DERIVABLE SCIENTIFIC DISCOVERY

CORNELIO CRISTINA - SAMSUNG AI

IN COLLABORATION WITH IBM RESEARCH: SANJEEB DASH, VERNON AUSTEL, TYLER R. JOSEPHSON, JOAO GONCALVES, KENNETH CLARKSON, NIMROD MEGIDDO, BACHIR EL KHADIR, LIOR HORESH

COLLOQUIUM PR[AI]RIE - OCTOBER 19, 2023
GOAL:
Discovering meaningful laws of nature in symbolic form from experimental data

Extraction of formulas that fit the data:
• NN and statistical regression:
  • good for discovery of patterns and relations in data
  • drawback: “black-box” models
• Standard regression:
  • the functional form is given, discovery = parameter fitting
• Symbolic regression:
  • the functional form is not given but is instead composed from the data
  • models are more “interpretable” and require less data

Discovery of scientifically meaningful formulas:
• Many expressions can be extracted for a given dataset, but not all are consistent with the known background theory
• Models that are derivable, and not merely empirically accurate, are appealing because they are arguably correct, predictive, and insightful
**IDEA**: unification of explicit symbolic model extraction from numerical data with formal reasoning

- **Verification capabilities**: providing a formal proof of the derivability of a formula produced from the data
- **When not derivable**: providing measures that indicate how close a formula is to a derivable formula.
SCIENTIFIC METHODS

Classic Scientific Discovery

ML/Regression

Al-Descartes

Hypothesis Generation from Theory

Testing on Data

New Formula

Hypothesis Generation from Data

New Formula

Hypothesis Generation from Data

Testing on Theory

New Formula
Given observational data, associated with a process or phenomenon, the goal is to discover an interpretable, universal, mathematical model in a symbolic form.
System overview

**INPUT: 4-tuples \( \langle B, C, D, M \rangle \)**

- **Background Knowledge \( B \):** domain-specific axioms (logic formulae)
  - Completeness assumption: \( B \) contains all the axioms necessary to comprehensively explain the phenomena under consideration
  - Consistency assumption: the axioms do not contradict one another
- **A Hypothesis Class \( C \):**
  - Grammar (set of admissible symbolic models)
  - Invariance constraints (e.g., \( X + Y \) is equivalent to \( Y + X \))
  - Constraints on the functional form (e.g., monotonicity)
- **Data \( D \):** a set of \( n \) examples
- **Modeler Preferences \( M \):** a set of numerical parameters (e.g., error bounds on accuracy)
SYMBOLIC REGRESSION

Define grammar so that every meaningful mathematical expression is a sentence of a formal language comprising:

- **Operators**
- **Variables**
- **Constants**

The grammar is encoded as a set of (decision) variables and constraints

Free-form search in the space of the sentences for ones that honors the data while minimizing expression complexity
SYMBOLIC REGRESSION

The optimization problem can be articulated at a high level as:

\[
\min_s \quad D(s(x), y) \quad \begin{array}{c}
\text{fidelity} \\
C(s) \leq \tau \quad \text{complexity} \\
\text{parse}(x) \in \mathcal{T} \quad \text{structural (grammar)} \\
s \in \Sigma^* \quad \text{symbols (primitive) choice} \\
s \in I \quad \text{invariances} \\
N(s(x)) \leq \delta \quad \text{non-linear numerical expression}
\end{array}
\]

Where:
- \( D \) - error model
- \( x \in \mathcal{X} \) - set of training datum
- \( y \in \mathcal{Y} \) - set of targets associated with \( x \)
- \( \Sigma \) - set of admissible symbols
- \( \Sigma^* \) - set of words in the language
- \( \mathcal{T} \) - a proper tree structure
- \( s \in \Sigma^* \) and \( \text{parse}(x) \in \mathcal{T} \) implies that \( s \in \mathcal{L}(\mathcal{T}, \Sigma) \) (belongs to the formal language)
- \( C \) - a measure of complexity
- \( \mathcal{N} \) - numerical function
- \( I \) - set of invariants
REASONING CAPABILITIES

1 – Derivability
Verify a formula from a set of axioms defining the background theory

2 – Reasoning measures
Relative/Absolute error between a formula (induced from data) and the variable of interest which represents a derivable formula deducible from the axioms

3 – Counterexample generation
Generation of new points that violates the current candidate formula, starting from the axioms

4 – Constraints pre-processing
Checking which of the candidate formulas support a set of constraints on the functional form:
• Monotonicity condition
• Conditions at the limit
• etc.
Two types of derivability

Direct derivability:

\[(C \land A) \rightarrow f\]

Existential derivability:

\[\exists c_1 \ldots c_n (C \land A) \rightarrow (f' \land C')\]

• f is replaced by f’ by introducing new existentially quantified variables for each numerical element in f.

Example:

**Formula extracted from data**

\[f = p/(0.709 \cdot p + 0.157)\]

**Formula to prove – Direct derivability**

\[(C \land A) \rightarrow p/(0.709 \cdot p + 0.157)\]

**Formula to prove – Existential derivability**

\[\exists c_1 c_2 (C \land A) \rightarrow p/(c_1 \cdot p + c_2)\]

Where:

• C = constraints
• A = axioms,
• \(C \cup A\) = background theory
• f = the formula we wish to prove
REASONING ERRORS

Distance between:

- A formula generated from the numerical data
- The derivable formula that is implicitly defined by the axiom set

Note: The derivable formula is defined by the variable of interest is not given explicitly, but only implicitly defined in the background theory.
**Pointwise reasoning error**

- measured by the $l_2$ or $l_\infty$ norm (holds for other norms as well)
- applied to the differences between the values of the numerically-derived formula and a derivable formula at the points in the dataset.

\[ \beta_2^i = \sqrt{\sum_{i=1}^{m} \left( \frac{f(X^i) - f_{\mathcal{B}}(X^i)}{f_{\mathcal{B}}(X^i)} \right)^2} \]
Generalization reasoning error

- Consider not only the specific datapoints but the interval where specific data points lie in
- Evaluate how much a formula generalizes between the data points

\[ \beta^\epsilon_\infty = \max_{1 \leq i \leq m} \left\{ \frac{|f(X^i) - f(\mathcal{B}(X^i))|}{|f(\mathcal{B}(X^i))|} \right\} \]
Variable dependence

Extension of the intervals beyond the order of magnitude of the data points in the dataset.

- Check if a formula generalizes even outside the space defined by the dataset

**Done by:**

- Increasing (or diminishing) the interval end (or start) point by one order of magnitude for a variable at a time.
SHOW CASES

Challenges:

• **Real data (noise)**
• **Few data points (~10 points)**
• **Kepler**: The masses involved are of very different magnitudes.
• **Langmuir**: Background theory contains material-dependent coefficients
• **Einstein**: Different background theories: Newtonian and relativistic
Symbolic regression

**BARON** as MINLP solver
- Supported operators: 
  - $+, -, \times, /, \exp, \log$
- Supported L-tree depth = 4 (~7 parsing tree)

**Reasoning**

**KeYmaera** as reasoning tool
- ATP for hybrid systems, which combines different types of reasoning: deductive, real-algebraic, and computer-algebraic reasoning.
- has an underlying CAD system

**Mathematica** for certain types of analysis of symbolic expressions
- e.g. constraints checking
Kepler’s law captures the relationship between the distance between two bodies and their orbital periods.

\[ p = \sqrt{\frac{4\pi^2d^3}{G(m_1 + m_2)}} \]

- \( p \) is the orbital period;
- \( G \) is the gravitational constant;
- \( m_1 \) and \( m_2 \) are the masses of the two bodies (e.g., the sun and a planet in the solar system).
# Kepler's Third Law of Planetary Motion

<table>
<thead>
<tr>
<th></th>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Solar system</strong></td>
<td>$\sqrt{0.1319d^3}$</td>
</tr>
<tr>
<td></td>
<td>$(0.03765d^3 + d^2)/(2 + d)$</td>
</tr>
<tr>
<td></td>
<td>$\sqrt{0.1316(d^3 + d)}$</td>
</tr>
<tr>
<td><strong>Exoplanets</strong></td>
<td>$\sqrt{0.1319d^3/m_1}$</td>
</tr>
<tr>
<td></td>
<td>$\sqrt{m_2^2m_3^3/d + 0.1319d^3/m_1}$</td>
</tr>
<tr>
<td></td>
<td>$\sqrt{(1 - 0.7362m_3)d^3/2}$</td>
</tr>
<tr>
<td><strong>Binary stars</strong></td>
<td>$1/(d^2m_1^2) + 1/(dm_2^2) - m_1^2m_2^2 + \sqrt{0.4787d^3/m_2 + d^2m_2^2}$</td>
</tr>
<tr>
<td></td>
<td>$(\sqrt{d^3 + m_1^2m_2}/\sqrt{d})/\sqrt{m_1 + m_2}$</td>
</tr>
<tr>
<td></td>
<td>$\sqrt{d^3}/(0.9967m_1 + m_2)$</td>
</tr>
</tbody>
</table>
KEPLER’S THIRD LAW OF PLANETARY MOTION

Background theory

K1. center of mass definition
K1. $m_1 \cdot d_1 = m_2 \cdot d_2$

K2. distance between bodies
K2. $d = d_1 + d_2$

K3. gravitational force
K3. $F_g = \frac{Gm_1 m_2}{d^2}$

K4. centrifugal force
K4. $F_c = m_2 d_2 w^2$

K5. force balance
K5. $F_g = F_c$

K6. period definition
K6. $p = \frac{2\pi}{w}$

K7. non-negativity constraints
K7. $m_1 > 0, m_2 > 0, p > 0, d_1 > 0, d_2 > 0$. 
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Candidate formula</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p =$</td>
<td>numerical error</td>
<td>$e^2_2$</td>
<td>$e^\infty_2$</td>
<td>$\beta^r_2$</td>
<td>$\beta^r_\infty$</td>
<td>$\text{gen. reas. error } \beta^r_{\infty,S}$</td>
<td>dependencies</td>
<td>$m_1$</td>
<td>$m_2$</td>
</tr>
<tr>
<td>solar</td>
<td>$\sqrt{0.1319 \cdot d^3}$</td>
<td>.01291</td>
<td>.006412</td>
<td>.0146</td>
<td>.0052</td>
<td>.0052</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\sqrt{0.1316 \cdot (d^3 + d)}$</td>
<td>1.9348</td>
<td>1.7498</td>
<td>1.9385</td>
<td>1.7533</td>
<td>1.7559</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$(0.03765 d^3 + d^2) / (2 + d)$</td>
<td>.3102</td>
<td>.2766</td>
<td>.3095</td>
<td>.2758</td>
<td>.2758</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>exoplanet</td>
<td>$\sqrt{0.1319d^3/m_1}$</td>
<td>.08446</td>
<td>.08192</td>
<td>.02310</td>
<td>.0052</td>
<td>.0052</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\sqrt{m_1 m_2^2 / d + 0.1319 d^3 / m_1}$</td>
<td>.1988</td>
<td>.1636</td>
<td>.1320</td>
<td>.1097</td>
<td>$&gt; 550$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$\sqrt{(1 - .7362 m_1) d^3 / 2}$</td>
<td>1.2246</td>
<td>.4697</td>
<td>1.2418</td>
<td>.4686</td>
<td>.4686</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>binary stars</td>
<td>$1 / (d^2 m_1^2) + 1 / (d m_2^2) - m_1^3 m_2^2 + \sqrt{.4787 d^3 / m_2 + d^2 m_2^2}$</td>
<td>.002291</td>
<td>.001467</td>
<td>.0059</td>
<td>.0050</td>
<td>timeout</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$(\sqrt{d^3 + m_1^3 m_2^2} / \sqrt{d}) / \sqrt{m_1 + m_2}$</td>
<td>.003221</td>
<td>.003071</td>
<td>.0038</td>
<td>.0031</td>
<td>timeout</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$\sqrt{d^3 / (0.9967 m_1 + m_2)}$</td>
<td>.005815</td>
<td>.005337</td>
<td>.0014</td>
<td>.0008</td>
<td>.0020</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
The Langmuir adsorption equation describes a chemical process in which gas molecules contact a surface, and relates the loading on the surface to the pressure of the gas.

\[ \frac{q}{q_{\text{max}}} = \frac{K_a \cdot p}{1 + K_a \cdot p} \]

- \( p \) is the pressure of the gas
- \( q \) is loading \( q \) on the surface
- \( q_{\text{max}} \) is the maximum loading
- \( K_a \) is the adsorption strength
LANGMUIR’S ADSORPTION EQUATION

Background theory

L1. Site balance: \( S_0 = S + S_a \)
L2. Adsorption rate model: \( r_{\text{ads}} = k_{\text{ads}} \cdot p \cdot S \)
L3. Desorption rate model: \( r_{\text{des}} = k_{\text{des}} \cdot S_a \)
L4. Equilibrium assumption: \( r_{\text{ads}} = r_{\text{des}} \)
L5. Mass balance on \( q \): \( q = S_a \)

\[ \mathcal{K} \quad \text{CONSTRAINTS} \]

C1. \( f(0) = 0 \)
C2. \( \forall p > 0 \) \( (f(p) > 0) \)
C3. \( \forall p > 0 \) \( (f'(p) \geq 0) \)
C4. \( 0 < \lim_{p \to 0} f'(p) < \infty \)
C5. \( 0 < \lim_{p \to \infty} f(p) < \infty \)
LANGMUIR’S ADSORPTION EQUATION

- $f_2$ and $g_2 \rightarrow$ provable
- $g_5$ $g_7 \rightarrow$ satisfy the constraints
Relativistic time dilation formula computes:
The frequency $f$ for a clock moving at speed $v$ is related to the frequency $f_0$ of a stationary clock by the formula

$$\frac{f - f_0}{f_0} = \sqrt{1 - \frac{v^2}{c^2}} - 1$$

Einstein's theory of relativity: the speed of light is constant
Two observers in relative motion to each other will experience time differently and observe different clock frequencies

Relativistic time dilation formula computes:
The frequency $f$ for a clock moving at speed $v$ is related to the frequency $f_0$ of a stationary clock by the formula

$$\frac{f - f_0}{f_0} = \sqrt{1 - \frac{v^2}{c^2}} - 1$$
### 2 Background theories

<table>
<thead>
<tr>
<th>Relativistic</th>
<th>Newtonian</th>
</tr>
</thead>
<tbody>
<tr>
<td>$dt_0 = 2 \cdot d/c$</td>
<td>$dt_0 = 2 \cdot d/c$</td>
</tr>
<tr>
<td>$dt = 2 \cdot L/c$</td>
<td>$dt = 2 \cdot L/\sqrt{v^2 + c^2}$</td>
</tr>
<tr>
<td>$L^2 = d^2 + (v \cdot dt/2)^2$</td>
<td>$L^2 = d^2 + (v \cdot dt/2)^2$</td>
</tr>
<tr>
<td>$f_0 = 1/dt_0$</td>
<td>$f_0 = 1/dt_0$</td>
</tr>
<tr>
<td>$f = 1/dt$</td>
<td>$f = 1/dt$</td>
</tr>
<tr>
<td>$df = f - f_0$</td>
<td>$df = f - f_0$</td>
</tr>
<tr>
<td>$d &gt; 0, v &gt; 0$</td>
<td>$d &gt; 0, v &gt; 0$</td>
</tr>
<tr>
<td>$c = 3 \times 10^8$</td>
<td>$c = 3 \times 10^8$</td>
</tr>
</tbody>
</table>

Relativistic axioms

Newtonian axioms

**AI-Descartes can identify which theory explains the phenomenon better**
### RELATIVISTIC TIME DILATION

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(-0.00563v^2)</td>
<td>.3822</td>
<td>.3067</td>
<td>37 (\leq v \leq 115)</td>
<td>37 (\leq v \leq 10^8)</td>
</tr>
<tr>
<td>(\frac{v}{1+0.00689v} - v)</td>
<td>.3152</td>
<td>.2097</td>
<td>37 (\leq v \leq 49)</td>
<td>37 (\leq v \leq 38)</td>
</tr>
<tr>
<td>(-0.00537\frac{v^2\sqrt{v+v^2}}{(v-1)})</td>
<td>.3027</td>
<td>.2299</td>
<td>37 (\leq v \leq 98)</td>
<td>37 (\leq v \leq 109)</td>
</tr>
<tr>
<td>(-0.00545\frac{v^4}{\sqrt{v^2+v^2(v-1)}})</td>
<td>.3238</td>
<td>.2531</td>
<td>37 (\leq v \leq 126)</td>
<td>37 (\leq v \leq 10^7)</td>
</tr>
</tbody>
</table>
COMPARISON WITH SOTA SYSTEMS

- **AI-Feynman**: deep learning based symbolic regression algorithm.
- **TuringBot**: simulated annealing method to find expressions that fit the input data.
- **PySR**: based on regularized evolution, simulated annealing, and gradient-free optimization.
- **Bayesian Machine Scientist (BMS)**: Markov chain Monte Carlo based method exploiting a prior learned from a large empirical corpus of mathematical expressions.

<table>
<thead>
<tr>
<th></th>
<th>AI-Descartes</th>
<th>AI-Feynman</th>
<th>PySR</th>
<th>BMS</th>
<th>TuringBot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.60</td>
<td>0.41</td>
<td>0.49</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>Accuracy (max 2 var)</td>
<td>0.87</td>
<td>0.80</td>
<td>0.73</td>
<td>0.80</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Accuracy on the Feynman synthetic dataset
CONCLUSION & FUTURE/ONGOING WORK

Strengths:
• Few data points / Real data
• Logical reasoning to distinguish the correct formula from a set of plausible formulas with similar error on the data

Limitations:
• Limitation of the tools
• Scalability to bigger formulas
• Rely on correctness & completeness of background theory

Main challenges:
• We need more real-data datasets (only synthetic datasets with non-realistic amount/type of noise)
• We need more numerical datasets with associated background theory

https://github.com/IBM/AI-Descartes

Combining data and theory for derivable scientific discovery with AI-Descartes

Cristina Cornelio1,2, Sanjeeb Dash1, Vernon Austel1, Tyler R. Josephson2,3, João Gonçalves4, Kenneth L. Clarkson5, Nimrod Megiddo5, Bashir El Khadir5 & Lior Hirsch1,3,5

Scientists aim to discover meaningful formulae that accurately describe experimental data. Mathematical models of natural phenomena can be manually created from domain knowledge and fitted to data, or, in contrast, created automatically from large datasets with machine-learning algorithms. The problem of incorporating prior knowledge expressed as constraints on the functional form of a learned model has been studied before, while finding models that are consistent with prior knowledge expressed via general logical axioms is an open problem. We develop a method to enable principled derivations of natural phenomena from axiomatic knowledge and experimental data by combining logical reasoning with symbolic regression. We demonstrate these concepts for Kepler's third law of planetary motion, Einstein's relativistic time-dilation law, and Langmuir's theory of adsorption. We show we can discover governing laws from few data points when logical reasoning is used to distinguish between candidate formulae having similar error on the data.

https://ai-descartes.github.io
REFERENCES

Papers:
• [Under submission] **AI-Hilbert**

Patents:
• **Symbolic Model Discovery based on a combination of Numerical Learning Methods and Reasoning**, C. Cornelio, L. Horesh, A. Fokoue-Nkoutche, S. Dash.
• **Experimental Design for Symbolic Model Discovery**, L. Horesh, K. Clarkson, C. Cornelio, S. Magliacane.
• **Logical and Statistical Composite Models**, L. Horesh, B. El Kadir, S. Dash, K. Clarkson, C. Cornelio