Data–Driven Patterns in Elliptic Curves

Dulce Rodriguez and Daniel Ramos
University of California Santa Barbara

What Are Elliptic Curves?

An elliptic curve over $\mathbb{Q}$ is a smooth cubic projective curve E defined over $\mathbb{Q}$, with at least one rational point $\mathcal{O} \in E(\mathbb{Q})$ that we call the origin. For simplicity, we concentrate on trying to find all rational points on a curve

$$E(\mathbb{Q}) : y^2 = x^3 + ax + b, \quad a, b \in \mathbb{Q}$$

with discriminant $\Delta = -16(4a^3 + 27b^2) \neq 0$ to ensure non-singularity. The set of rational points $E(\mathbb{Q})$, defined as

$$\{(x, y) \in E \mid x, y \in \mathbb{Q}\} \cup \{\mathcal{O}\}$$

where $\mathcal{O}$ = [0,1,0] is the point at infinity. This set forms a finite generated abelian group:

$$E(\mathbb{Q}) \cong \mathbb{Z}^r \oplus E(\mathbb{Q})_{\text{tors}}$$

where r is the rank (infinite-order points) and $E(\mathbb{Q})_{\text{tors}}$ is the finite torsion subgroup.
Elliptic curves play a central role in number theory, cryptography, and algebraic geometry.

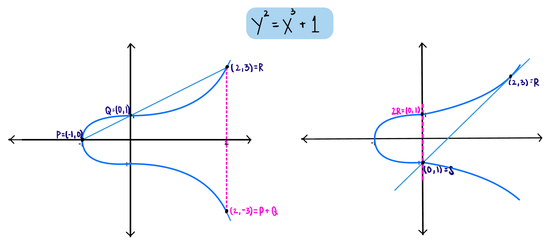
How Does Point Addition Work?

Point addition and doubling gives elliptic curves their group structure:
Point Addition: Given two distinct points P and Q on an elliptic curve, their sum $P + Q$ is defined as follows:

- Draw the straight line that intersects P and Q.
- This line will generally intersect the curve at a third point, say R.
- Reflect R across the x-axis to get $P + Q$.

Point Doubling ($2P$): If $P = Q$, we take the tangent line at P and repeat the same process:

- Compute where the tangent at P intersects the curve again.
- Reflect that intersection point across the x-axis to obtain $2P$.



Why Rank and Torsion Matter

When the **Rank** is zero, there are only a few rational points. On the other hand, if the rank is at least one, there are infinitely many, and these can be added together to create new points. The rank is at the center of the famous **Birch and Swinnerton-Dyer Conjecture**, which predicts how the rank connects to deeper properties of the curve.
Torsion subgroup is the part of the curve that contains points that eventually "loop back" to the identity when added to themselves a finite number of times (finite-order points). Over the rational numbers, the torsion subgroup is always one of just 15 possible types, thanks to a result called **Mazur's Theorem**. Both rank and torsion give us insight into the curve's structure. We explored patterns in torsion and discriminant values and their relation to rank.

Our Data Science Approach

Goal: Explore statistical and predictive relationships among rank, torsion, and discriminant.
Data Source: Sample of 1 million elliptic curves from the 3.8 million available in the LMFDB (L-Functions and Modular Forms Database), queried using the lmfdb-lite Python library.
Key Fields:

Rank (r), Discriminant (Δ), Torsion structure, Conductor (N), Maximal Rank

Workflow:

1. Connect to LMFDB mirror via 'psycopg2'
2. Extract and clean data using 'pandas'
3. Apply log-scaling to $|\Delta|$, one-hot encode torsion types
4. Build regression and classification models (Random Forest, Logistic Regression)

Statistical Modeling

Figure 1

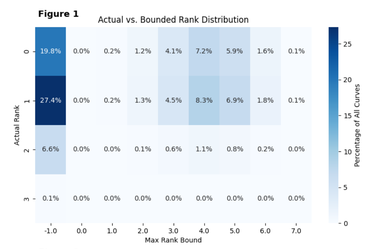


Figure 2

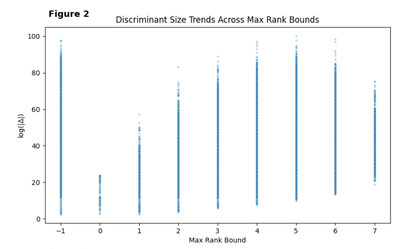


Figure 3

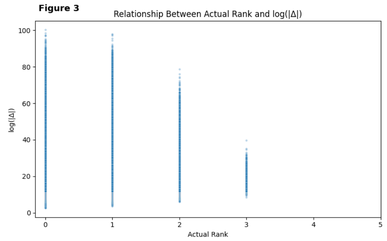
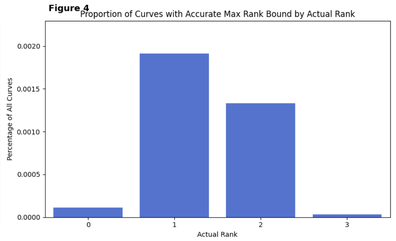


Figure 4



Graph Note: The "1" column appears only in the top two graphs and represents curves that don't satisfy the max rank condition. It's included for reference.

Model classification performance on elliptic curve rank prediction

Logistic Regression

Class	Precision	Recall	Support
0 (rank 0)	0.45	0.57	119,633
1 (rank > 0)	0.65	0.53	180,367
Accuracy	0.56		300,000
Weighted avg	0.57	0.55	300,000

Decision Tree

Class	Precision	Recall	Support
0 (rank 0)	0.44	0.64	119,633
1 (rank > 0)	0.66	0.46	180,367
Accuracy	0.53		300,000
Weighted avg	0.57	0.53	300,000

Max Rank

The **rank** is one of the most important properties describing the structure of an elliptic curve. There isn't a method or formula that computes the rank for elliptic curves due to its complexity. However, we can compute the **max rank** under certain conditions to create an interval given by:

$$0 \leq \operatorname{rank}(E) \leq \max \operatorname{rank}(E).$$

This allows us to narrow down our understanding of the elliptic curves rank. The bound is found by using the following statement. Let $E/\mathbb{Q}$ be any elliptic curve with a non-trivial point of 2-torsion, and let a (resp. m) be the number of primes of additive (resp. multiplicative) bad reduction of $E/\mathbb{Q}$. Then:

$$\operatorname{rank}_{\mathbb{Z}}(E(\mathbb{Q})) \leq m + 2a - 1.$$

The max rank used in the graphs was calculated using an algorithm we developed that involved the curves' numerical properties found in the database.

Findings and Discussion

Graph Discussion

Figure 1: We can see the majority of the curves in our sample fall in the "-1" column, further demonstrating how difficult obtaining knowledge about the rank can be. Interestingly, rank 0 and 1 curves had the highest frequency of bounded ranks in our sample, specifically a max range of 3, 4 or 5.
Figure 2: We can see the size of the "-1" column's Δ stretches the entire y-axis, exhibiting no apparent trend. However, the interval where the discriminant lies increases in size and value-wise as the max rank bound increases.
Figure 3: Here we see the graph exhibits a downwards trend. Large Δ tend to trigger more primes of bad reduction, the variables used in computing the max rank bound. This appears to overestimate their rank significantly.
Figure 4: This table illustrates the tremendous unlikelihood of an elliptic curves rank being equal to their bound. This refers to the **percentage** of curves in the sample whose actual rank matches their bound.

Table Insights

We applied several models that classified an elliptic curve as having rank 0 or greater than 0 based on the discriminant, max rank, and the torsion order. We chose to display the 2 most efficient models, a logistic regression and a binary decision tree. Advanced methods such as Support Vector Machines were not very effective as they require heavy computing power and smaller sample sizes were ineffective. The imbalance of curves with rank 0/rank above 0, limited us to methods that were efficient with large samples and accounted for imbalanced data. To facilitate **table interpretation**, we've included some definitions. **Precision:** out of the predictive positives, how many were actually correct. **Recall:** out of the actual positives, how many did the model guess correctly.
Limitations: Rank is difficult to predict precisely; the BSD conjecture remains unsolved.

Acknowledgments

Special thanks to our DRP mentor, Marcos Reyes, the UCSB Direct Reading Program, and the developers of LMFDB and 'lmfdb-lite'.

References

He, Y.-H., Lee, K.-H., Oliver, T. (2022). Machine-learning arithmetic curves. arXiv.https://arxiv.org/abs/2203.13705
Lozano-Robledo, Á. (2011). Elliptic curves, modular forms, and their L-functions (Vol. 58). American Mathematical Society.