Institutionen för datavetenskap
Umeå universitet

21 december 2010

Numeriska metoder för civilingenjörer 5DV040
Teknisk-vetenskapliga beräkningar 5DV005

Lösningar till tentamen

1. ***Detta problem skall endast göras av dem som ej gjort eller ej är godkända på duggan!***

(a) Gäller de associativa [$a+(b+c)=(a+b)+c$, $a(bc)=(ab)c$] respektive distributiva [$a(b+c)=ab+ac$] lagarna för beräkningar med flyttal? [2p]

(b) I standarden IEEE 754-2008 finns ett format för utökad precision, binär 128, som har ännu högre noggrannhet än binär 64 (dubbelprecision), det format vi gått igenom i kursen. Varje flyttal i det utökade formatet omfattar 128 bitar, varav 112 bitar används för att lagra decimaldelen, 15 bitar för exponenten och 1 bit för tecknet. Vad blir maskinepsilon i binär 128? [2p]

(c) Sant eller falskt: om man använder pivotering vid lösning av ekvationssystem förbättras konditionstalet. [2p]

(d) Man vill lösa ett system av icke-linjära ekvationer $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ där $\mathbf{x}, \mathbf{f} \in \mathbb{R}^n$. Vilken av metoderna Newtons metod eller fixpunktiterationer skulle du som tumregel rekommendera om (a) $n = 10$ (b) $n = 10^5$? Motivera. [2p]

(e) LU-faktorisera matrisen

$$A = \begin{pmatrix} -3 & 2 & 1 \\ 3 & -1 & 1 \\ -1 & 0 & 2 \end{pmatrix}.$$

[2p]

Svar. (a) Nej i samtliga fall.

(b) $\epsilon_M = 2^{-112}$ (avståndet mellan 1 och nästa större tal i flyttalssystemet).

(c) Nej, konditionstalet ändras inte vid pivotering.

(d) För fall (a) rekommenderas Netwon's metod, då denna kan förväntas konvergera snabbare mot lösningen. För fall (b) rekommenderas fixpunktsiteration, som för stora problem är mycket snabbare per iteration (inga ekvationssystem behöver lösas). Newtons metod kräver att ett linjärt ekvationssytem löses vid varje iteration, vilket är beräkningstungt (kräver många flyttalsoperationer och stort minnerutrymme) för stora problem. Dessutom kan det vara svårt/krävande att ens beräkna Jakobianmatrisen för stora problem.

(e)

$$A = \begin{pmatrix} -3 & 2 & 1 \\ 3 & -1 & 1 \\ -1 & 0 & 2 \end{pmatrix} \overset{\textcircled{1}\ \textcircled{-1/3}}{\sim} \begin{pmatrix} -3 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & -2/3 & 5/3 \end{pmatrix} \overset{\textcircled{2/3}}{\sim} \begin{pmatrix} -3 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{pmatrix}$$

Vilket ger oss att $A = LU$, där

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1/3 & -2/3 & 1 \end{pmatrix} \text{ och } U = \begin{pmatrix} -3 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{pmatrix}.$$

2. (a) Vi samplar en funktion i fyra punkter, sammanfattade i tabellen nedan.

x	0	1/3	2/3	1
$f(x)$	1.0	0.8	1.2	1.7

(i) Beräkna $\int_0^1 f(x)\mathrm{d}x$ med trapetsmetoden. Använd så många punkter som möjligt. [5p]

(ii) Nämn ett problem med att använda Simpsons metod för att beräkna ovanstående integral. [5p]

(b) Från tabellen framgår att funktionen f har ett minimum i närheten av $x = 1/3$. Använd interpolation för att på ett lämpligt sätt hitta en bättre approximation av var detta minimum befinner sig. [10p]

(c) En numerisk metod för integralberäkning kan skrivas

$$\int_a^b f(x)\,dx = I_h + \epsilon_h,$$

där I_h är den numeriska approximationen, ϵ_h diskretiseringsfelet och h steglängden. Antag att man vet att metodens noggrannhetsordning är p, d.v.s. det gäller för felet att $\epsilon_{2h} \approx 2^p \epsilon_h$. Visa att man kan utnyttja två beräkningar I_h och I_{2h} med steglängderna h respektive $2h$ för att uppskatta felet vid steglängden h till

$$\epsilon_h \approx \frac{I_h - I_{2h}}{2^p - 1}.$$ [10p]

Svar. (a) (i) Dela upp integralen över intervallen $[0,1/3]$, $[1/3,2/3]$ och $[2/3,1]$ och använd trapetsmetoden:

$$\int_0^1 f(x) = \sum_{n=1}^{3} \int_{(n-1)/3}^{n/3} f(x) \approx \sum_{n=1}^{3} \frac{1}{3}\frac{1}{2}\left(f\left(\frac{n-1}{3}\right) + f\left(\frac{n}{3}\right)\right)$$
$$= \frac{1}{6}(1.0 + 0.8 + 0.8 + 1.2 + 1.2 + 1.7) = \frac{6.7}{6}.$$

(ii) Funktionen som skall integreras är given i fyra punkter. För att Simpsons metod skall kunna användas behöver funktionen vara given i ett udda antal punkter.

(b) Minimum för f ligger någonstans i intervallet $[0,2/3]$. Interpolation av f i punkterna $0, 1/3, 2/3$ med ett kvadratiskt polynom ger $g(x) = 1 - 1.5x + 2.7x^2$. Vi antar att minimum av f ligger på samma plats som minimum av g, vilket är $x^* = 1.5/5.4 \approx 0.28$.

(c) Antagandet $\epsilon_{2h} \approx 2^p \epsilon_h$ ger att

$$I_h + \epsilon_h = \int_a^b f(x)\,dx = I_{2h} + \epsilon_{2h} \approx I_{2h} + 2^p \epsilon_h.$$

från vilket man kan lösa ut det efterfrågade uttrycket

$$\epsilon_h = \frac{I_h - I_{2h}}{2^p - 1}.$$

3. (a) Man vill lösa de kopplade ekvationssystemen

$$Bx + Ay = b,$$
$$Ax = c,$$

där A och B är kvadratiska matriser. Vi antar att vi har tillgång till LU-faktoriseringen $A = LU$. Visa hur man kan erhålla lösningarna x och y utan ytterligare faktoriseringar, d.v.s. med enbart lösningar av triangulära ekvationssystem samt matris–vektorprodukter och vektoradditioner. [10p]

(b) I tema 2 beräknade vi tryckskillnader mellan noderna i ett vattenledningsnätverk genom att lösa det linjära ekvationssystemet $K\bar{p} = \bar{s}$, där koefficientmatrisen $K = \bar{B}E\bar{B}^{\mathsf{T}}$. I ett av fallen hade matrisen K storlek 1544×1544 och konditionstal $\kappa(K) \approx 10^4$.

Antag att vi har en störning i högerledet, så istället för att räkna med $\bar{s}$ använder vi oss av det störda högerledet $\hat{s}$. Felet i högerledet kan begränsas enligt $\|\hat{s} - \bar{s}\| \leq \delta\|\bar{s}\|$. Hur stort får δ vara för att vi ska kunna garantera att det relativa felet hos den beräknande tryckskillnaden är mindre än 1%? [10p]

(c) Visa att för alla inverterbara matriser A gäller att konditionstalet $\kappa(A) \geq 1$. *Ledning:* Använd att $x = A^{-1}Ax$. [10p]

Svar: (a) Lös först $Ax = LUx = c$ genom fram- och bakåtsubstituering (d.v.s. lös först $Ld = c$ och sedan $Ux = d$). Substituera sedan lösningen x i första systemet och lös $Ay = LUy = b - Bx = \hat{b}$ genom fram- och bakåtsubstituering (d.v.s. lös först $Le = \hat{b}$ och sedan $Uy = e$).

(b) Det relativa felet i lösningen kan begränsas enligt

$$\frac{\|\hat{p} - \bar{p}\|}{\|p\|} \leq \kappa(K)\frac{\|\hat{s} - \bar{s}\|}{\|\bar{s}\|} \leq 10^4\delta$$

Från ovanstående ser vi att det relativa felet i trycket blir mindre än 1 procent då $10^4\delta < 0.01$. För att vi ska kunna garantera detta behöver vi kräva att $\delta < 10^{-6}$.

(c) Konditionstalet $\kappa(A)$ för en inverterbar matrix A är givet av $\kappa(A) = \|A\|\|A^{-1}\|$. Låt $x \neq 0$. Då gäller att

$$\|x\| = \|AA^{-1}x\| \leq \|A\|\|A^{-1}x\| \leq \|A\|\|A^{-1}\|\|x\|$$

vilket medför att $\kappa(A) = \|A\|\|A^{-1}\| \geq 1$.

ALT: För enhetsmatrisen har vi att $\|I\| = 1$, vidare gäller

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = \kappa(A)$$

4. (a) Skriv om följande problem på standardformen för begynnelsevärdesproblem:

$$\frac{d^2u}{\partial t^2} = v + t,$$
$$\frac{dv}{dt} = tu.$$

[5p]

(b) Implicita numeriska metoder för lösning av begynnelsevärdesproblem har oftast ett mycket större stabilitetsområde än explicita metoder, vilket innebär att man som regel kan välja tidssteg enbart utifrån den noggrannhet man behöver utan att ta hänsyn till stabilitetsbegränsningar. Varför använder man då inte alltid implicita metoder? [5p]

(c) Skriv en Matlabfunktion `heun` som numeriskt löser begynnelsevärdesproblemet

$$\begin{aligned} y' &= f(t,y), \\ y(t_0) &= 0; \end{aligned}$$

från t_0 till t_{final} med Heuns metod (den method som ni använde er av i tema 4). I denna uppgift antar vi att tillståndet y är endimensionellt. Funktionshuvudet ska vara:

```
function [t,y] = heun(f,tspan,y0,h)
%HEUN solve non-stiff differential equation with Heun's method
%   [T, Y] = HEUN(F,TSPAN,Y0,H) with TSPAN = [T0, TFINAL] integrates
%   the system of differential equations y' = f(t,y) from time T0 to
%   TFINAL with initial conditions Y0 employing a constant time step
%   H. F is a function handle. The solution vector Y holds the state
%   and each element corresponds to a time in output vector T.
```

[10p]

(d) Bestäm villkoret för stabilitet för Heuns metod när schemat tillämpas på problemet, $y' = \lambda y$, där $\lambda < 0$. [10p]

Svar (a) Definiera $p = du/dt$ och $\boldsymbol{y} = (u,p,v)^{\mathsf{T}}$. Då får vi att

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{y} = \frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} u \\ p \\ v \end{pmatrix} = \begin{pmatrix} p \\ v+t \\ tu \end{pmatrix} = \boldsymbol{f}(t,\boldsymbol{y}).$$

(b) Vid varje tidsteg krävs det många fler flyttalsberökningar för en implicit metod jämfört med motsvarande explicita metod. Den explicita metod kommer därför att vara effektivare än den implicita metoden utom i de fall då den explicita metodens stabilitetsvillkor är så restriktivt att de stabila tidsstegen är avsevärt mycket mindre än vad som är motiverat ur noggrannhetssynpunkt.

(c)
```
function [t,y] = heun(f,tspan,y0,h)
%HEUN solve non-stiff differential equation with Heun's method
%   [T, Y] = HEUN(F,TSPAN,Y0,H) with TSPAN = [T0, TFINAL] integrates
%   the system of differential equations y' = f(t,y) from time T0 to
%   TFINAL with initial conditions Y0 employing a constant time step
%   H. F is a function handle. The solution vector Y holds the state
%   and each element corresponds to a time in output vector T.
t = tspan(1):h:tspan(2);
y = zeros(size(t));
y(1) = y0;
for n = 1:length(t)-1
    kappa1 = f(t(n),y(n));
    kappa2 = f(t(n+1),y(n)+h*kappa1);
    y(n+1) = y(n) + h/2*(kappa1+kappa2);
end
```

(d) Heuns metod för $y' = \lambda y$, där $\lambda < 0$, blir

$$y_{n+1} = y_n + \Delta t \frac{1}{2}(\kappa_1+\kappa_2), \text{ där } \kappa_1 = f(t_n,y_n) = \lambda y_n \text{ och } \kappa_2 = f(t_{n+1}, y_n+\Delta t\kappa_1) = \lambda(y_n+\Delta_t \lambda y_n).$$

Alltså har vi att

$$y_{n+1} = y_n + \Delta t \frac{1}{2}\left(\lambda y_n + \lambda(y_n + \Delta t \lambda y_n)\right) = y_n + \Delta t \lambda y_n + \frac{1}{2}\Delta t^2 \lambda^2 y_n$$
$$= \frac{1}{2} y_n + \frac{1}{2} y_n \left(1 + 2\Delta t \lambda + \Delta t^2 \lambda^2\right) = y_n \left(\frac{1}{2} + \frac{1}{2}(1 + \Delta t \lambda)^2\right)$$

Metoden säges vara stabil om $|y_{n+1}| \leq |y_n|$, vilket i detta fall gäller om

$$1 \geq \left|\frac{1}{2} + \frac{1}{2}(1 + \Delta t \lambda)^2\right| = \frac{1}{2} + \frac{1}{2}(1 + \Delta t \lambda)^2,$$

vilket är sant om $(1 + \Delta t \lambda)^2 \leq 1$, d.v.s. om $-1 \leq 1 + \Delta t \lambda \leq 1$. Eftersom $\lambda < 0$ kan villkoret också skrivas $-1 \leq 1 - \Delta t |\lambda| \leq 1$. Den högra olikheten är uppfylld för alla Δt (som alltid är positiv; vi tar bara positiva tidssteg!) medan den vänstra olikheten är uppfylld för $\Delta t \leq 2/\lambda$, vilket är metodens stabilitetsvillkor för $\lambda < 0$.