HITTING THE PRIMES FOR THE k-TH TIME TAKES $k\log(k)$ DICE ROLLS (ON AVERAGE)

NOGA ALON, YAAKOV MALINOVSKY, LUCY MARTINEZ, AND DORON ZEILBERGER

Abstract. Recently, two of us (NA and YM) proved that it takes 2.43 dice rolls (on average) to reach the primes. But how long does it take (on average) until we visit them k times, for any specified k?

1. Results

Let $S = (d_1, d_2, d_3, \ldots)$ be an infinite sequence of rolls of independent fair dice. Thus the d_i are independent, identically distributed random variables, each uniformly distributed on the integers $\{1, 2, \ldots, 6\}$. For each $i \geq 1$ put $s_i = \sum_{j=1}^{i} d_j$. The sequence S *hits* a positive integer x if there exists an i so that $s_i = x$. In that case it hits x in step i.

For any positive integer k, let $L_k = L_k(S)$ be the random variable whose value is the smallest i so that the sequence S hits k primes during the first i steps (∞ if there is no such i, but it is easy to see that with probability 1 there is such i). The random variable L_1 is introduced and studied by Alon Noga and Yaakov Malinovsky [1], see also Lucy Martinez and Doron Zeilberger's work for several generalizations [4]. Shane Chern obtained the asymptotic expected hit time of the first prime for dice with a sufficiently large number of faces [3].

Here we consider the random variable L_k for larger values of k, focusing on the estimate of its expectation.

1.1. **Computational results.**

This article is accompanied by a Maple package `PRIMESk`, available from

`https://sites.math.rutgers.edu/~zeilberg/mamarim/mamarimhtml/primesk.html` ,

where there are also numerous output files.

Using our Maple package, we computed the following values of the expectation of L_k for $k \leq 30$.

k	L_k	k	L_k	k	L_k
1	2.428497914	11	48.14320555	21	106.3962997
2	5.712240468	12	53.61351459	22	112.5650207
3	9.498878119	13	59.16406655	23	118.7684092
4	13.65059271	14	64.79337350	24	125.0081994
5	18.05408931	15	70.50517127	25	131.2881683
6	22.64615402	16	76.30284161	26	137.6114097
7	27.42115902	17	82.18566213	27	143.9783110
8	32.37752852	18	88.14757626	28	150.3859881
9	37.50029903	19	94.17811256	29	156.8292462
10	42.76471868	20	100.2648068	30	163.3025173

The table suggests that the asymptotic value of this expectation is $(1+o(1))k\log k$, where the $o(1)$-term tends to zero as k tends to infinity, and the logarithm here and throughout the manuscript is in the natural basis. This is confirmed in the results stated and proved in Section 2.

The value of the standard deviation of L_k for $k \leq 30$ is given in the following table.

k	L_k	k	L_k	k	L_k
1	2.4985553	11	14.9184147	21	23.3873070
2	4.2393979	12	15.8185435	22	24.0816339
3	5.7679076	13	16.7109840	23	24.7769981
4	7.1185391	14	17.6115574	24	25.4821834
5	8.3598784	15	18.5197678	25	26.1952166
6	9.5715571	16	19.4227324	26	26.9055430
7	10.7618046	17	20.3022748	27	27.5997195
8	11.9062438	18	21.1419697	28	28.2678482
9	12.9824596	19	21.9329240	29	28.9080719
10	13.9823359	20	22.6771846	30	29.5276021

The value of the skewness of L_k for $k \leq 30$ is given in the following table.

k	L_k	k	L_k	k	L_k
1	3.3904247	11	0.7569428	21	0.5205173
2	2.1496468	12	0.7362263	22	0.5148284
3	1.6420771	13	0.7250716	23	0.5134409
4	1.3892778	14	0.7131387	24	0.5108048
5	1.2554076	15	0.6939289	25	0.5029053
6	1.1503502	16	0.6657344	26	0.4888319
7	1.0474628	17	0.6307374	27	0.4707841
8	0.9487703	18	0.5936550	28	0.4528198
9	0.8625227	19	0.5601812	29	0.4391145
10	0.7974496	20	0.5351098	30	0.4324204

The value of the kurtosis of L_k for $k \leq 30$ is given in the following table.

k	L_k	k	L_k	k	L_k
1	20.6214485	11	3.9630489	21	3.4553514
2	10.0475452	12	3.9427896	22	3.4675149
3	7.2098904	13	3.9031803	23	3.4566369
4	6.1044828	14	3.8308431	24	3.4199435
5	5.5085380	15	3.7314241	25	3.3679599
6	5.0273441	16	3.6223695	26	3.3183350
7	4.6151697	17	3.5254483	27	3.2873677
8	4.2993763	18	3.4590869	28	3.2835481
9	4.0978890	19	3.4312823	29	3.3051186
10	3.9989275	20	3.4359883	30	3.3414988

We end this section with some figures and a table of the **scaled** probability density functions for the number of rolls of a fair die until visiting the primes k times for various k values. (recall that the scaled version of a random variable X with expectation μ and variance σ^2 is $(X-\mu)/\sigma$).

k	Expectation	Standard Deviation	Skewness	Kurtosis
20	100.2648068	22.6771846	0.5351098	3.4359883
40	229.8903783	36.1271902	0.3777949	3.1278526
60	370.5241578	46.0245135	0.1406763	2.6164507
80	520.2899340	57.8152360	0.2910580	2.9707515

$k = 20$ $k = 40$

$k = 60$ $k = 80$

FIGURE 1. Scaled probability density function for the number of rolls of a fair die until visiting the primes k times.

1.2. Asymptotic results.

Theorem 1.1. *For any fixed positive reals* ε, δ *there exists* $k_0 = k_0(\varepsilon, \delta)$ *so that for all* $k > k_0$ *the probability that* $|L_k - k \log k| > \varepsilon k \log k$ *is smaller than* δ.

Theorem 1.2. *For any fixed* $\varepsilon > 0$ *and any* $k > k_0(\varepsilon)$, *the expected value of the random variable* L_k *satisfies* $|E(L_k) - k \log k| < \varepsilon k \log k$.

2. PROOFS

Lemma 2.1. *There are fixed positive* C *and* μ *so that the following holds. Let* $S = (d_1, d_2, \ldots)$ *be a random sequence as above. For any positive integer* x, *let* $p(x)$ *denote the probability that* S *hits* x. *Then* $|p(x) - 2/7| \leq C(1 - \mu)^x$, *that is, as* x *grows,* $p(x)$ *converges to the constant* $2/7$ *with an exponential rate.*

Proof. Define $p(-5) = p(-4) = p(-3) = p(-2) = p(-1) = 0$, $p(0) = 1$ and note that for every $i \geq 1$,

$$p(i) = \frac{1}{6}\sum_{j=1}^{6} p(i-j)$$

Indeed, S hits i if and only if the last number it hits before i is $i-j$ for some $j \in \{1, \ldots, 6\}$, and the die rolled after that gives the value j. The probability of this event for each specific value of j is $p(i-j) \cdot (1/6)$, providing the equation above. (Note that the definition of the initial values is consistent with this reasoning, as before any dice rolls the initial sum is 0). Thus, the sequence $(p(i))$ satisfies the homogeneous linear recurrence relation given above. The characteristic polynomial of that is

$$P(z) = z^6 - \frac{1}{6}(z^5 + z^4 + z^3 + z^2 + z + 1).$$

One of the roots of this polynomial is $z = 1$, and its multiplicity is 1 as the derivative of $P(z)$ does not vanish at 1. It is also easy to check that the absolute value of each of the other roots λ_j, $2 \leq j \leq 6$ of $P(z)$ is at most $1-\mu$ for some absolute positive constant μ. Therefore, there are constants c_j so that

$$p(i) = c_1 \cdot 1^i + \sum_{j=2}^{6} c_j \lambda_j^i,$$

implying that

$$|p(i) - c_1| \leq C(1-\mu)^i$$

for some absolute constant C. It remains to compute the value of c_1. By the last estimate, for any positive n,

$$|\sum_{i=1}^{n} p(i) - c_1 n| \leq C/(1-\mu).$$

Note that the sum $\sum_{i=1}^{n} p(i)$ is the expected number of integers in $[n] = \{1, 2, \ldots, n\}$ hit by the sequence S. For large n, this number is clearly $(1+o(1))(2/7)n$, by standard estimates for distributions of sums of independent bounded random variables, see, e.g., [2], Theorem A.1.16. Dividing by n and taking the limit as n tends to infinity shows that $c_1 = 2/7$, completing the proof. □

The next simple lemma shows that for any integers $x_1 < x_2 < \ldots < x_r$ that are far from each other, the events that the random sequence S hits x_i are nearly independent.

Lemma 2.2. *For any positive integers $x_1 < x_2 < \ldots < x_r$, the probability that the random sequence S hits all x_i is exactly*

$$p(x_1)p(x_2 - x_1)p(x_3 - x_2) \cdots p(x_r - x_{r-1}).$$

Therefore, if each difference $(x_i - x_{i-1})$ is at least s, then the probability of this event deviates from $(\frac{2}{7})^r$ by at most $(\frac{2}{7})^{r-1} Cr(1-\mu)^s$.

Proof. (Sketch) The conditional probability of the sequence to hit x_{i+1} given that it hit already $x_1, \ldots, x_i$ is exactly $p(x_{i+1} - x_i)$, since the sequence starting at x_i has the same distribution as S. The desired estimate follows from the assertion of Lemma 2.1. □

Using the two lemmas above we next show that the number of primes hit by the first f steps of the random sequence S is close to $2/7$ times the number of primes smaller than $3.5f$ with high probability.

Theorem 2.3. *Let* $\pi(x)$ *denote the number of primes smaller than* x*, and let* $Y(f)$ *denote the number of primes hit by the random sequence* S *during the first* f *steps. Then, for any fixed* $\varepsilon > 0$ *and any (large) constant* t*, and for any* $f > f_0(\varepsilon, t)$*, the probability that* $Y(f)$ *deviates from* $(2/7)\pi(3.5f)$ *by more than* $\varepsilon(2/7)\pi(3.5f)$ *is smaller than* f^{-t}*.*

Proof. (rough sketch) For large f, with probability larger than $1 - f^{-t}$ the sum $\sum_{i=1}^{f} d_i$ deviates from its expectation $3.5f$ by less than $(\varepsilon/3)3.5f$. Split all the primes smaller than $(1-\varepsilon/3)3.5f$ into, say, $\sqrt{f}$ groups of nearly equal sizes, where the difference between any two elements in the same group is at least $\sqrt{f}$. Using Lemma 2.2 it follows that for each fixed group of size g, the number of primes of the group hit by S is within an ε-fraction of its expectation. This is done by computing, say, the first $4t$ moments of this random variable, observing that these are very close to the same moments of a random variable which is the sum of g independent indicator random variables, each being 1 with probability 2/7. By considering the expectation of the $4t$-th power of the difference between this random variable and its expectation, this implies the desired concentration within each group, and the triangle-inequality supplies the required estimate for the union of all groups. The contribution of the primes between $(1-\varepsilon/3)3.5f$ and $(1+\varepsilon/3)3.5f$ is small, by the known results about the distribution of primes, and the contribution of the primes larger than $(1+\varepsilon/3)3.5f$ to the expectation is negligible, since the probability that the sequence reaches these numbers within the first f steps is tiny. This implies the assertion of the Theorem. $\square$

The assertions of Theorem 1.1 and Theorem 1.2 can be easily deduced from that of Theorem 2.3.

3. Concluding remarks and extensions

- **Extensions for biased r-sided dice and arbitrary subsets of the integers.** The proofs in the previous section use very little of the specific properties of the primes and the specific distribution of each d_i. It is easy to extend the result to any r-sided dice with an arbitrary discrete distribution on $[r]$ in which the values obtained with positive probabilities do not have any nontrivial common divisor. The constants 3.5 and 2/7 will then have to be replaced by the expectation of the random variable d_i and by its reciprocal, respectively. Similarly, we can replace the set of integers in which we count hits by an arbitrary set of positive integers, as long as its distribution satisfies some mild smoothness assumptions. We omit the details.
- **Heuristic suggestion for a more precise expression for $E(L_k)$.** We may state here a possible more precise expression for $E(L_k)$, which may be close to $k(\log k + \log\log k + c_1) + c_2$. This is justified very loosely by the heuristic argument described here together with the behavior of $\pi(n)$, and is also roughly consistent with the experimental evidence. Better to state it only as a possible guess and mainly raise the question of finding a more accurate estimate for the error term in the expectation.
- Can add the definition and conjecture that there are infinitely many AMMZ integers (after deciding if this is indeed the name we want to suggest). The known results and conjectures about prime gaps (specifically Cramér's Conjecture) and the fact that the difference between $E(L_{k+1})$ and $E(L_k)$ is close to $\log k$ (at least by our heuristics, supported by the computation), suggest that the function $2\pi(\mathrm{trunc}(\frac{7}{2} \cdot L[i]))$ will stay an even constant integer for some $(2/7)\log(L[i]) = (1+o(1))(2/7)\log i$ consecutive values of i around i for infinitely many values of i. As the denominator $7i$ will go through some $(2/7)\log i$ consecutive even values for this range, there is, possibly, a non-negligible chance that one of these will be equal to the numerator. Of course this depends on quite a few heuristic conjectures, including Cramér's. The computational results in the table we have do not handle sufficiently large i to test if this heuristics is valid. In particular, I think that in the range in the table

there are no two consecutive values in which the numerator in our formula is the same - this should happen later infinitely often.

References

[1] Noga Alon and Yaakov Malinovsky, Hitting a prime in 2.43 dice rolls (on average), The American Statistician, Volume 77 (2023), no. 3, 301-303. `https://arxiv.org/abs/2209.07698`

[2] Noga Alon and Joel H. Spencer, *The Probabilistic Method, Fourth Edition*, Wiley, 2016, xiv+375 pp.

[3] Shane Chern, Hitting a prime by rolling a die with infinitely many faces, The American Statistician, Volume 78 (2024), no. 3, 297-303. `https://arxiv.org/abs/2306.14073`

[4] Lucy Martinez and Doron Zeilberger, How many dice rolls would it take to reach your favorite kind of number?, Maple Transactions, Volume 3 (2023), no. 3, Article 15953. `https://mapletransactions.org/index.php/maple/article/view/15953`

(N. Alon) Department of Mathematics, Princeton University, Princenton, NJ 08544
E-mail address: `nalon@math.princeton.edu`

(Y. Malinovsky) Department of Mathematics and Statistics, University of Maryland, Baltimore, MD 21250
E-mail address: `yaakovm@umbc.edu`

(L. Martinez) Department of Mathematics, Rutgers University, Piscataway, NJ 08854
E-mail address: `lucy.martinez@rutgers.edu`

(D. Zeilberger) Department of Mathematics, Rutgers University, Piscataway, NJ 08854
E-mail address: `doronzeil@gmail.com`