Data-Driven Insights into the Rank of Elliptic Curves of Prime Conductors: Patterns, Classification, and Open Questions

University of Montreal, Faculty of Medicine

by Malik Amir

In collaboration with Rashed Atieh, Andreas Hatziiliou, and Eldar Sultanow.

August 25th 2023

An overview of the presentation I

1. Motivation
   - Basics of elliptic curves.
   - The BSD conjecture and the rank.
   - Some results at the intersection of data science and number theory.
2. The choice for prime conductor curves
   - Mazur, Elkies, Watkins and the rest of the dream team.
3. The BGR dataset as a new playground.
4. Old and new observations in the data
   - Some strange/interesting (?) patterns and a list of questions.

An overview of the presentation II

5. Classifying the rank
   - Some results so far in the literature.
   - Feature selection + weird feature construction + LightGBM = almost magic.
6. The end ... (if you have survived☺)

Introduction I : Nature and Patterns *à la Petit Poucet*

Question

Why machine learning for sciences ?

Quote

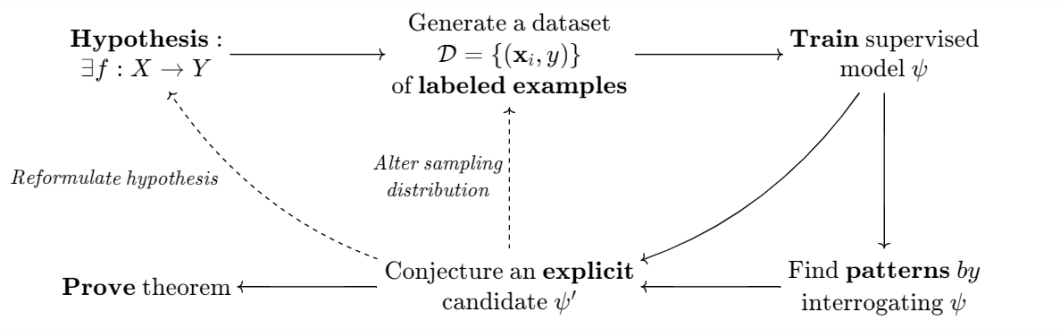
Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry.

-Richard P.Feynman

Introduction II : Guiding Intuition with ML

Definition

The machine learning of mathematical structures designates the use of machine learning to guide human intuition for mathematical research.



Elliptic Curves I : Definitions

Definition

An elliptic curve over $\mathbb{Q}$ is the graph of the cubic equation

$$E/\mathbb{Q} : y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6, \tag{1}$$

where $a_i \in \mathbb{Q}$.

Definition

The Mordell-Weil group of E is th algebraic object defined as

$$E(\mathbb{Q}) = \{(x, y) \in \mathbb{Q}^2 | y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6\} \cup \mathcal{O}.$$

A famous theorem of Mordell and Weil says that $E(\mathbb{Q})$ is a finitely generated Abelian group

$$E(\mathbb{Q}) = E(\mathbb{Q})_{\text{free}}^{r_E} \oplus E(\mathbb{Q})_{\text{tor}}.$$

Elliptic Curves II : L-Function

Definition

Let E be an elliptic curve over $\mathbb{Q}$ and p a prime. We define the Frobenius traces as

$$a_p = \begin{cases} p + 1 - \#E(\mathbb{F}_p) & \text{if } p \text{ is a prime of good reduction,} \\ -1, 0, 1 & \text{if } p \text{ is a prime of bad reduction.} \end{cases}$$

Definition

The L-function of E is the analytic object defined as the Euler product

$$L(E, s) := \prod_{\text{bad } p} \frac{1}{1 - a_p p^{-s}} \prod_{\text{good } p} \frac{1}{1 - a_p p^{-s} + p^{1-2s}} = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$$

for $\mathrm{Re}(s) > 3/2$.

Elliptic Curves III : The Conductor N_E

Theorem

The L-function $L(E, s)$ extends to an entire function on $\mathbb{C}$ and has a functional equation of the form

$$\Lambda(E, s) = \pm\Lambda(E, 2 - s),$$

where

$$\Lambda(E, s) := (2\pi)^{-s}\Gamma(s)N_E^{s/2}L(E, s).$$

The conductor of E is denoted N_E and the $\pm$ symbol is called the sign of the functional equation, also called the root number ε.

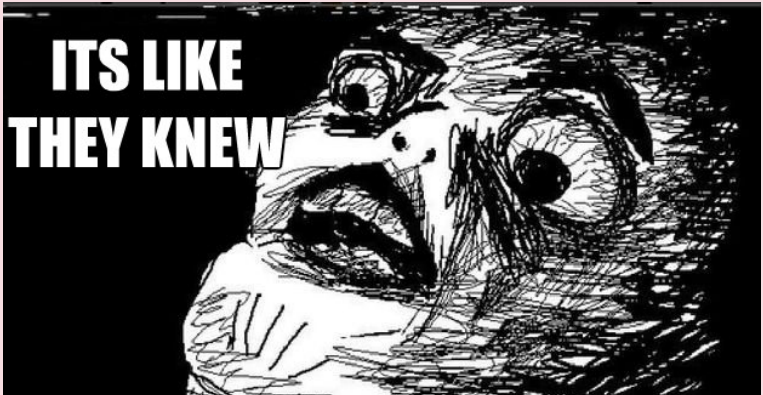
Remark

The conductor of an elliptic curve is a positive integer and it offers a way to order curves.

The BSD Conjecture I : Data-Driven Insights

Observation

The rank of the Mordell-Weil group $E(\mathbb{Q})$ seems to be reflected in the local properties of E through the asymptotic behaviour of the arithmetic function $p \mapsto N_p$. In particular, this tells us that the Frobenius traces *knew* about the rank...



The BSD Conjecture II : First Heuristics

Conjecture

There exists a constant C_E depending only on the choice of elliptic curve E such that

$$\lim_{x\to\infty} \prod_{p<x} \frac{N_p/p}{C_E \log^{r_E}(x)} = 1.$$

In a more informal way, if we evaluate formally the L-function's product at $s = 1$, we get the asymptotic

$$L(E,1) \sim \prod_{\text{good } p} \frac{p}{N_p}. \tag{2}$$

Hence, we can believe that the behaviour of the analytic object $L(E,s)$ near the critical line should capture the algebraic rank of the elliptic curve E...

The BSD conjecture III : Final Form

Conjecture (Strong Form)

Let E be an elliptic curve over $\mathbb{Q}$. Then

$$\lim_{s\to 1}(s-1)^{-r_E}L(E,s) = \frac{\Omega_E|Ш(E/\mathbb{Q})|R_E\prod_{bad\ p}c_p}{|E(\mathbb{Q})_{\text{tor}}|^2}.$$

Theorem (Kolyvagin)

The (weak) form of the BSD conjecture is true for elliptic curves of rank 0 and 1.

Question

What other information are *traces of Frobenius* hidding from us ? Do they know about the torsion order, the torsion structure, the existence of integral points, the order of the Tate-Shafarevich group ? Can we use machine learning to figure out such relationships and formulate explicit conjectures ?

Answer

Yes ! But no time to dive into anything else than the rank right now. Feel free to ask questions later on on these topics ☺.

Average Rank of Elliptic Curves : Tension Between Data and Conjecture

Bektemirov-Mazur-Stein-Watkins 2007

They discuss the minimalist conjecture which states that an elliptic curve of even parity has probability zero of having infinitely many rational points.

Conjecture

The number of even parity elliptic curves with infinitely many rational points and absolute discriminant or conductor less than X is asymptotically given by $c_1 X^{19/24} \log(X)^{3/8}$ for some positive computable constant c_1 as $X \to \infty$.

They present data from the Stein-Watkins database (bigger than LMFDB but not complete) and observe that for composite conductor curves, there is a large amount of curves with infinitely many rational points for conductors less than 10^8.

Observation

i.e. the average rank is ***increasing*** *in this range to around* 0.869.

Average Rank of Elliptic Curves II : Tension Between Data and Conjecture

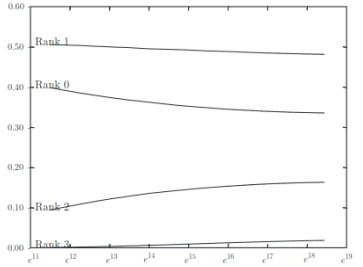


FIGURE 4. Rank Distribution of Stein-Watkins Curves with $N \leq 10^8$

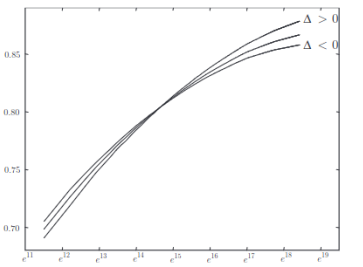


FIGURE 3. Average Rank of Stein-Watkins Curves of Conductor $\leq 10^8$

Average Rank of Elliptic Curves III : Tension Between Data and Conjecture

Observation

In the case of prime conductor curves, they notice that the average rank for conductor $N \leq 10^{10}$ is **decreasing** to about 0.965.

This was at the time one of the principle argument towards the minimalist conjecture.

-This large mass of rational points for elliptic curves of prime conductor $\leq 10^{10}$ is palpably there. We aren't in the dark about that. We are merely in the dark about how to give a satisfactory account of it being there, other than computing instances, one after another. We are, in a word, just at the very beginning of this story.

Remark

Alvaro Lozano-Robledo : Built a probabilistic model which mimics the initial bias in the data and which predicts that the average rank should be around 0.5032 in the range of ... 10^{100} and 0.500006 at 10^{200}.

Elliptic Curves and Bias

BSD invariants are still very mysterious. A natural question to ask is whether or not AI could help detect and understand the bias that may exist among them.

For example, we have seen the existence of this bulk of curves with infinitely many rational solutions.

Bennett-Gherga-Retchnizer Dataset I

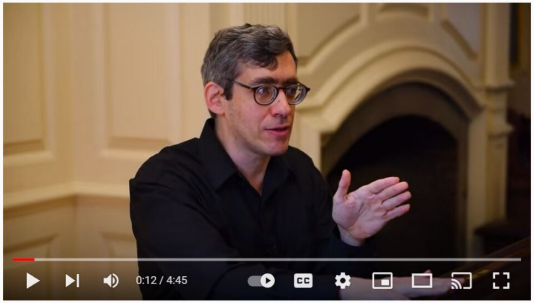
New dataset containing (with very high likelihood) the Weierstrass coefficients only of all prime conductor curves with $p \leq 2 \cdot 10^{13}$. Incompleteness of the data would imply a Hall ratio

$$\mathcal{H}_{c_4,c_6} = \frac{|c_4|^{1/2}}{|c_4^3 - c_6^2|} > 1.5 \times 10^6$$

Today's world record is own by no one else than ... (funny or not so funny joke to come)

Bennett-Gherga-Retchnizer Dataset II

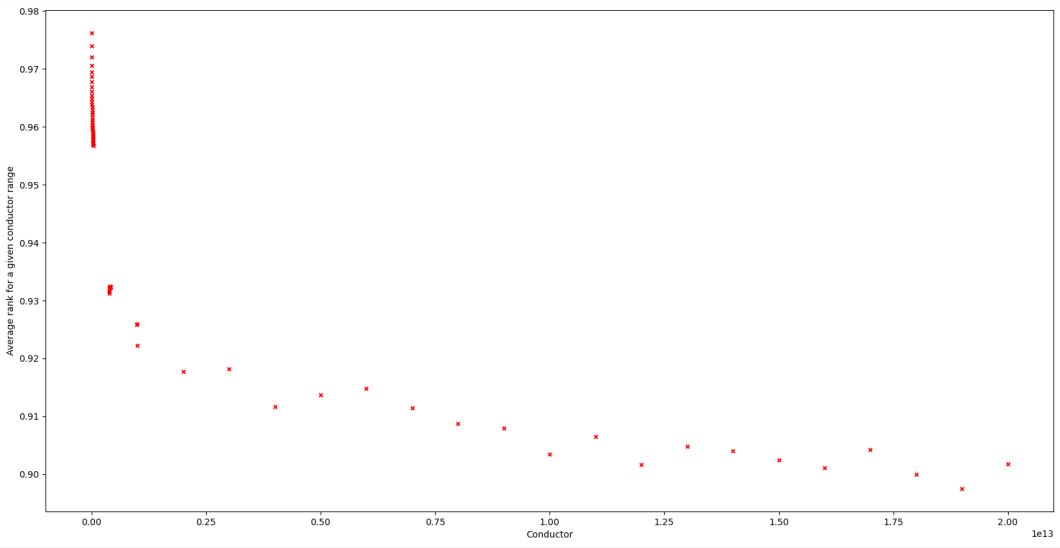
NOAMMMMMMM.... ELKIES ! With a Hall ratio of $\mathcal{H}_{c_4,c_6} \sim 46.600$



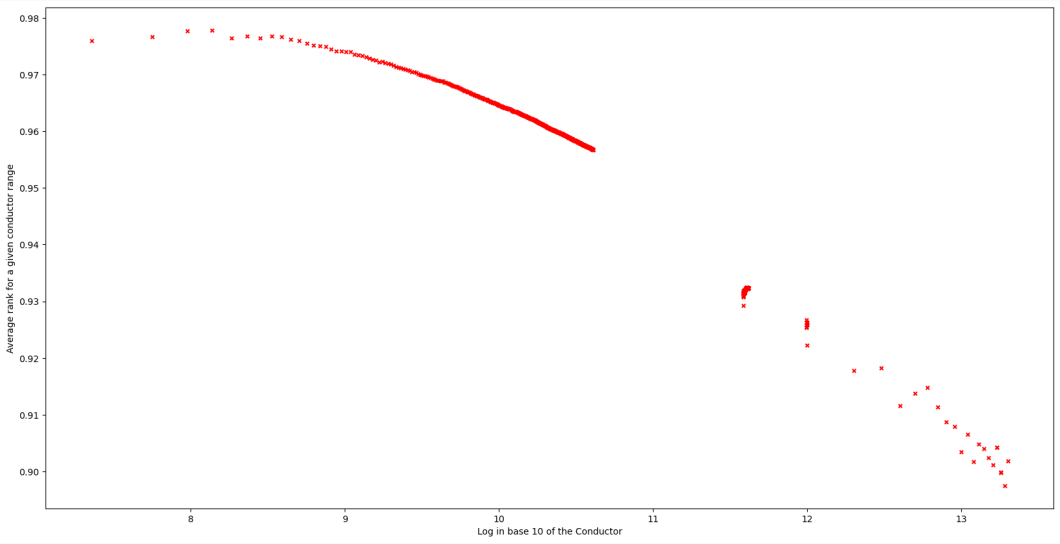
Quote

It is very important when you discover world records in mathematics to imitate Perelman and be as lowkey as possible.

Ranks of Elliptic Curves of Prime Conductor and the BGR Dataset I



Ranks of Elliptic Curves of Prime Conductor and the BGR Dataset II

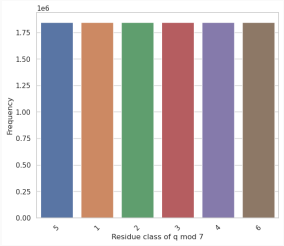


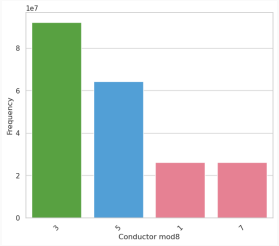
Bias in Conductors mod n I

Question

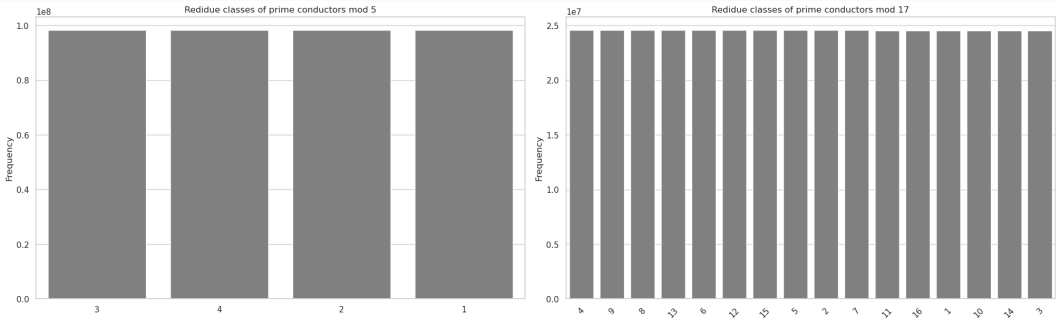
If we were to order elliptic curves by conductor (in our case $N_E \leq 2 \times 10^{12}$), would we observe a bias in the choice of prime conductors modulo a certain n ?

This seems to be the case !

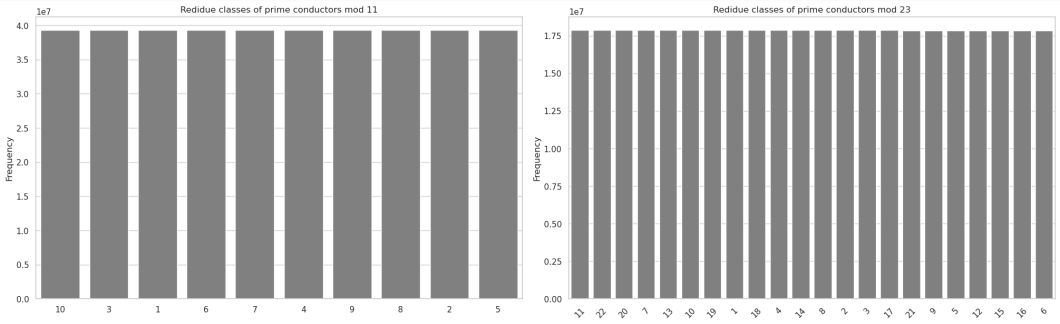




Bias in Conductors mod n II



Bias in Conductors mod n III

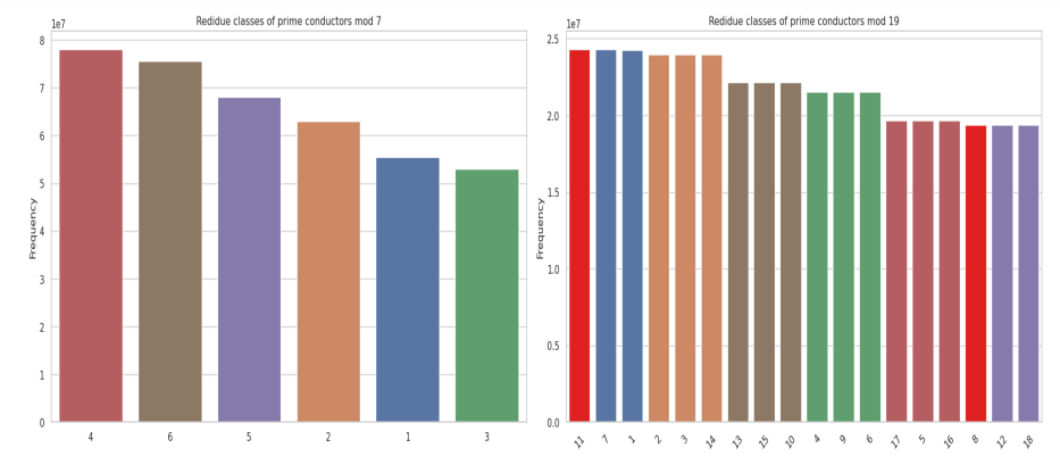


Bias in Conductors mod p I

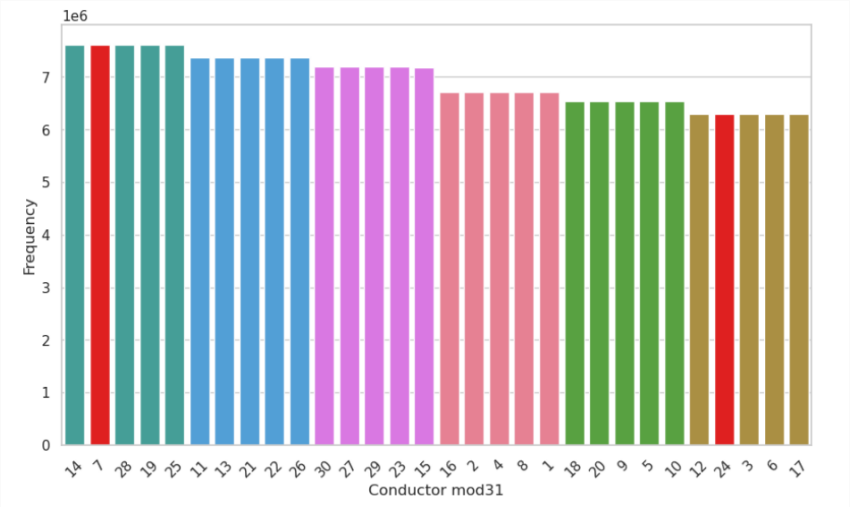
If we consider the distributions of conductors mod p, there are two types of behaviors that these distributions can take depending on whether or not the prime is 1 or 5 mod 6 (equivalently if -1 is a cube modulo p) .

- If p is 5 mod 6, we find that the distributions are perfectly uniform.
- If p is 1 mod 6, we find that distributions are not uniform. There are exactly 6 groups which appear that contain $\phi(p)/6$ residues mod p.

Bias in Conductors mod p II



Bias in Conductors mod p III



Bias in Conductors mod p IV

Question

When $p \equiv 7 \mod 12$, why is it the case that the coset containing the residue class $p - n$ is diametrically opposed to the coset of the class of n in terms of frequency?

Question

Why do the cosets appear with these frequencies? What distinguishes the most frequent coset from the other depending on p?

Question

The modularity theorem allows us to draw a parallel between an elliptic curve of conductor N_E and modular forms of weight 2 on the Hecke subgroup $\Gamma_0(N_E)$. Since N_E is a prime, these forms are newforms.

Hence, there is an intimate connection between the bias on the elliptic curves side that should be reflected on the modular forms side (the distribution of the dimensions of the spaces $S_2(\Gamma_0(N_E))$).

Claim

There is a higher likelihood of having a newform at level p when $p \equiv 4 \mod 7$ as opposed to $1 \mod 7$.

Machine Learning I : Learning the Rank using a_p's (again ...)

Question

Is it harder or simpler to classify the rank of prime conductor curves using a_p coefficients ? A priori, shouldn't be any difference except that curves may be more sparsed in terms of the conductor's value.

Machine Learning II : The experiment

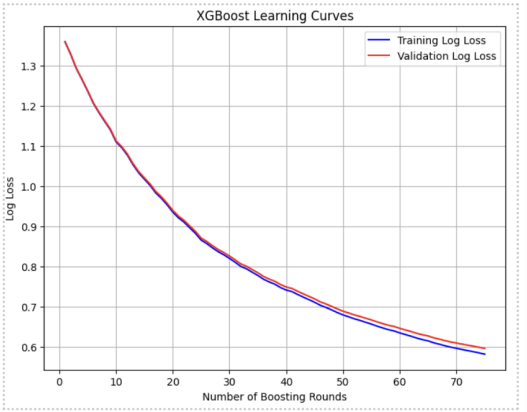
- Goal : classify rank 0,1,2,3 curves.
- Conductor range : $5 \times 10^9 \leq N_E \leq 1.1 \times 10^{10}$.
- Selection process : randomly picking 30 000 curves of each rank.
- Feature selection :
  - We consider the prime index coefficients a_p for $p < 10^4$. We replace each $a_p = 0$ value with the average of its column.
  - Then we use GBM's feature selection method to compute feature importance and pick the best 100 of them.
  - We used a buffer to inform our model about the relative position of the coefficients in the form of a weight : $-1 + 2n/303$ and $n = 1..303$ (much more values than necessary).
  - Then, we divide each selected a_p value column with the corresponding value of the buffer in absolute value $a_p/\sqrt{|-1 + 2n/303|}$.

Machine Learning III : The Model

- Model : LightGBM
- Why ? : Fast, memory efficient, don't need much data to generalize well, and boosting trees are incredible for tabular data.
- Split of the data : We use a train/test/val split of 40/30/30.

Machine Learning IV : Results

- Results : Accuracy : 98.20% and R^2 Score: 0.9705



Acknowledgements

Acknowledgements

Thank you for your attention and thank you to the organizing committee for this great workshop !

If you have any questions please let me know (unless these are difficult questions ☹).