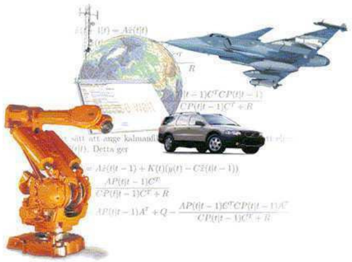
Convexity Issues in System Identification

State-of-the-art System Identification Revisited



Lennart Ljung with Tianshi Chen

Reglerteknik, ISY, Linköpings Universitet

Outline

- A review of the classical, conventional System Identification Setup With Special Emphasis on
  - Convexity Aspects
  - Bias – Variance
  - Regularization
  - Differential Algebra

System Identification in Short

A Typical Problem

Given Observed Input-Output Data: Find a Description of the System that Generated the Data [Simulator or Predictor. Linear System: Impulse response or Bode plot].

Basic Approach

Find a suitable Model Structure, Estimate its parameters, and compute the response of the resulting model

Techniques

Estimate the parameters by ML techniques/PEM (prediction error methods). Find the model structure by AIC, BIC or Cross Validation

More Formally

Models:

Model Structure: $\mathcal{M}$. Parameters: θ. Model: $\mathcal{M}(\theta)$.
Observed input–output (u, y) data up to time t: Z^t
Model described by predictor: $\mathcal{M}(\theta) : \hat{y}(t|\theta) = g(t, \theta, Z^{t-1})$.

Estimation: ML or PEM techniques

$-$ log likelihood function $V_N(\theta) = \sum_{t=1}^{N} |y(t) - \hat{y}(t|\theta)|^2$
$\hat{\theta}_N = \arg\min_\theta V_N(\theta)$

Model Structure (size) determination, AIC, BIC:

$\mathcal{M}(\hat{\theta}_N) = \arg\min_{\mathcal{M},\theta} [\log V_N(\theta) + g(N) \dim\theta]$
$g(N) = 2$ or $\log N$

Comment on Model Structure Selection

The model fit as measured by $\sum_{t=1}^{N} |y(t) - \hat{y}(t|\theta)|^2$ for a certain set of data will always improve as the model structure becomes larger (more parameters). The parameters will start adjusting also to the actual noise effects in the data ["Overfit"]

There are two ways of counteracting this effect:

- Compute the model on one set of (estimation) data and evaluate the fit on another (validation) data set. [Cross-Validation]
- Add a penalty term to the criterion which balances the overfit:

$$\mathcal{M}(\hat{\theta}_N) = \arg\min_{\mathcal{M},\theta} [\log V_N(\theta) + g(N)\text{dim}\theta]$$

$$AIC : g(N) = 2, \quad BIC : g(N) = \log(N)$$

AIC: Akaike's Information Criterion. BIC: Bayesian Information Criterion [= MDL: Minimum Description Length]

Model Estimate Properties

As the number of data, N, tends to infinity

- $\hat{\theta}_N \to \theta^* \sim \arg\min_\theta E|\varepsilon(t,\theta)|^2$ the best possible predictor in $\mathcal{M}$
- If $\mathcal{M}$ contains a true description of the system
  - Cov $\hat{\theta}_N = \frac{\lambda}{N}[E\psi(t)\psi^T(t)]^{-1}$ [$\psi(t) = \frac{d}{d\theta}\hat{y}(t|\theta), \lambda$: noise level]...
  - ... is the Cramér-Rao lower bound for any (unbiased) estimator.

E: Expectation. These are very nice optimal properties:

- The model structure is large enough: The ML/PEM estimated model is (asymptotically) the best possible unbiased one. Has smallest possible variance (Cramér- Rao)
- The model structure is not large enough: The ML/PEM estimate converges to the best possible approximation of the system. The limit model has the smallest possible bias.

Linear Models

General Description

$y(t) = G(q,\theta)u(t) + H(q,\theta)e(t)$, q : shift op. e : white noise

$$G(q,\theta)u(t) = \sum_{k=1}^{\infty} g_k u(t-k), \quad H(q,\theta)e(t) = 1 + \sum_{k=1}^{\infty} h_k e(t-k)$$

Predictor

$$\hat{y}(t|\theta) = G(q,\theta)u(t) + [I - H^{-1}(q,\theta)][y(t) - G(q,\theta)u(t)]$$

Asymptotics: [Φ_u, Φ_v: Spectra of input and additive noise $v = He$.]

$$\hat{\theta}_N \to \theta^* = \arg\min_\theta \int_{-\pi}^{\pi} |G(e^{i\omega},\theta) - G_0(e^{i\omega})|^2 \frac{\Phi_u(\omega)}{|H(e^{i\omega},\theta)|^2} d\omega$$

$$\text{Cov}G(e^{i\omega},\hat{\theta}_N) \sim \frac{n}{N}\frac{\Phi_v(\omega)}{\Phi_u(\omega)} \text{ as } n, N \to \infty \quad n : \text{model order}$$

Common Parameterizations:

BJ:

$$G(q,\theta) = \frac{B(q)}{F(q)}; \quad H(q,\theta) = \frac{C(q)}{D(q)}$$

$$B(q) = b_1 q^{-1} + b_2 q^{-2} + \ldots b_{nb} q^{-nb}$$

$$F(q) = 1 + f_1 q^{-1} + \ldots + f_{nf} q^{-nf}$$

$$\theta = [b_1, b_2, \ldots, f_{nf}]$$

ARX:

$$y(t) = \frac{B(q)}{A(q)}u(t) + \frac{1}{A(q)}e(t) \text{ or}$$

$$A(q)y(t) = B(q)u(t) + e(t) \text{ or}$$

or ARX:

$$y(t) + a_1 y(t-1) + \ldots + a_n y(t-n) = b_1 u(t-1) + \ldots + b_n u(t-n)$$

State-Space Models

State-Space:

$$x(t+1) = Ax(t) + Bu(t) + Ke(t)$$
$$y(t) = Cx(t) + e(t)$$

Corresponds to

$$G(q,\theta) = C(qI - A)^{-1}B. \quad H(q,\theta) = C(qI - A)^{-1}K + I$$

Continuous Time (CT) Models

Physical Model with unknown parameters

$$\dot{x}(t) = \mathcal{F}(\theta)x(t) + \mathcal{G}(\theta)u(t) + w(t)$$
$$y(t) = C(\theta)x(t) + D(\theta)u(t) + v(t)$$

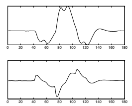
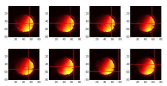
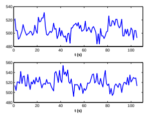
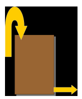
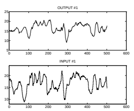
Sample it (with correct Input Intersample Behaviour):

$$x(t+1) = A(\theta)x(t) + B(\theta)u(t) + K(\theta)e(t)$$
$$y(t) = C(\theta)x(t) + e(t)$$

Now apply the discrete time formalism to this model, which is parameterized in terms of the CT parameters θ

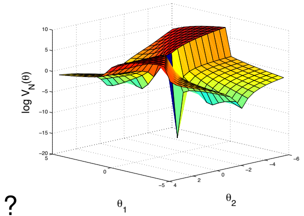
Status of the “Standard Framework”

- Well established statistical theory
- Optimal asymptotic properties
- Efficient software
- Many applications in very diverse areas. Some examples:

  - Aircraft Dynamics:  
  - Brain Activity (fMRI):  
  - Pulp Buffer Vessel:   ;

Time-out

This is a bright and rosy picture. Any issues and problems?



- $\arg\min_\theta V_N(\theta)$?
- Convexity issues!
- Small data sizes – complex systems (asymptotics do not apply): Well tuned bias–variance trade–off.

Bias – Variance Trade Off

Any estimated model is incorrect. The errors have two sources:

- Bias: The model structure is not flexible enough to contain a correct description of the system.
- Variance: The disturbances on the measurements affect the model estimate, and cause variations when the experiment is repeated, even with the same input.

Mean Square Error (MSE) = $|\text{Bias}|^2$ + Variance.
When model flexibility ↑,Bias ↓ and Variance ↑.
To minimize MSE is a good trade-off in flexibility.
In state-of-the-art Identification, this flexibility trade-off is governed primarily by model order. May need a more powerful tuning instrument for bias–variance trade-off.

Convexity Issues

For most model structures the criterion function $V_N(\theta) = \sum_{t=1}^{N} |y(t) - \hat{y}(t|\theta)|^2$ is non-convex and multi-modal (several local minima). *Evolutionary Minimization Algorithms* could be applied, but no major successes for identification problems have been reported.

Important observation for linear models

ARX can Approximate Any Linear System

Arbitrary Linear System: $y(t) = G_0(q)u(t) + H_0(q)e(t)$

ARX model order n, m : $A_n(q)y(t) = B_m(q)u(t) + e(t)$

as $N >> n, m \to \infty$

$[\hat{A}_n(q)]^{-1}\hat{B}_m(q) \to G_0(q),\ [\hat{A}_n(q)]^{-1} \to H_0(q)$

Convexity – Initial Estimates

The ARX-model Is a Linear Regression

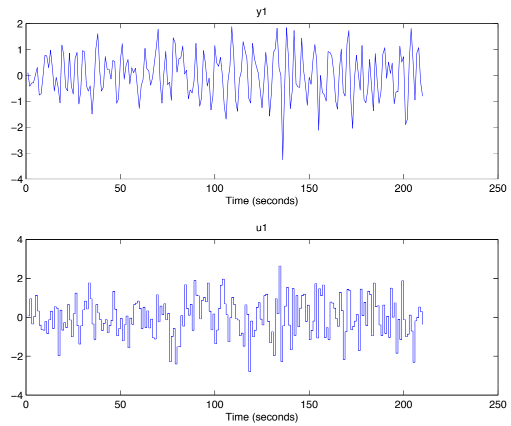
Note that the ARX-model is estimated as a linear regression $Y = \Phi\theta + E$, (Φ containing lagged y, u and θ containing a, b)
A convex estimation problem.

Virtually all methods to initialize the non-convex minimization of the ML criterion for linear models are based on an ARX-model of some kind.

In particular, so called *subspace methods* for state-space models can simplistically be seen as a high order ARX model that is reduced by Hankel-norm model order reduction. (Using SVD, so the algorithm is non-iterative.)

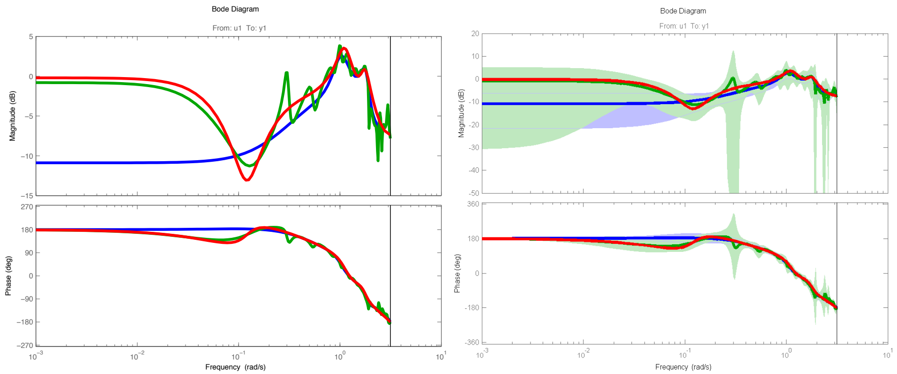
How High Orders are Required for the ARX Approximation?

Try to estimate the transfer function for the data:



How High Orders are Required for the ARX Approximation?

Estimate ARX-model of order 10 and 30: Bode plots of models together with true system:



Order 10. Order 30. True. The high order model picks up the true curves better, but seem more ”shaky”. Look at Uncertainty regions!

How to Curb Variance/Flexibility?

The ARX approximation property is valuable, but high orders come with high variance.
Can we curb the flexibility that causes high variance other than by lower order? Regularization

High Order Models – Regularization

Curb the freedom of the model by adding a regularization term to the Least Squares Criterion:

$$Y = \Phi\theta + E$$
$$\hat{\theta}_N{}^R = \arg\min_{\theta} |Y - \Phi\theta|^2 + \theta^T P^{-1}\theta$$

P is the Regularization Matrix. $\hat{\theta}_N^R = (R_N + P^{-1})^{-1}\Phi^T Y$ MSE:

$$\mathsf{E}[(\hat{\theta}_N^R - \theta_0)(\hat{\theta}_N^R - \theta_0)^T] = (R_N + P^{-1})^{-1} \times$$
$$(R_N + P^{-1}\theta_0\theta_0^T P^{-1})(R_N + P^{-1})^{-1} \qquad R_N = \Phi\Phi^T,\ \theta_0 = \text{true par}$$

Minimized by $P = \theta_0\theta_0^T$: MSE = $(R_N + P^{-1})^{-1}$ How to select P?

Regularization – Bayesian Interpretation

Suppose θ is a random variable, that *a priori* (before the measurement data have been observed) is assumed to be Gaussian with zero mean and covariance matrix P: $\theta^{prior} \in N(0, P)$

$Y = \Phi\theta + E$, so Y and θ are dependent variables. After Y has been measured, we know more about θ:

$$\theta^{post} \in N(\hat{\theta}_N^R, P^{post})$$

where $\hat{\theta}_N^R$ is the regularized LS estimate from the previous slide.

So, the *Maximum a posteriori (MAP)* estimate is equal to the regularized LS estimate with P as the regularization matrix.

So that is a natural way to think of a good regularization matrix: Let it mimic what is known or assumed about the parameter to be estimated. – It is the covariance matrix of the parameter vector.

Tuning the Regularization Matrix

θ is a Gaussian random vector with zero mean and covariance matrix P: $\theta \in N(0, P)$. The measured data in Φ is a known matrix, and the noise $E \in N(0, I)$. Then the output $Y = \Phi\theta + E$ is itself a Gaussian vector:

$$Y = \Phi\theta + E \in N(0, Z(P)), \quad Z(P) = \Phi P \Phi^T + I$$

So we know the pdf of Y given P, and P can be estimated by ML:

ML Estimate of P

$$\hat{P} = \arg\min_P Y^T Z(P)^{-1} Y + \log\det Z(P)$$

If P is parameterized by some hyperparameters α, $P(\alpha)$, these can be estimated by

ML Estimate of Hyperparameters

$$\hat{\alpha} = \arg\min_\alpha Y^T Z(P(\alpha))^{-1} Y + \log\det Z(P(\alpha))$$

ARX Model Priors

When estimating an ARX-model, we can think of the predictor

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

as made up of two impulse responses, A and B. The vector θ should thus mimic two impulse responses, both typically exponentially decaying and smooth.We can thus have a reasonable prior for θ:

$$P(\alpha_1, \alpha_2) = \begin{bmatrix} P^A(\alpha_1) & 0 \\ 0 & P^B(\alpha_2) \end{bmatrix} \qquad \text{Block Diagonal } A\&B$$

where the hyperparameters α describe decay and smoothness of the impulse responses. Typical choice:

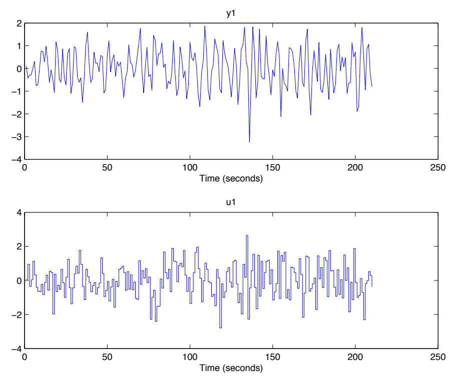
TC kernel

$P_{k,\ell} = C\min(\lambda^k, \lambda^\ell); \quad \alpha = [C, \lambda], \quad \lambda < 1$
$E|b_k|^2 = C\lambda^k$, $\text{corr}(b_k, b_{k+1}) = \sqrt{\lambda}$

An Example

Equipped with these tools, let us now test some data `z` (selected but not untypical). The example uses complex dynamics and few (210) data, so this is a case where asymptotic properties are not prevalent.

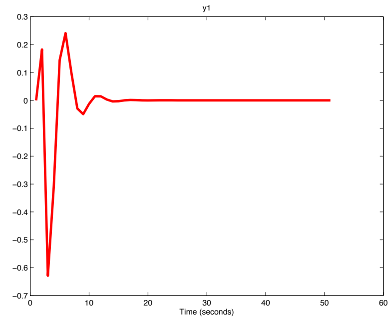
```
plot(z)
```



Estimate a Model: State-of-the-Art

We will try the state-of-the art approach: Estimate SS models of different orders. Determine the order by the AIC criterion.

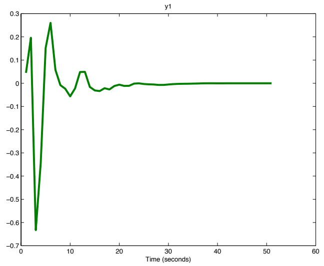
```
 for k=1:30
   m{k}= ssest(z,k);
end
(dum,n) =
min(aic(m{:}));
mss = m{n};
impulse(mss)
```



Estimate a Model: Regularized ARX

Now, let us try an ARX model with na=5, nb=60. Estimate a regularization matrix with the 'TC' kernel (2 parameters, C, λ each for the A and B parts):

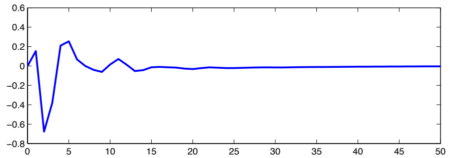
```
 aopt = arxOptions;
(L,R) = arxRegul(z,[5 60 0],'TC');
% "inv(P) = L*R"
aopt.Regularization.R = R;
aopt.Regularization.Lambda = L;
mr = arx(z,[5 60 0],aopt);
impulse(mr)
```



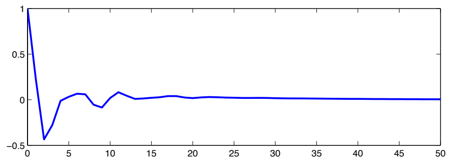
The Oracle

The examined data were obtained from a randomly generated model of order 30:

$$y(t) = G_0(q)u(t) + H_0(q)e(t)$$



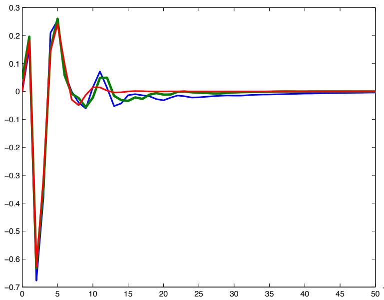
The input is Gaussian white noise with variance 1, and e is white noise with variance 0.1. The impulse responses of G_0 and H_0 are shown at the right.

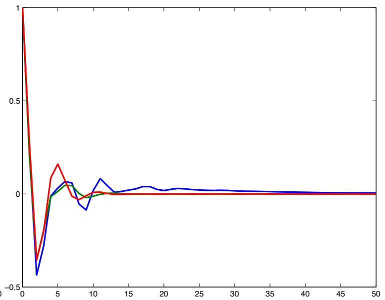


How Well Did Our Models `mss` and `mr` Do?

Blue curves: The true impulse responses.
Red curves: The selected SS-model `mss`
Green curves: The regularized ARX model `mr`

G: impulse response from u H: The impulse response from e





G : fit: **mss: 79.42%** **mr: 83.55%** H: fit **mss: 77.05%**, **mr: 81.59%**

Surprise ?

ML beaten by an "outsider algorithm"!:That is a surprise and embarrassment! There is a certain randomness in these data, but Monte-Carlo simulations substantiate the observed conclusion.

Even though ML is known to have the quoted optimal properties for bias and variance, the observation is still not a contradiction.

Recall: Mean Square Error (MSE) = $|\text{Bias}|^2$ + Variance.

ML: Bias $\approx 0 \Rightarrow$: MSE = Variance = CR Lower bound for unbiased estimators

But with some bias, Variance could be clearly smaller then CRB

Recall for Lin Reg: CRB = $(\Phi\Phi^T)^{-1} > (\Phi\Phi^T + P^{-1})^{-1}$ = MSE for best regularized estimated. More pronounced for short data

Objections?

Recall: mss fit 79.42%, mr fit 83.55 %

- We were just unlucky to pick order 3 (AIC). Other model selection criteria would have given better results.
  - If we ask the oracle what is the best possible state-space order for ML estimated model, the answer is **order 12 for G with a fit 82.95 %** and **order 3 for H with a fit 77.04%** So the regularized ARX -model gives better fit to both G and H than is at all possible for ML estimated state-space models [for these data].
- The R-ARX model is of order 60, and it is unfair to compare it with SS models of low order.
  - Try `mred = balred(mr,7)` to create a 7th order SS-model. It still has a G-fit of **83.56%** and outperforms the oracle-selected ML SS models.

Discussion

- In this case Regularized ARX gave a much better and more flexible bias–variance trade off through the continuously adjustable hyperparameters in the regularization matrix — Compared to the state-of-the art bias–variance trade off in terms of discrete model orders.
- Can we forget about `ssest` and move over to regularized ARX?
  - No, recall that the studied situation had quite few data, and the good trade-off is reached for rather large bias, not favoring ML.
  - But one should be equipped with regularized ARX in one's toolbox
- Regularized ARX (possible followed by `balred`) can be seen as a convexification of the state-of-the art SS model estimation techniques.
  NB: Tuning of hyperparameters normally non-convex

Algebraic Convexification of Model Structures

And Now for Something Completely Different:
Consider the following example, inspired by the Michaelis-Menten growth kinetic equations:

$$\dot{y} = \theta_1 \frac{y}{\theta_2 + y} - y + u$$

y: concentration of enzyme. u addition of nutrition substrate.
θ_1: Maximal growth rate. θ_2: the Michaelis constant.
Measure the concentration with some noise:

$$y_m(t_k) = y(t_k) + e(k)$$

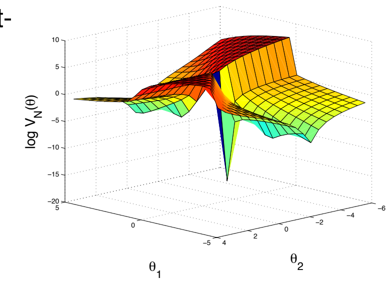
The Likelihood function

The Likelihood criterion function for estimating θ is defined from y_m and u as

$$V_N(\theta) = \sum_{k=1}^{N} [y_m(t_k) - \hat{y}(t_k|\theta)]^2$$

$$\dot{\hat{y}}(t|\theta) = \theta_1 \frac{\hat{y}(t|\theta)}{\theta_2 + \hat{y}(t|\theta)} - \hat{y}(t|\theta) + u(t)$$

It is depicted to the right for an impulse input:



$\arg\min$??!!

Is this complicated relationship between y, u and θ an inherent property of the model?

Algebraic Manipulations

Let us examine the relationship between y, u and θ in more detail:

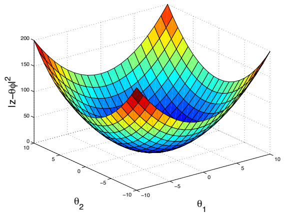
$$\dot{y} = \theta_1 \frac{y}{\theta_2 + y} - y + u$$

$$\dot{y}y + \theta_2\dot{y} = \theta_1 y - y^2 - \theta_2 y + uy + \theta_2 u$$

or

$$\dot{y}y + y^2 - uy = \begin{bmatrix}\theta_1 & \theta_2\end{bmatrix} \begin{bmatrix} y \\ u - \dot{y} - y \end{bmatrix}$$

or $z = \theta^T \phi$



This is not a reparameterization, but a reorganization of the original equations. z and ϕ are still measured, and they are related to θ as a linear regression. The criterion has been convexified

Is This a General Property?

Suppose we have a collection of physical model equations.
(u: input. y: output, z: latent variables (e.g. states) ,θ:parameters.):

A Differential Algebraic Equation (DAE) Model Structure

$$\mathcal{M} : g_i(y, u, z, \theta) = 0. \; i = 1, \ldots, p.$$

g_i are expression of the variables and their derivatives

Identifiability

Suppose that the structure is identifiable – no two different values of θ can give the same solution set y, u.

Allowed Model Manipulations

Form new model equations by adding, multiplying and differentiating the g_i. New equation sets can thus be formed that have the same solution set.

A General Property – Using Ritt's Algorithm

Convexifying Model Equations

Then by *Ritt's algorithm*, [differential – algebraic manipulations of the set of equations], the identifiable model structure can be transformed to

$$\mathcal{M}^* : \; \phi(y, u) = \theta\psi(y, u)$$

That is, the arbitrary, identifiable structure $\mathcal{M}$ can be convexified to the linear regression $\mathcal{M}^*$. (Cautions: ....)

Algebraic Convexification: Cautions

- If noise is assigned to the outputs y, the resulting linear regression need not be the ML criterion – The resulting LS parameters may be biased.
  - If the noise level is not too big, the bias can be small and provide a sufficiently good initial estimate for the numerical minimization of the ML criterion.
- In problems of practical sizes, the computational complexity of Ritt's algorithm may be forbidding.
  - It is active research area in Mathematics and Computer Science to develop more efficient general tools for symbolic equation manipulations.

Convexity Issues: Conclusions I

- The non-convexity of the criterion in state-of-the-art system identification is a source of concern
- For linear black-box models, the general approximation capability of ARX-models is a common ground for successful initialization of the numerical search for the estimate.
  - This includes the use of subspace methods like N4SID, MOESP, etc
  - It is a remaining unsolved problem to initialize by convex techniques structured linear grey-box models

Convexity Issues: Conclusions II

- Regularized ARX-models offer a fined tuned choice for efficient bias–variance trade-off and form a viable convex alternative to state-of-the-art ML techniques for linear black-box models.
  - This bias–variance tuning is potentially more powerful than by model order selection, since it involves a set of continuous hyper-parameters
  - Need to study good parameterizations of the regularization matrix that allows safe, preferably convex tuning
- Explicit convexification by differential-algebraic techniques is always possible for identifiable model structures. This is (at least) of conceptual interest.
  - Need to follow the computational development in symbolic equation manipulations.

References and Acknowledgments

- The regularization results were based on and inspired by:
  T. Chen, H. Ohlsson and L. Ljung: On the estimation of transfer functions, regularization and Gaussian Processes – Revisited. *Automatica*, Aug 2012.
- Funded by the ERC advanced grant LEARN

European Research Council
Established by the European Commission