

---

# MODEL IDENTIFICATION & DATA ANALYSIS

---

## PART 1

BY YANNICK GIOVANAKIS

April 7, 2018

# Contents

<b>0</b>	<b>Introduction</b>	<b>4</b>
0.1	Time-Series . . . . .	4
0.1.1	TS Applications . . . . .	4
0.2	I/O Systems . . . . .	5
0.2.1	I/O Applications . . . . .	5
0.3	Time Series vs I/O Systems . . . . .	6
0.4	Modelling structures . . . . .	7
0.5	Mathematical Models . . . . .	8
0.5.1	White Box Models . . . . .	8
0.5.2	Black Box Models . . . . .	8
0.5.3	White box vs Black box . . . . .	9
0.6	Stochastic Processes . . . . .	9
0.6.1	Characteristics . . . . .	9
0.6.2	Stationary Stochastic Processes . . . . .	10
0.6.3	White Noise . . . . .	11
0.7	Sample estimation of mean and covariance function . . . . .	12
0.7.1	Sample Mean . . . . .	12
0.7.2	Sample Covariance . . . . .	13
<b>1</b>	<b>Chapter 1</b>	<b>15</b>
1.1	Model classes . . . . .	15
1.1.1	Time-Series model classes . . . . .	15
1.1.2	Input/Output model classes . . . . .	17
1.2	Transfer function representation . . . . .	18
1.2.1	Z Operator . . . . .	18
1.2.2	Time domain to Transfer Function . . . . .	18
1.2.3	From $Z^-$ to $Z^+$ . . . . .	19
1.2.4	Importance of stationary property . . . . .	20
1.2.5	Pole,Zeros and Stability . . . . .	21
1.2.6	Stationary property and stability . . . . .	22
1.2.7	Poles and Zeros in MA & AR processes . . . . .	23

<b>2</b>	<b>Chapter 2 : Analysis of Stochastic Processes</b>	<b>24</b>
2.1	Probabilistic Representation . . . . .	24
2.1.1	Probabilistic representation of MA(n) . . . . .	24
2.1.2	Probabilistic representation of AR(1) . . . . .	25
2.1.3	AR/ARMA as MA( $\infty$ ) . . . . .	27
2.2	Frequency Representation . . . . .	28
2.3	Inverse Fourier Transform . . . . .	28
2.4	White Noise in the frequency domain . . . . .	30
2.5	Computation of the spectrum of a process generated as the output of a digital system . . . . .	31
2.5.1	Frequency Response of a linear system . . . . .	31
2.5.2	Spectrum computation with FR . . . . .	33
2.6	Equivalent representations of ARMA . . . . .	33
2.7	Example & Exercises . . . . .	34
<b>3</b>	<b>Chapter 3 : Prediction</b>	<b>41</b>
3.1	All-Pass Filter . . . . .	42
3.2	Canonical Representation . . . . .	43
3.3	Predictor . . . . .	45
3.3.1	Optimality . . . . .	46
3.3.2	1-step ahead prediction of MA(n) . . . . .	46
3.3.3	K-steps ahead predictor of MA(n) . . . . .	48
3.3.4	K-steps ahead predictor of general ARMA(m,n) . . . . .	48
3.3.5	K-steps ahead prediction of ARMAX(m,n,k+p) . . . . .	51
3.4	Examples & Exercises . . . . .	52
3.4.1	Example 1 . . . . .	52
3.4.2	Example 2 - Practical . . . . .	57
3.4.3	Example 3 - ARMAX & ARX . . . . .	59
3.4.4	Example ARMA with non-zero mean . . . . .	60
<b>4</b>	<b>Chapter 4 : Identification</b>	<b>64</b>
4.1	Identification of ARX models . . . . .	67
4.1.1	Loss function: Least Squares . . . . .	67
4.1.2	Example . . . . .	68

4.1.3	Example 2 . . . . .	71
4.2	Identification of ARMAX models . . . . .	73
4.2.1	Loss function :Maximum Likelihood Method . . . . .	74
4.2.2	Possible updating rules . . . . .	77

## 0 Introduction

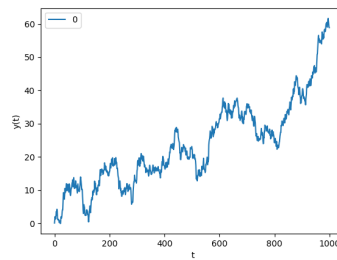
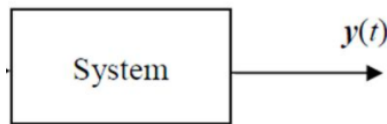
The course will deal with two types of situations :

1. Analysis and modelling of **Time-Series**
2. Analysis and modelling of **Input/Output Systems**

### 0.1 Time-Series

Time series consider vectors  $\{y(1), y(2), \dots, y(N)\}$  of **measured data** of cardinality  $N$  ( large , 1000 - 10000).

Said vectors are considered in the **time-domain** :  $y(t)$  is a signal or **stochastic process** generated by the system whose output is than sampled.



#### 0.1.1 TS Applications

TS are used for two problems :

1. **Prediction problem** :  $\{y(1) \dots y(N)\} \rightarrow \hat{y}(\frac{N+K}{N})$   
Given  $N$  measurements **estimate** the measurement  $K$  timesteps ahead
2. **Filtering problem** :  $\{x_1(t) \dots x_N(t)\} \rightarrow \hat{x}(\frac{t}{t})$   
Where  $\{x_1(t) \dots x_N(t)\}$  are internal variables of the system

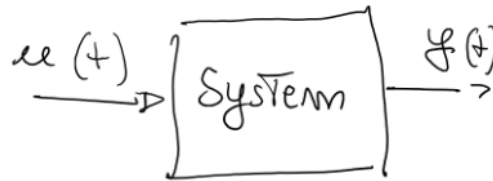
## 0.2 I/O Systems

I/O systems consider two measurements :

- **Input** :  $\{u(1)...u(N)\}$
- **Output**:  $\{y(1)...y(N)\}$

Resulting in two signals  $u(t)$  and  $y(t)$ . Input signal  $u(t)$  can be of two types:

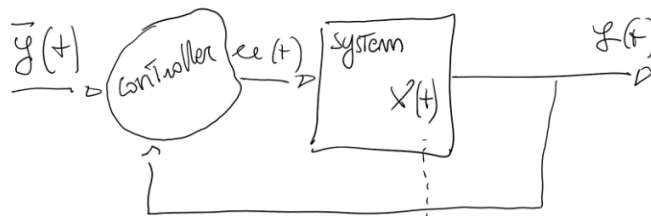
- **Controllable** : can be affected ( ex : voltage )
- **Uncontrollable** : cannot be affected ( ex : rain )



### 0.2.1 I/O Applications

I/O systems are used for three problems :

1. **Prediction problem** :  $\{y(1)...y(N)\} \rightarrow \hat{y}(\frac{N+K}{N})$   
Given N measurements **estimate** the measurement K timesteps ahead
2. **Filtering problem** :  $\{x_1(t)...x_N(t)\} \rightarrow \hat{x}(\frac{t}{t})$   
Where  $\{x_1(t)...x_N(t)\}$  are internal variables of the system
3. **System control problem** : given a desired output  $\bar{y}(t)$  , control  $u(t)$  so that  $y(t)$  is as close as possible to  $\bar{y}(t)$



### 0.3 Time Series vs I/O Systems

In prediction and filtering problems both I/O systems and TS can be used. How to choose which one to use?

#### Ex.1

- **System** : Electric Motor
- **Input** : Current , temperature of motor,electromagnetic fields nearby..
- **Output** : Torque

We can say that our main input variable (current) is responsible for 90% of the output.The other variables only have slight effects on the torque so they are considered **noise**

The best model to choose is the **I/O**

#### Ex.1

- **System** : Macro-Economic System
- **Input** : Too many
- **Output** : Stock prices of FCA

There are many thousand variables affecting the output. Listing and measuring them all would make the model too complex . In this case all the input variables are considered **noise** : the best model to choose is the **Time Series**

#### Ex.3

- **System** : Environment
- **Input** : Rain, wind, heatings, cars , temperature, pressure...
- **Output** : PM10 levels

In this case some main inputs variables can be selected ( ex :cars , heating and rain ) while the others are modelled as noise. In this case **I/O** model should be used.

It is not wrong to consider all the inputs as noise and model the problem as **Time Series**.

General rule:

	Advantages
TS	Only $y(t)$ must be measured
I/O	Better estimation

## 0.4 Modelling structures

Depending on the problem 2 modelling structures are used.

The TS are modelled with a **mathematical model** which outputs signal  $y(t)$ . An **imaginary** input  $e(t)$  called **white noise** is considered as **standard input** and it is **part of the model**.

The I/O system is modelled by two **mathematical models** which output signal  $y(t)$ . As above **white noise** is considered as input of one of the two models. The other model has input  $u(t)$  which is **not** part of the model.

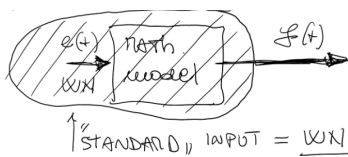


Figure 1: TS Model

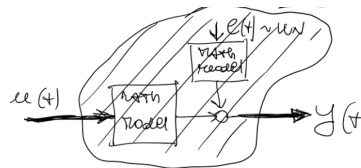


Figure 2: IO Model

All signals and systems are **time-discrete**. Analogue signals are converted to digital signals through **ADCs**.

Discrete time points are spaced evenly at pace  $\Delta T =$  sampling time



## 0.5 Mathematical Models

The mathematical models used to elaborate output functions are either **white boxes** or **black - boxes**.

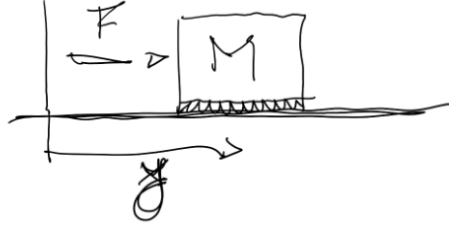


Figure 3: System to be modelled

### 0.5.1 White Box Models

Also called *first-principles models* assume that the parameters involved in the system are known and well defined. Using white box models we get a **physical interpretation of the model** which makes them useful if the aim is to design the system.

In the example we can derive laws that define our system's **transfer function** given as input a force  $\vec{F}$  and output  $y$  :

$$M\ddot{y} = F - c\dot{y} \rightarrow \text{Laplace} \rightarrow s^2My = F - scy$$

$$(s^2M + sc)y = F$$

$$y = \frac{1}{s^2M + sc}F$$

### 0.5.2 Black Box Models

In black box models we don't know the internal parameters that influence the system. In our example , we only know that by changing the input  $\vec{F}$  a corresponding change in output  $y(t)$  can be measured . By measuring the data we can derive a model :

$$y(t) = \frac{b_0Z^2 + b_1Z + b_2}{a_0Z^2 + a_1Z + a_2}F(t)$$

where  $a_0, \dots, a_2, b_0, \dots, b_2$  are the parameters.

### 0.5.3 White box vs Black box

Table 1: WB/BB Comparison

White Box	Black Box
-Get physical interpretation of the model and its parameters. -Useful for designing the system	-Very fast -Very accurate -Does not require know-how of the domain -Can be easily re-tuned

## 0.6 Stochastic Processes

### Random variable RV:

$v(s)$  is completely defined by its probability distribution ( Gaussian, Uniform...) which is related to its **probability density function** ( PDF )

### Stochastic Process:

is a sequence of **time-ordered random variables** defined at the same experiment  $S$

$$v(1, S), v(2, S), \dots, v(t, S)$$

where  $t$  is the time index. If the experiment is **fixed**  $S = \bar{S}$  , we get an instance , a **realisation** of the stochastic process :

$$v(1, \bar{S}), \dots, v(t, \bar{S})$$

resulting in a set of samples  $\{y(1), \dots, y(N)\} = \{y(1, \bar{S}), \dots, y(N, \bar{S})\}$

### 0.6.1 Characteristics

#### Mean value $m(t)$ :

expected value of a random variable  $v(t, S)$  at time  $t$

$$m(t) = E[v(t, S)]$$

### Covariance Function $\gamma(t_1, t_2)$ :

expected value of the **product** of two **unbiased** random variables at time instants  $t_1, t_2$  :

$$\gamma(t_1, t_2) : E[(v(t_1, S) - m(t_1))(v(t_2, S) - m(t_2))]$$

Removing the mean brings the signal closer to 0.

If  $t_1 = t_2 = t$  the covariance degenerates in **variance**:

$$\gamma(t) = E[(v(t, S) - m(t))^2]$$

### 0.6.2 Stationary Stochastic Processes

Has properties:

1.  $m(t) = m, \forall t$
2.  $\gamma(t_1, t_2)$  depends on  $\tau = |t_1 - t_2|$

This means that the covariance depends on the **distance in time** and not on specific considered samples.

$$\gamma(t_1, t_2) = \gamma(t_3, t_4) \rightarrow |t_1 - t_2| = |t_3 - t_4|$$

$\gamma(\tau) = E[(v(t) - m)(v(t - \tau) - m)]$  has properties :

- $\gamma(0) = E[(v(t) - m)^2] \rightarrow \text{variance}$
- $|\gamma(\tau)| \leq \gamma(0)$
- $\gamma(\tau) = \gamma(-\tau)$

### SSP Equivalence

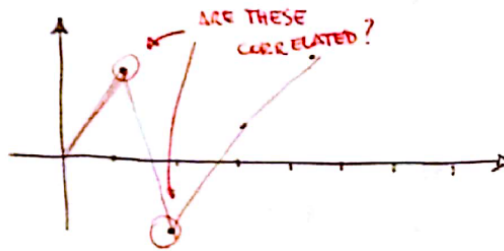
Two SSPs  $y_1(t), y_2(t)$  are equivalent in a **weak sense** if:

- $m_{y1} = m_{y2}$
- $\gamma_{y1}(\tau) = \gamma_{y2}(\tau), \forall \tau$

## Correlation Function

If the  $m = 0$  the  $\gamma(\tau)$  function degenerates in the **correlation function** :

$$E[v(t)v(t - \tau)]$$



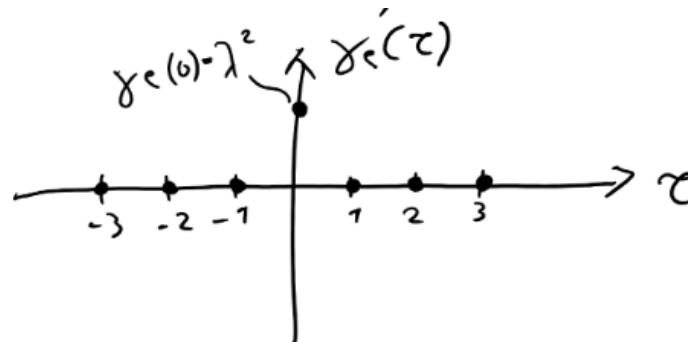
### 0.6.3 White Noise

$e(t)$  is SSP called **white noise** and is written as

$$e(t) \rightarrow WN(\mu, \lambda^2)$$

Properties:

- Mean value :  $E[e(t)] = \mu, \forall t$
- Variance :  $\gamma_e(0) = E[(e(t) - \mu)^2] = \lambda^2$
- Covariance :  $E[(e(t) - \mu)(e(t - \tau) - \mu)] = 0, \forall t, \forall \tau \neq 0$



No covariance means that the samples are **not related**

Considering a **Gaussian Distribution** :  $e(t) \rightarrow WGN(\mu, \lambda^2)$

## 0.7 Sample estimation of mean and covariance function

Dealing with samples it is useful to **estimate** the mean and covariance of the samples.

Output  $y(t)$  is a SSP :  $\{y(1), \dots, y(N)\}$  a particular realisation of  $\bar{S}$  with :

- Mean  $m = E[y(t)]$
- Covariance  $\gamma(\tau) = E[(y(t) - m)(y(t - \tau) - m)]$

This seems trivial but the computation of the expected value cannot be done because the **distribution of the process** is **unknown**.

These two can be **estimated**

### 0.7.1 Sample Mean

The sample mean is a good estimator for the mean  $m$  :

$$\hat{m}_n = \frac{1}{N} \sum_{t=1}^N y(t)$$

Properties of the estimator :

1.  $\hat{m}_n$  is **correct** if  $E[\hat{m}_n] = m$

**Proof:**  $E[\hat{m}_n] = E\left[\frac{1}{N} \sum_{t=1}^N y(t, s)\right] = \frac{1}{N} \sum_{t=1}^N E[y(t, s)] = \frac{1}{N} \sum_{t=1}^N m = m$

#### Example

$y(t, S) = \bar{v}(s) \rightarrow WN(0, 1)$  and  $S = \bar{S}, \{y(1, \bar{S}), \dots, y(N, \bar{S})\}$  so :

$$-\hat{m}_n = \frac{1}{N} \sum_{t=1}^N y(t, \bar{S}) = \frac{1}{N} \sum_{t=1}^N \bar{v}(\bar{S}) = \frac{1}{N} N \bar{v}(\bar{S}) \neq 0 \rightarrow \text{bad estimator}$$

$$-\check{m}_n = \frac{1}{N} \sum_{S=1}^N y(\bar{t}, S) = \frac{1}{N} v(S) \rightarrow 0 \rightarrow \text{good estimator}$$

2.  $\hat{m}_n$  is **consistent** if  $E[(\hat{m}_n - m)^2] \xrightarrow{N \rightarrow \infty} 0$

The **error variance** approaches 0 for large values of  $N$ : this means that with a lot of data  $N \rightarrow \infty$  we can estimate  $\hat{m}_n$  more effectively.

In general one can say that  $\hat{m}_n$  is consistent if  $\gamma(\tau) \xrightarrow{|\tau| \rightarrow \infty} 0$

**Example**

$$y(t, S) = \bar{V}(S) \rightarrow WN(0, 1)$$

$$\gamma(\tau) = E[(\gamma(\tau))(\gamma(t - \tau))] = E[\bar{V}(S)\bar{V}(S)] = E[\bar{V}(S)^2] = 1$$

**0.7.2 Sample Covariance**

$y(t)$  is a SSP with **zero mean**.

A good estimator for the covariance is the **sample covariance**:

$$\hat{\gamma}_N(\tau) = \frac{1}{N - \tau} \sum_{t=1}^{N-\tau} y(t)y(t + \tau)$$

$$0 \leq \tau \leq N - 1$$

It is important to notice that this approximation is good for  $\tau \ll N$  because the accuracy of  $\gamma_N(\tau)$  **decreases** with  $\tau$

Properties of the estimator :

1.  $\hat{\gamma}_N(\tau)$  is **correct** if  $E[\hat{\gamma}_N(\tau)] = \gamma(\tau)$

**Proof:**

$$E[\hat{\gamma}_N(\tau)] = E\left[\frac{1}{N-\tau} \sum_{t=1}^{N-\tau} y(t)y(t+\tau)\right] = \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} E[y(t)y(t+\tau)] = \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} \gamma(\tau) = \gamma(\tau)$$

2.  $\hat{\gamma}_N(\tau)$  is **consistent** if  $E[(\hat{\gamma}_N(\tau) - \gamma(\tau))^2] \xrightarrow[N \rightarrow \infty]{} 0$ ,

**True** if  $\gamma(\tau) \xrightarrow[|\tau| \rightarrow \infty]{} 0$

**Observation 1:**

$$\hat{\gamma}_N(\tau) = \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} y(t)y(t + \tau)$$

$$0 \leq \tau \leq N - 1, \tau \geq 0$$

but since  $y(t)$  is a SSP  $\gamma(\tau) = \gamma(-\tau)$  :

$$\hat{\gamma}_N(\tau) = \frac{1}{N - |\tau|} \sum_{t=1}^{N-|\tau|} y(t)y(t + |\tau|)$$

$$|\tau| \leq N - 1$$

**Observation 2:**

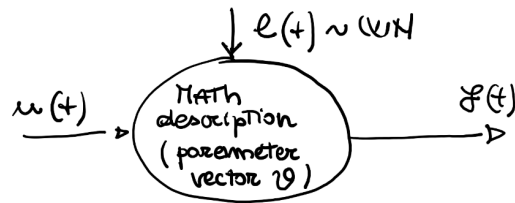
$$\hat{\gamma}'_N(\tau) = \frac{1}{N-|\tau|} \sum_{t=1}^{N-|\tau|} y(t)y(t+|\tau|) \rightarrow E[\hat{\gamma}'_N(\tau)] = \dots = \frac{1}{N}\gamma(\tau)(N-|\tau|)$$

As shown  $\hat{\gamma}'_N(\tau)$  **doest not** satisfy the **correct** property.

However for  $N \rightarrow \infty$  and  $\tau \ll N$  :  $\hat{\gamma}'_N(\tau)$  is **asimptotically correct**

# 1 Chapter 1

## 1.1 Model classes



$$\text{Mathematical model} = \begin{cases} u(t) & \text{input (I/O only)} \\ e(t) & \text{white noise} \\ y(t) & \text{output} \end{cases}$$

The mathematical model is described by **parametric parameter vector**  $\theta$  that is found using a **parametric supervised** identification approach.

The models can be described with:

- **Differential Equations** in time domain
- **Transfer functions**

### 1.1.1 Time-Series model classes

The following processes are modelled with **differential equations**

#### 1. Moving Average Models (MA):

A process  $y(t)$  **generated** by a WN  $e(t)$  is a moving average of order  $n$  **MA(n)** process if:

$$y(t) = c_0 e(t) + c_1 e(t-1) + \dots + c_n e(t-n)$$

with parameter vector  $\theta = \{c_0, \dots, c_n\}$



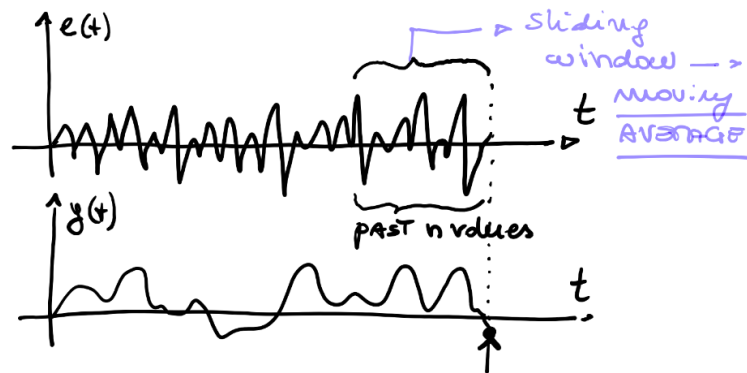


Figure 4:  $y(t)$  is linear combination of past  $n$   $e(t)$  values

## 2. Autoregressive Models (AR):

A process  $y(t)$  **generated** by a WN  $e(t)$  is an autoregressive of order  $m$  **AR(m)** process if:

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + \dots + a_m y(t-m) + c_0 e(t)$$

with parameter vector  $\theta = \{c_0, a_1, \dots, a_m\}$

## 3. Autoregressive Moving Average Models (ARMA):

A process  $y(t)$  **generated** by a WN  $e(t)$  is an ARMA of order  $(n, m)$  **ARMA(n, m)** process if:

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + \dots + a_m y(t-m) + c_0 e(t) + c_1 e(t-1) + \dots + c_n e(t-n)$$

with parameter vector  $\theta = \{c_0, \dots, c_n, a_1, \dots, a_m\}$

$\text{ARMA}(0, n) \rightarrow \text{MA}(n)$  :  $\text{MA}(n)$  is **subclass** of ARMA

$\text{ARMA}(m, 0) \rightarrow \text{AR}(m)$  :  $\text{AR}(m)$  is **subclass** of ARMA

### 1.1.2 Input/Output model classes

The following processes are modelled with **differential equations**

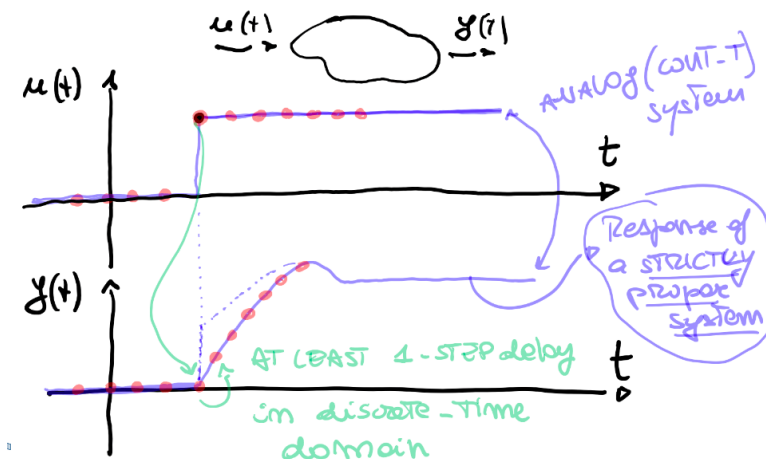
#### 1. Autoregressive Moving Average Exogenous (ARMAX):

A process  $y(t)$  **generated** by a WN  $e(t)$  and **exogenous** signal  $u(t)$  is an ARMAX of order  $(n, m, p+k)$  process if:

$$y(t) = a_1 y(t-1) + \dots + a_m y(t-m) + c_0 e(t) + \dots + c_n e(t-n) + b_0 u(t-k) + \dots + b_p u(t-k-p)$$

with parameter vector  $\theta = \{c_0, \dots, c_n, a_1, \dots, a_m, b_0, \dots, b_p\}$

$K \geq 1$  plays an important role : it represents the pure/intrinsic **delay** between  $y(t)$  and  $u(t)$ . If  $u(t)$  is a step the corresponding output  $y(t)$  is



shown in figure.

Sampling (red dots) gives a discrete approximation : when the input slope rises a sample is taken resulting in a high value. The corresponding output is still low : this causes a **1 step delay**

#### Example

$y(t) = \frac{1}{2}y(t-1) + \frac{1}{3}y(t-2) + e(t) + e(t-3) + u(t-2) + \frac{1}{2}u(t-4)$  The process is an ARMAX (2,3,2+2)

Observation : missing values as above can be present!

#### Remark

Armax models are the most general class models for **dynamic ,linear, time-invariant** systems.

**Non-Linear** N-ARMAX  $y(t) = f(y(t-1), \dots, y(t-m), e(t), \dots, e(t-n), u(t-k), \dots, u(t-k-p))$   
depend on **non-linear functions** : polynomials , splines ,NN ,Radial Basis Functions ,Fuzzy Sets.

## 1.2 Transfer function representation

The four models found above can be represented using **transfer functions**. To transform time domain equations into the equivalent transfer function representation the **Z operator** is introduced.

### 1.2.1 Z Operator

- The operator  $Z^{-1}$  is the **backward shift** operator :

$$Z^{-1}x(t) = x(t-1)$$

- The operator  $Z^{+1}$  is the **forward shift** operator :

$$Z^{+1}x(t) = x(t+1)$$

Both operators have properties :

- **Linearity** :  $Z^{-1}(ax(t)+by(t)) = Z^{-1}ax(t) + Z^{-1}by(t) = ax(t-1) + by(t-1)$
- **Recursion** :  $Z^{-1}(Z^{-1}...(Z^{-1}x(t))) = x(t-n) = Z^{-n}$

### 1.2.2 Time domain to Transfer Function

The Z operators are used to shift the equations of the time domain to be all at time **t**.

In case of a generic **ARMAX(m,n,p+k)** process

$$y(t) = a_1y(t-1) + \dots + a_my(t-m) + c_0e(t) + \dots + c_ne(t-n) + b_0u(t-k) + \dots + b_pu(t-k-p)$$

Applying the  $Z^{-1}$  operator:

$$y(t) = a_1Z^{-1}y(t) + \dots + a_mZ^{-m}y(t) + c_0e(t) + \dots + c_nZ^{-n}e(t) + b_0Z^{-k}u(t) + \dots + b_pZ^{-k-p}u(t)$$

Collecting :

$$y(t)[1 - a_1 Z^{-1} + \dots + a_m Z^{-m}] = [c_0 e + \dots + c_n Z^{-n}]e(t) + [b_0 Z^{-k} + \dots + b_p Z^{-k-p}]u(t)$$

Dividing :

$$y(t) = \frac{[c_0 e + \dots + c_n Z^{-n}]}{[1 - a_1 Z^{-1} + \dots + a_m Z^{-m}]}e(t) + \frac{[b_0 + \dots + b_p Z^{-p}]}{[1 - a_1 Z^{-1} + \dots + a_m Z^{-m}]}u(t)Z^{-k}$$

Defining :

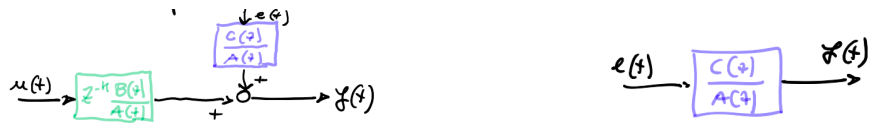
$$A(Z) = 1 - a_1 Z^{-1} + \dots + a_m Z^{-m}$$

$$B(Z) = b_0 + \dots + b_p Z^{-p}$$

$$C(Z) = c_0 e + \dots + c_n Z^{-n}$$

The resulting process using TF representation is :

$$y(t) = \frac{C(Z)}{A(Z)}e(t) + \frac{B(Z)}{A(Z)}u(t)Z^{-k}$$



### 1.2.3 From $Z^-$ to $Z^+$

The transfer functions can be written in negative, positive or mixed power of  $Z$ . The example explains how to get the positive power representation starting from a negative one :

$$y(t) = \frac{c_0 + c_1 Z^{-1} + \dots + c_n Z^{-n}}{1 - a_1 Z^{-1} - \dots - a_m Z^{-m}}e(t)$$

If  $m \geq n$  by multiplying by  $Z^{+m}$  :

$$y(t) = \frac{c_0 Z^m + c_1 Z^{m-1} + \dots + c_n Z^{m-n}}{Z^m - a_1 Z^{m-1} - \dots - a_m}e(t)$$

### Observation

Even if feasible and correct it is better to **avoid** the mixed representation!

### 1.2.4 Importance of stationary property

Transformation **Time Domain**  $\leftrightarrow$  **Transfer Functions** are **feasible** if the **stationary property** holds because otherwise the Z operator is not applicable.

$$y(t) = \frac{Z + \frac{1}{2}}{Z - \frac{1}{3}}e(t), e(t) \sim \text{WN}(0,1)$$

**In time domain**

$$\begin{aligned}(Z - \frac{1}{3})y(t) &= (Z + \frac{1}{2})e(t) \\ y(t+1) - \frac{1}{3}y(t) &= e(t+1) + \frac{1}{2}e(t) \\ y(t+1) &= \frac{1}{3}y(t) + e(t+1) + \frac{1}{2}e(t)\end{aligned}$$

Time shift to start a time "t" ( can be done in stationary conditions):

$$y(t) = \frac{1}{3}y(t-1) + e(t) + \frac{1}{2}e(t-1)$$

**Back to TF**

$$\begin{aligned}y(t) &= \frac{1}{3}Z^{-1}y(t) + e(t) + \frac{1}{2}Z^{-1}e(t) \\ [1 - \frac{1}{3}Z^{-1}]y(t) &= [1 + \frac{1}{2}Z^{-1}]e(t) \\ y(t) &= \frac{[1 + \frac{1}{2}Z^{-1}]}{[1 - \frac{1}{3}Z^{-1}]}e(t)\end{aligned}$$

### 1.2.5 Pole,Zeros and Stability



Considering a process with signals  $e(t)$  ,  $y(t)$  and a system  $W(Z)$  represented in **positive/null** power:

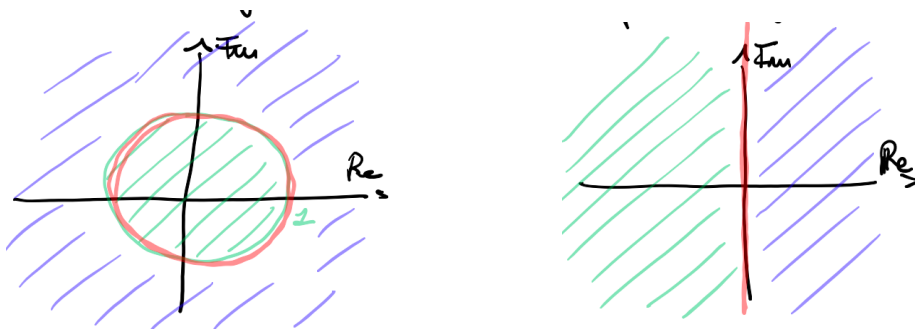
- **Poles** of  $W(Z)$  are the **roots** of the denominator
- **Zeros** of  $W(Z)$  are the **roots** of the nominator

A system is said to be **asymptotically stable** if and only if all the **poles** of  $W(Z)$  are **strictly inside** the unit circle (left graph).

Blue = unstable region

Red = simple stability region

Green = asymptotically stability region



**Note:** if we were dealing with **continuous** signals and processes instead of  $Z$  transformation we would apply **Laplace** . Also the stability region would change as seen on the right graph.

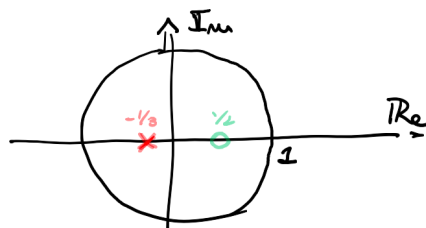
**Example:**

$$W(Z) = \frac{1 - \frac{1}{2}Z^{-1}}{1 + \frac{1}{3}Z^{-1}}$$

Move to positive power:

$$W(Z) = \frac{Z - \frac{1}{2}}{Z + \frac{1}{3}}$$

- **Pole** :  $Z = -\frac{1}{3}$
- **Zero** :  $Z = \frac{1}{2}$



The system is asymptotically stable since all poles are within the unit circle.

### 1.2.6 Stationary property and stability



In a stochastic process  $y(t)$  obtained as output of a system  $W(Z)$  fed with a stochastic process  $v(t)$ ,  $y(t)$  is a **stationary process SSP** if and only if:

1.  $v(t)$  is a **stationary stochastic process**
2.  $W(Z)$  is **asymptotically stable**

Checking the stationary property is usually very long, instead these two properties make it easy: input  $v(t)$  is usually a **white noise** which is a **stationary stochastic process**.

### 1.2.7 Poles and Zeros in MA & AR processes

#### MA(1)

$$y(t) = e(t) + \frac{1}{2}e(t-1)$$

$$y(t) = (1 + \frac{1}{2}Z^{-1})e(t)$$

$$y(t) = (\frac{Z + \frac{1}{2}}{Z})e(t)$$

- **Zero** :  $Z = -\frac{1}{2}$
- **Pole** :  $Z = 0$

#### AR(1)

$$y(t) = \frac{1}{2}y(t-1) + 3e(t)$$

$$y(t) = (\frac{3Z}{Z - \frac{1}{2}})e(t)$$

- **Zero** :  $Z = 0$
- **Pole** :  $Z = \frac{1}{2}$

#### General conclusion

A **MA(n)** process is generated by a TF having :

- n **generic** zeros
- n poles **all in 0** → **always stationary!**

It is also called **All-Zeros** process

An **AR(m)** process is generated by a TF having :

- m zero **all in 0**
- m **generic** poles It is also called **All-Poles** process



## 2 Chapter 2 : Analysis of Stochastic Processes

TS modelled with ARMA processes and I/O modelled with ARMAX models can be represented in 4 different ways :

- Time domain (Chap.1)
- Transfer function (Chap.1)
- Probabilistic representation
- Frequency representation

### 2.1 Probabilistic Representation

#### 2.1.1 Probabilistic representation of MA(n)

Time domain representation :  $y(t) = c_0e(t) + \dots + c_n e(t - n), e(t) \sim WN(0, \lambda^2)$ .

The process is **stationary** as all the poles are in the origin.

- Mean of y

$$m_y = E[y(t)] = E[c_0e(t) + \dots + c_n e(t - n)] = c_0E[e(t)] + \dots + c_nE[e(t - n)]$$

Because of stationary property  $E[e(t)] = \dots = E[e(t - n)] = 0$

$$\boxed{m_y = 0}$$

- Covariance of y

–  $\tau = 0$

$$\begin{aligned}\gamma_y(0) &= E[(y(t) - m_y)^2] = E[y(t)^2] = E[(c_0e(t) + \dots + c_n e(t - n))^2] = \\ &= c_0^2 E[e(t)^2] + \dots + c_n^2 E[e(t - n)^2] + 2c_0c_1 E[e(t)e(t - 1)] + \dots + 2c_{n-1}c_n E[e(t - n - 1)e(t - n)]\end{aligned}$$

where

$$E[e(t)^2] = E[e(t - 1)^2] = \dots = E[e(t - n)^2] = \lambda^2$$

$E[e(t)e(t - 1)] \dots = 0$  because not correlated

$$\boxed{\gamma_y(0) = \lambda^2(c_0^2 + \dots + c_n^2)}$$

–  $\tau = 1$

$$\gamma_y(1) = E[(y(t) - m_y)(y(t-1) - m_y)] = E[y(t)y(t-1)]$$

$$E[(c_0e(t) + \dots + e_n e(t-n))(c_0e(t-1) + \dots + c_n e(t-n-1))]$$

only terms at same time survive :

$$c_0c_1E[e(t-1)^2] + \dots + c_{n-1}c_nE[e(t-n)^2]$$

$$\text{where } E[e(t-i)^2] = \lambda^2$$

$$\boxed{\gamma_y(1) = (c_0c_1 + c_1c_2 + \dots + c_{n-1}c_n)\lambda^2}$$

–  $\tau = 2$

$$\boxed{\gamma_y(2) = (c_0c_2 + c_1c_3 + \dots + c_{n-2}c_n)\lambda^2}$$

– ...

–  $\tau = n$

$$\boxed{\gamma_y(n) = c_0c_n\lambda^2}$$

–  $|\tau| > n$

$$\boxed{\gamma_y(\tau) = 0, \tau > n}$$

Which means that **MA(n)** has a **finite memory** of n steps

### 2.1.2 Probabilistic representation of AR(1)

$$y(t) = ay(t-1) + e(t), e(t) \sim WN(0, \lambda^2)$$

Is  $y(t)$  a **SSP**?

**TF representation:**

$$y(t) = aZ^{-1}y(t) + e(t) \rightarrow y(t) = \frac{1}{1-aZ^{-1}}e(t) \rightarrow y(t) = \frac{Z}{Z-a}e(t)$$

So  $y(t)$  is a SSP if and only if  $|a| < 1$

- **Mean of y**

$$m_y = E[y(t)] = E[ay(t-1) + e(t)] = am_y + m_e$$

$$m_y(1-a) = m_e \rightarrow m_y = \frac{m_e}{1-a}$$

$m_e = 0$  so :

$$\boxed{m_y = 0}$$

This hold only if the general input  $v(t)$  has  $m_v = 0$  and the system  $\mathbf{W}(\mathbf{Z})$  is **asymptotically stable**.

- **Covariance of y**

–  $\tau = 0$

$$\gamma_y(0) = E[(y(t) - m_y)^2] = E[y(t)^2] = E[(ay(t-1) + e(t))^2]$$

$$\gamma_y(0) = a^2 E[y(t-1)^2] + E[e(t)] + 2aE[e(t)y(t-1)]$$

$$\gamma_y(0) = a^2 \gamma_y(0) + \lambda^2 + 0$$

**Observation:** the fact that  $E[e(t)y(t-1)] = 0$  is explained in 2.1.3!

$$\boxed{\gamma_y(0) = \frac{\lambda^2}{1 - a^2}}$$

–  $\tau = 1$

$$\gamma_y(1) = E[(y(t) - m_y)(y(t-1) - m_y)] = E[(ay(t-1) + e(t))y(t-1)]$$

$$\gamma_y(1) = E[ay(t-1)y(t-1)] + E[e(t)y(t-1)] = a\gamma_y(0) + 0$$

**Observation:** the fact that  $E[e(t)y(t-1)] = 0$  is explained in 2.1.3!

$$\boxed{\gamma_y(1) = a\gamma_y(0)}$$

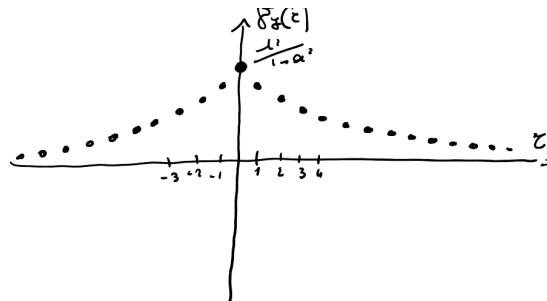
– ...

–  $\tau \neq 0$

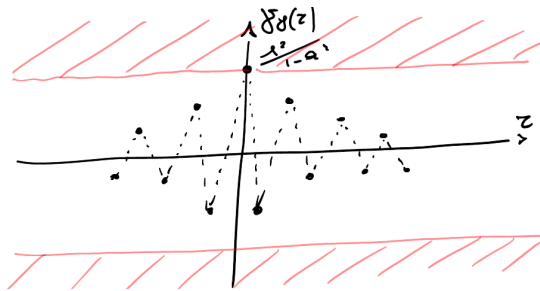
$$\boxed{a\gamma_y(\tau - 1)}$$

Which means that **AR(1)** has an **infinite memory**. The formula is also known as **Yule-Walker formula of order 1**

### 1. Plot of $\gamma_y(\tau)$ , $0 < a < 1$



### 2. Plot of $\gamma_y(\tau)$ , $-1 < a < 0$



#### 2.1.3 AR/ARMA as $\text{MA}(\infty)$

A general rule states that every AR/ARMA stationary stochastic process can be modelled as  $\text{MA}(\infty)$ . Example with  $\text{AR}(1)$  as above:

$$y(t) = \frac{1}{1-aZ^{-1}}e(t) \rightarrow y(t) = \sum_{k=0}^{\infty} (aZ^{-1})^k e(t)$$

A **geometric series of common ratio  $aZ^{-1}$**

$$y(t) = e(t)[1 + aZ^{-1} + a^2Z^{-2} + \dots]$$

$$\boxed{y(t) = e(t) + ae(t-1) + a^2e(t-2) + \dots}$$

Which is the  $\text{MA}(\infty)$  equivalent of  $\text{AR}(1)$ . This formula is very useful to demonstrate that in an  $\text{AR}(1)$   $E[e(t)y(t-1)] = 0$  by expressing  $y(t-1)$  in  $\text{MA}(\infty)$ :

$$E[e(t)(e(t-1) + ae(t-2) + a^2e(t-3) + \dots)] = E[e(t)e(t-1)] + E[e(t)ae(t-2) + \dots] = 0$$

Due to correlation all terms are equal to zero (WN property!).

## 2.2 Frequency Representation

The **power density** / **spectral density** / **spectrum** of a **SSP**  $y(t)$  :

$$\Gamma_y(w) = \sum_{\tau=-\infty}^{\infty} \gamma_y(\tau) e^{-jw\tau}$$

where  $\Gamma_y(w)$  is the **Discrete Fourier Transform**.

Properties :

1.  $\Gamma_y(w)$  is a **real** function of a **real** variable  $w$  which means that  $Im\{\Gamma_y(w)\} = 0$
2.  $\Gamma_y(w)$  is a **positive** function which means that  $\Gamma_y(w) \geq 0, \forall w \in \Re$
3.  $\Gamma_y(w)$  is an **even** function which means that  $\Gamma_y(w) = \Gamma_y(-w)$
4.  $\Gamma_y(w)$  is a **periodic** function of period  $2\pi$  which means that  $\Gamma_y(w) = \Gamma_y(w + k - 2\pi)$ .

## 2.3 Inverse Fourier Transform

Fourier Transform :

$$\Gamma_y(w) = F\{\gamma_y(\tau)\} = \sum_{t=-\infty}^{\infty} \gamma_y(\tau) e^{-jw\tau}$$

**Inverse Fourier Transform** :

$$\gamma_y(\tau) = F^{-1}\{\Gamma_y(w)\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Gamma_y(w) e^{jw\tau} dw$$

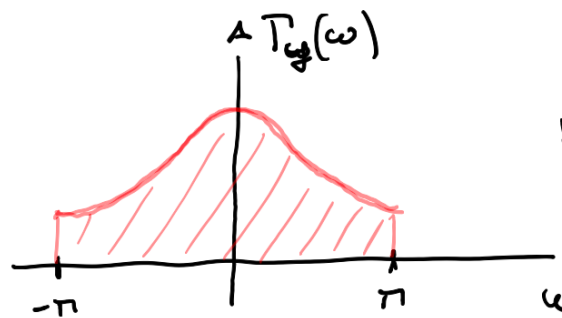
It is important to notice that  $\Gamma_y(w)$  and  $\gamma_y(\tau)$  contain the **same information** : passing from one to another does not result in **loss** or **gain** of information.

### Special IFT : Computation of variance

A special case of IFT is the computation of the variance , when  $\tau = 0$ :

$$\gamma_y(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Gamma_y(w) dw$$

which is the **area below the spectrum** between  $(-\pi, \pi)$  divided by  $2\pi$ .

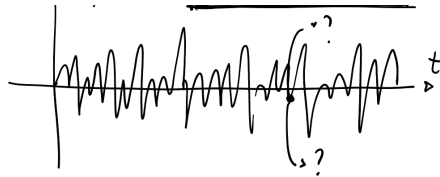


## 2.4 White Noise in the frequency domain

In case we are dealing with a WN :  $e(t) = WN(0, \lambda^2)$  we can consider it in three different domains.

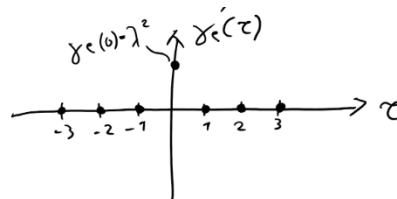
### 1. Time domain

WN is clearly **unpredictable**



### 2. Probabilistic domain

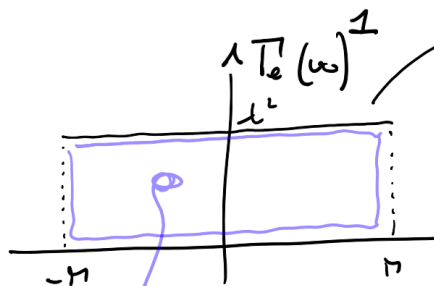
Considering the WN in the probabilistic domain and plotting its **variance** only  $\gamma_e(0) \neq 0$  : there is no **correlation** between  $e(t)$  and  $e(t \pm \tau)$



### 3. Frequency domain

Since the definition of FT relies on the definition of **covariance**  $\gamma_e(\tau)$  , as seen in point 2 only for  $\tau = 0 \rightarrow \gamma_e(\tau) \neq 0$  :

$$\Gamma_e(w) = \gamma_e(0)e^{jw0} = \gamma_e(0) = \lambda^2$$



The area is  $2\pi\lambda^2$  so the variance is  $\frac{area}{2\pi} = \lambda^2 = \gamma_e(0)$

The **energy** of the WN is **uniformly distributed** over all frequencies.

## 2.5 Computation of the spectrum of a process generated as the output of a digital system

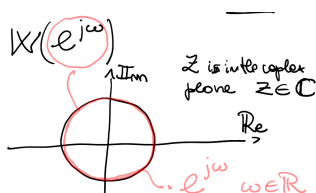
Problem : the **computation** of the  $\Gamma_y(w)$  is quite difficult most of the times. A **simpler** alternative can be found using the notion of **Frequency Response**.

### 2.5.1 Frequency Response of a linear system

Given two signals (**SSP**) input  $v(t)$  and output  $y(t)$  , where input passes through  $W(Z)$  the system or digital filter , then the **frequency response** is

$$W(e^{jw})$$

which corresponds to the evaluation of the **transfer function** on the **unit circumference**



The frequency response is used in system theory in the **Frequency Response Theorem**

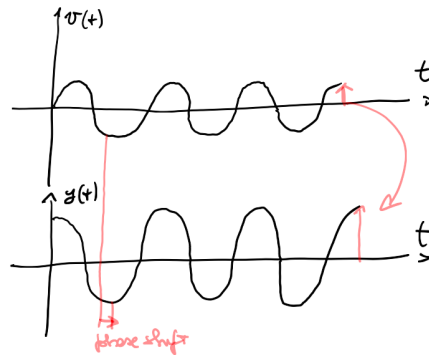
**FR Th.**

If  $W(Z)$  is **asymptotically stable** and  $v(t)$  is  $A \sin(\Omega t + \phi)$  , where  $A$  is the amplitude and  $\phi$  the phase of the sinusoid , the the output is a **pure sinusoid** with :



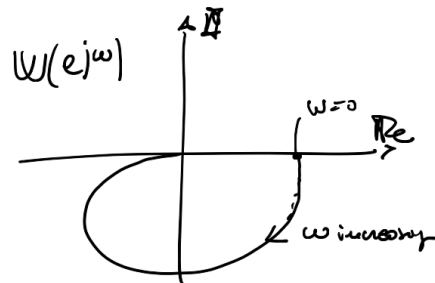
- the **same** angular speed  $\Omega$
- amplitude  $A|W(e^{j\Omega})| \rightarrow$  **gain**
- phase  $\phi + \angle W(e^{j\Omega}) \rightarrow$  **shift in phase**

$$y(t) = A|W(e^{j\Omega})| \sin(\Omega t + \phi + \angle W(e^{j\Omega}))$$



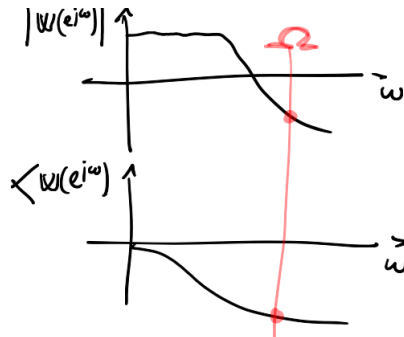
### 1.FR Nyquist plot

$W(e^{j\omega})$  is a complex function of a **real** variable.



## 2.FR Bode plot

Bode plot gives information about **magnitude** and **phase**



### 2.5.2 Spectrum computation with FR

If  $y(t)$  is output of a transfer function  $W(Z)$  which is **asymptotically stable** with input signal  $v(t)$ , then the spectrum is:

$$\Gamma_y(w) = |W(e^{jw})|^2 \Gamma_v(w)$$

The computation of  $\Gamma_v(w)$  still remains but most of the time signal  $v(t)$  is a **white noise**  $\sim (0, \lambda^2)$  which means that  $\Gamma_v(w) = \lambda^2$

## 2.6 Equivalent representations of ARMA

An ARMA SSP can be represented in 4 different but **equivalent** ways with  $e(t) \sim WN(0, 1)$ :

1. **Time domain**  $y(t) = a_1 y(t-1) + \dots + a_m y(t-m) + c_0 e(t) + \dots + c_n e(t-n)$
2. **Transfer function**  $y(t) = \frac{C(Z)}{A(Z)} e(t)$
3. **Probabilistic domain:**

$$\begin{cases} m_y &= E[y(t)] \\ \gamma_y(\tau) &= E[(y(t) - m_y)(y(t-\tau) - m_y)] \end{cases}$$

4. **Frequency domain**

$$\begin{cases} m_y &= E[y(t)] \\ \Gamma_y(w) &w \in \mathfrak{R} \end{cases}$$

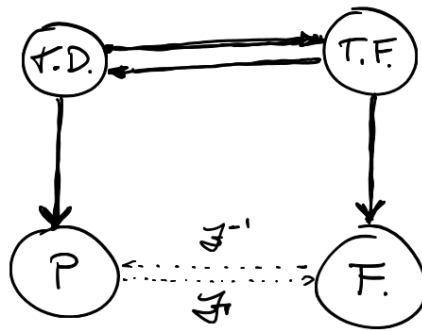
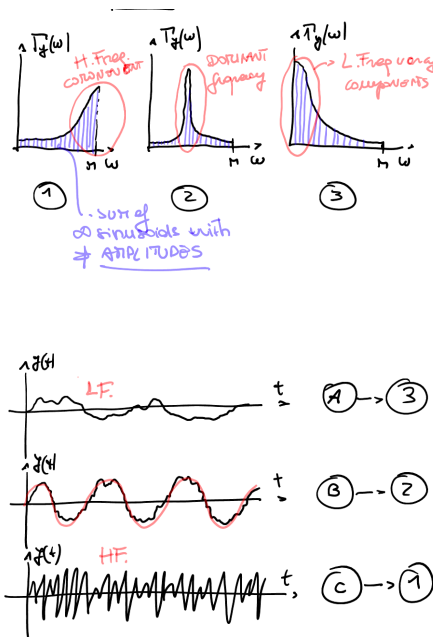


Figure 5: Bold : usual transformation , dotted : feasible but difficult

## 2.7 Example & Exercises

### Example 1

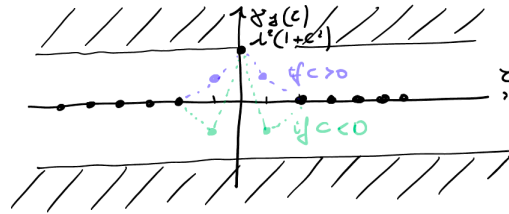
Given 3 output spectra match the corresponding time domain representation.



## Example 2

Given a **MA(1)** process  $y(t) = e(t) + ce(t-1)$ ,  $e \sim WN(0, \lambda^2)$ ,  $c \in \mathbb{R}$ .

-  $y(t)$  is **stationary** because MA(1) is always **asymptotically stable** -  
 $m_e = 0 \rightarrow m_y = 0$  -  $\gamma_y(0) = \lambda^2(1 + c^2)$  -  $\gamma_y(1) = \lambda^2 c$  -  $\gamma_y(\tau) = 0, |\tau| \geq 2$



1. Composition of  $\Gamma_y(w)$  with  $\lambda^2 = 1$ :

- **From definition**

$$\Gamma_y(w) = \sum_{\tau=-\infty}^{\infty} \gamma_y(\tau) e^{-jw\tau}$$

- For  $\tau = 0$  :  $(1 + c^2)$
- For  $\tau = 1$  :  $ce^{-jw}$
- For  $\tau = -1$  :  $c + e^{+jw}$
- For  $|\tau| \geq 2$  : 0

$$\Gamma_y(w) = 1 + c^2 + c(e^{-jw} + e^{+jw})$$

Recall :  $e^{-jw} + e^{jw} = \cos w - j \sin w + \cos w + j \sin w = 2 \cos w$

$$\Gamma_y(w) = (1 + c^2) + 2c \cos w$$

Which is **real, positive, even, periodic**

- **From frequency response**

The MA(1) transfer function is

$$y(t) = (1 + cZ^{-1})e(t)$$

$$\Gamma_y(w) = |W(e^{jw})|^2 \Gamma_e(w) = |1 + ce^{-jw}|^2 \cdot 1$$

Recall :  $|a + jb|^2 = \text{Im}[a + ib]^2 + \text{Re}[a + ib]^2 = a^2 + b^2 = (a + jb)(a - jb)$

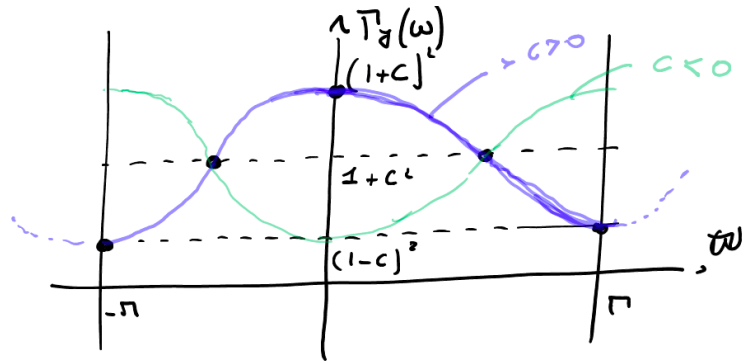
$$(1 + ce^{-jw})(1 - ce^{jw}) = 1 + c^2(e^{jw} \cdot e^{-jw}) + c(e^{-jw} + e^{jw}) = 1 + c^2 + 2c \cos w$$

2. Plotting of  $\Gamma_y(w)$  :

$$\Gamma_y(0) = (1 + c)^2$$

$$\Gamma_y\left(\frac{\pi}{2}\right) = 1 + c^2$$

$$\Gamma_y(\pi) = (1 - c)^2$$



3. Compute the variance  $\gamma_y(0)$  given  $\Gamma_y(w)$  :

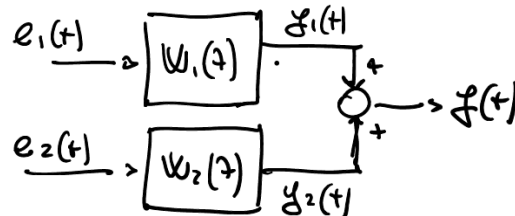
$$\gamma_y(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Gamma_y(w) dw = \frac{1}{2\pi} \int_{-\pi}^{\pi} (1 + c^2 + 2ccosw) dw$$

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} (1 + c^2) dw + \frac{1}{2\pi} \int_{-\pi}^{\pi} 2ccosw dw$$

$$\frac{1}{2\pi} [(1 + c^2)[w]_{-\pi}^{\pi} + 2c[senw]_{-\pi}^{\pi}] = \frac{1}{2\pi} [(1 + c^2)(2\pi)] =$$

$$1 + c^2$$

### Example 3



Consider the SSP  $y(t)$  generated by 2 inputs.

$W_1(t), W_2(t)$  are asymptotically stable.

$e_1(t) \sim W(0, \lambda_1^2), e_2(t) \sim W(0, \lambda_2^2)$

$e_1 \perp e_2 \rightarrow E[e_1(t)e_2(t - \tau)] = 0$

Calculate  $\gamma_y(\tau)$  and  $\Gamma_y(w)$

- $\gamma_y(\tau)$

$$\begin{aligned} \gamma_y(\tau) &= E[y(t)y(t - \tau)] = E[(y_1(t) + y_2(t))(y_1(t - \tau) + y_2(t - \tau))] \\ &= E[y_1(t)y_1(t - \tau)] + E[y_2(t)y_2(t - \tau)] + E[y_1(t)y_2(t - \tau)] + E[y_2(t)y_1(t - \tau)] \\ &= \gamma_{y_1}(\tau) + \gamma_{y_2}(\tau) + 0 + 0 \end{aligned}$$

Term 3 and 4 are  $= 0$  which is a result obtained by rewriting them as  $MA(\infty)$  and exploiting the hypothesis that  $e_1(t) \perp e_2(t)$ .

$$\boxed{\gamma_y(t) = \gamma_{y_1}(t) + \gamma_{y_2}(t)}$$

- $\Gamma_y(t)$

$$\Gamma_y(t) = \sum_{\tau=-\infty}^{\infty} \gamma_y(\tau)e^{-jw\tau} = \sum_{\tau=-\infty}^{\infty} \gamma_{y_1}(\tau)e^{-jw\tau} + \sum_{\tau=-\infty}^{\infty} \gamma_{y_2}(\tau)e^{-jw\tau}$$

$$\boxed{\Gamma_y(w) = \Gamma_{y_1}(w) + \Gamma_{y_2}(w)}$$

The result can be generalised to more than 2 inputs that are summed to form an SSP  $y(t)$ :

$$\boxed{\gamma_y(t) = \gamma_{y_1}(t) + \gamma_{y_2}(t) + \dots + \gamma_{y_k}(t)}$$

$$\boxed{\Gamma_y(w) = \Gamma_{y_1}(w) + \Gamma_{y_2}(w) + \dots + \Gamma_{y_k}(w)}$$

The result hold if all  $W_i(t)$  are asymptotically stable , all  $v_i(t)$  are ssp and uncorrelated

#### Example 4

Consider the following AR(1) SSP  $y(t) = \frac{1}{3}y(t-1) + e(t) + 2 \rightarrow e \sim WN(1, 1)$  which has a asymptotically stable TF.

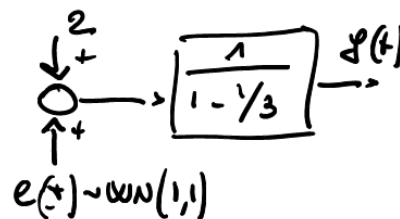
Calculate  $m_y$  and  $\gamma_y$ .

- Mean of y

- Method 1

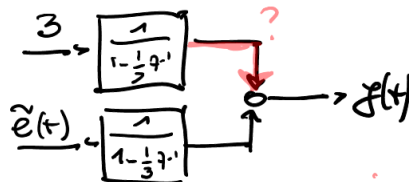
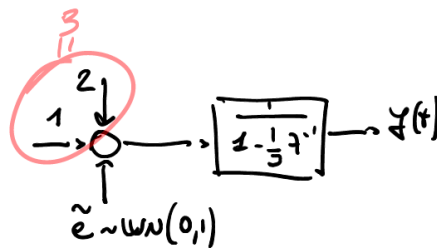
$$E[y(t)] = E[\frac{1}{3}y(t-1)e(t) + 2] \rightarrow (1 - \frac{1}{3})m_y = m_e + 2 \rightarrow m_y = \frac{9}{2}$$

- Method 2



$$e(t) = \tilde{e} + 1, \tilde{e} \sim WN(0, 1)$$

Using the superposition principle of LTI systems :



The constant value 3 can be seen as a **sinusoidal signal with  $w=0$**  so the **Frequency Response Theorem** can be applied:

$$3\left(\frac{1}{1-\frac{1}{3}Z^{-1}}\right) \text{ calculated in } z = e^{j0}$$

$$3\left(\frac{1}{1-\frac{1}{3}}\right) = \frac{9}{2} \text{ And since } m_{\tilde{e}} \text{ is a zero mean signal } \rightarrow m_y = \frac{9}{2}$$

- **Covariance of y**

- **Method 1 : BAD**

$$E[(y(t) - \frac{9}{2})^2] = E[(\frac{1}{3}y(t-1) + e(t) + 2 - \frac{9}{2})^2]$$

$$\gamma_y(0) = \frac{1}{9}E[y(t-1)^2] + E[e(t)^2] + \frac{25}{4} + \frac{2}{3}E[y(t-1)e(t)] - \frac{5}{3}E[y(t-1)] + 5E[e(t)]$$

---

**Remark:**  $E[(e(t) - m_e)^2] = \gamma_e(0) = E[e(t)^2] - 2E[e(t)m_e] + m_e^2$   
 $E[e(t)] = \gamma_e(0) + m_e^2$

Which can be generalised :

$$E[e(t)^2] = \gamma_e(0) + m_e^2$$

$$E[y(t)^2] = \gamma_y(0) + m_y^2$$

$$E[e(t)y(t-1)] = E[(e(t) - m_e)(y(t-1) - m_y)] + m_y m_e$$

$$E[e(t)y(t-1)] = m_e m_y$$

As the **de-biased signals are incorellated!**

---


$$\gamma_y(0) = \frac{1}{9}(\gamma_y(0) + m_y^2) + (\gamma_e(0) + m_e^2) + \frac{25}{4} + \frac{2}{3}(m_e m_y) - \frac{5}{3}m_y - 5m_e = \frac{9}{8}$$

Same computations for  $\gamma_y(1), \gamma_y(2)...$

- **Method 2: GOOD**

Define two new processes:

$$\tilde{y}(t) = y(t) - \frac{9}{2} \rightarrow m_{\tilde{y}=0}$$

$$\tilde{e}(t) = e(t) - 1 \rightarrow m_{\tilde{e}=0}$$

So  $y(t) = \tilde{y}(t) + \frac{9}{2}$  and  $e(t) = \tilde{e}(t) + 1$  :

$$\tilde{y}(t) + \frac{9}{2} = \frac{1}{3}(\tilde{y}(t-1) + \frac{9}{2}) + (\tilde{e} + 1) + 2$$

$$\tilde{y}(t) = \frac{1}{3}\tilde{y}(t-1) + \tilde{e}(t), \tilde{e} \sim WN(0, 1)$$



Where  $\tilde{y}(t)$  is the **de-biased process**

$$\gamma_{\tilde{y}(0)} = \frac{1}{1-\frac{1}{9}} = \frac{9}{8}$$

$$\gamma_{\tilde{y}(1)} = \frac{9}{8} \frac{1}{3} = \frac{3}{8}$$

$$\gamma_{\tilde{y}(2)} = \frac{3}{8} \frac{1}{3} = \frac{1}{8} \dots$$

Now that we found  $\gamma_{\tilde{y}(\tau)}$  we want to find  $\gamma_y(\tau)$   $\gamma_y(\tau)$ :

$$E[(y(t) - \frac{9}{2})(y(t - \tau) - \frac{9}{2})] = E[\tilde{y}(t)\tilde{y}(t - \tau)] = \gamma_{\tilde{y}}(\tau)$$

since  $m_{\tilde{y}} = 0$  Which can be generalised :

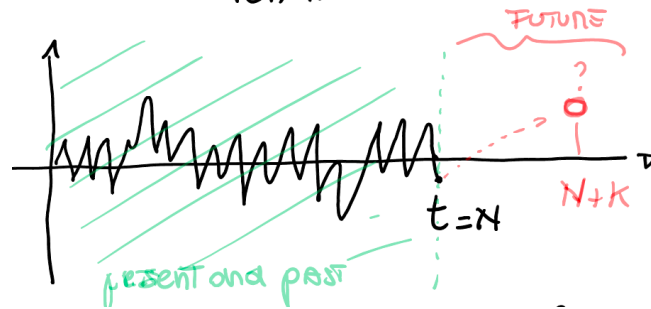
If  $y(t)$  and  $\tilde{y}(t)$  are two SSPs that **differ only from a constant value**  $y(t) = \tilde{y}(t) + k$  then :

$$\boxed{\gamma_y(\tau) = \gamma_{\tilde{y}}, \forall \tau}$$

$$\boxed{\Gamma_y(w) = \Gamma_{\tilde{y}}, \forall w}$$

### 3 Chapter 3 : Prediction

The prediction problem is to find the **best possible value** for  $\hat{y}(t+k|t)$  given the **measured data** up to time  $t$   $\{y(1), \dots, y(N)\}$



To obtain the **optimal** prediction :

1. We have to make a mathematical model for  $\{y(1), \dots, y(N)\}$
2. Using the model compute the optimal solution

To find the **best** mathematical model :

1. We select a class of models for time-series  $y(t) = W(z, \theta)e(t)$  where  $e(t)$  is a WN and  $\theta$  a parameter vector.
2. We compute the prediction of  $y(t)$  using the mathematical model :

$$\hat{y}(t+1|t; \theta)$$

3.  $\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left[ \frac{1}{N} \sum_{t=1}^N (y(t) - \hat{y}(t|t-1; \theta))^2 \right]$
4. Find  $y(t) = W(Z, \hat{\theta})e(t)$  which is the best model from prediction performance  
. Use this to compute  $\hat{y}(N+K|N)$

To create a **predictor** from an ARMA/ARMAX we need to define 2 tools:

- All-pass filter
- Canonical representation

### 3.1 All-Pass Filter

An All-Pass Filter is a **first-order, linear, digital** filter with a special **constrained** structure:

$$T(Z) = \frac{1}{a} \frac{Z + a}{Z + \frac{1}{a}}, a \in \mathbb{R}$$

that depends on only one parameter and has a **pole** in  $z = -\frac{1}{a}$  and zero in  $z = -a$   
Properties :

- **Magnitude**

$$|T(e^{jw})|^2 = \left| \frac{1}{a} \frac{e^{jw} + a}{e^{jw} + \frac{1}{a}} \right|^2 = \frac{1}{a^2} \left( \frac{e^{jw} + a}{e^{jw} + \frac{1}{a}} \right) \cdot \frac{1}{a} \left( \frac{e^{-jw} + a}{e^{-jw} + \frac{1}{a}} \right) = \frac{1}{a^2} \frac{1 + a^2 + 2a \cos w}{1 + \frac{1}{a^2} + \frac{2 \cos w}{a}} = 1$$

An all-pass filter is characterized by a **frequency response** having **unitary magnitude** :

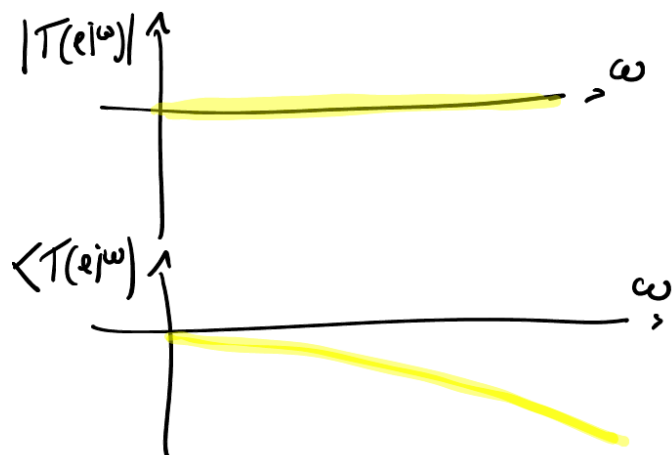
$$\Gamma_y(w) = |T(e^{jw})|^2 \Gamma_v(w)$$

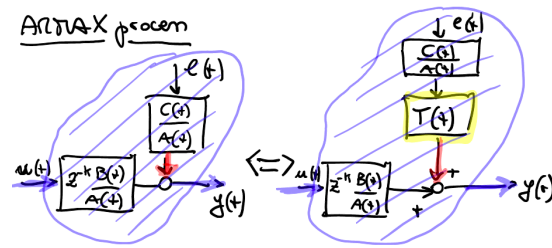
An all-pass filter does not alter the **spectrum** of its input  $v(t)$ . This does **not** mean that  $y(t) = v(t)$  but that they're **statistically equivalent** :

$$\Gamma_y(w) = \Gamma_v(w)$$

$$\gamma_y(w) = \gamma_v(w)$$

Input and output are **not** identical because although there is no change in amplitude, an all-pass filter makes a **distortion in phase**.





The two representations of the ARMAX process are **equivalent**. The phase distortion added to signal  $e(t)$  is **not relevant**. On the other hand, adding a  $T(Z)$  all-pass filter between  $u(t)$  and  $y(t)$  alters the behaviour of the system **critically!!**

### 3.2 Canonical Representation

An ARMA process can have  $\infty$  **equivalent** representations ( there is no way to represent it in a unique way).

There is a special representation called **Canonical Representation**:

Given a SSP  $y(t)$  that can be modellded as an ARMA process:

$$y(t) = \frac{C(Z)}{A(Z)}e(t)$$

,

$$\frac{C(Z)}{A(Z)}$$

is the **canonical representation** if:

1.  $C(Z)$  and  $A(Z)$  have **same degree** ( relative degree is 0)
2.  $C(Z)$  and  $A(Z)$  are **coprime** ( no common factors)
3.  $C(Z)$  and  $A(Z)$  are **monic** ( coefficient of max degree of both  $C(Z)$  and  $A(Z)$  is 1)
4. All roots of  $C(Z)$  and  $A(Z)$  are **strictly inside** the unit circle

#### Example

$$y(t) = \frac{1+3Z^{-1}}{2+Z^{-1}}e(t-2), e(t) \sim WN(0,1)$$

- **Type and order**

ARMA type process of order 1,3 = ARMA(1,3)

- **Canonical form**

$$y(t) = \frac{Z^{-2} + 3Z^{-3}}{2 + Z^{-1}} e(t)$$

1. Degree of C(Z) is 2 , degree of A(Z) is 0  $\rightarrow$  X
2. No common factors  $\rightarrow$  OK
3. C(Z) is monic , A(Z) is not monic  $\rightarrow$  X
4. Zero in -3  $\rightarrow$  X

By collecting and using an **All-Pass Filter**:

$$y(t) = \frac{Z^{-2}(1 + 3Z^{-1})}{2(1 + \frac{1}{2}Z^{-1})} \cdot 3 \frac{1 + \frac{1}{3}Z^{-1}}{1 + 3Z^{-1}} e(t)$$

$$y(t) = \frac{z^{-2} (1+3z^{-1})}{2(1+\frac{1}{2}z^{-1})} \cdot 3 \frac{1+\frac{1}{3}z^{-1}}{1+3z^{-1}} e(t)$$

$T(z)$

Defining  $\theta = \frac{3}{2}e(t-2)$ ,  $\theta \sim ?$

The variance of  $\theta = E[\theta(t)^2] = E[(\frac{3}{2}e(t-2))^2] = \frac{9}{4} \cdot 1$

$\theta \sim WN(0, \frac{9}{4})$ ?  $\rightarrow \Gamma_{\theta}(w) = |\frac{3}{2}e^{-2jw}|^2 \cdot 1 = \frac{9}{4}$  which is constant value so

$\theta \sim WN(0, \frac{9}{4})$  Finally we obtain :

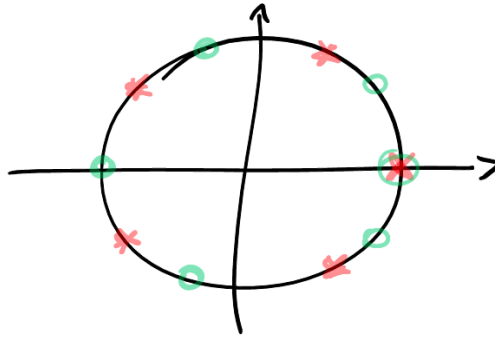
$$y(t) = \frac{1 + \frac{1}{3}Z^{-1}}{1 + \frac{1}{2}Z^{-1}} \theta(t), \theta \sim WN(0, \frac{9}{4})$$

Which is an **ARMA(1,1)**  $\rightarrow$  the **canonical representation** is the representation with **minimum order**!

### Remark

Does the canonical representation **always** exist?

Not always : the problem lies in the 4<sup>th</sup> condition . It is possible that the system has poles or zeros **on** the unitary circle.



If  $C(Z)$  has **zeros on the u.c**  $\rightarrow$  prediction from data is **not asymptotically stable**

If  $A(Z)$  has **roots in +1**  $\rightarrow$  **ARIMA models**:

$$y(t) = \frac{C(Z)}{(Z-1)^d A(Z)} e(t) \rightarrow \text{ARIMA}(m, d, n)$$

[Autoregressive Integrated Moving average ]

A special case of ARIMA  $\rightarrow$  **ARIMA(0,1,0)**:

$$y(t) = \frac{1}{1 - Z^{-1}} e(t)$$

$$y(t) = y(t-1) + e(t), e(t) \sim WN(0, \lambda^2)$$

Process  $y(t)$  is a **Random Walk** that uses as TF  $\frac{1}{1-Z^{-1}}$  which is an **integrator in discrete time**

ARIMA processes have an **asymptotically stable** predictor that can be used to model **not strictly stationary processes!!**

### 3.3 Predictor

The predictor at time  $t+k$  , given the data up to time  $t$  is:

$$\hat{y}(t+k|t)$$

The **prediction error** is:

$$\epsilon(t+k) = y(t+k) - \hat{y}(t+k|t)$$

So the **real value** is predictor + error:

$$y(t+k|t) = \hat{y}(t+k|t) + \epsilon(t+k)$$

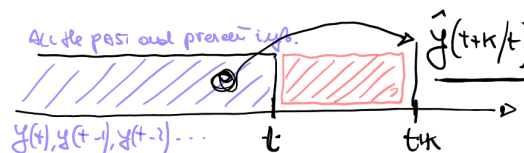
$$y(t) = \hat{y}(t|t-k) + \epsilon(t)$$

The formulas are equivalent because as per hypothesis  $y(t)$  is **stationary**.

### 3.3.1 Optimality

The predictor  $\hat{y}(t+k|t)$  is **optimal** if:

1.  $E[\hat{y}(t+k|t) \cdot \epsilon(t+k)] = 0$ , predictor and error must be **uncorrelated**
2.  $E[y(t) \cdot \epsilon(t+k)] = E[y(t-1) \cdot \epsilon(t+k)] \dots = 0$



The red part shows the **unpredictable** part of  $y(t+k)$ , which is the error  $\epsilon(t+k)$ . If error and predictor were **correlated** then some useful unused information about  $\hat{y}(t+k|t)$  would be in  $\epsilon(t+k)$  which means that the predictor is not optimal. The same goes for  $y(t), y(t-1) \dots$  of point 2): the error cannot contain information about the past/present information.

### 3.3.2 1-step ahead prediction of MA(n)

$$y(t) = e(t) + c_1 e(t-1) + \dots + c_n e(t-n), e(t) \sim WN(0, \lambda^2)$$

We assume the the MA(n) is represented in the **canonical representation**: we must make assumptions about the 4<sup>th</sup> property.

Given :

- **Present time** :  $t-1 \rightarrow c_1 e(t-1) + \dots + c_n e(t-n)$
- **Future** :  $t \rightarrow e(t)$

## Predictor from noise

The **optimal predictor** from **noise** is :

$$\hat{y}(t|t-1) = c_1 e(t-1) + \dots + c_n e(t-n)$$

with error

$$\epsilon(t) = y(t) - \hat{y}(t|t-1) = e(t)$$

Optimality :

- $E[\hat{y}(t|t-1)\epsilon(t)] = E[(c_1 e(t-1) + \dots + c_n e(t-n))(e(t))] = 0$
- $E[y(t-1)\epsilon(t)] = E[(e(t-1) + c_1 e(t-2) + \dots + c_n e(t-n-1))(e(t))] = \dots = 0$

Verified because of incorrelation of white noise.

Since WN is **unknown** and cannot be measured , a better predictor has to be chosen from **measurable data**.

## Predictor from data

TF:

$$y(t) = (1 + c_1 Z^{-1} + \dots + c_n Z^{-n})e(t)$$

Inverse TF (**Whitening Filter**):

$$e(t) = \frac{1}{1 + c_1 Z^{-1} + \dots + c_n Z^{-n}} y(t)$$

$$\hat{y}(t|t-1) = (1 + c_1 Z^{-1} + \dots + c_n Z^{-n})e(t) \rightarrow \hat{y}(t|t-1) = \frac{c_1 Z^{-1} + \dots + c_n Z^{-n}}{1 + c_1 Z^{-1} + \dots + c_n Z^{-n}} y(t)$$

Collecting  $Z^{-1}$  :

$$\hat{y}(t|t-1) = \frac{c_1 + \dots + c_n Z^{-n+1}}{1 + c_1 Z^{-1} + \dots + c_n Z^{-n}} y(t-1)$$

$$\hat{y}(t|t-1) = \underbrace{-c_1 \hat{y}(t-1|t-1) - c_2 \hat{y}(t-2|t-1) \dots - c_n \hat{y}(t-n|t-1)}_{\text{PAST PREDICTIONS}} + \underbrace{+ c_1 y(t-1) + c_2 y(t-2) + \dots + c_n y(t-n)}_{\text{PRESENT AND PAST MEASUREMENT}}$$

As seen in time-domain representation the prediction makes use of **present and past** data as well as **past predictions**.



### 3.3.3 K-steps ahead predictor of MA(n)

$$y(t) = e(t) + c_1 e(t-1) + \dots + c_{k-1} e(t-k+1) + c_k e(t-k) + \dots + c_n e(t-n)$$

Given:

-**Present time:**  $k \rightarrow c_k e(t-k) + \dots + c_n e(t-n)$

-**Future :**  $t \rightarrow e(t) + \dots + c_{k-1} e(t-k+1)$

**Predictor from noise**

$$\hat{y}(t|t-k) = c_k e(t-k) + \dots + c_n e(t-n)$$

with error:

$$\epsilon(t) = e(t) + \dots + c_{k-1} e(t-k+1)$$

**Predictor from data**

$$\hat{y}(t|t-k) = \frac{c_k + c_{k+1}Z^{-1} + \dots + c_n Z^{-n+k}}{1 + c_1 Z^{-1} + \dots + c_n Z^{-n}} y(t-k)$$

### 3.3.4 K-steps ahead predictor of general ARMA(m,n)

$$y(t) = \frac{C(Z)}{A(Z)} e(t), e(t) \sim WN(0, \lambda^2)$$

(Under the hypothesis of **canonical representation**)

The AR(m) part presents a recursion : need to introduce k-steps **polynomial division** between C(Z) and A(Z) obtaining :

- **E(Z)**  $\rightarrow$  result (quotient)

- **R(Z)**  $\rightarrow$  residual (remainder)

$$\begin{array}{r}
 \textcircled{1} + \frac{1}{2}z^{-1} \quad \textcircled{C(4)} \quad \textcircled{1 + \frac{1}{3}z^{-1}} \quad \textcircled{A(7)} \\
 - 1 - \frac{1}{3}z^{-1} \\
 \hline
 // \quad \textcircled{\frac{1}{6}z^{-1}} \\
 - \frac{1}{6}z^{-1} - \frac{1}{18}z^{-2} \\
 \hline
 // \quad \textcircled{-\frac{1}{18}z^{-2}} \quad \rightarrow \textcircled{R(3)} \\
 \quad \quad \quad \rightarrow \tilde{R}(z)
 \end{array}$$

$$C(Z) = E(Z)A(Z) + R(Z)$$

$$\boxed{\frac{C(Z)}{A(Z)} = E(Z) + \frac{R(Z)}{A(Z)}}$$

Noting that in k-steps division  $R(Z)$  can be rewritten by collecting  $Z^{-k}$ :

$$R(Z) = Z^{-k} \tilde{R}(Z)$$

$$\boxed{\frac{C(Z)}{A(Z)} = E(Z) + \frac{Z^{-k} \tilde{R}(Z)}{A(Z)}}$$

The new transfer function is:

$$y(t) = [E(Z) + \frac{Z^{-k} \tilde{R}(Z)}{A(Z)}]e(t)$$

$$y(t) = E(Z)e(t) + \frac{\tilde{R}(Z)}{A(Z)}e(t-k)$$

Where  $E(Z)e(t)$  is the **unpredictable part** as it depends on  $e(t), \dots, e(t-k+1)$

**Predictor from noise**

$$\boxed{\hat{y}(t|t-k) = \frac{\tilde{R}(Z)}{A(Z)}e(t-k)}$$

with error:

$$\boxed{\epsilon(t) = E(Z)e(t)}$$

### Predictor from data

$$y(t) = \frac{C(Z)}{A(Z)}e(t) \xrightarrow{\text{Whitening}} e(t) = \frac{A(Z)}{C(Z)}y(t)$$

$$\hat{y}(t|t+k) = \frac{\tilde{R}(Z)Z^{-k}}{A(Z)} \cdot \frac{A(Z)}{C(Z)}y(t)$$

$$\boxed{\hat{y}(t|t-k) = \frac{\tilde{R}(Z)}{C(Z)}y(t-k)}$$

### Remark 1

Both the predictor from noise and data work under the assumption of SSP . The stationary property is satisfied if both  $A(Z)$  and  $C(Z)$  have all roots (poles) strictly inside the unitary circle. But this is satisfied by the 4<sup>th</sup> condition of the canonical representation hypothesis.

### Remark 2

$\epsilon(t) = y(t) - \hat{y}(t|t-k) = E(Z)e(t)$  where  $E(Z)$  is a SSP of type **MA(k-1)**

### Remark 3

In the case of **K=1** the polynomial division result in :

-**E(Z)** = 1 as both  $C(Z), A(Z)$  are monic and have same degree

-**R(Z)** =  $C(Z)-A(Z)$

which results in

$$\boxed{\hat{y}(t|t-k) = \frac{C(Z) - A(Z)}{A(Z)}e(t)}$$

$$\boxed{\epsilon(t) = e(t)}$$

Instead of having term  $R(Z)$  , the formula is now  $C(Z)-A(Z)$ . As  $R(Z) = \tilde{R}(Z)Z^{-1}$  there is a hidden  $Z^{-1}$  in  $C(Z)-A(Z)$ .

### 3.3.5 K-steps ahead prediction of ARMAX(m,n,k+p)

$$y(t) = \frac{B(Z)}{A(Z)}u(t-k) + \frac{C(Z)}{A(Z)}e(t), e(t) \sim WN(0, \lambda^2)$$

Where:

$$A(Z) = 1 + a_1Z^{-1} + \dots + a_mZ^{-m}$$

$$B(Z) = b_0 + b_1Z^{-1} + \dots + b_pZ^{-p}$$

$$C(Z) = 1 + c_1Z^{-1} + \dots + c_nZ^{-n}$$

In the hypothesis that  $\frac{C(Z)}{A(Z)}$  is in **canonical representation** and keeping in mind that for  $\frac{B(Z)}{A(Z)}u(t-k)$  no **spectral equivalence** modifications can be made.

In an ARMAX(m,n,k+p) process the most interesting prediction that can be made is the **delay** between  $u(t)$  and  $y(t) \rightarrow k$  so we'll deal only with k-steps predictions.

#### Predictor from noise

Separate predictable from unpredictable part in  $\frac{C(Z)}{A(Z)}e(t)$

K-steps division  $\frac{C(Z)}{A(Z)} \rightarrow E(Z) + \frac{\tilde{R}(Z)}{A(Z)}$

$$y(t) = \frac{B(Z)}{A(Z)}u(t-k) + E(Z)e(t) + \frac{\tilde{R}(Z)}{A(Z)}e(t-k)$$

where

$$\frac{B(Z)}{A(Z)}u(t-k) \rightarrow \text{depends on } u(t-k), \dots, u(t-k-p) \rightarrow \text{predictable}$$

$$E(Z)e(t) \rightarrow \text{depends on } e(t), e(t-1), \dots, e(t-k+1) \rightarrow \text{unpredictable}$$

$$\frac{\tilde{R}(Z)}{A(Z)}e(t-k) \rightarrow \text{depends on } e(t-k), \dots, e(t-k-p) \rightarrow \text{predictable}$$

so

$$\hat{y}(t|t-k) = \frac{B(Z)}{A(Z)}u(t-k) + \frac{\tilde{R}(Z)}{A(Z)}e(t-k)$$

$$\epsilon(t) = y(t) - \hat{y}(t|t-k) = E(Z)e(t)$$

Which is optimal if

- $\epsilon(t) \perp \hat{y}(t|t-k)$
- $\epsilon(t) \perp y(t-k), y(t-k-1) \dots$

## Predictor from data

$$\begin{aligned}
 e(t) &= \frac{A(Z)}{C(Z)}y(t) - \frac{B(Z)}{A(Z)}u(t-k) \\
 \hat{y}(t|t-k) &= \frac{B(Z)}{A(Z)}u(t-k) + \frac{R(Z)}{A(Z)}\left[\frac{A(Z)}{C(Z)}y(t) - \frac{B(Z)}{A(Z)}u(t-k)\right] \\
 \hat{y}(t|t-k) &= \frac{R(Z)}{C(Z)}y(t) + \left[\frac{B(Z)}{A(Z)} - \frac{R(Z)B(Z)}{A(Z)C(Z)}\right]u(t-k) \\
 \hat{y}(t|t-k) &= \frac{B(Z)}{C(Z)}y(t) + \left[\frac{B(Z)(C(Z) - R(Z))}{A(Z)C(Z)}\right]u(t-k)
 \end{aligned}$$

Knowing that  $C(Z) = A(Z)E(Z) + R(Z) \rightarrow C(Z) - R(Z) = A(Z)E(Z)$

$$\boxed{\hat{y}(t|t-k) = \frac{B(Z)E(Z)}{C(Z)}u(t-k) + \frac{\tilde{R}(Z)}{C(Z)}y(t-k)}$$

$$\boxed{\epsilon = E(Z)e(t)}$$

Note that  $\frac{\tilde{R}(Z)}{C(Z)}y(t-k)$  is the exact ARMA predictor.

The prediction error is the same as in the **ARMA** process: the **exogenous** part does not add any **additional uncertainty**.

**Remark : Special case k=1**

$$y(t) = \frac{B(Z)}{A(Z)}u(t-1) + \frac{C(Z)}{A(Z)}e(t)$$

$$E(Z) = 1 \text{ and } R(Z) = C(Z) - A(Z)$$

$$\hat{y}(t|t-1) = \frac{B(Z)}{C(Z)}u(t-1) + \frac{C(Z) - A(Z)}{C(Z)}y(t)$$

## 3.4 Examples & Exercises

### 3.4.1 Example 1

Given a process

$$y(t) = \frac{Z+3}{2Z+1}e(t-1), e(t) \sim WN(0,1)$$

Since the pole of the TF is  $z = -\frac{1}{2}$  inside the unitary circle,  $W(Z)$  is asymptotically stable  $\rightarrow y(t)$  is **stationary**.

## 1. Compute $\gamma_y(0)$

NB.: To calculate the variance it is not important for the system to be in canonical representation

$y(t) = \frac{C(Z)}{A(Z)}e(t-1)$  is **not canonical** since it has

- $Z = -3$  not inside unitary circle
- $2Z$  in the  $A(Z)$  term
- $Z^{-1}e(t)$

Using an **All-Pass Filter**:

$$y(t) = \frac{Z+3}{2(Z+\frac{1}{2})}Z^{-1} \cdot 3\frac{Z+\frac{1}{3}}{Z+3}e(t)$$

$$\eta = \frac{3}{2}Z^{-1}e(t) \sim WN(0, \frac{9}{4})$$

$$y(t) = \frac{Z+\frac{1}{3}}{Z+\frac{1}{2}}\eta(t)$$

Passing in time domain :

$$y(t) = -\frac{1}{2}y(t-1) + \eta(t) + \frac{1}{3}\eta(t-1)$$

$$-m_y = E[y(t)] = -\frac{1}{2}E[y(t-1)] + \frac{4}{3}m_e \rightarrow 0$$

$$-\gamma_y(0) = E[y(t)^2] = E[(\frac{1}{2}y(t-1) + \eta(t) + \frac{1}{3}\eta(t-1))^2]$$

$$\gamma_y(0) = \frac{1}{4}\gamma_y(0) + \frac{9}{4} + \frac{1}{9}\frac{9}{4} - \frac{1}{3}E[y(t-1)\eta(t-1)]$$

$$\frac{3}{4}\gamma_y(0) = \frac{10}{4} - \frac{1}{3}E[(\frac{1}{2}y(t-2) + \eta(t-1) + \frac{1}{3}\eta(t-2))\eta(t-1)]$$

$$\frac{3}{4}\gamma_y(0) = \frac{10}{4} - \frac{1}{3}E[\eta(t-1)^2] \rightarrow \frac{3}{4}\gamma_y(0) = \frac{10}{4} - \frac{1}{3}\frac{9}{4}$$

$$\gamma_y(0) = \frac{7}{3}$$

## 2. Prediction for k=1

Using the canonical negative power representation

$$y(t) = \frac{1 + \frac{1}{3}Z^{-1}}{1 + \frac{1}{2}Z^{-1}}\eta(t)$$

Applying the k=1 prediction formula  $\hat{y}(t|t-1) = \frac{C(Z)-A(Z)}{C(Z)}y(t)$  :

$$\hat{y}(t|t-1) = \frac{1 + \frac{1}{3}Z^{-1} - 1 - \frac{1}{2}Z^{-1}}{1 + \frac{1}{3}Z^{-1}}y(t)$$

$$\hat{y}(t|t-1) = \frac{-\frac{1}{6}}{1 + \frac{1}{3}Z^{-1}}y(t-1)$$

In time domain :

$$\hat{y}(t|t-1) = -\frac{1}{3}\hat{y}(t-1|t-2) - \frac{1}{6}y(t-1)$$

$$\epsilon(t) = y(t) - \hat{y}(t|t-1) = E(Z)\eta(t) = \eta(t)$$

$$var[y(t) - \hat{y}(t|t-1)] = var[\eta(t)] = \frac{9}{4}$$

### 3. Prediction for k=2

$$\begin{array}{r|l}
 1 + \frac{1}{3}Z^{-1} & 1 + \frac{1}{2}Z^{-1} \\
 -1 - \frac{1}{2}Z^{-1} & 1 - \frac{1}{6}Z^{-1} \\
 \hline
 // -\frac{1}{6}Z^{-1} & \downarrow E(Z) \\
 +\frac{1}{6}Z^{-1} + \frac{1}{12}Z^{-2} & R(Z) \\
 \hline
 // \frac{1}{12}Z^{-2} & \tilde{R}(Z)
 \end{array}$$

$$\hat{y}(t|t-2) = \frac{R(Z)}{C(Z)}y(t) = \frac{\tilde{R}(Z)}{C(Z)}y(t-2)$$

$$\hat{y}(t|t-2) = \frac{\frac{1}{12}}{1 + \frac{1}{3}Z^{-1}}y(t-2)$$

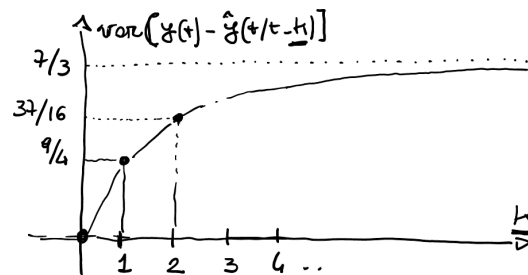
$$\hat{y}(t|t-2) = -\frac{1}{3}\hat{y}(t-1|t-3) + \frac{1}{12}y(t-2)$$

$$\epsilon(t) = y(t) - \hat{y}(t|t-2) = E(Z)\eta(t) = (1 - \frac{1}{6}Z^{-1})\eta(t)$$

$$\text{var}[y(t) - \hat{y}(t|t-2)] = \text{var}[(1 - \frac{1}{6}Z^{-1})\eta(t)] = \frac{37}{16}$$



#### 4. Properties of $\text{var}[\epsilon(t)]$ as function of $k$



- $k = 0 \rightarrow \text{var}[y(t) - \hat{y}(t|t-k)] = 0$
- $k = 1 \rightarrow \text{var}[y(t) - \hat{y}(t|t-k)] = \lambda^2$
- $k \rightarrow \infty \rightarrow \text{var}[y(t) - \hat{y}(t|t-k)] = \gamma_y(0)$  because when  $k \rightarrow \infty$  the prediction goes to zero!
- $\text{var}[y(t) - \hat{y}(t|t-k)]$  is a **monotonic (not strictly) increasing** function

#### 5. Prediction goodness

The **Error to signal ratio** is a useful prediction measure:

$$ESR(k) = \frac{\text{var}[y(t) - \hat{y}(t|t-k)]}{\text{var}[y(t)]}$$

For  $k=1$

$$ESR(1) = \frac{\frac{9}{4}}{\frac{7}{3}} = 0.97$$

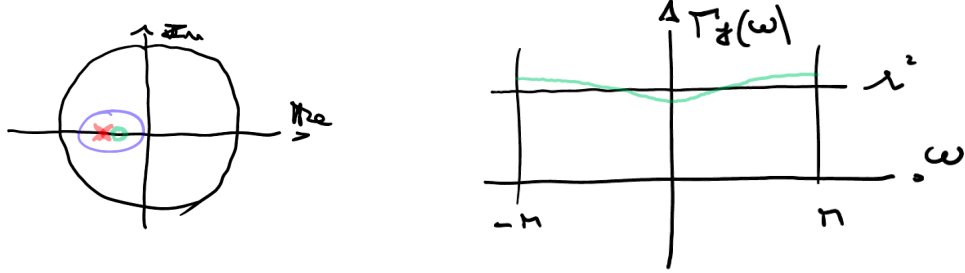
Which is a very bad prediction. The most trivial prediction that can be done is

$$\hat{y}(t|t-k) = m_y$$

(predicting the mean) which has **ESR(k)=1**.

For  $k=1$  we only have a 3% better prediction than the trivial one.

The predictor for  $k=1$  is **optimal** which means that **no better** prediction can be made : the bad prediction is an intrinsic property of the process  $y(t)$ .



By analysing the poles and zeros, it is easy to see that they're so close together that they almost cancel each other out.

The **spectrum**  $\Gamma_y(w)$  in green is very close to that of the **white noise**: this is the reason  $y(t)$  is hard to predict

### 3.4.2 Example 2 - Practical

We have measured 5 data points of a signal:

$$y(1) = 1, y(2) = \frac{1}{2}, y(3) = -\frac{1}{2}, y(4) = 0, y(5) = -\frac{1}{2}$$

With  $t=5$  represent the present time, make a prediction of  $\hat{y}(6|5)$ . To solve the problem we must make a mathematical modelling assumption. Since we're still not able to do this we need some interpretations models for this signal

Model A

$$y(t) = \frac{1}{2}y(t-1) + \frac{1}{4}y(t-2) + e(t), e(t) \sim WN(0, \lambda^2)$$

Model B

$$y(t) = e(t) + \frac{1}{2}e(t-1), e(t) \sim WN(0, \lambda^2)$$

To determine which model is better we compute the **optimal** model assuming the chosen model is right.

- **Assuming Model A right**

$$y(t) = \frac{1}{1 - \frac{1}{2}Z^{-1} - \frac{1}{4}Z^{-2}}e(t)$$

is an AR(2) process in canonical representation (check it always!).

Since we're dealing with a  $k=1$  prediction :

$$\hat{y}(t|t-1) = \frac{C(Z) - A(Z)}{C(Z)}y(t)$$

$$\hat{y}(t|t-1) = \frac{1 - 1 + \frac{1}{2}Z^{-1} + \frac{1}{4}Z^{-2}}{1}y(t)$$

$$\hat{y}(t|t-1) = \frac{1}{2}y(t-1) + \frac{1}{4}y(t-2)$$

Substituting the data points :

$$\hat{y}(6|5) = \frac{1}{2}y(5) + \frac{1}{4}y(4) = -\frac{1}{4}$$

- **Assuming Model B right**

$$y(t) = (1 + \frac{1}{2}Z^{-1})e(t)$$

is an MA(1) process in canonical representation.

Since we're dealing with a k=1 prediction :

$$\hat{y}(t|t-1) = \frac{C(Z) - A(Z)}{C(Z)}y(t)$$

$$\hat{y}(t|t-1) = \frac{1 + \frac{1}{2}Z^{-1} - 1}{1 + \frac{1}{2}Z^{-1}}y(t)$$

$$\hat{y}(t|t-1) = -\frac{1}{2}\hat{y}(t-1|t-2) + \frac{1}{2}y(t-1)$$

Substituting the data points :

$$\hat{y}(6|5) = -\frac{1}{2}\hat{y}(5|4) + \frac{1}{2}y(5) = -\frac{1}{2}\hat{y}(5|4) - \frac{1}{4}$$

To compute  $-\frac{1}{2}\hat{y}(5|4)$  we need to go back to the **initial condition** to compute all terms up to time =5:

$$-\hat{y}(2|1) = -\frac{1}{2}\hat{y}(1|0) + \frac{1}{2}y(1) \text{ by making the assumption that } -\frac{1}{2}\hat{y}(1|0) = m_y \rightarrow \frac{1}{2}$$

$$-\hat{y}(3|2) = -\frac{1}{2}\hat{y}(2|1) + \frac{1}{2}y(2) = 0$$

$$-\hat{y}(4|3) = -\frac{1}{2}\hat{y}(3|2) + \frac{1}{2}y(3) = -\frac{1}{4}$$

$$-\hat{y}(5|4) = -\frac{1}{2}\hat{y}(4|3) + \frac{1}{2}y(4) = \frac{1}{8}$$

$$-\hat{y}(6|5) = -\frac{1}{2}\hat{y}(5|4) + \frac{1}{2}y(5) = -\frac{5}{16}$$

Our final prediction for model B is  $\hat{y}(6|5) = -\frac{5}{16}$  which depends on the initial condition made assuming that  $-\frac{1}{2}\hat{y}(1|0) = m_y$ . Is the choice of the initial condition important? **If the system is asymptotically stable, and N is big the initial condition is not important as it will vanish.**

### 3.4.3 Example 3 - ARMAX & ARX

$$y(t) = (Z + 6Z^{-1})u(t-2) + \frac{2}{3 + \frac{3}{2}Z^{-1}}\eta(t-1), \eta \sim WN(0, 1)$$

Find predictor from data and the corresponding error with its variance.

$u(t-2) \rightarrow k=2$

Canonical form for ARMA part

$$\frac{2}{3(1 + \frac{1}{2}Z^{-1})}Z^{-1}\eta(t)$$

So

$$e(t) = \frac{2}{3}\eta(t-1), e(t) \sim WN(0, \frac{4}{9})$$

$$\frac{1}{1 + \frac{1}{2}Z^{-1}}e(t)$$

Substituting in the original process:

$$y(t) = (Z + 6Z^{-1})u(t-2) + \frac{1}{1 + \frac{1}{2}Z^{-1}}e(t), e(t) \sim WN(0, \frac{4}{9})$$

Is the term  $(Z + 6Z^{-1})u(t-2)$  also in canonical representation? Wrong question, there is nothing we can do about it!

We need the form :

$$y(t) = \frac{B(Z)}{A(Z)}u(t-k) + \frac{C(Z)}{A(Z)}e(t)$$

So rewriting :

$$y(t) = \frac{(2 + 6Z^{-1})(1 + \frac{1}{2}Z^{-1})}{(1 + \frac{1}{2}Z^{-1})}u(t-2) + \frac{1}{1 + \frac{1}{2}Z^{-1}}e(t)$$

Using a k-steps long division  $\frac{C(Z)}{A(Z)}$ :

$$\begin{array}{r} 2 - \frac{1}{2}Z^{-1} \\ (1 + \frac{1}{2}Z^{-1}) \overline{) 2 + 6Z^{-1}} \\ \underline{2 + 1Z^{-1}} \phantom{00} \\ -1Z^{-1} \phantom{00} \\ \underline{+ \frac{1}{2}Z^{-1}} \phantom{00} \\ -\frac{1}{2}Z^{-1} \phantom{00} \\ \underline{+ \frac{1}{4}Z^{-1}} \phantom{00} \\ -\frac{1}{4}Z^{-1} \phantom{00} \\ \underline{+ \frac{1}{8}Z^{-1}} \phantom{00} \\ -\frac{1}{8}Z^{-1} \phantom{00} \\ \vdots \end{array}$$

$\text{Quotient: } 2 - \frac{1}{2}Z^{-1} \quad \text{Remainder: } \frac{1}{2}Z^{-2}$   
 $Q(Z) = 2 - \frac{1}{2}Z^{-1}$   
 $R(Z) = \frac{1}{2}Z^{-2}$

$$\hat{y}(t|t-2) = \frac{(2 + 6Z^{-1})(1 + \frac{1}{2}Z^{-1})(1 - \frac{1}{2}Z^{-1})}{1}u(t-2) + \frac{\frac{1}{4}Z^{-2}}{1}y(t)$$

$$\boxed{\hat{y}(t|t-2) = 2u(t-2) + 6u(t-3) - \frac{1}{2}u(t-4) - \frac{3}{2}u(t-5) + \frac{1}{4}y(t-2)}$$

**No old prediction** is used  $\rightarrow$  process is **ARMAX(1,0,2+2)**  $\rightarrow$  **ARX(1,4)** model.

$$\boxed{\epsilon = E(Z)e(t) = (1 - \frac{1}{2})Z^{-1}e(t)}$$

The variance of  $\epsilon$  :

$$var[\epsilon(t)] = (1 + \frac{1}{4}) \cdot \frac{4}{9} = \frac{5}{4} \cdot \frac{4}{9} = \frac{5}{9}$$

### 3.4.4 Example ARMA with non-zero mean

$$y(t) = e(t) + 4e(t-1), e \sim WN(1, 1)$$

Compute  $\hat{y}(t|t-1)$  and  $\hat{y}(t|t-2)$  from data.

Canonical form representation

$$y(t) = (1 + 4Z^{-1})e(t) \rightarrow y(t) = (1 + 4Z^{-1})[4 \cdot \frac{1 + \frac{1}{4}Z^{-1}}{1 + 4Z^{-1}}]e(t)$$

Getting the new  $\eta(t)$ :

$$\eta(t) = 4e(t)$$

$$m_\eta = E[\eta(t)] = E[4e(t)] = 4$$

$$var[\eta] = E[(\eta(t) - 4)^2] = E[(4e(t) - 4)^2] = 16E[(e(t) - 1)^2] = 16$$

Canonical form :

$$\boxed{y(t) = (1 + \frac{1}{4}Z^{-1})\eta(t)}$$

$$\boxed{\eta(t) \sim WN(4, 16)}$$

### Method 1

**De-biasing technique:**

$$\tilde{y}(t) = y(t) - m_y$$

$$\tilde{\eta}(t) = \eta(t) - m_\eta$$

Mean of  $y$ :

$$E[y(t)] = E[(\eta(t) + \frac{1}{4}\eta(t-1))] \rightarrow m_y = \frac{5}{4}m_\eta = 5$$

So:

$$\tilde{y}(t) = y(t) - 5$$

$$\tilde{\eta}(t) = \eta(t) - 4 \rightarrow \tilde{\eta} \sim WN(0, 16)$$

Obtaining:

$$\tilde{y} + 5 = (\tilde{\eta}(t) + 4) + \frac{1}{4}(\tilde{\eta}(t-1) + 4)$$

$$\boxed{\tilde{y} = \tilde{\eta}(t) + \frac{1}{4}\tilde{\eta}(t-1)}$$

Now we can compute the predictions for  $\tilde{y}$  for  $k=1$ :

$$\hat{\tilde{y}}(t|t-1) = \frac{(1 + \frac{1}{4}Z^{-1}) - 1}{(1 + \frac{1}{4}Z^{-1})} \tilde{y}(t)$$

$$\boxed{\hat{\tilde{y}}(t|t-1) = \frac{\frac{1}{4}}{(1 + \frac{1}{4}Z^{-1})} \tilde{y}(t-1)}$$

$$\boxed{\epsilon(t) = \tilde{\eta}(t) = 16}$$

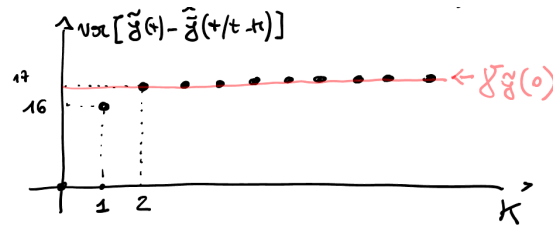
Now we can compute the prediction for  $\tilde{y}$  for  $k=2$ :

$$\begin{array}{r|l} 1 + \frac{1}{4}Z^{-1} & 1 \\ -1 & \hline \hline \frac{1}{4}Z^{-1} & 1 + \frac{1}{4}Z^{-1} \quad \equiv (+) \\ -\frac{1}{4}Z^{-1} & \\ \hline \hline \emptyset & \emptyset \quad \equiv (-) \end{array}$$

Which means that

$$\boxed{\hat{\tilde{y}}(t|t-2) = 0}$$

Because MA(1) process has a **finite memory** of 1-step only! So  $\tilde{\epsilon}(t) = \tilde{y}(t) - \hat{\tilde{y}}(t|t-2) = \tilde{y}(t)$  so the  $var[\tilde{\epsilon}(t)] = var[\tilde{y}(t)] = (1 + \frac{1}{16}) \cdot 16 = 17$



We need to go back to the original process because  $\hat{y}(t|t-1) \neq \hat{\hat{y}}(t|t-1)$  :  
- for **k=1**

$$\hat{y}(t|t-1) - 5 = \frac{\frac{1}{4}}{1 + \frac{1}{4}Z^{-1}}(y(t-1) - 5)$$

$$\hat{y}(t|t-1) = \frac{\frac{1}{4}}{1 + \frac{1}{4}Z^{-1}}y(t-1) + 5 - 5\frac{\frac{1}{4}}{1 + \frac{1}{4}Z^{-1}}$$

The term  $5\frac{\frac{1}{4}}{1 + \frac{1}{4}Z^{-1}}$  can be resolved by applying the **frequency response theorem** by taking in account that 5 is a sinusoid with frequency  $w = 0$  :

$$\frac{\frac{1}{4}}{1 + \frac{1}{4}e^{0j}} \cdot 5 = 1$$

so :

$$\hat{y}(t|t-1) = \frac{\frac{1}{4}}{1 + \frac{1}{4}Z^{-1}}y(t-1) + 4$$

- for **k=2**

$$\hat{\hat{y}}(t|t-2) = 0 \rightarrow \hat{y}(t|t-2) - 5 = 0$$

$$\hat{y}(t|t-2) = 5$$

$$var[\tilde{\eta}(t)] = var[\eta(t)]$$

## Method 2

De-bias technique only on  $\eta(t)$ :

$$\tilde{\eta}(t) = \eta(t) - 4$$

$$y(t) = \tilde{\eta}(t) + 4 + \frac{1}{4}(\tilde{\eta}(t) + 4)$$

$$y(t) = \tilde{\eta}(t) + \frac{1}{4}\tilde{\eta}(t-1) + 5$$

Which can be considered as an **ARMAX** process.

**-for k=1**

$$y(t) = u(t-1) + (1 + \frac{1}{4}Z^{-1})\tilde{\eta}(t), u(t) = 5\forall t$$

$u(t-1)$  is chosen arbitrarily because we need to compute  $\hat{y}(t|t-1)$ :

- k =1

-  $B(Z) = 1$

-  $C(Z) = 1 + \frac{1}{4}Z^{-1}$

-  $A(Z) = 1$

$$\hat{y}(t|t-1) = \frac{1 \cdot 1}{1 + \frac{1}{4}Z^{-1}}u(t-1) + \frac{(1 + \frac{1}{4}Z^{-1}) - 1}{1 + \frac{1}{4}Z^{-1}}y(t)$$

As  $u(t-1) = 5$  the system taking it as input can be simplified again using FR. Theorem :

$$\hat{y}(t|t-1) = \frac{\frac{1}{4}}{1 + \frac{1}{4}Z^{-1}}y(t-1) + 4$$

Which is the same result as for the first method.

**-for k=2**

$$y(t) = u(t-2) + (1 + \frac{1}{4}Z^{-1})\tilde{\eta}(t), u(t) = 5\forall t$$

Making a 2 step long division  $\frac{C(Z)}{A(Z)}$ :

-  $E(Z) = 1 + \frac{1}{4}Z^{-1}$  -  $R(Z) = 0$

$$\hat{y}(t|t-2) = \frac{1 \cdot (1 + \frac{1}{4}Z^{-1})}{1 + \frac{1}{4}Z^{-1}}u(t-2) + 0$$

$$\hat{y}(t|t-2) = 5$$

Which again is the same result as in the first method.



## 4 Chapter 4 : Identification

The focus of MIDA 1 are **parametric** identification or learning techniques. They are the most used and popular identification techniques but many non-parametric techniques are essential for identification ( ex: state-space identification ,spectrum estimation ,unsupervised learning...)

Any **parametric identification technique** is based on a **five step approach**:

1. **Experiment design & data collection**

This step deals with **designing** the experiment, selecting the **length N** of the dataset and **data pre-processing**.

2. **Selection of a class of parametric models**

This steps deals with the selection of **class** of parametric models  $m(\theta)$  where  $\theta$  is the unknown parameter vector. Our focus will be on :

-**discrete time**

-**dynamic**

-**linear**

-**time-invariant**

systems. As already seen **ARMAX & ARMA** are the most general class of models for these systems.

3. **Selection of a performance index**

A function  $J(\theta) \geq 0$  that tells the **ordering** of different models. The performance index assesses the **quality** of a model.

The **prediction error method** is the choice for our performance index:

$$J_N(\theta) = \frac{1}{N} \sum_{t=1}^N (y(t) - \hat{y}(t|t-1, \theta))^2$$

that represents the **sample variance** of the prediction error computed on the available dataset of length N.

The P.E.M assumes that the ability of a model to make a good prediction of the future is a good **quality index** for the model.

#### 4. Optimization

Optimization consists in **minimizing**  $J(\theta)$  with respect to  $\theta$  :

$$\hat{\theta} = \operatorname{argmin}_{\theta} \{J(\theta)\}$$

so that

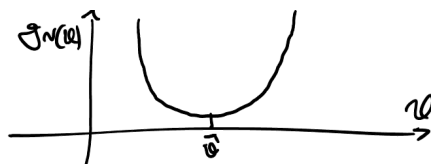
$$m(\hat{\theta})$$

is the **optimal model** on the class of models  $m(\theta)$ .

$$J_N(\theta) = R^{n_{\theta}} \rightarrow R^+$$

In optimisation 3 different situations can be found:

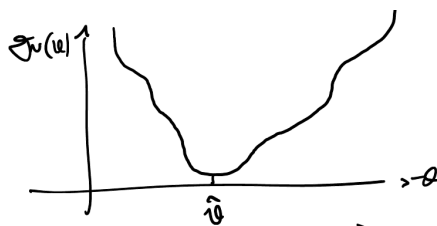
- $J(\theta)$  is **quadratic** function of  $\theta$



$J_N$  is a **quadratic** function of  $\theta$  : in this case it's usually easy to find the **global minimum** explicitly.

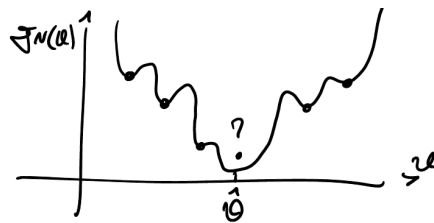
**AR & ARX** models are of this kind.

- $J(\theta)$  is **not** a quadratic function , **no local minima**



In this case the function has no local minima so the **unique solution** is guaranteed to be found using an **iterative algorithm**.

- $J(\theta)$  is **not** quadratic, **with local minima**



In this case the function has local minima so using an **iterative algorithm** is the best way to find the unique solution which is **not guaranteed** to be found.

**ARMAX & ARMA** models are of this kind.

## 5. Validation

The validation step checks if  $m(\hat{\theta})$  can be considered a **valid** model for our purposes. Usually a technique called **cross-validation** is used.

## 4.1 Identification of ARX models

Given an available dataset of length  $N$  :

$$\{u(1), u(2), \dots, u(N)\}$$

$$\{y(1), y(2), \dots, y(N)\}$$

An the model class **ARX(m,p+1)**:

$$y(t) = \frac{B(Z)}{A(Z)}u(t-1) + \frac{1}{A(Z)}e(t), e(t) \sim WN(0, \lambda^2)$$

where  $\theta = [a_1 \dots a_m b_0 \dots b_p]^T$  is the **parameter vector** of dimension  $n_\theta = m + p + 1$ .

### Remark

Using **k=1** is not a restriction but the **most general** case of an ARX. If the system has  $k > 1$  we will find out during the identification process.

#### 4.1.1 Loss function: Least Squares

The loss function for the ARX models is the **P.E.M.**:

$$J_N(\theta) = \frac{1}{N} \sum_{t=1}^N (y(t) - \hat{y}(t|t-1, \theta))^2$$

The predictor for the model , deriving it from the general ARMAX model, is :

$$\hat{y}(t|t-1; \theta) = \frac{B(Z)}{u}(t-1) + (1 - A(Z))y(t)$$

$$\hat{y}(t|t-1; \theta) = b_0 u(t-1) + \dots + b_p u(t-p-1) - a_1 y(t-1) - \dots - a_m y(t-m)$$

where we can define the **data vector**:

$$\phi = [-y(t-1), \dots - y(t-m), u(t-1), \dots u(t-p-1)]^T$$

so  $\hat{y}(t|t-1) = \phi(t)^T \theta$  , a linear function of  $\theta$ . Substituting in the loss function:

$$J_N(\theta) = \frac{1}{N} \sum_{t=1}^N (y(t) - \phi(t)^T \theta)^2$$

A **quadratic** function is obtained so the **unique solution** can be found explicitly using a minimization method. To find the minimum we differentiate wrt to the parameter vector  $\theta$  :

$$\begin{aligned}\frac{\partial J_N(\theta)}{\partial \theta} &= 0 \\ \frac{\partial J_N(\theta)}{\partial \theta} &= \frac{2}{N} \sum_{t=1}^N \phi(t)(y(t) - \phi(t)^T \theta) \\ \left( \sum_{t=1}^N \phi(t)\phi(t)^T \right) \theta &= \sum_{t=1}^N y(t)\phi(t)\end{aligned}$$

Assuming that the  $n_\theta \times n_\theta$  matrix  $\sum_{t=1}^N \phi(t)\phi(t)^T$  matrix is **non singular** and thus **invertible**:

$$\hat{\theta}_N = \left( \sum_{t=1}^N \phi(t)\phi(t)^T \right)^{-1} \left( \sum_{t=1}^N y(t)\phi(t) \right)$$

This is the **explicit** solution of the ARX identification problem also known as **Least Squares**

#### 4.1.2 Example

Consider a dataset of length  $N=10$  and

$$y(t) = \frac{b}{1 + aZ^{-1}}u(t-1) + \frac{1}{1 + aZ^{-1}}e(t), e(t) \sim WN(0, \lambda^2)$$

an ARX(1,1) general model class. Assuming that the process is in canonical representation (  $|a| < 1$  must hold). The predictor of the model is :

$$\hat{y}(t|t-1) = \frac{B(Z)}{1}u(t-1) + \frac{1 - A(Z)}{1}e(t)$$

$$\hat{y}(t|t-1) = bu(t-1) - ay(t-1)$$

and  $\theta = [a, b]^T$ .

#### Method 1

The loss function is :

$$J_{10}(\theta) = \frac{1}{10} \sum_{t=1}^{10} (y(t) - bu(t-1) + ay(t-1))^2$$

**Remark**

Since we don't have data points for  $t=0$  , starting at time  $t=1$  doesn't allow us to compute  $-bu(0) + ay(0)$  so a modified version of the performance index is used:

$$J_N(\theta) = \frac{1}{N-h} \sum_{t=h+1}^N (y(t) - \hat{y}(t|t-1))^2$$

where  $h = \max\{m, p+1\}$

In our case  $h = \max(1, 1) = 1$  so

$$J_{10}(\theta) = \frac{1}{9} \sum_{t=2}^{10} (y(t) - bu(t-1) + ay(t-1))^2$$

To obtain the best parameter vector  $\rightarrow \frac{\partial J_{10}(\theta)}{\partial \theta} = 0$  where  $\theta = [a, b]^T$  :

$$\frac{\partial J_{10}(\theta)}{\partial \theta} = \begin{cases} \frac{\partial J_{10}(\theta)}{\partial a} = \frac{2}{9} \sum_{t=2}^{10} (y(t) - bu(t-1) + ay(t-1)) \cdot y(t-1) = 0 \\ \frac{\partial J_{10}(\theta)}{\partial b} = \frac{2}{9} \sum_{t=2}^{10} (y(t) - bu(t-1) + ay(t-1)) \cdot (-u(t-1)) = 0 \end{cases}$$

Which can be rewritten as :

$$\begin{bmatrix} \sum_{t=2}^{10} y(t-1)^2 & -\sum_{t=2}^{10} y(t-1)u(t-1) \\ -\sum_{t=2}^{10} y(t-1)u(t-1) & \sum_{t=2}^{10} u(t-1)^2 \end{bmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -\sum_{t=2}^{10} y(t-1)y(t) \\ \sum_{t=2}^{10} y(t)u(t-1) \end{bmatrix}$$

so

$$\begin{bmatrix} \hat{a}_{10} \\ \hat{b}_{10} \end{bmatrix} = \begin{bmatrix} \sum_{t=2}^{10} y(t-1)^2 & -\sum_{t=2}^{10} y(t-1)u(t-1) \\ -\sum_{t=2}^{10} y(t-1)u(t-1) & \sum_{t=2}^{10} u(t-1)^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_{t=2}^{10} y(t-1)y(t) \\ \sum_{t=2}^{10} y(t)u(t-1) \end{bmatrix}$$

## Method 2

Consider the predictor  $\hat{y}(t|t-1) = bu(t-1) + ay(t-1)$  and the available data. Assuming that the predictor makes the **perfect prediction** on the measured data :

$$\begin{cases} -ay(1) + bu(1) = y(2) \\ -ay(2) + bu(2) = y(3) \\ \dots \\ -ay(9) + bu(9) = y(10) \end{cases}$$

Which can be separated into two matrices:

$$\Phi = \begin{bmatrix} -y(1) & u(1) \\ -y(2) & u(2) \\ \dots & \dots \\ -y(9) & u(9) \end{bmatrix} \quad Y = \begin{bmatrix} y(2) \\ y(3) \\ \dots \\ y(10) \end{bmatrix}$$

Obtaining a **linear** system of 9 equations and 2 unknowns :

$$\boxed{\Phi \cdot \theta = Y}$$

---

## Remark

- **Undetermined linear system**

Number of unknowns  $\neq$  number of equations  $\rightarrow$  **infinite solutions**

- **Square linear system**

Number of unknowns = number of equations  $\rightarrow$  **1! solution**

- **Over determined linear system**

Number of unknowns  $<$  number of equations  $\rightarrow$  **No solutions!**

---

Unfortunately our case is the last one with no solutions! In this case a **least squares (approximate)** solution can be found using the square matrix :

$$\Phi\theta = Y \rightarrow \Phi^T\Phi\theta = \Phi^TY$$

$$\boxed{\hat{\theta} = [\Phi^T\Phi]^{-1}\Phi^TY}$$

where  $[\Phi^T\Phi]^{-1}\Phi^T = \Phi^+$  is the **pseudo inverse** of  $\Phi$ . By creating the matrices above we obtain the **same** result for  $\hat{\theta}$

### 4.1.3 Example 2

Assume a dataset of 5 points collected from an SSP with zero mean.  $y(1) = \frac{1}{2}, y(2) = 0, y(3) = -1, y(4) = -\frac{1}{2}, y(5) = \frac{1}{4}$  and we want to make a prediction  $\hat{y}(6|5)$

#### 1. Build the model

We selected a class of model **AR(1)** (reason discussed further on ).

$$y(t) = ay(t-1) + e(t)e(t) \sim WN(0, \lambda^2)$$

is our  $m(\theta)$ . Transforming  $\rightarrow y(t) = \frac{1}{1-aZ^{-1}}e(t)$  which is in canonical representation if  $|a| < 1$ . The corresponding k=1 predictor is

$$\hat{y}(t|t-1) = \frac{C(Z) - A(Z)}{C(Z)}y(t) \rightarrow \hat{y}(t|t-1) = ay(t-1)$$

with **performance index**:

$$J_5(a) = \frac{1}{4} \sum_{t=2}^5 (y(t) - ay(t-1))^2$$

After some easy computation we find that  $J_5(a) = \frac{1}{4}[\frac{3}{2}a^2 - \frac{3}{4}a + \frac{21}{16}]$  which is the measure to be **minimized**:

$$\frac{\partial J_5(a)}{\partial a} = \frac{1}{4}[3a - \frac{3}{4}] = 0$$

Which has solution

$$\hat{a}_5 = \frac{1}{4}$$

So best model identified in AR(1) class is :

$$\boxed{y(t) = \frac{1}{1 - \frac{1}{4}Z^{-1}}e(t)}$$

#### 2. Compute prediction

$$\hat{y}(6|5) = \frac{1}{4}y(5) = \frac{1}{16}$$

#### 3. First remark

Is  $\hat{a} = \frac{1}{4}$  fine? Yes , it hold the condition that  $|a| < 1$ .

What if  $\hat{a} = 4$ ? In that case we have to find the **canonical form** of the system.



#### 4. Second remark

The variance of the white noise  $\lambda^2$  is not required for the model and prediction estimation. If we wanted to estimate also  $\lambda^2$  and not only  $a$  (which still is the most important parameter to estimate):

$$\lambda^2 = \text{var}[e(t)] \rightarrow \text{var}[\epsilon(t)]$$

Using an **approximate estimation** using the 5 available data points :

$$\hat{\lambda}_5^2 = J_5(\hat{a}_5)$$

## 4.2 Identification of ARMAX models

While ARX model estimation relies on *least squared method* , ARMAX model estimation relies on **maximum likelihood**.

Given the available N-long data vector  $\mathbf{u}$  and  $\mathbf{y}$  , the model class is  $m(\theta) : y(t) = \frac{B(Z)}{A(Z)}u(t-1) + \frac{C(Z)}{A(Z)}e(t)$  where

$$A(Z) = 1 + \dots + a_m Z^{-m}$$

$$B(Z) = b_0 + \dots + b_p Z^{-p}$$

$$C(Z) = 1 + \dots + c_n Z^{-n}$$

and working under the hypothesis that  $\frac{C(Z)}{A(Z)}$  is in **canonical representation**  $\rightarrow$  **ARMAX(m,n,p+1)** where k=1 is **not** a restriction but indeed the **most general case**.

Using the P.E.M. approach :

$$\hat{\theta} = \operatorname{argmin}_{\theta} = \{J_N(\theta)\}$$

with performance index

$$J_N(\theta) = \frac{1}{N} \sum_{t=1}^N (y(t) - \hat{y}(t|t-1, \theta))^2$$

where the parameter vector has dimension  $n_{\theta} = m + n + p + 1$ :

$$\theta = [a_1, \dots, a_m, b_0, \dots, b_p, c_1, \dots, c_m]^T$$

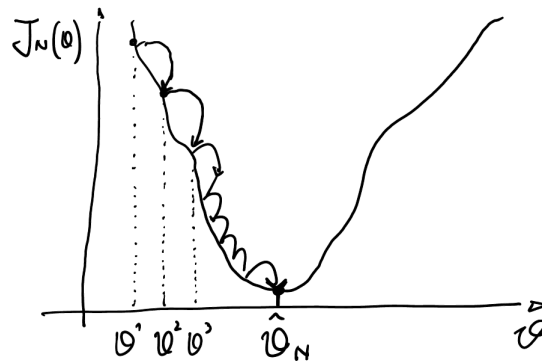
Since k=1  $\epsilon(t) = e(t) = \frac{A(Z)}{C(Z)}y(t) - \frac{B(Z)}{C(Z)}u(t-1)$  so the performance index

$$J_N(\theta) = \frac{1}{N} \sum_{t=1}^N \left( \frac{A(Z)}{C(Z)}y(t) - \frac{B(Z)}{C(Z)}u(t-1) \right)^2$$

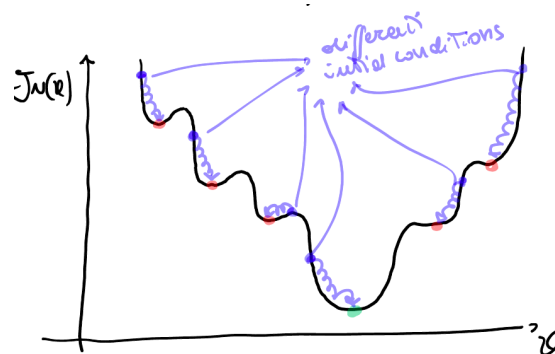
This performance index is now an issue as it kills the optimisation :  $C(Z)$  is a **non-linear** function of  $\theta$  so it is not a quadratic function of  $\theta$ . So the **minimization** requires an **iterative approach**.

### 4.2.1 Loss function :Maximum Likelihood Method

The **iterative** loss function of ARMAX models starts from an initial condition  $\theta^1 \rightarrow \theta^2 \dots \rightarrow \hat{\theta}_N$  until a final solution is reached



Unfortunately usually  $J_N(\theta)$  has lots of **local minima**:

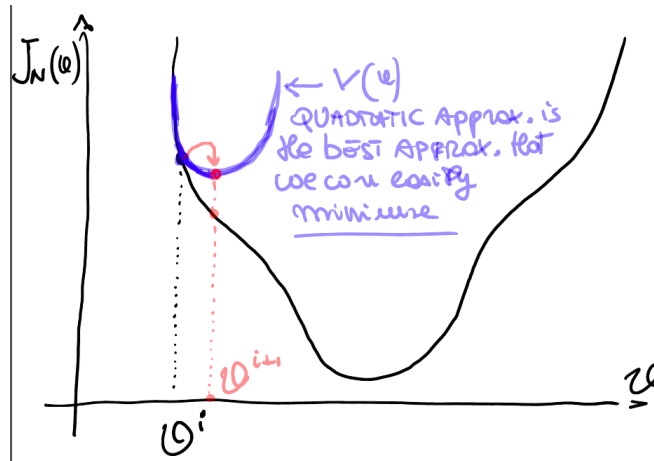


To attenuate this disturbance many **different** initial conditions  $\theta^1$  should be used, each time trying to reach the best solution. The problem is that there is no **theoretical guarantee** that solution converges to the global minimum.

The key problem for iterative methods is building the **step** :

$$\theta^i \rightarrow \theta^{i+1}$$

. The basic idea is to compute the **local quadratic approximation** of  $J_N(\theta)$  around the point  $\theta^i$  and **minimize the local approximation** using the **Newton method**



Using **Taylor's expansion** around  $\theta^i$ :

$$\gamma(\theta) = J_N(\theta^i) + \left[ \frac{\partial J_N(\theta)}{\partial \theta} \Big|_{\theta^i} \right] (\theta - \theta^i) + \frac{1}{2} (\theta - \theta^i)^T \left[ \frac{\partial^2 J_N(\theta)}{\partial^2 \theta} \Big|_{\theta^i} \right] (\theta - \theta^i) \dots$$

Minimizing the function  $\gamma(\theta) \rightarrow \frac{\partial \gamma(\theta)}{\partial \theta} = 0$ :

$$\left[ \frac{\partial J_N(\theta)}{\partial \theta} \Big|_{\theta^i} \right] + \frac{1}{2} \cdot 2 \left[ \frac{\partial^2 J_N(\theta)}{\partial^2 \theta} \Big|_{\theta^i} \right] (\theta - \theta^i) = 0$$

$$\theta = \theta^i - \left[ \frac{\partial^2 J_N(\theta)}{\partial^2 \theta} \Big|_{\theta^i} \right]^{-1} \cdot \left[ \frac{\partial J_N(\theta)}{\partial \theta} \Big|_{\theta^i} \right]$$

Where the first matrix is the inverse of the second order derivative ( **Hessian Matrix** ) of  $J_N(\theta)$  around  $\theta^i$  and the second matrix is the first order ( **Gradient Vector** ) of  $J_N(\theta)$  around  $\theta^i$ .

- **Gradient vector**

$$\frac{\partial J_N(\theta)}{\partial \theta} = \frac{2}{N} \sum_{t=1}^N \epsilon(t) \cdot \frac{\partial \epsilon(t)}{\partial \theta}$$

- **Hessian matrix**

$$\frac{\partial^2 J_N(\theta)}{\partial^2 \theta} = \frac{2}{N} \sum_{t=1}^N \frac{\partial \epsilon(t)}{\partial \theta} \cdot \frac{\partial \epsilon(t)}{\partial \theta}^T + \frac{2}{N} \sum_{t=1}^N \epsilon(t) \frac{\partial^2 \epsilon(t)}{\partial^2 \theta}$$

Which can be approximated to :

$$\frac{\partial^2 J_N(\theta)}{\partial^2 \theta} = \frac{2}{N} \sum_{t=1}^N \frac{\partial \epsilon(t)}{\partial \theta} \cdot \frac{\partial \epsilon(t)}{\partial \theta}^T$$

The approximation holds because of three reasons:

1. **Reason**

We can compute the Hessian using only  $\frac{\partial \epsilon(t)}{\partial \theta}$  without the burden of computing  $\frac{\partial^2 \epsilon(t)}{\partial \theta^2}$

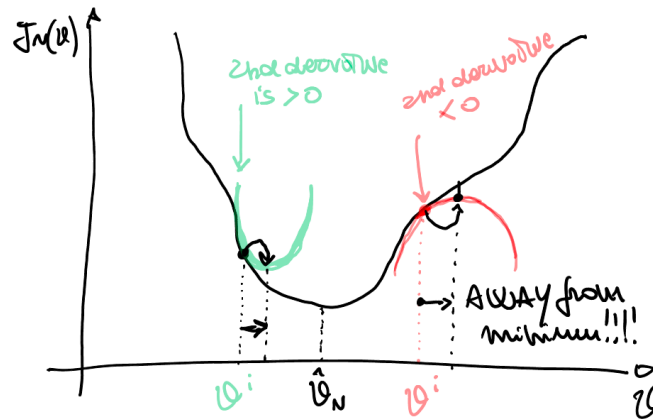
2. **Reason**

Since  $\epsilon(t, \theta) = y(t) - \hat{y}(t|t-1; \theta)$  notice that  $\frac{\partial^2 \epsilon(t)}{\partial \theta^2}$  does **not** depend on  $y(t)$  but only on  $\hat{y}(t|t-1; \theta) \rightarrow y(t-1), y(t-2), \dots$

Moreover if  $\theta^i$  is close to the **optimum**,  $\epsilon(t; \theta) \approx e(t)$ . So under these assumptions  $\epsilon(t; \theta^i)$  and  $\frac{\partial^2 \epsilon(t)}{\partial \theta^2} \Big|_{\theta^i}$  are **uncorrelated** (orthogonal).

3. **Reason**

By neglecting the second Hessian term we can guarantee that the approximation is **semi-definite positive** ( $\geq 0$ )



In conclusion the approximation **guarantees** that the updating step always goes in the **right direction**!

The final updating rule is :

$$\theta^{i+1} = \theta^i - \left[ \sum_{t=1}^N \frac{\partial \epsilon(t; \theta^i)}{\partial \theta} \cdot \frac{\partial \epsilon(t; \theta^i)}{\partial \theta}^T \right]^{-1} \cdot \left[ \sum_{t=1}^N \epsilon(t; \theta^i) \cdot \frac{\partial \epsilon(t; \theta^i)}{\partial \theta} \right]$$

### 4.2.2 Possible updating rules

There are different **updating rules** which have all the same general expression :

$$\theta^{i+1} = \theta^i - \bigcirc \cdot \left[ \frac{\partial J_N(\theta)}{\partial \theta} \right]$$

- **Gradient method**

$$\bigcirc = \mu$$

$\mu$  is a scalar number called **step**. Characteristics :

- + simplest method
- + correct direction **guaranteed**
- slow when near to minimum
- sensitive to choice of  $\mu$  (too small = slow, too big = instability)

This method is the backpropagation rule in NN.

- **Newton method**

$$\bigcirc = \text{Inverse of hessian matrix of performance index}$$

- **Quasi- Newton method**

$$\bigcirc = \text{Inverse of } \geq \text{approximation of the Hessian}$$

This is the one seen in 4.2.1.

Has all the positive aspects of Newton method ( variable step tuned to the specific point of optimisation )but can guarantee the right direction of the step.

To avoid the singularity of matrix  $\sum_{t=1}^N \frac{\partial \epsilon(t; \theta^i)}{\partial \theta} \cdot \frac{\partial \epsilon(t; \theta^i)^T}{\partial \theta}$  usually  $\delta I$  is added where  $\delta$  is a small number and  $I$  the identity matrix.