Hitting a prime in 2.43 dice rolls (on average)

Yaakov Malinovsky

University of Maryland, Baltimore County, USA

joint work with Noga Alon

Princeton University, USA and Tel Aviv University, Israel

Rutgers Experimental Mathematics Seminar

November 17, 2022

Outline

- The Puzzle
- Numerical Results
- Dynamic-Programming Algorithm
- Bounding the Remainders

The Puzzle

Bulletin of the IMS [a]

- Let $X_1, X_2, \ldots$ be independent Uniform$\{1,2,3,4,5,6\}$ random variables, and define

$$S_n = X_1 + \ldots + X_n, \text{ for } n = 1, 2, \ldots$$

- Denote by τ the discrete time in which S_n first hits the set of prime numbers P:

$$\tau = \min\{n \geqslant 1 : S_n \in P\}.$$

- The contributing Editor:
  - $E(\tau) > 2.3479$
  - Heuristic approximation $E(\tau) \approx 7.6$.
  - It is not known whether or not $Var(\tau) < \infty$.

[a] DasGupta, A. (2017). Solution to Puzzle 17. IMS Bulletin **4**6(5), page 9.
https://imstat.org/2017/07/14/student-puzzle-corner-18-and-solution-to-puzzle-17/

Numerical Results

Monte-Carlo Simulations of the Process

number of repetitions	$mean(\tau)$	$variance(\tau)$	$\max(\tau)$
10^6	2.4316	6.2735	49
$2 * 10^6$	2.4274	6.2572	67
$3 * 10^6$	2.4305	6.2372	70
$5 * 10^6$	2.4287	6.2418	64
10^7	2.4286	6.2463	65

Dynamic-Programming Algorithm

Rigorous Computation of $E(\tau)$ and $Var(\tau)$

We proceed with a rigorous computation of $E(\tau)$ and $Var(\tau)$ up to an additive error smaller than $1/10,000$. Not surprisingly, this computation shows that the simulations supply accurate values.

Note first that

$$E\left(\tau\right) = \sum_{k\geqslant 1} P(\tau \geqslant k), \quad E\left(\tau^2\right) = \sum_{k\geqslant 1} (2k-1)P(\tau \geqslant k). \tag{1}$$

Computation

For each integer $k \geqslant 1$ and each non-prime n satisfying $k \leqslant n \leqslant 6k$, let

$$p(k,n)$$
$$= P\left(X_1 + \ldots + X_k = n \text{ and } X_1 + \ldots + X_i \notin P,\ \forall i < k\right).$$

We have,

$$P(\tau \geqslant k+1) := p(k+1) = \sum_{\{n:\ k \leqslant n \leqslant 6k\}} p(k,n)$$

Dynamic-Programming Algorithm

Fix a parameter K (in our computation we later take $K = 1000$.)

DP algorithm for computing $p(k,n)$ precisely for all $1 \leqslant k \leqslant K$ and $k \leqslant n \leqslant 6k$.

1. $p(1,1) = p(1,4) = p(1,6) = 1/6$.

2. For $k = 2, \ldots, K$ and any non-prime n between k and $6k$

$$p(k,n) = \frac{1}{6} \sum_i p(k-1, n-i),$$

   where the sum ranges over all i between 1 and 6 so that $n - i$ is non-prime.

Lower Bounds of $E(\tau)$ and $Var(\tau)$

$$E(\tau) = E_K + R_K, \quad E(\tau^2) = E_K^{(2)} + R_K^{(2)},$$

where $E_K = \sum_{k=1}^{K} p(k)$, $R_K = \sum_{k \geqslant K+1} p(k)$, $E_K^{(2)} = \sum_{k=1}^{K} (2k-1)p(k)$, $R_K^{(2)} = \sum_{k \geqslant K+1} (2k-1)p(k)$. We also have

$$Var(\tau) = Var_K + RV_K,$$

where $Var_K = E_K^{(2)} - (E_K)^2$, $RV_K = R_K^{(2)} - 2E_K R_K - (R_K)^2$.

Applying the dynamic-programming algorithm in Matlab, with an execution time of less than 5 seconds, we obtain

$$E_{1000} = 2.428497913693504, \quad Var_{1000} = 6.242778668279075.$$

Bounding the Reminders

Proposition

Proposition 1 *For every k and every non-prime n*

$$p(k,n) < \frac{1}{3}\left(\frac{5}{6}\right)^{\pi(n)}, \tag{2}$$

where $\pi(n)$ is the number of primes smaller than n.

Proof (by Induction on k) of Proposition

Proof Note that (2) holds for $k = 1$:

$$1/6 = p(1,6) < (1/3)(5/6)^3, \quad 1/6 = p(1,4) < (1/3)(5/6)^2,$$
$$1/6 = p(1,1) < (1/3)(5/6)^0.$$

Assuming the inequality holds for $k-1$ (and every relevant n) we prove it for k. Suppose there are q primes in the set $\{n-6, \ldots, n-1\}$, then $\pi(n-i) \geqslant \pi(n) - q$ for all non-prime $n-i$ in this set. Thus, by the induction hypothesis

$$p(k,n) = \frac{1}{6}\sum_i p(k-1, n-i) < \frac{1}{6}\sum_i \frac{1}{3}\left(\frac{5}{6}\right)^{\pi(n-i)}$$
$$\leqslant \frac{1}{6}(6-q)\frac{1}{3}\left(\frac{5}{6}\right)^{\pi(n)-q} \leqslant \left(\frac{5}{6}\right)^q \frac{1}{3}\left(\frac{5}{6}\right)^{\pi(n)-q} = \frac{1}{3}\left(\frac{5}{6}\right)^{\pi(n)}.$$

Corollary

By the prime number theorem [a], for every $n > 1000$

$$\pi(n) > 0.9 \frac{n}{\ln n}$$

(with room to spare). Therefore, by the above estimate we get:

Corollary 1 *For every* $k > 1000$ *and every non-prime* $n (n \geqslant k)$,

$$p(k, n) < \frac{1}{3} \left(\frac{5}{6} \right)^{0.9 \frac{n}{\ln n}} . \tag{3}$$

[a] Hardy, G. H., Wright, E. M. (2008). An introduction to the theory of numbers. Sixth edition. Revised by D. R. Heath-Brown and J. H. Silverman, with a foreword by Andrew Wiles. Oxford University Press, Oxford.

Remaining Terms

Recall that

$$P(\tau \geqslant k+1) := p(k+1) = \sum_{\{n:\ k \leqslant n \leqslant 6k\}} p(k,n)$$

$$R_{1000} \equiv \sum_{k>1000} P(\tau \geqslant k) = \sum_{k>999} P(\tau \geqslant k+1) = \sum_{k>999} \sum_{\{n:\ k \leqslant n \leqslant 6k\}} p(k,n)$$

$$< \sum_{k>999} \sum_{\{n:\ k \leqslant n \leqslant 6k\}} \frac{1}{3} \left(\frac{5}{6}\right)^{0.9n/\ln n} = \sum_{n \geqslant 1000} \sum_{k=\max(1000,n/6)}^{n} \frac{1}{3} \left(\frac{5}{6}\right)^{0.9n/\ln n}$$

$$< \sum_{n \geqslant 1000} (n-999) \frac{1}{3} \left(\frac{5}{6}\right)^{0.9n/\ln n},$$

where the first inequality is obtained from Corollary 1.

Remaining Terms (cont'd)

Define $f(n) = (n-999)\frac{1}{3}\left(\frac{5}{6}\right)^{0.9n/\ln n}$, where n is an integer $\geqslant 1000$. It is easy to check that for $n \geqslant 1000$ the function $f(n)$ has a unique maximum at $n = 1049$. (To see it, it suffices to compute $f(n)$ precisely for all $1000 \leqslant n \leqslant 1100$ and observe that for $n > 1100$ $f(n)$ is far smaller than $f(1049)$.) It is also easy to check that for any $n \geqslant 1049$ $f(n + 13\ln n)/f(n) < 1/2$. Therefore,

$$R_{1000} < \sum_{n \geqslant 1000} (n-999)\frac{1}{3}\left(\frac{5}{6}\right)^{0.9n/\ln n} = \sum_{n \geqslant 1000} f(n)$$

$$< 50 f(1049) + 2(13\ln 1049) f(1049) < 7 \cdot 10^{-8}.$$

Remaining Terms (cont'd)

Similarly

$$R_{1000}^{(2)} \equiv \sum_{k>1000} (2k-1)P(\tau \geqslant k) = \sum_{k>1000} \sum_{\{n:\ k-1 \leqslant n \leqslant 6(k-1)\}} (2k-1)p(k-1,n)$$

$$< \sum_{k>1000} \sum_{\{n:\ k-1 \leqslant n \leqslant 6(k-1)\}} (2k-1)\left(\frac{5}{6}\right)^{0.9n/\ln n}$$

$$= \sum_{n \geqslant 1000} \frac{1}{3}\left(\frac{5}{6}\right)^{0.9n/\ln n} \sum_{k=\max(1001,n/6+1)}^{n+1} (2k-1)$$

$$< \sum_{n \geqslant 1000} \frac{1}{3}\left(\frac{5}{6}\right)^{0.9n/\ln n} ((n+1)^2 - 1000^2),$$

where the first inequality is obtained from Corollary 1.

Remaining Terms (cont'd)

Denote by $g(n) = ((n+1)^2 - 1000^2)\frac{1}{3}\left(\frac{5}{6}\right)^{0.9n/\ln n}$, where n is an integer $\geqslant 1000$. For $n \geqslant 10,000$ the function $g(n)$ has an unique maximum at $n = 10,000$ and also for any $n \geqslant 10,000$ $g(n + 13\ln n)/g(n) < 1/2$. Therefore,

$$R^{(2)}_{1000} < \sum_{n=1000}^{9999} g(n) + \sum_{n \geqslant 10000} g(n)$$

$$< 8.494 \cdot 10^{-5} + 2(13 \ln 10,000) g(10,000)$$

$$< 8.5 \cdot 10^{-5} + 3.4 \cdot 10^{-68} < 1/10,000.$$

Therefore, the error RV_{1000} in the variance estimation based on the first 1000 values is below $1/10,000$.

Thank You !