Über positive Lösungen linearer diophantischer Gleichungen und einen Satz von Teuffel

von
Wilfried Lex

Die Struktur monogener - oder auch "zyklischer" - Monoide ist wohlbekannt, s. etwa [1], 1.6, S. 19/20. Wir fragen hier ganz bescheiden zunächst nur nach der Struktur der von zwei Elementen erzeugten Untermonoide des freien monogenen Monoids: Dies führt unmittelbar auf die Frage nach denjenigen c aus $\mathbb{N}_0$, für die

$$(1) \qquad ax + by = c$$

mit $a,b \in \mathbb{N}_0$ eine nichtnegative ganzzahlige Lösung besitzt, wobei wir uns gleich auf den allein interessierenden Fall *teilerfremder natürlicher* a,b beschränken. Seit über hundert Jahren - zur Literatur sei auf [3] hingewiesen - ist die optimale untere Schranke für c derart, daß (1) stets in $\mathbb{N}_0 \times \mathbb{N}_0$ lösbar ist, als $(a-1)\cdot(b-1)$ bekannt.

In [3] wurde in leichter Verallgemeinerung die optimale untere Schranke für c bestimmt, die eine Lösung von (1) durch stets mindestens $n (\in \mathbb{N})$ Paare $(x,y) \in \mathbb{N}_0 \times \mathbb{N}_0$ garantiert (l.c., 3., Sätze 2 und 3, oder allgemeiner weiter unten Hilfssatz 1, b) und c)). - Für die unter der Bezeichnung *Frobeniussches Problem* bekannt gewordene, im Folgenden nicht tangierte, interessante und schwierige Frage nach entsprechenden Schranken bei Gleichungen der Form

$$\sum_{\nu=1}^{n} a_\nu x_\nu = c$$

mit $n \in \mathbb{N} \setminus \{1,2\}$ und $\mathrm{ggT}(a_1,\dots,a_n) = 1$ verweisen wir auf [2], [4] und [5].

Ferner wurde in [3] die Anzahl jener c berechnet, die unterhalb der genannten Schranke liegen und (1) mit genau n Paaren $(x,y) \in \mathbb{N}_0 \times \mathbb{N}_0$ erfüllen (l.c., 4., Satz 6, oder allgemeiner weiter unten Satz 2a). Wie l.c., 5., ohne Beweis erwähnt, hat *E. Teuffel* [6] - dem ich für die freundliche Erlaubnis, seine Ergebnisse und Beweise verwenden zu dürfen, herzlich danke - für n = 1 eine erstaunlich einfache Antwort auf die Frage nach der Summe der so charakterisierten Zahlen c gegeben ([3], 5., Satz 7). Im Folgenden wird dieses Resultat auf ganz elementare Weise in zwei Richtungen verallgemeinert: Zum einen werden ganzzahlige Lösungen x,y behandelt, die mindestens gleich $t \; (\in \mathbb{Z})$ sind (Satz 1), zum anderen das Analogon für n Lösungen (Satz 3).

Kleine (lateinische) Buchstaben mögen als Variable für ganze Zahlen fungieren; speziell seien a, b, n stets *positive ganze* Zahlen und a, b *teilerfremd* sowie $s \Leftrightarrow a+b$ und $p \Leftrightarrow (a-1)(b-1)$. Ferner sei

$$N(c,t) \Leftrightarrow |\{(x,y)|\ (1) \wedge x,y \geq t\}|$$

und

$$M(n,k) \Leftrightarrow |\{x \in [0,nab-s]\,|\,N(x,0) = k\}|.$$

Zum leichteren Verweisen stellen wir einige Ergebnisse aus [3] zusammen:

Hilfssatz 1: a) $N(c,0) = N(c+ts,\ t)$,

b) $c > nab + (t-1)s \implies N(c,t) \geq n$,

c) $N(nab+ts,\ t+1) = n-1$,

d) $c < (n-1)ab + ts \implies N(c,t) < n$,

e) $M(n,0) = \frac{p}{2} = M(n,n)$,

f) $n > 1 \implies M(n,k) = ab\ (k=1,\dots,n-1)$.

a) wird in [3], 3., als Hilfssatz 1 bewiesen, b) bzw. c) als Korollar b) zu Satz 2 bzw. zu Satz 3 und d) sowie e) bzw. f) l.c., 4., als Hilfssatz 2 sowie als Teil a) bzw. c) von Satz 6.

Als Abkürzungen benutzen wir

$$K(x) \leftrightharpoons \left[\frac{a}{b}(x-t+1)\right]$$

und

$$\rho \leftrightharpoons \sum_{\nu=1}^{b-1} \left\{\frac{a}{b}\nu\right\}\nu \, ,$$

wobei für $\xi \in \mathbb{R}$ wie üblich $[\xi] \leftrightharpoons \max\{x \mid x \leq \xi\}$ und $\{\xi\} \leftrightharpoons \xi - [\xi]$ sei. - Zur späteren Verwendung benötigen wir den - für $t=0$ von *Teuffel* [6] stammenden -

Hilfssatz 2:

a) $$\sum_{x=t}^{b+t-2} K(x) = \frac{\rho}{2} \, ,$$

b) $$\sum_{x=t}^{b+t-2} x\,K(x) = \frac{b-1}{6}\left(a(2b-1) + 3(t-1)(a-1)\right) - \rho \, ,$$

c) $$\sum_{x=t}^{b+t-2} K(x)^2 = \frac{a^2+1}{6b}(b-1)(2b-1) - \frac{2a}{b}\rho \, .$$

Beweis (der i.w. mit dem unveröffentlichten, von *Teuffel* l.c. gegebenen identisch ist):

a) Wegen $ggT(a,b) = 1$ ist bekanntlich (s. etwa [7], Kap. III, Aufg. 2a, S. 35 bzw. S. 106)

$$(2) \qquad \sum_{\nu=1}^{b-1} \left\{\frac{a}{b}\nu\right\} = \frac{b-1}{2}$$

und damit

$$\sum_{x=t}^{b+t-2} K(x) = \sum_{\nu=1}^{b-1} \frac{a}{b}\nu - \sum_{\nu=1}^{b-1} \left\{\frac{a}{b}\nu\right\} = \frac{b-1}{2}(a-1).$$

b) Mittels a) ergibt sich

$$\sum_{x=t}^{b+t-2} x\,K(x) = \sum_{\nu=1}^{b-1} \nu\left(\frac{a}{b}\nu - \left\{\frac{a}{b}\nu\right\}\right) + (t-1)\sum_{\nu=1}^{b-1}\left[\frac{a}{b}\nu\right] =$$

$$= \frac{b-1}{6}\left(a(2b-1) + 3(t-1)(a-1)\right) - \rho \, .$$

c) Analog zu (2) erhält man

$$\sum_{\nu=1}^{b-1} \{\tfrac{a}{b}\nu\}^2 = \sum_{\nu=1}^{b-1} b^{-2}\nu^2 = \frac{b-1}{6b}(2b-1)$$

und daher

$$\sum_{x=t}^{b+t-2} K(x)^2 = \frac{a^2}{b^2}\sum_{\nu=1}^{b-1}\nu^2 - 2\frac{a}{b}\sum_{\nu=1}^{b-1}\nu\{\tfrac{a}{b}\nu\} + \sum_{\nu=1}^{b-1}\{\tfrac{a}{b}\nu\}^2 =$$

$$= \frac{b-1}{6b}(2b-1)(a^2+1) - \frac{2a}{b}\rho .$$

Die uns zunächst interessierenden Summen sind

$$\sigma \leftrightharpoons \sum_{\substack{\nu=ts \\ N(\nu,t)\geq 1}}^{ab+(t-1)s} \nu \quad \text{und} \quad \bar{\sigma} \leftrightharpoons \sum_{\substack{\nu=ts \\ N(\nu,t)=0}}^{ab+(t-1)s} \nu \ ;$$

dabei ist die untere Summationsgrenze gemäß Hilfssatz 1d und $N(ts,t) = 1$ gerade das kleinste x mit $N(x,t) \geq 1$ und die obere die in Hilfssatz 1b genannte Schranke, die wegen Hilfssatz 1c optimal ist. In Verallgemeinerung der genannten Resultate von *Teuffel* [6] und i.w. parallel zu seinem Beweis ergibt sich

Satz 1: a) $\sigma = \frac{p}{12}(4ab + (6t-5)s + 1)$,

b) $\bar{\sigma} = \frac{p}{12}(2ab + (6t-1)s - 1)$.

Beweis:

a) O.B.d.A. sei $a \geq b$. - Für $y \geq t$ und

(3) $\qquad ax + by \leq ab + (t-1)s$

hat man $ax \leq ab + (t-1)a - b$, wegen $x \in \mathbb{Z}$ also

(4) $x \leq b + t - 2$.

Aus $t \leq x \leq b + t - 2$ folgt $b > 1$ sowie $0 < x - t + 1 < b$ und wegen $\mathrm{ggT}(a,b) = 1$ daher $\frac{a}{b}(x-t+1) \notin \mathbb{Z}$, mithin

(5) $\frac{a}{b}(x - t + 1) > K(x)$.

Mit (3) ergibt sich

$$by \leq ab - (x - t + 1)a + (t-1)b$$

und mit (5) wegen $y, K(x) \in \mathbb{Z}$ daraus

(6) $y \leq a + t - 2 - K(x)$.

Es gilt

$$\sigma = \sum_{\substack{ts \leq ax+by \leq ab+(t-1)s \\ t \leq x,y}} ax+by$$

und nach (4) und (6) (wegen $a \geq b$ sind die angegebenen Schranken optimal) daher

$$\sigma = \sum_{x=t}^{b+t-2} A(x) \quad \text{mit} \quad A(x) \Leftrightarrow \sum_{y=t}^{a+t-2-K(x)} ax + by ,$$

was

$$2\,A(x) = 2 \cdot \sum_{y=t}^{a+t-2-K(x)} ax + 2 \cdot \sum_{y=t}^{a+t-2-K(x)} by =$$

$$= b((a + t - 2)(a + t - 1) - (t - 1)t) + 2a(a - 1)x -$$

$$- b(2(a + t) - 3)K(x) - 2\,axK(x) + b\,K(x)^2 ,$$

wegen Hilfssatz 2 also

$$\sum_{x=t}^{b+t-2} 2\,A(x) = \frac{p}{2}(2\,ab - 2a - b) + tp(b + a) - \frac{b-1}{6}(2b - 1)(a^2 - 1)$$

und mithin die Behauptung bedingt.

b) Aus

$$\sigma + \overline{\sigma} = \sum_{\nu=ts}^{ab+(t-1)s} \nu = \frac{p}{2}\,(ab + (2t-1)s)$$

und a) folgt das Gewünschte.

Mit der Abkürzung

$$M(n,k,t) \leftrightharpoons |\{x \in [ts,\ nab + (t-1)s] \mid N(x,t) = k\}|$$

erhält man in Verallgemeinerung von Satz 6, [3], 4., den

Satz 2: a) $M(n,0,t) = \frac{p}{2} = M(n,n,t)$,

b) $n > 1 \Rightarrow M(n,k,t) = ab \quad (k=1,\dots,n-1)$,

c) $k > n \Rightarrow M(n,k,t) = 0$.

Beweis: Mit $n+1$ statt n ergibt Kontraposition von Hilfssatz 1d den Teil c). Nach Hilfssatz 1a gilt

$$M(n,k,t) = |\{x \in [0, nab - s] \mid N(x,0) = k\}| = M(n,k),$$

womit wegen Hilfssatz 1e und f) auch a) und b) bewiesen sind.

Als Vorbereitung zur zweiten angekündigten Verallgemeinerung beweisen wir noch

Hilfssatz 3: $c > (t-1)s \wedge k > -N(c,t) \Rightarrow N(c+kab,t) = N(c,t)+k$.

Beweis: Es sei $m \leftrightharpoons N(c,t)$. Um zunächst

(7) $$c > (t-1)s \Rightarrow N(c+ab,t) = m+1$$

zu zeigen, sei

a) $m = 0$: Aus $c > (t-1)s$ ergibt sich Hilfssatz 1b zufolge $N(c+ab,t) \geq 1$, also

(8) $$c > (t-1)s \Rightarrow N(c+ab,t) \geq m+1$$

für $m = 0$.

b) $m > 0$, etwa $ax_\mu + by_\mu = c$ und $x_\mu, y_\mu \geq t$ für $\mu = 1,\ldots,m$ und $x_\mu < x_{\mu+1}$ $(\mu = 1,\ldots,m-1)$. Mit

$$x'_\mu \leftrightharpoons x_\mu + b, \qquad y'_\mu \leftrightharpoons y_\mu \quad (\mu = 1,\ldots,m) ,$$

$$x'_{m+1} \leftrightharpoons x_1 , \qquad y'_{m+1} \leftrightharpoons y_1 + a$$

hat man $x'_\mu, y'_\mu \geq t$ sowie $x'_\mu \neq x'_\nu$ für $\mu \neq \nu$ und $ax'_\mu + by'_\mu = c + ab$ für $\mu,\nu = 1,\ldots,m+1$, also (8) auch für $m \geq 0$.

Wäre $N(c+ab,t) \geq m+2$, etwa $ax_\mu + by_\mu = c + ab$ mit $x_\mu, y_\mu \geq t$ für $\mu = 1,\ldots,m+2$ und $x_\mu < x_{\mu+1}$ $(\mu = 1,\ldots,m+1)$, so hätte man mit $x'_\mu \leftrightharpoons x_{\mu+1} - b$ und $y'_\mu \leftrightharpoons y_{\mu+1}$ wegen $ax'_\mu + by'_\mu = c$ und $x_2 - x_1 \geq b$, also $x'_\mu, y'_\mu \geq t$ für $\mu = 1,\ldots,m+1$, den Widerspruch $m \geq m+1$, womit (7) wegen (8) erwiesen ist.

Die eigentliche Behauptung für $k \in \mathbb{N}_0$ ergibt sich leicht mittels (7) durch Induktion nach k.

Für $k \in]-m,0[$ erhält man $m \geq 1-k$, nach Hilfssatz 1d also $c + kab > (t-1)s$, was wegen $-k \in \mathbb{N}$ nach dem schon Bewiesenen $m = N(c + kab,t) - k$ liefert.

Wäre schließlich $N(c + kab,t) \neq 0$ im Falle $k = -m < 0$, nach Hilfssatz 1d mithin $c + kab > (t-1)s$, so ergäbe sich nach dem Gezeigten $-k = m = N(c+kab,t) - k$, also der Widerspruch $N(c+kab,t) = 0$.

Es sei

$$S(n,k,t) \leftrightharpoons \sum_{\substack{\nu=ts \\ N(\nu,t)=k}}^{nab+(t-1)s} \nu \qquad ;$$

dann gilt:

Satz 3: a) $S(n,0,t) = \frac{p}{12}(2ab + (6t-1)s - 1)$,

b) $n > 1 \Rightarrow S(n,k,t) = \frac{ab}{2}(2kab - (2t-1)s)$ $(k=1,\ldots,n-1)$,

c) $S(n,n,t) = \frac{p}{12}(2(3n-1)ab + (6t-5)s + 1)$,

d) $k > n \Rightarrow S(n,k,t) = 0$.

Beweis:

d) Dies folgt unmittelbar aus Satz 2c.

a) Wegen Hilfssatz 1b ergibt sich $S(n,0,t) = \bar{\sigma}$, nach Satz 1b also das Gewünschte.

c) Mittels der Hilfssätze 1d und 3 sowie der Sätze 2a und 1a erhält man

$$S(n,n,t) = \sum_{\substack{\nu=(n-1)ab+ts \\ N(\nu-(n-1)ab,t)=1}}^{nab+(t-1)s} \nu = M(n,n,t)\cdot(n-1)ab + \sigma =$$

$$= \frac{p}{12}\,((6n-2)ab + (6t-5)s + 1).$$

d) Die Hilfssätze 1b und 1d liefern

(9) $$c \in\,]\,nab + (t-1)s,\ nab + ts\,[\implies N(c,t) = n.$$

Es sei $n > 1$. Nach (9) - mit $n-1$ statt n -, Hilfssatz 3 sowie Teil c), Satz 2a und Satz 1b gilt

$$S(n,n-1,t) = S(n-1,n-1,t) + \sum_{\nu=(n-1)ab+(t-1)s+1}^{(n-1)ab+ts-1} \nu + M(1,0,t)\cdot$$

$$\cdot(n-1)ab + \bar{\sigma} =$$

$$= \frac{p}{12}\,(12(n-1)ab + (12t-6)s) + \frac{s-1}{2}\,(2(n-1)ab + (2t-1)s)$$

und so

$$S(n,n-1,t) = \frac{ab}{2}\,(2(n-1)ab + (2t-1)s).$$

Unter Beachtung von Hilfssatz 1b erbringt Induktion nach n damit das Fehlende.

Literatur

1. A.H. Clifford, G.B. Preston: The Algebraic Theory of Semigroups I. Providence, R.I., 1977[3].

2. P. Erdös, R.L. Graham: On a linear diophantine problem of Frobenius. Acta Arithm., XXI (1972), S. 399 - 408.

3. W. Lex: Über Lösungsanzahlen linearer diophantischer Gleichungen. Math.-Phys. Semesterber., n.F., XXIV (1977), S. 254 - 279.

4. E.S. Selmer: On the linear diophantine problem of Frobenius. J. f. d. reine u. angew. Math., 293/294 (1977), S. 1 - 17.

5. C. Smoryński: Skolem's Solution to a Problem of Frobenius. The Math. Intel., 3 (1981), S. 123 - 132.

6. E. Teuffel: Über die Darstellung ganzer Zahlen in der Form ax + by. Unveröffentlichtes Manuskript.

7. I.M. Winogradow: Elemente der Zahlentheorie. München 1956[6].

Wilfried Lex
Institut für Informatik
der Technischen Universität Clausthal
Erzstr. 1
3392 Clausthal-Zellerfeld