Krister Svanberg, maj 2012

1 Ickelinjär optimering, speciellt problem utan bivillkor

En mycket allmän formulering av optimeringsproblem är följande:

$$\begin{aligned} \text{minimera} \quad & f(\mathbf{x}) \\ \text{då} \quad & \mathbf{x} \in \mathcal{F}, \end{aligned} \tag{1.1}$$

där $\mathbf{x} = (x_1, \ldots, x_n)^\mathsf{T} \in I\!R^n$ är en vektor innehållande problemets *variabler*.
$\mathcal{F}$ är en given delmängd av $I\!R^n$, och
f är en given reellvärd funktion som är definierad (åtminstone) på mängden $\mathcal{F}$.
f kallas *målfunktionen* och $\mathcal{F}$ kallas *tillåtna området* till problemet P_0.

Antag exempelvis att vi ska bestämma vilken längd, höjd och bredd en låda ska ha för att totala arean av lådans sex sidor ska bli så liten som möjligt under kravet att lådans volym måste vara minst 100 dm^3 och lådans rymddiagonal måste vara minst 9 dm. Genom att kalla längden, höjden och bredden för respektive x_1, x_2 och x_3 kan detta problem skrivas

$$\begin{aligned} \text{minimera} \quad & 2x_1x_2 + 2x_2x_3 + 2x_3x_1 \\ \text{då} \quad & x_1x_2x_3 - 100 \geq 0, \\ & x_1^2 + x_2^2 + x_3^2 - 81 \geq 0, \\ & x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0. \end{aligned} \tag{1.2}$$

Här är $\mathbf{x} = (x_1, x_2, x_3)^\mathsf{T}$, $f(\mathbf{x}) = 2x_1x_2 + 2x_2x_3 + 2x_3x_1$ och
$\mathcal{F} = \{\, \mathbf{x} \in I\!R^3 \mid x_1x_2x_3 - 100 \geq 0\,,\ x_1^2 + x_2^2 + x_3^2 - 81 \geq 0\,,\ x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0\,\}$.

Detta exempel faller inom den viktiga delklass av optimeringsproblem som går under namnet *ickelinjär optimering*, på engelska *nonlinear programming*, med den vedertagna förkortningen NLP. Inom denna problemklass är målfunktionen f en kontinuerligt deriverbar funktion, medan tillåtna området definieras av ett antal bivillkor av typen $g_i(\mathbf{x}) \leq 0$ (olikhetsbivillkor) och/eller $h_i(\mathbf{x}) = 0$ (likhetsbivillkor), där g_i och h_i är givna funktioner som är kontinuerligt deriverbara. NLP-problem är alltså på formen

$$\begin{aligned} \text{minimera} \quad & f(\mathbf{x}) \\ \text{då} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m_1 \\ & h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, m_2 \\ & \mathbf{x} \in I\!R^n, \end{aligned} \tag{1.3}$$

där minst en av funktionerna $f,\ g_1, \ldots, g_{m_1},\ h_1, \ldots, h_{m_2}$ är ickelinjär (annars är (1.3) ett *linjärt* optimeringsproblem, dvs ett LP-problem). Med beteckningar enligt (1.1) är alltså
$\mathcal{F} = \{\mathbf{x} \in I\!R^n \mid g_i(\mathbf{x}) \leq 0 \text{ för } i = 1, \ldots, m_1 \text{ och } h_i(\mathbf{x}) = 0 \text{ för } i = 1, \ldots, m_2\}$.

Ett inte helt ovanligt specialfall av problemet (1.3) erhålls om $m_1 = m_2 = 0$, varvid $\mathcal{F} = I\!R^n$. I detta specialfall är (1.3) ett NLP-problem *utan bivillkor*, på engelska *unconstrained* NLP. Annars är (1.3) ett NLP-problem *med bivillkor*, på engelska *constrained* NLP.

2 Teori för envariabel-optimering

I detta kapitel är f en reellvärd funktion av en reell variabel x, dvs $f : \mathbb{R} \to \mathbb{R}$.
Vi förutsätter genomgående att såväl (första)derivatan f' som andraderivatan f'' existerar och är kontinuerliga på $\mathbb{R}$.

Def 2.1. Punkten $\hat{x} \in \mathbb{R}$ är en *lokal* minpunkt till funktionen f om det finns ett tal $\delta > 0$ sådant att $f(\hat{x}) \leq f(x)$ för alla $x \in \mathbb{R}$ som uppfyller $\hat{x} - \delta < x < \hat{x} + \delta$.

Def 2.2. Punkten $\hat{x} \in \mathbb{R}$ är en *global* minpunkt till f om $f(\hat{x}) \leq f(x)$ för *alla* $x \in \mathbb{R}$.

Det är uppenbart att varje global minpunkt även är en lokal minpunkt, men det kan (för icke-konvexa funktioner) mycket väl finnas lokala minpunkter som inte är globala minpunkter.

Derivatan av f i en given punkt $\hat{x}$ ges per definition av

$$f'(\hat{x}) = \lim_{x \to \hat{x}} \frac{f(x) - f(\hat{x})}{x - \hat{x}}, \tag{2.1}$$

vilket betyder att det för varje tal $\varepsilon > 0$ finns ett tal $\delta > 0$ (som kan bero på ε) sådant att

$$-\varepsilon < \frac{f(x) - f(\hat{x})}{x - \hat{x}} - f'(\hat{x}) < \varepsilon \quad \text{för alla } x \text{ som uppfyller } 0 < |x - \hat{x}| < \delta. \tag{2.2}$$

Lemma 2.1. Om $f'(\hat{x}) > 0$ så finns det ett tal $\delta > 0$ sådant att
$f(x) > f(\hat{x})$ för alla x som uppfyller $\hat{x} < x < \hat{x} + \delta$, medan
$f(x) < f(\hat{x})$ för alla x som uppfyller $\hat{x} - \delta < x < \hat{x}$.

Bevis: Välj $\varepsilon = \frac{1}{2} f'(\hat{x}) > 0$ i (2.2). Då finns det alltså ett tal $\delta > 0$ sådant att

$$\tfrac{1}{2} f'(\hat{x}) < \frac{f(x) - f(\hat{x})}{x - \hat{x}} < \tfrac{3}{2} f'(\hat{x}) \quad \text{för alla } x \text{ som uppfyller } 0 < |x - \hat{x}| < \delta. \tag{2.3}$$

För alla x som uppfyller $0 < x - \hat{x} < \delta$ så gäller enligt (2.3) att
$\frac{1}{2} f'(\hat{x})(x - \hat{x}) < f(x) - f(\hat{x}) < \frac{3}{2} f'(\hat{x})(x - \hat{x})$.
Den vänstra av dessa olikheter medför att $f(x) > f(\hat{x})$.
För alla x som uppfyller $-\delta < x - \hat{x} < 0$ så gäller enligt (2.3) att
$\frac{1}{2} f'(\hat{x})(x - \hat{x}) > f(x) - f(\hat{x}) > \frac{3}{2} f'(\hat{x})(x - \hat{x})$.
Den vänstra av dessa olikheter medför att $f(x) < f(\hat{x})$. $\oplus$

Lemma 2.2. Om $f'(\hat{x}) < 0$ så finns det ett tal $\delta > 0$ sådant att
$f(x) < f(\hat{x})$ för alla x som uppfyller $\hat{x} < x < \hat{x} + \delta$, medan
$f(x) > f(\hat{x})$ för alla x som uppfyller $\hat{x} - \delta < x < \hat{x}$.

Bevis: Analogt med beviset av Lemma 2.1. Nu väljs i stället $\varepsilon = -\frac{1}{2} f'(\hat{x}) > 0$ i (2.2).

Sats 2.1. Om $\hat{x}$ är en lokal minpunkt till funktionen f så är $f'(\hat{x}) = 0$.

Bevis: Enligt Lemma 2.1 så är $\hat{x}$ inte en lokal minpunkt till f om $f'(\hat{x}) > 0$,
och enligt Lemma 2.2 så är $\hat{x}$ inte en lokal minpunkt till f om $f'(\hat{x}) < 0$.
Enda återstående möjligheten är då att $f'(\hat{x}) = 0$. $\oplus$

Vi påminner nu läsaren om följande variant av *Taylors formel*:

$$f(x) = f(\hat{x}) + f'(\hat{x})(x - \hat{x}) + \tfrac{1}{2}f''(\xi)(x - \hat{x})^2, \tag{2.4}$$

för något tal ξ mellan x och $\hat{x}$, dvs $\xi = \hat{x} + \theta \cdot (x - \hat{x})$ för något tal $\theta \in (0, 1)$.
Det exakta värdet på θ beror dels på vilka värden $\hat{x}$ och x har, dels på vilken funktion f som är inblandad. Men detta exakta värde behöver vi faktiskt inte känna till. Det räcker gott att veta att θ ligger mellan 0 och 1, dvs att ξ ligger mellan x och $\hat{x}$.

Ett intressant specialfall av (2.4) erhålls genom att låta $\hat{x} = 0$ och $x = 1$. Då övergår (2.4) till följande formel, som vi ska använda oss av i nästa kapitel.

$$f(1) = f(0) + f'(0) + \tfrac{1}{2}f''(\theta), \quad \text{för något tal } \theta \in (0, 1). \tag{2.5}$$

Lemma 2.3. Om $f'(\hat{x}) = 0$ och $f''(\hat{x}) > 0$ så finns det ett tal $\delta > 0$ sådant att
$f(x) > f(\hat{x})$ för alla $x \neq \hat{x}$ som uppfyller $\hat{x} - \delta < x < \hat{x} + \delta$.

Bevis: Eftersom f'' är kontinuerlig och $f''(\hat{x}) > 0$, så finns det ett tal $\delta > 0$ sådant att $f''(x) > 0$ för alla x som uppfyller $\hat{x} - \delta < x < \hat{x} + \delta$. Men om x uppfyller dessa olikheter så uppfylls de även av $\xi = \hat{x} + \theta \cdot (x - \hat{x})$, dvs $\hat{x} - \delta < \xi < \hat{x} + \delta$, och därmed är $f''(\xi) > 0$. Taylors formel (2.4) ger då (för dessa x) att $f(x) = f(\hat{x}) + \frac{1}{2}f''(\xi)(x - \hat{x})^2 \geq f(\hat{x})$, med likhet endast om $x = \hat{x}$. $\oplus$

Lemma 2.4. Om $f'(\hat{x}) = 0$ och $f''(\hat{x}) < 0$ så finns det ett tal $\delta > 0$ sådant att
$f(x) < f(\hat{x})$ för alla $x \neq \hat{x}$ som uppfyller $\hat{x} - \delta < x < \hat{x} + \delta$.

Bevis: Helt analogt med beviset av Lemma 2.3.

Sats 2.2. Nödvändiga (men inte tillräckliga) villkor för att $\hat{x}$ ska vara en lokal minpunkt till funktionen f är att $f'(\hat{x}) = 0$ och $f''(\hat{x}) \geq 0$.
Tillräckliga (men inte nödvändiga) villkor för att $\hat{x}$ ska vara en lokal minpunkt till funktionen f är att $f'(\hat{x}) = 0$ och $f''(\hat{x}) > 0$.

Bevis: Att $f'(\hat{x}) = 0$ och $f''(\hat{x}) \geq 0$ är nödvändiga villkor följer av Sats 2.1 och Lemma 2.4.
Att $f'(\hat{x}) = 0$ och $f''(\hat{x}) > 0$ är tillräckliga villkor följer av Lemma 2.3. $\oplus$

Att $f'(\hat{x}) = 0$ och $f''(\hat{x}) \geq 0$ inte är tillräckliga villkor för ett lokalt minimum kan inses av följande exempel. Låt f definieras av $f(x) = x^3$ och låt $\hat{x} = 0$. Då är $f'(\hat{x}) = 0$ och $f''(\hat{x}) = 0$, så att $f''(\hat{x}) \geq 0$ är uppfyllt. Men $\hat{x}$ är ändå ingen lokal minpunkt till f.
Att $f'(\hat{x}) = 0$ och $f''(\hat{x}) > 0$ inte är nödvändiga villkor för ett lokalt minimum kan inses av följande exempel. Låt f definieras av $f(x) = x^4$ och låt $\hat{x} = 0$. Då är $f'(\hat{x}) = 0$ och $f''(\hat{x}) = 0$, dvs $f''(\hat{x}) > 0$ är ej uppfyllt. Men $\hat{x}$ är ändå en lokal (och global) minpunkt till f.

3 Teori för flervariabel-optimering utan bivillkor

I detta kapitel behandlar vi problemet att minimera en given flervariabelfuntion *utan* några bivillkor, dvs problem på formen

$$\begin{aligned} \text{minimera} \quad & f(\mathbf{x}) \\ \text{då} \quad & \mathbf{x} \in I\!R^n, \end{aligned} \tag{3.1}$$

där f är en given reellvärd funktion på $I\!R^n$. Vi kommer genomgående att förutsätta att f är *två gånger kontinuerligt deriverbar*.

Att funktionen f är två gånger kontinuerligt deriverbar betyder dels att de n stycken partiella förstaderivatorna $\partial f/\partial x_j$ existerar och är kontinuerliga i hela $I\!R^n$, dels att de n^2 stycken partiella andraderivatorna $\partial^2 f/\partial x_i \partial x_j$ existerar och är kontinuerliga i hela $I\!R^n$.

Då definieras *gradienten* till f i punkten $\mathbf{x}$ som radvektorn

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}),\ \ldots\ , \frac{\partial f}{\partial x_n}(\mathbf{x}) \right). \tag{3.2}$$

Vidare definieras *Hessianen* till f i punkten $\mathbf{x}$ som den symmetriska $n \times n$ matris $\mathbf{F}(\mathbf{x})$ som i rad i och kolumn j har elementet $\dfrac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x})$, det vill säga

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1 \partial x_1}(\mathbf{x}) & \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{x}) & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}) \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{x}) & \dfrac{\partial^2 f}{\partial x_2 \partial x_2}(\mathbf{x}) & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}) & \dfrac{\partial^2 f}{\partial x_n \partial x_2}(\mathbf{x}) & \cdots & \dfrac{\partial^2 f}{\partial x_n \partial x_n}(\mathbf{x}) \end{bmatrix}, \tag{3.3}$$

Def 3.1. Punkten $\hat{\mathbf{x}} \in I\!R^n$ är en *lokal* minpunkt till funktionen f om det finns ett tal $\delta > 0$ sådant att $f(\hat{\mathbf{x}}) \leq f(\mathbf{x})$ för alla $\mathbf{x} \in I\!R^n$ som uppfyller $|\mathbf{x} - \hat{\mathbf{x}}| < \delta$.

Def 3.2. Punkten $\hat{\mathbf{x}} \in I\!R^n$ är en *global* minpunkt till f om $f(\hat{\mathbf{x}}) \leq f(\mathbf{x})$ för *alla* $\mathbf{x} \in I\!R^n$.

Det är uppenbart att varje global minpunkt även är en lokal minpunkt, men det kan (för icke-konvexa funktioner) mycket väl finnas lokala minpunkter som inte är globala minpunkter.

Låt $\hat{\mathbf{x}} \in I\!R^n$ vara en given punkt, och antag att vi vill avgöra huruvida $\hat{\mathbf{x}}$ är en lokal minpunkt till funktionen f eller inte. Vi ska då först undersöka hur målfunktionen uppför sig längs räta linjer genom $\hat{\mathbf{x}}$. Tag därför en godtyckligt vektor $\mathbf{d} \in I\!R^n$ ($\mathbf{d} \neq \mathbf{0}$) och låt

$$\mathbf{x}(t) = \hat{\mathbf{x}} + t\,\mathbf{d}, \quad \text{för } t \in I\!R. \tag{3.4}$$

Detta definierar en rät linje i $I\!R^n$ (på parameterform) som går genom punkten $\hat{\mathbf{x}}$ och som har en riktning given av $\mathbf{d}$. Speciellt är $\mathbf{x}(0) = \hat{\mathbf{x}}$.

Vi ska som sagt studera målfunktionen f längs denna linje och låter därför envariabelfunktionen φ definieras av

$$\varphi(t) = f(\mathbf{x}(t)) = f(\hat{\mathbf{x}} + t\,\mathbf{d}), \quad \text{för } t \in I\!R. \tag{3.5}$$

Eftersom f enligt förutsättningarna är två gånger kontinuerligt deriverbar så är även φ två gånger kontinuerligt deriverbar för alla $t \in I\!R$. Enligt kedjeregeln ges dess förstaderivata φ' och dess andraderivata φ'' av

$$\varphi'(t) = \nabla f(\mathbf{x}(t))\,\mathbf{d} \quad \text{och} \quad \varphi''(t) = \mathbf{d}^{\mathsf{T}}\mathbf{F}(\mathbf{x}(t))\,\mathbf{d}. \tag{3.6}$$

Speciellt är

$$\varphi'(0) = \nabla f(\hat{\mathbf{x}})\,\mathbf{d} \quad \text{och} \quad \varphi''(0) = \mathbf{d}^{\mathsf{T}}\mathbf{F}(\hat{\mathbf{x}})\,\mathbf{d}. \tag{3.7}$$

$\varphi'(0) = \nabla f(\hat{\mathbf{x}})\,\mathbf{d}$ kallas *riktningsderivatan* till funktionen f i punkten $\hat{\mathbf{x}}$ i riktningen $\mathbf{d}$.

Lemma 3.1. Om $\nabla f(\hat{\mathbf{x}})\,\mathbf{d} < 0$ så finns det ett tal $\delta > 0$ sådant att $f(\hat{\mathbf{x}} + t\,\mathbf{d}) < f(\hat{\mathbf{x}})$ för alla $t \in (0, \delta)$ (dvs för alla t som uppfyller $0 < t < \delta$).

Bevis: Tillämpa Lemma 2.2 på funktionen φ ovan.

Lemma 3.2. Om $\hat{\mathbf{x}}$ är en lokal minpunkt till funktionen f så gäller, för varje vektor $\mathbf{d} \in I\!R^n$, att $t = 0$ är en lokal minpunkt till funktionen φ definierad av (3.5).

Bevis: Övningsuppgift.

Observera att omvändningen till detta lemma inte gäller! Även om det för varje vektor $\mathbf{d} \in I\!R^n$ gäller att att $t = 0$ är en lokal minpunkt till funktionen φ definierad av (3.5), så är det inte säkert att $\hat{\mathbf{x}}$ är en lokal minpunkt till f. Detta illustreras av följande exempel.

Övningsuppgift:
Låt $n = 2$, $f(\mathbf{x}) = (x_2 - x_1^2)(x_2 - 3x_1^2) = x_2^2 - 4x_1^2x_2 + 3x_1^4$ och $\hat{\mathbf{x}} = \mathbf{0} = (0,0)^{\mathsf{T}}$.
(a) Visa att det för varje vektor $\mathbf{d} \in I\!R^2$ gäller att $t = 0$ är en lokal minpunkt till funktionen φ definierad av (3.5).
(b) Visa att $\hat{\mathbf{x}}$ inte är en lokal minpunkt till f.

Sats 3.2. Nödvändiga (men inte tillräckliga) villkor för att $\hat{\mathbf{x}}$ ska vara en lokal minpunkt till funktionen f är att $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}$ och att $\mathbf{F}(\hat{\mathbf{x}})$ är *positivt semidefinit*.

Bevis: Från Lemma 3.2, tillsammans med (3.7) och första halvan av Sats 2.2, följer att nödvändiga villkor för att $\hat{\mathbf{x}}$ ska vara en lokal minpunkt till f är att $\varphi'(0) = \nabla f(\hat{\mathbf{x}})\mathbf{d} = 0$ för alla $\mathbf{d} \in I\!R^n$, vilket medför att $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}$, och att $\varphi''(0) = \mathbf{d}^\mathsf{T}\mathbf{F}(\hat{\mathbf{x}})\,\mathbf{d} \geq 0$ för alla $\mathbf{d} \in I\!R^n$, vilket per definition innebär att $\mathbf{F}(\hat{\mathbf{x}})$ är positivt semidefinit. $\oplus$

För att kunna härleda *tillräckliga* villkor för att punkten $\hat{\mathbf{x}}$ ska vara en lokal minpunkt till f ska vi först ange flervariabelmotsvarigheten till envariabelformeln (2.4). Låt $\mathbf{x} \in I\!R^n$ ($\mathbf{x} \neq \hat{\mathbf{x}}$) vara en godtycklig punkt vars funktionsvärde $f(\mathbf{x})$ vi vill kunna jämföra med $f(\hat{\mathbf{x}})$, utan att explicit beräkna $f(\mathbf{x})$. Sätt $\mathbf{d} = \mathbf{x}-\hat{\mathbf{x}}$ och låt $\varphi(t) = f(\hat{\mathbf{x}} + t\,\mathbf{d}) = f(\hat{\mathbf{x}} + t\,(\mathbf{x}-\hat{\mathbf{x}}))$. Då är $\varphi(0) = f(\hat{\mathbf{x}})$ och $\varphi(1) = f(\mathbf{x})$. Specialfallet (2.5) av Taylors formel i envariabelfallet säger då att

$$\varphi(1) = \varphi(0) + \varphi'(0) + \tfrac{1}{2}\varphi''(\theta), \quad \text{för något tal } \theta \in (0,1), \tag{3.8}$$

vilket med hjälp av (3.6) och (3.7) kan skrivas

$$f(\mathbf{x}) = f(\hat{\mathbf{x}}) + \nabla f(\hat{\mathbf{x}})\,\mathbf{d} + \tfrac{1}{2}\mathbf{d}^\mathsf{T}\mathbf{F}(\mathbf{x} + \theta\cdot\mathbf{d})\,\mathbf{d}, \quad \text{för något tal } \theta \in (0,1), \tag{3.9}$$

eller, ekvivalent,

$$f(\mathbf{x}) = f(\hat{\mathbf{x}}) + \nabla f(\hat{\mathbf{x}})(\mathbf{x}-\hat{\mathbf{x}}) + \tfrac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^\mathsf{T}\mathbf{F}(\mathbf{x} + \theta\cdot(\mathbf{x}-\hat{\mathbf{x}}))(\mathbf{x}-\hat{\mathbf{x}}), \tag{3.10}$$

för något tal $\theta \in (0,1)$. Detta är flervariabelmotsvarigheten till envariabelformeln (2.4). Det är i själva verket en välkänd variant av Taylors formel i flervariabelfallet.

Vi behöver också följande resultat.

Lemma 3.3. Om Hessianen $\mathbf{F}(\hat{\mathbf{x}})$ är *positivt definit* i en given punkt $\hat{\mathbf{x}} \in I\!R^n$, dvs om $\mathbf{d}^\mathsf{T}\mathbf{F}(\hat{\mathbf{x}})\,\mathbf{d} > 0$ för alla $\mathbf{d} \neq \mathbf{0}$ ($\mathbf{d} \in I\!R^n$), så finns det ett tal $\delta > 0$ sådant att $\mathbf{F}(\mathbf{x})$ är positivt definit för alla $\mathbf{x} \in I\!R^n$ som uppfyller $|\mathbf{x}-\hat{\mathbf{x}}| < \delta$.

Bevis: Övningsuppgift.

Sats 3.3. Tillräckliga (men inte nödvändiga) villkor för att $\hat{\mathbf{x}}$ ska vara en lokal minpunkt till funktionen f är att $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}$ och att $\mathbf{F}(\hat{\mathbf{x}})$ är *positivt definit*.

Bevis: Antag att $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}$ och att $\mathbf{F}(\hat{\mathbf{x}})$ är positivt definit. Välj $\delta > 0$ enligt Lemma 3.3, så att $\mathbf{F}(\mathbf{x})$ är positivt definit för alla $\mathbf{x} \in I\!R^n$ som uppfyller $|\mathbf{x}-\hat{\mathbf{x}}| < \delta$.
För dessa $\mathbf{x}$ gäller då att även $\mathbf{F}(\hat{\mathbf{x}} + \theta\cdot(\mathbf{x}-\hat{\mathbf{x}}))$ är positivt definit för varje $\theta \in (0,1)$, eftersom $|(\hat{\mathbf{x}} + \theta\cdot(\mathbf{x}-\hat{\mathbf{x}})) - \hat{\mathbf{x}}| = |\theta\cdot(\mathbf{x}-\hat{\mathbf{x}})| = \theta\cdot|\mathbf{x}-\hat{\mathbf{x}}| \leq |\mathbf{x}-\hat{\mathbf{x}}| < \delta$.
Med hjälp av (3.10) erhålls då att för alla $\mathbf{x} \in I\!R^n$ som uppfyller $|\mathbf{x}-\hat{\mathbf{x}}| < \delta$ är

$$f(\mathbf{x}) = f(\hat{\mathbf{x}}) + \tfrac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^\mathsf{T}\mathbf{F}(\mathbf{x} + \theta\cdot(\mathbf{x}-\hat{\mathbf{x}}))(\mathbf{x}-\hat{\mathbf{x}}) \geq f(\hat{\mathbf{x}}), \tag{3.11}$$

med likhet endast om $\mathbf{x} = \hat{\mathbf{x}}$. Därmed är $\hat{\mathbf{x}}$ en (strikt) lokal minpunkt till f.

4 Konvexa funktioner

Ett optimeringsproblem är i viss mening "godartat" om den målfunktion som ska minimeras är en *konvex* funktion och det tillåtna område över vilket minimering ska utföras är en *konvex* mängd. En av flera trevliga egenskaper hos sådana problem är att varje lokal optimallösning även är en global optimallösning. I detta kapitel listas några viktiga egenskaper hos konvexa funktioner.

Det öppna intervallet mellan 0 och 1 beteckas $(0,1)$, dvs $(0,1) = \{t \in I\!R \mid 0 < t < 1\}$.

Def 4.1. En given mängd $C \subseteq I\!R^n$ är *konvex* om $(1-t)\hat{\mathbf{x}} + t\,\mathbf{x} \in C$ för varje val av $\hat{\mathbf{x}} \in C$, $\mathbf{x} \in C$ och $t \in (0,1)$.

Def 4.2. Låt $C \subseteq I\!R^n$ vara en konvex mängd och f en reellvärd funktion definierad på C. f är en *konvex* funktion på C om följande olikhet är uppfylld för varje val av $\hat{\mathbf{x}} \in C$, $\mathbf{x} \in C$ och $t \in (0,1)$:

$$f((1-t)\hat{\mathbf{x}} + t\mathbf{x}) \leq (1-t)f(\hat{\mathbf{x}}) + tf(\mathbf{x}). \tag{4.1}$$

En ofta användbar ekvivalent form på ovanstående olikhet är följande.

$$f(\hat{\mathbf{x}} + t\,(\mathbf{x} - \hat{\mathbf{x}})) \leq f(\hat{\mathbf{x}}) + t\,(f(\mathbf{x}) - f(\hat{\mathbf{x}})). \tag{4.2}$$

Man kan visa att om $f_1, \ldots, f_m$ är konvexa funktioner på C och $\gamma_1, \ldots, \gamma_m$ är icke-negativa reella konstanter, så är funktionen f definierad av $f(\mathbf{x}) = \gamma_1 f_1(\mathbf{x}) + \ldots + \gamma_m f_m(\mathbf{x})$ konvex på C, liksom funktionen f definierad av $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})\}$. (Ett envariabelexempel: Om $f_1(x) = 1 - x$ och $f_2(x) = x - 5$ så är $\max\{f_1(x), f_2(x)\} = |x-3| - 2$, som är konvex.) Vidare är då mängden $\{\mathbf{x} \in I\!R \mid f_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m\}$ en konvex mängd.

Definitionen av konvexitet kan geometriskt tolkas som att varje linjär interpolation av en konvex funktion ger till resultat en *överskattning*. (Exempelvis $\frac{1}{2}f(\hat{\mathbf{x}}) + \frac{1}{2}f(\mathbf{x}) \geq f(\frac{1}{2}(\hat{\mathbf{x}} + \mathbf{x}))$.) Följande sats visar att för kontinuerligt deriverbara konvexa funktioner ligger varje tangentplan till funktionens graf *under* grafen, dvs varje (första ordningens) linjärisering av en konvex funktion ger till resultat en *underskattning*.

Sats 4.1. Antag att $C \subseteq I\!R^n$ är en given konvex mängd och att f är en funktion som är kontinuerligt deriverbar på C. Då är f konvex på C *om och endast om* följande olikhet är uppfylld för varje val av $\hat{\mathbf{x}} \in C$ och $\mathbf{x} \in C$:

$$f(\mathbf{x}) \geq f(\hat{\mathbf{x}}) + \nabla f(\hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}}). \tag{4.3}$$

En kontinuerligt deriverbar envariabelfunktion f är konvex på ett givet intervall $[a,b] \subseteq I\!R$ om och endast om dess derivata f' är icke avtagande på intervallet, dvs om och endast om $(f'(x_1) - f'(x_2))(x_1 - x_2) \geq 0$ för alla x_1 och x_2 i intervallet. Nästa sats utgör en generalisering av detta till flervariabelfallet.

Sats 4.2. Antag att $C \subseteq I\!R^n$ är en given konvex mängd och att f är en funktion som är kontinuerligt deriverbar på C. Då är f konvex på C *om och endast om* följande olikhet är uppfylld för varje val av $\hat{\mathbf{x}} \in C$ och $\mathbf{x} \in C$:

$$(\nabla f(\mathbf{x}) - \nabla f(\hat{\mathbf{x}}))(\mathbf{x} - \hat{\mathbf{x}}) \geq 0. \tag{4.4}$$

En två gånger kontinuerligt deriverbar envariabelfunktion f är konvex på ett givet intervall $[a, b] \subseteq I\!R$ om och endast om dess andraderivata f'' är icke-negativ på intervallet, dvs om och endast om $f''(x) \geq 0$ för alla $x \in [a, b]$. Följande sats utgör en generalisering av detta till flervariabelfallet.

Sats 4.3. Antag att $C \subseteq I\!R^n$ är en given konvex mängd och att f är en funktion som är två gånger kontinuerligt deriverbar på C. Då är f konvex på C *om och endast om* följande olikhet är uppfylld för varje val av $\hat{\mathbf{x}} \in C$ och $\mathbf{x} \in C$:

$$(\mathbf{x} - \hat{\mathbf{x}})^{\mathsf{T}} \mathbf{F}(\hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}}) \geq 0. \tag{4.5}$$

För specialfallet att $C = I\!R^n$ får vi följande karakterisering av konvexa funktioner på $I\!R^n$:

Sats 4.4. Antag att f är en funktion som är två gånger kontinuerligt deriverbar på $I\!R^n$. Då är f konvex på $I\!R^n$ *om och endast om* Hessianen $\mathbf{F}(\mathbf{x})$ är positivt semidefinit i varje punkt $\mathbf{x} \in I\!R^n$.

Här kommer en optimeringstillämpning av Sats 4.1:

Sats 4.5. Antag att funktionen f är kontinuerligt deriverbar och konvex på $I\!R^n$. Då är $\hat{\mathbf{x}} \in I\!R^n$ en *global* minpunkt till f *om och endast om* $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}^{\mathsf{T}}$.

Bevis: Antag först att $\hat{\mathbf{x}}$ är en global minpunkt till f.
Då är $\hat{\mathbf{x}}$ även en lokal minpunkt till f, och då följer från Sats 3.2 att $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}^{\mathsf{T}}$.
Antag nu omvänt att $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}^{\mathsf{T}}$.
Då ger Sats 4.1 att $f(\mathbf{x}) \geq f(\hat{\mathbf{x}}) + \nabla f(\hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}}) = f(\hat{\mathbf{x}})$ för alla $\mathbf{x} \in I\!R^n$,
vilket innebär att $\hat{\mathbf{x}}$ är en global minpunkt till f.

5 Newtons metod för minimering av en konvex flervariabelfunktion

I detta avsnitt antar vi att $f : \mathbb{R}^n \to \mathbb{R}$ är en given funktion som är två gånger kontinuerligt deriverbar och vars Hessian $\mathbf{F}(\mathbf{x})$ (dvs matris med andraderivator) är positivt definit för alla $\mathbf{x} \in \mathbb{R}^n$. Vi antar dessutom att $f(\mathbf{x}) \to \infty$ då $|\mathbf{x}| \to \infty$.

Dessa antaganden medför att f är en strikt konvex funktion med en unik global minpunkt $\hat{\mathbf{x}} \in \mathbb{R}^n$. Denna unika minpunkt karakteriseras av att $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}$, och vi ska i detta avsnitt beskriva hur man med hjälp av Newtons metod kan bestämma $\hat{\mathbf{x}}$ numeriskt. Metoden är iterativ, så det räcker att beskriva hur man utgående från iterationspunkten $\mathbf{x}^{(k)}$ genererar nästa iterationspunkt $\mathbf{x}^{(k+1)}$. Startpunkten $\mathbf{x}^{(1)}$ får användaren välja efter bästa förmåga.

Givet iterationspunkten $\mathbf{x}^{(k)}$ så beräknas först gradienten $\nabla f(\mathbf{x}^{(k)})$ och Hessianen $\mathbf{F}(\mathbf{x}^{(k)})$. Om $\nabla f(\mathbf{x}^{(k)}) = \mathbf{0}^\mathsf{T}$ så har vi hittat vår eftersökta minpunkt $\hat{\mathbf{x}}$ och kan avbryta sökandet. Fortsättningsvis antar vi därför att $\nabla f(\mathbf{x}^{(k)}) \neq \mathbf{0}^\mathsf{T}$.

Andra ordningens Taylor-approximation av funktionen f i punkten $\mathbf{x}^{(k)}$, uttryckt i variabelvektorn $\mathbf{d} = \mathbf{x} - \mathbf{x}^{(k)} \in \mathbb{R}^n$, ges av

$$f(\mathbf{x}^{(k)} + \mathbf{d}) \approx f(\mathbf{x}^{(k)}) + \nabla f(\mathbf{x}^{(k)})\,\mathbf{d} + \tfrac{1}{2}\mathbf{d}^\mathsf{T}\mathbf{F}(\mathbf{x}^{(k)})\,\mathbf{d}. \tag{5.1}$$

Högerledet i (5.1) är en strikt konvex kvadratisk funktion som minimeras av den unika lösningen till följande ekvationssystem i $\mathbf{d} \in \mathbb{R}^n$:

$$\mathbf{F}(\mathbf{x}^{(k)})\,\mathbf{d} = -\nabla f(\mathbf{x}^{(k)})^\mathsf{T}. \tag{5.2}$$

Låt den unika lösningen till (5.2) heta $\mathbf{d}^{(k)}$. Eftersom $\nabla f(\mathbf{x}^{(k)}) \neq \mathbf{0}^\mathsf{T}$ så är $\mathbf{d}^{(k)} \neq \mathbf{0}$. Vidare är riktningsderivatan $\nabla f(\mathbf{x}^{(k)})\,\mathbf{d}^{(k)} = -\mathbf{d}^{(k)\mathsf{T}}\mathbf{F}(\mathbf{x}^{(k)})\,\mathbf{d}^{(k)} < 0$, vilket visar att $\mathbf{d}^{(k)}$ är en riktning i vilken f avtar (se Lemma 3.1), dvs en så kallad avtaganderiktning till f i punkten $\mathbf{x}^{(k)}$. Observera att detta bygger på antagandet att $\mathbf{F}(\mathbf{x}^{(k)})$ är positivt definit!

Den naturliga kandidaten till nästa iterationspunkt är nu $\mathbf{x}^{(k)} + \mathbf{d}^{(k)}$, som alltså minimerar den kvadratiska Taylor-approximationen av f i $\mathbf{x}^{(k)}$, men då kan man inte vara säker på att sekvensen av iterationspunkter kommer att konvergera. Bättre är att låta nästa iterationspunkt bestämmas av

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t_k\,\mathbf{d}^{(k)}, \tag{5.3}$$

där t_k väljs (exempelvis) som det största av talen $1, \frac{1}{2}, \frac{1}{4}, \ldots$ för vilket

$$f(\mathbf{x}^{(k)} + t_k\mathbf{d}^{(k)}) < f(\mathbf{x}^{(k)}). \tag{5.4}$$

För att kunna bevisa att sekvensen $\mathbf{x}^{(k)}$ garanterat konvergerar mot den sökta optimala lösningen $\hat{\mathbf{x}}$ behöver man faktiskt kräva lite mer än (5.4). I praktiken (och på tentor i denna kurs!) duger dock (5.4).

För specialfallet $n = 1$, som betyder att f är en envariabelfunktion, så övergår ekvationssystemet (5.2) till den enkla ekvationen $f''(x^{(k)})\,d = -f'(x^{(k)})$, varvid (5.3) övergår till

$$x^{(k+1)} = x^{(k)} - t_k \cdot \frac{f'(x^{(k)})}{f''(x^{(k)})}, \tag{5.5}$$

med t_k enligt ovan.

6 Newtons metod för minimering av icke-konvexa funktioner

Newtons metod används även vid minimering av funktioner som inte är konvexa. I princip kan man även då använda metoden som den beskrevs i förra avsnittet, med en viktig modifiering:

Eftersom f inte är konvex kan det hända att Hessianen $\mathbf{F}(\mathbf{x}^{(k)})$ inte är positivt definit, och då är det inte säkert att lösningen $\mathbf{d}^{(k)}$ till ekvationssystemet (5.2) är en avtaganderiktning till målfunktionen. Om Hessianen $\mathbf{F}(\mathbf{x}^{(k)})$ inte är positivt definit måste man därför modifiera ekvationssystemet (5.2) genom att i detta ersätta matrisen $\mathbf{F}(\mathbf{x}^{(k)})$ med en ny matris som dels är positiv definit, dels approximerar $\mathbf{F}(\mathbf{x}^{(k)})$ på ett rimligt sätt. Det kanske enklaste exemplet på hur man kan göra är att ersätta $\mathbf{F}(\mathbf{x}^{(k)})$ med $\mathbf{F}(\mathbf{x}^{(k)}) + \mu_k \mathbf{I}$, där $\mathbf{I}$ är enhetsmatrisen och skalären μ_k väljs "tillräckligt stor" för att $\mathbf{F}(\mathbf{x}^{(k)}) + \mu_k \mathbf{I}$ ska bli positivt definit, men ändå inte "onödigt stor". De tekniska detaljerna för hur man väljer μ_k avstår vi från här.

Vid minimering av en icke-konvex funktion (med exempelvis Newtons metod enligt ovan) så måste man också vara medveten om att även om iterationspunkterna $\mathbf{x}^{(k)}$ konvergerar snyggt mot en punkt $\hat{\mathbf{x}}$, så är det inte säkert att $\hat{\mathbf{x}}$ är en global minpunkt till $f(\mathbf{x})$. I normalfallet är $\hat{\mathbf{x}}$ åtminstone en lokal minpunkt, men om man har riktig otur med valet av startpunkt kan det hända att $\hat{\mathbf{x}}$ bara är exempelvis en sadelpunkt. Man bör därför i praktiken helst låta datorn upprepa metoden från några olika startpunkter

7 Newton–Raphsons metod för ickelinjära ekvationssystem

Vi påminner i detta kapitel om en välkänd numerisk metod för ickelinjära ekvationssystem. Betrakta följande ekvationssystem i variabelvektorn $\mathbf{x} = (x_1, \ldots, x_n)^\mathsf{T} \in I\!R^n$:

$$\begin{aligned} h_1(\mathbf{x}) &= 0, \\ &\vdots \\ h_n(\mathbf{x}) &= 0, \end{aligned} \tag{7.1}$$

där $h_1, \ldots, h_n$ är givna reellvärda funktioner som antas vara kontinuerligt deriverbara. Minst en av dessa h_i antas vara ickelinjär (annars är ju (7.1) ett *linjärt* ekvationssystem). Observera det viktiga antagandet att systemet har lika många ekvationer som variabler!

Newton–Raphsons metod för att söka en lösning till detta system är iterativ så det räcker att beskriva hur man kommer från en iterationspunkt $\mathbf{x}^{(k)}$ till nästa iterationspunkt $\mathbf{x}^{(k+1)}$. I den givna iterationspunkten $\mathbf{x}^{(k)}$ linjäriserar man först de enskilda funktionerna h_i, dvs approximerar varje $h_i(\mathbf{x})$ med första ordningens Taylorpolynom i $\mathbf{x}^{(k)}$:

$$h_i(\mathbf{x}) \approx h_i(\mathbf{x}^{(k)}) + \nabla h_i(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)}),\ i = 1, \ldots, n. \tag{7.2}$$

Med variabelbytet $\mathbf{d} = \mathbf{x} - \mathbf{x}^{(k)}$, dvs $\mathbf{x} = \mathbf{x}^{(k)} + \mathbf{d}$, kan detta skrivas

$$h_i(\mathbf{x}^{(k)} + \mathbf{d}) \approx h_i(\mathbf{x}^{(k)}) + \nabla h_i(\mathbf{x}^{(k)})\,\mathbf{d},\ i = 1, \ldots, n. \tag{7.3}$$

Låt $\mathbf{h}(\mathbf{x}) \in I\!R^n$ vara en kolonnvektor med komponenterna $h_1(\mathbf{x}), \ldots, h_n(\mathbf{x})$, och låt $\nabla\mathbf{h}(\mathbf{x})$ vara en $n \times n$–matris med raderna $\nabla h_1(\mathbf{x}), \ldots, \nabla h_n(\mathbf{x})$, dvs

$$\mathbf{h}(\mathbf{x}) = \begin{pmatrix} h_1(\mathbf{x}) \\ \vdots \\ h_n(\mathbf{x}) \end{pmatrix} \quad \text{och} \quad \nabla\mathbf{h}(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial h_1}{\partial x_1}(\mathbf{x}) & \cdots & \dfrac{\partial h_1}{\partial x_n}(\mathbf{x}) \\ \vdots & & \vdots \\ \dfrac{\partial h_n}{\partial x_1}(\mathbf{x}) & \cdots & \dfrac{\partial h_n}{\partial x_n}(\mathbf{x}) \end{bmatrix}. \tag{7.4}$$

Då kan ekvationssystemet (7.1) kortfattat skrivas

$$\mathbf{h}(\mathbf{x}) = \mathbf{0}, \tag{7.5}$$

medan approximationerna (7.3) kan skrivas på formen

$$\mathbf{h}(\mathbf{x}^{(k)} + \mathbf{d}) \approx \mathbf{h}(\mathbf{x}^{(k)}) + \nabla\mathbf{h}(\mathbf{x}^{(k)})\,\mathbf{d}. \tag{7.6}$$

I Newton–Raphsons metod löser man nu ekvationssystemet $\mathbf{h}(\mathbf{x}^{(k)}) + \nabla\mathbf{h}(\mathbf{x}^{(k)})\,\mathbf{d} = \mathbf{0}$, som ekvivalent kan skrivas

$$\nabla\mathbf{h}(\mathbf{x}^{(k)})\,\mathbf{d} = -\mathbf{h}(\mathbf{x}^{(k)}). \tag{7.7}$$

Detta är ett linjärt ekvationssystem i vektorn $\mathbf{d} \in I\!R^n$. Det har en unik lösning $\mathbf{d}^{(k)}$ om den kvadratiska matrisen $\nabla\mathbf{h}(\mathbf{x}^{(k)})$ är ickesingulär (dvs om dess kolonner är linjärt oberoende, vilket de är i "normalfallet").

Om $\mathbf{d}^{(k)}$ är lösningen till (7.7) så ges nästa iterationspunkt idealt av

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{d}^{(k)}, \tag{7.8}$$

så att $\mathbf{h}(\mathbf{x}^{(k)}) + \nabla\mathbf{h}(\mathbf{x}^{(k)})(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) = \mathbf{0}$.

Men om man använder uppdateringsformeln (7.8), och har en dålig startpunkt, så kan det på vissa problem inträffa att sekvensen av iterationspunkter divergerar. Därför brukar man i stället för (7.8) använda sig av formeln

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t_k\mathbf{d}^{(k)}, \tag{7.9}$$

där t_k exempelvis väljs som det största av talen $1, \frac{1}{2}, \frac{1}{4}, \ldots$ för vilket

$$|\mathbf{h}(\mathbf{x}^{(k)} + t_k\mathbf{d}^{(k)})|^2 < |\mathbf{h}(\mathbf{x}^{(k)})|^2. \tag{7.10}$$

Om den kvadratiska matrisen $\nabla\mathbf{h}(\mathbf{x}^{(k)})$ är ickesingulär och $\mathbf{h}(\mathbf{x}^{(k)}) \neq \mathbf{0}$ så är (7.10) uppfyllt för alla tillräckligt små $t_k > 0$.

Om man startar "tillräckligt nära" en lösning till ekvationssystemet (7.1) så är Newton–Raphsons metod ofta mycket effektiv. Men i det allmänna fallet är det inte säkert att sekvensen av iterationspunkter konvergerar mot en lösning till ekvationssystemet (7.1), även om det finns en sådan lösning. Det kan exempelvis hända att sekvensen konvergerar mot en punkt $\hat{\mathbf{x}}$ som är en lokal men inte global minpunkt till $|\mathbf{h}(\mathbf{x})|^2$, med $\mathbf{h}(\hat{\mathbf{x}}) \neq \mathbf{0}$. I detta fall är $\nabla\mathbf{h}(\hat{\mathbf{x}})$ singulär.

8 Ickelinjära minsta-kvadratproblem och Gauss–Newtons metod

I detta avsnitt ska vi behandla så kallade ickelinjära minsta-kvadratproblem. Dessa dyker upp i många olika tillämpningar, bland annat när man ska anpassa en matematisk modell till givna mätdata. Man behöver då bestämma värden på vissa parametrar i modellen, och detta ska göras på ett sådant sätt att avvikelserna mellan modell och mätdata blir så små som möjligt. Detta uttrycks ofta som att man vill minimera en kvadratsumma av avvikelserna.

Om parametrar ingår linjärt i modellen får man ett linjärt minsta-kvadratproblem. Denna (relativt enkla) typ av problem har vi redan behandlat i kursavsnittet Kvadratisk optimering. Om däremot parametrarna ingår på ett ickelinjärt sätt i modellen (vilket inte är ovanligt) får man ett ickelinjärt minsta-kvadratproblem. Sådana problem kan skrivas på formen

$$\text{minimera } \ f(\mathbf{x}) = \tfrac{1}{2}\sum_{i=1}^{m}(h_i(\mathbf{x}))^2, \tag{8.1}$$

där $h_1, \ldots, h_m$ är givna funktioner från $\mathbb{R}^n$ till $\mathbb{R}$. Faktorn $\frac{1}{2}$ är instoppad för att förenkla vissa kommande uttryck. Vanligtvis är m betydligt större än n, så att antalet funktioner h_i är betydligt större än antalet variabler x_j. Vid modellanpassning svarar detta mot att antalet mätdata är betydligt större än antalet parametar som ska bestämmas.

En ofta relevant tolkning av optimeringsproblemet (8.1) är att vi egentligen skulle vilja lösa det ickelinjära ekvationssystemet

$$\begin{aligned} h_1(\mathbf{x}) &= 0, \\ &\vdots \\ h_m(\mathbf{x}) &= 0, \end{aligned} \tag{8.2}$$

men eftersom antalet ekvationer (m st) är större än antalet variabler (n st) så finns det typiskt ingen lösning till detta system. Då är det naturligt att söka efter en lösning $\mathbf{x}$ som bryter "så lite som möjligt" mot (8.2), exempelvis en optimal lösning till (8.1). (Notera att om systemet (8.2) skulle råka ha en lösning $\hat{\mathbf{x}}$ så är $\hat{\mathbf{x}}$ en globalt optimal lösning till problemet (8.1).)

Här följer ett konkret exempel på ett ickelinjärt minsta-kvadratproblem:
Antag att man vill bestämma koordinaterna $(x_1,\ x_2)$ för en punkt P_0 med hjälp av uppmätta avstånd $b_1, \ldots, b_m$ till P_0 från var och en av m st referenspunkter $P_1, \ldots, P_m$ som har kända koordinater. Antag speciellt att $m = 4$, att punkterna P_1, P_2, P_3 och P_4 har koordinaterna $(-40,\ 30)$, $(40,\ 30)$, $(-30, -40)$ resp $(30, -40)$, och att motsvarande uppmätta avstånd är $b_1 = 51,\ b_2 = 52,\ b_3 = 48$ resp $b_4 = 49$.
Helst skulle vi här vilja bestämma x_1 och x_2 så att

$$\begin{aligned} h_1(\mathbf{x}) &= \sqrt{(x_1+40)^2 + (x_2-30)^2} - 51 = 0, \\ h_2(\mathbf{x}) &= \sqrt{(x_1-40)^2 + (x_2-30)^2} - 52 = 0, \\ h_3(\mathbf{x}) &= \sqrt{(x_1+30)^2 + (x_2+40)^2} - 48 = 0, \\ h_4(\mathbf{x}) &= \sqrt{(x_1-30)^2 + (x_2+40)^2} - 49 = 0. \end{aligned} \tag{8.3}$$

Men eftersom de uppmätta avstånden b_i inte är exakta utan innehåller okända mätfel så går det inte att bestämma $\mathbf{x} = (x_1,\, x_2)^{\mathsf{T}}$ som exakt uppfyller alla fyra ekvationerna $h_i(\mathbf{x}) = 0$. I stället betraktar vi minsta-kvadratproblemet (MK-problemet)

$$\text{minimera} \quad f(\mathbf{x}) = \tfrac{1}{2}(h_1(\mathbf{x})^2 + h_2(\mathbf{x})^2 + h_3(\mathbf{x})^2 + h_4(\mathbf{x})^2). \tag{8.4}$$

Eftersom funktionerna h_i är ickelinjära i x_1 och x_2 så är (8.4) ett ickelinjärt MK-problem.

Även om man i princip kan använda Newtons metod (eventuellt modifierad enligt avsnitt 6 ovan) för att minimera $f(\mathbf{x})$ definierad av (8.1), så är det oftast både enklare och effektivare att använda den så kallade Gauss–Newtons metod, som bättre utnyttjar den speciella struktur som problemet har. Denna metod kan tolkas på två olika sätt. Vi ska här ge bägge dessa tolkningar, och börjar med den tekniskt enklaste.

Metoden är iterativ, så det räcker att beskriva hur man kommer från en iterationspunkt $\mathbf{x}^{(k)}$ till nästa iterationspunkt $\mathbf{x}^{(k+1)}$.
I den givna iterationspunkten $\mathbf{x}^{(k)}$ linjäriserar man först de enskilda funktionerna h_i, dvs approximerar varje $h_i(\mathbf{x})$ med första ordningens Taylorpolynom i $\mathbf{x}^{(k)}$:

$$h_i(\mathbf{x}) \approx h_i(\mathbf{x}^{(k)}) + \nabla h_i(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)}),\ \ i = 1, \ldots, m. \tag{8.5}$$

Med variabelbytet $\mathbf{d} = \mathbf{x} - \mathbf{x}^{(k)}$, dvs $\mathbf{x} = \mathbf{x}^{(k)} + \mathbf{d}$, kan detta skrivas

$$h_i(\mathbf{x}^{(k)} + \mathbf{d}) \approx h_i(\mathbf{x}^{(k)}) + \nabla h_i(\mathbf{x}^{(k)})\,\mathbf{d},\ \ i = 1, \ldots, m. \tag{8.6}$$

Låt $\mathbf{h}(\mathbf{x}) \in I\!R^m$ vara en kolonnvektor med komponenterna $h_1(\mathbf{x}), \ldots, h_m(\mathbf{x})$, och låt $\nabla\mathbf{h}(\mathbf{x})$ vara en $m \times n$–matris med raderna $\nabla h_1(\mathbf{x}), \ldots, \nabla h_m(\mathbf{x})$, dvs

$$\mathbf{h}(\mathbf{x}) = \begin{pmatrix} h_1(\mathbf{x}) \\ \\ \vdots \\ \\ h_m(\mathbf{x}) \end{pmatrix} \quad \text{och} \quad \nabla\mathbf{h}(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial h_1}{\partial x_1}(\mathbf{x}) & \cdots & \dfrac{\partial h_1}{\partial x_n}(\mathbf{x}) \\ \\ \vdots & & \vdots \\ \\ \dfrac{\partial h_m}{\partial x_1}(\mathbf{x}) & \cdots & \dfrac{\partial h_m}{\partial x_n}(\mathbf{x}) \end{bmatrix}. \tag{8.7}$$

Då kan målfunktionen kortfattat skrivas $f(\mathbf{x}) = \frac{1}{2}\,|\mathbf{h}(\mathbf{x})|^2$, medan approximationerna (8.6) kan skrivas på formen

$$\mathbf{h}(\mathbf{x}^{(k)} + \mathbf{d}) \approx \mathbf{h}(\mathbf{x}^{(k)}) + \nabla\mathbf{h}(\mathbf{x}^{(k)})\,\mathbf{d}. \tag{8.8}$$

Motsvarande approximation av målfunktionen $f(\mathbf{x})$ är då

$$f(\mathbf{x}^{(k)} + \mathbf{d}) = \tfrac{1}{2}\,|\mathbf{h}(\mathbf{x}^{(k)} + \mathbf{d})|^2 \approx \tfrac{1}{2}\,|\mathbf{h}(\mathbf{x}^{(k)}) + \nabla\mathbf{h}(\mathbf{x}^{(k)})\mathbf{d}|^2 = \tfrac{1}{2}\,|\mathbf{A}^{(k)}\mathbf{d} - \mathbf{b}^{(k)}|^2, \tag{8.9}$$

där vi infört matrisen $\mathbf{A}^{(k)} = \nabla\mathbf{h}(\mathbf{x}^{(k)})$ och vektorn $\mathbf{b}^{(k)} = -\mathbf{h}(\mathbf{x}^{(k)})$. I Gauss–Newtons metod minimerar man nu högerledet i (8.9), dvs löser följande problem i variabelvektorn $\mathbf{d} \in I\!R^n$:

$$\text{minimera} \ \ \tfrac{1}{2}\,|\mathbf{A}^{(k)}\mathbf{d} - \mathbf{b}^{(k)}|^2. \tag{8.10}$$

Men detta är ett linjärt minsta-kvadratproblem, som enligt avsnittet om kvadratisk optimering är ekvivalent med normalekvationerna $\mathbf{A}^{(k)\mathsf{T}}\mathbf{A}^{(k)}\mathbf{d} = \mathbf{A}^{(k)\mathsf{T}}\mathbf{b}$, dvs

$$\nabla\mathbf{h}(\mathbf{x}^{(k)})^{\mathsf{T}}\nabla\mathbf{h}(\mathbf{x}^{(k)})\,\mathbf{d} = -\nabla\mathbf{h}(\mathbf{x}^{(k)})^{\mathsf{T}}\mathbf{h}(\mathbf{x}^{(k)}). \tag{8.11}$$

Om $\mathbf{d}^{(k)}$ är en lösning till (8.11) så ges nästa iterationspunkt av

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t_k \mathbf{d}^{(k)}, \tag{8.12}$$

där t_k exempelvis väljs som det största av talen $1, \frac{1}{2}, \frac{1}{4}, \ldots$ för vilket

$$f(\mathbf{x}^{(k)} + t_k \mathbf{d}^{(k)}) < f(\mathbf{x}^{(k)}). \tag{8.13}$$

Om kolonnerna i $\nabla \mathbf{h}(\mathbf{x}^{(k)})$ är linjärt oberoende (vilket är uppfyllt i "normalfallet" eftersom $m > n$) så är matrisen $\nabla \mathbf{h}(\mathbf{x}^{(k)})^\mathsf{T} \nabla \mathbf{h}(\mathbf{x}^{(k)})$ positivt definit, varvid lösningen $\mathbf{d}^{(k)}$ till (8.11) är unik. Om dessutom $\mathbf{d}^{(k)} \neq \mathbf{0}$ så är (8.13) uppfyllt för alla tillräckligt små $t_k > 0$.

Den andra tolkningen av Gauss–Newtons metod utgår från att gradienten och Hessianen av målfunktionen $f(\mathbf{x}) = \frac{1}{2} |\mathbf{h}(\mathbf{x})|^2$ kan skrivas på följande form (efter en del kalkyler som vi hoppar över här):

$$\nabla f(\mathbf{x}) = \mathbf{h}(\mathbf{x})^\mathsf{T} \nabla \mathbf{h}(\mathbf{x}), \tag{8.14}$$

$$\mathbf{F}(\mathbf{x}) = \nabla \mathbf{h}(\mathbf{x})^\mathsf{T} \nabla \mathbf{h}(\mathbf{x}) + \sum_{i=1}^{m} h_i(\mathbf{x}) \mathbf{H}_i(\mathbf{x}), \tag{8.15}$$

där $\mathbf{H}_i(\mathbf{x})$ är Hessianen till funktionen h_i.

Om nu $\mathbf{F}(\mathbf{x}^{(k)})$ är positivt definit, där $\mathbf{x}^{(k)}$ är den aktuella iterationspunkten, så kan man använda Newtons metod, dvs bestämma en sökriktning $\mathbf{d}^{(k)}$ ur ekvationssystemet:

$$\left(\nabla \mathbf{h}(\mathbf{x}^{(k)})^\mathsf{T} \nabla \mathbf{h}(\mathbf{x}^{(k)}) + \sum_{i=1}^{m} h_i(\mathbf{x}^{(k)}) \mathbf{H}_i(\mathbf{x}^{(k)}) \right) \mathbf{d} = -\nabla \mathbf{h}(\mathbf{x}^{(k)})^\mathsf{T} \mathbf{h}(\mathbf{x}^{(k)}). \tag{8.16}$$

Om $\mathbf{F}(\mathbf{x}^{(k)})$ inte är positivt definit så måste man modifiera (8.16) enligt kapitel 6 ovan.
I många fall är det dock rimligt att göra approximationen

$$\mathbf{F}(\mathbf{x}^{(k)}) \approx \nabla \mathbf{h}(\mathbf{x}^{(k)})^\mathsf{T} \nabla \mathbf{h}(\mathbf{x}^{(k)}), \tag{8.17}$$

och allså bortse från summan $\sum_i h_i(\mathbf{x}^{(k)}) \mathbf{H}_i(\mathbf{x}^{(k)})$ i (8.16). Vid exempelvis modellanpassning är det troligt att varje $h_i(\mathbf{x})$ är "ganska liten", åtminstone efter några iterationer om modellen ansluter väl till data. Det kan också i vissa fall vara så att funktionerna h_i är "nästan linjära", så att andraderivatorna är små. Om vi använder approximationen (8.17) så övergår ekvationssystemet (8.16) till ekvationssystemet

$$\nabla \mathbf{h}(\mathbf{x}^{(k)})^\mathsf{T} \nabla \mathbf{h}(\mathbf{x}^{(k)})\, \mathbf{d} = -\nabla \mathbf{h}(\mathbf{x}^{(k)})^\mathsf{T} \mathbf{h}(\mathbf{x}^{(k)}), \tag{8.18}$$

vilket är detsamma som (8.11) ovan.
En viktig fördel med Gauss–Newtons metod (8.18), jämfört med Newtons metod (8.16), är förstås att man inte behöver beräkna några andraderivator.

Vi återvänder nu till exemplet i början av kapitlet, där funktionerna h_i gavs av (8.3). Detta problem ska angripas med Gauss–Newtons metod.

Antag att vi startar i $\mathbf{x}^{(1)} = (0,\, 0)^\mathsf{T}$. Då är $\mathbf{h}(\mathbf{x}^{(1)}) = (-1, -2,\, 2,\, 1)^\mathsf{T}$ och $f(\mathbf{x}^{(1)}) = 5$.

Vidare är exempelvis $\dfrac{\partial h_3}{\partial x_2}(\mathbf{x}) = \dfrac{x_2 + 40}{\sqrt{(x_1+30)^2 + (x_2+40)^2}}$, så att $\dfrac{\partial h_3}{\partial x_2}(\mathbf{x}^{(1)}) = \dfrac{40}{50} = 0.8$.

Motsvarande kalkyler för övriga förstaderivatorna ger att $\nabla\mathbf{h}(\mathbf{x}^{(1)}) = \begin{bmatrix} 0.8 & -0.6 \\ -0.8 & -0.6 \\ 0.6 & 0.8 \\ -0.6 & 0.8 \end{bmatrix}$.

Därmed är $\nabla\mathbf{h}(\mathbf{x}^{(1)})^\mathsf{T}\nabla\mathbf{h}(\mathbf{x}^{(1)}) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ och $\nabla\mathbf{h}(\mathbf{x}^{(1)})^\mathsf{T}\mathbf{h}(\mathbf{x}^{(1)}) = \begin{pmatrix} 1.4 \\ 4.2 \end{pmatrix}$.

Systemet (8.11) blir då $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = -\begin{pmatrix} 1.4 \\ 4.2 \end{pmatrix}$, med lösningen $\mathbf{d}^{(1)} = \begin{pmatrix} -0.7 \\ -2.1 \end{pmatrix}$.

Vi prövar enhetssteget ($t_1 = 1$) och sätter $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + t_1\mathbf{d}^{(1)} = \mathbf{x}^{(1)} + \mathbf{d}^{(1)} = \begin{pmatrix} -0.7 \\ -2.1 \end{pmatrix}$.
Då är $\mathbf{h}(\mathbf{x}^{(2)}) = (-0.2565, -0.1647, -0.0949, -0.2260)^\mathsf{T}$ och $f(\mathbf{x}^{(2)}) = 0.0765 < f(\mathbf{x}^{(1)})$, så enhetssteget gick bra. Därmed har vi utfört en fullständig iteration.
De fortsatta räkningarna är jobbiga att utföra för hand, så vi avbryter här.

Man bör vara medveten om att problemet (8.1) i allmänhet inte är ett konvext problem. Därför kan man normalt inte vara säker på att hitta en *globalt* optimal lösning till (8.1).