A. M. Ostrowski
Über Fehlerabschaetzungen a priori und a posteriori

Acta Universitatis Carolinae. Mathematica et Physica, Vol. 15 (1974), No. 1-2, 111--115

Persistent URL: http://dml.cz/dmlcz/142337

1974 ACTA UNIVERSITATIS CAROLINAE - MATHEMATICA ET PHYSICA NO. 1-2 PAG .111-115

Über Fehlerabschaetzungen a priori und a posteriori

A. M. OSTROWSKI
Mathematisches Institut der Universität, Basel

1. Bei numerischer Auswertung iterativer Prozesse wird in der Regel bereits mit dem Konvergenzbeweis eine Fehlerabschätzung hergeleitet, die allerdings gewöhnlich äusserst konservativen Charakter hat und dem Rechner keine Handhabe zur Beurteilung der effektiven Geschwindigkeit des Rechenprozesses liefert. Dies ist die *Fehlerabschätzung a priori.*

Praktisch wird jedoch der Rechner anhand der im Laufe der Rechnung erhaltenen numerischen Ergebnisse sich während der Rechnung ein Urteil über die erzielte Genauigkeit zu bilden versuchen, wenn auch dieses Urteil nur zu oft darauf hinausläuft, dass eine weitere Rechnung mit der benutzten Anzahl der Dezimalstellen keine Verbesserung des Rechenresultats liefern kann und daher man sich mit der erzielten Annäherung begnügen sollte — eine theoretisch höchst bedenkliche Überlegung, wenn sie auch in der Praxis des Rechners und auch in der wissenschaftlichen Literatur viel häufiger vorkommt, als man annehmen würde.

Anderseits ist es in vielen, häufig in der Praxis vorkommenden Fällen, in der Tat möglich, aus der Kenntnis einer sehr schlechten theoretischen a priori Abschätzung des Fehlers, eine häufig vollständig ausreichende Abschätzung des wirklichen Fehlers zu erhalten, anhand der während der Rechnung erhaltenen Teilergebnisse — der *Fehler a posteriori.*

2. Ich gebe dieses Verfahren gleich im allgemeinen Fall eines metrischen Raumes mit der Distanzfunktion $|a, b|$ an.

Es sei von einer Folge $\{x_\nu\}$ bekannt, dass sie gegen ein ζ konvergiert, und es möge ferner eine rekurrente Fehlerabschätzung bekannt sein:

$$|x_{\nu+1}, \zeta| \leqslant \varphi_\nu |x_\nu, \zeta| \,. \tag{1}$$

Über die Koeffizienten φ_ν nehmen wir an, dass

$$0 < \varphi_\nu < 1 \,, \quad \prod_{\nu=1}^{\infty} \varphi_\nu = 0 \,. \tag{2}$$

Dann gelten, behaupte ich, allgemein die Ungleichungen

$$\frac{1}{1+\varphi_\nu} \leqslant \frac{|x_\nu, \zeta|}{|x_{\nu+1}, x_\nu|} \leqslant \frac{1}{1-\varphi_\nu} \,, \tag{3}$$

$$|x_{\nu+1}, \zeta| \leqslant \frac{\varphi_\nu}{1-\varphi_\nu} |x_{\nu+1}, x_\nu| \,. \tag{4}$$

3. Der Beweis ergibt sich fast unmittelbar. Es folgt, wegen der Dreiecksungleichung, aus (1)

$$|x_{\nu+1}, x_\nu| \leqslant |\zeta, x_\nu| + |\zeta, x_{\nu+1}| \leqslant (1 + \varphi_\nu)\, |\zeta, x_\nu| ,$$

$$|x_{\nu+1}, x_\nu| \geqslant |\zeta, x_\nu| - |\zeta, x_{\nu+1}| \geqslant (1 - \varphi_\nu)\, |\zeta, x_\nu| ,$$

und daraus, durch Division, (3) und ferner, unter Benützung von (1), (4).

Wenn zum Beispiel $\varphi_\nu = 1 - \frac{1}{\nu}$ ist, was an sich eine für numerische Zwecke unbrauchbar langsame Konvergenz bedeutet, würde aus (3) und (4) folgen:

$$\frac{\nu}{2\nu - 1} \leqslant \frac{|x_\nu, \zeta|}{|x_{\nu+1}, x_\nu|} \leqslant \nu ,$$

und weiter

$$|x_{\nu+1}, \zeta| \leqslant (\nu - 1)\, |x_{\nu+1}, x_\nu| ,$$

was für die Rechenpraxis ohne weiteres ausreicht, falls die Folge $\{x_\nu\}$ zum Beispiel geometrisch konvergiert.

4. Falls bereits die theoretische Abschätzung eine geometrische Konvergenz liefert, sodass $\varphi_\nu = q$, $0 < q < 1$, gesetzt werden kann, folgt insbesondere

$$\frac{1}{1 + q} \leqslant \frac{|x_\nu, \zeta|}{|x_\nu, x_{\nu+1}|} \leqslant 1 - q , \tag{5}$$

$$|x_{\nu+1} - \zeta| \leqslant \frac{q}{1 - q}\, |x_{\nu+1} - x_\nu| . \tag{6}$$

Im speziellen Fall, dass die Folge $\{x_\nu\}$ durch sukzessive Anwendung eines kontrahierenden Operators entsteht, sind die Abschätzungen (5) und (6) aus den Abschätzungen im Banach'schen Beweis seines Satzes und in den Beweisen der Verfeinerungen dieses Satzes bekannt (siehe z. B. Weissinger [6]).

5. Wir wollen nun sehen, wie sich die Anwendung der obigen Überlegungen im Falle einer Matrix-Iteration gestaltet.
Es möge eine Folge von Vektoren $\xi_\nu \in R^n$ durch homogene Matrizen-Iteration mit der konstanten Matrix A entstehen und gegen den Vektor ζ konvergieren:

$$\xi_{\nu+1} = A\xi_\nu \quad (\nu = 0, 1, \ldots) , \quad \xi_\nu \to \zeta .$$

Dann gilt offenbar $\xi_{\nu+1} - \zeta = A(\xi_\nu - \zeta)$, und daher für irgendeine Vektornorm und die dadurch induzierte Matrixnorm:

$$|\xi_{\nu+1} - \zeta| \leqslant |A|\, |\xi_\nu - \zeta| .$$

Nun ist aber für die Konvergenz der ξ_ν (für beliebige ξ_0) notwendig und hinreichend, dass *der Spektralradius von A kleiner als* 1 *ist*. Das bedeutet aber keineswegs, dass die Norm $|A| < 1$ ist, sodass hier unsere Überlegungen nicht direkt anwendbar zu sein brauchen.

6. Man kann aber natürlich in diesem Falle direkt schreiben

$$\xi_\nu - \zeta = A^\nu(\xi_0 - \zeta)\,, \tag{7}$$

und das Verhalten der Norm $|A^\nu|$ untersuchen. Hierfür gilt nun die Relation

$$|A^\nu| \sim c\nu^{k-1}\lambda_A^\nu\,, \quad c \neq 0 \ (\nu \to \infty)\,, \tag{8}$$

wo λ_A der Spektralradius von A ist und k die maximale Mehrfachheit eines Elementarteilers von A, der dem Eigenwert mit dem absoluten Betrag λ_A entspricht (siehe Ostrowski [3]). Die Anwendung dieser Formel, sogar als rein asymptotische Formel aufgefasst, verlangt aber in der Regel eingehende theoretische Untersuchung der Matrix A.

7. Man kann aber auch anders verfahren. Aus unserem Ansatz folgt, unter E die Einheitsmatrix verstanden,

$$\xi_\nu - \xi_{\nu+1} = (E - A)\,\xi_\nu\,, \quad \xi_\nu - \zeta = (E - A)^{-1}\,(\xi_\nu - \xi_{\nu+1})\,.$$

Daraus folgt, wenn wir nunmehr als Vektornorm die euklidische Länge betrachten und die dadurch induzierte Matrixnorm mit dem Index e versehen,

$$\frac{1}{|E - A|_e} \leqslant \frac{|\xi_\nu - \zeta|}{|\xi_{\nu+1} - \zeta_\nu|} \leqslant |(E - A)^{-1}|_e\,. \tag{9}$$

Man beachte, dass die rechtsseitige Schranke, falls $|A|_e < 1$ ist, sich abschätzen lässt durch $1/(1 - |A|)$, doch ist diese Annahme durchaus unnötig, wenn man die Formel (9) benutzen will, sofern es gelingt, die in ihr vorkommenden Schranken anders abzuschätzen.

8. Wir bezeichnen die Eigenwerte von A mit λ_ν und ordnen sie so, dass

$$|\lambda_1| \leqslant \ldots \leqslant |\lambda_n| =: \lambda_A\,.$$

Die praktisch bequemste Matrixnorm ist die sogenannte Frobenius-Norm,

$$|A|_F := \sqrt{\sum_{i,k} |a_{i,k}|^2} \geqslant \sum_\nu |\lambda_\nu|^2\,,$$

wo $a_{i,k}$ die Elemente von A sind. Die letzte Ungleichung wird zur Gleichung nur im Falle der *normalen Matrizen*, während sonst als „das Normalitätsmass„ die Grösse

$$\Delta_A = \sqrt{|A|_F^2 - \sum_\nu |\lambda_\nu|^2}$$

zu benutzen ist.

Gelingt es, Δ_A und λ_A abzuschätzen, so kann man die Ungleichungen benutzen:

$$|E - A|_e \leqslant \Delta_A + \lambda_A + 1\,,$$

$$|E - A|_e \leqslant |A|_e + 1 \leqslant |A|_F + 1\,.$$

Die letzte Abschätzung ist wichtig, weil $|A|_e$ sich eigentlich erst durch die Auflösung einer Gleichung n-ten Grades ergibt.

9. Für die rechtsseitige Schranke in (9) gelten die Abschätzungen

$$|(E-A)^{-1}|_e \leqslant \sqrt{e}(1+|A|_F/\sqrt{n})^{n-1}/\,|\det(E-A)|\,, \tag{10}$$

$$|(E-A)^{-1}|_e \leqslant \frac{1}{1-\Delta_A-\lambda_A} \quad (\Delta_A+\lambda_A<1)\,,$$

$$|(E-A)^{-1}|_e \leqslant \frac{1}{\Delta_A+\lambda_A-1}\left(\frac{\Delta_A}{1-\lambda_A}\right)^n \quad (\Delta_A+\lambda_A>1,\ \lambda_A<1)\,,$$

$$|(E-A)^{-1}|_e \leqslant \frac{n}{1-\lambda_A} \quad (\Delta_A+\lambda_A\leq 1)$$

$$|(E-A)^{-1}|_e \leqslant \frac{n}{1-\lambda_A}\left(\frac{\Delta_A}{1-\lambda_A}\right)^n \quad (\Delta_A+\lambda_A\leq 1,\ \lambda_A<1)\,.$$

Die obigen Abschätzungen hängen allerdings an der Möglichkeit, Δ_A abzuschätzen.

10. Nun besteht aber eine weitere Eigenschaft der normalen Matrizen A darin, dass $A^*A-AA^*=0$ ist — dadurch wurden die normalen Matrizen durch I. Schur überhaupt definiert. Es gelang nun P. Henrici die folgende Ungleichung zu erhalten, die eine „rationale„ numerische Abschätzung von Δ_A liefert:

$$\Delta_A \leqslant 4\sqrt{\frac{n^3-n}{12}}\ \sqrt{|A^*A-AA^*|_F}\,. \tag{11}$$

Zugleich wurde durch P. Eberlein und andere gezeigt, dass diese Ungleichung auch die „richtige Grössenordnung„ für Δ_A liefert.

Die Anwendung der Formel (10) verlangt allerdings die Berechnung, oder wenigstens die Abschätzung nach unten, von $|\det A|$, eine Aufgabe, die im allgemeinen Fall einen relativ grossen Rechenaufwand voraussetzt, aber in vielen Fällen sich auch theoretisch durchführen lässt.

11. Wir wollen noch unsere Abschätzungen auf die Diskussion der zyklischen Einzelschritt-Iteration (der sogenannten Gauss-Seidel-Iteration) anwenden. Man betrachte das lineare System n-ter Ordnung

$$U\xi=\alpha\,, \qquad U=(u_{ik})\,, \qquad \det U\neq 0\,, \tag{12}$$

und zerlege die Matrix U in die Summe:

$$U=L+D+R\,. \tag{13}$$

Hier ist D die zu U gehörende Diagonalmatrix. L entsteht, indem man in U alle Elemente links von der Hauptdiagonale beibehält und sämtliche übrigen Elemente durch Nullen ersetzt, während R die analog gebildete Matrix mit den Elementen von U rechts von der Hauptdiagonale ist.

12. Setzt man nun

$$A:=-(D+L)^{-1}R\,, \tag{14}$$

so läuft der allgemeine Zyklus der zyklischen Einzelschritt-Iteration angewandt auf (12), auf die Iteration

$$\xi_{\nu+1} = A\xi_\nu + (D + L)^{-1}\,\alpha \tag{15}$$

hinaus. Dies ist eine *inhomogene* Iteration mit der konstanten Matrix A. Für die Konvergenz dieser in der Praxis viel gebrauchten Iteration sind nur einige spezielle Kriterien bekannt. Trotzdem lassen sich auch hier unsere Abschätzungen anwenden.

13. Sei ζ die Lösung von (12). Aus (15) folgt:

$$\xi_{\nu+1} - \zeta = A(\xi_\nu - \zeta)\,, \quad \xi_{\nu+1} - \xi_\nu = (A - E)\,(\xi_\nu - \zeta)\,,$$
$$\xi_\nu - \zeta = (A - E)^{-1}(\xi_{\nu+1} - \xi_\nu)\,. \tag{16}$$

Nun gilt aber wegen (14), wenn mit E die Einheitsmatrix bezeichnet wird,

$$E - A = (D + L)^{-1}\,U\,,$$

und daher:

$$\xi_\nu - \zeta = -\,U^{-1}\,(D + L)\,(\xi_{\nu+1} - \xi_\nu)\,. \tag{17}$$

14. Bei der Herleitung von (17) haben wir die Konvergenz der Iteration (15) *nicht vorausgesetzt*. Für die Konvergenz (mit allgemeinem α) ist natürlich notwendig und hinreichend, dass die rechtsstehende Matrix in (17) den Spektralradius < 1 hat.

Aus der Ungleichung (17) folgt aber unabhängig von der Konvergenz

$$|\xi_\nu - \zeta| \leqslant |U^{-1}\,(D + L)|\;|\xi_{\nu+1} - \xi_\nu|\,, \tag{18}$$

wenn hier die Matrixnorm die durch die zugehörige Vektornorm induzierte ist. Ist der Wert dieser Matrixnorm bekannt, so liefert (18) eine Fehlerschranke, sobald $|\xi_{\nu+1} - \xi_\nu|$ klein genug ist, unabhängig von jeder Konvergenzannahme.

15. Wollen wir insbesondere die euklidischen Normen benutzen, so ergibt sich speziell die Abschätzung

$$|\xi_\nu - \zeta| \leqslant \frac{|D + L|_e}{|\det U|}\left(\frac{|U|_F}{\sqrt{n-1}}\right)^{n-1}|\xi_{\nu+1} - \xi_\nu|\,. \tag{19}$$

In der Praxis wird man in dieser Formel wohl $|D + L|_e$ durch $|D + L|_F$ ersetzen. Zur Anwendung von (19) ist allerdings die Kenntnis einer Abschätzung von $|\det U|$ nach unten unerlässlich.

Literatur

[1] Eberlein, P. J.: On Measures of Nonnormality for Matrices. Amer. Math. Monthly 72, 995 (1965).

[2] Henrici, P.: Bounds for Iterates, Inverses, Spectral Variation and Fields of Values of Non-normal Matrices. Numer. Math. 4, 24 (1962).

[3] Ostrowski, A. M.: Über Normen von Matrizen. Math. Z. 63, 12 (1955).

[4] Ostrowski, A. M.: Les estimations des erreurs a posteriori dans les procédés itératifs. C.R. Acad. Sc. Paris 275 (A), 275 (1972).

[5] Ostrowski, A. M.: A posteriori error estimates in iterative procedures. SIAM J. Numer. Anal. 10, 290 (1973).

[6] Weissinger, J.: Über das Iterationsverfahren. Z. Angew. Math. Mech. 31, 245 (1951).