3 Priemgetal-heuristiek
Benne de Weger

3.1 Inleiding

Wiskunde wordt traditioneel nog wel eens gezien als het bewijzen van stellingen. Je wilt toch graag wel zeker weten wat er aan de hand is, en een bewijs geeft je wat wel gezien wordt als de hoogste soort zekerheid, die gebaseerd op logica. Daar is natuurlijk veel voor te zeggen, maar er zijn ook wel beroemde vermoedens in de wiskunde, zoals het Vermoeden van Goldbach waarf Frits het over heeft, en de Hypothese van Riemann die centraal staat in de bijdrage van Roland. Kurt Gödel heeft weliswaar aangetoond dat er meer ware beweringen zijn dan bewijsbare, maar de meeste wiskundigen blijven positief ingesteld en zeggen dat we die bekende vermoedens *nóg* niet bewezen hebben, bijvoorbeeld omdat we daar met z'n allen nog niet aan toe zijn. Er is veel 'echte', d.w.z. bewezen, wiskunde ontwikkeld gestuurd door het verlangen om een bekend vermoeden te ontwikkelen, zoals in het geval van de Laatste Stelling van Fermat: de zoektocht die uiteindelijk leidde tot het bewijs van Wiles en Taylor heeft een enorme hoeveelheid fraaie wiskunde opgeleverd in de algebraische getaltheorie en arithmetische meetkunde.

Vermoedens spelen dus zeker een belangrijke rol in de wiskunde. Er is meer dan alleen stellingen en bewijzen. Er zijn zelfs vele stellingen van de vorm "Als Vermoeden X waar is, dan is ook Vermoeden Y waar", en het is zeker geen schande zo'n stelling te bewijzen. Maar je moet ook weer niet zomaar een of andere bewering verzinnen en die een "Vermoeden" gaan noemen, het is dan wel de bedoeling dat je *aannemelijk* maakt dat jouw vermoeden wel eens waar zou kunnen zijn. Dan kom je in het gebied van de *heuristiek* terecht. Het Griekse 'heuristein' betekent 'vinden' (denk aan 'Eureka'), heuristiek is dan zoiets als 'de kunst van het vinden'. Van Dale zegt ook 'de wetenschap die langs methodische weg tot ontdekkingen leert komen' en 'de methode die de leerling bij het onderwijs gelegenheid geeft zelf waarheden en regels te vinden'. In de praktijk betekent het in de getaltheorie vaak een 'probabilistische' redenering die via het opstellen van een kansmodel tot ideeën komt over wat wel eens waar zou kunnen

zijn. In dit deel van de cursus gaan we dat doen voor diverse telproblemen in de verzameling van de priemgetallen.

3.2 De Priemgetalstelling bekeken vanuit de kansrekening

De priemgetalstelling, al door Frits besproken, is er in een nauwkeuriger maar wat abstractere vorm:

$$\pi(x) \sim \mathrm{li}(x) = \int_2^x \frac{1}{\log t}\,\mathrm{d}t \quad \text{voor} \quad x \to \infty,$$

en in een wat 'slordiger' maar makkelijker te hanteren vorm:

$$\pi(x) \sim \frac{x}{\log x} \quad \text{voor} \quad x \to \infty.$$

Slordiger in de zin dat voor eindige x de benadering over het algemeen minder nauwkeurig is: de fout $|\pi(x) - \mathrm{li}(x)|$ is, als je de Riemann-Hypothese gelooft, kleiner dan $\sqrt{x}\log x$ voor x groot genoeg, terwijl de andere fout $\left|\pi(x) - \dfrac{x}{\log x}\right|$ van de grootteorde $\dfrac{x}{(\log x)^2}$ is, veel groter dus.
Van belang is om op te merken dat de Priemgetalstelling een *asymptotisch* resultaat is, dus iets zegt over de situatie waarbij je x naar ∞ laat gaan. Vaak wil je iets meer weten over hoe snel $\pi(x)$ naar $\mathrm{li}(x)$ of naar $\dfrac{x}{\log x}$ gaat, en wil je een verfijning van de asymptotische formule met een foutschatting. Het is bijvoorbeeld zo dat uit de Riemann-hypothese volgt dat

$$\pi(x) = \frac{x}{\log x} + \frac{x}{(\log x)^2} + \text{ een fout van grootteorde } \frac{x}{(\log x)^3} \text{ voor } x \to \infty.$$

Dit is nog steeds een asymptotische bewering. Maar ook wil je wel enige richtlijn hebben over wat er bij x van een bepaalde grootte ongeveer gebeurt. Als cryptoloog wil je bijvoorbeeld redelijke zekerheid (niet perse een bewijs) hebben over de vraag hoeveel priemgetallen er in een bepaald interval zullen zijn. Daar kan de Priemgetalstelling geen bewezen uitspraken over doen, maar wel een heuristiek geven, bijvoorbeeld door uit een asymptotische uitspraak simpelweg het gedeelte over de fout en het "voor $x \to \infty$" weg te laten. Dan mag je er formeel geen '='-teken meer bij gebruiken, en schrijven we "$\approx$", daarbij in het midden latend hoe groot de fout is in de benadering. Dus:

$$\pi(x) \approx \frac{x}{\log x} + \frac{x}{(\log x)^2}.$$

Dit lijkt aardig te kloppen, ook al voor redelijk kleine x, zie Tabel 3.1.

x	$\pi(x)$	$\frac{x}{\log x}$	$\frac{x}{\log x} + \frac{x}{(\log x)^2}$	$\epsilon(x)$
10^8	5761455	5428681	5723387	2.38
10^9	50847534	48254942	50583482	2.35
10^{10}	455052511	434294482	453155652	2.32
10^{11}	4118054813	3948131654	4104009089	2.28
10^{12}	37607912018	36191206825	37501010277	2.26
10^{13}	346065536839	334072678387	345233133832	2.23
10^{14}	3204941750802	3102103442166	3198333899825	2.21
10^{15}	29844570422669	28952965460217	29791239669157	2.20

Tabel 3.1: Benaderingen voor $\pi(x)$, $\epsilon(x)$ is de relatieve fout t.o.v. $\frac{x}{(\log x)^3}$,

$$\text{dus } \epsilon(x) = \frac{\pi(x) - \frac{x}{\log x} - \frac{x}{(\log x)^2}}{\frac{x}{(\log x)^3}}.$$

Een manier om hier tegenaan te kijken is met behulp van kansrekening. Frits deed dat al een beetje toen hij zei dat de 'dichtheid' van de priemgetallen ter grootte X ongeveer $1/\log X$ is. Dat betekent dat in de buurt van X ongeveer 1 op de $\log X$ getallen een priemgetal is. Je kunt ook zeggen dat de kans dat een willekeurig getal in de buurt van X een priemgetal is, ongeveer $1/\log X$ is. Met behulp van de vuistregel 'kans = aantal gunstige gebeurtenissen gedeeld door aantal mogelijke gebeurtenissen' kunnen we het zo uit de Priemgetalstelling afleiden: neem een interval $(X-\Delta, x+\Delta)$ rondom X, waarbij $\Delta \ll X$, dan zijn er ongeveer 2Δ getallen, waarvan $\pi(X+\Delta) - \pi(X-\Delta)$ priemgetallen zijn. De kans dat een willekeurig getrokken getal uit het interval een priemgetal is, is dan

$$\frac{\pi(X+\Delta) - \pi(X-\Delta)}{2\Delta} \approx \frac{\frac{X+\Delta}{\log(X+\Delta)} - \frac{X-\Delta}{\log(X-\Delta)}}{2\Delta}.$$

Hier kunnen we grip op krijgen door te gebruiken $\log(X \pm \Delta) = \log X + \log\left(1 + \frac{\pm\Delta}{X}\right)$, waarbij $\frac{\pm\Delta}{X}$ nu in de buurt van 0 zit, zodat we maar een heel kleine fout maken als we binnen de $\log(X \pm \Delta)$ de Δ verwaarlozen. Dan zien we meteen

$$\frac{\pi(X+\Delta) - \pi(X-\Delta)}{2\Delta} \approx \frac{\frac{X+\Delta}{\log X} - \frac{X-\Delta}{\log X}}{2\Delta} = \frac{1}{\log X}.$$

Opgave 3.2.1. *Maak een soortgelijke redenering door nu* $\mathrm{li}(x)$ *te gebruiken in plaats van* $\dfrac{x}{\log x}$. *Gebruik dat de functie* $\log x$ *op het interval* $(X - \Delta, X + \Delta)$ *vrijwel constant is.*

Opgave 3.2.2. *Frits heeft het over gaten tussen de priemgetallen:* $g_n = p_{n+1} - p_n$, *waarbij* p_n *het* n*-e priemgetal is, waarbij* g_n *zo klein als* 2 *kan zijn, ook vaak groter dan* $\log p_n$, *maar vermoedelijk (Cramér) niet groter dan* $C(\log p_n)^2$. *Leid uit de Priemgetalstelling een heuristiek af voor de grootte van het gemiddelde gat* g_n, *in termen van* p_n.

Opgave 3.2.3. *Cryptologen willen graag weten dat er voldoende priemgetallen zijn van een bepaalde grootte, bv. die in binaire schrijfwijze een voorgeschreven aantal bits hebben, zeg* b. *Geef een zo eenvoudig mogelijke formule die het aantal priemgetallen telt in het interval* $[2^{b-1}, 2^b - 1]$. *Het aantal elementaire deeltjes in het universum wordt geschat op* 3×10^{80}. *Hoe groot moet je* b *kiezen om op ongeveer evenveel priemgetallen van* b *bits uit te komen? Ter vergelijking: de meeste websites gebruiken voor hun beveiliging tegenwoordig* $b = 1024$.

3.3 Priemgetallen in congruentieklassen: het blijkt niet eerlijk te zijn

We gaan een stapje verder. Frits noemt de stellingen van Dirichlet en Siegel-Walfisz. Laat $q > 1$ een positief geheel getal zijn, en a een geheel getal ≥ 1 en $\leq q - 1$, dat relatief priem is met q. Dan tellen we het aantal $\pi(x, q, a)$ van de priemgetallen $p \leq x$ zodat $p \equiv a \pmod{q}$. De genoemde stellingen zeggen dan in feite dat

$$\pi(x, q, a) \sim \frac{1}{\phi(q)} \pi(x).$$

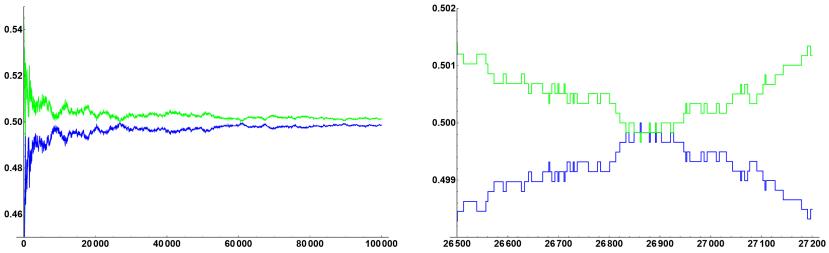
Merk op dat $\phi(q)$ precies het aantal mogelijke a's is.

Dit betekent dus een asymptotisch eerlijke verdeling van de priemgetallen over de congruentieklassen modulo q, de gegeven uitdrukking hangt niet van a af.

In termen van kansen zeggen we dan: voor een gegeven a zijn de kansen op de gebeurtenissen "x is een priemgetal" en "$x \equiv a \pmod{q}$" onafhankelijk, dus de kans dat beide gebeurtenissen tegelijk optreden is het product van de afzonderlijke kansen, respectievelijk $\dfrac{1}{\log x}$ en $\dfrac{1}{\phi(q)}$. Er is kennelijk geen reden om een afhankelijkheid van die gebeurtenissen te veronderstellen.

Pari opgave 3.3.1. *Met Pari kunnen we dit wel testen. De opdracht*
`tel(q,x) = t=vector(q-1); forprime(p=1,x,if(gcd(p,q)==1, t[p%q]++)); t`
definieert een functie `tel(q,s)` *die een lijstje geeft van de aantallen priemgetallen* $\leq x$ *in de congruentieklassen* $1, 2, \ldots, q-1 \pmod q$. *Bijvoorbeeld,* `tel(12,10000)` *geeft* `[300, 0, 0, 0, 309, 0, 311, 0, 0, 0, 307]`. *Speel hier wat mee; begin bijvoorbeeld met* $q = 3$ *of* $q = 4$ *en probeer verschillende waarden van* x*, je kunt rustig tot 1 miljoen gaan. Doe het ook eens met* $q = 10$*, dat geeft tellingen van priemgetallen met een gegeven laatste decimaal.*
Valt je iets bijzonders op?

In Figuur 3.1 staan de grafieken van $\dfrac{\pi(x,4,1)}{\pi(x)-1}$ (blauw) en $\dfrac{\pi(x,4,3)}{\pi(x)-1}$ (groen), op het interval $(2, 10^5)$, en ingezoomd rond 26900.



Figuur 3.1: $\dfrac{\pi(x,4,1)}{\pi(x)-1}$ (blauw) en $\dfrac{\pi(x,4,3)}{\pi(x)-1}$ (groen).

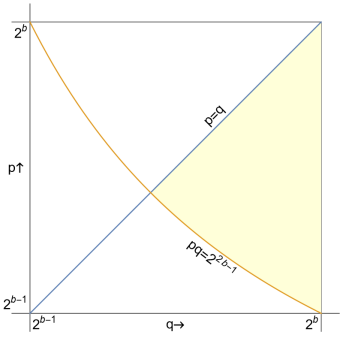
Wat nou eerlijk verdeeld. Het lijkt erop dat 3 (mod 4) systematisch meer van de priemgetallenkoek krijgt dan 1 (mod 4). Dit verschijnsel treedt bij alle moduli in meer of mindere mate op. het is niet in tegenspraak met de stelling van Siegel-Walfisz, want die zegt alleen iets over $x \to \infty$, en die heeft geen verder uitgewerkte foutterm.

Dit verschijnsel staat bekend als *Chebyshev's bias*. Rubinstein en Sarnak wisten in 1994 er een verklaring voor te geven, door te laten zien dat de fouttermen in de asymptotische ontwikkelingen van $\pi(x,q,a)$ niet allemaal dezelfde vorm hebben voor alle a. Het voert voor deze cursus veel te ver daar dieper op in te gaan. maar het verschijnsel is wel aan den lijve te ondervinden, als je gereedschap als Pari hebt.

3.4 RSA-moduli

Voor een RSA-sleutelpaar heb je twee grote priemgetallen p, q nodig, zo laat Frits zien. Cryptologen willen ze graag ongeveer even groot hebben, zeg elk van b bits (dus in het interval $(2^{b-1}, 2^b)$), en ook nog zodanig dat hun product $N = pq$ precies $2b$ bits heeft. En ze moeten natuurlijk verschillend zijn. Dan leveren ze een goede RSA-modulus N op.

Hoeveel goede RSA-moduli van $2b$ bits zijn er? Dat kunnen we met behulp van de Priemgetalstelling en onze heuristische aanpak goed schatten. We kunnen zonder verlies van algemeenheid stellen dat $p < q$. De eisen zijn dat p, q priemgetallen zijn, met $2^{b-1} < p < q < 2^b$, en $2^{2b-1} < pq < 2^{2b}$. Zie Figuur 3.2.



Figuur 3.2: Het toegestane gebied V (geel) voor goede RSA-moduli.

Voor ieder punt $(p, q) \in V$ hebben we nu als eis dat zowel p als q priemgetallen moeten zijn, met kansen respectievelijk $\dfrac{1}{\log p}$ en $\dfrac{1}{\log q}$, die niet overal in het gebied dezelfde zijn. Maar per punt zijn die twee kansen wel onafhankelijk, althans, we zien geen reden om anders te veronderstellen. We moeten dus $\dfrac{1}{(\log p)(\log q)}$ voor alle (p, q) in het gebied optellen. Je kunt proberen van deze dubbele som een tweevoudige integraal te maken en die uit te rekenen; soms gaat dat goed, maar in dit geval is het helaas nogal lastig.

Daarom gebruiken we voor p en q een overschatting van 2^b elk, dat zal een onderschatting geven van de kansen en dan zitten we aan de veilige kant. Dan hebben we opeens een kans die niet meer van de variabelen p en q afhangt, en om de som uit te rekenen hoeven we dan alleen het aantal punten in V met gehele coördinaten te hebben. Een goede benadering

daarvan is de oppervlakte van V, en die is uit te rekenen:

$$\int_{2^{b-\frac{1}{2}}}^{2^b} \left(q - \frac{2^{2b-1}}{q}\right) \mathrm{d}q = \left(\frac{1}{2}q^2 - 2^{2b-1}\log q\right)\Big|_{2^{b-\frac{1}{2}}}^{2^b} = 2^{2b-2}(1-\log 2).$$

Vermenigvuldigd met de schattingen voor de kansen, elk $\dfrac{1}{\log 2^b}$, krijgen we nu als ondergrens voor het aantal goede RSA-moduli

$$\frac{2^{2b}}{b^2}\frac{1-\log 2}{4(\log 2)^2} \approx 0.16\frac{2^{2b}}{b^2}.$$

Om een bovengrens te krijgen gebruiken we onderschattingen, 2^{b-1} voor p en q (beetje grof...). Eenzelfde redenering geeft dan een bovengrens voor het aantal goede RSA-moduli:

$$\frac{2^{2b}}{(b-1)^2}\frac{1-\log 2}{4(\log 2)^2} \approx 0.16\frac{2^{2b}}{(b-1)^2}.$$

Omdat we in de cryptografie tegenwoordig pas bij $b = 1024$ beginnen, is de bovengrens maar ongeveer 0.2% groter dan de ondergrens. Dat valt weg in de afronding naar 0.16 die we toch al maakten. Met $b = 1024$ is het aantal goede RSA-moduli dus $0.16 \times 2^{2028} \approx 5 \times 10^{609}$. Meer dan we als mensheid ooit zullen kunnen opmaken.

Een andere vraag die van groot belang is in de cryptografie is hoe we dit soort grote priemgetallen kunnen maken. Een theoretisch wiskundige kan gewoon zeggen: kies een priemgetal van 1024 bits. Maar een toegepast wiskundige heeft daar niet zoveel aan, die heeft een methode nodig die zowel effectief is als efficiënt. De methode die in de praktijk gebruikt wordt is in essentie verbijsterend simpel: genereer een willekeurige rij bits van 1022 lang, zet er een 1-bit voor en een 1-bit achter zodat je een getal krijgt dat echt 1024 bits heeft en niet minder, en oneven is; laat daar de priemtest van Rabin op los (zie de bijdrage van Frits), en als daar uitkomt dat het getal samengesteld is, dan gooi je het weg en begin je opnieuw.

Deze methode is efficiënt om twee redenen: Rabin is heel goed te doen voor getallen van 1024 bits (alleen een paar machtsverheffingen zijn nodig maar Frits heeft laten zien dat dat snel kan), en het verwachte aantal keren dat je opnieuw moet beginnen is 1 gedeeld door de kans op succes, en volgens de Priemgetalstelling is die kans $\dfrac{2}{\log 2^{1024}} \approx \dfrac{1}{355}$. Dus je verwacht dat je gemiddeld zo'n 300 keer hoeft te proberen. Fluitje van een cent.

De methode is ook effectief. Hoewel Rabin geen zekerheid geeft over een correct antwoord, is de kans op een misser (een getal dat door de test

heenkomt als vermoedelijk priem maar in feite samengesteld is) zo klein te maken dat je je daar echt geen zorgen over hoeft te maken. Dat klein maken doe je door de test een aantal keren te herhalen: de test is gerandomiseerd dus een paar keer toepassen op hetzelfde getal geeft toch iedere keer een andere test. Men kan aantonen dat die testresultaten met hoge kans onafhankelijk zijn, en dat je het maar een stuk of 6 ker hoeft te doen om de kans op een misser astronomisch klein te maken. Naar verwachting gaat het pas voor de eerste keer (wereldwijd) mis als de zon al uitgeblust is.

Pari opgave 3.4.1. *De volgende Pari-code maakt priemgetallen zoals zojuist beschreven.*

```
maakpriem(b) = x=0; t=0; while(!ispseudoprime(x),t++; x=2*
random(2^(b-2))+2^(b-1)+1); printf(``aantal pogingen:  %d'',t);x
```

Experimenteer er maar wat mee. De code laat zien hoeveel verschillende getallen getest zijn. Gebruik de met deze methode gemaakte priemgetallen ook eens bij Opgave 1.7.2 van Frits.

3.5 Priemtweelingen en Sophie Germain-priemgetallen

Een priemtweeling is een paar priemgetallen met een gat van slechts 2, dus $p, p+2$. Het is niet bekend of er oneindig veel van zijn, maar het sterke vermoeden is dat dat wel zo is. Een argument is de heuristiek aan de hand van de Priemgetalstelling. Uitgaande van de veronderstelling dat het priem zijn van p en $p+2$ onafhankelijke gebeurtenissen zijn, zou de kans erop $\dfrac{1}{(\log p)^2}$ zijn (eigenlijk $\dfrac{1}{(\log p)(\log(p+2))}$, maar dat is vrijwel

x	$\pi_2(x)$	$c(x)$
10^3	35	1.06562
10^4	205	1.27805
10^5	1224	1.29672
10^6	8169	1.30806
10^7	58980	1.32546
10^8	440312	1.32016
10^9	3424506	1.32002
10^{10}	27412679	1.32038

x	$\pi_2(x)$	$c(x)$
10^{11}	224376048	1.32037
10^{12}	1870585220	1.32034
10^{13}	15834664872	1.32033
10^{14}	135780321665	1.32032
10^{15}	1177209242304	1.32032
10^{16}	10304195697298	1.32032
10^{17}	90948839353159	1.32032
10^{18}	808675888577436	1.32032

Tabel 3.2: Priemtweelingen tellen, $c(x) = \dfrac{\pi_2(x)}{\mathrm{li}_2(x)}$.

hetzelfde), en zou de priemtweelingtelfunctie $\pi_2(x)$, het aantal priemgetallen $p \leq x$ zodanig dat ook $p+2$ priem is, asymptotisch equivalent zijn met $\mathrm{li}_2(x) = \displaystyle\int_2^x \frac{1}{(\log t)^2}\mathrm{d}t$, oftewel ook met $\dfrac{x}{(\log x)^2}$. Zie Tabel 3.2.

Hier is iets vreemds aan de hand: we verwachtten dat $c(x)$ naar 1 zou gaan convergeren, maar dat lijkt niet te gebeuren, en wel tamelijk overtuigend. Misschien is onze onafhankelijkheidsaanname niet goed?

Inderdaad is dat hier het probleem. Laten we beginnen met een priemgetal p, groter dan 2, dus oneven. Maar dan heeft $p+2$ al geen keus meer tussen even en oneven, en zal dus een grotere kans hebben om een priemgetal te zijn, en wel met een factor 2. En datzelfde kunnen we doen, niet alleen voor modulus 2 (even/oneven), maar voor iedere priem q: als we al weten dat q geen deler is van p, dan is de kans dat q ook geen deler is van $p+2$ niet meer $\dfrac{q-1}{q}$, maar $\dfrac{q-2}{q-1}$. Want we weten al dat er voor p slechts $q-1$ mogelijkheden $\pmod q$ zijn, namelijk $1, 2, \ldots, q-3, q-2, q-1$, maar niet 0. En dan zijn er voor $p+2 \pmod q$ ook $q-1$ mogelijkheden, namelijk $3, 4, \ldots, q-1, 0, 1$, maar niet 2. De kans dat $p+2$ niet deelbaar is door q, gegeven dat p dat niet is, is dus $\dfrac{q-2}{q-1}$. Daar staat tegenover dat de kans dat een volledig willekeurig getal niet deelbaar door q is, gelijk is aan $\dfrac{q-1}{q}$, en we moeten nu de impliciet aanwezige factor $\dfrac{q-1}{q}$ in de kans vervangen door $\dfrac{q-2}{q-1}$. Dat geeft dus een wijzigingsfactor van $\dfrac{q-2}{q-1}/\dfrac{q-2}{q-1} = \dfrac{q(q-2)}{(q-1)^2} = 1 - \dfrac{1}{(q-1)^2}$. Voor de priemgetallen $3, 5, 7, 11, \ldots$ geeft dit extra factoren $\dfrac{3}{4}, \dfrac{15}{16}, \dfrac{35}{36}, \dfrac{99}{100}, \ldots$.

Dit suggereert dat we de hierboven gebruikte heuristiek kunnen corrigeren, met een factor 2 voor het even/oneven-effect, en voor ieder oneven priemgetal met een factor $1 - \dfrac{1}{(q-1)^2}$. Dit leidt tot het invoeren van de *priemtweelingconstante*:

$$C_2 = \prod_{q \text{ oneven priem}} \left(1 - \frac{1}{(q-1)^2}\right).$$

Is dit oneindige product convergent? Ja, want je vermenigvuldigt altijd met een getal > 0 en < 1. Maar dan loop je het risico dat er 0 uitkomt. Neem de logarithme om daar achter te komen:

$$\log C_2 = \sum_{q \text{ oneven priem}} \log\left(1 - \frac{1}{(q-1)^2}\right),$$

en omdat $-\log\left(1 - \frac{1}{x}\right) = \log\left(1 + \frac{1}{x-1}\right) < \frac{1}{x-1}$ voor $x \geq 2$, vinden we

$$\begin{aligned} -\log C_2 &< \sum_{q \text{ oneven priem}} \frac{1}{(q-1)^2 - 1} < \sum_{n=3}^{\infty} \frac{1}{(n-1)^2 - 1} \\ &= \sum_{n=3}^{\infty} \frac{1}{n(n-2)} = \frac{1}{2} \sum_{n=3}^{\infty} \left(\frac{1}{n-2} - \frac{1}{n}\right) = \frac{3}{4}. \end{aligned}$$

Dus is $C_2 > e^{-3/4} = 0.47\ldots$. Een precieze berekening geeft

$$C_2 = 0.66016\ldots,$$

en dus is ons vermoeden nu dat

$$\pi_2(x) \sim 2C_2 \cdot \mathrm{li}_2(x) = 1.32032\ldots \cdot \mathrm{li}_2(x).$$

Tabel 3.2 geeft duidelijk ondersteuning voor dit vermoeden. Kennelijk hebben we nu de afhankelijkheden in de kansen goed te pakken. Maar let wel, bewezen is hier niets. Het blijft pure heuristiek.

Frits keek naar de som van de omgekeerde priemgetallen $\sum_{p \text{ priem}} \frac{1}{p}$, en uit de divergentie ervan concludeerde hij dat er dus oneindig veel priemgetallen bestaan. Dat was een net bewijs, maar vanuit de heuristiek is dit ook aannemelijk te maken: omdat het n-e priemgetal ongeveer $n \log n$ zal zijn, kunnen we de som benaderen met $\sum_{n=1}^{\infty} \frac{1}{n \log n}$, en dus met de integraal $\int_2^{\infty} \frac{1}{x \log x} \mathrm{d}x$. Een primitieve van $\frac{1}{x \log x}$ is $\log \log x$, en voor $x \to \infty$ divergeert dit inderdaad.

Net zo kunnen we nu kijken naar de som van de omgekeerde priemtweelingen: $\sum_{p \text{ priem en } p+2 \text{ ook}} \frac{1}{p}$ (of, zo je wilt, $\sum_{p \text{ priem en } p+2 \text{ ook}} \left(\frac{1}{p} + \frac{1}{p+2}\right)$). Maar de heuristiek zegt ons dat nu de n-e priemtweeling bij $n(\log n)^2$ in de buurt zal liggen, en de oneindige som zal zich gedragen als $\sum_{n-1}^{\infty} \frac{1}{n(\log n)^2}$, en

dus als de integraal $\int_2^\infty \frac{1}{x(\log x)^2}\mathrm{d}x$. Maar een primitieve van $\frac{1}{x(\log x)^2}$ is $-\frac{1}{\log x}$, en deze som zal dus wel convergeren. Dat dit echt zo is is bewezen door Brun. Brun's constante is

$$\sum_{p \text{ priem en } p+2 \text{ ook}} \left(\frac{1}{p} + \frac{1}{p+2}\right) = 1.90216\ldots.$$

We kunnen niet uit dit bewijs van Brun concluderen dat er eindig veel priemtweelingen zijn, want een convergerende som kan eindig maar ook oneindig veel termen hebben. Gelukkig geeft ons vermoeden over $\pi_2(x)$ wel forse steun aan het vermoeden dat er oneindig veel zijn.

Dan noemen we nog even kort de Sophie Germain-priemgetallen. Een priemgetal p is vernoemd naar Sophie Germain (1776-1831, een van de eerste vrouwen die bekend werden in de wiskunde) als ook $2p+1$ een priemgetal is. Deze priemgetallen zijn nuttig in de cryptologie, omdat $\phi(2p+1) = 2p$ weinig delers heeft, en delers van $\phi(q)$ voor een priemgetal q zitten de veiligheid nog wel eens in de weg. Zo'n priem $2p+1$ wordt daarom ook wel een 'veilige priem' genoemd.

Opgave 3.5.1. *Bepaal de eerste* 10 *Sophie Germain-priemgetallen. Als je lui bent mag je Pari gebruiken.*

Dan willen we natuurlijk ook weten hoeveel er van zijn.

Opgave 3.5.2. *We voeren de Sophie Germanpriemtelfunctie $\pi_{SG}(x)$ in als het aantal priemgetallen $p \leq x$ waarvoor ook $2p+1$ priem is. Laat zien dat de volgende heuristiek geldt:*

$$\pi_{SG}(x) \sim 2C_2 \cdot \mathrm{li}_2(x),$$

waar C_2 de priemtweelingconstante is.

3.6 Patronen in opeenvolgende priemgetallen

Je kunt natuurlijk veel verder gaan met varianten bedenken op het thema 'priemtweelingen', zo zijn er de 'priemneefjes' $p, p+4$, de 'sexy priemgetallen' $p, p+6$, of het algemener maken met bijvoorbeeld het Vermoeden van Bunyakovsky (Zie de bijdrage van Frits). Enkele jaren geleden keken Robert Lemke Oliver en Kannan Soundararajan naar patronen van restklassen modulo q die optreden in rijtjes van opeenvolgende priemgetallen. Voor een vector $\mathbf{a} = (a_1, \ldots, a_r)$ van restklassen modulo q (met

alle $\gcd(a_i, q) = 1$) keken ze naar de priemtelfunctie $\pi(x, q, \mathbf{a})$ die het aantal priemgetallen $p \leq x$ telt zodat voor de opeenvolgende priemgetallen $p_1 = p, p_2, \ldots, p_r$ geldt dat $p_i \equiv a_i \pmod q$. Bijvoorbeeld, $\pi(x, 10, (f, g))$ telt het aantal priemgetallen $p \leq x$ zodat p als laatste decimaal f heeft, en het eerstvolgende priemgetal na p als laatste decimaal g heeft. Je zou verwachten dat de boel weer een beetje eerlijk verdeeld is, dat zou dan zijn

$$\pi(x, q, \mathbf{a}) \sim \frac{1}{\phi(q)^r} \operatorname{li}(x),$$

maar tot hun verbazing vonden ze experimenteel sterke afwijkingen, en tot ieders verbazing vonden ze er ook een verklaring voor. Tabel 3.3 geeft een indruk van wat er uit experimenten kwam.

	$g = 1$	$g = 3$	$g = 7$	$g = 9$
$f = 1$	4623042	7429438	7504612	5442345
$f = 3$	6010982	4442562	7043695	7502896
$f = 7$	6373981	6755195	4439355	7431870
$f = 9$	7991431	6372941	6012739	4622916

Tabel 3.3: $\pi(x_0, 10, (f, g))$ met $\pi(x_0) = 10^8 + 2$ voor $f, g \in \{1, 3, 7, 9\}$.

Dat zijn wel erg grote afwijkingen. Hun vermoeden is als volgt:

$$\begin{aligned} \pi(x, q, \mathbf{a}) \quad &\sim \quad \frac{1}{\phi(q)^r} \operatorname{li}(x) \left(1 + c_1(q, \mathbf{a}) \frac{\log\log x}{\log x} + c_2(q, \mathbf{a}) \frac{1}{\log x} \right. \\ &+ \quad \text{een foutterm van ordegrootte } \left. \frac{1}{(\log x)^{7/4}} \right), \end{aligned}$$

waarbij $c_1(q, \mathbf{a})$ en $c_2(q, \mathbf{a})$ alleen van q en $\mathbf{a}$ afhangen. Ze geven expliciete formules voor deze constanten, maar die zijn nogal gecompliceerd. Hoe dan ook, asymptotisch blijft het vermoeden $\pi(x, q, \mathbf{a}) \sim \frac{1}{\phi(q)^r} \operatorname{li}(x)$ fier overeind staan, maar er is wel degelijk meer aan de hand.

3.7 Alda-priemgetallen

Toen ik mijn echtgenote Alda, geen wiskundige, uitlegde wat priemtweelingen zijn, vroeg ze ogenblikkelijk of er ook priemtweelingen bestaan zodanig dat de eerstvolgende twee priemgetallen ook een priemtweeling zijn. Dus definieer ik een *Alda-priemgetal*[1] als een priemgetal p zodanig dat het met

[1] Later kwam ik er achter dat het ook wel een *priemtweelingcluster van orde* 2 heet, maar ik verkies dat te negeren.

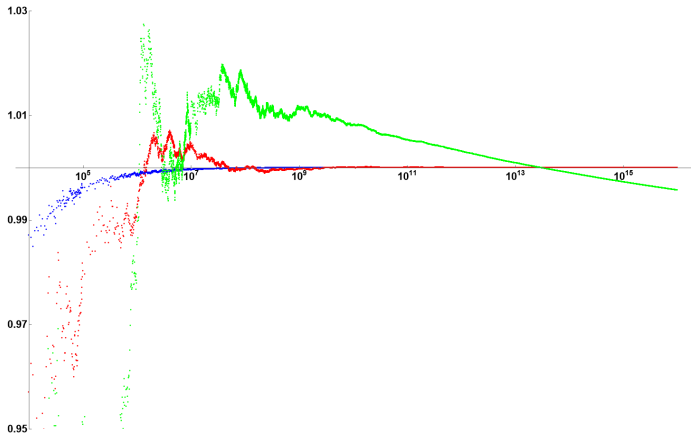
de eerstvolgende drie priemgetallen twee priemtweelingen vormt. De kleinste is 5, en 11 is de enige andere onder de 100, maar 101 is er ook weer eentje.

We definiëren nu uiteraard de Alda-priemgetaltelfunctie $\pi_{\mathrm{A}}(x)$ als het aantal Alda-priemgetallen $\leq x$. Ik legde dit probleem voor aan Pieter Moree, die schakelde zijn collega Efthymios Sofos in, en die kwam met de volgende heuristiek:

$$\pi_{\mathrm{A}}(x) \sim 4C_2^2 \cdot \mathrm{li}_3(x), \text{ waarbij } \mathrm{li}_3(x) = \int_2^x \frac{1}{(\log t)^3}\mathrm{d}t \sim \frac{x}{(\log x)^3}.$$

Opgave 3.7.1. *Kun je deze heuristiek verklaren? Waarom een derdemacht van* $\log x$ *terwijl er toch 4 priemgetallen zijn? Waarom de* C_2 *in het kwadraat?*

Toch is er iets vreemds aan de hand. Ik heb het "high performance cluster" van de TU/e zo'n 2065 uur Alda-priemgetallen laten tellen, en kwam tot $x < 10^{16}$. Deze berekeningen zijn onafhankelijk gecontroleerd dooe Alexander Weisse uit Bonn. Figuur 3.3 laat een vergelijking zien tussen $\dfrac{\pi(x)}{\mathrm{li}(x)}$ (priemgetallen, blauw), $\dfrac{\pi_2(x)}{2C_2 \cdot \mathrm{li}_2(x)}$ (priemtweelingen, rood), en $\dfrac{\pi_{\mathrm{A}}(x)}{4C_2^2 \cdot \mathrm{li}_3(x)}$ (Alda-priemgetallen, groen); alle drie de grafieken zouden naar hoogte 1 moeten convergeren.



Figuur 3.3: Alda-priemgetallen tellen vergeleken met gewone priemgetallen en priemtweelingen.

Hieruit blijkt dat de Alda-priemgetallen zich niet aan de heuristiek lijken

te houden. Of hebben we gewoon nog niet ver genoeg gerekend, en gaat de groene grafiek alsnog omhoog? Later heb ik ook nog een slordige 1000 uur Alda-priemgetallen in intervallen ter grootte 10^{10} en 10^{11} voor x tot aan 10^{100} geteld, en met veel goede wil zie je daar een vage trend de goede kant uit. We hebben nog geen goed idee over wat er hier an de hand is.

Het probleem van Alda-priemgetallen is lastiger dan de priemgetallen die Lemke Oliver en Soundararajan bekeken, omdat nu het gat tussen de twee priemtweelingen niet vastgelegd is.

We kunnen natuurlijk ook *sexy Alda-priemgetallen* bedenken: een viertal priemgetallen $p, p+6, q, q+6$ waarbij er tussen p en $p+6$ wel een priemgetal mag zitten, zo ook tussen q en $q+6$, maar niet tussen $p+6$ en q. Hun telfunctie $\pi_{6\mathrm{A}}(x)$ heeft als heuristiek

$$\pi_{6\mathrm{A}}(x) \sim 16C_2^2 \cdot \mathrm{li}_3(x),$$

en hier vinden we ook de onverklaarde afwijkingen die we bij $\pi_{\mathrm{A}}(x)$ tegenkwamen.

Het is overigens best aardig iets te laten zien van hoe deze tellingen uitgevoerd worden. Allereerst worden priemgetallen in intervallen opgedeeld, van bv. lengte 10^{10}. De priemgetallen worden niet zelf opgeslagen, maar we werken met een positiesysteem: elk getal krijgt een bit op een makkelijk terug te vinden positie toegewezen, en een priemgetal krijgt dan een 1-bit, een samengesteld getal een 0-bit. Omdat computers nu eenmaal met bytes van 8 bits werken, en omdat $n = 30$ het grootste getal is met $\phi(n) = 8$, slaan we alleen informatie op over getallen die relatief priem zijn met 30, dus alleen getallen $1, 7, 11, 13, 17, 19, 23, 29 \pmod{30}$, dan vergeten we alleen de priemgetallen $2, 3, 5$. Dus per 30 opeenvolgende getallen hoeven we slechts 1 byte op te slaan. Voor een interval van 10^{10} getallen kost dat slechts 318 MB.

We beginnen met alle bits op 1 te zetten, en voeren dan gewoon de Zeef van Eratosthenes uit, om alle samengestelde getallen er uit te zeven. Als we zo alle priemgetallen in het interval hebben, gaan we ze allemaal af, met het bijhouden van een toestand (p, g, b), hierbij is p het vorige bezochte priemgetal, g het gat tussen het vorige bezochte priemgetal en het priemgetal waar we nu zijn, en b een bit dat 1 is als we al een priemtweeling hadden gezien, 0 anders. Dan voeren we het volgende uit:

- als $g = 2$ en $b = 0$ dan hebben we een nieuwe priemtweeling gevonden en wordt $b = 1$,
- anders: als $g = 2$ en $b = 1$ dan hebben we een nieuw Alda-priemgetal gevonden, en hogen we de teller met 1 op,
- anders: als $g > 2$ en $b = 0$ dan doen we niets,

- anders: als $g > 2$ en $b = 1$ dan leidt de priemtweeling die we al gezien hadden niet tot een Alda-priemgetal, en wordt $b = 0$,
- als $g = 2$ ga naar het volgende priemgetal, en pas p en g aan,
- ga naar het volgende priemgetal, en pas p en g aan.

Aan het einde van een interval moet je even goed opletten dat je de toestand opslaat en doorgeeft als begintoestand voor het nieuwe interval. Een Alda-priemgetal kan immers net op de rand van een interval zitten.

3.8 Om het af te leren

We sluiten af met een aantal opgaven. Je mag in deze opgaven best een beetje grof redeneren, bv. een makkelijk gemiddelde nemen van een $\log x$ over een interval (bv. op $[1000, 2000]$ stijgt $\log x$ van ongeveer 6.9 naar ongeveer 7.6, doe dan net alsof dat over het hele interval ongeveer 7.3 is).

Opgave 3.8.1. *Een palindroomgetal is een getal dat van achter naar voren hetzelfde getal is, zoals* 123454321*. Geef een heuristiek voor het aantal palindroom-priemgetallen van n cijfers.*

Opgave 3.8.2. *Geef een heuristiek voor het aantal priemgetallen van n cijfers waar het cijfer* 3 *niet in voorkomt.*

Opgave 3.8.3. *Een Mersenne-priemgetal is een priemgetal van de vorm $2^n - 1$. Het is eenvoudig in te zien dat dat alleen kan als n zelf een priemgetal is. Denk je dat er eindig of oneindig veel Mersenne-priemgetallen zijn?*

Opgave 3.8.4. *Een Fermat-priemgetal is een priemgetal van de vorm $2^n + 1$. Het is eenvoudig in te zien dat dat alleen kan als $n = 2^k$. Denk je dat er eindig of oneindig veel Fermat-priemgetallen zijn?*

Opgave 3.8.5. *Geef een heuristiek voor het gemiddelde aantal manieren waarop je een even getal n kunt schrijven als de som van twee priemgetallen. Zie wat Frits schrijft over het Goldbach-vermoeden.*

Opgave 3.8.6. *Geef een heuristiek voor het aantal priemgetallen $\leq x$ die van de vorm $n^2 + 1$ zijn (zie de bijdrage van Frits over het Vermoeden van Bunyakovsky). Maak je geen zorgen over de constante, alleen over de juiste grootteorde.*

Pari opgave 3.8.7. *Maak een priemgetal dat begint met je eigen telefoonnummer. Bedenk van te voren hoeveel cijfers je verwacht er aan toe te moeten voegen.*

Literatuur

Paul Levrie en Rudi Penne – De pracht van priemgetallen, Prometheus / Bert Bakker, 2014.

Richard Crandall and Carl Pomerance – Prime Numbers, A Computational Perspective, Springer, 2nd. Ed., 2005.

Jean-Marie De Koninck and Nicolas Doyon – The Life of Primes in 37 Episodes, AM. Math. Soc., 2020.