Class Number problem

First we introduce the class number in one of many ways. Consider a quadratic form with integer coefficients a, b, c

$$Q(x, y) = ax^2 + bxy + cy^2,$$

$a, c > 0$ and we also assume it is primitive: that is the greatest common divisor of a, b, c denoted by $(a, b, c) = 1$. We recall the discriminant

$$d = b^2 - 4ac.$$

We shall **only focus** on $d < 0$ in this note, that is when the quadratic form $Q(x, y)$ is positive definite. When $d > 0$ the situation is equally fascinating and deep with consequences to prime geodesics in hyperbolic surfaces by virtue of a clever observation in Peter Sarnak's thesis(J. Number Theory 1981). First note that by the definition $d = 4k$ or $d = 4k + 1$. Since $d < 0$ it is trivial that we should have $a, c > 0$ in $Q(x, y)$.

Definition: A discriminant d is said to be fundamental iff d is square free in its prime factorization or if it contains a square $q^2 > 1$ in its factorization, then d/q^2 is not a discriminant.

We shall only restrict ourselves to fundamental discriminants in what follows.

Now we apply a transformation to $Q(x, y)$ via the matrix in $SL(2, Z)$ that is a 2×2 matrix with determinant 1 having integer entries.

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}.$$

Here $\alpha\delta - \beta\gamma = 1$ and the matrix above has integer entries.

One can check the discriminant does not change under the transformation and the new coefficients are given by

$$A^t \begin{pmatrix} a & \frac{b}{2} \\ \frac{b}{2} & c \end{pmatrix} A,$$

where

$$A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix},$$

and A^t is the transpose of A. Quadratic forms that are related by such a transformation of $SL(2, Z)$ are said to be equivalent.

An Example: Consider $d = 4k + 1$ and the two quadratic forms,

$$Q(x, y) = x^2 - xy + \frac{1 + |d|}{4} y^2, \; R(X, Y) = X^2 + XY + \frac{1 + |d|}{4} Y^2.$$

Note both quadratic forms have the same discriminant d. Next consider the transformation

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}, \; x = X + Y, y = Y.$$

Apply this transformation to $Q(x, y)$ and we get on simplification

$$(X+Y)^2 - (X+Y)Y + \frac{1+|d|}{4}Y^2 = X^2 + XY + \frac{1+|d|}{4}Y^2.$$

Thus the two quadratic forms $Q(x, y)$ and $R(X, Y)$ are in fact equivalent.

Gauss obtained the following easy result, a part of Reduction theory(taken forward by Borel and Harishchandra (1962) in general) that every quadratic form can be brought to a standard form by using matrices in $SL(2, Z)$. The new quadratic form has coefficients that satisfy the inequalities:

$$-a \leq b \leq a, a < c. \text{ Or } 0 \leq b \leq a = c.$$

For a given d thus

$$|d| = -d = 4ac - b^2 > 3a^2.$$

So $a < \sqrt{|d|/3}$. Hence $|b| < a < \sqrt{|d|/3}$. And so because (a being a natural number is ≥ 1) $c < 4ac < |d| + b^2 < 4|d|/3$, we conclude that there are only finitely many non-equivalent quadratic forms for a given discriminant d.

Proof of Gauss's result on Reduction: We consider the two matrices
$E_n, F \in SL(2, Z)$,

$$E_n = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}, \; F = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Clearly E_n is associated with the transformation

$$x = X + nY, y = Y,$$

and F with

$$x = Y, y = -X.$$

We apply the two transformations E_n and F in tandem to the quadratic form $Q(x, y)$. We first apply E_n. We get

$$a(X+nY)^2 + b(X+nY)Y + cY^2 = aX^2 + (b+2an)XY + (c+an^2+bn)Y^2.$$

By choosing n appropriately we can arrange $|b| \leq a$. We can use F to switch $c + an^2 + bn$ with a if the coefficient of Y^2 is smaller than a to arrive at

$$(c+an^2+bn)X^2 - (b+2an)XY + aY^2 = AX^2 + BXY + CY^2.$$

Thus applying E_n and F in tandem we see we have

$$|B| + A < a + |b|,$$

where $a, A > 0, b, B$ are integers. This process has to terminate in a finite number of steps. When it terminates, since the discriminant is preserved, $A \neq 0$ as the quadratic form has to remain positive definite and $A \leq C$, or we may apply E_n and F again to obtain a contradiction to the minimality. If $A = C$ and $B < 0$, we may apply F at which point B is transformed to $-B$ and $A = C$ still. We have now proved all parts of the claim.

Definition: The class number $h(d)$ is the number of non-equivalent quadratic forms of discriminant d.

Remark: Note $h(d) \geq 1$, since one has the quadratic forms,

$$x^2 + \frac{|d|}{4}y^2,\ x^2 + xy + \frac{1+|d|}{4}y^2,$$

with discriminant d when $d = 4k$ and $d = 4k+1$ respectively($d = 4k+1, (1+|d|)/4$ is a natural number!).

The class number first appears in a formula of Dirichlet in his famous theorem of primes in arithmetic progression. This link can be established by using Eisenstein series and the Dedekind zeta function. This point of view is central to the proof of Heilbronn-Linfoot. By assuming there are 11 such d with $h(d) = 1$ they concoct a Landau-Siegel zero using this link and this leads to a contradiction. The rest of the proof is a deft use of the Euler-Maclaurin formula.

Gauss computed the discriminant d for which $h(d) = 1$ and found 9 such values for d. In 1933 Heilbronn and Linfoot showed that apart for the 9 values found by Gauss, there is possibly a 10th one and no more.

In the 1950s a schoolteacher in Germany, Heegner claimed to have a proof that the list of Gauss was complete but he was not believed. In 1967 two proofs one by Stark and another by Alan Baker(using his theorem of Linear Logarithms) established that the list of Gauss was complete. Moreover Stark pointed out that Heegner's proof was in fact correct but by then Heegner was dead. Baker showed that there are 19 discriminants with $h(d) = 2$. The situation for $h(d) \geq 4$ is still open. Oesterle has classified the discriminants for $h(d) = 3$ in recent times.

It can be easily shown that

Proposition:

$$h(d) \leq C\sqrt{|d|}\log|d|.$$

Proof: First note by the discussion above

$$h(d) \leq \sum_{a<\sqrt{|d|/3}} \rho_d(a),$$

where $\rho_d(a)$ denotes the number of solutions $b \bmod(2a)$ (since $-a \leq b \leq a$) to the quadratic congruence

$$b^2 \equiv d(\text{mod } 4a).$$

It is easily seen that $\rho_d(a) \leq C\tau(a)$ where $\tau(a)$ is the divisor function, the number of divisors of a. Thus our inequality above can be re-written as

$$h(d) \leq \sum_{a<\sqrt{|d|/3}} \rho_d(a) \leq C \sum_{a<\sqrt{|d|/3}} \tau(a).$$

Using the fact from Dirichlet's theorem(see the end of these notes) that

$$\sum_{a<x} \tau(a) = O(x \log x),$$

we immediately conclude

$$h(d) \leq C\sqrt{|d|} \log |d|.$$

A finer result can be obtained, see Iwaniec-Kowalski [3].

The opposite inequality is also valid under the assumption that the L function $L(s, \chi_d)$ (where the character χ_d is given by the Jacobi symbol $(\frac{d}{n})$) satisfies

$$L(\frac{1}{2}, \chi_d) \geq 0.$$

This hypothesis is tantamount to the Generalized Riemann Hypothesis(GRH). The idea of the proof is to evaluate a residue at $s = 1/2$ in a representation of the Dedekind zeta function in terms of Eisenstein series. See [3].

Through the works of Erich Hecke, Max Deuring, Edmund Landau and Heilbronn, it was established that

$$h(d) \to \infty, |d| \to \infty.$$

Dorian Goldfeld showed that for an absolute and effective constant C one has

$$h(d) \geq C \log |d|.$$

If the constant is allowed to be ineffective, one has for any positive ϵ,

$$h(d) \geq C_\epsilon |d|^{\frac{1}{2}-\epsilon}.$$

A major problem in Modern Number Theory is to obtain an effective constant in the last inequality. This problem is hard and related to Landau-Siegel zeros of L-functions. If there is no such spurious zero, the problem will be solved. See remarks above for the Heilbronn-Linfoot theorem.

We now state a cute theorem of Rabinovitch(1913) on prime spitting polynomials. We consider discriminants of the form $4k + 1$ in the rest of this note.

First a remark. the discriminant d by definition can only be of the form $4k$ or $4k+1$. By a little thought if $h(d) = 1$ then the discriminant d has to be prime if $|d| > 15$ and also $(1 + |d|)/4$ has to be prime.

We set

$$q = \frac{1 + |d|}{4}.$$

Consider the polynomial

$$f_q(x) = x^2 + x + q.$$

Theorem: $f_q(x)$ is a prime for all integers x, $0 \leq x \leq q - 2$ if and only if $h(d) = 1$. That is the polynomial spits out primes.

By Gauss and Heilbronn-Linfoot, Baker, Stark the complete list of fundamental d for which $h(d) = 1$ is the list

$$-1, -2, -3, -7, -11, -19, -43, -67, -163.$$

Thus for example if we take $d = -163$ we get the polynomial

$$x^2 + x + 41.$$

Then if we evaluate the polynomial for $0 \leq x \leq 39$ we get primes. There are no other such polynomials of the type

$$x^2 + x + q,$$

from the discussion above.

We shall now apply the previous discussion to prove a theorem of Fermat.

Theorem(Fermat): Any prime p of the form $4k + 1$ or 2 can be written as a sum of two squares

$$p = A^2 + B^2.$$

We may now apply a special case of an identity that goes back to Brahmagupta and prove

Theorem: Given any number n which has a prime factorization

$$n = p_1^{a_1} p_2^{a_2} \cdots p_n^{a_n},$$

where the primes are either 2 or all of the form $4k + 1$, then

$$n = A^2 + B^2.$$

This theorem is a special case of Lagrange's four-square theorem which states that any number n can always be written as a sum of four squares

$$n = A^2 + B^2 + C^2 + D^2.$$

We first prove the second theorem as a corollary of Fermat's theorem. We use Brahmagupta's identity

$$(A^2 + B^2)(a^2 + b^2) = (aA + bB)^2 + (aB - bA)^2.$$

Now any product of primes p, q either distinct or the same and either 2 or of the type $4k + 1$ using Fermat's theorem and Brahmagupta can be written as

$$pq = (a^2 + b^2)(A^2 + B^2) = (aA + bB)^2 + (aB - bA)^2.$$

Using induction (on the number of primes in the factorization) the theorem is now proved.

We now prove Fermat's theorem. First we note there is only one non-equivalent quadratic form of discriminant $d = -4$. We have already observed that,

$$|b| \leq \sqrt{|d|/3}.$$

Thus if $d = -4$, $b = \pm 1, 0$. Next since $d = -4$,

$$d = -4 = b^2 - 4ac,\ 4ac = b^2 + 4.$$

If $b = \pm 1$, then $4ac = 5$. But 4 cannot divide 5. Thus $b = 0$ and $ac = 1$ and so $a = c = 1$. Thus the unique quadratic form with discriminant -4 is

$$x^2 + y^2.$$

Since 2 can obviously be written as

$$2 = 1^2 + 1^2$$

we only focus on primes of the form $4k + 1$ in Fermat's theorem. Let us take for granted an immediate consequence of Euler's criterion the fact that one can always find B so that for primes p of the form $4k + 1$ one has

$$B^2 + 1 = kp.$$

Multiplying this by 4 we get

$$(2B)^2 + 4 = 4kp.$$

Set $b = 2B$ and we get

$$b^2 + 4 = 4kp. \qquad (\star)$$

Thus using $(\star)$ the quadratic form

$$px^2 + bxy + ky^2,$$

has discriminant -4. But $X^2 + Y^2$ is the only quadratic form in standard form with discriminant -4, so both $X^2 + Y^2$ and $px^2 + bxy + ky^2$ must be related by a $SL(2, Z)$ transformation,

$$X = \alpha x + \beta y, Y = \gamma x + \delta y.$$

Thus plugging this into $X^2 + Y^2$ we get

$$X^2 + Y^2 = (\alpha x + \beta y)^2 + (\gamma x + \delta y)^2.$$

Equating coefficients for x^2 we easily get

$$p = \alpha^2 + \gamma^2.$$

This finishes the proof of Fermat's theorem. Euler's criterion is completely elementary and left to the reader to prove.

Euler's Criterion: For p a prime, the congruence

$$x^2 \equiv a(\mathrm{mod} p)$$

has a solution, if and only if

$$a^{\frac{p-1}{2}} \equiv 1(\mathrm{mod} p).$$

In our problem related to Fermat's theorem we are trying to solve

$$B^2 \equiv -1(\mathrm{mod} p).$$

There is a solution provided

$$(-1)^{\frac{p-1}{2}} \equiv 1(\mathrm{mod} p).$$

This is certainly true if p is a prime of type $4k + 1$.

Remark: The proof of Lagrange's four square theorem stated above has a similar flavor. First one proves that any prime p of the form $4k + 3$ can be written as the sum of four squares. Then one applies a form of Brahmagupta's identity for four squares. Many proofs of Lagrange's theorem and Fermat's theorem are available. One such proof proceeds by applying Minkowski's theorem on lattices in convex domains.

Now we will give an application to counting Lattice points on a circle and eigenvalues of the Laplace operator, a problem that appears in Quantum Mechanics and Engineering, to heat flow in plates, vibration problems of plates and so on. We state these as a theorem and prove them.

Definition: A lattice point in the plane $\mathbf{R}^2$ is defined as a point P whose coordinates (m, n) are both given by integers.

Our question is: Given a circle $x^2 + y^2 = r$, does this circle have any lattice points on it. Obviously if it does, then since $m^2 + n^2 = r$, r must be a natural number. We have

Theorem: A circle $x^2 + y^2 = n$ where n is a natural number will have lattice points on the circle, provided one can factor n into prime factors that contain only 2 or primes of the type $4k + 1$ in the prime factorization. Let $\tau(n)$ denote the number of divisors of n. The number of lattice points on the circle will be exactly $4\tau(n)$, when n is odd, and exactly 4 lattice points when $n = 2^a$ and $4\tau(m)$, when $n = 2^a m$ with m odd.

Proof: The first part of the theorem is immediate from our previous theorem that, numbers n that have prime factors that are 2 or primes of the form $4k + 1$ can be written as

$$n = k^2 + j^2,$$

with k, j integers. Thus the circle $x^2 + y^2 = n$ has the lattice point (k, j) on it. Thus the only question is, in how many ways can we split n as a sum of two squares. This is answered by a theorem of Jacobi at the end of these notes.

Lastly we give our application to Quantum Mechanics and Engineering. Consider the Laplace operator on a domain which is a square S of edge length π. We study the eigenvalue and eigenfunction problem

$$-\Delta u(x, y) = \lambda u(x, y), \ u\Big|_{\partial S} = 0.$$

This problem has eigenfunctions via separation of variables,

$$u(x, y) = \sin kx \sin jy,$$

with corresponding eigenvalue $n = \lambda = k^2 + j^2$. Thus from the previous theorem we conclude, that the eigenvalues λ that can appear are natural numbers that are exactly those natural numbers n that have only 2 and primes of the type $4k + 1$ in their prime factorization. Moreover each eigenspace has dimension exactly equal to $2\tau(n)$. We have to take 2 as $(-k, -j)$ lead to the same eigenfunction as does $(-k, j)$ etc.

However primes $4k + 3$ can also appear as eigenvalues, provided they are exactly squares in the prime factorization. We collect all this information and state a theorem.

Theorem: Every n of the form

$$n = p_1^{a_1} \cdots p_k^{a_k} P_1^{2A_1} \cdots P_m^{2A_m},$$

where p_i is either 2 or a prime of the form $4k + 1$ and P_i a prime of the form $4k + 3$ and the set p_i is non-empty, is an eigenvalue for the vibrating square. Moreover the dimension of the eigenspace is $2\tau(N)$ where

$$N = p_1^{a_1} \cdots p_k^{a_k},$$

where $\tau(N)$ is the number of divisors of N and where N is odd. The dimension of the eigenspace is 1 when $N = 2^a$ a odd and 0 when a even and $4\tau(M)$ when $N = 2^a M$, with M odd.

Examples: Circles $x^2 + y^2 = 35,\ x^2 + y^2 = 77$ will contain no lattice points and neither will $35, 77$ be eigenvalues/vibrating frequencies of a clamped plate, since the numbers $35, 77$ contain 7 in their prime factorization which is a prime of the form $4k + 3$. But note that $7^2 \times 5$ can be written as a sum of two squares, for example,

$$7^2 \times 5 = 7^2 2^2 + 7^2.$$

7^2 by itself cannot be an eigenvalue as the corresponding eigenfunction vanishes

$$7^2 = 7^2 + 0^2,\ \sin 7x \sin 0y \equiv 0.$$

Note: Linfoot moved to Optics after this work with Heilbronn. He partnered with Wolff to write books on Optics. This is the same Wolff who wrote a classic textbook on Optics with Max Born. Wolff moved to the University of Rochester presumably so that he could be near Kodak. Linfoot was heavily involved in war related work in WW2 related to Optics. After the war he developed telescopes and headed important Govt. positions in UK regarding science administration. Linfoot can be seen coming out to field with Hardy as captain in a famous match: Hardy's Mathematicals vs Rest of the World at the British Science Association meeting at Oxford around 1929. Linfoot was a student of Hardy.

The Theorem of Jacobi and Lattice Points

We will derive in these notes the theorem of Jacobi via a simple proof. Jacobi used his beautiful and deep identities on theta functions to derive this theorem. Our proof is very simple.

Let us consider any n, with its prime factorization

$$n = p_1^{a_1} \cdots p_k^{a_k}.$$

Proposition: The number of divisors $\tau(n)$ is given by

$$\tau(n) = (a_1 + 1)(a_2 + 1) \cdots (a_k + 1) = \Pi_{i=1}^{k}(a_i + 1).$$

Proof: The proof is trivial. For a single prime p^a, note the divisors are

$$1 = p^0, p^1, , p^2, \cdots p^a.$$

These constitute $a+1$ divisors. Thus applying this observation to each of the prime factors of n, we easily get our result.

Now we will prove the theorem of Jacobi in a special case by an elementary argument.

Theorem(Jacobi): Let

$$n = p_1^{a_1} \cdots p_k^{a_k},$$

where each p_i is a prime of the form $4k + 1$. Then the number of lattice points on the circle

$$x^2 + y^2 = n,$$

is $4\tau(n)$, when n is odd. It is 4 when $n = 2^a$ and $4\tau(m)$ when $n = 2^a m$, m odd. The lattice points lie on the axes when $n = 2^a$ with a even.

Proof: We first take up the case of n odd. By Fermat's theorem each of the prime factors that divide n can be expressed as a sum of two squares. Focusing on an arbitrary prime p_i in this list we have

$$p_i = m^2 + n^2 = (m + \sqrt{-1}n)(m - \sqrt{-1}n).$$

The prime p_i occurs a_i times. We now arrange the number $m+\sqrt{-1}n$ in the first column of a table where the number of rows is exactly a_i. The second column of the table consists of the **complex conjugate** $m - \sqrt{-1}n$ also in a_i rows. See Figure 1. Note that multiplying all the entries in Figure 1. we get $p_i^{a_i}$. Next we multiply the entries in the first column to get $z = M_0 + \sqrt{-1}N_0$. Multiplying the entries in the second column yields the conjugate $M_0 - \sqrt{-1}N_0$. Further note (M_0, N_0) is one desired lattice point since obviously $p_i^{a_i} = M_0^2 + N_0^2$. Next we interchange the entries in row 1 to get Figure 2. We perform the operation of multiplying the first column again and get $M_1 + \sqrt{-1}N_1$, with $p_i^{a_i} = M_1^2 + N_1^2$. This gets us the lattice point (M_1, N_1). We next interchange the elements of the second

row and multiply the first column again to get $M_2 + \sqrt{-1}N_2$ leading to the lattice point (M_2, N_2) since $p_i^{a_i} = M_2^2 + N_2^2$. The process ends when we interchange the elements of the last row at which point the second column of Figure 1 will become the first column at the end of the process. We multiply out the first column leading to the lattice point (M_s, N_s), where $s = a_i + 1$. Thus $p_i^{a_i}$ has provided us $a_i + 1$ lattice points.

If we perform the operations described above for two primes p_1, p_2 arranging their factorization $p_1 = (m + \sqrt{-1}n)(m - \sqrt{-1}n), p_2 = (r + \sqrt{-1}s)(r - \sqrt{-1}s)$ in two columns with a_1 and a_2 rows respectively, we find we get $(a_1 + 1)(a_2 + 1)$ lattice points. In general applying this process for the primes that appear in the factorization of n, we see the number of lattice points we get by this process is exactly $\tau(n)$.

Now given any lattice point, represented by $M + iN$ we can rotate the lattice point by $\pi/2$ and π by multiplication by $\pm\sqrt{-1}, -1$ to get additional lattice points. Thus in all we have exactly $4\tau(n)$ lattice points.

For the case $x^2 + y^2 = 2$, the lattice points are $(1, 1), (1, -1), (-1, -1), (-1, 1)$. These are 4 lattice points and if $n = 2^a$, we will still get 4 lattice points if a is even and 4 when a is odd. When a is even the only lattice points are the trivial ones $(\pm 2^{\frac{a}{2}}, 0)$ which do not contribute to the eigenfunctions. This finishes the proof.

Appendix I: The Divisor Function and Dirichlet's Theorem

Since the divisor function $\tau(n)$ appeared in the formula for the dimension of the eigenspaces, we end these notes with an appendix that establishes bounds for $\tau(n)$ and discuss a statistical average for $\tau(n)$ obtained by Dirichlet and some open problems. We note that the divisor function is very wild and fluctuates a lot from being $\tau(p) = 2$ for p a prime to being large for composite numbers. Typically in Number Theory, fluctuations can be smoothed out by averaging.

Proposition: Given any $\epsilon > 0$, there exists a constant C_ϵ only dependent on ϵ, such that

$$\tau(n) \leq C_\epsilon n^\epsilon.$$

This proposition gives us a bound on the dimension of the eigenspace in terms of the eigenvalue n.

Proof: We use the formula for the number of divisors in the last section. For

$$n = \Pi_{i=1}^k p_i^{a_i},\ \tau(n) = \Pi_{i=1}^k (1 + a_i).$$

We take a function $f(n)$ to be specified later and consider those primes in the decomposition for n such that $p_i \geq f(n)$. Then for such primes p_i, $i \in S$ we have,

$$f(n)^{\sum_{i \in S} a_i} \leq \Pi_{i \in S} p_i^{a_i} \leq n.$$

Thus

$$\log f(n) \sum_{i \in S} a_i \leq \log n \tag{1}.$$

For those indices i not in S, we have $p_i \leq f(n)$. The number of such indices, by the Prime number theorem is bounded by

$$\frac{f(n)}{\log f(n)}. \tag{2}$$

We have using $p_i \geq 2$

$$2^{a_i} \leq p_i^{a_i} \leq n.$$

Thus,

$$1 + a_i \leq C \log n,$$

So for $i \notin S$ we get using (2)

$$\sum_{i \notin S} \log(1 + a_i) \leq C \frac{f(n) \log \log n}{\log f(n)} \tag{3}$$

Now

$$\log \tau(n) \leq \sum_{i \notin S} \log(1 + a_i) + \sum_{i \in S} \log(1 + a_i)$$

$$\leq \sum_{i \notin S} \log(1 + a_i) + C \sum_{i \in S} a_i.$$

Using (1) and (3) we get

$$\log \tau(n) \leq C \frac{f(n) \log \log n}{\log f(n)} + C \frac{\log n}{\log f(n)}.$$

We equalize the two terms by choosing

$$f(n) = \frac{\log n}{\log \log n}.$$

Thus,

$$\log \tau(n) \leq C \frac{\log n}{\log \log n}.$$

This yields,

$$\tau(n) \leq n^{C/\log \log n},$$

which proves the proposition.

Now we pass to Dirichlet's theorem which is proved by counting the lattice points under a hyperbola, a method invented by Dirichlet.

Theorem(Dirichlet): We have

$$\sum_{n \leq x} \tau(n) = x \log x + (2\gamma - 1)x + O(x^{1/2}),$$

where γ is the Euler-Mascheroni constant. Dirichlet himself noted that the error bound is not sharp. In fact Voronoi improved the error bound to $O(x^{1/3})$. The theorem of Voronoi can be proved using the Fourier series for the periodic sawtooth function and then applying the classical Van der Corput lemma to estimate oscillatory sums. Voronoi's result has been improved by Kolesnik, Iwaniec and Mozzochi and others. Hardy proved in 1912 that one cannot improve the error term beyond $O(x^{1/4})$. It remains an open problem to obtain Hardy's bound. Dirichlet's theorem shows that statistically on the average the divisor function behaves like a logarithm which should be compared with the proposition we proved above.

Appendix II: Voronoi's Theorem

We shall prove Voronoi's theorem with a slight loss. We prove

Theorem: Let $\epsilon > 0$ be any positive number. Let $\tau(n)$ be the number of divisors of n. Then we have

$$\sum_{n \leq x} \tau(n) = x \log x + (2\gamma - 1)x + O(x^{\frac{1}{3}+\epsilon}).$$

Our proof is essentially self-contained except we assume the bound given by Van der Corput's lemma, the easy proof of which can be found in the book by Iwaniec and Kowalski [3].

Notation: We shall use the notation $[x]$ for the integer part of x. $\{x\}$ denotes the fractional part of x. Thus

$$x = [x] + \{x\}.$$

We will also need to avail of the sawtooth and periodic function, with period 1(see Fig. 4.)

$$\psi(x) = x - [x] - \frac{1}{2}.$$

As is standard in Number Theory we use the notation

$$e^{2\pi\imath nx} = e(nx), \imath = \sqrt{-1}.$$

Lastly we denote the distance of a point x to the nearest integer as

$$||x|| = \inf_{m} |x - m|, \ m \in Z.$$

Lemma 1: Since $\psi(x)$ is periodic with period 1, it has the Fourier expansion,

$$\psi(x) \sim -\frac{1}{2\pi\imath} \sum_{n \neq 0} \frac{e(nx)}{n}.$$

Proof: This is straightforward. Note that

$$\int_0^1 \psi(x)\, dx = 0,$$

and the Fourier coefficients are given by

$$c_n = \int_0^1 \psi(x) e(-nx)\, dx.$$

This easily yields the Lemma.

Lemma 2: We have for any $H > 0$

$$\psi(x) = -\frac{1}{2\pi\imath} \sum_{0 \neq |n| \leq H} \frac{e(nx)}{n} + O(\frac{1}{1 + H||x||}).$$

Proof: We first note by summing a finite Geometric series,

$$|\sum_{M \leq |n| \leq M+N} e(nx)| \leq C \min(N, \frac{1}{||x||}).$$

Thus by summation by parts,

$$|\frac{1}{2\pi\imath} \sum_{|n|>H} \frac{e(nx)}{n}| \leq \frac{C}{||x||} \sum_{|n|>H} \frac{1}{n^2}.$$

This yields,

$$|\frac{1}{2\pi\imath} \sum_{|n|>H} \frac{e(nx)}{n}| \leq \frac{C}{H||x||}.$$

So it is enough to consider the situation $H||x|| < 1$ to estimate the error and finish the proof of the lemma. We write the error as

$$\psi(x) + \frac{1}{2\pi\imath} \sum_{0 \neq |n| \leq H} \frac{e(nx)}{n}.$$

$\psi(x)$ is already bounded by 1 . Thus we show the finite sum above is uniformly bounded when $H||x|| < 1$. To do that we observe that the finite sum above is

$$\frac{1}{\pi} \sum_{0<n\leq H} \frac{\sin 2\pi nx}{n}.$$

Next the Dirichlet kernel is given by

$$D_H(t) = \frac{1}{2} + \sum_{n=1}^{H} \cos(2\pi nt).$$

By summing a finite geometric series, we see,

$$|D_H(t)| \leq C \min(H, \frac{1}{||t||}). \tag{1}$$

So we have

$$\frac{1}{\pi} \sum_{0<n\leq H} \frac{\sin 2\pi nx}{n} = \int_{[x]}^{x} (D_H(t) - \frac{1}{2}) dt.$$

By changing variables $t \to u + [x]$ and (1) we obtain,

$$|\frac{1}{\pi} \sum_{0<n\leq H} \frac{\sin 2\pi nx}{n}| \leq C \int_0^{||x||} H\, dt \leq CH||x|| \leq C.$$

This proves the Lemma.

We next need a standard device in Number Theory, the Euler-Maclaurin formula.

Lemma 3: We have,

$$\sum_{a<n\leq b} f(n) = \frac{f(b)-f(a)}{2} + \int_a^b f(t)\, dt + \int_a^b f'(t)\psi(t)\, dt.$$

Proof: By a straightforward integration by parts we see,

$$\int_n^{n+1} f'(t)(t-n-\frac{1}{2})\, dt = \frac{f(n)+f(n+1)}{2} - \int_n^{n+1} f(t)\, dt.$$

The left side can be written as

$$\int_n^{n+1} f'(t)\psi(t)\, dt.$$

Now summing over $a \leq n \leq b-1$ we obtain

$$\sum_{a<n<b} f(n) + \frac{f(b)+f(a)}{2} = \int_a^b f(t)\, dt + \int_a^b f'(t)\psi(t)\, dt.$$

From the identity above the lemma follows immediately.

We draw a consequence from the previous Lemma.

Lemma 4:

$$\sum_{n=1}^{M} \frac{1}{n} = \log M + \gamma + \frac{1}{2M} + O(M^{-2}),$$

where γ is the Euler-Mascheroni constant.

Proof: We apply the Euler-Maclaurin formula to $f(t) = 1/t$, to get

$$\sum_{n=1}^{M} \frac{1}{n} = \frac{1}{2M} + \frac{1}{2} + \int_1^M \frac{dt}{t} - \int_1^M \frac{\psi(t)}{t^2}\, dt.$$

Now

$$\gamma = \frac{1}{2} - \int_1^\infty \frac{\psi(t)}{t^2} dt.$$

Thus we get,

$$\sum_{n=1}^{M} \frac{1}{n} = \log M + \gamma + \frac{1}{2M} + \int_M^\infty \frac{\psi(t)}{t^2} dt. \tag{2}$$

We integrate by parts the integral on the right, remembering by periodicity that

$$\Big| \int_A^B \psi(t) dt \Big| \leq 1,$$

to get

$$\Big| \int_M^\infty \frac{\psi(t)}{t^2} \Big| \leq C \int_M^\infty \frac{dt}{t^3} = O(M^{-2}).$$

Inserting the last bound into (2) we obtain our lemma.

We now come to a key lemma which is obtained by the method of Dirichlet by counting lattice points under a hyperbola. See Fig.5

Lemma 5: We have

$$\sum_{n=1}^{[x]} \tau(n) = x \log x + (2\gamma - 1)x - 2 \sum_{n=1}^{[\sqrt{x}]} \psi(\frac{x}{n}).$$

Corollary(Dirichlet):

$$\sum_{n=1}^{[x]} \tau(n) = x \log x + (2\gamma - 1)x + O(x^{\frac{1}{2}})$$

Proof: The Corollary follows by observing that $|\psi(t)| \leq 1$ and using this fact in Lemma 5.

We now prove Lemma 5. It is helpful to look at the figures.

Proof: We count the lattice points under the hyperbola $XY = x$. The lattice points can be obtained by counting the lattice points in Region I and Region II in Fig 5. But this double counts the lattice points in Region III in Fig. 6. Thus the lattice points in Region III have to subtracted. We obtain by this geometric argument

$$\sum_{n=1}^{[x]} \tau(n) = 2 \sum_{n=1}^{[\sqrt{x}]} \left[\frac{x}{n} \right] - [\sqrt{x}]^2. \tag{3}$$

Now by definition

$$\left[\frac{x}{n} \right] = \frac{x}{n} - \psi(\frac{x}{n}) - \frac{1}{2}.$$

On inserting the identity above into (3) we get

$$2x\sum_{n=1}^{[\sqrt{x}]}\frac{1}{n}-2\sum_{n=1}^{[\sqrt{x}]}\psi(\frac{x}{n})-[\sqrt{x}]-[\sqrt{x}]^2. \tag{4}$$

We next use Lemma 4 with $M=[\sqrt{x}]$ to get

$$\sum_{n=1}^{[\sqrt{x}]}\frac{1}{n}=\log[\sqrt{x}]+\gamma+\frac{1}{2[\sqrt{x}]}+O(x^{-1}). \tag{5}$$

Inserting (5) into (4) we get,

$$\sum_{n=1}^{[x]}\tau(n)=2x\log[\sqrt{x}]+2\gamma x-2\sum_{n=1}^{[\sqrt{x}]}\psi(\frac{x}{n})+\frac{x}{[\sqrt{x}]}-[\sqrt{x}]-[\sqrt{x}]^2+O(1). \tag{6}$$

Now

$$x=x^{1/2}x^{1/2}=[\sqrt{x}]^2+O([\sqrt{x}]).$$

Thus the right side of (6) simplifies to

$$2x\log[\sqrt{x}]+2\gamma x-2\sum_{n=1}^{[\sqrt{x}]}\psi(\frac{x}{n})-[\sqrt{x}]^2+O(1). \tag{7}$$

Next we note that

$$[\sqrt{x}]^2=(\sqrt{x}+\{\sqrt{x}\})^2=x+2\sqrt{x}\{\sqrt{x}\}+O(1). \tag{8}$$

Lastly we observe using (8) again

$$2x\log[\sqrt{x}]=x\log[\sqrt{x}]^2=x\log(x+2\sqrt{x}\{\sqrt{x}\}+O(1))=x\log x+2\sqrt{x}\{\sqrt{x}\}+O(1)). \tag{9}$$

Inserting (8) and (9) into (7) we obtain our Lemma.

Next note that

$$\sum_{n=1}^{[\sqrt{x}]}\psi(\frac{x}{n})=\sum_{n\le x^{1/3}}\psi(\frac{x}{n})+\sum_{x^{1/3}<n<x^{1/2}}\psi(\frac{x}{n}).$$

The first sum on the right in view of the fact that $|\psi(t)|\le 1$ is already $O(x^{1/3})$. Therefore to obtain the result of Voronoi we will prove

Lemma 6:

$$\Big|\sum_{x^{1/3}<n<x^{1/2}}\psi(\frac{x}{n})\Big|\le Cx^{\frac{1}{3}+\epsilon},$$

where C only depends on ϵ.

A key ingredient of the proof is the Lemma of Van der Corput which is a simple consequence of Poisson summation and any elementary version of the stationary phase formula.

Lemma 7(Van der Corput): Let $|f''(t)| \geq \Lambda$ on $[a,b]$. Then,

$$\Big| \sum_{a \leq n \leq b} e(f(n)) \Big| \leq C(b-a)\Lambda^{1/2}.$$

Lemma 6 will follow from the next Lemma

Lemma 8: For $N = 2^k x^{1/3}$ $k = 1, 2, \cdots k_0$, where $2^{k_0} x^{1/3} = x^{1/2}$ we have, for any $\epsilon > 0$

$$\big| \sum_{N<n<2N} \psi(\frac{x}{n}) \big| \leq C x^{\frac{1}{3}+\epsilon}.$$

Lemma 6 is an immediate consequence since

$$\big| \sum_{x^{1/3}<n<x^{1/2}} \psi(\frac{x}{n}) \big| \leq \sum_{1 \leq k \leq k_0} \big| \sum_{2^k x^{1/3} \leq n < 2^{k+1} x^{1/3}} \psi(\frac{x}{n}) \big| \leq C x^{\frac{1}{3}+\epsilon}.$$

Proof(Lemma 8): We have via Lemma 2

$$\sum_{N<n<2N} \psi(\frac{x}{n}) = \sum_{0<|h|\leq H} \sum_{N \leq n < 2N} \frac{e(\frac{hx}{n})}{h} + O\Big(\sum_{N \leq n < 2N} \frac{1}{1+H||\frac{x}{n}||} \Big). \tag{10}$$

We will apply the Van der Corput Lemma 7 to the inner sum in n to the first term on the right with $f(t) = hx/t$. Observe on $[N, 2N]$, $|f''(t)| \geq 2|h|x/N^3$ so Lemma 7 gives

$$\Big| \sum_{0<|h|\leq H} \sum_{N \leq n < 2N} \frac{e(\frac{hx}{n})}{h} \Big| \leq C \sum_{1 \leq |h| \leq H} \frac{x^{\frac{1}{2}}}{|h|^{1/2} N^{1/2}} \leq C \frac{H^{1/2} x^{1/2}}{N^{1/2}}. \tag{11}$$

We next estimate the error term in (9). For $k \in Z$ such that $|\frac{x}{n} - k| < \frac{1}{2}$ we obtain

$$||\frac{x}{n}|| = |\frac{x}{n} - k| < \frac{1}{2}, |x - kn| < \frac{n}{2} < N < x^{1/2}.$$

Setting $kn = m$ we obtain,

$$|m| \leq |x - m| + x < 2x.$$

Thus for fixed $kn = m$ there are $\tau(m)$ solutions to $|x - kn| < x$. But $\tau(m) \leq C m^{\epsilon} < C x^{\epsilon}$ by the previous section and so the error term is bounded by

$$C N x^{\epsilon} \sum_{0 \leq m < x} \frac{1}{1 + H|x-m|} \leq C \frac{N x^{\epsilon}}{H}, \tag{12}$$

where we used

$$\frac{1}{H||\frac{x}{n}||} = \frac{1}{H|\frac{x}{n} - k|} = \frac{n}{H|x - kn|} \leq \frac{N}{H|x - m|}.$$

Combining (11), (12) we get

$$\Big| \sum_{N<n<2N} \psi(\frac{x}{n}) \Big| \leq C\frac{H^{1/2}x^{1/2}}{N^{1/2}} + C\frac{Nx^{\epsilon}}{H}.$$

Equalizing the two terms by setting $H = N/x^{1/3} > 1$ we obtain our lemma and Voronoi's theorem.

Remarks: G. H. Hardy [1] has shown that in the theorem we have established we cannot obtain a better error term than $O(x^{1/4})$. Though there have been improvements to Voronoi's result by Kolesnik, H. Iwaniec and Mozzochi and others, Hardy's optimal error estimate has not been attained. See also Ingham[2] for the optimal error bounds.

Appendix III: Proof of Rabinovitch's Theorem

In this appendix we shall prove Rabinovitch's theorem on prime spitting polynomials that was stated earlier. We have a quadratic field $K = Q(\sqrt{d})$ where Q denotes the rationals and $d = 4k + 1$ and we denote by $\mathcal{O}_K$ the ring of integers associated with this field. We also recall that elements of $\mathcal{O}_K$ are of the form

$$m + n\omega,\ \omega = \frac{1+\sqrt{d}}{2},\ m, n \in Z.$$

We shall denote ideals I in $\mathcal{O}_K$ by displaying the generators

$$I = [n, m(b+\omega)], m, n \in Z. \tag{1}$$

Recall the norm of the ideal $N(I) = mn$ and is equal to the number of elements in $\mathcal{O}_K/I$. With these preliminary facts out of the way, we can set ourselves to prove the theorem.

Proof: Recall we had set $q = \frac{1+|d|}{4}$ and

$$f_q(x) = x^2 + x + q.$$

First observe that

$$f_q(q-2) = (q-2)^2 + (q-2) + q = q^2 - 2q + 2 < q^2.$$

Assume by contradiction the polynomial does not spit out primes, i.e there exists $a \leq q-2$ such that $f_q(a)$ is not a prime. Let a prime p then divide $f_q(a)$. It follows from the computation above that $p < q^2$. Thus w.l.o.g $p \leq q-1$. Consider the ideal in $\mathcal{O}_K$ given by

$$I = [p, a+\omega].$$

Notice that $N(a+\omega) = a^2 + a + q$. By our contradiction assumption $p/N(a+\omega)$ and thus I is indeed an ideal and clearly $N(I) = p$. We now establish that I is **not** a principal ideal. Assume by contradiction it is. Then there exists $\gamma \in \mathcal{O}_K$ such that

$$I = [p, a+\omega] = (\gamma).$$

Then we have equating norms $N(\gamma) = p$. Since γ is of the type given in (1), we obtain

$$N(\gamma) = (m + \frac{n}{2})^2 + \frac{n^2|d|}{4} = p \leq q - 1 = \frac{1+|d|}{4} - 1 < \frac{|d|}{4}.$$

This can only hold if $n = 0$. But then the prime p is a perfect square,

$$p = (m + \frac{n}{2})^2.$$

This is nonsense. Thus the ideal I is not principal and so $h(d) \neq 1$. This proves that if $h(d) = 1$, then $f_q(x)$ is prime for $0 \leq x \leq q-2$.

We now prove the converse. Assume that $f_q(x)$ is prime for $0 \leq x \leq q-2$. Note from the beginning parts of these notes on Reduction theory we can show that all ideals in $\mathcal{O}_K$ are equivalent to ideals whose norm is less than $\sqrt{|d|}/3$ since we showed that quadratic forms via a $SL(2,Z)$ transformation can be brought to a quadratic form with $a < \sqrt{|d|/3}$. We want to establish all such equivalent ideals are principal and thus $h(d) = 1$. The reduction above can also be done via Minkowski's theorem on lattice points in convex bodies. Moreover we can even reduce the situation to prime ideals. Thus we consider a non-principal prime ideal I such that

$$I = [p, a + \omega], a \in Z$$

where p is a prime. We can also with no loss of generality assume $0 \leq a \leq p-1$. Now by our reduction above

$$N(I) = p < \sqrt{|d|/3} \leq \frac{1+|d|}{4} - 2 = q - 2, |d| > 3. \tag{2}$$

For the small values of d our theorem is a direct check and so we can assume that $|d| > 3$ with no loss of generality. Since I is an ideal, $p/N(a+\omega) = f_q(a)$. But $a < p \leq q-2$ and so by hypothesis, $f_q(a)$ must be prime. Thus,

$$f_q(a) = p.$$

Next $f_q(0) = q$. The polynomial $f_q(x)$ is monotonic increasing for $x > 0$ by elementary calculus and so

$$\frac{1+|d|}{4} = q = f_q(0) < f_q(a) = p < \sqrt{|d|/3}.$$

But this is nonsense. Thus the ideal I has to be principal.

We thank Po-lam Yung for useful remarks that improved the presentation.

REFERENCES

[1] G. H. Hardy, On Dirichlet's Divisor Problem, Proc. London Math. Society, **15**(2), (1917), 1-25

[2] A. E. Ingham, On Two classical Lattice Point Problems, Proc. Cambridge Philosophical Soc., **36**(1940),131-138.

[3] H. Iwaniec and E. Kowalski, *Analytic Number Theory*, American Mathematical Society, Colloquium Publications.

Sagun Chanillo
Deptt. of Math., Rutgers Univ.,
110 Frelinghuysen Rd.,
Piscataway, NJ 08854, USA
chanillo@math.rutgers.edu

Column 1	Column 2
$m+\sqrt{-1}n$	$m-\sqrt{-1}n$
$m+\sqrt{-1}n$	$m-\sqrt{-1}n$
⋮	⋮
$m+\sqrt{-1}n$	$m-\sqrt{-1}n$

a_i rows.

multiply all of these to get $z = M_0 + \sqrt{-1}N_0$

Fig. 1.

Transposing the entries in Row 1. once

$m-\sqrt{-1}n$	$m+\sqrt{-1}n$
$m+\sqrt{-1}n$	$m-\sqrt{-1}n$
$m+\sqrt{-1}n$	$m-\sqrt{-1}n$
⋮	⋮
$m+\sqrt{-1}n$	$m-\sqrt{-1}n$

a_1 rows.

→ multiply all of these to get

$z = M_1 + \sqrt{-1}N_1$.

Fig. 2.

$m-\sqrt{-1}n$	$m+\sqrt{-1}n$
$m-\sqrt{-1}n$	$m+\sqrt{-1}n$
$m+\sqrt{-1}n$	$m-\sqrt{-1}n$
$m+\sqrt{-1}n$	$m-\sqrt{-1}n$
⋮	⋮
$m+\sqrt{-1}n$	$m-\sqrt{-1}n$

a_1 rows

Fig. 3.

multiply all of these to get $M_2 + \sqrt{-1}N_2$.

The process ends when Column 2 in Fig. 1. becomes Column 1.

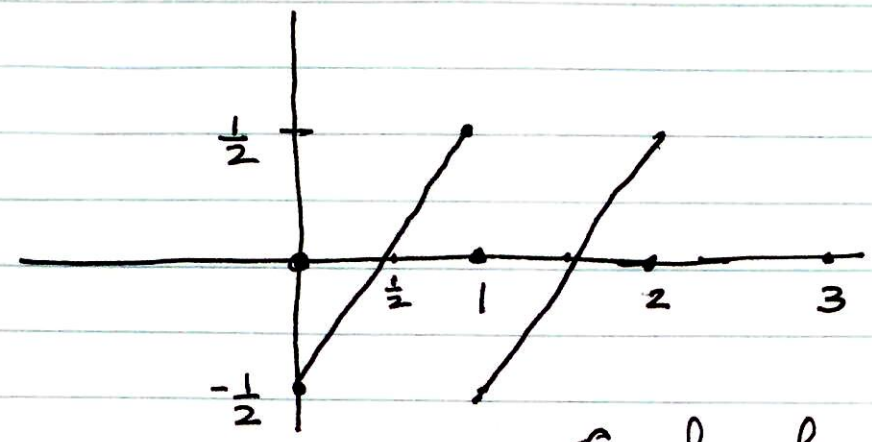
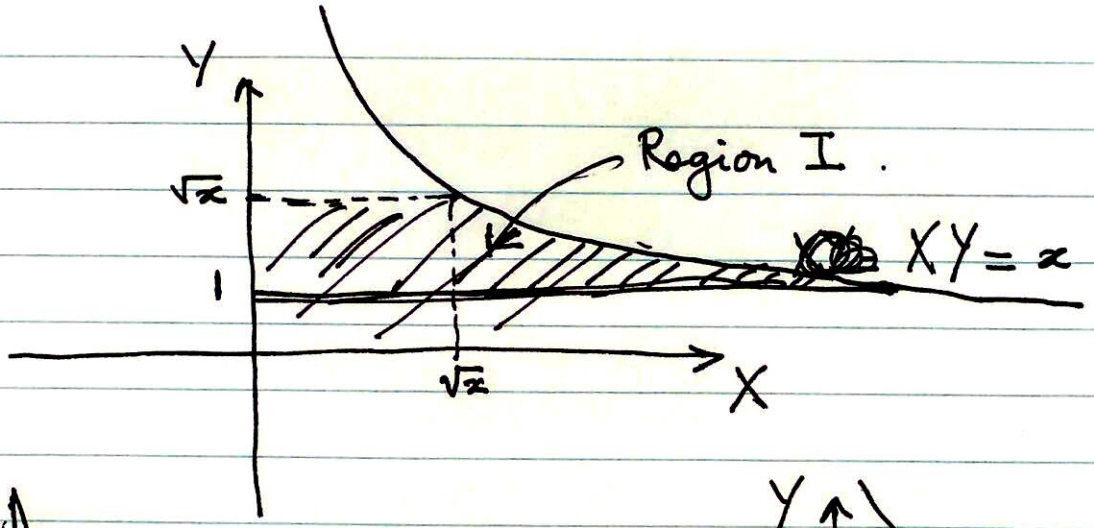
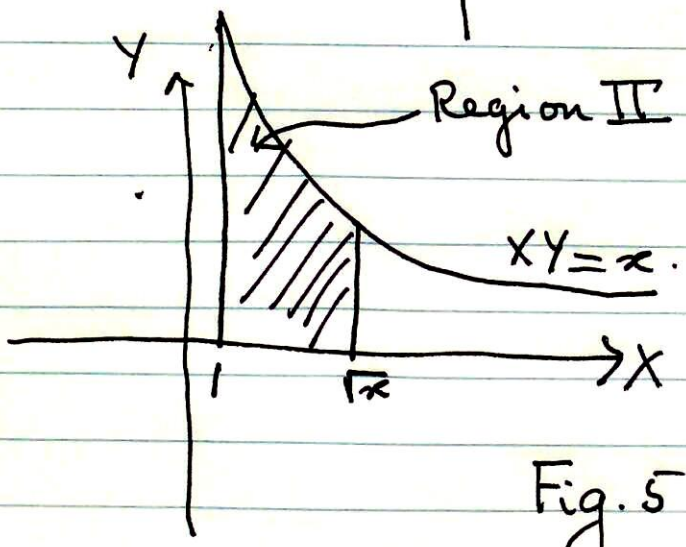


Fig. 4.

Graph of the saw-tooth fn. $\psi(t)$.





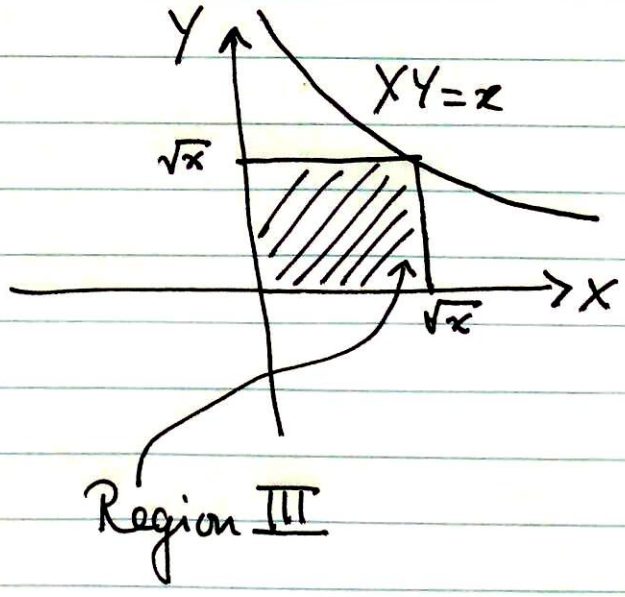


Fig. 5.