Diskrete Mathematik

Götz Kersting, SS 2000

Die Diskrete Mathematik behandelt ‚diskrete', insbesondere endliche Objekte und Strukturen. Der Name betont den Gegensatz zu den kontinuierlichen Ansätzen und Methoden in der Mathematik. Die Grenzwertbetrachtungen der Analysis treten in den Hintergrund, wichtig sind Prozeduren, die nach endlich vielen Schritten zu einem Resultat führen, also **Algorithmen**. Die Diskrete Mathematik verdankt ihrem Aufschwung den Computerwissenschaften, für die sie eine wichtige Rolle spielt.

Literatur

M. Aigner (1994): Diskrete Mathematik, Vieweg

L. Childs (1979): A Concrete Introduction to Higher Algebra, Springer

T. Ihringer (1993): Diskrete Mathematik, Teubner

D.E. Knuth (1973,1981): The Art of Computer Programming, Vol. I,II, Addison-Wesley

N. Koblitz (1994): A Course in Number Theory and Cryptography, Springer

J.H. van Lint (1992): Introduction to Coding Theory, Springer

C. Papadimitriou (1994): Computational Complexity, Addison Wesley

Inhaltsverzeichnis

Kapitel 1

Der Euklidische Algorithmus

1.1 Größte gemeinsame Teiler und der Euklidische Algorithmus

Der Euklidische Algorithmus gehört zu den ältesten Rechenverfahren, er war schon Eudoxus (375 v.Chr.) bekannt. Er ist von fundamentaler Bedeutung und kommt in vielen Rechenprozeduren zur Anwendung. Der Algorithmus dient zur Bestimmung von größten gemeinsamen Teilern. Wir betrachten zunächst den Ring $\mathbb{Z}$ der ganzen Zahlen. Zur Notation: Wir schreiben $a \mid b$ für ganze Zahlen a, b, falls a Teiler von b ist, d.h., falls es ein ganze Zahl c gibt, so daß $ac = b$.

Definition 1.1. *$d \in \mathbb{Z}$ heißt **größter gemeinsamer Teiler**, kurz* ggT *der ganzen Zahlen $a_1, \dots, a_n \neq 0$, falls*

i) $d \mid a_1, \dots, d \mid a_n$,

ii) $z \mid a_1, \dots, z \mid a_n \quad \Rightarrow \quad z \mid d \quad$ *für alle* $z \in \mathbb{Z}$.

*Wir schreiben dann $d = \mathrm{ggT}(a_1, \dots, a_n)$. Gilt $1 = \mathrm{ggT}(a_1, \dots, a_n)$, so sagt man, $a_1, \dots, a_n$ sind **relativ prim** oder **teilerfremd**.*

Aus $z \mid d$ folgt $|z| \leq |d|$ (wegen $a_1, \dots, a_n \neq 0$ ist $d = 0$ ausgeschlossen). In diesem Sinne ist d am größten unter allen Teilern von $a_1, \dots, a_n$. Ist d' ein weiterer ggT für $a_1, \dots, a_n$, so teilen sich d und d' gegenseitig, so daß $|d'| = |d|$, also $d' = \pm d$ folgt. Der ggT von $a_1, \dots, a_n$ ist daher bis auf das Vorzeichen eindeutig bestimmt.

Weniger evident ist, daß immer ein größter gemeinsamer Teiler existiert – etwa für zwei natürliche Zahlen a und b. Eine Möglichkeit besteht darin, a und b in Primfaktoren zu zerlegen. Seien $p_1, \dots, p_r$ die Primzahlen, die in a oder b

als Teiler enthalten sind. Dann gibt es ganze Zahlen $e_1, \ldots, e_r, f_1, \ldots, f_r \geq 0$ so daß $a = p_1^{e_1} \ldots p_r^{e_r}$ und $b = p_1^{f_1} \ldots p_r^{f_r}$ ($e_i = 0$ bedeutet, daß p_i kein Teiler von a ist). Die gemeinsamen Teiler von a und b sind dann von der Gestalt $z = \pm p_1^{g_1} \ldots p_r^{g_r}$ mit $g_i \in \{0, 1, \ldots, m_i\}$, $m_i = \min(e_i, f_i)$, und die größten gemeinsamen Teiler von a und b sind $d = \pm p_1^{m_1} \ldots p_r^{m_r}$. Diese Überlegung macht davon Gebrauch, daß man ganze Zahlen in eindeutiger Weise in Primfaktoren zerlegen kann, darauf kommen wir in Kapitel 5 zu sprechen. Vom Standpunkt des Rechnens ist dieses Vorgehen nicht befriedigend, denn die Zerlegung einer Zahl in ihre Primfaktoren ist sehr rechenaufwendig.

Wir gehen hier anders vor und klären die Existenzfrage, indem wir ein Rechenverfahren angeben, daß größte gemeinsame Teiler liefert. Es beruht auf einer grundlegenden Eigenschaft ganzer Zahlen, der **Division mit Rest**: Zu ganzen Zahlen $a, b \neq 0$ gibt es ganze Zahlen m, r, so daß

$$a \;=\; mb + r \;\text{ und }\; 0 \;\leq\; r \;<\; |b|.$$

Die Idee des Euklidischen Algorithmus ist es, aus zwei Zahlen den ggT schrittweise herauszudividieren.

Euklidischer Algorithmus.

Eingabe: ganze Zahlen $a, b \neq 0$.

Ausgabe: $r_{j-1} = \mathrm{ggT}(a, b)$.

Verfahren: Setze $r_{-1} = a, r_0 = b$. Bestimme durch Division mit Rest sukzessive ganze Zahlen $r_1, \ldots, r_{j-1}$ mit $|b| > r_1 > \ldots > r_{j-1} > r_j = 0$, bis kein Rest mehr bleibt. r_{i+1} sei also der Rest, der bei Division von r_{i-1} durch r_i entsteht:

$$\begin{aligned}
r_{-1} &= m_1 r_0 + r_1, \\
r_0 &= m_2 r_1 + r_2, \\
&\;\vdots \\
r_{i-1} &= m_{i+1} r_i + r_{i+1}, \\
&\;\vdots \\
r_{j-3} &= m_{j-1} r_{j-2} + r_{j-1}, \\
r_{j-2} &= m_j r_{j-1},
\end{aligned}$$

mit ganzen Zahlen $m_1, \ldots, m_j$. □

Da die Divisionsreste r_i strikt fallen, bricht das Verfahren nach endlich vielen Schritten ab. Es ist korrekt: r_{j-1} teilt der Reihe nach $r_{j-2}, r_{j-3}, \ldots, r_0 = b$ und $r_{-1} = a$, wie sich sukzessive aus den Gleichungen $r_{i-1} = m_{i+1} r_i + r_{i+1}$ ergibt. Teilt umgekehrt z sowohl a wie b, so teilt z der Reihe nach $r_1, \ldots, r_{j-1}$, wie aus den Gleichungen $r_{i+1} = r_{i-1} - m_{i+1} r_i$ folgt.

Beispiel. Für $a = 1736$, $b = 1484$ ergibt das Verfahren

$$\begin{aligned} 1736 &= 1 \cdot 1484 + 252 \\ 1484 &= 5 \cdot 252 + 224 \\ 252 &= 1 \cdot 224 + 28 \\ 224 &= 8 \cdot 28\ (+0). \end{aligned}$$

Also

$$28 = \mathrm{ggT}(1736, 1484).$$

Liest man die Gleichungen von unten nach oben, so erkennt man: $28 \mid 224$, $28 \mid 252$, $28 \mid 1484$, $28 \mid 1736$. Liest man von oben nach unten, so ist klar, daß jeder Teiler von 1736 und 1484 auch 28 teilt. □

Aus der Gleichung $r_{j-3} = m_{j-1}r_{j-2} + r_{j-1}$ läßt sich r_{j-2} mittels $r_{j-2} = r_{j-4} - m_{j-2}r_{j-3}$ eliminieren. Genauso lassen sich $r_{j-3}, \ldots, r_1$ eliminieren, und wir erhalten r_{j-1} als ganzzahlige Linearkombination von a und b. Im Beispiel sieht das so aus:

$$28 = 252 - 224 = 6 \cdot 252 - 1484 = 6 \cdot 1736 - 7 \cdot 1484.$$

Damit haben wir im Fall $n = 2$ den **Satz von Bézout** bewiesen.

Satz 1.2. *Beliebige ganze Zahlen $a_1, \ldots, a_n$ besitzen einen größten gemeinsamen Teiler d, und es gibt ganze Zahlen $z_1, \ldots, z_n$, so daß*

$$d = z_1 a_1 + \ldots + z_n a_n.$$

Beweis. Der Fall $n = 2$ ist schon geklärt. Den Fall $n > 2$ kann man induktiv behandeln: Sei $d' = z'_1 a_1 + \ldots + z'_{n-1} a_{n-1}$ ggT von $a_1, \ldots, a_{n-1}$ und $d = sd' + ta_n$ ggT von d' und a_n. Dann ist d ggT von $a_1, \ldots, a_n$ und als ganzzahlige lineare Kombination von $a_1, \ldots, a_n$ darstellbar. □

Durch Erweiterung des Euklidischen Algorithmus kann man gleichzeitig mit dem ggT zweier Zahlen a und b auch seine Darstellung nach dem Satz von Bézout gewinnen. Man bestimme dazu ganze Zahlen $s_{-1}, \ldots, s_{j-1}, t_{-1}, \ldots, t_{j-1}$ rekursiv aus den Gleichungen

$$\begin{aligned} s_{-1} &= 1,\ s_0 = 0, & t_{-1} &= 0,\ t_0 = 1, \\ s_{i-1} &= m_{i+1}s_i + s_{i+1}, & t_{i-1} &= m_{i+1}t_i + t_{i+1}, \end{aligned}$$

unter Benutzung der vom Euklidischen Algorithmus gewonnenen ganzen Zahlen $m_1, \ldots, m_j$. Dann gilt

$$\mathrm{ggT}(a, b) = r_{j-1} = as_{j-1} + bt_{j-1}.$$

Beweis. Es gilt sogar $r_i = as_i + bt_i$ für alle $-1 \leq i < j$, wie sich per Induktion nach i ergibt: Für $i = -1, 0$ folgt dies nach Wahl von s_{-1}, s_0, t_{-1} und t_0, und der Induktionsschritt folgt aus

$$\begin{aligned} r_{i+1} &= r_{i-1} - m_{i+1} r_i \\ &= as_{i-1} + bt_{i-1} - m_{i+1}(as_i + bt_i) \\ &= as_{i+1} + bt_{i+1}. \end{aligned}$$

Beispiel. a=1736, b=1484.

i	r_{i-1}	r_i	m_{i+1}	s_{i-1}	s_i	t_{i-1}	t_i
0	1736	1484	1	1	0	0	1
1	1484	252	5	0	1	1	-1
2	252	224	1	1	-5	-1	6
3	224	$\boxed{28}$	8	-5	$\boxed{6}$	6	$\boxed{-7}$

Also

$$28 = 6 \cdot 1736 - 7 \cdot 1484.$$

□

1.2 Laufzeit des Euklidischen Algorithmus

Der Euklidische Algorithmus führt schnell zum Resultat, die Anzahl der benötigten Divisionen hat eine Größe, die nur logarithmisch von den Eingabedaten a und b abhängt. Dies zeigt eine *worst-case* Analyse des Algorithmus. Wir schätzen die Laufzeit des Algorithmus ab, indem wir Eingaben a, b betrachten, für die besonders viele Divisionen anfallen. Ohne Einschränkung sei $a > b > 0$ an. (Im Fall $|b| > |a|$ werden im ersten Schritt a und b vertauscht.) Das kleinste Paar a, b, für das j Divisionen benötigt werden, ergibt sich, indem man r_{j-1} und die m_i möglichst klein wählt: $m_1 = \ldots = m_{j-1} = 1$, $m_j = 2$ und $r_{j-1} = 1$ ($m_j = 1$ ist ausgeschlossen). Die Gleichungen

$$r_{j-1} = 1, \quad r_{j-2} = 2, \quad r_{i-1} = r_i + r_{i+1} \text{ für } i = j-2, \ldots, 0$$

bestimmen dann $r_{-1} = a$ und $r_0 = b$ eindeutig. Es ergibt sich ein Zusammenhang zur Folge der Fibonacci-Zahlen 0,1,1,2,3,5,8,13,...

Definition 1.3. *Die **Fibonacci-Zahlen** $F_0, F_1, \ldots$ sind rekursiv definiert durch $F_0 = 0, F_1 = 1$ und $F_i = F_{i-1} + F_{i-2}$ für $i \geq 2$.*

Ein Vergleich zeigt $r_{j-1} = F_2, r_{j-2} = F_3, \ldots, r_0 = F_{j+1}, r_{-1} = F_{j+2}$. Bei der Eingabe $a = F_{j+2}, b = F_{j+1}$ benötigt der Euklidische Algorithmus also genau j Divisionen. Gilt allgemeiner $0 < b < a$ und $b \leq F_{j+1}$, so zeigt unsere Überlegung, daß der Algorithmus höchstens j Divisionen benötigt.

Für die Fibonacci-Zahlen gilt die bemerkenswerte Formel

$$F_i \;=\; \frac{1}{\sqrt{5}}\Big[\Big(\frac{1+\sqrt{5}}{2}\Big)^i - \Big(\frac{1-\sqrt{5}}{2}\Big)^i\Big].$$

Der Beweis ergibt sich leicht per Induktion. Die Fälle $i = 0$ und 1 prüft man direkt nach. Außerdem erfüllt der angegebene Ausdruck für F_i die rekursiven Gleichungen der Fibonacci-Zahlen. Der Grund ist, daß die Zahlen

$$\phi_1 \;:=\; \frac{1+\sqrt{5}}{2}, \quad \phi_2 \;:=\; \frac{1-\sqrt{5}}{2}$$

die Gleichung $\phi^2 = \phi + 1$ und folglich $\phi^i = \phi^{i-1} + \phi^{i-2}$ erfüllen.

Satz 1.4. *Der Euklidische Algorithmus benötigt bei der Eingabe $a > b > 0$ höchstens $c \ln(b\sqrt{5})$ Divisionen, mit $c = (\ln \frac{1+\sqrt{5}}{2})^{-1} \approx 2,08$.*

Beweis. Wähle $j \in \mathbb{N}$ so, daß $F_j < b \leq F_{j+1}$. Dann sind höchstens j Divisionen durchzuführen. Aus der Darstellung für F_j folgt

$$\frac{1}{\sqrt{5}}\Big(\frac{1+\sqrt{5}}{2}\Big)^j \;\leq\; F_j + 1 \;\leq\; b$$

(es gilt $\phi_2 \approx -0,618$) bzw.

$$j \,\ln\frac{1+\sqrt{5}}{2} \;\leq\; \ln(b\sqrt{5}).$$

□

1.3 Ein binärer Algorithmus

Für den Computer ist Euklids Algorithmus nicht das günstigste Verfahren. Divisionen mit Rest sind relativ rechenaufwendig. Der folgende Algorithmus benötigt nur Divisionen durch 2. Sie sind auf einem Rechner besonders schnell zu realisieren, da er Zahlen in Dualdarstellung verarbeitet. Die Idee ist, a und b so zu verkleinern, daß sich der ggT in kontrollierbarer Weise verändert. Im ersten Schritt nimmt man aus a und b den gemeinsamen geraden Anteil hinaus, gemäß der Regel

$$a, b \text{ gerade},\ d = \mathrm{ggT}(a,b) \quad\Rightarrow\quad d = 2\cdot \mathrm{ggT}(a/2, b/2).$$

Anschließend verkleinert man a und b schrittweise, ohne den ggT noch zu verändern. Ist a gerade und b ungerade, so ersetzt man a durch $a/2$ und läßt b unverändert. Der ggT ändert sich nicht, denn

$$a \text{ gerade, } b \text{ ungerade, } d = \mathrm{ggT}(a,b) \quad \Rightarrow \quad d = \mathrm{ggT}(a/2, b).$$

Der Fall a ungerade, b gerade ist analog. Sind a und b beide ungerade, so benutzen wir die Regel

$$d = \mathrm{ggT}(a,b) \quad \Rightarrow \quad d = \mathrm{ggT}(a-b, b).$$

Wir ziehen dann die kleinere der Zahlen von der größeren ab. Diese Regeln kommen abwechselnd zu Zuge, wie das folgende Beispiel zeigt.

$$\begin{aligned}
\mathrm{ggT}(1736, 1484) &= 4 \cdot \mathrm{ggT}(434, 371) = 4 \cdot \mathrm{ggT}(217, 371) \\
&= 4 \cdot \mathrm{ggT}(217, 154) = 4 \cdot \mathrm{ggT}(217, 77) = 4 \cdot \mathrm{ggT}(140, 77) \\
&= 4 \cdot \mathrm{ggT}(35, 77) = 4 \cdot \mathrm{ggT}(35, 42) = 4 \cdot \mathrm{ggT}(35, 21) \\
&= 4 \cdot \mathrm{ggT}(14, 21) = 4 \cdot \mathrm{ggT}(7, 21) = 4 \cdot \mathrm{ggT}(7, 14) \\
&= 4 \cdot \mathrm{ggT}(7, 7) = 28.
\end{aligned}$$

Das Verfahren läßt sich leicht auf dem Computer implementieren.

1.4 Der Euklidische Algorithmus für Polynome

Der binäre Algorithmus ist an die Computerarithmetik angepaßt und benutzt spezielle Eigenschaften der ganzen Zahlen. Die Stärke des Euklidischen Algorithmus besteht darin, daß er sich auch in anderen Rechenbereichen als dem der ganzen Zahlen anwenden läßt. Wir betrachten nun **Polynome** in der Variablen x mit rationalen Koeffizienten (oder allgemeiner mit Koeffizienten in einem Körper K). Ein rationales Polynom $f = f(x)$ ist gegeben durch einen Ausdruck der Gestalt

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0$$

mit $n \in \mathbb{N}$ und $a_i \in \mathbb{Q}$, $i = 0, \ldots, n$. Glieder $a_i x^i$ mit dem Koeffizienten $a_i = 0$ dürfen aus der Summe weggelassen bzw. zur Summe beliebig hinzugefügt werden. Wir betrachten also zwei Polynome als identisch, falls sie dieselben Glieder haben, abgesehen von Summanden mit dem Koeffizienten 0.

Bemerkung. Es ist hier weder nötig noch angemessen, sich Polynome wie in der Analysis als Funktionen vorzustellen. Wir fassen Polynome als formale Ausdrücke auf. Dieser Unterschied in der Auffassung macht sich zwar für rationale Polynome nicht bemerkbar, sehr wohl jedoch für Polynome mit Koeffizienten in anderen Körpern. So gilt (wie wir später genauer erläutern) $f(0) = f(1) = 0$ für das Polynom $f(x) = x(x+1) = x^2 + x$ mit Koeffizienten aus dem Körper $\mathbb{Z}_2 = \{0, 1\}$, es nimmt nur den Wert 0 an. Dennoch erweist es sich nicht als sinnvoll, dieses Polynom mit dem Nullpolynom gleichzusetzen. □

Mit Polynomen kann man rechnen wie mit ganzen Zahlen. Zwei Polynome $f(x) = a_n x^n + \ldots + a_0$ und $g(x) = b_m x^m + \ldots + b_0$ lassen sich addieren und multiplizieren:

$$(f+g)(x) \;:=\; (a_n + b_n)x^n + \ldots + (a_0 + b_0)$$

(o.E.d.A. können wir $m = n$ annehmen) und

$$(fg)(x) \;:=\; c_{m+n}x^{m+n} + \ldots + c_0, \qquad \text{mit } c_i \;:=\; \sum_{j+k=i} a_j b_k.$$

Es bereitet keine Mühe, die bekannten Rechenregeln für Zahlen auf Polynome zu übertragen, wie Assoziativität, Kommutativität, Distributivität. Die Rolle der Zahl 0 übernimmt das **Nullpolynom**, dessen Koeffizienten alle verschwinden. Insgesamt haben wir den Ring $\mathbb{Q}[x]$ der rationalen Polynome konstruiert (zur Definition eines Ringes vergleiche den folgenden Abschnitt).

Um eine Division mit Rest für Polynome einzuführen, definieren wir den **Grad** $\deg(f)$ eines vom Nullpolynom verschiedenen Polynoms $f(x) = a_n x^n + \ldots + a_0$. Er ist die größte Zahl i, so daß $a_i \neq 0$. Das zugehörige a_i bezeichnen wir als den **Anfangskoeffizienten** oder **höchsten Koeffizienten** von f.

Seien nun $f, g \neq 0$ Polynome vom Grade n und m, mit Anfangskoeffizienten a_n und b_m. Falls $n \geq m$, können wir g aus f hinausdividieren, also das Polynom

$$h(x) \;=\; f(x) - a_n b_m^{-1} x^{n-m} g(x)$$

bilden. Offenbar hat h einen kleineren Grad als f. Gilt $\deg(h) \geq m$, so kann g aus h ein weiteres Mal hinausdividiert werden. Dies läßt sich fortsetzen, bis ein Polynom r vom Grade kleiner m oder aber das Nullpolynom übrigbleibt. Wir erhalten also für Polynome wie für die ganzen Zahlen eine **Division mit Rest**: Zu rationalen Polynomen $f(x), g(x) \neq 0$ existieren rationale Polynome $m(x), r(x)$, so daß

$$f(x) \;=\; m(x)g(x) + r(x), \text{ mit } \deg(r) < \deg(g) \text{ oder } r(x) = 0.$$

Der Euklidische Algorithmus läßt sich nun auch auf rationale Polynome anwenden. Da sich der Grad der Restpolynome bei jeder Division verkleinert, bricht er nach endlich vielen Schritten ab, seine Laufzeit ist durch den Grad des Polynoms g beschränkt. Wie für ganze Zahlen können wir also feststellen: Zwei rationale Polynome $f(x), g(x) \neq 0$ besitzen einen ggT $d(x)$, und es gibt rationale Polynome $s(x)$ und $t(x)$, so daß

$$d(x) \;=\; s(x)f(x) + t(x)g(x).$$

Beispiel. Seien $a \geq b \geq 0$ natürliche Zahlen. Wir wollen den ggT von $x^a - 1$ und $x^b - 1$ berechnen. Sei $a = mb + r$ die Division von a durch b mit Rest r. Für die Division von $x^a - 1$ durch $x^b - 1$ erhalten wir dann

$$x^a - 1 \;=\; (x^{a-b} + x^{a-2b} + \ldots + x^{a-mb})(x^b - 1) + x^r - 1.$$

Der Rest ist $x^r - 1$, er verschwindet genau dann, wenn $r = 0$ ist, und ist von demselben Typ wie die beiden Polynome, von denen wir ausgegangen sind. Der Euklidische Algorithmus, angewandt auf $x^a - 1$ und $x^b - 1$, läuft daher völlig parallel ab zum Euklidischen Algorithmus, angewandt auf a und b. Insbesondere folgt

$$x^{\mathrm{ggT}(a,b)} - 1 \;=\; \mathrm{ggT}(x^a - 1, x^b - 1).$$

□

Diese Überlegungen gelten nicht nur für rationale Polynome, die Koeffizienten können genauso gut reelle oder komplexe Zahlen sein. Für die Division mit Rest langt es, daß die Koeffizienten einem Körper K angehören, denn in Körpern kann man wie in den reellen Zahlen addieren, multiplizieren und dividieren (die Definition eines Körpers wiederholen wir im nächsten Abschnitt). Später werden insbesondere endliche Körper wichtig.

Dagegen hat man im Ring $\mathbb{Z}[x]$ aller Polynome $f(x) = a_n x^n + \ldots + a_0$, deren Koeffizienten a_i ganze Zahlen sind, im allgemeinen keine Division mit Rest, denn der Kehrwert b_m^{-1} des Anfangskoeffizienten von $g(x)$ läßt sich in den ganzen Zahlen nur bilden, falls $b_m = \pm 1$ ist.

1.5 Euklidische Ringe

Eine Division mit Rest hat man nicht nur für die ganzen Zahlen und für Polynome. Allgemein nennt man die Bereiche, in denen eine Division mit Rest möglich ist, Euklidische Ringe. Die Terminologie stammt aus der Algebra, wir wollen sie abschließend kurz rekapitulieren.

Sei G eine Menge, die mit einer binären Verknüpfung versehen ist. Je zwei Elementen a, b aus G soll also ein Element ab (ihr ‚Produkt') zugeordnet sein. Gilt Assoziativität $(ab)c = a(bc)$, so heißt G (zusammen mit der Verknüpfung) eine **Halbgruppe**. Ein **neutrales Element** $e \in G$ ist durch die Eigenschaft $a = ea = ae$ für alle $a \in G$ ausgezeichnet. G kann höchstens ein neutrales Element besitzen, denn für neutrale Elemente e, e' folgt $e = ee' = e'$. Gibt es zu $a \in G$ ein $b \in G$, so daß $ab = ba = e$, so heißt b das **inverse Element** von a. Ist auch b' invers zu a, so folgt $b' = b'(ab) = (b'a)b = b$, das inverse Element ist also eindeutig. Man schreibt $b = a^{-1}$. Enthält eine Halbgruppe G ein neutrales Element, und besitzt jedes $a \in G$ ein inverses Element, so nennt man G eine **Gruppe**. Es gilt dann $(ab)^{-1} = b^{-1}a^{-1}$. Ist in G zusätzlich noch Kommutativität $ab = ba$ für alle a, b gegeben, so heißt G eine **abelsche Gruppe**.

Sei nun R eine Menge, die wie die ganzen Zahlen mit zwei binären Operationen $a + b$ und ab (‚Addition' und ‚Multiplikation') versehen ist. Dann heißt R (zusammen mit $+$ und $\cdot$) ein **Ring**, falls gilt:

i) R ist in Bezug auf die Addition eine abelsche Gruppe, mit neutralem Element 0. Jedes $a \in R$ besitzt also ein ‚entgegengesetztes' Element $-a$, so daß $a + (-a) = 0$ (das inverse Element bzgl. der Addition).

ii) R ist in Bezug auf die Multiplikation eine Halbgruppe.

iii) Es gilt Distributivität: $a(b + c) = ab + ac$, $(a + b)c = ac + bc$.

Wie üblich schreiben wir $a - b$ für $a + (-b)$ und $a \mid b$, wenn es ein $c \in R$ gibt, so daß $ac = b$. Es gilt

$$a \cdot 0 = 0 \cdot a = 0$$

(beachte $a \cdot 0 + a \cdot a = a(0 + a) = a \cdot a$), ein Produkt ist also 0, falls ein Faktor 0 ist. Umgekehrt gilt dies im allgemeinen nicht. Wir werden später Ringe kennenlernen, in denen $ab = 0$ gelten kann für zwei Elemente $a, b \neq 0$. a und b heißen dann **Nullteiler**. Ringe ohne Nullteiler haben den Vorteil, daß man in ihnen kürzen darf: Aus $ac = bc$ folgt $(a - b)c = 0$ und damit $a = b$, sofern c kein Nullteiler ist.

In manchen Ringen ist das Kommutativitätsgesetz verletzt. Ist die Multiplikation kommutativ, $ab = ba$, so spricht man von einem **kommutativen Ring**. Ein kommutativer Ring ohne Nullteiler heißt **Integritätsbereich**.

Wir werden uns nur mit Ringen befassen, die ein Einselement enthalten. Ein **Einselement** $1 \in R - \{0\}$ in einem Ring R ist ein neutrales Element bzgl. der Multiplikation in $R - \{0\}$, es erfüllt also $1a = a1 = a$ für alle $a \neq 0$. Diejenigen $a \in R - \{0\}$, die bzgl. der Multiplikation ein Inverses a^{-1} besitzen, heißen die **Einheiten** des Ringes. Die Menge R^* aller Einheiten ist eine Gruppe. Für $a, b \in R^*$ ist nämlich $b^{-1}a^{-1}$ invers zu ab, so daß $ab \in R^*$. Gilt für einen kommutativen Ring R mit Einselement $R^* = R - \{0\}$, besitzt also jedes $a \neq 0$ ein Inverses a^{-1}, so nennt man R einen **Körper**.

Beispiel. Sei R ein Ring. Dann ist auch die Menge aller Polynome $f(x) = a_n x^n + \ldots + a_1 x + a_0$ mit $a_i \in R$ ein Ring, der **Polynomring** $R[x]$. Zwei Polynome gelten als gleich, wenn sie sich nur um Summanden unterscheiden, deren Koeffizienten 0 sind. Addition und Multiplikation werden wie in Abschnitt 1.4 definiert. R ist in $R[x]$ durch die Polynome vom Grad 0 eingebettet. Ein Einselement in R ist auch Einselement in $R[x]$, und es gilt $R[x]^* = R^*$. Man sagt, $R[x]$ entsteht aus R *durch Adjunktion der Unbestimmten* x. □

Definition 1.5. *Ein Integritätsbereich R heißt **Euklidischer Ring**, falls jedem $a \in R - \{0\}$ eine nicht-negative ganze Zahl $g(a)$ zugeordnet ist, und falls zu beliebigen $a, b \in R - \{0\}$ Ringelemente m, r existieren, so daß gilt: $a = mb + r$ und entweder $r = 0$ oder $g(r) < g(b)$.*

Unsere Überlegungen über größte gemeinsame Teiler übertragen sich vollständig auf Euklidische Ringe. Man hat erneut den Euklidischen Algorithmus zur Verfügung, daher existieren größte gemeinsame Teiler im Sinne der Definition 1.1, und es gilt der Satz von Bézout: Für $a, b \neq 0$ und $d = \mathrm{ggT}(a, b)$ gibt es $s, t \in R$, so daß

$$d = sa + tb.$$

Außerdem enthält jeder Euklidische Ring R ein Einselement. Um es zu erhalten, wählen wir ein $b \in R - \{0\}$, für das $g(b)$ minimal ist. Wenn wir $a \in R$ durch b mit Rest teilen, folgt $a = mb$ für ein $m \in R$, denn wegen der Minimalität von $g(b)$ verschwindet bei der Division der Rest. Insbesondere können wir b durch sich selbst teilen: $b = be$ für ein $e \neq 0$. e ist das gesuchte Einselement, denn es folgt $a = mb = mbe = ae$. Man überzeugt sich leicht, daß größte gemeinsame Teiler bis auf Einheiten eindeutig bestimmt ist.

Beispiele.

1) $\mathbb{Z}$ ist ein Euklidischer Ring, mit $g(a) := |a|$.

2) Sei K ein Körper (etwa $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ oder $\mathbb{Z}_2 = \{0, 1\}$). Dann ist der Polynomring $K[x]$ ein Euklidischer Ring, mit $g(f) := \deg(f)$ (vergleiche Abschnitt 1.4).

3) Die *ganzen Gaußschen Zahlen*, also die komplexen Zahlen $z = x + iy$ mit $x, y \in \mathbb{Z}$, bilden bzgl. der komplexen Addition und Multiplikation einen Euklidischen Ring, mit $g(z) := |z|^2 = x^2 + y^2$ (Übung). □

Kapitel 2

Kettenbrüche

2.1 Kettenbruchdarstellung rationaler Zahlen

Der Euklidische Algorithmus ist nicht nur ein effizientes Verfahren zur Berechnung größter gemeinsamer Teiler. Man kann ihn gleichzeitig dazu benutzen, rationale Zahlen als Kettenbrüche auszudrücken. Die Divisionen $r_{i-1} = m_{i+1}r_i + r_{i+1}$ des Algorithmus schreiben wir dazu um zu

$$\frac{r_{i-1}}{r_i} = m_{i+1} + \left(\frac{r_i}{r_{i+1}}\right)^{-1}. \tag{2.1}$$

Diese Gleichungen kann man ineinander einsetzen. Ausgehend von $r_{-1}/r_0 = a/b$ entsteht so die Kettenbruchdarstellung

$$\frac{a}{b} = m_1 + \left(\frac{r_0}{r_1}\right)^{-1} = m_1 + \cfrac{1}{m_2 + \left(\cfrac{r_1}{r_2}\right)^{-1}} = \ldots$$

Zum Beispiel ist

$$\frac{1736}{1484} = 1 + \cfrac{1}{5 + \cfrac{1}{1 + \cfrac{1}{8}}}$$

Allgemein schreiben wir

$$\frac{a}{b} = m_1 + \frac{1}{m_2+}\,\frac{1}{m_3+}\cdots\frac{1}{m_{j-1}+}\,\frac{1}{m_j}, \tag{2.2}$$

dabei benutzen wir die Schreibweise

$$\begin{aligned} & m_1 + \frac{1}{m_2+}\,\frac{1}{m_3+}\cdots\frac{1}{m_{j-1}+}\,\frac{1}{m_j} \\ := \;& m_1 + \cfrac{1}{m_2 + \cfrac{1}{m_3 + \cfrac{1}{\ddots \, \cfrac{}{m_{j-1} + \cfrac{1}{m_j}}}}} \end{aligned}$$

Für die Berechnung der Kettenbrüche braucht man nicht auf den Euklidischen Algorithmus zurückzugreifen. Wir formen (2.1) um zu der Gleichung

$$\alpha_i \;=\; m_{i+1} + \alpha_{i+1}^{-1},\; i = 0, \ldots, j-2, \quad \text{mit } \alpha_i := \frac{r_{i-1}}{r_i}.$$

Da die Reste r_i strikt fallen, gilt $\alpha_{i+1} > 1$ für $i \geq 0$, und es folgt

$$m_{i+1} \;=\; [\alpha_i], \quad \alpha_{i+1} \;=\; (\alpha_i - m_{i+1})^{-1}.$$

Ausgehend vom Startwert $\alpha_0 = a/b$ kann man so $m_1, m_2, \ldots$ rekursiv berechnen. Diese Prozedur heißt **Kettenbruchalgorithmus**.

2.2 Näherungsbrüche reeller Zahlen

Auch irrationale Zahlen lassen sich in Kettenbrüche entwickeln. In diesem Abschnitt lernen wir den Kettenbruchalgorithmus als Verfahren kennen, das für reelle Zahlen α besonders gute rationale Näherungen a/b ermittelt. Das folgende Beispiel zeigt, daß es dabei für die Approximationsgüte Grenzen gibt.

Beispiel. Sei a/b eine rationale Näherung von $\sqrt{2}$, deren Abweichung geringer als $1/b$ ist: $\sqrt{2} - 1/b \leq a/b \leq \sqrt{2} + 1/b$. Da $\sqrt{2}$ bekanntlich eine Irrationalzahl ist, gilt

$$|\sqrt{2} - a/b| \cdot |\sqrt{2} + a/b| \;=\; |2 - (a/b)^2| \;=\; |2b^2 - a^2|/b^2 \;\geq\; 1/b^2.$$

Mit $\sqrt{2} + a/b \leq 2\sqrt{2} + 1/b \leq 4$ folgt

$$|\sqrt{2} - a/b| \;\geq\; 1/(4b^2). \qquad \square$$

Bemerkenswerterweise läßt sich die durch das Beispiel nahegelegte Approximationsgüte immer realisieren.

Proposition 2.1. *Sei α eine irrationale Zahl. Dann gibt es unendlich viele rationale Zahlen a/b, so daß $|\alpha - a/b| < 1/b^2$.*

Beweis. Wähle $n \in \mathbb{N}$. Die $n+1$ Zahlen $i\alpha - [i\alpha], i = 0, \ldots, n$, liegen alle im Intervall $[0,1) = \bigcup_{k=1}^{n} [(k-1)/n, k/n)$. Eines der Intervalle $[(k-1)/n, k/n)$ muß daher zwei der Zahlen enthalten (‚Dirichletsches Schubfachprinzip'), also gibt es $0 \leq i < j \leq n$ mit

$$|j\alpha - i\alpha - [j\alpha] + [i\alpha]| \ < \ 1/n.$$

Setzen wir $a = [j\alpha] - [i\alpha], b = j - i \leq n$, so folgt

$$|\alpha - a/b| \ < \ 1/(bn) \ \leq \ 1/b^2.$$

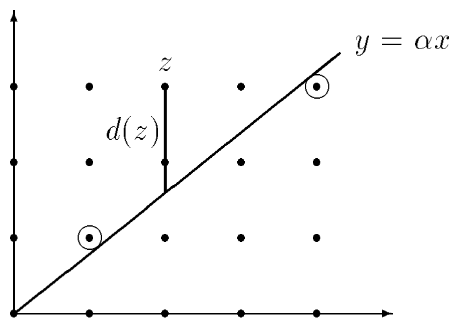
Mit wachsendem n erhält man immer bessere Approximationen. Daher gibt es eine unendliche Folge von Näherungen der behaupteten Güte. $\square$

Dies ist ein charmanter Beweis mit einem gravierenden Mangel: Es handelt sich um einen reinen Existenzbeweis, er gibt keinen Anhaltspunkt, wie die guten Näherungsbrüche effizient gefunden werden können.

Dies leistet nun der Kettenbruchalgorithmus. Wir leiten ihn aus einer geometrischen Betrachtung ab: Gesucht werden Punkte $z = (b, a)$ in der Ebene mit ganzzahligen Koordinaten b und a, die besonders nahe an der durch die Gleichung $y = \alpha x$ gegebenen Geraden liegen, so daß also

$$d(z) \ := \ |a - \alpha b|$$

besonders klein wird (bzw. $d(z)/\sqrt{1+\alpha^2}$, der Abstand zwischen (b, a) und der Geraden).



Ein systematisches Verfahren zur Konstruktion besserer Näherungspunkte aus zwei vorgegebenen Punkten liefert die folgende geometrische Variante einer Division mit Rest. Ohne Einschränkung der Allgemeinheit sei von nun an $\alpha > 0$.

Konstruktion. *$z_1 = (b_1, a_1)$, $z_2 = (b_2, a_2)$ seien zwei Punkte in $\mathbb{R}^2_+$, die nicht auf derselben Seite der Geraden $y = \alpha x$ liegen, also*

$$\frac{a_1}{b_1} < \alpha < \frac{a_2}{b_2} \quad \textit{oder} \quad \frac{a_1}{b_1} > \alpha > \frac{a_2}{b_2}.$$

Setze $z_3 = z_1 + mz_2$, wobei m die größte ganze Zahl sei, so daß z_3 noch auf derselben Seite der Geraden liegt wie z_1. Dann gilt

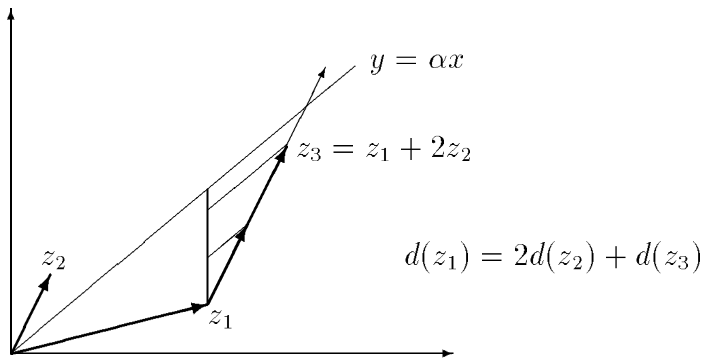
$$d(z_1) \;=\; md(z_2) + d(z_3), \quad 0 \leq d(z_3) < d(z_2).$$

Im Falle $d(z_1) \geq d(z_2)$, also $m \geq 1$, gilt

$$\frac{a_1}{b_1} < \frac{a_3}{b_3} \leq \alpha \quad \textit{bzw.} \quad \frac{a_1}{b_1} > \frac{a_3}{b_3} \geq \alpha.$$

□

$d(z_3)$ berechnet sich aus $d(z_1)$ und $d(z_2)$ nach demselben Schema wie bei der Division mit Rest für die ganzen Zahlen. Die beiden letzten Ungleichungen ergeben sich daraus, daß die Steigung des Vektors z_3 zwischen α und der Steigung von z_1 liegt. Das folgende Bild verdeutlicht die Situation im Fall $a_1/b_1 < \alpha < a_2/b_2$.



Durch wiederholte Anwendung der Konstruktion entsteht aus den Startwerten $(1, 0)$ und $(0, 1)$ eine ganze Folge von Näherungen, die **Näherungsbrüche** von α.

Konstruktion der Näherungsbrüche.

Eingabe: $\alpha > 0$

Ausgabe: a_1/b_1, $a_2/b_2, \ldots$

Verfahren: Setze $z_{-1} = (1,0)$, $z_0 = (0,1)$. Bestimme $m_1, m_2, \ldots \in \mathbb{N}_0$, $r_1, r_2, \ldots \geq 0$ und $z_1 = (b_1, a_1)$, $z_2 = (b_2, a_2) \ldots \in \mathbb{N}_0^2$ rekursiv aus den Gleichungen

$$\begin{aligned} d(z_{i-1}) &= m_{i+1} d(z_i) + r_{i+1}, \quad 0 \leq r_{i+1} < d(z_i) \quad \text{und} \\ z_{i+1} &= m_{i+1} z_i + z_{i-1}, \end{aligned}$$

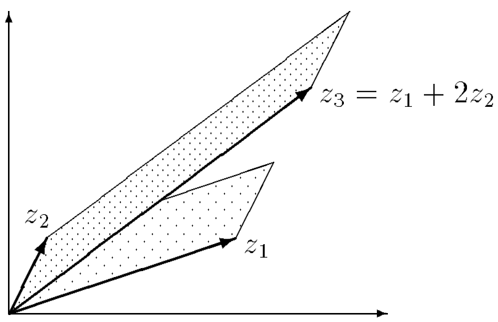
$i = 0, 1 \ldots$ Gilt $r_j = 0$, so beende die Prozedur. □

Wegen $d(z_{-1}) = \alpha$ und $d(z_0) = 1$ ist $m_1 = [\alpha]$ und $z_1 = (1, [\alpha])$. Für ganzzahliges α bricht das Verfahren mit der Ausgabe $a_1/b_1 = \alpha$ bereits ab. Andernfalls gilt $d(z_0) = 1 > d(z_1) = \alpha - [\alpha]$, und unsere Konstruktion kommt zum Zuge. Dann folgt $d(z_{i+1}) = r_{i+1} < d(z_i)$, und folglich nach unserer Konstruktionsvorschrift $m_2, m_3, \ldots \geq 1$. Es entstehen immer bessere Näherungsbrüche, die abwechselnd oberhalb und unterhalb von α liegen,

$$\frac{a_1}{b_1} < \frac{a_3}{b_3} < \frac{a_5}{b_5} < \ldots \leq \alpha \leq \ldots < \frac{a_4}{b_4} < \frac{a_2}{b_2}.$$

Das Verfahren bricht ab, falls $r_j = d(z_j) = 0$ für ein j, also $a_j/b_j = \alpha$. Für irrationales α ist dies nicht möglich, dann entstehen unendlich viele Näherungsbrüche. Im rationalen Fall $\alpha = a/b$ mit $a, b \in \mathbb{N}$ sind $bd(z_i) = |ba_i - ab_i|$ ganzzahlig, und die Gleichungen $bd(z_{i-1}) = m_{i+1} bd(z_i) + bd(z_{i+1})$ sind Divisionen mit Rest, wie sie beim Euklidischen Algorithmus auftreten. Daher bricht im rationalen Fall unser Verfahren genau wie der Euklidische Algorithmus ab.

Als nächstes zeigen wir, daß es sich bei den a_i/b_i um gekürzte Brüche handelt. Wir machen uns zunutze, daß das von den Vektoren z_i und z_{i+1} aufgespannte Parallelogramm $P(z_i, z_{i+1})$ die Fläche 1 hat. Denn: $P(z_{-1}, z_0)$ hat als Einheitsquadrat die Fläche 1, und wegen $z_{i+1} = z_{i-1} + m_{i+1} z_i$ entsteht $P(z_{i+1}, z_i)$ durch Scherung aus $P(z_{i-1}, z_i)$.



Algebraisch ausgedrückt heißt dies

$$\begin{aligned}\det\begin{pmatrix} b_i & b_{i+1} \\ a_i & a_{i+1}\end{pmatrix} &= \det\begin{pmatrix} b_i & b_{i-1}+m_{i+1}b_i \\ a_i & a_{i-1}+m_{i+1}a_i\end{pmatrix} \\ = \det\begin{pmatrix} b_i & b_{i-1} \\ a_i & a_{i-1}\end{pmatrix} &= -\det\begin{pmatrix} b_{i-1} & b_i \\ a_{i-1} & a_i\end{pmatrix},\end{aligned}$$

also

$$\begin{aligned} a_{i+1}b_i - b_{i+1}a_i &= -(a_ib_{i-1} - b_ia_{i-1}) = \ldots \\ = (-1)^{i+1}(a_0b_{-1} - b_0a_{-1}) &= (-1)^{i+1}. \end{aligned} \tag{2.3}$$

Insbesondere teilt jeder Teiler von a_i und b_i auch 1, daher folgt wie behauptet

$$1 = \mathrm{ggT}(a_i, b_i).$$

Bemerkung. Ist $\alpha = a/b$ rational, so bricht das Verfahren ab und es gilt $a/b = a_j/b_j$ für ein $j \geq 1$. Da a_j und b_j teilerfremd sind, folgt $a = da_j$ und $b = db_j$ mit $d = \mathrm{ggT}(a, b)$. Aus $a_jb_{j-1} - b_ja_{j-1} = \pm 1$ ergibt sich durch Multiplikation mit d die folgende Version des Satzes von Bézout:

$$ab_{j-1} - ba_{j-1} = \mathrm{ggT}(a, b).$$

□

Wir können nun leicht den **Satz von Lagrange** beweisen, nach welchem die Näherungsbrüche die erwünschte gute Approximationsgüte haben.

Satz 2.2. *Es gilt für alle Näherungsbrüche a_i/b_i von α*

$$\left|\alpha - \frac{a_i}{b_i}\right| < \frac{1}{b_i^2}.$$

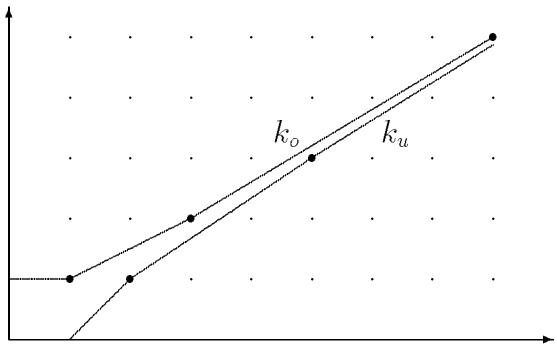
Beweis. Da α zwischen a_i/b_i und a_{i+1}/b_{i+1} liegt, folgt unter Beachtung von (2.3) und $b_{i+1} = m_{i+1}b_i + b_{i-1} > b_i$

$$\left|\alpha - \frac{a_i}{b_i}\right| \leq \left|\frac{a_{i+1}}{b_{i+1}} - \frac{a_i}{b_i}\right| = \frac{1}{b_{i+1}b_i} < \frac{1}{b_i^2}.$$

□

Die Abschätzung zeigt, daß wir besonders gute Näherungen erhalten, wenn b_{i+1} sehr viel größer als b_i ist, d.h. m_{i+1} einen großen Wert annimmt (vgl. dazu das nachfolgende Beispiel 1)).

Wir beenden unsere geometrische Betrachtung, indem wir zeigen, daß unsere Konstruktion keine guten Näherungspunkte ausläßt. Dazu betrachten wir die Kurve k_u in $\mathbb{R}^2_+$, die sich ergibt, indem wir die Punkte (1,0), z_1, z_3, z_5 durch Strecken miteinander verbinden. Sie liegt unterhalb der Geraden $y = \alpha x$. Analog sei k_o die Kurve, die sich durch lineare Interpolation von (0,1), z_2, z_4,... ergibt. Sie liegt oberhalb von $y = \alpha x$. Das folgende Bild illustruiert dies für $\alpha = (\sqrt{5} - 1)/2$.



Nach Konstruktion hat die Verbindungsstrecke zwischen z_{i-1} und z_{i+1} die Steigung a_i/b_i, und diese Zahlen wachsen für ungerades i und fallen für gerades i. Daher ist k_u eine konkave und k_o eine konvexe Kurve. Auf diesen Kurven können weitere Punkte mit ganzzahligen Komponenten liegen, man kann auch sie als gute Näherungspunkte für α betrachten. Wesentlich ist, daß *zwischen* den beiden Kurven keine Punkte mit ganzzahligen Komponenten liegen (abgesehen von (0,0)). In diesem Sinne findet unsere Konstruktion *alle* guten Näherungen. Zum Beweis zerlegen wir den Bereich zwischen k_u und k_o in die Dreiecke Δ_i, $i = 0, 1, \ldots$, mit den Ecken z_{i-1}, z_i und z_{i+1}. Auf Δ_i betrachten wir die lineare Funktion

$$z = (x, y) \ \mapsto \ \ell(z) := xa_i - yb_i.$$

Es gilt $\ell(z_i) = 0$ und $\ell(z_{i-1}) = \ell(z_{i+1}) = \pm 1$ nach (2.3). Daher hat ℓ auch auf der Strecke zwischen z_{i-1} und z_{i+1} den Wert ± 1. Sonst kann ℓ auf Δ_i keine ganzzahligen Werte annehmen, deswegen gibt es dort keine Punkte mit ganzzahligen Koordinaten. Dies ist unsere Behauptung. (Wir haben sie für irrationales α formuliert. Im rationalen Fall, in dem das Verfahren nach endlich vielen Schritten abbricht, ist die Behauptung geeignet abzuändern.)

Geometrisch gesehen ist unsere Konstruktion 2-dimensional, sie macht plausibel, warum die Näherungsbrüche besonders nahe bei α liegen. Rechnerisch geht es um die Bestimmung der Zahlen m_i, dies leistet ein 1-dimensionaler

Algorithmus, der Kettenbruchalgorithmus des letzten Abschnitt. Dazu formen wir die Gleichungen $d(z_{i-1}) = m_{i+1}d(z_i) + d(z_{i+1})$ um zu

$$\alpha_i \;=\; m_{i+1} + \alpha_{i+1}^{-1}, \qquad \text{mit } \alpha_i \;:=\; \frac{d(z_{i-1})}{d(z_i)}.$$

Da $d(z_i)$ für $i \geq 0$ strikt monoton fällt, ist $\alpha_{i+1} > 1$, also

$$m_{i+1} \;=\; [\alpha_i], \quad \alpha_{i+1} \;=\; (\alpha_i - m_{i+1})^{-1}.$$

Unter Beachtung von $\alpha_0 = \alpha$ können wir daher unsere geometrische Konstruktion der Näherungsbrüche zu dem folgenden arithmetischen Verfahren umformulieren.

Kettenbruchalgorithmus für Näherungsbrüche.

Eingabe: $\alpha > 0$

Ausgabe: Die Näherungsbrüche $a_1/b_1, a_2/b_2, \ldots$ von α

Verfahren: Setze $\alpha_0 = \alpha$, $a_{-1} = b_0 = 0$ und $b_{-1} = a_0 = 1$. Bestimme Zahlen $m_1, m_2, \ldots$ und $\alpha_1, \alpha_2, \ldots$ aus den Gleichungen $m_{i+1} = [\alpha_i], \alpha_i = m_{i+1} + 1/\alpha_{i+1}$. Berechne $a_1, a_2, \ldots, b_1, b_2, \ldots$ rekursiv aus den Gleichungen $a_{i+1} = m_{i+1}a_i + a_{i-1}$, $b_{i+1} = m_{i+1}b_i + b_{i-1}$. Im Falle $m_j = \alpha_j$ breche die Prozedur ab. □

Dieser Algorithmus ist nicht nur analog zum Kettenbruchalgorithmus für gebrochene Zahlen aus Abschnitt 2.1, er liefert uns gleichzeitig Kettenbruchdarstellungen aller Näherungsbrüche von α. Dazu betrachten wir auch alle Zahlen α' zwischen α und a_i/b_i. Aus unseren geometrischen Überlegungen folgt, daß zwischen den Geraden mit Steigungen α bzw. α' keine ganzzahligen Punkte liegen, wenn man sich auf den Bereich beschränkt, der links und unterhalb von z_i liegt. Der Kettenbruchalgorithmus liefert also in den ersten i Schritten für α und α' dieselben Näherungspunkte $z_1, \ldots, z_i$ und Zahlen $m_1, \ldots, m_i$. Daher gilt auch

$$\alpha'_j = m_{j+1} + 1/\alpha'_{j+1}, \quad j = 0, \ldots, i-1$$

für die analogen Größen $\alpha'_j := d'(z_{j-1})/d'(z_j)$. Setzen wir diese Gleichungen wieder ineinander ein, so folgt

$$\alpha'_0 \;=\; m_1 + \frac{1}{m_2+}\,\frac{1}{m_3+}\cdots\frac{1}{m_{i-1}+}\,\frac{1}{m_i+}\,\frac{1}{\alpha'_i}.$$

Lassen wir nun α' gegen a_i/b_i streben, so konvergiert $\alpha'_0 = \alpha'$ gegen a_i/b_i und α'_i gegen ∞ wegen $d'(z_i) \to 0$, und wir erhalten die **Kettenbruchdarstellung für die Näherungsbrüche**

$$\frac{a_i}{b_i} \;=\; m_1 + \frac{1}{m_2+}\,\frac{1}{m_3+}\cdots\frac{1}{m_{i-1}+}\,\frac{1}{m_i}. \tag{2.4}$$

Wie wir gesehen haben, konvergieren im irrationalen Fall die Näherungsbrüche gegen α, man spricht dann von der **Kettenbruchentwicklung** von α und schreibt

$$\alpha \;=\; m_1 + \frac{1}{m_2+}\,\frac{1}{m_3+}\cdots$$

Anders als Dezimal- oder Dualbruchdarstellungen reeller Zahlen beruht die Kettenbruchentwicklung einer irrationalen Zahl nicht auf der Wahl einer Basis. Sie spiegelt daher Eigenschaften der Zahl eher wieder als etwa die Dezimalbruchentwicklung. Zum Rechnen sind sie dagegen weniger gut geeignet.

Bemerkung. Die Kettenbruchdarstellung einer rationalen Zahl ist nicht eindeutig, wie man der Formel

$$m_1 + \frac{1}{m_2+}\,\frac{1}{m_3+}\cdots\frac{1}{m_{i-1}+}\,\frac{1}{m_i+}\,\frac{1}{1} \;=\; m_1 + \frac{1}{m_2+}\,\frac{1}{m_3+}\cdots\frac{1}{m_{i-1}+}\,\frac{1}{m_i+1}$$

entnimmt. Daher stimmen die Entwicklungen (2.2) und (2.4) im allgemeinen nicht überein. Geometrisch entspricht dem (wie eine genauere Analyse zeigt) der Sachverhalt, daß man sich dem Punkt z_i in dem beschriebenen Grenzübergang von oben oder von unten nähern kann. □

Beispiele.

1) Die Kettenbruchentwicklung der Zahl $\pi = 3,14159265\ldots$ ist

$$\pi \;=\; 3 + \frac{1}{7+}\,\frac{1}{15+}\,\frac{1}{1+}\,\frac{1}{292+}\,\frac{1}{1+}\cdots$$

Wegen $m_5 = 292$ ist die Näherung $3 + \frac{1}{7+}\,\frac{1}{15+}\,\frac{1}{1} = 3 + \frac{16}{113} = 3,14159292\ldots$ besonders gut (vg. die Bemerkung nach Satz 2.2).

2) Wir wollen die Kettenbruchdarstellung von $\sqrt{2}$ berechnen. Nach dem Algorithmus für Näherungsbrüche gilt

$$\begin{aligned} m_1 &= [\sqrt{2}] = 1, \quad \alpha_1 = (\sqrt{2}-1)^{-1} = \sqrt{2}+1, \\ m_2 &= [\alpha_2] = 2, \quad \alpha_2 = (\sqrt{2}-1)^{-1}, \\ &\cdots \end{aligned}$$

Diese Rechnung reproduziert sich in offensichtlicher Weise, daher folgt

$$\sqrt{2} \;=\; 1 + \frac{1}{2+}\frac{1}{2+}\frac{1}{2+}\cdots$$

3) $\phi = \frac{1}{2}(\sqrt{5}-1)$ erfüllt die Gleichung $\phi = 1 + \phi^{-1}$, daher gilt $0 = m_1$, $1 = m_2 = m_3 = \ldots$ und

$$\frac{1}{2}(\sqrt{5}-1) \;=\; \frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\cdots \qquad \square$$

Bei den beiden letzten Beispielen handelt es sich um einfache Fälle von periodischen Kettenbrüchen, bei denen sich die Zahlen m_i schließlich periodisch wiederholen. In der Zahlentheorie wird gezeigt, daß sich genau die quadratischen Irrationalzahlen, die Lösungen quadratischer Gleichungen in periodische Kettenbrüche entwickeln lassen.

Literatur

M.C. Irwin (1989): Geometry of Continued Fractions, Amer. Math. Monthly 96

O. Perron (1954): Die Lehre von den Kettenbrüchen

H.M. Stark (1970): An Introduction to Number Theory

Kapitel 3

Kongruenzen

3.1 Der Restklassenring $\mathbb{Z}_m$

Beim Rechnen mit ganzen Zahlen ist es häufig von Vorteil, nicht mit den Zahlen selbst zu operieren, sondern mit den Resten, die beim Teilen der Zahlen durch ein fest vorgegebenes $m \in \mathbb{Z}$ übrigbleiben. m wird dann als **Modul** bezeichnet. Entscheidend ist, daß man mit den Resten genauso rechnen kann wie mit den Zahlen selbst, die Rechenregeln bleiben zu einem wesentlichen Teil bestehen. Dies macht sich zum Beispiel die bekannte ‚Neunerprobe' zur Kontrolle von Multiplikationen zunutze. Der Vorteil des Rechnens mit Resten liegt auf der Hand: Die Reste sind im allgemeinen viel kleiner als die Zahlen selbst.

Definition 3.1. *Sei $m \in \mathbb{N}$ und $a, b \in \mathbb{Z}$.*

i) a und b heißen ***kongruent modulo m****, falls $m \mid b - a$, falls also a und b denselben Rest bei Division durch m haben. Man schreibt dann*

$$a \equiv b \bmod m.$$

ii) Die Menge $\overline{a} := \{a + zm \ : \ z \in \mathbb{Z}\} = a + m\mathbb{Z}$ heißt ***Restklasse*** *von a modulo m. Eine andere Schreibweise für die Restklasse ist $a \bmod m$. Die Menge aller Restklassen wird mit $\mathbb{Z}/m\mathbb{Z}$ oder $\mathbb{Z}_m$ bezeichnet.*

Es gilt $a \equiv b \bmod m \Leftrightarrow \overline{a} = \overline{b}$. Die Kongruenz ist also eine Äquivalenzrelation, und $\mathbb{Z}$ zerfällt modulo m in m disjunkte Restklassen. Der folgende Satz präzisiert die Aussage, daß sich das Rechnen mit ganzen Zahlen im wesentlichen auf die Restklassen überträgt.

Satz 3.2. *Es gilt*

$$a \equiv a',\ b \equiv b' \bmod m \quad \Rightarrow \quad a+b \equiv a'+b',\ ab \equiv a'b' \bmod m,$$

so daß $\overline{a}+\overline{b} := \overline{a+b}, \overline{a}\cdot\overline{b} := \overline{ab}$ *wohldefinierte Verknüpfungen in* $\mathbb{Z}_m$ *sind. Damit wird* $\mathbb{Z}_m$ *zu einem kommutativen Ring mit Einselement.*

Beweis. $m \mid (a-a')$ und $m \mid (b-b')$ implizieren $m \mid (a+b-(a'+b'))$ und, wegen $ab-a'b' = a(b-b')+(a-a')b'$, auch $m \mid (ab-a'b')$. Dies ergibt die erste Behauptung. Die Rechenregeln übertragen sich von $\mathbb{Z}$ unmittelbar auf $\mathbb{Z}_m$, z.B.

$$\begin{aligned} \overline{a}+(\overline{b}+\overline{c}) &= \overline{a}+\overline{b+c} = \overline{a+(b+c)} \\ &= \overline{(a+b)+c} = \overline{a+b}+\overline{c} = (\overline{a}+\overline{b})+\overline{c}. \end{aligned}$$

Das Nullelement ist $\overline{0}$, das Einselement $\overline{1}$. □

Beispiel. Seien $a = \sum a_i 10^i$, $b = \sum b_i 10^i$ ganze Zahlen in Dezimaldarstellung und $c = \sum c_i 10^i$ ihr Produkt. Dann gilt wegen $10 \equiv 1 \bmod 9$

$$\sum a_i \sum b_i \equiv ab \equiv c \equiv \sum c_i \bmod 9.$$

Die **Neunerprobe** zur Kontrolle von Multiplikationen (bei der eine Zahl durch die Summe der Ziffern aus ihrer Dezimaldarstellung ersetzt wird) beruht auf dieser Feststellung. □

$\mathbb{Z}_m$ heißt **Restklassenring modulo m**. Diese Ringe haben Besonderheiten, wie man sie von $\mathbb{Z}$ nicht kennt:

Nullteiler. Haben a und m einen gemeinsamen Teiler d, also $a = cd$ und $m = bd$, so folgt $\overline{a}\cdot\overline{b} = \overline{cdb} = \overline{cm} = \overline{0}$. Ist $\overline{a}, \overline{b} \neq \overline{0}$, so heißen $\overline{a}, \overline{b}$ Nullteiler. Zum Beispiel gilt $\overline{2}\cdot\overline{3} = \overline{0}$ in $\mathbb{Z}_6$, jedoch $\overline{2}, \overline{3} \neq \overline{0}$. Anders als $\mathbb{Z}$ enthält $\mathbb{Z}_m$ Nullteiler, abgesehen von dem Fall, daß m eine Primzahl ist.

In Ringen mit Nullteilern kann man im allgemeinen nicht kürzen. So gilt $\overline{4}\cdot\overline{2} = \overline{4}\cdot\overline{5}$, aber $\overline{2} \neq \overline{5}$ in $\mathbb{Z}_6$.

Einheiten. Seien nun a und m teilerfremd. Dann gibt es Zahlen s und t, so daß $as + mt = 1$, also $\overline{1} = \overline{a}\cdot\overline{s} + \overline{mt} = \overline{a}\cdot\overline{s} + \overline{0} = \overline{a}\cdot\overline{s}$. Es ist also $\overline{a}$ eine Einheit in $\mathbb{Z}_m$ und $\overline{s} = \overline{a}^{-1}$. Inverse Elemente lassen sich mit dem (erweiterten) Euklidischen Algorithmus berechnen. In $\mathbb{Z}_7$ gilt $\overline{2}^{-1} = \overline{4}$, $\overline{3}^{-1} = \overline{5}$ und $\overline{6}^{-1} = \overline{6}$.

Eine besondere Rolle spielen die Restklassenringe modulo einer Primzahl p. In $\mathbb{Z}_p$ besitzt jede Restklasse ungleich $\overline{0}$ eine inverse Restklasse, es gilt also der folgende Satz.

Satz 3.3. *Ist p eine Primzahl, so ist $\mathbb{Z}_p$ ein Körper.*

3.2 Der Chinesische Restsatz

Es ist eine allgemeine Strategie zur Bewältigung von umfangreichen Rechnungen, das Problem in kleinere, vom Rechenaufwand her überschaubare Teile zu zerlegen (‚divide et impera'). Dazu hat sich das Rechnen mit Restklassen als besonders wirksam erwiesen. In der **modularen Arithmetik** wird eine ganze Zahl a durch ein Zahlentupel $(a_1, \ldots, a_k)$ ersetzt. Man gibt sich paarweise teilerfremde Moduln $m_1, \ldots, m_k$ vor und bestimmt die a_i als die Reste, die bei Division von a durch m_i übrigbleiben. Im letzten Abschnitt haben wir gesehen, daß sich das Rechnen mit a in kanonischer Weise auf die a_i überträgt. Gerechnet wird daher soweit wie möglich mit den Resten modulo m_i.

Diese Strategie setzt voraus, daß man a aus $(a_1, \ldots, a_k)$ wieder zurückgewinnen kann. Sei also a' eine weitere ganze Zahl, die bei Division durch m_i die Reste a_i ergibt. Dann gilt $m_i \mid (a - a')$ für alle i. Wegen der Eindeutigkeit der Primfaktorzerlegung in den ganzen Zahlen folgt $m \mid (a - a')$ bzw. $a \equiv a' \bmod m$, mit $m = m_1 \cdot \ldots \cdot m_k$, denn die m_i sind als paarweise teilerfremd angenommen. Ist daher $-m/2 < a < m/2$, so ist a durch $(a_1, \ldots, a_k)$ eindeutig festgelegt. Man wird also die m_i so wählen, daß m ausreichend groß ist.

Es bleibt die Frage, wie man am Ende der Rechnung a aus den Resten a_i wieder rekonstruiert. Dazu muß man das System von Kongruenzen

$$x \equiv a_1 \bmod m_1, \quad x \equiv a_2 \bmod m_2, \quad \ldots, \quad x \equiv a_k \bmod m_k$$

in der Unbekannten x lösen. Unsere bisherigen Überlegungen zeigen, daß dieses Gleichungssystem modulo m eindeutig lösbar ist, d.h. zwei Lösungen a und a' zu derselben Restklasse modulo m gehören.

Um eine Lösung zu finden, konstruieren wir Basislösungen $e_i, i = 1, \ldots, k$. e_i soll das Gleichungssystem in dem speziellen Fall $a_i = 1,\ a_j = 0$ für $j \neq i$ lösen. Setze

$$m_i' := \prod_{j \neq i} m_j = m/m_i.$$

Da die Modulen $m_1, \ldots, m_k$ paarweise teilerfremd sind, gilt $1 = \mathrm{ggT}(m_i, m_i')$. Mit dem erweiterten Euklidischen Algorithmus gewinnt man ganze Zahlen

s_i, t_i, so daß

$$1 = m_i s_i + m_i' t_i.$$

Wir setzen nun

$$e_i := 1 - m_i s_i = m_i' t_i.$$

Dann gilt nach Definition von m_i' wie gewünscht

$$e_i \equiv \begin{cases} 1 \bmod m_i \\ 0 \bmod m_j & \text{für } j \neq i. \end{cases}$$

Aus den Basislösungen können wir eine Lösung

$$a = \sum a_i e_i$$

für das ursprüngliche Kongruenzensystem zusammensetzen. Insgesamt erhalten wir den **Chinesischen Restsatz** (der in einem speziellen Fall bereits Sun Tsu etwa 300 n.Chr. bekannt war).

Satz 3.4. *Seien $m_1, \ldots, m_k$ paarweise teilerfremde natürliche Zahlen und seien $a_1, \ldots, a_k$ ganze Zahlen. Dann existiert eine Lösung a des Systems von Kongruenzen*

$$x \equiv a_1 \bmod m_1,\ x \equiv a_2 \bmod m_2,\ \ldots\ , x \equiv a_k \bmod m_k,$$

und sie ist modulo $m = m_1 \cdot \ldots \cdot m_k$ eindeutig. Mit anderen Worten: Es gibt eine ganze Zahl a, so daß das Gleichungssystem äquivalent ist zu der Gleichung

$$x \equiv a \bmod m.$$

In Restklassen ausgedrückt bedeutet der Satz, daß die Zuordnung

$$a \bmod m \mapsto (a \bmod m_1, \ldots, a \bmod m_k)$$

eine Bijektion zwischen $\mathbb{Z}_m$ und $\mathbb{Z}_{m_1} \times \ldots \times \mathbb{Z}_{m_k}$ definiert. Die Eindeutigkeitsaussage bedeutet Injektivität der Abbildung, die Existenzaussage Surjektivität. Da beide Mengen $m = m_1 \cdot \ldots \cdot m_k$ Elemente enthalten, bedingen sich Injektivität und Surjektivität gegenseitig. So gesehen ist die Existenzaussage des Chinesischen Restsatzes eine Folgerung aus seiner Eindeutigkeitsaussage, und umgekehrt.

Beispiel. Gesucht ist eine Lösung des Systems

$$x \equiv 2 \bmod 3, \; x \equiv 3 \bmod 5, \; x \equiv 4 \bmod 7.$$

Es gilt $m_1' = 5 \cdot 7 = 35$, $m_2' = 3 \cdot 7 = 21$,$m_3' = 3 \cdot 5 = 15$. Aus

$$\begin{array}{rcccl} 1 & = & \mathrm{ggT}(3,35) & = & 12 \cdot 3 - 1 \cdot 35 \\ 1 & = & \mathrm{ggT}(5,21) & = & -4 \cdot 5 + 1 \cdot 21 \\ 1 & = & \mathrm{ggT}(7,15) & = & -2 \cdot 7 + 1 \cdot 15 \end{array}$$

erhalten wir das Basissystem $e_1 = -35, e_2 = 21, e_3 = 15$. Es ist also

$$-2 \cdot 35 + 3 \cdot 21 + 4 \cdot 15 \;=\; 53,$$

Lösung, und das Gleichungssystem ist äquivalent zu $x \equiv 53 \bmod 105$. □

3.3 Ein probabilistischer Gleichheitstest

Zwei Personen an den Enden eines Nachrichtenkanals wollen zwei natürliche Zahlen $a, b < 2^{10.000}$ auf Gleichheit hin überprüfen. Um Übertragungsfehler zu vermeiden, möchten sie die Zahlen nicht vollständig übermitteln. Das folgende Verfahren erlaubt einen Vergleich der beiden Zahlen, dabei werden anstelle der 10.000 Bits einer Zahl nur $k \cdot 101$ Bits gesendet. Wir werden sehen, daß schon für $k = 1$ das Verfahren höchste Sicherheit garantiert.

Probabilistischer Gleichheitstest.

Ausgabe: ‚$a = b$' oder ‚$a \neq b$'

Verfahren: Wähle zufällig Primzahlen $p_1, \ldots, p_k$ zwischen 2^{100} und 2^{101}. Übertrage a modulo p_i für alle $i = 1, \ldots, k$. Falls $a \not\equiv b \bmod p_i$ für ein i, gilt ‚$a \neq b$'. Andernfalls treffe die Entscheidung ‚$a = b$'. □

Das Verfahren setzt voraus, daß man sich die benötigten Primzahlen leicht verschaffen kann. Darauf kommen wir später zurück.

Wir schätzen nun die Wahrscheinlichkeit ab, daß das Verfahren zu einer Fehlentscheidung führt. Nehmen wir dazu zunächst an, daß es 100 verschiedene Primzahlen $q_1, \ldots, q_{100} > 2^{100}$ gibt, für die alle $a \equiv b \bmod q_i$ gilt. Nach dem chinesischen Restsatz folgt $a \equiv b \bmod m$, mit $m = q_1 \cdot \ldots \cdot q_{100}$. Es gilt $m > (2^{100})^{100} = 2^{10.000}$, nach Annahme folgt daher $a = b$. In diesem Fall ist ein Übertragungsfehler ausgeschlossen.

Eine Fehlentscheidung ist also nur möglich, falls es weniger als 100 Primzahlen $q > 2^{100}$ mit der Eigenschaft $a \equiv b \bmod q$ gibt, und eine Fehlentscheidung tritt nur dann ein, wenn das Verfahren zufälligerweise nur derartige Primzahlen auswählt. Nach dem berühmten Primzahlsatz gilt für die Anzahl $\pi(x)$ aller Primzahlen $q < x$ die asymptotische Formel $\pi(x) \sim x/\ln x$. Zwischen 2^{100} und 2^{101} gibt es daher approximativ

$$\pi(2^{101}) - \pi(2^{100}) \;\approx\; \frac{2^{101}}{\ln 2^{101}} - \frac{2^{100}}{\ln 2^{100}} \;\approx\; \frac{2^{100}}{100 \ln 2} \;\approx\; 9,90 \cdot 10^{27}$$

Primzahlen. Bei k-facher unabhängiger Wahl einer Primzahl ist die Fehlerwahrscheinlichkeit also höchstens

$$\left(\frac{99}{2^{93}}\right)^k \;\approx\; 10^{-26 \cdot k}.$$

Schon für $k = 1$ ist dies ein verschwindend kleiner Wert.

3.4 Exakte Lösung ganzzahliger Gleichungssysteme

Wir lösen nun lineare Gleichungssysteme mit modularer Arithmetik. Sei $A = (a_{ij})$ eine ganzzahlige $n \times n$ Matrix mit $\det A \neq 0$, und sei $b = (b_1, \ldots, b_n)^t$ ein ganzzahliger Vektor. Dann ist das lineare Gleichungssystem

$$Ax \;=\; b$$

mit $x = (x_1, \ldots, x_n)^t$ eindeutig in den rationalen Zahlen lösbar. Das bekannteste Lösungsverfahren ist das Gaußsche Eliminationsverfahren. Es kann in den Zwischenrechnungen zu so großen Werten führen, daß es nicht mehr in exakter Weise gelingt. Bei Verwendung von Gleitkomma-Arithmetik kann es sehr ungenaue Ergebnisse liefern.

Es liegt daher nahe, das Gaußsche Verfahren in modularer Arithmetik durchzuführen. Der Hauptanteil der arithmetischen Operationen kann so mit kleinen Zahlen vollzogen werden. Unser Ausgangspunkt ist die Cramersche Regel, nach der die Lösung durch

$$x_j \;=\; (\det A)^{-1} \sum_i b_i \det A_{ij}$$

gegeben ist; dabei bezeichnet A_{ij} die Matrix, die aus A durch Streichen der i-ten Zeile und der j-ten Spalte entsteht. Es genügt daher, $\det A$ und den ganzzahligen Vektor $y = \det A \cdot x$ zu bestimmen. Man erhält y als eindeutige ganzzahlige Lösung des Gleichungssystems

$$Ay \;=\; \det A \cdot b.$$

Das Verfahren besteht aus mehreren Schritten.

1) Wähle Primzahlen $p_1, \dots, p_k$ (und zwar an der Grenze der verfügbaren Zahlen: Hat der Computer Wortlänge 64, so wähle man die p_i zwischen 2^{63} und 2^{64}).

2) Löse des Gleichungssystems modulo p_i. Bringe dazu die erweiterte Koeffizientenmatrix (A, b) modulo p_i auf obere Dreiecksform:

$$(A, b) \;\mapsto\; (A', b') = \begin{pmatrix} a'_{11} & a'_{12} & \dots & a'_{1n} & b'_1 \\ 0 & a'_{22} & \dots & a'_{2n} & b'_2 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & a'_{nn} & b'_n \end{pmatrix}$$

Dies läßt sich erreichen, indem man das Gaußsche Eliminationsverfahren modulo p_i durchführt. Das Verfahren besteht bekanntlich aus der Addition von Vielfachen einer Matrixzeile zu einer anderen Zeile und dem Austausch von Zeilen. Diese Operationen lassen sich in Vektorräumen mit beliebigem Skalarbereich durchführen. In unserem Fall ist der Skalarbereich der Körper $\mathbb{Z}_{p_i}$. Die Addition einer Zeile zu einer anderen läßt $\det A$ unverändert, während Zeilenvertauschungen das Vorzeichen von $\det A$ wechseln. Für den Rest d_i von $a'_{11}a'_{22}\dots a'_{nn}$ modulo p_i folgt also

$$\pm d_i \;\equiv\; \det A \mod p_i$$

(+ bei einer geraden, – bei einer ungeraden Anzahl von Zeilenvertauschungen). Eine Lösung $y^{(i)} = (y_1^{(i)}, \dots, y_n^{(i)})$ des Gleichungssystems modulo p_i erhält man nun schrittweise aus den Gleichungen

$$\begin{aligned} y_n^{(i)} &\equiv \pm d_i b'_n / a'_{nn} \bmod p_i \\ &\;\vdots \\ y_1^{(i)} &\equiv (\pm d_i b'_1 - \sum_{j=2}^{n} a'_{1j} y_j^{(i)}) / a'_{11} \bmod p_i \end{aligned}$$

Für die Divisionen werden die inversen Elemente von $a'_{nn}, \dots, a'_{11}$ in $\mathbb{Z}_{p_i}$ benötigt. Sie existieren, falls $d_i \not\equiv 0 \bmod p_i$. Der Ausnahmefall $d_i \equiv 0 \bmod p_i$ tritt nur für Primteiler p_i von $\det A$ ein, man verwerfe dann p_i.

3) Nur in diesem Schritt sind Rechnungen mit großen Zahlen erforderlich. Setze $m = p_1 \dots p_k$. Bestimme nach dem Chinesischen Restsatz ganze Zahlen d und $y'_1, \dots, y'_n$ zwischen $-m/2$ und $m/2$, so daß für alle i

$$d \equiv d_i \bmod p_i \text{ und } y'_j \equiv y_j^{(i)} \bmod p_i.$$

Dann folgt

$$\pm d \equiv \det A \bmod m \text{ und } y' \equiv \det A \cdot b \bmod m,$$

4) Die Zahl $\pm d$ und der Vektor y' sind Kandidaten für $\det A$ und y. Mit dem probabilistischen Gleichheitstest aus Abschnitt 3.3 kann man dies überprüfen. Man wählt dazu zufällig eine weitere Primzahl p_{k+1} und testet, ob $Ay' \equiv \pm d \cdot b \bmod p_{r+1}$ gilt. Bestehen $\pm d$ und y' diesen Test, so wird die Prozedur mit der

$$\textit{Ausgabe:} \quad x = \pm d^{-1} \cdot y'$$

beendet. Andernfalls wird die Prozedur mit der erweiterten Primzahlfolge $p_1, \dots, p_{k+1}$ wiederholt. Die Fehlerwahrscheinlichkeit dieses Tests ist vernachlässigbar klein, ähnlich wie die des Gleichheitstests im letzten Abschnitt.

3.5 Ideale und Faktorringe

Die Grundidee der letzten Abschnitte war es, die ganzen Zahlen so in Klassen aufzuteilen, daß man mit diesen Klassen in ähnlicher Weise wie mit den Zahlen rechnen kann. Man muß nur darauf achten, daß sich die Klassenbildung mit der Addition und Multiplikation ganzer Zahlen verträgt, im Sinne von Satz 3.2. Diese Idee ist so einfach wie schlagend, daß man sich herausgefordert fühlt, sie auch in anderen Rechenbereichen auszuprobieren. Für Polynomringe wird sie sich später als ähnlich nützlich erweisen. In diesem Abschnitt wollen wir die Restklassenbildung in beliebigen Ringen betrachten.

Sei $\equiv$ eine Äquivalenzrelation in einem Ring R. Die Äquivalenzklasse, in der a liegt, bezeichnen wir wieder mit $\overline{a}$. Wir setzen voraus, daß sich die Relation mit Addition und Multiplikation verträgt, daß also

$$a_1 \equiv a_2,\ b_1 \equiv b_2 \quad \Rightarrow \quad a_1 + b_1 \equiv a_2 + b_2,\ a_1 b_1 \equiv a_2 b_2 \tag{3.1}$$

gilt. Dann können wir wie in den ganzen Zahlen durch $\overline{a} + \overline{b} := \overline{a+b}$, $\overline{a} \cdot \overline{b} := \overline{ab}$ in wohldefinierter Weise Addition und Multiplikation von Äquivalenzklassen einführen. Die Äquivalenzklassen bilden einen Ring $\overline{R}$. Es ist offensichtlich, daß sich Assoziativ-, Distributiv- und Kommutativgesetze von R auf $\overline{R}$ übertragen, wir haben dies bereits für den Restklassenring $\mathbb{Z}_m$ festgestellt. Das Nullelement in $\overline{R}$ ist die Restklasse $\overline{0}$, denn $\overline{a} + \overline{0} = \overline{a}$, und aus $a \equiv b$ folgt $-a \equiv (-a)+b+(-b) \equiv (-a)+a+(-b) \equiv -b$, so daß $-\overline{a} := \overline{-a}$ wohldefiniert ist. Besitzt R ein Einselement 1, so ist $\overline{1}$ Einselement in $\overline{R}$.

Wir wollen nun die Äquivalenzrelationen charakterisieren, die (3.1) erfüllen. Es gilt $a \equiv b$ genau dann, wenn $a-b \equiv 0$, daher ist die Äquivalenzrelation bereits durch die Teilmenge $I := \overline{0}$ von R eindeutig gekennzeichnet. Als Nullelement von $\overline{R}$ hat I folgende Eigenschaften:

$$\begin{array}{ll}
a, b \in I \;\Rightarrow\; a+b \in I, & \text{denn } a \equiv 0, b \equiv 0 \;\Rightarrow\; a+b \equiv 0+0 = 0, \\
a \in I \;\Rightarrow\; -a \in I, & \text{denn } a \equiv 0 \;\Rightarrow\; -a \equiv -0 = 0, \\
a \in I, b \in R \;\Rightarrow\; ab, ba \in I, & \text{denn } a \equiv 0 \;\Rightarrow\; ab \equiv 0b = 0.
\end{array}$$

Definition 3.5. *Eine nichtleere Teilmenge I eines Ringes R heißt **Ideal**, falls gilt:*

$$\begin{aligned}
a, b \in I \quad &\Rightarrow \quad a+b \in I, \\
a \in I \quad &\Rightarrow \quad -a \in I \\
a \in I, b \in R \quad &\Rightarrow \quad ab,\ ba \in I.
\end{aligned}$$

Ist umgekehrt I ein Ideal, so wird, wie man leicht nachprüft, in R durch $a \equiv b \;:\Leftrightarrow\; a - b \in I$ eine Äquivalenzrelation erklärt. Wir schreiben dann

$$a \equiv b \bmod I$$

und nennen a und b **kongruent modulo *I***. Die Relation ist mit Addition und Multiplikation verträglich, denn aus $a_1 - b_1,\ a_2 - b_2 \in I$ folgt $a_1 + a_2 - (b_1 + b_2) \in I$ und $a_1a_2 - b_1b_2 = a_1(a_2 - b_2) + (a_1 - b_1)b_2 \in I$. Insgesamt erkennen wir, daß eine eineindeutige Beziehung besteht zwischen Idealen und denjenigen Äquivalenzrelationen, die (3.1) erfüllen. Die Äquivalenzklassen nennt man **Restklassen**. Sie sind gegeben durch $\overline{a} = a + I$, $a \in R$, wir schreiben sie auch als $a \bmod I$. Der Ring der Restklassen wird als **Faktorring** R/I bezeichnet. Speziell erhalten wir bei der Wahl $R = \mathbb{Z}$ und $I = m\mathbb{Z}$ den Restklassenring $\mathbb{Z}_m$.

Die Abbildung $a \mapsto \overline{a}$ von R nach R/I ist, wie wir gesehen haben, mit Addition und Multiplikation verträglich. Man spricht von einem Homomorphismus.

Definition 3.6. *Seien R, R' Ringe. Eine Abbildung $\phi : R \to R'$ heißt **Ring-Homomorphismus**, falls $\phi(a+b) = \phi(a) + \phi(b)$ und $\phi(ab) = \phi(a) \cdot \phi(b)$ für alle $a, b \in R$ gilt. Ist ϕ bijektiv, so spricht man von einem **Isomorphismus** und schreibt $R \cong R'$.*

Ein Homomorphismus ϕ von R nach R' induziert seinerseits in R eine Äquivalenzrelation: $a \equiv b :\Leftrightarrow \phi(a) = \phi(b)$. Hier entsprechen Restklassen Elementen aus R', das zugehörige Ideal ist $\ker(\phi) := \{a \in R \,:\, \phi(a) = 0\}$, der Kern der Abbildung. Wir erhalten so den **Homomorphiesatz für Ringe**.

Satz 3.7. *Sei $\phi : R \to R'$ ein Homomorphismus, dann ist $\ker(\phi)$ ein Ideal und $R/\ker(\phi) \cong \operatorname{bild}(\phi) := \{\phi(a) \,:\, a \in R\}$.*

Wir beweisen nun eine allgemeine Version des Chinesischen Restsatzes. Dazu vereinbaren wir die folgende Sprechweise: Zwei Ideale I_1, I_2 heißen **teilerfremd**, falls

$$R \;=\; I_1 + I_2$$

gilt, mit $I_1 + I_2 \;:=\; \{r + s \,:\, r \in I_1, s \in I_2\}$.

Satz 3.8. *Sei R ein Ring mit Einselement 1 und seien $I_1, \ldots, I_k \subset R$ paarweise teilerfremde Ideale. Dann gibt es zu $a_1, \ldots, a_k \in R$ ein $a \in R$, so daß das Kongruenzensystem*

$$x \equiv a_1 \bmod I_1, \ldots, x \equiv a_k \bmod I_k$$

äquivalent ist zu der Kongruenz

$$x \equiv a \bmod I.$$

Dabei bezeichnet I das Ideal $I_1 \cap \ldots \cap I_k$.

Beweis. Wir gehen wie im ganzzahligen Fall vor. Da nach Voraussetzung $I_i + I_j = R$ für $i \neq j$, gibt es $b_{ij} \in I_i, c_{ij} \in I_j$, so daß $b_{ij} + c_{ij} = 1$. Setze

$$e_i \;:=\; \prod_{j \neq i} c_{ij} \;=\; \prod_{j \neq i} (1 - b_{ij}).$$

Dann gilt $e_i \equiv 0 \mod I_j$ für $i \neq j$ und $e_i \equiv 1 \mod I_i$. Folglich löst $a = \sum_i a_i e_i$ das Gleichungssystem. Ist a' eine weitere Lösung des Systems, so gilt $a' - a \in I_i$ für alle i, d.h. $a' - a \in I$ bzw. $a' \equiv a \mod I$. □

Der Satz läßt sich zu einer Isomorphieaussage von Ringen umformulieren. Wir benötigen dazu den Begriff des direkten Produktes von Ringen. Seien $R_1, \ldots, R_k$ Ringe und $R = R_1 \times \ldots \times R_k$ ihr kartesisches Produkt. Durch die Vereinbarung

$$\begin{aligned}(a_1, \ldots, a_k) + (b_1, \ldots, b_k) \;&:=\; (a_1 + b_1, \ldots, a_k + b_k),\\ (a_1, \ldots, a_k) \cdot (b_1, \ldots, b_k) \;&:=\; (a_1 b_1, \ldots, a_k b_k)\end{aligned}$$

wird offenbar R zum Ring. R heißt das **direkte Produkt** von $R_1, \ldots, R_k$.

Wir betrachten nun die Abbildung

$$a \bmod I \;\mapsto\; (a \bmod I_1, \ldots, a \bmod I_k)$$

von R/I nach $R/I_1 \times \ldots \times R/I_k$. Sie ist wohldefiniert, denn $a \bmod I = b \bmod I$ bzw. $b - a \in I$ impliziert $b - a \in I_i$ bzw. $a \bmod I_i = b \bmod I_i$ für alle i. Umgekehrt folgt $a \bmod I = b \bmod I$ aus $a \bmod I_i = b \bmod I_i$ für alle i, die Abbildung ist also injektiv. Nach dem letzten Satz gibt es zu $a_1, \ldots, a_k$ ein a, so daß $(a_1 \bmod I_1, \ldots, a_k \bmod I_k) = (a \bmod I_1, \ldots, a \bmod I_k)$. Daher ist die Abbildung surjektiv. Schließlich ist sie offenbar mit Addition und Multiplikation verträglich. Wir können daher den Chinesischen Restsatz auch als Aussage über die Isomorphie von Ringen ansehen,

$$R/\bigcap_{i=1}^{k} I_i \;\cong\; R/I_1 \times \ldots \times R/I_k.$$

Zum Abschluß gehen wir auf den Begriff Hauptideal ein.

Definition 3.9. *Sei R ein Integritätsbereich. Ein Ideal I heißt **Hauptideal**, wenn es ein $m \in R$ gibt, so daß $I = mR := \{mr \,:\, r \in R\}$. Dieses von m erzeugte Ideal wird mit (m) bezeichnet. Ist jedes Ideal in R ein Hauptideal, so heißt R **Hauptidealring**.*

Für ein Hauptideal $I = mR$ gilt $a \equiv b \mod I$ genau dann, wenn $m \mid b - a$; dies entspricht völlig der Situation in $\mathbb{Z}$. Jeder Hauptidealring besitzt ein Einselement. Zum Beweis bemerke man das R selbst ein Ideal ist. Es gibt also ein $m \in R$ mit $R = mR$ und damit ein Element $1 \in R$ mit $m = m1$. Außerdem gibt es zu jedem $a \in R$ ein $b \in R$, so daß $a = bm$. Es folgt $a = bm = bme = a1$, 1 ist also Einselement.

Proposition 3.10. *Jeder Euklidische Ring R ist ein Hauptidealring.*

Beweis. Das Nullideal $\{0\} = 0 \cdot R$ ist ein Hauptideal. Sei also $I \subset R$ ein Ideal ungleich $\{0\}$. Wähle $m \in I - \{0\}$ derart, daß $g(m)$ auf $I - \{0\}$ minimal ist. Für jedes $b \in I$ gibt es dann $s, r \in R$, so daß $b = sm + r$, mit $r = 0$ oder $g(r) < g(m)$. Es folgt $r \in I$, nach Wahl von m kann also nur $r = 0$ gelten. Daher gilt $b = sm$ und $I = mR$. $\square$

Von einem algebraischen Standpunkt unterscheiden sich Hauptidealringe nur unwesentlich von Euklidischen Ringen. So existieren in einem Hauptidealring R zu Elementen a und b immer ein größter gemeinsamer Teiler. Zum Beweis bilden wir das Ideal

$$I \;:=\; \{as + bt \,:\, s, t \in R\}.$$

Nach Annahme ist es von der Gestalt $I = dR$ für ein $d \in R$. Wegen $a, b \in I$ ist d Teiler von a und b. Andererseits gibt es $s, t \in R$, so daß $d = as + bt$. Jeder Teiler von a und b teilt daher auch d. d ist also ein größter gemeinsamer Teiler von a und b. Dieser Beweis bestätigt auch die Gültigkeit des Satzes von Bézout in Hauptidealringen. Ist d' ein weiterer ggT von a und b, so gibt es $e, e' \in R$, so daß $d = e'd'$ und $d' = ed$. Es folgt $d = e'ed$ und durch Kürzen $1 = e'e$. Daher sind e und e' Einheiten. Ein ggT ist also bis auf Einheiten eindeutig festgelegt. Schließlich stellen wir noch fest, daß $R = aR + bR$ genau dann gilt, wenn $1 = \mathrm{ggT}(a, b)$, wenn also a und b teilerfremd sind. Teilerfremdheit von Idealen, wie oben eingeführt, hat in Hauptidealringen daher die übliche Bedeutung.

Von einem algorithmischen Standpunkt sind dagegen Euklidische Ringe und Hauptidealringe wesentlich verschieden. Der Euklidische Algorithmus steht in Hauptidealringen nicht zur Verfügung.

Kapitel 4

Geheimer Nachrichtenaustausch

4.1 Prime Restklassen

Das Rechnen mit Restklassen hat wichtige Anwendungen in der Kryptographie gefunden, der Lehre von der geheimen Nachrichtenübertragung. Wie wir im letzten Kapitel festgestellt haben, ist eine Restklassen $\overline{a}$ in $\mathbb{Z}_m$ entweder Einheit oder Nullteiler. Wir untersuchen nun die Gruppe der Einheiten.

Definition 4.1. *Sei $a \in \mathbb{Z}$. Die Restklasse $\overline{a} \in \mathbb{Z}_m - \{0\}$ heißt* ***prime Restklasse*** *modulo m, falls $1 = \mathrm{ggT}(a, m)$. Die Menge der primen Restklassen modulo m wird mit $\mathbb{Z}_m^*$ bezeichnet, ihre Anzahl mit $\phi(m)$. ϕ heißt* ***Eulersche ϕ-Funktion****.*

Es gilt $\mathbb{Z}_m^* = \mathbb{Z}_m - \{0\}$ und $\phi(m) = m - 1$ genau dann, wenn m prim ist. Wir zeigen nun, daß $\mathbb{Z}_m^*$ bzgl. der Restklassenmultiplikation eine Gruppe ist.

Satz 4.2.

i) $\overline{a}, \overline{b} \in \mathbb{Z}_m^* \quad \Rightarrow \quad \overline{a} \cdot \overline{b} \in \mathbb{Z}_m^*$.

ii) Jedes $\overline{a} \in \mathbb{Z}_m^*$ besitzt ein inverses Element $\overline{a}^{-1} \in \mathbb{Z}_m^*$.

iii) Die Gleichung $\overline{a} \cdot \overline{x} = \overline{b}$ ist in $\mathbb{Z}_m^*$ eindeutig lösbar.

Beweis. Zu i): Aus $1 = \mathrm{ggT}(a, m) = \mathrm{ggT}(b, m)$ folgt $1 = \mathrm{ggT}(ab, m)$. Zu ii): Wegen $1 = \mathrm{ggT}(a, m)$ gibt es nach dem Satz von Bézout ganze Zahlen r, s, so daß $ar + ms = 1$. Es folgt $1 = \mathrm{ggT}(r, m)$ und $ar \equiv 1 \bmod m$, also $\overline{r} = \overline{a}^{-1}$. Zu iii): $\overline{x} = \overline{a}^{-1} \cdot \overline{b}$. □

Von fundamentaler Bedeutung ist das **Theorem von Euler**. Wir schreiben wie üblich

$$\overline{a}^{\,n} \;=\; \underbrace{\overline{a}\cdot\ldots\cdot\overline{a}}_{n\text{-mal}}.$$

Satz 4.3. *Für alle $\overline{a} \in \mathbb{Z}_m^*$ gilt*

$$\overline{a}^{\,\phi(m)} \;=\; \overline{1}.$$

Anders ausgedrückt: $\overline{a}^{\,-1} = \overline{a}^{\,\phi(m)-1}$.

Beweis. Seien $\overline{a}_1, \ldots, \overline{a}_{\phi(m)}$ alle Restklassen in $\mathbb{Z}_m^*$. Dann durchlaufen auch $\overline{a}\cdot\overline{a}_1, \ldots, \overline{a}\cdot\overline{a}_{\phi(m)}$ alle Restklassen in $\mathbb{Z}_m^*$, denn aus $\overline{a}\cdot\overline{a}_i = \overline{a}\cdot\overline{a}_j$ folgt durch Multiplikation mit $\overline{a}^{\,-1}$ die Gleichung $\overline{a}_i = \overline{a}_j$ und damit $i = j$. Daher gilt

$$\overline{a}_1\cdot\ldots\cdot\overline{a}_{\phi(m)} = \overline{a}^{\,\phi(m)}\cdot\overline{a}_1\cdot\ldots\cdot\overline{a}_{\phi(m)}$$

und, indem wir mit $\overline{a}_i^{\,-1}$, $i = 1, \ldots, \phi(m)$, multiplizieren,

$$\overline{a}^{\,\phi(m)} = \overline{1}.$$

□

Im Fall eines primen Moduls geht diese Aussage auf Fermat zurück.

Korollar 4.4. (Theorem von Fermat) *Sei p eine Primzahl. Dann gilt für jede ganze Zahl a*

$$a^p \equiv a \bmod p.$$

Beweis. Für $a \equiv 0 \bmod p$ ist die Aussage offensichtlich, und für $a \not\equiv 0$ gilt wegen $\phi(p) = p - 1$

$$a^{p-1} \equiv 1 \bmod p.$$

□

Die Eulersche ϕ-Funktion läßt sich leicht berechnen, vorausgesetzt wir kennen die Primfaktorzerlegung von m.

Proposition 4.5. *Seien $p_1, \ldots, p_k$ die unterschiedlichen Primteiler von m. Dann gilt*

$$\phi(m) \;= m\, \prod_{i=1}^{k}\left(1 - \frac{1}{p_i}\right).$$

Beweis. Ist $m = p^r$ mit primem p, so sind genau die Zahlen $0, p, 2p, \ldots, (p^{r-1} - 1)p$ Vertreter von Restklassen, die nicht zu $\mathbb{Z}_m^*$ gehören. Also:

$$\phi(p^r) \;=\; p^r - p^{r-1} \;=\; p^r\left(1 - \frac{1}{p}\right).$$

Sei nun $m = m_1 \cdot \ldots \cdot m_k$ mit paarweise teilerfremden Zahlen $m_1, \ldots, m_k$. Dann gilt $1 = \mathrm{ggT}(a, m)$ genau dann, wenn $1 = \mathrm{ggT}(a, m_i)$ für alle i. Nach dem Chinesischen Restsatz ist daher

$$a \bmod m \;\mapsto\; (a \bmod m_1, \ldots, a \bmod m_k)$$

eine Bijektion von $\mathbb{Z}_m^*$ nach $\mathbb{Z}_{m_1}^* \times \ldots \times \mathbb{Z}_{m_k}^*$. Es folgt

$$\phi(m) \;=\; \phi(m_1) \cdot \ldots \cdot \phi(m_k).$$

□

4.2 Öffentliche Chiffriersysteme

Die **Kryptographie** ist die Lehre von den Chiffriersystemen. Stellen wir uns vor, daß eine Person A eine geheime Nachricht an die Person B übermitteln möchte. A kodiert sie deswegen mittels einer bijektiven Kodierabbildung

$$\kappa : \mathcal{N} \to \mathcal{K}.$$

Anstelle der Nachricht $a \in \mathcal{N}$ im Klartext sendet A die chiffrierte Nachricht $\kappa(a)$. Der Empfänger dekodiert mittels der inversen Abbildung

$$\kappa^{-1} : \mathcal{K} \to \mathcal{N}.$$

Ein bekanntes Verfahren beruht auf der Addition modulo 2. Wir nehmen an, daß die Nachrichten als 01-Folgen der Länge n vorliegen, wählen also $\mathcal{N} = \mathcal{K} = \{0, 1\}^n$. Wir fassen die Folgen als Vektoren der Läne n über dem Körper $\mathbb{Z}_2$ auf. Zum Kodieren wird ein 01-String $k = k_1 k_2 \ldots k_n$ verwendet. Wir setzen

$$\kappa(a) \;:=\; a + k,$$

die beiden 01-Folgen a und k werden also komponentenweise modulo 2 addiert. Es gilt $\kappa^{-1} = \kappa$, denn $k + k$ ist die nur aus Nullen bestehende Folge. Nachrichten werden daher nach demselben Verfahren dekodiert. Dieses klassische Verfahren hat den Namen ‚One-time-pad'. Es gilt $a + \kappa(a) = k$, gelingt es daher, eine kodierte Nachricht $k(a)$ zu entschlüsseln, so kennt man k und damit bereits die vollständige Kodier- und Dekodiervorschrift. Dies bedeutet, daß das Verfahren sicher ist, man hat keine Chance, eine Nachricht zu dechiffrieren, wenn man auf k keinen Zugriff hat. Ist k rein zufällig aus $\{0, 1\}^n$ gewählt, so gilt dies auch für $\kappa(a)$.

Das RSA-Schema

Für die klassischen Chiffrierverfahren besteht ein Sicherheitsproblem darin, daß man nicht nur die Dekodiervorschrift κ^{-1} geheimhalten muß, sondern auch das Kodierverfahren κ. Wie das Beispiel des One-time-pad zeigt, lassen sich die Nachrichten mit Hilfe der Chiffrier- wie der Dechiffriervorschrift leicht entschlüsseln. Diffie und Hellman haben daher 1976 einen (damals beherzten) Vorschlag gemacht: Man solle Kodierabbildungen κ benutzen, für die $b = \kappa(a)$ aus a leicht berechnet werden kann, umgekehrt aber die Berechnung von $a = \kappa^{-1}(b)$ aus b typischerweise mit einem immensen Rechenaufwand verbunden ist, der praktisch nicht bewältigt werden kann (selbst wenn einem die Abbildung κ bekannt ist!). In dieser asymmetrischen Situation darf man das Chiffrierverfahren öffentlich bekannt machen, ohne die Sicherheit zu gefährden. Dies ist die Grundidee der modernen, computergestützten **öffentlichen Kodiersysteme**. Theoretisch ist es schwer zu begründen, daß solche Abbildungen κ, man spricht von **Einweg- (one-way) Abbildungen**, wirklich existieren. Für praktische Zwecke haben sich jedoch verschiedene Abbildungen als brauchbar erwiesen. Wir werden Abbildungen betrachten, bei denen das Dekodieren erst dann praktisch durchführbar ist, wenn man über einen zusätzlichen ‚Schlüssel' verfügt. Dann spricht man von **Falltür- (trapdoor) Abbildungen**.

Zum Beispiel kann man schnell modulo m potenzieren:

$$a^{20} = a^{16+4} = (((a^2)^2)^2)^2 \cdot (a^2)^2.$$

Allgemein berechnet man a^c modulo m, indem man den Exponenten als binäre Zahl darstellt, $c = d_0 + d_1 \cdot 2 + \ldots + d_k \cdot 2^k$ mit $d_i \in \{0,1\}$, in k Schritten rekursiv die Potenzen $a^{2^i} = (a^{2^{i-1}})^2$ modulo m berechnet und schließlich diejenigen Potenzen modulo m multipliziert, für die $d_i = 1$ gilt. Die Anzahl der Operationen ist von der Ordnung $O(k) = O(\log c)$. – Eine allgemeine Methode, mit der man schnell (schneller als Durchprobieren) die Gleichung $b \equiv a^x \bmod m$ nach x auflöst (‚diskreter Logarithmus'), kennt man dagegen nicht.

Wir beschreiben nun das bekannteste öffentliche Chiffriersystem, das von Rivest, Shamir und Adleman 1978 vorgeschlagene **RSA-System**. Es baut darauf auf, daß es schwer ist, eine Zahl m in ihre Primfaktoren zu zerlegen.

- Als Nachrichtenmenge $\mathcal{N} = \mathcal{K}$ wird $\mathbb{Z}_m^*$ gewählt. Dabei sei $m = pq$ das Produkt zweier sehr großer Primzahlen p und q. Wir gehen also davon aus, daß die Nachricht als natürliche Zahl $a < m$ dargestellt ist.

- Zum Kodieren wird eine natürliche Zahl $s < \phi(m)$ gewählt, die teilerfremd zu $\phi(m)$ ist. Die Kodierung der Nachricht $a \in \mathcal{N}$ ist dasjenige

$\kappa(a) \in \mathcal{N}$, so daß

$$\kappa(a) \equiv a^s \bmod m.$$

- Zum Dekodieren benötigt man einen Schlüssel, eine Zahl t mit der Eigenschaft

$$t \cdot s \equiv 1 \bmod \phi(m).$$

  Eine solche Zahl existiert, denn nach Wahl von s gilt $\overline{s} \in \mathbb{Z}^*_{\phi(m)}$. Aus $st = 1 + k\phi(m)$ folgt dann unter Beachtung des Theorems von Euler

$$(a^s)^t \equiv a \cdot a^{k\phi(m)} \equiv a \bmod m.$$

  Es gilt also

$$\kappa^{-1}(b) \equiv b^t \bmod m. \qquad \square$$

Man darf dieses Verfahren so einrichten, daß die Nachrichtenmenge $\mathcal{N}$, die Zahl s und auch die kodierte Nachricht $\kappa(a)$ öffentlich zugänglich sind. Nur der Schlüssel t muß geheim bleiben. Bei Kenntnis von $\phi(m)$ läßt sich t natürlich mühelos mit Euklids Algorithmus aus s errechnen. Daher muß dafür gesorgt sein, daß $\phi(m)$ praktisch nicht berechenbar ist. Diese Aufgabe ist aber im wesentlichen so aufwendig wie das Zerlegen von m in seine beiden Primfaktoren. Einerseits gilt nämlich

$$\phi(m) = (p-1)(q-1),$$

andererseits kann man m bei Kenntnis von $\phi(m)$ sofort faktorisieren: p und q erhält man aus den Gleichungen

$$\begin{aligned} p + q &= m - \phi(m) + 1, \\ p - q &= \pm\sqrt{(p+q)^2 - 4pq} \\ &= \pm\sqrt{(m - \phi(m) + 1)^2 - 4m}. \end{aligned}$$

Das RSA-System steht und fällt damit, daß die Primfaktoren p und q geheim bleiben. Die Erfahrung zeigt, daß die Faktorisierung von m einen hohen Rechenaufwand erfordert und praktisch nicht mehr gelingt, wenn p und q ausreichend groß sind. Dies ist ein *Grundpostulat der Kryptographie*: Man geht davon aus, daß es keinen schnellen Algorithmus zum Zerlegen einer Zahl in seine Primfaktoren gibt. Man nennt deshalb in der Kryptographie ein Chiffriersystem ‚beweisbar sicher', wenn es letztlich auf die Faktorisierung großer Zahlen zurückgeführt werden kann.

Signatur von Nachrichten

Das RSA-Schema eignet sich gut zum geheimen Austausch von Nachrichten innerhalb einer Gruppe von Teilnehmern. Jeder Teilnehmer A erhält von einer Zentrale einen öffentlichen Schlüssel $(m(A), s(A))$ und seinen persönlichen geheimen Schlüssel $t(A)$. A teilt B eine geheime Nachricht a als $b := \kappa_B(a) \equiv a^{s(B)} \bmod m(B)$ mit, B dekodiert mittels $\kappa_B^{-1}(b) \equiv b^{t(B)} \bmod m(B)$. Ein vorheriger Kontakt zwischen A und B ist nicht erforderlich, A braucht lediglich den öffentlich zugänglichen Schlüssel $s(B)$ von B. Ein wichtiger Vorteil des Systems ist, daß es den Teilnehmern erlaubt, Mitteilungen fälschungssicher zu signieren (‚Unterschriftensystem'). A beglaubigt eine öffentliche Nachricht a durch die öffentliche Mitteilung von $c = \kappa_A^{-1}(a) \equiv a^{t(A)} \bmod m(A)$. Alle anderen Teilnehmer können verifizieren, daß $a \equiv c^{s(A)} \bmod m(A)$ gilt.

Randomisiertes RSA

Um ein einzelnes bit $a = 0$ oder 1 geheim zu übertragen, muß man das RSA-Schema modifizieren. Dann gilt $a^s = a$, und die Nachricht bleibt unverschlüsselt. Stattdessen kann man so vorgehen: Man wähle rein zufällig eine gerade Zahl $2u$ aus $\{0, 1, \ldots, (m-1)\}$ und übertragt anstelle von a mit dem RSA-Schema die Zahl $2u + a$, die verschlüsselte Nachricht ist also $(2u + a)^s$. Der Empfänger kann mit seinem geheimen Schlüssel die Nachricht als $2u + a$ dekodieren und a als das letzte bit dieser Nachricht zurückgewinnen. Alexi, Chor, Goldreich und Schnorr haben gezeigt, daß dieses Verfahren genauso sicher wie das RSA-Schema ist.

4.3 Ein probabilistischer Primzahltest

Für das RSA-Schema wie für die Anwendungen der modularen Arithmetik im letzten Kapitel braucht man ein effektives Verfahren zur Gewinnung großer Primzahlen. Nach dem Primzahlsatz ist in dem Intervall $[n, 2n]$ im Mittel jede $(\ln n)$-te Zahl eine Primzahl. Durch zufällige Wahl von Zahlen aus dem Intervall wird man also ausreichend oft auf Primzahlen treffen. Um sie zu erkennen, benötigt man einen effektiven Primzahl-Test. Das folgende probabilistische Verfahren testet, ob die Zahl m die Fermatsche Identität $a^{m-1} \equiv 1 \bmod m$ erfüllt. Nach dem Fermatschen Theorem genügen Primzahlen dieser Identität, sie sind aber nicht die einzigen Zahlen.

Definition 4.6. *Die ganze Zahl m heißt **Carmichael-Zahl**, falls m nicht prim ist, und falls für alle $a \in \mathbb{Z}_m^*$ gilt*

$$a^{m-1} \equiv 1 \bmod m.$$

Die kleinste Carmichael-Zahl ist $561 = 3 \cdot 11 \cdot 17$. Computerberechnungen haben gezeigt, daß sie sehr viel seltener sind als Primzahlen. Lehmer und Poulet haben festgestellt, daß es unter den ersten 100.000.000 Zahlen 5.761.455 Primzahlen gibt, aber nur 252 Carmichael-Zahlen. – Zahlen die weder prim noch Carmichael-Zahlen sind, lassen sich mit ausreichender Sicherheit erkennen.

Proposition 4.7. *Ist m weder prim noch Carmichael-Zahl, so gilt*

$$\operatorname{card}\{a \in \mathbb{Z}_m^* \,:\, a^{m-1} \equiv 1 \bmod m\} \leq \phi(m)/2.$$

Beweis. Seien $a_1, \ldots, a_k$ alle Elemente aus $\mathbb{Z}_m^*$ mit $a_i^{m-1} \equiv 1 \bmod m$. Nach Voraussetzung gibt es ein $b \in \mathbb{Z}_m^*$ mit $b^{m-1} \not\equiv 1 \bmod m$. Dann gilt für $b_i = ba_i$

$$b_i^{m-1} = a_i^{m-1} b^{m-1} \not\equiv 1 \bmod m,$$

es gibt daher mindestens ebenso viele Elemente in $\mathbb{Z}_m^*$, die die Fermatsche Identität nicht erfüllen. Also gilt $k \leq \phi(m)/2$. □

Ein probabilistischer Primzahltest.

Eingabe: Eine natürliche Zahl m.

Ausgabe: ‚m ist prim' oder ‚m ist nicht prim'.

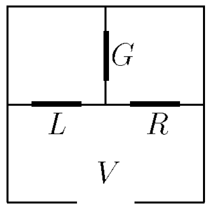
Verfahren: Wähle rein zufällig Zahlen $a_1, \ldots, a_k < m$, die teilerfremd zu m sind. Gilt $a_i^{m-1} \equiv 1 \bmod m$ für alle $i = 1, \ldots, k$, so entscheide ‚m ist prim'. Andernfalls gilt ‚m ist nicht prim'. □

Dieser Test erkennt sicher Primzahlen. Ein Zahl, die weder prim noch Carmichal-Zahl ist wird mit ausreichender Sicherheit erkannt, die Fehlerwahrscheinlichkeit ist kleiner als 2^{-k}. Nur Carmichael-Zahlen bleiben unerkannt. Praktisch ist dies für das Verfahren bedeutungslos, weil Carmichael-Zahlen so überaus selten auftreten. – Ein raffinierteres Verfahren von Rabin beseitigt diese letzte Unsicherheit (vgl. die Monographie von Knuth Vol. II, 3. Ausgabe, S. 395).

4.4 Zero-Knowledge Beweise

Eine Person möchte beweisen (und sich dadurch ausweisen), daß sie ein Geheimnis (eine Geheimzahl) kennt, ohne von ihrem Wissen irgend etwas preiszugeben. Wie ist das möglich?

Um die Idee des Verfahrens zu erläutern, betrachten wir das folgende Szenario: Die Person A möchte einer zweiten Person B demonstrieren, daß sie den Schlüssel zu einer Geheimtür G besitzt, die zwei Räume verbindet. Den Schlüssel möchte A nicht vorzeigen, ja nicht einmal einen Blick in die Räume gestatten. Deswegen einigen sich A und B auf folgendes Vorgehen: A geht durch den Vorraum V und dann rein zufällig entweder durch die Tür R in den rechten Raum, oder die Tür L in den linken Raum.



Nun betritt B den Vorraum. Sie weiß nicht, in welchem der beiden Räume sich A befindet. Sie wählt zufällig eine der Türen R oder L aus und fordert A auf, durch diese Tür wieder herauszukommen. Dies ist A immer möglich, wenn sie die Geheimtür öffnen kann. Andernfalls besteht sie den Test nur mit Wahrscheinlichkeit 1/2, und bei r-facher unabhängiger Wiederholung sogar nur mit Wahrscheinlichkeit 2^{-r}. Auch für kleines r bietet dieses Verfahren schon hohe Sicherheit.

Der Fiat-Shamir-Algorithmus

Ganz ähnlich funktioniert der Algorithmus, den Fiat und Shamir 1988 vorgeschlagen haben, und bei dem der Schlüssel eine geheime Zahl ist. Der Algorithmus beruht darauf, daß es sehr aufwendig ist, die Wurzel eines quadratischen Restes modulo einem Primzahlprodukt $m = pq$ zu berechnen. Zuerst erläutern wir diesen Sachverhalt. a heißt **quadratischer Rest modulo m**, falls die Kongruenz

$$x^2 \equiv a \bmod m \tag{4.1}$$

eine Lösung b besitzt, wenn also die Restklassengleichung $\overline{b}^2 = \overline{a}$ gilt. $\overline{b}$ nennen wir dann eine **Wurzel** von $\overline{a}$.

Beispiel. $\mathbb{Z}_{15}^* = \{\overline{1}, \overline{2}, \overline{4}, \overline{7}, \overline{8}, \overline{11}, \overline{13}, \overline{14}\}$ enthält die quadradischen Reste $\overline{1}$ und $\overline{4}$:

$$\begin{array}{ccccccccc} \overline{1}^2 & = & \overline{4}^2 & = & \overline{11}^2 & = & \overline{14}^2 & = & \overline{1}, \\ \overline{2}^2 & = & \overline{7}^2 & = & \overline{8}^2 & = & \overline{13}^2 & = & \overline{4}. \end{array}$$

In dem Beispiel besitzt die Gleichung (4.1) für die quadratischen Reste $\overline{1}$ und $\overline{4}$ vier verschiedene Lösungen. Wir zeigen nun, daß dies allgemein für Moduln $m = pq$ mit zwei Primzahlen $p, q > 2$ gilt. Die Gleichung $\overline{x}^2 = \overline{a} = \overline{b}^2$ in $\mathbb{Z}_m^*$ ist äquivalent zu der Aussage $pq \mid (x-b)(x+b)$. Daher teilt p entweder $x-b$ oder $x+b$, und ebenso q entweder $x-b$ oder $x+b$. Es sind vier Fälle möglich:

$$\begin{array}{lll} x \equiv b \bmod p, & x \equiv b \bmod q, & \text{oder} \\ x \equiv -b \bmod p, & x \equiv -b \bmod q, & \text{oder} \\ x \equiv b \bmod p, & x \equiv -b \bmod q, & \text{oder} \\ x \equiv -b \bmod p, & x \equiv b \bmod q. & \end{array}$$

Nach dem Chinesischen Restsatz sind alle Fälle eindeutig lösbar, und die Lösungen sind voneinander verschieden. Die beiden ersten Kongruenzen haben offenbar die Lösungen b und $-b$, die beiden anderen Kongruenzen führen zu weiteren Lösungen c und $-c$. b und c unterscheiden sich nicht nur um das Vorzeichen, sie heißen **wesentlich verschiedene Lösungen**.

Wir begründen nun, daß es genauso rechenintensiv ist, (4.1) im Fall $m = pq$ zu lösen (selbst wenn bekannt ist, daß a quadratischer Rest ist), wie m in seine Primfaktoren zu zerlegen. Nehmen wir an, daß wir über einen Algorithmus verfügen, der zu jedem quadratischen Rest modulo m eine Wurzel berechnet. Wir können dann auch leicht zwei wesentlich verschiedene Wurzeln eines quadratischen Restes finden. Dazu wähle man rein zufällig ein $\overline{b}$ aus $\mathbb{Z}_m^*$ und bilde daraus den quadratischen Rest $\overline{a} = \overline{b}^2$. Mit dem Algorithmus erhalten wir eine Wurzel $\overline{c}$ von $\overline{a}$, die mit Wahrscheinlichkeit $1/2$ wesentlich verschieden von $\overline{b}$ ist (im gegenteiligen Fall wähle zufällig ein neues $\overline{b}$). In $\mathbb{Z}_m^*$ gilt dann $(\overline{b}+\overline{c})(\overline{b}-\overline{c}) = \overline{0}$, d.h. $pq \mid (b+c)(b-c)$. Sind $\overline{b}$ und $\overline{c}$ wesentlich verschieden, so können p und q nicht gleichzeitig $b+c$ teilen, und auch nicht gleichzeitig $b-c$ teilen. Es gilt also $\mathrm{ggT}(m, b+c) = p$ oder $\mathrm{ggT}(m, b+c) = q$. Daher läßt sich p oder q mit dem Euklidischen Algorithmus berechnen und m problemlos faktorisieren.

Das Fazit ist: Das Ziehen von Quadratwurzeln in $\mathbb{Z}_m^*$ ist genauso rechenintensiv wie das Faktorisieren von m. Ist es ausgeschlossen, m in Primfaktoren zu zerlegen, so kann man auch keine Quadratwurzeln modulo m berechnen.

Dies macht sich der Fiat-Shamir Algorithmus zu Nutze. Verschiedene Personen können sich beteiligen. Wir stellen uns vor, daß eine Zentrale geheime Schlüssel verteilt. Sie wählt zwei Primzahlen p und q und veröffentlicht $m = pq$. p und q sollen geheim bleiben und müssen deswegen so groß sein, daß sich m praktisch nicht in seine Primteiler zerlegen läßt. Die Zentrale weist jedem Teilnehmer eine geheime Zahl b zu, die zu einer primen Restklasse modulo m gehört. Gleichzeitig wird die Restklasse $\overline{a} = \overline{b}^2$ veröffentlicht.

A möchte nun B davon überzeugen, daß sie die Geheimzahl b kennt. Dazu führen A und B den folgenden ‚Dialog':

- A wählt rein zufällig $\overline{r} \in \mathbb{Z}_m^*$ und bildet $\overline{s} = \overline{r}^{\,2}$. A teilt B die Zahl s mit.
- B wählt rein zufällig die Zahl $t = 0$ oder 1. Falls $t = 0$, fragt er A nach einer Wurzel $\overline{w}$ von $\overline{s}$, andernfalls fragt er nach einer Wurzel $\overline{w}$ von $\overline{a} \cdot \overline{s}$.
- Ist $t = 0$, so gibt A an B die Antwort $\overline{w} = \overline{r}$, ist dagegen $t = 1$, so gibt A als Antwort $\overline{w} = \overline{b} \cdot \overline{r}$.
- B überprüft die Antwort: $\overline{r}^{\,2} = \overline{s}$ im Falle $t = 0$, und $(\overline{b} \cdot \overline{r})^{\,2} = \overline{a} \cdot \overline{s}$ im Falle $t = 1$.

Kennt A die Geheimzahl, so bereitet ihm sein Teil des Dialogs kein Problem. Kennt A die Geheimzahl dagegen nicht, so wird er höchstens eine der Fragen von B beantworten können. Sonst hätte A nämlich sowohl eine Wurzel $\overline{x}$ von $\overline{s}$ als auch eine Wurzel $\overline{y}$ von $\overline{a} \cdot \overline{s}$ zur Verfügung, die sich zu einer Wurzel $\overline{x}^{\,-1}\overline{y}$ von $\overline{a}$ zusammensetzen ließen. Das Ziehen von Wurzeln quadratischer Reste modulo m ist aber, wie wir gesehen haben, genau so schwer wie das Faktorisieren von m. Ein Schwindler A wird daher r Dialogrunden mit dem Verifizierer B höchstens mit Wahrscheinlichkeit 2^{-r} bestehen.

Schießlich wird in dem Dialog keine Information über b preisgegeben. Wie kann man dies einsehen? Protokolliert B den Dialog, so kann das Protokoll nur die Zahlen s, t, w enthalten. Sie haben die folgenden Eigenschaften:

- $\overline{s}$ ist Quadrat eines rein zufälligen Elements aus $\mathbb{Z}_m^*$, ein zufälliger quadratischer Rest modulo m.
- t ist rein zufällig aus $\{0, 1\}$ gewählt.
- $\overline{w}^{\,2} = \overline{s}$ oder $\overline{a} \cdot \overline{s}$, je nachdem $t = 0$ oder 1 ist.

Solch ein Protokoll kann nun B aber auch ohne die Hilfe von A simulieren, nämlich so:

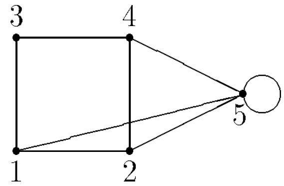
- B wählt rein zufällig $t \in \{0, 1\}$ und ein $\overline{w} \in \mathbb{Z}_m^*$. B bildet daraus $\overline{s} = \overline{w}^{\,2}$ bzw. $\overline{w}^{\,2} \cdot \overline{a}^{\,-1}$, je nachdem, ob $t = 0$ oder 1 ist.
- B erstellt das Protokoll (s, t, w).

Dieses Protokoll ist ohne Kenntnis von b erstellt und hat dieselben Eigenschaften wie das Protokoll aus dem Dialog zwischen A und B. Deswegen kann auch der Dialog keine Informationen über b tragen. Man sagt, der Dialog besitzt die **Zero-Knowledge-Eigenschaft**.

Zero-Knowledge Beweis für ein NP-vollständiges Problem

Wir betrachten nun ein Entscheidungsproblem, von dem man gute Gründe hat anzunehmen, daß es zu den Problemen zählt, die den höchsten Rechenaufwand erfordern. Es handelt sich darum zu entscheiden, ob ein gegebener endlicher Graphen G 3-färbbar ist, d.h. ob man seine Ecken so mit drei verschiedenen Farben anmalen kann, daß benachbarte Ecken immer verschieden gefärbt sind.

Ein **Graph** $G = (E, K)$ besteht bekanntlich aus einer Menge E, den **Ecken**, und einer Menge $K \subset \{\{a, b\} \ : \ a, b \in E\}$, den **Kanten**, die jeweils zwei Ecken verbinden. Ecken $a \neq b$ heißen **benachbart**, falls sie durch eine Kante verbunden sind, falls also $\{a, b\} \in K$ gilt. Für $E = \{1, 2, 3, 4, 5\}$ und $K = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{2, 5\}, \{3, 4\}, \{4, 5\}, \{5, 5\}\}$ ergibt sich das folgende Bild



Ob ein Graph 2-färbbar ist, läßt sich schnell feststellen, man probiert es, ausgehend von einer Ecke, einfach aus. Nur ganz am Anfang steht einem die Wahl der Farbe offen. Bei 3 Farben ist dies anders, man muß beim Färben der Ecken immer wieder zwischen Alternativen entscheiden, und führt ein Versuch nicht zum Erfolg, so bedeutet dies noch nicht, daß der Graph nicht 3-färbbar ist. Es stellt sich heraus, daß die 3-Färbbarkeit von Graphen ein ***NP*-vollständiges Entscheidungsproblem** ist. Dies ist die Klasse der Entscheidungsprobleme, die nach dem heutigen Kenntnisstand den größten Rechenaufwand erfordern. Wir gehen darauf später genauer ein.

Wir beschreiben nun, wie eine Person A eine andere Person B mit beliebig hoher Sicherheit davon überzeugen kann, daß sie eine 3-Färbung eines Graphen G kennt, ohne von ihrem Wissen irgend etwas zu verraten. Dieser Dialog erfordert von A eine sorgfältige Vorbereitung.

- Für jede Ecke $e \in E$ des Graphen erzeugt A eine RSA-Kodierung $m_e = p_e q_e, s_e, t_e$. Sie gibt ihrem Gegenüber m_e und s_e für alle e bekannt.

- Weiter generiert A eine zufällige Permutation der drei Farben und damit eine neue zufällige 3-Färbung des Graphen. Die permutierte Farbe jeder Ecke e läßt sich durch 2 bits $b_e b'_e$ beschreiben. A kodiert diese bits mit dem randomisierten RSA-Schema und gibt sie kodiert als $y_e \equiv (2u + b_e)^{s_e}$ bzw. $y'_e \equiv (2u' + b'_e)^{s_e}$ modulo m_e bekannt.

- Nun wählt B zufällig zwei benachbarte Ecken $e(1), e(2)$, zu denen A ihm die Schlüssel $t_{e(1)}$ und $t_{e(2)}$ überläßt. B kann dann die Farben $b_{e(1)}b'_{e(1)}$ und $b_{e(2)}b'_{e(2)}$ dekodieren und feststellen, ob sie, wie von A behauptet, verschieden sind.

Hat A eine 3-Färbung des Graphen, so kann sie die Frage von B immer beantworten. Dagegen wird ein Schwindler, der keine 3-Färbung kennt, mit einer Wahrscheinlichkeit von mindestens k^{-1} überführt, wobei k die Anzahl der Kanten bezeichne. r unabhängige Runden übersteht er nur mit Wahrscheinlichkeit $(1-k^{-1})^r \leq e^{-r/k}$.

B lernt aus diesem Dialog nichts. Er erfährt von A ein zufälliges Paar von verschiedenen Farben, das zudem Runde für Runde unabhängig ist. Solche Paare kann er sich auch verschaffen, ohne eine 3-Färbung des Graphen zu kennen. Folglich trägt der Dialog keine Information über die 3-Färbung.

4.5 Gruppen

Die Sätze von Fermat und Euler sind Spezialfälle eines allgemeinen Satzes der Gruppentheorie, den wir nun behandeln wollen. Im folgenden sei G eine Gruppe mit neutralem Element e.

Definition 4.8. *$H \subset G$ heißt **Untergruppe** von G, falls $a \cdot b \in H$ und $a^{-1} \in H$ für alle $a, b \in H$. Als **Links-** und **Rechtsnebenklassen** von H bezeichnet man die Mengen*

$$aH := \{ab : b \in H\} \quad \textit{bzw.} \quad Ha := \{ba : b \in H\},$$

wobei a durch G läuft.

Beispiel. Die Menge H der quadratischen Reste bilden eine Untergruppe von $\mathbb{Z}_m^*$. Gilt nämlich $\overline{a} = \overline{b}^2$ und $\overline{a'} = \overline{b'}^2$, so folgt $\overline{a} \cdot \overline{a'} = (\overline{b} \cdot \overline{b'})^2$ und $\overline{a}^{-1} = (\overline{b}^{-1})^2$. Wegen $\overline{1}^2 = \overline{-1}^2$ ist $H \neq \mathbb{Z}_m^*$, sofern $m > 2$. □

Zwei Linksnebenklassen a_1H und a_2H sind entweder disjunkt oder gleich: Nehmen wir an daß $a_1H \cap a_2H \neq \emptyset$. Dann gilt $a_1b_1 = a_2b_2$ für geeignete $b_1, b_2 \in H$. Es folgt $a_1b = a_2b_2b_1^{-1}b \in a_2H$ für alle $b \in H$. Deswegen gilt $a_1H \subset a_2H$, und analog $a_2H \subset a_1H$.

Es ist also G Vereinigung von disjunkten Nebenklassen. Außerdem sind alle Nebenklassen gleichmächtig,

$$\operatorname{card} aH = \operatorname{card} H,$$

denn die Abbildung $b \mapsto ab$ ist eine Bijektion von H nach aH (die Umkehrabbildung ist $b \mapsto a^{-1}b$).

Dies führt uns unmittelbar zum **Satz von Lagrange**. Sei dazu $|G|$ die **Ordnung** von G, d.h. die Anzahl ihrer Elemente, und sei $[G:H]$ der **Index** der Untergruppe H, d.h. die Anzahl der verschiedenen (disjunkten) Nebenklassen von H in G.

Satz 4.9. *Sei G endliche Gruppe und H Untergruppe. Dann gilt*

$$|G| \;=\; [G:H]\cdot|H|.$$

Insbesondere ist $|G|$ ein Vielfaches von $|H|$.

Beispiel. Die Anzahl der quadratischen Reste ist für $m > 2$ ein echter Teiler der Euler-Funktion $\phi(m)$. □

Wir betrachten nun spezielle Untergruppen. Sei $a \in G$. Dann ist die kleinste Untergruppe von G, die a enthält, gegeben durch

$$H_a \;:=\; \{\ldots, a^{-2}, a^{-1}, a^0 = e, a, a^2, \ldots\}.$$

Wir nehmen an, daß H_a endlich ist. Dann gibt es ganze Zahlen $s < t$, so daß $a^s = a^t$, also $a^{t-s} = e$. Ist k die kleinste natürliche Zahl mit der Eigenschaft

$$a^k \;=\; e,$$

so folgt

$$H_a \;=\; \{a, a^2, \ldots, a^{k-1}, e\},$$

insbesondere $(a^j)^{-1} = a^{k-j}$ und $a^s = e \Leftrightarrow k \mid s$. k heißt die **Ordnung** von a und wird mit $\mathrm{ord}(a)$ bezeichnet. $\mathrm{ord}(a)$ ist offenbar die Ordnung von H_a. Nach dem Satz von Lagrange ist die Ordnung von a ein Teiler der Ordnung von G. Daher folgt

$$a^{|G|} \;=\; e. \tag{4.2}$$

Im Fall $G = \mathbb{Z}_m^*$ ergibt sich erneut den **Satz von Euler**:

$$\overline{a}^{\,\phi(m)} \;=\; \overline{1}.$$

Ist $\mathrm{ord}(a) = |G|$, also $H_a = G$, so heißt G **zyklische Gruppe** und a ein **erzeugendes Element** von G.

Kapitel 5

Primfaktorzerlegung

5.1 Faktorielle Ringe

In diesem Abschnitt betrachten wir Ringe mit der Eigenschaft, daß sich ihre Elemente eindeutig in Primfaktoren zerlegen lassen. Wir unterscheiden zwischen zwei Eigenschaften von Primteilern.

Definition 5.1. *Sei R ein Integritätsbereich mit Einselement und sei R^* die Menge seiner Einheiten.*

i) $p \in R$ heißt ***irreduzibel****, falls $p \neq 0, p \notin R^*$ und für alle $a, b \in R$*

$$p = ab \;\Rightarrow\; a \in R^* \text{ oder } b \in R^*.$$

p heißt ***prim*** *oder* ***Primelement****, falls $p \neq 0, p \notin R^*$ und für alle $a, b \in R$*

$$p \mid ab \;\Rightarrow\; p \mid a \text{ oder } p \mid b.$$

ii) R heißt ***faktoriell****, falls jedes Element, welches ungleich 0 und keine Einheit ist, Primelement oder Produkt von endlich vielen Primelementen ist.*

Beispiel. Der Integritätsbereich $\mathbb{Z}[\sqrt{-5}] := \{x + y\sqrt{-5} \;:\; x, y \in \mathbb{Z}\}$ ist nicht faktoriell. Es gilt z.B.

$$6 \;=\; 2 \cdot 3 \;=\; (1 + \sqrt{-5}) \cdot (1 - \sqrt{-5}).$$

Die Faktoren sind irreduzibel, aber nicht prim. Zum Nachweis dieser Behauptung ziehen wir die ‚Norm' $N(x + y\sqrt{-5}) := |x + y\sqrt{-5}|^2 = x^2 + 5y^2$ des Elementes $x + y\sqrt{-5}$ heran. Sie ist multiplikativ, $N(ab) = N(a)N(b)$, und sie nimmt nur ganzzahlige Werte an. Nun gilt $N(2) = 4, N(3) = 9$ und $N(1+\sqrt{-5}) = N(1-\sqrt{-5}) = 6$. Daher teilt etwa $1+\sqrt{-5}$ zwar das Produkt $2 \cdot 3$, aber weder 2 noch 3 (denn $N(1+\sqrt{-5})$ teilt weder $N(2)$ noch $N(3)$). Es handelt sich also nicht um Primelemente. Andererseits prüft man leicht nach, daß die Faktoren sich nicht weiter zerlegen lassen (denn Elemente der Norm 2 oder 3 gibt es nicht), die Faktoren sind also irreduzibel.

In diesem Ring hat man im allgemeinen keine größten gemeinsamen Teiler: 6 und $2+2\sqrt{-5}$ haben die gemeinsamen Teiler 2 und $1+\sqrt{-5}$, die ihrerseits teilerfremd sind. $\square$

Im Allgemeinen ist es also nicht gleichbedeutend, daß ein Element irreduzibel oder prim ist. Es bestehen die folgenden Zusammenhänge.

Proposition 5.2. *Es gilt:*

i) Jedes Primelement ist irreduzibel.

ii) In faktoriellen Ringen ist jedes irreduzible Element prim.

iii) Eine Zerlegung eines Elements a in Primfaktoren ist eindeutig, bis auf Einheiten und die Reihenfolge. Genauer: Hat a die Primfaktorzerlegungen $a = p_1 \cdot \ldots \cdot p_r = q_1 \cdot \ldots \cdot q_s$, so folgt $r = s$, und nach einer geeigneten Vertauschung der q_i untereinander gilt $p_i = e_i q_i$ für alle i, mit Einheiten e_i.

Beweis. i) Ist $p = ab$ prim, so folgt $p \mid a$ oder $p \mid b$. Im ersten Fall (der zweite ist analog) ergibt sich $a = pc$ und folglich $p = pbc$. Da man in Integritätsringen kürzen darf, folgt $1 = bc$ und $b \in R^*$.
ii) Sei a irreduzibel und $a = p_1 \cdot \ldots \cdot p_r$ Primfaktordarstellung von a. Wegen der Irreduzibilität müssen die p_i bis auf einen Faktor Einheiten sein. Das ist nur im Falle $r = 1$ möglich, da ein Primelement per Definition keine Einheit ist.
iii) Sei $p_1 \cdot \ldots \cdot p_r = q_1 \cdot \ldots \cdot q_s$ mit Primelementen p_i, q_j. Nach Definition eines Primelements folgt, daß p_1 ein q_{j_1} teilt, also $p_1 = q_{j_1} e_1$. Aus der Irreduzibilität von p_1 folgt, daß e_1 Einheit ist. Wir können nun q_{j_1} aus der Gleichung kürzen und mit p_2 analog verfahren. Schrittweise ergibt sich $p_i = q_{j_i} e_i$ mit paarweise verschiedenen j_i und Einheiten e_i. Insbesondere folgt $r \leq s$. Aus Symmetriegründen gilt $r = s$. $\square$

In faktoriellen Ringen lassen sich daher auch größte gemeinsame Teiler bilden. Einen ggT von a und b kann man aus den gemeinsamen Primfaktoren von a und b zusammensetzen. Eine alternative Charakterisierung faktorieller Ringe mittels irreduzibler Elemente liefert die folgende Proposition.

Proposition 5.3. *Ein Integritätsbereich R mit Einselement ist genau dann faktoriell, falls sich jedes Element, das ungleich 0 und keine Einheit ist, in (bis auf Reihenfolge und Einheiten) eindeutiger Weise als Produkt von endlich vielen irreduziblen Elementen darstellen läßt.*

Beweis. Nach Proposition 5.2 gilt diese Zerlegungseigenschaft für faktorielle Ringe. Umgekehrt langt es, aus dieser Eigenschaft zu folgern, daß jedes irreduzible Element p bereits prim ist. Nehmen wir an, daß p das Produkt $c = ab$ teilt, es gilt also $pd = c$ für ein $d \in R$. Indem wir a, b und d in irreduzible Faktoren zerlegen, erhalten wir zwei irreduzible Zerlegungen für c. Nach Voraussetzung sind sie gleich, insbesondere taucht p in der Zerlegung auf, die sich aus $c = ab$ ergibt. Daher ist p bis auf eine Einheit einer der irreduziblen Teiler von a oder von b. p teilt also a oder b und ist damit Primelement. □

Satz 5.4. *Jeder Hauptidealring R ist faktoriell.*

Beweis. Wir zeigen zunächst, daß in einem Hauptidealring jedes irreduzible Element prim ist. Sei p ein irreduzibles Element, daß ab teilt. Wir müssen zeigen, daß p dann entweder a oder b teilt. Dazu bilden wir das von p und a erzeugte Ideal

$$I := \{rp + sa : r, s \in R\}.$$

Nach Voraussetzung ist I ein Hauptideal, es gibt also ein c, so daß

$$I = \{tc : t \in R\}.$$

c teilt p und a, denn $p, a \in I$. Insbesondere gibt es ein $t \in R$ mit $p = tc$. Da p irreduzibel ist, ist entweder t oder c Einheit. Im ersten Fall gilt $c = t^{-1}p$, p teilt also c und folglich a. Im zweiten Fall folgt $1 \in I$, d.h. $1 = rp + sa$ für geeignete $r, s \in R$. In diesem Fall ist p ein Teiler von $b = rpb + sab$.

Sei nun $a \neq 0$. Entweder ist a prim, oder es zerfällt in zwei Faktoren, die beide keine Einheiten sind. Wenn diese Faktoren nicht beide prim sind, können wir sie weiter zerlegen etc. Diese Prozedur bricht nach endlich vielen Schritten mit einer Primfaktorzerlegung von a ab. Andernfalls würde man nämlich eine unendliche Folge $a = a_1, a_2, \ldots$ von Elementen finden, so daß $a_{j+1} \mid a_j$, aber $a_j \nmid a_{j+1}$ für alle j. Dies führt zu einem Widerspruch: Die Hauptideale (a_j) bilden dann eine strikt wachsende Folge. Daher ist auch $I_\infty := \bigcup_j (a_j)$ ein Ideal. In der Tat: Gilt $a, b \in I_\infty$, so gibt es wegen der Monotonie der (a_j) ein i, so daß $a, b \in (a_i)$, und es folgt $a + b \in (a_i) \subset I_\infty$. Die anderen Forderungen an ein Ideal ergeben sich analog. Nach Annahme ist I_∞ ein Hauptideal: $I_\infty = (b)$ für ein $b \in R$. Es folgt $b \in (a_j)$ für ein j und damit $I_\infty \subset (a_j) \subset (a_{j+1}) \subset \ldots \subset I_\infty$. Die Folge der Ideale wäre also doch nicht strikt wachsend. □

Der Beweis dieses Satzes gibt keinen Hinweis, wie man die Primfaktorzerlegung von Elementen eines Hauptidealringes erhält. Wir wollen uns nun der Frage zuwenden, wie man ganze Zahlen in Primfaktoren zerlegt.

5.2 Zerlegung von ganzen Zahlen

Wie wir bereits gesehen haben, gibt effektive Verfahren, die es erlauben, eine Zahl als Primzahl zu erkennen. Die Zerlegung einer Zahl n in ihre Primfaktoren ist dagegen eine sehr viel aufwendigeres Unternehmen.

Indem man der Reihe nach durch die Primzahlen $p = 2, 3, 5, 7, 11, 13, \ldots$ teilt, findet man schnell die kleinen Primteiler von n. Nimmt man einen Computer zur Hilfe, so wird man die Reihe der Primzahlen durch eine leicht zu generierende Testreihe ersetzen, etwa $2, 3, 5, 7, 11, 13, 17, 19, 23, 25, 29, 31, \ldots$, aus der alle Vielfachen von 2 und 3 entfernt sind. Für große Primteiler benötigt diese Methode viel Rechenzeit. Ist n zerlegbar, so gibt es einen Primteiler $p \leq \sqrt{n}$. Man muß sich also auf größenordnungsmäßig $\sqrt{n}$ Divisionen einstellen, bevor man sicher auf den ersten Primteiler stößt.

Es gibt effizientere Methoden zur Zerlegung ganzer Zahlen.

Die ρ-Methode

Die ρ-Methode ist ein Verfahren, daß in einer systematischen Weise nach Teilern einer natürlichen Zahl n sucht. Dazu benötigt man ein Polynom $f(x)$ mit ganzzahligen Koeffizienten, etwa $f(x) = x^2 + 1$, und einen ganzzahligen Startwert a_0. Damit erzeugt man die Zahlen $a_1 = f(a_0), a_2 = f(f(a_0))$,$a_3 = f(f(f(a_0))), \ldots$, allgemein

$$a_i \ := \ f(a_{i-1}),$$

und prüft dann, ob es unter den $a_i - a_j$ mit $j < i$ Zahlen gibt, die einen nicht-trivialen gemeinsamen Teiler mit n besitzen. Man bestimmt also im i-ten Schritt mit dem Euklidischen Algorithmus größte gemeinsame Teiler der Zahlenpaare $(a_i - a_0, n), (a_i - a_1, n), \ldots, (a_i - a_{i-1}, n)$. Findet man einen ggT ungleich $\pm 1, n$, so hat man einen Teiler von n.

Dies ist eine erste Variante eines Verfahrens, das von J.M. Pollard vorgeschlagen wurde. Es erscheint auf den ersten Blick wenig einleuchtend. Warum darf man hoffen, auf diese Weise schneller auf Teiler von n zu stoßen? Hierfür gibt es eine heuristische Begründung. Wir stellen ein stochastisches Modell der ρ-Methode auf. In diesem Modell wird das Polynom $f(x)$ durch eine Zufallsabbildung F von $\mathbb{Z}_n$ nach $\mathbb{Z}_n$ ersetzt. Das bedeutet, daß die Zahlen $F(1), F(2), \ldots, F(n)$ sich modulo n verhalten wie Zahlen, die rein zufällig und unabhängig voneinander aus $\{1, 2, \dots, n\}$ gezogen wurden. Die a_i werden ersetzt durch zufällige Restklassen A_i, die rekursiv durch $A_i = F(A_{i-1})$ und einen Startwert A_0 gegeben sind. – Die Erfahrung hat gezeigt (und Computerexperimente belegen es), daß dieses stochastische Modell eine zutreffende Vorstellung über die Verteilung der Zahlen $f(1), f(2), \dots, f(n)$ modulo n geben, etwa für das Polynom $f(x) = x^2 + 1$.

Sei nun m ein Teiler von n. Aus unserem stochastischen Ansatz gewinnen wir eine Abschätzung für die Schrittzahl, die die Methode braucht, um mit großer Wahrscheinlichkeit m zu entdecken. Aus unserer Annahme folgt, daß $F(1), \dots, F(n)$ auch modulo m rein zufällig gewählte, unabhängige Elemente von $\mathbb{Z}_m$ sind. Der Algorithmus identifiziert nach i Schritten m als Teiler von n, falls $m \mid A_i - A_j$, d.h. falls $A_i \equiv A_j \bmod m$ für ein $j < i$ gilt. Umgekehrt hat das Verfahren m nach i Schritten noch nicht entdeckt, falls $A_0, A_1, \dots, A_i$ alle in verschiedenen Restklassen modulo m liegen. Dieses Ereignis hat nach unserem Ansatz die Wahrscheinlichkeit

$$p_i \;=\; \frac{m-1}{m}\frac{m-2}{m}\cdots\frac{m-i}{m}.$$

Unter Benutzung der Ungleichung $1 - x \le \exp(-x)$ folgt

$$p_i \;\le\; \exp\left(-\frac{1}{m} - \dots - \frac{i}{m}\right),$$

und schließlich wegen $1/m + \dots + i/m = i(i+1)/(2m) > i^2/(2m)$

$$p_i \;\le\; \exp\left(-\frac{i^2}{2m}\right).$$

Anders ausgedrückt: Nach $\sqrt{2\lambda m}$ Schritten hat die ρ-Methode den Teiler m mit einer Wahrscheinlichkeit von höchstens $\exp(-\lambda)$ noch nicht erkannt. Unsere heuristische Überlegung sagt uns also, daß wir mit größenordnungsmäßig $\sqrt{m}$ Schritten rechnen müssen.

Dies ist noch kein Gewinn gegenüber dem schlichten Verfahren, der Reihe nach auszuprobieren, welche Zahlen n teilen. Um den Teiler m zu erkennen, braucht man dann zwar der Größenordnung nach m (statt $\sqrt{m}$) Schritte (zumindest, wenn m eine Primzahl ist), andererseits muß man bei der ρ-Methode im i-ten Schritt i-mal den Euklidischen Algorithmus anwenden. Dies macht den Gewinn wieder zunichte. Nun hat aber Pollard seine Methode effizienter gestaltet. Zunächst ist klar, daß die a_i nur modulo n berechnet zu werden brauchen, dies beschränkt die Größe der zu berechnenden Zahlen. Weiter kann man die Methode ohne wesentlichen Effizienzverlust so umgestalten, daß pro Schritt nur noch ein größter gemeinsamer Teiler berechnet wird.

Die ρ-Methode.

Eingabe: Eine ganze Zahl n.

Ausgabe: Ein Teiler m von n.

Verfahren: Berechne sukzessive aus dem Startwert $a = a_0$ modulo n die Zahlen $a_i = f(a_{i-1}), i = 1, 2, \ldots$. Bestimme $d_i = \mathrm{ggT}(a_i - a_k, n)$, wobei $k = 2^r - 1$ sei für $i = 2^r, 2^r + 1, \ldots, 2^{r+1} - 1$. Setze $m = d_j$ für das kleinste j mit $d_j \neq \pm 1$. □

Dieser Algorithmus bestimmt pro Schritt nur einen ggT. Er kann eine Zahl $a_i - a_j$, die einen gemeinsamen Teiler mit n besitzt, zunächst übergehen. Das Verfahren erkennt diesen Teiler m jedoch rechtzeitig. Es ist nämlich $a_i - a_j$ ein Teiler von $a_i^r - a_j^r$ für $r = 0, 1, 2, \ldots$, dies folgt aus der Formel $a_i^r - a_j^r = (a_i - a_j)(a_i^{r-1} + a_i^{r-2} a_j + \ldots + a_j^{r-1})$. Daher teilt $a_i - a_j$ auch $f(a_i) - f(a_j)$ und folglich $a_{i+l} - a_{j+l}$ für alle $l \geq 1$. Diese Zahlen enthalten also alle den Teiler m, den der Algorithmus zunächst nicht erkannt hat. Gilt nun $2^r \leq i < 2^{r+1}$, so wähle $l = 2^{r+1} - j - 1$, also $l > 0$. Dann gilt $j' = j + l = 2^{r+1} - 1$ und $2^{r+1} \leq i' = i + l < 2 \cdot 2^{r+1}$. Daher testet die ρ-Methode $a_{i'} - a_{j'}$ und entdeckt so den in $a_i - a_j$ verborgenen Teiler. Außerdem gilt $i' < 2^{r+2} \leq 4i$. Die Schrittzahl des Verfahrens vergrößert sich daher auch in ungünstigen Fällen höchstens um den Faktor 4. Dieser Verlust ist geringfügig im Vergleich zu dem Gewinn, daß pro Schritt nur ein ggT zu bestimmen ist.

Beispiel. Sei $n = 1517 = 37 \cdot 41$, $f(x) = x^2 + 1$ und $a_0 = 2$. Dann ist

$$\begin{array}{ll} a_1 = 5, & \mathrm{ggT}(a_1 - a_0, n) = \mathrm{ggT}(3, 1517) = 1, \\ a_2 = 26, & \mathrm{ggT}(a_2 - a_1, n) = \mathrm{ggT}(21, 1517) = 1, \\ a_3 = 677, & \mathrm{ggT}(a_3 - a_1, n) = \mathrm{ggT}(672, 1517) = 1, \\ a_4 \equiv 196 \bmod n, & \mathrm{ggT}(a_4 - a_3, n) = \mathrm{ggT}(-481, 1517) = 37. \end{array}$$

□

Fermats Methode

Die Methode von Fermat führt das Faktorisieren einer Zahl n zurück auf ihre Darstellung als Differenz zweier Quadratzahlen. Sei $n = a \cdot b$ eine ungerade Zahl. Dann sind auch a und b ungerade, und wir können die Zahlen

$$s = \frac{a-b}{2}, \quad t = \frac{a+b}{2}$$

bilden. Es folgt

$$n = t^2 - s^2.$$

Umgekehrt erhalten wir aus dieser Darstellung die Zerlegung

$$n = (t+s)(t-s).$$

Dies führt zu einem ersten Algorithmus. Man bilde die Zahlen $t_0 = [\sqrt{n}] + 1, t_1 = [\sqrt{n}] + 2, \ldots$ und überprüfe der Reihe nach, ob $t_0^2 - n, t_1^2 - n, \ldots$ Quadratzahlen sind.

Beispiel. Für $n = 1767$ ist $t_0^2 - n = 43^2 - 1767 = 82$ keine Quadratzahl und $t_1^2 - n = 44^2 - 1767 = 169 = 13^2$. Wir erhalten die Zerlegung $1767 = (44+13)\cdot(44-13) = 57 \cdot 31$. □

Das Verfahren findet zuerst die großen Teiler von n und führt schnell zum Ziel, wenn sich n in zwei ähnlich große Faktoren zerlegen läßt. Manchmal läßt sich dies dadurch erreichen, daß man nicht n sondern rn zerlegt, etwa mit $r = 3$.

Schon Fermat gestaltete seine Methode effizienter, indem er die Überprüfung auf Quadratzahl modular durchführte. Man wählt dazu teilerfremde Moduln $m_1, \ldots, m_k$ und betrachtet zunächst $t_i^2 - n$ als Rest modulo m_j. Liegt kein quadratischer Rest vor, so kann $t_i^2 - n$ keine Quadratzahl sein. Nur in dem Fall, daß es sich für jeden Modul $m_1, \ldots, m_k$ um einen quadratischen Rest handelt, muß man nachprüfen, ob man auf eine Quadratzahl gestoßen ist.

In der praktischen Durchführung stellt man eine Liste (ein ‚Sieb') der quadratischen Reste auf. Knuth faktorisiert in seiner Monographie auf diese Weise die Zahl $n = 8.616.460.799$ (von der der englische Ökonom und Logiker W.S. Jevons 1874 unvorsichtigerweise behauptet hat, daß man sie niemals wird zerlegen können, vgl. Knuth Vol. II, S. 388). Er stellt dazu die folgende Tabelle auf:

m_i	t mod m_i	t^2 mod m_i	$(t^2 - n)$ mod m_i
3	$0, 1, 2$	$0, 1, 1$	$1, 2, 2$
5	$0, 1, 2, 3, 4$	$0, 1, 4, 4, 1$	$1, 2, 0, 0, 2$
7	$0, 1, 2, 3, 4, 5, 6$	$0, 1, 4, 2, 2, 4, 1$	$5, 6, 2, 0, 0, 2, 6$
8	$0, 1, 2, 3, 4, 5, 6, 7$	$0, 1, 4, 1, 0, 1, 4, 1$	$1, 2, 5, 2, 1, 2, 5, 2$
11	$0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$	$0, 1, 4, 9, 5, 3, 3, 5, 9, 4, 1$	$10, 0, 3, 8, 4, 2, 2, 4, 8, 3, 0$

Ein Vergleich der beiden letzten Spalten zeigt, daß nur dann $t^2 - n$ eine

Quadratzahl sein kann, falls gilt

$$\begin{array}{lcll} t & \equiv & 0 & \bmod 3 \\ t & \equiv & 0,2 \text{ oder } 3 & \bmod 5 \\ t & \equiv & 2,3,4 \text{ oder } 5 & \bmod 7 \\ t & \equiv & 0 \text{ oder } 4 & \bmod 8 \\ t & \equiv & 1,2,4,7,9 \text{ oder } 10 & \bmod 11. \end{array}$$

Die Suche nach einem passenden t wird so erheblich eingeschränkt. Nach den Bedingungen modulo 3 und modulo 8 muß t ein Vielfaches von $3 \cdot 4 = 12$ sein. Das kleinste t größer als $[\sqrt{n}] + 1 = 92825$ mit dieser Eigenschaft ist 92832. Es hat modulo 5 den Rest 2, modulo 7 den Rest 5 und modulo 11 den Rest 3. Die Liste zeigt, daß dieses t kein Kandidat zum Faktorisieren ist. Erhöhen wir t um den Betrag 12, so ändert sich sein Residuum jeweils um 2 modulo 5, um 5 modulo 7 und um 1 modulo 11. Wie eine kurze Rechnung zeigt, ist das erste t, daß alle aufgestellten Bedingungen erfüllt, gleich 92880. Es folgt $92880^2 - n = 10233601 = 3199^2$. Wir haben also eine Lösung $s = 3199, t = 92880$ gefunden und können n zerlegen:

$$8.616.460.799 \;=\; (t-s)(t+s) \;=\; 89681 \cdot 96079.$$

Allgemein gesprochen wird zum Faktorisieren von n mit den Moduln $m_1, \ldots, m_k$ eine *Siebtabelle* $s(i,t), 1 \leq i \leq k, 0 \leq t < m_i$ aufgestellt. Es ist $s(i,t) = 1$ oder 0, je nachdem ob $q = t^2 - n$ quadratischer Rest modulo m_i ist oder nicht.

Faktorisieren mit Sieben.

Eingabe: n und Moduln $m_1, \ldots, m_k$.

Ausgabe: Ein Teiler m von n.

Verfahren: Stelle die Siebtafel auf. Berechne sukzessive modulo $m_1, \ldots, m_k$ die Zahlen $q_t = t^2 - n, t > [\sqrt{n}]$ (also $q_{t+1} = q_t + 2t + 1$). Bestimme mit der Siebtafel das kleinste t, so das q_t für alle j quadratischer Rest modulo m_j ist. Teste, ob q_t eine Quadratzahl s^2 ist. In diesem Fall ist $m = t - s$ Teiler von n. Andernfalls setze die Suche nach einem geeigneten t fort. □

Faktorisieren mit Basen

Die Idee, auf der das Verfahren von Fermat beruht, läßt sich weiter ausbauen. Zum Faktorisieren von n langt es, ganze Zahlen s, t zu finden, so daß

$$s^2 \;\equiv\; t^2 \bmod n, \quad s \;\not\equiv\; \pm t \bmod n.$$

Dann ist $(t+s)(t-s) = t^2 - s^2$ Vielfaches von n, d.h. es gibt eine Zerlegung $n = n_1 n_2$ mit $n_1 \mid t+s, n_2 \mid t-s$. Andererseits sind $t+s$ und $t-s$ keine Vielfachen von n, d.h. $n_1, n_2 \neq n$. Wenn wir also $\mathrm{ggT}(t+s, n)$ berechnen, erhalten wir einen echten Teiler von n.

Es gibt eine allgemeine Strategie, um sich Lösungen s, t der Kongruenz zu verschaffen. Wir vereinbaren dazu die folgenden Sprechweisen. Eine Menge $B = \{p_1 = -1, p_2, \ldots, p_h\}$ heißt **Faktorbasis**, falls $p_2, \ldots, p_h$ Primzahlen sind. Eine ganze Zahl a heißt ***B*-Zahl**, falls der kleinste absolute Rest modulo n von a^2 sich als Produkt von Zahlen aus B schreiben läßt. Mit dem kleinsten absoluten Rest ist die Zahl r zwischen $-n/2$ und $n/2$ gemeint, die kongruent a^2 modulo n ist.

Nehmen wir nun an, wir verfügen über verschiedene B-Zahlen a_i und den zugehörigen kleinsten absoluten Resten r_i ihrer Quadrate modulo n. Dann gilt

$$r_i = \prod_{j=1}^{h} p_j^{\alpha_{ij}}$$

mit nicht-negativen ganzen Zahlen α_{ij}. Nehmen wir weiter an, daß wir aus diesen B-Zahlen Zahlen $a_1, \ldots, a_k$ aussondern können, so daß $\sum_{i=1}^{k} \alpha_{ij}$ für alle $j = 1, \ldots, h$ geradzahlig ist. Dann können wir

$$s = \prod_{i=1}^{k} a_i, \quad t = \prod_{j=1}^{h} p_j^{\frac{1}{2}\sum_i \alpha_{ij}}.$$

bilden. Nach Konstruktion sind s^2 und t^2 kongruent modulo n. Da s und t auf ganz verschiedene Weise konstruiert sind, besteht eine gute Chance , daß $s \not\equiv \pm t \bmod n$ gilt. Allerdings ist der gegenteilige Fall nicht ausgeschlossen, dann sind s und t zur Zerlegung von n ungeeignet.

Es bleibt die Frage, wie man zu einer vorgegebenen Zahl n eine Faktorbasis B und ausreichend viele B-Zahlen a_i findet. Dazu gibt es verschiedene Ansätze. Sehr erfolgreich ist das *quadratische Sieb* von C. Pomerance, das z.B. in der 3. Auflage des Buches von Koblitz behandelt wird. Wir wollen eine andere, auf dem Kettenbruchalgorithmus basierende Methode besprechen. Die Idee ist, zunächst Zahlen zu konstruieren, deren Quadrate modulo n einen besonders kleinen absoluten Rest besitzen. Dazu benutzt man die Näherungsbrüche von $\sqrt{n}$. Anschließend versucht man, die Faktorbasis B an diese Zahlen anzupassen.

Proposition 5.5. *Sei n eine natürliche Zahl, die keine Quadratzahl ist, und seien a_i/b_i die Näherungsbrüche aus der Kettenbruchentwicklung von $\sqrt{n}$. Dann gilt für den kleinste absolute Rest r_i von a_i^2 modulo n*

$$|r_i| < 2\sqrt{n}.$$

Beweis. Nach dem Satz von Lagrange gilt $|a_i/b_i - \sqrt{n}| < 1/b_ib_{i+1}$ (vgl. den Beweis von Satz 2.2). Es folgt

$$\begin{aligned}
|a_i^2 - nb_i^2| &= b_i^2\,|a_i/b_i - \sqrt{n}| \cdot |a_i/b_i + \sqrt{n}| \\
&< \frac{b_i}{b_{i+1}}\left(2\sqrt{n} + \frac{1}{b_ib_{i+1}}\right) \\
&= 2\sqrt{n}\left(\frac{b_i}{b_{i+1}} + \frac{1}{2\sqrt{n}b_{i+1}^2}\right) \\
&< 2\sqrt{n}\left(\frac{b_i}{b_{i+1}} + \frac{1}{b_{i+1}}\right) \\
&< 2\sqrt{n}.
\end{aligned}$$

Wegen $|r_i| \le |a_i^2 - nb_i^2|$ folgt die Behauptung. □

Es besteht also Grund zu der Erwartung, daß man für die Primfaktorzerlegung der absolut kleinsten Reste von a_i^2 modulo n vergleichweise kleine und wenige Primfaktoren benötigt.

Faktorisieren mit Kettenbrüchen.

Eingabe: Eine ganze Zahl n

Ausgabe: Ein Teiler m von n

Verfahren: Setze $a_0 = 1$, $a_1 = [\sqrt{n}]$ und $\alpha_1 = \sqrt{n}$. Wähle eine natürliche Zahl k und bestimme die Zahlen $m_1, m_2, \ldots, m_k, a_1, a_2, \ldots, a_k, \alpha_2, \alpha_3, \ldots, \alpha_k$ aus den Gleichungen $m_i = [\alpha_i], a_i = m_ia_{i-1} + a_{i-2}$ und $\alpha_i = m_i + 1/\alpha_{i+1}$. Die a_i brauchen nur modulo n berechnet zu werden. Bilde den kleinsten absoluten Rest r_i von a_i^2 modulo n. Wähle die r_i aus, die in ein Produkt kleiner Primzahlen zerfallen und fasse diese Primzahlen zusammen mit -1 zu einer Faktorbasis B zusammen. Bilde eine Liste der B-Zahlen unter den betrachteten a_i und versuche, mit ihnen n zu faktorisieren. Gelingt dies nicht, so vergrößere k. □

Beispiel. Es sei $n = 9073$. Wir erhalten die Liste

i	1	2	3	4	5
m_i	95	3	1	26	2
$a_i \bmod n$	95	286	381	1119	2619
$a_i^2 \bmod n$	-48	139	-7	87	-27

Es liegt nahe, $B = \{-1, 2, 3, 7\}$ zu wählen, die zugehörigen B-Zahlen sind dann $-48 = -2^4 \cdot 3$, -7 und $-27 = -3^3$. Die erste und dritte Zahl sind für das Faktorisieren von n geeignet. Wir wählen also $s = 95 \cdot 2619 \equiv 3834 \bmod 9073$ und $t = 2^2 \cdot 3^2 = 36$. Da $3834 \not\equiv \pm 36 \bmod 9073$, erhalten wir den nichttrivialen Faktor $\mathrm{ggT}(3834 + 36, 9073) = 43$. Es folgt $9073 = 43 \cdot 211$. $\square$

Kapitel 6

Faktorzerlegung in Polynomringen

6.1 Faktorielle Polynomringe

Sei R ein faktorieller Ring und $R[x]$ der Ring der Polynome in der Unbestimmten x mit Koeffizienten in R. Wir wollen in diesem Abschnitt zeigen, daß dann auch $R[x]$ ein faktorieller Ring ist.

Dies bedarf keiner zusätzlichen Begründung, falls R ein Körper ist, dann ist $R[x]$ ein Euklidischer Ring und damit Hauptidealring. Alle Elemente $a \in R-\{0\}$ sind Einheiten, in R wie in $R[x]$. Es folgt, daß die Polynome vom Grad 1 sämtlich prim sind. In einigen Fällen sind dies bereits alle Primpolynome. Der folgende Satz beschreibt, wann ein Polynom einen Teiler vom Grad 1 besitzt. Wir benötigen dazu den Begriff der Nullstelle. $s \in R$ heißt **Nullstelle** des Polynoms $f(x)$, falls sich durch Einsetzen von s in die Unbestimmte x der Wert Null ergibt, also $f(s) = a_n s^n + \ldots + a_1 s + a_0 = 0$ gilt.

Satz 6.1. *Sei $f(x)$ ein Polynom mit Koeffizienten aus einem Ring mit Einselement. Dann teilt das Polynom $x - s$ genau dann $f(x)$, wenn s Nullstelle von $f(x)$ ist.*

Beweis. Wir teilen $f(x)$ durch $x - s$ mit Rest,

$$f(x) \;=\; m(x)(x-s) + r$$

mit $r \in R$. Durch Einsetzen von s folgt $f(s) = r$. Es teilt also $x - s$ genau dann $f(x)$, wenn $r = f(s) = 0$ ist. Man bemerke: Die Division von $f(x)$ durch $x - s$ ist in jedem Ring mit Einselement möglich, denn $x - s$ hat 1 als Anfangskoeffizienten. □

Offensichtlich kann ein Polynom vom Grade n als Teiler höchstens n verschiedene Polynome vom Grad 1 besitzen. Dies begrenzt die Anzahl von Nullstellen eines Polynoms.

Korollar 6.2. *Ein Polynom ungleich dem Nullpolynom vom Grade n mit Koeffizienten aus einem Ring mit Einselement hat höchstens n verschiedene Nullstellen.*

Beispiele.

1) Bekanntlich hat nach dem ‚Hauptsatz der Algebra' jedes Polynom mit komplexen Koeffizienten eine Nullstelle. Daher zerfällt jedes komplexe Polynom in lineare Polynome. Die einzigen irreduziblen Elemente von $\mathbb{C}[x]$ sind die Polynome vom Grad 1.

2) Es hat also auch jedes Polynom $f(x) = a_n x^n + \ldots + a_0$ mit reellen Koeffizienten eine komplexe Nullstelle $s = s_1 + is_2$, $s_1, s_2 \in \mathbb{R}$. Wir unterscheiden zwei Fälle. Entweder s ist eine reelle Zahl. Dann hat $f(x)$ das reelle Polynom $x - s$ als Teiler. Oder aber $s_2 \neq 0$. Dann gilt $f(\overline{s}) = \overline{f(s)} = \overline{0} = 0$, mit $\overline{s} = s_1 - is_2$. In diesem Fall wird $f(x)$ nach Satz 6.1 von dem reellen Polynom
$$g(x) \ = \ (x-s)(x-\overline{s}) = (x-s_1)^2 + s_2^2$$
geteilt. $g(x)$ ist ein Polynom zweiten Grades ohne Nullstellen, daher ein irreduzibles Polynom in $\mathbb{R}[x]$. Insgesamt können wir feststellen: Jedes Polynom in $\mathbb{R}[x]$ zerfällt in Polynome des Grades 1 oder 2. Die irreduziblen Polynome in $\mathbb{R}$ sind die linearen Polynome $ax + b$ und die quadratischen Polynome $ax^2 + bx + c$ ohne reellen Nullstellen .

3) Sei p eine Primzahl. Wir betrachten das Polynom
$$f(x) \ = \ x^p - x$$
als Polynom in $\mathbb{Z}_p[x]$. Nach dem Satz von Fermat hat $f(x)$ die Nullstellen $0, 1, \ldots, p-1$, daher gilt in $\mathbb{Z}_p[x]$ die Identität
$$x^p - x \ = \ x(x-1)\ldots(x-p+1).$$
Dies ist eine wichtige Polynomidentität, die uns später gute Dienste leisten wird. Vergleich des Koeffizienten von x führt zu dem *Satz von Wilson*:
$$1 \cdot 2 \cdot \ldots (p-1) \ \equiv \ -1 \bmod p.$$
$f(x)$ ist ein Beispiel für ein Polynom, das nicht das Nullpolynom ist, bei dem sich beim Ersetzen der Unbestimmten durch ein Element des Koeffizientenbereichs aber immer den Wert 0 ergibt. □

Wir wollen nun den Ring $\mathbb{Z}[x]$ der ganzzahligen Polynome betrachten. $\mathbb{Z}[x]$ ist kein Hauptidealring; zum Beispiel ist das von 2 und x erzeugte Ideal

$$I \;:=\; \{2a_0 + a_1x + \ldots + a_nx^n \;:\; n \geq 0,\; a_i \in \mathbb{Z}\}$$

kein Hauptideal. Es bedarf also eines gesonderten Beweises, daß die ganzzahligen Polynome einen faktoriellen Ring bilden.

Satz 6.3. *$\mathbb{Z}[x]$ ist ein faktorieller Ring.*

Beweis. Zum Beweis benötigen wir den Begriff eines primitiven Polynoms. Ein Polynom heißt *primitiv*, falls seine von Null verschiedenen Koeffizienten teilerfremd sind. Wir können jedes ganzzahlige Polynom $f(x)$ darstellen als $f(x) = d \cdot p(x)$ mit einer ganzen Zahl d und einem primitiven Polynom $p(x)$. Diese Darstellung ist eindeutig bis auf Einheiten in $\mathbb{Z}$.

Wir zeigen nun, daß ein Produkt $h(x) = f(x)g(x) = \sum c_kx^k$ genau dann primitiv ist, wenn $f(x) = \sum a_ix^i$ und $g(x) = \sum b_jx^j$ primitive Polynome sind (diese Aussage wird ‚Gaußsches Lemma' genannt). Ist nämlich d etwa ein Teiler aller a_i, so ist d auch Teiler aller $c_k = \sum_{i+j=k} a_ib_j$. Sind umgekehrt $f(x)$ und $g(x)$ primitiv, so gibt es zu jeder Primzahl p Koeffizienten a_i und b_j, die p nicht als Primteiler enthalten. Wählen wir i und j so klein wie möglich, so teilt p auch

$$c_{i+j} \;=\; a_0b_{i+j} + \ldots + a_ib_j + \ldots + a_{i+j}b_0$$

nicht (denn p teilt alle Summanden bis auf a_ib_j). Also ist $h(x)$ primitiv.

Es ist nun leicht, ein beliebiges ganzzahliges Polynom $f(x)$ in irreduzible Polynome aufzuspalten. In einem ersten Schritt zerlegen wir $f(x)$ in eine ganze Zahl d und ein primitives Polynom $p(x)$. d läßt sich in $\mathbb{Z}$ in Primfaktoren zerlegen. Zerfällt $p(x)$ in Faktoren, so sind sie wegen der Primitivität von $p(x)$ ebenfalls primitiv und von einem Grad größer als 0. Die Anzahl der Faktoren, in die $p(x)$ zerfallen kann, ist daher durch den Grad von $p(x)$ beschränkt, so daß $p(x)$ schließlich in irreduzible Faktoren zerlegt ist.

Nach Proposition 5.3 bleibt zu zeigen, daß diese Zerlegung eindeutig ist. Dazu machen wir von der Eindeutigkeit der Primfaktorzerlegung im Euklidischen Ring der rationalen Polynome Gebrauch. Zunächst beweisen wir, daß ein irreduzibles ganzzahliges Polynom $f(x)$ auch in dem Ring $\mathbb{Q}[x]$ irreduzibel ist. Wir fassen also $f(x)$ als Polynom mit rationalen Koeffizienten auf und nehmen an, daß es in $\mathbb{Q}[x]$ in Faktoren zerfällt: $f(x) = r(x)s(x)$. Durch Multiplikation mit einer passenden natürlichen Zahl m können wir die rationalen Polynome $r(x), s(x)$ in ganzzahlige Polynome verwandeln und diese dann durch primitive Polynome $p(x), q(x)$ darstellen. Insgesamt erhalten wir die

Gleichung $mf(x) = dp(x)q(x)$ mit $d \in \mathbb{Z}$. Als irreduzibles Polynom ist $f(x)$ primitiv, außerdem ist $p(x)q(x)$ primitives Polynom. Daher stimmen m und d bis auf Einheiten überein, und wir erhalten die Gleichung $f(x) = \pm p(x)q(x)$. Da nun aber $f(x)$ als irreduzibel angenommen wurde, muß entweder $p(x)$ oder $q(x)$ ein Einheit in $\mathbb{Z}[x]$ sein, und damit ein Polynom vom Grad 0. Es ist folglich auch $r(x)$ oder $s(x)$ Polynom vom Grade 0 und damit Einheit in $\mathbb{Q}[x]$, und $f(x)$ ist irreduzibel auch im Ring der rationalen Polynome.

Seien nun $f(x) = d_1 f_1(x) \cdot \ldots \cdot f_r(x) = d_2 g_1(x) \cdot \ldots \cdot g_s(x)$ zwei Zerlegungen von $f(x)$ in $\mathbb{Z}[x]$ mit ganzen Zahlen d_1, d_2 und irreduziblen Polynomen $f_i(x), g_j(x)$ vom Grade grösser als Null. Es folgt $d_1 = d_2$, denn die Polynome f_i, g_j sind primitiv. Wegen der Eindeutigkeit der Primfaktorzerlegung in $\mathbb{Q}[x]$ gibt es weiter zu jedem $f_i(x)$ ein zugehöriges $g_j(x)$ und eine rationale Zahl $r = a/b$, so daß $f_i(x) = rg_j(x)$ bzw. $af_i(x) = bg_j(x)$. Da die Polynome primitiv sind, folgt $f_i(x) = \pm g_j(x)$. Die Zerlegung von $f(x)$ ist also auch im Ring der ganzzahligen Polynome eindeutig. $\square$

Dieser Beweis ist für beliebige faktorielle Ringe R gültig, denn man kann R wie die ganzen Zahlen in einen Körper einbetten. Wir rekapitulieren kurz die Konstruktion dieses **Quotientenkörpers** Q_R von R, die sich für einen beliebigen Integritätsbereich mit Einselement durchführen läßt. Um Quotienten von Ringelementen zu erhalten, definieren wir eine Äquivalenzrelation

$$(a, b) \sim (c, d) \;:\Leftrightarrow\; ad = bc$$

auf der Menge $R \times (R - \{0\})$. Die Transitivität ergibt sich so: $ad = bc, cf = de$ impliziert $adf = bcf = bde$ und damit $af = eb$, denn wir dürfen in R kürzen. Den Bruch a/b definieren wir nun als die Äquivalenzklasse, in der (a, b) liegt. Brüche werden gemäß der Vereinbarung

$$\frac{a}{b} + \frac{c}{d} \;:=\; \frac{ad + bc}{bd}, \quad \frac{a}{b} \cdot \frac{c}{d} \;:=\; \frac{ac}{bd}$$

addiert und multipliziert. Man zeigt leicht, daß diese Definition unabhängig von der Wahl der Repräsentanten $(a, b), (c, d)$ ist. Wir definieren $0 := 0/1$, $1 := 1/1$, $-(a/b) := (-a)/b$ und, falls $a, b \neq 0$, $(a/b)^{-1} := (b/a)$. Es ist nicht schwer nachzurechnen, daß die Menge Q_R aller Brüche mit diesen Vereinbarungen ein Körper wird. Die Abbildung $a \mapsto a/1$ bettet R homomorph in Q_R ein, wir dürfen also Q_R als Erweiterungskörper von R auffassen. Algebraisch läßt sich der Quotientenkörper (bis auf Isomorphie) charakterisieren als kleinster Körper, der in allen R umfassenden Körpern enthalten ist.

Beispiel. Sei R ein Integritätsbereich mit Einselement. Dann gilt dies auch für den Polynomring $R[x]$. Der zugehörige Quotientenkörper wird mit $R(x)$ bezeichnet. Seine Elemente heißen **rationale Funktionen** und lassen sich als Quotienten von Polynomen schreiben. □

Offenbar können wir nun den Beweis des letzten Satzes auf faktorielle Ringe übertragen.

Satz 6.4. *Sei R ein faktorieller Ring. Dann ist auch $R[x]$ faktoriell.*

Beispiel. Der Ring $\mathbb{Z}[x, y]$ der ganzen Polynome $\sum_{i,j} a_{ij}x^i y^j$ in zwei Unbestimmten x, y ist faktoriell, denn

$$\mathbb{Z}[x, y] \;\cong\; (\mathbb{Z}[x])[y].$$

Allgemeiner ist der Ring $R[x_1, \ldots, x_n]$ aller Polynome in mehreren Unbestimmten $x_1, \ldots, x_n$ mit Koeffizienten in einem faktoriellen Ring R wieder faktoriell. □

Wir wenden uns nun dem Faktorisieren von Polynomen zu. Um für ein Polynom $f(x) = \sum_{i=0}^{n} a_i x^i$ Faktoren $g(x) = \sum_{j=0}^{l} b_j x^j$ und $h(x) = \sum_{k=0}^{m} c_k x^k$ zu finden, wobei ohne Einschränkung $l \leq m$ sei, muß man die Gleichungen

$$\begin{aligned}
a_0 &= b_0c_0 \\
a_1 &= b_1c_0 + b_0c_1 \\
&\;\;\vdots \\
a_l &= b_lc_0 + b_{l-1}c_1 + \ldots + b_0c_l \\
&\;\;\ddots \\
a_m &= b_lc_{m-l} + \ldots + b_1c_{m-1} + b_0c_m \\
&\;\;\ddots \\
a_{n-1} &= b_lc_{m-1} + b_{l-1}c_m \\
a_n &= b_lc_m
\end{aligned}$$

erfüllen. Diese Aufgabe ist in vielen Fällen gar nicht so aufwendig, wie sie auf den ersten Blick erscheinet. Manchmal ist es leicht zu sehen, daß sich $f(x)$ nicht faktorisieren läßt. Bekannt ist das **Irreduzibilitätskriterium von Eisenstein**.

Proposition 6.5. *Sei R faktorieller Ring und $f(x) = a_nx^n + \ldots + a_0$ Polynom in $R[x]$. Dann ist $f(x)$ irreduzibel, falls es ein Primelement $p \in R$ gibt, so daß*

$$p \mid a_0,\; p \mid a_1, \ldots, p \mid a_{n-1},\; p \nmid a_n \;\textit{und}\; p^2 \nmid a_0.$$

Beweis. Nehmen wir an, $f(x)$ läßt sich zerlegen, so daß die angegebenen Gleichungen gelten. Aus $p \mid a_0, p^2 \nmid a_0$ folgt $p \mid b_0, p \nmid c_0$ oder $p \mid c_0, p \nmid b_0$. Im ersten Fall folgt aus den Gleichungen (von oben nach unten gelesen) $p \mid b_1, \ldots, p \mid b_l$, im zweiten Fall $p \mid c_1, \ldots, p \mid c_m$. In jedem Fall ergibt die letzte Gleichung $p \mid a_n$, im Widerspruch zur Voraussetzung. Also ist $f(x)$ irreduzibel. □

Beispiele.

1) Für eine Primzahl p ist $x^n - p$ ein irreduzibles Polynom in $\mathbb{Z}[x]$ und damit auch in $\mathbb{Q}[x]$ (vgl. den Beweis von Satz 6.3). Folglich besitzt das Polynom keine rationale Nullstelle, und $\sqrt[n]{p}$ ist eine irrationale Zahl.

2) Für eine Primzahl p ist

$$f(x) \;=\; x^{p-1} + x^{p-2} + \ldots + x + 1$$

in $\mathbb{Z}[x]$ irreduzibel. Zum Beweis ersetzen wir in der Identität $(x-1)f(x) = x^p - 1$ die Unbestimmte x durch $x+1$ und erhalten

$$xf(x+1) \;=\; (x+1)^p - 1 \;=\; x\Big(x^{p-1} + \binom{p}{p-1}x^{p-2} + \ldots + p\Big).$$

Nach dem Eisensteinschen Kriterium ist $f(x+1)$ irreduzibel, denn $\binom{p}{k}$ ist für primes p und $0 < k < p$ durch p teilbar. Daher ist auch $f(x)$ ein irreduzibles Polynom. □

Wie steht es nun mit dem Zerlegen eines ganzzahligen Polynoms $f(x)$ in ganzzahlige Primpolynome? Wie schon Kronecker bemerkt hat, läßt sich diese Aufgabe in endlich vielen Schritten durchführen. Ausgangspunkt ist die Beobachtung, daß es zu ganzen Zahlen $s_0 < s_1 < \ldots < s_d, t_0, t_1 \ldots, t_d$ genau ein Polynom $g(x) \in \mathbb{Q}[x]$ mit Grad $\deg(g) \le d$ gibt, so daß sich bei Einsetzen von s_i in die Unbestimmte x der Wert t_i ergibt: $g(s_i) = t_i$. Das Polynom

$$g(x) \;:=\; \sum_{i=0}^{d} t_i \prod_{j \ne i} \frac{x - s_j}{s_i - s_j}$$

leistet dies (Lagrange-Interpolation). Erfüllt auch das Polynom $h(x)$ vom Grad $\deg(h) \le d$ die Gleichungen $h(s_i) = t_i$, so hat $g(x) - h(x)$ mindestens $d+1$ Nullstellen, nämlich $s_0, \ldots, s_d$. Nach Korollar 6.2 ist $g(x) - h(x)$ das Nullpolynom, $g(x)$ ist also eindeutig bestimmt.

Ist nun $g(x)$ in $\mathbb{Z}[x]$ ein Teiler von $f(x)$, so gilt $g(s) \mid f(s)$ für jede ganze Zahl s. Man kann diese Beobachtung zum Ausgangspunkt machen, um einen

Teiler $g(x)$ von $f(x)$ zu finden, indem man ihn unter denjenigen Polynomen sucht, welche an der Stelle s einen Wert annehmen, der $f(s)$ teilt. Macht man dies für ausreichend viele s, so läßt sich $g(x)$ mittels Lagrange-Interpolation konstruieren. Ohne Einschränkung können wir voraussetzen, daß $g(x)$ einen Grad $\deg(g) \leq \deg(f)/2$ hat.

Verfahren von Kronecker.

1) Wähle ganze Zahlen $s_0, \ldots, s_d$ mit $d = [\deg(f)/2]$.

2) Für jedes Zahlentupel $(t_0, \ldots, t_d)$ mit $t_i \mid f(s_i)$ bestimme das eindeutige $g(x) \in \mathbb{Q}[x]$ mit $g(s_i) = t_i$, $\deg(g) \leq d$.

3) Prüfe für dieses $g(x)$, ob es ganzzahlig ist und ob es $f(x)$ teilt, etwa durch Division mit Rest.

4) Mit Teilern von $f(x)$ verfahre analog. $\square$

Dieses Verfahren führt zum Ziel, es ist aber immens rechenaufwendig und schon ab dem Grad $n = 5$ kaum noch durchführbar.

Im allgemeinen ist es sehr viel effizienter, ganzzahlige Polynome zunächst modulo m, d.h. in $\mathbb{Z}_m[x]$ zu zerlegen. Häufig gelingt es, dann auf die Zerlegung in $\mathbb{Z}[x]$ zurückzuschließen. Der Einfachheit halber betrachten wir **normierte Polynome**, also Polynome $f(x) \in \mathbb{Z}[x]$, deren Anfangskoeffizient 1 ist,

$$f(x) \;=\; x^n + a_{n-1}x^{n-1} + \ldots + a_0.$$

Wir fassen nun $f(x)$ auch als Polynom in $\mathbb{Z}_m[x]$ auf (a_i wird zur Restklasse $a_i \bmod m$). Da $f(x)$ als normiert angenommen ist, bleibt der Grad von $f(x)$ unverändert. Gibt es eine nicht-triviale Zerlegung $f(x) = g(x)h(x)$, so können wir auch $g(x)$ und $h(x)$ als normiert annehmen, und die Zerlegung bleibt in $\mathbb{Z}_m[x]$ gültig. Anders ausgedrückt: Ist das normierte Polynom $f(x)$ in $\mathbb{Z}_m[x]$ irreduzibel, so ist $f(x)$ auch in $\mathbb{Z}[x]$ irreduzibel.

Beispiel. Das Polynom

$$f(x) \;=\; x^8 + x^6 - 3x^4 - 3x^3 + 8x^2 + 2x - 5$$

zerfällt in $\mathbb{Z}_2[x]$ und $\mathbb{Z}_{13}[x]$ in die irreduziblen Faktoren

$$\begin{aligned} f(x) &\equiv (x^6 + x^5 + x^4 + x + 1)(x^2 + x + 1) \bmod 2 \\ f(x) &\equiv (x^4 + 2x^3 + 3x^2 + 4x + 6)(x^3 + 8x^2 + 4x + 12)(x + 3) \bmod 13, \end{aligned}$$

wie man mit der im folgenden Abschnitt beschriebenen Berlekamp-Methode findet. Eine Faktorisierung von $f(x)$ in $\mathbb{Z}[x]$ müßte mit diesen beiden Zerlegungen vereinbar sein, was offenbar nicht möglich ist. Zerfiele $f(x)$ etwa in irreduzible Faktoren vom Grad 6 und 2, so müßten diese Faktoren modulo 13 weiter zerfallen. Ein irreduzibler Teiler vom Grad 4 modulo 13 kann dabei nicht entstehen. Folglich ist $f(x)$ in $\mathbb{Z}[x]$ irreduzibel. $\square$

Die Erfahrung zeigt, daß sich auf diese Weise Irreduzibilität häufig feststellen läßt.

6.2 Faktorisieren in $\mathbb{Z}_p[x]$

Wir behandeln nun die Faktorisierung von Polynomen in $\mathbb{Z}_p[x]$ für eine Primzahl p. Die Situation unterscheidet sich deutlich vom Zerlegen ganzer Zahlen. Das Faktorisieren ganzer Zahlen haben wir im wesentlichen als mehr oder weniger geschicktes Suchen nach Teilern kennengelernt, am Anfang ist schwer abzusehen, wie schnell man fündig wird. Polynome in $\mathbb{Z}_p[x]$ lassen sich dagegen mit der **Methode von Berlekamp** (1967) in ganz systematischer Weise zerlegen. Wir werden dazu die folgenden Polynom-Identitäten benötigen.

Proposition 6.6. *Sei p eine Primzahl und seien $h(x)$ und $k(x)$ Polynome in $\mathbb{Z}_p[x]$. Dann gilt*

$$\begin{aligned}
(h(x)+k(x))^p &= h(x)^p + k(x)^p,\\
h(x^p) &= h(x)^p,\\
h(x)^p - h(x) &= h(x)(h(x)-1)(h(x)-2)\dots(h(x)-p+1).
\end{aligned}$$

Beweis. Der Binomialkoeffizient $\binom{p}{k} = \frac{p(p-1)\cdots(p-k+1)}{1\cdot 2\cdots k}$ ist für primes p und $0<k<p$ ein Vielfaches von p, daher gilt

$$\begin{aligned}
(h(x)+k(x))^p &= h(x)^p + \binom{p}{1} h(x)^{p-1}k(x) + \ldots + k(x)^p\\
&\equiv h(x)^p + k(x)^p \bmod p.
\end{aligned}$$

Dies ist die erste Behauptung. Zusammen mit dem Fermatschen Theorem impliziert sie die zweite Behauptung,

$$h(x)^p = (\sum a_i x^i)^p = \sum a_i^p x^{ip} = \sum a_i (x^p)^i = h(x^p).$$

Die letzte Aussage ergibt sich schließlich, indem wir in der Identität

$$x^p - x \equiv x(x-1)\cdots(x-p+1) \bmod p$$

x durch $h(x)$ ersetzen. $\square$

Quadratische Anteile eines Polynoms lassen sich mithilfe seiner (formalen) Ableitung identifizieren. Die **Ableitung** eines Polynoms $f(x) = a_n x^n + \ldots + a_1 x + a_0$ ist definiert als

$$f'(x) \;:=\; n \cdot a_n \, x^{n-1} + (n-1) \cdot a_{n-1} \, x^{n-2} + \ldots + 2 \cdot a_2 \, x + a_1.$$

Dabei benutzen wir für ein Element a aus dem Koeffizientenbereich und eine natürliche Zahl n die Schreibweise

$$n \cdot a := \underbrace{a + a + \ldots + a}_{n-\mathrm{mal}}.$$

Diese Definition einer Ableitung macht Sinn in beliebigen Polynomringen. Die aus der Analysis bekannte Interpretation einer Ableitung als Steigung läßt sich natürlich nicht mehr aufrechterhalten. Wesentlich ist, daß die bekannten Rechenregeln für Ableitungen erhalten bleiben. Es gilt (Übung)

$$\begin{aligned} f(x) = h(x) + k(x) \quad &\Rightarrow \quad f'(x) = h'(x) + k'(x), \\ f(x) = h(x)k(x) \quad &\Rightarrow \quad f'(x) = h'(x)k(x) + h(x)k'(x). \end{aligned}$$

Proposition 6.7. *Sei p prim und sei $f(x) \in \mathbb{Z}_p[x]$. Dann gilt:*

i) $f'(x) = 0 \;\Rightarrow\; f(x) = h(x)^p$ *für ein* $h(x) \in \mathbb{Z}_p[x]$,

ii) $f'(x) \neq 0 \;\Rightarrow\; f(x) = [\mathrm{ggT}(f(x), f'(x))]^2 h(x)$, *mit einem* $h(x) \in \mathbb{Z}_p[x]$, *das keine quadratischen Teiler mehr enthält.*

Beweis. i): Die Koeffizienten $i \cdot a_i$ von $f'(x)$ verschwinden genau dann für alle i, falls $a_i = 0$ für alle $i \not\equiv 0 \bmod p$. Dann gibt es offenbar ein Polynom $h(x)$, so daß $f(x) = h(x^p)$. Die Behauptung folgt nun aus Proposition 6.6.

ii): Gilt $f(x) = s(x)^2 t(x)$, so teilt $s(x)$ auch

$$f'(x) \;=\; 2s(x)s'(x)t(x) + s(x)t'(x).$$

Sei umgekehrt $s(x)$ ein Teiler von $f(x)$ und $f'(x)$. Wir nehmen an, daß $s(x)$ irreduzibel in $\mathbb{Z}_p$ ist, so daß nach i) $s'(x) \neq 0$ gilt. Aus $f(x) = s(x)k(x)$ folgt $f'(x) = s'(x)k(x) + s(x)k'(x)$, folglich teilt $s(x)$ auch $s'(x)k(x)$. Da $s'(x) \neq 0$ einen Grad kleiner als $s(x)$ besitzt, muß $s(x)$ bereits $k(x)$ teilen. Folglich ist sogar $s(x)^2$ Teiler von $f(x)$. Insgesamt können wir feststellen, daß $s(x)$ genau dann $f(x)$ und $f'(x)$ teilt, falls $s(x)^2$ ein Teiler von $f(x)$ ist. Dies ergibt die zweite Behauptung. $\square$

Da sich größte gemeinsame Teiler problemlos mit dem Euklidischen Algorithmus berechnen lassen, können quadratische Anteile eines Polynoms in $\mathbb{Z}_p$ leicht ermittelt werden. Es bleibt die Aufgabe, ein Polynom zu zerlegen, dessen irreduzible Teiler alle verschieden sind. Dazu ist die folgende Aussage nützlich.

Proposition 6.8. *Sei p Primzahl und sei $f(x) \in \mathbb{Z}_p[x]$. Dann sind äquivalent:*

i) Es gilt $f(x) = f_1(x)f_2(x)$ mit teilerfremden Polynomen $f_i(x)$, $i = 1, 2$, so daß $0 < \deg(f_i) < \deg(f)$.

ii) Es gibt ein Polynom $g(x) \in \mathbb{Z}_p[x]$ mit $0 < \deg(g) < \deg(f)$, so daß $f(x)$ ein Teiler von $g(x)^p - g(x)$ ist.

Es ist dann

$$f(x) \;=\; \prod_{i=0}^{p-1} \mathrm{ggT}(f(x), g(x) - i)$$

eine nicht-triviale Faktorisierung von $f(x)$.

Beweis. Hat $g(x)$ die geforderten Eigenschaften, so ist wegen $\deg(g) \geq 1$ zunächst $g(x)^p - g(x) \neq 0$. Die Polynome $g(x), g(x)-1, \ldots, g(x)-p+1$ sind paarweise teilerfremd. Es folgt daher aus $f(x) \mid g(x)^p - g(x)$

$$\begin{aligned} f(x) &= \mathrm{ggT}(f(x), g(x)^p - g(x)) \\ &= \prod_{i=0}^{p-1} \mathrm{ggT}(f(x), g(x) - i). \end{aligned}$$

Mit $g(x)-i$ sind auch die Polynome $\mathrm{ggT}(f(x), g(x)-i)$ paarweise teilerfremd und haben einen Grad kleiner als $f(x)$. Mindestens zwei haben einen Grad größer 0, die man noch zu zwei teilerfremden Faktoren zusammenfassen kann. Dies ergibt die gewünschte Zerlegung von $f(x)$.

Sei umgekehrt $f(x) = f_1(x)f_2(x)$ eine Zerlegung von $f(x)$ in teilerfremde Faktoren. Wir betrachten das Kongruenzensystem

$$g(x) \;\equiv i \mod f_i(x), \quad i = 1, 2$$

im Polynomring $\mathbb{Z}_p[x]$. Nach dem Chinesischen Restsatz gibt es ein solches Polynom $g(x)$ von einem Grad kleiner als $\deg(f)$. Da $\deg(f_i) > 0$, kann $g(x)$ kein Polynom vom Grade 0 sein. Es folgt, daß $f_i(x)$ ein Teiler von $g(x) - i$ ist. Nach Proposition 6.6 teilt $f_i(x)$ auch $g(x)^p - g(x)$, und da die $f_i(x)$ teilerfremd sind, ist schließlich $f(x)$ ein Teiler von $g(x)^p - g(x)$. □

Damit man die letzte Proposition zum Faktorisieren eines Polynoms nutzen kann, muß man passende Polynome $g(x)$ bestimmen können. Es stellt sich heraus, daß diese Aufgabe auf das Lösen eines linearen Gleichungssystems führt. Wir machen den Ansatz

$$g(x) \;=\; b_{n-1}x^{n-1} + \ldots + b_1 x + b_0$$

mit $n = \deg(f)$. Dann folgt

$$g(x)^p \;=\; b_{n-1}x^{p(n-1)} + \ldots + b_1 x^p + b_0.$$

Division von x^{pi} durch $f(x)$ ergibt

$$x^{ip} \;=\; f(x)m_i(x) + r_i(x)$$

mit einem Rest $r_i(x)$ von einem Grad kleiner als n. Es folgt

$$g(x)^p \;\equiv\; b_{n-1}r_{n-1}(x) + \ldots + b_0 r_0(x) \bmod f(x).$$

$f(x)$ teilt also genau dann $g(x)^p - g(x)$, wenn es

$$b_{n-1}r_{n-1}(x) + \ldots + b_0 r_0(x) - b_{n-1}x^{n-1} - \ldots - b_0$$

teilt, falls dieses Polynom also verschwindet. Koeffizientenvergleich führt zu einem linearen Gleichungssystem für die b_i: Sind $r_{j,i}$ die Koeffizienten von $r_i(x)$,

$$r_i(x) \;=\; r_{n-1,i}x^{n-1} + \ldots + r_{0,i},$$

so ergibt sich in $\mathbb{Z}_p$ das lineare Gleichungssystem

$$b_j \;=\; \sum_{i=0}^{n-1} r_{j,i}b_i.$$

In Matrix-Schreibweise lautet es

$$(R - Id)\cdot b \;=\; 0,$$

mit

$$R \;:=\; \begin{pmatrix} r_{00} & \ldots & r_{0,n-1} \\ \vdots & & \vdots \\ r_{n-1,0} & \ldots & r_{n-1,n-1} \end{pmatrix}, \quad b \;:=\; \begin{pmatrix} b_0 \\ \vdots \\ b_{n-1} \end{pmatrix}.$$

Beispiel. Wir wollen das Polynom

$$f(x) \;=\; x^6 + x^5 + x^4 + x^3 + x^2 + x + 1$$

in $\mathbb{Z}_2[x]$ zerlegen. (In $\mathbb{Z}[x]$ ist $f(x)$ irreduzibel, wie wir oben aus dem Eisensteinschen Kriterium gefolgert haben.) Die Ableitung ist

$$f'(x) \;=\; 6x^5 + 5x^4 + 4x^3 + 3x^2 + 2x + 1 \;=\; x^4 + x^2 + 1.$$

Division von $f(x)$ durch $f'(x)$ mit Rest ergibt

$$f(x) = (x^2 + x)f'(x) + 1.$$

Der ggT von $f(x)$ und $f'(x)$ ist 1, $f(x)$ enthält daher keine quadratischen Anteile. Nun berechnen wir die Restpolynome $r_i(x)$:

$$\begin{array}{rcl@{\qquad}rcl}
x^0 &=& f(x)\cdot 0 + 1 & r_0(x) &=& 1 \\
x^2 &=& f(x)\cdot 0 + x^2 & r_1 &=& x^2 \\
x^4 &=& f(x)\cdot 0 + x^4 & r_2 &=& x^4 \\
x^6 &=& f(x) + (x^5 + \ldots + 1) & r_3(x) &=& x^5 + x^4 + x^3 + x^2 + x + 1 \\
x^8 &=& f(x)(x^2 + x) + x & r_4(x) &=& x \\
x^{10} &=& f(x)(x^4 + x^3) + x^3 & r_5(x) &=& x^3,
\end{array}$$

also

$$R = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Für b erhalten wir das Gleichungssystem

$$\begin{array}{rcl@{\quad}rcl@{\quad}rcl}
b_0 &=& b_0 + b_3, & b_2 &=& b_1 + b_3, & b_4 &=& b_2 + b_3, \\
b_1 &=& b_3 + b_4, & b_3 &=& b_3 + b_5, & b_5 &=& b_3,
\end{array}$$

das sich zu

$$b_1 = b_2 = b_4, \quad b_3 = b_5 = 0$$

zusammenfassen läßt. b_0 ist beliebig. Aus der Bedingung $\deg(g) > 0$ in Proposition 6.8 folgt, daß nur die Wahl $b_1 = 1$ zu einer Faktorisierung von $f(x)$ führt. Wir erhalten also

$$g(x) = x^4 + x^2 + x + b_0,$$

und die Faktorisierung

$$\begin{aligned}
f(x) &= \mathrm{ggT}(f(x), x^4 + x^2 + x) \cdot \mathrm{ggT}(f(x), x^4 + x^2 + x + 1) \\
&= (x^3 + x + 1)(x^3 + x^2 + 1).
\end{aligned}$$

Beide Faktoren haben in $\mathbb{Z}_2[x]$ keine Nullstellen und sind damit irreduzibel. $\square$

Die Methode von Berlekamp ist nicht das einzige Faktorisierungsverfahren in $\mathbb{Z}_p[x]$. Eine andere Methode beruht auf dem folgenden bemerkenswerten Satz, dessen Beweis wir später kennenlernen werden.

Satz 6.9. *Sei p Primzahl und n natürliche Zahl. Dann ist $x^{p^n}-x$ das Produkt aller normierten irreduziblen Polynome in $\mathbb{Z}_p[x]$ von einem Grade m, der n teilt.*

Indem man größte gemeinsame Teiler von $f(x)$ und $x^{p^n}-x$ für verschiedene n bildet, lassen sich Teiler von $f(x)$ herausfiltern. Details finden sich in der Monographie von Knuth, Vol. II.

Kapitel 7

Algorithmische Probleme für Boolesche Funktionen

7.1 Boolesche Funktionen

Boolesche Funktionen haben über die Logik hinaus, der sie entstammen, vielfältige Anwendungen.

Definition 7.1. *Eine Abbildung $f : \{0,1\}^n \to \{0,1\}$ heißt **n-wertige Boolesche Funktion**.*

In der Logik stehen 1 und 0 für die Wahrheitswerte *wahr* und *falsch*. Man kann aber auch an einen elektrischen Schaltkreis mit n Schaltern, einem Eingang und einem Ausgang denken. Dann stehen 1 und 0 für *geschlossen* bzw. *offen*. Der Schaltkreis wird durch eine Boolesche Funktion $f(x_1, \ldots, x_n)$ beschrieben, die angibt, für welche Schalterstellungen $x_1, \ldots, x_n$ die Verbindung zwischen Eingang und Ausgang offen oder geschlossen ist.

Nach dem Vorbild der Logik setzt man Boolesche Funktionen häufig aus den folgenden Bausteinen zusammen. Die n-wertigen Funktionen **Konjunktion** $\bigwedge$ und **Disjunktion** $\bigvee$ sind gegeben durch

$$\begin{aligned}
\bigwedge_{i=1}^{n} x_i &= x_1 \wedge x_2 \wedge \ldots \wedge x_n := \begin{cases} 1, & \text{falls } x_1 = \ldots = x_n = 1, \\ 0 & \text{sonst,} \end{cases} \\
\bigvee_{i=1}^{n} x_i &= x_1 \vee x_2 \vee \ldots \vee x_n := \begin{cases} 0, & \text{falls } x_1 = \ldots = x_n = 0, \\ 1 & \text{sonst,} \end{cases}
\end{aligned}$$

die 1-wertige **Negation** $\neg$ ist definiert als

$$\neg x := \begin{cases} 1, & \text{falls } x = 0, \\ 0, & \text{falls } x = 1. \end{cases}$$

Zwischen diesen Funktionen bestehen verschiedene Zusammenhänge, wie die de Morganschen Regeln

$$\neg \bigwedge_i x_i \;=\; \bigvee_i \neg x_i, \quad \neg \bigvee_i x_i \;=\; \bigwedge_i \neg x_i.$$

Es ist eine nützliche Tatsache, daß sich jede Boolesche Funktion aus Konjunktionen, Disjunktionen und Negationen zusammensetzen läßt.

Proposition 7.2. *Jede Boolesche Funktion f läßt sich darstellen als*

$$f(x_1, \ldots, x_n) \;=\; \bigvee_{i=1}^{k} \Big(\bigwedge_{j=1}^{l_i} y_{ij} \Big),$$

dabei steht jedes y_{ij} für einen der Ausdrücke $x_1, \neg x_1, \ldots, x_n, \neg x_n$. Genauso besitzt f eine Darstellung der Gestalt

$$f(x_1, \ldots, x_n) \;=\; \bigwedge_{i=1}^{s} \Big(\bigvee_{j=1}^{t_i} z_{ij} \Big),$$

mit $z_{ij} = x_1, \neg x_1, \ldots, x_n$ oder $\neg x_n$.

Man spricht von Darstellungen in **disjunktiver Normalform** bzw. **konjunktiver Normalform**, kurz DNF und KNF, die Unterausdrücke $\bigwedge_{j=1}^{l_i} y_{ij}$ bzw. $\bigvee_{j=1}^{t_i} z_{ij}$ heißen **Klauseln**. Die Grundbausteine sind $x_1, \ldots, x_n, \neg x_1, \ldots, \neg x_n$, sie heißen **Literale**. Diese Darstellungen sind nicht eindeutig.

Beweis. Sei D die Menge aller 01-Folgen $x_1, \ldots, x_n$, für die f den Wert 1 annimmt, und sei $(d_{11}, \ldots, d_{1n}), (d_{21}, \ldots, d_{2n}), \ldots, (d_{k1}, \ldots, d_{kn})$ eine Aufzählung der Elemente von D. Wir setzen $y_{ij} = x_j$, falls $d_{ij} = 1$, und $y_{ij} = \neg x_j$, falls $d_{ij} = 0$. Mit diesen y_{ij} (und $l_i = n$) gilt offenbar die erste Behauptung der Proposition. Genauso können wir $\neg f(x_1, \ldots, x_n)$ in disjunktiver Normalform darstellen. Die Darstellung von f in konjunktiver Normalform ergibt sich dann aus den de Morganschen Regeln. $\square$

Beispiel. Ein Netzwerk von Rechnern läßt sich durch einen zusammenhängenden Graphen G mit Eckenmenge E und Kantenmenge K darstellen. Ein Graph heißt **zusammenhängend**, falls es zwischen zwei beliebigen Ecken a und b einen Verbindungsweg gibt, also eine Folge von Ecken $e_0 = a, e_1, e_2, \ldots, e_m = b$, so daß e_{i-1} und e_i für alle $i = 1, \ldots, m$ benachbart sind.

Die Ecken symbolisieren die Rechner, die Kanten Verbindungen zwischen Rechnern. Wenn einzelne Verbindungen ausfallen, kann der Zusammenhang des Netzwerkes verlorengehen. Dies legt es nahe, dem Netzwerk eine Boolesche Funktion f in den Variablen x_k, $k \in K$ zuzuordnen. x_k nimmt die Werte 1 oder 0 an, falls die durch die Kante k symbolisierte Verbindung im Netzwerk ausgefallen ist bzw. noch steht. f soll genau dann den Wert 1 annehmen, wenn das Netzwerk nicht mehr zusammenhängt. Um für f eine Formel zu erhalten, bezeichnen wir für jede Teilmenge $F \subset E$ mit $K(F)$ die Menge der Kanten, deren einer Endpunkt in F und deren anderer Endpunkt in $E - F$ liegt. Sind alle Kanten zwischen F und $E - F$ ausgefallen, so geht der Zusammenhang verloren, daher gilt

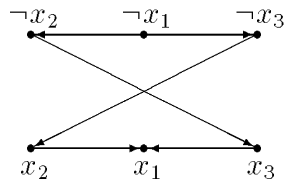
$$f((x_k)_{k \in K}) = \bigvee_{F \subset E} \Big(\bigwedge_{k \in K(F)} x_k \Big).$$

Dies ist eine Darstellung von f in disjunktiver Normalform. □

7.2 Das KNF-Entscheidungsproblem

Die Frage, ob eine vorgegebene Boolesche Funktion f erfüllbar ist, ob es also eine 01-Folge $x_1, \ldots, x_n$ gibt, so daß $f(x_1, \ldots, x_n) = 1$ gilt, ist eine klassische Aufgabenstellung der Komplexitätstheorie. Daß dies im Allgemeinen schwer entscheidbar ist, erkennt man schon daran, daß es insgesamt 2^{2^n} n-wertige Boolesche Funktionen gibt (jeder Teilmenge von $\{0,1\}^n$ entspricht genau eine Funktion). Ist f in DNF gegeben, bestehen keine Probleme: Dann ist f erfüllbar, wenn mindestens eine ihrer Klauseln erfüllt werden kann, und eine Klausel ist (als Konjunktion) erfüllbar, wenn sie nicht gleichzeitig x_i und $\neg x_i$ für ein i enthält.

Dagegen kann man Boolesche Funktionen in KNF als typisch für unser Entscheidungsproblem ansehen. Wir betrachten als erstes das **2SAT-Problem**, die Frage, ob eine Boolesche Funktion f in KNF, deren Klauseln aus genau 2 Literalen bestehen, erfüllbar ist. Für dieses Problem gibt es einen einfachen schnellen Algorithmus. Dazu repräsentieren wir f als gerichteten Graphen $G = G_f$, dessen Menge E der Ecken aus allen Literalen besteht. Die Menge der Kanten ist bei einem gerichteten Graphen eine Teilmenge $K \subset E \times E$, dabei steht $(a, b) \in K$ für eine gerichtete Kante von a nach b. Wir verbinden die Literalen a und b durch die Kante (a, b), wenn die Klausel $(\neg a \vee b)$ (bzw. $(b \vee \neg a)$) in f auftritt (dabei setzen wir $\neg\neg x = x$). Jeder Klausel gibt so Anlaß zu zwei Kanten. Ein gerichteter Weg von a nach b besteht aus ein Folge $(a_0, a_1), (a_1, a_2), \ldots, (a_{k-1}, a_k)$ von Kanten mit $a_0 = a$ und $a_k = b$. Es ist leicht zu sehen, daß es speziell im Graphen G_f genau dann einen Weg von a nach b gibt, falls ein Weg von $\neg b$ nach $\neg a$ existiert. Der folgende Graph stellt $f(x) = (x_1 \vee \neg x_2) \wedge (x_2 \vee x_3) \wedge (x_1 \vee \neg x_3)$ dar.



Behauptung. *Sei f in KNF gegeben, so daß jede Klausel genau zwei Literale enthält. Dann ist f genau dann unerfüllbar, wenn es für ein Literal x Wege von x nach $\neg x$ und von $\neg x$ nach x gibt.*

Beweis. Geben wir dem Literal a den Wert 1 und ist (a, b) Kante in G_f, so müssen wir auch b den Wert 1 geben, andernfalls ist die Klausel $(\neg a \vee b)$ nicht erfüllt. Legen wir also den Wert von a als 1 fest, so müssen alle Literale den Wert 1 erhalten, die von a aus in G_f erreichbar sind. Dies führt zu einem Konflikt, falls von a aus sowohl b wie auch $\neg b$ erreicht werden können. Dann gibt es auch Wege von $\neg b$ und b nach $\neg a$ und schließlich einen Weg von a nach $\neg a$. Gibt es keinen solchen Weg, so kann man a und allen Literalen, die von a aus erreichen kann, den Wert 1 geben, und die negierten Literale den Wert 0.

Analog ist die Situation, wenn man den Wert von a gleich 0 setzt. Dann muß man allen Literalen den Wert 1 geben, die von $\neg a$ aus erreicht werden können. Hier kommt es zum Konflikt, wenn ein Weg von $\neg a$ nach a existiert. Es ist nun leicht zu sehen, daß die angegebene Bedingung notwendig und hinreichend dafür ist, daß f nicht erfüllt werden kann. □

Diese Überlegung läßt sich zu einem effizienten Algorithmus ausbauen, der das 2SAT-Problem löst. Es ist nicht schwer zu sehen, daß seine Schrittzahl polynomiell ist, daß heißt von der Größenordnung $O(n^c)$ für eine geeignetes $c > 0$. Man sagt, 2SAT gehört zur Klasse $\boldsymbol{P}$.

Für 3SAT, die Frage, ob eine Boolesche Funktion in KNF, deren Klauseln alle aus 3 Literalen bestehen, ist die Situation völlig anders. Von Cook stammt das folgende Satz, der in der Komplexitätstheorie eine zentrale Rolle einnimmt.

Satz 7.3. 3SAT *ist ein NP-vollständiges Problem.*

Diesen Satz können wir hier nicht beweisen, wir wollen nur seinen anschaulichen Inhalt erläutern (die formalen Aspekte, insbesondere die genauen Definitionen der Komplexitätsklassen P und NP findet man in der Monographie von Papadimitriou). Entscheidungsprobleme, auf die die Antwort entweder *ja* oder *nein* lautet, werden in der Komplexitätstheorie in verschiedene Klassen eingeteilt. P enthält Probleme, für deren Entscheidung schnelle Algorithmen existieren. Die Klasse $\boldsymbol{NP}$ umfaßt sie. Anschaulich gesprochen enthält sie alle Probleme, die die folgende Eigenschaft haben: Lautet die Antwort *ja*, so gibt es dafür *Zeugen*. Das KNF-Problem gehört in diese Klasse. Ist f erfüllbar, so ist eine 01-Folge $x_1, \ldots, x_n$ mit $f(x_1, \ldots, x_n) = 1$ ein Zeuge für die Erfüllbarkeit. Die 3-Färbbarkeit von Graphen G gehört ebenfalls zu NP, hier ist ein Zeuge eine Färbung der Ecken, so daß benachbarte Ecken verschiedene Farben haben. Auch die Eigenschaft, daß eine Zahl n in verschiedene Primteiler zerfällt, gehört zu NP. Hier sind die Primteiler von n Zeugen für die Teilbarkeit. Wichtig ist, daß man schnell, in polynomieller Zeit nachprüfen kann, ob ein Zeuge vorliegt oder nicht. Dagegen ist nicht gefordert, daß man Zeugen schnell und effizient finden kann. Im Gegenteil, man geht davon aus, daß NP Entscheidungsprobleme enthält, die nur mit extrem hohen Rechenaufwand zu klären sind. Mathematisch handelt es sich um die berühmte Vermutung $\boldsymbol{NP} \neq \boldsymbol{P}$, die bis heute nicht bewiesen werden konnte.

Interessant ist, daß man Probleme in NP angeben kann, die mindestens so hart sind, wie jedes andere Problem in NP. Gemeint ist damit das Folgende: Besitzt man für ein solches Problem einen Algorithmus, so kann man mit diesem Algorithmus auch jedes andere Problem in NP entscheiden, mit einem zusätzlichen Rechenaufwand von vernachlässigbarer (polynomieller) Größe. Man zeigt also, daß sich jedes Problem in NP auf das betrachtete Problem reduzieren läßt (ähnlich wie wir in Abschnitt 4.4 das Faktorisieren von $m = pq$ zurückgeführt haben auf das Berechnen von Wurzeln quadratischer Reste modulo $m = pq$). Entscheidungsprobleme in NP mit dieser Eigenschaft heißen **$\boldsymbol{NP}$-vollständig**. Cooks bemerkenswertes Resultat ist, daß 3SAT zu diesen harten Problemen gehört. Ein anderes NP-vollständiges Problem ist das in Abschnitt 4.4 betrachtete 3-Färben von Graphen. Man kennt heute eine Vielzahl NP-vollständiger Probleme.

Nachdem es also schwer ist zu entscheiden, ob man alle Klauseln einer Booleschen Funktion f in KNF erfüllen kann, betrachten wir ein einfacheres Problem. Sei $0 < \alpha < 1$. Wir fragen, ob f eine *α-approximierbare Funktion* ist, ob es also eine Belegung der Variablen $x_1, \ldots, x_n$ gibt, so daß mindestens ein Anteil von α Klauseln von f erfüllt werden können. Wir setzen wieder voraus, daß jede Klausel aus genau 3 Literalen besteht, wobei in einer Klausel nicht gleichzeitig x_i und $\neg x_i$ enthalten seien. Mit $n_f(x_1, \ldots, x_n)$ bezeichnen wir die Anzahl der Klauseln von f, die bei der Belegung $x_1, \ldots, x_n$ der Variablen erfüllt sind.

Eine erste Beobachtung ist, daß jede solche Boolesche Funktion 7/8-

approximierbar sind. Zum Beweis ersetzen wir die Variablen durch unabhängige Zufallsvariablen $X_1, \ldots, X_n$, die die Werte 0 und 1 mit jeweils Wahrscheinlichkeit 1/2 annehmen. Dann ist jede Klausel mit Wahrscheinlichkeit 1/8 nicht erfüllt, und aus den bekannten Eigenschaften von Erwartungswerten folgt für den Erwartungswert der Zufallsvariablen $N_f := n_f(X_1, \ldots, X_n)$

$$\mathbf{E}N_f \;=\; \frac{7}{8}s,$$

wobei s die Anzahl der Klauseln in f bezeichne. Insbesondere muß es also, wie behauptet, eine Belegung der Variablen $x_1, \ldots, x_n$ geben, so daß $n_f(x_1, \ldots, x_n) \geq \frac{7}{8}s$. Wie kann man eine solche Belegung finden? Es ist naheliegend, sie durch eine Zufallswahl der x_i zu finden, man kann aber auch einen effizienten deterministischen Algorithmus angeben. Dazu definieren wir die (bedingten) Erwartungen

$$E(x_1, \ldots, x_i) \;:=\; \mathbf{E}(N_f \mid X_1 = x_1, \ldots, X_i = x_i),$$

die sich ergeben, wenn wir die ersten i Variablen gleich $x_1, \ldots, x_i$ setzen und nur $X_{i+1}, \ldots, X_n$ zufällig wählen. Diese Erwartungswerte lassen sich leicht ausrechnen: Man bestimme für jede Klausel in f die Wahrscheinlichkeit, daß sie erfüllt ist (es kommen dafür die Werte 0, 1/8, 1/4, 1/2 und 1 in Frage), und summiere diese Wahrscheinlichkeiten dann auf. Nach den bekannten Regeln für Erwartungswerte gilt

$$\mathbf{E}N_f \;=\; \frac{1}{2}\,E(0) + \frac{1}{2}\,E(1),$$

daher ist entweder $E(0)$ oder $E(1) \geq \frac{7}{8}s$. Im ersten Fall setzen wir $x_1 = 0$, andernfalls $x_1 = 1$. Nun benutzen wir die Gleichung

$$E(x_1) \;=\; \frac{1}{2}\,E(x_1, 0) + \frac{1}{2}\,E(x_1, 1)$$

und wählen entsprechend x_2 so, daß $E(x_1, x_2) \geq \frac{7}{8}s$. Nach n Schritten erhalten wir $x_1, \ldots, x_n$ derart, daß

$$n_f(x_1, \ldots, x_n) \;=\; E(x_1, \ldots, x_n) \;\geq\; \frac{7}{8}s.$$

Dieses effiziente Verfahren heißt **Algorithmus von Johnson**. Man sagt, daß er aus dem probabilistischen Verfahren durch **Derandomisieren** gewonnen ist.

Es fragt sich, ob man ähnlich effizient auch für größere Werte als 7/8 vorgehen kann. Das folgende Resultat von Hastad verneint dies.

Satz 7.4. *Sei f eine Boolesche Funktion in KNF mit Klauseln der Länge 3. Dann ist das Entscheidungsproblem, ob f α-approximierbar ist, für alle $\alpha > 7/8$ NP-vollständig.*

7.3 Das DNF-Zählproblem

Abzählprobleme sind im Allgemeinen aufwendiger als Entscheidungsprobleme. Wie wir gesehen haben, bereitet die Frage, ob eine Boolesche Funktion f in DNF erfüllbar ist, keine Schwierigkeiten. Dies ändert sich, wenn wir danach fragen, wieviele Belegungen der Variablen $x_1, \ldots, x_n$ es gibt, so daß $f(x_1, \ldots, x_n) = 1$ gilt. Dies ist das DNF-Zählproblem (es wird auch als ♯SAT bezeichnet). Zur Motivation erinnern wir an die Boolesche Funktion in DNF aus dem Beispiel in Abschnitt 7.1. Die (relative) Anzahl der erfüllenden Belegungen dieser Funktion kann als Maßzahl dafür dienen, wie gut das Netzwerk gegen den Ausfall einzelner Verbindungen zwischen Rechnern geschützt ist. (Einschränkungen an die Anzahl der Literale pro Klausel, wie im letzten Abschnitt, stellen wir nun nicht mehr.)

Die Klasse der Abzählprobleme, die NP entspricht, wird mit $\sharp P$ bezeichnet. Informell gesprochen besteht sie aus Problemen, bei denen die Größe einer Teilmenge T in einer Grundmenge Q zu bestimmen ist; dabei ist gefordert, daß man für jedes $x \in Q$ polynomiell schnell entscheiden kann, ob x zu T gehört oder nicht. Dies bedeutet noch nicht, daß die Größe von T effektiv berechnet werden kann. Ganz im Gegenteil geht man davon aus, daß es in $\sharp P$ wie in NP rechnerisch völlig unzugängliche Probleme gibt. Ein Kandidat dafür ist das DNF-Abzählproblem, denn es gilt das folgende Resultat.

Satz 7.5. *Das DNF-Abzählbarkeitsproblem ist $\sharp P$-vollständig.*

Das heißt, jedes Abzählproblem in $\sharp P$ kann algorithmisch auf das DNF-Abzählproblem reduziert werden, mit einem vernachlässigbaren zusätzlichen Rechenaufwand von polynomialer Größe.

An ein exaktes Abzählen aller erfüllenden Belegungen einer Booleschen Funktion f ist also im Allgemeinen nicht zu denken. Um so überraschender ist, daß es effiziente probabilistische Algorithmen gibt, die das DNF-Zählproblem zwar nicht exakt, dafür aber mit hoher Wahrscheinlichkeit lösen, bis auf einen relativen Fehler, den man beliebig klein machen kann.

Die Grundidee ist, $f(X_1, .., X_n)$ für eine zufällige Belegung $X_1, \ldots, X_n$ auszuwerten. Ist $(X_1, \ldots, X_n)$ gleichförmig aus $\{0,1\}^n$ gezogen, so ist die Anzahl aller erfüllenden Belegungen von f gleich $2^n \cdot p_f$, wobei p_f die Wahrscheinlichkeit bezeichne, daß $f(X_1, \ldots, X_n)$ den Wert 1 annimmt. Nun kann man versuchen p_f durch eine relative Häufigkeit zu schätzen, indem man die Zufallswahl mehrfach unabhängig wiederholt. Gibt es nur wenige Belegungen, die f erfüllen, so ist p_f exponentiell klein, und man muß ein exponentielle Anzahl von Zufallsversuchen durchführen, um eine einigermaßen verläßliche

Schätzung von p_f zu erhalten. Daher ist dieses Verfahren im Allgemeinen nicht brauchbar.

Eine bessere Möglichkeit ist von Karp, Luby und Madras (1989) vorgeschlagen worden. Wir gehen wieder davon aus, daß f in DNF vorliegt, mit insgesamt s Klauseln. Sei K_i die Menge aller Belegungen $x = (x_1, \ldots, x_n)$, die die i-te Klausel erfüllen. Wir betrachen die Menge

$$K_1 \oplus \ldots \oplus K_s \;:=\; \{(i,x)\;:\; i = 1, \ldots, s,\; x \in K_i\}$$

(die *Multimengenvereinigung* der K_i) und deren Teilmenge

$$V \;:\; \{(i,x) \in K_1 \oplus \ldots \oplus K_s \;:\; x \notin K_j \text{ für alle } j < i\}.$$

Offenbar gilt $|V| = |\bigcup_i K_i|$. Für eine Boolesche Funktion in DNF ist $\bigcup_i K_i$ die Menge aller erfüllenden Belegungen ist, daher langt es, die Mächtigkeit von V abzuschätzen. Sei dazu (I, X) ein zufälliges, gleichförmig aus $K_1 \oplus \ldots \oplus K_s$ gezogenes Element und q_f die Wahrscheinlichkeit, daß (I, X) in V liegt. Dann ist die Anzahl der erfüllenden Belegungen von f gegeben durch $q_f \cdot |K_1 \oplus \ldots \oplus K_s|$. $|K_1 \oplus \ldots \oplus K_s| = |K_1| + \ldots + |K_s|$ ist leicht zu bestimmen, denn es gilt $|K_i| = 2^{n-r_i}$, wobei r_i die Anzahl der Literale in der i-ten Klausel bezeichnet (den uninteressanten Fall, daß eine Klausel nicht erfüllt werden kann, daß sie also x_j und gleichzeitig $\neg x_j$ für ein j enthält, lassen wir beiseite).

Dieser Ansatz ist gut, da q_f nicht mehr beliebig klein werden kann, bei vorgegebener Anzahl s von Klauseln. Es gilt nämlich $q_f \geq 1/s$. Zum Beweis bemerke man, daß $|V| \geq \max_i |K_i|$ und $|K_1 \oplus \ldots \oplus K_s| \leq s \cdot \max_i |K_i|$. Um also q_f verläßlich zu schätzen, muß man die Zufallswahl nicht allzu häufig wiederholen. (Eine quantitative Analyse erfordert die exponentielle Ungleichung von Chernoff, wir lassen dies beiseite.) Es bleibt die Frage, wie man in effizienter Weise aus $K_1 \oplus \ldots \oplus K_s$ ein gleichförmig verteiltes Element auswählen kann. Dies kann man in zwei Stufen durchführen.

- Wähle ein zufälliges Element I aus $\{1, 2, \ldots, s\}$, und zwar i mit Wahrscheinlichkeit $|K_i| \cdot (|K_1| + \ldots + |K_s|)^{-1}$.

- Ist i ausgewählt, so ziehe uniform ein Element $X = (X_1, \ldots, X_n)$ aus K_i. Dies bedeutet, daß man die Variablen, die in der i-ten Klausel auftreten, so festlegen muß, daß diese Klausel erfüllt ist. Die anderen Variablen wähle man voneinander unabhängig per Münzwurf.

Kapitel 8

Hamming-Kodes

8.1 Ein fehlerkorrigierender Kode

Bei der Übertragung von Nachrichten durch einen gestörten Kanal entstehen Fehler, die der Empfänger gern erkennen würde. Man sendet daher die Nachricht in redundanter Form, die Information wird teilweise wiederholt oder in eine spezielle Form gebracht, die es erlaubt, Fehler zu entdecken. Sei zum Beispiel $ab \dots e$ eine binäre, aus den Bits 0 und 1 zusammengesetzte Nachricht. Man kann ihr das Prüfbit $f \equiv a + b + \dots + e \bmod 2$ anhängen und $ab \dots ef$ senden. Der Empfänger kann dann überprüfen, ob $a + b + \dots + e + f \equiv 0 \bmod 2$ gilt (parity check), und so einen Fehler erkennen. Solche Prüfzeichen sind in den Strichkodes auf Warenpackungen, den Internationalen Standard-Buchnummern (ISBN) und den Nummern der deutschen Geldscheine enthalten.

Manchmal ist es wichtig, Übertragungsfehler auch korrigieren zu können, man denke etwa an Satellitendaten, die nur einmal gesendet werden. Zum Beispiel kann man jedes bit a 3-fach senden: aaa. Der Empfänger entscheidet sich für das Bit, das in der Nachricht überwiegt. Dieser Kode korrigiert auf drei Buchstaben einen Fehler, allerdings trägt nur jedes dritte bit Information. Fehlerkorrigierende Kodes finden u.a. Anwendung in CD-Spielern zur Beseitigung von Ablesefehlern.

Wir beschreiben nun einen binären Kode, der in 7 bits einen Fehler korrigiert, dabei 4 bits Information überträgt. Die Worte sind Tupel aus 1en und 0en, jedes Tupel der Länge 4 wird in ein Tupel der Länge 7 verwandelt. Die Buchstaben werden als Elemente von $\mathbb{Z}_2$ aufgefaßt. Setze

$$H := \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

H ist spaltenweise aus allen 01-Tripeln aufgebaut, abgesehen vom Tripel aus drei Nullen.

Kodieren. Soll die Botschaft $w = abcd$ übertragen werden, so wird der Vektor

$$k = (xyzabcd)$$

gesendet, für den

$$H \cdot k^t = 0,$$

gilt. In Gleichungen in $\mathbb{Z}_2$ ausgedrückt bedeutet dies

$$\begin{aligned} x &= a+b+d, \\ y &= a+c+d \\ z &= b+c+d. \end{aligned}$$

Dekodieren. Wird $\widetilde{k} = (XYZABCD)$ empfangen, so bestimmt man $H \cdot \widetilde{k}^t$. Ist $H \cdot \widetilde{k}^t = 0$, so wird an $\widetilde{k}$ keine Änderung vorgenommen. Andernfalls ist $H \cdot \widetilde{k}^t$ eine der Spalten von H. Dann wird $\widetilde{k}$ an der entsprechenden Stelle korrigiert.

Tritt ein einziger Fehler auf, so enthält der Vektor $f = \widetilde{k} - k$ genau eine 1 und sechs 0en, und

$$H \cdot \widetilde{k}^t = H \cdot f^t$$

taucht als Spalte in H an der Position auf, an der es in $\widetilde{k}$ zu dem Übertragungsfehler gekommen ist. Dies ermöglicht es dem Empfänger, den Fehler zu korrigieren. Für $\widetilde{k} = (1011011)$ etwa gilt $H \cdot \widetilde{k}^t = (111)^t$, der Fehler ist (vorausgesetzt es liegt nur ein Fehler vor!) an der letzten Stelle aufgetreten, und der Empfänger dekodiert $k = (1011010)$.

Dieser von Hamming 1950 vorgeschlagene Kode ist in der Korrektur von Fehlern recht wirksam. Ist $p = 0,01$ die Wahrscheinlichkeit, daß ein Bit falsch empfangen wird, und treten Fehler unabhängig voneinander auf, so ist die Wahrscheinlichkeit, daß die dekodierte Nachricht Fehler enthält (d.h. mindestens zwei Übertragungsfehler vorliegen), gleich $1-(1-p)^7 - 7p(1-p)^6 = 0,0020$. Im Kontrast dazu ist die Wahrscheinlichkeit, daß bei Übertragung von 4 bits mindestens ein Fehler auftritt, fast 20-mal so groß, nämlich $1-(1-p)^4 = 0,039$.

Durch Modifikation läßt sich erreichen, daß der Hamming-Kode zwei Fehler erkennen und einen korrigieren kann. Sei

$$\overline{H} := \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & \dots & 1 \\ 0 & & \\ 0 & H & \\ 0 & & \end{pmatrix}.$$

Das Wort $abcd$ wird als der Vektor $k = (vxyzabcd)$ kodiert, der die Gleichung

$$\overline{H} \cdot k^t \;=\; 0$$

erfüllt. x, y, z werden wie oben berechnet, ferner gilt $v+x+y+z+a+b+c+d = 0$, was zu der weiteren Gleichung

$$v \;=\; a + b + c$$

in $\mathbb{Z}_2$ führt. Der Empfänger bestimmt bei Empfang von $\widetilde{k} = (VXYZABCD)$ wieder $\overline{H} \cdot \widetilde{k}^t$. Wir unterscheiden 3 Fälle.

Fall 1. Tritt kein Übertragungsfehler ein, so ist $\overline{H} \cdot \widetilde{k}^t = 0$.

Fall 2. Bei einem Fehler enthält der Vektor $f = \widetilde{k} - k$ genau eine 1, und $\overline{H} \cdot \widetilde{k}^t = \overline{H} \cdot f^t$ ist wieder die Spalte in $\overline{H}$ an der Stelle, an der in $\widetilde{k}$ der Fehler steckt.

Fall 3. Bei zwei Fehlern enthält f genau zwei 1en. Dann ist der oberste Eintrag von $\overline{H} \cdot \widetilde{k}^t$ eine 0, und $\overline{H} \cdot \widetilde{k}^t \neq 0$, nämlich die Summe zweier Spalten aus $\overline{H}$. Man erkennt so, daß es zwei Übertragungsfehler gibt, kann ihre Position jedoch nicht mehr erkennen.

Hammings Konstruktion läßt sich verallgemeinern.

Definition 8.1. *Sei H eine binäre $l \times (2^l - 1)$ Matrix, die als Spalten alle 01-Tupel der Länge l besitzt, abgesehen vom Tupel aus lauter Nullen. Der zugehörige binäre Kode heißt* ***(n,m)-Hamming-Kode****, mit $n = 2^l - 1$ und $m = n - l$.*

Diese Kodes, die Worte der Länge m als Worte der Länge $n = m + l$ kodieren, können alle einen Fehler korrigieren. Man kann sie erneut so erweitern, so daß sie 2 Fehler erkennen. Von praktischer Bedeutung ist allerdings nur der (7,4)-Hamming-Kode. Das Problem besteht darin, daß Hamming-Kodes nicht nur Fehler korrigieren, sondern auch zusätzliche Fehler erzeugen können, wenn es nämlich pro Wort zu mehr als einem Fehler kommt. Diese Tendenz verstärkt sich mit wachsendem l. Die folgende Tabelle belegt dies. Sie enthält den Erwartungswert E der Fehler, die ein Kodewort eines (n, m)-Hamming-Kode enthält, unter der Annahme, daß Fehler pro bit mit Wahrscheinlichkeit $p = 0,02$ und unabhängig voneinander auftreten.

	l	E	$(2^l - 1)p$
$(7,4)$ − Kode	3	0,024	0,14
$(31,26)$ − Kode	5	0,41	0,62
$(127,120)$ − Kode	7	3,06	2,54

Zum Vergleich ist die erwartete Anzahl von Fehlern in einer unkodierten Nachricht der Länge $2^l - 1$ angegeben. Das Resultat ist ernüchternd: Für $l = 7$ erzeugt der Kode mehr Fehler, als er beseitigt. Genaueres dazu findet sich in Abschnitt 8.3.

Wir wollen nun dem (7,4)-Hamming-Kode eine alternative Gestalt geben. Es wird sich in Kapitel 10 zeigen, daß dieser algebraische Ansatz ganz andere Möglichkeiten der Verallgemeinerung bietet.

8.2 Ein endlicher Körper

$\mathbb{Z}_2$ ist nicht der einzige Körper, der für binäre Kodes nützlich ist. Wir betrachten das irreduzible Polynom

$$g(x) \;=\; x^3 + x + 1$$

in dem Euklidischen Ring $\mathbb{Z}_2[x]$ und das zugehörige Hauptideal

$$I \;=\; I_g \;:=\; \{f(x)g(x) \;:\; f(x) \in \mathbb{Z}_2[x]\}.$$

Wie früher beim Übergang von $\mathbb{Z}$ zu $\mathbb{Z}_p$ gehen wir jetzt von $\mathbb{Z}_2[x]$ zu dem Restklassenring

$$K \;:=\; \mathbb{Z}_2[x]/I$$

über. $g(x)$ irreduzibel, denn es hat keine Nullstellen. Daher ist K (wie im Fall $\mathbb{Z}_p$) ein Körper. Inverse Elemente von Restklassen verschafft man sich wie schon in $\mathbb{Z}_p^*$ mit dem Satz von Bézout.

K ist ein endlicher Körper. Seine Elemente werden von den Polynomen in $\mathbb{Z}_2[x]$ vom Grade höchstens 2 repräsentiert, von denen es 8 gibt. Ist α die Restklasse, in der das Monom x in $\mathbb{Z}_2[x]$ liegt, so besteht K aus den Elementen

$$a + b\alpha + c\alpha^2 \quad \text{mit } a, b, c \in \{0, 1\}.$$

Diese Ausdrücke werden in der üblichen Weise addiert und multipliziert, unter Berücksichtigung der Gleichung

$$g(\alpha) \;=\; 0 \quad \text{bzw.} \quad \alpha^3 \;=\; \alpha + 1.$$

Die Gruppe $K^* = K - \{0\}$ ist zyklisch mit Erzeuger α. Wir können nämlich die Elemente von K^* darstellen als:

$$\begin{array}{ll}
\alpha & \alpha^5 = \alpha^3 + \alpha^2 = \alpha^2 + \alpha + 1 \\
\alpha^2 & \alpha^6 = \alpha^3 + \alpha^2 + \alpha = \alpha^2 + 1 \\
\alpha^3 = \alpha + 1 & \alpha^7 = \alpha^3 + \alpha = 1 \\
\alpha^4 = \alpha^2 + \alpha &
\end{array}$$

Man sagt, α ist **primitives Element** von K. Nach Proposition 6.6 gilt $g(\alpha^2) = g(\alpha)^2 = 0$ und $g(\alpha^4) = g(\alpha^2)^2 = 0$. Es sind also auch α^2 und α^4 Nullstellen von $g(x)$, was man leicht direkt nachrechnet. $g(x)$ zerfällt damit in $K[x]$ vollständig in lineare Polynome. Man bezeichnet K deswegen als den Zerfällungskörper von $g(x)$. K ist zugleich Zerfällungskörper von $x^3 + x^2 + 1$, dem zweiten irreduziblen Polynoms dritten Grades in $\mathbb{Z}_2[x]$. Dieses Polynom hat die Nullstellen $\alpha^6 = \alpha^{-1}$, $\alpha^5 = \alpha^{-2}$ und $\alpha^3 = \alpha^{-4}$.

Wir benutzen nun den Körper K zum Kodieren einer binären Nachricht.

Kodieren. Wir wollen das binäre Wort $w = abcd$ kodieren. Dazu bilden wir das Polynom

$$k_1(x) \;=\; ax^6 + bx^5 + cx^4 + dx^3$$

und teilen es in $\mathbb{Z}_2[x]$ mit Rest durch $g(x)$:

$$k_1(x) \;=\; m(x)g(x) + k_2(x) \;=\; m(x)g(x) + rx^2 + sx + t$$

mit $r, s, t \in \mathbb{Z}_2[x]$. Da $1 = -1$ in $\mathbb{Z}_2$, gilt

$$\begin{aligned} k(x) &= ax^6 + bx^5 + cx^4 + dx^3 + rx^2 + sx + t \\ &= k_1(x) + k_2(x) \;=\; m(x)g(x). \end{aligned}$$

Die Nachricht $abcd$ kodieren wir nun als $k = (abcdrst)$.

Dekodieren. Zum Dekodieren fassen wir die Polynome als Elemente von $K[x]$ auf. Dies ist möglich, da $\mathbb{Z}_2$ in K enthalten ist. Wir können also die Polynome in α auswerten. Es gilt $g(\alpha) = 0$ und daher auch

$$k(\alpha) \;=\; 0.$$

Dies ist die Kontrollgleichung zum Dekodieren. Erhält der Empfänger also die Nachricht $\widetilde{k} = (ABCDRST)$, so bildet er das Polynom

$$\widetilde{k}(x) = Ax^6 + Bx^5 + Cx^4 + Dx^3 + Rx^2 + Sx + T$$

und überprüft, ob $\widetilde{k}(\alpha) = 0$ gilt. Wir betrachten zwei Fälle:

Fall 1. Wenn kein Fehler aufgetreten ist, ist $k(x) - \widetilde{k}(x) = 0$ und $\widetilde{k}(\alpha) = 0$.

Fall 2. Wenn es einen Übertragungsfehler gibt, dann gilt $k(x) - \widetilde{k}(x) = x^e$. e ist die fehlerhafte Stelle in der Nachricht. Es folgt

$$\widetilde{k}(\alpha) \;=\; \alpha^e,$$

und diese Größe bestimmt e eindeutig, da α ein Erzeuger von K^* ist. Der Empfänger kann also e rekonstruieren.

Dieses Kodierungsschema ist offenbar zum (7,4)-Hamming-Kode äquivalent, wir werden dies in Kapitel 10 weiter präzisieren.

Beispiel. Um 1101 zu kodieren, teilt man $x^6 + x^5 + x^3$ in $\mathbb{Z}_2[x]$ durch das Polynom $x^3 + x + 1$:

$$(x^6 + x^5 + x^3) \; = \; (x^3 + x^2 + x + 1)(x^3 + x + 1) + 1.$$

Es gilt

$$k(x) \; = \; x^6 + x^5 + x^3 + 1,$$

die kodierte Nachricht ist also 1101001. – Bei Empfang von 1$\underline{0}$01001 berechnet der Empfänger

$$\alpha^6 + \alpha^3 + 1 \; = \; (\alpha^2 + 1) + (\alpha + 1) + 1 \; = \; \alpha^5.$$

Der Fehler befindet sich an der 2. Stelle, und die Nachricht heißt korrigiert 1$\underline{1}$01001. □

8.3 Die mittlere Fehlerzahl

In diesem Anhang bestimmen wir die erwartete Fehlerzahl pro Kodewort des (n, m)-Hamming-Kodes. Die kodierte Nachricht hat die Länge $n = 2^l - 1$. Wir nehmen an, daß die bits unabhängig voneinander jeweils mit Wahrscheinlichkeit p falsch übertragen werden.

Die zulässigen Kodewörter k sind durch die Eigenschaft $H \cdot k^t = 0$ charakterisiert. Besitzt die empfangene Nachricht $\widetilde{k}$ diese Eigenschaft nicht, so verwandelt der Hamming-Kode die Nachricht in ein zulässiges Kodewort $\widehat{k}$, das im Allgemeinen nicht mit der gesendeten Nachricht k übereinstimmen wird. Die Anzahl der Fehler zwischen gesendeter und dekodierter Nachricht ist die Anzahl der Einsen in der Differenz $\delta = k - \widehat{k}$. Es gilt $H \cdot \delta^t = H \cdot k^t - H \cdot \widehat{k}^t = 0$, δ ist also ebenfalls ein zulässiges Kodewort.

Sei umgekehrt δ ein Vektor aus i Einsen und $n - i$ Nullen, so daß $H \cdot \delta^t = 0$. δ kann bei der Übertragung auf drei verschiedene Weisen als Differenz von k und $\widehat{k}$ entstehen. Entweder gibt δ wirklich die Stellen an, an denen es zu Übertragungsfehlern gekommen ist. Wegen $H \cdot \delta^t = 0$ gilt dann $H \cdot \widetilde{k}^t = 0$. Der Hamming-Kode erkennt keinen Fehler und nimmt keine Korrekturen vor. Dieser Fall tritt mit Wahrscheinlichkeit $p^i q^{n-i}$, $q = 1 - p$ ein. Oder aber es liegt ein weiterer Übertragungsfehler vor, an einer Stelle s, an der in δ eine 0 steht. Dann ist

$$H \cdot \widetilde{k}^t \; = \; H \cdot (\widetilde{k} - k - \delta)^t$$

die s-te Spalte von H, denn $\widetilde{k} - k - \delta$ besteht an der s-ten Stelle aus einer 1 und sonst aus Nullen. Der Hamming-Kode korrigiert also diesen Fehler. Der zusätzliche Fehler kann an den $n - i$ Positionen stehen, an denen δ

eine 0 aufweist, dieser Fall tritt also mit Wahrscheinlichkeit $(n-i)p^{i+1}q^{n-i-1}$ ein. Schließlich ist es möglich, daß nur $i-1$ Fehler aufgetreten sind, einer der i Stellen, die in δ eine 1 aufweisen, ist nicht falsch beim Empfänger angekommen. Auch in diesem Fall ist

$$H \cdot \widetilde{k}^t \;=\; H \cdot (\widetilde{k} - k - \delta)^t$$

die entsprechende Spalte in H. In diesem Fall fügt also der Hamming-Kode an der entsprechenden Stelle einen weiteren Fehler ein. Dieser letzte Fall tritt mit Wahrscheinlichkeit $ip^{i-1}q^{n-i+1}$ ein. Setzen wir

$$d_i \;:=\; \text{Anzahl der zulässigen Kodewörter mit } i \text{ Einsen},$$

so ergibt sich für die erwartete Anzahl der Fehler pro Kodewort, die der Hamming-Kode produziert, die Formel

$$E_l \;=\; \sum_{i=0}^{n} i d_i \{p^i q^{n-i} + (n-i)p^{i+1}q^{n-i-1} + ip^{i-1}q^{n-i+1}\}.$$

Wir stellen nun eine Rekursionsformel für die d_i auf. Sei $u = (u_1 u_2 \ldots u_n)$ ein Vektor aus i Einsen und $n-i$ Nullen. Wir betrachten den Vektor $v = H \cdot u^t$. Es bestehen verschiedene Möglichkeiten. Ist $v = 0$, so ist u ein zulässiges Kodewort. Ist dagegen $v \neq 0$ so taucht v als Spalte in H auf, etwa als s-te Spalte. Dann ändern wir u an der s-ten Position ab und erhalten einen Vektor u' mit der Eigenschaft $H \cdot u' = v \pm v = 0$. Nun ist also u' ein zulässiges Kodewort mit $i+1$ oder $i-1$ Einsen.

Geht man nach diesem Schema durch alle Vektoren u mit i Einsen, so erhält man offenbar jedes Kodeworte mit i Einsen einmal. Die Kodeworte mit $i+1$ Einsen entstehen dagegen $(i+1)$-mal, denn an jeder der $i+1$ Positionen, in der in u' eine Eins steht, kann beim Übergang von u nach u' eine 0 in eine 1 verwandelt werden. Analog erhält man die Kodeworte mit $i-1$ Einsen $(n-i+1)$-mal. Es folgt

$$\binom{n}{i} \;=\; d_i + (i+1)d_{i+1} + (n-i+1)d_{i-1}, \quad i = 0, \ldots, n,$$

mit der Vereinbarung $d_{-1} = d_{n+1} = 0$. Diese rekursiven Gleichungen legen die d_i eindeutig fest. Um sie nach d_i aufzulösen, benutzen wir die Methode der erzeugenden Funktionen und betrachten die **erzeugende Funktion**

$$d(t) \;:=\; d_0 + d_1 t + \ldots + d_n t^n, t \in \mathbb{R},$$

der Zahlen d_i. Multiplizieren wir die Rekursionsgleichungen mit t^i und summieren über $i = 0, \ldots, n$, so ergibt sich

$$\begin{aligned}(1+t)^n &= \sum_{i=0}^{n} d_i t^i + \sum_{i=0}^{n} (i+1) d_{i+1} t^i + \sum_{i=0}^{n} (n-i+1) d_{i-1} t^i \\ &= d(t) + d'(t) + ntd(t) - t^2 d'(t).\end{aligned}$$

Dies ist eine Differentialgleichung erster Ordnung, ihre Lösung ist eindeutig durch $d(0) = d_0 = 1$ bestimmt. Man rechnet nach, daß die Lösung gegeben ist durch

$$d(t) = \frac{1}{n+1}(1+t)^n + \frac{n}{n+1}(1+t)^{\frac{n-1}{2}}(1-t)^{\frac{n+1}{2}}.$$

Mit dieser Formel können wir nun E_l explizit bestimmen. Eine einfache Rechnung zeigt

$$E_l = q^n \{(1 + t + (n-1)t^2) d'(t) + (t - t^3) d''(t)\} \quad \text{mit } t = \frac{p}{q}.$$

Mit der Formel für $d(t)$ erhält man nach Ausführung der Differentiationen schließlich

$$\begin{aligned}E_l &= \left(1 + (n-1)\frac{p^2}{q}\right)\left(\frac{n}{n+1}[1 - (1 + (n-1)p)(1-2p)^{\frac{n-1}{2}}]\right) \\ &\quad + \frac{p(1-2p)}{q}\frac{n(n-1)}{n+1}\left(1 + (np^2 - 3p^2 + 4p - 1)(1-2p)^{\frac{n-3}{2}}\right).\end{aligned}$$

Um einen übersichtlicheren Ausdruck zu erhalten, setzen wir $p = \lambda/n$. Dann ist λ die mittlere Fehlerzahl in einer Nachricht der Länge n. Lassen wir nun l gegen ∞ gehen, so folgt

$$E_l \to \lambda + 1 - (1 + 2\lambda)\exp(-\lambda).$$

Dieser Formel kann man entnehmen, daß ein Hamming-Kode für großes l im Mittel mehr als λ Fehler produziert, falls $(1 + 2\lambda)\exp(-\lambda) < 1$ gilt. Für $\lambda > 1,26$ ist dies der Fall. Dies ist ein wahrhaft enttäuschendes Resultat: Falls pro Kodewort im Mittel mehr als 1,26 Fehler auftreten, erzeugt ein Hamming-Kode für großes l im Durchschnitt mehr Fehler, als er korrigiert. Dies unterstreicht, daß die (n, m)-Hamming-Kodes für großes l unbrauchbar sind.

Abschließend leiten wir noch eine explizite Formel für die Zahlen d_i ab. Da $n = 2^l - 1$ eine ungerade Zahl ist, gilt

$$
\begin{aligned}
d(t) &= \frac{1}{n+1}(1+t)^n + \frac{n}{n+1}(1-t)(1-t^2)^{\frac{n-1}{2}} \\
&= \frac{1}{n+1}\sum_{k=0}^{n}\binom{n}{k}t^k + \frac{n}{n+1}(1-t)\sum_{j=0}^{\frac{n-1}{2}}\binom{\frac{n-1}{2}}{j}(-1)^j t^{2j} \\
&= \frac{1}{n+1}\sum_{j=0}^{\frac{n-1}{2}}\left(\binom{n}{2j}t^{2j} + \binom{n}{2j+1}t^{2j+1}\right) \\
&\quad + \frac{n}{n+1}\sum_{j=0}^{\frac{n-1}{2}}\left((-1)^j\binom{\frac{n-1}{2}}{j}t^{2j} - (-1)^j\binom{\frac{n-1}{2}}{j}t^{2j+1}\right).
\end{aligned}
$$

Ein Koeffizientenvergleich ergibt

$$
\begin{aligned}
d_{2j} &= \frac{1}{n+1}\binom{n}{2j} + \frac{n}{n+1}(-1)^j\binom{\frac{n-1}{2}}{j}, \\
d_{2j+1} &= \frac{1}{n+1}\binom{n}{2j+1} - \frac{n}{n+1}(-1)^j\binom{\frac{n-1}{2}}{j}.
\end{aligned}
$$

Kapitel 9

Endliche Körper

Für jede Primzahl p gibt es einen endlichen Körper $\mathbb{Z}_p$ mit p Elementen, wir haben ihn aus den ganzen Zahlen durch Restklassenbildung konstruiert. Diese Körper haben sich in ganz unterschiedlicher Weise als nützlich erwiesen, es stellt sich daher die Frage nach weiteren endlichen Körpern. Wir werden sehen, daß es einen Körper mit q Elementen genau dann gibt, falls q eine Primzahlpotenz ist: $q = p^n$. Man gewinnt ihn ebenfalls durch Restklassenbildung, nun nicht mehr aus den ganzen Zahlen, sondern aus geeigneten Polynomringen. Diese Körper haben vielfältige Anwendungen gefunden. Sie sind eindeutig bis auf Isomorphie, wir werden sie mit $\mathbb{F}_q$ bezeichnen.

9.1 Die Struktur endlicher Körper

Wir werden später alle endlichen Körper als Oberkörper der Körper $\mathbb{Z}_p$ konstruieren. Vorbereitend betrachten wir zunächst Unterkörper eines endlichen Körpers K. Den kleinsten Unterkörper in K, der $\alpha \in K$ enthält, bezeichnen wir mit $K(\alpha)$. Wir wollen diese Körper genauer beschreiben.

Sei zunächst $\alpha = 1$ das Einselement in K. Da 1 in jedem Unterkörper von K enthalten ist, ist $K(1)$ der kleinste Unterkörper, er heißt **Primkörper** von K. Er enthält alle Elemente

$$i \cdot 1 \;:=\; \underbrace{1 + \ldots + 1}_{i \text{ Summanden}}, \quad i \in \mathbb{N}.$$

Da K endlich ist, gibt es $i < j$, so daß $i \cdot 1 = j \cdot 1$. Für $p = j - i$ folgt $p \cdot 1 = 0$. Wir wählen $p > 0$ minimal mit dieser Eigenschaft. Dann sind die Körperelemente $i \cdot 1$ für $k = 1, 2, \ldots, p$ voneinander verschieden, und es liegt nahe, $\mathbb{Z}_p$ in K einzubetten, indem wir der Restklasse $i \bmod p$ das Körperelement $i \cdot 1$ zuordnen. Wegen $p \cdot 1 = 0$ ist diese Bijektion offenbar mit

Addition und Multiplikation verträglich. Da K keine Nullteiler enthält, ist auch $\mathbb{Z}_p$ nullteilerfrei. Folglich ist p eine Primzahl und $\mathbb{Z}_p$ ein Körper. $K(1)$ ist zu diesem Körper isomorph und besteht genau aus den Elementen $i \cdot 1$, $i = 1, \ldots, p$. p heißt die **Charakteristik** von K.

Wir können nun K (und jeden seiner Unterkörper) als Vektorraum über $\mathbb{Z}_p$ auffassen. Die Addition von Elementen in K ist uns vorgegeben und die Skalarmultiplikation ak mit $a \in \mathbb{Z}_p$ und $k \in K$ erhalten wir aus der Multiplikation in K und der Einbettung von $\mathbb{Z}_p$ in K. Da K endlich ist, ist auch seine Dimension n über $\mathbb{Z}_p$ endlich. Nach den Resultaten der Linearen Algebra läßt sich also jedes $k \in K$ eindeutig in der Gestalt

$$k \;=\; a_1\alpha_1 + a_2\alpha_2 + \ldots + a_n\alpha_n, \quad a_i \in \mathbb{Z}_p$$

darstellen, mit einer Vektorraumbasis $\alpha_1, \ldots, \alpha_n \in K$. Insbesondere können wir eine Aussage über die Anzahl der Elemente von K machen, die wir im folgenden mit q bezeichnen.

Proposition 9.1. *Sei K ein endlicher Körper der Charakteristik p. Dann ist die Anzahl der Elemente in K gleich $q = p^n$, mit einer natürliche Zahl n.*

Der Unterkörper $K(\alpha)$, $\alpha \in K$, enthält die Elemente $a_0 + a_1\alpha + \ldots + a_r\alpha^r$. Es liegt daher nahe, eine Vektorraumbasis aus Potenzen von α zu bilden.

Proposition 9.2. $1, \alpha, \alpha^2, \ldots, \alpha^{m-1}$ *ist eine Basis von $K(\alpha)$, wobei m die Dimension von $K(\alpha)$ über $\mathbb{Z}_p$ bezeichne.*

Beweis. Sei m die größte natürliche Zahl, so daß $1, \alpha, \ldots, \alpha^{m-1}$ linear unabhängig sind. Wir zeigen, daß dann die Menge K' aller Körperelemente der Gestalt

$$f(\alpha) \;=\; a_0 + a_1\alpha + \ldots + a_{m-1}\alpha^{m-1}, \quad a_i \in \mathbb{Z}_p$$

bereits ganz $K(\alpha)$ ist, mit dem Polynom $f(x) := a_0 + a_1x + \ldots + a_{m-1}x^{m-1}$. Offenbar ist K' unter Addition abgeschlossen und enthält 0 und 1. Nach Definition von m gibt es $b_0, b_1, \ldots, b_{m-1} \in \mathbb{Z}_p$, so daß

$$\alpha^m \;=\; -b_0 - b_1\alpha - \ldots - b_{m-1}\alpha^{m-1}.$$

Wenn wir daher zwei Elemente aus K' multiplizieren, können wir mit dieser Gleichung Potenzen α^r mit $r \geq m$ durch kleinere Potenzen ersetzen. Dies zeigt, daß K' unter Multiplikation abgeschlossen ist.

Um zu zeigen zeigen, daß jedes Element $f(\alpha)$ in K' ein Inverses besitzt, benutzen wir (ähnlich wie früher für $\mathbb{Z}_p$) den Satz von Bézout. Wir bilden das normierte Polynom $g(x) := b_0 + b_1x + \ldots + b_{m-1}x^{m-1} + x^m$. Es ist irreduzibel: Aus einer Zerlegung $g(x) = g_1(x)g_2(x)$ folgt nämlich wegen $g(\alpha) = 0$ die

Gleichung $g_1(\alpha) = 0$ (bzw. $g_2(\alpha) = 0$). Nach Wahl von m muß daher $g_1(x)$ vom Grade m und $g_2(x)$ vom Grade 0 sein (oder umgekehrt). $g(x)$ hat daher keine nicht-trivialen Teiler. Da $f(x)$ einen kleineren Grad als $g(x)$ hat, sind $f(x)$ und $g(x)$ teilerfremde Polynome. Daher gibt es Polynome $s(x)$ und $t(x) \in \mathbb{Z}_p$, so daß $1 = s(x)f(x) + t(x)g(x)$. Wegen $g(\alpha) = 0$ folgt $1 = s(\alpha)f(\alpha)$, $s(\alpha)$ ist also das inverse Element von $f(\alpha)$. Andererseits gehört $s(\alpha)$ zu K', denn Potenzen α^r mit $r \geq m$ können wir erneut durch kleinere Potenzen ersetzen.

Es gilt also $K' = K(\alpha)$. m ist dann offenbar die Dimension von $K(\alpha)$ über $\mathbb{Z}_p$. □

Das in dem Beweis gebildete normierte irreduzible Polynom $g(x) = x^m + b_{m-1}x^{m-1} + \ldots + b_1x + b_0 \in \mathbb{Z}_p[x]$ nennt man Minimalpolynom von α über $\mathbb{Z}_p$. Unsere Überlegungen sind von speziellen Wahl der Unterkörper unabhängig. Allgemeiner kann man einen jeden Körper K als Vektorraum über jeden Unterkörper $K_0 \subset K$ auffassen. Ist K endlich, so kann man für jedes $\alpha \in K$ ein normiertes Polynom $g(x)$ mit Koeffizienten in K_0 und von minimalem Grade (ungleich dem Nullpolynom) bilden, so daß $g(\alpha) = 0$ gilt. Wie wir gesehen haben, ist es irreduzibel. $g(x)$ heißt das **Minimalpolynom** von α über K_0. Für ein Polynom $h(x) \in K_0[x]$ gilt dann

$$h(\alpha) = 0 \quad \Leftrightarrow \quad g(x) \mid h(x).$$

Die Division mit Rest $h(x) = m(x)g(x)+r(x)$ zeigt nämlich, daß aus $h(\alpha) = 0$ die Gleichung $r(\alpha) = 0$ folgt, daher ist $r(x)$ das Nullpolynom, und $g(x)$ teilt $h(x)$. Insbesondere folgt

$$g(x) \mid x^q - x,$$

denn $q - 1$ ist die Ordnung (die Anzahl der Elemente) der Gruppe $K^* = K - \{0\}$, so daß nach (4.2) $\alpha^{q-1} = 1$ für alle $\alpha \in K^*$ und damit

$$\alpha^q \;=\; \alpha \tag{9.1}$$

für alle $\alpha \in K$ gilt. Nach Korollar 6.2 zerfällt $x^q - x$ in $K[x]$ vollständig in Linearfaktoren,

$$x^q - x \;=\; \prod_{a \in K}(x - a). \tag{9.2}$$

Es stellt sich heraus, daß wir uns bei der Untersuchung von $K(\alpha)$ mit keinen speziellen Unterkörpern von K befaßt haben, jeder Unterkörper von K ist von dieser Gestalt. Dies folgt aus dem nächsten Resultat.

Proposition 9.3. *Ist K ein endlicher Körper, so ist K^* bzgl. der Multiplikation eine zyklische Gruppe.*

Es gibt also ein $\alpha \in K^* := K - \{0\}$, so daß α^k für $k = 1, \ldots, q-1$ alle Elemente von K^* durchläuft. α wird **primitives Element** des Körpers genannt. Der Beweis beruht auf dem folgenden gruppentheoretischen Sachverhalt.

Lemma 9.4. *Sei G eine endliche Gruppe mit neutralem Element e, und sei m das Maximum der Ordnungen aller Gruppenelemente. Dann gilt $a^m = e$ für alle $a \in G$.*

Beweis. Sei $a \in G$ und sei r seine Ordnung, also die kleinste natürliche Zahl, so daß $a^r = e$. Zu zeigen ist, daß r ein Teiler von m ist. Wir führen einen Widerspruchsbeweis und nehmen an, daß $r = st$ gilt, mit $s > 1$ und $1 = \mathrm{ggT}(s, m)$. Dann hat $b := a^t$ offenbar die Ordnung s. Sei weiter $c \in G$ ein Element der Ordnung m und sei u die Ordnung von bc. Dann folgt $b^{um} = (bc)^{um} = e$ und folglich $s \mid um$ bzw. $s \mid u$, da m und s teilerfremd sind. Genauso folgt $c^{us} = (bc)^{us} = e$ und $m \mid u$. Insgesamt erhalten wir $u \geq sm$ im Widerspruch dazu, daß m die maximale Ordnung ist. $\square$

Beweis von Proposition 9.3. Für die maximale Ordnung m der Elemente aus K^* gilt nach dem letzten Lemma $a^m = 1$ für alle $a \in K^*$. Das Polynom $x^m - 1 \in K[x]$ hat daher $q-1$ verschiedene Nullstellen, wobei q die Anzahl der Elemente von K bezeichne. Nach Korollar 6.2 ist dies nur möglich, falls $m \geq q-1$ gilt. Andererseits ist die Ordnung eines Elementes höchstens $q-1$. Daher muß es Elemente der Ordnung $q-1$ geben. $\square$

Eine wichtige Konsequenz ist, daß zwei endliche Körper gleicher Mächtigkeit sich strukturell nicht unterscheiden.

Proposition 9.5. *Zwei endliche Körper K und K' gleicher Mächtigkeit sind isomorph.*

Beweis. Nach Proposition 9.1 haben K und K' dieselbe Charakteristik. Sei $g(x)$ das Minimalpolynom von $\alpha \in K$ über $\mathbb{Z}_p$. Es teilt $x^q - x$. Da $x^q - x$ über K' nach (9.2) vollständig in Linearfaktoren zerfällt, hat $g(x)$ auch in K' eine Nullstelle α', und ist wegen seiner Irreduzibilität das Minimalpolynom von α'. Wählen wir nun α als primitives Element von K, so hat $g(x)$ den Grad n, und wir erhalten eine Bijektion

$$a_0 + a_1\alpha + \ldots + a_{n-1}\alpha^{n-1} \;\mapsto\; a_0 + a_1\alpha' + \ldots + a_{n-1}(\alpha')^{n-1}, \quad a_i \in \mathbb{Z}_p,$$

von K nach K'. Diese Abbildung ist ein Isomorphismus, sie verträgt sich mit der Addition wie der Multiplikation, die in den beiden Körpern durch die sich entsprechenden Gleichungen $g(\alpha) = 0$ bzw. $g(\alpha') = 0$ gegeben ist. $\square$

9.2 Konstruktion von endlichen Körpern

Wir setzen die Überlegungen des letzten Abschnitts nun um in eine Konstruktion endlicher Körper als Oberkörper von $\mathbb{Z}_p$. Nullstellen von irreduziblen Polynomen stehen nicht mehr zur Verfügung, deswegen werden wir sie uns verschaffen müssen.

Allgemein sieht diese Konstruktion der Algebra so aus: Sei $g(x) = x^n + b_{n-1}x^{n-1} + \ldots + b_0$ ein normiertes Polynom vom Grade n, mit Koeffizienten in einem Körper K_0. Wir nehmen $g(x)$ als irreduzibel an, insbesondere besitzt es keine Nullstellen in K_0. Wir ‚erfinden' nun eine Nullstelle α von $g(x)$, die nicht zu K_0 gehören soll, und konstruieren einen Erweiterungskörper K von K_0 formal aus den Elementen

$$a_0 + a_1\alpha + a_2\alpha^2 + \ldots + a_{n-1}\alpha^{n-1}, \quad a_i \in K_0,$$

als Vektorraum über K_0 der Dimension n mit der Basis $1, \alpha, \ldots, \alpha^{n-1}$. Zusätzlich fordern wir die Gleichung

$$g(\alpha) \; = 0 \; , \quad \text{bzw. } \alpha^n \; = \; -b_{n-1}\alpha^{n-1} - \ldots - b_0.$$

Die Addition und Multiplikation zweier Körperelementen geschieht nach den üblichen Regeln. Bei der Multiplikation können dabei Potenzen α^r entstehen mit einem Exponenten $r \geq n$. Diese Potenzen werden mit Hilfe der Gleichung für α^n eliminiert, bis nur noch Potenzen von α mit Exponenten kleiner als n vorhanden sind. Es entsteht so, wir man leicht nachprüft, ein Ring. Aus der Irreduzibilität von $g(x)$ folgt wie im Beweis von Proposition 9.2, daß jedes von 0 verschiedene Element ein Inverses besitzt. K ist also ein Körper. Man spricht von der **Adjunktion** von α an den Körper K_0.

Beispiel. Die komplexen Zahlen $\mathbb{C}$ entstehen aus den reellen Zahlen $\mathbb{R}$ durch Adjunktion einer Nullstelle i des irreduziblen reellen Polynoms $x^2 + 1$. $\mathbb{C}$ besteht aus den Elementen $a + bi,\ a, b \in \mathbb{R}$. i erfüllt die für die imaginäre Zahl charakteristische Gleichung $i^2 = -1$. □

Diese Konstruktion läßt sich auch als Restklassenbildung in dem Polynomring $K_0[x]$ nach dem von $g(x)$ erzeugten Hauptideal I begreifen. Dann ist $\alpha = \overline{x}$, die von dem Monom x erzeugte Restklasse. Die Gleichung $g(\alpha) = 0$ wird nun nicht postuliert, sondern ergibt sich durch Restklassenbildung, denn die Aussagen $g(x) \equiv 0 \bmod I$ bzw. $\overline{g(x)} = 0$ und $g(\overline{x}) = 0$ sind äquivalent. Der konstruierte Körper ist $K_0[x]/I$. – Insgesamt erhalten wir die folgende Aussage.

Proposition 9.6. *Sei K_0 ein Körper und $g(x)$ ein irreduzibles Polynom mit Koeffizienten in K_0. Dann gibt es einen K_0 umfassenden Körper K, in dem $g(x)$ eine Nullstelle besitzt.*

Mit diesem Resultat konstruieren wir nun alle endlichen Körper. Dazu benutzen wir die folgende Aussage.

Proposition 9.7. *In einem Körper der Charakteristik p gilt für alle Elemente a, b*

$$(ab)^p = a^p b^p, \quad (a \pm b)^p = a^p \pm b^p.$$

Beweis. Die erste Aussage ist offenbar, die zweite ergibt sich aus der Binomialformel

$$(a \pm b)^p = \sum_{i=0}^{p} \binom{p}{i} \cdot 1 \; a^i (\pm b)^{p-i},$$

denn $\binom{p}{i}$ ist für $i \neq 0, p$ ein Vielfaches von p, und die entsprechenden Summanden verschwinden. □

Proposition 9.8. *Sei p eine Primzahl und n eine natürliche Zahl. Dann gibt es einen Körper K mit $q = p^n$ Elementen.*

Beweis. Wir konstruieren K als Körper, in dem das Polynoms $f(x) = x^q - x$ vollständig in Linearfaktoren zerfällt. Dazu zerlegen wir $f(x)$ in $\mathbb{Z}_p[x]$ in irreduzible Faktoren $g_1(x), \ldots, g_r(x)$. Nach der letzten Proposition gibt es einen Erweiterungskörper K_1, in dem $g_1(x)$, und damit $f(x)$ eine Nullstelle hat. Daher besitzt $f(x)$ in $K_1[x]$ einen Linearfaktor, möglicherweise mehrere, die wir von $f(x)$ abspalten. (Ist $g_1(x)$ bereits linear, so erübrigt sich dieser Schritt.) Den Rest von $f(x)$ zerlegen wir erneut in irreduzible Faktoren und konstruieren einen Körper $K_2 \supset K_1$ in dem $f(x)$ weitere lineare Faktoren abspaltet. Nach maximal q Erweiterungen zerfällt $f(x)$ schließlich in einem Körper K' vollständig in Linearfaktoren.

Wir definieren nun K als die Menge aller Nullstellen von $f(x)$, also derjenigen $a \in K'$ mit $a^q = a$. Wegen $q \cdot 1 = 0$ gilt $f'(x) = q \cdot 1 \; x^{q-1} - 1 = -1$, daher sind alle Nullstellen von $f(x)$ voneinander verschieden (vgl. Proposition 6.7), so daß K insgesamt q Elemente enthält. K enthält 0 und 1, und für $a, b \in K$ gilt

$$(ab)^q = a^q b^q = ab,$$

daher folgt $ab \in K$ und $a^{-1} \in K$. Weiter folgt aus der vorangehenden Proposition

$$(a \pm b)^{p^n} = ((a \pm b)^p)^{p^{n-1}} = (a^p \pm b^p)^{p^{n-1}} = \ldots = a^{p^n} \pm b^{p^n} = a \pm b.$$

Es gehört also auch $a \pm b$ zu K. Damit ist K ein Körper mit q Elementen. □

Der konstruierte Körper heißt **Zerfällungskörper** von $x^q - x$. Allgemeiner bezeichnet man als Zerfällungkörper eines Polynoms $f(x) \in K_0[x]$ jeden minimalen Körper $K \supset K_0$, so daß $f(x)$ in $K[x]$ vollständig in Linearfaktoren zerfällt. In der Algebra wird gezeigt, daß Zerfällungkörper immer existieren und bis auf Isomorphie eindeutig sind. – Wir fassen unsere Resultate in dem folgenden Satz zusammen.

Satz 9.9. *Es gibt genau dann einen endlichen Körper K mit q Elementen, wenn q Potenz einer Primzahl ist. K ist dann bis auf Isomorphie eindeutig bestimmt.*

Diesen Körper nennt man das **Galois-Feld** $\mathbb{F}_q$. Abschließend beweisen wir eine bemerkenswerte Konsequenz unserer Überlegungen.

Satz 9.10. *Sei p prim und $q = p^n$ für ein $n \in \mathbb{N}$. Dann ist das Polynom $x^q - x \in \mathbb{Z}_p[x]$ gleich dem Produkt aller normierten, irreduziblen Polynome aus $\mathbb{Z}_p[x]$, deren Grad n teilt.*

Beweis. $x^q - x$ hat keine quadratischen Anteile, da die Ableitung gleich -1 ist (vgl. Proposition 6.7).

Sei $g(x)$ ein irreduzibles Polynom vom Grad m, so daß $m \mid n$. Nach Proposition 9.8 besitzt $g(x)$ eine Nullstelle α in einem Körper aus p^m Elementen. Es folgt $\alpha^{p^m} = \alpha$ und, da n ein Vielfaches von m ist, $\alpha^{p^n} = (\ldots(\alpha^{p^m})^{p^m}\ldots)^{p^m} = \alpha$. α ist also auch Nullstelle von $x^q - x$. Da $g(x)$ das Minimalpolynom von α ist, ist $g(x)$ Teiler von $x^q - x$.

Sei umgekehrt $g(x)$ ein normierter, irreduzibler Teiler von $x^q - x$ vom Grade m. Dann hat $g(x)$ eine Nullstelle α in $\mathbb{F}_q$, und der Teilkörper $K(\alpha)$ hat p^m Elemente. Wir fassen nun $\mathbb{F}_q$ als Vektorraum über $K(\alpha)$ auf, mit der Basis $\beta_1, \ldots, \beta_d$. Dann hat jedes Element aus $\mathbb{F}_q$ eine eindeutige Darstellung

$$b_1\beta_1 + \ldots + b_d\beta_d, \quad b_i \in K(\alpha).$$

$\mathbb{F}_q$ hat also $(p^m)^d = p^{md}$ Elemente, und es folgt $n = md$. □

Beispiel. Sei n Primzahl. Dann hat $x^q - x$ nur irreduzible Teiler vom Grad 1 oder n. Sind $g_1(x), \ldots, g_r(x) \in \mathbb{Z}_p[x]$ die normierten, irreduziblen Teiler vom Grad n, so gilt

$$x^{p^n} - x \;=\; (x^p - x)g_1(x) \cdot \ldots \cdot g_r(x).$$

Es folgt $p^n = p + rn$. Die Anzahl der normierten, irreduziblen Polynome vom Grad p^n ist für primes p und primes n daher gegeben durch

$$r \;=\; \frac{p^n - p}{n}.$$

Für $p = 7, n = 5$ ist $r = 3360$. Man kann diese Formel auch auf nichtprimes n verallgemeinern (vgl. das Lehrbuch von Childs). □

9.3 Elliptische Kurven über endlichen Körpern

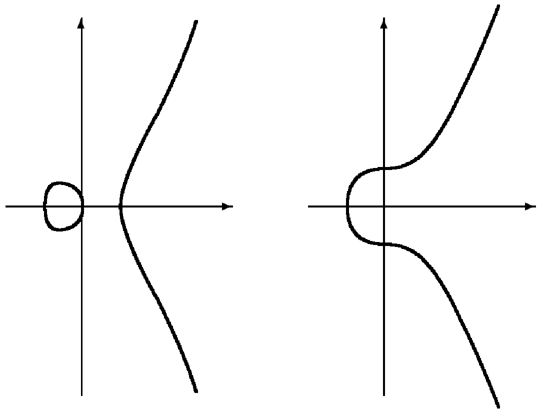
Die wesentliche Eigenschaft einer elliptischen Kurve ist, daß sie eine Gruppenstruktur trägt, was verschiedene interessante Anwendungen gefunden hat, z.B. in der Kryptographie. In diesem Abschnitt geben wir eine kurze Einführung in elliptische Kurven über endlichen Körpern.

Definition 9.11. *Sei K ein Körper mit einer Charakteristik $p \neq 2, 3$, und sei $x^3 + rx + s \in K[x]$ ein kubisches Polynom ohne mehrfache Nullstelle. Die zu diesem Polynom gehörige* ***elliptische Kurve über K*** *besteht aus allen Punkten $(x, y) \in K \times K$, die die Gleichung*

$$y^2 = x^3 + rx + s$$

erfüllen, zusammen mit einem weiteren Element O, dem ‚unendlich fernen Punkt'.

Die elliptischen Kurven $y^2 = x^3 - x$ und $y^2 = x^3 + 1$ über den reellen Zahlen sind in der folgenden Zeichnung skizziert.



Bemerkungen.

1. Die Bedingung, daß keine mehrfachen Nullstellen vorkommen, läßt sich in r und s ausdrücken. Sind $\alpha_1, \alpha_2, \alpha_3$ die Nullstellen, so folgt durch Koeffizientenvergleich aus $x^3 + rx + s = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3)$

$$\alpha_1 + \alpha_2 + \alpha_3 = 0, \quad \alpha_1\alpha_2 + \alpha_2\alpha_3 + \alpha_3\alpha_1 = r, \quad \alpha_1\alpha_2\alpha_3 = -s.$$

Eine Rechnung zeigt, daß sich die *Diskriminante* $D = \prod_{i<j}(\alpha_i - \alpha_j)^2$ aus diesen Ausdrücken zusammensetzen läßt (vgl. van der Waerden, Algebra 1, §33),

$$-D = 4r^3 + 27s^2.$$

$x^3 + rx + s$ hat also genau dann keine mehrfachen Nullstellen, falls $4r^3 + 27s^2 \neq 0$ gilt.

2. In Körpern der Charakteristik 2 oder 3 ist die Bedingung an eine elliptische Kurve passend abzuwandeln (vgl. das Lehrbuch von Koblitz). □

Die fundamentale Eigenschaft von elliptischen Kurven ist, daß man sie als abelschen Gruppe betrachten kann. Das Additionsgesetz läßt sich kurz zu der folgenden Aussage zusammenfassen: *Die 2 oder 3 Punkte einer elliptischen Kurve, die auf einer vorgegebenen Geraden ℓ liegen, summieren sich zu O auf, dem unendlich fernen Punkt*, dabei sind die Punkte in ihrer Vielfachheit zu berücksichtigen. Wir präzisieren dies nun. Unter einer Geraden ℓ verstehen wir alle Punkte (x, y), die einer Gleichung der Gestalt $ax + by + c = 0$ genügen mit $a \neq 0$ oder $b \neq 0$ $(a, b, c \in K)$. Es gibt verschiedene Fälle zu unterscheiden.

1. Sei $b = 0$. Dann können wir ohne Einschränkung $a = 1$ wählen, ℓ ist also durch die Gleichung $x + c = 0$ gegeben. Der Schnitt von ℓ mit der elliptischen Kurve enthält die Punkte (x, y) mit $y^2 = -c^3 - rc + s$. Wenn der Schnitt nicht leer ist, enthält er genau zwei Punkte $P = (x, y)$ und $Q = (x, -y)$. Wir setzen dann $P + Q = O$. Im Fall $y = 0$ fallen P und Q zusammen, dann lautet die Gleichung $P + P = O$.

2. Sei nun $b \neq 0$. Dann können wir ohne Einschränkung $b = -1$ setzen, also die Gerade $y = ax + c$ betrachten. Seien weiter $P = (x_1, y_1) \neq Q = (x_2, y_2)$ zwei Punkte der elliptischen Kurve, die auf der Geraden liegen. Es folgt $x_1 \neq x_2$ und
$$a = \frac{y_2 - y_1}{x_2 - x_1}, \quad c = y_1 - \frac{y_2 - y_1}{x_2 - x_1} x_1.$$
In diesem Fall gibt es auf der Geraden einen weiteren Punkt der elliptischen Kurve. Alle Schnittpunkte (x, y) erfüllen nämlich die Gleichung
$$f(x) := x^3 + rx + s - (ax + c)^2 = 0.$$
Dieses kubische Polynom hat die beiden Nullstellen x_1, x_2. Es zerfällt daher vollständig in Linearfaktoren und besitzt eine weitere Nullstelle x_3, und auf ℓ liegt genau ein weiterer Punkt $R = (x_3, -y_3)$ der elliptischen Kurve (für spätere Zwecke erhält y_3 ein Minuszeichen). Seine Koordinaten lassen sich aus der Gleichung $x_1 + x_2 + x_3 = a^2$ (dem Koeffizient von x^2 in dem Polynom $f(x)$) bestimmen:

$$
\begin{aligned}
x_3 &= \left(\frac{y_2 - y_1}{x_2 - x_1}\right)^2 - x_1 - x_2, \\
y_3 &= -ax_3 - c \\
&= \left(\frac{y_2 - y_1}{x_2 - x_1}\right)(x_1 - x_3) - y_1.
\end{aligned} \tag{9.3}
$$

Insgesamt befinden sich auf der Geraden drei Punkte P, Q, R der elliptischen Kurve, die sich zu O addieren: $P + Q + R = O$. Es ist möglich, daß sich $R = P$ oder $R = Q$ ergibt, dann wird der entsprechende Punkt in der Gleichung $R + P + Q = O$ zweimal aufgeführt.

3. Die soeben betrachtete Gerade ist durch die beiden Punkten $P \neq Q$ bestimmt. Jetzt behandeln wir den Fall einer Geraden $y = ax + c$, deren Lage durch einen Punkt $P = (x_1, y_1)$ gegeben ist, dem eine mehrfache Nullstelle x_1 des kubischen Polynoms $f(x)$ entspricht. Dann gilt zusätzlich zu $f(x_1) = 0$ die Gleichung $0 = f'(x_1) = 3x_1^2 + r - 2(ax_1 + c)a$, also
$$
a = \frac{3x_1^2 + r}{2y_1}, \quad c = y_1 - \frac{3x_1^2 + r}{2y_1}\, x_1.
$$
(Im Fall $K = \mathbb{R}$ bedeutet die Bedingung, daß die Gerade eine Tangente der elliptischen Kurve ist.) Wie oben hat nun $f(x)$ neben der doppelten Nullstelle x_1 eine weitere Nullstelle x_3. Wir können sie aus der Gleichung $2x_1 + x_3 = a^2$ bestimmen,
$$
\begin{aligned}
x_3 &= \left(\frac{3x_1^2 + r}{2y_1}\right)^2 - 2x_1, \\
y_3 &= \left(\frac{3x_1^2 + r}{2y_1}\right)(x_1 - x_3) - y_1.
\end{aligned} \tag{9.4}
$$
Die Gleichung für die Punkte $P = (x_{1,}y_1)$ und $R = (x_3, -y_3)$ der elliptischen Kurve lautet nun $P + P + R = O$, P wird also doppelt berücksichtigt.
□

Die Sprechweise vom unendlich fernen Punkt begründet sich aus der Vorstellung, daß O unendlich fern auf allen senkrechten (durch $b = 0$ gegebenen) Geraden liegt. Das hat den Vorteil, daß nun auf *allen* Geraden genau drei Punkte liegen (gezählt in der jeweiligen Vielfachheit). Dieser Vorstellung kann man einen mathematisch präzisen Sinn geben, das Stichwort ist ‚projektiver Abschluß' der Kurve (vgl. das Buch von Koblitz).

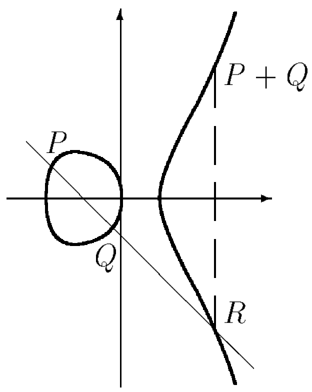
Wir wollen nun Punkte auf einer elliptischen Kurve so addieren, daß die angegebenen Gleichungen gültig bleiben, dabei soll O die Rolle des neutralen Elements übernehmen. Aus 1. ergibt sich der entgegengesetzte Punkt eines Punktes. Wir setzen

$$-P \;:=\; (x, -y), \quad \text{falls } P \;=\; (x, y).$$

$-P$ entsteht aus P durch Spiegelung an der x-Achse. Die Summe von P und Q erhalten wir durch Umformung der Gleichung $P + Q + R = O$ zu $P + Q = -R$. Wir setzen also

$$P + Q \;:=\; (x_3, y_3), \quad \text{falls } P = (x_1, y_1),\; Q = (x_2, y_2),$$

dabei sind x_3 und y_3 durch die Formeln (9.3) bzw. (9.4) gegeben (je nachdem ob $P \neq Q$ oder $P = Q$ gilt).



Beispiel. Die Punkte $P = (-3, 9)$ und $Q = (-2, 8)$ liegen auf der elliptischen Kurve $y^2 = x^3 - 36x$ über den rationalen Zahlen. Es gilt $P + Q = (6, 0)$ und $2 \cdot P = (25/4, -35/8)$. □

Daß man eine elliptische Kurve auf diese Weise zu einer abelschen Gruppe macht, ist nicht offensichtlich, insbesondere ist nicht evident, daß das Assoziativitätsgesetz $(P + Q) + R = P + (Q + R)$ gilt. Man kann es aus den angegebenen Formeln in etwas mühsamer Rechnung bestätigen. Für den Körper der reellen oder komplexen Zahlen gibt es andere Beweise. Sie benutzen Argumente aus der projektiven Geometrie oder der Funktionentheorie und gewähren einen tieferen Einblick in die Gruppenstruktur elliptischer Kurven. Wir fassen die Diskussion in einem Satz zusammen.

Satz 9.12. *Jede elliptische Kurve wird durch die in (9.3) bzw. (9.4) gegebene Addition eine Gruppe. Das neutrale Element ist der unendlich ferne Punkt O, und das inverse Element von (x, y) ist $(x, -y)$.*

Elliptische Kurven über den komplexen wie über den rationalen Zahlen sind ein klassischer Untersuchungsgegenstand der Mathematik. Wir betrachten im folgenden elliptische Kurven über den Galois-Feldern $\mathbb{F}_q$. Die Anzahl N der Punkte einer elliptischen Kurve über $\mathbb{F}_q$ ist offensichtlich höchstens $2q+1$. Neben O gibt es nämlich für jedes $x \in \mathbb{F}_q$ höchstens zwei Punkte (x, y) und $(x, -y)$ auf der elliptischen Kurve. Dazu muß $x^3 + rx + s$ ein quadriertes Element von $\mathbb{F}_q$ sein. Da in dem Fall, daß q ungerade ist, genau die Hälfte der Elemente von $\mathbb{F}_q^*$ Quadrate sind, wird man erwarten, daß eine elliptische Kurve über $\mathbb{F}_q$ etwa q Elemente enthält. Genauer gilt der **Satz von Hasse**:

$$|N - (q+1)| \ \leq \ 2\sqrt{q}.$$

Eine elliptische Kurve zusammen mit einem Punkt $P = (x, y)$ kann man sich leicht verschaffen. Man wähle zufällige Elemente x, y, r aus $\mathbb{F}_q$ und setze $s = y^2 - (x^3 + rx)$. Dann ist offenbar $P = (x, y)$ ein Punkt auf der Kurve, die durch die Gleichung $y^2 = x^3 + rx + s$ gegeben ist. Es besteht eine gute Chance, daß die Diskriminante $4r^3 + 27s^2 \neq 0$ ist, andernfalls treffe man eine neue Wahl.

Das Schlüsselsystem von Diffie-Hellman

Elliptischen Gruppen über endlichen Körpern sind in der Kryptographie auf Interesse gestoßen. Ein Grund ist, daß man mit ihrer Hilfe Abbildungen konstruieren kann, die leicht zu berechnen sind, die sich aber nur mit erheblichen Rechenaufwand umkehren lassen. Die Situation ist analog zum diskreten Logarithmus im Restklassenring $\mathbb{Z}_m$. In elliptischen Kurven benötigt man zur Berechnung des Punktes

$$i \cdot P \ := \ \underbrace{P + \ldots + P}_{i-\text{mal}}$$

größenordnungsmäßig $O(\log i)$ Rechenschritte. Wie beim Potenzieren modulo m schreibe man dazu i in Dualdarstellung, $i = \sum d_j 2^j$ mit $d_j = 0, 1$, berechne $P_j = 2^j \cdot P = 2 \cdot P_{j-1}$ und summiere alle P_j mit $d_j = 1$ auf. Dagegen kann man nur in speziellen Fällen schneller als in $O(i)$ Schritten zu einem vorgegebenen Q ein Punkt P bestimmen, so daß $Q = i \cdot P$ gilt. Dies macht man sich in der Kryptographie zunutze.

Öffentliche Kryptosysteme sind im Vergleich zu klassischen Chiffriersystemen relativ rechenaufwendig. Deswegen versucht man, beide Systeme zu kombinieren. Nachrichten werden mit einem schnellen klassischen Chiffrierystem ausgetauscht, bei dem man zum Kodieren und Dekodieren der Nachrichten einen Schlüssel braucht. Für den Austausch des Schlüssels benutzt man ein sicheres öffentliches Kryptosystem. Einen ersten detaillierten Vorschlag haben Diffie und Hellman gemacht.

Das System von Diffie und Hellman basiert in seiner ursprünglichen Form auf dem diskreten Logarithmus. Wir nehmen an, daß der Schlüssel Element eines großen endlichen Körper $\mathbb{F}_q$ ist. q ist öffentlich bekannt. Weiter einigt man sich auf ein Element g in $\mathbb{F}_q$, idealerweise ein primitives Element von $\mathbb{F}_q^*$. Auch g wird öffentlich bekanntgemacht. Schlüssel werden aus g gebildet: Ein Teilnehmer A wählt sich zufällig eine natürliche Zahl a zwischen 0 und $q-1$, die er geheim hält. Gleichzeitig veröffentlich er das Element g^a in $\mathbb{F}_q^*$. Ein weiterer Teilnehmer B macht dasselbe, er wählt seine Geheimzahl b und veröffentlicht g^b. Der Geheimschlüssel für den Nachrichtenaustausch zwischen A und B ist dann g^{ab}. Beide können ihn leicht berechnen. A etwa potenziert die öffentlich zugängliche Größe g^b mit seiner Geheimzahl a, dazu sind keinerlei Absprachen zwischen A und B nötig. Eine dritte Partei hat dagegen auf g^{ab} keinen Zugriff. Man kann versuchen, sich a und b aus g^a und g^b zu verschaffen, um dann g^{ab} zu berechnen. Im allgemeinen ist aber die Berechnung solcher ‚diskreter Logarithmen' bei ausreichend großer Wahl von q rechnerisch nicht zu bewältigen.

In neuerer Zeit hat man anstelle von $\mathbb{F}_q$ auch elliptische Kurven über $\mathbb{F}_q$ herangezogen. An die Stelle von g tritt ein Punkt P der Kurve, und g^a wird ersetzt durch $a \cdot P$. Man verspricht sich dadurch Vorteile in der Sicherheit des Verfahrens. Für binäre Kodes ist es praktisch, einen Körper $\mathbb{F}_{2^r}$ zugrundezulegen. Dagegen spricht, daß in diesen speziellen Körpern diskrete Logarithmen mit vergleichsweise geringerem Rechenaufwand bestimmt werden können, r müßte also sehr groß gewählt werden. Nach dem heutigen Kenntnisstand sind dagegen geeignete elliptische Kurven über $\mathbb{F}_{2^r}$ sicher. Ob das RSA-Schema oder aber das Schema von Diffie und Hellman eine größere Sicherheit bietet, muß die Zukunft entscheiden.

Literatur

N. Koblitz (1994): A Course in Number Theory and Cryptography

Kapitel 10

Fehlerkorrigierende Kodes

Wir setzen nun die Untersuchung von fehlerkorrigierenden Kodes fort.

10.1 Grundbegriffe

In diesem Abschnitt betrachten wir Kodes, die Wörter der Länge m in Wörter der Länge n verwandeln. Als Alphabeth, in dem die Nachrichten geschrieben werden, wählen wir ein Galois-Feld $\mathbb{F}_q$, es gibt also $q = p^n$ verschiedene Buchstaben. Die unchiffrierte Nachricht $w = (w_0 w_1 \dots w_{m-1})$ und seine Kodierung $k = (k_0 k_1 \dots k_{n-1})$ können wir dann als Vektoren aus $\mathbb{F}_q^m$ bzw. $\mathbb{F}_q^n$ auffassen.

Ein $\boldsymbol{(n,m)}$**-Blockkode** ist eine injektive Abbildung

$$\kappa : \mathbb{F}_q^m \to \mathbb{F}_q^n .$$

Wir werden häufig den Kode mit der Menge $K = \kappa(\mathbb{F}_q^m)$ seiner Kodewörter identifizieren (und damit das Kodieren der Information beiseite lassen). Ein Kode heißt **linear**, falls K linearer Teilraum von $\mathbb{F}_q^n$ ist. m ist dann die Dimension von K.

Lineare Kodes lassen sich mit Matrizen beschreiben. Eine $m \times n$ Matrix G heißt **erzeugende Matrix** von K, falls ihre m Zeilen eine Vektorraumbasis von K bilden. Dann kann man (wie die lineare Algebra lehrt) mit der linearen Abbildung

$$w = (w_0 w_1 \dots w_{m-1}) \ \mapsto \ \kappa(w) := w \cdot G$$

kodieren.

Aus Sicht des Empfängers ist es günstig, K mit einer $(n-m) \times n$ Matrix H zu beschreiben, so daß

$$k \in K \ \Leftrightarrow \ H \cdot k^t = 0$$

gilt. Wie beim Hamming-Kode kann er dann nachprüfen, ob es Übertragungsfehler gibt. H heißt **Kontrollmatrix** von K.

Wichtig für die Güte eines Kodes ist der gegenseitige Abstand seiner Kodewörter. Seien $k = (k_0 k_1 \dots k_{n-1})$, $k' = (k'_0 k_1 \dots k'_{n-1})$ zwei Vektoren der Länge n. Dann wird der **Hamming-Abstand** von k und k' definiert als

$$d(k, k') \; := \; \text{Anzahl der } i \text{ mit } k_i \neq k'_i.$$

Als **Gewicht** von k bezeichnet man

$$g(k) \; := \; d(k, 0),$$

die Anzahl der Buchstaben $k_i \neq 0$ in k. Der **Minimalabstand** des Kodes K ist definiert als

$$d_{\min}(K) \; := \; \min\{d(k, k') \; : \; k, k' \in K, k \neq k'\}.$$

Für lineare Kodes gilt wegen $d(k, k') = g(k - k')$ offenbar

$$d_{\min}(K) \; = \; g_{\min}(K) \; := \; \min\{g(k) \; : \; k \in K, k \neq 0\}.$$

Kommt es bei der Übertragung der Nachricht zu Fehlern, so gehört das empfangene Wort $\widetilde{k}$ möglicherweise nicht mehr zu K. Man dekodiert dann üblicherweise nach dem **Maximum-Likelihood-Prinzip**: Man entscheidet sich für ein Kodewort k , bei dem es zu dem Übertragungsfehler $f = \widetilde{k} - k$ mit möglichst großer Wahrscheinlichkeit kommt. Typischerweise läuft das darauf hinaus, dasjenige Kodewort zu wählen, daß von dem empfangenen Wort minimalen Hamming-Abstand hat.

Ein **Dekodierschema** für einen Kode K ist eine Familie

$$D = (D_k)_{k \in K}$$

mit den Eigenschaften

$$\begin{array}{ll} D_k \subset \mathbb{F}_q^n & \text{für alle } k, \\ k \in D_k & \text{für alle } k, \\ D_k \cap D_{k'} = \emptyset & \text{für alle } k \neq k'. \end{array}$$

Liegt ein empfangenes Wort in D_k, so wird es als die Nachricht k dekodiert. Man sagt, D **entdeckt den Übertragungsfehler** $f \in \mathbb{F}_q^n$, $f \neq 0$, falls für alle $k' \in K$

$$k' + f \notin \bigcup_k D_k,$$

und D **korrigiert den Fehler** $f \neq 0$, falls für alle $k \in K$

$$k + f \in D_k.$$

Proposition 10.1. *Sei $K \subset \mathbb{K}_q^n$ ein Kode und $s \in \mathbb{N}$.*

i) Genau dann existiert ein Dekodierungsschema, das alle Fehler vom Gewicht höchstens s entdeckt, wenn $s < d_{\min}(K)$ ist.

ii) Genau dann existiert ein Dekodierungsschema, das alle Fehler vom Gewicht höchstens s korrigiert, wenn $2s < d_{\min}(K)$ gilt.

Beweis. i) Einen Fehler $f \neq 0$, für den zwei Kodeworte k, k' mit $k + f = k'$ existieren, bleibt beim Empfang von k' unerkannt. Nur diejenigen Fehler f, deren Gewicht $g(f) < d_{\min}(K)$ ist, können sicher entdeckt werden. Das Dekodierschema $D_k = \{k\}$ leistet dies.

ii) Ein Dekodierschema (D_k), daß alle Fehler vom Gewicht höchstens s korrigiert, hat die Eigenschaft, daß für alle $k \in K$ die ,Kugeln'

$$B_s(k) \;:=\; \{v \in \mathbb{F}_q^n \;:\; d(v,k) \leq s\}$$

in D_k enthalten sind. Es folgt $B_s(k) \cap B_s(k') = \emptyset$, daher muß $d(k,k') > 2s$ für beliebige Kodeworte $k \neq k'$ gelten. Umgekehrt ist unter der angegebene Bedingung $D_k := B_s(k)$ ein Dekodierschema, das s Fehler korrigiert. $\square$

Beispiel: Hadamard-Kodes. Sei n geradzahlig. Eine $n \times n$ Matrix $M = (m_{ij})$ mit den Einträgen 1 oder -1 heißt **Hadamard-Matrix**, falls gilt

$$M \cdot M^t = n \cdot Id.$$

Geometrisch besagt diese Bedingung, daß zwei Zeilen $v \neq v'$ von M zueinander orthogonal sind. Für Vektoren aus 1en und -1en bedeutet dies, daß die Zahl der übereinstimmenden und der unterschiedlichen Komponenten in v und v' gleich sind, nämlich $n/2$. Die $2n$ Zeilen der Matrix, die aus

$$\begin{pmatrix} M \\ -M \end{pmatrix}$$

durch Ersetzen jeder -1 durch eine 0 entsteht, haben daher in $\mathbb{F}_2^n$ alle mindestens einen Hamming-Abstand $n/2$ voneinander. Sie bilden einen Kode aus $2n$ Kodewörtern der Länge n, der bis zu $[\frac{n-2}{4}]$ Fehler korrigieren kann.

Geeignete Matrizen erhält man durch folgende Beobachtung (Übung): Ist M' eine weitere Hadamard-Matrix, so gilt dies auch für das ,Kronecker-Produkt'

$$M \otimes M' \;:=\; \begin{pmatrix} m_{11}M' & \cdots & m_{1n}M' \\ \vdots & & \vdots \\ m_{n1}M' & \cdots & m_{nn}M' \end{pmatrix}.$$

Ausgehend von

$$M_2 := \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

kann man so Hadamard Matrizen mit $n = 2^r$ Zeilen konstruieren. – Die Mariner-Sonde benutzte 1962 für die Bildübertragung vom Mars einen Hadamard-Kode. Er bestand aus 64 Kodewörtern der Länge 32 und konnte pro Wort bis zu 7 Fehlern korrigieren. □

Kodes, bei denen die Kodewörter völlig gleichmäßig in $\mathbb{F}_q^n$ verteilt sind, nennt man perfekt.

Definition 10.2. *Ein Kode $K \subset \mathbb{F}_q^n$ heißt* ***perfekt****, falls es eine natürliche Zahl s gibt, so daß für die Kugeln $B_s(k) = \{v \in \mathbb{F}_q^n : d(v,k) \leq s\}$, $k \in K$ gilt:*

$$B_s(k) \cap B_s(k') = \emptyset \text{ für alle } k \neq k', \text{ und } \bigcup_{k \in K} B_s(k) = \mathbb{F}_q^n.$$

Alle (n,m)-Hamming-Kodes sind perfekt: Jede Kugel vom Radius 1 um ein Kodewort enthält genau $n + 1 = 2^l$ Worte der Länge n. Diese Kugeln sind disjunkt, denn die Hamming-Kodes korrigieren einen Fehler. Die Dimension des Kodes ist $m = n - l$; dies bedeutet, daß es 2^{n-l} verschiedene Kodeworte gibt. Wegen

$$2^l \cdot 2^{n-l} = 2^n$$

schöpfen die Kugeln vom Radius 1 um die Kodeworte bereits den gesamten $\mathbb{F}_2^n$ aus.

Sonst treten perfekte Kodes nur sporadisch auf. Zum Beispiel gibt es nur einen nicht-trivialen perfekten binären Kode, der 3 Fehler korrigiert. Es handelt sich um den bekannten **Golay-Kode**, einen (23,12)-Blockkode.

Literatur

J.H. van Lint (1992): Introduction to Coding Theory, Springer

10.2 Zyklische Kodes

Ein zyklischer Kode ist ein Kode, bei dem Kodewörter bei zyklischer Vertauschung der Buchstaben in Kodewörter übergehen.

Definition 10.3. *Ein linearer Kode $K \subset \mathbb{F}_q^n$ heißt* ***zyklisch****, falls gilt*

$$(k_0 k_1 \dots k_{n-1}) \in K \;\Rightarrow\; (k_{n-1} k_0 \dots k_{n-2}) \in K.$$

Die Struktur zyklischer Kodes läßt sich übersichtlich mit Polynomen beschreiben. Wir ordnen dazu jedem Kodewort $k = (k_0 k_1 \dots k_{n-1})$ das Polynom

$$k(x) = k_0 + k_1 x + \dots + k_{n-1} x^{n-1}$$

zu. Diese Polynome vom Grade kleiner als n nennen wir die zu K gehörigen Kodepolynome. Wir können nun die zyklische Vertauschung von Kodewörtern als Multiplikation mit x modulo $x^n - 1$ ausdrücken,

$$\begin{aligned} & k_{n-1} + k_0 x + \dots + k_{n-2} x^{n-1} \\ & \quad \equiv \; x(k_0 + k_1 x + \dots + k_{n-1} x^{n-1}) \; = \; x k(x) \bmod (x^n - 1). \end{aligned}$$

Genauso entspricht der r-fache zyklische Vertauschung von Buchstaben die Multiplikation der Kodepolynome mit dem Monom x^r modulo $x^n - 1$. Da sich jedes Polynom linear aus Monomen zusammensetzt, erhalten wir die folgende Eigenschaft zyklischer Kodes: Gilt für $k'(x) = k'_0 + k'_1 x + \dots + k'_{n-1} x^{n-1} \in \mathbb{F}_q[x]$

$$k'(x) \;\equiv\; k(x) w(x) \bmod (x^n - 1) \tag{10.1}$$

mit einem Kodepolynom $k(x)$ und ein beliebiges Polynom $w(x) \in \mathbb{F}_q[x]$, so ist auch $k'(x)$ Kodepolynom.

Sei nun $g(x)$ ein Kodepolynom von K von minimalem Grad (ungleich dem Nullpolynom). Dividieren wir das Kodepolynom $k(x)$ durch $g(x)$ mit Rest

$$r(x) \;=\; k(x) - w(x) g(x),$$

so folgt nach (10.1), daß auch $r(x)$ Kodepolynom ist. Nach Wahl von $g(x)$ ist $r(x)$ das Nullpolynom. Die Kodepolynome sind also gerade die Vielfachen von $g(x)$, die einen Grad kleiner als n haben. Insbesondere ist $g(x)$ bis auf Einheiten eindeutig bestimmt. Vereinbaren wir, daß $g(x)$ ein normiertes Polynom ist, so ist $g(x)$ eindeutig festgelegt. $g(x)$ heißt das **Basispolynom** oder **erzeugende Polynom** des Kodes K.

Dividieren wir $x^n - 1$ durch $g(x)$ mit Rest

$$q(x) \;=\; x^n - 1 - m(x) g(x),$$

so folgt nach (10.1), daß $q(x)$ ebenfalls ein Kodepolynom ist. Nach Wahl von $g(x)$ ist $q(x)$ das Nullpolynom, das Basispolynom ist also ein Teiler von $x^n - 1$.

Umgekehrt induziert jeder normierte Teiler $g(x)$ von $x^n - 1$ einen zyklischen Kode. Seine Kodepolynome sind alle Vielfachen von $g(x)$ vom Grade kleiner n. Für die Polynome $k(x) = k_0 + k_1 x + \dots + k_{n-1} x^{n-1}$ und

$k'(x) = k_{n-1} + k_0 x + \ldots + k_{n-2} x^{n-1}$ mit zyklisch vertauschten Koeffizienten gilt

$$k'(x) = xk(x) - k_{n-1}(x^n - 1),$$

mit $k(x)$ ist daher auch $k'(x)$ Kodepolynom, und der Kode ist zyklisch. Ist l der Grad von $g(x)$, so wird der Vektorraum der Kodepolynome durch die linear unabhängigen Polynome $g(x)$, $xg(x)$,...,$x^{m-1}g(x)$ aufgespannt, mit $m = n - l$. Insgesamt erhalten wir das folgende Resultat.

Satz 10.4. *Die zyklischen Kodes $K \subset \mathbb{F}_q^n$ entsprechen eineindeutig den normierten Teilern $g(x)$ des Polynoms $x^n - 1$ in $\mathbb{F}_q[x]$. $k = (k_0 k_1 \ldots k_{n-1})$ ist genau dann ein Kodewort in K, wenn $k(x) = k_0 + \ldots + k_{n-1} x^{n-1}$ ein Vielfaches von $g(x)$ ist. Unter diesen Polynomen ist $g(x)$ ausgezeichnet als normiertes Polynom minimalen Grades. Ist l der Grad von $g(x)$, so handelt es sich bei K um einen (n,m)-Blockkode, mit $m = n - l$.*

Die Kodewörter entstehen aus $g(x)$ durch Multiplikation mit Polynomen $w(x)$. Wir drücken dies nun in Matrixschreibweise aus. Dazu leiten wir aus dem erzeugenden Polynom

$$g(x) = g_0 + \ldots + g_l x^l \qquad (l < n)$$

die **erzeugende Matrix**

$$G = \begin{pmatrix} g_0 & g_1 & \cdots & g_l & & & \\ & g_0 & g_1 & \cdots & g_l & & \\ & & \ddots & & & \ddots & \\ & & & g_0 & g_1 & \cdots & g_l \end{pmatrix}$$

mit $m = n - l$ Zeilen und n Spalten ab. Sie enthält in jeder Zeile die Koeffizienten von $g(x)$, die restlichen Eintragungen sind 0. Es ist unmittelbar einsichtig, daß der Polynommultiplikation $k(x) = w(x)g(x)$ die Matrixmultiplikation

$$w = (w_0 w_1 \ldots w_{m-1}) \;\mapsto\; k = (k_0 k_1 \ldots k_{n-1}) := w \cdot G$$

entspricht, wobei $w_0, w_1, \ldots, w_{m-1}$ die Koeffizienten des Polynoms $w(x)$ seien. Dies ist eine Kodiervorschrift, die ein Wort w in das Kodewort k transformiert. Die Zeilen von G sind die Kodierungen der Worte $(10\ldots0), (01\ldots0)$, $\ldots, (00\ldots1)$. Da die Matrix G linear unabhängige Zeilen hat, ist die Abbildung injektiv, es handelt sich um einen (n,m)-Blockkode.

Eine duale Beschreibung des Kodes K, vorteilhaft für das Dekodieren, gelingt mit dem normierten Polynom $h(x)$ vom Grad $\deg(h) = m = n - l$, gegeben durch die Gleichung

$$g(x)h(x) = x^n - 1.$$

Aus $k(x) = w(x)g(x)$ folgt $k(x)h(x) = w(x)(x^n - 1)$ und umgekehrt. $k(x)$ ist also genau dann Vielfaches von $g(x)$, wenn $k(x)h(x)$ Vielfaches von $x^n - 1$ ist. Dies ergibt eine alternative Charakterisierung von Kodewörtern.

Proposition 10.5. *Für jeden zyklischen Kode K gibt es ein eindeutig bestimmtes normiertes Polynom $h(x)$, einen Teiler von $x^n - 1$, so daß gilt:*

$$k = (k_0 k_1 \ldots k_{n-1}) \in K \;\Leftrightarrow\; k(x)h(x) \equiv 0 \bmod (x^n - 1).$$

$h(x)$ heißt das **Kontrollpolynom** des Kodes. Auch die Kontrollbedingung $k(x)h(x) \equiv 0 \bmod (x^n - 1)$ läßt sich in Matrixschreibweise umsetzen. Sie bedeutet, daß sich im Polynom $k(x)h(x)$ die Koeffizienten von x^i und x^{n+i} in der Summe aufheben,

$$(k_0 h_i + \ldots + k_i h_0) + (k_{i+1} h_{n-1} + \ldots + k_{n-1} h_{i+1}) \;=\; 0$$

für $i < n$, dabei haben wir $h_{m+1} = \ldots = h_{n-1} = 0$ gesetzt. Bilden wir die $(n - m) \times n$-Matrix

$$H \;=\; \begin{pmatrix} h_m & h_{m-1} & \ldots & h_0 & & & & \\ & h_m & h_{m-1} & \ldots & h_0 & & & \\ & & \ddots & & & & \ddots & \\ & & & & h_m & h_{m-1} & \ldots & h_0 \end{pmatrix},$$

die zeilenweise neben den Koeffizienten von $h(x)$ nur aus Nullen besteht, so folgt aus den angegebenen Kontrollgleichungen

$$H \cdot k^t \;=\; 0 \quad \text{für alle } k \in K.$$

Da H den Rang $n - m$ hat, folgt nach der bekannten Dimensionsformel der linearen Algebra, daß die Lösungen des homogenen Gleichungssystems $H \cdot y^t = 0$ einen Raum der Dimension m bilden. Da K von der Dimension m ist, gibt es außerhalb K keine weiteren Vektoren k, für die $H \cdot k^t = 0$ gilt. H ist also Kontrollmatrix des Kodes K.

Beispiel. In $\mathbb{Z}_2[x]$ gilt (vgl. Abschnitt 6.2)

$$x^7 - 1 = (x - 1)(x^3 + x + 1)(x^3 + x^2 + 1).$$

Wählen wir $g(x) = x - 1$ und $h(x) = x^6 + x^5 + x^4 + x^3 + x^2 + x + 1$, so gilt $m = 6$ und H besteht aus der einzigen Zeile (1111111). Es ergibt sich der einfache Parity-Check-Kode $k_0 + \ldots + k_6 = 0 \bmod 2$.

Wählen wir $g(x) = x^3 + x + 1, h(x) = x^4 + x^2 + x + 1$, so erhalten wir als Kontrollmatrix

$$H = \begin{pmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

Dies ist bis auf die Reihenfolge der Spalten die Kontrollmatrix des (7,4)-Hamming-Kode. Man sagt, die beiden Kodes sind äquivalent. □

10.3 BCH-Kodes

Wir betrachten nun eine weitere Möglichkeit zur Konstruktion einer Kontrollmatrix für einen zyklischen Kode K. Sie ist theoretisch wichtig, für das praktische Dekodieren ist sie weniger geeignet. Sei $g(x) = g_1(x)g_2(x)\dots g_k(x)$ das Basispolynom von K, zerlegt in seine irreduziblen Teiler $g_1(x), \dots, g_k(x) \in \mathbb{F}_q[x]$. Wir können dann $g_1(x), \dots, g_k(x)$ als Minimalpolynome von Elementen $\alpha_1, \dots, \alpha_l$ in einem Erweiterungskörper von $\mathbb{F}_q$ auffassen. Wir setzen nun voraus, daß $g_1(x), \dots, g_k(x)$ alle voneinander verschieden sind. Nach den Eigenschaften von Minimalpolynomen ist dann $k(x) = k_0 + k_1 x + \dots + k_{n-1}x^{n-1}$ genau dann ein Vielfaches von $g(x)$, falls

$$k(\alpha_1) = k(\alpha_2) = \dots = k(\alpha_l) = 0$$

gilt. In Matrizenschreibweise lautet diese Bedingung

$$H \cdot k^t = 0,$$

mit $k = (k_0 k_1 \dots k_{n-1})$ und der Matrix

$$H := \begin{pmatrix} 1 & \alpha_1 & \alpha_1^2 & \cdots & \alpha_1^{n-1} \\ 1 & \alpha_2 & \alpha_2^2 & \cdots & \alpha_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \alpha_l & \alpha_l^2 & \cdots & \alpha_l^{n-1} \end{pmatrix}.$$

Nach Satz 10.4 ist H Kontrollmatrix des Kodes.

Bose, Chaudhuri und Hocquenghem haben nun vorgeschlagen, solche zyklischen Kodes zu betrachten, für die $\alpha_1 = \alpha, \alpha_2 = \alpha^2, \dots, \alpha_l = \alpha^l$ für ein geeignetes α gilt.

Definition 10.6. *Ein zyklischer Kode $K \subset \mathbb{F}_q^n$ mit erzeugendem Polynom $g(x)$ heißt **BCH-Kode**, falls ein α in einem Erweiterungskörper von $\mathbb{F}_q$ und eine natürliche Zahl $l < n$ existieren, so daß gilt:*

i) Die Ordnung von α ist n: $\alpha^n = 1$ und $\alpha^j \neq 1$ für $j < n$.

ii) $g(x)$ ist das Produkt der Minimalpolynome von $\alpha, \alpha^2, \ldots, \alpha^l$, wobei jedes Minimalpolynom in $g(x)$ als Faktor nur einmal erscheinen soll.

Für einen BCH-Kode ist also

$$H \ := \ \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n-1} \\ 1 & \alpha^2 & \alpha^4 & \cdots & \alpha^{2(n-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \alpha^l & \alpha^{2l} & \cdots & \alpha^{(n-1)l} \end{pmatrix}.$$

Kontrollmatrix. Dies erlaubt den Beweis des folgenden Satzes.

Satz 10.7. *Der Minimalabstand eines BCH-Kodes ist mindestens $l+1$.*

Für den Beweis benötigen wir die **Vandermonde-Determinante**.

Proposition 10.8. *Für Elemente $\beta_1, \ldots, \beta_l$ in einem Körper gilt*

$$\det \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_l \\ \beta_1^2 & \beta_2^2 & \cdots & \beta_l^2 \\ \vdots & \vdots & & \vdots \\ \beta_1^l & \beta_2^l & \cdots & \beta_l^l \end{pmatrix} \ = \ \prod_i \beta_i \cdot \prod_{i<j} (\beta_j - \beta_i).$$

Beweis. Sei A_{ij} die $(l-1) \times (l-1)$-Matrix, die aus einer $l \times l$ Matrix $A = (a_{ij})$ durch Streichen der i-ten Zeile und der j-ten Spalte entsteht. Dann gilt bekanntlich der Laplacesche Entwicklungssatz

$$\det(A) \ = \ \sum_{i=1}^{l} (-1)^{i+l} a_{il} \det(A_{il}).$$

Wir betrachten das Polynom

$$f(x) \ := \ \sum_{i=1}^{l} (-1)^{i+l} \det(A_{il})\, x^i \ = \ \det \begin{pmatrix} a_{11} & \cdots & a_{1,l-1} & x \\ a_{21} & \cdots & a_{2,l-1} & x^2 \\ \vdots & & \vdots & \vdots \\ a_{l1} & \cdots & a_{l,l-1} & x^l \end{pmatrix}.$$

Es hat im Fall der Vandermonde-Determinante ($a_{ij} = \beta_i^j$) die Nullstellen $0, \beta_1, \ldots, \beta_{l-1}$, daher gilt

$$f(x) \ = \ \det(A_{ll}) x (x - \beta_1) \cdots (x - \beta_{l-1}).$$

Es folgt

$$\det\begin{pmatrix} \beta_1 & \cdots & \beta_l \\ \vdots & & \vdots \\ \beta_1^l & \cdots & \beta_l^l \end{pmatrix} \;=\; f(\beta_l) \;=\; \det(A_{ll})\beta_l(\beta_l-\beta_1)\dots(\beta_l-\beta_{l-1}).$$

Die Behauptung folgt nun per Induktion über l, denn $\det(A_{ll})$ ist wieder eine Vandermonde-Determinante. □

Beweis von Satz 10.7. Nach Voraussetzung des Satzes sind die Elemente der ersten Zeile von H alle voneinander verschieden, damit sind je l Spalten von H linear unabhängig (Vandermonde-Determinante), und man kann keine l Spalten in nicht-trivialer Weise linear zu 0 kombinieren. Gilt daher $H \cdot k^t = 0$ für ein $k \neq 0$, so muß k mindestens das Gewicht $l+1$ haben. Da K ein linearer Kode ist, folgt die Behauptung aus den Bemerkungen in Abschnitt 10.1. □

Beispiel. Ein Kode, der zwei Fehler korrigiert: Wir wollen einen binären BCH-Kode konstruieren. Das Polynom

$$m(x) \;:=\; x^4+x+1$$

ist in $\mathbb{Z}_2[x]$ irreduzibel. (Es hat in $\mathbb{Z}_2$ keine Nullstellen und könnte höchstens in zwei Polynome vom Grad 2 zerfallen. Da $m'(x)=1$, müßten die beiden Faktoren nach Prop. 6.7 verschieden sein. Es gibt aber in $\mathbb{Z}_2[x]$ nur ein irreduzibles Polynom vom Grad 2, nämlich x^2+x+1.) Nach Satz 9.10 teilt $m(x)$ das Polynom $x^{16}-x$, daher besitzt $m(x)$ in $\mathbb{F}_{16}$ eine Nullstelle α. Nach Proposition 9.2 läßt sich jedes Element aus $\mathbb{F}_{16}$ in eindeutiger Weise als Linearkombination von $1, \alpha, \alpha^2$ und α^3 mit Koeffizienten 0 oder 1 darstellen. Es ist $\alpha^{15}=1$, jedoch $\alpha^3 \neq 1$ und $\alpha^5 = \alpha^2+\alpha \neq 0$. Daher ist α primitives Element in $\mathbb{K}_{16}$. Als irreduzibles Polynom ist $m(x)$ das Minimalpolynom von α. Gleichzeitig ist $m(x)$ auch das Minimalpolynom von α^2 und α^4, denn es gilt (vgl. Proposition 6.6) $m(\alpha^2) = m(\alpha)^2 = 0$ und $m(\alpha^4) = m(\alpha^2)^2 = 0$. Wir bestimmen nun das Minimalpolynom von α^3 durch den Ansatz

$$0 \;=\; a + b\alpha^3 + c\alpha^6 + d\alpha^9 + e\alpha^{12}$$

mit $a, \dots, e \in \{0,1\}$ (da $\alpha^{15}=1$, ist es höchstens vom Grad 4). Aus $\alpha^4 = \alpha+1$ ergibt sich

$$\begin{aligned} \alpha^6 &= \alpha^4\alpha^2 \;=\; \alpha^3+\alpha^2, \\ \alpha^9 &= (\alpha^4)^2\alpha \;=\; (\alpha+1)^2\alpha \;=\; \alpha^3+\alpha, \\ \alpha^{12} &= ((\alpha)^4)^3 \;=\; (\alpha+1)^3 \;=\; \alpha^3+\alpha^2+\alpha+1. \end{aligned}$$

Es folgt $a+e=0, d+e=0, c+e=0, b+c+d+e=0$, d.h. $a=b=c=d=e$. Es gibt also genau ein Polynom $n(x) \neq 0$ vom Grad höchstens 4, so daß $n(\alpha^3)=0$, nämlich

$$n(x) \; := \; x^4+x^3+x^2+x+1,$$

das Minimalpolynom von α^3.

$$g(x) \; := \; m(x)n(x) \; = \; x^8+x^7+x^6+x^4+1$$

ist damit das erzeugende Polynom eine BCH-Kodes mit $d=5$, der Kode kann folglich 2 Fehler korrigieren. Die Ordnung von α ist 15, und die Dimension von K ist $m=n-\deg(g)=7$. Nach Definition eines BCH-Kodes liegt ein (15,7)-Kode vor, der Kode sendet Wörter der Länge 15, die 7 Informationsstellen haben. k ist genau dann Kodewort, wenn für das zugehörige Polynom $k(x)$

$$k(\alpha) \; = \; k(\alpha^3) \; = \; 0$$

gilt. Genau dann wird $k(x)$ von den Minimalpolynomen $m(x)$ und $n(x)$ geteilt, und damit von $g(x)$.

Kodieren. Für das Kodieren von Wörtern der Länge 7 bestehen verschiedene Möglichkeiten. Dies kann mit der oben angegebenen erzeugenden Matrix geschehen. Übersichtlich ist die folgende Methode. Dem Wort $w=(w_0,\ldots,w_6)$ ordnen wir das Polynom

$$k_1(x) \; = \; w_0x^8+w_1x^9+\ldots+w_6x^{14}$$

zu. Wir ergänzen es mit einem Polynom $k_2(x)$ vom Grad höchstens 7 zu

$$k(x) \; = \; k_1(x)+k_2(x),$$

so daß $k(x)$ ein Vielfaches von $g(x)$ wird. Dann gilt für ein Polynom $q(x)$

$$k_1(x) \; = \; q(x)g(x)-k_2(x) \; = \; q(x)g(x)+k_2(x).$$

$k_2(x)$ ist also eindeutig bestimmt als der Rest, der bei Division von $k_1(x)$ durch $g(x)$ übrig bleibt.

Dekodieren. Sei $f=(f_0,\ldots,f_{14})$ der Vektor der Fehler, es wird anstelle von k also die Nachricht $\widetilde{k}=k+f$ empfangen. Für die zugehörigen Polynome $f(x)$ und $\widetilde{k}(x)$ gilt dann

$$f(\alpha) \; = \; \widetilde{k}(\alpha), \; f(\alpha^2) \; = \; \widetilde{k}(\alpha^2), \; f(\alpha^3) \; = \; \widetilde{k}(\alpha^3)$$

Der Empfänger kann nun mithilfe des Polynoms

$$p(x) \; = \; \widetilde{k}(\alpha)x^2+\widetilde{k}(\alpha^2)x+\widetilde{k}(\alpha^3)+\widetilde{k}(\alpha)\widetilde{k}(\alpha^2)$$

Fehler korrigieren. Wir unterscheiden drei Fälle.

Fall 1. Gibt es keine Übertragungsfehler, so ist $f(x) = 0$ und $p(x) = 0$.

Fall 2. Bei einem Übertragungsfehler ist $f(x)$ von der Gestalt x^r und damit

$$\begin{aligned} p(x) &= \alpha^r x^2 + \alpha^{2r} x + \alpha^{3r} + \alpha^r \alpha^{2r} \\ &= \alpha^r x(x + \alpha^r). \end{aligned}$$

Fall 3. Bei zwei Übertragungsfehlern gilt $f(x) = x^r + x^s$ mit $r \neq s$. In diesem Fall ergibt sich

$$\begin{aligned} p(x) &= (\alpha^r + \alpha^s)x^2 + (\alpha^{2r} + \alpha^{2s})x \\ &\quad + \alpha^{3r} + \alpha^{3s} + (\alpha^r + \alpha^s)(\alpha^{2r} + \alpha^{2s}) \\ &= (\alpha^r + \alpha^s)(x + \alpha^r)(x + \alpha^s). \end{aligned}$$

Der Empfänger kann diese Fälle unterscheiden, indem er die Nullstellen von $p(x)$ bestimmt. Gleichzeitig kann er in den beiden letzten Fällen r bzw. r, s aus den Nullstellen bestimmen, da α ein primitives Element ist. Er kann also die falsch übertragenen Bits lokalisieren und korrigieren. $\square$

10.4 Spezielle Fälle von BCH-Kodes

Hamming-Kodes

Die Hamming-Kodes ergeben sich bei der Wahl $q = 2$, also $\mathbb{F}_q = \mathbb{Z}_2$, und $n = 2^l - 1$. Wähle α als primitives Element von $\mathbb{F}_{2^l}$, mit Minimalpolynom $g(x) \in \mathbb{Z}_2[x]$. Die Elemente von $\mathbb{F}_{2^l}$ lassen sich eindeutig als Linearkombinationen von $1, \alpha, \alpha^2, \ldots, \alpha^{l-1}$ mit Koeffizienten 0 oder 1 darstellen. Daher gibt es eindeutige $h_{ij} \in \mathbb{Z}_2$, so daß die Gleichungen

$$\alpha^j = \sum_{i=0}^{l-1} \alpha^i h_{ij},$$

$0 \leq j \leq n-1$ gelten. Kodeworte sind durch $k(\alpha) = 0$ charakterisiert, und es gilt

$$k(\alpha) = \sum_{j=0}^{n-1} k_j \alpha^j = \sum_{i=0}^{l-1} \left(\sum_{j=0}^{n-1} h_{ij} k_j \right) \alpha^i = 0$$

genau dann, wenn

$$H \cdot k^t = \begin{pmatrix} h_{00} & \cdots & h_{0,n-1} \\ \vdots & & \vdots \\ h_{l-1,0} & \cdots & h_{l-1,n-1} \end{pmatrix} \cdot \begin{pmatrix} k_0 \\ \vdots \\ k_{n-1} \end{pmatrix} = 0.$$

H ist also die Kontrollmatrix des BCH-Kodes mit erzeugendem Polynom $g(x)$. Da α primitives Element ist, durchlaufen die Spalten von H alle möglichen Vektoren aus Nullen und Einsen der Länge l, abgesehen vom Nullvektor. H ist daher bis auf Reihenfolge der Spalten die Kontrollmatrix des $(2^l - 1, 2^l - l - 1)$-Hamming-Kodes, die beiden Kodes sind äquivalent.

Neben $g(\alpha) = 0$ gilt $g(\alpha^2) = g(\alpha)^2 = 0$. Nach Satz 10.7 ist der Mindestabstand des Kodes mindestens $d = 3$, ein Sachverhalt, der uns bereits bekannt ist.

Reed-Solomon-Kodes

BCH-Kodes, für die $n = q - 1$ gilt, heißen Reed-Solomon-Kodes. Nun ist α primitives Element von $\mathbb{F}_q$, und es gilt

$$g(x) \;=\; (x-\alpha)(x-\alpha^2)\dots(x-\alpha^l).$$

Proposition 10.9. *Reed-Solomon-Kodes sind (n, m)-Blockkodes mit $m = n - l$. Für ihren Mindestabstand gilt*

$$d_{\min} \;=\; l+1 \;=\; n-m+1.$$

Beweis. Die erste Behauptung folgt aus $l = \deg(g) = n - m$ nach Satz 10.4. Nach Satz 10.7 gilt $d_{\min} \geq l + 1$. Umgekehrt gilt für *jeden* linearen (n, m)-Kode die Ungleichung

$$d_{\min} \;\leq\; n-m+1.$$

Zum Beweis sei $k^i = (k_0^i k_1^i \dots k_{n-1}^i), i = 1, \dots, m$ eine Vektorraumbasis von K. Wir betrachten das homogene Gleichungssystem

$$\sum_{i=1}^{m} k_j^i x_i \;=\; 0, \quad j = 0, \dots, m-2,$$

in den m Unbekannten $x_1, \dots, x_m$. Es besteht aus $m - 1$ Gleichungen, deswegen besitzt es eine von 0 verschiedene Lösung $(\lambda_1, \dots, \lambda_m)$. Daher ist auch das Kodewort

$$k \;=\; \sum_{i=1}^{m} \lambda_i k^i$$

von 0 verschieden. Nach Konstruktion sind seine ersten $m - 1$ Buchstaben $k_0, k_1, \dots, k_{m-2}$ gleich 0. Der Hamming-Abstand zwischen k und dem Kodewort $(00\dots0)$ ist daher höchstens $n - m + 1$. Dies ergibt die Behauptung. □

(n, m)-Kodes mit $d_{\min} = n - m + 1$ nennt man **optimal**.

Ein Reed-Solomon-Kode, der s Fehler korrigiert, hat Worte der Länge $n = q-1$ und eine Dimension $m = n-d_{\min}+1 = n-2s$. Wenn wir $q = 2^k$ wählen, können wir ihn in einen binären Kode umwandeln. Jedes Element aus $\mathbb{F}_q$ kann man dann als k-Vektor mit Symbolen aus $\mathbb{Z}_2$ schreiben. Es entsteht so ein binärer Kode mit Wörtern der Länge kn und der Dimension $k(n-2s)$, ein $(kn, k(n-2s))$-Kode. Dieser Kode hat die günstige Eigenschaft, lange Serien aufeinanderfolgender Fehler (‚bursts') korrigieren zu können, und zwar pro Wort einen Serienfehler, dessen Länge kleiner als $(s-1)k$ ist. Eine solche Serie beeinflussen im Originalkode höchstens s aufeinanderfolgende Buchstaben und wird daher korrigiert. Sind die Fehler dagegen über das Wort verstreut, so ist nur garantiert, daß s Fehler korrigiert werden. Reed-Solomon-Kodes haben z.B. in CD-Spielern Anwendung gefunden.