# Statistics in Number Theory

Kimball Martin

Osaka Metropolitan University

June 12, 2025

# Outline

Explore statistical behavior of discrete collections of objects in number theory (**arithmetic statistics**):

I. Primes (GL(1))
II. Elliptic Curves (GL(2))
III. Modular Forms (GL(2))

Themes:

- Compare actual behavior with random models
- Computational data is useful, but (almost) never enough!

# I. Primes: basic distribution

$\mathcal{P} = \{\text{primes } p\} = \{2, 3, 5, 7, 11, ...\}$
— Primes are building blocks of arithmetic, but there is no simple formula to describe $\mathcal{P} \subset \mathbb{N}$
— Study distribution and patterns

- **Conjecture** (Legendre $\approx 1797$):

$$\pi(x) := \# \{p \in \mathcal{P} : p < x\} \sim \frac{x}{\log x}$$

- **Prime Number Theorem** (Hadamard, de la Vallée Pousson 1896): true!
- equivalently, the $n$-th prime $p_n \approx n \log n$

# I. Primes: a probabilistic "model"

- **Prime Number Theorem:**

$$\pi(x) := \# \{p \in \mathcal{P} : p < x\} \sim \frac{x}{\log x}$$

i.e., the $n$-th prime $p_n \approx n \log n$

- **Random model** (Cramér 1936):

$$\mathrm{Prob}(p \in \mathcal{P}) \approx \frac{1}{\log x} \quad \text{if} \quad p \approx x$$

- model suggests *Cramér's conjecture*: $p_{n+1} - p_n \ll (\log n)^2$ (gaps between primes cannot get too big, still open)

(Asymptotic inequality notation: $f(n) \ll g(n)$ means $f(n) = O(g(n))$, i.e., there exists $C, N$ such that $f(n) \leq Cg(n)$ for $n > N$)

# I. Primes: distribution mod $m$

Random model predicts: primes are "equidistributed" mod $m$

$m = 10$ :

| ⋆ | | ⋆ | | | | ⋆ | | ⋆ | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | **2** | **3** | 4 | **5** | 6 | **7** | 8 | 9 | 10 |
| **11** | 12 | **13** | 14 | 15 | 16 | **17** | 18 | **19** | 20 |
| 21 | 22 | **23** | 24 | 25 | 26 | 27 | 28 | **29** | 30 |
| **31** | 32 | 33 | 34 | 35 | 36 | **37** | 38 | 39 | 40 |
| **41** | 42 | **43** | 44 | 45 | 46 | **47** | 48 | 49 | 50 |
| 51 | 52 | **53** | 54 | 55 | 56 | 57 | 58 | **59** | 60 |
| **61** | 62 | 63 | 64 | 65 | 66 | **67** | 68 | 69 | 70 |
| **71** | 72 | **73** | 74 | 75 | 76 | 77 | 78 | **79** | 80 |

**Prime Number Theorem for arithmetic progressions (1896):**
Let $\phi(m) = \#\{1 \leq a < m : \gcd(a,m) = 1\}$. If $\gcd(a,m) = 1$, then

$$\#\{p \in \mathcal{P} : p < x \text{ and } p \equiv a \bmod m\} \sim \frac{1}{\phi(m)} \frac{x}{\log x}$$

# I. Primes: biases mod $m$

- **Chebyshev's bias** (1853): numerically, there are more primes 3 mod 4 than 1 mod 4 (up to any given $x$)

| $x$ | $\#\{p < x : 1 \bmod 4\}$ | $\#\{p < x : 3 \bmod 4\}$ |
|---:|---:|---:|
| 100 | 11 | 13 |
| 500 | 44 | 50 |
| 1000 | 80 | 87 |
| 5000 | 329 | 339 |
| 10000 | 609 | 619 |

- first counterexample at $x = 26861$, next at $616841$
- Littlewood (1914): there are infinitely many counterexamples
- **Conjecture** (Knapowski–Turán, 1962): For 100% of $x \in \mathbb{N}$, Chebyshev's bias holds, i.e.,

$$\#\{p < x : p \equiv 3 \bmod 4\} > \#\{p < x : p \equiv 1 \bmod 4\}$$

- **Conjecture** (Knapowski–Turán, 1962): For 100% of $x \in \mathbb{N}$, Chebyshev's bias holds, i.e.,

$$\#\{p < x : p \equiv 3 \bmod 4\} > \#\{p < x : p \equiv 1 \bmod 4\}$$

- Kaczorwoski (1996), Sarnak: False!

$$\frac{\#\{x \leq N : \text{Chebyshev's bias holds}\}}{N} \quad \text{has no limit}$$

- Rubinstein–Sarnak (1994): $S = \{x \in \mathbb{N} : \text{Chebyshev's bias holds}\}$

$$\frac{\sum_{x \leq N : x \in S} \frac{1}{x}}{\sum_{x \leq N} \frac{1}{x}} \to 0.9959\ldots$$

i.e., Chebyshev's bias holds for $\approx 99.59\%$ of $x$ when we use "logarithmic measure" on $\mathbb{N}$

# I. Primes: biases mod $m$ (cont'd $\times 2$)

- Rubinstein–Sarnak (1994): $S = \{x \in \mathbb{N} : \text{Chebyshev's bias holds}\}$

$$\frac{\sum_{x \le N : x \in S} \frac{1}{x}}{\sum_{x \le N} \frac{1}{x}} \to 0.9959\ldots$$

  i.e., Chebyshev's bias holds for $\approx 99.59\%$ of $x$ when we use "logarithmic measure" on $\mathbb{N}$

What is the *reason* for this bias?

– for any odd $n$, $n^2 \equiv 1 \mod 4$

– i.e., numbers 1 mod 4 must contain all odd squares, making them slightly less likely to be prime

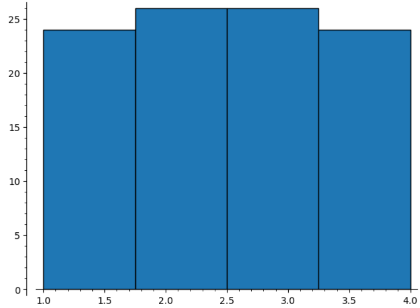
– similar phenomena mod $m$ for any $m \ge 3$*

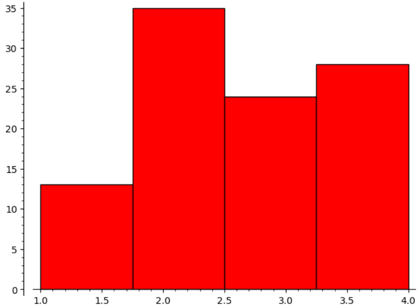
*see: Prime number races by Andrew Granville and Greg Martin, *Amer. Math. Monthly*, 2006

# I. Primes: variance mod $m$

- Cowan (*arXiv:2504.20691*): primes are more equidistributed mod $m$ than random!

Histograms of first 100 primes mod 5 (excluding $p = 5$) versus 100 random numbers coprime to 5




Why???

# II. Elliptic curves: definition

Many definitions...

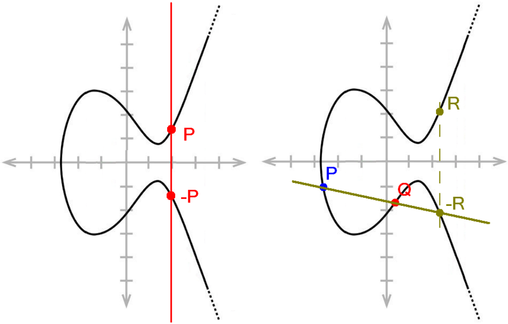
An **elliptic curve** over a field $F$ is any of the following:

- $E : y^2 = x^3 + ax + b$, where $a, b \in F^\dagger$ such that the discriminant $\Delta_E = -(4a^3 + 27b^2) \neq 0$
- a smooth projective cubic curve $E/F$ with a marked point ($O = \infty$ in the above Weierstrass form)
- a smooth projective curve $E/F$ of genus 1, together with a marked point $O$ (defined over $F$)
- a smooth projective curve $E/F$ with an (additive) group structure (addition and negation should be given by rational functions)

† Assuming char $F \neq 2, 3$, otherwise the general form is more complicated

# II. Elliptic curves: group law

$E : y^2 = x^3 + ax + b$ cubic curve
— points form an abelian group




Three points on line sum to $0 = (0, \infty)$:

$P, -P$ on same vertical line $\iff$ $0 + P + (-P) = 0$

$P + Q + (-R) = 0, \quad R = P + Q$

(Images from Wikimedia Commons)

# II. Elliptic curves: motivation

Elliptic curves arise in many interesting number theory problems...

- (Pépin, Lucas, Sylvester $\approx 1879$) Sums of 2 cubes:

$$x^3 + y^3 = n \quad \longleftrightarrow \quad y^2 = x^3 - 432n^2$$

- (Fermat 1637, Frey 1986, Wiles/Taylor–Wiles 1995) Fermat's Last Theorem: $x^n + y^n = z^n$ has no solutions in positive integers for $n > 2$
- (Gauss 1801, Goldfeld 1976) Class number problems, e.g., for which $D$ does $\mathbb{Z}[\sqrt{D}]$ have unique factorization?
- integer factorizations and cryptography
- ...

# II. Elliptic curves: counting

To count elliptic curves over $\mathbb{Q}$

$$E : y^2 = x^3 + ax + b, \quad \Delta_E = -(4a^3 + 27b^2) \neq 0$$

we need to order them...

—first change variables to assume $a, b \in \mathbb{Z}$
—but changing variables changes discriminant $\Delta_E$, so it is not an invariant...

Three standard options:

- Order by height, e.g., $H_E = \max\{|a|, |b|\}$
- Order by absolute discriminant $|\Delta_E|$
- Order by **conductor** $N = N_E$

Conductor is most natural from arithmetic/geometry
Precise definition is technical, but $N$ divides $\Delta_E$

# II. Elliptic curves: minimal example

$$E : y^2 = x^3 + ax + b, \quad \Delta_E = -(4a^3 + 27b^2) \neq 0$$

"Smallest" example:

$$E = E_{11a3} \simeq X_0(11) : y^2 = x^3 - 432x + 8208$$

$$\Delta_E = -2^8 \cdot 3^{12} \cdot 11$$

Actually, has another cubic expression with smaller discriminant

$$E' : y^2 + y = x^3 - x^2, \quad \Delta_{E'} = -11$$

Conductor:

$$N_E = N_{E'} = 11$$

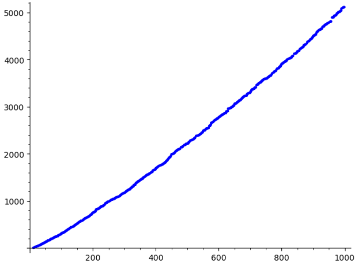
Meaning: $E$ mod 11 has a nodal singularity

# II. Elliptic curves: computing

- Cremona (1992): algorithm for enumerating all *modular* elliptic curves with conductor $N \leq X$
- Wiles/Taylor–Wiles (1995), Breuil–Conrad–Diamond–Taylor (2001): all elliptic curves are modular
- Cremona's 1992 database: 5113 curves with $N \leq 999$
- current status: 3,064,705 curves with conductor $N \leq$ 500,000 (complete database)

# II. Elliptic curves: counting (cont'd)

Graph $\#\{E : N \leq X\}$ on Cremona's original database ($X < 1000$)




- looks like number of curves with $N \leq X$ grows faster than $X$
- best fit exponent for $\#\{E : N \leq X\} \approx cX^d$ is $d \approx 1$
- Duke–Kowalski (2000): upper bound $\#\{E : N \leq X\} \ll X^{1+\varepsilon}$

# II. Elliptic curves: counting (cont'd $\times 2$)

**Conjecture:** (Brumer–McGuinness 1990/Watkins 2008):

$$\#\{E : N \le X\} \sim cX^{5/6}$$

- Idea: generically expect $N \le |\Delta_E| \le CN$ for some constant $C$
- so expect: $\#\{E : N \le X\} \sim c\#\{E : |\Delta_E| \le X\}$
- for $E : y^2 = x^3 + ax + b$ $(a, b \in \mathbb{Z})$, expect

$$|\Delta_E| = \left|4a^3 + 27b^2\right| \ll X \overset{??}{\iff} \max\left\{|a|^3, b^2\right\} \ll X$$
$$\iff |a| \ll X^{1/3}, \quad |b| \ll X^{1/2}$$

- so get $\ll X^{1/3} \cdot X^{1/2} = X^{5/6}$ possibilities for $\{(a, b)\}$
- $\#\{E : N \le X\} \gg X^{5/6}$ is easy

# II. Elliptic curves: counting – summary

**Conjecture:** (Brumer–McGuinness 1990/Watkins 2008):

$$\#\{E : N \leq X\} \sim cX^{5/6}$$

- Theoretical bounds

$$X^{5/6} \ll \#\{E : N \leq X\} \ll X^{1+\varepsilon}$$

- data suggests closer to $X^1$
- extensive data for prime $N <$ 2,000,000,000 (3,218,940 curves) agrees with $X^{5/6}$ (Bennett–Gherga–Rechnitzer 2019)
- we believe convergence to $X^{5/6}$ is very slow because of many "excess curves" of "small" conductor
- tension between data and theory/heuristics is prominent in this area

# II. Elliptic curves: distributions of conductors

- Are conductors $N$ distributed randomly?

– no, there are some restrictions on $N$ (e.g., $p^3 \nmid N$ for $p \geq 5$)

– conductors tend to cluster together (many elliptic curves with same $N$)

– numerically there are many more even conductors than odd conductors

# II. Elliptic curves: counting by rank

Count elliptic curves (f.g. abelian groups!) with certain properties, such as:
**rank**$(E) := r$ where $E(\mathbb{Q}) \simeq \mathbb{Z}^r \oplus$ (finite group)

- **Minimalist Conjecture** (1980s?): 50% of elliptic curves have rank 0, and 50% have rank 1
  – numerically it appears many more have rank 1 than rank 0, and a positive proportion have rank 2, 3 or 4
- **conjecture** (Néron 1950): ranks of elliptic curves are bounded
  – rank $\geq 4$ was known (Wiman 1945)
- **conjecture** (Cassels 1966, Tate 1974, Mestre 1982): ranks of elliptic curves are unbounded
  – rank $\geq 12$ was known (Mestre 1982)
- **Conjecture** (Park–Poonen–Voight–Wood 2019): ranks of elliptic curves are bounded
  – rank $\geq 28$ was known (Elkies 2006)
  – now $\geq 29$ is known (Elkies–Klagsbrun 2024)

# III. Modular forms: modular curves

$\mathfrak{H} = \{z \in \mathbb{C} : \mathsf{Im}(z) > 0\}$ — upper half plane
$\mathrm{SL}_2(\mathbb{R})$ acts on $\mathfrak{H}$ by linear fractional transformations:
$\begin{pmatrix} a & b \\ c & d \end{pmatrix} z = \frac{az+b}{cz+d}$
(orientation-preserving isometries of the hyperbolic plane)

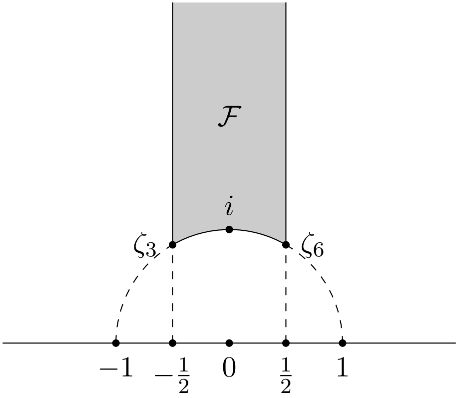
**Congruence subgroups:**

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) : N \text{ divides } c \right\}$$

**Modular curves:**

$Y_0(N) = \Gamma_0(N) \backslash \mathfrak{H}$
$X_0(N) = Y_0(N) \cup \{\text{cusps}\}$ — compact Riemann surface
— parametrize elliptic curves$/\mathbb{C}$ with a cyclic subgroup of order $N$

# III. Modular forms: $X_0(1) = \mathrm{SL}_2(\mathbb{Z})\backslash\mathfrak{H}$

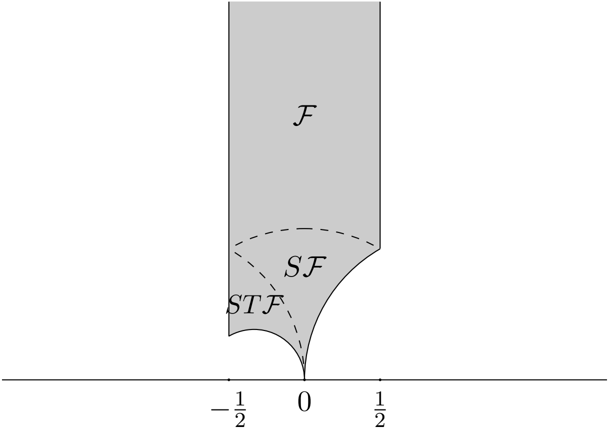



$$\mathrm{SL}_2(\mathbb{Z}) = \langle S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \rangle$$

$$z \mapsto -1/z, \qquad z \mapsto z+1$$

Cusps: $i\infty$

# III. Modular forms: $X_0(2) = \Gamma_0(2)\backslash\mathfrak{H}$




$$\Gamma_0(2) = \left\{ \begin{pmatrix} a & b \\ 2c & d \end{pmatrix} : ad - 2bc = 1 \right\} \subset \mathrm{SL}_2(\mathbb{Z})$$

Cusps: $0, i\infty$

# III. Modular forms: definition

$M_k(N)$ – (holomorphic) modular forms of weight $k$ and level $N$:

$$f : \mathfrak{H} \to \mathbb{C}$$
$$f(\begin{pmatrix} a & b \\ c & d \end{pmatrix} z) = (cz+d)^k f(z), \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N)$$

$+$ growth conditions

$M_k(N)$ – finite dimensional vector space, trivial unless $k \geq 2$ even

$f(z+1) = f(\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} z) = f(z) \rightsquigarrow$ Fourier expansion:

$$f(z) = \sum_{n=0}^{\infty} a_n q^n, \quad q = e^{2\pi i z}.$$

Fourier coefficients are arithmetically interesting...

# III. Modular forms: examples

**(Eisenstein series)** $k \geq 4$, $\sigma_j(n) = \sum_{d|n} d^j$

$$E_k(z) = \sum_{(c,d) \in \mathbb{Z}^2 - 0} \frac{1}{(cz+d)^k}$$
$$= 2\zeta(k) + 2\frac{(2\pi)^k}{(k-1)!} \sum_{n=1}^{\infty} \sigma_{k-1}(n) q^n \in M_k(1)$$

**(theta series)** $r_{2k}(n) = \# \left\{ \text{integer solns to } x_1^2 + \cdots + x_{2k}^2 = n \right\}$

$$\vartheta^{2k}(n) = (\sum_{n \in \mathbb{Z}} q^{n^2})^{2k} = \sum_{n \geq 0} r_{2k}(n) q^n \in M_k(4)$$

Sample corollary: $r_8(n) = 16(\sigma_3(n) - 2\sigma_3(\frac{n}{2}) + 16\sigma_3(\frac{n}{4}))$

# III. Modular forms: weight 2

$$f : \mathfrak{H} \to \mathbb{C}$$

$$f(\begin{pmatrix} a & b \\ c & d \end{pmatrix} z) = (cz+d)^2 f(z), \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N)$$

$$M_2(N) = \mathrm{Eis}_2(N) \oplus S_2(N)$$

$S_2(N)$ - cusp forms (holomorphic differentials on $X_0(N)$)

- there is a canonical generating set[‡] $S_2(N)$ consisting of **primitive** (or **new**) forms $f = \sum a_n q^n$
- the Fourier coefficients $a_n$'s of primitive forms are multiplicative: $a_{mn} = a_m a_n$ when $\gcd(m, n) = 1$

[‡]a basis if $N$ is prime

# III. Modular forms and elliptic curves

$$f = \sum a_n q^n = q + a_2 q^2 + a_3 q^3 + \dots \quad \text{primitive}$$

- **rationality (Fourier coefficient) field:** $K_f = \mathbb{Q}(a_2, a_3, \dots)$
- $[K_f : \mathbb{Q}]$ finite

**Theorem (Modularity, Breuil–Conrad–Diamond–Taylor 2001)**

$$\{f \in S_2(N) : \textit{primitive}, K_f = \mathbb{Q}\} \longleftrightarrow \{E : \textit{ell. curve of conductor } N\}$$
$$a_p = p + 1 - \#\{E \bmod p\} \quad (p \nmid N)$$

What about other (weight 2) forms?

- Shimura (1959): $f \in S_2(N)$ primitive $\implies$ *abelian variety* of dimension $d = [K_f : \mathbb{Q}]$ with multiplication by $K_f$ and conductor $N^d$
- More general modularity (Ribet 2004, Khare–Winterberger 2009, Kisin 2009): $\Longleftarrow$

$L$-functions and Modular Forms DataBase (LMFDB) – *lmfdb.org*

| Label | Dim | $A$ | Field | Traces | | | | Fricke sign | $q$-expansion |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $a_2$ | $a_3$ | $a_5$ | $a_7$ | | |
| 11.2.a.a | 1 | 0.088 | $\mathbb{Q}$ | $-2$ | $-1$ | 1 | $-2$ | $-$ | $q - 2q^2 - q^3 + 2q^4 + q^5 + 2q^6 - 2q^7 + \cdots$ |
| 14.2.a.a | 1 | 0.112 | $\mathbb{Q}$ | $-1$ | $-2$ | 0 | 1 | $-$ | $q - q^2 - 2q^3 + q^4 + 2q^6 + q^7 - q^8 + \cdots$ |
| 15.2.a.a | 1 | 0.120 | $\mathbb{Q}$ | $-1$ | $-1$ | 1 | 0 | $-$ | $q - q^2 - q^3 - q^4 + q^5 + q^6 + 3q^8 + \cdots$ |
| 17.2.a.a | 1 | 0.136 | $\mathbb{Q}$ | $-1$ | 0 | $-2$ | 4 | $-$ | $q - q^2 - q^4 - 2q^5 + 4q^7 + 3q^8 - 3q^9 + \cdots$ |
| 19.2.a.a | 1 | 0.152 | $\mathbb{Q}$ | 0 | $-2$ | 3 | $-1$ | $-$ | $q - 2q^3 - 2q^4 + 3q^5 - q^7 + q^9 + 3q^{11} + \cdots$ |
| 20.2.a.a | 1 | 0.160 | $\mathbb{Q}$ | 0 | $-2$ | $-1$ | 2 | $-$ | $q - 2q^3 - q^5 + 2q^7 + q^9 + 2q^{13} + \cdots$ |
| 21.2.a.a | 1 | 0.168 | $\mathbb{Q}$ | $-1$ | 1 | $-2$ | $-1$ | $-$ | $q - q^2 + q^3 - q^4 - 2q^5 - q^6 - q^7 + \cdots$ |
| 23.2.a.a | 2 | 0.184 | $\mathbb{Q}(\sqrt{5})$ | $-1$ | 0 | $-2$ | 2 | $-$ | $q - \beta q^2 + (-1+2\beta)q^3 + (-1+\beta)q^4 + \cdots$ |
| 24.2.a.a | 1 | 0.192 | $\mathbb{Q}$ | 0 | $-1$ | $-2$ | 0 | $-$ | $q - q^3 - 2q^5 + q^9 + 4q^{11} - 2q^{13} + \cdots$ |
| 26.2.a.a | 1 | 0.208 | $\mathbb{Q}$ | $-1$ | 1 | $-3$ | $-1$ | $-$ | $q - q^2 + q^3 + q^4 - 3q^5 - q^6 - q^7 + \cdots$ |
| 26.2.a.b | 1 | 0.208 | $\mathbb{Q}$ | 1 | $-3$ | $-1$ | 1 | $-$ | $q + q^2 - 3q^3 + q^4 - q^5 - 3q^6 + q^7 + \cdots$ |
| 27.2.a.a | 1 | 0.216 | $\mathbb{Q}$ | 0 | 0 | 0 | $-1$ | $-$ | $q - 2q^4 - q^7 + 5q^{13} + 4q^{16} - 7q^{19} + \cdots$ |
| 29.2.a.a | 2 | 0.232 | $\mathbb{Q}(\sqrt{2})$ | $-2$ | 2 | $-2$ | 0 | $-$ | $q + (-1+\beta)q^2 + (1-\beta)q^3 + (1-2\beta)q^4 +$ |

— $a_2$'s start off negative [Farmer–Koutsoliotas 2016]

— Fricke sign tends to be $-1$ [M 2018, 2023]

— rationality (Fourier coefficient) field is often $\mathbb{Q}$ [misleading]

# III. Modular forms: counting by rationality field

- **Question:** Fix $K/\mathbb{Q}$. How many primitive *minimal* weight 2 $f$ are there with level $N < X$ with $K_f = K$?
- **Conjecture** (Brumer–McGuinness 1990/Watkins 2008): $\sim cX^{5/6}$ if $K = \mathbb{Q}$

## Conjecture (Cowan–M)

*If* $[K : \mathbb{Q}] = d$*, this count is*

$$\begin{cases} \ll X^{2/3+\varepsilon} & d = 2 \\ \ll X^{1/2+\varepsilon} & d = 3 \\ \ll X^{1/3+\varepsilon} & d = 4 \\ \ll X^{1/6+\varepsilon} & d = 5 \\ \ll X^{\varepsilon} & d = 6 \\ \textit{finite} & d \geq 7 \end{cases}$$