

Kapitel 6

Newton's metod

6.1 Numerisk beräkning av derivata

Låt $f : I \rightarrow \mathbf{R}$ vara en deriverbar funktion (I är ett intervall). Derivatans definition är

$$(6.1) \quad f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}.$$

Det betyder att vi kan approximera derivatan med en differenskvot:

$$(6.2) \quad f'(x) \approx \frac{f(x+h) - f(x)}{h} \quad \text{med } h \approx 0.$$

Vi ska nu diskutera felet i denna approximation. Vi måste då ta hänsyn till att datorn beräknar med ändlig precision, dvs räknar med ändligt många siffror och att detta leder till avrundningsfel. Datorn beräknar alltså en approximation

$$(6.3) \quad \tilde{f}(x) \approx f(x)$$

där avrundningsfelet är

$$(6.4) \quad e_f(x) = \tilde{f}(x) - f(x)$$

med begränsningen

$$(6.5) \quad |e_f(x)| = |\tilde{f}(x) - f(x)| \leq \delta_f, \quad x \in I.$$

I MATLAB, som räknar med cirka 16 siffror, och om $|f(x)| \approx 1$, kan vi antaga att $\delta_f \approx 10^{-15}$. Det vi beräknar är alltså

$$(6.6) \quad f'(x) \approx \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} \quad \text{med } h \approx 0.$$

Felet är

$$\begin{aligned} \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} - f'(x) &= \frac{f(x+h) - f(x)}{h} - f'(x) + \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} - \frac{f(x+h) - f(x)}{h} \\ &= \left(\frac{f(x+h) - f(x)}{h} - f'(x) \right) + \frac{e_f(x+h) - e_f(x)}{h}. \end{aligned}$$

Triangelolikheten ger

$$(6.7) \quad \left| \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} - f'(x) \right| \leq \left| \frac{f(x+h) - f(x)}{h} - f'(x) \right| + \left| \frac{e_f(x+h) - e_f(x)}{h} \right|.$$

Den första termen är diskretiseringsfelet och den andra är avrundningsfelet. För avrundningsfelet använder vi (6.5)

$$(6.8) \quad \left| \frac{e_f(x+h) - e_f(x)}{h} \right| \leq \frac{|e_f(x+h)| + |e_f(x)|}{h} \leq \frac{2\delta_f}{h}.$$

För diskretiseringsfelet använder vi en formel för linjäriseringsfelet (se Definition 8 och Theorem 11 i Adams 4.9)

$$(6.9) \quad f(x) = f(a) + f'(a)(x-a) + E(x) \quad \text{med felet } E(x) = \frac{1}{2}f''(s)(x-a)^2,$$

där s är en (obekant) punkt mellan x och a . Vi använder denna genom att byta ut x mot $x+h$ och a mot x :

$$(6.10) \quad f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(s)h^2 \quad \text{med } s \text{ mellan } x \text{ och } x+h.$$

Genom att dividera med h får vi

$$(6.11) \quad \left| \frac{f(x+h) - f(x)}{h} - f'(x) \right| = \frac{1}{2}|f''(s)|h.$$

Eftersom s är okänd måste vi antaga att vi har en begränsning för f'' ,

$$(6.12) \quad |f''(x)| \leq K_f, \quad x \in I.$$

Vi får då:

$$(6.13) \quad \left| \frac{f(x+h) - f(x)}{h} - f'(x) \right| = \frac{1}{2}|f''(s)|h \leq \frac{1}{2}K_fh.$$

För totala felet i (6.7) får vi med (6.13) och (6.8):

$$(6.14) \quad \left| \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} - f'(x) \right| \leq \frac{1}{2}K_fh + \frac{2\delta_f}{h}.$$

Detta gäller för alla $x \in I$ och $x+h \in I$. Vi vill välja h så att felet blir minimalt. Vi ser att den första termen i $\frac{1}{2}K_fh + \frac{2\delta_f}{h}$ minskar då h minskar medan den andra ökar. Vi får minimum då båda är lika:

$$(6.15) \quad \frac{1}{2}K_fh = \frac{2\delta_f}{h},$$

dvs

$$(6.16) \quad h^2 = \frac{4\delta_f}{K_f}, \quad h = 2\sqrt{\frac{\delta_f}{K_f}}.$$

(Man kan också visa detta genom att derivera $\frac{1}{2}K_fh + \frac{2\delta_f}{h}$ med avseende på h och sätta derivatan = 0.) Minimala felet blir då

$$(6.17) \quad \left| \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} - f'(x) \right| \leq \frac{1}{2}K_fh + \frac{2\delta_f}{h} = K_fh = 2\sqrt{K_f}\sqrt{\delta_f}.$$

I MATLAB med $\delta_f \approx 10^{-15}$ och $|f'(x)| \approx 1$ och $K_f \approx 1$ (till exempel) får vi ungefär

$$(6.18) \quad h \approx 2\sqrt{\delta_f} \approx 2 \cdot 10^{-7.5} \approx 10^{-7},$$

$$(6.19) \quad \left| \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} - f'(x) \right| \approx 2\sqrt{\delta_f} \approx 10^{-7}.$$

Dvs vi får ungefär 7 korrekta decimaler.

En bättre approximation

Vi får en bättre approximation om vi ersätter den ensidiga differenskvoten i (6.2) med den symmetriska differenskvoten

$$(6.20) \quad f'(x) \approx \frac{f(x+h) - f(x-h)}{2h} \quad \text{med } h \approx 0.$$

Det totala felet blir nu istället för (6.7)

$$(6.21) \quad \left| \frac{\tilde{f}(x+h) - \tilde{f}(x-h)}{2h} - f'(x) \right| \leq \left| \frac{f(x+h) - f(x-h)}{2h} - f'(x) \right| + \left| \frac{e_f(x+h) - e_f(x-h)}{2h} \right|.$$

Man kan visa att detta begränsas av

$$(6.22) \quad \left| \frac{\tilde{f}(x+h) - \tilde{f}(x-h)}{2h} - f'(x) \right| \leq \frac{1}{6} M_f h^2 + \frac{\delta_f}{h},$$

där δ_f är som förut och M_f är en begränsning av f''' :

$$(6.23) \quad |f'''(x)| \leq M_f, \quad x \in I.$$

Vi gör en överslagsberäkning baserad på antagandena $\delta_f \approx 10^{-15}$, $M_f \approx 1$. Feluppskattningen i (6.22) blir approximativt minimum då båda termerna är approximativt lika:

$$(6.24) \quad h^2 \approx \frac{\delta_f}{h}, \quad h^3 \approx \delta_f, \quad h \approx \delta_f^{1/3} \approx 10^{-5},$$

och då blir minimala felet ungefär

$$(6.25) \quad \left| \frac{\tilde{f}(x+h) - \tilde{f}(x-h)}{2h} - f'(x) \right| \approx h^2 \approx \delta_f^{2/3} \approx 10^{-10}.$$

Jämfört med (6.19) har vi cirka 3 decimaler noggrannare approximation av derivatan. Men man ska komma ihåg att (6.19) och (6.25) beror också på K_f och M_f , vilket kan påverka jämförelsen om någon av dessa är stor. (Exakt värde i (6.24) är $h = (3\delta_f/M_f)^{1/3}$ och i (6.25) $cM_f^{1/3}\delta_f^{2/3}$ med $c \approx 1$.)

Bevis av (6.22)

(Överkurs, kan skippas.) Avrundningsfelet uppskattas som förut i (6.8). För diskretiseringsfelet använder vi Taylors formel (se Theorem 12 i Adams 4.10, som vi kommer att studera i lp 2)

$$(6.26) \quad f(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + E(x) \quad \text{med felet } E(x) = \frac{1}{6}f'''(s)(x-a)^3,$$

där s är en (obekant) punkt mellan x och a . Vi använder denna genom att byta ut x mot $x+h$ och a mot x :

$$(6.27) \quad f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + \frac{1}{6}f'''(s_1)h^3,$$

$$(6.28) \quad f(x-h) = f(x) + f'(x)(-h) + \frac{1}{2}f''(x)(-h)^2 + \frac{1}{6}f'''(s_2)(-h)^3,$$

med s_1 mellan x och $x+h$ och s_2 mellan $x-h$ och x . Detta ger

$$(6.29) \quad \frac{f(x+h) - f(x-h)}{2h} = f'(x) + \frac{1}{12}f'''(s_1)h^2 + \frac{1}{12}f'''(s_2)h^2.$$

Diskretiseringsfelet blir

$$(6.30) \quad \left| \frac{f(x+h) - f(x-h)}{2h} - f'(x) \right| \leq \frac{1}{12}|f'''(s_1)|h^2 + \frac{1}{12}|f'''(s_2)|h^2.$$

Eftersom s_1, s_2 är okända måste vi antaga att vi har en begränsning för f''' ,

$$(6.31) \quad |f'''(x)| \leq M_f, \quad x \in I.$$

Vi får då:

$$(6.32) \quad \left| \frac{f(x+h) - f(x-h)}{2h} - f'(x) \right| \leq \frac{1}{6} M_f h^2.$$

Detta visar (6.22).

6.2 Newtons metod

(Adams 4.2) Låt $f : I \rightarrow \mathbf{R}$ vara en deriverbar funktion (I är ett intervall). Vi erinrar oss att ekvationen $f(x) = 0$ kan skrivas som en fixpunktsekvation $x = g(x)$ genom omskrivningen

$$x = x - \alpha(x)f(x),$$

dvs med

$$g(x) = x - \alpha(x)f(x).$$

Det gäller att hitta $\alpha(x)$ så att $g : I \rightarrow I$ blir en kontraktion. Då konvergerar fixpunktsiterationen

$$(6.33) \quad x_{k+1} = g(x_k).$$

Detta är inte lätt att åstadkomma, men Newtons metod ger oss ett systematiskt sätt att göra detta.

Antag att vi har en approximativ lösning x_k och vi vill hitta en bättre approximation x_{k+1} som i (6.33). Vi bildar linjäriseringen av funktionen f i x_k (Definition 8 i Adams 4.9):

$$(6.34) \quad L(x) = f(x_k) + f'(x_k)(x - x_k)$$

och löser $L(x) = 0$ istället för $f(x) = 0$. Dvs

$$(6.35) \quad f(x_k) + f'(x_k)(x - x_k) = 0$$

med lösningen

$$(6.36) \quad x = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Detta får bli nästa approximation:

$$(6.37) \quad x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Detta är på formen (6.33) med

$$g(x) = x - \frac{f(x)}{f'(x)}, \quad \text{dvs } \alpha(x) = \frac{1}{f'(x)}.$$

Iterationen (6.37) kallas Newtons metod (eller Newton-Raphsons metod). För att Newtons metod ska fungera måste $f'(x_k) \neq 0$. Vi antar därför att

$$f'(x) \neq 0 \quad \text{för alla } x \in I.$$

Kom ihåg att ekvationen för tangenten till grafen $y = f(x)$ i $(x_k, f(x_k))$ är $y = L(x)$, dvs

$$y = f(x_k) + f'(x_k)(x - x_k).$$

Geometriskt betyder (6.35) att vi följer tangenten och hittar x_{k+1} där denna skär x -axeln.

Exempel 1. Med $f(x) = x^2 - 2$ får vi

$$g(x) = x - \frac{x^2 - 2}{2x} = x - \frac{1}{2}x + \frac{1}{x} = \frac{1}{2}x + \frac{1}{x}.$$

Newtons iteration blir

$$x_{k+1} = \frac{1}{2}x_k + \frac{1}{x_k}.$$

Vi har tidigare sett att g är en kontraktion med $L = 1/2$ på $[1, 2]$ och denna iteration konvergerar snabbt mot $\sqrt{2}$ om vi väljer $x_0 \in [1, 2]$. $\square$

Antag nu att $\bar{x} \in I$ är en rot, dvs

$$f(\bar{x}) = 0.$$

Om vi deriverar $g(x)$ får vi

$$g'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2},$$

så att

$$g'(\bar{x}) = \frac{f(\bar{x})f''(\bar{x})}{f'(\bar{x})^2} = 0$$

och $g'(x) \approx 0$ på varje litet intervall nära roten. Det betyder att g har en mycket liten Lipschitz-konstant på varje sådant intervall. Det innebär att Newtons metod konvergerar mycket snabbt.

Följande sats handlar om detta. Det är Theorem 2 i Adams 4.2.

Sats 1. Antag att $x_k, x_{k+1}, \bar{x} \in I$ och att vi har begränsningarna

$$(6.38) \quad \frac{1}{|f'(x)|} \leq M, \quad \forall x \in I,$$

$$(6.39) \quad |f''(x)| \leq K, \quad \forall x \in I.$$

Då gäller

$$(6.40) \quad |x_k - \bar{x}| \leq M|f(x_k)|,$$

$$(6.41) \quad |x_{k+1} - \bar{x}| \leq \frac{1}{2}MK|x_{k+1} - x_k|^2,$$

$$(6.42) \quad |x_{k+1} - \bar{x}| \leq \frac{1}{2}MK|x_k - \bar{x}|^2.$$

I Adams skrivs (6.38) som $|f'(x)| \geq L$, så att $M = 1/L$.

Begränsningarna (6.38) och (6.39) innebär att tangentens lutning inte får vara alltför nära 0 och inte ändra sig för mycket på intervallet I .

Feluppskattningen (6.40) relaterar felet i x till felet i $f(x)$, dvs relaterar felet $x_k - \bar{x}$ till residualen $f(x_k)$. Residualen anger hur väl x_k uppfyller ekvationen $f(x) = 0$. Om $f'(x)$ är liten på I så blir M stor (L liten) och då kan felet vara stort även om residualen $f(x_k)$ är liten. Det betyder att ekvationen är svår att lösa om $f'(x)$ är liten.

De andra uppskattningarna visar att x_k konvergerar mycket snabbt om den konvergerar alls. Till exempel, om x_k har kommit så nära roten att $|x_k - \bar{x}| < 1$, så ser vi i (6.42) att felet i nästa steg är proportionellt mot kvadraten på det tidigare felet, dvs det minskar mycket snabbt. Vi säger att felet konvergerar kvadratisk mot 0. Vi erinrar oss att bisektionsmetoden och fixpunktsiterationen i allmänhet bara konvergerar linjärt.

En nackdel med Newtons metod är att det är svårt att ange ett intervall I där ovanstående begränsningar gäller och sådant att iterationen stannar kvar i intervallet. I praktiken får man

oftast nöja sig med att säga att metoden konvergerar om man väljer x_0 tillräckligt nära en rot och prova sig fram med olika x_0 .

Beviset kan skippas om du inte orkar mer teori.

Bevis. Medelvärdessatsen (Theorem 13 i Adams 2.8) ger

$$(6.43) \quad f(x_k) - f(\bar{x}) = f'(s_1)(x_k - \bar{x}),$$

där $f(\bar{x}) = 0$ och s_1 är en punkt mellan x_k och $\bar{x}$. Vi får med hjälp av begränsningen (6.38)

$$(6.44) \quad |x_k - \bar{x}| = \frac{|f(x_k)|}{|f'(s_1)|} \leq M|f(x_k)|,$$

vilket är (6.40). Linjärisering med felterm (Theorem 11 i Adams 4.9) ger

$$(6.45) \quad f(x_{k+1}) = f(x_k) + f'(x_k)(x_{k+1} - x_k) + \frac{1}{2}f''(s_2)(x_{k+1} - x_k)^2 = \frac{1}{2}f''(s_2)(x_{k+1} - x_k)^2$$

där s_2 är en punkt mellan x_k och x_{k+1} , och där vi använde

$$f(x_k) + f'(x_k)(x_{k+1} - x_k) = 0$$

från (6.37). Tillsammans med (6.40) och begränsningen (6.39) ger detta

$$(6.46) \quad |x_{k+1} - \bar{x}| \leq M|f(x_{k+1})| = \frac{1}{2}M|f''(s_2)|(x_{k+1} - x_k)^2 \leq \frac{1}{2}MK(x_{k+1} - x_k)^2,$$

vilket är (6.41). Linjärisering med felterm ger även

$$(6.47) \quad f(\bar{x}) = f(x_k) + f'(x_k)(\bar{x} - x_k) + \frac{1}{2}f''(s_3)(\bar{x} - x_k)^2,$$

där s_3 är en punkt mellan x_k och $\bar{x}$, och där $f(\bar{x}) = 0$ så att

$$(6.48) \quad x_k - \bar{x} = \frac{f(x_k)}{f'(x_k)} + \frac{1}{2} \frac{f''(s_3)}{f'(x_k)}(x_k - \bar{x})^2.$$

Enligt (6.37) har vi

$$(6.49) \quad \frac{f(x_k)}{f'(x_k)} = -(x_{k+1} - x_k),$$

så att

$$(6.50) \quad x_k - \bar{x} = -(x_{k+1} - x_k) + \frac{1}{2} \frac{f''(s_3)}{f'(x_k)}(x_k - \bar{x})^2,$$

vilket leder till

$$(6.51) \quad x_{k+1} - \bar{x} = \frac{1}{2} \frac{f''(s_3)}{f'(x_k)}(x_k - \bar{x})^2$$

och sedan

$$(6.52) \quad |x_{k+1} - \bar{x}| = \frac{1}{2} \frac{|f''(s_3)|}{|f'(x_k)|}(x_k - \bar{x})^2 \leq \frac{1}{2}MK(x_k - \bar{x})^2,$$

vilket är (6.42). □

När stoppar man?

Vi vill stoppa iterationen när felet är mindre än en given tolerans:

$$(6.53) \quad |x_k - \bar{x}| \leq \text{TOL}.$$

Feluppskattningarna i satsen är inte användbara eftersom vi inte vill använda de svårbestämda och grova begränsningarna M och K . Men (6.50) ger

$$x_k - \bar{x} = -(x_{k+1} - x_k) + \frac{1}{2} \frac{f''(s_3)}{f'(x_k)} (x_k - \bar{x})^2,$$

så att

$$|x_k - \bar{x}| \leq |x_{k+1} - x_k| + \frac{1}{2} MK |x_k - \bar{x}|^2.$$

Om iterationen konvergerar så är den sista termen snart mycket mindre än de andra termerna och vi har

$$(6.54) \quad |x_k - \bar{x}| \approx |x_{k+1} - x_k|.$$

Stoppvilkoret

$$(6.55) \quad |x_{k+1} - x_k| \leq \text{TOL}$$

garanterar alltså (6.53) approximativt. Det betyder att vi accepterar x_k om ändringen i nästa iteration är mindre än toleransen.

Algoritmen

Algoritmen är mycket enkel.

```
while |h| > TOL
beräkna residualen:  b = -f(x)
beräkna derivatan:  a = f'(x)
beräkna ändringen:  h = b/a
uppdatera x:  x = x + h
```

Derivatan beräknas lämpligen numeriskt. Newtons metod är nämligen okänslig för fel i derivatan.