17. BIASES IN THE DISTRIBUTION OF CONSECUTIVE PRIMES

ROBERT J. LEMKE OLIVER AND FRANK THORNE

We now look at a very different sort of bias to Chebyshev's. Let (a_1, a_2) denote a pair of reduced residue classes modulo q (not necessarily distinct), and define

$$\pi(x; q, (a_1, a_2)) := \#\{p_n \leq x \ : \ p_n \equiv a_1, \ p_{n+1} \equiv a_2 \pmod q\}.$$

If $\mathbf{a}$ is an r-tuple of reduced residue classes, we define $\pi(x; q, \mathbf{a})$ similarly.

Naturally we expect, for each r-tuple $\mathbf{a}$, that

$$\pi(x; q, \mathbf{a}) \sim \frac{1}{\phi(q)^r} \mathrm{Li}(x). \tag{17.1}$$

Here is some numerical data, where x is chosen with $\pi(x) = 10^8$.

a	b	$\pi(x_0; 10, (a, b))$	a	b	$\pi(x_0; 10, (a, b))$
1	1	4,623,042	7	1	6,373,981
	3	7,429,438		3	6,755,195
	7	7,504,612		7	4,439,355
	9	5,442,345		9	7,431,870
3	1	6,010,982	9	1	7,991,431
	3	4,442,562		3	6,372,941
	7	7,043,695		7	6,012,739
	9	7,502,896		9	4,622,916

That is *not* what we were expecting. As good analytic number theorists, we demand an explanation. Although we don't have a proof, we do have a solid conjecture, thanks to the joint work of the first author and Soundararajan (henceforth 'LO-S'):

Conjecture 17.1. *We have*

$$\pi(x; q, \mathbf{a}) = \frac{\mathrm{Li}(x)}{\phi(q)} \left(1 + c_1(q; \mathbf{a}) \frac{\log \log x}{\log x} + c_2(q; \mathbf{a}) \frac{1}{(\log x)} + O\left(\frac{1}{(\log x)^{7/4}}\right)\right),$$

for explicit constants $c_1(q; \mathbf{a})$ and $c_2(q; \mathbf{a})$.

The constants are somewhat complicated. We have

$$c_1(q; \mathbf{a}) := \frac{\phi(q)}{2} \left(\frac{r-1}{\phi(q)} - \#\{1 \leq i < r \ : a_i \equiv a_{i+1} \pmod q\}\right),$$

and $c_2(q; \mathbf{a})$ is more complicated still. The constant simplifies in some cases however; in particular, if $\mathbf{a} = (a, a)$ we have

$$c_2(q; (a, a)) = \frac{\phi(q) \log(q/2\pi) + \log(2\pi)}{2} - \frac{\phi(q)}{2} \sum_{p | q} \frac{\log p}{p - 1}.$$

Date: December 11, 2019.

In particular, (17.1) *is* (expected to be) correct, but there are large secondary terms in the data, of size only logarithmically smaller than the main term. It is this phenomenon which we want to explain.

17.1. **Review and extension of the Hardy-Littlewood conjectures.** We recall the Hardy-Littlewood conjectures, in a somewhat different formulation used by LO-S. These conjectures are very unproven, especially in the strong quantitative form in which we cite them – which means we cannot hope for a proof of Conjecture 17.1 just yet.

Let $\mathcal{H} = \{h_1, \dots, h_k\}$ be a k-tuple of integers, and let $\mathbf{1}_{\mathcal{P}}$ denote the characteristic function of the primes. Then a strong form of the Hardy-Littlewood conjecture asserts that

$$\sum_{n \leq x} \prod_{h \in \mathcal{H}} \mathbf{1}_{\mathcal{P}}(n+h) = \mathfrak{S}(\mathcal{H}) \int_2^x \frac{dt}{(\log t)^k} + O(x^{1/2+\epsilon}),$$

with

$$\mathfrak{S}(\mathcal{H}) = \prod_p \left(1 - \frac{\#(\mathcal{H} \,(\mathrm{mod}\, p))}{p}\right) \left(1 - \frac{1}{p}\right)^{-|\mathcal{H}|}.$$

Note the form of the integral above; we have that

$$\frac{x}{(\log x)^k} \sim \int_2^x \frac{dt}{(\log t)^k},$$

but not up to an error of $O(x^{1/2+\epsilon})$! We recall also Gallagher's result from Section [???], to the effect that

$$\sum_{\substack{\mathcal{H} \subseteq [1,h] \\ |\mathcal{H}|=k}} \mathfrak{S}(\mathcal{H}) \sim \binom{h}{k} \sim \frac{h^k}{k!}, \tag{17.2}$$

i.e. that the singular series is 1 on average.

We will require some slight modifications of all of the above. We first introduce (following Montgomery and Soundararajan) a modified singular series

$$\mathfrak{S}_0(\mathcal{H}) := \sum_{\mathcal{T} \subseteq \mathcal{H}} (-1)^{|\mathcal{H} \setminus \mathcal{T}|} \mathfrak{S}(\mathcal{T}), \quad \text{so that} \quad \mathfrak{S}(\mathcal{H}) := \sum_{\mathcal{T} \subset \mathcal{H}} \mathfrak{S}_0(\mathcal{T}). \tag{17.3}$$

(Note that our definitions impose that $\mathfrak{S}(\emptyset) = \mathfrak{S}_0(\emptyset) = 1$.) This modified singular series arises naturally in the following version of the Hardy-Littlewood conjecture:

$$\sum_{n \leq x} \prod_{h \in \mathcal{H}} \left(\mathbf{1}_{\mathcal{P}}(n+h) - \frac{1}{\log n}\right) = \mathfrak{S}(\mathcal{H}) \int_2^x \frac{dt}{(\log t)^k} + O(x^{1/2+\epsilon}). \tag{17.4}$$

We have the following result of Montgomery and Soundararajan:

Proposition 17.2. *We have*

$$\sum_{\substack{\mathcal{H} \subseteq [1,h] \\ |\mathcal{H}|=k}} \mathfrak{S}_0(\mathcal{H}) = \frac{\mu_k}{k!} \left(-h \log h + Ah\right)^{k/2} + O_k(h^{k/2 - 1/(7k) + \epsilon}), \tag{17.5}$$

where μ_k is a constant depending on k (in particular, the kth moment of the standard Gaussian, with $\mu_k = 0$ if k is odd), and A is a constant independent of k.

Although it might not be initially obvious, this refines Gallagher's asymptotic (17.2). By definition, we have

$$\begin{aligned}\sum_{\substack{\mathcal{H}\subseteq[1,h]\\|\mathcal{H}|=k}}\mathfrak{S}_0(\mathcal{H}) &= \sum_{|\mathcal{H}|=k}\sum_{\mathcal{T}\subseteq\mathcal{H}}(-1)^{|\mathcal{H}|\setminus|\mathcal{T}|}\mathfrak{S}(\mathcal{T})\\ &= (-1)^k\sum_{|\mathcal{H}|=k}\sum_{\mathcal{T}\subseteq\mathcal{H}}(-1)^{|\mathcal{T}|}\mathfrak{S}(\mathcal{T}).\end{aligned}$$

Now, for each $\mathcal{T}$ with $|\mathcal{T}|=j$, there are $\binom{h-j}{k-j}$ choices of $|\mathcal{H}|$ of size k which contain $\mathcal{T}$, so that

$$\sum_{\substack{\mathcal{H}\subseteq[1,h]\\|\mathcal{H}|=k}}\mathfrak{S}_0(\mathcal{H}) = (-1)^k\sum_{j=1}^{k}(-1)^j\binom{h-j}{k-j}\sum_{\substack{\mathcal{T}\subseteq[1,h]\\|\mathcal{T}|=j}}\mathfrak{S}(\mathcal{T}).$$

Now, by applying Gallagher's result (17.2) to the inner sum, and then cleaning up the ensuing sums of binomial coefficients, one may *formally* conclude an asymptotic formula for (17.5). Note, however, one needs a more sophisticated treatment of the error term to obtain an asymptotic in (17.5), let alone the power saving error term. Nevertheless, this should give the reader a sense of where Proposition 17.2 comes from.

We now present a modification of the Hardy-Littlewood conjectures, taking into account congruence conditions modulo q. For any integer $q\geq 1$, define

$$\mathfrak{S}_q(\mathcal{H}) = \prod_{p\nmid q}\left(1-\frac{\#(\mathcal{H}\,(\mathrm{mod}\,p))}{p}\right)\left(1-\frac{1}{p}\right)^{-|\mathcal{H}|},$$

and also define $\mathfrak{S}_{q,0}(\mathcal{H})$ analogously to (17.3). For each arithmetic progression $a\,(\mathrm{mod}\,q)$, for which $a+h$ is coprime to q for each $h\in\mathcal{H}$, we expect that

$$\sum_{\substack{n<x\\ n\equiv a(\mathrm{mod}\,q)}}\prod_{h\in\mathcal{H}}\mathbf{1}_{\mathcal{P}}(n+h)\sim\mathfrak{S}_q(\mathcal{H})\left(\frac{q}{\phi(q)}\right)^{|\mathcal{H}|}\frac{1}{q}\int_2^x\frac{dt}{(\log t)^{|\mathcal{H}|}}.$$

This is the analogue of the Hardy-Littlewood prime tuple conjecture for arithmetic progressions. The factor of $1/q$ is there because n is restricted to a single residue class modulo q; the factor of $\left(\frac{q}{\phi(q)}\right)^{|\mathcal{H}|}$ arises because the condition on a already guarantees that each of the $n+h$ will be coprime to q.

Similarly to (17.4), we expect that

$$\sum_{\substack{n<x\\ n\equiv a(\mathrm{mod}\,q)}}\prod_{h\in\mathcal{H}}\left(\mathbf{1}_{\mathcal{P}}(n+h)-\frac{q}{\phi(q)\log n}\right)\sim\mathfrak{S}_{q,0}(\mathcal{H})\left(\frac{q}{\phi(q)}\right)^{|\mathcal{H}|}\frac{1}{q}\int_2^x\frac{dt}{(\log t)^{|\mathcal{H}|}}. \tag{17.6}$$

17.2. **The main conjecture for $r=2$.**

Let a and b be two reduced residue classes $(\mathrm{mod}\,q)$, and let h be a positive integer with $b-a\,(\mathrm{mod}\,q)$. We formulate a conjecture for the number of primes $n\leq x$ with $n\equiv a\,(\mathrm{mod}\,q)$, and such that the next prime after n is $n+h$.

This is expected to be

$$\begin{aligned} &\sum_{\substack{n\le x\\ n\equiv a(\mathrm{mod}\, q)}} \mathbf{1}_{\mathcal{P}}(n)\mathbf{1}_{\mathcal{P}}(n+h) \prod_{\substack{0<t<h\\ (t+a,q)=1}} (1-\mathbf{1}_{\mathcal{P}}(n+t)) \\ &= \sum_{\substack{n\le x\\ n\equiv a(\mathrm{mod}\, q)}} \mathbf{1}_{\mathcal{P}}(n)\mathbf{1}_{\mathcal{P}}(n+h) \prod_{\substack{0<t<h\\ (t+a,q)=1}} \left(1-\frac{q}{\phi(q)\log(n+t)} - \widetilde{\mathbf{1}}_{\mathcal{P}}(n+t)\right). \end{aligned} \tag{17.7}$$

Here $\widetilde{\mathbf{1}}_{\mathcal{P}}$ is defined so that the second equation is a tautology, and so that it is zero on average. Giving the first two $\widetilde{\mathbf{1}}_{\mathcal{P}}$'s the same treatment, expanding the product, and approximating $\log(n+t)\approx \log n$, we obtain[1]

$$\sum_{\mathcal{A}\subset\{0,h\}} \sum_{\substack{\mathcal{T}\subset[1,h-1]\\ (t+a,q)=1\forall t\in\mathcal{T}}} (-1)^{|\mathcal{T}|} \sum_{\substack{n\le x\\ n\equiv a(\mathrm{mod}\, q)}} \left(\frac{q}{\phi(q)\log n}\right)^{2-|\mathcal{A}|} \prod_{\substack{t\in[1,h-1]\\ (t+a,q)=1\\ t\notin\mathcal{T}}} \left(1-\frac{q}{\phi(q)\log n}\right) \prod_{t\in\mathcal{A}\cup\mathcal{T}} \tilde{\mathbf{1}}_{\mathcal{P}}(n+t). \tag{17.8}$$

Given reduced residue classes a and b, and a positive $h\equiv b-a\,(\mathrm{mod}\, q)$, we may write

$$\#\{0<t<h:\ (t+a,q)=1\} = \frac{\phi(q)}{q}h+\epsilon_q(a,b), \tag{17.9}$$

where $\epsilon_q(a,b)$ is independent of h. We also write for convenience

$$\alpha(y) = 1-\frac{q}{\phi(q)\log y}. \tag{17.10}$$

Appealing now to the (conjectural) modified Hardy-Littlewood heuristic (17.6), we are led to hypothesize that the quantity in (17.7) (and (17.8)) is

$$\sim \sum_{\mathcal{A}\subset\{0,h\}} \sum_{\substack{\mathcal{T}\subset[1,h-1]\\ (t+a,q)=1\forall t\in\mathcal{T}}} (-1)^{|\mathcal{T}|}\mathfrak{S}_{q,0}(\mathcal{A}\cup\mathcal{T})\Big(\frac{1}{q}\int_2^x \Big(\frac{q}{\phi(q)\log y}\Big)^{2+|\mathcal{T}|}\alpha(y)^{h\phi(q)/q+\epsilon_q(a,b)-|\mathcal{T}|}dy\Big). \tag{17.11}$$

Now we just sum this expression over all positive integers $h\equiv b-a\,(\mathrm{mod}\, q)$ and we're done – *in principle.* We conjecture that

$$\pi(x;q,(a,b)) \sim \frac{1}{q}\int_2^x \alpha(y)^{\epsilon_q(a,b)}\Big(\frac{q}{\phi(q)\log y}\Big)^2\mathcal{D}(a,b;y)dy, \tag{17.12}$$

say, where

$$\mathcal{D}(a,b;y) = \sum_{\substack{h>0\\ h\equiv b-a(\mathrm{mod}\, q)}} \sum_{\mathcal{A}\subset\{0,h\}} \sum_{\substack{\mathcal{T}\subset[1,h-1]\\ (t+a,q)=1\forall t\in\mathcal{T}}} (-1)^{|\mathcal{T}|}\mathfrak{S}_{q,0}(\mathcal{A}\cup\mathcal{T})\Big(\frac{q}{\phi(q)\alpha(y)\log y}\Big)^{|\mathcal{T}|}\alpha(y)^{h\phi(q)/q}. \tag{17.13}$$

It is now time to practice the analytic number theorist's most-cherished discipline: to stare down a messy, complicated expression and figure out what is 'actually going on'.

[1] copy-and-paste from original paper!

Terms with $\mathcal{T} = \emptyset$. This is a suitable starting place. The contribution of these terms to $\mathcal{D}(a, b; y)$ is

$$\begin{aligned}\mathcal{D}_0(a, b, y) :=& \sum_{\substack{h>0\\ h\equiv b-a (\operatorname{mod} q)}} \sum_{\mathcal{A}\subset\{0,h\}} \mathfrak{S}_{q,0}(\mathcal{A})\alpha(y)^{h\phi(q)/q}\\ =& \sum_{\substack{h>0\\ h\equiv b-a (\operatorname{mod} q)}} \big(1 + \mathfrak{S}_q(\{0, h\}\big)\alpha(y)^{h\phi(q)/q},\end{aligned}$$

because $\mathfrak{S}_{q,0}$ is 1 for the empty set, and 0 for a singleton. The contribution to (17.12) is

$$\begin{aligned}&\frac{1}{q}\int_2^x \alpha(y)^{\epsilon_q(a,b)}\Big(\frac{q}{\phi(q)\log y}\Big)^2 \mathcal{D}_0(a, b; y)dy\\ =&\frac{q}{\phi(q)^2}\int_2^x \frac{\alpha(y)^{\epsilon_q(a,b)}}{(\log y)^2}\mathcal{D}_0(a, b; y)dy.\end{aligned}$$

Note that

$$\sum_{\substack{h>0\\ h\equiv b-a (\operatorname{mod} q)}} \alpha(y)^{h\phi(q)/q} = \alpha(y)^{h_0\phi(q)/q}\cdot\frac{1}{1-\alpha(y)^{\phi(q)}},$$

where h_0 is the smallest positive $h \equiv b - a \pmod q$.

In particular, we have

$$\frac{1}{1-\alpha(y)^{\phi(q)/q}} = \frac{1}{1-\Big(1-\frac{q}{\phi(q)\log y}\Big)^{\phi(q)}} \asymp \frac{\log y}{q}dy,$$

so that the contribution from 1 is roughly

$$\frac{q}{\phi(q)^2}\int_2^x \frac{\alpha(y)^{\epsilon_q(a,b)}}{(\log y)^2}\cdot\frac{\log y}{q} \sim \frac{1}{\phi(q)^2}\int_2^x \frac{1}{\log y}dy,$$

the expected main term!