Krister Svanberg, maj 2012

1 Ickelinjär optimering under bivillkor

Hittills har vi behandlat optimeringsproblem där alla variabler x_j kunnat röra sig fritt, oberoende av varann, och anta hur stora eller små värden som helst. I många (de flesta) tillämpningar av optimering finns inte denna frihet. Fortsättningsvis ska vi behandla problem där variablerna måste uppfylla vissa så kallade *bivillkor* (på engelska *constraints*).
Vi inleder med optimeringsproblem på den generella formen

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

där $\mathcal{F}$ är en given delmängd av $I\!R^n$ och f är en given reellvärd funktion definierad (åtminstone) på hela $\mathcal{F}$. Mängden $\mathcal{F}$ kallas för det tillåtna området till problemet (1.1). Längre fram kommer $\mathcal{F}$ att ges en mer explicit form, exempelvis för problem med likhetsbivillkor:

$$\mathcal{F} = \{\, \mathbf{x} \in I\!R^n \mid h_i(\mathbf{x}) = 0, \ \text{för } i = 1, \ldots, m\}, \tag{1.2}$$

där $h_1, \ldots, h_m$ är givna funktioner från $I\!R^n$ till $I\!R$, eller för problem med olikhetsbivillkor:

$$\mathcal{F} = \{\, \mathbf{x} \in I\!R^n \mid g_i(\mathbf{x}) \leq 0, \ \text{för } i = 1, \ldots, m\}, \tag{1.3}$$

där $g_1, \ldots, g_m$ är givna funktioner från $I\!R^n$ till $I\!R$.

Def 1.1. Punkten $\mathbf{x} \in I\!R^n$ är en tillåten lösning till problemet (1.1) om $\mathbf{x} \in \mathcal{F}$.
Punkten $\hat{\mathbf{x}} \in \mathcal{F}$ är en *lokal* optimallösning till problemet (1.1) om det finns ett tal $\delta > 0$ sådant att $f(\hat{\mathbf{x}}) \leq f(\mathbf{x})$ för alla $\mathbf{x} \in \mathcal{F}$ som uppfyller $|\mathbf{x} - \hat{\mathbf{x}}| < \delta$.
Punkten $\hat{\mathbf{x}} \in \mathcal{F}$ är en *global* optimallösning till problemet (1.1) om $f(\hat{\mathbf{x}}) \leq f(\mathbf{x})$ för *alla* $\mathbf{x} \in \mathcal{F}$.

Det är uppenbart att varje global optimallösning även är en lokal optimallösning, men det kan för vissa problem finnas lokala optimallösningar som inte är globala optimallösningar.

Def 1.2. Vektorn $\mathbf{d} \in I\!R^n$ är en *tillåten riktning* till problemet (1.1) i punkten $\mathbf{x} \in \mathcal{F}$ om det finns ett tal $\delta > 0$ sådant att $\mathbf{x} + t\,\mathbf{d} \in \mathcal{F}$ för alla $t \in (0, \delta)$.
Vektorn $\mathbf{d} \in I\!R^n$ är en *tillåten avtaganderiktning* till problemet (1.1) i $\mathbf{x} \in \mathcal{F}$ om det finns ett tal $\delta > 0$ sådant att dels $\mathbf{x} + t\,\mathbf{d} \in \mathcal{F}$ för alla $t \in (0, \delta)$, dels $f(\mathbf{x} + t\,\mathbf{d}) < f(\mathbf{x})$ för alla $t \in (0, \delta)$.

Lemma 1.1. Antag att $\hat{\mathbf{x}} \in \mathcal{F}$ är en lokal optimallösning till problemet (1.1).
Då finns det *inte* någon tillåten avtaganderiktning $\mathbf{d}$ i $\hat{\mathbf{x}}$.

Bevis: Om det finns en tillåten avtaganderiktning $\mathbf{d}$ i $\hat{\mathbf{x}}$ så finns det alltså ett tal $\delta > 0$ sådant att det för *varje* $t \in (0, \delta)$ gäller dels att $\hat{\mathbf{x}} + t\,\mathbf{d} \in \mathcal{F}$, dels att $f(\hat{\mathbf{x}} + t\,\mathbf{d}) < f(\hat{\mathbf{x}})$. Men det innebär att det godtyckligt nära $\hat{\mathbf{x}}$ finns tillåtna punkter $\mathbf{x} = \hat{\mathbf{x}} + t\,\mathbf{d}$ som har ett lägre målfunktionsvärde än $\hat{\mathbf{x}}$, dvs $f(\mathbf{x}) < f(\hat{\mathbf{x}})$, vilket i sin tur medför att $\hat{\mathbf{x}}$ *inte* är en lokal optimallösning till problemet (1.1).

1.1 Optimalitetsvillkor för problem med enbart likhetsbivillkor

I detta avsnitt antar vi att mängden $\mathcal{F}$ definieras av ett antal likhetsbivillkor, dvs

$$\mathcal{F} = \{\mathbf{x} \in \mathbb{R}^n \mid h_i(\mathbf{x}) = 0, \text{ för } i = 1, \dots, m\}, \tag{1.4}$$

där $h_1, \dots, h_m$ är givna funktioner från $\mathbb{R}^n$ till $\mathbb{R}$.
Problemet (1.1) övergår då till följande problem:

$$\begin{aligned} \text{minimera} \quad & f(\mathbf{x}) \\ \text{då} \quad & h_i(\mathbf{x}) = 0, \text{ för } i = 1, \dots, m. \end{aligned} \tag{1.5}$$

Såväl bivillkorsfunktionerna h_i som målfunktionen f antas vara kontinuerligt deriverbara.
I normalfallet är $m < n$, vilket betyder att bivillkoren definierar ett (ickelinjärt) ekvationssystem med fler obekanta (n st) än ekvationer (m st). Oftast har detta ekvationssystem oändligt många lösningar, och optimeringsproblemet består i att bland dessa bestämma en lösning $\mathbf{x}$ med så lågt målfunktionsvärde $f(\mathbf{x})$ som möjligt.

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

$$\mathbf{h}(\mathbf{x}) = \begin{pmatrix} h_1(\mathbf{x}) \\ \vdots \\ h_m(\mathbf{x}) \end{pmatrix} \quad \text{och} \quad \nabla\mathbf{h}(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial h_1}{\partial x_1}(\mathbf{x}) & \cdots & \dfrac{\partial h_1}{\partial x_n}(\mathbf{x}) \\ \vdots & & \vdots \\ \dfrac{\partial h_m}{\partial x_1}(\mathbf{x}) & \cdots & \dfrac{\partial h_m}{\partial x_n}(\mathbf{x}) \end{bmatrix}. \tag{1.6}$$

Def 1.1.1. En tillåten lösning $\mathbf{x} \in \mathcal{F}$ är en *regulär punkt* till problemet (1.5) om raderna i matrisen $\nabla\mathbf{h}(\mathbf{x})$ ovan är linjärt oberoende, dvs om gradientvektorerna $\nabla h_i(\mathbf{x})$ är linjärt oberoende, dvs om det *inte* finns några skalärer u_i, för $i = 1, \dots, m$, sådana att:

$$(u_1, \dots, u_m) \neq (0, \dots, 0) \quad \text{och} \quad \sum_{i=1}^{m} u_i \nabla h_i(\mathbf{x}) \;=\; \mathbf{0}^\mathsf{T}.$$

Eftersom matrisen $\nabla\mathbf{h}(\mathbf{x})$ i normalfallet har färre rader än kolonner, $m < n$, så är "nästan alltid" dess rader linjärt oberoende. Man ska i viss mening ha "otur" för att stöta på en icke regulär punkt. Men ibland har man ju otur.

Lemma 1.1.1. Antag att $\hat{\mathbf{x}} \in \mathcal{F}$ är både en regulär punkt och en lokal optimallösning till (1.5). Då finns det ingen vektor $\mathbf{d} \in \mathbb{R}^n$ som samtidigt uppfyller att $\nabla f(\hat{\mathbf{x}})\,\mathbf{d} < 0$ och $\nabla h_i(\hat{\mathbf{x}})\,\mathbf{d} = 0$ för alla $i = 1, \dots, m$.

Beviset av Lemma 1.1.1 är ganska komplicerat och utelämnas (det bygger bland annat på den så kallade Implicita funktionssatsen.)

Utgående från Lemma 1.1.1 är det dock ganska lätt att bevisa följande viktiga sats:

Sats 1.1.1. Antag att $\hat{\mathbf{x}} \in \mathcal{F}$ är både en regulär punkt och en lokal optimallösning till (1.5). Då finns det en vektor $\hat{\mathbf{u}} \in I\!R^m$ som tillsammans med $\hat{\mathbf{x}}$ uppfyller

$$\nabla f(\hat{\mathbf{x}}) + \sum_{i=1}^{m} \hat{u}_i \nabla h_i(\hat{\mathbf{x}}) = \mathbf{0}^{\mathsf{T}}. \tag{1.7}$$

Bevis: Antag att $\hat{\mathbf{x}} \in \mathcal{F}$ är både en regulär punkt och en lokal optimallösning till (1.5).
Enligt Lemma 1.1.1 finns det då ingen vektor $\mathbf{d}$ sådan att $\nabla f(\hat{\mathbf{x}})\,\mathbf{d} < 0$ och $\nabla \mathbf{h}(\hat{\mathbf{x}})\,\mathbf{d} = \mathbf{0}$.
Men då finns det heller inte någon vektor $\mathbf{d}$ sådan att $\nabla f(\hat{\mathbf{x}})\,\mathbf{d} > 0$ och $\nabla \mathbf{h}(\hat{\mathbf{x}})\,\mathbf{d} = \mathbf{0}$, ty i såfall skulle vektorn $-\mathbf{d}$ uppfylla $\nabla f(\hat{\mathbf{x}})(-\mathbf{d}) < 0$ och $\nabla \mathbf{h}(\hat{\mathbf{x}})(-\mathbf{d}) = \mathbf{0}$, vilket strider mot Lemma 1.1.1. Alltså är $\nabla f(\hat{\mathbf{x}})\,\mathbf{d} = 0$ för alla $\mathbf{d}$ som uppfyller $\nabla \mathbf{h}(\hat{\mathbf{x}})\,\mathbf{d} = \mathbf{0}$.
Det betyder att (kolonn-)vektorn $\nabla f(\hat{\mathbf{x}})^{\mathsf{T}} \in I\!R^n$ är ortogonal mot varje vektor $\mathbf{d} \in I\!R^n$ som ligger i nollrummet till matrisen $\nabla \mathbf{h}(\hat{\mathbf{x}})$, dvs att $\nabla f(\hat{\mathbf{x}})^{\mathsf{T}} \in \mathcal{N}(\nabla \mathbf{h}(\hat{\mathbf{x}}))^{\perp}$.
Men $\mathcal{N}(\nabla \mathbf{h}(\hat{\mathbf{x}}))^{\perp} = \mathcal{R}(\nabla \mathbf{h}(\hat{\mathbf{x}})^{\mathsf{T}})$, så detta är ekvivalent med att $\nabla f(\hat{\mathbf{x}})^{\mathsf{T}} \in \mathcal{R}(\nabla \mathbf{h}(\hat{\mathbf{x}})^{\mathsf{T}})$, vilket i sin tur är ekvivalent med att $\nabla f(\hat{\mathbf{x}})^{\mathsf{T}} = \nabla \mathbf{h}(\hat{\mathbf{x}})^{\mathsf{T}}(-\hat{\mathbf{u}})$ för någon vektor $\hat{\mathbf{u}} \in I\!R^m$, vilket kan skrivas $\nabla f(\hat{\mathbf{x}})^{\mathsf{T}} + \nabla \mathbf{h}(\hat{\mathbf{x}})^{\mathsf{T}}\hat{\mathbf{u}} = \mathbf{0}$, eller $\nabla f(\hat{\mathbf{x}}) + \hat{\mathbf{u}}^{\mathsf{T}}\nabla \mathbf{h}(\hat{\mathbf{x}}) = \mathbf{0}^{\mathsf{T}}$, som är (1.7).

Man kan notera att (1.7) är n st ekvationer. Tillsammans med de m st bivillkoren $h_i(\hat{\mathbf{x}}) = 0$ har man därmed ett (ickelinjärt) ekvationssystem med $n+m$ st ekvationer som ska uppfyllas av de $m+n$ st obekanta $\hat{x}_1, \ldots, \hat{x}_n, \hat{u}_1, \ldots, \hat{u}_m$. På kompakt form kan detta system skrivas:

$$\begin{aligned} \nabla f(\hat{\mathbf{x}})^{\mathsf{T}} \;+\; \nabla \mathbf{h}(\hat{\mathbf{x}})^{\mathsf{T}}\hat{\mathbf{u}} &= \mathbf{0}\,, \\ \mathbf{h}(\hat{\mathbf{x}}) &= \mathbf{0}\,. \end{aligned} \tag{1.8}$$

Exempel 1.1.1.
Vid kvadratisk optimering under linjära likhetsbivillkor har man problem på formen

$$\begin{aligned} \text{minimera} \quad & \tfrac{1}{2}\,\mathbf{x}^{\mathsf{T}}\mathbf{H}\mathbf{x} + \mathbf{c}^{\mathsf{T}}\mathbf{x} + c_0 \\ \text{då} \quad & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{aligned} \tag{1.9}$$

där $\mathbf{A} \in I\!R^{m\times n}$ och $\mathbf{H} \in I\!R^{n\times n}$ är givna matriser med $\mathbf{H}$ symmetrisk.
Detta är ett specialfall av (1.5), med $f(\mathbf{x}) = \frac{1}{2}\,\mathbf{x}^{\mathsf{T}}\mathbf{H}\mathbf{x} + \mathbf{c}^{\mathsf{T}}\mathbf{x} + c_0$ och $\mathbf{h}(\mathbf{x}) = \mathbf{b} - \mathbf{A}\mathbf{x}$.
Här blir $\nabla f(\mathbf{x})^{\mathsf{T}} = \mathbf{H}\mathbf{x} + \mathbf{c}$ och $\nabla \mathbf{h}(\mathbf{x}) = -\mathbf{A}$,
varvid systemet (1.8) övergår till till följande *linjära* ekvationssystem i $\hat{\mathbf{x}}$ och $\hat{\mathbf{u}}$:

$$\begin{bmatrix} \mathbf{H} & -\mathbf{A}^{\mathsf{T}} \\ \mathbf{A} & \mathbf{0} \end{bmatrix} \begin{pmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} -\mathbf{c} \\ \mathbf{b} \end{pmatrix}. \tag{1.10}$$

1.2 Optimalitetsvillkor för problem med enbart olikhetsbivillkor

I detta avsnitt antar vi att mängden $\mathcal{F}$ definieras av ett antal olikhetsbivillkor, dvs

$$\mathcal{F} = \{\mathbf{x} \in I\!R^n \mid g_i(\mathbf{x}) \leq 0, \text{ för } i = 1, \ldots, m\}, \tag{1.11}$$

där $g_1, \ldots, g_m$ är givna funktioner från $I\!R^n$ till $I\!R$.
Problemet (1.1) övergår då till följande problem:

$$\begin{aligned} \text{minimera} \quad & f(\mathbf{x}) \\ \text{då} \quad & g_i(\mathbf{x}) \leq 0, \text{ för } i = 1, \ldots, m. \end{aligned} \tag{1.12}$$

Såväl bivillkorsfunktionerna g_i som målfunktionen f antas vara kontinuerligt deriverbara. När vi i förra kapitlet behandlade problem med likhetsbivillkor, påstod vi att det i normalfallet gällde att $m < n$. När vi i detta avsnitt ska behandla problem med olikhetsbivillkor kan vi inte göra motsvarande påstående. Nu kan det mycket väl gälla att $m > n$ utan att problemet behöver vara urartat. Men det kan också gälla att $m < n$, eller att $m = n$. Vi behöver inte skilja på dessa tre olika fall i den fortsatta framställningen.

Def 1.2.1. För $\mathbf{x} \in \mathcal{F}$ låter vi $\mathcal{I}_a(\mathbf{x})$ beteckna indexmängden för de *aktiva* bivillkoren i punkten $\mathbf{x}$, dvs $\mathcal{I}_a(\mathbf{x}) = \{i \in \{1, ..., m\} \mid g_i(\mathbf{x}) = 0\}$.

Speciellt om $\mathbf{x} \in \mathcal{F}$ och $\mathcal{I}_a(\mathbf{x}) = \emptyset$ (tomma mängden) så är $g_i(\mathbf{x}) < 0$ för alla i, dvs samtliga bivillkor är uppfyllda med strikt olikhet. Sådana punkter $\mathbf{x}$ ligger i det *inre* av det tillåtna området $\mathcal{F}$, och de är oftast enklare att analysera, med avseende på optimalitet, än randpunkter till $\mathcal{F}$ som har $\mathcal{I}_a(\mathbf{x}) \neq \emptyset$.

Sats 1.2.1. Antag att $\hat{\mathbf{x}} \in \mathcal{F}$ med $\mathcal{I}_a(\hat{\mathbf{x}}) = \emptyset$ är en lokal optimallösning till (1.12).
Då är $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}$.

Bevis: När $\mathcal{I}_a(\hat{\mathbf{x}}) = \emptyset$ så är varje vektor $\mathbf{d} \in I\!R^n$ en tillåten riktning i $\hat{\mathbf{x}}$,
ty för varje i är $g_i(\hat{\mathbf{x}}) < 0$, vilket medför att $g_i(\hat{\mathbf{x}} + t\,\mathbf{d}) < 0$ för alla tillräckligt små $t > 0$ eftersom g_i är kontinuerlig.
Antag nu att $\nabla f(\hat{\mathbf{x}}) \neq \mathbf{0}^\mathsf{T}$.
Låt $\mathbf{d} = -\nabla f(\hat{\mathbf{x}})^\mathsf{T} \neq \mathbf{0}$. Enligt ovan är $\mathbf{d}$ en tillåten riktning i $\hat{\mathbf{x}}$.
Riktningsderivatan av funktionen f i punkten $\hat{\mathbf{x}}$ i riktningen $\mathbf{d}$ blir då
$\nabla f(\hat{\mathbf{x}})\,\mathbf{d} = -\nabla f(\hat{\mathbf{x}})\nabla f(\hat{\mathbf{x}})^\mathsf{T} = -|\nabla f(\hat{\mathbf{x}})^\mathsf{T}|^2 < 0$,
vilket betyder att $f(\hat{\mathbf{x}} + t\,\mathbf{d}) < f(\hat{\mathbf{x}})$ för alla tillräckligt små $t > 0$.
Sammantaget ger detta att $\mathbf{d}$ är en tillåten avtaganderiktning i $\hat{\mathbf{x}}$,
vilket enligt Lemma 1.1 medför att $\hat{\mathbf{x}}$ inte är en lokal optimallösning.

Fortsättningsvis ska vi analysera punkter med med $\mathcal{I}_a(\hat{\mathbf{x}}) \neq \emptyset$.

Lemma 1.2.1. Antag att $\hat{\mathbf{x}} \in \mathcal{F}$ med $\mathcal{I}_a(\hat{\mathbf{x}}) \neq \emptyset$ är en lokal optimallösning till (1.12).
Då finns det ingen vektor $\mathbf{d} \in I\!R^n$ som uppfyller samtliga följande olikheter:

$\nabla f(\hat{\mathbf{x}})\,\mathbf{d} < 0$ och $\nabla g_i(\hat{\mathbf{x}})\,\mathbf{d} < 0$ för $i \in \mathcal{I}_a(\hat{\mathbf{x}})$.

Bevis: Antag att vektorn $\mathbf{d} \in I\!\!R^n$ uppfyller samtliga ovanstående olikheter.
För varje $i \in \mathcal{I}_a(\hat{\mathbf{x}})$ är då $g_i(\hat{\mathbf{x}} + t\,\mathbf{d}) < g_i(\hat{\mathbf{x}}) = 0$ för alla tillräckligt små $t > 0$, eftersom riktningsderivatan $\nabla g_i(\hat{\mathbf{x}})\,\mathbf{d} < 0$ för $i \in \mathcal{I}_a(\hat{\mathbf{x}})$.
Men även för varje $i \notin \mathcal{I}_a(\hat{\mathbf{x}})$ är $g_i(\hat{\mathbf{x}} + t\,\mathbf{d}) < 0$ för alla tillräckligt små $t > 0$, eftersom g_i är kontinuerlig och $g_i(\hat{\mathbf{x}}) < 0$ för $i \notin \mathcal{I}_a(\hat{\mathbf{x}})$.
Alltså är $\mathbf{d}$ en tillåten riktning i $\hat{\mathbf{x}}$.
Vidare gäller att $f(\hat{\mathbf{x}} + t\,\mathbf{d}) < f(\hat{\mathbf{x}})$ för alla tillräckligt små $t > 0$, eftersom riktningsderivatan $\nabla f(\hat{\mathbf{x}})\,\mathbf{d} < 0$.
Sammantaget ger detta att $\mathbf{d}$ är en tillåten avtaganderiktning i $\hat{\mathbf{x}}$, vilket enligt Lemma 1.1 medför att $\hat{\mathbf{x}}$ inte är en lokal optimallösning.

Nu behöver vi följande resultat som är en enkel konsekvens av Farkas lemma:

Lemma 1.2.2. Antag att de m st vektorerna $\mathbf{a}_1, \ldots, \mathbf{a}_m \in I\!\!R^n$ är givna. Då har *exakt ett* av följande bägge system, i $\mathbf{d} \in I\!\!R^n$ respektive $\mathbf{v} \in I\!\!R^m$, minst en lösning.

$$\text{Antingen} \quad \mathbf{a}_i^\mathsf{T}\mathbf{d} < 0,\ i = 1, \ldots, m \quad \text{eller} \quad \begin{cases} \displaystyle\sum_{i=1}^{m} \mathbf{a}_i v_i = \mathbf{0}, \\ \displaystyle\sum_{i=1}^{m} v_i > 0, \\ v_i \geq 0, \quad i = 1, \ldots, m. \end{cases} \tag{1.13}$$

Lemma 1.2.3. Antag att $\hat{\mathbf{x}} \in \mathcal{F}$ med $\mathcal{I}_a(\hat{\mathbf{x}}) \neq \emptyset$ är en lokal optimallösning till (1.12). Då finns det skalärer $v_0 \geq 0$ och $v_i \geq 0$, för $i \in \mathcal{I}_a(\hat{\mathbf{x}})$, sådana att:

$$v_0 + \sum_{i \in \mathcal{I}_a(\hat{\mathbf{x}})} v_i > 0 \quad \text{och} \quad v_0 \nabla f(\hat{\mathbf{x}}) + \sum_{i \in \mathcal{I}_a(\hat{\mathbf{x}})} v_i \nabla g_i(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}.$$

Bevis: Kombinera Lemma 1.2.1 med Lemma 1.2.2.

Def 1.2.2. En tillåten lösning $\mathbf{x} \in \mathcal{F}$ med $\mathcal{I}_a(\mathbf{x}) \neq \emptyset$ är en *regulär punkt* till problemet (1.12) om det *inte* finns några skalärer $v_i \geq 0$, för $i \in \mathcal{I}_a(\mathbf{x})$, sådana att:

$$\sum_{i \in \mathcal{I}_a(\mathbf{x})} v_i > 0 \quad \text{och} \quad \sum_{i \in \mathcal{I}_a(\mathbf{x})} v_i \nabla g_i(\mathbf{x}) = \mathbf{0}^\mathsf{T}.$$

En tillåten lösning $\mathbf{x} \in \mathcal{F}$ med $\mathcal{I}_a(\mathbf{x}) = \emptyset$ är alltid en regulär punkt till (1.12).

Ett ekvivalent sätt att uttrycka denna definition är enligt Lemma 1.2.2 följande:

Def 1.2.3. En tillåten lösning $\mathbf{x} \in \mathcal{F}$ med $\mathcal{I}_a(\mathbf{x}) \neq \emptyset$ är en *regulär punkt* till problemet (1.12) om det finns minst en vektor $\mathbf{d} \in I\!\!R^n$ sådan att $\nabla g_i(\mathbf{x})\,\mathbf{d} < 0$ för alla $i \in \mathcal{I}_a(\mathbf{x})$.
En tillåten lösning $\mathbf{x} \in \mathcal{F}$ med $\mathcal{I}_a(\mathbf{x}) = \emptyset$ är alltid en regulär punkt till (1.12).

Anmärkning: Det följer från Def 1.2.2 ovan att ett tillräckligt (men inte nödvändigt) villkor för att $\mathbf{x} \in \mathcal{F}$ ska vara en regulär punkt är att gradienterna $\nabla g_i(\mathbf{x})$ för $i \in \mathcal{I}_a(\mathbf{x})$, dvs gradienterna svarande mot de aktiva bivillkoren i $\mathbf{x}$, är *linjärt oberoende.* Detta används i vissa böcker som definition på regulär punkt, men det är en onödigt inskränkt definition.

Lemma 1.2.4. Antag att $\hat{\mathbf{x}} \in \mathcal{F}$ med $\mathcal{I}_a(\hat{\mathbf{x}}) \neq \emptyset$ är både en regulär punkt och en lokal optimallösning till (1.12). Då finns det skalärer $y_i \geq 0$, för $i \in \mathcal{I}_a(\hat{\mathbf{x}})$,

sådana att $\nabla f(\hat{\mathbf{x}}) \ + \sum_{i \in \mathcal{I}_a(\hat{\mathbf{x}})} y_i \nabla g_i(\hat{\mathbf{x}}) \ = \ \mathbf{0}^\mathsf{T}$.

Bevis: Av Def 1.2.2 följer att v_0 i Lemma 1.2.3 inte kan vara $= 0$ när $\hat{\mathbf{x}}$ är en regulär punkt. Eftersom v_0 alltså är ett strikt positivt tal kan man dividera optimalitetsvillkoret $v_0 \nabla f(\hat{\mathbf{x}}) \ + \sum_{i \in \mathcal{I}_a(\hat{\mathbf{x}})} v_i \nabla g_i(\hat{\mathbf{x}}) \ = \ \mathbf{0}^\mathsf{T}$ med v_0 och kalla kvoterna v_i/v_0 för y_i.

Lemma 1.2.5. Antag att $\hat{\mathbf{x}} \in \mathcal{F}$ med $\mathcal{I}_a(\hat{\mathbf{x}}) \neq \emptyset$ är både en regulär punkt och en lokal optimallösning till (1.12). Då finns det ingen vektor $\mathbf{d} \in I\!R^n$ som uppfyller samtliga följande olikheter:

$\nabla f(\hat{\mathbf{x}})\, \mathbf{d} < 0$ och $\nabla g_i(\hat{\mathbf{x}})\, \mathbf{d} \leq 0$ för $i \in \mathcal{I}_a(\hat{\mathbf{x}})$.

Bevis: Kombinera Lemma 1.2.4 med Farkas lemma.

Följande sats utgör detta avsnitts höjdpunkt. De optimalitetsvillkor (1)–(4) som ingår brukar kallas *Karush-Kuhn-Tucker villkoren* eller kortare KKT–villkoren.

Sats 1.2.2. Antag att $\hat{\mathbf{x}} \in \mathcal{F}$ är både en regulär punkt och en lokal optimallösning till problemet (1.12). Då finns det en vektor $\hat{\mathbf{y}} \in I\!R^m$ som tillsammans med $\hat{\mathbf{x}}$ uppfyller samtliga följande villkor:

(1): $\nabla f(\hat{\mathbf{x}}) + \sum_{i=1}^{m} \hat{y}_i \nabla g_i(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}$,

(2): $g_i(\hat{\mathbf{x}}) \leq 0$, för $i = 1, \ldots, m$,

(3): $\hat{y}_i \geq 0$, för $i = 1, \ldots, m$,

(4): $\hat{y}_i \, g_i(\hat{\mathbf{x}}) = 0$, för $i = 1, \ldots, m$.

Bevis: Antag först att $\hat{\mathbf{x}} \in \mathcal{F}$ med $\mathcal{I}_a(\hat{\mathbf{x}}) \neq \emptyset$ är både en regulär punkt och en lokal optimallösning till problemet (1.12). Då gäller enligt Lemma 1.2.4 att det finns skalärer $y_i \geq 0$, för $i \in \mathcal{I}_a(\hat{\mathbf{x}})$, sådana att $\nabla f(\hat{\mathbf{x}}) + \sum_{i \in \mathcal{I}_a(\hat{\mathbf{x}})} y_i \nabla g_i(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}$.
Låt $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_m)^\mathsf{T}$, där $\hat{y}_i = y_i$ för $i \in \mathcal{I}_a(\hat{\mathbf{x}})$ och $\hat{y}_i = 0$ för $i \notin \mathcal{I}_a(\hat{\mathbf{x}})$.
Då uppfyller $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ villkoren (1)–(4).

Antag sedan att $\hat{\mathbf{x}} \in \mathcal{F}$ med $\mathcal{I}_a(\hat{\mathbf{x}}) = \emptyset$ är en lokal optimallösning till problemet (1.12). Då gäller enligt Sats 1.2.1 att $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}$.
Låt nu $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_m)^\mathsf{T} = (0, \ldots, 0)^\mathsf{T} = \mathbf{0}$. Då uppfyller $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ villkoren (1)–(4).

Lemma 1.2.6. I Sats 1.2.2 ovan kan villkoren att $\hat{y}_i \, g_i(\hat{\mathbf{x}}) = 0$ för $i = 1, \ldots, m$

bytas ut mot villkoret att $\sum_{i=1}^{m} \hat{y}_i \, g_i(\hat{\mathbf{x}}) = 0$.

Bevis: Antag först att $\hat{y}_i\, g_i(\hat{\mathbf{x}}) = 0$ för $i = 1, \ldots, m$. Då är förstås $\sum_{i=1}^{m} \hat{y}_i\, g_i(\hat{\mathbf{x}}) = 0$.

Antag omvänt att $\sum_{i=1}^{m} \hat{y}_i\, g_i(\hat{\mathbf{x}}) = 0$. Enligt KKT–villkoren (2) och (3) är $\hat{y}_i\, g_i(\hat{\mathbf{x}}) \leq 0$ för alla i, och en summa av icke-positiva termer är noll om och endast om varje term i summan är noll. Alltså är $\hat{y}_i\, g_i(\hat{\mathbf{x}}) = 0$ för $i = 1, \ldots, m$.

Låt $\mathbf{g}(\mathbf{x}) \in I\!R^m$ vara en kolonnvektor med komponenterna $g_1(\mathbf{x}), \ldots, g_m(\mathbf{x})$, och låt $\nabla \mathbf{g}(\mathbf{x})$ vara en $m \times n$–matris med raderna $\nabla g_1(\mathbf{x}), \ldots, \nabla g_m(\mathbf{x})$, dvs

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_m(\mathbf{x}) \end{pmatrix} \quad \text{och} \quad \nabla \mathbf{g}(\mathbf{x}) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial g_1}{\partial x_n}(\mathbf{x}) \\ \vdots & & \vdots \\ \frac{\partial g_m}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial g_m}{\partial x_n}(\mathbf{x}) \end{bmatrix}. \tag{1.14}$$

Då kan Sats 1.2.2 skrivas på följande mer kompakta form:

Sats 1.2.3. Antag att $\hat{\mathbf{x}} \in \mathcal{F}$ är både en regulär punkt och en lokal optimallösning till problemet (1.12). Då finns det en vektor $\hat{\mathbf{y}} \in I\!R^m$ som tillsammans med $\hat{\mathbf{x}}$ uppfyller samtliga följande villkor:

(1): $\nabla f(\hat{\mathbf{x}})^{\mathsf{T}} + \nabla \mathbf{g}(\hat{\mathbf{x}})^{\mathsf{T}} \hat{\mathbf{y}} = \mathbf{0}$,

(2): $\mathbf{g}(\hat{\mathbf{x}}) \leq \mathbf{0}$,

(3): $\hat{\mathbf{y}} \geq \mathbf{0}$,

(4): $\hat{\mathbf{y}}^{\mathsf{T}} \mathbf{g}(\hat{\mathbf{x}}) = 0$.

Bevis: Kombinera Sats 1.2.2 med Lemma 1.2.6.

Exempel 1.2.1.
Vid kvadratisk optimering under linjära olikhetsbivillkor har man problem på formen

$$\begin{aligned} \text{minimera} \quad & \tfrac{1}{2}\,\mathbf{x}^{\mathsf{T}}\mathbf{H}\mathbf{x} + \mathbf{c}^{\mathsf{T}}\mathbf{x} + c_0 \\ \text{då} \quad & \mathbf{A}\mathbf{x} \geq \mathbf{b}, \end{aligned} \tag{1.15}$$

där $\mathbf{A} \in I\!R^{m \times n}$ och $\mathbf{H} \in I\!R^{n \times n}$ är givna matriser med $\mathbf{H}$ symmetrisk.
Detta är ett specialfall av (1.12), med $f(\mathbf{x}) = \frac{1}{2}\,\mathbf{x}^{\mathsf{T}}\mathbf{H}\mathbf{x} + \mathbf{c}^{\mathsf{T}}\mathbf{x} + c_0$ och $\mathbf{g}(\mathbf{x}) = \mathbf{b} - \mathbf{A}\mathbf{x}$.
Här blir $\nabla f(\mathbf{x})^{\mathsf{T}} = \mathbf{H}\mathbf{x} + \mathbf{c}$ och $\nabla \mathbf{g}(\mathbf{x}) = -\mathbf{A}$,
varvid KKT–villkoren i Sats 1.2.3 övergår till följande system:

(1): $\mathbf{H}\hat{\mathbf{x}} + \mathbf{c} - \mathbf{A}^{\mathsf{T}}\hat{\mathbf{y}} = \mathbf{0}$,

(2): $\mathbf{A}\hat{\mathbf{x}} \geq \mathbf{b}$,

(3): $\hat{\mathbf{y}} \geq \mathbf{0}$,

(4): $\hat{\mathbf{y}}^{\mathsf{T}}(\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}) = 0$.

1.3 Optimalitetsvillkor för konvexa optimeringsproblem

I detta avsnitt ska vi behandla en synnerligen välartad klass av optimeringsproblem, nämligen så kallade *konvexa* problem. För dessa kan man härleda mycket starkare optimalitetsvillkor än för generella ickelinjära problem.

Betrakta först det generellt formulerade optimeringsproblemet

$$\begin{aligned} \text{minimera} \quad & f(\mathbf{x}) \\ \text{då} \quad & \mathbf{x} \in \mathcal{F}, \end{aligned} \tag{1.16}$$

där $\mathcal{F}$ ("tillåtna området") är en given delmängd av $\mathbb{R}^n$ och f ("målfunktionen") är en given reellvärd funktion på $\mathcal{F}$.

Def 1.3.1. Problemet (1.16) är ett *konvext optimeringsproblem* om $\mathcal{F}$ är en konvex mängd och f är en konvex funktion på $\mathcal{F}$.

För konvexa optimeringsproblem har man följande trevliga ekvivalenser

Lemma 1.3.1. Antag att (1.16) är ett konvext optimeringsproblem och att $\hat{\mathbf{x}}$ är en given tillåten lösning till (1.16), dvs $\hat{\mathbf{x}} \in \mathcal{F}$. Då är följande tre påståenden ekvivalenta, dvs om ett av dem är sant så är alla tre sanna:
(1) $\hat{\mathbf{x}}$ är en lokal optimallösning till (1.16).
(2) $\hat{\mathbf{x}}$ är en global optimallösning till (1.16).
(3) Det finns ingen tillåten avtaganderiktning $\mathbf{d}$ i $\hat{\mathbf{x}}$.

Bevis: Att (2) implicerar (1) följer direkt från definitionerna i avsnitt 1 av lokal respektive global optimallösning. Att (1) implicerar (3) följer från Lemma 1.1. Det återstår då bara att visa att (3) implicerar (2). Det gör vi genom att visa att om $\hat{\mathbf{x}}$ *inte* är en global optimallösning så *finns* det en tillåten avtaganderiktning $\mathbf{d}$ i $\hat{\mathbf{x}}$.

Antag alltså att $\hat{\mathbf{x}} \in \mathcal{F}$ inte är en global optimallösning.
Då finns det (minst) en punkt $\mathbf{x} \in \mathcal{F}$ sådan att $f(\mathbf{x}) < f(\hat{\mathbf{x}})$.
Vi ska visa att vektorn $\mathbf{d} = \mathbf{x} - \hat{\mathbf{x}}$ är en tillåten avtaganderiktning i $\hat{\mathbf{x}}$.
Låt $\mathbf{x}(t) = \hat{\mathbf{x}} + t\,\mathbf{d}$ för $t \in (0,1)$, där alltså $\mathbf{d} = \mathbf{x} - \hat{\mathbf{x}}$, dvs $\mathbf{x}(t) = \hat{\mathbf{x}} + t\,(\mathbf{x} - \hat{\mathbf{x}})$.
Att $\mathcal{F}$ är en konvex mängd medför att $\mathbf{x}(t) \in \mathcal{F}$ för alla $t \in (0,1)$,
vilket i sin tur medför att $\mathbf{d}$ är en tillåten riktning i $\hat{\mathbf{x}}$.
Att f är en konvex funktion medför att $f(\mathbf{x}(t)) = f(\hat{\mathbf{x}} + t\,(\mathbf{x} - \hat{\mathbf{x}})) \leq$
$\leq f(\hat{\mathbf{x}}) + t\,(f(\mathbf{x}) - f(\hat{\mathbf{x}})) < f(\hat{\mathbf{x}})$ för alla $t \in (0,1)$.
Därmed har vi visat att $\mathbf{d}$ är en tillåten avtaganderiktning i $\hat{\mathbf{x}}$.

Eftersom lokala och globala optimallösningar är samma sak för konvexa problem brukar man bara säga "optimallösning" eller "optimal lösning", utan tillägget "lokal" eller "global". Detta känner läsaren kanske igen från kursavsnitten om linjär optimering (där alla problem var konvexa även om detta inte alltid betonades) och konvex kvadratisk optimering.

Fortsättningsvis i detta avsnitt ska vi behandla problem på formen

$$\begin{aligned} \text{minimera} \quad & f(\mathbf{x}) \\ \text{då} \quad & g_i(\mathbf{x}) \le 0, \ \text{för } i = 1, \ldots, m, \end{aligned} \tag{1.17}$$

där såväl målfunktionen f som alla bivillkorsfunktionerna g_i antas vara *konvexa* funktioner. Dessutom antas f och alla g_i vara kontinuerligt deriverbara.

Tillåtna området $\mathcal{F}$ till detta problem (1.17) ges av

$$\mathcal{F} = \{\mathbf{x} \in I\!R^n \mid g_i(\mathbf{x}) \le 0, \ \text{för } i = 1, \ldots, m\} = \{\mathbf{x} \in I\!R^n \mid \mathbf{g}(\mathbf{x}) \le \mathbf{0}\}. \tag{1.18}$$

Följande lemma visar att (1.17) är ett konvext optimeringsproblem i enlighet med Def 1.3.1.

Lemma 1.3.2. Att $g_1, \ldots, g_m$ är konvexa funktioner på $I\!R^n$ medför att tillåtna området $\mathcal{F}$, definierat av (1.18), är en konvex mängd i $I\!R^n$.

Bevis: Övningsuppgift.

Följande viktiga sats visar att för konvexa problem på formen (1.17) är KKT–villkoren tillräckliga villkor för ett globalt optimum.

Sats 1.3.1. Antag att f och alla g_i i (1.17) är konvexa och kontinuerligt deriverbara. Antag vidare att vektorerna $\hat{\mathbf{x}} \in I\!R^n$ och $\hat{\mathbf{y}} \in I\!R^m$ tillsammans uppfyller KKT–villkoren

(1): $\nabla f(\hat{\mathbf{x}}) + \sum_{i=1}^{m} \hat{y}_i \nabla g_i(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}$,

(2): $g_i(\hat{\mathbf{x}}) \le 0$, för $i = 1, \ldots, m$,

(3): $\hat{y}_i \ge 0$, för $i = 1, \ldots, m$,

(4): $\hat{y}_i \, g_i(\hat{\mathbf{x}}) = 0$, för $i = 1, \ldots, m$.

Då är $\hat{\mathbf{x}}$ en (globalt) optimal lösning till problemet (1.17).

Bevis: Antag att $\hat{\mathbf{x}}$ och $\hat{\mathbf{y}}$ uppfyller (1)–(4) ovan.

Låt funktionen $\ell : I\!R^n \to I\!R$ definieras av $\ell(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^{m} \hat{y}_i \, g_i(\mathbf{x}) = f(\mathbf{x}) + \hat{\mathbf{y}}^\mathsf{T}\mathbf{g}(\mathbf{x})$.
Observera att $\hat{\mathbf{y}}$ här hålls fixt medan $\mathbf{x}$ är en variabelvektor.
Enligt (3) är alla $\hat{y}_i \ge 0$, vilket tillsammans med antagandena ovan om f och g_i medför att ℓ är en konvex och kontinuerligt deriverbar funktion (av $\mathbf{x}$) på $I\!R^n$.
Villkoret (1) ovan är ekvivalent med att $\nabla \ell(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T}$, vilket medför att $\hat{\mathbf{x}}$ är en global minpunkt till funktionen ℓ, dvs $\ell(\hat{\mathbf{x}}) \le \ell(\mathbf{x})$ för alla $\mathbf{x} \in I\!R^n$.
Villkoret (2) medför att $\hat{\mathbf{x}}$ är en tillåten lösning till problemet (1.17), dvs $\hat{\mathbf{x}} \in \mathcal{F}$.
Låt $\mathbf{x} \in \mathcal{F}$ vara en godtycklig tillåten lösning till (1.17), dvs $g_i(\mathbf{x}) \le 0$ för alla i.
Vi behöver visa att $f(\hat{\mathbf{x}}) \le f(\mathbf{x})$.

Men detta följer av att $f(\hat{\mathbf{x}}) = \ell(\hat{\mathbf{x}}) \le \ell(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^{m} \hat{y}_i \, g_i(\mathbf{x}) \le f(\mathbf{x})$,
där $f(\hat{\mathbf{x}}) = \ell(\hat{\mathbf{x}})$ följer av villkor (4), $\ell(\hat{\mathbf{x}}) \le \ell(\mathbf{x})$ gäller enligt ovan (pga (1)), och $\ell(\mathbf{x}) \le f(\mathbf{x})$ följer av att $\hat{y}_i \ge 0$ (pga (3)) och $g_i(\mathbf{x}) \le 0$ (pga att $\mathbf{x} \in \mathcal{F}$).

Enligt denna sats 1.3.1 är alltså KKT–villkoren *tillräckliga* optimalitetsvillkor för konvexa problem. Vi ska närmast visa att om det aktuella konvexa problemet är "regulärt", så är KKT–villkoren även *nödvändiga* optimalitetsvillkor.

Def 1.3.2. Det konvexa problemet (1.17) är *regulärt* om det finns en punkt $\mathbf{x}_0 \in I\!R^n$ sådan att $g_i(\mathbf{x}_0) < 0$ för alla $i = 1, \ldots, m$.

Lemma 1.3.3. Om (1.17) är ett regulärt konvext problem så är varje tillåten lösning $\mathbf{x} \in \mathcal{F}$ en regulär punkt i enlighet med Def 1.2.3 (och Def 1.2.2).

Bevis: Låt $\mathbf{x} \in \mathcal{F}$ vara en tillåten lösning till (1.17).
Antag först att $\mathcal{I}_a(\mathbf{x}) = \emptyset$. Då är $\mathbf{x}$ en regulär punkt enligt Def 1.2.3.
Antag nu att $\mathcal{I}_a(\mathbf{x}) \neq \emptyset$. Då ska vi visa att det finns en vektor $\mathbf{d} \in I\!R^n$ sådan att $\nabla g_i(\mathbf{x})\,\mathbf{d} < 0$ för alla $i \in \mathcal{I}_a(\mathbf{x})$.
Låt $\mathbf{d} = \mathbf{x}_0 - \mathbf{x}$, där $\mathbf{x}_0$ ges av Def 1.3.2.
För alla $i \in \{1, \ldots, m\}$ gäller då att
$0 > g_i(\mathbf{x}_0) \geq g_i(\mathbf{x}) + \nabla g_i(\mathbf{x})(\mathbf{x}_0 - \mathbf{x}) = g_i(\mathbf{x}) + \nabla g_i(\mathbf{x})\,\mathbf{d}$.
Speciellt för $i \in \mathcal{I}_a(\mathbf{x})$ är $g_i(\mathbf{x}) = 0$, vilket ger att $\nabla g_i(\mathbf{x})\,\mathbf{d} < 0$ för alla $i \in \mathcal{I}_a(\mathbf{x})$.

Sats 1.3.2. Antag att (1.17) är ett regulärt konvext problem och att $\hat{\mathbf{x}}$ är en optimal lösning till detta problem. Då finns det en vektor $\hat{\mathbf{y}} \in I\!R^m$ som tillsammans med $\hat{\mathbf{x}}$ uppfyller samtliga följande villkor:

$$(1):\quad \nabla f(\hat{\mathbf{x}}) + \sum_{i=1}^{m} \hat{y}_i \nabla g_i(\hat{\mathbf{x}}) = \mathbf{0}^\mathsf{T},$$

$$(2):\quad g_i(\hat{\mathbf{x}}) \leq 0, \quad \text{för } i = 1, \ldots, m,$$

$$(3):\quad \hat{y}_i \geq 0, \quad \text{för } i = 1, \ldots, m,$$

$$(4):\quad \hat{y}_i\, g_i(\hat{\mathbf{x}}) = 0, \quad \text{för } i = 1, \ldots, m.$$

Bevis: Kombinera Sats 1.2.2 med Lemma 1.3.3.

Sammanfattningsvis har vi därmed visat att om (1.17) är ett regulärt konvext problem så är $\hat{\mathbf{x}} \in I\!R^n$ en (globalt) optimal lösning till (1.17) *om och endast om* det finns en vektor $\hat{\mathbf{y}} \in I\!R^m$ som tillsammans med $\hat{\mathbf{x}}$ uppfyller KKT–villkoren (1)–(4).

Exempel 1.3.1. Betrakta följande problem i variabelvektorn $\mathbf{x} \in I\!R^n$:

$$\begin{aligned} \text{minimera} \quad & |\,\mathbf{x}-\mathbf{p}\,|^2 \\ \text{då} \quad & |\,\mathbf{x}-\mathbf{q}\,|^2 \le 1, \end{aligned}$$

där $\mathbf{p}$ och $\mathbf{q}$ är två givna vektorer i $I\!R^n$ som uppfyller $|\,\mathbf{p}\,|^2 = 1$, $|\,\mathbf{q}\,|^2 = 1$ och $\mathbf{p}^\mathsf{T}\mathbf{q} = 0$.
Låt $f(\mathbf{x}) = |\,\mathbf{x}-\mathbf{p}\,|^2 = (\mathbf{x}-\mathbf{p})^\mathsf{T}(\mathbf{x}-\mathbf{p}) = \mathbf{x}^\mathsf{T}\mathbf{x} - 2\,\mathbf{p}^\mathsf{T}\mathbf{x} + 1$,
och $g(\mathbf{x}) = |\,\mathbf{x}-\mathbf{q}\,|^2 - 1 = (\mathbf{x}-\mathbf{q})^\mathsf{T}(\mathbf{x}-\mathbf{q}) - 1 = \mathbf{x}^\mathsf{T}\mathbf{x} - 2\,\mathbf{q}^\mathsf{T}\mathbf{x}$.

Då kan vårt givna problem skrivas: minimera $f(\mathbf{x})$ då $g(\mathbf{x}) \le 0$.

Gradienterna ges av $\nabla f(\mathbf{x}) = 2\,\mathbf{x}^\mathsf{T} - 2\,\mathbf{p}^\mathsf{T}$ och $\nabla g(\mathbf{x}) = 2\,\mathbf{x}^\mathsf{T} - 2\,\mathbf{q}^\mathsf{T}$.
Vidare ges andraderivatsmatriserna av $\nabla^2 f(\mathbf{x}) = 2\,\mathbf{I}$ och $\nabla^2 g(\mathbf{x}) = 2\,\mathbf{I}$.
Eftersom båda dessa är positivt definita så är både f och g konvexa funktioner.
Om man (exempelvis) sätter $\mathbf{x}_0 = \mathbf{q}$ så gäller att $g(\mathbf{x}_0) < 0$.
Vårt optimeringsproblem är alltså ett regulärt konvext problem, vilket medför att KKT–villkoren är både nödvändiga och tillräckliga villkor för ett globalt optimum!
Eftersom problemet enbart har ett enda bivillkor så blir KKT–villkoren:

KKT-1: $\nabla f(\mathbf{x}) + y\nabla g(\mathbf{x}) = \mathbf{0}^\mathsf{T}$,
KKT-2: $g(\mathbf{x}) \le 0$,
KKT-3: $y \ge 0$,
KKT-4: $y\,g(\mathbf{x}) = 0$.

som i vårt speciella fall kan skrivas

KKT-1: $\mathbf{x} - \mathbf{p} + y\cdot(\mathbf{x}-\mathbf{q}) = \mathbf{0}$,
KKT-2: $\mathbf{x}^\mathsf{T}\mathbf{x} - 2\,\mathbf{q}^\mathsf{T}\mathbf{x} \le 0$,
KKT-3: $y \ge 0$,
KKT-4: $y\cdot(\mathbf{x}^\mathsf{T}\mathbf{x} - 2\,\mathbf{q}^\mathsf{T}\mathbf{x}) = 0$.

Vi ska undersöka två fall: $y = 0$ och $y > 0$. (Fallet $y < 0$ utesluts av KKT-3).
Antag först att $y = 0$.
Då ger KKT-1 att $\mathbf{x} = \mathbf{p}$, som insatt i vänsterledet i KKT-2 ger att
$\mathbf{x}^\mathsf{T}\mathbf{x} - 2\,\mathbf{q}^\mathsf{T}\mathbf{x} = \mathbf{p}^\mathsf{T}\mathbf{p} - 2\,\mathbf{q}^\mathsf{T}\mathbf{p} = 1 - 0 = 1$, vilket bryter mot olikheten i KKT-2.
Alltså finns det ingen lösning till KKT–villkoren med $y = 0$.

Antag fortsättningsvis att $y > 0$.
Då ger KKT-4 att $\mathbf{x}^\mathsf{T}\mathbf{x} - 2\,\mathbf{q}^\mathsf{T}\mathbf{x} = 0$, medan KKT-1 ger att $\mathbf{x} = \dfrac{\mathbf{p} + \mathbf{q}\,y}{1+y}$.
Tillsammans ger dessa relationer att
$0 = \mathbf{x}^\mathsf{T}\mathbf{x} - 2\,\mathbf{q}^\mathsf{T}\mathbf{x} = \dfrac{(\mathbf{p}+\mathbf{q}\,y)^\mathsf{T}(\mathbf{p}+\mathbf{q}\,y)}{(1+y)^2} - \dfrac{2\,\mathbf{q}^\mathsf{T}(\mathbf{p}+\mathbf{q}\,y)}{1+y} = \dfrac{1+y^2}{(1+y)^2} - \dfrac{2\,y}{1+y} = \dfrac{1-2y-y^2}{(1+y)^2}$.
Vi får alltså ekvationen $y^2 + 2y - 1 = 0$, med lösningarna $y = -1 + \sqrt{2}$ och $y = -1 - \sqrt{2}$.
Men den andra av dessa är negativ, så den enda återstående lösningen är $\hat{y} = \sqrt{2} - 1$.
Motsvarande $\mathbf{x}$ ges av $\hat{\mathbf{x}} = \dfrac{\mathbf{p}+\mathbf{q}\,\hat{y}}{1+\hat{y}} = \dfrac{1}{\sqrt{2}}\,(\mathbf{p} + \mathbf{q}(\sqrt{2}-1)) = \mathbf{q} + \dfrac{1}{\sqrt{2}}\cdot(\mathbf{p}-\mathbf{q})$.
Då uppfyller $\hat{\mathbf{x}}$ och $\hat{y}$ KKT–villkoren ovan. Eftersom såväl f som g är konvexa funktioner så betyder det att $\hat{\mathbf{x}}$ är en globalt optimal lösning till det betraktade problemet.

1.4 En Newtonmetod för optimeringsproblem med likhetsbivillkor

I detta avsnitt återvänder vi till problemet (1.5), dvs

$$\begin{aligned} \text{minimera} \quad & f(\mathbf{x}) \\ \text{då} \quad & h_i(\mathbf{x}) = 0, \ \text{för } i = 1, \ldots, m, \end{aligned} \tag{1.19}$$

där såväl bivillkorsfunktionerna h_i som målfunktionen f antas vara kontinuerligt deriverbara. Enligt Sats 1.1.1 måste varje regulär optimallösning till (1.19) uppfylla optimalitetsvillkoren

$$\begin{aligned} \nabla f(\mathbf{x})^{\mathsf{T}} \; + \; \nabla \mathbf{h}(\mathbf{x})^{\mathsf{T}} \mathbf{u} \; &= \; \mathbf{0}\,, \\ \mathbf{h}(\mathbf{x}) \; &= \; \mathbf{0}\,. \end{aligned} \tag{1.20}$$

Detta är ett ickelinjärt ekvationssystem, så en naturligt ide är att försöka använda Newton–Raphsons metod för att lösa (1.20). Låt därför

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{u} \end{pmatrix} \in I\!R^{n+m} \quad \text{och} \quad \mathbf{q}(\mathbf{z}) = \begin{pmatrix} \nabla f(\mathbf{x})^{\mathsf{T}} + \nabla \mathbf{h}(\mathbf{x})^{\mathsf{T}} \mathbf{u} \\ \mathbf{h}(\mathbf{x}) \end{pmatrix} \in I\!R^{n+m}. \tag{1.21}$$

Då kan optimalitetsvillkoren (1.20) kortfattat skrivas

$$\mathbf{q}(\mathbf{z}) = \mathbf{0}. \tag{1.22}$$

Newton–Raphsons metod innebär att man i den aktuella iterationspunkten $\mathbf{z}^{(k)}$ approximerar $\mathbf{q}(\mathbf{z})$ med första ordningens Taylorpolynom i $\mathbf{z}^{(k)}$, dvs

$$\mathbf{q}(\mathbf{z}) \approx \mathbf{q}(\mathbf{z}^{(k)}) + \nabla \mathbf{q}(\mathbf{z}^{(k)})(\mathbf{z} - \mathbf{z}^{(k)}). \tag{1.23}$$

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

$$\mathbf{q}(\mathbf{z}^{(k)} + \mathbf{p}) \approx \mathbf{q}(\mathbf{z}^{(k)}) + \nabla \mathbf{q}(\mathbf{z}^{(k)})\, \mathbf{p} \tag{1.24}$$

Sedan löser man ekvationssystemet $\mathbf{q}(\mathbf{z}^{(k)}) + \nabla \mathbf{q}(\mathbf{z}^{(k)})\, \mathbf{p} = \mathbf{0}$, som ekvivalent kan skrivas

$$\nabla \mathbf{q}(\mathbf{z}^{(k)})\, \mathbf{p} = -\mathbf{q}(\mathbf{z}^{(k)}). \tag{1.25}$$

Detta är ett linjärt ekvationssystem i vektorn $\mathbf{p} \in I\!R^{n+m}$.
Om $\mathbf{p}^{(k)}$ är en lösning till (1.25) så ges nästa iterationspunkt av

$$\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} + t_k \mathbf{p}^{(k)}, \tag{1.26}$$

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

$$|\, \mathbf{q}(\mathbf{z}^{(k)} + t_k \mathbf{p}^{(k)})\, |^2 < |\, \mathbf{q}(\mathbf{z}^{(k)})\, |^2. \tag{1.27}$$

Nu ska vi uttrycka och tolka ovanstående metod i $\mathbf{x}$, $\mathbf{u}$, f och $\mathbf{h}$ från (1.20).

Först konstateras att vektorerna $\mathbf{z}^{(k)}$ och $\mathbf{q}(\mathbf{z}^{(k)})$ ges av

$$\mathbf{z}^{(k)} = \begin{pmatrix} \mathbf{x}^{(k)} \\ \mathbf{u}^{(k)} \end{pmatrix} \quad \text{och} \quad \mathbf{q}(\mathbf{z}^{(k)}) = \begin{pmatrix} \nabla f^{(k)\mathsf{T}} + \nabla \mathbf{h}^{(k)\mathsf{T}} \mathbf{u}^{(k)} \\ \mathbf{h}^{(k)} \end{pmatrix}, \tag{1.28}$$

där $\mathbf{h}^{(k)} = \mathbf{h}(\mathbf{x}^{(k)})$, $\nabla f^{(k)} = \nabla f(\mathbf{x}^{(k)})$ och $\nabla \mathbf{h}^{(k)} = \nabla \mathbf{h}(\mathbf{x}^{(k)})$.

Efter en del deriveringar erhålls vidare att matrisen $\nabla \mathbf{q}(\mathbf{z}^{(k)})$ ges av

$$\nabla \mathbf{q}(\mathbf{z}^{(k)}) = \begin{bmatrix} \mathbf{L}^{(k)} & \nabla \mathbf{h}^{(k)\mathsf{T}} \\ \nabla \mathbf{h}^{(k)} & \mathbf{O} \end{bmatrix}, \tag{1.29}$$

med $\mathbf{L}^{(k)} = \mathbf{F}(\mathbf{x}^{(k)}) + \sum_{i=1}^{m} u_i^{(k)} \mathbf{H}_i(\mathbf{x}^{(k)})$ där $\mathbf{F}(\mathbf{x}^{(k)})$ och $\mathbf{H}_i(\mathbf{x}^{(k)})$ är Hessianen av f resp h_i i $\mathbf{x}^{(k)}$. Med $\mathbf{p} \in I\!R^{n+m}$ uppdelad i $\mathbf{d} \in I\!R^n$ och $\mathbf{w} \in I\!R^m$ kan systemet (1.25) då skrivas

$$\begin{bmatrix} \mathbf{L}^{(k)} & \nabla \mathbf{h}^{(k)\mathsf{T}} \\ \nabla \mathbf{h}^{(k)} & \mathbf{O} \end{bmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} -\nabla f^{(k)\mathsf{T}} - \nabla \mathbf{h}^{(k)\mathsf{T}} \mathbf{u}^{(k)} \\ -\mathbf{h}^{(k)} \end{pmatrix}. \tag{1.30}$$

Om man först adderar den konstanta vektorn $\begin{pmatrix} \nabla \mathbf{h}^{(k)\mathsf{T}} \mathbf{u}^{(k)} \\ \mathbf{0} \end{pmatrix}$ till bägge leden i (1.30), därefter inför vektorn $\mathbf{v} = \mathbf{u}^{(k)} + \mathbf{w} \in I\!R^m$, samt slutligen multiplicerar de m st nedersta ekvationerna i (1.30) med -1, så ehålls följande ekvivalenta system:

$$\begin{bmatrix} \mathbf{L}^{(k)} & \nabla \mathbf{h}^{(k)\mathsf{T}} \\ -\nabla \mathbf{h}^{(k)} & \mathbf{O} \end{bmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} -\nabla f^{(k)\mathsf{T}} \\ \mathbf{h}^{(k)} \end{pmatrix}. \tag{1.31}$$

Om $(\mathbf{d}^{(k)}, \mathbf{v}^{(k)})$ är en lösning till (1.31) så ges nästa iterationspunkt $(\mathbf{x}^{(k+1)}, \mathbf{u}^{(k+1)})$ av

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t_k \mathbf{d}^{(k)} \quad \text{och} \quad \mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + t_k \mathbf{w}^{(k)}, \tag{1.32}$$

där $\mathbf{w}^{(k)} = \mathbf{v}^{(k)} - \mathbf{u}^{(k)}$ och t_k väljs exempelvis som det största av talen $1, \frac{1}{2}, \frac{1}{4}, \ldots$ för vilket

$$\begin{aligned} &|\, \nabla f(\mathbf{x}^{(k)} + t_k \mathbf{d}^{(k)})^{\mathsf{T}} + \nabla \mathbf{h}(\mathbf{x}^{(k)} + t_k \mathbf{d}^{(k)})^{\mathsf{T}} (\mathbf{u}^{(k)} + t_k \mathbf{w}^{(k)}) \,|^2 + |\, \mathbf{h}(\mathbf{x}^{(k)} + t_k \mathbf{d}^{(k)}) \,|^2 \\ &\quad < \; |\, \nabla f(\mathbf{x}^{(k)})^{\mathsf{T}} + \nabla \mathbf{h}(\mathbf{x}^{(k)})^{\mathsf{T}} \mathbf{u}^{(k)} \,|^2 + |\, \mathbf{h}(\mathbf{x}^{(k)}) \,|^2. \end{aligned} \tag{1.33}$$

Om man startar "tillräckligt nära" en lokalt optimal lösning till problemet (1.19) så är denna Newton-metod ofta mycket effektiv. Men i det allmänna fallet är det inte säkert att sekvensen av iterationspunkter konvergerar mot en lösning till systemet (1.20), även om det finns en sådan lösning. Vidare kan (1.20) vara uppfyllt även i ett lokalt *maximum* till problemet, och om man råkar starta nära en sådan lokalt maximum så kan det hända att metoden konvergerar mot detta!

I nästa avsnitt beskrivs en modifiering av metoden som gör den mer pålitlig.

1.5 En SQP-metod för optimeringsproblem med likhetsbivillkor

En intressant och viktig observation är att ekvationssystemet (1.31) utgör optimalitetsvillkor till ett visst kvadratiskt optimeringsproblem med linjära likhetsbivillkor. Genom att i (1.9) och (1.10) sätta $\mathbf{H} = \mathbf{L}^{(k)}$, $\mathbf{A} = -\nabla\mathbf{h}^{(k)}$, $\mathbf{c} = \nabla f^{(k)\mathsf{T}}$, $c_0 = f^{(k)} = f(\mathbf{x}^{(k)})$ och $\mathbf{b} = \mathbf{h}^{(k)}$, så ser vi att (1.31) utgör optimalitetsvillkor till följande QP-problem (QP = Quadratic Programming) i variabelvektorn $\mathbf{d} \in I\!R^n$:

$$\begin{aligned} &\text{minimera} \quad f^{(k)} + \nabla f^{(k)}\mathbf{d} + \tfrac{1}{2}\,\mathbf{d}^{\mathsf{T}}\mathbf{L}^{(k)}\mathbf{d} \\ &\qquad\text{då} \quad -\nabla\mathbf{h}^{(k)}\mathbf{d} = \mathbf{h}^{(k)} \quad (\text{dvs} \;\; \mathbf{h}^{(k)} + \nabla\mathbf{h}^{(k)}\mathbf{d} = \mathbf{0}). \end{aligned} \tag{1.34}$$

Man kan tolka detta QP-problem som att man gör naturliga linjära approximationer av bivillkoren, medan målfunktionen approximeras med en kvadratisk funktion som även innehåller viktade andragradstermer från bivillkoren.

Vi vet från avsnittet om kvadratisk optimering att en förutsättning för att $\mathbf{d}^{(k)}$ erhållen från systemet (1.31) verkligen ska vara en optimal lösning till QP-problemet (1.34) är att matrisen $\mathbf{L}^{(k)}$ är positivt semidefinit på nollrummet till matrisen $\nabla\mathbf{h}^{(k)}$ (dvs $\mathcal{N}(\nabla\mathbf{h}^{(k)})$). För att $\mathbf{d}^{(k)}$ ska vara en *unik* optimal lösning till (1.34) så måste $\mathbf{L}^{(k)}$ vara positivt definit på $\mathcal{N}(\nabla\mathbf{h}^{(k)})$. Man kan visa att detta i normalfallet är uppfyllt "nära" en lokalt optimal lösning till ursprungsproblemet (1.19), men i allmänhet kan man inte räkna med att detta krav är uppfyllt.

Om det i en given iterationpunkt gäller att $\mathbf{L}^{(k)}$ *inte* är positivt definit på $\mathcal{N}(\nabla\mathbf{h}^{(k)})$ så modifierar man därför $\mathbf{L}^{(k)}$ så att den *blir* positivt definit på $\mathcal{N}(\nabla\mathbf{h}^{(k)})$, innan man löser QP-problemet (1.34), dvs innan man löser ekvationssystemet (1.31). De tekniska detaljerna för denna modifiering av $\mathbf{L}^{(k)}$ avstår vi från här.

Om $(\mathbf{d}^{(k)}, \mathbf{v}^{(k)})$ är en lösning till det modifierade systemet (1.31) så ges nästa iterationspunkt $(\mathbf{x}^{(k+1)}, \mathbf{u}^{(k+1)})$ av

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t_k\mathbf{d}^{(k)} \quad \text{och} \quad \mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + t_k\mathbf{w}^{(k)}, \tag{1.35}$$

där $\mathbf{w}^{(k)} = \mathbf{v}^{(k)} - \mathbf{u}^{(k)}$ och t_k väljs exempelvis som det största av talen $1, \frac{1}{2}, \frac{1}{4}, \ldots$ för vilket

$$f(\mathbf{x}^{(k)} + t_k\mathbf{d}^{(k)}) + \rho\cdot\sum_{i=1}^{m} |\, h_i(\mathbf{x}^{(k)} + t_k\mathbf{d}^{(k)})\,| < f(\mathbf{x}^{(k)}) + \rho\cdot\sum_{i=1}^{m} |\, h_i(\mathbf{x}^{(k)})\,|, \tag{1.36}$$

där ρ är ett "stort" tal.

Man kan visa att om $\rho > \max_i\{\,|\,v_i^{(k)}|\,\}$ så utgör $\mathbf{d}^{(k)}$ (bestämd enligt ovan) en avtagande-riktning till funktionen $f(\mathbf{x}) + \rho\cdot\sum_{i=1}^{m} |\, h_i(\mathbf{x})\,|$ i punkten $\mathbf{x}^{(k)}$.

Eftersom man löser en sekvens av kvadratiska optimeringsproblem så brukar metoden som beskrivits i detta avsnitt kallas sekvensiell kvadratisk optimering, på engelska Sequential Quadratic Programming med den vedertagna förkortningen SQP.

1.6 En SQP-metod för optimeringsproblem med olikhetsbivillkor

I detta avsnitt betraktar vi problemet (1.12), dvs

$$\begin{aligned} \text{minimera} \quad & f(\mathbf{x}) \\ \text{då} \quad & g_i(\mathbf{x}) \leq 0, \ \text{för } i = 1, \ldots, m, \end{aligned} \tag{1.37}$$

där såväl bivillkorsfunktionerna g_i som målfunktionen f antas vara kontinuerligt deriverbara. Enligt Sats 1.2.3 måste varje regulär optimallösning till (1.37) uppfylla optimalitetsvillkoren

$$\begin{aligned} &(1): \quad \nabla f(\hat{\mathbf{x}})^{\mathsf{T}} + \nabla \mathbf{g}(\hat{\mathbf{x}})^{\mathsf{T}} \hat{\mathbf{y}} = \mathbf{0}, \\ &(2): \quad \mathbf{g}(\hat{\mathbf{x}}) \leq \mathbf{0}, \\ &(3): \quad \hat{\mathbf{y}} \geq \mathbf{0}, \\ &(4): \quad \hat{\mathbf{y}}^{\mathsf{T}} \mathbf{g}(\hat{\mathbf{x}}) = 0. \end{aligned}$$

Vi ska här presentera en lösningsmetod för problem på formen (1.37). Metoden är iterativ, så det räcker att beskriva hur man utgående från den aktuella iterationspunkten $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$, som alltid uppfyller $\mathbf{y}^{(k)} \geq \mathbf{0}$, genererar nästa iterationspunkt $(\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)})$.

Eftersom den "enda" skillnaden mellan problemen (1.37) och (1.19) är att likhetsbivillkoren $h_i(\mathbf{x}) = 0$ i (1.19) ersatts med olikhetsbivillkor $g_i(\mathbf{x}) \leq 0$ i (1.37), är det naturligt att i problemet (1.34) ersätta de linjäriserade likhetsbivillkoren $h_i(\mathbf{x}^{(k)}) + \nabla h_i(\mathbf{x}^{(k)})\mathbf{d} = 0$ med linjäriserade olikhetsbivillkor $g_i(\mathbf{x}^{(k)}) + \nabla g_i(\mathbf{x}^{(k)})\mathbf{d} \leq 0$. Problemet (1.34) ersätts då med följande problem:

$$\begin{aligned} \text{minimera} \quad & f^{(k)} + \nabla f^{(k)}\mathbf{d} + \tfrac{1}{2}\,\mathbf{d}^{\mathsf{T}}\mathbf{L}^{(k)}\mathbf{d} \\ \text{då} \quad & -\nabla \mathbf{g}^{(k)}\mathbf{d} \geq \mathbf{g}^{(k)} \quad (\text{dvs } \ \mathbf{g}^{(k)} + \nabla \mathbf{g}^{(k)}\mathbf{d} \leq \mathbf{0}). \end{aligned} \tag{1.38}$$

Här är $f^{(k)} = f(\mathbf{x}^{(k)})$, $\mathbf{g}^{(k)} = \mathbf{g}(\mathbf{x}^{(k)})$, $\nabla f^{(k)} = \nabla f(\mathbf{x}^{(k)})$, $\nabla \mathbf{g}^{(k)} = \nabla \mathbf{g}(\mathbf{x}^{(k)})$ och
$\mathbf{L}^{(k)} = \mathbf{F}(\mathbf{x}^{(k)}) + \displaystyle\sum_{i=1}^{m} y_i^{(k)} \mathbf{G}_i(\mathbf{x}^{(k)})$, där $\mathbf{F}(\mathbf{x}^{(k)})$ och $\mathbf{G}_i(\mathbf{x}^{(k)})$ är Hessianen av f resp g_i i $\mathbf{x}^{(k)}$.

Antag tills vidare att matrisen $\mathbf{L}^{(k)}$ är positivt definit, vilket den är om exempelvis alla bivillkorsfunktioner g_i är konvexa och målfunktionen f är (strikt) konvex med positivt definit hessian. Då kan problemet (1.38) lösas med exempelvis en så kallad "active set"–metod.

Låt $\mathbf{d}^{(k)}$ vara den optimala lösningen till (1.38) och låt $\mathbf{v}^{(k)}$ vara motsvarande vektor med "lagrangemultiplikatorer" som tillsammans med $\mathbf{d}^{(k)}$ uppfyller optimalitetsvillkoren till (1.38), dvs

$$\begin{aligned} &(1): \quad \mathbf{L}^{(k)}\mathbf{d}^{(k)} + \nabla f^{(k)\mathsf{T}} + \nabla \mathbf{g}^{(k)\mathsf{T}}\mathbf{v}^{(k)} = \mathbf{0}, \\ &(2): \quad \mathbf{g}^{(k)} + \nabla \mathbf{g}^{(k)}\mathbf{d}^{(k)} \leq \mathbf{0}, \\ &(3): \quad \mathbf{v}^{(k)} \geq \mathbf{0}, \\ &(4): \quad \mathbf{v}^{(k)\mathsf{T}}(\mathbf{g}^{(k)} + \nabla \mathbf{g}^{(k)}\mathbf{d}^{(k)}) = 0. \end{aligned}$$

Då ges nästa iterationspunkt $(\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)})$ i metoden av

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t_k \mathbf{d}^{(k)} \quad \text{och} \quad \mathbf{y}^{(k+1)} = \mathbf{y}^{(k)} + t_k \mathbf{w}^{(k)}, \tag{1.39}$$

där $\mathbf{w}^{(k)} = \mathbf{v}^{(k)} - \mathbf{y}^{(k)}$ och t_k väljs exempelvis som det största av talen $1, \frac{1}{2}, \frac{1}{4}, \ldots$ för vilket

$$f(\mathbf{x}^{(k)} + t_k \mathbf{d}^{(k)}) + \rho \cdot \sum_{i=1}^{m} g_i(\mathbf{x}^{(k)} + t_k \mathbf{d}^{(k)})_+ \; < \; f(\mathbf{x}^{(k)}) + \rho \cdot \sum_{i=1}^{m} g_i(\mathbf{x}^{(k)})_+ \tag{1.40}$$

där $g_i(\mathbf{x})_+$ betecknar det största av de bägge talen 0 och $g_i(\mathbf{x})$, medan ρ är ett "stort" tal.

Man kan visa att om $\rho > \max_i \{ v_i^{(k)} \}$ så utgör $\mathbf{d}^{(k)}$ (bestämd enligt ovan) en avtagande-riktning till funktionen $f(\mathbf{x}) + \rho \cdot \sum_{i=1}^{m} g_i(\mathbf{x})_+$ i punkten $\mathbf{x}^{(k)}$.

Metoden ovan bygger på förutsättningen att $\mathbf{L}^{(k)}$ är positivt definit. Om det i en given iteration skulle visa sig att $\mathbf{L}^{(k)}$ inte är positivt definit så måste man därför modifiera $\mathbf{L}^{(k)}$ så att den *blir* positivt definit (exempelvis genom att addera en lämplig multipel av enhetsmatrisen till $\mathbf{L}^{(k)}$). Vi avstår från vidare detaljer om hur denna modifiering av $\mathbf{L}^{(k)}$ görs i praktiken.

Även metoden som beskrivits i detta avsnitt kallas sekvensiell kvadratisk optimering, på engelska Sequential Quadratic Programming med den vedertagna förkortningen SQP, eftersom den bygger på att man löser en sekvens av kvadratiska optimeringsproblem.