is $\mathbb{E}(X) = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{pmatrix}$.

→ N.B.: the expectation may not exist (e.g. X such that $X_i = 2i$ with probability $2i$).

- If X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.
- The conditional pdf of Y given X is $T_{Y|X}(y|x) := T_{X,Y}(x,y)/T_X(x)$.
- Bayes theorem : if $P(B) \neq 0$, $P(A|B) = P(B|A)P(A)/P(B)$ for events.
- For pdf's, the Bayes theorem is $T_{X|Y}(x|y) = T_{Y|X}T_X(x)/T_Y(y)$.
- Law of total expectation : $\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X))$.
- The covariance between rv's X and Y is $Voc(X,Y) = \mathbb{E}\left((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))^T\right)$.
- The covariance matrix of a random vector $X = (X_1 \dots X_n)^T$ is The matrix C such that $C_{ij} := C(X_i, X_j)$.
- If $C_{XY} = 0$, then X and Y are said to be uncorrelated, or orthogonal.
- The correlation coefficient of rv's X and Y is $\rho_{XY} = C_{XY}/(\sigma_X\sigma_Y)$.
- The Gaussian pdf (normal distribution), with mean $mu \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n\times n}$ positive definite, is given by

$$T_X(X_1, \dots, x_n) = \frac{1}{\sqrt{(2\pi)^n(\det\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{1.1}$$

- Conditional pdf of bivariate Gaussian : if x and y are distributed according to bivariate Gaussian pdf with mean vector $\left(\mathbb{E}(x) \quad \mathbb{E}(y))^T\right)$ and covariance matrix

$$C = \begin{pmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{pmatrix} \tag{1.2}$$

so that

$$p(x,y) = \frac{1}{2\pi \det^{1/2}(C)} \exp\left(-\frac{1}{2}\begin{pmatrix} x - \mathbb{E}(x) \\ y - \mathbb{E}(Y) \end{pmatrix}^T C^{-1} \begin{pmatrix} x - \mathbb{E}(x) \\ y - \mathbb{E}(Y) \end{pmatrix}\right) \tag{1.3}$$

then the conditional pdf $p(y|x)$ is also Gaussian and

$$\mathbb{E}(y|x) = \mathbb{E}(y) + \frac{C_{XY}}{C_{YY}}(x - \mathbb{E}(x)) \qquad C_{YY|X} = C_{YY} - \frac{C_{XY}^2}{C_{XX}} \tag{1.4}$$

This result can be generalized : if x and y are jojontly Gaussian, where x is $k \times 1$ and y is $l \times 1$, with mean vector $[E(x)^T \quad E(y)^T]$ and partitioned covariance matrix

$$C = \begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix} = \begin{pmatrix} k \times k & k \times l \\ l \times k & l \times l \end{pmatrix} \tag{1.5}$$

then the conditional pdf $p(y|x)$ is also Gaussian and

$$\begin{cases} \mathbb{E}(y|x) = \mathbb{E}(y) + C_{yx}C_{xx}^{-1}(x - \mathbb{E}(x)) \\ C_{Y|X} = C_{YY} - C_{YX}C_{XX}^{-1}C_{XY} \end{cases} \tag{1.6}$$

- The uniform pdf over an interval $[a, b]$ is

$$T_X(x) = \begin{cases} \frac{1}{b-a} \text{ for } a \leq x \leq b \\ 0 \text{ otherwise} \end{cases} \tag{1.7}$$

and we write $X \sim \mathcal{U}(a, b)$.

- The exponential pdf with rate parameter $\lambda > 0$ is

$$T_X(x) = \begin{cases} \lambda e^{-\lambda x} \text{ si } x \geq 0 \\ 0 \text{ otherwise} \end{cases} \tag{1.8}$$

- The Laplace pdf with parameters $\mu \in \mathbb{R}$ and $b > 0$ is

$$T_X(x) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right) \tag{1.9}$$

1.1.1 Functions of a random variable

Let $g : \mathbb{R} \to \mathbb{R}$ be a strictly monotone function, let X be a random variable and let $Y = g(X)$. Then

$$T_Y(y) = \left.\frac{T_X(x)}{|g'(x)|}\right|_{x=g^{-1}(y)} \tag{1.10}$$

if $y \in g(\mathbb{R})$ and 0 otherwise.

Let X be an n-dimensional random vector, A a constant $n \times n$ matrix, and b a constant n-dimensional vector. Let $Y = AX + b$. Then

$$T_Y(y) = T_X(A^{-1}(y - b))\frac{1}{|\det(A)|} \tag{1.11}$$

Let $X_1, \ldots, X_n$ be a finite set of rv's with pdf's $T_{X_1}, \ldots, T_{X_n}$. Let $w_1, \ldots, , w_n$ be a set of wiehgts such that $w_i \geq 0$ and $\sum_{k=1}^{n} w_k = 1$. Then, the mixture rv Z is the rv with mixture distirbution:

$$T_Z(z) = w_1 T_{X_1}(z) + \cdots + w_n T_{X_n}(z) \tag{1.12}$$

→ N.B.: the distribution of a sum of rv's is not the sum of the distributions of the rv's.

Minimum Variance Unbiased Estimation and Cramer-Rao bound

2.1 The mathematical estimation problem

Given the N-point data set $\{x[0], \dots, x[N-1]\}$, which depends on an unkown parameter $\theta \in \mathbb{R}^p$, we wish to determine an estimator $\hat{\theta}$ of θ based on the data : $\hat{\theta} = g(x[0], \dots, x[N-1])$, where g is some function to be determined. We then have a parametric model, which is a stochastic description of the data as function of the parameter that generated it.

The first step is to model the data: because they are inherently random, we describe it by its pdf : $p(x[0], \dots, x[N-1]; \theta)$.

→ N.B.: an estimator is a random variable. Thus it can only be completely described statistically or by its pdf.

→ N.B.: a white Gaussian noise (WGN) is such that each sample $w[n] \sim \mathcal{N}(0, \sigma^2)$.

We must differentiate the Bayesian and Fisher estimation: it is Fisher when θ is deterministic, i.e. when there is no prior knowledge on θ involved in the estimation of g; and Bayesian when θ is a realization of a random vector.

2.2 Assessing Estimator Performance

To estimate the performance of an estimator, we can define its expectation and variance:

- Its expectation changes according to how it is defined based on $x[i]$, but it is considered good when close to the actual parameter θ. We call $b(\theta) = \mathbb{E}(\hat{\theta}) - \theta$ the bias.
- From its variance we can conclude about the uncertainty; the smaller it is, the less uncertain the model is.

2.3 Minimum Variance Unbiased Estimation

The estimator $\hat{\theta}$ of θ is unbiased if

$$\mathbb{E}(\hat{\theta}_i) = \theta_i \qquad \forall a_i < \theta < b_i \tag{2.1}$$

where (a_i, b_i) is the range of possible values of θ_i. The bias of the estimator is $b(\theta) = \mathbb{E}(\hat{\theta}) - \theta$.

2.3.1 Minimum variance criterion

An approach to find the optimal estimator is to constrain the bias to be zero and find the estimator which minimizes the variance, to have an optimal MSE ($MSE = \mathbb{V}(\hat{\theta}) + b^2(\theta)$.

The unbiased estimator $\hat{\theta}$ is called minimum variance unbiased (MVU) if $\mathbb{V}(\hat{\theta}_i)$, for $i \in \{1, \dots, p\}$, is minimum among all unbiased estimators.

→ N.B.: it is possible that a MVU estimator does not exist.

2.4 Cramer-Rao Lower bound

The Fisher information matrix $I(\theta) \in \mathbb{R}^{p \times p}$ is

$$I(\theta) = -\mathbb{E}\left(\frac{\partial^2 \ln\left(p(x;\theta)\right)}{\partial \theta^2}\right) \qquad I_{k,j}(\theta) = -\mathbb{E}\left(\frac{\partial^2 \ln\left(p(x;\theta)\right)}{\partial \theta_k \partial \theta_j}\right) \tag{2.2}$$

Assume that $\theta \in \mathbb{R}$ and

$$\mathbb{E}\left(\frac{\partial \ln\left(p(x;\theta)\right)}{\partial \theta}\right) = 0 \qquad \forall \theta \in \mathbb{R} \tag{2.3}$$

Then, the variance of any unbiased estimators satisfies: $\mathbb{V}(\hat{\theta} \geq 1/I(\theta)$. Moreover, an MVU estimator that attains the bound can be found iff

$$\frac{\partial \ln\left(p(x;\theta)\right)}{\partial \theta} = I(\theta)(g(x) - \theta) \tag{2.4}$$

for som eg. That estimator is $\hat{\theta} = g(x)$, its variance is $1/I(\theta)$ and it is called efficient.

2.4.1 Vectorisation

Assume that

$$\mathbb{E}\left(\frac{\partial \ln\left(p(x;\theta)\right)}{\partial \theta}\right) = 0 \qquad \forall \theta \in \mathbb{R}^p \tag{2.5}$$

Then, the variance of any unbiased estimators satisfies

$$Cov(\hat{\theta}) - I^{-1}(\theta) \succeq 0 \tag{2.6}$$

Moreover, an MVU estimator that attains the bound can be found iff

$$\frac{\partial \ln\left(p(x;\theta)\right)}{\partial \theta} = I(\theta)(g(x) - \theta) \tag{2.7}$$

for some g. That estimator is $\hat{\theta} = g(x)$, its covariance matrix is $I^{-1}(\theta)$.

2.5 Stochastic convergence notions

Let $X_k \in \mathbb{R}^n$, $k \in \mathbb{Z}_+$ be a random sequence.

- X_k is said to converge in distribution to X if

$$\lim_{k\to\infty} P(X_k \in A) = P(X \in A) \tag{2.8}$$

for every A which is a continuity set of X, i.e. every A such that $P(X \in \partial A) = 0$.

We write $X_k \xrightarrow[\mathcal{D}]{k\to\infty}$.

- X_k is said to converge in probability to X if

$$\forall \epsilon > 0 \ : \ \lim_{k\to\infty} P(\|X_k - X\| > \epsilon) = 0 \tag{2.9}$$

We write $p\lim_{k\to\infty} X_k = X$.

- X_k is said to converge in mean square to X if

$$\lim_{k\to\infty} \mathbb{E}\left(\|X_k - X\|^2\right) = 0 \tag{2.10}$$

- X_k is said to converge to X almost surely if

$$P\left(X_k(x) \xrightarrow[k\to\infty]{X} (w)\right) = 1 \iff x_k(w) \xrightarrow[k\to\infty]{x} (w) \qquad \forall w \in A \subset W \qquad P(A) = 1 \tag{2.11}$$

2.6 Asymptotic properties of estimators

The estimator $\hat{\theta} = g(x^N)$, based on the vector of observations $x^N = (x[0], \ldots, x[N-1])$ of increasing dimension, is said asymptotically unbiased if

- Fisher framework: $\lim_{N\to\infty} \mathbb{E}(g(x^N)) = \theta$.
- Bayesian framework : $\lim_{N\to\infty} \mathbb{E}(g(x^N)) = \mathbb{E}(\theta)$.

A sequence $\{\hat{\theta}_N(x)\}_{N\in\mathbb{N}}$ of estimators is called consistent if, for all θ,

$$p\lim_{N\to\infty} \hat{\theta}_N(x) = \theta \tag{2.12}$$

A sequence $\{\hat{\theta}_N(x)\_N \in \mathbb{N}$ of consistent estimators is called asymptotically normal if

$$\sqrt{N}(\hat{\theta}_N(x) - \theta) \xrightarrow[\mathcal{D}]{N\to\infty} \mathcal{N}(0, \Sigma) \tag{2.13}$$

for some positive-definite matrix Σ. It is a best asymptotically normal estimator if Σ is minimal in the class of asymptotically normal estimators.

Fisher estimators

3.1 Linear models

Consider the linear model $x[n] = A + Bn + Cn^2 + \cdots + w[n]$, where $w[n]$ is WGN. In vector form, we have $x = H\theta + w$, with $\theta = (A, B, \dots)^T$.

We have: x is $N \times 1$, H is $N \times p$ ($N > p$) and its rank is p, w is $N \times 1$ with pdf $\mathcal{N}(0, \mathbf{C})$. Then, the MVU estimators is

$$\hat{\theta} = \left(H^T C^{-1} H\right)^{-1} H^T C^{-1} x \qquad C_{\hat{\theta}} = \sigma^2 \left(H^T C^{-1} H\right)^{-1} \tag{3.1}$$

Moreover, this estimator is the MVU estimator and efficient. And, since it is a linear transformation of a Gaussian vector, $\hat{\theta} \sim \mathcal{N}(\theta, C_{\hat{\theta}})$.

3.2 Best Linear Unbiased Estimator (BLUE)

It is frequent that the MVU estimator does not exist or at least cannot be found. We can find a first solution to that problem by looking for a linear estimator instead. The BLUE restricts the estimator to be linear in the data: $\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n]$, where a_n are coefficients to be determined. The BLUE is defined to be the estimator with linear structure that is unbiased and has minimum variance.
As we want the blue to be unbiased and we want to minimise the variance, we write:

$$\mathbb{E}(\hat{\theta}) = \sum_{n=0}^{N-1} a_n \mathbb{E}(x[n]) = \theta \tag{3.2}$$

$$\mathbb{V}(\hat{\theta}) = \mathbb{E}\left(\left(\sum_{n=0}^{N-1} a_n x[n] - \mathbb{E}\left(\sum_{n=0}^{N-1} a_n x[n]\right)\right)^2\right) = \mathbb{E}\left(\left(a^T x - a^T \mathbb{E}(x)\right)^2\right) = a^T C a \tag{3.3}$$

with $a = (a_0, a_1, \dots, a_{N-1})^T$, and $C := (x - \mathbb{E}(x))(x - \mathbb{E}(x))^T$

→ N.B.: for the bias to be null, we must have a linear expectation, i.e. $\mathbb{E}(x[n]) = s[n]\theta$.

Combining both expressions, we find that the BLUE can be computed by solving

$$\min_a a^T C a \qquad a^T s = 1 \tag{3.4}$$

The solution is $a_{opt} = \frac{C^{-1}s}{s^TC^{-1}s}$.

Assume that the expectation is linear. The BLUE of θ is given by

$$\hat{\theta} = \frac{s^TC^{-1}x}{s^TC^{-1}s} \tag{3.5}$$

Where $C = (x - \mathbb{E}(x))(x - \mathbb{E}(x))^T$, and has minimum variance

$$\mathbb{V}(\hat{\theta}) = \frac{1}{s^TC^{-1}s} \tag{3.6}$$

Equivalently, we can say the following:
Assume that $x[n] = \theta s[n] + w[n]$, where w has zero mean and covariance C. The BLUE is

$$\hat{\theta} = \frac{s^TC^{-1}x}{s^TC^{-1}s} \tag{3.7}$$

with $C = (x - \mathbb{E}(x))(x - \mathbb{E}(x))^T$, and has minimum variance

$$\mathbb{V}(\hat{\theta}) = \frac{1}{s^TC^{-1}s} \tag{3.8}$$

Gauss-Markov Theorem :
Assume that the data is generated by the linear model $x = H\theta + w$, where H is $N \times p$, θ is $p \times 1$ and w is $N \times 1$ with zero mean and covariance C, but arbitrary distribution. The BLUE is

$$\hat{\theta} = (H^TC^{-1}H)^{-1}H^TC^{-1}x \qquad Cov(\hat{\theta}) = (H^TC^{-1}H)^{-1} \tag{3.9}$$

3.3 Maximum Likelihood Estimation

The Maximum Likelihood Estimator (MLE) is defined as

$$\hat{\theta} := \arg\max_{\theta} p(x;\theta) \tag{3.10}$$

If the pdf $p(x;\theta)$ satisfies some regularity conditions, then the MLE

- is consistent, i.e. $p\lim_{N\to\infty} \hat{\theta}_N = \theta$.
- is asymptotically normal and efficient:

$$\sqrt{N}(\hat{\theta}_N - \theta) \xrightarrow[\mathcal{D}]{N\to\infty} \mathcal{N}(0, I^{-1}(\theta)) \tag{3.11}$$

where $I^{-1}(\theta)$ is the Fisher information matrix evaluated at θ.

Assume $x = H\theta + w$, where the pdf of w is $\mathcal{N}(0, C)$. Then, the MLE of θ is

$$\hat{\theta} = (H^TC^{-1}H)^{-1}H^TC^{-1}x \tag{3.12}$$

$\hat{\theta}$ is an efficient estimator and the MVU. Moreover, its pdf is $\hat{\theta} \sim \mathcal{N}(\theta, (H^TC^{-1}H)^{-1})$.

3.4 Least squares

For the system

$$x[n] = s[n;\theta] + w[n] \tag{3.13}$$

w being the noise of the system, The least squares (LS) estimator is defined as

$$\hat{\theta}^{LS}(z) := \arg\min_{\theta} \sum_{n=0}^{N-1} (x[n] - s[n;\theta])^2 := J(\theta) \tag{3.14}$$

3.4.1 LS for linear model

Assume that the data is generated by the linear model $x = H\theta + w$, where H is $N \times p$, ($N > p$) of rank p, θ is $p \times 1$ and w is $N \times 1$ a deterministic signal. The LSE is found by minimizing

$$J(\theta) = (x - H\theta)^T W (x - H\theta) \tag{3.15}$$

We find

$$\hat{\theta} = (H^T W^{-1} H)^{-1} H^T W^{-1} x \qquad J_{\min} = x^T (I - WH(H^T WH)^{-1} H^T W) x \tag{3.16}$$

→ N.B.: if we assume $w[n]$ to be stochastic with zero mean and covariance W, then we recover the BLUE.

→ N.B.: if $w \sim \mathcal{N}(0, W)$, then we recover the MVU.

Bayesian Estimator

In Bayesian estimation, we treat θ as a random variable, and our objective is to estimate the particular realization of θ that generated the data. The Bayes estimator selects $\hat{\theta}$ such that

$$\hat{\theta} = \arg\min_{\theta} \mathbb{E}[C(\varepsilon)] \tag{4.1}$$

with $\varepsilon = \theta - \hat{\theta}$ and C a cost function. We take the expectation with respct to the joint distribution in x and θ $p(x, \theta)$. There are several possibilities for the cost function:

- $C(\varepsilon) = \varepsilon^2$, we call it the Bayesian MSE;
- $C(\varepsilon) = |\varepsilon|$;
- Hit-or-miss: $C(\varepsilon) = \begin{cases} 0 \text{ si } |\varepsilon| < \delta \\ 1 \text{ si } |\varepsilon| \geq \delta \end{cases}$.

We define the Bayesian risk :

$$\mathcal{R} = \mathbb{E}(C(\varepsilon)) = \int \left(\int C(\theta - \hat{\theta}) p(\theta|x) \right) p(x) dx \tag{4.2}$$

The estimator that miniminzes the Bayesian MSE $C(\varepsilon) = \varepsilon^2$ is called minimum mean square error (MMSE) estimator and is given by

$$\hat{\theta} = \mathbb{E}(\theta|x) \tag{4.3}$$

The estimator that minimizes the Bayesian absolute error $\mathbb{E}(C(\varepsilon))$, $C(\varepsilon) = |\varepsilon|$ is given by

$$\hat{\theta} = \text{median}(p(\theta|x)) \tag{4.4}$$

The estimator that minimizes the Bayesian hit-or-miss $\mathbb{E}(C\varepsilon))$, $C(\varepsilon) = \text{hit-or-miss}(\varepsilon)$, with $\delta \to 0$ is given by

$$\hat{\theta} = \text{mode}(p(\theta|x)) \tag{4.5}$$

with the mode being the point at which $p(\theta|x)$ is maximum.

4.1 Minimum MSE Estimator

The vector MMSE estimator $\mathbb{E}(\theta|x)$ minimizes $\mathbb{E}(\theta_i - \hat{\theta}_i)^2)$ for all i and is given by

$$\hat{\theta}_i = [\mathbb{E}(\theta|x)]_i \tag{4.6}$$

where the expectation is with respect to $p(\theta|x)$. Moreover, the minimum Bayesian MSE is

$$Bmse(\hat{\theta}_i) = \int [C_{\theta|x}]_i i p(x) dx \qquad i = 1, \dots, p \tag{4.7}$$

where

$$C_{\theta|x} = \mathbb{E}_{\theta|x} \left((\theta - \mathbb{E}(\theta|x))(\theta - \mathbb{E}(\theta|x))^T \right) \tag{4.8}$$

4.2 MMSE estimators for Gaussian pdf's

Assume that θ and x are jointly Gaussian. Then, the MMSE is

$$\mathbb{E}(\theta|x) = \mathbb{E}(\theta) + C_{\theta x} C_{xx}^{-1} (x - \mathbb{E}(x)) \tag{4.9}$$

$$Bmse(\hat{\theta}) = C_{\theta|x} = C_{\theta\theta} - C_{\theta x} C_{xx}^{-1} C_{x\theta} \tag{4.10}$$

4.3 Bayesian Linear Model

By application of the theorem in the previous section, let us assume that $x = H\theta + w$, where x is a $N \times 1$ data, H is known $N \times p$ matrix, θ is $p \times 1$ random vector with prior pdf $\mathcal{N}(\mu_\theta, C_\theta)$, and w is an noise $N \times 1$ vector with pdf $\mathcal{N}(0, C_w)$ and independent of θ. Then, the posterior pdf $p(\theta|x)$ is Gaussian with mean

$$\mathbb{E}(\theta|x) = \mu_\theta + C_\theta H^T (H C_\theta H^T + C_w)^{-1} + (x - H\mu_\theta) \tag{4.11}$$

and covariance

$$C_{\theta|x} = C_\theta - C_\theta H^T (H C_\theta H^T + C_w)^{-1} H C_\theta \tag{4.12}$$

4.4 Maximum A Posteriori Estimators

In the MAP estimation approach, we choose

$$\hat{\theta} = \arg\max_\theta p(\theta|x) \tag{4.13}$$

Equivalently,

$$\hat{\theta} = \arg\max_\theta (\ln p(x|\theta) + \ln p(\theta)) \tag{4.14}$$

4.5 Performance Description

In Bayesian estimators, it makes sense to assess the performance by determining the pdf of the error ε. We can show that if ε is Gaussian, then

$$\varepsilon \sim \mathcal{N}(0, Bmse(\hat{\theta})) \tag{4.15}$$

4.6 Linear Bayesian Estimators

Assume $\hat{\theta}_i = \sum_{n=0}^{N-1} a_{in} x[n] + a_{iN}$ for $i = 1, \ldots, p$. The LMMSE estimator that minimizes $Bmse(\hat{\theta}_i) = \mathbb{E}\left((\theta_i - \hat{\theta}_i)^2\right)$ is given by

$$\hat{\theta}\mathbb{E}(\theta) + C_{\theta x} C_{xx}^{-1}(x - \mathbb{E}(x)) \tag{4.16}$$

where $C_{\theta x}$ is a $p \times N$ matrix. The Bayesian MSE matrix is

$$M_{\hat{\theta}} = \mathbb{E}((\theta - \hat{\theta})(\theta - \hat{\theta})^T) = C_{\theta\theta} - C_{\theta x} C_{xx}^{-1} C_{x\theta} \tag{4.17}$$

where $C_{\theta\theta}$ is the $p \times p$ covariance matrix. The minimum Bayesian MSE is

$$Bmse(\hat{\theta}_i) = (M_{\hat{\theta}})_{ii} \tag{4.18}$$

Furthermore, if x is linear $x = H\theta + w$, we have the Bayesian Gauss-Markov Theorem: The LMMSE estimator of θ is

$$\begin{aligned} \hat{\theta} &= \mathbb{E}(\theta) + C_{\theta\theta} H^T (H C_{\theta\theta} H^T + C_w)^{-1} (x - H\mathbb{E}(\theta)) && (4.19) \\ &= \mathbb{E}(\theta) + (C_{\theta\theta}^{-1} + H^T C_w^{-1} H)^{-1} H^T C_w^{-1} (x - H\mathbb{E}(\theta)) && (4.20) \end{aligned}$$

The performance of the estimator is measured by the error $\varepsilon = \theta - \hat{\theta}$ whose mean is zero and whose covariance matrix is

$$\begin{aligned} C_\varepsilon &= E_{x,\theta}(\varepsilon \varepsilon^T) && (4.21) \\ &= C_{\theta\theta} - C_{\theta\theta} H^T (H C_{\theta\theta} H^T + C_w)^{-1} H C_{\theta\theta} && (4.22) \\ &= (C_{\theta\theta}^{-1} + H^T C_w^{-1} H)^{-1} && (4.23) \end{aligned}$$

The error covariance matrix is also the minimum MSE matrix $M_{\hat{\theta}}$ whose diagonal elements yield the minimum Bayesian MSE

$$[M_{\hat{\theta}}]_{ii} = [C_\varepsilon]_{ii} = Bmse(\hat{\theta}_i) \tag{4.24}$$

Dynamic state estimation

5.1 State Estimation Problem

Until now, we have often used static models, such as $x[n] = A + w[n]$. In this part, we will consider dynamic models, i.e. the successive time samples of $x[n]$ can be correlated.

5.1.1 The vector Gauss-Markov Model

The Gauss-Markov model for a $p \times 1$ vector signal $s[n]$ is

$$s[n] = As[n-1] + Bu[n] \tag{5.1}$$

$$x[n] = H[n]s[n] + w[n] \tag{5.2}$$

for $n = 0, 1, \ldots$, where

- $s[n]$ is a $p \times 1$ vector signal (state model) that cannot be measured.
- $x[n]$ is a $m \times 1$ vector model (observations).
- $u[n]$ is a $r \times 1$ vector of WGN: $\mathcal{N}(0, \mathcal{Q})$, the model noise.
- $w[n]$ is a $m \times 1$ vector of WGN: $\mathcal{N}(0, C[n])$, the output noise (it is independent from sample to sample).
- $A \in \mathbb{R}^{p\times p}$, $B \in \mathbb{R}^{p\times r}$ and $H[n] \in \mathbb{R}^{M\times p}$ are the system matrices.
- the initial state vector $s[-1] \sim \mathcal{N}(\mu_s, C_S)$ and independent of $u[n]$.

The objective of optimal filtering is to construct a sequential (i.e. at every n) MMSE estimator of $s[n]$ based on the data $\{x[0], \ldots, x[n]\}$.

The objective of optimal prediction is to construct a sequential MMSE estimator of $s[k], k > n$ based on the data $\{x[0], \ldots, x[n]\}$.

The objective of optimal smoothing is to construct a sequential MMSE estimator of $s[k], k < n$ based on the data $\{x[0], \ldots, x[n]\}$.

5.2 Kalman Filter (scalar)

We begin with the scalar version of the Gauss-Markov model :

$$s[n] = as[n-1] + u[n] \tag{5.3}$$
$$x[n] = s[n] + w[n] \tag{5.4}$$

→ N.B: the esimator of $s[n]$ based on the observations $\{x[0], \ldots, x[m]\}$ will be denoted by $\hat{s}[n|m]$.

The optimality criterion will be the Bayesian MSE:

$$\mathbb{E}\left((s[n] - \hat{s}[n|n])^2\right) \tag{5.5}$$

We defined the innovation as

$$\tilde{x}[n] = x[n] - \hat{x}[n|n-1] \tag{5.6}$$

It is the part of $x[]$ that is uncorrelated with the previous samples $\{x[0], \ldots, x[n-1]\}$. Let $X[n] = (x[0], \ldots, x[n])^T$.
We can rewrite the MMSE estimator as

$$\hat{s}[n|n] = \mathbb{E}(s[n]|X[n-1], \hat{x}[n]) = \hat{s}[n|n] = \underbrace{\mathbb{E}(s[n]|X[n-1])}_{\hat{s}[n|n-1]} + \mathbb{E}(s[n]|\tilde{x}[n]) \tag{5.7}$$

where

$$\hat{s}[n|n-1] = a\hat{s}[n-1|n-1] \tag{5.8}$$

What is a??We define $K[n]$ such that

$$\mathbb{E}(s[n]|\tilde{x}[n]) = K[n]\tilde{x}[n] \tag{5.9}$$

and we have

$$K[n] = \frac{\mathbb{E}(s[n](x[n] - \hat{s}[n|n-1]))}{\mathbb{E}((x[n] - \hat{s}[n|n-1]]^2)} \tag{5.10}$$

We have two properties useful in the derivation of the gain $K[n]$:

$$\mathbb{E}(s[n](x[n] - \hat{s}[n|n-1])) = \mathbb{E}((s[n|n-1])(x[n] - \hat{s}[n|n-1]))$$
$$\mathbb{E}(w[n](s[n] - \hat{s}[n|n-1])) = 0 \tag{5.11}$$

Using those, we have

$$K[n] = \frac{\mathbb{E}((s[n] - \hat{s}[n|n-1])^2)}{\sigma_n^2 + \underbrace{\mathbb{E}((s[n] - \hat{s}[n|n-1])^2)}_{=:M[n|n-1]}} = \frac{M[n|n-1]}{\sigma_n^2 + M[n|n-1]} \tag{5.12}$$

where σ_n^2 is the variance of σ_n^2 and $M[n|n-1]$ the minimum prediction MSE matrix and $M[n|n]$ the minimum MSE matrix.

We now need to derive an expression for $M[n|n-1]$. We have by its definition and by some property

$$M[n|n-1] = \mathbb{E}\left((a(s[n-1]-\hat{s}[n-1|n-1]) + u[n])^2\right) \tag{5.13}$$

$$M[n|n-1] = a^2 M[n-1|n-1] + \sigma_u^2 \tag{5.14}$$

From this, we can find an expression for $M[n|n]$:

$$M[n|n] = \mathbb{E}\left((s[n]-\hat{s}[n|n])^2\right) = (1-K[n])M[n|n-1] \tag{5.15}$$

The final equations needed for the Kaman filter are the following:

$$\begin{aligned}
\hat{s}[n|n-1] &= a\hat{s}[n-1|n-1] && (5.16)\\
M([n|n-1] &= a^2 M[n-1|n-1] + \sigma_u^2 && (5.17)\\
K[n] &= \frac{M[n|n-1]}{\sigma_n^2 + M[n|n-1]} && (5.18)\\
\hat{s}[n|n] &= \hat{s}[n|n-1] + K[n](x[n] - \hat{s}[n|n-1]) && (5.19)\\
M[n|n] &= (1-K[n])M[n|n-1] && (5.20)
\end{aligned}$$

With the initial conditions:

$$\begin{cases} s[-1] \sim \mathcal{N}(\mu_s, \sigma_s^2) \\ \hat{s}[-1|-1] = \mathbb{E}(s[-1]) = \mu_s \\ M[-1|-1] = \sigma_s^2 \end{cases} \tag{5.21}$$

5.3 Kalman Filter (vector)

The model now is

$$\begin{cases} \mathbf{s}[n] = \mathbf{A}\mathbf{s}[n-1] + \mathbf{B}\mathbf{u}[n] \\ \mathbf{x}[n] = \mathbf{H}[n]\mathbf{s}[n] + \mathbf{w}[n] \end{cases} \tag{5.22}$$

with the same definitions as in Equation 5.1.

The MMSE estimator of $s[n]$ based on $\{x[0], \ldots, x[n]\}$ is

$$\hat{s}[n|n] = \mathbb{E}(s[n]|x[0], \ldots, x[n]) \tag{5.23}$$

and can be computed sequentially in time using the following recursion:

$$\begin{aligned}
&\hat{s}[n|n-1] = A\hat{s}[n-1|n-1] && (5.24)\\
&M[n|n-1] = AM[n-1|n-1]A^T + BQB^T && (5.25)\\
&K[n] = M[n|n-1]H^T[n]\left(C[n] + H[n]M[n|n-1]H^T[n]\right)^{-1} && (5.26)\\
&\hat{s}[n|n] = \hat{s}[n|n-1] + K[n](x[n] - H[n]\hat{s}[n|n-1]) && (5.27)\\
&M[n|n] = (I - K[n]H[n])M[n|n-1] && (5.28)\\
& && (5.29)
\end{aligned}$$

The recursion is initialized by $\hat{s}[-1|-1] = \mu_s$ and $M[-1|-1] = C_s$.

5.4 Bayes filters

Bayes filters aim to compute the entire distribution $p(s[n]|X[n])$ instead of just $\mathbb{E}(s[n]|X[n])$. We will asume the following:

- $s[n]$ given $s[n-1]$ is independent of anything that has happened before the time step $n-1$. Namely,
$$p(s_n|s_{0:n-1}, x_{0:n-1}) = p(s_n|s_{n-1}) \tag{5.30}$$
- The past is independent of the future given the present:
$$p(s_{n-1}|s_{n:T}, x_{n:T}) = p(s_{n-1}|s_n) \tag{5.31}$$
- The current measurement x_n given the current state s_n is conditionnally independent of the measurement and state histories:
$$p(x_n|s_{1:n}, x_{1:n-1}) = p(x_n|s_n) \tag{5.32}$$

From the previous chapter, we can determine all the equations appearing in the theorem:

For all $n \geq 0$, the folllowing algorithm (the Bayesian filter) returns $g_{n|n}(s_n) = p(s_n|x_{0:n})$.

- Initialization: $g_{-1|-1}(s_{-1}) = p(x_{-1})$.
- Prediction:
$$g_{n|n-1}(s_n) = \int p(s_n|s_{n-1})g_{n-1|n-1}(s_{n-1})ds_{n-1} \tag{5.33}$$
- Update:
$$g_{n|n}(s_n) = \frac{1}{Z_n}p(x_n|s_n)g_{n|n-1}(s_n) \tag{5.34}$$

where $Z_n = \int p(x_n|s_n)p(s_n|x_{0:n-1})ds_n$.

5.4.1 For a non linear model

For a non linear model such as the following:

$$\begin{cases} s[n] = F(s[n-1]) + u[n-1] \\ x[n] = G(s[n]) + w[n] \end{cases} \tag{5.35}$$

The Bayesian filter is given by

- Initialization: $g_{-1|-1}(s_{-1}) = p(x_{-1})$.
- Prediction:
$$g_{n|n-1}(s_n) = \int p_{u[n-1]}(s[n] - F(s[n-1]))g_{n-1|n-1}(s_{n-1})ds_{n-1} \tag{5.36}$$
- Update:
$$g_{n|n}(s_n) = \frac{1}{Z_n}p_{w[n]}(x[n] - G(s[n]))g_{n|n-1}(s_n) \tag{5.37}$$

where $Z_n = \int p(x_n|s_n)p(s_n|x_{0:n-1})ds_n$.

5.5 Particle filter

5.5.1 Monte Carlo approximations

Monte Carlo methods is a general class of methods where densities, expected values, ... are replaced by drawing samples from the distirbution and estimating the quantities by sample quantities.
Assume we want to estimate the pdf $p(s)$ (here representing $p(s_n|x_{0:n})$). We then draw N idnependent random samples $s^{(i)} \sim p(s), i = 1, \ldots, N$ called particles and estimate the pdf as

$$\hat{p}(s) := \frac{1}{N} \sum_{i=1}^{N} \delta\left(s - s^{(i)}\right) \tag{5.38}$$

Assume we want to estimate the general expectation $\mathbb{E}(\varphi(s))$. We draw N independent random samples $s^{(i)} \sim p(s), i = 1, \ldots, N$ and

$$\hat{\mathbb{E}}(\varphi(s)) := \frac{1}{N} \sum_{i=1}^{N} \varphi(s^{(i)}) \tag{5.39}$$

Properties:

- Unbiased: $\mathbb{E}(\hat{\mathbb{E}}(\varphi(s))) = \mathbb{E}(\varphi(s))$;
- Consistent: $\hat{\mathbb{E}}(\varphi(s)) \xrightarrow{N\to\infty} \mathbb{E}(\varphi(s))$;
- Asymptotically normal: if $\mathbb{V}(\varphi(s))\sigma_\varphi < \infty$, then by the central limit theorem (CLT)

$$\frac{\sqrt{N}\left(\hat{\mathbb{E}}(\varphi(s)) - \mathbb{E}(\varphi(s))\right)}{\sigma_\varphi} \xrightarrow{N\to\infty} \mathcal{N}(0,1) \tag{5.40}$$

Hence, the error in the Monte Carlo estimate decreases as $\mathcal{O}\left(N^{-1/2}\right)$.

5.5.2 Importance sampling

In practice, it is impossible to obtain samples from $p(s_n|x_{0:n})$. In importance sampling, we use an approximate distribution called the importance distribution $\pi(s|x_{0:n})$, from which we can easily draw samples. Assume we want to estimate $\mathbb{E}(\varphi(s))$. IS is based on the following decomposition:

$$\mathbb{E}(\varphi(s)) = \int \varphi(s)p(s)ds = \int \left(\varphi(s)\frac{p(s)}{\pi(s)}\right)\pi(s)ds \tag{5.41}$$

assuming that $pi(s)$ is nonzero whenever $p(s)$ is nonzero. If $s^{(i)} \sim \pi(s)$, we can compute the following:

$$\hat{\mathbb{E}}(\varphi(s)) = \sum_{i=1}^{N} \underbrace{\frac{1}{N}\frac{p\left(s^{(i)}\right)}{\pi\left(s^{(i)}\right)}}_{=:\tilde{w}^{(i)}} \varphi\left(s^{(i)}\right) \tag{5.42}$$

Similarly,

$$\hat{p}(s) = \sum_{i=1}^{N} \tilde{w}^{(i)} \delta\left(s - s^{(i)}\right) \tag{5.43}$$

We call $s^{(i)}$ the particle and $\tilde{w}^{(i)}$ the weight of the particle.

→ N.B.: We can choose $\pi(s)$, and we need to choose it as close to $p(s)$ as possible.

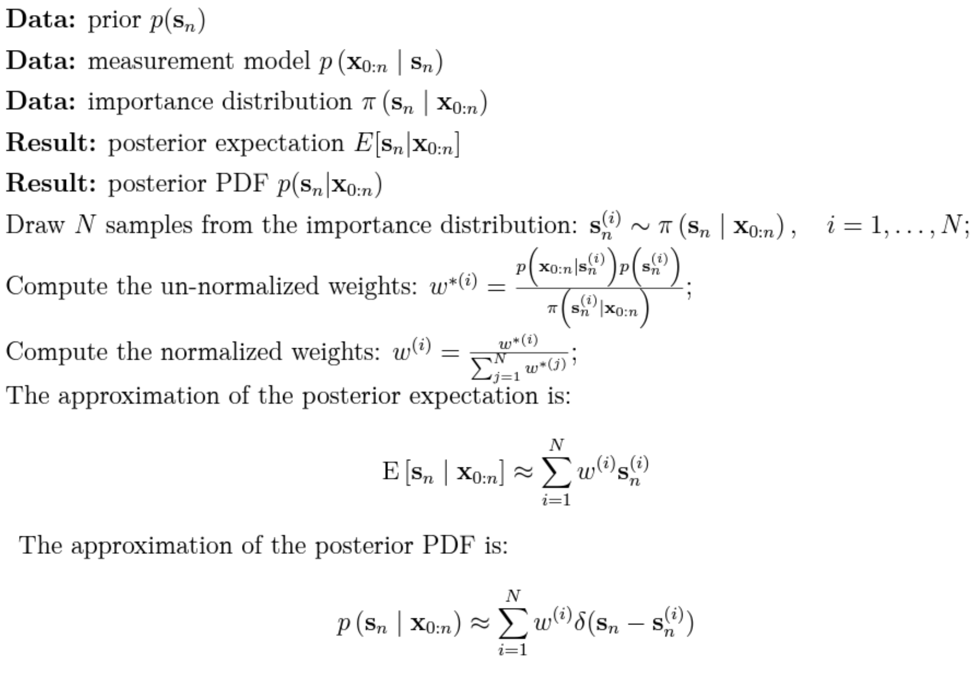
The disadvantage of this technique is that we need to know $p(s^{(i)})$ to compute the weights. However, by using Bayes rule, we can solve that problem:

$$\mathbb{E}(\varphi(s)|x_{0:n}) = \sum_{i=1}^{N} \underbrace{\left(\frac{\frac{p\left(x_{0:n}|s^{(i)}\right)p\left(s^{(i)}\right)}{\pi\left(s^{(i)}|x_{0:n}\right)}}{\sum_{j=1}^{N} \frac{p\left(x_{0:n}|s^{(j)}\right)p\left(s^{(j)}\right)}{\pi\left(s^{(j)}|x_{0:n}\right)}} \right)}_{=:w^{(i)}} \varphi\left(s^{(i)}\right) \tag{5.44}$$

To simplify notations, we will use

$$\begin{cases} w^{*(i)} = \frac{p(x_{0:n}|s^{(i)})p\left(s^{(i)}\right)}{\pi(s^{(i)}|x_{0:n})} \\ w^{(i)} = \frac{w^{*(i)}}{\sum_{j=1}^{N} w^{*(j)}} \end{cases} \tag{5.45}$$

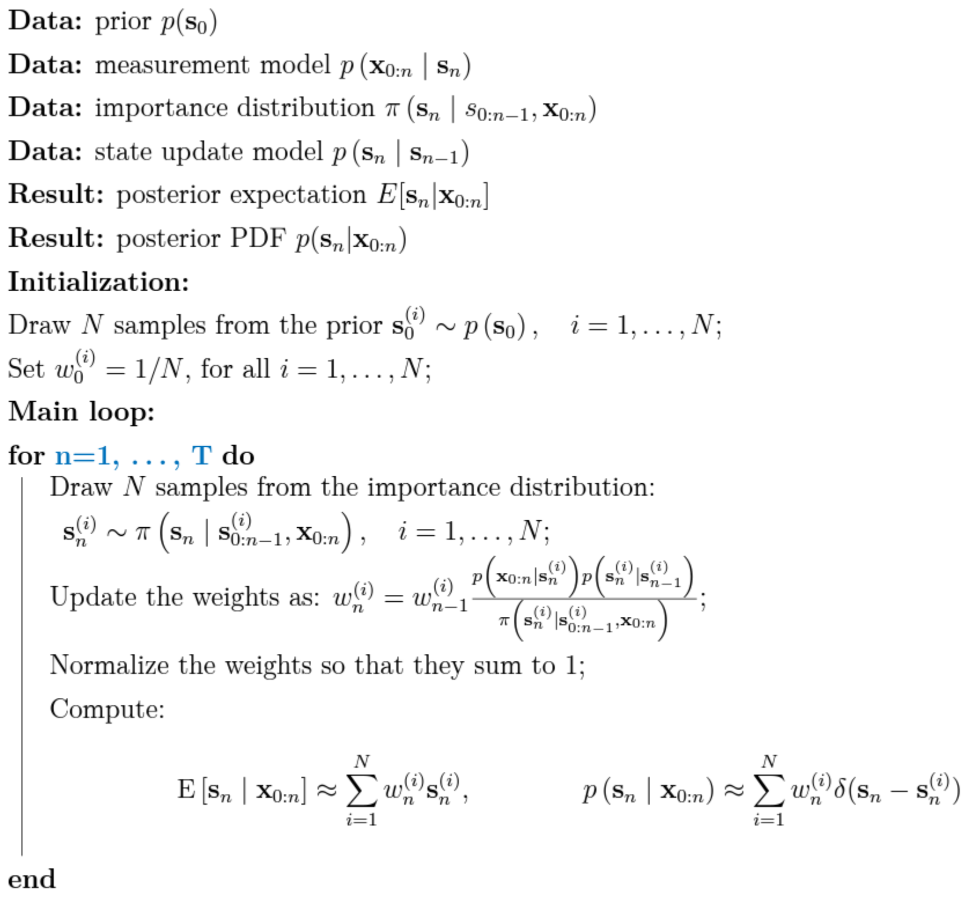
Here is an algorithm for the importance sampling:



```
Data: prior $p(\mathbf{s}_n)$
Data: measurement model $p(\mathbf{x}_{0:n} \mid \mathbf{s}_n)$
Data: importance distribution $\pi(\mathbf{s}_n \mid \mathbf{x}_{0:n})$
Result: posterior expectation $E[\mathbf{s}_n|\mathbf{x}_{0:n}]$
Result: posterior PDF $p(\mathbf{s}_n|\mathbf{x}_{0:n})$
Draw $N$ samples from the importance distribution: $\mathbf{s}_n^{(i)} \sim \pi(\mathbf{s}_n \mid \mathbf{x}_{0:n}), \quad i = 1, \ldots, N$;
Compute the un-normalized weights: $w^{*(i)} = \frac{p\left(\mathbf{x}_{0:n}|\mathbf{s}_n^{(i)}\right)p\left(\mathbf{s}_n^{(i)}\right)}{\pi\left(\mathbf{s}_n^{(i)}|\mathbf{x}_{0:n}\right)}$;
Compute the normalized weights: $w^{(i)} = \frac{w^{*(i)}}{\sum_{j=1}^{N} w^{*(j)}}$;
The approximation of the posterior expectation is:
    $\mathrm{E}[\mathbf{s}_n \mid \mathbf{x}_{0:n}] \approx \sum_{i=1}^{N} w^{(i)} \mathbf{s}_n^{(i)}$
The approximation of the posterior PDF is:
    $p(\mathbf{s}_n \mid \mathbf{x}_{0:n}) \approx \sum_{i=1}^{N} w^{(i)} \delta(\mathbf{s}_n - \mathbf{s}_n^{(i)})$
```

5.5.3 Sequential importance sampling

Is requires N samples $s_n^{(i)}$. Since we are dealing with dynamical systems in this part, we can only collect one sample before time advances, leading to the sample at the next time step. Sequential importance sampling allows to account for this temporal variability. The method is the same as for the IS, but the weights vary with time:

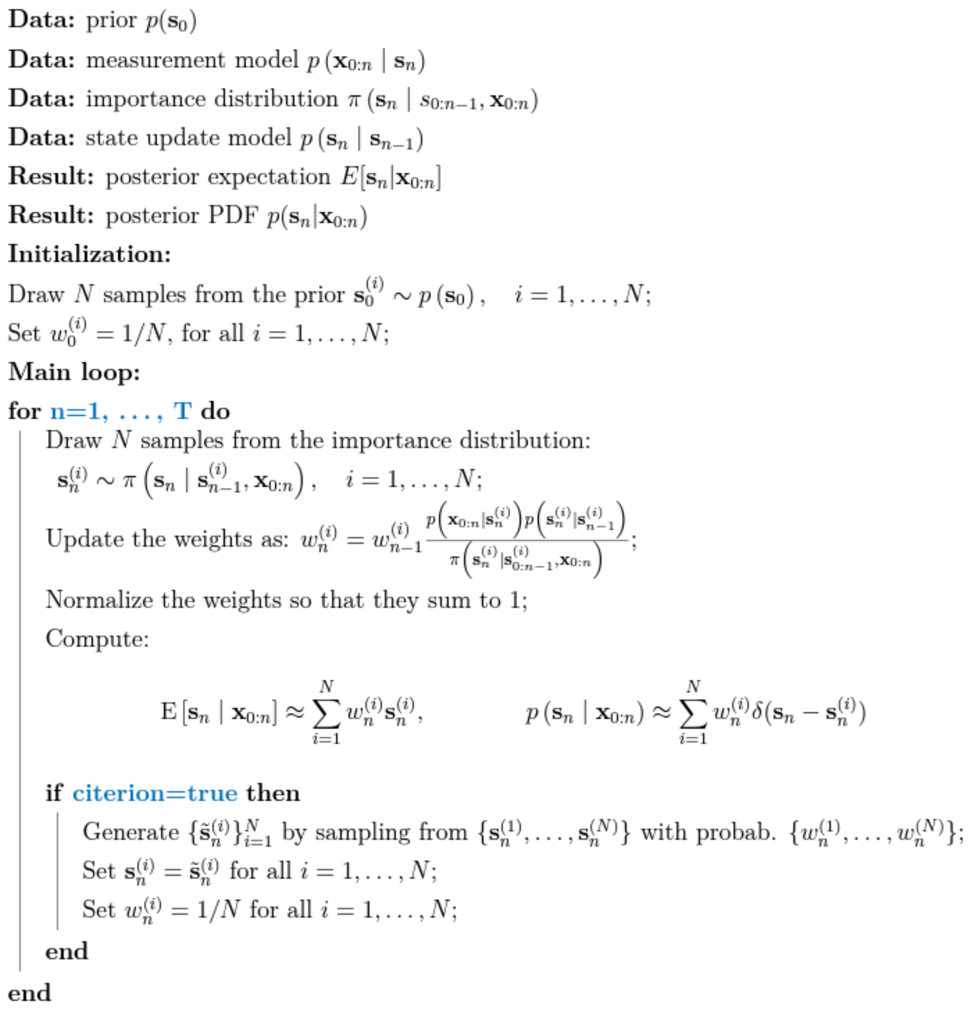


```
Data: prior $p(\mathbf{s}_0)$
Data: measurement model $p\left(\mathbf{x}_{0:n} \mid \mathbf{s}_n\right)$
Data: importance distribution $\pi\left(\mathbf{s}_n \mid s_{0:n-1}, \mathbf{x}_{0:n}\right)$
Data: state update model $p\left(\mathbf{s}_n \mid \mathbf{s}_{n-1}\right)$
Result: posterior expectation $E[\mathbf{s}_n|\mathbf{x}_{0:n}]$
Result: posterior PDF $p(\mathbf{s}_n|\mathbf{x}_{0:n})$
Initialization:
Draw $N$ samples from the prior $\mathbf{s}_0^{(i)} \sim p\left(\mathbf{s}_0\right), \quad i = 1, \ldots, N$;
Set $w_0^{(i)} = 1/N$, for all $i = 1, \ldots, N$;
Main loop:
for n=1, ..., T do
    Draw $N$ samples from the importance distribution:
    $\mathbf{s}_n^{(i)} \sim \pi\left(\mathbf{s}_n \mid \mathbf{s}_{0:n-1}^{(i)}, \mathbf{x}_{0:n}\right), \quad i = 1, \ldots, N$;
    Update the weights as: $w_n^{(i)} = w_{n-1}^{(i)} \frac{p\left(\mathbf{x}_{0:n}|\mathbf{s}_n^{(i)}\right) p\left(\mathbf{s}_n^{(i)}|\mathbf{s}_{n-1}^{(i)}\right)}{\pi\left(\mathbf{s}_n^{(i)}|\mathbf{s}_{0:n-1}^{(i)}, \mathbf{x}_{0:n}\right)}$;
    Normalize the weights so that they sum to 1;
    Compute:
        $\mathrm{E}\left[\mathbf{s}_n \mid \mathbf{x}_{0:n}\right] \approx \sum_{i=1}^{N} w_n^{(i)} \mathbf{s}_n^{(i)}, \qquad p\left(\mathbf{s}_n \mid \mathbf{x}_{0:n}\right) \approx \sum_{i=1}^{N} w_n^{(i)} \delta(\mathbf{s}_n - \mathbf{s}_n^{(i)})$
end
```

Notice that we now have a loop for what was previously done one time. It is here done for every time step.

5.5.4 Sequential importance resampling

The degeneracy probvlem is when we converge to a situation where almost all the particles have zero or nearly zero weights. To solve it, we can draw N new samples from the discrete distribution defined by the weights and replace the old set of N samples with this new set.



```
Data: prior $p(\mathbf{s}_0)$
Data: measurement model $p(\mathbf{x}_{0:n} \mid \mathbf{s}_n)$
Data: importance distribution $\pi(\mathbf{s}_n \mid s_{0:n-1}, \mathbf{x}_{0:n})$
Data: state update model $p(\mathbf{s}_n \mid \mathbf{s}_{n-1})$
Result: posterior expectation $E[\mathbf{s}_n|\mathbf{x}_{0:n}]$
Result: posterior PDF $p(\mathbf{s}_n|\mathbf{x}_{0:n})$
Initialization:
Draw $N$ samples from the prior $\mathbf{s}_0^{(i)} \sim p(\mathbf{s}_0)\,, \quad i = 1, \ldots, N$;
Set $w_0^{(i)} = 1/N$, for all $i = 1, \ldots, N$;
Main loop:
for n=1, ..., T do
    Draw $N$ samples from the importance distribution:
     $\mathbf{s}_n^{(i)} \sim \pi\left(\mathbf{s}_n \mid \mathbf{s}_{n-1}^{(i)}, \mathbf{x}_{0:n}\right), \quad i = 1, \ldots, N$;
    Update the weights as: $w_n^{(i)} = w_{n-1}^{(i)} \frac{p\left(\mathbf{x}_{0:n}|\mathbf{s}_n^{(i)}\right) p\left(\mathbf{s}_n^{(i)}|\mathbf{s}_{n-1}^{(i)}\right)}{\pi\left(\mathbf{s}_n^{(i)}|\mathbf{s}_{0:n-1}^{(i)}, \mathbf{x}_{0:n}\right)}$;
    Normalize the weights so that they sum to 1;
    Compute:
        $\mathrm{E}[\mathbf{s}_n \mid \mathbf{x}_{0:n}] \approx \sum_{i=1}^{N} w_n^{(i)} \mathbf{s}_n^{(i)}, \qquad p(\mathbf{s}_n \mid \mathbf{x}_{0:n}) \approx \sum_{i=1}^{N} w_n^{(i)} \delta(\mathbf{s}_n - \mathbf{s}_n^{(i)})$
    if citerion=true then
        Generate $\{\tilde{\mathbf{s}}_n^{(i)}\}_{i=1}^{N}$ by sampling from $\{\mathbf{s}_n^{(1)}, \ldots, \mathbf{s}_n^{(N)}\}$ with probab. $\{w_n^{(1)}, \ldots, w_n^{(N)}\}$;
        Set $\mathbf{s}_n^{(i)} = \tilde{\mathbf{s}}_n^{(i)}$ for all $i = 1, \ldots, N$;
        Set $w_n^{(i)} = 1/N$ for all $i = 1, \ldots, N$;
    end
end
```

The variable "criterion" describes how frequently resampling is performed. If we define

$$P_n^{eff} = \frac{1}{\sum_{i=1}^{N} \left(w_n^{(i)}\right)^2} \tag{5.46}$$

we can perform resampling when this value reaches the criterion, e.g. $P_n^{eff} < N/10$.

Notion of random function

6.1 Random vector

6.1.1 Real random variable

$\rightarrow$ N.B. : We denote X the rv and x one realization.

A rv is a correspondance between the result of an experiment and a real/complex scalar value. It is completely described by its probability density function (pdf), denoted $T_X(x)$. Its associated moments are

- Mean : $m_X = \mathbb{E}(X) = \int_x xT_X(x)dx$
- Variance : $\sigma_X^2 = Var(X) = \int_x (x - m_X)^2 T_X(x)dx$

6.1.2 Vectorization

We denote the random vector V of size 2 by $V = (X \quad Y)^T$. Each entry is a random variable. The random vector is fully characterized by the joint probability density function denoted by $T_{XY}(x, y)$. We define the marginal probability density functions:

$$T_X(x) = \int_y T_{XY}(x, y)dy \tag{6.1}$$

$$T_Y(y) = \int_x T_{XY}(x, y)dx \tag{6.2}$$

$$\tag{6.3}$$

The moments of the random vector are

$$m_V = (m_X \quad m_Y)^T$$

$$\begin{cases} m_X = \int_x \int_y xT_{XY}(x, y)dxdy = \int_x xT_X(x)dx \\ m_Y = \int_x \int_y yT_{XY}(x, y)dxdy = \int_y yT_Y(y)dy \end{cases} \quad C_V = \begin{pmatrix} \sigma_X^2 & C_{XY} \\ C_{YX} & \sigma_Y^2 \end{pmatrix} C_{XY} = \mathbb{E}((X - m_X)(Y - m_Y)) =$$

We also define the correlation:

$$R_{XY} = \mathbb{E}(XY) = \int_{x,y} T_{XY}dxdy \tag{6.5}$$

Properties of the correlation:

- Decorrelation if $C_{XY} = 0$
- Independence if $T_{XY}(x, y) = T_X(x)T_Y(y)$
- Independence → Decorrelation
- Condition probability density function:

$$T_{X|Y}(x|y) = \frac{T_{XY}(x,y)}{T_Y(y)} \Longrightarrow \begin{cases} T_{XY}(x,y) = T_{X|Y}(x|y)T_Y(y) \\ T_X(x) = \int_y T_{X|Y}(x|y)T_Y(y)dy \end{cases} \tag{6.6}$$

6.1.3 Complex random variable

We note a compex random variable $Z = X + iY$. It is modeled as a particular case of a random vector of size 2: the statistical properties of Z are those of X and Y considered jointly.
Properties:

- By linearity of the expectation, $m_Z = m_X + im_Y$
- $C_{Z_1Z_2} = \mathbb{E}((Z_1 - m_{Z_1})(Z_2 - m_{Z_2})^*)$
- $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$

6.1.4 Generalization to size N

The concept of random vector can be generalized to size N: it is fully characterized by the joint pdf

$$T_V(v) = T_{X_1,\dots,X_N}(x_1, \dots, x_N) \tag{6.7}$$

and only partially characterized by the marginal pdfs of each entry:

$$T_{X_1}(x_1), \dots, T_{X_N}(x_N) \tag{6.8}$$

The mean is given by $m_V = (m_{X_1} \quad \dots \quad m_{X_N})^T$, and the covariance matrix is

$$C_V = \begin{pmatrix} \sigma_{X_1}^2 & \dots C_{X_1,X_N} & \\ \vdots & \ddots & \vdots \\ C_{X_N,X_1} & \dots & \sigma_{X_N}^2 \end{pmatrix} \tag{6.9}$$

6.2 Random function

A random function is the correspondance between the result of an experiment (=draw) and a function $X(t)$, whose output at each instant t is not a scalar and deterministic number, but rather a random variable, whose pdf could be different for each instant t. Let us consider n different instants, denoted $t_1, \dots, t_n$. The marginal pdfs associated with these different instants are denoted by

$$T_X(x(t_1)), \dots, T_X(x(t_n)) \tag{6.10}$$

To fully characterize n instants simultaneously, one uses once again the joint probability density functions:

$$T_X(x(t_1), \ldots, x(t_n)) \tag{6.11}$$

For a random function which is continuous time, we have $t \in (-\infty, \infty)$ and thus an infinite number of time instants need to be considered. Therefore, to fully characterise the random function, one should know the joint pdfs of all orders:

$$\lim_{n \to \infty} T_X(x(t_1, \ldots, x(t_n)) \tag{6.12}$$

This means that a continuous time random function could be seen as a random vector of infinite size.

The random function will be said to have independent values for instants $t_1, \ldots, t_n$ if

$$T_X(x(t_1, \ldots, x(t_n)) = T_X(x(t_1)) \times \cdots \times T_X(x(t_n)) \tag{6.13}$$

Let us assume that the instants $t_1, \ldots, t_n$ are in chronological order, that is $t_1 < t_2 < \cdots < t_n$. Instant n is independent of past values if

$$T_X(x(t_n)|x(t_1), \ldots, x(t_{n-1})) = T_X(x(t_n)) \tag{6.14}$$

Moments of a random function

The mean is a deterministic function which, for t, provides the mean of the rv $X(t)$ associated with that instant:

$$m_X(t) = \mathbb{E}(X(t)) = \int_{-\infty}^{\infty} x T_X(x(t)) dx \tag{7.1}$$

The variance has the same interpretation as for random variables, for $X(t) = X_r(t) + jX_i(t)$.

$$\sigma_X^2(t) = \sigma_{X_r}^2(t) + \sigma_{X_i}^2(t) \tag{7.2}$$

Assuming two instants t_1 and t_2, the covariance is

$$\begin{aligned} C_X(t_1, t_2) &= \mathbb{E}((X(t_1) - m_X(t_1))(X(t_2) - m_X(t_2))^*) & (7.3) \\ &= \int_{x_1} \int_{x_2} (x_1 - m_X(t_1))(x_2 - m_X(t_2))^* T_X(x(t_1), x(t_2)) dx_1 dx_2 & (7.4) \end{aligned}$$

<u>Covariance properties :</u>

- $\sigma_X(t) = C_X(t, t)$
- For a real random function, $C_X(t_1, t_2) = C_X(t_2, t_1)$
- For a complex random function, $C_X(t_1, t_2) = C_X^*(t_2, t_1)$

A random function has uncorrelated values if, for all pairs t_1, t_2,

$$C_X(t_1, t_2) = \sigma_X^2 \delta(t_1 - t_2) \tag{7.5}$$

7.0.1 Cross Variance properties

Assume two random functions $X(t), Y(t)$. The cross-variance between these two functions is given by

$$\begin{aligned} C_{XY}(t_1, t_2) &= \mathbb{E}((X(t_1) - m_X(t_1))(Y(t_2) - m_Y(t_2))^*) & (7.6) \\ &= \int_x \int_y (x - m_X(t_1)(y - m_Y(t_2))^* T_{XY}(x(t_1), y(t_2)) dx dy & (7.7) \end{aligned}$$

We also have

$$C_{XY}(t_1, t_2) = C_{YX}^*(t_2, t_1) \tag{7.8}$$

Properties of random functions

8.1 Stationary signals

A signal is said to be strong-sense stationary (SSS), if the joint pdfs of all orders do not depend on the time origin $t = 0$:

$$T_X(x(t_1), x(t_2), \dots, x(t_n)) = T_X(x(0), x(t_2 - t_1), \dots, x(t_n - t_1)) \tag{8.1}$$

This means that there is a dependence of the joint pdfs to only the time differences.

→ N.B.: a second order stationary signal is a SSS signal, but limited to order $n = 2$.

A signal is said to be weak-sense stationary (WSS) if

- The mean is time-independent: $m_X(t) = m_X$;
- The covariance function only depends on the difference between the two time instants: $C_X(t_1, t_2) = C_X(\tau = t_1 - t_2, 0) = C_X(\tau)$

8.2 Ergodic signals

Sometimes, we only have one single recorded realisation $x_1(t)$ of the process. To estimate the mean in that case, we assume the signal to be WSS.

Ergodicity is the framework enabling to investigate whether a signal is ergodic, i.e. under which conditions ensemble averaging[1] can be replaced by mean averaging.

8.2.1 Process ergodic in the mean

Let $\eta_T(t_0)$ be the time-average of the random function $X(t)$ over the time interval T:

$$\eta_T(t_0) = \frac{1}{T} \int_{t_0}^{t_0+T} X(t)dt \tag{8.2}$$

It is therefore a random variable depending in the general case on T and t_0. The random process is said to be ergodic in the mean when, for $T \to \infty$, $\eta_T(t_0)$ becomes independent of t_0 and T, deterministic, and equal to m_X:

$$\lim_{T\to\infty} \eta_T(t_0) = m_X \qquad \forall t_0 \tag{8.3}$$

[1] ensemble averaging means averging on the random nature of the function

A necessary and sufficient condition to benefit from egodicity in the mean is

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \left(1 - \frac{\tau}{T}\right) C_X(\tau) d\tau = 0 \tag{8.4}$$

A condition which is sufficient to benefit from ergodicity in the mean is

$$\lim_{\tau\to\infty} C_X(\tau) = 0 \tag{8.5}$$

For discrete-time processes, the time average is given by

$$\eta_N(n_0) = \frac{1}{N} \sum_{n=n_0}^{n_0+N} X(n) \tag{8.6}$$

and we have ergodicity in the mean if

$$\lim_{N\to\infty} \eta_N(n_0) = m_X \qquad \forall n_0 \tag{8.7}$$

8.2.2 Process ergodic in the auto-correlation function

The time auto-correlation function $\phi_X(\tau, t_0)$ of a WSS signal $X(t)$ is given by

$$\phi_T(\tau, t_0) = \frac{1}{T} \int_{t_0}^{t_0+T} \left(X(t+\tau) - \eta_T(t_0+\tau)\right) \left(X(t) - \eta_T(t_0)\right)^* dt \tag{8.8}$$

$\phi_T(\tau, t_0)$ is a random variable, function of t_0 and τ. There is ergodicity with respect to the auto-correlation function if

- ergodicity in the mean is fulfilled;
- for $T \to \infty$, the time auto-correlation function becomes independent of t_0, deterministic and equal to the covariance matrix:

$$\lim_{T\to\infty} \phi_T(\tau, t_0) = C_X(\tau) \tag{8.9}$$

Properties

- The ergodicity in the auto-correlation functiuon of a random signal $X(t)$ actually corresponds to the ergodicity in the mean of

$$\left(X(t+\tau) - \eta_T(t_0+\tau)\right) \left(X(t) - \eta_T(t_0)\right)^* \tag{8.10}$$

- For gaussian distributed signals, the higher order moments only depend on order 1 and 2 moments.
- A necessary and sufficient condition ot benefit from ergodicity in the auto-correlation function for all gaussian distributed and WSS signals:

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \left(1 - \frac{\theta}{T}\right) \left(C_X^2(\theta) + C_X(\theta+\tau) C_X(\theta-\tau)\right) d\theta = 0 \qquad \forall\tau \tag{8.11}$$

- A sufficient condition for similar signals is $\lim_{\tau\to\infty} C_X(\tau) = 0$.
- For sequences (discrete signals), the correlation writes

$$\phi_N(k, n_0) = \frac{1}{N} \sum_{n=n_0}^{n_0+N} \left(X(n+k) - \eta_N(n_0+k)\right) \left(X(n) - \eta_N(n_0)\right)^* \tag{8.12}$$

Corollary

$$m_X = \lim_{T\to\infty} \eta_T(t_0) = \lim_{T\to\infty} \frac{1}{T}\int_{t_0}^{t_0+T} X(t)dt \tag{8.13}$$

$$\sigma_X^2 = \lim_{T\to\infty} \phi_T(0,t_0) = \lim_{T\to\infty} \frac{1}{T}\int_{t_0}^{t_0+T} |X(t)-\eta_T(t_0)|^2 dt \tag{8.14}$$

In electricity, with ergodicity in the mean, the statistical mean represents the DC component. With ergodicity in the auto-correlation function, the variance represents the power of the AC component.

To conclude, an ergodic WSS process is a WSS process for which the pdf that should be obtained from ensemble averages, i.e. scanning the range of possible values at a given time t, can be reconstructed or estimated by having observations over time, which enable to scan the range of values similarly to what would be done by staying at instant t.

Spectral description of random functions

9.1 Spectral representation of random signals

Let us assume that $X(t)$ is a random signal and $X_c(t) = X(t) - m_X(t)$ is its centered version. By means of the Fourier transform applied to $X_c(t)$, we obtain a new random function named random spectrum:

$$\mathcal{X}(\omega) = \int_{-\infty}^{\infty} e^{-j\omega t} X_c(t) dt \tag{9.1}$$

$$X_c(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{j\omega t} \mathcal{X}(\omega) d\omega \tag{9.2}$$

$$\tag{9.3}$$

By linearity, and as the mean of $X_c(t)$ is zero, $\mathcal{X}(\omega)$ is zero-mean too.
Its covariance is given by the following equation:

$$C_{\mathcal{X}}(\omega, \omega') = \mathbb{E}(\mathcal{X}(\omega)\mathcal{X}^*(\omega')) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C_X(t, t') e^{-j\omega t} e^{j\omega' t'} dt dt' \tag{9.4}$$

And if the process is WSS:

$$C_{\mathcal{X}}(\omega, \omega') = \underbrace{\int_{-\infty}^{\infty} C_X(\tau) e^{-j\omega\tau} d\tau}_{=:\gamma_X(\omega)} \underbrace{\int_{-\infty}^{\infty} e^{-j(\omega-\omega')t'} dt'}_{2\pi\delta(\omega-\omega')} \tag{9.5}$$

9.2 Cramèr Loève theorem

$$\boxed{C_{\mathcal{X}}(\omega, \omega') = \gamma_X(\omega) 2\pi\delta(\omega - \omega')} \tag{9.6}$$

where $\gamma_X(\omega)$ is the Fourier trasform of the covariance $C_X(\tau)$, named power spectral density (psd).This means that $\mathcal{X}(\omega)$ has uncorrelated values. Hence, the Fourier transform is the change of basis that transforms the WSS random signal $X(t)$ with correlated values in another random signal $\mathcal{X}(\omega)$, with uncorrelated values.
the relation between covariance and power spectral density is

$$\begin{cases} \gamma_X(\omega) = \int_{-\infty}^{\infty} e^{-j\omega\tau} C_X(\tau) d\tau \\ C_X(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \gamma_X(\omega) d\omega \end{cases} \tag{9.7}$$

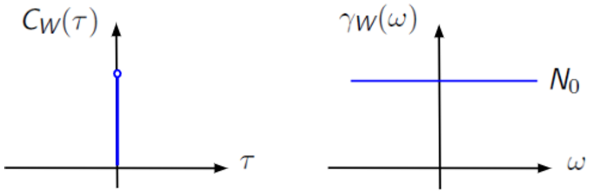


Figure 9.1: White noise covariance and psd

The variance can also be computed from the psd:

$$\sigma_X^2 = \int_{-\infty}^{\infty} \gamma_X(2\pi f) df \tag{9.8}$$

Thus the psd can be interpreted as a description of how the power is distributed along the frequency axis.

9.3 White noise

A WSS random signal is said to be a white noise if the values of this signal are uncorrelated. For such a process, the covariance and the psd are the following:

$$C_W(\tau) = N_0 \delta(\tau)$$

$$\gamma_W(\omega) = N_0 \qquad \text{for } -\infty \leq \omega \leq \infty \tag{9.9}$$

As the psd is flat, it means that the average white noise power is the same for all frequencies. For discrete time signals, a WSS sequence $X(n)$ is a white noise if

$$C_W(n) = N_0 \delta(n)$$

$$\gamma_W(e^{j\Omega}) = N_0 \qquad \text{for } 0 \leq \Omega \leq 2\pi \tag{9.10}$$

9.3.1 Additive White Gaussian Noise

A white noise signal $W(t)$ is said to be an Additive White Gaussian Noise (AWGN) if it fulfills the following two conditions:

- $W(t)$ is added to a deterministic $x(t)$ and corrupts it, and is independent of $x(t)$.
- $W(t)$ is Gaussian distributed for all instants t.

Filtering by an LTI system

10.1 Filtering by means of an LTI system

We study here the filtering of a WSS random process $X(t)$ by a system. The hypotheses on the system are

- The system is deterministic;
- the system is LTI;
- its impulse response is noted $g(t)$.

The input/output relation of the system is

$$Y(t) = \int_v X(t-v)g(v)dv \tag{10.1}$$

10.1.1 Mean of the output

The mean of the system output is given by

$$\mathbb{E}(Y(t)) = \mathbb{E}\left(\int_v X(t-v)g(v)dv\right) = \int_v \mathbb{E}(X(t-v))g(v)dv \tag{10.2}$$

$$\int_v m_X g(v)dv = m_X \underbrace{\int_v g(v)e^{-j0v}dv}_{G(\omega=0} \tag{10.3}$$

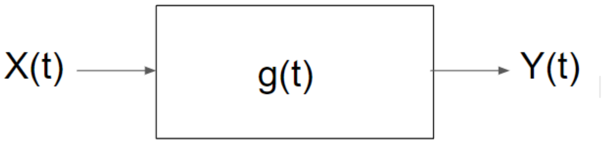
$$m_X G(0) \tag{10.4}$$

where $G(\omega)$ is the transfer function of the system, and $G(0)$ is named static gain.

10.1.2 Covariance of the system ouput

The covariance of the system output is given by

$$C_Y(t,t') = \mathbb{E}(Y(t)Y^*(t')) \quad = \int_v \int_{v'} \mathbb{E}\left(X_c(t-v)X_c^*(t'-v')\right) g(v)g^*(v')dv'dv \tag{10.5}$$



For a WSS $X(t)$, the last relation leads to

$$C_Y(t,t') = \int_v \int_{v'} C_X(t - t' + v' - v) g(v) g^*(v') dv' dv \quad = C_Y(t - t') = C_Y(\tau) \tag{10.6}$$

This means that the system output is also WSS.

10.1.3 Power spectral density of the system output - Wiener-Khintchine theorem

In the spectral domain, we have:

$$\begin{aligned} \gamma_Y(\omega) &= \int_s C_Y(s) e^{-j\omega s} ds && (10.7) \\ &= G(\omega) G^*(\omega) \gamma_X(\omega) && (10.8) \\ &= |G(\omega)|^2 \gamma_X(\omega) && (10.9) \end{aligned}$$

10.2 Generate a signal

Let $W(t)$ be a white noise. We have $C_W(\tau) = N_0 \delta(\tau)$ and $\gamma_W(\omega) = N_0$, $\forall \omega$. By WK theorem, if the input to an LTI system with a transfer function $H(\omega)$, the psd of the output will be given by

$$\gamma_Y(\omega) = |H(\omega)|^2 N_0 \tag{10.10}$$

We need to find the transfer function the filter such that $|H(\omega)|^2 = \gamma_Y(\omega)/N_0$. This is called spectral factorisation.

10.3 Generalisation to joint densities

Joint densities have the following properties:

$$\begin{aligned} \gamma_{Y,X}(\omega) &= H(\omega) \gamma_X(\omega) && (10.11) \\ \gamma_{X,Y}(\omega) &= \gamma^*_{Y,X}(\omega) = \gamma_X(\omega) H^*(\omega) && (10.12) \end{aligned}$$

where $\gamma_{X,Y}(\omega)$ and $\gamma_{Y,X}(\omega)$ are the Fourier transforms of their respective cross-covariances.

For discrete time processes,

$$\begin{aligned} \gamma_Y(z) &= H(z) \gamma_X(z) H^*(1/z^*) && (10.13) \\ \gamma_{Y,X}(z) &= H(z) \gamma_X(z) && (10.14) \\ \gamma_{X,Y}(z) &= \gamma_X(z) H^*(1/z^*) && (10.15) \end{aligned}$$

where $\gamma_X(z)$ is the psd of process $X(n)$ which is discrete time and expressed by means of the z-transform:

$$\gamma_X(z) = \sum_{p=-\infty}^{\infty} C_X(p) z^{-p} \tag{10.16}$$

Wiener filtering theory

11.1 Objective of Wiener problems

The signal $Y(k)$ is an observation of an original signal $X(k)$ and therefore correlated with it. It is a noisy or distorted version of $X(k)$. Our goal here is to estimate $X(k)$ based on the knowledge of $Y(k)$.

The idea of the theory of Wiener is to process $Y(k)$ by means of filter $w(k)$ to obtain an estimate of $X(k)$:

$$\hat{x}(k) = \sum_{l=l_1}^{l=l_2} w(l)y(k-l) \tag{11.1}$$

And we choose $w(k)$, the deterministic filter coefficients, in order to minimize the mean square error between $X(k)$ and $\hat{W}(k)$.

11.1.1 Filtering

$$\hat{x}(k) = \sum_{l=0}^{l=l_2} w(l)y(k-l) \qquad l_2 > 0 \tag{11.2}$$

The estimation of $x(k)$ is done using observations $y(j)$ until instant $j = k$. It relies on present and past observations.

11.1.2 Prediction

$$\hat{x}(k) = \sum_{l=l_1}^{l=l_2} w(l)y(k-l) \qquad 0 < l_1 < l_2 \tag{11.3}$$

The estimation of $x(k)$ is done using observations $y(j)$ until instant $j < k$. It relies on past observations only.

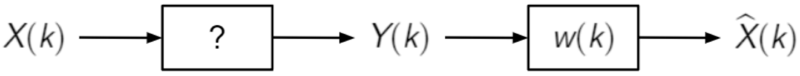


Figure 11.1: Wiener problem

11.1.3 Smoothing

$$\hat{x}(k) = \sum_{l=l_1}^{l=l_2} w(l)y(k-l) \qquad l_1 < 0 < l_2 \tag{11.4}$$

The estimation of $x(k)$ is done using observations $y(j)$ until instant $j > k$. It relies on past, present and future observations.

→ N.B.: we assume the process $Z(k) = \begin{pmatrix} X(k) \\ Y(k) \end{pmatrix}$ to be WSS.

11.2 Orthogonality principle

To find the filter coefficients $w(n)$, we minimize the MSE:

$$\zeta = \mathbb{E}\left(|E(k)|^2\right) = \mathbb{E}\left(|X(k) - \hat{X}(k)|^2\right) \tag{11.5}$$

is called the Wiener criterion.

→ N.B.: the coefficients $w(n)$ can be complex : $w(k) = w_r(k) + w_i(k)$

For the filter to be optimum, we require

$$\frac{\partial\zeta}{\partial w_r(l)} = 0 \qquad l = l_1, \ldots, l_2 \frac{\partial\zeta}{\partial w_i(l)} = 0 \qquad l = l_1, \ldots, l_2 \tag{11.6}$$

And the second derivatives should be positive for minimum.

Computing the derivatives, we find the orthogonality principle:

$$\mathbb{E}\left(E(k)Y^*(k-l)\right) = 0 \qquad \text{for } l = l_1, \ldots, l_2 \tag{11.7}$$

Corollary:

$$R_{\hat{X}E}(0) = \mathbb{E}\left(\hat{X}(k)E^*(k)\right) = 0 \tag{11.8}$$

with R the correlation. Since the estimate $\hat{X}(k)$ is a linear combinsation of the observations, it turns out that the mutual correlation $R_{\hat{X}E}(0)$, between the error and the estimate, is also equal to zero.

11.3 Finite impulse response (FIR) filters

11.3.1 FIR Filter of order N

Here, we will expand the orthogonality principle in the case of a filtering problem, with a finite number of coefficients: $l_1 = 0$ and $l_2 = N - 1$. Let us first replace $E(k)$ with

$$E(k) = X(k) - \hat{X}(k) = X(k) - \sum_{j=0}^{N-1} w(j)Y(k-j) \tag{11.9}$$

in the orthogonality principle:

$$\mathbb{E}\left((X(k)-\hat{X}(k))Y^*(k-l)\right)=0 \tag{11.10}$$

$$\iff \mathbb{E}\left(X(k)Y^*(k-l)\right) = \sum_{j=0}^{N-1} w(j)\mathbb{E}\left(Y(k-j)Y^*(k-l)\right) \tag{11.11}$$

Which gives the Wiener-Hopf equations:

$$R_{XY}(l)=\sum_{j=0}^{N-1} w(j)R_Y(l-j) \qquad 0\le l\le N-1 \tag{11.12}$$

By developping this equation, we obtain a linear system:

$$\begin{bmatrix} R_Y(0) & R_Y(-1) & \dots & R_Y(-N+1) \\ R_Y(1) & R_Y(0) & \dots & R_Y(-N+2) \\ \dots & \dots & \ddots & \vdots \\ R_Y(N-1) & \dots & \dots & R_Y(0) \end{bmatrix} \begin{bmatrix} w(0) \\ w(1) \\ \vdots \\ w(N-1) \end{bmatrix} = \begin{bmatrix} R_{XY}(0) \\ R_{XY}(1) \\ \vdots \\ R_{XY}(N-1) \end{bmatrix} \tag{11.13}$$

As the correlation function is hermitian, i.e. $R_Y(-k)=R_Y^*(k)$, then

$$\underbrace{\begin{bmatrix} R_Y(0) & R_Y^*(1) & \dots & R_Y^*(N-1) \\ R_Y(1) & R_Y(0) & \dots & R_Y^*(N-2) \\ \dots & \dots & \ddots & \vdots \\ R_Y(N-1) & \dots & \dots & R_Y(0) \end{bmatrix}}_{R_Y} \underbrace{\begin{bmatrix} w(0) \\ w(1) \\ \vdots \\ w(N-1) \end{bmatrix}}_{\mathbf{w}} = \underbrace{\begin{bmatrix} R_{XY}(0) \\ R_{XY}(1) \\ \vdots \\ R_{XY}(N-1) \end{bmatrix}}_{R_{XY}} \tag{11.14}$$

The optimal solution then is

$$\mathbf{w}_o = R_Y^{-1}R_{XY} \tag{11.15}$$

11.3.2 FIR - Analysis of the error

For the optimal filter, we have

$$\zeta_{\min} = \mathbb{E}\left(E(k)E^*(k)\right) \tag{11.16}$$

$$= \mathbb{E}\left(X(k)X^*(k)\right) - \mathbb{E}\left(\hat{X}(k)X^*(k)\right) \tag{11.17}$$

$$= R_X(0) - w_o^T R_{XY}^* \tag{11.18}$$

$$= R_X(0) - R_{XY}^T R_Y^{-T} R_{XY}^* \tag{11.19}$$

For a non optimal coefficient vector, the error is

$$\zeta = \zeta_{\min} + \sum_k \lambda_k |\nu_k|^2 \tag{11.20}$$

with λ_k the eigenvalues of R_Y, and ν_k the k-th component of the vector $\nu = M(w-w_o)$, M being the eigenvector matrix of R_Y.

Spectral factorization

The spectral factorization consists in finding, for a given $\gamma(\omega)$, $L(\omega)$ such that

$$\gamma(\omega) = L(\omega)L^*(\omega)\sigma^2 \tag{12.1}$$

12.1 Continuous case

A solution $L(s)$ exists if the Paley-Wiener condition is fulfilled:

$$\int_{-\infty}^{\infty} \frac{|\ln \gamma_Y(\omega)|}{1+\omega^2} d\omega < \infty \tag{12.2}$$

We focus here on a real and rational process. Therefore, $C_X(\tau)$ is real and even, and thus $\gamma(\omega)$ too. This means that $\gamma(\omega)$ should be function of ω^2 and rational. We can then write it this way:

$$\gamma_Y(\omega) = \frac{N(\omega^2)}{D(\omega^2)} \tag{12.3}$$

and with the Laplace transforms:

$$\gamma_Y(s) = \frac{N(-s^2)}{D(-s^2)} \tag{12.4}$$

As $\gamma(\omega)$ is real, its roots are either real of complex conjugate, meaning they are simetrically distributed in the complex plane. We therefore choose for $L(s)$ all roots with negative real part:

$$\gamma_Y(s) = \frac{N(-s^2)}{D(-s^2)} = \frac{C(s)C(-s)}{A(s)A(-s)} = L(s)L(-s) \tag{12.5}$$

defining

$$L(s) := \frac{C(s)}{A(s)} \tag{12.6}$$

With this definition, we can demonstrate that $L(s)$ is causal and stable, with an inverse also causal and stable.We say that is is a minimal phase filter.

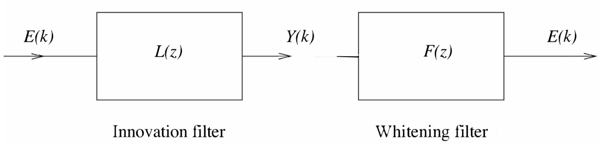


Figure 12.1: Discrete spectral factorization

12.2 Discrete case

In the discrete state, we have the same relation between $\gamma(e^{j\Omega})$ and $L(e^{j\Omega})$ as in continous time:

$$\gamma_Y\left(e^{j\Omega}\right) = L\left(e^{j\Omega}\right)L^*\left(e^{j\Omega}\right)\sigma^2 \tag{12.7}$$

The Paley Wiener criterion is also identical:

$$\int_{-\pi}^{\pi} |\ln \gamma_Y(e^{j\Omega})| d\Omega < \infty \tag{12.8}$$

The factorization becomes

$$\gamma_Y(e^{j\Omega}) = L(e^{j\Omega})L^*(e^{j\Omega})\sigma_E^2 = |L(e^{j\Omega})|^2\sigma_E^2 \tag{12.9}$$
$$\gamma_Y(z) = L(z)L^*(1/z)\sigma_E^2 \tag{12.10}$$
$$\tag{12.11}$$

As we did is the continuous case, we can find a $L(z)$ that is causal and stable, and we can normalise it such that $\lim_{z\to\infty} L(z) = 1$. This means that

$$L(z) = \sum_{i=0}^{\infty} l(i)z^i \tag{12.12}$$

and it is said to be monic. Therefore, the random process can be expressed as

$$Y(k) = \sum_{j=0}^{\infty} l(j)E(k-j) = E(k) + \sum_{j=1}^{\infty} l(j)E(k-j) \tag{12.13}$$

defining the innovation E(k) based on the inverse filter:

$$E(k) = \sum_{j=0}^{\infty} f(j)Y(k-j) = Y(k) + \sum_{j=1}^{\infty} f(j)Y(k-j) \tag{12.14}$$

The relations between all those signals is the following Thus, if we know $Y(k)$, we know $E(k)$ and that means that knowing one is equivalent to knowing the other. However, working with the innovation is more interesting: the expressions of the predictors are simpler and the innovation elements are uncorrelated, hence containing less redundancy and enabling more compact encoding.

12.2.1 Horizon 1 predictor

We know that

$$Y(k) = E(k) - \sum_{j=1}^{\infty} f(j)Y(k-j) \tag{12.15}$$

Assuming Y is observed until $k-1$ and we want to predict $y(k)$, then we choose the condition expectation of $Y(k)$ as a predictor:

$$\mathbb{E}(Y(k)|\{Y(j), j \leq k-1\}) \quad = \mathbb{E}(E(k)|\{Y(j), j \leq k-1\}) - \sum_{j=1}^{\infty} f(j) y(k-j) = -\sum_{j=1}^{\infty} f(j) y(k-j) \tag{12.16}$$

As the expectation of $Y(k)$ does not depend on $E(k)$, observing Y until $k-1$ means that we miss the information associated with $E(k)$.

12.2.2 Real process with rational psd

For the reason stated above, we have

$$\gamma_Y(z) = L(z)L(1/z)\sigma_E^2 = \frac{C(z)C(1/z)}{A(z)A(1/z)}\sigma_E^2 \tag{12.17}$$

and we have the reccurence equation

$$\sum_{i=0}^{n} a_i Y(k-i) = \sum_{i=0}^{n} c_i E(k-i) \tag{12.18}$$

- If $a_k = 0$ for all k, then we have a moving average on the y and call it MA.
- If $c_k = 0$ for all k, then we call it autoregressive (AR).
- If $a_k \neq 0$ and $c_j \neq 0$ for at least one k and one j then it is ARMA (autoregressive and moving average).

→ N.B.: a first order autoregressive system is $Y(k) = E(k) + \alpha Y(k-1)$.