Konvergenzdiskussion und Fehlerabschätzung für die Newton'sche Methode bei Gleichungssystemen.

Autor(en): **Ostrowski, Alexander**

Objekttyp: **Article**

Zeitschrift: **Commentarii Mathematici Helvetici**

Band (Jahr): **9 (1936-1937)**

PDF erstellt am: **22.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-10171

Konvergenzdiskussion und Fehlerabschätzung für die Newton'sche Methode bei Gleichungssystemen

Von ALEXANDER OSTROWSKI, Basel

Einleitung. Um nach der Newton'schen Methode eine Lösung des Gleichungssystems

$$f(x,y)=0,\ g(x,y)=0 \tag{1}$$

zu finden, bildet man, von einer angenäherten Lösung x_0, y_0 ausgehend, eine Folge von Näherungen (x_n, y_n) mit Hilfe von Gleichungen

$$\begin{aligned} f'_x(x_n, y_n)\,(x_{n+1}-x_n)+f'_y(x_n, y_n)\,(y_{n+1}-y_n)+f(x_n, y_n)&=0\\ g'_x(x_n, y_n)\,(x_{n+1}-x_n)+g'_y(x_n, y_n)\,(y_{n+1}-y_n)+g(x_n, y_n)&=0\,. \end{aligned} \tag{2}$$

Während es im Falle einer Variabel eine Reihe von Untersuchungen über die Konvergenz und die Fehlerabschätzung bei der Newton'schen Methode gibt[1]), sind mir nur zwei Stellen in der Literatur bekannt, die darauf für den Fall der *Gleichungssysteme* eingehen, allerdings ohne die Untersuchung bis zu den für den Rechner brauchbaren Regeln durchzuführen[2]).

[1]) Zum Beispiel: *A. L. Cauchy*, Leçons sur le calcul différentiel, Paris 1829 (Note sur la détermination approximative des racines d'une équation algébrique ou transzendante), wiederabgedruckt in Oeuvres complètes (II), Tome 4, pp. 573—609.

J. B. Fourier, Analyse des équations déterminées, Paris 1831; (Ostwalds Klassiker); Nr. 127.

G. Faber, J. f. d. r. u. a. M., 138 (1910), pp. 1—21.

A. Ostrowski, Sur la convergence et l'estimation des erreurs dans quelques procédés pour la résolution des équations numériques. Erscheint in der Festschrift für *D. A. Grawe*.

[2]) *C. Runge*, Separation und Approximation der Wurzeln; I B 3a, Enzyklopädie der Mathematischen Wissenschaften, I[1] (1899), insbesondere pp. 446—448.

Fr. A. Willers, Methoden der praktischen Analysis; Berlin 1928, insbesondere pp. 176—178.

C. Runge sagt a a. O., daß das Verfahren (quadratisch) konvergiert, wenn die Ausgangsnäherung gut genug ist, ohne die hinreichende Bedingung durchzurechnen und insbesondere den Bereich abzuschätzen, in dem „alle in Betracht kommenden Werte" liegen. Ferner wird bei *Runge* angedeutet, wie man aus der Kleinheit des „Einsetzungsresultats" von (x_0, y_0) in f und g auf die Güte der Näherung rückschließen kann, wobei aber wiederum eine klare Abgrenzung der „in Betracht kommenden Werte" fehlt. Die Tendenz dieser Andeutungen wird durch das Korollar 1 zu unserem Satz I in Nr. 4 bestätigt.

Hr. *Willers* gibt a. a. O. den Weg an, auf dem hinreichende Konvergenzbedingungen berechnet werden können. Allerdings hängen die so zu berechnenden Schranken nicht nur von den ersten und zweiten, sondern auch von den *dritten* Ableitungen der Funktionen f und g ab.

Wir werden nun im folgenden zuerst die *Konvergenz* des Newton'schen Verfahrens unter den folgenden Voraussetzungen beweisen:

Es sei $\mathfrak{R}'$ ein axenparalleles, abgeschlossenes Quadrat, in dem eine Lösung (ξ, η) von (1) liegt. Es sei $\mathfrak{R}^*$ das „verdoppelte" $\mathfrak{R}'$, d. h. das axenparallele Quadrat mit dem gleichen Mittelpunkt wie $\mathfrak{R}'$, aber doppelt so großen Kanten — oder irgend ein Bereich, der dieses „verdoppelte" $\mathfrak{R}'$ enthält. f und g seien in $\mathfrak{R}^*$ mit ihren ersten und zweiten Ableitungen stetig, und die Funktionaldeterminante von f und g sei in $\mathfrak{R}^*$ von 0 verschieden.

Es sei $\mathfrak{M}$ eine obere Schranke der absoluten Beträge der drei Ableitungen *zweiter* Ordnung von f und g in $\mathfrak{R}^*$, M analog eine obere Schranke der absoluten Beträge der Ableitungen *erster* Ordnung von f und g in $\mathfrak{R}^*$, $m > 0$ endlich eine untere Schranke des absoluten Betrages der Funktionaldeterminante von f und g in $\mathfrak{R}^*$. Ferner sei

$$\delta_n = \mathrm{Max}\,(|\,\xi - x_n\,|, |\,\eta - y_n\,|). \tag{3}$$

Dann gilt für $n \geqq 0$, *wenn der Punkt* $P_0\,(x_0, y_0)$ in $\mathfrak{R}'$ *liegt und*

$$\delta_0 < \frac{m}{4\,M\,\mathfrak{M}} \tag{4}$$

ist:

$$\delta_{n+1} \leqq \frac{4\,M\,\mathfrak{M}}{m}\,\delta_n^2\,, \tag{5}$$

und das Verfahren konvergiert. (Vgl. Nr. 1.)

Man sieht, daß in diesem Falle das Verfahren „quadratisch" konvergiert, d. h. die Anzahl der richtigen Dezimalen bei den Produkten $\frac{4\,M\,\mathfrak{M}}{m}\,\xi$, $\frac{4\,M\,\mathfrak{M}}{m}\,\eta$ sich praktisch bei jedem Schritt verdoppelt.

Da man allerdings ξ und η nicht kennt, muß man, um die Konvergenz sicherzustellen, das Quadrat $\mathfrak{R}'$ so klein annehmen, daß seine Seiten kleiner als $\frac{m}{4\,M\,\mathfrak{M}}$ sind.

Für den Rechner ist freilich eine andere Fragestellung wichtiger. Er hört ja im allgemeinen mit der weiteren Rechnung auf, sobald die letzte Korrektur, also auch

$$d_n = \mathrm{Max}\,(|\,x_{n+1} - x_n\,|, |\,y_{n+1} - y_n\,|) \tag{6}$$

so klein ist, daß sie die verlangte Genauigkeit praktisch nicht mehr zu beeinflussen vermag. Es ist aber wichtig, auch in diesem Falle *genaue*

Fehlerabschätzungen zu haben, d. h. also Abschätzungen für δ_{n+1} in Abhängigkeit von d_n. *Eine solche, unter der Voraussetzung* $8M\mathfrak{M}\,\delta_n \leqq m$ *geltende, sehr gut brauchbare Abschätzung ist* (Vgl. die Formeln (2, 8), (2, 9)):

$$\delta_{n+1} \leqq \frac{48\,M\,\mathfrak{M}}{m}\,d_n^2\,. \quad — \tag{7}$$

Die oben angeführten Resultate sind allerdings nur dann anwendbar, wenn man die *Existenz* einer Lösung in $\mathfrak{R}'$ bewiesen hat, womit sich der Rechner in der Praxis wohl in den seltensten Fällen abgeben wird. Daher ist eine Formulierung von Interesse, bei der aus der relativen Kleinheit der Werte von $f(x_0, y_0)$ und $g(x_0, y_0)$ oder aus der relativen Kleinheit von d_0 direkt auf die Existenz einer Lösung in $\mathfrak{R}'$ geschlossen werden kann.

Das einfachste und brauchbarste Resultat ergibt sich mit Hilfe des Wertes von d_0. Man bilde nämlich das Quadrat

$$|x - x_0| \leqq 2d_0\,,\quad |y - y_0| \leqq 2d_0 \tag{8}$$

und bestimme in ihm obere Schranken M, $\mathfrak{M}$ für die absoluten Beträge der ersten bzw. der zweiten Ableitungen von f und g, sowie eine positive untere Schranke m für den absoluten Betrag der Funktionaldeterminante von f und g. Wenn dann

$$8\,\mathfrak{M}\,M\,d_0 < m \tag{9}$$

ist, so liegt im obigen Quadrat eine Lösung (ξ, η) von (1), gegen die die nach dem Verfahren (2) bestimmten Näherungspunkte konvergieren (Satz I).

Bei der Anwendung dieses Resultats wird unter Umständen die Ermittlung eines positiven Wertes von m Schwierigkeiten bereiten. Ist aber Δ_0 der absolute Betrag der Funktionaldeterminante im Punkte x_0, y_0, so läßt sich die obige Bedingung (9) ersetzen durch

$$16\,\mathfrak{M}\,M\,d_0 < \Delta_0\,. \tag{10}$$

Es gilt dann

$$\delta_1 < \frac{16\,M\,\mathfrak{M}}{\Delta_0}\,d_0^2 \tag{11}$$

(Satz II).

Das so formulierte Ergebnis kommt wohl den Bedürfnissen des Rechners am besten entgegen, da in ihm mit d_0 und Δ_0 sowieso nur die Größen benutzt werden, die er bei seiner Rechnung zu ermitteln hat, und sich

zugleich eine sehr brauchbare Fehlerschranke ergibt. Man wird dieses Resultat in der Praxis so anwenden, daß man, wenn es auf die Ausgangsnäherung nicht anwendbar ist, zunächst weiter rechnet, bis man zu einer Näherung kommt, auf die, als eine neue Ausgangsnäherung, dieses Resultat anwendbar wird.

Daneben ist es natürlich von Interesse, ein ähnliches Kriterium für den Fall anzugeben, wo die Werte von $f(x_0, y_0)$ und $g(x_0, y_0)$ relativ klein sind. Denn die Berechnung dieser Werte entspricht ja dem sehr natürlichen Bestreben des Rechners, die Genauigkeit der errechneten Näherungen durch „Einsetzen" nachzuprüfen. Hierüber ergibt sich aus dem Satz I ein einfaches, im Text als Korollar 1 zum Satz I formuliertes Ergebnis.

Sodann gehen wir auf eine weitere, für den Rechner wichtige Frage ein. Es wird nämlich gelegentlich empfohlen, bei der Anwendung der Näherungsformel (2) die Werte der Ableitungen von f und g nicht jedesmal neu zu berechnen, sondern bei jedem einzelnen Schritt die im Ausgangspunkt ermittelten Werte zu benutzen. Wir zeigen nun, daß auch das so abgeänderte Verfahren konvergiert, wenn der Ausgangspunkt die Lösung gut genug approximiert. Die Konvergenz ist allerdings wesentlich schwächer, da dabei die Anzahl der richtig ermittelten Dezimalen der Anzahl der Näherungsschritte proportional ist. Trotzdem ist es in den meisten Fällen am günstigsten, das Verfahren (2) mit dem so abgeänderten Verfahren zu kombinieren. Wir zeigen, daß man im allgemeinen mit einem Mindestmaß an Rechnungen auskommt, wenn man bei Anwendung der Formeln (2) die partiellen Ableitungen nur *jedes dritte Mal* neu berechnet. Wenn allerdings die Ausgangsnäherung bereits gut genug ist, und man nicht mehr als eine Versechsfachung der Anzahl der genauen Dezimalen beabsichtigt, kann durchweg mit den im Ausgangspunkt ermittelten Werten der Ableitungen gerechnet werden.

Dabei machen wir zum Vergleich der Rechenarbeit Gebrauch von einer, nur sehr konventionell aufzufassenden Einheit für die Rechenarbeit, die wir als „Horner" bezeichnen. Es ist dies die Rechenarbeit, die zur Ermittlung der Werte eines Funktionenpaares in einem vorgegebenen Punkt dient[3]). Es braucht kaum betont zu werden, daß die Benutzung einer solchen Einheit nur zum Vergleich der Rechenarbeit

[3]) Ein dem bekannten Hornerschen Schema nachgebildetes Verfahren für die Berechnung der Werte eines Polynoms in zwei Variabeln findet sich ausführlich geschildert bei *H. Scheffler*, Die Auflösung der algebraischen und transzendenten Gleichungen mit einer und mehreren Unbekannten in reellen und komplexen Zahlen nach neuen und zur praktischen Anwendung geeigneten Methoden, Braunschweig 1859, pp. 56—61.

innerhalb eines bestimmten Problems angebracht ist und auch dann nur unter bestimmten Voraussetzungen.

Im wesentlichen kann man *dann* die Rechenarbeit in Horner überschlagen, wenn man mit der Rechenmaschine rechnet. Falls es sich dagegen um mehr oder weniger gut tabulierte Funktionswerte handelt, wird man die sparsamste Anordnung für die Verwendung der besprochenen Rechenverfahren sich jedesmal den Umständen entsprechend neu zu überschlagen haben — und dasselbe gilt, wenn die Vergrößerung der Anzahl der Dezimalen in ihrer Wirkung auf die Rechenarbeit zu berücksichtigen ist.

1. Für eine Funktion $F(x, y)$ schreiben wir allgemein auch $F(Q)$, wo Q der Punkt mit den Koordinaten x, y ist. Um nun die Formeln (5) und (7) zu beweisen, nehmen wir an, es liege im Mittelpunkt eines axenparallelen, abgeschlossenen Quadrats $\mathfrak{R}$ der (x, y)-Ebene eine Lösung $P(\xi, \eta)$ des Gleichungssystems (1). Über f und g setzen wir voraus, daß sie in $\mathfrak{R}$ mit ihren ersten und zweiten Ableitungen stetig sind und daß die Funktionaldeterminante

$$\Delta(x, y) = f'_x g'_y - f'_y g'_x$$

in $\mathfrak{R}$ nicht verschwindet. Setzen wir dann allgemein:

$$u_n = \xi - x_n , \quad v_n = \eta - y_n ,$$

so folgt aus (2), wenn wir die Punkte (x_ν, y_ν) mit P_ν bezeichnen und P_n in $\mathfrak{R}$ liegt:

$$f'_x(P_n) u_{n+1} + f'_y(P_n) v_{n+1} = f(P_n) + f'_x(P_n) u_n + f'_y(P_n) v_n = A . \quad (1,1)$$

Für A aber folgt aus der bekannten Restgliedformel, wenn mit P^* ein geeigneter Punkt auf der Strecke (P, P_n) bezeichnet wird:

$$-A = f(P) - f(P_n) - f'_x(P_n) u_n - f'_y(P_n) v_n = \tfrac{1}{2} (f''_{xx}(P^*) u_n^2 + 2 f''_{xy}(P^*) u_n v_n + f''_{yy}(P^*) v_n^2).$$

Daher gilt unter Benutzung von (3)

$$|A| \leqq \tfrac{1}{2} \delta_n^2 \left(|f''_{xx}(P^*)| + 2 |f''_{xy}(P^*)| + |f''_{yy}(P^*)| \right) . \quad (1,2)$$

Analog ergibt sich aus der zweiten der Gleichungen (2)

$$g'_x(P_n) u_{n+1} + g'_y(P_n) v_{n+1} = B , \quad (1,3)$$

wo für ein P^{**} auf der Strecke (P, P_n)

$$|B| \leqq \tfrac{1}{2} \delta_n^2 \left(|g''_{xx}(P^{**})| + 2 |g''_{xy}(P^{**})| + |g''_{yy}(P^{**})| \right) \quad (1,2^0)$$

ist.

Durch Auflösung von (1,1) und (1,3) nach u_{n+1} und v_{n+1} folgt:

$$u_{n+1} = \alpha_1(P_n)A + \beta_1(P_n)B\ , \qquad v_{n+1} = \alpha_2(P_n)A + \beta_2(P_n)B, \tag{1,4}$$

wo

$$\alpha_1(Q) = \frac{g'_y(Q)}{\Delta(Q)}\ , \qquad \beta_1(Q) = \frac{-f'_y(Q)}{\Delta(Q)}$$
$$\alpha_2(Q) = \frac{-g'_x(Q)}{\Delta(Q)}\ , \qquad \beta_2(Q) = \frac{f'_x(Q)}{\Delta(Q)} \tag{1,5}$$

ist.

Es sei nun μ eine obere Schranke der absoluten Beträge aller vier Funktionen (1,5) in $\mathfrak{R}$. Offenbar kann man, wenn M eine obere Schranke für die absoluten Beträge der ersten Ableitungen von f und g in $\mathfrak{R}$ und m eine positive untere Schranke für den absoluten Betrag der Funktionaldeterminante von f und g in $\mathfrak{R}$ ist,

$$\mu = \frac{M}{m} \tag{1,6}$$

setzen. Andererseits sei

$$K = \operatorname*{Max}_{Q \prec \mathfrak{R}} \left(|f''_{xx}(Q)| + 2|f''_{xy}(Q)| + |f''_{yy}(Q)|\ ,\ |g''_{xx}(Q)| + 2|g''_{xy}(Q)| + |g''_{yy}(Q)| \right). \tag{1,61}$$

Offenbar gilt, wenn $\mathfrak{M}$ eine obere Schranke für die absoluten Beträge der zweiten Ableitungen von f und g in $\mathfrak{R}$ ist,

$$K \leqq 4\,\mathfrak{M}. \tag{1,7}$$

Dann folgt aus (1,4), (1,2), (1,2^0)

$$|u_{n+1}| \leqq \mu K \delta_n^2\ , \qquad |v_{n+1}| \leqq \mu K \delta_n^2\ ,$$

$$\delta_{n+1} \leqq \mu K \delta_n^2,\ (\mu K \delta_{n+1}) \leqq (\mu K \delta_n)^2 \leqq \cdots \leqq (\mu K \delta_0)^{2^{n+1}}\ , \tag{1,8}$$

woraus unter der Bedingung

$$\mu K \delta_0 < 1 \tag{1,9}$$

die Konvergenz des Verfahrens folgt. Denn wegen $\mu K \delta_n < 1$ ist dann sicher $\delta_{n+1} < \delta_n$, und P_{n+1} (und damit alle P_ν) liegen sicher in $\mathfrak{R}$. *Hieraus folgt aber das in der Einleitung ausgesprochene Resultat* (5) *unter der Voraussetzung* (4) *ohne weiteres.*

2. Es handelt sich nun darum, eine Abschätzung von δ_{n+1} durch d_n zu finden, also *eine Abschätzung des Fehlers*. Zu dem Zwecke gehen wir wieder unter den Voraussetzungen von Nr. 1 und der Annahme (1,9) von den Relationen (2) aus, wenden aber jetzt auf

$$-f(P_n) = f(P) - f(P_n) \quad \text{und} \quad -g(P_n) = g(P) - g(P_n)$$

den gewöhnlichen Mittelwertsatz an. Dann folgt:

$$f'_x(P_n)(x_{n+1} - x_n) + f'_y(P_n)(y_{n+1} - y_n) = f'_x(Q)u_n + f'_y(Q)v_n ,$$
$$g'_x(P_n)(x_{n+1} - x_n) + g'_y(P_n)(y_{n+1} - y_n) = g'_x(Q')u_n + g'_y(Q')v_n ,$$

wo Q, Q' auf der Strecke (P, P_n) liegen. Aus diesen beiden Gleichungen folgt aber

$$u_n = \frac{Z_1(x_{n+1} - x_n) + Z_2(y_{n+1} - y_n)}{N} , \tag{2,1}$$

wo

$$N = f'_x(Q)g'_y(Q') - f'_y(Q)g'_x(Q') ,$$
$$Z_1 = f'_x(P_n)g'_y(Q') - g'_x(P_n)f'_y(Q) , \quad Z_2 = f'_y(P_n)g'_y(Q') - g'_y(P_n)f'_y(Q)$$

gesetzt ist. Für N gilt nun

$$N - \Delta(Q) = f'_x(Q)\left(g'_y(Q') - g'_y(Q)\right) - f'_y(Q)\left(g'_x(Q') - g'_x(Q)\right)$$

und daher, wenn der Mittelwertsatz auf die Differenzen in den Klammern angewandt wird,

$$|N - \Delta(Q)| \leqq 4M\mathfrak{M}\delta_n . \tag{2,2}$$

Wenn also

$$\delta_n \leqq \frac{m}{8M\mathfrak{M}} \tag{2,3}$$

ist, gilt wegen $|\Delta(Q)| \geqq m$:

$$|N| \geqq \frac{m}{2} . \tag{2,4}$$

Hier haben M, m, $\mathfrak{M}$ für $\mathfrak{N}$ die gleiche Bedeutung wie in Nr. 1.

Ziehen wir nun auf den beiden Seiten von (2,1) $x_{n+1} - x_n$ ab, so folgt

$$u_{n+1} = \frac{Z_1 - N}{N}(x_{n+1} - x_n) + \frac{Z_2}{N}(y_{n+1} - y_n) . \tag{2,5}$$

Hier ist aber

$$Z_1 - N = g'_y(Q')\,(f'_x(P_n) - f'_x(Q)) - f'_y(Q)\,(g'_x(P_n) - g'_x(Q'))$$

und daher nach dem Mittelwertsatz

$$|\,Z_1 - N\,| \leqq 4M\,\mathfrak{M}\,\delta_n,$$

und ebenso folgt

$$|\,Z_2\,| \leqq 4M\,\mathfrak{M}\,\delta_n.$$

Daher liefert (2,5)

$$u_{n+1} \leqq \frac{16M\,\mathfrak{M}}{m}\,d_n \cdot \delta_n\ .$$

Da aus Symmetriegründen die gleiche Abschätzung für v_{n+1} gilt, folgt wegen (2,3)

$$\delta_{n+1} \leqq \frac{16M\,\mathfrak{M}}{m}\,d_n\,\delta_n \leqq 2d_n\ . \tag{2,6}$$

Wegen $|\,\delta_{n+1} - \delta_n\,| \leqq d_n$ folgt weiter

$$\delta_n \leqq 3d_n \tag{2,7}$$

und daher aus (2,6)

$$\delta_{n+1} \leqq \frac{48M\,\mathfrak{M}}{m}d_n^2\ . \tag{2,8}$$

Eine schärfere Abschätzung erhalten wir, wenn wir auf beiden Seiten von (2,8) d_n addieren. Dann ergibt sich wegen (2,6) allgemein

$$\delta_n \leqq \delta_{n+1} + d_n \leqq d_n\left(1 + \frac{48M\,\mathfrak{M}}{m}\,d_n\right),$$

$$\delta_{n+1} \leqq \frac{16M\,\mathfrak{M}}{m}d_n^2\left(1 + \frac{48M\,\mathfrak{M}}{m}d_n\right), \quad n = 0,\ 1\cdots\ . \tag{2,9}$$

Die Abschätzung (7) der Einleitung folgt unter den dort genannten Voraussetzungen aus (2,8) unmittelbar.

3. An die obigen Entwicklungen mögen einige Bemerkungen geknüpft werden.

a) Wir notieren zuerst die leicht zu beweisende Ungleichung

$$d_n - \delta_{n+1} \leqq \delta_n \leqq d_n + \delta_{n+1}\ , \tag{3,1}$$

deren rechte Seite bereits oben benutzt wurde. Da für $d_n \to 0$ wegen (2,9) $\delta_{n+1} = o(d_n)$ ist, folgt aus (3,1) unter der Voraussetzung (2,3) für $n = 0$ mit $n \to \infty$

$$\delta_n \sim d_n. \tag{3,2}$$

b) Was die Bedingung (4) anbetrifft, so kann man, wie schon in der Einleitung bemerkt wurde, ihr Erfülltsein zunächst nur dann feststellen, wenn die Kante von $\mathfrak{R}'$ kleiner als die rechte Seite von (4) ist. Es sei s die Kante von $\mathfrak{R}'$, dann ist $\delta_0 \leqq s$ und

$$\frac{4M\mathfrak{M}}{m}\delta_n \leqq \left(\frac{4M\mathfrak{M}}{m}s\right)^{2^n}. \tag{3,3}$$

Die Bedingung (2,3) ist daher bereits für δ_0 und daher für alle δ_n erfüllt, wenn

$$\theta = \frac{4M\mathfrak{M}}{m}s \leqq \tfrac{1}{2}$$

ist.

Ist aber $\theta > \frac{1}{2}$, so wird man n in (2,3) und daher in (2,6) und (2,9) der Bedingung unterwerfen, daß $\theta^{2^n} \leqq \frac{1}{2}$ ist. Einfacher ist es aber, in diesem Falle die Formel (2,4) etwas abzuschwächen. Setzen wir

$$\theta^{2^n} = \theta_n\,, \tag{3,4}$$

so folgt aus (2,2) und (3,3)

$$\begin{aligned} |\,N - \Delta(Q)\,| &\leqq \theta_n m, \\ |\,N\,| &\geqq (1-\theta_n)\,m. \end{aligned} \tag{3,5}$$

Mit dieser Abschätzung von $|\,N\,|$ wird aber (2,6) zu

$$\delta_{n+1} \leqq \frac{8M\mathfrak{M}}{(1-\theta_n)m}d_n\delta_n \leqq \frac{2\theta_n}{1-\theta_n}d_n\,. \tag{3,6}$$

Durch Addition von d_n wird hieraus

$$\delta_n \leqq \frac{1+\theta_n}{1-\theta_n}d_n\,, \tag{3,7}$$

$$\delta_{n+1} \leqq \frac{8M\mathfrak{M}}{m}\,\frac{1+\theta_n}{(1-\theta_n)^2}\,d_n^2\,. \tag{3,8}$$

c) Dem Rechner wird zumeist die Bestimmung von m Schwierigkeiten bereiten. Man kann nun m und zugleich M folgendermaßen abschätzen, wenn $\mathfrak{R}$ klein genug ist, d. h. wenn die Lösung bereits gut genug angenähert wurde.

Man berechnet ja auf jeden Fall $\Delta(P_0)$, da dies zur Bestimmung von x_1 und y_1 nötig ist. Es sei nun $|\Delta(P_0)| = \Delta_0$ gesetzt. Dann folgt

$$\Delta'_x(x, y) = f'''_{xx} g'_y + f'_x g'''_{xy} - f'''_{xy} g'_x - f'_y g'''_{xx},$$

$$|\Delta'_x(x, y)| \leqq 4 M \mathfrak{M},$$

$$|\Delta'_y(x, y)| \leqq 4 M \mathfrak{M},$$

wenn (x, y) in $\mathfrak{R}$ liegt. Es sei nun die Kante von $\mathfrak{R}' \leqq s$. Dann gilt in $\mathfrak{R}^*$

$$|\Delta(x, y) - \Delta(x_0, y_0)| \leqq 16 s M \mathfrak{M},$$

$$|\Delta(x, y)| \geqq \Delta_0 - 16 s M \mathfrak{M}. \tag{3,9}$$

Wenn daher

$$s \leqq \frac{\Delta_0}{32 M \mathfrak{M}} \tag{3,91}$$

ist, gilt

$$m \geqq \frac{\Delta_0}{2}. \tag{3,92}$$

Offenbar ergibt sich analog, wenn

$$M_0 = \mathrm{Max}\left(|f'_x(P_0)|, |f'_y(P_0)|, |g'_x(P_0)|, |g'_y(P_0)|\right)$$

gesetzt wird,

$$M \leqq M_0 + 4 s \mathfrak{M}. \tag{3,93}$$

Die Anwendbarkeit der Formel (3,92) ist aber an die Bedingung (3,91) gebunden.

d) Endlich sei noch darauf aufmerksam gemacht, daß wir in der Formulierung der Einleitung *nicht* ausdrücklich vorausgesetzt haben, $P(\xi, \eta)$ sei die einzige Lösung von (1) in $\mathfrak{R}'$. Da aber unter der Annahme (4) oder, wenn die Kante von $\mathfrak{R}'$ mit s bezeichnet wird, unter der Annahme

$$s < \frac{m}{4 M \mathfrak{M}}, \tag{3,94}$$

die Folge P_n gegen *jede* Lösung von (1) in $\mathfrak{R}'$ konvergieren muß, folgt, *daß es in* $\mathfrak{R}'$ *nur eine Lösung von* (1) *geben kann, sobald* (3,94) *erfüllt ist.*

4. In der Praxis wird man allerdings in den seltensten Fällen der Rechnung einen Existenzbeweis für eine Lösung in $\mathfrak{R}'$ vorausschicken. Man wird vielmehr, sobald die Werte von $f(P_0)$ und $g(P_0)$ klein genug sind, annehmen, daß es in der Nähe von P_0 eine Lösung von (1) gibt, und versuchen, sie als Grenze der Punktfolge P_n zu berechnen. Wir wollen nun die Frage beantworten, in welchem Falle man aus der „Kleinheit" von $f(P_0)$ und $g(P_0)$ auf die Konvergenz des Verfahrens schließen kann.

Es ist allerdings nicht praktisch, zur Beurteilung der Güte einer Näherung die Werte der Funktionen f und g zu berechnen, da man damit im allgemeinen ein Drittel der Arbeit zu leisten hat, die zur Berechnung einer weiteren, in der Regel besseren Newton'schen Näherung zu leisten ist. Bezieht man sich dagegen auf die Werte von d_0 (bzw. $d_1, d_2, \ldots$), so ergeben sich zugleich einfachere Formeln. Wir beweisen daher zuerst einen auf d_0 bezüglichen Satz. Wir wollen dabei und später von der Bezeichnung Gebrauch machen:

$$k_n = \mathrm{Max}\big(|f(P_n)|, |g(P_n)|\big). \tag{4,1}$$

I. *$f(x, y), g(x, y)$ seien zwei im Quadrat*

$$(\mathfrak{R}_0) \quad |x - x_0| \leqq 2d_0, \quad |y - y_0| \leqq 2d_0$$

mit ihren ersten und zweiten Ableitungen stetige Funktionen, wobei d_0 durch (6) *definiert ist. $\mathfrak{M}$ sei eine obere Schranke der absoluten Beträge der zweiten Ableitungen von f und g in $\mathfrak{R}_0$, M eine obere Schranke der absoluten Beträge der ersten Ableitungen von f und g in $\mathfrak{R}_0$, m endlich eine als positiv vorausgesetzte untere Schranke des absoluten Betrages der Funktionaldeterminante von f und g in $\mathfrak{R}_0$. Gilt dann*

$$8\,\mathfrak{M}\,M\,d_0 < m, \tag{4,2}$$

so liegt in $\mathfrak{R}_0$ eine und nur eine Lösung der Gleichungen (1), *gegen die die Newton'sche Punktfolge P_n, $n = 0, 1, \ldots$, von P_0 aus gebildet, konvergiert. Zugleich gilt, $|\Delta(P_n)| = \Delta_n$ gesetzt,*

$$d_{n+1} \leqq \frac{4\,\mathfrak{M}\,M}{\Delta_{n+1}}\, d_n^2\,, \tag{4,3}$$

$$d_{n+1} \leqq \frac{m}{4\,\mathfrak{M}\,M}\left(\frac{4\,\mathfrak{M}\,M}{m}\, d_0\right)^{2^{n+1}}. \tag{4,4}$$

Beweis. Die Punkte P_0, P_1 liegen sicher in $\mathfrak{R}_0$. Daher folgt aus (2) für $n=0$, 1 durch Auflösung nach $x_{n+1}-x_n$ und $y_{n+1}-y_n$ unter Benutzung der Bezeichnung (4,1)

$$d_n \leqq \frac{2M k_n}{\Delta_n} \leqq \frac{2M}{m} k_n \,. \tag{4,5}$$

Ferner folgt aus der Restgliedsformel wegen (2) für $n=0$

$$f(P_1)=\tfrac{1}{2}\left(f''_{xx}(P^*)\,(x_1-x_0)^2+2f''_{xy}(P^*)\,(x_1-x_0)\,(y_1-y_0)+f''_{yy}(P^*)\,(y_1-y_0)^2\right),$$

wo P^* auf der Strecke P_0, P_1 und daher in $\mathfrak{R}_0$ liegt. Daher gilt

$$|f(P_1)| \leqq 2\,\mathfrak{M}\,d_0^2\,,$$

und daher, da die gleiche Abschätzung für $g(P_1)$ gilt,

$$k_1 \leqq 2\,\mathfrak{M}\,d_0^2\,, \tag{4,51}$$

und folglich, wegen (4,5),

$$d_1 \leqq \frac{4\mathfrak{M}M}{\Delta_1} d_0^2 \leqq \frac{4\mathfrak{M}M}{m} d_0^2 \,. \tag{4,6}$$

Wegen (4,2) folgt aber hieraus

$$d_1 \leqq \tfrac{1}{2} d_0,$$

und daher liegt das Quadrat

$$(\mathfrak{R}_1) \quad |x-x_1| \leqq 2d_1, \quad |y-y_1| \leqq 2d_1$$

innerhalb $\mathfrak{R}_0$. In $\mathfrak{R}_1$ liegt aber der nächste Näherungspunkt P_2. Setzt man die gleiche Überlegung weiter fort, so gelangt man zu einer Folge von in einander geschachtelten Quadraten, deren Kanten gegen 0 konvergieren und in denen bzw. die Punkte P_n liegen. Daher konvergieren die Punkte P_n gegen einen Grenzpunkt $P(\xi, \eta)$, der in $\mathfrak{R}_0$ liegt und in dem f und g verschwinden müssen. Das letztere folgt unmittelbar aus der Relation

$$k_{n+1} \leqq 2\,\mathfrak{M}\,d_n^2, \tag{4,61}$$

die mit (4,51), angewandt auf d_n, identisch ist.

(4,3) ist ferner mit (4,6), angewandt auf d_n, identisch, und (4,4) folgt aus (4,3).

Wir beweisen endlich, daß in $\mathfrak{R}_0$ nur eine Lösung von (1) liegen kann. Denn es möge neben $P(\xi, \eta)$ in $\mathfrak{R}_0$ noch eine zweite Lösung $P'(\xi', \eta')$ liegen. Dann verschwinden die beiden Funktionen von t

$$f(t\xi+(1-t)\,\xi', t\eta+(1-t)\,\eta'),\ g(t\xi+(1-t)\,\xi', t\eta+(1-t)\,\eta')$$

für $t=0$ und $t=1$. Nach dem Rolle'schen Satze folgt hieraus die Existenz zweier Punkte Q, Q' auf der Strecke von P nach P', in denen

$$f_x'(Q)\ (\xi-\xi')+f_y'(Q)\ (\eta-\eta')=0$$

und

$$g_x'(Q')\,(\xi-\xi')+g_y'(Q')\,(\eta-\eta')=0$$

gilt. Das Verschwinden der Determinante dieser Gleichungen liefert

$$\Delta(Q)=\begin{vmatrix} f_x'(Q) & f_y'(Q) \\ g_x'(Q)-g_x'(Q') & g_y'(Q)-g_y'(Q') \end{vmatrix}.$$

Die Elemente der zweiten Zeile der Determinante rechts sind aber absolut $\leqq 2\,\mathfrak{M}\cdot 2d_0$, so daß die Determinante rechts absolut genommen den Betrag von $8\,\mathfrak{M}\,M d_0$ nicht überschreitet, während $|\,\Delta(Q)\,| \geqq m$ nach (4,2) größer als $8\,\mathfrak{M}\,M d_0$ ist. Daher kann P' von P nicht verschieden sein, und der Satz I ist vollständig bewiesen.

Korollar 1 zum Satz I. *Ersetzt man in den Voraussetzungen des Satzes* I *das Quadrat* $\mathfrak{R}_0$ *durch das Quadrat*

$$|\,x-x_0\,| \leqq \frac{4Mk_0}{m}\ ,\quad |\,y-y_0\,| \leqq \frac{4Mk_0}{m} \tag{4,7}$$

und die Annahme (4,2) *durch*

$$16\,\mathfrak{M}\,M^2 k_0 < m^2, \tag{4,71}$$

so bleiben die Behauptungen des Satzes I *gültig.*

In der Tat folgt aus (4,5), daß das Quadrat $\mathfrak{R}_0$ in (4,7) enthalten ist. Andererseits ist auch (4,2) zugleich mit (4,71) wegen (4,5) erfüllt.

Korollar 2 zum Satz I. *Unter den Voraussetzungen des Satzes I bzw. des Korollars* 1 *dazu gilt:*

$$\delta_n < \frac{4M\mathfrak{M}}{m}\,d_{n-1}^2\,\frac{1}{1-\left(\frac{4M\mathfrak{M}}{m}\,d_{n-1}\right)^2}\ . \tag{4,8}$$

In der Tat gilt offenbar, $\varepsilon = \frac{4M\mathfrak{M}}{m} d_{n-1}$ gesetzt,

$$\frac{4M\mathfrak{M}}{m}\delta_n \leqq \frac{4M\mathfrak{M}}{m}(d_n + d_{n+1} + \cdots) ,$$

dies ist aber wegen (4,3)

$$\leqq \varepsilon^2 + \varepsilon^4 + \varepsilon^8 + \cdots = \frac{\varepsilon^2}{1-\varepsilon^2} ,$$

woraus (4,8) unmittelbar folgt.

Es sei noch bemerkt, daß für die Gültigkeit von I und der Korollare dazu über $\Delta(x, y)$ nur $|\Delta(P_n)| \geqq m$, $n = 0, 1, \ldots$, vorausgesetzt zu werden braucht.

Wenn für den Ausgangspunkt P_0 die obigen Bedingungen nicht erfüllt sind, so wird man zunächst nach der Newton'schen Methode weitere Näherungspunkte $P_1, P_2, \ldots$ suchen. Sobald eine der Zahlen d_ν, k_ν genügend klein geworden ist, läßt sich das entsprechende Resultat anwenden und man kann zugleich die obigen Abschätzungen benutzen. Sonst aber ist der kleine Wert von k_0 oder d_0 nicht durch die Nähe einer Lösung bedingt.

5. Bei der Anwendung der in Nr. 4 bewiesenen Resultate hat man die Größenordnung der ersten und zweiten Ableitungen in einem gegebenen Gebiet zu überschlagen, was im Falle, daß f und g Polynome sind, nicht schwer ist. Dagegen ist die Abschätzung von $|\Delta(x, y)|$ *nach unten* oft sehr umständlich. Daher liegt es nahe, nur den Wert $\Delta(P_0)$, der ja sowieso berechnet werden muß, in Betracht zu ziehen und die Werte $\Delta(P_n)$ nach unten abzuschätzen.

Nun folgt, wenn M und $\mathfrak{M}$ auf der Strecke $P_n P_{n+1}$ die gleiche Bedeutung haben wie oben, ähnlich wie in Nr. 3 unter c):

$$|\Delta(P_{n+1}) - \Delta(P_n)| \leqq 4M\mathfrak{M}(|x_{n+1} - x_n| + |y_{n+1} - y_n|) \leqq 8M\mathfrak{M}d_n ,$$

und daher

$$\Delta_{n+1} \geqq \Delta_n - 8M\mathfrak{M}d_n . \tag{5,1}$$

Es möge nun in den Voraussetzungen des Satzes I die Annahme (4,2) durch

$$16\mathfrak{M}Md_0 < \Delta_0,\ 16\mathfrak{M}Md_0 \leqq \alpha\Delta_0 ,\ \alpha < 1 \tag{5,2}$$

ersetzt werden. Dann gilt wegen (5,1)

$$\Delta_1 \geqq \Delta_0 - 8\mathfrak{M}Md_0 \geqq \left(1 - \frac{\alpha}{2}\right)\Delta_0 > \frac{\Delta_0}{2} . \tag{5,3}$$

Andererseits folgt wie bei (4,5)

$$d_n \leqq \frac{2 M k_n}{\Delta_n} \quad , \quad (n = 0,1)$$

und daher insbesondere wegen (4,51)

$$d_1 \leqq \frac{2 M k_1}{\Delta_1} \leqq \frac{4 \mathfrak{M} M d_0^2}{\Delta_1} \quad , \tag{5,31}$$

$$\frac{16 \mathfrak{M} M d_1}{\Delta_1} \leqq \left(\frac{16 \mathfrak{M} M d_0}{\Delta_0}\right)^2 \left(\frac{\Delta_0}{2 \Delta_1}\right)^2 < \left(\frac{16 \mathfrak{M} M d_0}{\Delta_0}\right)^2 \leqq \alpha^2 \quad , \tag{5,4}$$

$$16 \mathfrak{M} M d_1 < \Delta_1 \quad ,$$

$$2 d_1 \leqq \frac{16 \mathfrak{M} M d_0}{\Delta_0} \frac{\Delta_0}{2 \Delta_1} d_0 < d_0 \quad .$$

Daher liegt das Quadrat $\mathfrak{R}_1$ innerhalb $\mathfrak{R}_0$ und für den Punkt P_1 gelten die analogen Annahmen wie für P_0, so daß sich unsere Überlegung ad infinitum fortsetzen läßt. Daher konvergieren die Punkte P_ν gegen einen Punkt (ξ, η), der innerhalb $\mathfrak{R}_0$ liegt und eine Lösung von (1) darstellt.

Zugleich gilt, wenn die Formeln (5,31) für d_{n-1} und d_n, (5,4) für Δ_{n-1} und Δ_n benutzt werden:

$$\delta_n \leqq 2 d_n \leqq \frac{8 \mathfrak{M} M d_{n-1}^2}{\Delta_n} \quad , \tag{5,51}$$

$$\delta_n \leqq \frac{16 \mathfrak{M} M d_{n-1}^2}{\Delta_{n-1}} \frac{\Delta_{n-1}}{2 \Delta_n} < \frac{16 \mathfrak{M} M}{\Delta_{n-1}} d_{n-1}^2 \quad . \tag{5,52}$$

Für eine Überschlagsbetrachtung ist es allerdings wichtig, in diesen Formeln anstatt Δ_{n-1} und Δ_n einen von Δ_0 abhängigen Ausdruck einzuführen.

Nun muß α wegen (5,4) für Δ_n durch α^{2^n} ersetzt werden. Daher folgt durch sukzessive Anwendung von (5,3)

$$\Delta_n \geqq \left(1 - \frac{\alpha}{2}\right) \left(1 - \frac{\alpha^2}{2}\right) \left(1 - \frac{\alpha^4}{2}\right) \cdots \left(1 - \frac{\alpha^{2^{n-1}}}{2}\right) \Delta_0 \quad ,$$

und dies ist, wie man durch vollständige Induktion leicht beweist, für $0<\alpha<1$ größer als $\left(1-\alpha+\frac{\alpha}{2^n}\right)\Delta_0$. Daraus folgt aber

$$\Delta_n > (1-\alpha)\,\Delta_0\,, \quad [4] \tag{5,53}$$

und daher nach (5,51)

$$\delta_n \leqq \frac{1}{1-\alpha}\,\frac{8\mathfrak{M}M}{\Delta_0}d_{n-1}^2\,. \tag{5,6}$$

Für δ_1 ergibt sich aus (5,51) und (5,1) insbesondere die schärfere Relation

$$\delta_1 \leqq \frac{1}{1-\frac{\alpha}{2}}\,\frac{8\mathfrak{M}M}{\Delta_0}d_0^2 < \frac{16\mathfrak{M}Md_0^2}{\Delta_0}\,. \tag{5,7}$$

Zusammenfassend erhalten wir als unser wichtigstes Ergebnis:

II. *$f(x, y), g(x, y)$ mögen für ein $d_0>0$ im Quadrat*

$$(\mathfrak{R}) \quad |x-x_0|<2d_0,\ |y-y_0|<2d_0,$$

mit den ersten und zweiten Ableitungen stetig sein. Es sei

$$\Delta_0 = |\Delta(x_0, y_0)| = \left|\frac{\partial(f(x_0, y_0),\ g(x_0, y_0))}{\partial(x_0, y_0)}\right| > 0\,,$$

und es seien M, $\mathfrak{M}$ positive Zahlen mit (5,2). Sind dann in $\mathfrak{R}$ die ersten Ableitungen von f und g absolut $\leqq M$ und die zweiten Ableitungen absolut $\leqq \mathfrak{M}$ und gilt für die vermöge der Formeln (2) berechneten x_1-x_0, y_1-y_0

$$|x_1-x_0|\leqq d_0,\ |y_1-y_0|\leqq d_0,$$

so liegt in $\mathfrak{R}$ eine Lösung (ξ, η) von (1), gegen die die Folge der vermöge (2) berechneten Punkte $P_n(x_n, y_n)$ konvergiert. Zugleich gelten unter Benutzung der Bezeichnungen (3), (6) und $|\Delta(P_n)| = \Delta_n$ die Abschätzungen (5,52), (5,6), (5,7), wo sicher für $n = 0, 1, \ldots$

---

[4]) Man kann zum Beweis von (5,53) statt der erwähnten Ungleichung auch die folgende Identität benutzen:

$$\left(1-\alpha^{2^n}\right)\prod_{\nu=0}^{n-1}\left(1-\frac{\alpha^{2^\nu}}{2}\right) = (1-\alpha)\prod_{\nu=0}^{n-1}\left(1+\frac{\alpha^{2^\nu}(1-\alpha^{2^\nu})}{2}\right). \tag{5,531}$$

$$d_n \leqq \frac{\Delta_n}{16 M \mathfrak{M}} \left(\frac{16 M \mathfrak{M}}{\Delta_0} d_0 \right)^{2^n} = \frac{\Delta_n}{16 M \mathfrak{M}} \alpha^{2^n} \tag{5,8}$$

$$d_{n+1} \leqq \frac{8 M \mathfrak{M}}{\Delta_n} d_n^2 \leqq \frac{\Delta_n}{32 \mathfrak{M} M} \alpha^{2^{n+1}} . \tag{5,9}$$

gilt.

In der Tat folgt (5,8) aus (5,4) durch wiederholte Anwendung unmittelbar. Um aber (5,9) zu erhalten, beachte man, daß (5,31), für d_n und d_{n+1} geschrieben,

$$d_{n+1} \leqq \frac{4 \mathfrak{M} M d_n^2}{\Delta_{n+1}} = \frac{8 \mathfrak{M} M d_n^2}{\Delta_n} \frac{\Delta_n}{2 \Delta_{n+1}} < \frac{8 \mathfrak{M} M d_n^2}{\Delta_n}$$

liefert, wegen der aus (5,3), für Δ_n und Δ_{n+1} geschrieben, folgenden Relation $2\Delta_{n+1} > \Delta_n$.

6. Bei der zahlenmäßigen Berechnung der sukzessiven Näherungspunkte P_ν besteht in der Regel der Hauptteil der Rechenarbeit in der Berechnung der Werte von

$$f(P_n) , \; g(P_n) , \; f'_x(P_n) , \; g'_x(P_n) , \; f'_y(P_n) , \; g'_y(P_n) . \tag{6,1}$$

Die daneben weiter auszuführenden Rechnungen sind, wenigstens beim Gebrauch einer Rechenmaschine oder der Logarithmen, zu vernachlässigen. Trifft dies nun für das Gleichungssystem (1) zu, so kann man durch eine leichte Abänderung der Gleichungen (2) eine nicht unwesentliche Ersparnis in der Rechenarbeit erzielen.

Wir wollen im folgenden als eine, natürlich sehr konventionelle Einheit für Rechenarbeit das Berechnen der Werte eines Funktionenpaares in einem gegebenen Punkt ansehen und dafür die Bezeichnung: ein „Horner" benutzen[5]). Dann verlangt der Übergang von P_n zu P_{n+1} vermöge der

[5]) Da man im Falle, daß f und g Polynome sind, im allgemeinen gleich leicht bzw. gleich schwer die Werte dieser Polynome und ihrer Ableitungen berechnet, brauchen die für die Spezifizierung eines Horner zugrunde gelegten Polynome nicht genauer angegeben zu werden. Anders kann es natürlich sein, wenn f und g keine Polynome sind. Unter Umständen wird man z. B. die Werte der Ableitungen viel leichter berechnen können als diejenigen der Funktionen, und dann ist das Verfahren (2) in seiner ursprünglichen Gestalt unbedingt vorzuziehen. Ferner wird bei der obigen Definition eines Horner die Anzahl der Dezimalstellen nicht weiter festgelegt. Es ist dies bei Rechnungen mit der Rechenmaschine nicht sehr wesentlich, wird aber wichtig, wenn die Anzahl der Dezimalen über diejenige der benutzten Rechenmaschine hinausgeht. Diese Vorbehalte darf man unter keinen Umständen aus den Augen verlieren, wenn man einen Überschlag mit der Anzahl der Horner macht.

Gleichungen (2) drei Horner. Da aber, wenn sich die Punkte P_n nicht mehr sehr stark vom Grenzpunkt unterscheiden, auch die Variationen der Werte der vier Ableitungen

$$f'_x(P_n)\,,\ f'_y(P_n)\,,\ g'_x(P_n)\,,\ g'_y(P_n) \tag{6,2}$$

nur sehr klein sind, liegt es nahe, von einem Näherungspunkt an weiterhin mit festen, diesem Näherungspunkt entsprechenden Werten der vier Ableitungen (6,2) zu rechnen und nur die Werte von $f(P_n)$ und $g(P_n)$ bei jedem Schritt weiter auszurechnen.

Dieses Verfahren läuft, wenn P_0 als der Ausgangspunkt angesehen wird, auf die Benutzung des Gleichungssystems

$$\begin{aligned} f'_x(P_0)\,(x_{n+1}-x_n) + f'_y(P_0)\,(y_{n+1}-y_n) + f(P_n) &= 0 \\ g'_x(P_0)\,(x_{n+1}-x_n) + g'_y(P_0)\,(y_{n+1}-y_n) + g(P_n) &= 0 \end{aligned} \tag{6,3}$$

hinaus. Wir wollen daher zunächst untersuchen, wie sich die sukzessiven Näherungspunkte bei diesem Verfahren verhalten.

Zu dem Zwecke benutzen wir die gleichen Voraussetzungen und Bezeichnungen wie in der Nr. 1, sowie die Bezeichnungen (3) und (6). Dann kann die erste der Gleichungen (6,3) in der Form geschrieben werden

$$f'_x(P_0)\,(u_n-u_{n+1}) + f'_y(P_0)\,(v_n-v_{n+1}) + f(P_n) = 0$$

oder

$$f'_x(P_0)\,u_{n+1} + f'_y(P_0)\,v_{n+1} = f'_x(P_0)\,u_n + f'_y(P_0)\,v_n + f(P_n) = A'\,. \tag{6,31}$$

Für A' ergibt sich aber unter Benutzung der Bezeichnung (1,1)

$$A' = A + \left(f'_x(P_0)-f'_x(P_n)\right)\,u_n + \left(f'_y(P_0)-f'_y(P_n)\right)\,v_n = A + A_1\,.$$

Hier gilt für A die Abschätzung (1,2), wo P^* ein Punkt der Strecke (P, P_n) ist, also in $\mathfrak{R}$ liegt. Für A_1 aber ergibt sich in analoger Weise durch Anwendung des Mittelwertsatzes die Abschätzung

$$|\,A_1\,| \leqq 2\,\mathfrak{M}(|\,x_n - x_0\,| + |\,y_n - y_0\,|)\,\delta_n$$

oder, da der Klammerausdruck rechts $\leqq 2\,(\delta_0+\delta_n)$ ist,

$$|\,A_1\,| \leqq 4\,\delta_n^2\,\mathfrak{M} + 4\,\delta_0\,\delta_n\,\mathfrak{M}\,.$$

Für A aber folgt aus (1,2) und (1,7)

$$|A| \leqq 2\mathfrak{M}\delta_n^2$$

und daher

$$|A'| \leqq 6\mathfrak{M}\delta_n^2 + 4\mathfrak{M}\delta_0\delta_n .$$

Eine zu (6,31) analoge Relation ergibt sich aus der zweiten der Gleichungen (6,3). Durch die Auflösung der beiden so entstehenden Gleichungen nach u_{n+1}, v_{n+1} folgt

$$\delta_{n+1} \leqq \frac{4M\mathfrak{M}}{m}\delta_n(3\delta_n + 2\delta_0) . \tag{6,4}$$

Wir setzen nun

$$\frac{12M\mathfrak{M}}{m} = \varkappa$$

und nehmen an, δ_0 sei bereits so klein, daß

$$\varkappa\delta_0 = \frac{12M\mathfrak{M}}{m}\delta_0 < 1 \tag{6,5}$$

ist. Dann ist die Voraussetzung (4) auf jeden Fall erfüllt, so daß, für $n = 0$, wo ja (6,3) auf (2) hinausläuft, aus (5) folgt

$$\delta_1 \leqq \frac{\varkappa}{3}\delta_0^2 < \frac{\delta_0}{3} . \tag{6,50}$$

Ist nun, für $\nu = 1, 2, \ldots, n$,

$$\delta_n < \delta_{n-1} < \delta_{n-2} < \cdots < \delta_1 < \frac{\delta_0}{3} , \tag{6,51}$$

so folgt aus (6,4)

$$\delta_{n+1} < \varkappa\delta_0\delta_n , \quad n = 1, 2, \ldots , \tag{6,6}$$

und daher $\delta_{n+1} < \delta_n$, so daß (6,51) allgemein gilt.

Aus (6,6) folgt nun, wegen (6,5), daß δ_n für $n \to \infty$ gegen 0 strebt, so daß also unter der Voraussetzung (6,5) auch das vereinfachte Verfahren (6,3) eine gegen die in $\mathfrak{R}$ liegende Lösung von (1) konvergierende Punktfolge liefert. Dagegen ist die Konvergenz im allgemeinen, wie man sagt, „linear", so daß die Anzahl der sicheren Dezimalen proportional der Anzahl der Näherungsschritte ist. Gegen diese Verschlechterung der Konvergenz fällt dann natürlich die Verminderung der Rechenarbeit auf ein Drittel bei jedem auf den ersten folgenden Schritt *auf die Dauer* nicht mehr ins Gewicht.

7. Trotzdem ist es aber möglich, durch Kombination der Ansätze (2) und (6,3) zu einer Vereinfachung zu kommen. Nehmen wir an, wir berechnen, von P_n ausgehend, wofür die Relation

$$\varkappa\, \delta_n < 1$$

gilt, den nächsten Näherungspunkt vermöge der Gleichungen (2), von da an aber die weiteren Näherungspunkte $P_{n+\nu}$ vermöge der Gleichungen

$$f'_x(P_n)\,(x_{n+\nu} - x_{n+\nu-1}) + f'_y(P_n)\,(y_{n+\nu} - y_{n+\nu-1}) + f(P_{n+\nu-1}) = 0$$

$$g'_x(P_n)\,(x_{n+\nu} - x_{n+\nu-1}) + g'_y(P_n)\,(y_{n+\nu} - y_{n+\nu-1}) + g(P_{n+\nu-1}) = 0\ .$$

Dann ergibt sich aus den obigen Relationen (5), (6,50) und (6,6) nach der entsprechenden Änderung der Bezeichnungen

$$\delta_{n+1} \leqq \frac{\varkappa}{3}\delta_n^2\ , \quad \delta_{n+2} \leqq \frac{\varkappa^2}{3}\delta_n^3\ , \quad \delta_{n+3} \leqq \frac{\varkappa^3}{3}\delta_n^4\ , \ \ldots$$

und allgemein

$$\delta_{n+\nu} \leqq \frac{\varkappa^\nu}{3}\delta_n^{\nu+1}\ .$$

Andererseits ist die Anzahl der zur Berechnung von $P_{n+\nu}$ nötigen Horner gleich $2+\nu$. Für $\nu = 4$, d. h. nach Leistung von 6 Horner, ergibt sich dann $\frac{\varkappa^4}{3}\delta_n^5$ als Fehlerschranke. Wendet man dagegen diese 6 Horner so auf, daß man aus P_{n+1} wiederum vermöge des Ansatzes (2) die nächste Näherung berechnet, so erhält man durch zweimalige Anwendung der Formel (4) als Schranke für den Fehler

$$\frac{\varkappa}{3}\left(\frac{\varkappa}{3}\delta_n^2\right)^2 = \frac{\varkappa^3}{27}\delta_n^4\ ,$$

und dies ist von niedrigerem Grade in δ_n und auch, sobald $\varkappa\,\delta_n < \frac{1}{9}$ wird, numerisch größer als der obige Wert.

Man erhält daher auf jeden Fall mit demselben Aufwand an Rechenarbeit eine schnellere Konvergenz, wenn man nach jedesmaliger Anwendung des Ansatzes (2) *dreimal* den Ansatz (6,3) benutzt, oder kürzer gesagt, wenn man beim Übergang von P_n zu P_{n+1} die Werte der ersten Ableitungen von f und g nur bei jedem *vierten* Schritt neu berechnet.

Noch besser ist es allerdings, wenn man nach jedesmaliger Anwendung

des Ansatzes (2) den Ansatz (6,3) *zweimal* benutzt. In der Tat gelangt man auf diese Weise durch Aufwendung von 5 Horner zur Größenordnung δ_n^4, und daher durch Anwendung von $5n$ Horner zur Größenordnung $\delta_n^{4^n}$. Werden aber Ableitungen von f und g nur bei jedem vierten Schritt neu berechnet, so wird damit nach Anwendung von $6m$ Horner die Größenordnung von $\delta_n^{5^n}$ erzielt.

Will man nun z. B. 30 Horner aufwenden, so erhält man auf den beiden angegebenen Wegen, bzw. $\delta_n^{4^6}$, $\delta_n^{5^5}$ und 4^6 ist größer als 5^5.

Man könnte natürlich auch die Ableitungen von f und g bei jedem zweiten Schritt neu berechnen. Dann würde man nach Aufwendung von $4n$ Horner bis zur Größenordnung $\delta_n^{3^n}$ gelangen. Dies ist aber ungünstiger als das vorhin angegebene Verfahren. Im allgemeinen läßt sich die folgende *Faustregel* angeben:

Man soll die Ableitungen von f und g bei jedem dritten Schritt neu berechnen. Will man dann noch einen oder zwei Horner aufwenden, so ist der Ansatz (6,3) zu benutzen. Sollen dagegen 3 Horner aufgewendet werden, so ist es am günstigsten, wenn man zuletzt zweimal abwechselnd die Ansätze (2) und (6,3) benutzt. Wenn endlich noch 4 Horner aufzuwenden sind, so ist zuerst vom Ansatz (2) und sodann vom Ansatz (6,3) Gebrauch zu machen.

Auf die Begründung dieser Regel gehen wir nicht ein, da ja wohl der Rechner sowieso in jedem einzelnen Falle sich die günstigste Kombination aus den obigen Angaben zurecht legen kann.

Im übrigen überlegt man leicht, daß wenn man nicht *mehr* will, als die gegebene Anzahl der richtigen Dezimalstellen zu versechsfachen, man am einfachsten zunächst einmal den Ansatz (2) und sodann hinreichend oft den Ansatz (6,3) anwenden soll.

8. Wir wollen nun die obigen Ergebnisse an einigen Beispielen erläutern, die wir der einschlägigen Literatur entnehmen.

Unser erstes Beispiel ist dem Lehrbuch von *Whittaker* und *Robinson*[6]) entnommen. Es handelt sich um das Beispiel von 2 Gleichungen:

$$f \equiv x^3 + 2y^2 - 1 = 0$$
$$g \equiv 5y^3 + x^2 - 2xy - 4 = 0 .$$

Durch einige Versuche wird festgestellt, daß eine Wurzel ξ, η im Quadrat

[6]) *E. T. Whittaker* and *G. Robinson*, The Calculus of Observations; a Treatise on Numerical Mathematics; London 1926, pp. 88—90.

$$-0{,}650 < x < -0{,}649; \quad 0{,}798 < y < 0{,}799$$

liegt. Von den Werten

$$x_0 = -0{,}6494; \quad y_0 = 0{,}7981$$

ausgehend, werden die ersten Korrekturen bestimmt. Es ergibt sich

$$f(x_0, y_0) = 0{,}620 \cdot 10^{-4}; \; g(x_0, y_0) = 0{,}959 \cdot 10^{-4}$$

$$x_1 - x_0 = -0{,}1597 \cdot 10^{-4}; \; y_1 - y_0 = -0{,}1310 \cdot 10^{-4}.$$

Wie genau werden nun die Lösungen durch x_0 bzw. y_0 approximiert? Man erhält zunächst durch einfache Überschlagsrechnung im obigen Quadrat

$$M < 11, \; \mathfrak{M} < 24, \; m > 20.$$

Andererseits ist offenbar

$$\delta_0 < 10^{-3}, \; d_0 \leqq 1{,}6 \cdot 10^{-5}.$$

Da die Bedingung (4) offenbar erfüllt ist, ergibt sich aus (5)

$$\delta_1 \leqq \frac{4 \cdot 11 \cdot 24}{20} \cdot 10^{-6} = 52{,}8 \cdot 10^{-6},$$

so daß danach x_1 und y_1 bis auf 6 Einheiten der fünften Dezimalen genau sind.

Viel bessere Abschätzungen ergeben sich aus der Formel (2,9), da die Bedingung (2,3) hier erfüllt ist:

$$\delta_1 \leqq \frac{16 \cdot 11 \cdot 24}{20} \cdot 2{,}56 \cdot 10^{-10} \cdot \left(1 + \frac{48 \cdot 11 \cdot 24}{20} \cdot 1{,}6 \cdot 10^{-5}\right),$$

$$\delta_1 < 5{,}5 \cdot 10^{-8},$$

so daß danach x_1, y_1 bis auf 5 Einheiten der achten Dezimalen richtig sind.

Dabei haben wir als gegeben vorausgesetzt, daß im obigen Quadrat eine Lösung des Gleichungssystems enthalten ist. Dies ergibt sich aber unmittelbar aus dem Satze I oder dem Korollar 1 dazu. Denn um das Korollar 1 zum Satz I anzuwenden, beachte man, daß $k_0 < 10^{-4}$ ist. Bildet man nun mit den obigen Schranken für M und m das Quadrat (4,7) ($M = 11$, $m = 20$), so ergibt eine einfache Überschlagsrechnung,

daß in ihm die Funktionaldeterminante von f und g absolut $> 20 = m$ und die absoluten Beträge der ersten Ableitungen $< 11 = M$ sind. Die absoluten Beträge der zweiten Ableitungen sind aber in ihm kleiner als

$$24 < 10^3 < \frac{10^4 \cdot 20^2}{16 \cdot 11^2} < \frac{m^2}{16 M^2 \cdot k_0} \cdot$$

Daher liegt im Quadrat (4,7) eine Lösung der Gleichungen (1), gegen die die von x_0, y_0 aus gebildeten Näherungspunkte konvergieren und für die dann die obigen Abschätzungen gelten.

Ein besseres Resultat ergibt sich mit Hilfe des Satzes I. Denn danach haben wir nur das axenparallele Quadrat zu betrachten mit dem Mittelpunkt x_0, y_0, dessen Kantenlänge $4d_0 < 4 \cdot 1{,}6 \cdot 10^{-5} < 6{,}4 \cdot 10^{-5}$ ist. Wegen

$$\mathfrak{M} < 24 < 10^3 < \frac{m}{8 M d_0}$$

liegt nach dem Satze I im Quadrat $\mathfrak{R}_0$ eine Lösung der Gleichungen (1).

Das nächste Beispiel sei einem Werk von *Runge* und *König*[7]) entnommen. Das Gleichungssystem

$$f \equiv 2x^3 - y^2 - 1 = 0\,, \quad g \equiv xy^3 - y - 4 = 0$$

hat eine Lösung in der Nähe von P_0: $x_0 = 1{,}2$; $y_0 = 1{,}7$. Durch Einsetzen dieser Werte ergibt sich

$$f(P_0) = -0{,}434 \quad ; \quad g(P_0) = 0{,}1956.$$

Durch die Anwendung des Newton'schen Verfahrens folgt

$$x_1 - x_0 = 0{,}0349 \quad ; \quad y_1 - y_0 = -0{,}0390\,.$$

Für den neuen Näherungspunkt P_1 gilt

$$f(P_1) = 74{,}70 \cdot 10^{-4} \quad ; \quad g(P_1) = -19{,}87 \cdot 10^{-4}.$$

Und von hier aus ergibt sich die zweite Korrektur

$$x_2 - x_1 = -6{,}253 \cdot 10^{-4} \quad ; \quad y_2 - y_1 = 5{,}263 \cdot 10^{-4},$$

$$f(P_2) = 252{,}8 \cdot 10^{-8} \quad ; \quad g(P_2) = -54{,}8 \cdot 10^{-8},$$

$$x_3 - x_2 = -215{,}7 \cdot 10^{-8} \quad ; \quad y_3 - y_2 = 166{,}6 \cdot 10^{-8}.$$

---

[7]) *C. Runge* und *H. König*, Vorlesungen über numerisches Rechnen; Berlin 1924, pp. 178—179.

Um die Genauigkeit der entstehenden Näherungen beurteilen zu können, überschlagen wir zunächst im Quadrat

$$(\mathfrak{R}^*) \quad 1 < x < 1{,}4; \quad 1{,}5 < y < 1{,}9$$

obere Schranken M, $\mathfrak{M}$ für die ersten und zweiten Ableitungen von f und g, sowie eine untere Schranke m für die Funktionaldeterminante von f und g. Es ergibt sich z. B.

$$M = 16; \quad \mathfrak{M} = 19; \quad m = 38.$$

Es handelt sich zunächst darum, festzustellen, ob es in der Nähe von P_0 wirklich eine Lösung des Gleichungssystems gibt. Indessen reichen hierzu der Satz I und das Korollar 1 dazu, auf P_0 angewandt, nicht aus, da weder $k_0 = 0{,}434$ noch $d_0 = 0{,}039$ hinreichend klein sind. Wir gehen daher vom nächsten Näherungspunkt P_1 aus. Dafür gilt

$$k_1 < 76 \cdot 10^{-4} \quad ; \quad d_1 < 7{,}6 \cdot 10^{-4}.$$

Nunmehr ist der Satz I sowie das Korollar 1 dazu anwendbar. Denn für das Korollar hat man mit den obigen Werten von M und m das Quadrat (4,7) zu betrachten, das im folgenden Quadrat liegt:

$$(\mathfrak{R}) \quad |x - x_1| \leqq 0{,}0128, \quad |y - y_1| \leqq 0{,}0128.$$

In diesem Quadrat gelten nun die obigen Schranken und zugleich gilt in ihm offenbar

$$19 = \mathfrak{M} \leqq \mathfrak{M}^* = \frac{38^2 \cdot 10^4}{16 \cdot 16^2 \cdot 76};$$

daher liegt in $\mathfrak{R}$ eine Lösung unseres Gleichungssystems. Engere Schranken ergeben sich mit Hilfe des Satzes I. Denn dann hat man nur das Quadrat

$$(\mathfrak{R}_1) \quad |x - x_1| \leqq 2d_1 < 15{,}2 \cdot 10^{-4} \; ; \; |y - y_1| < 15{,}2 \cdot 10^{-4}$$

zu betrachten, in dem die obigen Schranken M, $\mathfrak{M}$, m gültig sind und nachzuprüfen, daß die Relation gilt:

$$19 = \mathfrak{M} < \mathfrak{M}_1^* = \frac{m}{8 M d_1} = \frac{38 \cdot 10^5}{8 \cdot 16 \cdot 76} \cdot$$

Nunmehr ergibt sich aber, da

$$\delta_1 \leqq 2d_1 < 15{,}2 \cdot 10^{-4}$$

und die Bedingung (4) für δ_1 erfüllt ist, die Abschätzung für δ_2

$$\delta_2 \leqq \frac{4M\mathfrak{M}}{m}\delta_1^2 = 32\,\delta_1^2 \leqq 10^{-4}\ ,$$

so daß x_2 und y_2 in den ersten vier Dezimalen nach dem Komma richtig sind.

In der Tat ergeben $x_3 - x_2$ und $y_3 - y_2$ nur Korrekturen, die sich auf die sechste Dezimale beziehen. Man erhält so

$$x_3 = 1{,}234272173 \quad ; \quad y_3 = 1{,}661527966.$$

Um die Genauigkeit dieser Werte abzuschätzen, benutzen wir (2,9):

$$\delta_3 \leqq \frac{16M\mathfrak{M}}{m}\,d_2^2\left(1 + \frac{48M\mathfrak{M}}{m}d_2\right) < 7 \cdot 10^{-10}\ .$$

Daher ist der Fehler bei unseren Werten kleiner als eine Einheit der neunten Dezimalen, und wir erhalten für die genauen Werte:

$$\xi = 1{,}234272173 \pm 10^{-9} \quad ; \quad \eta = 1{,}661527966 \pm 10^{-9}.$$

Basel, 19. Mai 1936.

(Eingegangen den 22. Mai 1936.)