# The Riemann Hypothesis for Elliptic Curves

Jasbir S. Chahal and Brian Osserman

## 1 INTRODUCTION

The Riemann zeta function $\zeta(s)$ is defined, for $\mathrm{Re}(s) > 1$, by

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}, \tag{1}$$

and extended analytically to the whole complex plane by a functional equation (see [8, p. 14]). The original Riemann hypothesis asserts that the nonreal zeros of the Riemann zeta function $\zeta(s)$ all lie on the line $\mathrm{Re}(s) = 1/2$. In his monumental paper [11] of 1859, Riemann made this assertion in order to derive an expression for the deviation of the exact number of primes $\leq x$, which is denoted by $\pi(x)$, from the estimate $x/\log x$ that had been conjectured by Gauss, Legendre, and others. Riemann alluded to returning to this matter later by saying that he was "setting it aside for the time being." Apparently Riemann did not live long enough to do that. To this day, no one has been able to prove the Riemann hypothesis despite overwhelming numerical evidence in its favor. However, many generalizations and analogs of the Riemann zeta function have been formulated by, among others, Dirichlet, Dedekind, E. Artin, F. K. Schmidt, and Weil, and the Riemann hypothesis has been shown to be true in some of these cases.

One such case is the Riemann hypothesis for elliptic curves, originally conjectured by E. Artin (see [1, pp. 1–94]) and proved by Hasse, and therefore also known as Hasse's theorem. We begin by laying out the statement of this result in Section 2 below. We then turn to the two main topics of this article: i) a brief explanation of the fact that these two Riemann hypotheses are not only closely analogous, but indeed two examples of a single more general framework; and ii) an elementary proof of the Riemann hypothesis

for elliptic curves over finite fields. This is carried out in Sections 3 and 4 respectively, and these may be read independently of one another.

Our proof is based on an idea of Manin. The presentation is self-contained except for an appeal to the "Basic Identity," which is a technical lemma stated in (19) below. The proof of the Basic Identity, although somewhat complicated, is completely elementary (see [5]) and is the least illuminating part of our proof of this Riemann hypothesis.

# 2 THE STATEMENT

Fix a prime $p$. For each integer $r \geq 1$, there is a unique finite field $\mathbb{F}_q$ having $q = p^r$ elements. For simplicity of exposition, from now on we assume that the prime $p$ is not 2 or 3. We may then take our elliptic curve to be defined by the *Weierstrass equation*

$$y^2 = x^3 + ax + b \quad (a, b \in \mathbb{F}_q) \tag{2}$$

with $4a^3 + 27b^2 \neq 0$, so that the cubic has no multiple roots (this ensures that the corresponding curve has no singularities).

The Riemann hypothesis for curves over finite fields has several equivalent formulations, of which we give two here (see [14, Section 9.6] for additional ones). If we denote an elliptic curve by $E$, we can (provisionally) define its zeta function by the following formula:

$$Z_E(t) = \frac{1 - a_q(E)t + qt^2}{(1-t)(1-qt)}, \tag{3}$$

where the dependence of $Z_E(t)$ on $E$ appears in the coefficient $a_q = a_q(E)$ of the numerator of $Z_E(t)$. Here $a_q = q - N_q$, with $N_q =$ the number of solutions of (2) in $\mathbb{F}_q$. [We give an equivalent formula for $Z_E(t)$ more obviously connected to the Riemann zeta function in Section 3 below.]

The *Riemann hypothesis* for $E$ is then the assertion that if $Z_E(q^{-s}) = 0$, then $\mathrm{Re}(s) = 1/2$. However, in order to prove this assertion, we will want to rephrase the Riemann hypothesis as a bound on $a_q$.

Why is it natural to expect a bound on $a_q$? Suppose that the values of $x^3 + ax + b$ were evenly distributed over $\mathbb{F}_q$ as $x$ varied. We would get one point of $E$ when $x^3 + ax + b = 0$. Because $q$ is odd, half of the $q - 1$ nonzero values of $x^3 + ax + b$ would be nonsquares, and give no points of $E$. For the

other half, we would have $(\pm y)^2 = x^3 + ax + b$ for some $y \in \mathbb{F}_q$, giving two points of $E$ for each of the $\frac{q-1}{2}$ values of $x$. Thus the expected value of $N_q$ is $1 + 2 \cdot \frac{q-1}{2} = q$, and $a_q$ is the deviation of $N_q$ from its expected value.

With this motivation, we claim that in fact the bound

$$|N_q - q| \leq 2\sqrt{q} \tag{4}$$

for $a_q$ is equivalent to the Riemann hypothesis for $E$.

Indeed, if $Z_E(q^{-s}) = 0$, then we see that $q^s$ is a root of the polynomial

$$f(u) = u^2 - a_q u + q.$$

Note that the inequality (4) holds if and only if the discriminant $a_q^2 - 4q$ of $f(u)$ is $\leq 0$, which is true if and only if the two roots $u_1$, $u_2$ of $f(u)$ are either strictly complex, or equal to one another. This in turn is equivalent to having $|u_1| = |u_2|$. Since the constant term $q$ of $f(u)$ is the product $u_1 u_2$, we see (4) holds if and only if both roots of $f(u)$ have absolute value $\sqrt{q}$, if and only if for all $s$ with $Z_E(q^{-s}) = 0$, we have $|q^s| = \sqrt{q}$, and thus $\mathrm{Re}(s) = 1/2$.

Broadly speaking, it is the existence of such geometric interpretations that has allowed versions of the Riemann hypothesis in algebraic geometry to be proved, while the case of Riemann's original zeta function remains so intractable.

# 3 THE GLOBAL ZETA FUNCTION

With the provisional definition (3) we have given for the zeta function of an elliptic curve, it is entirely unclear that it is in any way related to the Riemann zeta function. However, they are both special cases of zeta functions of global fields. A *global field* is a field of one of the following two types, introduced by Dedekind and E. Artin respectively.

1. A *number field*, that is, a subfield of $\mathbb{C}$ whose dimension as a vector space over $\mathbb{Q}$ is finite. [Recall that every subfield of $\mathbb{C}$ contains $\mathbb{Q}$ as a subfield. Thus $\mathbb{Q}$ is the smallest number field.]

2. The *function field* of a curve defined by

$$F(x, y) = 0, \tag{5}$$

where $F(x, y)$ is an irreducible polynomial over (i.e., with coefficients in) a finite field $\mathbb{F}_q$ of $q$ elements. By definition, the function field of the curve (5) is the quotient field of the integral domain $\mathbb{F}_q[x, y]/(F(x, y))$.

We remark that although the two cases above may seem at first glance to be unrelated, one can give a uniform definition, they have many important parallels, and it is often the case that conjectures and results in one setting work equally well in the other. The classical Riemann hypothesis and its formulation for elliptic curves is only one of many examples of this phenomenon.

The most down-to-earth and natural way to define the Dedekind zeta function, that is, the zeta function of a number field, is in terms of its integral ideals. But, because of the issue of points at infinity, this definition becomes awkward in the case of the function field of a curve. The approach that is perhaps best suited for working with both cases at once is to work with valuations on global fields, and we will follow this approach.

## 3.1 VALUATIONS

Suppose $K$ is a field. We denote the set of nonzero elements of $K$ by $K^\times$. A (discrete) *valuation* on $K$ is a map $v : K^\times \to \mathbb{Z}$ such that

1. $v(xy) = v(x) + v(y)$,
2. $v(x + y) \geq \min(v(x), v(y))$.

By convention, we always extend a valuation to a map $v : K \to \mathbb{Z} \cup \{+\infty\}$ by setting $v(0) = +\infty$. *Throughout this paper, we exclude from consideration the trivial valuation given by the zero map.* Two valuations on $K$ are *equivalent* if they can be rescaled to give the same valuation. Note that every valuation can be normalized uniquely to an equivalent valuation which surjects onto $\mathbb{Z}$. We will work with valuations up to equivalence, and always assume that our valuations have been normalized in this manner. We denote by $V_K$ the set of valuations of $K$.

**Example 1** *p-adic valuation.*

Suppose $K = \mathbb{Q}$. Fix a prime number $p = 2, 3, 5, \ldots$. If $x \neq 0$ is in $\mathbb{Q}$, we write

$$x = p^{v_p(x)} \cdot \frac{a}{b},$$

where $v_p(x) \in \mathbb{Z}$ and the nonzero integers $a, b$ satisfy $(p, ab) = 1$. In other words, $v_p(x)$ is the uniquely determined integer, positive, negative, or zero, which is the exponent of $p$ in the factorization of the rational number $x$ into powers of distinct primes. The group homomorphism $v_p : \mathbb{Q}^\times \to \mathbb{Z}$ gives a valuation on $\mathbb{Q}$, known as the *p-adic valuation*.

## 3.2 THE DEFINITION

Our starting point is the Euler product formula

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1} \tag{6}$$

for the Riemann zeta function. The product in (6) is over all primes $p$.

Our task is to translate this formula into an equivalent one expressed purely in terms of valuations on $\mathbb{Q}$. This will allow us to associate a zeta function to any global field, recovering the Riemann zeta function when the field is $\mathbb{Q}$ itself.

The first step is fairly straightforward: a theorem of Ostrowski states that every valuation on $\mathbb{Q}$ is equivalent to one of the $p$-adic valuations $v_p$ discussed above. Therefore, instead of indexing the product in (6) with the set of primes $p$ of $\mathbb{Q}$, we can consider the product to be over the set of valuations of $\mathbb{Q}$. However, to proceed further we need some additional definitions.

Suppose we have a field $K$, and a valuation $v$ on $K$. We can define two subsets $\mathcal{O}_v, \mathfrak{p}_v$ of $K$ as follows:

$$\mathcal{O}_v := \{x \in K : v(x) \geq 0\};$$

$$\mathfrak{p}_v := \{x \in K : v(x) > 0\}.$$

We see immediately from the definition of valuation that both of these constitute additive subgroups of $K$. In fact, one sees that $\mathcal{O}_v$ is a subring of $K$, and $\mathfrak{p}_v$ is a prime ideal of $\mathcal{O}_v$, but this is not important for our immediate purposes. All that matters is that we can form the quotient $\mathcal{O}_v/\mathfrak{p}_v$. When this quotient is finite, we can define the *norm* of $v$ by $N_v := \#(\mathcal{O}_v/\mathfrak{p}_v)$.

We then claim that we have $p = N_{v_p}$ for the $p$-adic valuation $v_p$:

**Example 2** *p-adic valuation revisited.*

We see that $\mathcal{O}_{v_p}$ is the set of fractions whose denominators are relatively prime to $p$, and $\mathfrak{p}_{v_p}$ is the subset of these whose numerators are a multiple of $p$. One can then check that $\mathcal{O}_{v_p}/\mathfrak{p}_{v_p}$ is made up of the equivalence classes of $0, 1, \ldots, p-1$, so that $N_{v_p} = p$, as asserted.

We can thus rewrite the Riemann zeta function as

$$\zeta(s) = \prod_{v \in V_{\mathbb{Q}}} \left(1 - \frac{1}{N_v^s}\right)^{-1}. \tag{7}$$

This formulation will immediately generalize to global fields. Indeed, the key fact about global fields is that for any valuation $v$, the quotient $\mathcal{O}_v/\mathfrak{p}_v$ has finitely many elements, so the norm $N_v$ is well defined. Given a global field $K$, we therefore define the *global zeta function* $\zeta_K(s)$ of $K$ by the formula

$$\zeta_K(s) = \prod_{v \in V_K} \left(1 - \frac{1}{N_v^s}\right)^{-1}. \tag{8}$$

We have seen that $\zeta_{\mathbb{Q}}(s) = \zeta(s)$, so this is the promised generalization of Riemann's zeta function.

## 3.3 DEDEKIND ZETA FUNCTION

All the various incarnations of the Riemann hypothesis for global fields bound the deviation from the predicted number of certain objects. In the case of the Riemann hypothesis for an elliptic curve $E$, the bound involves the number of points of $E$, as expressed by (4). For the classical Riemann hypothesis, as asserted earlier, the bound involves the number of prime numbers up to a given size. This may be generalized to the case of number fields, where the Riemann hypothesis can be expressed as a bound on the deviation from the expected value of the number of *prime ideals* of the field; see [2, Section 8.7]. We briefly explain how the zeta function of a number field may be rewritten in terms of prime ideals.

For a number field $K$, we have a natural subring $\mathcal{O}_K$, called the ring of integers of $K$. It consists of the roots in $K$ of monic polynomials with coefficients in $\mathbb{Z}$; in particular, Gauss' lemma says that $\mathcal{O}_{\mathbb{Q}} = \mathbb{Z}$ (note that it is by no means obvious even that $\mathcal{O}_K$ constitutes a subring of $K$!). It turns out that the notion of a $p$-adic valuation extends to $\mathcal{O}_K$, except that

instead of working with a prime number $p$, one is forced to work with a prime ideal $\mathfrak{p}$. Ostrowski's theorem extends (see [4, p. 45]) to show that the only nontrivial valuations on $K$ are the $\mathfrak{p}$-adic valuations $v_{\mathfrak{p}}$. Furthermore, one checks rather easily that although $\mathcal{O}_{v_{\mathfrak{p}}}$ is much larger than $\mathcal{O}_K$, we have the identity $\#(\mathcal{O}_K/\mathfrak{p}) = \#(\mathcal{O}_{v_{\mathfrak{p}}}/\mathfrak{p}_{v_{\mathfrak{p}}})$, so if we set $N(\mathfrak{p}) := \#(\mathcal{O}_K/\mathfrak{p})$, we can rewrite the zeta function (called in this case the *Dedekind zeta function* $\zeta_K(s)$ of (the number field) $K$) as follows:

$$\zeta_K(s) = \prod_{\mathfrak{p}} \left(1 - \frac{1}{N(\mathfrak{p})^s}\right)^{-1}, \tag{9}$$

the product being over all nonzero prime ideals $\mathfrak{p}$ of $K$. Here again we see that the Riemann zeta function coincides with the Dedekind zeta function $\zeta_{\mathbb{Q}}(s)$.

In order to make a Riemann hypothesis, we still need to extend the zeta function to the entire plane. This is accomplished via the functional equation for $\zeta_K(s)$. Although the well-known functional equation for $\zeta_{\mathbb{Q}}(s)$ was proved by Riemann, it was already known, when $s = 2, 4, 6, 8$, to Euler (see [16, Chap. 3, Section XX]), who had exploited $\zeta_{\mathbb{Q}}(s)$ for $s$ real to reprove Euclid's theorem on the infinitude of primes. However, it was not until 1917 that Hecke generalized the functional equation for the Riemann zeta function to arbitrary number fields $K$ (see [9]). The generalized Riemann hypothesis (GRH) for Dedekind zeta functions then asserts that *the nonreal zeros of* $\zeta_K(s)$ *all lie on the line* $\mathrm{Re}(s) = 1/2$ when $K$ is a number field.

## 3.4 CURVES OVER FINITE FIELDS AND THEIR ZETA FUNCTIONS

Now consider a curve $C$, defined by an irreducible polynomial

$$F(x, y) = 0 \tag{10}$$

over a finite field $\mathbb{F}_q$ of $q$ elements. Suppose $K$ is the function field of this (plane) curve $C$, i.e., the quotient field of the integral domain $\mathbb{F}_q[x, y]/(F(x, y))$.

It turns out that the valuations of $K$ are closely related to the points on $C$, where we allow points to have coordinates in any finite field containing $\mathbb{F}_q$. The basic idea is quite simple: we can view the elements of $K$ as rational functions on $C$, since if we have rational functions $\frac{G_1(x,y)}{H_1(x,y)}$ and $\frac{G_2(x,y)}{H_2(x,y)}$ whose

numerators and denominators agree modulo $F(x, y)$, we can think of them as defining the same function on $C$. If we choose a point $P \in C$, we can obtain a valuation $v_P$ on $K$ by looking at the order of vanishing or pole at $P$ of each nonzero function in $K$. However, there are some subtleties to consider. First, $C$ must be complete, meaning that we need to include the "points at infinity" on $C$, and not simply the points of $C$ in the affine plane. Second, $C$ has to be everywhere nonsingular, including at the points at infinity. This is a technical condition insuring that the order of vanishing of a function at a point is well defined. Accordingly, from now on we always assume our curves $C$ to be complete and nonsingular. Finally, and most substantively, we will see that different points can give the same valuation on $C$.

In the case of elliptic curves, the first two issues do not pose a problem: we add a single point at infinity, which is considered to have coordinates in $\mathbb{F}_q$, and all points will be nonsingular. However, the third issue is a real one, and arises even in the case of the line:

**Example 3** *Valuations on the line.*

For the sake of concreteness, we consider the case that $F(x, y) = y$, and $q = 3$. Here we have $\mathbb{F}_3[x, y]/(y) \cong \mathbb{F}_3[x]$, so that $K \cong \mathbb{F}_3(x)$. This corresponds simply to the affine line over $\mathbb{F}_3$, viewed as the $x$-axis in the plane. Points on this curve are determined uniquely by a value of $x$, in $\mathbb{F}_3$ or any field extension. [Strictly speaking, we should also include the single point at infinity to obtain the complete nonsingular curve $\mathbb{P}^1_{\mathbb{F}_3}$, the projective line over $\mathbb{F}_3$. However, this will not affect the content of the example.] Note that $\mathbb{F}_3$ does not have a square root of $-1$, so if we let $i$ denote an abtract square root of $-1$, we will have $\mathbb{F}_3[i] \cong \mathbb{F}_{3^2}$. The issue is that the rational functions making up $K$ all have coefficients in $\mathbb{F}_3$. This means that if we look at the valuation $v_i$ obtained by considering order of vanishing at $i$, we get a perfectly good valuation (for instance, $v_i(x^2 + 1) = 1$), but as a valuation on $K$, it will be the same as $v_{-i}$, since any rational function with coefficients in $\mathbb{F}_3$ must have the same number of factors of $x + i$ as of $x - i$.

More generally, the correct statement is that for any curve $C$, every point of $C$ gives a valuation on $K$, and every valuation on $K$ arises in this way, but with the caveat that if a point $P \in C$ has coordinates in $\mathbb{F}_{q^m}$, but not in any smaller field, then there are a total of $m$ points giving rise to the valuation $v_P$. One can prove that for such a point, we have $N_{v_P} = q^m$.

It is then a simple exercise to give the following equivalent formula for the zeta function, in terms of counting points on $C$:

$$\zeta_K(s) = \exp\left(\sum_{m=1}^{\infty} N_m(C)\frac{q^{-ms}}{m}\right), \tag{11}$$

where $N_m(C)$ denotes the number of points of $C$ with coordinates in $\mathbb{F}_{q^m}$. Thus, computing the zeta function of $C$ is essentially equivalent to computing the number of points of $C$ over all finite extensions of $\mathbb{F}_q$. When we express $\zeta_K(s)$ in terms of counting points on $C$, it is customary to write it instead as $Z_C(t)$ with $t = q^{-s}$, and we follow this convention throughout.

We can use this definition to compute the zeta function in the case that $C = \mathbb{P}^1_{\mathbb{F}_q}$, the projective line over $\mathbb{F}_q$ discussed above in the case $q = 3$. The number of points of $\mathbb{P}^1_{\mathbb{F}_q}$ with coordinates in $\mathbb{F}_{q^m}$ is $q^m + 1$, as each point is either given by an $x$-value in $\mathbb{F}_{q^m}$, or is the point at infinity. If we therefore set $N_m(C) = q^m + 1$ in (11), it is an easy exercise to compute that the zeta function in this case is given by:

$$Z_{\mathbb{P}^1_{\mathbb{F}_q}}(t) = \frac{1}{(1-t)(1-qt)}. \tag{12}$$

Although it is far less trivial, it can also be shown that when our curve is the elliptic curve $E$ given by equation (2), its zeta function turns out to be as in formula (3), which we used as a provisional definition. For a beautiful and fairly elementary proof see [15, Proposition 12.1].

It is true but much harder to prove that for all curves $C$, the zeta function is a rational function in $q^{-s}$, so we have an extension to a meromorphic function on all of $\mathbb{C}$, and we can state the generalized Riemann hypothesis for function fields, which asserts that *all zeros of* $\zeta_K(s)$ *lie on the line* $\mathrm{Re}(s) = 1/2$ when $K$ is the function field of a curve over $\mathbb{F}_q$.

## 3.5 WEIL CONJECTURES

In fact, the picture we have sketched is not limited to curves. Indeed, we can generalize to algebraic varieties $V$ over $\mathbb{F}_q$ of higher dimension by taking (11) as the definition of the zeta function $Z_V(t)$, replacing each $q^{-s}$ by $t$. In 1948, Weil conjectured:

1. $Z_V(t)$ is a rational function in $t$;

2. $Z_V(t)$ satisfies a functional equation of a certain prescribed form;

3. $Z_V(t)$ has an explicitly described form, which implies that the zeroes of $Z_V(q^{-s})$ lie on the lines $\mathrm{Re}(s) = (2j-1)/2$, for $j = 1, \ldots, \dim V$, i.e., that the (analogue of the) *Riemann hypothesis* holds.

The rationality of the zeta functions of curves was established in 1931 by F. K. Schmidt [12], and A. Weil proved the Riemann hypothesis for them in 1948 (simpler proofs were given in [3] and [13]; see also [14, Part II]). The rationality of $Z_V(t)$ was proved in 1960 by Dwork [7]. Grothendieck discovered a method of applying ideas from algebraic topology to abstract algebraic varieties, and this approach culminated in the proof of the most difficult part of the Weil conjecture – the Riemann hypothesis for higher dimensional varieties – in 1974 by Deligne [6], for which he was awarded the Fields Medal.

# 4 AN ELEMENTARY PROOF OF HASSE'S THEOREM

We now prove the Hasse theorem (inequality (4)), that is, the Riemann hypothesis for elliptic curves over finite fields. The proof is essentially that of Manin [10], which in itself is based on the original one by Hasse. To begin with let us assume that $k$ is any field that does not contain $\mathbb{F}_2$ or $\mathbb{F}_3$ as a subfield. For us, an *elliptic curve $E$ over $k$* is a curve

$$y^2 = x^3 + ax + b \quad (a, b \in k) \tag{13}$$

with $4a^3 + 27b^2 \neq 0$.

If $K$ is any field containing $k$, then the set $E(K)$ consisting of points on (13) with coordinates in $K$ together with a point $O$ at infinity forms an abelian group. For $k = \mathbb{Q}$ and $K = \mathbb{R}$, assuming $x^3 + ax + b$ has only one real root, it looks as in Figure 1.

The point $O$ at infinity is on each end of every vertical line. The sum of two points $P_1, P_2$ is the reflection in the $x$-axis of the third point of intersection of the line through $P_1$ and $P_2$ (tangent to (13) at $P$ if $P_1 = P_2 = P$) with the cubic (13). One then checks that $O$ is the zero of the group, and that the inverse of a point $(X, Y)$ is given simply by $(X, -Y)$.

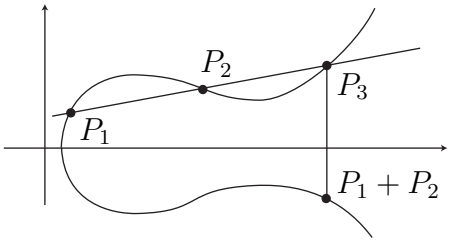


Figure 1:

## 4.1 Twists

To prove the Riemann hypothesis for elliptic curves over finite fields, that is the inequality (4), we shall be working with another elliptic curve closely related to $E$. It is defined over the function field $K = \mathbb{F}_q(t)$ by

$$\lambda y^2 = x^3 + ax + b \tag{14}$$

where $\lambda = \lambda(t) = t^3 + at + b$. The elliptic curve $E_\lambda$ given by equation (14) is a *twist* of $E$.

If $x(P)$ denotes the $x$-coordinate of a point $P$, we compute $x(P_1 + P_2)$ for $P_1, P_2$ in $E_\lambda(K) = \{(x, y) \in K^2 \mid \lambda y^2 = x^3 + ax + b\} \cup \{O\}$. This formula for $x(P_1 + P_2)$ in terms of $x(P_1)$ and $x(P_2)$ plays a dominant role in the proof of inequality (4). We leave aside certain cases (such as $x(P_1) = x(P_2)$; $P_1$ or $P_2 = O$) that we do not need for our proof.

Suppose $P_j = (X_j, Y_j) \in E_\lambda(K)$ for $j = 1, 2$. To compute $x(P_1 + P_2)$ we write the equation of the line through $P_1$ and $P_2$, which is

$$y = \left(\frac{Y_1 - Y_2}{X_1 - X_2}\right) x + \ell. \tag{15}$$

To find the $x$-coordinate $X_3$ of the third point $P_3$ of intersection of this line with the cubic (14), we substitute for $y$ from (15) in (14) to get

$$x^3 - \lambda \left(\frac{Y_1 - Y_2}{X_1 - X_2}\right)^2 x^2 + \cdots = 0. \tag{16}$$

Since $X_1, X_2, X_3$ are the three solutions of (16), the left side of (16) is

$$\begin{aligned}(x - X_1)(x - X_2)(x - X_3) \\ = x^3 - (X_1 + X_2 + X_3)x^2 + \cdots .\end{aligned} \tag{17}$$

Comparing the coefficient of $x^2$ in (16) and (17), we get

$$x(P_1 + P_2) = X_3 = \lambda \left( \frac{Y_1 - Y_2}{X_1 - X_2} \right)^2 - (X_1 + X_2) \,. \tag{18}$$

## 4.2 Frobenius map

A crucial ingredient in the proof of inequality (4) is the *Frobenius map* $\Phi$ and its elementary properties. For a fixed $q$, let $K$ be any field containing $\mathbb{F}_q$ as a subfield. We define $\Phi = \Phi_q : K \to K$ as the function given by $\Phi(X) = X^q$.

We summarize the properties of the Frobenius map we need in the following theorem:

**Theorem** *The Frobenius map $\Phi(X) = X^q$ has the following properties:*

*i)* $(XY)^q = X^q Y^q$.

*ii)* $(X + Y)^q = X^q + Y^q$.

*iii)* $\mathbb{F}_q = \{\alpha \in K | \, \Phi(\alpha) = \alpha\}$.

*iv) For $\phi(t)$ in $\mathbb{F}_q(t)$, $\Phi(\phi(t)) = \phi(t^q)$.*

Although it is not used directly in this proof of the Hasse inequality, it is worth noting that iii) above implies that $E(\mathbb{F}_q)$ consists precisely of points fixed by $\Phi$. Other proofs of the Hasse inequality use this fact directly.

*Proof.* i) is trivial.

ii) We use induction on $r = \log_p q$. If $r = 1$, $q = p$ and

$$(X + Y)^p = \sum_{j=0}^{p} \binom{p}{j} X^j Y^{p-j}.$$

For $0 < j < p$, the binomial coefficient $\binom{p}{j}$ satisfies

$$\binom{p}{j} = \frac{p!}{j!(p-j)!} = p \cdot m$$

for some positive integer $m$, because nothing in the denominator can cancel $p$ in the numerator and $\binom{p}{j}$ is a whole number. Since $p\alpha = 0$ for all $\alpha$ in $K$, ii) follows. For $r > 1$, by the induction hypothesis

$$\begin{aligned}(X+Y)^q &= ((X+Y)^{p^{r-1}})^p \\ &= (X^{p^{r-1}} + Y^{p^{r-1}})^p \\ &= X^q + Y^q.\end{aligned}$$

iii) The set $\mathbb{F}_q^\times$ of nonzero elements of $\mathbb{F}_q$ is a multiplicative group of order $q-1$. Therefore, by elementary group theory, $\alpha^{q-1} = 1$ for all $\alpha$ in $\mathbb{F}_q^\times$. In other words, each of the $q$ elements of $\mathbb{F}_q$ is a root of the polynomial $t^q - t = t(t^{q-1} - 1)$ of degree $q$. Since a polynomial of degree $q$ cannot have more than $q$ roots, $\mathbb{F}_q$ consists precisely of the elements of $K$ which are roots of $t^q - t$. This proves iii).

iv) follows at once from i), ii) and iii).

## 4.3 Counting points on elliptic curves

Returning to the situation that $K = \mathbb{F}_q(t)$, we now show how we can use the properties of the Frobenius map $\Phi_q$ and of the elliptic curve $E_\lambda(K)$ to count the number of solutions of the equation $y^2 = x^3 + ax + b$ ($a, b \in \mathbb{F}_q$, $q = p^r$, $4a^3 + 27b^2 \neq 0$) with $x, y$ in $\mathbb{F}_q$.

Clearly $(t, 1)$ and its negative $-(t, 1) = (t, -1)$ are in $E_\lambda(K)$. Using the properties of $\Phi_q$, it is also clear that the point

$$P_0 = (t^q, (t^3 + at + b)^{(q-1)/2})$$

is in $E_\lambda(K)$.

We now define a degree function $d$, which we will ultimately show to be a quadratic polynomial with nonreal roots. Its discriminant plays a central role in the proof of inequality (4). For $n \in \mathbb{Z}$, let

$$P_n = P_0 + n(t, 1),$$

the addition being the one on $E_\lambda(K)$. Define $d : \mathbb{Z} \to \{0, 1, 2, \ldots\}$ by

$$d(n) = d_n = \begin{cases} 0, & \text{if } P_n = O; \\ \deg(\mathrm{num}(x(P_n))), & \text{otherwise.} \end{cases}$$

Here num$(X)$ is the numerator of a rational function $X \in \mathbb{F}_q(t)$, taken in the lowest form. The values of this degree function on three consecutive integers satisfy (for a proof, see [5]) the following identity:

**Basic Identity**

$$d_{n-1} + d_{n+1} = 2d_n + 2. \tag{19}$$

The crux of the proof of (4) is the following theorem relating the degree function to the number $N_q$ of solutions of $y^2 = x^3 + ax + b$ ($a, b \in \mathbb{F}_q, 4a^3 + 27b^2 \neq 0$) with $x, y$ in $\mathbb{F}_q$.

**Theorem 1**

$$d_{-1} - d_0 - 1 = N_q - q. \tag{20}$$

*Proof.* Let $X_n = x(P_n)$. Since $P_0 \neq (t, 1)$, we have $P_{-1} \neq O$, so $d_{-1} = \deg(\text{num}(X_{-1}))$. We therefore compute $X_{-1}$ and look at the degree of its numerator when it is in the lowest form. By (18),

$$\begin{aligned} X_{-1} &= \frac{(t^3 + at + b)\left[(t^3 + at + b)^{(q-1)/2} + 1\right]^2}{(t^q - t)^2} - (t^q + t) \\ &= \frac{t^{2q+1} + \text{ lower terms}}{(t^q - t)^2}, \end{aligned} \tag{21}$$

where the last expression is obtained by putting the previous one over the common denominator $(t^q - t)^2$ and using property iv) of the Frobenius map. We wish to cancel any common factors in the last expression. Since the term $t^q + t$ has no denominator, it suffices to compute the cancellation in the first term of the previous expression.

Property iii) of the Frobenius map is, as noted in the proof, equivalent to the fact that $\mathbb{F}_q$ consists precisely of the $q$ roots of $t^q - t$. Hence

$$t^q - t = \prod_{\alpha \in \mathbb{F}_q} (t - \alpha),$$

so to compute $d_{-1}$ we wish to cancel all common factors of the fraction

$$\frac{(t^3 + at + b)\left[(t^3 + at + b)^{(q-1)/2} + 1\right]^2}{\prod_{\alpha \in \mathbb{F}_q} (t - \alpha)^2}.$$

The only factors to cancel from the denominator of this quotient are either

i) $(t-\alpha)^2$ with $(\alpha^3+a\alpha+b)^{(q-1)/2}=-1$, or

ii) $t-\alpha$ with $\alpha^3+a\alpha+b=0$.

[Recall that $t^3+at+b$ has no repeated root by hypothesis.] Let

$$m = \text{the number of factors of the first kind,}$$
$$n = \text{the number of factors of the second kind.}$$

Since factors of the first kind are coprime to the factors of the second kind,

$$d_{-1} = 2q+1-2m-n.$$

Since $d_0 = q$, this gives

$$d_{-1} - d_0 - 1 = q - 2m - n. \tag{22}$$

Now an $\alpha$ in $\mathbb{F}_q$ with $\alpha^3+a\alpha+b$ equal to a nonzero square in $\mathbb{F}_q$ will give two solutions of $y^2 = x^3+ax+b$, whereas there is only one solution of this equation when $\alpha^3+a\alpha+b=0$. Moreover, Euler's criterion says that $\alpha^3+a\alpha+b$ is a nonsquare if and only if $(\alpha^3+a\alpha+b)^{(q-1)/2}=-1$, so $m$ counts the number of $\alpha$ which do not correspond to any solution of $y^2=x^3+ax+b$. Hence

$$N_q = 2q-n-2m.$$

or

$$N_q - q = q - 2m - n. \tag{23}$$

Equation (20) follows from (22) and (23).

**Theorem 2** *The degree function $d(n)$ is a polynomial of degree 2 in $n$. In fact,*

$$d(n) = n^2 - (d_{-1}-d_0-1)n + d_0. \tag{24}$$

*Proof.* By induction on $n$. For $n=-1$ and 0, (24) is a triviality. By the Basic Identity and the induction hypothesis,

$$\begin{aligned} d_{n+1} &= 2d_n - d_{n-1} + 2 \\ &= 2\left[n^2-(d_{-1}-d_0-1)n+d_0\right] \\ &\quad -\left[(n-1)^2-(d_{-1}-d_0-1)(n-1)+d_0\right]+2 \\ &= (n+1)^2-(d_{-1}-d_0-1)(n+1)+d_0. \end{aligned}$$

The induction step in the other direction can be carried out in a similar manner.

**Proof of the Riemann hypothesis.** We consider the roots $x_1, x_2$ of the quadratic polynomial

$$d(x) = x^2 - (N_q - q)x + q.$$

Suppose that (4) fails to hold, so that the discriminant $(N_q - q)^2 - 4q$ is positive. Then $x_1, x_2$ are distinct real numbers, say $x_1 < x_2$. By the way it is constructed, $d(x)$ takes only nonnegative integer values on $\mathbb{Z}$, so there must exist some $n \in \mathbb{Z}$ such that

$$n \leq x_1 < x_2 \leq n + 1. \tag{25}$$

Since the coefficients of $d(x)$ are in $\mathbb{Z}$, we have $x_1 + x_2$, $x_1 \cdot x_2 \in \mathbb{Z}$. Hence

$$(x_1 - x_2)^2 = (x_1 + x_2)^2 - 4x_1x_2 \in \mathbb{Z},$$

and for (25) to hold, we must have $x_1 = n$, $x_2 = n + 1$. But we note that $x_1x_2 = q$ is a prime power, so this could only happen if $q = 2$ and $n = 1$ or $-2$, which is a contradiction since we have assumed throughout that $p \neq 2$. We thus conclude that (4) must hold, as desired.

## Acknowledgements

We would like to thank Sudhir Ghorpade for suggesting a substantial simplification in our argument.

# References

1. E. Artin, *Collected Papers*, Addison-Wesley, Reading, MA, 1965.

2. E. Bach and J. Shallit, *Algorithmic Number Theory*, vol. 1, MIT Press, Cambridge, MA, 1996.

3. E. Bombieri, Counting points over finite fields (d'apres S.A. Stepanov), *Seminaire Bourbaki*, vol. 1972/1973, exp. 430, in Lecture Notes in Mathematics, vol. 383, Springer, Berlin, 1974.

4. J. W. S. Cassels and A. Fröhlich, *Algebraic Number Theory*, Academic Press, London, 1967.

5. J. S. Chahal, Manin's proof of the Hasse inequality revisited, *Nieuw Arch. Wiskd.* **13** (1995) 219–232.

6. P. Deligne, La conjecture de Weil, *Publ. Math. I.H.E.S.* **43** (1974) 273–307.

7. B. Dwork, On the rationality of the zeta function of an algebraic variety, *Amer. J. Math.* **82** (1960), 631–648.

8. H. M. Edwards, *Riemann's Zeta Function*, Dover, Mineola, NY, 2001.

9. E. Hecke, *Math. Werke*, Vandenhoeck & Ruprecht, Göttingen, 1959.

10. Ju. I. Manin, On cubic congruences to a prime modulus, *Izv. Akad. Nauk USSR, Math. Ser.* **20** (1956) 673–678.

11. B. Riemann, Über die Anzahl der Primzahlen unter einer gegebnen Grösse, in *Gesammelte Werke*, Teubner, Leipzig, 1892.

12. F. K. Schmidt, Analytische Zahlentheorie in Körpern der Charakteristik $p$, *Math. Z.* **33** (1931) 1–32.

13. W. M. Schmidt, Zur Methode von Stepanov, *Acta Arithm.* **24** (1973) 347–367.

14. ———, *Equations over Finite Fields: An Elementary Approach*, Kendrick, Heber City, UT, 2004.

15. L. Washington, *Elliptic Curves*, Chapman & Hall, Boca Raton, FL, 2003.

16. A. Weil, *Number Theory: An Approach through History*, Birkhäuser, Boston, MA, 1984.

**JASBIR S. CHAHAL** is a professor of mathematics at Brigham Young University with research interests in number theory. He has written a number of papers on algebraic groups and on the arithmetic of elliptic curves. His book *Topics in Number Theory* was published by Plenum in 1988.
*Department of Mathematics, Brigham Young University, Provo, UT 84602*

*jasbir@math.byu.edu*

**BRIAN OSSERMAN** received his PhD. from MIT in 2004, and spent the spring of that year as a Japan Society for the Promotion of Science fellow in Kyoto, Japan. He is currently an NSF Poctdoctoral Fellow at the University of California, Berkeley, with research interests in algebraic and arithmetic geometry. A retired freelance video games reviewer, he spends his weekends practicing aikido and crashing fencing tournaments.

*Department of Mathematics, University of California, Berkeley, CA 94720*
*osserman@math.berkeley.edu*