MATHEMATISCH CENTRUM,
Rekenafdeling,
A m s t e r d a m - O.

9285

Mitteilung MR 4.

Grundsätzliche Problemen über Abründungsfehler.

durch

A. van Wijngaarden.

Vortrag für die Darmstädter GaMM-Tagung.
April 1950.

Die hier im Auszug gegebene Theorie ist entwickelt in Zusammenarbeit mit meinen vormaligen Mitarbeiter W.L. Scheen.

Die Arbeit des numerischen Rechners ist bekanntlich ein fortwährender Kampf gegen Fehler. Sind auch eine Anzahl dieser Fehler, die eigentliche Irrtümer, durch Sorgfalt und Kontrollen zu vermeiden, es bleiben dennoch zwei Fehlerquellen immer da, nämlich die Vernachlässigung von Resttermen in Reihenentwicklungen und das Abründen der Zahlen. Die erstgenannten Fehler kann man manchmal genau abschatzen aber die zweitgenannten Fehler sind besonders hinderlich durch ihre scheinbare Gesetzlosigkeit wodurch ihrer Einflusz auf das Endergebnis der Rechnung schwer abzuschätzen ist. Meistens ist es wohl möglich eine obere und untere Schranke zu finden aber meistens ist eine derartige Abschätzung viel zu grob um nützlich zu sein. Dies ist besonders der Fall wenn es sich handelt um Operationen an grosze Reihen von Zahlen, für welche es manchmal ganz "unwahrscheinlich" ist, dasz die darin vorhandene Abründungsfehler so beschaffen sein wären, dasz der Fehler im Endergebnis auch nur von derselben Gröszenordnung wäre als der maximal "mögliche" Fehler. Unter gewisse Umstände können jedoch Aussagen gemacht werden über die Frequenz des Auftretens eines Fehlers von vorgegebener Grösze und es sind solche Aussagen welche näher betrachtet werden sollen.

Einfachheitshalber betrachten wir jede abgerundete Zahl als ganz, was durch passende Definition von Funktionen immer erreicht werden kann. Unter Einführung von dem zum Einheitsoperator komplementären Abründungsoperator A und Bruchteiloperator B sind der abgeründete Wert Af und der Bruchteil Bf einer reellen Zahl f definiert durch

$$Af + Bf = f,$$
$$Af = 0 \pmod 1, \quad |Bf| < \tfrac{1}{2} \text{ wenn } f \neq \tfrac{1}{2} \pmod 1,$$
$$Af = 0 \pmod 2, \quad |Bf| = \tfrac{1}{2} \text{ wenn } f = \tfrac{1}{2} \pmod 1.$$

Diese Definition ist symmetrisch in so weit als a priori gleich oft nach oben und nach unten abgerundet wird.

Sei gegeben eine unendliche Folge Z von Zahlen $f_k (k = 1,2,\dots)$ und sei gefragt eine Funktion f von n auffolgenden Glieder von Z zu bestimmen, also $f = F(f_j, \dots f_{j+n-1})$. Sind nur Af_k gegeben, so kann man nicht besser tun als mit den Af_k statt mit den f_k in F ein zu gehen. Aber auch die Funktion F wird im allgemeinen Konstanten enthalten, welche nur mit ihrem abgerundeten Wert gebraucht werden können. Somit kann nur mit einer "abgeründeten Funktion" AF auf die Af_k operiert werden. Das Ergebnis sei $g = (AF)(Af_j, \dots Af_{j+n-1})$, und die Diskrepanz $f - g = \psi$. Sei die Rechnung wiederholt für N auffolgende Werte von j und dann bestimmt die Anzahl $M(N,\xi)$ der Fälle dasz $\psi \leq \xi$. Definitionsgemäsz ist dann die Verteilung $v(\psi) = \lim M(N,\psi)/N$. Die Frequenzdichte sei $w(\psi) = dv(\psi)d\psi$ und das Verteilungsintegral

Manchmal ist nicht f sondern Af zu bestimmen und die Rechnung leistet $Ag = A(AF)(Af_j, \ldots Af_{j+n-1})$. Die Diskrepanz sei $P = Af - Ag$, eine ganze Zahl. Sei $M'(N,Q)$ nun die Anzahl der Fälle unter N wo $P = Q$, dann ist definitionsgemäsz die Frequenz $\Omega(P) = \lim M'(N,P)/N$.

Für die Bestimmung dieser Frequenzgröszen ist es notwendig, neben natürlich Kenntnis der Funktion F, etwas zu wissen über das Verhalten der Bf_k. Sei unter N auffolgende f_k von Z die Anzahl der Glieder für welche $-\frac{1}{2} \leq Bf_k \leq \xi \leq \frac{1}{2}$ gleich $M''(N,\xi)$. Gilt nun $\lim M''(N,\xi)/N = \xi + \frac{1}{2}$ für $|\xi| \leq \frac{1}{2}$, dann sind definitionsgemäsz die Bruchteile Bf_k oder auch Z gleichverteilt. Sei weiter Z' eine Teilfolge von Z, welche alle Glieder f_k von Z enthält, für welche in Z die m forhergehende Glieder f_{k-j} $(j = 1,2 \ldots m)$ mit vorgegebenen Konstanten a_j, b_j die Gleichungen $-\frac{1}{2} \leq a_j \leq Bf_{k-j} \leq b_j \leq \frac{1}{2}$ genügen. Wenn die Teilfolge Z' gleichverteilt ist für jede Wahl der a_j und b_j so ist definitionsgemäsz das Bruchteil Bf_k unabhängig von seinen m Vorgängern oder auch m+1 auffolgende Bruchteile Bf_k sind unabhängig.

Est ist im allgemeinen auszerordentlich schwierig fest zu stellen ob eine Folge gleichverteilt ist oder nicht. Es ist bekannt z.B. wenn f_k die Funktionswerte für $x = kh$ sind eines Polynoms in x mit wenigstens einen irrationalen Koeffizient, die Folge gleichverteilt ist, aber es ist weit davon dasz für eine willkürliche Funktion eine Aussage gemacht kan werden. Die Frage nach der Abhängigkeit ist desto mehr unlöslich im allgemeinen. Uber den einzigen uns interessierenden Fall dasz die f_k Tafelwerte für $x = k\,h$ einer mehrfach differenzierbaren reellen Funktion sind, kann jedoch etwas gesagt werden. Für eine n-fach kontinu differenzierbare reelle Funktion gilt dann ja doch $\Delta_1^n f = h^n f^{(n)}(\xi)$ mit $x_1 \leq \xi \leq x_{n+1}$. Ist $f^{(n)}(\xi)$ beschränkt im Bereich der Tafel, dann kann also h so klein genommen werden, dass $\Delta_1^n f$ willkürlich klein ist. Fast jede Tafel einer Funktion im Praxis ist mit so kleinem Argumentschritt Ausgeführt dasz wenigstens eine Anzahl von Differenzen (der nicht abgerundeten Funktion) von meistens ganz niedriger Ordnung praktisch verschwinden. Sei allgemeiner die Differenz von der niedrigsten Ordnung deren Bruchteil sich praktisch nicht ändert im den ganzen betrachteten Bereich, die kritische Differenz genannt, $\Delta^m f$. Nun ist

$$\Delta^m Bf_1 = \Delta^m f_1 - \Delta^m Af_1 = \sum_{k=1}^{m+1} (-1)^{m-k+1} \binom{m}{k-1} Bf_k ,$$

also

$$Bf_{m+1} = B(B\,\Delta^m f_1 - \sum_{k=1}^{m} a_k\, Bf_k), \text{ mit } a_k = (-1)^{m-k+1}\binom{m}{k-1} .$$

Weil $B\,\Delta^m f_1$ nun praktisch eine Konstante ist, so ist offenbar Bf_{m+1} abhängig von seinen m Vorgängern.

Nun ist es nicht so schlimm dass die Frage nach Gleichverteilung und Unabhängigkeit nicht oder schwer zu beantworten ist. Es stehen ja doch nimmer unendliche Folgen f_k zur Verfügung und praktisch kommt es nur darauf an ein ungefähr richtiges Ergebnis zu versichern. Es genügt dann auch zu wissen ob die Gleichverteilung und in wie weit die Unabhängigkeit zutrifft. Nun stellt es sich heraus dass in vielen Fällen die Gleichverteilung ganz gut zutrifft und ebenso die Unabhängigkeit von nicht mehr als m Bruchteilen wo m wieder die Ordnung der kritischen Differenz ist. Natürlich sollen Polynomen mit rationalen Koeffizienten und derartigen Funktionen dann beiseite gelassen werden müssen, ganz abgesehen noch von Tafeln von Primzahlen u.s.w. Im folgenden ist nun vorausgesetzt dass es sich nur handelt um gleichverteilte Funktionen, dass m auffolgende Bruchteile unabhängig sind und dass für mehr als m Bruchteile Abhängigkeit besteht und zwar im funktionalen Sinne, dass nämlich ein Bruchteil durch seine m Vorgängern praktisch vollständig festgelegt wird. Ob eine bestimmte Tafel diesen Förderungen genügt wird weiter nicht mehr berücksichtigt. Eine weitere Beschränkung sei dass als aus den f_k gebildeten Funktionen $F(f_1 \ldots f_n)$ nur homogene lineare Funktionen mit exakt zu verwenden Koeffizienten zugelassen werden. Dies findet seine Berechtigung erstens hierin dass manche wichtige Operationen wie Differenzenbildung, Summation, Interpolation, Subtabellierung, numerische Differentiation und Integration sich auf solche Linearkombinationen zurückführen lassen, weiter weil teilweise ganz merkwürdige und unerwartete Resultate gefunden werden und letztens weil die Behandlung auch nun schon recht kompliziert ist.

Die Bestimmung von $v(\psi)$ und $\Omega(P)$ sei nun kurz skizziert. Erstens sei der Fall betrachtet, dass $n \leq m$, dass also alle in der Linearkombination begriffene Bruchteile unabhängig sind. Sei $f = \sum_{k=1}^{n} a_k f_k$, $g = \sum_{k=1}^{n} a_k A f_k$ und $\psi = f-g$ die zu untersuchen Diskrepanz. Unter Vorübergehung der vorhergehende Rechnung findet man die zweiseitig Laplace-transformierte der Frequenzdichte $w(\psi)$:

$$L_{II} w(\psi) = \int_{-\infty}^{\infty} e^{-p} w(\psi) d\psi = \prod_{k=1}^{n} \frac{e^{pa_k/2} - e^{-pa_k/2}}{pa_k}$$

Die Bestimmung der $w(\psi)$ kann hieraus auf zwei Wege erfolgen. Für grosse n ist es angebracht $w(\psi)$ zu entwickeln nach Hermiteschen Funktionen, wobei die Entwicklungskoeffizienten besonders einfach aus der Laplace-transformierten abgeleitet werden können. Unter gewisse Bedingungen, welche z.B. erfüllt sind, wenn die a_k Binomialkoeffizienten $\binom{n}{k}$ sind, findet man selbst nach n asymptotische Entwicklungen.

Für kleinere n kann man besser in geschlossenem Forme zurücktransformieren. Unter Benützung der 2^n Grössen $\lambda_j = \sum_{k=1}^{n} \pm a_k/2$ und des Symbols $[x]^n$, das für $x \geqslant 0$ gleich x^n und für $x < 0$ gleich null sein soll, findet mann

$$w(\psi) = \frac{1}{(n-1)! \prod_k a_k} \sum_{\lambda_j} \pm [\psi + \lambda_j]^{n-1},$$

wo in den Summanden das + resp. - Zeichen verwendet wird je nachdem in der Berechnung der betreffenden λ_j ein gerades oder ungerades Anzahl-Zeichen verwendet worden ist.

Um die Verteilung $v(\psi)$ resp. das Verteilungsintegral $u(\psi)$ zu bestimmen hat man im obigen Formel nur n-1 zu verwandeln in n resp. n+1.

Jetzt sei zu untersuchen die Diskrepanz P = Af - Ag. Diese ist viel verwickelter. Sind alle a_k rational, so seien sie reduziert zu Brüchen mit dem kleinst möglichen gemeinschaftlichen Nenner q. Ist mindestens eine a_k irrational, so ist $q = \infty$ zu setzen. Sind weiter die zentrale Differenzenoperatoren δ, δ^2 definiert durch $\delta f(x) = f(x + \frac{1}{2}) - f(x - \frac{1}{2})$ und $\delta^2 f(x) = f(x+1) - 2f(x) + f(x-1)$ und sind B_{2k} und $B_{2k}(x)$ Bernoullische Zahlen und Polynomen, dann gibt es unter der Voraussetzung, dass die Af_k alle ganzen Werte mit gleichen Frequenz annehmen:

1) $q = 0 \pmod 2$

a) alle $q\lambda_j = 0 \pmod 1$ und $n = 2p$ oder $2p + 1$, oder alle $q\lambda_j = \frac{1}{2} \pmod 1$ und $n = 2p$:

$$\Omega(P) = \delta^2 \sum_{k=0}^{p} \frac{B_{2k}}{(2k)!\, q^{2k}}\, u^{(2k)}(P),$$

b) alle $q\lambda_j = \frac{1}{2} \pmod 1$ und $n = 2p+1$:

$$\Omega(P) = \delta^2 \left\{ \sum_{k=0}^{p} \frac{B_{2k}}{(2k)!\, q^{2k}}\, u^{(2k)}(P) - \frac{B_{2p+2}(\frac{1}{2})}{(2p+2)!\, q^{2m+2}}\, u^{(2p+2)}(P) \right\}.$$

2) $q = 1 \pmod 2$

a) alle $q\lambda_j = 0 \pmod 1$ und $n=2p$ oder $2p+1$, oder alle $q\lambda_j = \frac{1}{2} \pmod 1$ und $n=2p$:

$$\Omega(P) = \delta^2 \sum_{k=0}^{p} \frac{B_{2k}(\frac{1}{2})}{(2k)!\, q^{2k}}\, u^{(2k)}(P),$$

b) alle $q\lambda_j = \frac{1}{2} \pmod 1$ und $n = 2p + 1$:

$$\Omega(P) = \delta^2 \left\{ \sum_{k=0}^{p} \frac{B_{2k}(\frac{1}{2})}{(2k)!\, q^{2k}}\, u^{(2k)}(P) - \frac{B_{2p+2}}{(2p+2)!\, q^{2p+2}}\, u^{(2p+2)}(P) \right\}.$$

3) $q = \infty$

$$\Omega(P) = \delta^2 u(P)$$

Eine andere meistens weniger nützliche Form des selben Ergebnisses ist:

1) $q = 0 \pmod 2$

$$\Omega(P) = \frac{1}{q}\delta^2\left\{\sum_{k=-\infty}^{qP} v\left(\frac{k}{q}\right) - \tfrac{1}{2}v(P)\right\},$$

2) $q = 0 \pmod 1$

$$\Omega(P) = \frac{1}{q}\delta^2 \sum_{k=-\infty}^{qP} v\left(\frac{k-\frac{1}{2}}{q}\right)$$

Jedoch ergibt sich hieraus gerade für den anderen extremen Fall, dass alle a_k ganz sind, also q=1, ein einfacher Ausdruck, nämlich $\Omega(P) = \delta v(P)$.

Wieder viel verwickelter wird die Sache wenn $n > m$, wenn also die in der Linearkombination eingehenden Bruchteile nicht mehr abhängig sind. Obwohl dennoch rein analytische Methoden das gewünschte Ergebnis erzielen können, empfehlt es sich eine geometrische Betrachtung zu benützen. Sei R_n ein n-dimensionaler Raum mit Koordinaten Bf_k. Die Grenzen $|Bf_k| = \frac{1}{2}$ definieren den Einheitswürfel W um Ursprung. Sind die Bf_k gleichverteilt und unabhängig, dann ist die Frequenzdichte eines Punktes innerhalb W einfach konstant und gleich eins. Die Gleichung $\psi = \Sigma a_k Bf_k$ definiert einen R_{n-1}, Ψ, in R_n und $v(\psi)$ ist das Volumen des Teilraumes von W begrenzt durch Ψ und $Bf_k = -\frac{1}{2}$. Die Konstanten λ_j, welche eine so hervorragende Rolle im Vorgehenden spielen, korrespondieren mit den Werten von ψ, wofür die zugehörige Ψ eine Ecke von W enthält.

Sind nun n-m Nebenbedingungen zufolge der Abhängigkeit gegeben, so korrespondieren hiermit n-m R_{n-1}'s, Φ_i, und der zur Verfügung stehende Teilraum ist nur der zu den Φ_i gemeinschaftlichen R_m, Φ. Sind übrigens die Bf_k gleichverteilt, dann ist $v(\psi)$ das Volumen von Φ begrenzt durch Ψ und $Bf_k = -\frac{1}{2}$ geteilt durch das Volumen von Φ begrenzt durch $|Bf_k| = \frac{1}{2}$.

In dieser Weise können auch Problemen mit Abhängigkeit der Bruchteile durchgerechnet werden. Obwohl manchmal durch passende Projektion und durch Benützung etwaiger Symmetrieeigenschaften der Koeffizienten a_k weitgehende Vereinfachungen in der Rechnung angebracht werden können, sind dennoch die praktische Schwierigkeiten bei grösserem n-m fast unüberwindlich.

Frequenz Ω (P) für die dritten Differenz.

P	$\Delta^3 f$ superkritisch: $B\Delta f \approx$						$\Delta^3 f$ superkritisch: $B\Delta^2 f \approx$						$\Delta^3 f$ kritisch: $B\Delta^3 f \approx$						$\Delta^3 f$ subkritisch	P
	0	.1	.2	.3	.4	.5	0	.1	.2	.3	.4	.5	0	.1	.2	.3	.4	.5		
<-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	>4
-4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.0003	3
-3	0	0	0	0	0	0	0	.0025	.0100	.0225	.0367	.0417	.0185	.0135	.0095	.0064	.0040	.0023	.0226	3
-2	0	.1	.2	.3	.4	.5	.1667	.1592	.1367	.0992	.0567	.0417	.1111	.1000	.0890	.0783	.0679	.0579	.1114	2
-1	0	0	0	0	0	0	.1667	.1742	.1967	.2342	.2767	.2917	.2222	.2111	.2000	.1889	.1778	.1667	.2216	1
0	1	.7	.4	.1	0	0	.3333	.3283	.3133	.2883	.2600	.2500	.2963	.2952	.2921	.2873	.2809	.2731	.2882	0
1	0	.2	.4	.6	.4	0	.1667	.1742	.1967	.2342	.2767	.2917	.2222	.2333	.2441	.2546	.2643	.2731	.2216	-1
2	0	0	0	0	.2	.5	.1667	.1592	.1367	.0992	.0567	.0417	.1111	.1222	.1333	.1444	.1556	.1667	.1114	-2
3	0	0	0	0	0	0	0	.0025	.0100	.0225	.0367	.0417	.0185	.0246	.0317	.0397	.0484	.0579	.0226	-3
4	0	0	0	0	0	0	0	0	0	0	0	0	0	.0000	.0001	.0005	.0012	.0023	.0003	-4
>4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	<-4
	0	-.1	-.2	-.3	-.4	-.5	0	-.1	-.2	-.3	-.4	-.5	0	-.1	-.2	-.3	-.4	-.5	$\Delta^3 f$ subkritisch	P
	$B\Delta f \approx$						$B\Delta^2 f \approx$						$B\Delta^3 f \approx$							
	$\Delta^3 f$ superkritisch												$\Delta^3 f$ kritisch							

Intervalfraktion p. Wenn die Af_k in dem betrachteten Bereich genügend varieren, was fast auch immer der Fall sein wird, wird $\Omega(P)$ vorgestellt durch die verwickelten Funktionen, welche wir gegeben haben und welche den gemeinschaftlichen Nenner q enthalten. Dies hat zur Folge dass $\Omega(P)$ eine ganz pathologische Funktion von p ist. Ist p rational dann sind es auch alle a_k und q hat einen endlichen Wert und $\Omega(P)$ weicht um einen endlichen Betrag ab von $\delta^2 u(P)$, welchen Wert sie annimmt für einen willkürlich nähen irrationalen Wert von p. Diese Erscheinung tritt klar zu Tage in die folgende Tabelle, wo in die erste Zeile $\Omega(0)$ gegeben ist für den exakt-gegebenen rationalen Wert von p während in die zweite Zeile $\Omega(0)$ steht für einen infinitesimal verschiedenen irrationalen Wert von p.

p	Zweipunkts-Interpolation		Dreipunkts-Interpolation		Vierpunkts-Interpolation	
0	1,0000	0,7500	1,0000	0,7500	1,0000	0,7500
0,1	0,7722	0,7741	0,7521	0,7521	0,7642	0,7642
0,2	0,8000	0,7979	0,7583	0,7582	0,7792	0,7792
0,3	0,8119	0,8143	0,7680	0,7680	0,7942	0,7942
0,4	0,8333	0,8278	0,7810	0,7808	0,8062	0,8062
0,5	0,7500	0,8333	0,7917	0,7951	0,8103	0,8114

Speziell bei p=0 und 0,5 ist der Effekt so gross, dass man es leicht experimentell bestätigt, durch zum Beispiel eine Tafel zu interpolieren mit p = 0,01 abgerundet und genau um P zu bestimmen und die so gefundene $\Omega(0)$ zu vergleichen mit dem Wert für p = 0, der natürlich gleich eins ist. Nur soll man darauf achten das Experiment zu erstrecken über eine Anzahl Interpolationen welche mindestens von der selben Ordnung ist als der reziproke Wert der Differenz der zwei betrachteten p-Werte um die Voraussetzungen der Theorie zu genügen.

Sehr verwickelt wird die Sache sobald die kritische oder auch superkritische Differenzen im Interpolationsformel vorkommen. Noch ohne besondere Mühe ist dann der jedoch praktisch wenig interessante Fall der linearen Interpolation einer linearen Funktion zu bewaltigen. Im allgemeinen macht nun die Rationalität oder Irrationalität von p nicht mehr aus! Dagegen ist nun wichtig ob $B\Delta f$ rational ist oder nicht, während auch Af eine grosse Rolle spielt. Ist z.B. $B\Delta f$ irrational dann ist jedenfalls $\Omega(0) > \frac{1}{2}$, während bei rationalen $B\Delta f$ selbst $\Omega(0) = 0$ sein kann. Man betrachte z.B. die Funktionen $f_k = 40k$ und $f_k = 40k + 0,4$. Interpolation mit p = 0,01 gibt im ersten Falle immer ein richtiges, im zweiten Falle dagegen immer ein falsches Ergebnis!

Andere auf Linearkombinationen beruhende Prozessen wie numerische Differentiation, Summation oder Inte gration in einem beschränkten Bereich sind grundsätzlich völlig ähnlich zur Interpolation. Bei Summation oder Inte gration in einem grossen Bereich treten dagegen neue Gesichtspunkte auf. Fast immer sind dann praktisch alle Bf_k abhängig ($n \gg m$). Manchmal wird jedoch diese Abhängigkeit für das Ergebnis ohne Einfluss sein, sodass man rechnen kann mit den entwickelten Formeln für unabhängige Bf_k und zwar wird dies eintreten wenn nur n so gross ist, dass auch in jedem Folge von n Bruchteilen praktisch Gleichverteilung zutrifft. Trifft dies jedoch nicht zu dann sind grosse Abweichungen zu erwarten.