When this occurs $g(n) \geq |F| \geq (1-\epsilon)\mu$ but again we must worry about $g(n)$ being considerably larger than $|F|$. Here we use only that $p = n^{-2/3+o(1)}$. Note that the number of representations of $n = x+y+z$ with a given x is the number of representations $m = y+z$ of $m = n-x$.
Lemma 3.3. Almost surely no sufficiently large m has four (or more) representations as $m = y+z$, $y, z \in S$.
Proof. Here $\mu = \Theta(m^{-1/3})$ so the expected number of 4-tuples of representatives is $O(m^{-4/3})$ and so the probability of having four representatives is $O(m^{-4/3})$. Apply Borel-Cantelli. □

Now almost surely there is a C so that *no* m has more than C representations $m = y+z$. Let S be such that this holds and that all maximal disjoint families of solutions F have

$$K(1-\epsilon)\log n < |F| < K(1+\epsilon)\log n$$

Each triple $x, y, z \in S$ with $x+y+z = n$ must include one of the at most $3K(1+\epsilon)\log n$ elements of sets of F and each such element is in less than C such triples so that $g(n) < 3CK(1+\epsilon)\log n$. Take $c_1 = K(1-\epsilon)$ and $c_2 = 3KC(1+\epsilon)$.

With additional work one can prove Theorem 3.2 with $c_1 = K(1-\epsilon'), c_2 = K(1+\epsilon')$ for arbitrarily small ϵ' and K dependent only on ϵ'.

The full result of Erdős and Tetali was that for each k there is a set S and constants c_1, c_2 so that the number of representations of n as the sum of k terms of S lies between $c_1 \log n$ and $c_2 \log n$ for all sufficiently large n. Proof. Define S randomly by

$$\Pr[x \in S] = p_x = \min\left[10\left(\frac{\ln x}{x^2}\right)^{1/3}, \frac{1}{2}\right]$$

Fix n. Now $g(n)$ is a random variable and

$$\mu = E[g(n)] = \sum_{x+y+z=n} p_x p_y p_z$$

Careful asymptotics give

$$\mu \sim 10^3 \ln n \int_{x=0}^{1} \int_{y=0}^{1-x} \frac{dx\,dy}{[xy(1-x-y)]^{2/3}} = K \ln n$$

where K is large. (We may make K arbitrarily large by increasing "10".) We apply the Janson inequality. Here $\epsilon = 1/8$ as all $p_x \le 1/2$. Also

$$\Delta = \sum p_x p_y p_z p_{y'} p_{z'},$$

the sum over all five-tuples with $x + y + z = x + y' + z' = n$. Roughly there are n^3 terms, each $\sim p_n^5 = n^{-10/3+o(1)}$ so that the sum is $o(1)$. Care must be taken that those terms with one (or more) small variables don't contribute much to the sum.

Now we emulate the argument of Theorem 5.3.1. Call F a maximal disjoint family of solutions if F is a family of sets $\{x_i, y_i, z_i\}$ with all x_i, y_i, z_i distinct, all $x_i + y_i + z_i = n$, all $x_i, y_i, z_i \in S$ and so that there is no $x, y, z \in S$ with $x + y + z = n$ and x, y, z distinct from all x_i, y_i, z_i. Let $Z^{(s)}$ denote the number of maximal disjoint families of solutions of size s. As in Theorem 5.3.1 when $s < \log^2 n$

$$E[Z^{(s)}] < \frac{\mu^s}{s!} e^{-\mu(1+o(1))}$$

while for $s \ge \log^2 n$

$$E[Z^{(s)}] < \mu^s / s!$$

so that $\sum^* E[Z^{(s)}] = o(n^{-10})$, say, where $\sum^*$ is over those s with $|s - \mu| > \epsilon\mu$. (Here ϵ is fixed and K must be sufficiently large.) With probability $1 - o(n^{-10})$ there is an F with $|s - \mu| < \epsilon\mu$.

Proof. Define S randomly by

$$\Pr[x \in S] = p_x = \min\left[10\sqrt{\frac{\ln x}{x}}, 1\right]$$

Fix n. Now $f(n)$ is a random variable with mean

$$\mu = E[f(n)] = \sum_{x+y=n} p_x p_y$$

Roughly there are n addends with $p_x p_y > p_n^2 = 100\frac{\ln n}{n}$. We have $p_x p_x = \Theta(\frac{\ln n}{n})$ except in the regions $x = o(n), y = o(n)$ and care must be taken that those terms don't contribute significantly to μ. Careful asymptotics (and first year Calculus!) yield

$$\mu \sim (100 \ln n) \int_0^1 \frac{dx}{\sqrt{x(1-x)}} = 100\pi \ln n$$

The negligible effect of the $x = o(n), y = o(n)$ terms reflects the finiteness of the indefinite integral at poles $x = 0$ and $x = 1$. The possible representations $x + y = n$ are mutually independent events so that from basic Large Deviation results

$$\Pr[|f(n) - \mu| > \epsilon\mu] < 2(1-\delta)^{\mu}$$

for constants ϵ, δ. To be specific we take $\epsilon = .9, \delta = .1$ and

$$\Pr[|f(n) - \mu| > .9\mu] < .9^{314 \ln n} < n^{-1.1}$$

for n sufficiently large. Take $c_1 < .1(100\pi)$ and $c_2 > 1.9(100\pi)$.

Let A_n be the event that $c_1 \ln n \leq f(n) \leq c_2 \ln n$ does *not* hold. We have $\Pr[A_n] < n^{-1.1}$ for n sufficiently large. The Borel Cantelli Lemma applies, almost always all A_n fail for n sufficiently large. Therefore there exists a specific point in the probability space, i.e., a specific set S, for which $c_1 \ln n \leq f(n) \leq c_2 \ln n$ for all sufficiently large n. □

Now for a given set S of natural numbers let $g(n) = g_S(n)$ denote the number of representations $n = x + y + z$, $x, y, z \in S$, all unequal.

Theorem 3.2. (Erdős, Tetali[1990]) There is a set S and a positive constants c_1, c_2 so that

$$c_1 \log n \leq g(n) \leq c_2 \log n$$

for all sufficiently large n.

Thus $e^{-\mu} < n^{-100+o(1)}$. The addends of Δ break into two parts, those $\Pr[A_F \wedge A_{F'}]$ with $|F \cap F'| = 1$ and those with $|F \cap F'| = 2$. The bounds on $r_3(n)$ give that there are at most $n^{3/2+o(1)}$ pairs F, F' of the first type and each has

$$\Pr[F \cap F'] = p^7 = n^{-7/4+o(1)}$$

The bounds on $r_2(n)$ give that there are at most $n^{1+o(1)}$ pairs F, F' of the second type and each has

$$\Pr[F \cap F'] = p^6 = n^{-3/2+o(1)}$$

Hence

$$\Delta \leq n^{3/2+o(1)-7/4+o(1)} + n^{1+o(1)-3/2+o(1)} = o(1)$$

Thus

$$\Pr[\wedge_{F \in \mathcal{F}_n} \overline{A}_F] \leq (1+o(1))e^{-\mu} \leq n^{-100+o(1)}$$

As $\sum n^{-100+o(1)}$ converges the Borel-Cantelli lemma gives that almost always all sufficiently large $n \not\equiv 0(\text{mod}4)$ will be the sum of four elements of X.

Remark The constant "10" could be made smaller as long as the exponent of n here is less than -1.

Let X be a particular set having the above properties. (As customary, the probabilistic method does not actually "construct" X.) Suppose all $n \geq n_0$, $n \not\equiv 0(\text{mod}4)$ are the sum of four elements of X. Add to X all squares up to n_0. This does not affect the asymptotics of $N_X(x)$ and now all $n \not\equiv 0(\text{mod}4)$ are the sum of four elements of X. Finally, replace X by $X \cup 4X \cup 4^2X \cup 4^3X \cup \ldots$. This affects the asyptotics of $N_X(x)$ only by a constant and now *all* integers are the sum of four elements of X.

3 Counting Representations

For a given set S of natural numbers let (for every $n \in N$) $f(n) = f_S(n)$ denote the number of representations $n = x + y$, $x, y \in S, x \neq y$. For many years it was an open question whether there existed an S with $f(n) \geq 1$ for all sufficiently large n and yet $f(n) \leq n^{o(1)}$.

Theorem 3.1. (Erdős (1956)) There is a set S for which $f(n) = \Theta(\ln n)$. That is, there is a set S and constants c_1, c_2 so that for all sufficiently large n

$$c_1 \ln n \leq f(n) \leq c_2 \ln n$$

$N_X(x) = \Omega(x^{1/4})$. Our object here is to give a quick proof of the following result of Wirsing.

Theorem. There is a set $X \subseteq S$ such that every $n \geq 0$ can be expressed as the sum of four elements of X and

$$N_X(x) = O(x^{1/4}(\ln x)^{1/4})$$

In 1828 Jacobi showed that the number $r_4(n)$ of solutions in integers to $n = a^2 + b^2 + c^2 + d^2$ is given by eight times the sum of those $d|n$ with $d \not\equiv 0(\text{mod} 4)$. In 1801 Gauss found an exact expression for the number $r_2(n)$ of solutions in integers to $n = a^2 + b^2$. We will need only $r_2(n) = n^{o(1)}$ which follows easily from his results. From this the number $r_3(n)$ of solutions to $n = a^2 + b^2 + c^2$ is $O(n^{1/2+o(1)})$. Now suppose $n \not\equiv 0(\text{mod} 4)$. Then $r_4(n) > 8n$ so, excluding order there are at least $n/48$ different solutions to $n = a^2 + b^2 + c^2 + d^2$ in nonnegative integers. From $r_2(n) = n^{o(1)}$ it follows that there are $O(n^{1/2+o(1)})$ solutions with $a = b$. Hence there are at least $(1 + o(1))n/48$ *sets* F of four squares adding to n .

Define a random subset $X \subseteq S$ by

$$\Pr[y \in X] = p_y = 10(\ln y)^{1/4} y^{-1/4}$$

for $y \in S$, $y \geq 10^8$. For definiteness say $\Pr[y \in X] = p_y = 1$ for $y \in S$, $y < 10^8$. Then

$$E[N_X(x)] = \sum_{i=0}^{x^{1/2}} \Pr[i^2 \in X] = O(x^{1/4}(\ln x)^{1/4})$$

and large deviation results give $N_X(x) = O(x^{1/4}(\ln x)^{1/4})$ almost always.

For any given $n \not\equiv 0(\text{mod} 4)$, $n \geq 10^8$, let $\mathcal{F}_n$ denote the family of sets F of four squares adding to n. For each $F \in \mathcal{F}_n$ let A_F be the event $F \subseteq X$. We apply Janson's Inequality to give an upper bound to $\Pr[\wedge_{F \in \mathcal{F}_n} \overline{A_F}]$. Observe that this probability increases when the p_y decrease so, as the function p_y is decreasing in y, we may make the simplifying assumption

$$p_y = p = 10(\ln n)^{1/4} n^{-1/4}$$

for all $y \in S$, $y \leq n$. Then

$$\Pr[A_F] = p^4 = 10^4(\ln n)/n$$

and

$$\mu \geq (1 + o(1))(n/48)10^4(\ln n)/n \geq (100 + o(1))(\ln n)$$

With p, q distinct primes, $X_pX_q = 1$ if and only if $p|x$ and $q|x$ which occurs if and only if $pq|x$. Hence

$$\begin{array}{l} Cov[X_p, X_q] = E[X_p]E[X_q] - E[X_pX_q] \\ = \frac{\lfloor n/pq \rfloor}{n} - \frac{\lfloor n/p \rfloor}{n}\frac{\lfloor n/q \rfloor}{n} \\ \leq \frac{1}{pq} - (\frac{1}{p} - \frac{1}{n})(\frac{1}{q} - \frac{1}{n}) \\ \leq \frac{1}{n}(\frac{1}{p} + \frac{1}{q}) \end{array}$$

Thus

$$\sum_{p \neq q} Cov[X_p, X_q] \leq \frac{1}{n}\sum_{p \neq q} \frac{1}{p} + \frac{1}{q} = \frac{\pi(n) - 1}{n}\sum_{p} \frac{2}{p}$$

where $\pi(n) \sim \frac{n}{\ln n}$ is the number of primes $p \leq n$. So

$$\sum_{p \neq q} Cov[X_p, X_q] < \frac{(n/\ln n)}{n}(2 \ln\ln n) = o(1)$$

That is, the covariances do not affect the variance, $Var[X] \sim \ln\ln n$ and Chebyschev's Inequality actually gives

$$\Pr[|v(n) - \ln\ln n| > \lambda\sqrt{\ln\ln n}] < \lambda^{-2} + o(1)$$

for any constant λ. □

In a classic paper Paul Erdős and Marc Kac [1940] showed, essentially, that X does behave like a normal distribution with mean and variance $\ln\ln n + o(1)$. Here is their precise result.

The Erdős-Kac Theorem. Let λ be fixed, positive, negative or zero. Then

$$\lim_{n \to \infty} \frac{1}{n}|\{x : 1 \leq x \leq n, v(x) \geq \ln\ln n + \lambda\sqrt{\ln\ln n}\}| = \int_{\lambda}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt$$

We do not prove this result here.

2 Four Squares with Few Squares

The classic theorem of Lagrange states that every nonnegative integer n is the sum of four squares. How "sparse" can a set of squares be and still retain the four square property. For any set X of nonnegative integers set $N_X(x) = |\{i \in X, i \leq x\}|$. Let $S = \{0, 1, 4, 9, \ldots\}$ denote the squares. If $X \subseteq S$ and every $n \geq 0$ can be expressed as the sum of four elements of X then how slow can be the growth rate of $N_X(x)$? Clearly we must have

Lecture 6: A Number Theory Interlude

We take a break from Graph Theory and explore applications of these methods to Number Theory.

1 Prime Factors

The second moment method is an effective tool in number theory. Let $v(n)$ denote the number of primes p dividing n. (We do not count multiplicity though it would make little difference.) The following result says, roughly, that "almost all" n have "very close to" $\ln \ln n$ prime factors. This was first shown by Hardy and Ramanujan in 1920 by a quite complicated argument. We give the proof of Paul Turan [1934] a proof that played a key role in the development of probabilistic methods in number theory.
Theorem 1.1 Let $\omega(n) \to \infty$ arbitrarily slowly. Then the number of x in $\{1, \ldots, n\}$ such that

$$|v(x) - \ln \ln n| > \omega(n)\sqrt{\ln \ln n}$$

is $o(n)$.
Proof. Let x be randomly chosen from $\{1, \ldots, n\}$. For p prime set

$$X_p = \begin{cases} 1 & \text{if } p|x \\ 0 & \text{otherwise} \end{cases}$$

and set $X = \sum X_p$, the summation over all primes $p \leq n$, so that $X(x) = v(x)$. Now

$$E[X_p] = \frac{\lfloor n/p \rfloor}{n}$$

As $y - 1 < \lfloor y \rfloor \leq y$

$$E[X_p] = 1/p + O(1/n)$$

By linearity of expectation

$$E[X] = \sum_{p \leq n} \frac{1}{p} + O(\frac{1}{n}) \sim \ln \ln n$$

Now we bound the variance

$$Var[X] \leq (1 + o(1)) \ln \ln n + \sum_{p \neq q} Cov[X_p, X_q]$$