Das Proportionalwahl-Problem als diophantische Näherungsaufgabe

Von

A. Stoll (Zürich)

Das Problem

Bei der Proportionalwahl handelt es sich bekanntlich darum, eine gegebene Anzahl m von Mandaten auf eine ebenfalls gegebene Anzahl p von, im allgemeinen, ungleich grossen Parteien oder «Listen», bzw. Gruppen von solchen, derart zu verteilen, dass die p Anteile M_i, $i = 1$ bis p, zu den ebenfalls gegebenen Stimmenzahlen s_i, $i = 1$ bis p, der p Parteien proportional sind. Es ist offenbar

$$M_i = m \,.\, s_i \,/\, \Sigma s_i, \; i = 1 \text{ bis } p, \tag{1}$$

wobei Σs_i die Summe der p Zahlen s_i bedeutet.

Nun sind aber die Zahlen M_i im allgemeinen nicht ganz. (Grosse Buchstaben sollen im Folgenden stets beliebige reelle Zahlen bezeichnen, während kleine Buchstaben ganze Zahlen bedeuten sollen.) Da die Mandate nicht geteilt werden können, müssen deshalb an Stelle der M_i p ganze Zahlen m_i gesucht werden, welche erstens die gleiche Summe haben:

$$\Sigma m_i = \Sigma M_i = m, \tag{2}$$

und welche zweitens den M_i möglichst nahe kommen.

Zu jeder der Zahlen M_i gibt es eine nächste ganze Zahl. Es ist dies entweder die nächstkleinere $[M_i]$ oder die nächstgrössere $[M_i] + 1$. Wenn nun diese nächsten ganzen Zahlen die Bedingung (2) erfüllen, dann sind dies offenbar die gesuchten Zahlen. Wenn aber die Bedingung (2) nicht erfüllt ist, dann muss eine oder müssen mehrere jener nächsten ganzen Zahlen durch zweit- und drittnächste usw. ersetzt werden. Welcher Ersatz ist nun der beste? Das lässt sich ohne weitere, willkürliche Festsetzungen nicht sagen.

Die präzise Definition einer besten ganzzahligen Näherung bildet den einen springenden Punkt des Problems. Der andere besteht in der Auffindung einer Methode zur Ermittlung jener besten Näherung.

Beste ganzzahlige Näherung bei gegebener Summe

Es seien p reelle Zahlen A_i, $i = 1$ bis p, gegeben. Die nach wachsendem Index geordnete Gruppe dieser p Zahlen sei kürzehalber Punkt genannt und mit (A_i) oder mit $\mathfrak{A}$ bezeichnet. A_i sind die Koordinaten von $\mathfrak{A}$ (etwa in einem p-dimensionalen orthogonalen Koordinatensystem).

Für jeden Index i sei $\overline{a}_i$ die nächste ganze Zahl zu A_i (oder die eine der beiden, wenn A_i genau in der Mitte zwischen zwei ganzen Zahlen liegt). Der ganzzahlige oder Gitterpunkt $\overline{\mathfrak{a}} = (\overline{a}_i)$ ist dann im gewöhnlichen Sinne der nächste Gitterpunkt von $\mathfrak{A}$.

Um nun die Abweichung irgend eines Gitterpunktes $\mathfrak{x} = (x_i)$ von $\mathfrak{A}$ zu definieren, ist zweierlei nötig: ein Mass für die Abweichung jeder einzelnen Koordinate x_i von A_i und ausserdem ein Mass für die Gesamtabweichung.

Beide Masse brauchen nur für ganze Zahlen x definiert zu sein. Für die Einzel-Abweichungsmasse $F_i(x)$ ist es in Hinsicht auf ein Lösungsverfahren zweckmässig zu verlangen, dass sie nach unten konvex oder doch nirgends konkav seien:

$$2 \,.\, F_i(x) \leqq F_i(x-1) + F_i(x+1) \text{ für alle } x. \tag{3}$$

Es entspricht dies auch ihrer Bedeutung, zusammen mit der weiteren, naheliegenden Forderung, dass sie für $\overline{a}_i$ oder wenigstens für die eine oder andere der beiden Zahlen $[A_i]$ und $[A_i]+1$ minimal sein sollen.

Als Gesamtabweichungsmass $F(\mathfrak{x})$ werde naheliegender- und zweckmässigerweise die Summe der Einzelabweichungen gewählt:

$$F(\mathfrak{x}) = \Sigma F_i(x_i) \tag{4}$$

Als beste ganzzahlige Näherung mit der Summe n gelte nun derjenige unter allen Gitterpunkten mit der Koordinatensumme n, für den die Funktion F minimal wird. Ein solcher Punkt heisse Minimalpunkt der Summe n.

Folgende Funktionen sind Beispiele für Einzelabweichungsmasse:

$$U_i = |x - A_i| \tag{5}$$

$$V_i = (x - A_i)^2 \tag{6}$$

$$Y_i = (x - A_i)^2 : |A_i| \tag{7}$$

$$Z_i = \begin{cases} \infty & \text{, für } x < [A_i] \\ (x - A_i + 0{,}5)^2 : |A_i| & \text{, für } x \geqq [A_i] \end{cases} \tag{8}$$

Die Funktionen U_i und V_i verwenden den Absolutfehler, Y_i einen Relativfehler. Die sonderbar anmutende Funktion Z_i ist asymmetrisch; die Koordinaten des Minimalpunktes können nicht kleiner sein als $[A_i]$. Die Funktion $Z = \Sigma Z_i$ entspricht dem Zürcher Wahlverfahren.

Bei der obigen Definition der besten Näherung wurde absichtlich die Summe n nicht an die Summe der A_i gebunden. Durch die so gewonnene Ellbogenfreiheit wird ein einfaches Lösungsverfahren möglich. Es muss aber darauf hingewiesen werden, dass bei grosser Differenz zwischen n und ΣA_i und bei ungeeigneter Wahl der Abweichungsmasse, zum Beispiel bei den Funktionen U_i, Lösungen (n_i) möglich werden, die man mit dem besten Willen auch nicht «einigermassen proportional» zu den A_i nennen kann.

Minimalketten

Es seien die Funktionen $F_i(x)$, $i = 1$ bis p, gegeben. Sie sollen alle der Konvexitätsbedingung (3) genügen und ausserdem für alle endlichen x endlich sein. Funktionen wie Z_i seien also noch ausgeschlossen.

Zu jedem Gitterpunkt $\mathfrak{x}$ gehört nun ein bestimmter Wert der Funktion $F(\mathfrak{x})$. Vergrössert man die i-te Koordinate x_i um 1, so soll das ein Rechtsschritt heissen. Er bewirkt eine Änderung von F um die Änderung von $F_i(x_i)$, die mit $R_i(x_i)$ bezeichnet und Grösse des Schrittes genannt werde. Unter den p Änderungen $R_i(x_i)$ gibt es eine kleinste $R(\mathfrak{x})$. Ist ein Rechtsschritt geschehen, so können weitere Rechtsschritte folgen. Wegen der Konvexitätsbedingung sind diese niemals kleiner als $R(\mathfrak{x})$.

Analog soll die Verkleinerung einer Koordinate um 1 ein Linksschritt heissen. Die p möglichen Linksschritte bewirken p Änderungen $L_i(x_i)$. Unter ihnen gibt es einen kleinsten $L(\mathfrak{x})$. Auch die späterhin möglichen Linksschritte sind wegen der Konvexitätsbedingung nie kleiner als $L(\mathfrak{x})$.

$L_i(x_i) + R_k(x_k) < 0$ ist für $i = k$ wegen der Konvexitätsbedingung ausgeschlossen. Für $i \neq k$ ist es dann ausgeschlossen, wenn $\mathfrak{x}$ Minimalpunkt ist. Denn dann würde die Kombination des i-ten Linksschrittes mit dem k-ten Rechtsschritt zu einem neuen Punkt mit gleicher Koordinatensumme, aber mit kleinerem F führen. Für einen Minimalpunkt ist somit notwendig

$$L(\mathfrak{x}) + R(\mathfrak{x}) \geqq 0 \qquad (9)$$

Sei andererseits $\mathfrak{x}$ ein Punkt, für den (9) gilt. Irgend ein anderer Gitterpunkt mit der gleichen Koordinatensumme wird von $\mathfrak{x}$ aus erreicht, indem in jeder Koordinate eine gewisse Anzahl (auch null) von lauter Linksschritten oder von lauter Rechtsschritten ausgeführt wird, im ganzen aber ebensoviele Links- wie Rechtsschritte, etwa je s. Die dadurch bewirkte totale Änderung von F ist dann nach obigem sicher nicht kleiner als $s \, . \, L(\mathfrak{x}) + s \, . \, R(\mathfrak{x})$ und daher wegen (9) nicht negativ. $\mathfrak{x}$ ist also Minimalpunkt.

Somit ist die Ungleichung (9) für einen Minimalpunkt charakteristisch.

Sei nun $\mathfrak{x}$ ein Minimalpunkt der Summe n und $R_1(x_1) = R(\mathfrak{x})$ ein kleinster Rechtsschritt. Er führt zu einem neuen Punkt $\mathfrak{x}'$, für den die Rechtsschritte R_2 bis R_p dieselben sind wie für $\mathfrak{x}$, während $R_1(x_1 + 1)$ nicht kleiner ist als $R_1(x_1)$. Also ist $R(\mathfrak{x}') \geqq R(\mathfrak{x})$. Die Linksschritte L_2 bis L_p sind für $\mathfrak{x}'$ ebenfalls dieselben wie für $\mathfrak{x}$, während $L_1(x_1 + 1) = - R_1(x_1) = - R(\mathfrak{x})$ ist, und dies ist wegen der Minimalpunktsforderung $\leqq L(\mathfrak{x})$ und also nicht grösser als L_2 bis L_p. Daher ist $L(\mathfrak{x}') = - R(\mathfrak{x})$ und somit $L(\mathfrak{x}') + R(\mathfrak{x}') \geqq - R(\mathfrak{x}) + R(\mathfrak{x}) = 0$. Das heisst $\mathfrak{x}'$ ist ein Minimalpunkt der Summe $n + 1$.

Auf gleiche Weise erkennt man, dass ein kleinster Linksschritt zu einem Minimalpunkt der Summe $n - 1$ führt. Wiederholung ergibt das Resultat:

Wenn für irgendeine Summe ein Minimalpunkt existiert, dann existiert eine nicht abbrechende Kette von Minimalpunkten, deren Summe die Reihe der ganzen Zahlen durchläuft, eine Minimalkette.

Kennt man irgendein Glied einer Minimalkette, so findet man jedes andere durch endlich viele Minimalschritte. Besitzen die Funktionen F_i ein Minimum, etwa für c_i, $i = 1$ bis p, dann ist $F(\mathfrak{c})$ Minimum von F und $\mathfrak{c}$ absoluter Minimalpunkt und also auch Minimalpunkt für die Summe seiner Koordinaten.

Damit ist ein allgemein brauchbares Verfahren zur Ermittlung der durch F definierten besten Näherung mit gegebener Summe gefunden.

Wenn die Funktionen F_i nur für beschränkte Bereiche u_i bis v_i endlich, ausserhalb aber unendlich sind, wie zum Beispiel die Funktionen Z_i, dann bewirken Schritte, die über eine Schranke hinausführen, eine unendlich grosse Änderung von F. Als Minimalschritte fallen sie von selber aus der Konkurrenz, so dass Minimalketten nicht über die Schranken hinausführen können und in den äussersten Ecken (u_i) und (v_i) abbrechen müssen. Im übrigen bleiben die angestellten Überlegungen unverändert richtig. Die Ecken (u_i) und (v_i) selber sind Minimalpunkte. Denn für (u_i) sind alle Linksschritte positiv unendlich und alle Rechtsschritte endlich, so dass $L + R$ positiv wird. Analog für (v_i).

Ein- und Mehrdeutigkeit

Wenn für eine bestimmte Koordinatensumme nur ein einziger Minimalpunkt existiert, dann soll dieser «einzeln» heissen.

Für einen Einzel-Minimalpunkt darf die Links-Rechtsschritt-Summe $L_i(x_i) + R_k(x_k)$ höchstens für einen einzigen Index $i = k = e$ verschwinden, für alle andern muss sie positiv ausfallen. Denn andernfalls gäbe es einen Links- und einen Rechtsschritt mit ungleichem Index, die zusammen zu einem neuen Minimalpunkt gleicher Summe führen würden.

Wenn umgekehrt $L_i(x_i) + R_k(x_k)$ für alle Indexkombinationen, ausgenommen höchstens für $i = k = e$, positiv ist, dann führt wegen der Konvexitätsbedingung von $\mathfrak{x}$ aus zu jedem andern Punkt mit gleicher Koordinatensumme eine Auswahl von solchen jetzt und anschliessend möglichen Linksschritten und ebensovielen Rechtsschritten anderer Dimension, dass deren Gesamtwirkung die Funktion $F(\mathfrak{x})$ sicher vergrössert, so dass also $\mathfrak{x}$ Einzel-Minimalpunkt ist.

Unter dem Rechtsraum von $\mathfrak{x}$ sollen alle und nur die Punkte verstanden werden, deren Koordinaten nicht kleiner sind als die entsprechenden Koordinaten von $\mathfrak{x}$, unter dem Linksraum diejenigen, deren Koordinaten nicht grösser sind. Zu einem Gitterpunkt im Rechtsraum führen lauter Rechtsschritte, zu einem solchen im Linksraum lauter Linksschritte.

Sei nun $\mathfrak{x}$ ein Einzel-Minimalpunkt und $\mathfrak{y}$ ein beliebiger Punkt, der aber weder im Links- noch im Rechtsraum von $\mathfrak{x}$ liegen soll. Der Weg von $\mathfrak{x}$ nach $\mathfrak{y}$ enthält dann mindestens einen Linksschritt, es sei L_1, und mindestens einen Rechtsschritt anderer Dimension, es sei R_2. Dabei ist $L_1 + R_2 > 0$, weil $\mathfrak{x}$ ein einzelner Minimalpunkt ist, und für den durch diese zwei Schritte erreichten Punkt ist $R_1(x_1 - 1) = L_1$ und $L_2(x_2 + 1) = -R_2$.

Enthält der Weg nach $\mathfrak{y}$ ausser L_1 und R_2 keine weiteren Schritte in diesen Dimensionen, dann ist sicher $L(\mathfrak{y}) \leqq -R_2$ und $R(\mathfrak{y}) \leqq -L_1$. Folglich ist $L(\mathfrak{y}) + R(\mathfrak{y}) \leqq -R_2 - L_1 < 0$, also $\mathfrak{y}$ kein Minimalpunkt. Enthält aber der Weg in der ersten Dimension noch weitere Linksschritte, dann sind diese nicht kleiner als L_1 und daher die neu auftretenden Rechtsschritte nicht grösser als der Rechtsschritt $-L_1$, den sie ersetzen. Analog sind bei weiteren Rechtsschritten in der zweiten Dimension die neu auftretenden Linksschritte nicht grösser als $-R_2$. Also ist $L(\mathfrak{y}) + R(\mathfrak{y})$ erst recht < 0 und $\mathfrak{y}$ kein Minimalpunkt.

Es gilt somit: Wenn $\mathfrak{x}$ ein Einzel-Minimalpunkt ist, dann gibt es weitere Minimalpunkte (mit andern Koordinatensummen) nur im Links- und Rechtsraum von $\mathfrak{x}$.

Nunmehr sei $\mathfrak{x}$ ein Minimalpunkt der Summe x. Unter seinen Rechtsschritten gebe es mehr als einen kleinsten, etwa (bei passender Numerierung) $R_i(x_i) = R$ für $i = 1$ bis q. Sie führen zu q Minimalpunkten der Summe $x + 1$. Geschieht R_1, dann fällt für einen nächsten Rechtsschritt die erste Dimension aus der Wahl oder nicht, je nachdem $R_1(x_1 + 1)$ grösser ist als R oder gleich gross.

Wenn es in der i-ten Dimension r_i, aber nicht mehr, aufeinanderfolgende Rechtsschritte der Grösse R gibt, dann kann man einerseits durch Minimalschritte sämtliche Gitterpunkte des Quaders, der dem Rechtsraum von $\mathfrak{x} = (x_i)$ und dem Linksraum von $\mathfrak{x} + \mathfrak{r} = (x_i + r_i)$ gemeinsam ist, erreichen, und andererseits muss der Weg zu $\mathfrak{x} + \mathfrak{r}$ führen, weil der Minimalschritt nicht grösser werden kann als R ehe nicht alle $r_1 + r_2 + \ldots + r_q$ Möglichkeiten erschöpft sind.

Für $\mathfrak{x} + \mathfrak{r}$ ist erstmals $R(\mathfrak{x} + \mathfrak{r}) > R$. Seine Linksschritte sind in den Dimensionen 1 bis q gleich $-R$, und in den übrigen Dimensionen können sie nicht kleiner sein. Denn sonst wären sie es schon für $\mathfrak{x}$ gewesen, und es wäre $L(\mathfrak{x}) + R(\mathfrak{x}) < -R + R = 0$, was der Minimalpunktsforderung widerspricht. Es ist also $L(\mathfrak{x} + \mathfrak{r}) = -R$ und somit $L(\mathfrak{x} + \mathfrak{r}) + R(\mathfrak{x} + \mathfrak{r}) > 0$, d. h. $\mathfrak{x} + \mathfrak{r}$ ist Einzel-Minimalpunkt.

Aus dem bisherigen erhellt: Sämtliche Minimalpunkte fügen sich im wesentlichen zu einer einzigen Minimalkette zusammen.

Soweit diese aus Einzel-Minimalpunkten besteht, ist sie glatt, und jeder Punkt liegt im Rechtsraume seines Vorgängers. Mehrdeutigkeit tritt dort auf, wo die Kette «Knollen» hat. Diese bestehen aus allen Gitterpunkten von je einem zwei- bis p-dimensionalen Quader, dessen Schnitte mit den Ebenen konstanter Koordinatensumme alle möglichen besten Näherungen der betreffenden Summe repräsentieren. Ihre äussersten Minimalpunkte links und rechts (kleinste und grösste Summe) sind einzeln.

Wenn die Funktionen F_i keine flachen Stellen haben, dann sind alle $r_i = 1$, und die Quader sind Würfel.

Absolutfehler-Verfahren

Es seien die positiven Zahlen M_i mit der Summe m gegeben. Sie sollen durch ganze Zahlen m_i mit der gleichen Summe approximiert werden. Als Näherungskriterium diene die Summe U der Funktionen U_i der Gleichung (5):

$$U_i = |x_i - M_i| \qquad (10)$$

Daher ist:

$$L_i(x_i) = \begin{Bmatrix} 1 \\ 2(M_i - [M_i]) - 1 \\ -1 \end{Bmatrix}, \text{ für } x_i \begin{Bmatrix} < \\ = \\ > \end{Bmatrix} [M_i] + 1 \qquad (11)$$

$$R_i(x_i) = \begin{Bmatrix} -1 \\ 1 - 2(M_i - [M_i]) \\ 1 \end{Bmatrix}, \text{ für } x_i \begin{Bmatrix} < \\ = \\ > \end{Bmatrix} [M_i]. \qquad (12)$$

Für $\mathfrak{x} = [\mathfrak{M}] = ([M_i])$ ist $L_i = 1$ und, wegen $M_i - [M_i] < 1$, $R_i > -1$, also $L(\mathfrak{x}) + R(\mathfrak{x}) > 0$. Dasselbe gilt für $\mathfrak{x} = [\mathfrak{M}] + \mathfrak{E} = ([M_i] + 1)$, da hier $R_i = 1$ und $L_i > -1$ ist, sofern wenigstens M_i nicht ganz ist, in welchem Falle an Stelle des Zeichens $>$ das Gleichheitszeichen treten würde.

$[\mathfrak{M}]$ ist also Einzel-Minimalpunkt, und der gesuchte Minimalpunkt $\mathfrak{m}$ liegt in seinem Rechtsraum; d. h. kein m_i ist kleiner als $[M_i]$, somit auch keines negativ. $[\mathfrak{M}] + \mathfrak{E}$ ist Minimalpunkt und sogar einzeln, wenn kein M_i ganz ist.

$[\mathfrak{M}]$ ist mit $[\mathfrak{M}] + \mathfrak{E}$ durch mindestens eine Kette von lauter minimalen Rechtsschritten verbunden, so dass mindestens eine Lösung $\mathfrak{m}$ durch einen Punkt des Einheitswürfels mit den Gegenecken $[\mathfrak{M}]$ und $[\mathfrak{M}] + \mathfrak{E}$ dargestellt wird. Wenn kein M_i ganz ist, kann es nur solche Lösungen geben, weil dann $\mathfrak{m}$ im Linksraum von $[\mathfrak{M}] + \mathfrak{E}$ liegen muss. Andernfalls kann es noch Lösungen ausserhalb des genannten Würfels, d. h. mit Koordinaten, die teilweise um mehr als 1 von dem entsprechenden M_i abweichen, geben.

Beispielshalber sei noch erwähnt, dass der ganze Linksraum von $[\mathfrak{M}]$ einen sich ins Unendliche erstreckenden p-dimensionalen Knollen bildet. Ebenso der ganze Rechtsraum von $[\mathfrak{M}] + \mathfrak{E}$ oder eines um eine oder mehrere Dimensionen vorgeschobenen Punktes.

Zur Auffindung von $\mathfrak{m}$ kann man nach Belieben bei $[\mathfrak{M}]$ oder $[\mathfrak{M}] + \mathfrak{E}$ starten oder, noch zweckmässiger weil mit passenderer, wenn nicht schon richtiger Koordinatensumme, beim $\mathfrak{M}$ am nächsten liegenden Gitterpunkt $\overline{\mathfrak{m}}$, der absoluter, aber nicht notwendig einzelner Minimalpunkt ist.

Über die Auswahl der notwendigen Minimalschritte entscheidet beim Start in $[\mathfrak{M}]$ gemäss Gleichung (12) jedesmal das Minimum von $1 - 2(M_i - [M_i])$, d. h. das Maximum von $M_i - [M_i]$. Mit andern Worten: zuerst werden alle M_i abgerundet und dann der Reihe nach die nötige Anzahl der am meisten abgerundeten M_i aufgerundet. Beim Start in $[\mathfrak{M}] + \mathfrak{E}$ entscheidet über die Linksschritt-Wahl gemäss Gleichung (11) das Minimum von $2(M_i - [M_i]) - 1$, bzw. von $M_i - [M_i]$. Mit andern Worten: zuerst werden alle M_i aufgerundet und dann der Reihe nach die nötige Anzahl der am stärksten aufgerundeten M_i abgerundet. Startet man aber in $\overline{\mathfrak{m}}$, dann geschieht nötigenfalls das eine oder andere, jedoch in geringerer Zahl.

Das geschilderte Start- und Auswahlverfahren ist genau das, was ein kluger mathematischer Laie im vorliegenden Falle machen würde. Man kann deshalb dieses Verfahren füglich als das Verfahren des gesunden Menschenverstandes bezeichnen.

Wählt man an Stelle der Gesamtabweichungsfunktion U die Summe V der Funktionen V_i der Gleichung (6):

$$V_i = (x_i - M_i)^2, \qquad (13)$$

dann ist

$$L_i(x_i) = 1 - 2(x_i - M_i) \qquad (14)$$

und

$$R_i(x_i) = 1 + 2(x_i - M_i). \qquad (15)$$

Für $x_i = [M_i]$ ist $L_i \geqq 1$ und $R_i > -1$, und für $x_i = [M_i] + 1$ ist $L_i \geqq -1$ und $R_i > 1$. Daher ist sowohl für $[\mathfrak{M}]$ als auch für $[\mathfrak{M}] + \mathfrak{E}$ $L + R > 0$, d. h. beides sind Einzel-Minimalpunkte. $\overline{\mathfrak{m}}$ ist auch jetzt absoluter, aber nicht notwendig einzelner Minimalpunkt. Die möglichen Startpunkte sind also die gleichen wie bei der Funktion U. Da aber jetzt in allen Fällen $[\mathfrak{M}] + \mathfrak{E}$ Einzel-Minimalpunkt ist, gibt es ausserhalb des Würfels mit den Gegenecken $[\mathfrak{M}]$ u. $[\mathfrak{M}] + \mathfrak{E}$ keine Lösungen.

Für die Schrittwahl ist massgebend: bei Start in $[\mathfrak{M}]$, nach Gleichung (15) das Minimum von $1 + 2([M_i] - M_i)$, d. h. das Maximum von $M_i - [M_i]$, und bei Start in $[\mathfrak{M}] + \mathfrak{E}$ das Minimum von $1 - 2([M_i] - M_i)$, d. h. das Minimum von $M_i - [M_i]$, bei Start in $\overline{\mathfrak{m}}$ aber das eine oder das andere, also genau wie bei der Funktion U. Die Gesamtabweichungsfunktionen U und V führen zu den gleichen Lösungen.

Die unterschiedliche Wirkung der beiden Funktionen zeigt sich im wesentlichen erst ausserhalb des genannten Einheitswürfels. Aus (14) und (15) erkennt man leicht, dass jeder Punkt $[\mathfrak{M}] + n\mathfrak{E} = ([M_i] + n)$ für alle ganzen Zahlen n Einzel-Minimalpunkt ist, so dass die Minimalkette auf der Kette der Einheitswürfel zwischen $[\mathfrak{M}] + n\mathfrak{E}$ und $[\mathfrak{M}] + (n+1)\mathfrak{E}$ verläuft. Weiter erkennt man, dass das Schrittwahl-Kriterium innerhalb jedes Würfels dasselbe ist, so dass die Kettenstücke auf allen Würfeln kongruent sind.

Ein Relativfehler-Verfahren

M_i seien wieder p gegebene positive, m_i p gesuchte ganze Zahlen mit der gleichen Summe m wie jene, und die jene approximieren gemäss der Summe Y der Funktionen Y_i der Gleichung (7):

$$Y_i = (x_i - M_i)^2 / M_i \,. \tag{16}$$

Dann ist

$$L_i(x_i) = \frac{1 - 2(x_i - M_i)}{M_i} \tag{17}$$

und

$$R_i(x_i) = \frac{1 + 2(x_i - M_i)}{M_i} \,. \tag{18}$$

Auch jetzt ist $\overline{\mathfrak{m}}$ absoluter Minimalpunkt, wie (16) unmittelbar zeigt, und wie (17) und (18) bestätigen, da die Zähler nie negativ sind. Dagegen sind $[\mathfrak{M}]$ und $[\mathfrak{M}] + \mathfrak{E}$ im allgemeinen nicht Minimalpunkte.

Der Nullpunkt aber ist Einzel-Minimalpunkt, da für ihn $L_i + R_k = 1/M_i + 1/M_k$ für alle i und k positiv ausfällt. $\mathfrak{m}$ liegt daher auf alle Fälle im Rechtsraum des Nullpunktes, negative m_i sind nicht zu befürchten.

$\overline{\mathfrak{m}}$ ist der gegebene Startpunkt. Ist dann seine Koordinatensumme $\overline{m}$ kleiner als m, so ist ein minimaler Rechtsschritt nötig. Seine Auswahl erfolgt nach (18) gemäss dem Minimum der p Quotienten $(1 + 2\overline{m}_i)/M_i$, oder — da alle positiv sind — gemäss dem Maximum von $M_i/(2\overline{m}_i + 1)$, oder endlich — da die M_i zu den s_i proportional sind — gemäss dem Maximum von $s_i/(2\overline{m}_i + 1)$.

Liegt dieses Maximum etwa für $i = 1$ vor und ist noch einmal ein Rechtsschritt nötig, dann konkurriert an Stelle von $s_1/(2\overline{m}_1 + 1)$ die kleinere Zahl $s_1/(2\overline{m}_1 + 3)$, und zwar durchaus nicht zum vornherein aussichtslos, sondern nicht selten mit Erfolg. Dies im Gegensatz zum Absolutfehler-Verfahren, wo nach einem Minimalschritt zuerst alle andern Dimensionen an die Reihe kommen.

Ist aber $\overline{m}$ grösser als m, dann ist ein minimaler Linksschritt nötig. Seine Auswahl erfolgt nach (17) gemäss dem Minimum der p Quotienten $(1 - 2\overline{m}_i)/M_i$, oder gemäss dem Maximum von $(2\overline{m}_i - 1)/M_i$. Unter diesen Quotienten ist mindestens einer positiv, weil die $\overline{m}_i$ nicht negativ sind, und weil, wenn alle null wären, ein Linksschritt überhaupt nicht in Frage käme. Nur die positiven bleiben in der Konkurrenz, und unter ihnen entscheidet das Minimum von $M_i/(2\overline{m}_i - 1)$ oder von $s_i/(2\overline{m}_i - 1)$.

Liegt dieses Minimum etwa für $i = 1$ vor, so konkurriert nunmehr an Stelle von $s_1/(2\overline{m}_1 - 1)$ die neue Zahl $s_1/(2\overline{m}_1 - 3)$, sofern sie noch positiv ist. Andernfalls fällt diese Dimension aus der Wahl.

Die Rechenvorschrift für dieses Relativfehler-Verfahren lautet also: Ermittlung von $\overline{\mathfrak{m}}$ durch Division der Produkte ms_i durch die totale Stimmenzahl s, Bestimmung der jeweiligen Reste und Erhöhung des Quotienten um 1 falls Rest grösser als $s/2$. Bestimmung der Koordinatensumme $\overline{m}$; wenn kleiner als m dann Rechtsschritte nötig, wenn grösser als m dann Linksschritte. Auswahl der Rechtsschritte gemäss dem

Maximum der (stets positiven) Quotienten $s_i/(2x_i+1)$, Auswahl der Linksschritte gemäss den positiven unter den Zahlen $s_i/(2x_i-1)$, wobei für x_i die Koordinaten des jeweilen erreichten Punktes einzusetzen sind. Wiederholung bis zur Erreichung der Koordinatensumme m.

Die Rechnung ist etwas umständlicher als beim Absolutfehler-Verfahren, weil ausser den Quotienten und Resten zur Bestimmung von $\overline{m}$ noch die eben genannten Quotienten berechnet werden müssen. Es ist aber immer noch einfach genug, um für die Praxis der Proportionalwahl in Frage zu kommen.

Das Zürcher Verfahren

Es wird hier so genannt, weil es in Zürich gesetzliche Gültigkeit hat. Was für Überlegungen bei der Aufstellung seiner Rechenvorschriften massgebend waren, ist dem Verfasser nicht genau bekannt. Dem hier eingenommenen allgemeinen Gesichtspunkt lässt es sich unterordnen wie folgt.

Gesamtabweichungsfunktion ist hier die Summe Z der Funktionen (8):

$$Z_i = \begin{cases} +\infty & \text{, für } x_i < [M_i] \\ \dfrac{(x_i - M_i + 1/2)^2}{M_i} & \text{, für } x_i \geqq [M_i]. \end{cases} \tag{19}$$

Daher ist

$$L_i(x_i) = \begin{cases} \text{nicht erklärt} & \text{für } x_i < [M_i] \\ +\infty & \text{, für } x_i = [M_i] \\ -\dfrac{2(x_i - M_i)}{M_i} & \text{, für } x_i > [M_i] \end{cases} \tag{20}$$

und

$$R_i(x_i) = \begin{cases} \text{nicht erklärt} & \text{für } x_i < [M_i] \\ \dfrac{2(x_i - M_i + 1)}{M_i} & \text{, für } x_i \geqq [M_i]. \end{cases} \tag{21}$$

$\overline{m}$ ist hier im allgemeinen nicht Minimalpunkt, ebensowenig $[\mathfrak{M}] + \mathfrak{E}$. Hingegen ist $[\mathfrak{M}]$ stets absoluter und einzelner Minimalpunkt, denn für ihn sind alle L_i und alle R_i positiv.

$[\mathfrak{M}]$ kann somit als Startpunkt dienen, und es kommen höchstens Rechtsschritte in Frage. Ihre Auswahl geschieht nach (21) gemäss dem Minimum von $(x_i+1)/M_i$, oder — da alle positiv sind — gemäss dem Maximum von $M_i/(x_i+1)$, bzw. von $s_i/(x_i+1)$, zunächst für $x_i = [M_i]$.

Es sei darauf hingewiesen, dass nach dieser — gesetzlichen — Auswahlregel jeweilen von den p durch einen Rechtsschritt erzeugbaren neuen Rechtsabweichungen $(x_i+1)/M_i$ die kleinste realisiert wird, ohne Rücksicht darauf, ob die dadurch ersetzte Abweichung um viel oder wenig kleiner war, weshalb nicht die Summe dieser Abweichungen, sondern eben die Funktion Z möglichst klein bleibt.

Die gesetzliche Rechenvorschrift realisiert noch eine bemerkenswerte Abkürzung, indem durch einen sehr geschickten Kunstgriff der Startpunkt möglichst weit nach rechts verlegt wird, in jeder Koordinate jedoch höchstens um 1 und im ganzen so, dass die Koordinatensumme nicht grösser wird als m.

Anstatt nämlich die m_i zu berechnen durch Multiplikation der s_i mit m/s bzw. durch Division mit s/m, d. h. der auf ein Mandat entfallenden Stimmenzahl, werden vielmehr die s_i durch eine etwas kleinere Zahl w, die sogenannte Wahlzahl, dividiert, und die Quotienten auf die nächst kleinere ganze Zahl abgerundet:

$$w = [s/(m+1)] + 1, \tag{22}$$

$$k_i = [s_i/w], \quad \mathfrak{k} = (k_i). \tag{23}$$

Da w nach (22) grösser ist als $s/(m+1)$, ist $s_i/w < s_i(m+1)/s = M_i(m+1)/m = M_i + M_i/m < M_i + 1$, und daher nach (23):

$$k_i \leqq [M_i] + 1. \tag{24}$$

und

$$k = \Sigma k_i \leqq [(\Sigma s_i)/w] = [s/w] < m + 1 \leqq m. \tag{25}$$

Wegen (25) werden niemals Linksschritte nötig. Dass aber $\mathfrak{k}$ selber Minimalpunkt ist, zeigt folgende Überlegung.

In der Praxis der Proportionalwahl ist s nicht kleiner als $m(m+1)$. Dann aber ist

$$s/(m+1) = s/m - s/m(m+1) \leqq s/m - 1,$$

daher

$$w = [s/(m+1)] + 1 \leqq s/m,$$

und somit

$$k_i = [s_i/w] \geqq [s_i m/s] = [M_i].$$

Ist nun etwa $k_1 = [M_1] + 1$, aber $k_2 = [M_2]$, so ist

$$s_1/w \geqq k_1 = [M_1] + 1$$

und

$$s_2/w < k_2 + 1 = [M_2] + 1,$$

folglich

$$s_1/([M_1] + 1) \geqq w > s_2/([M_2] + 1),$$

d. h. die Auswahlregel verlangt, ausgehend von [$\mathfrak{M}$], den Rechtsschritt R_1 vor R_2 und führt somit sukzessive zwangsläufig zu $\mathfrak{k}$. $\mathfrak{k}$ ist also Minimalpunkt.

Für den Fall, dass s kleiner ist als $m(m+1)$, lässt sich beweisen, dass $\mathfrak{k}$ zusammen mit der Auswahlregel zum Minimalpunkt [$\mathfrak{M}$] führt, doch ist dies hier ohne Bedeutung und sei daher unterdrückt.

Vergleich der drei Verfahren

Von den drei Verfahren lässt das Absolutfehler-Verfahren an Einfachheit und Durchsichtigkeit nichts zu wünschen übrig. Dem gegenüber sind die andern beiden zwar in der Handhabung noch ziemlich einfach, aber das Relativfehler-Verfahren ist für den Laien bereits undurchsichtig, und das Zürcher Verfahren muss ihm vollends unverständlich bleiben. Für die Anwendung bei Wahlen ist dies nicht unwesentlich.

Der wichtigste Unterschied zwischen dem Absolutfehler-Verfahren und dem Relativfehler-Verfahren besteht in der Grösse der möglichen Abweichungen der m_i von den genauen, aber nicht realisierbaren Zuteilungszahlen M_i.

Beim Absolutfehler-Verfahren sind diese Abweichungen stets kleiner als eins, und wenn etwa M_l zufällig ganz und also realisierbar ist, dann ist $m_l = M_l$. Für die stattfindenden Auf- oder Abrundungen sind allein die Überschüsse der M_i über die nächst kleineren ganzen Zahlen $[M_i]$ massgebend, nicht aber die Grösse der letzteren. Grosse und kleine Parteien werden in diesem Sinne gleich behandelt.

Anders beim Relativfehler-Verfahren. Hier kann, nach (17), ein Linksschritt selbst dann kleiner sein als alle andern, wenn $x_i - M_i$ bereits null oder sogar negativ geworden ist; es ist dazu nur ein genügend grosses M_i erforderlich. Analog kann, nach (18), ein Rechtsschritt auch dann kleiner sein als alle andern, wenn sein x_i bereits grösser ist als sein M_i. Die grossen Parteien dürfen und sollen also grössere Abweichungen haben als die kleinen.

Diese unterschiedliche Behandlung der grossen und kleinen Parteien bedeutet keine Bevorzugung der Grossen. Denn die Abweichungen können ebensowohl positiv wie negativ sein, und über die Notwendigkeit des einen oder des andern entscheidet der Wahlzufall und nicht die Parteigrösse. Es wird hier einfach der Gedanke realisiert, dass ein Mandat mehr oder weniger für eine grosse Partei weniger ausmacht als für eine kleine.

Das Zürcher Verfahren hingegen begünstigt dezidiert die grossen Parteien. Gemäss den Abweichungsfunktionen (19) bleibt $\mathfrak{m}$ auf den Rechtsraum von [$\mathfrak{M}$] beschränkt, die Unterschreitungen sind also alle kleiner als eins. Es wird im Prinzip von [$\mathfrak{M}$], also von den kleinsten Unterschreitungen ausgegangen, und es werden so viele Unterschreitungen in Überschreitungen verwandelt, oder bereits vorhandene Überschreitungen vergrössert bis die Koordinatensumme m erreicht ist. Die Auswahl der dazu nötigen Rechtsschritte erfolgt wie beim Relativfehler-Verfahren nach Massgabe relativer Abweichungen, d. h. grosse Parteien dürfen grössere absolute Abweichungen haben.

Da dies einseitig nur bei Rechtsschritten geschieht, bedeutet es eine Begünstigung der grossen Parteien gegenüber den kleinen. Das braucht kein Fehler zu sein; denn die Wahl der Abweichungsfunktion ist frei. Es können politische Gründe dafür massgebend sein. Solche stehen hier nicht zur Diskussion. Unbefriedigend ist der Umstand, dass $\overline{\mathfrak{m}}$ nicht notwendig Minimalpunkt ist.

Zahlenbeispiele

Die folgenden Stimmenzahlen sind den Ergebnissen der letzten Zürcher Gemeinderatswahlen entnommen.

Die Zuteilung der Mandate wurde nach den drei geschilderten Verfahren berechnet und in leicht verständlicher Weise mit A, R und Z unterschieden. Ausserdem geschah die Berechnung nicht nur entsprechend den geschehenen und nicht in allen Wahlkreisen gleichen Listenverbindungen, sondern auch ohne solche, weil durch sie der Zusammenhang mit den Parteistimmenzahlen teilweise verwischt wird.

Die folgenden zwei Tabellen enthalten die Zahlen der Wahlkreise 5 und 6:

Wahlkreis 5

Partei	Stimmen-zahlen	Mandate ohne L'verb. A	R	Z	Mandate bei L'verb. A	R	Z
B	618	0	0	0	0	0	0
C	2 216	1	1	0	1	1	1
D	506	0	0	0	0	0	0
E	—	—	—	—	—	—	—
F	1 781	0	0	0	0	0	0
K	7 134	2	2	2	2	2	2
L	719	0	0	0	0	0	0
S	11 027	2	2	4	2	2	3
U	2 471	1	1	0	1	1	0

Wahlkreis 6

Partei	Stimmen-zahlen	Mandate ohne L'verb. A	R	Z	Mandate bei L'verb. A	R	Z
B	10 931	1	1	1	1	1	1
C	19 908	1	2	2	2	2	2
D	9 293	1	1	0	1	1	1
E	3 933	0	0	0	0	0	0
F	37 631	3	3	4	3	3	4
K	11 945	1	1	1	1	1	1
L	5 253	1	0	0	0	0	0
S	38 630	4	3	4	4	4	3
U	43 676	4	5	4	4	4	4

Man sieht, dass sich die Ergebnisse nach den drei Verfahren in der Regel gar nicht oder nur um 1 unterscheiden. Ein einziges mal bei 9 Parteien und 11 Wahlkreisen wuchs die Differenz auf 2 an, nämlich bei der Partei S im Wahlkreis 5. Bei hinreichend grossem Unterschied der Parteistärken könnte freilich die Differenz noch grösser werden.

Bei der nächsten Tabelle sind die Zahlen über alle 11 Wahlkreise summiert. Die Mandatzahlen m sind wieder ohne und mit Listenverbindung berechnet, ausserdem aber noch ohne Unterteilung in Wahlkreise (o. W'k.) und ohne Listenverbindung, entsprechend den totalen Parteistimmenzahlen.

Partei	Totale Stimmen-zahlen	m ohne L'verb. A	R	Z	m bei L'verb. A	R	Z	m o. W'k. A	R	Z
B	53 235	9	8	2	8	8	4	6	6	6
C	107 617	13	13	12	14	13	14	12	12	12
D	46 723	5	5	2	6	5	4	5	5	5
E	21 060	0	0	0	0	0	0	2	2	2
F	165 500	20	20	22	18	19	25	19	19	18
K	176 754	19	19	21	18	19	19	20	20	20
L	30 087	1	0	0	0	0	0	3	3	3
S	325 523	36	37	44	37	38	38	36	36	37
U	193 402	22	23	22	24	23	21	22	22	22

Wie man sieht, weichen die nach dem A- und R-Verfahren berechneten Mandatzahlen höchstens um 1 voneinander ab, da ihre Abweichungen zufälliger Art sind und sich bei der Summation ausgleichen. Die Abweichungen beim Z-Verfahren sind dagegen systematischer Art und vergrössern sich daher bei der Summation. Siehe die 44 Mandate der S-Partei beim Z-Verfahren und bei Fehlen von Listenverbindung gegenüber den 36 Mandaten beim A-Verfahren und den 37 beim R-Verfahren.

Durch die geschehenen Listenverbindungen wurde diese Vorzugsstellung der grössten Partei gebrochen. Sie war mit der K-Partei verbunden, welche nichts davon profitierte. Hingegen wurde diesen beiden durch unterschiedliche Verbindungen der Parteien B bis F 8 Mandate weggenommen. Die Parteien L und U blieben unverbunden.

Die Tabelle zeigt ausserdem, besonders deutlich durch ihre drei letzten Kolonnen, dass die Unterteilung in Wahlkreise eine sehr viel stärkere Wirkung hat als die Wahl des Verfahrens. Sie beeinflusst das Endresultat so stark, dass man nur noch bedingt von Proportionalität reden kann. Doch dies gehört nicht mehr in den Rahmen dieser Untersuchung.