Machine learning and moduli spaces of curves

Elira Curri and Tony Shaska

Abstract. We propose new methods to apply machine learning to various databases which have emerged in the study of the moduli spaces of algebraic curves. We find that with such methods one can learn many significant quantities to astounding accuracy in a matter of minutes and can also predict unknown results making this approach a valuable tool in pure mathematics.

Introduction

Artificial Intelligence and Machine Learning are some of the most active and exciting branches of science of the last few decades. These new technologies have made their way into economy, including engineering, medical science, finance, cybersecurity, etc. Can they be used for mathematical research?

The question is not new. After all science is all about collecting data and deducing conclusions. Machine Learning is about gathering data, training the data, and drawing conclusions. Depending on the kind of data we use different methods for machine learning: supervised learning, unsupervised learning, or a combination of the two.

So the first step is to gather the data. There have always been databases in mathematics, but the most famous databases of the XX-century were the Atlas of Finite Simple Groups, Cremona tables of elliptic curves, database of elliptic curves compiled by Birch and Swinnerton-Dyer which led to the famous Birch and Swinnerton-Dyer conjecture; see [2]. With the development of computer algebra toward the last quarter of the XX-century we saw different databases which had a huge impact on mathematics, for example the Small Library of Groups in Gap, the list of Calabi-Yau hypersurfaces, etc.

The goal of this work is to use new tools of machine learning to study the moduli space $\mathcal{M}_g$ of genus $g \geq 2$ curves defined over a field k. The moduli space of algebraic curves has been the focal point of algebraic geometry for the last few decades. With the development of the new computational tools it became necessary in the last few decades to reconsider the theory of invariants, in its classical form or in the framework of the theory developed by Mumford with the intention of studying the arithmetic of moduli spaces. Naturally some of the first attempts focused on $\mathcal{M}_2$; see for example [6] and attempts to generalize to $g > 2$ [5]. From these attempts the concept of weighted Weil height was born; see [7].

The moduli space $\mathcal{M}_2$ as a case study.

The moduli space $\mathcal{M}_2$ of genus 2 curves is the most understood moduli space among all moduli spaces. This is mostly due to two main facts; first all genus two curves are hyperelliptic and therefore studying them it is easier than general curves, secondly even among hyperelliptic curves the curves of genus two have a special place since they correspond to binary sextics which, from the computational point of view, are relatively well understood compared to higher degree binary forms.

One of the main questions related to $\mathcal{M}_2$ has been to recover a nice equation for any point $\mathfrak{p} \in \mathcal{M}_2$. Since $\mathcal{M}_2$ is a coarse moduli space, such equation is not always defined over the field of moduli of $\mathfrak{p}$. Can we find a universal equation for genus two curves over their minimal field of definition? Can such equation provide a minimal model for the curve? Does the height of this minimal model has any relation to the projective height of the corresponding moduli point $\mathfrak{p} \in \mathcal{M}_2$? What is the distribution in $\mathcal{M}_2$ of points $\mathfrak{p}$ for which the field of moduli is not a field of definition? The answers to these questions are still unknown.

In [1] we provide a database of genus 2 curves which contains all curves with height $\mathrm{h} \leq 5$, curves with moduli height $\mathfrak{h} \leq 20$, and curves with automorphism and height ≤ 101. They are organized in three Python directories $\mathcal{L}_i$, $i = 1, 2, 3$. The database is build with the idea of better understanding $\mathcal{M}_2$, the distribution of points in $\mathcal{M}_2$ based on the moduli height, the distribution of points for which the field of moduli is not a field of definition. Even in genus $g = 2$ there are many technical issues that need to be addressed.

Let $\mathcal{X}$ be a genus two curve defined over $\mathbb{Q}$. The moduli point in $\mathcal{M}_2$ corresponding to $\mathcal{X}$ is given by $\mathfrak{p} = (i_1, i_2, i_3)$, where i_1, i_2, i_3 are absolute invariants as in [1]. Since i_1, i_2, i_3 are rational functions in terms of the coefficients of $\mathcal{X}$, then $i_1, i_2, i_3 \in \mathbb{Q}$. The converse isn't necessarily true. Let $\mathfrak{p} = (i_1, i_2, i_3) \in \mathcal{M}_2(\mathbb{Q})$. The universal equation of a genus 2 curve corresponding to $\mathfrak{p}$ is determined in [6], which is defined over a quadratic number field K. The main questions we want to consider is what percentage of the rational moduli points are defined over $\mathbb{Q}$? How can we determine a minimal equation for such curves?

For every point $\mathfrak{p} \in \mathcal{M}_2$ such that $\mathfrak{p} \in \mathcal{M}_2(k)$, for some number field K, there is a pair of genus-two curves $\mathcal{C}^\pm$ given by $\mathcal{C}^\pm : \; y^2 = \sum_{i=0}^6 a_{6-i}^\pm \, x^i$, corresponding to $\mathfrak{p}$, such that $a_i^\pm \in K(d)$, $i = 0, \dots, 6$; see [6].

In [1] were created three Python dictionaries: $\mathcal{L}_1$: curves with height ≤ 10, $\mathcal{L}_2$: curves with extra involutions, $\mathcal{L}_3$: curves with small moduli height. There are 20 697 curves in $\mathcal{L}_2$, such that for each h we have roughly $4\,\mathrm{h}$ curves. So it is expected that the number of curves of height $\leq h$, defined over $\mathbb{Q}$ is $\leq 4\frac{\mathrm{h}(\mathrm{h}+1)}{2}$. Let $\mathfrak{p} \in \mathcal{M}_2(\mathbb{Q})$ be such that $\mathrm{Aut}(\mathfrak{p}) \cong V_4$. There is a genus 2 curve $\mathcal{X}$ corresponding to $\mathfrak{p}$ with equation $y^2 \, z^4 = f(x^2, z^2)$. We pick $f \in \mathbb{Z}[x, z]$, such that $f(x, z)$ is a reduced binary form. From 20 292 such curves we found only 57 which do not have minimal absolute height. $\mathcal{L}_3$ is a list of all moduli points $[x_0 : x_1 : x_2 : x_3]$ of projective height $\leq \mathfrak{h}$ in $\mathbb{P}^3(\mathbb{Q})$, for some integer $\mathfrak{h} \geq 1$. Each such point correspond to the point $[J_2^5 : J_4 J_2^3 : J_6 J_2^2 : J_{10}]$.

What percentage of rational points $\mathfrak{p} \in \mathcal{M}_2(\mathbb{Q})$ with a fixed moduli height $\mathfrak{h}$ have $\mathbb{Q}$ as a field of definition, when $\mathfrak{h}$ becomes arbitrarily large? We confirm, as expected, that for large moduli height $\mathfrak{h} \in \mathcal{M}_2(\mathbb{Q})$, the majority of genus 2 curves not defined over $\mathbb{Q}$ and they don't have extra automorphisms.

Higher moduli:

Can we generalize the approach above to $\mathcal{M}_g$ for $g > 2$? Moreover, can we train a machine learning model to obtain reliable results for $g \geq 2$?

The moduli space $\mathcal{M}_2$ is a very good model for the hyperelliptic moduli $\mathcal{H}_g$. Many of the results of $g = 2$ have been realized to higher genus hyperelliptic curves already and we now know many general theorems for $\mathcal{H}_g$; see [3], [5], etc.

Moreover, generalizing from hyperelliptic curves to superelliptic curves gives a very important tool in understanding $\mathcal{M}_g$; see [5] for details. Using results from [4] and previous work of these authors we can determine fully the list of automorphisms groups and inclusions among the loci for any genus, hence obtaining a full stratification of the moduli space $\mathcal{M}_g$. About 75-80% of all cases come from superelliptic curves, for which we know a great deal.

A very important development in understanding $\mathcal{M}_g$ is the discovery of the weighted height on the weighted projective spaces. Hence, the most efficient way to create a database of points in $\mathcal{M}_g$ is to consider the corresponding weighted moduli space $\mathcal{W}_g$ and sort the points in this space via their weighted heights.

A great learning example is the case $g = 3$ for many reasons. It is the first case that we have non-hyperelliptic curves, so it is more general than $g = 2$, but also it is still a case that we fully understand. For example, we explicitly know invariants of binary octavics, which classify hyperelliptic genus 3 curves, and invariants of ternary quartics which classify non-hyperelliptic genus 2 curves. We have a full understanding of the list of groups of automorphisms and in each case we can write an explicit parametric equation for the corresponding family. There has been work in the last decade by several authors on the field of moduli of genus 3 curves and we can recover the equation of the curve over a minimal field of definition.

It needs to be pointed out that in this general approach the biggest difficulty comes from arithmetic invariant theory in the sense that we don't know an explicit way of describing a moduli point $\mathfrak{p} \in \mathcal{M}_g$. While GIT provides an elegant theoretical framework, explicit results are missing even for genus g as small as 4 or 5.

Conclusion

Our general philosophy is to build the skeleton of $\mathcal{M}_g$ using the superelliptic curves. After all, the majority of points in $\mathcal{M}_g$ with nontrivial $\mathrm{Aut}(\mathfrak{p})$ are superelliptic points. We can say a lot on these superelliptic points on the problem of field of moduli versus field of definition, determine if they have complex multiplication, and write down explicit equations for them.

In this talk we will describe what can be achieved and what are the challenges for fully understanding the arithmetic of the moduli space. Our goal is to bring this topic to the attention to mathematicians specialized on machine learning and artificial intelligence techniques and hopefully involve more people in this ambitious but exciting project.

References

1. L. Beshaj, R. Hidalgo, S. Kruk, A. Malmendier, S. Quispe, and T. Shaska, *Rational points in the moduli space of genus two*, Higher genus curves in mathematical physics and arithmetic geometry, Contemp. Math., vol. 703, Amer. Math. Soc., [Providence], RI, 2018, pp. 83–115. MR 3782461
2. B. J. Birch and H. P. F. Swinnerton-Dyer, *Notes on elliptic curves. II*, J. Reine Angew. Math. **218** (1965), 79–108. MR 179168
3. J. Gutierrez and T. Shaska, *Hyperelliptic curves with extra involutions*, LMS J. Comput. Math. **8** (2005), 102–115. MR **2135032 (2006b:**14049)
4. K. Magaard, T. Shaska, S. Shpectorov, and H. Völklein, *The locus of curves with prescribed automorphism group*, Sūrikaisekikenkyūsho Kōkyūroku (2002), no. 1267, 112–141, Communications in arithmetic fundamental groups (Kyoto, 1999/2001). MR 1954371
5. A. Malmendier and T. Shaska, *From hyperelliptic to superelliptic curves*, Albanian J. Math. **13** (2019), no. 1, 107–200. MR 3978315
6. Andreas Malmendier and Tony Shaska, *A universal genus-two curve from siegel modular forms*, SIGMA Symmetry Integrability Geom. Methods Appl. **13** (2017), Paper No. 089, 17. MR 3731039
7. Sajad Salami and Tony Shaska, *Local and global heights on weighted projective varieties and Vojta's conjecture*, arXiv preprint arXiv:2204.01624 (2022).

Elira Curri
Department of Computer Science
Research Institute of Science and Technology
9405, Vlora, Albania.
e-mail: `elira@risat.org`

Tony Shaska
Department of Computer Science
Research Institute of Science and Technology
9405, Vlora, Albania.
e-mail: `shaska@risat.org`