On the nth Record Gap Between Primes in an Arithmetic Progression

Alexei Kourbatov
www.JavaScripter.net/math
akourbatov@gmail.com

Abstract

Let $q > r \geq 1$ be coprime integers; let $R(n,q,r)$ be the nth record gap between primes in the arithmetic progression r, $r+q$, $r+2q, \ldots$, and denote by $N_{q,r}(x)$ the number of such records below x. For $x \to \infty$, we heuristically argue that if the limit of $N_{q,r}(x)/\log x$ exists, then the limit is 2; we also conjecture that $R(n,q,r) = O_q(n^2)$. Numerical evidence supports the conjectural a.s. upper bound

$$R(n,q,r) < \varphi(q)n^2 + (n+2)q\log^2 q.$$

The median (over r) of $R(n,q,r)$ grows like a quadratic function of n; so do the mean and quartile points of $R(n,q,r)$. For fixed values of $q \gtrsim 200$ and $n \approx 10$, the distribution of $R(n,q,r)$ is skewed to the right and close to both Gumbel and lognormal distributions; however, the skewness appears to slowly decrease as n increases. The existence of a limiting distribution of $R(n,q,r)$ is an open question.

Keywords: arithmetic progression, Cramér conjecture, distribution of records, prime gap, residue class, Resnick duality theorem, Shanks conjecture.

1 Introduction

This paper is a sequel to arXiv:1610.03340 [12]. Let $q > r \geq 1$ be coprime integers, and consider the arithmetic progression

$$r, \ r+q, \ r+2q, \ r+3q, \ldots \tag{1}$$

Dirichlet [4] proved that there are infinitely many primes in such progressions. Let $R(n, q, r)$ be the size of the nth record gap between primes in the arithmetic progression (1), and denote by $N_{q,r}(x)$ the number of such records below x. The prime number theorem for arithmetic progressions [6] guarantees that, for any fixed coprime pair (q, r), the sequence of records $R(n, q, r)$ is infinite. Our present goal is to investigate the behavior of $R(n, q, r)$ and $N_{q,r}(x)$ statistically.

In [12] we already studied maximal gaps between primes *below* x in progression (1). We empirically found that, for a fixed q, the histogram of appropriately rescaled maximal gaps between primes $r + kq \leq x$ is very close to the Gumbel extreme value distribution. The nature of empirical results in [12] is akin to probabilistic results for *sample maxima*.

For i.i.d. random variables, Resnick's theorems [18, 19] establish that the limit law of sample maxima *cannot be the same* as the limit law of the nth record. In particular, if the limit law of sample maxima is the Gumbel distribution, then the *normal* distribution is a possible limit law for the nth record. Three generally possible limiting distributions of the nth records in sequences of i.i.d. random variables are the normal, lognormal, and negative lognormal distributions [1, 19]. However, beyond i.i.d. settings, one can also encounter situations where the Gumbel distribution itself may be the limit law for records [1, p. 193].

Probabilistic results [1, 13, 18, 19] are not directly applicable to record prime gaps. Nevertheless, it is natural to look at the growth and distribution of the record gaps between primes in progressions (1) — and investigate whether any results similar to Resnick's duality theorem might also be true for record gaps $R(n, q, r)$. Does the number of records $N_{q,r}(x)$ behave like it would in an i.i.d. sequence, i.e., does $N_{q,r}(x)$ grow about as fast as $\log x$? What is the order of magnitude of the nth record gap, as a function of n? What are statistical properties of the nth record gap? Is the actual distribution of $R(n, q, r)$ approximately normal or lognormal or Gumbel — or none of the above? We will attempt to answer these questions using heuristics and statistical analysis of numerical results. Still, we have to remember that *prime numbers* are neither random nor independent [7, 17]; likewise, *prime gaps* are neither random nor independent. So any statistical observations and heuristic reasoning about prime gaps should be used with a lot of caution.

2 Heuristic predictions

2.1 The nth record gap between primes

Denote by $G(x)$ the maximal gap between primes below x. Let $R(n)$ be the nth record prime gap; $R(n) =$ A005250(n) in the *Online Encyclopedia of Integer Sequences* (OEIS) [22]. Suppose that x is so large that there have been *many record gaps* between primes below x. Cramér [3] used probabilistic reasoning to conjecture that

$$G(x) = O(\log^2 x). \tag{2}$$

Moreover, Shanks [21] heuristically found

$$G(x) \sim \log^2 x \qquad \text{as } x \to \infty \qquad [21, \text{p.} 648]. \tag{3}$$

Let $\tau = \tau(x)$ be a smooth function estimating the number of record prime gaps with endpoints in $[x, ex]$. We postulate the existence of such a function τ and, in accordance with our earlier observations [12, section 3.4], assume that

(A) $\tau \geq 1$ as $x \to \infty$ (that is, prime gap records occur more often than records in a sequence of i.i.d. random variables, for which we would have $\lim\limits_{x\to\infty} (\log \text{li}(ex) - \log \text{li}\, x) = 1$);

(B) $\tau = o(\log x)$ as $x \to \infty$ (this means that only a zero proportion of positive integers are terms of sequence A005250, i.e. values of the $R(n)$ function);

(C) τ is a non-decreasing, slowly varying function[1] of x.

Assume further that the actual number of records in $[x, ex]$ does not differ much from τ. Then there are about $\bar{\tau} \log x$ records below x, where $\bar{\tau}$ is the average value of τ on the interval $[1, x]$. Denoting by n the number of records up to x, we have

$$\frac{n}{\bar{\tau}} \sim \log x. \tag{4}$$

This, together with the Cramér and Shanks conjectures, implies that

$$R(n) = G(x) \sim \log^2 x \sim \frac{n^2}{\bar{\tau}^2} \leq n^2 \quad \text{ as } x \to \infty. \tag{5}$$

Granville's correction [7, p. 24] to the Cramér and Shanks conjectures might imply an additional numerical constant in the above estimate; still we have $R(n) = O(n^2)$.

Reality check. Computations of Oliveira e Silva, Herzog and Pardi [15, 16] established the actual size of the nth record prime gap $R(n) = G(x)$, for $x \leq 4 \cdot 10^{18}$ and $n \leq 75$. The actual prime gaps indeed turn out to satisfy

$$\begin{aligned} G(x) &\lesssim \log^2 x && \text{for } x \leq 4 \cdot 10^{18}, && (6)\\ R(n) &\leq n^2 && \text{for } n \leq 75, && (7)\\ R(n) &\approx 0.25n^2 + 0.5n && \text{for } n \leq 75, && (8) \end{aligned}$$

which suggests that in the available data range we can take $1/\bar{\tau}^2 \approx 0.25$ and $\bar{\tau} \approx 2$.

[1] Here we do not assume that τ tends to a finite limit as $x \to \infty$; but see sect. 2.3.

2.2 The nth record gap $R(n, q, r)$

Now consider the general case: gaps between primes in the arithmetic progression (1). Suppose that x is so large that we have already observed *many record gaps* between primes $\le x$ in progression (1). Let $G_{q,r}(x)$ be the maximal gap between primes $r + kq \le x$, $k \in \mathbb{N}^0$.

Instead of the Cramér and Shanks conjectures, we will now need the following more general statements [12, sections 5.2, 5.3]:

Generalized Cramér conjecture. Almost all maximal gaps $G_{q,r}(x)$ satisfy

$$G_{q,r}(x) \ < \ \varphi(q) \log^2 x \tag{9}$$

for any coprime $q > r \ge 1$. Here $\varphi(q)$ is *Euler's totient function*.

Generalized Shanks conjecture. Almost all maximal gaps $G_{q,r}(x)$ satisfy

$$G_{q,r}(x) \ \sim \ \varphi(q) \log^2 x \qquad \text{as } x \to \infty. \tag{10}$$

The heuristic reasoning then proceeds similar to the previous subsection. Let τ be a smooth function estimating the number of record gaps between primes $p = r + kq$ with end-of-gap primes $p \in [x, ex]$. As before, for any fixed q, let $\tau = \tau(q, x)$ obey the heuristic assumptions (A), (B), (C) of sect. 2.1. There are about $\bar{\tau} \log x$ record gaps between primes $r + kq \le x$; denoting by n the "typical" number of records up to x we have

$$n \sim \bar{\tau} \log x \qquad \text{as } x \to \infty. \tag{11}$$

Using eqs. (9), (10), (11) we estimate the "typical" nth record gap:

$$R(n, q, r) \ = \ G_{q,r}(x) \ \lesssim \ \varphi(q) \log^2 x \ \sim \ \varphi(q) \frac{n^2}{\bar{\tau}^2}. \tag{12}$$

The above is valid for *large* n and x. To make estimate (12) applicable to moderate n, we add a semi-empirical correction term of size $O_q(n)$ (motivated in part by heuristics of [14]):

$$R(n, q, r) \ \lesssim \ \varphi(q) \frac{n^2}{\bar{\tau}^2} + q(\log q)^2 (n + 2). \tag{13}$$

Roughly speaking, the correction term takes into account that in progression (1) the very first prime $p = r + kq$ might occur unusually late[2] and then subsequent primes occur less frequently than usual.

But we do not have any precise knowledge of τ. Therefore, let us eliminate τ using assumption (A) $\tau \ge 1$; we thus heuristically arrive at the a.s. upper bound

$$R(n, q, r) \ < \ \varphi(q) n^2 + q(\log q)^2 (n + 2). \tag{14}$$

For large x, computations of [12, section 3.4] suggest that τ is strictly greater than one; with this in mind, we expect at most finitely many exceptions to inequality (14) for any fixed q. In section 3 we will compare this heuristic prediction with results of computations.

[2] Cf. [14, section 2]; note the heuristic formula for $\limsup P(q)$ describing the behavior of the first prime in progression (1).

2.3 The limit of $N_{q,r}(x)/\log x$

As before, let $\tau = \tau(q,x)$ be an estimator for $N_{q,r}(ex) - N_{q,r}(x)$, the number of record gaps between primes $p = r + kq$, with $p \in [x, ex]$. As a function of x, let τ obey the heuristic assumptions (A), (B), (C) of sect. 2.1. Below we heuristically argue[3] that if the limit $\lim_{x\to\infty} N_{q,r}(x)/\log x = \lim_{x\to\infty} \tau$ exists, then the limit is 2. This will justify the estimate for $N_{q,r}(ex) - N_{q,r}(x)$ given in [12]; see eq. (18).

Suppose that the following limits exist and are equal to some positive number τ_*:

$$\lim_{x\to\infty} \frac{N_{q,r}(x)}{\log x} = \lim_{x\to\infty} \frac{\mathrm{mean}_r\, N_{q,r}(x)}{\log x} = \lim_{x\to\infty} \tau(q,x) = \tau_* > 0.$$

Let n be a "typical" number of records up to x. For large x, eq. (11) gives

$$n \sim \tau_* \log x. \tag{15}$$

Define $\Delta R(n,q,r) = R(n+1,q,r) - R(n,q,r)$. By formula (12), for $x \to \infty$ we have

$$\begin{aligned}
\operatorname*{mean}_r R(n,q,r) &\sim \varphi(q)\frac{n^2}{\tau_*^2}, \\
\operatorname*{mean}_r \Delta R(n,q,r) &= \operatorname*{mean}_r \big(R(n+1,q,r) - R(n,q,r)\big) \\
&\sim \frac{\varphi(q)}{\tau_*^2}((n+1)^2 - n^2) \sim \frac{2n\varphi(q)}{\tau_*^2}.
\end{aligned}$$

Combining this with (15) we find

$$\operatorname*{mean}_r \Delta R(n,q,r) \sim \frac{2}{\tau_*}\varphi(q)\log x. \tag{16}$$

On the other hand, heuristically we expect that, on average, two consecutive record gaps should differ by the "local" average gap between primes in progression (1):

$$\operatorname*{mean}_r \Delta R(n,q,r) \sim \varphi(q)\log x \quad \text{(average gap near } x\text{)}. \tag{17}$$

Together, eqs. (16) and (17) imply that

$$\tau_* = 2.$$

Remark. From numerical experiments [12] we already know the empirical estimate

$$\tau(q,x) \approx \operatorname*{mean}_r(N_{q,r}(ex) - N_{q,r}(x)) \approx 2 - \frac{\kappa(q)}{\log x - \delta(q)} \qquad \text{[12, sect. 3.4]}, \tag{18}$$

which agrees with the above heuristic prediction $\lim\limits_{x\to\infty} \tau = \tau_* = 2$.

[3] A similar argument for prime k-tuples leads to the number of records $N_k(x) \sim (k+1)\log x$.

3 Numerical results

Using a modified version of PARI/GP code from [12] we have computed the first fourteen record gaps $R(n,q,r)$ for all $q \le 2000$. Twenty or more records were computed for selected small values of q (see e. g. Fig. 1). Records $R(n,q,r)$ were also computed for selected larger values of q up to 80000. We used all admissible values of $r \in [1, q-1]$, $\gcd(q,r) = 1$, to assemble a complete data set of record gaps for given q and n. (As an example, Table 1 gives the $R(n,q,r)$ data set for $n = 10$, $q = 50$.) For each data set, we computed its largest and smallest values, mean, median, standard deviation, skewness, and quartile points. This section summarizes our numerical results.

Table 1. The 10th record gap between primes $r + kq$, $q = 50$

r	Record gap $R(10, 50, r)$	Start of gap	End of gap
1	1150	158551	159701
3	1950	504953	506903
7	1950	959207	961157
9	1950	1229359	1231309
11	1150	56911	58061
13	1600	211663	213263
17	1400	404267	405667
19	1950	794669	796619
21	2300	6534071	6536371
23	1350	266023	267373
27	2100	1286777	1288877
29	1150	145879	147029
31	1150	289381	290531
33	3000	8314433	8317433
37	1950	1336637	1338587
39	1650	706039	707689
41	1650	1061591	1063241
43	1400	668543	669943
47	750	39847	40597
49	1150	241249	242399

3.1 Conjectural (a.s.) upper bound for $R(n,q,r)$

All record gaps that we have computed turn out to satisfy the heuristic inequality (14):

$$R(n,q,r) \ < \ \varphi(q)n^2 + q(\log q)^2(n+2) \quad \text{for all coprime } r < q \le 2000,\ n < 15.$$

While we expect a finite number of exceptions at least for some values of q, thus far we have not seen any at all. However, if we use a smaller correction term $\varphi(q)(\log q)^2(n+2)$, then there are a couple of exceptions, e. g. for $q = 20$ and $q = 23$.

Table 2. Median nth record gap between primes $r + kq$, $q = 11, 17, 50$

n	median $R(n, 11, r)$	median $R(n, 17, r)$	median $R(n, 50, r)$
1	33	68	75
2	66	136	175
3	110	221	275
4	176	306	450
5	231	374	675
6	275	493	775
7	319	612	950
8	407	680	1100
9	539	850	1350
10	616	1071	1625
11	748	1139	1850
12	825	1309	2025
13	935	1513	2300
14	1177	1700	2550
15	1232	1870	2725
16	1342	2057	3125
17	1540	2227	3250
18	1639	2448	3750
19	1958	2822	4375
20	2046	3281	4525

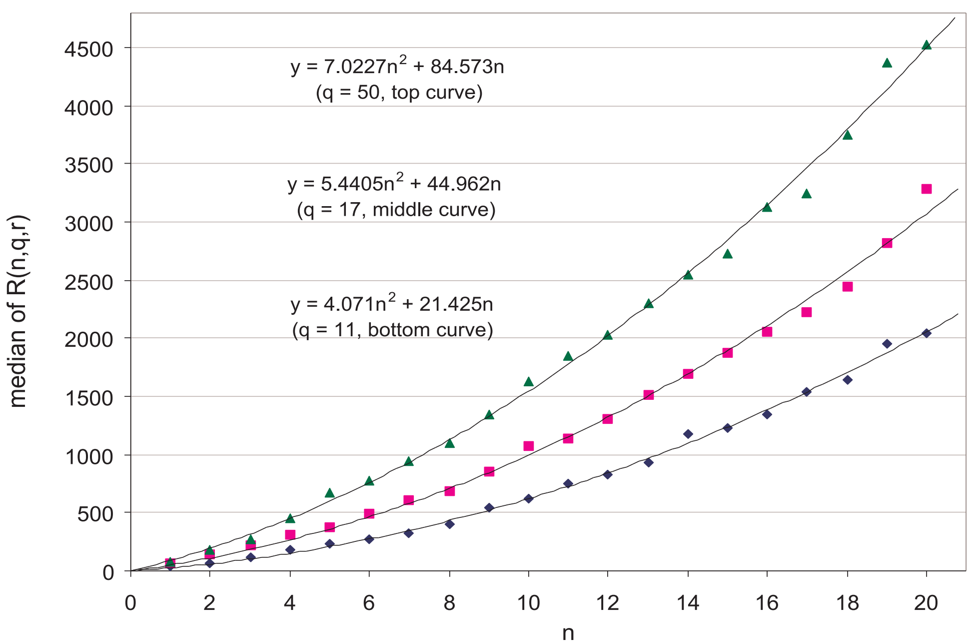

Figure 1: Median nth record gap between primes $r + kq$, $q = 11, 17, 50$. Smooth curves are quadratic approximations to median $R(n, q, r)$, eq. (19).

3.2 The growth trend of $R(n, q, r)$

For a fixed pair (q, r), the sequence $R(n, q, r)$ is a strictly increasing function of n; as n increases, the records $R(n, q, r)$ seem to grow somewhat erratically. But consider the *median* of $R(n, q, r)$ over all admissible r with $\gcd(q, r) = 1$. Table 2 and Figure 1 show that the growth of this median is described quite accurately by a quadratic function of n:

$$\operatorname*{median}_{\substack{r \in [1,q] \\ q,r \text{ coprime}}} R(n, q, r) \;\approx\; A_q n^2 + B_q n. \tag{19}$$

Rough empirical estimates for the coefficients A_q and B_q in (19) are

$$A_q \;\approx\; 0.3\varphi(q), \tag{20}$$
$$B_q \;<\; \varphi(q)(\log q)^2. \tag{21}$$

Quadratic approximations similar to (19) also work quite well for the mean value, least value and quartile points of $R(n, q, r)$, as shown in Figure 2. For the largest value of the nth record, a three-term quadratic approximation is suitable (Fig. 2, top curve):

$$\max_{\substack{r \in [1,q] \\ q,r \text{ coprime}}} R(n, q, r) \;\approx\; \alpha_q n^2 + \beta_q n + \gamma_q. \tag{22}$$

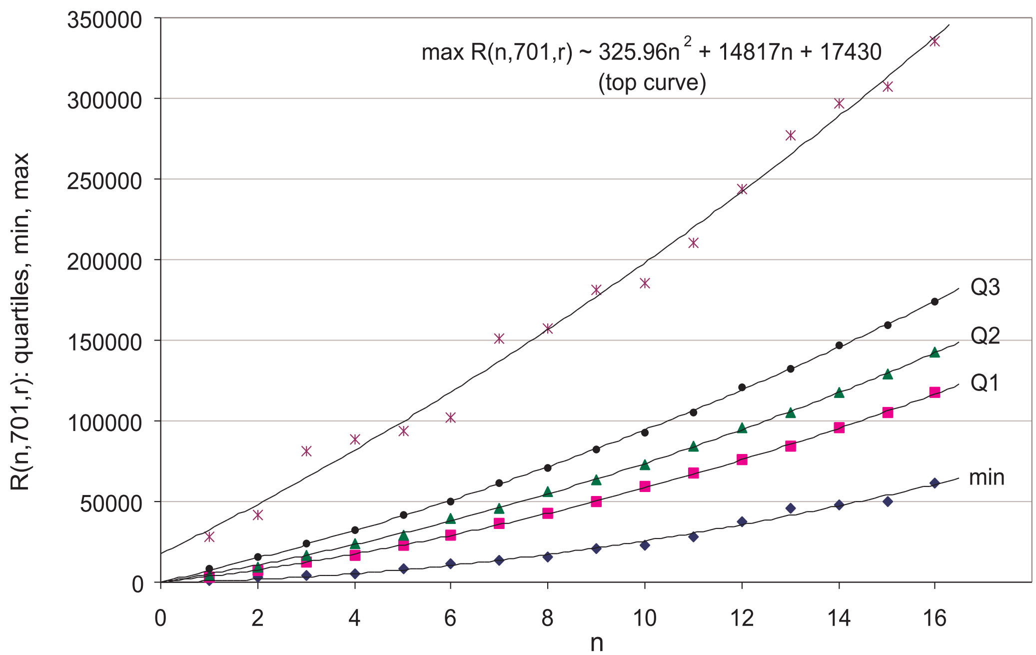

Figure 2: Quartile points, smallest and largest values of $R(n, q, r)$ for $q = 701$. $Q1$: lower quartile, $Q2$: median, $Q3$: upper quartile. Smooth curves are quadratic approximations.

Remark. The quadratic approximation (19) appears to remain valid for large n, with the leading coefficient A_q stabilizing near some positive constant value; this suggests that the function τ of section 2 tends to a finite limit as $x \to \infty$. On the other hand, if the coefficient A_q in (19) were to decrease to zero when we attempt to approximate $R(n, q, r)$ for larger and larger n, this would mean that τ increases without bound. (In the special case of *record prime gaps* $R(n)$, the quadratic approximation (8) remains valid for a wide range of n, at least up to $n = 75$, which suggests that $\lim_{x\to\infty} \tau$ does exist; and the limit might be about 2.)

3.3 The distribution of $R(n, q, r)$

In the previous section we have seen that, for a fixed q, the median value of $R(n, q, r)$ grows like a quadratic function of n. Now let us look at the distribution of the $R(n, q, r)$ values around the median. Figure 3 shows the histograms of $R(n, q, r)$ computed for $q = 9001$, $n = 6, 8, 10, 12$. The histograms are clearly skewed to the right. We see that for moderate values of n both the Gumbel and lognormal distributions are good approximations for the $R(n, q, r)$ histograms.

However, note that the actual $R(n, q, r)$ data sets appear to have slowly decreasing skewness as n increases (see Fig. 4), whereas the Gumbel distribution has constant skewness independent of the distribution's scale and mode:

$$\text{Gumbel distribution skewness} \;=\; \frac{12\sqrt{6}\,\zeta(3)}{\pi^3} \;=\; 1.139547\ldots$$

In this respect, the lognormal distribution is a better fit to the data. Indeed, the best-fit lognormal distributions do reflect the decreasing skewness observed in the data. The existence of a limiting distribution of $R(n, q, r)$ is an open question.

Remarks.

(i) For smaller q, the skewness of $R(n, q, r)$ data exhibits a lot of fluctuations. Such fluctuations may mask the general trend of decreasing skewness; nevertheless, this trend becomes apparent for larger q.

(ii) The decrease in skewness of $R(n, q, r)$ is satisfactorily described by a power law (Fig. 4). If the skewness continues to decrease all the way to zero, then it is possible that the *normal distribution* turns out to be the limit law for $R(n, q, r)$ as $n \to \infty$. (A sequence of lognormal distributions with vanishing skewness becomes indistinguishable from the normal distribution.) So a certain analog of the "Gumbel/normal case" of Resnick's theorems [18, 19] might be valid for the $R(n, q, r)$ limit law (*if one exists*); however, convergence of records to the hypothetical normal limit law is exceedingly slow. For example, at the rate shown in Fig. 4, we would need to observe over ten thousand consecutive records in order for the skewness to go down to 0.1. Anyway, for moderate values of n practically attainable in computation, the $R(n, q, r)$ histograms are quite far from the normal distribution suggested by the "Gumbel/normal case" of Resnick's theorems.

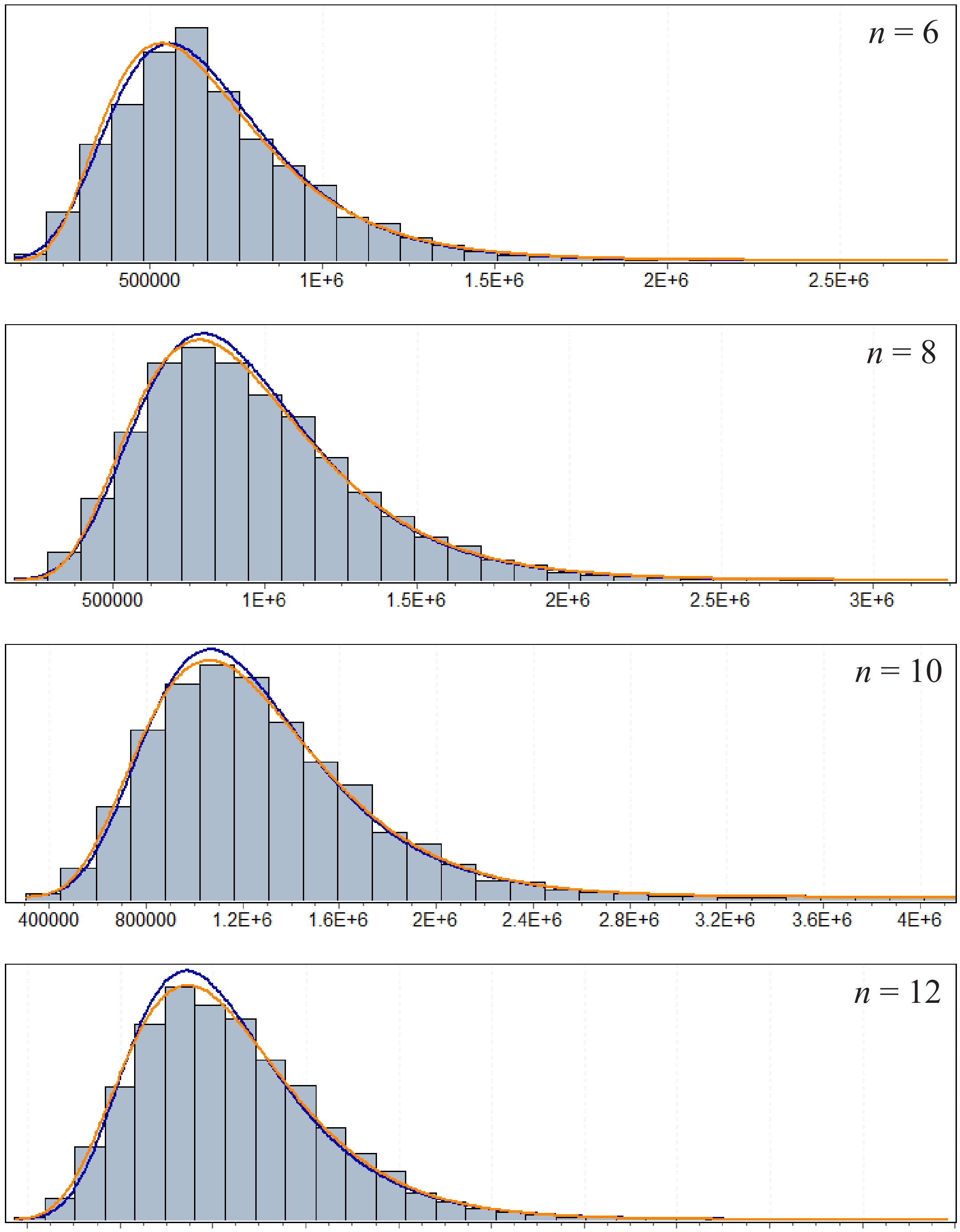

Figure 3: Histograms of the nth record gap $R(n, q, r)$ between primes in progression (1) for $q = 9001$, $n = 6, 8, 10, 12$. Orange curve: best-fit lognormal pdf; dark blue curve: best-fit Gumbel pdf.

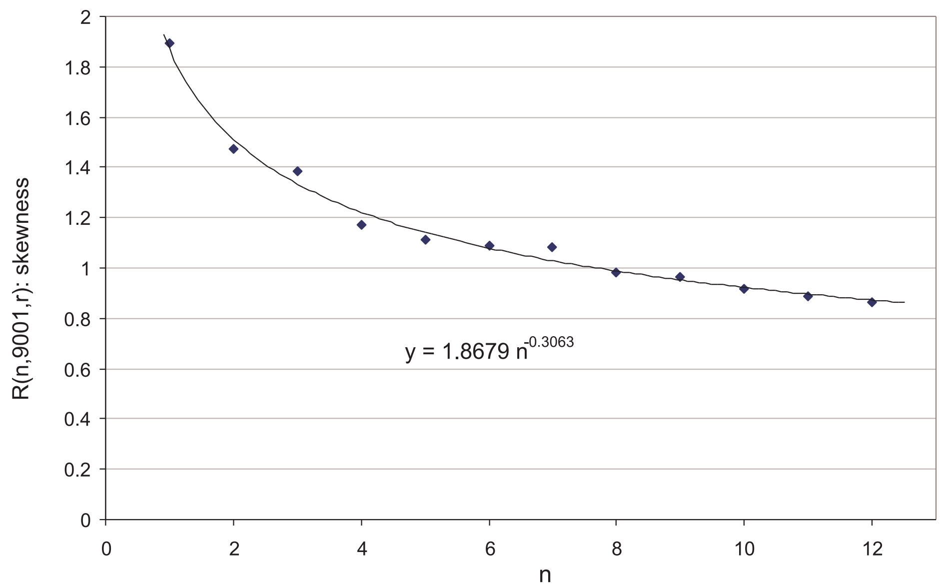

Figure 4: Skewness of the $R(n, 9001, r)$ data sets for $n = 1, 2, 3, \ldots, 12$.

4 Acknowledgments

I am grateful to all contributors and editors of the websites *OEIS.org* and *PrimePuzzles.net*.

References

[1] B. C. Arnold, N. Balakrishnan, H. N. Nagaraja, *Records*, Wiley, 1998. `http://dx.doi.org/10.1002/9781118150412`

[2] R. C. Baker, G. Harman, J. Pintz, The difference between consecutive primes, II. *Proceedings of the London Mathematical Society.* **83** (3) (2001), 532–562.

[3] H. Cramér, On the order of magnitude of the difference between consecutive prime numbers, *Acta Arith.* **2** (1936), 23–46.

[4] P. G. L. Dirichlet, Beweis des Satzes, dass jede unbegrenzte arithmetische Progression, deren erstes Glied und Differenz ganze Zahlen ohne gemeinschaftlichen Factor sind, unendlich viele Primzahlen enthält, *Abhandlungen der Königlichen Preußischen Akademie der Wissenschaften zu Berlin.* **48** (1837), 45–71. English translation: arXiv:0808.1408

[5] K. Ford, B. Green, S. Konyagin, J. Maynard, T. Tao, Long gaps between primes, arXiv preprint arXiv:1412.5029 (2014). `http://arxiv.org/abs/1412.5029`

[6] J. B. Friedlander and D. A. Goldston, Variance of distribution of primes in residue classes, *Quart. J. Math. Oxford* **47** (1996), 313–336.

[7] A. Granville, Harald Cramér and the distribution of prime numbers, *Scandinavian Actuarial Journal* **1** (1995), 12–28.

[8] E. J. Gumbel, *Statistics of Extremes*, Columbia University Press, 1958. Dover, 2004.

[9] A. Kourbatov, Maximal gaps between prime k-tuples: a statistical approach, *Journal of Integer Sequences* **16** (2013), Article 13.5.2. http://arxiv.org/abs/1301.2242

[10] A. Kourbatov, Is there a limiting distribution of maximal gaps between primes? Poster presentation at *DIMACS Conference on Challenges of Identifying Integer Sequences*, Oct. 9-10, 2014. http://www.javascripter.net/math/publications/AKourbatovOEIS50poster.pdf

[11] A. Kourbatov, The distribution of maximal prime gaps in Cramér's probabilistic model of primes, *Int. Journal of Statistics and Probability* **3** (2) (2014), 18–29. arXiv:1401.6959

[12] A. Kourbatov, On the distribution of maximal gaps between primes in residue classes, arXiv preprint. http://arxiv.org/abs/1610.03340 (2016)

[13] J. Krug, Records in a changing world, *J. Stat. Mech.: Theory and Experiment* P07001 (2007). http://iopscience.iop.org/article/10.1088/1742-5468/2007/07/P07001

[14] J. Li, K. Pratt, G. Shakan, A lower bound for the least prime in an arithmetic progression, arXiv preprint. http://arxiv.org/abs/1607.02543 (2016).

[15] T. R. Nicely, First occurrence prime gaps, preprint, 2014. Available at http://www.trnicely.net/gaps/gaplist.html.

[16] T. Oliveira e Silva, S. Herzog, and S. Pardi, Empirical verification of the even Goldbach conjecture and computation of prime gaps up to $4 \cdot 10^{18}$, *Math. Comp.* **83** (2014), 2033-2060. http://www.ams.org/journals/mcom/2014-83-288/S0025-5718-2013-02787-1/

[17] J. Pintz, Cramér vs Cramér: On Cramér's probabilistic model of primes. *Functiones et Approximatio*, **37** (2) (2007), 361–376.

[18] S. I. Resnick, Record values and maxima, *The Annals of Probability*, **1** (1973), 650–662. http://dx.doi.org/10.1214/aop/1176996892

[19] S. I. Resnick, Limit laws for record values, *Stochastic Processes and Their Applications*, **1** (1973), 67–82. http://doi.org/10.1016/0304-4149(73)90033-1

[20] C. Rivera (ed.), Conjecture 77. Gaps between primes of the form $p = qn + r$, 2016. Available at http://www.primepuzzles.net/conjectures/conj_077.htm.

[21] D. Shanks, On maximal gaps between successive primes. *Math. Comp.* **18** (1964), 646–651.

[22] N. J. A. Sloane (ed.), *The On-Line Encyclopedia of Integer Sequences*, 2015. Published electronically at http://oeis.org/.

[23] M. Wolf, Some heuristics on the gaps between consecutive primes, arXiv preprint. http://arxiv.org/abs/1102.0481 (2011)

[24] M. Wolf, Nearest neighbor spacing distribution of prime numbers and quantum chaos, *Phys. Rev. E* **89**, 022922 (2014). `http://arxiv.org/abs/1212.3841`

2010 *Mathematics Subject Classification*: 11N05.
Keywords: arithmetic progression, Cramér conjecture, distribution of records, prime gap, residue class, Resnick duality theorem, Shanks conjecture.

(Concerned with sequences A005250, A084162, A235402, A235492, A268799, A268925, A268928, A268984, A269234, A269238, A269261, A269420, A269424, A269513, A269519.)
