Kapitel 6. Metoder att lösa olinjära problem

Olinjära fenomen är mycket vanliga i naturen och nuförtiden föremål för ett intensivt studium. Vi behöver bara tänka t.ex. på energikonsumtionen, eller befolkningsutvecklingen, som följer en exponentiell lag. Om vi har populationsmodeller med flere komponenter, får vi ett system av olinjära ekvationer, som inte går att lösa med de enkla metoder för linjära ekvationssystem, som vi behandlat tidigare. Olinjära ekvationer behöver inte ha entydiga lösningar, eller ens en lösning i sluten form.

Vi har redan tidigare studerat *iterativa* metoder för att lösa en olinjär ekvation. Om vi har ett olinjärt ekvationssystem, blir det givetvis ännu svårare att finna en approximativ lösning. Ofta är funktionen mycket komplicerad, varför räknemetoden planeras så att beräkningen konvergerar med ett så litet antal iterationer som möjligt.

6.1. Lösning av olinjära ekvationssystem

Betrakta ett system av n olinjära ekvationer med n obekanta:

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0 \\ f_2(x_1, \ldots, x_n) = 0 \\ \qquad\qquad \ldots \\ f_n(x_1, \ldots, x_n) = 0 \end{cases}$$

Om vi inför beteckningarna $x = (x_1, \ldots, x_n)$ och $f(x) = (f_1(x), \ldots, f_n(x))$ så kan ekvationssystemet skrivas som *en* olinjär ekvation:

$$f(x) = 0.$$

Detta gör det möjligt att tillämpa liknande resonemang som vid lösningen av rötter till olinjära ekvationer i en variabel. Analogt med behandlingen av Newtons metod för en enkel olinjär ekvation kan vi i det allmänna fallet serieutveckla $f(x)$ i en Taylors serie kring en approximation $x^{(k)}$ till systemets lösning x^*:

$$f(x^{(k)} + h) = f(x^{(k)}) + J(x^{(k)})h + \ldots.$$

Här betecknar $J(x^{(k)}) = \nabla f(x^{(k)})$ jakobianen av f, tagen i punkten $x^{(k)}$, som är en matris, vars element kan uttryckas

$$J(x^{(k)})_{ij} = \left(\frac{\partial f_i(x)}{\partial x_j} \right)_{x=x^{(k)}}$$

För enkelhetens skull skall vi använda beteckningarna $J^{(k)} \equiv J(x^{(k)})$ och $f^{(k)} \equiv f(x^{(k)})$. Om man försummar alla termer av högre ordning än den första i Taylorserien för f, får vi den linjära funktionsapproximationen

$$f(x^{(k)} + h) \approx y = f^{(k)} + J^{(k)} h.$$

En ny approximation till lösningen finner vi genom att sätta $y = 0$:

$$J^{(k)} h = -f^{(k)},$$

lösa vektorn h ur detta ekvationssystem, samt beräkna $x^{(k+1)} = x^{(k)} + h$, dvs

$$x^{(k+1)} = x^{(k)} - \left(J^{(k)} \right)^{-1} f^{(k)}.$$

Detta är Newtons metod för att lösa ett olinjärt ekvationssystem.

För enkelhetens skull skall vi börja med att studera ett ekvationssystem med två obekanta:

$$\begin{cases} f_1(x,y) = 0 \\ f_2(x,y) = 0 \end{cases}$$

Om vi utvecklar funktionerna i Taylors serie och medtar endast första ordningens termer, får vi

$$f_1(x+h, y+k) \approx f_1(x,y) + hf_{1x}(x,y) + kf_{1y}(x,y)$$
$$f_2(x+h, y+k) \approx f_2(x,y) + hf_{2x}(x,y) + kf_{2y}(x,y)$$

Genom att sätta högra membrum av dessa ekvationer lika med noll får vi lösningarna

$$h = \frac{-f_1 f_{2y} + f_2 f_{1y}}{f_{1x} f_{2y} - f_{1y} f_{2x}}$$
$$k = \frac{f_1 f_{2x} - f_2 f_{1x}}{f_{1x} f_{2y} - f_{1y} f_{2x}}$$

där alla funktionsvärden och derivator beräknats i punkten (x, y).

För ett ekvationssystem med två obekanta kan Newtons metod således uttryckas på följande sätt:

$$x_{n+1} = x_n + h; \qquad y_{n+1} = y_n + k,$$

där h och k beräknas ur formlerna ovan.

Som ett enkelt exempel skall vi tillämpa Newtons metod på skärningspunkten mellan ellipsen $4x^2+y^2 = 4$ och kurvan $x^2y^3 = 1$. Ett sätt att lösa detta ekvationssystem skulle vara att eliminera x^2 ur ekvationerna, vilket leder till en femtegradens ekvation, varur y kan lösas. Denna metod skall inte användas här eftersom avsikten är att visa hur Newtons metod kan tillämpas på ett ekvationssystem.

En av lösningarna är nära punkten $(0.4, 1.8)$, som vi därför kan använda som utgångspunkt. De partiella derivatorna av funktionerna $f_1(x, y) = 4x^2 + y^2 - 4$ och $f_2(x, y) = x^2y^3 - 1$ är

$$\begin{aligned} f_{1x} &= 8x\ , \qquad f_{1y} = 2y \\ f_{2x} &= 2xy^3, \quad f_{2y} = 3x^2y^2 \end{aligned}$$

I punkten $x_0 = (0.4, 1.8)$ är funktionsvärdena och derivatorna

$$\begin{aligned} f_1 &= 4(0.4)^2 + 1.8^2 - 4 = -0.12, \qquad f_{1x} = 3.2 \qquad f_{1y} = 3.6 \\ f_2 &= (0.4)^2(1.8)^3 - 1 = -0.06688, \quad f_{2x} = 4.6656 \quad f_{2y} = 1.5552 \end{aligned}$$

Således får vi $h = 4.5808967 \cdot 10^{-3}$ och $k = 2.9261425 \cdot 10^{-2}$, samt därav $x_1 = 0.4045809$ och $y_1 = 1.8292614$.

Upprepade iterationer ger approximationerna

0.404149564420627	1.82938603854765
0.404149457020688	1.82938592581215
0.404149457020644	1.82938592581218

Konvergensen är som synes mycket rask. Den andra lösningen får man lika enkelt genom att välja utgångspunkten $(0.8, 1.2)$.

Newtons metod konvergerar kvadratiskt, och sålunda snabbt (om den konvergerar). Å andra sidan måste man lösa ett linjärt ekvationssystem vid varje iteration, och beräkna n^2 partiella derivator, vilket kan vara tidskrävande. Derivatorna kan emellertid beräknas medels numeriska approximationer, eller också kan man använda sekantmetoden, i vilket fall iterationerna dock inte längre konvergerar kvadratiskt.

Ifall jakobianen är nästan singulär, uppstår liknande problem som i det endimensionella fallet, och iterationerna bör därför kontrolleras noga.

Ett sätt att göra detta är att kräva att

$$\|f(x^{(k+1)})\|_2 < \|f(x^{(k)})\|_2$$

är uppfyllt för alla värden av k. På detta sätt kan man garantera att man närmar sig ett nollställe av $f(x)$.

För att en approximation inte skall skilja sig alltför mycket från en tidigare approximation (vilket kan leda till oscillationer, eller t.o.m. divergens) kan man fordra, att $\|h\|_2 \leq \delta$, där δ är någon lämplig övre gräns för iterationssteget. Denna övre gräns anger huru pass tillförlitlig den linjära approximationen $f(x^{(k)} + h) \approx f(x^{(k)}) + J^{(k)}h$ kan anses vara. Om approximationen är bra, kan man använda ett stort värde av δ, men om den är dålig, duger bara ett litet värde. Mängden $\{h : \|h\|_2 \leq \delta\}$ brukar kallas tillförlitlighetsområdet.

Ibland går det inte att uppfylla båda villkoren, utan man måste kompromissa. Ett sätt är att avstå från kravet att h skall vara en lösning till $Jh = -f$, men bibehålla villkoret $\|h\|_2 \leq \delta$. Man uppställer m.a.o. villkoret : Minimera $\|J^{(k)}h + f^{(k)}\|_2$ under antagandet att $\|h\|_2 \leq \delta$.

En algoritm för att lösa detta problem kan uttryckas på följande sätt:

1. Avsluta räkningen, om $f(x_k) \approx 0$.
2. Beräkna h genom att lösa det begränsade minimeringsproblemet ovan.
3. Om $\|f(x^{(k)} + h)\|_2 \leq \|f(x^{(k)})\|_2$ godkänns h, och man kan välja $x^{(k+1)} \leftarrow x^{(k)} + h$, och $k \leftarrow k + 1$, samt gå till steg 1.
4. I annat fall minskas gränsen δ, och man väljer därpå $x^{(k+1)} \leftarrow x^{(k)}$, och $k \leftarrow k + 1$, samt går till steg 1.

Om $\|h\|_2$ är tillräckligt litet, kommer Taylorseriens två första termer att vara en god approximation till funktionsvärdet i en omgivning av $f(x^{(k)})$, varför man vanligen endast behöver minska δ några få gånger. Med en sådan algoritm kan man garantera endast en lokal lösning till ekvationssystemet, inte en global lösning. Detta skall vi studera närmare i nästa avsnitt.

6.2. Allmänt om minimering av funktioner

Antag, att f är en reellvärd funktion av n reella variabler (eller parametrar):

$$f(x_1, x_2, \ldots, x_n) \equiv f(x).$$

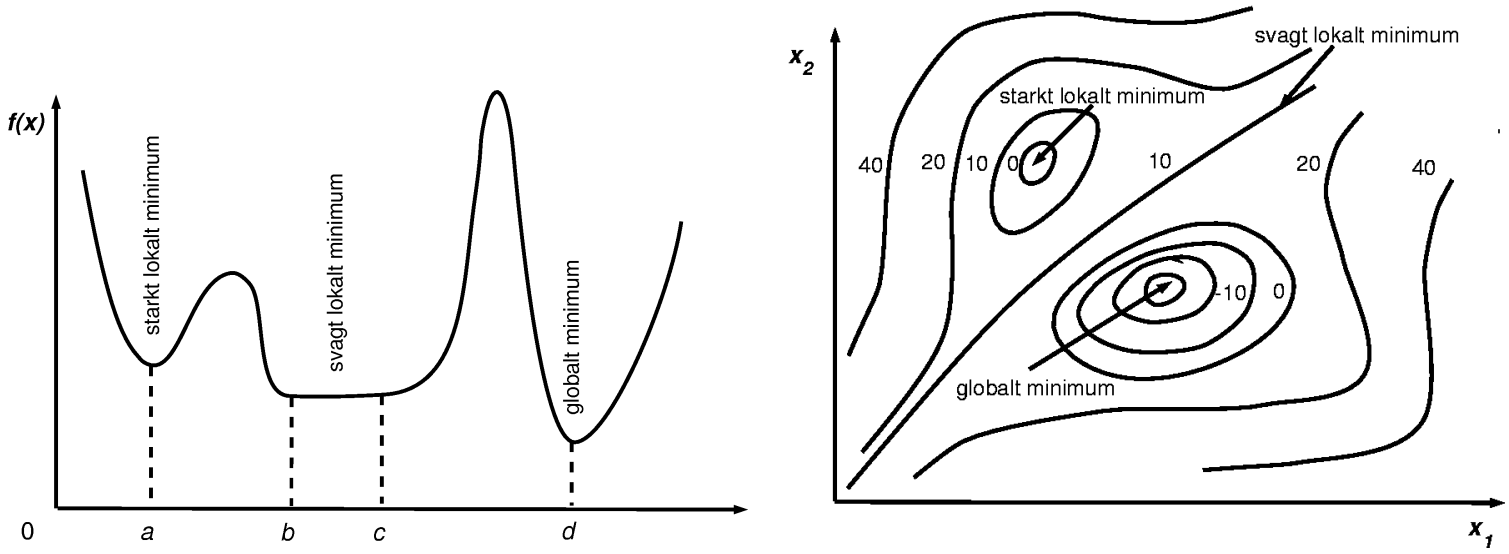
Vi vill bestämma x_0 så, att $f(x_0)$ är ett minimum. I allmänhet skiljer man mellan tre slag av minimer:

(a) x_0 är ett starkt lokalt minimum av $f(x)$, ifall f är definierad inom en δ–omgivning av x_0, och det existerar ett sådant ϵ, $(0 < \epsilon < \delta)$, att $f(x_0) < f(x)$ gäller för alla x, som uppfyller villkoret $0 < \|x_0 - x\| < \epsilon$.

(b) x_0 sägs vara ett svagt lokalt minimum av $f(x)$, om f är definierad inom en δ–omgivning av x_0, och det existerar ett sådant ϵ, $(0 < \epsilon < \delta)$, att $f(x_0) \leq f(x)$ gäller för alla x, som uppfyller villkoret $0 < \|x_0 - x\| < \epsilon$.

(c) x_0 sägs vara ett globalt minimum av $f(x)$, om villkoret $f(x_0) \leq f(x)$ gäller för varje x, som tillhör det n–dimensionella euklidiska rummet.

Dessa olika slag av minimer kan åskådliggöras grafiskt genom att man betraktar en funktion av en variabel samt en funktion av två variabler:

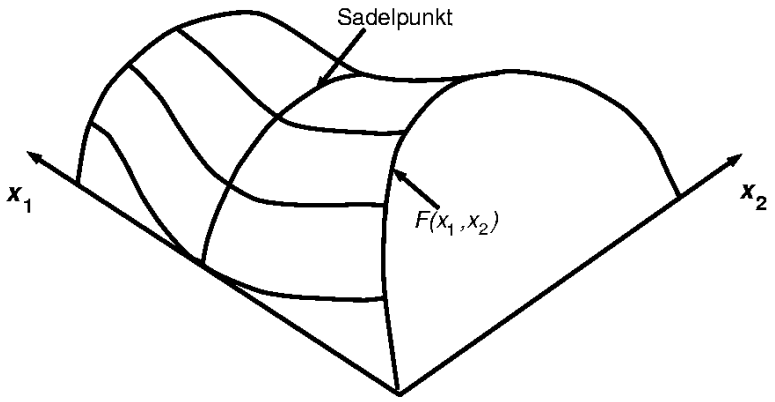


Resultatet av en numerisk minimeringsprocedur är i allmänhet ett lokalt minimum (som vi konstaterade redan i föregående avsnitt). Det globala minimet är svårare att bestämma, men det är inte heller alltid eftersträvansvärt.

Vid den praktiska beräkningen av funktionsminimet ställer man ofta vissa krav på funktionen f, vanligen att den är *kontinuerlig*, och att åtminstone första derivatorna (dvs. *gradienten*) är kontinuerliga. Ofta förutsätts även derivatorna av andra ordningen vara kontinuerliga.

En vektor sägs vara stationär för f, ifall gradienten försvinner, dvs. $g \equiv \nabla_x f(x) = 0$. Ett nödvändigt villkor för att x_0 skall vara ett lokalt extremum av en funktion f, vars första derivator är kontinuerliga, är att x_0 är stationär.

Detta villkor är emellertid inte tillräckligt, vilket leder oss till följande definition: En punkt x_0 sägs vara en sadelpunkt av f, ifall x_0 är stationär, men inte ett lokalt extremum. Detta begrepp åskådliggörs i nedanstående bild:



Om vi ytterligare antar, att f är en funktion, vars derivator av andra ordningen är kontinuerliga, så gäller att x_0 är ett starkt lokalt minimum av f ifall

$$g(x_0) = \nabla_x f(x_0) = 0,$$

och H, Hesses matris (eller hessianen) för funktionen f, är positivt definit i punkten x_0 (dvs. den har enbart positiva egenvärden). Hesses matris (som är uppkallad efter *Ludwig Otto Hesse* (1811-1874)) konstrueras ur de partiella derivatorna av andra ordningen:

$$H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} \quad (i, j = 1, 2, \ldots, n).$$

Dessa villkor för ett funktionsminimum finner man genom att studera Taylorserien för f i avseende på punkten x_0.

Vi skall till en början anta, att gradienten är kontinuerlig. Då kan funktionen utvecklas i en Taylor-serie av formen

$$f(x_0 + \epsilon y) = f(x_0) + \epsilon y^T g + \mathcal{O}(\epsilon^2),$$

där y är en godtycklig kolonnvektor, ϵ ett litet reellt tal, och $y^T g = \sum_i y_i g_i = \sum_i y_i \frac{\partial f(x)}{\partial x_i}\Big|_{x=x_0}$.

Om ϵ är ett tillräckligt litet tal, så följer av olikheten $f(x_0) \leq f(x_0 + \epsilon y)$ att $y^T g = 0 \ \forall y$. Detta leder till villkoret $g(x_0) = 0$.

Om vi därpå antar, att funktionen f har kontinuerliga partiella derivator av andra ordningen, så kan dess Taylor–utveckling skrivas i formen

$$f(x_0 + \epsilon y) = f(x_0) + \tfrac{1}{2}\epsilon^2 y^T H y + \mathcal{O}(\epsilon^3),$$

där

$$y^T H y = \sum_{i,j} y_i y_j \frac{\partial^2 f(x)}{\partial x_i \partial x_j}\bigg|_{x=x_0}.$$

Om H *inte* är en positivt definit (eller semidefinit) matris, så existerar det en vektor y som uppfyller villkoret

$$y^T H y \leq 0.$$

Om ϵ är tillräckligt litet, så gäller alltså

$$f(x_0 + \epsilon y) < f(x_0),$$

varav följer, att H åtminstone måste vara positivt semidefinit för att x_0 skall kunna vara ett lokalt minimum, och om H är positivt definit, så är x_0 ett *starkt* lokalt minimum.

Härnäst skall vi studera olika typer av funktioner. En funktion av typen

$$f(x) = \tfrac{1}{2} x^T A x + b^T x,$$

där A är en symmetrisk $n \times n$–matris och b en kolonnvektor (båda med konstanta element) kallas kvadratisk. Gradienten för en dylik funktion är

$$g = Ax + b,$$

varav följer, att minimet kan beräknas ur ekvationen

$$Ax_0 + b = 0.$$

Om A är icke-singulär, så finns det en stationär punkt, och ifall A är positivt definit, så är $x_0 = -A^{-1}b$ ett starkt lokalt minimum.

En funktion av en variabel $f(x)$, som har endast *ett* minimum inom ett bestämt intervall av x sägs vara unimodal. Formellt kan man säga, att f är unimodal, ifall x_0 är det enda värde av x för vilket $f(x) \leq f(y)$ gäller för *alla* y inom *varje* intervall, som innehåller x. Denna definition kan tillämpas både på kontinuerliga och diskontinuerliga funktioner.

En sådan funktion är också konvex. Geometriskt är en funktion konvex, ifall en rät linje som förbinder två godtyckliga punkter på kurvan ligger ovanför densamma. Formellt sett är en funktion (av en variabel) konvex, ifall olikheten

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

är uppfylld för alla x och y samt för $0 < \lambda < 1$. Ifall olikheten gäller *utan* likhetstecknet, så är f strängt konvex. En funktion f är konkav om $-f$ är konvex.

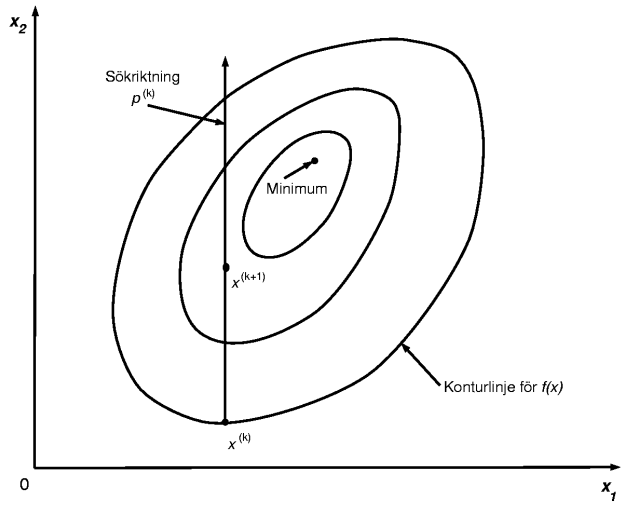
Dessa definitioner kan utsträckas till n dimensioner genom att man tolkar x och y som vektorer. Konvexa funktioner har intressanta egenskaper. Om f är en strängt konvex funktion med kontinuerliga partiella derivator av andra ordningen, så är H positivt semidefinit (dvs. $y^T H y \geq 0 \ \forall y$) inom det euklidiska rummet. Om f är en kvadratisk funktion, och H är positivt definit, så är f strängt konvex. Om f är en konvex funktion med ett starkt lokalt minimum, så är detta entydigt och globalt.

Som en enkel tillämpning skall vi studera minimering av en funktion av *en variabel*, ett problem som ofta uppträder vid optimering av en funktion av flera variabler. Man kan nämligen ofta definiera en parameter s, som minimerar $f(x(s))$. En typisk algoritm av detta slag är t.ex. följande:

Vid den k:te iterationen bestämmer man

(a) en "sökriktning" $p^{(k)}$, samt
(b) en skalär $s^{(k)}$, som minimerar funktionen $f(x^{(k)} + sp^{(k)})$ med avseende på s;
(c) varpå man sätter $x^{(k+1)} = x^{(k)} + s^{(k)}p^{(k)}$.

Att finna ett minimum av en funktion i en viss riktning brukar man vanligen kalla för en linjär sökmetod. I praktiken använder man härvid antingen funktionsvärden, som jämförs med varandra, eller någon funktionsapproximeringsmetod. Proceduren åskådliggörs i nedanstående figur.



Vi skall här endast beskriva den förstnämnda metoden. Vi antar, att funktionen endast har *ett* minimum, dvs att den är unimodal. Vi antar vidare, att vi har beräknat funktionsvärdena i fyra punkter a_1, a_2, a_3 och a_4, och att vi på grund av detta vet att minimet befinner sig inom intervallet (a_1, a_4) (vi antar dessutom, att $a_1 < a_2 < a_3 < a_4$ gäller). Utgångsvärdena kan bestämmas genom att man beräknar funktionsvärdet i en viss punkt, och därpå ger ett tillskott till argumentet tills funktionen börjar växa. Om detta inte sker, går man i motsatt riktning. Härtill kan bisektionsmetoden användas, men det finns effektivare metoder.

I den s.k. gyllene snitt–metoden utgår man från fyra dylika a–värden som begränsar minimet och uppfyller ekvationerna $a_3 - a_1 = a_4 - a_2 = \gamma(a_4 - a_1)$, där $\gamma = 2/(1+\sqrt{5}) \approx 0.618034\ldots$ (se figuren nedan). Genom att testa funktionsvärdena $f(a_1)$, $f(a_2)$, $f(a_3)$, $f(a_4)$ är det möjligt att ta reda på inom vilket av de två lika stora intervallen (a_1, a_3) eller (a_2, a_4) minimet ligger.

Vi kan för enkelhetens skull anta att det ligger inom intervallet (a_1, a_3). Funktionen beräknas därpå i en ny punkt a_5, som uppfyller villkoret $a_5 - a_1 = a_3 - a_2$. I detta fall gäller $a_3 - a_5 = \gamma(a_3 - a_1)$. Vi inser detta, om $\gamma^{-1} = \gamma + 1$ tillämpas i formeln ovan, varpå vi får $a_4 - a_1 = \gamma(a_3 - a_1) + a_3 - a_1$, dvs $\gamma(a_3 - a_1) = a_4 - a_3 = a_2 - a_1 = a_3 - a_5$. Om vi nu betecknar punkterna $a'_1 = a_1$, $a'_2 = a_5$, $a'_3 = a_2$ och $a'_4 = a_3$, så finner vi situationen motsvara utgångspunkten med undantag av att intervallängden reducerats med beloppet γ: $a'_4 - a'_1 = \gamma(a_4 - a_1)$. Som av bilden framgår, kommer alltså minimet att inneslutas mellan allt trängre gränser.