Discovery of exact equations via finding a vanishing ideal basis

Boštjan Gec[1,2], Sašo Džeroski[1], Ljupčo Todorovski[3,1]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[2] Jozef Stefan International Postgraduate School, Ljubljana, Slovenia
[3] Faculty of Mathematics and Physics, University of Ljubljana, 1000 Ljubljana, Slovenia

Keywords: machine learning; symbolic regression; equation discovery; Gröbner basis; Diophantine equations; online encyclopedia of integer sequences, OEIS

In this abstract, we propose a method for addressing the machine learning task of equation discovery (also known as symbolic regression) [Tod10]. We demonstrate its strong connection to a fundamental concept in algebraic geometry: the correspondence between an affine variety and a vanishing ideal. More specifically, we propose that the problem of calculating the Gröbner basis [Buc06] of a vanishing ideal is related to and can be beneficial to the task of equation discovery. While algorithms derived from Buchberger's algorithm have been previously utilized to discern systems of ordinary differential equations [KH16], we believe that the realm of equation discovery, particularly the segment about algebraic equations (including the so-called exact equations), has not yet been explored by such algorithms. To illustrate the potential of this approach, we focus on an application of equation discovery that aligns more closely with the perspective of a mathematician. This particular application is chosen also to demonstrate the potential relevance and utility of such approaches in their respective research work. This potential is exemplified by recent advancements in applying machine learning techniques in theoretical algebraic geometry [HLOP22]. For this purpose, we delve into the realm of integer sequences to discover or reconstruct already known exact (recursive) equations, such as the one defined by Fibonacci, using only the given sequence terms.

Equation discovery, also known as symbolic regression, is a discipline of machine learning focused on the computational discovery of the most fitting and simplest mathematical expression that describes a given data set. Typically, the input data are presented in tabular format as a matrix $A \in \mathbb{R}^{m \times n}$, where each row A_i signifies a data point (also referred to as an example within the data set), and each column A^j represents the values of a corresponding variable x_j. The objective is to find a mathematical equation

$$E(x_1, ..., x_n) = 0$$

involving variables x_j that minimizes two criteria:

- The error-of-fit (e.g., root-mean-square-error (RMSE) of the expression against the data set A)
- The complexity of the expression $E(x_1, ..., x_n)$ (e.g., number of nodes in its abstract syntax tree)

This task is guided by the Occam's razor principle [Grü07], which advocates for the simplest explanation. The goal is not only to predict the data as accurately as possible but also to extract meaningful knowledge from it. This aligns with the increasing emphasis on explainability and causality in the machine learning community, where understanding the underlying patterns is a priority. If the complexity criterion is ignored and the usual equation form (as in the footnote 4) is sought, we obtain the classical (supervised) machine learning task of regression [JWH+23]. However, traditional machine learning algorithms approach such tasks by assuming a specific equation via the assumption of a certain equation form (e.g., linear, polynomial, neural network) and subsequently optimizing the parameters to minimize the error of fit. In contrast, symbolic regression is more about searching for the form of the equation than the parameters. Symbolic regression aims to automate the scientific discovery process, reminiscent of the manual search for equations that best describe data, as practiced by scientists like Kepler, Newton, or Maxwell.

Algebraic geometry [Sha13] on the other side studies solutions to sets of polynomial equations, known as algebraic varieties. These varieties, and the ideals associated with them in a ring (an algebraic structure extending the concept of integers), can be infinitely large. However, the elegance of algebraic geometry lies in its ability to simplify even such complex structures. Through the concept of generators – a finite set of polynomials – entire ideals can be represented, effectively condensing the infinite into a manageable form. This fundamental process of simplification allows mathematicians to delve into the intricate relationships between ideals and varieties, unlocking the rich landscape of algebraic structures.

In the interest of precision, without delving into all the mathematical intricacies [Sha13], we aim to define the vanishing ideal, denoted $I(X)$, corresponding to a subset X within the so-called affine space $K^n = K \times \cdots \times K$. Here, K represents an algebraically closed field, such as $\mathbb{C}$. The vanishing ideal is defined as the collection of all multivariate polynomials on $X_1, ..., X_n$, that is, $\mathbb{K}[X_1, ..., X_n]$, which equate to zero (i.e., vanish) at every point $(x_1, ..., x_n)$ within the specified subset $X \subset K^n$. In other words,

$$I(X) := \{f \in \mathbb{K}[X_1, ..., X_n] \mid f(x_1, ..., x_n) = 0 \text{ for all } (x_1, ..., x_n) \in X\} \ .$$

It is worth mentioning that the set X is termed as the variety of the ideal $I(X)$, provided it includes all common zeros of all elements of $I(X)$. The link between equation discovery and algebraic geometry can be elucidated as follows: According to the definition above, for each element of the ideal $I(X)$, the equation

$$f(x_1, ..., x_n) = 0$$

is valid for all points in X. Consequently, if the set X is viewed as the input data set and only polynomial expressions are taken into account, the equation discovery problem can be reinterpreted as the task of identifying the simplest element of the associated vanishing ideal. While the scope of equation discovery is not confined solely to polynomials, the utilization of algebraic geometry tools is still a viable approach, particularly in the absence of any background knowledge or constraints on the data.

The simplification of the vanishing ideal, as previously discussed, involves identifying a finite set of generators that not only form a basis for the ideal but also satisfy specific conditions. A prime example of such a basis is the Gröbner basis [Buc06]. This basis is characterized by additional conditions defined by a so-called monomial order, which essentially necessitates the lowest degrees of the polynomials in the basis. Calculating the Gröbner basis is a well-established problem in commutative algebra, typically resolved using Buchberger's algorithm [Buc06]. Given the uniqueness of the Gröbner basis (subject to the choice of a monomial order), it emerges as a natural solution to the equation discovery problem. The Gröbner basis computation process can be viewed as a nonlinear multivariate extension of both Euclid's algorithm for determining the greatest common polynomial divisors and the Gaussian elimination for linear systems. As a fundamental tool in computer algebra, it finds applications in numerous areas of mathematics and computer science, including but not limited to algebraic geometry, cryptography, and automated theorem proving [SSM$^+$09].

Due to its precise nature, Buchberger's algorithm may not be suitable for numerical data. The presence of any noise in the data could lead to overfitting, resulting in meaningless high-degree terms appearing in the polynomials. However, recognizing the potential to extract geometric information from data, there is an expanding body of research dedicated to developing approximate versions of the algorithm to handle numerical data. One of the groundbreaking ideas in this field was to search for polynomials f that "almost" vanish on the given data, i.e. $f(x_1, ..., x_n) < \varepsilon$ for a chosen threshold ε. This approach [HKPP09] led to the development of the so-called *Approximate Vanishing Ideal* (AVI), which has seen a wide range of applications in machine learning. One such application is the construction of new variables similar to principal component analysis, termed *vanishing component analysis* [LLS$^+$13].

The Approximate Vanishing Ideal algorithm has seen a wealth of improvements and modifications, including the introduction of the Monomial Agnostic Variational Ideal (MAVI) in paper [KH23]. These adaptations have found applications also in modeling differential equations and system identification tasks, which can be viewed as equation discovery tasks where systems of differential equations are the target. Although various vanishing ideal approaches [KNTB20,KH16] have been used to discover differential equations, to the best of our knowledge, AVI-based approaches have not been applied to find algebraic (non-differential) equations.

We used data from prior experimental studies which were sourced from the Online Encyclopedia of Integer Sequences (OEIS [SI20]) focusing on the first 200 terms of 164 *core* integer sequences identified by OEIS maintainers as most significant. Data include *trivial* sequences, like the sequence of natural numbers, *simple* sequences with known closed-form and recursive equations, e.g., the Fibonacci and Catalan numbers, and *others* for which no equations are known or available, such as the sequence of prime numbers. We conducted a loose comparison of MAVI performance with our previous experiments using Diofantos [Gec] and SINDy [KdSF$^+$22] methods. It is important to note that this comparison is biased in favor of MAVI due to the comparative analysis methodology employed in our previous studies. For instance, certain parameters (e.g., the number of sequence elements considered) were not altered as they were in the MAVI experiments. In contrast, the comparison may of course also be unfair to MAVI due to our unfamiliarity with the method, as we did not adjust parameters such as the scale factor α.

Table 1. Preliminary comparison of the performance of MAVI against Diofantos and SINDy on the *core* data set. Results are grouped by categories. Note, that the comparison is unfair, as discussed in the abstract.

Seq. category	#Sequences	MAVI	Diofantos	SINDy
Trivial	4	4	4	4
Simple	66	25	32	31
Other	94	0	0	0
Σ	164	29	36	35

The results of our preliminary study, presented in Table 1, show promise. On one side, among 164 sequences, our prior methods successfully identified equations for all 4 *trivial* and over 30 *simple* sequences but found none for the remaining *other* sequences. On contrast, MAVI underperformed compared to other methods for 7 *simple* sequences. We hypothesize that the failure of these seven sequences may be related to large terms within the sequences. This hypothesis stems from the observation that when considering only the first 20 terms of each sequence, the maximum terms in all the failing sequences strictly exceed 10^8, while other non-trivial sequences have a maximum term below 10^7. For instance, factorial sequences naturally exhibit exponential term growth. Furthermore, all 7 failing sequences have a ground truth equation of degree 2 or higher, whereas MAVI only reconstructed equations of degree 1. MAVI underperformed for 7 simple sequences, potentially due to large terms. Notably, these failing sequences all have maximum terms exceeding 10^8 in the first 20 terms, unlike other sequences (e.g., factorials) with maxima below that threshold. Furthermore, all failing sequences have higher-degree ground truth equations (≥ 2) compared to MAVI's reconstructed degree-1 equations. With meticulous parameter tuning and the implementation of the algorithm's exact version, we anticipate achieving results that are not only comparable but potentially superior.

One of the key advantages of the MAVI approach over other state-of-the-art symbolic regression methods is its ability to capture the geometry of data in the algebraic geometry sense, as well as its capacity to discover equations in implicit form. To date, the only tool developed capable of discovering equations that can handle implicit equations is SINDy-PI [KKB20]. The scarcity of tools capable of discovering implicit equations is because the symbolic regression task is typically defined in explicit form, as described in equation 1 in footnote[4]. If a typical symbolic regression tool was used with an assumed implicit equation, the tool would automatically find the trivial solution $E = 0$ and miss the actual one.

While equation discovery traditionally pertains to numerical measurements of physical phenomena, it can also be applied to noise-free or exact data, such as those describing mathematical objects. In such cases, an approximate version or modification of the algorithm becomes redundant. Consequently, the original Buchberger algorithm should theoretically outperform its approximate versions when dealing with exact data. In future work, we intend to incorporate computational experiments to validate this hypothesis. Despite this hypothesis, we selected the data from integer sequences as our testing ground for the MAVI method, conducting preliminary computational experiments. This choice was influenced by our prior experience with this specific data set.

The development of exact symbolic regression methods holds significant value, particularly considering the potential contributions to the field of mathematics. Machine learning has gained considerable momentum in various mathematical fields, including theoretical algebraic geometry.

[4] Trivial solution $E := 0$ obviously satisfies the equality E = 0 with minimal complexity. But since it does not give us any new information, we are not interested in it. Usually equation of the form

$$x_j = E(x_1, ..., x_n) \tag{1}$$

$x_j = E(x_1, ..., x_n)$ is considered, where x_j is any of the variables and is excluded from the variables appearing on the right-hand side. This is equivalent to finding an equation of the form $E = 0$ with an additional constraint of x_j combining linearly in E.

In the study [HL19], for instance, researchers innovatively employed a multi-layer perceptron (a compact artificial neural network) to determine whether a Calabi-Yau variety is elliptically fibered. In a subsequent paper [HL21], the same team successfully undertook a regression task to predict Hodge numbers associated with Calabi-Yau varieties with remarkable precision. Fascinatingly, the authors put forward the possibility of an undiscovered formula for the Hodge number $h^{3,1}$, an idea sparked by the success of their machine learning approach. The endeavor of uncovering such a formula appears to be ideally suited for exact equation discovery methods. In addition to the studies mentioned above, there have been numerous successful attempts to leverage machine learning for novel discoveries in the field of mathematics. These include but are not limited to, studies [DVB+21], [HLOP22], [CKV23b], and [CKV23a], among others.

References

Buc06. Bruno Buchberger. Bruno buchberger's phd thesis 1965: An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal. *Journal of Symbolic Computation*, 41(3):475–511, 2006. Logic, Mathematics and Computer Science: Interactions in honor of Bruno Buchberger (60th birthday).

CKV23a. Tom Coates, Alexander M. Kasprzyk, and Sara Veneziale. Machine learning detects terminal singularities. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

CKV23b. Tom Coates, Alexander M. Kasprzyk, and Sara Veneziale. Machine learning the dimension of a fano variety. *Nature comunications*, 14:2041–1723, 9 2023.

DVB+21. Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, Marc Lackenby, Geordie Williamson, Demis Hassabis, and Pushmeet Kohli. Advancing mathematics by guiding human intuition with ai. *Nature*, 600:70–74, 12 2021.

Gec. Boštjan Gec. Exact equation discovery for integer sequences. GitHub repository.

Grü07. Peter D. Grünwald. *The Minimum Description Length Principle (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

HKPP09. Daniel Heldt, Martin Kreuzer, Sebastian Pokutta, and Hennie Poulisse. Approximate computation of zero-dimensional polynomial ideals. *Journal of Symbolic Computation*, 44(11):1566–1591, 2009. In Memoriam Karin Gatermann.

HL19. Yang-Hui He and Seung-Joo Lee. Distinguishing elliptic fibrations with AI. *Physics Letters B*, 798:134889, nov 2019.

HL21. Yang-Hui He and Andre Lukas. Machine learning calabi-yau four-folds. *Physics Letters B*, 815:136139, apr 2021.

HLOP22. Yang-Hui He, Kyu-Hwan Lee, Thomas Oliver, and Alexey Pozdnyakov. Murmurations of elliptic curves, 2022.

JWH+23. G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor. *An Introduction to Statistical Learning: with Applications in Python*. Springer Texts in Statistics. Springer Cham, 2023.

KdSF+22. Alan A. Kaptanoglu, Brian M. de Silva, Urban Fasel, Kadierdan Kaheman, Andy J. Goldschmidt, Jared Callaham, Charles B. Delahunt, Zachary G. Nicolaou, Kathleen Champion, Jean-Christophe Loiseau, J. Nathan Kutz, and Steven L. Brunton. Pysindy: A comprehensive python package for robust sparse system identification. *Journal of Open Source Software*, 7(69):3994, 2022.

KH16. Hiroshi Kera and Yoshihiko Hasegawa. Noise-tolerant algebraic method for reconstruction of nonlinear dynamical systems. *Nonlinear Dynamics*, 85(1):675–692, 7 2016.

KH23. Hiroshi Kera and Yoshihiko Hasegawa. Monomial-agnostic computation of vanishing ideals, 2023.

KKB20. Kadierdan Kaheman, J. Nathan Kutz, and Steven L. Brunton. Sindy-pi: a robust algorithm for parallel implicit sparse identification of nonlinear dynamics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2242), 2020.

KNTB20. Artur Karimov, Erivelton G. Nepomuceno, Aleksandra Tutueva, and Denis Butusov. Algebraic method for the reconstruction of partially observed nonlinear systems using differential and integral embedding. *Mathematics*, 8(2), 2020.

LLS+13. Roi Livni, David Lehavi, Sagi Schein, Hila Nachliely, Shai Shalev-Shwartz, and Amir Globerson. Vanishing component analysis. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 597–605, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Sha13. Igor R. Shafarevich. *Basic Algebraic Geometry 1*. Springer Berlin, Heidelberg, third edition, 2013.

SI20. Neil J. A. Sloane and The OEIS Foundation Inc. The on-line encyclopedia of integer sequences, 2020.

SSM+09. Massimiliano Sala, Shojiro Sakata, Teo Mora, Carlo Traverso, and Ludovic Perret, editors. *Gröbner Bases, Coding, and Cryptography*. Springer Berlin, Heidelberg, 2009.

Tod10. Ljupčo Todorovski. *Equation Discovery*, pages 327–330. Springer US, Boston, MA, 2010.