More Mathematics Mandatory in Data Science

Bart De Moor

KU Leuven, Belgium
Dept.EE: ESAT - STADIUS

bart.demoor@kuleuven.be

KU LEUVEN

I have deeply regretted
that I did not proceed far enough
at least to understand something
of the great leading principles of mathematics
for men thus endowed
seem to have an extra sense

Charles Darwin

Outline

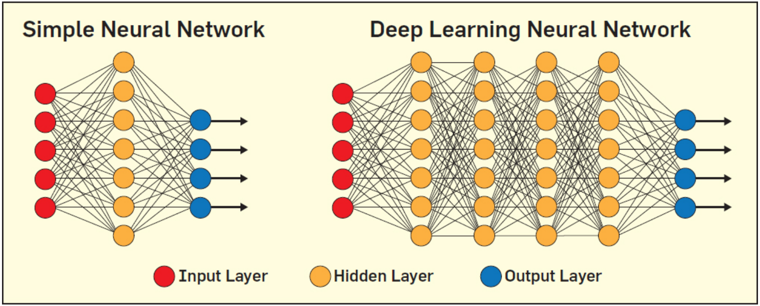
Basic modelling loop:

1. Collect data (preprocess, wrangle, clean, ...)
2. Select model class parametrized by unknown parameters
3. Select an approximation criterion
4. 'Solve' using nonlinear optimization
5. Validate the results
6. Re-iterate when necessary

What do we mean by **'solved'** ?

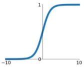
- Result of nonlinear optimization ? Trouble with:
  1. Starting points (feasibility);
  2. Convergence (step sizes, rate, stopping criteria,...)
  3. Local minima
- Solved = *convex* or *set of linear equations* or *eigenvalue problem* !

Outline

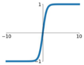


Activation Functions

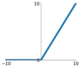
Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$



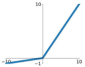
tanh
$\tanh(x)$



ReLU
$\max(0, x)$

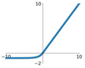


Leaky ReLU
$\max(0.1x, x)$



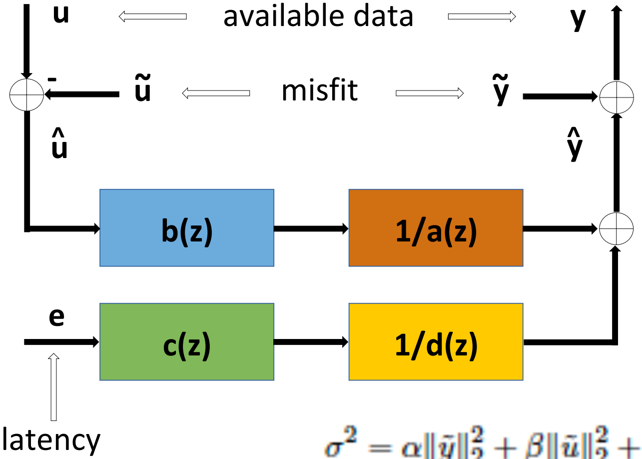
Maxout
$\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU
$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$



$\boxed{\text{Tackled by nonlinear optimization}}$

Misfit-Latency LTI Models



$$\sigma^2 = \alpha\|\tilde{y}\|_2^2 + \beta\|\tilde{u}\|_2^2 + \gamma\|\bar{e}\|_2^2,$$

Errors using inadequate data are much less than those using no data at all.
Charles Babbage.

KU LEUVEN

How nonlinear is least squares linear system identification ?

	Nonlinearity	'Heuristic' remedy
State space	$x_{k+1} = \mathbf{Ax_k} + Bu_k$ Unknown $A \times x_k$	Subspace: Oblique projection and SVD
EIV	Unknown parameters $\times$ misfits $\tilde{u}, \tilde{y}$	Instrumental Variables
PEM	Unknown parameters $\times$ latency input e	Nonlinear optimization

Tackled by nonlinear optimization

Outline

Scalar smooth objective function $f(x) \in \mathbb{R}$:

$$\min_{x \in \mathbb{R}^n} f(x)$$

Gradient flow:

$$\dot{x} = -\frac{\partial f}{\partial x}$$

Lyapunov function:

$$\frac{\partial f}{\partial t} = (\frac{\partial f}{\partial x})^T \frac{\partial x}{\partial t} = -(\frac{\partial f}{\partial x})^T \frac{\partial f}{\partial x} = -\|\frac{\partial f}{\partial x}\|_2^2 \leq 0$$

Convergence to local or global minimum (depends on $x(0)$):

$$\frac{\partial f}{\partial x} = 0$$

Discretization (e.g. forward Euler: $\dot{x} \approx (x_{k+1} - x_k)/\tau_k$):

$$x_{k+1} = x_k - \tau_k \frac{\partial f}{\partial x}(x_k)$$

Scalar smooth objective function $f(x) \in \mathbb{R}$:

$$\min_{x \in \mathbb{R}^n} f(x)$$

Weighted gradient flow: $W(x) = W(x)^T$ nonnegative definite:

$$\dot{x} = -W(x)\frac{\partial f}{\partial x}$$

Lyapunov function (weighted 2-norm):

$$\frac{\partial f}{\partial t} = (\frac{\partial f}{\partial x})^T W(x) \frac{\partial x}{\partial t} = -(\frac{\partial f}{\partial x})^T W(x) \frac{\partial f}{\partial x} \leq 0$$

Convergence to local or global minimum (depends on $x(0)$):

$$\frac{\partial f}{\partial x} = 0$$

Discretization (e.g. forward Euler: $\dot{x} \approx (x_{k+1} - x_k)/\tau_k$ with $W(x) = H(x)^{-1}$ inverse Hessian):

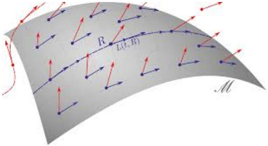
$$x_{k+1} = x_k - \tau_k W(x_k) \frac{\partial f}{\partial x}(x_k)$$

Scalar objective function $f(x) \in \mathbb{R}$, p constraints $g(x) \in \mathbb{R}^p$:

$$\min_{x \in \mathbb{R}^n} f(x) \text{ subject to } g(x) = 0$$

Projected gradient flow:

$$\dot{x} = -\frac{\partial f}{\partial x} + \frac{\partial g}{\partial x} l$$



Columns of $\frac{\partial g}{\partial x} \in \mathbb{R}^{n \times p}$ are normals to tangent space $T_M(x)$ of manifold M generated by $g(x)$ in x.

$$\dot{x} \in T_M(x) \Longrightarrow (\frac{\partial g}{\partial x})^T \dot{x} = 0 \Longrightarrow l = [(\frac{\partial g}{\partial x})^T \frac{\partial g}{\partial x}]^{-1} (\frac{\partial g}{\partial x})^T \frac{\partial f}{\partial x}$$

Project gradient flow

$$\dot{x} = -[I - \frac{\partial g}{\partial x} [(\frac{\partial g}{\partial x})^T \frac{\partial g}{\partial x}]^{-1} (\frac{\partial g}{\partial x})^T] \frac{\partial f}{\partial x} = -\Pi_M(x) \frac{\partial f}{\partial x}$$

Constrained optimization: 'Lagrangean' with Lagrange multipliers:

$$\mathcal{L}(x,l) = f(x) - l^T g(x)$$

First order optimality conditions: $n+p$ eqs. in $n+p$ unknowns:

$$\begin{aligned} \frac{\partial L}{\partial x} &= \frac{\partial f}{\partial x} - \frac{\partial g}{\partial x} l = 0 \\ \frac{\partial L}{\partial l} &= g(x) = 0 \end{aligned}$$

What if $f(x)$ (e.g. least squares) and constraints $g(x)$ are **multivariate polynomial** ? Then

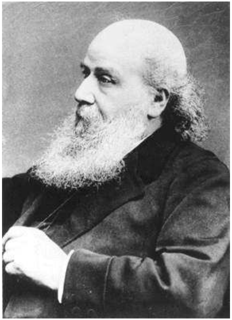
$\boxed{\frac{\partial f}{\partial x} = \frac{\partial g}{\partial x} l \text{ and } g(x) = 0 \text{ are multivariate polynomial !}}$

Solutions ('roots'): local/global minima/maxima, and saddlepoints. The global minimum of $f(x)$ is multivariate polynomial in one of the roots.

$\boxed{\text{How to find the roots of a set of multivariate polynomials ?}}$

KU LEUVEN

Rooting a set of multivariate polynomials is an eigenvalue problem !



James Joseph Sylvester

- Algebra (fundamental theorem)
- Numerical linear algebra (power method and derivates, multiparameter eig. problem (MEVP), SVD)
- (Commutative) Algebraic geometry (ideals and varieties)
- Optimization theory (Lagrangean)
- System theory (state space, realization theory)
- nD system theory (nD realization)
- Operator theory (shift-invariant spaces)
- Interpolation theory (moment problems)

Outline

Example: Univariate polynomial of degree 3:

$$x^3 + a_1 x^2 + a_2 x + a_3 = 0,$$

having three distinct roots x_1, x_2 and x_3

$$\begin{bmatrix} a_3 & a_2 & a_1 & 1 & 0 & 0 \\ 0 & a_3 & a_2 & a_1 & 1 & 0 \\ 0 & 0 & a_3 & a_2 & a_1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ x_1^2 & x_2^2 & x_3^2 \\ x_1^3 & x_2^3 & x_3^3 \\ x_1^4 & x_2^4 & x_3^4 \\ x_1^5 & x_2^5 & x_3^5 \end{bmatrix} = 0$$

- Banded Toeplitz; linear homogeneous equations
- Null space: (Confluent) Vandermonde structure
- Corank (nullity) = number of solutions

Two univariate polynomials: common roots ?

$$\begin{aligned} f(x) &= x^3 - 6x^2 + 11x - 6 = (x-1)(x-2)(x-3) \\ g(x) &= -x^2 + 5x - 6 = -(x-2)(x-3) \end{aligned}$$

$$\begin{array}{l} \\ f(x) = 0 \\ x \cdot f(x) = 0 \\ g(x) = 0 \\ x \cdot g(x) = 0 \\ x^2 \cdot g(x) = 0 \end{array} \quad \begin{array}{c} \begin{array}{ccccc} 1 & x & x^2 & x^3 & x^4 \end{array} \\ \begin{bmatrix} -6 & 11 & -6 & 1 & 0 \\ & -6 & 11 & -6 & 1 \\ -6 & 5 & -1 & & \\ & -6 & 5 & -1 & \\ & & -6 & 5 & -1 \end{bmatrix} \end{array} \begin{bmatrix} 1 & 1 \\ x_1 & x_2 \\ x_1^2 & x_2^2 \\ x_1^3 & x_2^3 \\ x_1^4 & x_2^4 \end{bmatrix} = 0$$

where $x_1 = 2$ and $x_2 = 3$ are the common roots of f and g

- Sylvester matrix = double banded Toeplitz
- Null space = (confluent) Vandermonde structure
- Null space = intersection of null spaces of two banded Toeplitz matrices
- Nullity = number of common zeros

KU LEUVEN

The vectors in the Vandermonde kernel K obey a 'shift structure':

$$\begin{bmatrix} 1 & 1 \\ x_1 & x_2 \\ x_1^2 & x_2^2 \\ x_1^3 & x_2^3 \end{bmatrix} \begin{bmatrix} x_1 & 0 \\ 0 & x_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ x_1^2 & x_2^2 \\ x_1^3 & x_2^3 \\ x_1^4 & x_2^4 \end{bmatrix}$$

or

$$\underline{K}.D = \overline{K}$$

The Vandermonde structure K is not available directly, instead we compute Z, for which $ZV = K$. We now have

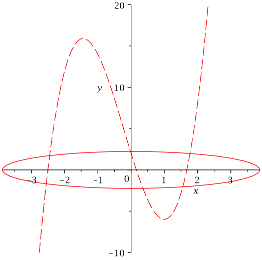
$$\begin{aligned} \underline{K}.D &= \overline{K} \\ \underline{Z}.V.D &= \overline{Z}.V \end{aligned}$$

- Generalized EVP with eigenvalues in D and eigenvectors the columns of V.
- Null space $\mathbf{R}(K) = \mathbf{R}(Z)$ is shift-invariant.

Two polynomials in two variables

- Consider

$$\begin{cases} p(x,y) &= x^2 + 3y^2 - 15 = 0 \\ q(x,y) &= y - 3x^3 - 2x^2 + 13x - 2 = 0 \end{cases}$$



- Fix a monomial order, e.g., $1 < x < y < x^2 < xy < y^2 < x^3 < x^2y < \ldots$

- Construct quasi-Toeplitz Macaulay matrix M:

$$\begin{array}{c} \\ p(x,y) \\ q(x,y) \\ x \cdot p(x,y) \\ y \cdot p(x,y) \end{array} \begin{array}{c} \begin{array}{cccccccccc} 1 & x & y & x^2 & xy & y^2 & x^3 & x^2y & xy^2 & y^3 \end{array} \\ \left[\begin{array}{cccccccccc} -15 & & & 1 & & 3 & & & & \\ -2 & 13 & 1 & -2 & & & -3 & & & \\ & -15 & & & & & 1 & & 3 & \\ & & -15 & & & & & 1 & & 3 \end{array}\right] \end{array} \begin{pmatrix} 1 \\ x \\ y \\ x^2 \\ xy \\ \vdots \\ xy^2 \\ y^3 \end{pmatrix} = 0$$

KU LEUVEN

$$\begin{cases} p(x,y) &= x^2 + 3y^2 - 15 = 0 \\ q(x,y) &= y - 3x^3 - 2x^2 + 13x - 2 = 0 \end{cases}$$

Continue to enlarge M ('quasi-Toeplitzification'):

it #	form	1	x	y	x^2	xy	y^2	x^3	x^2y	xy^2	y^3	x^4	x^3y	x^2y^2	xy^3	y^4	x^5	x^4y	x^3y^2	x^2y^3	xy^4	y^5	$\rightarrow$
$d=3$	p	-15			1		3																
	xp		-15					1		3													
	yp			-15					1		3												
	q	-2	13	1	-2			-3															
$d=4$	x^2p				-15							1		3									
	xyp					-15							1		3								
	y^2p						-15							1		3							
	xq		-2		13	1		-2				-3											
	yq			-2		13	1		-2				-3										
$d=5$	x^3p							-15									1		3				
	x^2yp								-15									1		3			
	xy^2p									-15									1		3		
	y^3p										-15									1		3	
	x^2q				-2			13	1				-2				-3						
	xyq					-2			13	1				-2				-3					
	y^2q						-2			13	1				-2				-3				
	$\downarrow$							$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$

- $\#$ rows grows faster than $\#$ cols $\Rightarrow$ overdetermined system
- If solution exists: rank deficient by construction!

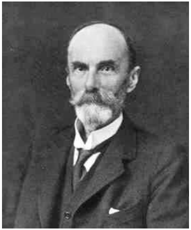
- Macaulay matrix M:

$$M = \begin{bmatrix} \times & \times & \times & \times & 0 & 0 & 0 \\ 0 & \times & \times & \times & \times & 0 & 0 \\ 0 & 0 & \times & \times & \times & \times & 0 \\ 0 & 0 & 0 & \times & \times & \times & \times \end{bmatrix}$$

- Solutions generate vectors in kernel of M:

$$MK = 0$$

- Number of solutions s follows from rank decisions



Francis Sowerby Macaulay

Vandermonde null space K built from s solutions (x_i, y_i):

$$\left[\begin{array}{c|c|c|c} 1 & 1 & \dots & 1 \\ \hline x_1 & x_2 & \dots & x_s \\ y_1 & y_2 & \dots & y_s \\ \hline x_1^2 & x_2^2 & \dots & x_s^2 \\ x_1 y_1 & x_2 y_2 & \dots & x_s y_s \\ y_1^2 & y_2^2 & \dots & y_s^2 \\ \hline x_1^3 & x_2^3 & \dots & x_s^3 \\ x_1^2 y_1 & x_2^2 y_2 & \dots & x_s^2 y_s \\ x_1 y_1^2 & x_2 y_2^2 & \dots & x_s y_s^2 \\ y_1^3 & y_2^3 & \dots & y_s^3 \\ \hline x_1^4 & x_2^4 & \dots & x_4^4 \\ x_1^3 y_1 & x_2^3 y_2 & \dots & x_s^3 y_s \\ x_1^2 y_1^2 & x_2^2 y_2^2 & \dots & x_s^2 y_s^2 \\ x_1 y_1^3 & x_2 y_2^3 & \dots & x_s y_s^3 \\ y_1^4 & y_2^4 & \dots & y_s^4 \\ \hline \vdots & \vdots & \vdots & \vdots \end{array}\right]$$

Setting up an eigenvalue problem in x

- Choose s linear independent rows in K

$$S_1 K$$

- This corresponds to finding linear dependent columns in M

$$\left[\begin{array}{c|c|c|c}
1 & 1 & \ldots & 1 \\
\hline
x_1 & x_2 & \ldots & x_s \\
y_1 & y_2 & \ldots & y_s \\
\hline
x_1^2 & x_2^2 & \ldots & x_s^2 \\
x_1 y_1 & x_2 y_2 & \ldots & x_s y_s \\
y_1^2 & y_2^2 & \ldots & y_s^2 \\
\hline
x_1^3 & x_2^3 & \ldots & x_s^3 \\
x_1^2 y_1 & x_2^2 y_2 & \ldots & x_s^2 y_s \\
x_1 y_1^2 & x_2 y_2^2 & \ldots & x_s y_s^2 \\
y_1^3 & y_2^3 & \ldots & y_s^3 \\
\hline
x_1^4 & x_2^4 & \ldots & x_4^4 \\
x_1^3 y_1 & x_2^3 y_2 & \ldots & x_s^3 y_s \\
x_1^2 y_1^2 & x_2^2 y_2^2 & \ldots & x_s^2 y_s^2 \\
x_1 y_1^3 & x_2 y_2^3 & \ldots & x_s y_s^3 \\
y_1^4 & y_2^4 & \ldots & y_s^4 \\
\hline
\vdots & \vdots & \vdots & \vdots
\end{array}\right]$$

Shifting the selected rows gives (shown for 3 columns)

$$
\left[\begin{array}{c|c|c}
1 & 1 & 1 \\ \hline
x_1 & x_2 & x_3 \\
y_1 & y_2 & y_3 \\ \hline
x_1^2 & x_2^2 & x_3^2 \\
x_1 y_1 & x_2 y_2 & x_3 y_3 \\
y_1^2 & y_2^2 & y_3^2 \\ \hline
x_1^3 & x_2^3 & x_3^3 \\
x_1^2 y_1 & x_2^2 y_2 & x_3^2 y_3 \\
x_1 y_1^2 & x_2 y_2^2 & x_3 y_3^2 \\
y_1^3 & y_2^3 & y_3^3 \\ \hline
x_1^4 & x_2^4 & x_4^4 \\
x_1^3 y_1 & x_2^3 y_2 & x_3^3 y_3 \\
x_1^2 y_1^2 & x_2^2 y_2^2 & x_3^2 y_3^2 \\
x_1 y_1^3 & x_2 y_2^3 & x_3 y_3^3 \\
y_1^4 & y_2^4 & y_3^4 \\ \hline
\vdots & \vdots & \vdots
\end{array}\right]
\longrightarrow \text{“shift with } x\text{”} \longrightarrow
\left[\begin{array}{c|c|c}
1 & 1 & 1 \\ \hline
x_1 & x_2 & x_3 \\
y_1 & y_2 & y_3 \\ \hline
x_1^2 & x_2^2 & x_3^2 \\
x_1 y_1 & x_2 y_2 & x_3 y_3 \\
y_1^2 & y_2^2 & y_3^2 \\ \hline
x_1^3 & x_2^3 & x_3^3 \\
x_1^2 y_1 & x_2^2 y_2 & x_3^2 y_3 \\
x_1 y_1^2 & x_2 y_2^2 & x_3 y_3^2 \\
y_1^3 & y_2^3 & y_3^3 \\ \hline
x_1^4 & x_2^4 & x_4^4 \\
x_1^3 y_1 & x_2^3 y_2 & x_3^3 y_3 \\
x_1^2 y_1^2 & x_2^2 y_2^2 & x_3^2 y_3^2 \\
x_1 y_1^3 & x_2 y_2^3 & x_3 y_3^3 \\
y_1^4 & y_2^4 & y_3^4 \\ \hline
\vdots & \vdots & \vdots
\end{array}\right]
$$

so that:

$$
\left[\begin{array}{c|c|c}
1 & 1 & 1 \\
x_1 & x_2 & x_3 \\
y_1 & y_2 & y_3 \\
x_1 y_1 & x_2 y_2 & x_3 y_3 \\
x_1^3 & x_2^3 & x_3^3 \\
x_1^2 y_1 & x_2^2 y_2 & x_3^2 y_3
\end{array}\right]
\begin{bmatrix} x_1 & & \\ & x_2 & \\ & & x_3 \end{bmatrix}
=
\left[\begin{array}{c|c|c}
x_1 & x_2 & x_3 \\
x_1^2 & x_2^2 & x_3^2 \\
x_1 y_1 & x_2 y_2 & x_3 y_3 \\
x_1^2 y_1 & x_2^2 y_2 & x_3^2 y_3 \\
x_1^4 & x_2^4 & x_4^4 \\
x_1^3 y_1 & x_2^3 y_2 & x_3^3 y_3
\end{array}\right]
$$

KU LEUVEN

Finding the x-roots

Let $D_x = \mathrm{diag}(x_1, x_2, \ldots, x_s)$, then

$$S_1 K D_x = S_x K,$$

where S_1 selects linear independent rows of K and S_x the ones 'hit' by the shift x.

Generalized Vandermonde K is not known as such, instead a null space basis Z is calculated, which is a linear transformation of K:

$$ZV = K$$

which leads to the generalized eigenvalue problem

$$(S_1 Z) V D_x = (S_x Z) V$$

Here, V is the matrix with eigenvectors, D_x contains the roots x as eigenvalues.

Setting up an eigenvalue problem in y

It is possible to shift with y as well. . .

We find

$$S_1 K D_y = S_y K$$

with D_y diagonal matrix of y-components of roots, leading to

$$(S_y Z)V = (S_1 Z)V D_y$$

Some interesting observations:

- same eigenvectors V!
- $(S_x Z)^{-1}(S_1 Z)$ and $(S_y Z)^{-1}(S_1 Z)$ commute
  $\Longrightarrow$ 'commutative algebra'

KU LEUVEN

Single shift invariance

$$\Gamma = \begin{pmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{p-2} \\ CA^{p-1} \end{pmatrix} \Longrightarrow \underline{\Gamma} A = \overline{\Gamma}$$

- Null space of Toeplitz or Sylvester
- Single shift (only one A)
- Cayley-Hamilton
- Shift-invariant $\mathbf{R}(\Gamma)$ fixed by $\lambda(A)$
- 1D observability
- 1D realization theory
- 1D Beurling-Lax
- 'Block' when $C =$ matrix

Multi-shift invariance $(n = 2)$

$$\Gamma = \left(\begin{array}{c} C \\ \hline CA_1 \\ CA_2 \\ \hline CA_1^2 \\ CA_1A_2 \\ CA_2^2 \\ \hline \vdots \\ \hline CA_1^{p-1} \\ CA_1^{p-2}A_2 \\ \vdots \\ CA_2^{p-1} \end{array}\right) \Longrightarrow \begin{array}{rl} \underline{\Gamma}\ A_1 & = S_1\Gamma \\ \underline{\Gamma}\ A_2 & = S_2\Gamma \end{array}$$

- Null space of Macaulay
- n shifts A_1, A_2: $A_1A_2 = A_2A_1$
- nD Cayley-Hamilton (new)
- Multi-shift invariant $\mathbf{R}(\Gamma)$ fixed by $\lambda(A_1)$ and $\lambda(A_2)$
- nD observability
- nD realization theory
- nD Beurling-Lax
- 'Block' when $C =$ matrix

Not treated here:

- Deflate roots at infinity
- Algorithms: kernel-driven versus data-driven (QR), SVD for rank decisions, ...
- Cayley-Hamilton (in 1D and nD)
- 1D and nD system theoretic interpretations of the null space (1D and nD observability matrices) based on 1D/nD state space models (possibly singular (roots at infinity))
- ...

Finding the minimum of univariate polynomial

$$p(x) = \alpha_0 x^n + \alpha_1 x^{n-1} + \ldots + \alpha_n$$

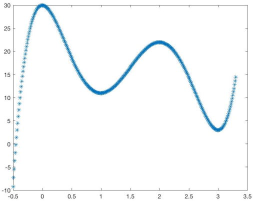
$$\min_{\sigma} \sigma = p(x) \text{ subject to } p'(x) = 0$$

Construct Sylvester matrix M with $\sigma = p(x)$ and $p'(x) = 0$:

$$\begin{pmatrix} M_{11} & M_{12} \\ M_{21} - \sigma I & M_{22} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = 0$$

$$(M_{21} - M_{22} M_{12}^{-1} M_{11}) u = u\sigma$$

$$
\begin{aligned}
p(x) &= 6x^5 - 45x^4 + 110x^3 - 90x^2 + 30 \\
p'(x) &= 30x^4 - 180x^3 + 330x^2 - 180x
\end{aligned}
$$



Fifth degree polynomial

$$
\left(\begin{array}{cccc|ccccc}
0 & -180 & 330 & -180 & 30 & 0 & 0 & 0 & 0 \\
0 & 0 & -180 & 330 & -180 & 30 & 0 & 0 & 0 \\
0 & 0 & 0 & -180 & 330 & -180 & 30 & 0 & 0 \\
0 & 0 & 0 & 0 & -180 & 330 & -180 & 30 & 0 \\
0 & 0 & 0 & 0 & 0 & -180 & 330 & -180 & 30 \\
\hline
30-\sigma & 0 & -90 & 110 & -45 & 6 & 0 & 0 & 0 \\
0 & 30-\sigma & 0 & -90 & 110 & -45 & 6 & 0 & 0 \\
0 & 0 & 30-\sigma & 0 & -90 & 110 & -45 & 6 & 0 \\
0 & 0 & 0 & 30-\sigma & 0 & -90 & 110 & -45 & 6
\end{array}\right)
\begin{pmatrix} 1 \\ x \\ x^2 \\ x^3 \\ x^4 \\ x^5 \\ x^6 \\ x^7 \\ x^8 \end{pmatrix} = 0
$$

Eigenvalues of $(M_{21} - M_{22}M_{12}^{-1}M_{11})$

$$
\lambda \begin{pmatrix} 30 & -54 & 45 & -10 \\ 0 & -30 & 56 & -15 \\ 0 & -90 & 135 & -34 \\ 0 & -204 & 284 & -69 \end{pmatrix} = (30, 22, 11, 3)
$$

KU LEUVEN

Minimal eigenvalue from inverse power method !

Generalizes to multivariate polynomial optimization problems:

$$\min_{x \in \mathbb{R}^n} f(x) \text{ subject to } g(x) = 0,$$

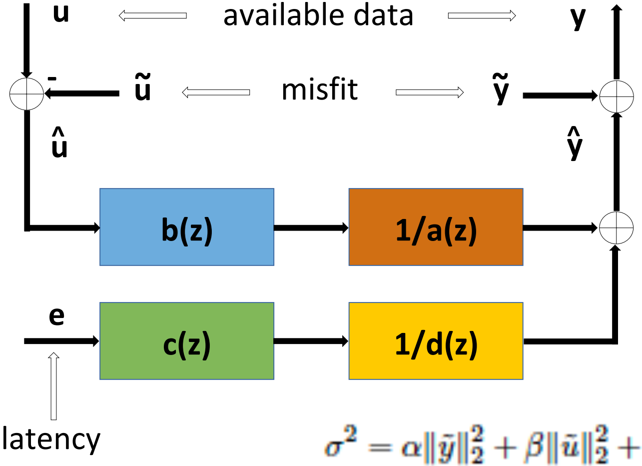
with $f(.)$ scalar multivariate polynomial objective function
$g(x) \in \mathbb{R}^p$ p multivariate polynomial constraints:

- Sylvester matrix $\longrightarrow$ (block) Macaulay matrix
- Null space shift-invariant $\longrightarrow$ multi-shift invariant
- 1 parameter EVP $\longrightarrow$ multi-parameter EVP
- Critical value global minimum σ as smallest eigenvalue $\longrightarrow$ Critical value global minimum $\sigma = f(x^*)$ where elements of x^* are eigenvalues of commuting matrices $A_1, A_2, \ldots$ obtained from multi-shift invariant null space.

Outline

KU LEUVEN

Misfit-Latency LTI Models



$$\sigma^2 = \alpha\|\tilde{y}\|_2^2 + \beta\|\tilde{u}\|_2^2 + \gamma\|\tilde{e}\|_2^2,$$

SISO transfer function (with $\mathcal{Z}\{x_k\} = x(z)$), e.g. ARMAX:

$$y(z) = \frac{b(z)}{a(z)}u(z) + \frac{c(z)}{a(z)}e(z),$$

with polynomial $a(z)$ (monic), $b(z)$, $c(z)$ (monic) of degree n_a, n_b, n_c.

Corresponding difference equation with $\alpha_i, \beta_i, \gamma_i \in \mathbb{R}$:

$$\begin{aligned} y_{k+n_a} + \alpha_1 y_{k+n_a-1} + \ldots + \alpha_{n_a} y_k \quad &= \quad \beta_0 u_{k+n_b} + \beta_1 y_{k+n_b-1} + \ldots + \alpha_{n_b} u_k \\ &\quad\ \ + e_{k+n_c} + \gamma_1 e_{k+n_c-1} + \ldots + \gamma_{n_c} e_k \end{aligned}$$

Algebraic representation, e.g. ARMAX.

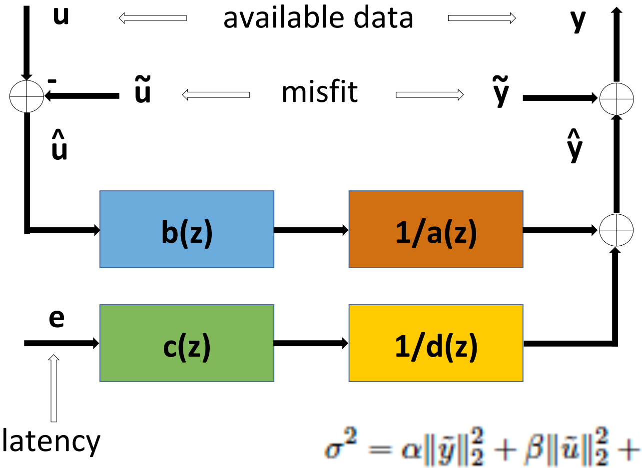
$$T_a y = T_b u + T_c e$$

where $y^T = (y_0\ y_1\ \dots\ y_N)$ and e, u alike.

T_a, T_b, T_c are banded Toeplitz convolution operators, e.g. T_c:

$$\begin{pmatrix} \gamma_{n_c} & \gamma_{n_c-1} & \dots & \dots & \gamma_1 & 1 & 0 & 0 & \dots & 0 \\ 0 & \gamma_{n_c} & \gamma_{n_c-1} & \dots & \gamma_2 & \gamma_1 & 1 & 0 & \dots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \dots & \dots & \dots & \dots & \dots & \gamma_{n_c} & \gamma_{n_c-1} & \dots & \dots & 1 \end{pmatrix}$$

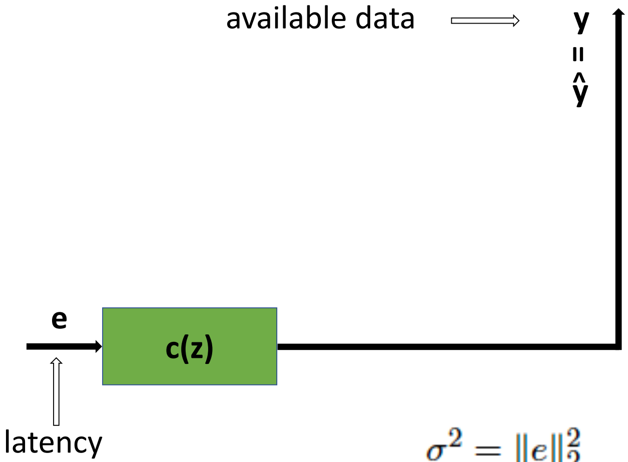
KU LEUVEN



Misfit-Latency LTI Models

$$\sigma^2 = \alpha\|\tilde{y}\|_2^2 + \beta\|\tilde{u}\|_2^2 + \gamma\|\tilde{e}\|_2^2,$$

Latency: MA
Moving Average



$$y = T_c e$$

$$\sigma^2 = \|e\|_2^2$$

Latency case: Moving average: Given $y \in \mathbb{R}^N$.

$$\min_{e \in \mathbb{R}^{N+n_c}, \gamma_i \in \mathbb{R}} \sigma^2 = \|e\|_2^2 \text{ subject to } y = T_c e.$$

$T_c \in \mathbb{R}^{N \times (N+n_c)}$ = banded Toeplitz of full row rank (monic: $\gamma_0 = 1$). $e \in \mathbb{R}^{N+n_c}$ because of n_c initial conditions.
Underdetermined set of linear equations: minimum norm solution

$$e = T_c^\dagger y = T_c^T (T_c T_c^T)^{-1} y,$$

so that

$$\sigma^2 = \|e\|_2^2 = e^T e = y^T (T_c T_c^T)^{-1} y = y^T D_c^{-1} y \ ,$$

where D_c is symm. pos. def. banded Toeplitz, quadratic in the γ_i.

Interpretation: We look for a metric D_c^{-1} in which the weighted norm of y is minimal. $T_c^\dagger$ is a 'whitening' filter.

KU LEUVEN

First order optimality conditions from $\sigma^2 = y^T D_c^{-1} y$:

$$\frac{\partial \sigma^2}{\partial \gamma_i} = y^T \frac{\partial D_c^{-1}}{\partial \gamma_i} y = -y^T D_c^{-1} \frac{\partial D_c}{\partial \gamma_i} D_c^{-1} y = 0 \ , \ i = 1, \ldots, n_c. \quad (1)$$

These are n_c 'nonlinear' equations in the n_c unknowns γ_i.
Since

$$D_c^{-1} = \mathrm{adj}(D_c) / \det(D_c),$$

where the adjugate matrix $\mathrm{adj}(D_c)$ is multivariate polynomial in the γ_i, equations (1) constitute n_c multivariate polynomials in n_c variables γ_i:

$$\frac{\partial \sigma^2}{\partial \gamma_i} = 0 = y^T \mathrm{adj}(D_c) \frac{\partial D_c}{\partial \gamma_i} \mathrm{adj}(D_c) y \ , i = 1, \ldots, n_c.$$

The γ_i are the roots of a set of n_c multivariate polynomials in n_c unkowns.

Call $f = D_c^{-1} y$, then, with $\sigma^2 = y^T D_c^{-1} y$:

$$\begin{pmatrix} D_c & y \\ y^T & \sigma^2 \end{pmatrix} \begin{pmatrix} f \\ -1 \end{pmatrix} = 0. \tag{2}$$

First order optimality conditions: Chain rule with $D_c^{\gamma_i} = \partial D_c / \partial \gamma_i$, $f^{\gamma_i} = \partial f / \partial \gamma_i$ and $\partial \sigma^2 / \partial \gamma_i = 0$:

$$\begin{pmatrix} D_c^{\gamma_i} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} f \\ -1 \end{pmatrix} + \begin{pmatrix} D_c & y \\ y^T & \sigma^2 \end{pmatrix} \begin{pmatrix} f^{\gamma_i} \\ 0 \end{pmatrix} = 0 \ , \ i = 1, \ldots, n_c. \tag{3}$$

$(N+1)(n_c+1)$ **equations:** $N+1$ in (2) and $n_c.(N+1)$ in (3).
$(N+1)(n_c+1)$ **unknowns:** N (f) + $n_c.N$ (f^{γ_i}) + n_c (γ_i) + 1 (σ^2).

The last row of (2) defines σ^2.

The last row of (3) defines n_c orthogonality relations $y^T f^{\gamma_i} = 0, i = 1, \ldots, n_c$.

KU LEUVEN

Latency case: MA ($n_c = 1$)

$$\begin{pmatrix} D_c^\gamma & D_c & 0 \\ D_c & 0 & y \\ 0 & y^T & 0 \end{pmatrix} \begin{pmatrix} f \\ f^\gamma \\ -1 \end{pmatrix} = 0.$$

For $N = 4$:

$$\left(\begin{array}{cccc|cccc|c}
2\gamma & 1 & 0 & 0 & 1+\gamma^2 & \gamma & 0 & 0 & 0 \\
1 & 2\gamma & 1 & 0 & \gamma & 1+\gamma^2 & \gamma & 0 & 0 \\
0 & 1 & 2\gamma & 1 & 0 & \gamma & 1+\gamma^2 & \gamma & 0 \\
0 & 0 & 1 & 2\gamma & 0 & 0 & \gamma & 1+\gamma^2 & 0 \\
\hline
1+\gamma^2 & \gamma & 0 & 0 & 0 & 0 & 0 & 0 & y_0 \\
\gamma & 1+\gamma^2 & \gamma & 0 & 0 & 0 & 0 & 0 & y_1 \\
0 & \gamma & 1+\gamma^2 & \gamma & 0 & 0 & 0 & 0 & y_2 \\
0 & 0 & \gamma & 1+\gamma^2 & 0 & 0 & 0 & 0 & y_3 \\
\hline
0 & 0 & 0 & 0 & y_0 & y_1 & y_2 & y_3 & 0
\end{array}\right)
\left(\begin{array}{c} f_0 \\ f_1 \\ f_2 \\ f_3 \\ \hline f_0^\gamma \\ f_1^\gamma \\ f_2^\gamma \\ f_3^\gamma \\ \hline -1 \end{array}\right) = 0.$$

Regroup as **quadratic eigenvalueproblem** and 'linearize' :

$$(A_2\gamma^2 + A_1\gamma + A_0)z = 0 \text{ with } z = \begin{pmatrix} -1 \\ f \\ f^\gamma \end{pmatrix} \Longrightarrow \begin{pmatrix} 0 & I \\ A_0 & A_1 \end{pmatrix} \begin{pmatrix} z \\ z\gamma \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & -A_2 \end{pmatrix} \begin{pmatrix} z \\ z\gamma \end{pmatrix} \gamma.$$

$$\boxed{\text{Block shift invariant null space} \Longrightarrow \text{EVP}}$$

KU LEUVEN

Latency case MA ($n_c = 2$)

$$\begin{pmatrix} D_c^{\gamma_i} & D_c & 0 \\ D_c & 0 & y \\ 0 & y^T & 0 \end{pmatrix} \begin{pmatrix} f \\ f^{\gamma_i} \\ -1 \end{pmatrix} = 0 \ , \ i = 1, 2.$$

Regroup in a multi-parameter eigenvalueproblem with $z^T = (-1 \ f^T \ (f^{\gamma_1})^T \ (f^{\gamma_2})^T \ )$:

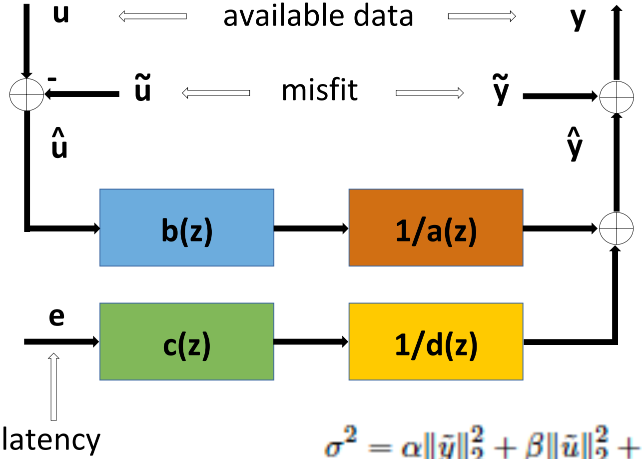
$$(A_{00} + A_{10}\gamma_1 + A_{01}\gamma_2 + A_{20}\gamma_1^2 + A_{11}\gamma_1\gamma_2 + A_{02}\gamma_2^2) \left(\begin{array}{c} z \\ \hline z\gamma_1 \\ z\gamma_2 \\ \hline z\gamma_1^2 \\ z\gamma_1\gamma_2 \\ z\gamma_1^2 \end{array}\right) = 0.$$

and build up block Macaulay recursively (quasi-Toeplitz-ify) until 'mind-the-gap' starts in the null space, which is **block multi-shift invariant**:

$$\begin{array}{c} \\ 1 \\ \times\gamma_1 \\ \times\gamma_2 \\ \times\gamma_1^2 \\ \times\gamma_1\gamma_2 \\ \times\gamma_2^2 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccccccccc} 1 & \gamma_1 & \gamma_2 & \gamma_1^2 & \gamma_1\gamma_2 & \gamma_2^2 & \gamma_1^3 & \gamma_1^2\gamma_2 & \gamma_1\gamma_2^2 & \gamma_2^3 & \gamma_1^4 & \dots \end{array} \\ \begin{pmatrix} A_{00} & A_{10} & A_{01} & A_{20} & A_{11} & A_{02} & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & A_{00} & 0 & A_{10} & A_{01} & 0 & A_{20} & A_{11} & A_{02} & 0 & 0 & \dots \\ 0 & 0 & A_{00} & 0 & A_{10} & A_{01} & 0 & A_{20} & A_{11} & A_{02} & 0 & \dots \\ 0 & 0 & 0 & A_{00} & 0 & 0 & A_{10} & A_{01} & 0 & 0 & A_{20} & \dots \\ 0 & 0 & 0 & 0 & A_{00} & 0 & 0 & A_{10} & A_{01} & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & A_{00} & 0 & 0 & A_{10} & A_{01} & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \end{array} \left(\begin{array}{c} z \\ \hline z\gamma_1 \\ z\gamma_2 \\ \hline z\gamma_1^2 \\ z\gamma_1\gamma_2 \\ z\gamma^2 \\ \hline z\gamma_1^3 \\ \vdots \end{array}\right) = 0$$

$$\boxed{\text{Block multi-shift invariant null space} \implies \text{Multiparameter EVP}}$$

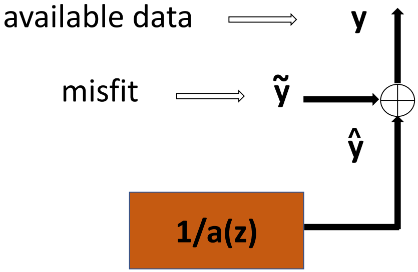
Misfit case: Least squares realization (n_a)



Misfit-Latency LTI Models

$$\sigma^2 = \alpha\|\tilde{y}\|_2^2 + \beta\|\tilde{u}\|_2^2 + \gamma\|\tilde{e}\|_2^2,$$

Misfit case: Least squares realization (n_a)



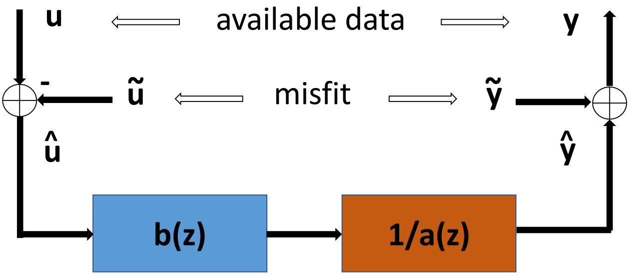
Misfit:
LS realization

$$T_a \hat{y} = 0$$

$$\sigma^2 = \|\tilde{y}\|_2^2$$

Block multi-shift invariant null space $\Longrightarrow$ Multiparameter EVP

Misfit case: Dynamic Total Least Squares (n_a, n_b)



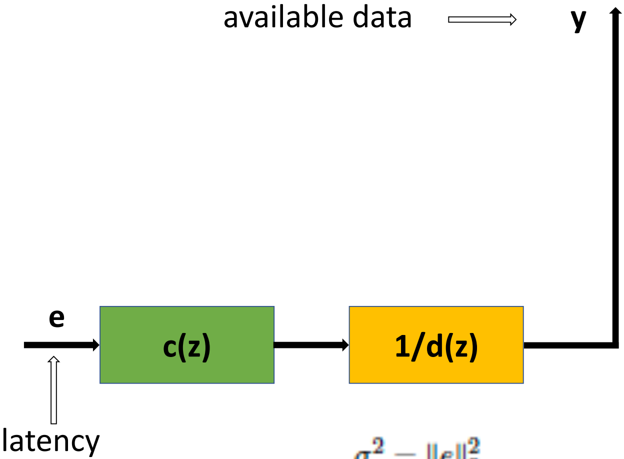
Misfit:
Dynamic Total LS

$$T_a \hat{y} + T_b \hat{u} = 0$$

$$\sigma^2 = \|\tilde{u}\|_2^2 + \|\tilde{y}\|_2^2$$

Block multi-shift invariant null space $\Longrightarrow$ Multiparameter EVP

Latency case: ARMA (n_a, n_c)



Latency: ARMA
Autoregressive
Moving Average

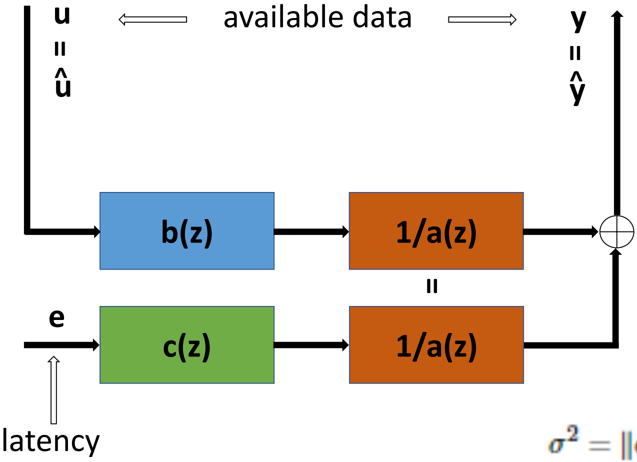
$$T_a y = T_c e$$

$$\sigma^2 = \|e\|_2^2$$

Block multi-shift invariant null space $\Longrightarrow$ Multiparameter EVP

KU LEUVEN

Latency case: ARMAX (n_a, n_b, n_c)



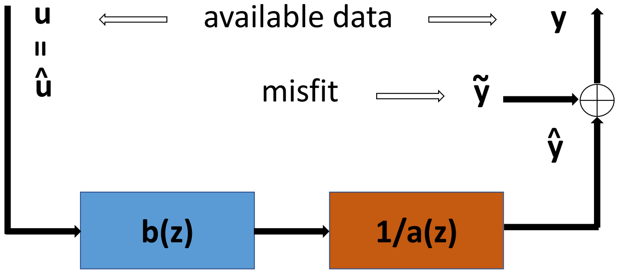
Latency: ARMAX ARMA + eXoge-neous inputs

$$T_a y = T_b u + T_c e$$

$$\sigma^2 = \|e\|_2^2$$

Block multi-shift invariant null space $\Longrightarrow$ Multiparameter EVP

Misfit case: Output Error (n_a, n_b)



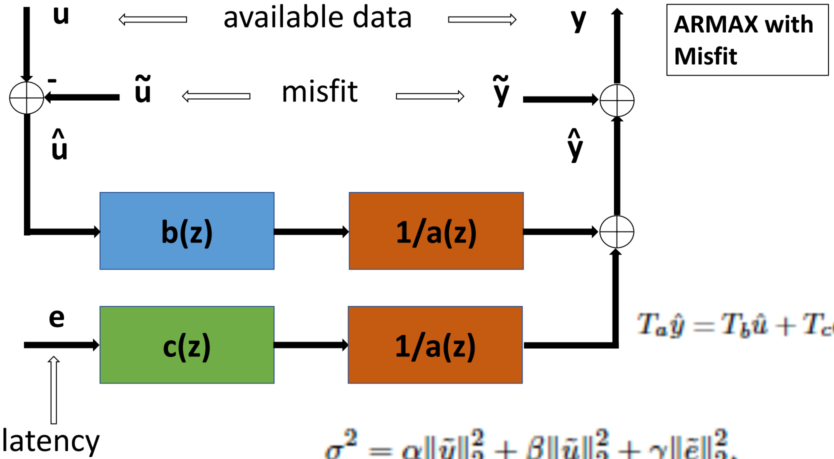
Misfit: OE
Output error

$$T_a \hat{y} = T_b u$$

$$\sigma^2 = \|\tilde{y}\|_2^2$$

Block multi-shift invariant null space $\Longrightarrow$ Multiparameter EVP

Misfit+Latency case: ARMAX with I/O Misfit(n_a, n_b, n_c)

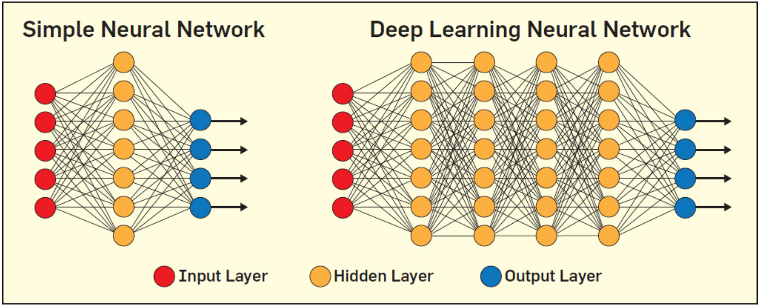


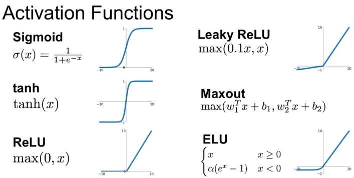
$$T_a \hat{y} = T_b \hat{u} + T_c e$$

$$\sigma^2 = \alpha \|\tilde{y}\|_2^2 + \beta \|\tilde{u}\|_2^2 + \gamma \|\tilde{e}\|_2^2,$$

Block multi-shift invariant null space $\Longrightarrow$ Multiparameter EVP

Name	u	e	α	β	γ	a	b	c	d
Exact data									
Autonomous system	0	0	∞	∞	∞	a	1	1	1
Exact FIR	u	0	∞	∞	∞	1	b	1	1
Diff. eq.	u	0	∞	∞	∞	a	b	1	1
$\vdots$									
Latency									
MA	0	e	∞	∞	1	1	1	c	1
AR	0	e	∞	∞	1	1	1	1	d
ARMA	0	e	∞	∞	1	1	1	c	d
ARMAX	u	e	∞	∞	1	a	b	c	a
$\vdots$									
Misfit									
LS Realization	0	0	1	∞	∞	a	1	1	1
OE FIR	u	0	1	∞	∞	1	b	1	1
IE FIR	u	0	∞	1	∞	1	b	1	1
IE+OE FIR	u	0	α	β	∞	1	b	1	1
OE	u	0	1	∞	∞	a	b	1	1
IE	u	0	∞	1	∞	a	b	1	1
Dynamic TLS	u	0	α	β	∞	a	b	1	1
$\vdots$									
Misfit + Latency									
ARMAX with M+L	u	e	α	β	γ	a	b	c	a
$\vdots$									





If

- Activation functions = polynomial function of sum of weighted (= parameters) inputs
- Objective function is multivariate polynomial (e.g. least squares)

then, in principle, **training a neural network is finding the minimal eigenvalue of a (large) matrix**.

Outline

Main conclusions:

- Multivariate polynomial problems are ubiquitous (many applications!)
- Multivariate polynomial optimization problems are eigenvalue problems
- Path goes over
  - Affinely structured matrices: Toeplitz, Sylvester, Macaulay and block versions
  - Multiparameter eigenvalue problems
  - Null spaces that are (block) (multi-)shift invariant
  - Roots follow from the multi-shift invariance via nD realization theory
  - Only one minimizing root needs to be calculated (e.g. inverse power method)
- Misfit/latency identification of LTI dynamical systems = solved !
- Patiently studying mathematics (over different fields), inventing new math and deploying it into mathematical engineering, pays off.
- This talk: many details omitted !

Future work

- Numerical algorithms (large scale structure exploiting iterative algorithms)
- Explore system theoretic properties
- Explore numerical issues: conditioning, sensitivity, etc.
- Applicability to neural nets / machine learning ?

Bolzmann at 55: "When I look back on all the scientific developments and revolutions that occurred since the beginning of my career, I feel like a monument of ancient scientific memories. I would go further and say that I am the only one left who still grasped the old doctrines with unreserved enthusiasm - at any rate I am the only one who still fights for them as far as I can. I regard as my life's task to help to ensure, by as clear and logically ordered an elaboration as I can give of the results of the classical theory, that the great portion of valuable and permanently usable material that in my view is contained in it need not be rediscovered one day, which would not be the first time that such an event had happened in science. I therefore present myself to you as a reactionary, one who has stayed behind and remains enthusiastic for the old classical doctrines as against the men of today; but I do not believe that I am narrow-minded or blind for the advantages of the new doctrines."

Bolzmann at 55: "When I look back on all the scientific developments and revolutions that occurred since the beginning of my career, I feel like a monument of ancient scientific memories. I would go further and say that I am the only one left who still grasped the old doctrines with unreserved enthusiasm - at any rate I am the only one who still fights for them as far as I can. I regard as my life's task to help to ensure, by as clear and logically ordered an elaboration as I can give of the results of the classical theory, that the great portion of valuable and permanently usable material that in my view is contained in it need not be rediscovered one day, which would not be the first time that such an event had happened in science. I therefore present myself to you as a reactionary, one who has stayed behind and remains enthusiastic for the old classical doctrines as against the men of today; but I do not believe that I am narrow-minded or blind for the advantages of the new doctrines."

At the end of the day, the only thing we really understand, is linear algebra