Institut für Geometrie
und Praktische Mathematik
Prof. Dr. A. Reusken, Dr.-Ing. T. Kremp

RWTH
RHEINISCH-WESTFÄLISCHE TECHNISCHE HOCHSCHULE AACHEN

Numerische Mathematik I (Maschinenbau) frequently asked questions (FAQ), Klausur 25.03.08

- *Welche Themen sind klausurrelevant?*

  Klausurrelevant sind alle in der Vorlesung und Übung behandelten Begriffe und Verfahren (wobei ein **Multiple-Choice-Teil** vorkommen wird für den es **50% der Gesamtpunktzahl** geben wird — siehe auch Online-Übungsaufgaben):

  Kondition, Stabilität, gestörte lineare Gleichungssysteme, Gauß-Elimination, LR-Transformation, Skalierung, Pivotisierung, (keine Nachiteration), LDL^T-Transformation, QR-Zerlegung (Givens, Householder), Aufwand, linearer Ausgleich (QR und Normalgleichungen), nichtlineare skalare Gleichungen (Bisektion, Sekantenverfahren), nichtlineare Gleichungssysteme (Banach'sche Fixpunktiteration, Newton-Verfahren, vereinfachtes Newton-Verfahren, gedämpftes Newton-Verfahren), nichtlinearer Ausgleich (Gauss-Newton mit QR und Normalgleichungen, Levenberg-Marquardt), Konvergenzordnung, Interpolation (Lagrange, Newton, Neville-Aitken), numerische Differentiation, Quadratur (Newton-Cotes, auch summiert, Gauß-Quadratur, keine Romberg-Extrapolation).

- *Mit wie vielen Stellen soll man in der Klausur rechnen?*

  Falls in der Aufgabenstellung die Verwendung einer bestimmten Anzahl von Stellen gefordert ist, hat man sich natürlich daran zu halten (Erinnerung: nach **jeder** Operation entsprechend runden und mit dem gerundeten Wert weiterrechnen). Eine große Anzahl von Stellen ("Rechnen Sie in 10-stelliger GPA") wird jedoch sicher nicht gefordert werden, da dies für eine zweistündige Klausur ungeeignet wäre.

  Falls in der Aufgabenstellung keine Vorschrift zur Gleitpunktarithmetik gemacht ist, gelten folgende Faustregeln:

  4–5 Stellen reichen bei fast allen klausurrelevanten Aufgaben. Hierbei ist es nicht notwendig, eine bestimmte Gleitpunktarithmetik konsequent durchzuhalten (d. h. mühsames Runden und neues Eintippen nach jeder einzelnen Rechenoperation). Um Zeit zu sparen verwendet man stattdessen jeweils entweder die alten noch im Speicher befindlichen (z. B. 12-stelligen) Werte, oder tippt nicht mehr vorhandene Werte neu ein, wobei 4–5 Stellen reichen. Nur bei hochgenauen Rechnungen benötigt man mehr Stellen, d. h. etwa ≥ 8.

  Beim Romberg-Quadratur-Schema beispielsweise muss von Anfang an, d. h. im **gesamten** Schema, mit ausreichend vielen Stellen (etwa ≥ 8) gerechnet werden, da sich die Rundungsfehler von oben links ja bis ins Ergebnis unten rechts fortpflanzen.

  Da es sich bei dem (ebenfalls hochgenauen, da quadratisch konvergenten) Newton-Verfahren hingegen um ein iteratives (d. h. alte Rundungsfehler werden im Falle der Konvergenz automatisch korrigiert) Verfahren handelt, ist es dort ausreichend, erst sukzessive mit jedem weiteren Iterationsschritt immer mehr Stellen zur Verfügung zu stellen, um tatsächlich die quadratische Konvergenz zu sehen und nicht nur Rundungsfehler zu iterieren.

- *Wie ausführlich sind die Voraussetzungen des Banachschen Fixpunktsatzes darzustellen?*

  Bei der Untersuchung der (bei uns mindestens stetig differenzierbaren) Iterationsfunktion $F(x)$ (mit $x^* = F(x^*)$) auf Selbstabbildung sind die Extrema von F zu bestimmen, wobei Randwerte und Extremwerte von F in Frage kommen. Nur wenn man glaubhaft vermitteln kann, dass F monoton ist, braucht man keine Extremwertbetrachtung von F durchzuführen, d. h. man braucht nur die Randwerte einzusetzen und nicht noch zusätzlich alle Extremwerte $F(x_0)$ zu berechnen mit $F'(x_0) = 0$. Beispielsweise glauben wir der Aussage "$F(x) = \mathrm{e}^x$ ist monoton in $[0,1] \Rightarrow$ untersuche nur Randwerte" ohne weiteren Beweis. Die Aussage "$-1 \leq F(x) := \sin(x) \leq 1 \Rightarrow F$ ist selbstabbildend auf jedem Intervall $[a,b] \supseteq [-1,1]$" glauben wir ebenfalls ohne weiteren Beweis. Jedoch glauben wir

einer Aussage wie "$F(x) = 5x^4 + 3x^3 - 2x^2 + 10x - 3$ ist monoton in $[a, b]$" nicht ohne Beweis, d. h. es ist erst zu untersuchen, ob Stellen x_0 mit $F'(x_0) = 0$ in $[a, b]$ liegen, und — falls dies der Fall ist — sind auch alle Extremwerte $F(x_0)$ auszurechnen und mit den Randwerten zu vergleichen.

Hat man eine **abgeschlossene** (unbedingt erwähnen!) Menge E gefunden mit $F(E) =: \widetilde{E} \subseteq E$ (d. h. F bildet E auf sich selbst ab), so ist nun die Kontraktivität zu untersuchen, d. h. ob $\max_{x \in E} |F'(x)| =: L < 1$ gilt. Da F stetig differenzierbar ist, kommen hierbei auch wieder Rand- und Extremwerte in Frage, so dass man bei monotonem F' (wieder begründen, siehe oben) nur Randwerte von F' untersuchen muss. Man hat also praktisch dieselben Überlegungen wie bei der Selbstabbildung durchzuführen, allerdings alles "um eine Ableitung höher".

Im mehrdimensionalen Fall muss E **konvex** (unbedingt erwähnen!) sein, damit $\max_{x \in E} ||F'(x)|| =: L < 1$ hinreichend ist für Kontraktivität. (Erinnerung: Eine Menge E ist konvex, wenn die Verbindungsstrecke je zweier beliebiger Punkte aus E stets **ganz** in E liegt. Ein n-dimensionaler Quader (n-dimensionales Intervall) ist trivialerweise konvex, ebenso eine n-dimensionale Kugel, nicht jedoch ein L-förmiges Gebiet.) Bei der Normbildung wird man, wie in der Übung, in der Regel erst die Elemente der Jacobi-Matrix F' einzeln betragsmäßig abschätzen und dann die Norm bilden, was zwar gröber, aber sehr viel weniger aufwändig ist als wenn man erst $||F'(x)||$ für allgemeines x hinschreibt und dann bzgl. x maximiert. Hierbei kann man sich entweder für die 1- oder ∞-Norm entscheiden, wobei im Falle $\max_{x \in E} ||F'(x)||_\infty \approx \max_{x \in E} ||F'(x)||_1$ meist die ∞-Norm am geschicktesten ist, da ja stets $||x^1 - x^0||_\infty \leq ||x^1 - x^0||_1$ gilt, so dass die a-priori-Abschätzung dann günstiger wird als mit der 1-Norm. Für die 2-Norm ist dieses vorherige komponentenweise Abschätzen nicht zulässig, da nicht sicher ist, dass ein Vergrößern der Beträge der Komponenten von F' auch ein Vergrößern von $||F'||_2$ zur Folge hat. Dies ist aber keine nennenswerte Einschränkung, da die 2-Norm aufgrund ihres höheren Aufwandes (Eigenwertbestimmung von $(F')^T F'$) hierbei ohnehin kaum verwendet wird.

Da der gesuchte Fixpunkt x^* stets in $\widetilde{E}$ liegt, reicht es auch aus, wenn man die Kontraktivitätsuntersuchung auf dem **kleineren** Bereich $\widetilde{E}$ durchführt. Dies hat zwar den Nachteil, dass die Zahlenwerte oft etwas unschöner werden, jedoch den erheblichen Vorteil, dass die Lipschitzkonstante L dadurch meist kleiner wird. Man muss allerdings beachten, dass die sich daraus ergebende a-priori-Abschätzung ("nach $n = 17$ Schritten gilt sicher $||x^n - x^*|| < \varepsilon$") nur gilt, wenn auch der (evtl. vorgeschriebene) Startwert in $\widetilde{E}$ liegt. Liegt der Startwert nicht $\widetilde{E}$, so braucht man eben einen Schritt mehr (um nämlich von diesem Startwert aus E nach $\widetilde{E}$ zu kommen), d. h. es gilt sicher $||x^{n+1} - x^*|| < \varepsilon$.

Sowohl a-priori- als auch a-posteriori-Abschätzungen können falsch sein, wenn die für die Differenzen $||x^1 - x^0||$ bzw. $||x^n - x^{n-1}||$ verwendete Vektornorm nicht mit der für die Lipschitzkonstante L verwendeten Matrixnorm verträglich ist. Dies ist der Fall, wenn man verschiedene Normen verwendet, z. B. die $||\cdot||_1$-Norm für L und die $||\cdot||_\infty$-Norm für $||x^1 - x^0||$, und wird mit Punktabzügen bewertet.

- *Welche Richtung hat das Ungleichheitszeichen bei der a-priori-Abschätzung?*

  Unter den obigen Voraussetzungen gilt die a-priori-Abschätzung $||x^n - x^*|| \leq \frac{L^n}{1-L} ||x^1 - x^0||$. Um die Forderung $||x^n - x^*|| \leq \varepsilon$ (z. B. $\varepsilon := 10^{-3}$) sicher zu erfüllen, ist also $\frac{L^n}{1-L} ||x^1 - x^0|| \leq \varepsilon$ hinreichend, was wegen $\ln(L) < 0$ äquivalent ist zu $n \geq \ln\left(\frac{\varepsilon(1-L)}{||x^1 - x^0||}\right) / \ln(L)$. Ein "$\leq$" statt des "$\geq$" wäre hier offensichtlich völlig falsch und wird ebenfalls mit Punktabzügen bewertet.

- *Welche Funktion und welches Vorzeichen muss man beim nichtlinearen Ausgleich wählen?*

  Beim nichtlinearen Ausgleich ist eine Funktion $F : \mathbb{R}^n \to \mathbb{R}^m$, $n < m$ (deshalb leider i. a. $F(x) \neq 0$) bzgl. der 2-Norm zu minimieren, d. h. im (Gaußschen) Sinne minimaler Quadrate der Komponenten von F: $||F(x)||_2 \to \min$, d. h. Beträge der einzelnen Zeilen gegeneinander "ausgleichen". Bei der Gauß-Newton-Methode wird dieses nichtlineare Problem durch eine Iterationsfolge linearer Probleme ersetzt, indem man F jeweils im aktuellen Punkt x^k linear approximiert (deshalb "Newton") und das sich daraus ergebende lineare Ausgleichsproblem $||F'(x^k)\Delta x^k + F(x^k)||_2 \to \min$ löst (mit $x^{k+1} := x^k + \Delta x^k$). Hierzu wird das System $[F'(x^k) \mid -F(x^k)]$ (also inkl. rechter Seite) QR-transformiert und der obere $(n \times n)$-Teil anschließend durch Rückwärtseinsetzen exakt gelöst. Das sich aus den letzten $m - n$ Zeilen der transformierten rechten Seite ergebende Residuum ist das Residuum $||F'(x^k)\Delta x^k + F(x^k)||_2$ der linearen Näherung, das sowohl größer als auch kleiner als das "echte" Residuum $||F(x^{k+1})||_2$ sein kann. Arbeitet man hingegen mit Normalgleichungen, so ist das System $F'(x^k)^T F'(x^k)\Delta x^k = -F'(x^k)^T F(x^k)$ mit LDL^T-Transformation (oder bei kleinen m, n auch mit Gaußelemination) zu lösen, wobei man sich wegen $\kappa_2(F'(x^k)^T F'(x^k)) = \kappa_2(F'(x^k))^2$ jedoch eine Quadrierung und damit Verschlechterung der Kondition einhandelt.

  Ebenso wie beim linearen Ausgleich kann auch beim nichtlinearen Ausgleich die zu minimierende Funktion $F(x)$ sowohl explizit als auch implizit von den Werten in der Messwertetabelle abhängen. Wir beschränken uns im Folgenden auf den Fall, dass m Messwerte**paare** (t_i, y_i), $i = 1, \ldots, m$, gegeben sind.

  Im expliziten Fall ist eine explizite analytische Funktion $y(t; x)$ gegeben für den Messwert y, der von dem anderen Messwert t sowie den unbekannten Parametern x_j, $j = 1, \ldots, n < m$, abhängt, z. B. $y(t; x) = x_1 \exp(-x_2 t)$ mit $n = 2$. Man wählt dann beispielsweise $F(x) := [y(t_i; x) - y_i]_{i=1,\ldots,m}$, also ergibt sich als $(m \times n)$-Jacobi-Matrix $F'(x) = [\partial y(t_i; x)/\partial x_j]_{i=1,\ldots,m;j=1,\ldots,n}$ und als rechte Seite $-F(x) = [y_i - y(t_i; x)]_{i=1,\ldots,m}$.

  Im impliziten Fall ist ein impliziter analytischer Zusammenhang $f(t, y; x) = 0$ gegeben, der die Messwerte t und y miteinander verknüpft und von den unbekannten Parametern x_j, $j = 1, \ldots, n < m$, abhängt, z. B. $f(t, y; x) = (t/x_1)^2 + (y/x_2)^2 - 1 = 0$ für eine Ellipse in Normallage in der (t, y)-Ebene mit $n = 2$. Man wählt dann einfach $F(x) := [f(t_i, y_i; x)]_{i=1,\ldots,m}$, also ergibt sich als $(m \times n)$-Jacobi-Matrix $F'(x) = [\partial f(t_i, y_i; x)/\partial x_j]_{i=1,\ldots,m;j=1,\ldots,n}$ und als rechte Seite $-F(x) = [-f(t_i, y_i; x)]_{i=1,\ldots,m}$.