Litt mer om differensligninger og avrundingsfeil

Knut Mørken

16. oktober 2003

1 Innledning

Seksjon 4.2 i kompendiet for MAT-INF 1100 dreier som om simulering av differensligninger og hvordan avrundingsfeil kan ødelegge beregningene. Ligningen vi ser på der er

$$x_n = \frac{10}{3}x_{n-1} - x_{n-2} \qquad \text{med } x_0 = 1 \text{ og } x_1 = 1/3 \tag{1}$$

som har den eksakte løsningen $x_n = 1/3^n$. Når vi simulerer dette på datamaskin med 64-bits flyttall (`double`-variable i Java) får vi til å begynne med oppførselen som vi forventer og x_n avtar pent og pyntelig, men fra og med $n = 18$ begynner så verdiene å stige igjen. I kompendiet påstår jeg at årsaken er at vi på grunn av avrundingsfeil ikke får den eksakte løsningen, men heller en løsning på formen

$$z_n = (1 + \epsilon_1)\frac{1}{3^n} + \epsilon_2\, 3^n \tag{2}$$

der ϵ_1 og ϵ_2 er små tall. Når n blir stor vil åpenbart det andre leddet dominere så sant $\epsilon_2 > 0$, selv om den er aldri så liten, siden ϵ_2 er fast mens 3^n bare blir større og større når n vokser.

Noen har påpekt at det ikke er noen grunn til at løsningen vi genererer ved simulering skal være på formen (2), noe som i og for seg er helt riktig. I kompendiet presenterer jeg dette som en forklaring uten å gjøre noe forsøk på å motivere det. Her skal vi se litt nøyere på fenomenet, men en fullstendig forklaring går utover det som er naturlig å gi seg i kast med i et kurs som MAT-INF 1100.

2 Modellen stemmer

La oss først se om det er riktig at løsningen vi genererer ved simulering stemmer godt overens med formen gitt ved (2). Hvis vi betegner de simulerte

verdiene med $\{\tilde{x}_n\}$ så finner vi at $\tilde{x}_0$ og $\tilde{x}_1$ er helt riktige (inntil maskinnøyaktighet) mens vi har

$$\tilde{x}_2 = 0.1111111111111112, \qquad x_2 = 1/9 \approx 0.1111111111111111,$$
$$\tilde{x}_3 = 0.03703703703703725, \qquad x_3 = 1/27 \approx 0.03703703703703704.$$

La oss nå forsøke å bestemme ϵ_1 og ϵ_2 i (2). Siden vi har to ukjente størrelser er det rimelig å kreve at $\tilde{x}_n = z_n$ for to verdier av n. Siden de to første simulerte verdiene stemmer med de eksakte verdiene med full flyttallsnøyaktighet må vi minst gå ut til $n = 2$. Vi kunne bruke de to verdiene for $n = 2$ og $n = 3$, men det viser seg at det ikke fungerer så bra, så vi velger å bruke $n = 3$ og $n = 20$ slik at vi krever $z_2 = \tilde{x}_2$ og $z_{20} = \tilde{x}_{20}$. Dette gir oss to ligninger med de to ukjente ϵ_1 og ϵ_2 som har løsningen

$$\epsilon_1 = -1.976017040613457 \cdot 10^{-16}, \tag{3}$$
$$\epsilon_2 = 8.60743290299588 \cdot 10^{-18} \tag{4}$$

(her har vi løst ligningssystemet med flyttall så det er nok en liten avrundingsfeil). Når vi nå har bestemt de to ϵ'ene kan vi se om z_n stemmer overens med de simulerte verdiene $\tilde{x}_n$ slik jeg har påstått. Noen tilfeldige verdier er

$$\tilde{x}_3 = 0.03703703703703725, \qquad z_3 = 0.03703703703703725,$$
$$\tilde{x}_6 = 0.001371742112489128, \qquad z_6 = 0.001371742112489128,$$
$$\tilde{x}_{12} = 1.88168099750167 \cdot 10^{-6}, \qquad z_{12} = 1.88168099750167 \cdot 10^{-6},$$
$$\tilde{x}_{25} = 7.29298103548869 \cdot 10^{-6}, \qquad z_{25} = 7.29298103548869 \cdot 10^{-6},$$
$$\tilde{x}_{100} = 4.436077429413149 \cdot 10^{30}, \qquad z_{100} = 4.436077429413155 \cdot 10^{30},$$

slik de regnes ut med flyttall. Som vi ser er det svært god overenstemmelse mellom verdiene, og vi må kunne konkludere med at følgen $\{z_n\}$ gitt ved (2) med ϵ_1 og ϵ_2 gitt ved (3) beskriver meget godt de tallene som vi beregner når differensligningen (1) simuleres med 64-bits flyttall.

3 Hvorfor stemmer modellen?

Overenstemmelsen mellom den simulerte følgen $\{\tilde{x}_n\}$ og modellen gitt ved (2) er så god at vi tydeligvis må ha fått med oss den vesentlige oppførselen i beregningene, men hvorfor er det tilfelle? Det vil føre for langt å gå inn på et formelt bevis for dette, men det er ikke så vanskelig å få en liten ide om at det er rimelig.

Det vi ønsker å regne ut i n'te iterasjon er $x_n = 10x_{n-1}/3 - x_{n-2}$. Når vi regner med flyttall kan vi ikke representere konstanten 10/3 eksakt, isteden får vi et tall $b = \frac{10}{3}(1+\delta_n)$ der δ_n er et tall av størrelsesorden 10^{-16} som angir den relative feilen i tilnærmingen til 10/3. Når vi så multipliserer b med x_{n-1} og så subtraherer x_{n-2} gjør vi også en liten feil slik at resultatet blir

$$\tilde{x}_n = \Big(\frac{10}{3}(1+\delta_n)\tilde{x}_{n-1} - \tilde{x}_{n-2}\Big)(1+\gamma_n)$$

der γ_n også er et lite tall av samme størrelsesorden som δ_n som angir den relative feilen i multiplikasjonen og subtraksjonen (vi kunne tatt med relativ feil i hver av disse to operasjonene, men ville ikke tjene noe på det i et kvalitativt resonnement). Skriver vi dette ut litt tydeligere så har vi

$$\tilde{x}_n = \frac{10}{3}(1+\delta_n)(1+\gamma_n)\tilde{x}_{n-1} - (1+\gamma_n)\tilde{x}_{n-2} \tag{5}$$

Koeffisienten 10/3 har altså blitt erstattet med $(10/3)(1+\delta_n)(1+\gamma_n)$ mens koeffisienten -1 har blitt erstattet med $-(1+\gamma_n)$. Siden

$$\frac{10}{3}(1+\delta_n)(1+\gamma_n) = \frac{10}{3}\big(1 + (\delta_n + \gamma_n) + \delta_n\gamma_n\big)$$

og både δ_n og γ_n er små så ser vi at koeffisientene vi bruker er svært nær de korrekte. Det betyr at den karakteristiske ligningen for våre beregninger er

$$r^2 - \frac{10}{3}(1+\delta_n)(1+\gamma_n)r + (1+\gamma_n) = 0 \tag{6}$$

som er svært nær den korrekte ligningen $r^2 - 10\,r/3 + 1 = 0$. Det er da mulig å vise at røttene $\tilde{r}_1$ og $\tilde{r}_2$ i (6) må ligge nær røttene 1/3 og 3 til den opprinnelige differensligningen slik at vi har

$$\tilde{r}_1 = \frac{1}{3} + \alpha,$$
$$\tilde{r}_2 = 3 + \beta,$$

der α og β er av samme størrelsesorden som δ_n og γ_n. Den generelle løsningen av (5) vil derfor være

$$\tilde{x}_n = C_1\Big(\frac{1}{3} + \alpha\Big)^n + C_2(3+\beta)^n.$$

Hvis vi tilpasser koeffisientene C_1 og C_2 til startverdiene $x_0 = 1$ og $x_1 = 1/3$, regnet om til nærmeste flyttall ($x_0 = 1$ klarer vi eksakt), får vi en løsning som vil være

$$\tilde{x}_n = (1+\tilde{\epsilon}_1)\Big(\frac{1}{3} + \alpha\Big)^n + \tilde{\epsilon}_2(3+\beta)^n$$

der $\tilde{\epsilon}_1$ og $\tilde{\epsilon}_2$ begge er av samme størrelsesorden som α og β.

I dette siste uttrykket har vi nå fire tilnærminger i forhold til den eksakte løsningen $x_n = 1/3^n$: Den første roten $1/3$ har blitt tilnærmet med $1/3 + \alpha$, den andre roten 3 med $3 + \beta$, den første koeffisienten 1 har blitt tilnærmet med $1 + \tilde{\epsilon}_1$ og den andre koeffisienten 0 har blitt tilnærmet med ϵ_2. Den eneste av disse tilnærmingene som er kritisk er den siste fordi det er den eneste som gir en stor relativ feil. Dette betyr at i forhold til denne kan vi godt ignorere de andre og si at

$$\tilde{x}_n \approx \frac{1}{3^n} + \tilde{\epsilon}_2\, 3^n.$$

Vi ser at dette minner om uttrykket (2) for z_n bortsett fra at vi der har tatt med en tilnærmet koeffisient foran $1/3^n$ også. Argumentet vi nettopp har gjennomgått viser at det faktisk ikke er nødvendig, men fra et pedagogisk synspunkt synes jeg det var enklest å bruke formen i (2).