Hybrid system identification and Learning theory

Fabien Lauer
https://members.loria.fr/FLauer/
fabien.lauer@loria.fr
joint work with Louis Massucci and Marion Gilson

Université de Lorraine, CNRS, LORIA, France

Workshop conjoint GdR-MACS & COMET-SCA sur l'Automatique et l'IA

Outline

Outline

(a subset of) System identification

Problem statement

- (Discrete-time SISO ARX) dynamical system:

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i$$

$$\text{output at time } i = \text{function of } \overbrace{[\underbrace{y_{i-1}, \ \ldots, y_{i-n_a}}_{\text{past outputs}}, \underbrace{u_i}_{\text{input}}, \underbrace{\ldots, u_{i-n_b}}_{\text{past inputs}}]^T}^{\boldsymbol{x}_i} + \underbrace{\varepsilon_i}_{\text{noise}}$$

- Given a sequence of input–output data $((u_i, y_i))_{1 \le i \le n_0}$, estimate f

(a subset of) System identification

Problem statement

- (Discrete-time SISO ARX) dynamical system:

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i$$

$$\text{output at time } i = \text{function of } \overbrace{[\underbrace{y_{i-1}, \ldots, y_{i-n_a}}_{\text{past outputs}}, \underbrace{u_i}_{\text{input}}, \underbrace{\ldots, u_{i-n_b}}_{\text{past inputs}}]^T}^{\boldsymbol{x}_i} + \underbrace{\varepsilon_i}_{\text{noise}}$$

- Given a sequence of input–output data $((u_i, y_i))_{1 \le i \le n_0}$, estimate f

Literature

- Has focused on *linear* systems for many years
- Mostly based on estimation theory
  - parametric models
  - asymptotic results
  - with assumptions on the true system and the noise ε_i
- Aims at taking into account the sequential/dynamical nature of the data

Machine Learning

Problem statement

- Observed phenomenon:

$$y \approx f(\boldsymbol{x})$$
$$\text{label} \approx \text{function of the input pattern}$$

- Given a training set $((\boldsymbol{x}_i, y_i))_{1 \leq i \leq n}$ of examples, learn a model f that can predict y given $\boldsymbol{x}$

Machine Learning

Problem statement

- Observed phenomenon:

$$y \approx f(\boldsymbol{x})$$
$$\text{label} \approx \text{function of the input pattern}$$

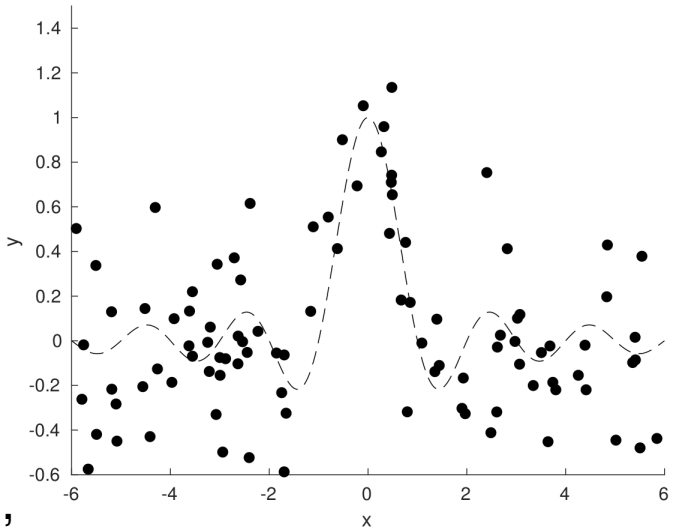
- Given a training set $((\boldsymbol{x}_i, y_i))_{1 \le i \le n}$ of examples, learn a model f that can predict y given $\boldsymbol{x}$

Literature

- Focuses on *nonlinear* models
- Mostly based on statistical learning theory
  - nonparametric models
  - non-asymptotic results
  - distribution-free results (without assumptions on the true relationship or the noise)
- Heavily builds on the assumption of independence of the data
  (via concentration arguments)

Machine Learning

Regression



- Random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$
- Unknown joint distribution P
- Learning:
  from a realization $((\boldsymbol{x}_i, y_i))_{1 \leq i \leq n}$
  of n independent copies $(\boldsymbol{X}_i, Y_i)$ of $(\boldsymbol{X}, Y)$,
  find the function f minimizing the risk

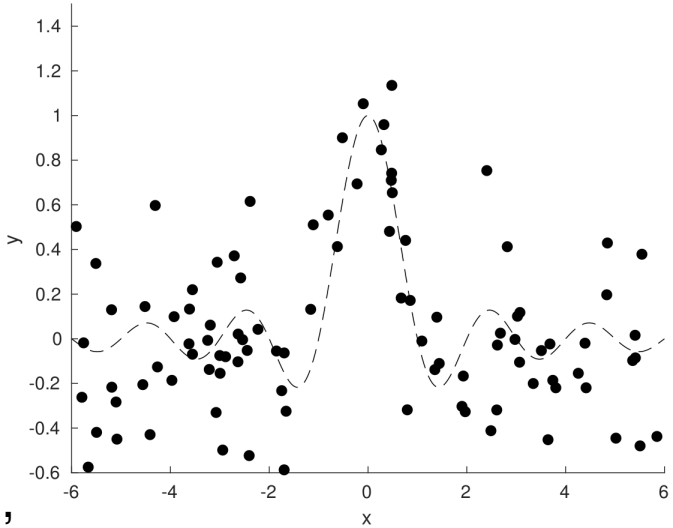
$$L(f) = MSE(f) = \mathbb{E}(Y - f(\boldsymbol{X}))^2$$

- Optimal target function $f^* = \underset{f:\mathcal{X}\to\mathcal{Y}}{\operatorname{argmin}} L(f)$ is

$$\forall \boldsymbol{x} \in \mathcal{X}, \quad f^*(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$$

Machine Learning

Regression



- Random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$
- Unknown joint distribution P
- Learning:
  from a realization $((\boldsymbol{x}_i, y_i))_{1 \leq i \leq n}$
  of n independent copies $(\boldsymbol{X}_i, Y_i)$ of $(\boldsymbol{X}, Y)$,
  find the function f minimizing the risk

$$L(f) = MSE(f) = \mathbb{E}(Y - f(\boldsymbol{X}))^2$$

- Optimal target function $f^* = \underset{f:\mathcal{X}\to\mathcal{Y}}{\operatorname{argmin}} L(f)$ is

$$\forall \boldsymbol{x} \in \mathcal{X}, \quad f^*(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$$

- *but cannot be computed without knowledge of P*
- In practice: minimize the empirical risk estimating $L(f)$ over a predefined function class $\mathcal{F}$:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2$$

Hybrid system identification / Switching regression

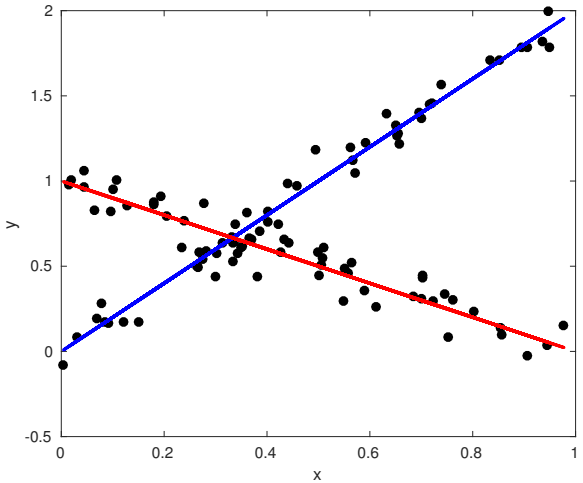
$$y_i = f^*_{q_i}(\boldsymbol{x}_i) + \varepsilon_i$$

$q_i \in [C] = \{1, \ldots, C\}$: mode of point $\boldsymbol{x}_i$

f^*_k: smooth target model for mode k

ε_i: noise term

q_i *evolves independently of* $\boldsymbol{x}_i$



Goal

- Estimate $\{f_k\}_{k=1}^{C}$ and $\{q_i\}_{i=1}^{n}$
- For hybrid systems: $\boldsymbol{x}_i = [y_{i-1}, \ldots, y_{i-n_a}, u_i, \ldots, u_{i-n_b}]^T$

Hybrid system identification / Switching regression

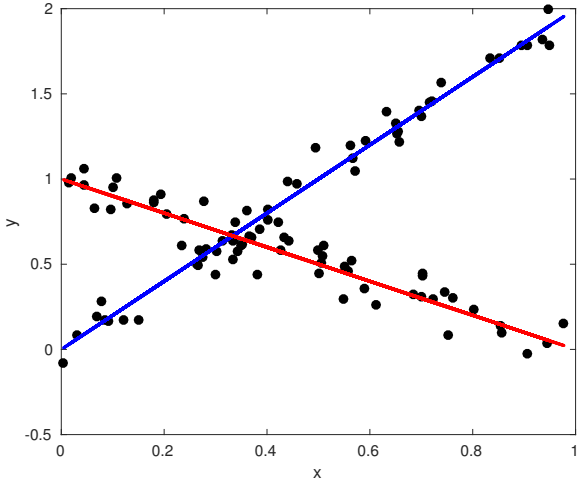
$$y_i = f^*_{q_i}(\boldsymbol{x}_i) + \varepsilon_i$$

$q_i \in [C] = \{1, \ldots, C\}$: mode of point $\boldsymbol{x}_i$

f^*_k: smooth target model for mode k

ε_i: noise term

q_i *evolves independently of* $\boldsymbol{x}_i$



Goal

- Estimate $\{f_k\}_{k=1}^C$ and $\{q_i\}_{i=1}^n$
- For hybrid systems: $\boldsymbol{x}_i = [y_{i-1}, \ldots, y_{i-n_a}, u_i, \ldots, u_{i-n_b}]^T$

Learning problem

- Minimize over $\mathcal{F}^C$, the switching risk of $f = (f_k)_{1 \le k \le C}$:

$$L_{\text{switch}}(f) = \mathbb{E} \min_{k \in [C]} (Y - f_k(\boldsymbol{X}))^2$$

- In practice, minimize the *empirical* switching risk:

$$\min_{f \in \mathcal{F}^C} \frac{1}{n} \sum_{i=1}^{n} \min_{k \in [C]} (y_i - f_k(\boldsymbol{x}_i))^2$$

Hybrid system identification and Machine learning

Main issues and machine learning solutions

- **Optimization**: nonconvex/combinatorial problem (even for *linear* hybrid systems)
  algorithms inspired by clustering methods [Lauer, 2013]
- **Nonlinear hybrid systems**: estimate unknown nonlinearities
  kernel methods and nonparametric models
  [Le et al., 2011, Le et al., 2013]
- **Theoretical guarantees**: statistical accuracy of the estimated model
  risk bounds and statistical learning theory
  [Lauer, 2020, Massucci et al., 2020]
- **Model selection**: estimating the number of modes/components
  structural risk minimization principle [Massucci et al., 2020]

Outline

Statistical learning theory [Vapnik, 1995]

Goal

- Derive statistical guarantees for models learned from data
- Risk bounds (for regression):

$$P\left\{ \forall f \in \mathcal{F}, \quad \underbrace{\mathbb{E}(Y - f(\boldsymbol{X}))^2}_{\text{risk/expected error}} \leq \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\boldsymbol{X}_i))^2}_{\text{empirical risk/measured error}} + \underbrace{\epsilon(n, \mathcal{F}, \delta)}_{\text{confidence interval}} \right\} \geq 1-\delta$$

- Non-asymptotic: holds with a finite number of data n
- Distribution-free: no assumption on P or the optimal f required
- Uniform ($\forall f \in \mathcal{F}$): independent of the algorithm

Using tools such as...

- Concentration inequalities
- Capacity measures of function classes
  covering numbers, Rademacher complexities...
- A few other more specific tricks
  symmetrization, ghost samples,...

Risk bounds for linear and kernel regression

Empirical Rademacher complexity of $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$

$$\hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(\boldsymbol{X}_i) = \mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{f \in \mathcal{F}} \frac{1}{n} \left\langle \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_n \end{bmatrix}, \begin{bmatrix} f(\boldsymbol{X}_1) \\ \vdots \\ f(\boldsymbol{X}_n) \end{bmatrix} \right\rangle$$

$\mathbf{X}_n = (\boldsymbol{X}_i)_{1 \leq i \leq n}$: sequence of n independent copies of $\boldsymbol{X} \in \mathcal{X}$
$\boldsymbol{\sigma}_n = (\sigma_i)_{1 \leq i \leq n}$: sequence of n iid. uniform r.v. in $\{-1, +1\}$

Risk bounds for linear and kernel regression

Empirical Rademacher complexity of $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$

$$\hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(\boldsymbol{X}_i) = \mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{f \in \mathcal{F}} \frac{1}{n} \left\langle \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_n \end{bmatrix}, \begin{bmatrix} f(\boldsymbol{X}_1) \\ \vdots \\ f(\boldsymbol{X}_n) \end{bmatrix} \right\rangle$$

$\mathbf{X}_n = (\boldsymbol{X}_i)_{1 \le i \le n}$: sequence of n independent copies of $\boldsymbol{X} \in \mathcal{X}$
$\boldsymbol{\sigma}_n = (\sigma_i)_{1 \le i \le n}$: sequence of n iid. uniform r.v. in $\{-1, +1\}$

Risk bound for regression

With probability at least $1 - \delta$ (on the random draw of the $(\boldsymbol{X}_i, Y_i)$'s),

$$\forall f \in \mathcal{F}, \quad \mathbb{E}(Y - \bar{f}(\boldsymbol{X}))^2 \le \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{f}(\boldsymbol{X}_i))^2 + 4\hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

$\bar{f} = f$ with saturated output within $\mathcal{Y} = [-1/2, 1/2]$

Risk bounds for linear and kernel regression

Empirical Rademacher complexity of $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$

$$\hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(\boldsymbol{X}_i) = \mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{f \in \mathcal{F}} \frac{1}{n} \left\langle \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_n \end{bmatrix}, \begin{bmatrix} f(\boldsymbol{X}_1) \\ \vdots \\ f(\boldsymbol{X}_n) \end{bmatrix} \right\rangle$$

$\mathbf{X}_n = (\boldsymbol{X}_i)_{1 \le i \le n}$: sequence of n independent copies of $\boldsymbol{X} \in \mathcal{X}$
$\boldsymbol{\sigma}_n = (\sigma_i)_{1 \le i \le n}$: sequence of n iid. uniform r.v. in $\{-1, +1\}$

Risk bound for regression

With probability at least $1 - \delta$ (on the random draw of the $(\boldsymbol{X}_i, Y_i)$'s),

$$\forall f \in \mathcal{F}, \quad \mathbb{E}(Y - \bar{f}(\boldsymbol{X}))^2 \le \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{f}(\boldsymbol{X}_i))^2 + 4\hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

$\bar{f} = f$ with saturated output within $\mathcal{Y} = [-1/2, 1/2]$

For (regularized) linear regression [Bartlett and Mendelson, 2002]

- Model class: $\mathcal{F} = \{f \in \mathbb{R}^{\mathcal{X}} : f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle, \ \|\boldsymbol{w}\| \le \Lambda\}$
- Rademacher complexity: $\hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}) \le \frac{\Lambda \sqrt{\sum_{i=1}^{n} \|\boldsymbol{X}_i\|^2}}{n}$

Risk bounds for linear and kernel regression

Empirical Rademacher complexity of $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$

$$\hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(\boldsymbol{X}_i) = \mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{f \in \mathcal{F}} \frac{1}{n} \left\langle \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_n \end{bmatrix}, \begin{bmatrix} f(\boldsymbol{X}_1) \\ \vdots \\ f(\boldsymbol{X}_n) \end{bmatrix} \right\rangle$$

$\mathbf{X}_n = (\boldsymbol{X}_i)_{1 \le i \le n}$: sequence of n independent copies of $\boldsymbol{X} \in \mathcal{X}$
$\boldsymbol{\sigma}_n = (\sigma_i)_{1 \le i \le n}$: sequence of n iid. uniform r.v. in $\{-1, +1\}$

Risk bound for regression

With probability at least $1 - \delta$ (on the random draw of the $(\boldsymbol{X}_i, Y_i)$'s),

$$\forall f \in \mathcal{F}, \quad \mathbb{E}(Y - \bar{f}(\boldsymbol{X}))^2 \le \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{f}(\boldsymbol{X}_i))^2 + 4\hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

$\bar{f} = f$ with saturated output within $\mathcal{Y} = [-1/2, 1/2]$

For (regularized) kernel regression [Bartlett and Mendelson, 2002]

- Reproducing kernel Hilbert space $\mathcal{H}$ of kernel K:
  $\forall f \in \mathcal{H},\ f(\boldsymbol{x}) = \langle f, K(\boldsymbol{x}, \cdot) \rangle$
- Model class: $\mathcal{F} = \{f \in \mathcal{H} : \|f\| \le \Lambda\}$
- Rademacher complexity: $\hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}) \le \frac{\Lambda \sqrt{\sum_{i=1}^{n} K(\boldsymbol{X}_i, \boldsymbol{X}_i)}}{n}$

Risk bounds for switching regression

Setting

- Assumption: iid training sample $((\boldsymbol{X}_i, Y_i))_{1 \leq i \leq n}$
- C independent submodels: $\mathcal{F} = \mathcal{F}_0^C$
  - linear submodels: $\mathcal{F}_0 = \{f \in \mathbb{R}^{\mathcal{X}} : f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle \,,\ \|\boldsymbol{w}\| \leq \Lambda\}$
  - kernel submodels: $\mathcal{F}_0 = \{f \in \mathcal{H} : \|f\| \leq \Lambda\}$

Theorem [Lauer, 2020]

With probability at least $1 - \delta$, for all $f \in \mathcal{F}$,

$$\mathbb{E} \min_{k \in [C]} (Y - \bar{f}_k(\boldsymbol{X}))^2 \leq \frac{1}{n} \sum_{i=1}^{n} \min_{k \in [C]} (Y_i - \bar{f}_k(\boldsymbol{X}_i))^2 + 4 \frac{C\Lambda \sqrt{\sum_{i=1}^{n} K(\boldsymbol{X}_i, \boldsymbol{X}_i)}}{n} + 3 \sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

with $K(\boldsymbol{X}_i, \boldsymbol{X}_i) = \|\boldsymbol{X}_i\|^2$ for linear submodels

Risk bounds for dynamical system identification

- Data typically comes from a single trajectory (or a few)
- Sequential observations of outputs y_i are not independent
- The main assumption of statistical learning is violated

Risk bounds for dynamical system identification

- Data typically comes from a single trajectory (or a few)
- Sequential observations of outputs y_i are not independent
- The main assumption of statistical learning is violated

Independent block sequence approach [Yu, 1994]

- For β-mixing data sequences

  the dependence between two data points decreases with the time interval between them
- Consider a sequence of well-separated blocks instead of a sequence of consecutive points

  $\Rightarrow$ decrease the dependence between two consecutive objects
- Quantify/control the error made when performing the analysis as if the blocks were independent

  For μ blocks of length a separated by a time steps and a function F of blocks bounded by $\overline{F}$,

$$|\mathbb{E}F(\text{separated block sequence}) - \mathbb{E}F(\text{independent block sequence})| \leq (\mu - 1)\overline{F}\beta(a)$$

Risk bounds for dynamical system identification

- Data typically comes from a single trajectory (or a few)
- Sequential observations of outputs y_i are not independent
- The main assumption of statistical learning is violated

Independent block sequence approach [Yu, 1994]

- For β-mixing data sequences
  the dependence between two data points decreases with the time interval between them
- Consider a sequence of well-separated blocks instead of a sequence of consecutive points
  $\Rightarrow$ decrease the dependence between two consecutive objects
- Quantify/control the error made when performing the analysis as if the blocks were independent
  For μ blocks of length a separated by a time steps and a function F of blocks bounded by $\overline{F}$,

$$|\mathbb{E}F(\text{separated block sequence}) - \mathbb{E}F(\text{independent block sequence})| \leq (\mu - 1)\overline{F}\beta(a)$$

- Risk bound for regression with non-iid. data
  [Mohri and Rostamizadeh, 2009]: With probability at least $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad \mathbb{E}(Y - \bar{f}(\boldsymbol{X}))^2 \leq \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{f}(\boldsymbol{X}_i))^2 + 4\hat{\mathcal{R}}_{\mathbf{X}_\mu}(\mathcal{F}) + 3\sqrt{\frac{\log\frac{4}{\delta - 4(\mu-1)\beta(a)}}{2\mu}}$$

The number of blocks μ replaces the number of data n

Outline

Model selection for hybrid system identification

Estimating the number of modes/components C

- A major issue in hybrid system identification
- Typically done by estimating a model for all possible numbers C

Structural risk minimization principle [Vapnik, 1995]

- For each class of a sequence of model classes $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_m$
- Find the model within that class that minimizes the empirical risk
- Select the model that minimizes an upper bound on the risk

Model selection for hybrid system identification

Estimating the number of modes/components C

- A major issue in hybrid system identification
- Typically done by estimating a model for all possible numbers C

Structural risk minimization principle [Vapnik, 1995]

- For each class of a sequence of model classes $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_m$
- Find the model within that class that minimizes the empirical risk
- Select the model that minimizes an upper bound on the risk

SRM-based algorithm [Massucci et al., 2020]

For switched linear systems:

$$\hat{C} = \underset{C \in \{1,\dots,C_{max}\}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \min_{k \in [C]} (y_i - \bar{f}_k(\boldsymbol{x}_i))^2 + \frac{4C\tilde{\Lambda}\sqrt{\sum_{i=1}^{\mu} \|\boldsymbol{x}_{2a(i-1)+1}\|^2}}{\mu}$$
$$+3\sqrt{\frac{\log(C_{max}K) + \log \frac{4}{\delta - 4C_{max}K(\mu-1)\beta(a)}}{2\mu}}$$

Good news: the optimization of the bound wrt. C does not require the computation/knowledge of the mixing coefficient $\beta(a)$

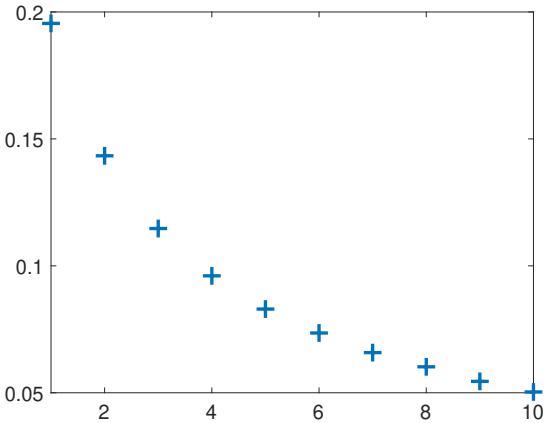
A numerical example

- Switched ARX system with $C = 3$ modes of order $n_a = n_b = 2$
- $n = 4.10^5$ data
- Gaussian noise ε, SNR$=$ 10 dB

A numerical example

- Switched ARX system with $C = 3$ modes of order $n_a = n_b = 2$
- $n = 4.10^5$ data
- Gaussian noise ε, SNR$=$ 10 dB

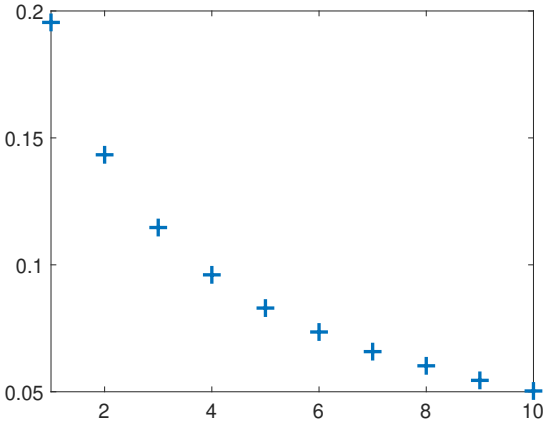
Empirical risk vs. C



A numerical example

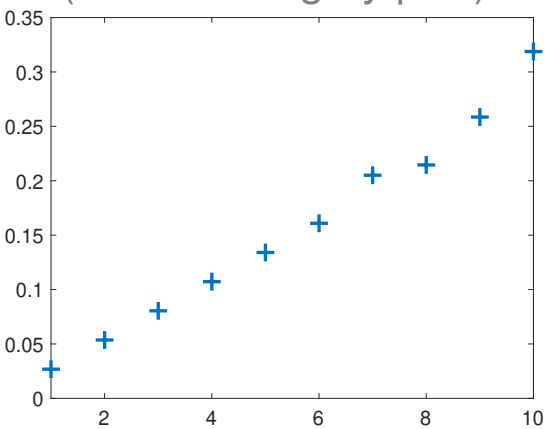
- Switched ARX system with $C = 3$ modes of order $n_a = n_b = 2$
- $n = 4.10^5$ data
- Gaussian noise ε, SNR$=$ 10 dB

Empirical risk vs. C



Confidence interval

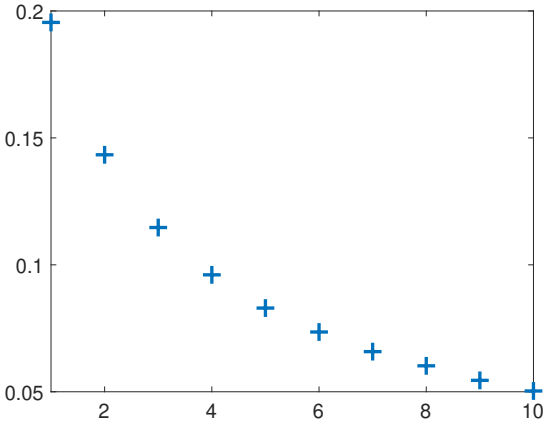
(without the grey part)



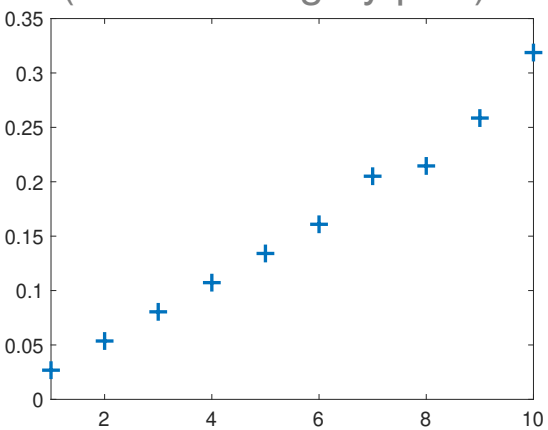
A numerical example

- Switched ARX system with $C = 3$ modes of order $n_a = n_b = 2$
- $n = 4.10^5$ data
- Gaussian noise ε, SNR$=$ 10 dB

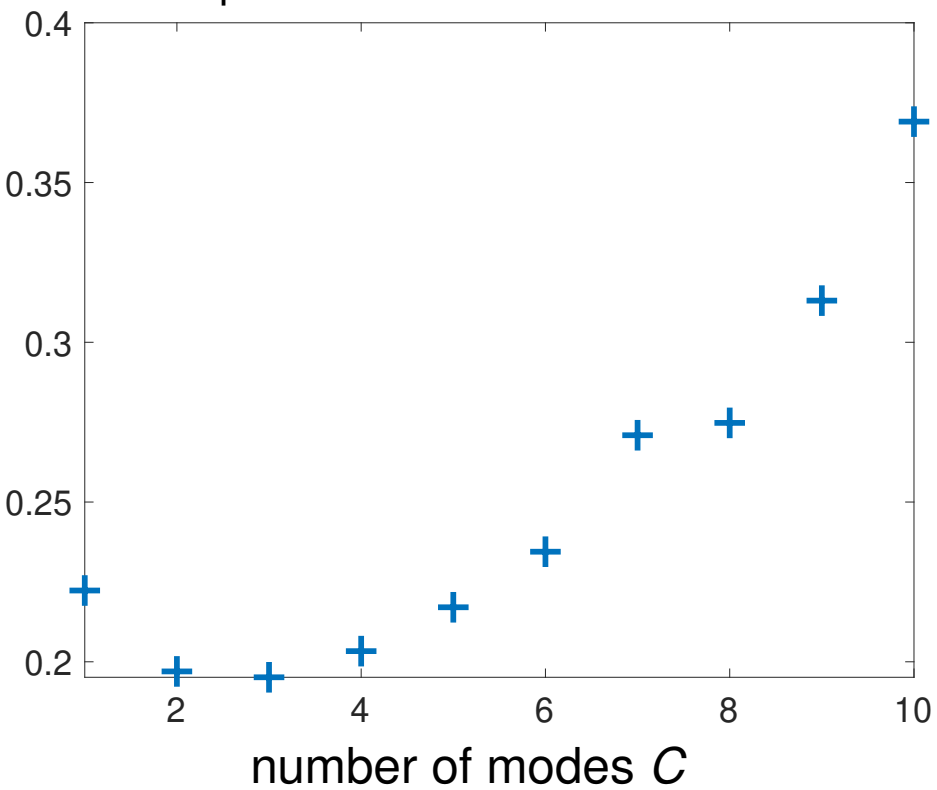
Empirical risk vs. C



Confidence interval
(without the grey part)



Risk bound (without the grey part)
$=$ Empirical risk + confidence interval



Outline

Conclusions

Hybrid system identification can benefit from learning theory

- For statistical guarantees
- For model selection (estimating the number of modes)

Machine learning can also benefit from hybrid system identification

- Clustering problems enjoy similar results
- Can lead to model selection methods to tune the number and dimensions of subspaces in subspace clustering

Conclusions

Hybrid system identification can benefit from learning theory

- For statistical guarantees
- For model selection (estimating the number of modes)

Machine learning can also benefit from hybrid system identification

- Clustering problems enjoy similar results
- Can lead to model selection methods to tune the number and dimensions of subspaces in subspace clustering

Future work

- Estimate the mixing coefficient $\beta(a)$
- Tighter bounds for other forms of regularization [Massucci et al., 2021]
- Learning without mixing [Simchowitz et al., 2018]

References I

Bartlett, P. and Mendelson, S. (2002).

Rademacher and Gaussian complexities: Risk bounds and structural results.

Journal of Machine Learning Research, 3:463–482.

Lauer, F. (2013).

Estimating the probability of success of a simple algorithm for switched linear regression.

Nonlinear Analysis: Hybrid Systems, 8:31–47.

Lauer, F. (2020).

Error bounds for piecewise smooth and switching regression.

IEEE Transactions on Neural Networks and Learning Systems, 31(4):1183–1195.

Le, V. L., Bloch, G., and Lauer, F. (2011).

Reduced-size kernel models for nonlinear hybrid system identification.

IEEE Transactions on Neural Networks, 22(12):2398–2405.

Le, V. L., Lauer, F., Bako, L., and Bloch, G. (2013).

Learning nonlinear hybrid systems: from sparse optimization to support vector regression.

In *Proc. of the 16th ACM Int. Conf. on Hybrid Systems: Computation and Control (HSCC), Philadelphia, PA, USA*, pages 33–42.

References II

Massucci, L., Lauer, F., and Gilson, M. (2020).

Structural risk minimization for switched system identification.

In *Proc. of the 59th Int. Conf. on Decision and Control (CDC)*.

Massucci, L., Lauer, F., and Gilson, M. (2021).

Regularized switched system identification: a statistical learning perspective.

In *Proc. of the 7th IFAC Conf. on Analysis and Design of Hybrid Systems (ADHS)*.

Mohri, M. and Rostamizadeh, A. (2009).

Rademacher complexity bounds for non-iid processes.

In *Advances in Neural Information Processing Systems*, pages 1097–1104.

Simchowitz, M., Mania, H., Tu, S., Jordan, M., and Recht, B. (2018).

Learning without mixing: Towards a sharp analysis of linear system identification.

In *Proc. of the 31st Annual Conference on Learning Theory (COLT)*, pages 439–473.

Vapnik, V. (1995).

The Nature of Statistical Learning Theory.

Springer-Verlag.

Yu, B. (1994).

Rates of convergence for empirical processes of stationary mixing sequences.

The Annals of Probability, 22(1):94–116.