Distribution of the Prime Factors of $p + d$

Neil O'Connell, Thomas Womack*
Basic Research Institute in the Mathematical Sciences
HP Laboratoires Bristol
HPL-BRIMS-97-26
November, 1997

Poisson-Dirichlet,
GEM,
random partition,
prime factors,
p+1

For each positive integer m, there is a natural representation for the factorisation of m as a partition of the unit interval. The elements of this partition can be arranged in non-increasing order, and represented as a point $V(m)$ on the infinite-dimensional simplex. In 1972, Billingsley proved that, if N is a randomly chosen positive integer less than n, then for large n, the law of $V(N)$ can be approximated by the Poisson-Dirichlet distribution (with parameter 1). We prove the following: if P is a randomly chosen prime less than n, and d is a fixed non-zero integer, then for large n the distribution of $V(P + d)$ can be approximated by the same Poisson-Dirichlet distribution. We will discuss some implications of this result in cryptography.

*Oxford University, Oxford, United Kingdom

We will begin by introducing the Poisson-Dirichlet distribution, and related GEM distribution. Denote by

$$C = \prod_{i=1}^{\infty} [0,1]$$

the infinite dimensional unit cube, by

$$\Delta = \left\{ \mathbf{x} \in C : \sum_{i=1}^{\infty} x_i = 1 \right\}$$

the simplex of vectors in C with unit sum, and by

$$T = \{ \mathbf{x} \in \Delta : x_1 \geq x_2 \geq \cdots \}$$

the set of vectors in Δ with non-increasing components. We endow C with the product Euclidean topology and Borel σ-algebra; Δ and T are equipped with the topologies and σ-algebras they inherit as subsets of C. The *GEM* and *Poisson-Dirichlet* distributions are supported on Δ and T, respectively, and are constructed as follows. Let $U_1, U_2, \ldots$ be a sequence of independent random variables, each uniformly distributed on $[0,1]$, and set

$$X_1 = U_1, \; X_2 = (1-U_1)U_2, \; X_3 = (1-U_1)(1-U_2)U_3, \; \ldots$$

The law of the random vector $\mathbf{X} = (X_1, X_2, \ldots)$, which we denote by γ, is the so-called GEM distribution with parameter 1. Note that γ is supported on the simplex Δ. If we write $X_{(1)} \geq X_{(2)} \geq \cdots$ for the non-increasing rearrangement of components of the vector $\mathbf{X}$, then the law of $\mathbf{X}_{()} = (X_{(1)}, X_{(2)}, \ldots)$, which we denote by π, is Poisson-Dirichlet with parameter 1.

We remark that, if the U_i were taken to be independent with common density given by $f(u) = \theta(1-u)^{\theta - 1}$ on $[0,1]$, the distributions γ and π would

be, respectively, GEM and Poisson-Dirichlet with parameter θ. These distributions arise naturally in a variety of contexts, from the study of random permutations and random mappings to Brownian excursion theory and neutral population genetics. References are given in [2].

Now, for each positive integer m, there is a corresponding partition of unity:

$$\frac{\log p_1}{\log m} + \cdots + \frac{\log p_{\Omega(m)}}{\log m} = 1.$$

Here, $\Omega(m)$ is the number of (not necessarily distinct) prime factors $p_1, \ldots, p_{\Omega(m)}$ of m. The elements of this partition can be arranged in non-increasing order and an infinite string of zero's attached to the right, giving an element $V(m)$ of T.

Billingsley [1] proved the following beautiful result.

Theorem 1 (Billingsley) *If N_n is an integer chosen uniformly at random from $\{1, \ldots, n\}$, then the law of the $V(N_n)$ is asymptotically Poisson-Dirichlet with parameter 1.*

Donnelly and Grimmett [2] give an alternative proof, using the GEM construction outlined above. We will also make use of the GEM construction, to prove the following.

Theorem 2 *If d is a non-zero integer, and P_n is a prime chosen uniformly at random from all primes less than n, then the law of $V(P_n + d)$ is asymptotically Poisson-Dirichlet with parameter 1.*

Our motivation for proving such a result comes from cryptography, where primes are quite useful and it is often desireable, for reasons of security, to

use primes p with the property that $p+1$ and/or $p-1$ are hard to factor (contain large primes). A 'corollary' of Theorem 2 is that most primes have this property.

Proof of Theorem 2. Our proof follows closely the proof of Donnelly and Grimmett for the uniform case, up to a certain point, where we make use of the following essential lemma, due to Dirichlet (see for example [4]). Denote by $\pi_{a,b}(n)$ the number of primes less than n which are $\equiv a \pmod{b}$, and by $\phi(n)$ the number of positive integers less than and prime to n.

Lemma 1 *As* $n \to \infty$,

$$\pi_{a,b}(n) \sim \frac{n}{\log(n)\phi(b)}.$$

The integer $P_n + d$, which we will denote by $N \equiv N(n)$, has a prime factorisation in the form

$$N = \prod_p p^{A(p,n)}$$

where $A(p,n)$ is the multiplicity of the prime p. (The product is over the set of all primes.) If $N(n) = 1$, then $A(p,n) = 0$ for all p. Set

$$M(n) = \Omega(N(n)) = \sum_p A(p,n).$$

Just as in [2], we place the prime factors $\alpha_1, \ldots, \alpha_{M(n)}$ of N in random order by *size-biased sampling*: the first term, which we denote by $D_1(n)$, is chosen at random from the sequence α_j, with each α_j being chosen with probability proportional to $\log \alpha_j$. Having chosen the first term, the second term, $D_2(n)$, is chosen similarly from the remaining divisors, and so on. In this way we obtain a sequence $D_1, D_2, \ldots, D_{M(n)}$ of prime divisors of N.

For $i \leq M(n)$, set

$$R_i(n) = \frac{N(n)}{D_1(n)D_2(n)\cdots D_{i-1}(n)}$$

and

$$B_i(n) = \frac{\log D_i(n)}{\log R_i(n)};$$

for $i > M(n)$ we set $B_i(n) = 0$. Note that the vector $\mathbf{B}(n) \in C$. It suffices to prove, just as in [2], that the law of $\mathbf{B}(n)$ converges weakly to Lebesgue measure on C. (That is, the B_i are asymptotically a sequence of independent, uniformly distributed random variables on $[0,1]$.) This would follow if, for each k, the law of the vector $\mathbf{B}_k(n) = (B_1(n), \ldots, B_k(n))$ converges weakly to Lebesgue measure on $[0,1]^k$. Donnelly and Grimmett argue that, to establish this, it suffices to prove that, for any $\mathbf{0} < \mathbf{a} < \mathbf{b} < \mathbf{1}$ in $[0,1]^k$ (the usual partial ordering),

$$\liminf_{n\to\infty} P(\mathbf{a} < \mathbf{B}_n(n) \leq \mathbf{b}) \geq \prod_{i=1}^{k}(b_i - a_i).$$

We refer the reader to the paper of Donnelly and Grimmett for a detailed justification of this claim.

Set

$$Q = \{\mathbf{x} \in [0,1]^k : \ \mathbf{a} < \mathbf{x} \leq \mathbf{b}\}.$$

Now, $\mathbf{B}_k \in Q$ iff $R_i^{a_i} < D_i \leq R_i^{b_i}$ for $1 \leq i \leq k$. So

$$P(\mathbf{B}_k \in Q) = \sum_{\mathbf{p},m} P(N = m, \mathbf{D}_k = \mathbf{p}),$$

where $\mathbf{p}$ is the set of vectors with prime elements and m is restricted by the inequalities

$$(\frac{m}{\prod_{j=1}^{i-1} p_j})^{a_i} \leq p_i \leq (\frac{m}{\prod_{j=1}^{i-1} p_j})^{b_i}.$$

Note that the probabilities summed will very frequently be zero, and that, for certain $\mathbf{p}$ and Q, the inequalities will prevent any values of m from being valid.

It is convenient to fix $\epsilon > 0$ and restrict to $m > \epsilon n$ (losing at most $O(\epsilon)$ to the value of the summand, by the prime number theorem). We can also assume that $n > \max(1, \epsilon^{-2})$, to avoid problems with divergence for n small. Reversing the order, the summation is now over the whole of $\epsilon n < m < n$, but we are restricting the vector $\mathbf{p}$ by requiring $n_i^{a_i} < p_i < (\epsilon n_i)^{b_i}$, $n_i = n/\prod_{j=1}^{i-1} p_j$.

For these $\mathbf{p}$ and m, and for n sufficiently large

$$P(N = m, \mathbf{D} = \mathbf{p}) \geq (1 - \epsilon)\frac{\log n}{n} \prod_{i=1}^{k} \frac{\log p_i}{\log(m/p_1 p_2 \dots p_{i-1})}$$

whenever $p_1 p_2 \dots p_k$ actually divides m, and zero otherwise. Here, the first factor is to account for $m - d$ being prime (and uses the prime number theorem), and the second arises from the size-biased choice process.

The inequality is preserved by replacing m with n in the denominator of the second factor, yielding

$$P(N = m, \mathbf{D} = \mathbf{p}) \geq (1 - \epsilon)\frac{\log n}{n} \prod_{i=1}^{k} \frac{\log p_i}{\log n_i}.$$

We now perform the summation over values of m of the form $p + d$ and with $p_1 p_2 \dots p_k$ dividing m. Since the summand is independent of m, this ammounts to counting the number of terms in the summation which, by Dirichlet's lemma (Lemma 1) is of order

$$\frac{1}{\phi(\prod p_i)} \left(\frac{n}{\log n} - \frac{\epsilon n}{\log n + \log \epsilon} \right);$$

for n large enough, this at least $(1-2\epsilon)n\log n/\phi(\prod p_i)$. Since

$$\phi(\prod p_i) = \prod(p_i - 1) < \prod p_i,$$

we can conclude that, for n sufficiently large,

$$P(\mathbf{B}_k(n) \in Q) \geq (1-\epsilon)(1-2\epsilon)\prod_{i=1}^{k}\sum_{\mathbf{p}}\frac{\log p_i}{p_i \log n_i} + O(\epsilon)$$

Recall that the sum is restricted to $\mathbf{p}$ with $n_i^{a_i} \leq p_i \leq n_i^{b_i}$, $1 \leq i \leq k$. Observe that, since $p_i = n_i/n_{i+1}$, we have $n_i \geq n^{\nu}$ where $\nu = \prod_{i=1}^{k}(1-b_i)$. It is a standard fact (see, for example, [3]) that

$$\sum_{p<k}\frac{\log p}{p} = \log k + O(1)$$

as $k \to \infty$; thus, there exists a L such that, for all $l > L$,

$$\sum_{l^{a_i}<p<l^{b_i}}\frac{\log p}{p} \geq (b_i - a_i - \epsilon)\log l.$$

Thus, for $n > L^{1/\nu}$, we have

$$\sum_{n_i^{a_i}<p<n_i^{b_i}}\frac{\log p}{p\log n_i} \geq (b_i - a_i - \epsilon).$$

Now, recall that

$$P(\mathbf{B}_k(n) \in Q) \geq (1-\epsilon)(1-2\epsilon)\prod_{i=1}^{k}\sum_{\mathbf{p}}\frac{\log p_i}{p_i \log n_i} + O(\epsilon).$$

for n sufficiently large. Performing the product starting with $i = k$ and letting i decrease—this ensures that the $\mathbf{p}$ are well-defined at each step—we obtain

$$P(\mathbf{B}_k(n) \in Q) \geq (1-\epsilon)(1-2\epsilon)\prod_{i=1}^{k}(b_i - a_i - \epsilon) + O(\epsilon),$$

and the result follows by first letting $n \to \infty$ and then $\epsilon \to 0$.

References

[1] P. Billingsley. On the distribution of large prime factors. *Period. Math. Hungar.* 2 (1972) 283–289.

[2] Peter Donnelly and Geoffrey Grimmett. On the asymptotic distribution of large prime factors. *J. London Math. Soc.* (2) 47 (1993) 395–404.

[3] G.H. Hardy and E.M. Wright. *An Introduction to the Theory of Numbers.* Clarendon Press, Oxford (1979).

[4] Gérald Tenenbaum. *Introduction to Analytic and Probabilistic Number Theory.* Cambridge University Press (1995).