Equation discovery as vanishing ideal basis problem

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

[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

In this abstract, we offer a different perspective on a well-known machine learning task that is central to the field of Equation Discovery [**?**]. 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 Groebner basis of a vanishing ideal is related to, and can be beneficial for, the task of equation discovery. While algorithms derived from Buchberger's algorithm have been previously utilized to discern systems of ordinary differential equations, we believe that the realm of equation discovery, particularly the segment pertaining to algebraic equations (including the so-called exact equations), has not yet been explored by such algorithms. To illustrate the potential of this approach, we present two use cases. In the second, we examine the classical problem of equation discovery in the field of chemistry, more specifically, material science, where the discovery of algebraic equations is targeted. In the first use case, we focus on an application of equation discovery that aligns more closely with a mathematician's perspective. This particular application is chosen also with the aim to demonstrate the potential relevance and utility of such approaches in their respective research work. For this purpose, we delve into the realm of integer sequences with the goal of discovering or reconstructing already known exact (recursive) equations, such as the one defined by Fibonacci, using only the given sequence terms. As a side effect, we highlight the potential of using equation discovery techniques, as well as machine learning in general, in the field of mathematics. This field has recently gained a lot of traction, including in the area of theoretical algebraic geometry.

Equation discovery, also known as symbolic regression, is a machine learning discipline focused on the computational discovery of the most fitting and simplest mathematical expression that describes a given dataset.

Typically, the input data is presented in a 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 dataset), and each column A_j represents the values of a corresponding feature x_j. The objective is to find a mathematical equation

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

involving some of the features x_j that minimizes two criteria:

- The error-of-fit (e.g., root-mean-square-error (RMSE) of the expression against the dataset)
- The complexity of the expression (e.g., number of nodes in its abstract syntax tree)

This task is guided by the Occam's razor principle ([**?**]), 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 underlying patterns is a priority. If the complexity criterion is ignored, we obtain the classical machine learning task of regression. However, traditional machine learning algorithms approach such tasks by assuming a specific equation via 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 rather 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.

In the following sections, we will delve into the motivation behind this approach and present the results of applying equation discovery in the fields of chemistry and mathematics. In next section, we will explore the connection between equation discovery and algebraic geometry.

Algebraic geometry is on the other side a branch of mathematics that studies solutions of sets of polynomial equations, known as algebraic varieties. In this field, one of the key concepts is the ideal, which is a special subset of a ring, an algebraic structure that extends the familiar concept of integers.

The variety X may be infinite, representing an infinite set of solutions to a system of polynomial equations. Similarly, an ideal associated with this variety may also be infinite, encompassing an infinite set of polynomials that vanish on X.

However, the beauty of algebraic geometry lies in its ability to simplify these seemingly complex structures. Despite the potential infiniteness of an ideal, it can be represented by a finite set of generators. These generators form a smaller set of polynomials whose ideal is the same as the original, effectively simplifying the infinite ideal into a more manageable, finite form.

This process of simplification is a fundamental aspect of algebraic geometry, allowing mathematicians to study and understand the intricate relationships between ideals and varieties, and explore the rich and fascinating 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 as $I(X)$, 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 $f \in \mathbb{K}[X_1, ..., X_n]$, that equate to zero (i.e., vanish) at every point $(x_1, ..., x_n)$ within the specified subset $X \subset K^n$. I.e.,

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

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: As per 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 remains a viable approach for our task, particularly in the absence of any additional 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 ([**?**]). 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. The computation of the Gröbner basis is a well-established problem in comutative algebra, typically resolved using Buchberger's algorithm. Given the uniqueness of the Gröbner basis (subject to the choice of a monomial order), it emerges as a natural solution for equation discovery. The process of Gröbner basis computation can be viewed as a multivariate, non-linear extension of both Euclid's algorithm for determining polynomial greatest common divisors and 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.

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 features similar to Principal Component Analysis, termed as *Vanishing Components Analysis*. This method was published at the International Conference on Machine Learning (ICML), one of the three primary high-impact machine learning conferences.

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]. papers ... 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. While various vanishing ideal approaches ([**?**] ...) have been used to discover differential equations, to the best of our knowledge, MAVI has not been applied to find algebraic (non-differential) equations.

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]). While SINDy-PI is primarily designed to discover differential equations, it can also be used for implicit algebraic equations.

The scarcity of tools capable of discovering implicit equations is due to the fact that 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 implicit equation assumed, the tool would automatically find the trivial solution $E = 0$.

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 integer sequences' data as our testing ground for the MAVI method, conducting preliminary computational experiments. This choice was influenced by our prior experience with this specific dataset.

The data used in our experiments were obtained from our previous studies and sourced from the Online Encyclopedia of Integer Sequences (OEIS ([**?**])). The first 200 terms of selected sequences were stored in a CSV file and made publicly accessible in our repository [**?**]. We opted for a set of 164 "core" sequences, identified by OEIS maintainers as the most significant and representative ones.

We conducted a loose comparison of MAVI's performance against our previous experiments using the Diofantos and SINDy methods. It's important to note that this comparison is skewed 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. Conversely, the comparison may of course also be unfair to MAVI due to our unfamiliarity with the method, as we did not tune parameters such as the scale factor α.

The preliminary study's results, as shown in Table 1, are promising.

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 paper.

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

Although initial results indicate that MAVI underperforms when compared to other methods, we remain optimistic. 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.

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.

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 forth 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 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.

Beyond computational experiments on integer sequences data, our poster will also showcase computational experiments in the field of material science. Here, we plan to apply the MAVI method to discover algebraic equations from chemical data, demonstrating the method's potential applicability in material science. Such a discovery could revolutionize the operational mode of a specific modern measurement device, broadening its applicability.

[4]

[KH23]
[Gec]

References

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 and 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.

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.

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):20200279, 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.

[4] Trivial solution $E := 0$ obviously satisfies the equality $\mathrm{E} = 0$ with the minimal complexity. But since it does not give us any new information, we are not interested in it. Actually, 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 features and is excluded from the variables appearing on the right hand side. This is equivalent to finding the equation of the form $E = 0$ with additional constrain of x_j combining linearly in E.