Aplikace matematiky

Petr Liebl

Einige Bemerkungen zur numerischen Stabilität von Matrizeniterationen

Aplikace matematiky, Vol. 10 (1965), No. 3, 249–254

Persistent URL: http://dml.cz/dmlcz/102959

EINIGE BEMERKUNGEN ZUR NUMERISCHEN STABILITÄT VON MATRIZENITERATIONEN

PETR LIEBL

(zum Thema b)

1. Betrachten wir stationäre Iterationen in einem endlichdimensionalen linearen Raum

$$x_{p+1} = A(x_p)\,, \tag{1}$$

wo x_p, $p = 0, 1, \ldots$ Elemente eines N-dimensionalen linearen metrischen Raumes $\mathscr{R}$ sind und A ein (im allgemeinen nichtlinearer) Operator in diesem Raum ist. Setzen wir voraus, daß ein derartiges Element $y \in \mathscr{R}$ existiert, für welches

$$y = A(y) \tag{2}$$

gilt. Wählen wir eine bestimmte Basis in $\mathscr{R}$. Die Koordinaten des Elementes x in dieser Basis bezeichnen wir mit ξ_i, $i = 1, 2, \ldots, N$, die Koordinaten des Elementes $A(x)$ mit $a_i(\xi_1, \xi_2, \ldots, \xi_N)$, i $= 1, 2, \ldots, N$. Ordnen wir (unter der Voraussetzung, daß die Ableitungen existieren) dem Operator A seine Jacobische Matrix $A'(x)$ zu,

$$A'(x)_{ik} = \frac{\partial a_i(\xi_1, \xi_2, \ldots, \xi_N)}{\partial \xi_k}\,, \tag{3}$$

die quadratisch und N-ter Ordnung ist. Die Eigenschaften dieser Matrix hängen mit der Konvergenz des Verfahrens (1) zusammen. Setzen wir nämlich voraus, man kann die Funktionen a_i in einer genügend kleinen Umgebung des Punktes y unter Hinweglassung Glieder höherer Ordnung durch das lineare Glied ersetzen

$$a_i(\xi_1, \xi_2, \ldots, \xi_N) = a_i(\eta_1, \eta_2, \ldots, \eta_N) + \sum_{k=1}^{N} \frac{\partial a_i(\eta_1, \eta_2, \ldots, \eta_N)}{\partial \eta_k} (\xi_k - \eta_k)$$

oder in Vektorform

$$A(x) = A(y) + A'(y)(x - y)\,,$$

wo mit $A'(y)$ die Jacobische Matrix (3) für $x = y$ bezeichnet ist. Mit (1) und (2) haben wir dann

$$x_{p+1} - y = A'(y)(x_p - y) \tag{4}$$

und es ist zu sehen, wie z.B. die Iteration (1) konvergiert, wenn alle Eigenwerte der Matrix $A'(y)$ im Betrag kleiner als 1 sind und x_0 (oder im allgemeinen ein x_p für ein beliebiges $p \geqq 0$) in einer bestimmten, genügend kleinen Umgebung des Punktes y liegt.

Die Abbildung, die jedem (differenzierbaren) Operator F seine Jacobische Matrix F' zuordnet, ist linear

$$(\alpha F + \beta G)' = \alpha F' + \beta G' . \tag{5}$$

Außerdem gilt

$$(F(G))' = F'G' , \tag{6}$$

d.h. die Jacobische Matrix des zusammengesetzten Operators, der die aufeinanderfolgende Anwendung von G und dann F bedeutet, ist gleich dem Produkt der Jacobischen Matrizen der Operatoren F und G (in dieser Reihenfolge). Schließlich ist

$$A' = 0 , \quad I' = E , \tag{7}$$

d.h. die Jacobische Matrix des konstanten Operators ist die Nullmatrix und die des identischen Operators die Einheitsmatrix.

2. Für die weiteren Überlegungen brauchen wir den Begriff des Tensorproduktes zweier Matrizen. Es seien A und B Matrizen vom Typus $m \times n$ resp. $r \times s$. Als ihr Tensorprodukt bezeichnen wir die Matrix vom Typus $mr \times ns$

$$A \times B = \begin{pmatrix} a_{11}B, & a_{12}B, & \ldots, & a_{1n}B \\ a_{21}B, & a_{22}B, & \ldots, & a_{2n}B \\ \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots \\ a_{m1}B, & a_{m2}B, & \ldots, & a_{mn}B \end{pmatrix}, \tag{8}$$

deren Element an der Stelle $(\alpha - 1)r + \beta, (\gamma - 1)s + \delta$ gleich $a_{\alpha\gamma} \cdot b_{\beta\delta}$ ist. Es gilt dann

$$\begin{aligned} &(A \times B) + (A \times C) = A \times (B + C) , \quad (A \times B)(C \times D) = AC \times BD , \\ &(A \times B) + (C \times B) = (A + C) \times B , \quad (A \times B)^T = A^T \times B^T . \end{aligned} \tag{9}$$

3. Beschränken wir uns nun auf Matrizeniterationen und Matrizenfunktionen, d. h. $\mathfrak{R}$ sei der $N = n^2$-dimensionale Raum der quadratischen Matrizen der Ordnung n. Als Basis wählen wir das System der n^2 Matrizen $e_i e_j^T$, $i, j = 1, 2, \ldots, n$, wo mit e_t der t-te (Spalten-) Einheitsvektor bezeichnet ist, und zwar in der Reihenfolge

$e_1e_1^T, e_1e_2^T, \ldots, e_1e_n^T, e_2e_1^T, e_2e_2^T, \ldots$. Der Koeffizient α_{ij} bei dem Basiselement $e_ie_j^T$ bei der Zerlegung der Matrix $A = \|a_{ij}\|_1^n$ in dieser Basis ist einfach gleich a_{ij}.

Es seien nun F, U, V Matrizenfunktionen, d.h. Operatoren, die jeder quadratischen Matrix n-ter Ordnung X eine Matrix $F(X)$, $U(X)$ resp. $V(X)$ zuordnen. Setzten wir voraus, daß alle vorkommenden Ableitungen existieren. Wenn nun der Operator F so definiert ist, daß die Matrix $F(X)$ gleich dem Produkt der Matrizen $U(X)$ und $V(X)$ ist

$$F = UV, \tag{10}$$

so gilt

$$F' = (U \times E)\, V' + (E \times V^T)\, U', \tag{11}$$

wo F', U', V' die Jacobischen Matrizen der Funktionen F, U, V sind; E ist hier die Einheitsmatrix n-ter Ordnung, V^T die Transponierte zu V. (Für $n = 1$ geht (11) in die gewöhnliche Formel für die Ableitung des Produktes zweier Funktionen über.) Mit Hilfe der Formeln (5), (6), (7), (11) können die Jacobischen Matrizen der einfacheren Matrizenfunktionen leicht abgeleitet werden.

4. Untersuchen wir als Beispiel iterative Formeln zur Berechnung der Quadratwurzel einer symmetrischen, positiv definiten Matrix A

$$X_{p+1} = X_p + d_1(A - X_p^2), \qquad X_0 = 2d_1A, \tag{12}$$

$$Y_{p+1} = Y_p + d_2(E - A^{-1}Y_p^2), \quad Y_0 = 2d_2A, \tag{13}$$

(d_1, d_2 sind noch unbestimmte numerische Konstanten),

$$Z_{p+1} = \tfrac{1}{2}(Z_p + AZ_p^{-1}), \quad Z_0 = E. \tag{14}$$

Beachten wir, daß alle so definierten X_p, Y_p, Z_p in dem n-dimensionalen Raum $\mathfrak{R}_A$ der mit A vertauschbaren Matrizen liegen. Es sei Λ die Diagonalmatrix der Eigenwerte λ_i der Matrix A, U die Ortogonalmatrix, für die

$$UAU^T = \Lambda$$

ist. Die Matrizen

$$UX_pU^T, UY_pU^T, UZ_pU^T \tag{15}$$

sind auch Diagonalmatrizen mit den Diagonalelementen $\xi_i^{(p)}$ resp. $\eta_i^{(p)}$ resp. $\zeta_i^{(p)}$, für die

$$\xi_i^{(p+1)} = \xi_i^{(p)} + d_1(\lambda_i - (\xi_i^{(p)})^2), \tag{12'}$$

$$\eta_i^{(p+1)} = \eta_i^{(p)} + d_2(1 - (\eta_i^{(p)})^2\, \lambda_i^{-1}), \tag{13'}$$

$$\zeta_i^{(p+1)} = \tfrac{1}{2}(\zeta_i^{(p)} + \lambda_i(\zeta_i^{(p)})^{-1}), \tag{14'}$$

$$i = 1, 2, \ldots, n,$$

$$p = 0, 1, \ldots$$

gilt. Hier handelt es sich offenbar in jeder der drei Formeln um n nebeneinander verlaufende skalare Iterationsprozesse, die für $i = 1, 2, \ldots, n$ zu $\sqrt{\lambda_i}$ konvergieren sollen. Die Ableitung der Funktion rechts im Falle (12′) ist gleich

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\left(\xi + d_1(\lambda_i - \xi^2)\right) = 1 - 2d_1\xi\,.$$

Nach Einsetzung von $\sqrt{\lambda_i}$ für ξ ist ersichtlich, daß die Wahl

$$d_1 = \frac{1}{2\sqrt{\lambda_{\max}}} \tag{16}$$

zur besten Gesamtkonvergenz führt; die n Ableitungen, die eigentlich Eigenwerte des Prozesses (12) sind, sind dann gleich

$$1 - \sqrt{\frac{\lambda_i}{\lambda_{\max}}}\,.$$

Ebenso finden wir für (13)

$$d_2 = \frac{\sqrt{\lambda_{\min}}}{2} \tag{17}$$

und die Eigenwerte

$$1 - \sqrt{\frac{\lambda_{\min}}{\lambda_i}}\,.$$

Die Eigenwerte des Prozesses (14′) sind für $i = 1, \ldots, n$ gleich Null, was auf quadratische Konvergenz des Verfahrens (14) schließen läßt.

5. Die Erwägungen des Absatzes 4 haben guten Sinn, solange alle Beziehungen (12), (13), (14) genau gelten. In dem praktisch wichtigen Falle aber, wenn die Berechnungen numerisch durchgeführt werden, muß man mit Rundungsfehlern rechnen. Das bedeutet z.B., daß die Matrizen X_p, Y_p, Z_p nicht in $\mathfrak{R}_A$ liegen und die Matrizen (15) nicht diagonal sind. Es handelt sich dann nicht mehr um Iterationen in dem n-dimensionalen Raum $\mathfrak{R}_A$, sondern in dem n^2-dimensionalen Raum aller quadratischer Matrizen n-ter Ordnung.[1]) Die Jacobische Matrix n^2-ter Ordnung des Prozesses (12) im Punkte der Lösung ist gleich

$$(E \times E) - d_1\left[(E \times A^{\frac{1}{2}}) + (A^{\frac{1}{2}} \times E)\right]. \tag{18}$$

[1]) Im Falle (12) bleiben alle Iterationen, dank dem speziellen Charakter der Rundungsfehler, in dem $\frac{1}{2}(n+1)$ n-dimensionalen Raum der symmetrischen Matrizen. In diesem Falle machen wir davon nicht Gebrauch.

Um ihre Eigenwerte zu berechnen, multiplizieren wir sie von links resp. rechts mit der Ortogonalmatrix $U \times U$ resp. $U^T \times U^T$. So erhalten wir eine Diagonalmatrix mit den Diagonalelementen x_{ik}, $i = 1, \ldots, n$, $k = 1, \ldots, n$,

$$x_{ik} = 1 - d_1(\sqrt{\lambda_i} + \sqrt{\lambda_k}) .$$

Für (16) ist dann

$$x_{ik} = 1 - \frac{1}{2}\left(\sqrt{\frac{\lambda_i}{\lambda_{\max}}} + \sqrt{\frac{\lambda_k}{\lambda_{\max}}}\right);$$

alle diese Eigenwerte liegen im Intervall $\langle 0, 1)$. Das bedeutet, daß der Prozess zu $\sqrt{A}$ konvergiert, auch wenn wir als Ausgangsnäherung eine beliebige (z.B. nicht symmetrische) Matrix wählen, die der Matrix $\sqrt{A}$ genügend nahe ist.

Vollständig anders ist die Lage bei den Prozessen (13), (14). Ganz analog wie (18) finden wir für (13) die Jacobische Matrix

$$(E \times E) - d_2[(A^{-1} \times A^{\frac{1}{2}}) + (A^{-\frac{1}{2}} \times E)]$$

mit den Eigenwerten

$$y_{ik} = 1 - d_2\left(\frac{\sqrt{\lambda_k}}{\lambda_i} + \frac{1}{\sqrt{\lambda_i}}\right).$$

Für (17) ist dann

$$y_{ik} = 1 - \frac{1}{2}\left(\frac{\sqrt{\lambda_k \lambda_{\min}}}{\lambda_i} + \sqrt{\frac{\lambda_{\min}}{\lambda_i}}\right).$$

Bei $\lambda_i = \lambda_{\min}$ ist nun

$$y_{ik} = \frac{1}{2}\left(1 - \sqrt{\frac{\lambda_k}{\lambda_{\min}}}\right),$$

und das ist für schlecht bedingte Matrizen für einige k eine im Betrag große, negative Zahl. Das bedeutet, daß, sobald die Näherungen Y_p den Unterraum $\mathfrak{R}_A$ (dem die n Eigenwerte y_{ii} entsprechen) verlassen, der Prozess (13) mit (17) nicht konvergieren kann. Erst eine Wahl

$$d_2' = \frac{\lambda_{\min}}{\sqrt{\lambda_{\max}}} \tag{19}$$

bringt alle Eigenwerte in das Intervall $(-1, 1)$.

Bei dem Prozess (14) schließlich ist die Jacobische Matrix gleich

$$\tfrac{1}{2}((E \times E) - (A^{\frac{1}{2}} \times A^{-\frac{1}{2}}))$$

mit den Eigenwerten

$$z_{ik} = \frac{1}{2}\left(1 - \sqrt{\frac{\lambda_i}{\lambda_k}}\right).$$

(Die dem Unterraum $\mathscr{R}_A$ entsprechenden Eigenwerte sind gleich Null, wie schon im Absatz 4 erwähnt wurde.) Auch dieser Prozess muß für schlecht bedingte Matrizen divergieren.

6. Weisen wir kurz auf praktische Folgerungen aus den Schlüssen des Absatzes 5 hin. Im Falle (12) kann erwartet werden, daß der Prozess, ausgehend von $2d_1A$, auch bei numerischer Durchführung zu $\sqrt{A}$ konvergiert. Das durch Rundungsfehler bedingte Heraustreten der X_p aus $\mathscr{R}_A$ wird keine weitere Auswirkungen haben und kompensiert sich im weiteren Verlauf der Iteration.

Die pessimistischen Resultate über die Methoden (13), (14) bedeuten nicht etwa, daß diese nicht brauchbar sind. Die Komponenten in den Richtungen der kritischen Eigenvektoren, die den großen Eigenwerten entsprechen, sind am Anfang von der Größenordnung der Rundungsfehler. Bei genügender Geschwindigkeit der Konvergenz in $\mathscr{R}_A$ (dessen Basis diejenigen Eigenvektoren bilden, die den Eigenwerten y_{ii} resp. z_{ii} entsprechen) kann es gelingen, daß diese divergenten Komponenten nicht Zeit haben, allzu groß zu werden. Es ist nicht einmal zu empfehlen, bei (13) d_2 nach (19) statt nach (17) zu wählen, da dadurch die Konvergenz in $\mathscr{R}_A$ wesentlich verlangsamt würde.

Es ist nur notwendig, von diesen divergenten Komponenten zu wissen und den Prozess rechtzeitig abzubrechen. Bei der Durchführung der Rechnung auf Rechenautomaten bedeutet das, bei programmierter Beendigung der Iteration nach Erreichung vorgeschrieben kleiner Residuen, die Anforderungen an die Genauigkeit nicht allzu streng zu stellen, da bei langer Iteration die divergierenden Komponenten Oberhand gewinnen können und der Prozess scheitert.

Petr Liebl, Matematický ústav ČSAV, Žitná 25, Praha 1, ČSSR.