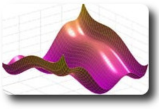
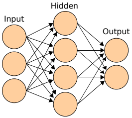
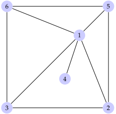
Polynomial optimization methods for machine learning

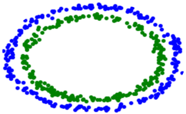
Victor Magron, CNRS LAAS
Chair of Polynomial optimization for Machine Learning

Spot ANITI Days, 12 November 2022

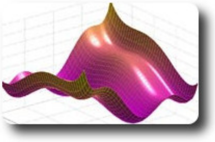


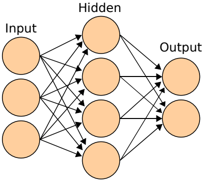




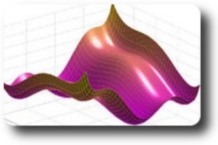


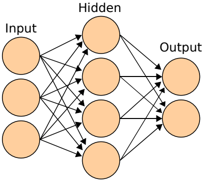
Context: what and who?





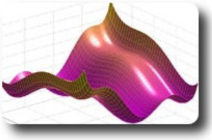
Context: what and who?

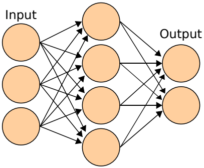




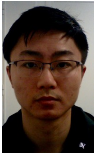
⇝ use **polynomial optimization** (POP) for machine learning

Context: what and who?





⇝ use **polynomial optimization** (POP) for machine learning

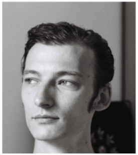


Tong Chen

Milan Korda

Jean-Bernard Lasserre

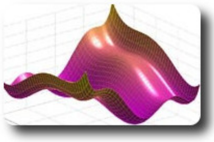
Alexey Lazarev

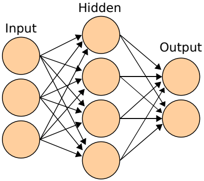


Victor Magron

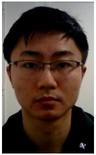
Edouard Pauwels

Context: what and who?





⇝ use **polynomial optimization** (POP) for machine learning

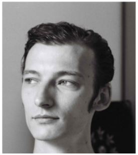


Tong Chen 2019

Milan Korda

Jean-Bernard Lasserre

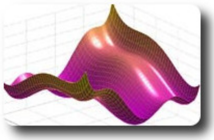
Alexey Lazarev

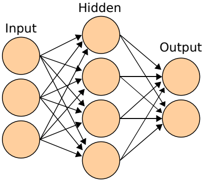


Victor Magron

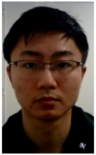
Edouard Pauwels

Context: what and who?





⇝ use **polynomial optimization** (POP) for machine learning

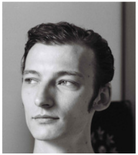


Tong Chen 2019

Milan Korda

Jean-Bernard Lasserre

Alexey Lazarev 2021



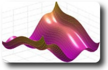
Victor Magron

Edouard Pauwels

The POP hammer

NP-hard NON CONVEX Problem $f_{\min} = \inf f(\mathbf{x})$

Theory

(Primal)		(Dual)
$\inf \int f \, d\mu$		$\sup \; b$
with μ proba $\Rightarrow$	**infinite-dim**	$\Leftarrow$ with $f - b \geqslant 0$

The POP hammer

NP-hard NON CONVEX Problem $f_{\min} = \inf f(\mathbf{x})$

Practice

(Primal **Relaxation**)  (Dual **Strengthening**)

moments $\int \mathbf{x}^\alpha \, d\mu$ $f - b =$ **sum of squares**

finite number $\Rightarrow$ **finite-dim** $\Leftarrow$ **fixed** degree

[Lasserre '01] HIERARCHY of **CONVEX Problems** $\uparrow f_{\min}$
Based on representation of positive polynomials [Putinar '93]

The POP hammer

NP-hard NON CONVEX Problem $f_{\min} = \inf f(\mathbf{x})$

Practice

(Primal **Relaxation**)

moments $\int \mathbf{x}^\alpha \, d\mu$

(Dual **Strengthening**)

$f - b =$ **sum of squares**

finite number $\Rightarrow$ **finite-dim** $\Leftarrow$ **fixed** degree

[Lasserre '01] Hierarchy of **CONVEX Problems** $\uparrow f_{\min}$

Based on representation of positive polynomials [Putinar '93]

Attracted a lot of attention in optimization, applied mathematics, quantum computing, engineering, theoretical computer science

The POP hammer

NP-hard NON CONVEX Problem $f_{\min} = \inf f(\mathbf{x})$

Practice

(Primal **Relaxation**)

moments $\int \mathbf{x}^\alpha \, d\mu$

(Dual **Strengthening**)

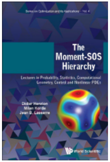
$f - b =$ **sum of squares**

finite number $\Rightarrow$ **finite-dim** $\Leftarrow$ **fixed** degree

[Lasserre '01] Hierarchy of **CONVEX Problems** $\uparrow f_{\min}$

Based on representation of positive polynomials [Putinar '93]

Attracted a lot of attention in optimization, applied mathematics, quantum computing, engineering, theoretical computer science



Emerging applications: quantum information theory, deep learning & power systems

Sparse hierarchies

Structure exploitation POP $\inf f(\mathbf{x})$ with "SPARSE" f

Sparse hierarchies

Structure exploitation POP inf $f(\mathbf{x})$ with "SPARSE" f

Correlative sparsity: few variable products in f

Sparse hierarchies

Structure exploitation POP $\inf f(\mathbf{x})$ with "SPARSE" f

Correlative sparsity: few variable products in f

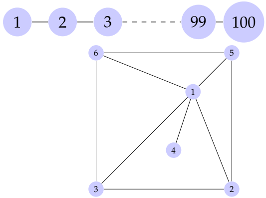
$\rightsquigarrow f = x_1x_2 + x_2x_3 + \cdots + x_{99}x_{100}$

Sparse hierarchies

Structure exploitation POP $\inf f(\mathbf{x})$ with "SPARSE" f

Correlative sparsity: few variable products in f

$\rightsquigarrow f = x_1 x_2 + x_2 x_3 + \cdots + x_{99} x_{100}$

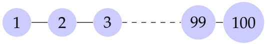


Sparse hierarchies

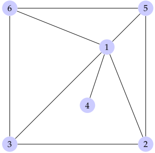
Structure exploitation POP $\inf f(\mathbf{x})$ with "SPARSE" f

Correlative sparsity: few variable products in f

$\rightsquigarrow f = x_1 x_2 + x_2 x_3 + \cdots + x_{99} x_{100}$



Term sparsity: few terms in f



Sparse hierarchies

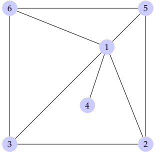
Structure exploitation POP $\inf f(\mathbf{x})$ with "SPARSE" f

Correlative sparsity: few variable products in f

$\rightsquigarrow f = x_1 x_2 + x_2 x_3 + \cdots + x_{99} x_{100}$

Term sparsity: few terms in f

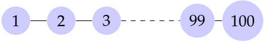
$\rightsquigarrow f = x_1^{99} x_2 + x_1 x_2^{100}$

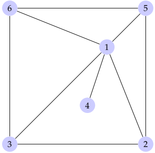


Sparse hierarchies

Structure exploitation POP $\inf f(\mathbf{x})$ with "SPARSE" f

Correlative sparsity: few variable products in f
$\rightsquigarrow f = x_1 x_2 + x_2 x_3 + \cdots + x_{99} x_{100}$

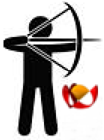




Term sparsity: few terms in f
$\rightsquigarrow f = x_1^{99} x_2 + x_1 x_2^{100}$

PERFORMANCE

VS



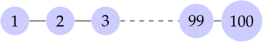
ACCURACY

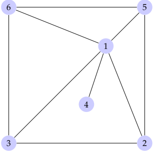
Sparse hierarchies

Structure exploitation POP $\inf f(\mathbf{x})$ with "SPARSE" f

Correlative sparsity: few variable products in f

$\rightsquigarrow f = x_1x_2 + x_2x_3 + \cdots + x_{99}x_{100}$



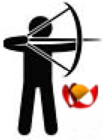


Term sparsity: few terms in f

$\rightsquigarrow f = x_1^{99}x_2 + x_1x_2^{100}$

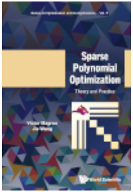
PERFORMANCE

VS



ACCURACY

Tons of applications: roundoff error bounds, entanglement, optimal power-flow, analysis of dynamical systems



Koopman operators

Koopman operators for nonlinear dynamical system $\mathbf{x}^+ = f(\mathbf{x})$

Koopman operators

Koopman operators for nonlinear dynamical system $\mathbf{x}^+ = f(\mathbf{x})$

Represented by an ∞-dim **linear** operator

Koopman operators

Koopman operators for nonlinear dynamical system $\mathbf{x}^+ = f(\mathbf{x})$

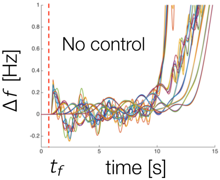
Represented by an ∞-dim **linear** operator
💡 truncate its spectrum $\implies$ **finite-dim linear** programming

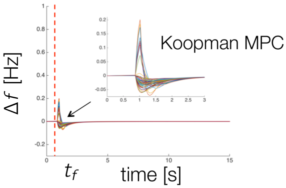
Koopman operators

Koopman operators for nonlinear dynamical system $\mathbf{x}^+ = f(\mathbf{x})$

Represented by an ∞-dim **linear** operator
truncate its spectrum $\implies$ **finite-dim linear** programming

Applications to model predictive control





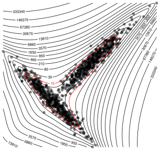
Christoffel-Darboux kernels

Old tool well-known in approximation theory & orthogonal polynomials

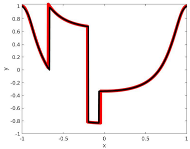
Christoffel-Darboux kernels

Old tool well-known in approximation theory & orthogonal polynomials

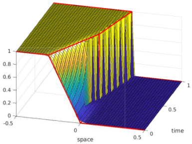
Outlier detection



Interpolation



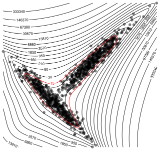
Recovery



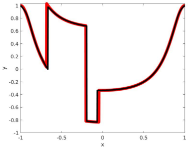
Christoffel-Darboux kernels

Old tool well-known in approximation theory & orthogonal polynomials

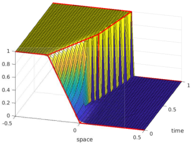
Outlier detection



Interpolation



Recovery

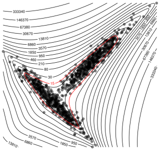


💡 Defined easily from the input data

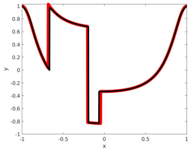
Christoffel-Darboux kernels

Old tool well-known in approximation theory & orthogonal polynomials

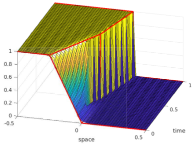
Outlier detection



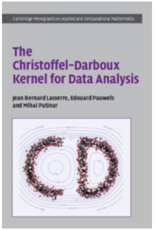
Interpolation



Recovery



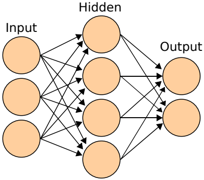
Defined easily from the input data



Chair activities

ROBUSTNESS CERTIFICATION OF NEURAL NETWORKS

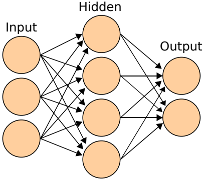
💡 sparse polynomial optimization



Chair activities

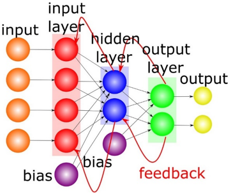
Robustness certification of neural networks

💡 sparse polynomial optimization

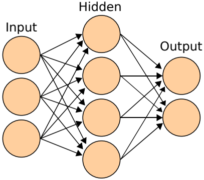


Stability analysis of recurrent networks

💡 copositive programming/integral constraints, Koopman

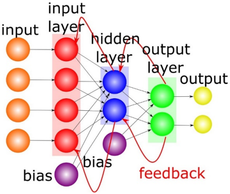


Chair activities



Robustness certification of neural networks

💡 sparse polynomial optimization



Stability analysis of recurrent networks

💡 copositive programming/integral constraints, Koopman

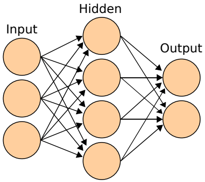
Training/classification

💡 Christoffel-Darboux kernels

Chair activities

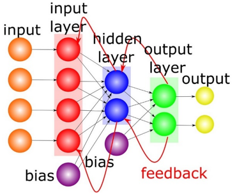
ROBUSTNESS CERTIFICATION OF NEURAL NETWORKS

💡 sparse polynomial optimization



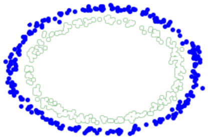
STABILITY ANALYSIS OF RECURRENT NETWORKS

💡 copositive programming/integral constraints, Koopman



TRAINING/CLASSIFICATION

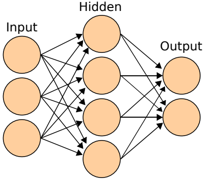
💡 Christoffel-Darboux kernels



Chair activities

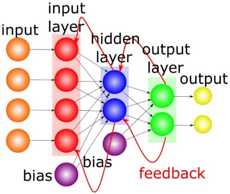
Robustness certification of neural networks

💡 sparse polynomial optimization



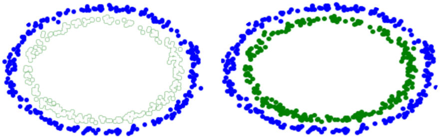
Stability analysis of recurrent networks

💡 copositive programming/integral constraints, Koopman



Training/classification

💡 Christoffel-Darboux kernels



Zoom: robustness of NN [PhD Chen '19-22]

[SIAM News March '21]

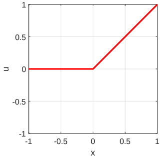
"Yet DL has an Achilles' heel. Current implementations can be highly unstable, meaning that a certain small perturbation to the input of a trained neural network can cause substantial change in its output. This phenomenon is both a nuisance and a major concern for the safety and robustness of DL-based systems in critical applications—like healthcare—where reliable computations are essential"

Zoom: robustness of NN [PhD Chen '19-22]

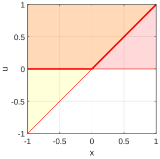
$$\mathbf{x}_0 \in \mathbb{R}^p \longrightarrow \mathbf{z}_0 \in \mathbb{R}^p \longrightarrow \mathbf{z}_1 \in \mathbb{R}^{p_1} \longrightarrow \cdots \longrightarrow \mathbf{z}_m \in \mathbb{R}^{p_m}$$

$$\mathbf{z}_i = \mathbf{A}_i \mathrm{ReLU}(\mathbf{z}_{i-1}) + \mathbf{b}_i$$

ReLU (left) & its "semialgebraicity" (right)



$u = \max\{x, 0\}$



$u(u - x) = 0, u \geq x, u \geq 0$

Zoom: robustness of NN [PhD Chen '19-22]

"Direct" certification of a classifier with 1 hidden layer

$$\max_{\mathbf{x},\mathbf{z}} \quad (\mathbf{C}^{i,:} - \mathbf{C}^{k,:})\mathbf{z}$$

$$\text{s.t.} \quad \begin{cases} \mathbf{z} = \mathrm{ReLU}(\mathbf{A}\mathbf{x} + \mathbf{b}) \\ ||\mathbf{x} - \mathbf{x}_0|| \leq \epsilon \end{cases}$$

Zoom: robustness of NN [PhD Chen '19-22]

- "Direct" certification of a classifier with 1 hidden layer

$$\max_{\mathbf{x},\mathbf{z}} \quad (\mathbf{C}^{i,:} - \mathbf{C}^{k,:})\mathbf{z}$$
$$\text{s.t.} \quad \begin{cases} \mathbf{z} = \mathrm{ReLU}(\mathbf{A}\mathbf{x} + \mathbf{b}) \\ ||\mathbf{x} - \mathbf{x}_0|| \leq \epsilon \end{cases}$$

- Monotone equilibrium networks [Winston Kolter '20]

$$\mathbf{z} = \mathrm{ReLU}(\mathbf{A}\mathbf{x} + \mathbf{b}) \to \mathbf{z} = \mathrm{ReLU}(\mathbf{W}\mathbf{z} + \mathbf{A}\mathbf{x} + \mathbf{b})$$

Zoom: robustness of NN [PhD Chen '19-22]

- "Direct" certification of a classifier with 1 hidden layer

$$\max_{\mathbf{x},\mathbf{z}} \quad (\mathbf{C}^{i,:} - \mathbf{C}^{k,:})\mathbf{z}$$
$$\text{s.t.} \quad \begin{cases} \mathbf{z} = \mathrm{ReLU}(\mathbf{A}\mathbf{x} + \mathbf{b}) \\ ||\mathbf{x} - \mathbf{x}_0|| \leq \epsilon \end{cases}$$

- Monotone equilibrium networks [Winston Kolter '20]

$$\mathbf{z} = \mathrm{ReLU}(\mathbf{A}\mathbf{x} + \mathbf{b}) \to \mathbf{z} = \mathrm{ReLU}(\mathbf{W}\mathbf{z} + \mathbf{A}\mathbf{x} + \mathbf{b})$$

- "Indirect" with Lipschitz constant/ellipsoid approximation

Zoom: robustness of NN [PhD Chen '19-22]

- "Direct" certification of a classifier with 1 hidden layer

$$\max_{\mathbf{x},\mathbf{z}} \quad (\mathbf{C}^{i,:} - \mathbf{C}^{k,:})\mathbf{z}$$
$$\text{s.t.} \quad \begin{cases} \mathbf{z} = \mathrm{ReLU}(\mathbf{A}\mathbf{x} + \mathbf{b}) \\ ||\mathbf{x} - \mathbf{x}_0|| \leq \epsilon \end{cases}$$

- Monotone equilibrium networks [Winston Kolter '20]

$$\mathbf{z} = \mathrm{ReLU}(\mathbf{A}\mathbf{x} + \mathbf{b}) \to \mathbf{z} = \mathrm{ReLU}(\mathbf{W}\mathbf{z} + \mathbf{A}\mathbf{x} + \mathbf{b})$$

- "Indirect" with Lipschitz constant/ellipsoid approximation

- Go between 1ST & 2ND stair in SPARSE hierarchy

Collaborations & output

Polynomial optimization for ML

POP ANITI
MAC control
DISCO chair of Travé-Massuyès
UT3 Pauwels

TU Chemnitz formal verif
CREATE project Descartes
UCSB Koopman
CTU Prague control
CWI ideal sparsity
Kyushu University RNN

Collaborations & output

Polynomial optimization for ML

POP ANITI
MAC control
DISCO chair of Travé-Massuyès
UT3 Pauwels

TU Chemnitz formal verif
CREATE project Descartes
UCSB Koopman
CTU Prague control
CWI ideal sparsity
Kyushu University RNN

Julia packages **github:InterRelax** and **github:TSSOS**

Collaborations & output

Polynomial optimization for ML

- **POP** ANITI
- **MAC** control
- **DISCO** chair of Travé-Massuyès
- **UT3** Pauwels

- **TU Chemnitz** formal verif
- **CREATE** project Descartes
- **UCSB** Koopman
- **CTU Prague** control
- **CWI** ideal sparsity
- **Kyushu University** RNN

Julia packages **github:InterRelax** and **github:TSSOS**

$\sim$ 40 publications including 3 books, 2 NeurIPS, 20 journals