Technische Universität Berlin

Lecture Notes

Nonlinear Optimisation

Prof. Dr. Dietmar Hömberg, Mathieu Rosière,

Summersemester 2020

Contents

Mitschrift von Viktor Glombik.

Zuletzt am April 12, 2022 geändert.

List of Figures

1 Optimierungsaufgaben

20.04.2020

Wir untersuchen in diesem Kurs die Aufgabe, ein Minimum einer gegebenen Funktion $f\colon \mathbb{R}^n \supset D \to \mathbb{R}$ zu berechnen.

Remark 1.0.1 Alle Lösungen y von $\arg\max_{x \in X} f(x)$ erfüllen $f(y) \geqslant f(x)$ für alle $x \in X$. Die Ungleichung ist äquivalent zu $-f(y) \leqslant -f(x)$ also $g(y) \leqslant g(x)$ für alle $x \in X$ mit $g(x) := -f(x)$. Deshalb können wir uns ohne Beschränkung der Allgemeinheit auf Minima beschränken.

1.1 Grundlagen

Im Weiteren sei $D \subset \mathbb{R}^n$ eine offene Menge, $\Omega \subset D$ und $f\colon D \to \mathbb{R}$. Wir betrachten das Optimierungsproblem

$$\min_{x \in \Omega} f(x) \qquad \text{(P)}$$

DEFINITION 1.1.1 (GRUNDBEGRIFFE)
Die Funktion f heißt Zielfunktion und Ω der zulässige Bereich.

Ist $\Omega = D$, so ist (P) unrestringiert bzw. frei. Ist Ω durch Nebenbedingungen gegebene, heißt (P) restringiert. Die Elemente von Ω sind zulässige Punkte.

Example 1.1.2 (Grundbegriffe)
Bei $\min_{x \geqslant 1} x^3$ handelt es sich um ein Problem mit lineare Ungleichungsrestriktion für $D := \mathbb{R}$, $\Omega := [1, \infty)$. Die Lösung ist $\hat{x} = 1$. ◇

DEFINITION 1.1.3 ((STRENGE) LOKALE / GLOBALE MINIMA)
Ein Punkte $x \in \Omega$ heißt

- lokales Minimum von f auf Ω oder lokale Lösung von (P), wenn ein $r > 0$ existiert, sodass $f(y) \geqslant f(x)$ für alle $y \in \Omega \cap B(x, r)$ gilt. Für ein strenges lokales Minimum ersetzt man $B(x, r)$ durch $B(x, r) \backslash \{x\}$ und $\geqslant$ durch $>$. lokales Minimum
- globales Minimum oder globale Lösung, wenn $f(y) \geqslant f(x)$ für alle $x \in \Omega$ gilt. Ein strenges globales Minimum ist analog zu einem lokalen definiert. globales Minimum

Example 1.1.4 (Lokale und globale Minima)
Die Funktion $x \cdot \sin\left(\frac{1}{x}\right)$ für $x \neq 0$ und 0 sonst, besitzt abzählbar viele lokale Minima, jedoch kein globales Optimum. ◇

Example 1.1.5 (Lineare Regression I)
Gesucht ist eine lineare Funktion $y(x) := g(x, a, b) := ax + b$ mit unbekannten Koeffizienten $a, b \in \mathbb{R}$, welche am besten zu gegebenen Messwerten

$(x_k, y_k)_{k=1}^m$ passen. Dazu minimieren wir die quadratische Zielfunktion

$$f(a,b) := \sum_{k=1}^m (y_k - g(x,a,b))^2 = \sum_{k=1}^m (y_k - a x_k - b)^2. \tag{1}$$

◇

Anhand diesen Beispiels lassen sich die zentralen Fragestellungen dieser Vorlesung erklären:

- Existenz und Eindeutigkeit von Lösungen,
- Notwendige Optimalitätsbedingungen,
- Hinreichende Optimalitätsbedingungen,
- Numerische Lösungen / Lösungsverfahren.

Example 1.1.6 (Machine Learning: Support vector machine) The goal is to learn a classification function g from labelled training data $((x_i, y_i))_{i=1}^m$, where $x_i \in \mathbb{R}^n$ is a feature vector and $y_i \in \{\pm 1\}$ are the labels. When, for instance, comparing cats and dogs, one could obtain features such as snout length (x_1) and ear length (x_2) from them. We can thus identify a cat with the vector (x_1, x_2) and the dog similarly. We now want to learn a decision boundary (hyperplane) that separates the vectors of the dogs and the vectors of the cats.

The widest-street-approach aims to separate both classes by two lines and considers the line between as the decision boundary. We want to maximise the distance d between the two thick red lines in figure Figure 1.

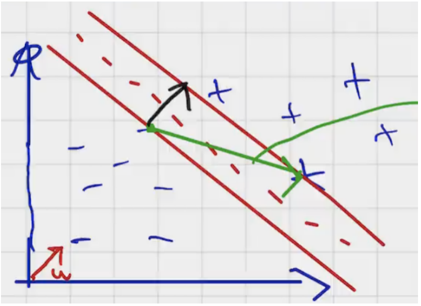


Figure 1: The pluses and minuses are called support vectors and the green vector is $x^+ - x^-$.

The decision rule w is orthogonal to the "centre line of the street". For all pluses, we demand $\langle w, x \rangle + b \geqslant 0$. We have to find w and b.

For the test samples, we demand

$$\langle w, x^+ \rangle + b \geqslant 1 \quad \text{and} \quad \langle w, x^- \rangle + b \leqslant -1$$

(without loss generality, we choose 1) and $= 1$ for support vectors. We can combine those inequalities by introducing

$$y_i = \begin{cases} 1, & \text{for } +, \\ -1, & \text{for } -. \end{cases}$$

which yields

$$y_i(\langle w, x_i \rangle + b) \geqslant 1 \quad \forall i \in \{1, \dots, m\}.$$

To find d, we project $x^+ - x^-$ onto w, which yields

$$\begin{aligned} d &= \left\langle x^+ - x^-, \frac{w}{|w|} \right\rangle = \frac{1}{|w|} \left(\langle x^+, w \rangle - \langle x^-, w \rangle \right) \\ &= \frac{1}{|w|} \left((1 \cancel{-b}) - (-1 \cancel{-b}) \right) = \frac{2}{|w|}. \end{aligned}$$

Our (quadratic) optimisation problem (with linear inequality constraints), called linear support vector machine, is **(TODO: WHAT IS μ?!)**

$$\begin{cases} \min\limits_{w,\xi,b} \frac{1}{2}|w|^2 + \mu \sum_{i=1}^m \xi_i, \\ \text{subject to} \quad y_i(\langle w, x_i \rangle + b) \geqslant 1 - \xi_i \quad \text{and} \quad \xi_i \geqslant 0 \ \forall i \in \{1, \dots, m\}, \end{cases}$$

where the ξ_i are slack variables. We have

$$\Omega := \{(w,\xi,b) \in \mathbb{R}^m \times \mathbb{R}^m_{\geqslant 0} \times \mathbb{R} : y_i(\langle w, x_i \rangle + b) \geqslant 1 - \xi_i,\ \forall i \in \{1,\dots,m\}\}.$$

For deciding if an unseen image depicts a cat or a dog, we use the classification function

$$g(x) := \operatorname{sign}(\langle w, x \rangle + b). \qquad \diamond$$

1.2 Existenz von Lösungen

23.04.2020

Sei $f\colon \mathbb{R}^n \supset D \to \mathbb{R}$ eine stetige Funktion.

Definition 1.2.1 (Niveaumengen)
Für $a \in \mathbb{R}$ ist $\mathcal{N}(f,a) := \{x \in D : f(x) \leqslant a\}$ eine Niveaumenge von f.

Niveaumenge

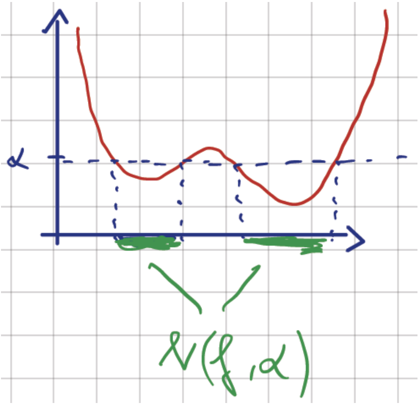


Figure 2: A level set of a function.

Theorem 1.2.1: Kompakte Niveaumengen und Minima

Sei Ω abgeschlossen. Existiert ein $\omega \in \Omega$, sodass $\mathcal{N}(f, f(\omega))$ kompakt ist, existiert ein globales Minimum von f auf Ω.

Proof. Sei $a := \inf_{x\in\Omega} f(x) \leqslant f(\omega)$. Da Ω abgeschlossen ist, ist $N := \Omega \cap \mathcal{N}(f, f(\omega))$ kompakt und es gilt $a = \inf_{x\in N} f(x)$. Nach dem Satz von Weierstrass existiert ein $\hat{x} \in \Omega$ mit $\inf_{x\in N} f(x) = f(\hat{x})$. □

Corollary 1.2.2 (Koerzivivät)
Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ stetig und koerzitiv, das heißt $\lim_{\|x\|\to\infty} f(x) = \infty$. Dann existiert ein globales Minimum auf Ω.

koerzitiv

Proof. Sei $\omega \in \mathbb{R}$. Weil f koerzitiv ist, existiert ein $M > 0$, sodass $f(x) \geqslant f(\omega)$ für alle $x \in \mathbb{R}^n \setminus B_M(0)$ gilt. Somit ist $\mathcal{N}(f, f(\omega)) \subset B_M(0)$ beschränkt.

Da f stetig ist, sind alle Niveaumengen $\mathcal{N}(f, f(\omega)) = f^{-1}((-\infty, f(\omega)])$ abgeschlossen sind und somit kompakt. Die Aussage folgt aus Satz 1.2.1. □

Lemma 1.2.3
Ist f stetig und koerzitiv sowie Ω abgeschlossen, so existiert ein globales Minimum von f.

Proof. Sei $(x_k)_{k\in\mathbb{N}} \subset \Omega$ eine Minimalfolge, d.h. $f(x_k) \to \inf_{x\in\Omega} f(x)$.

① Wäre $(x_k)_{k\in\mathbb{N}}$ unbeschränkt, so gäbe es eine Teilfolge $(x_{k_j})_{j\in\mathbb{N}}$ mit $\|x_{k_j}\| \to \infty$. Wegen der Koerzivität müsste dann $f(x_{k_j}) \to \infty$ gelten, was eine Widerspruch dazu darstellt, dass $(x_k)_{k\in\mathbb{N}}$ eine Minimalfolge ist. Somit ist $(x_k)_{k\in\mathbb{N}}$ beschränkt.

② Da $(x_k)_{k\in\mathbb{N}}$ beschränkt ist, existiert eine konvergente Teilfolge $(x_{k_j})_{j\in\mathbb{N}}$, welche gegen x konvergiert. Da Ω abgeschlossen ist, gilt $x \in \Omega$.

③ Wegen der Stetigkeit von f gilt

$$f(x) = \lim_{j\to\infty} f(x_{k_j}) = \lim_{k\to\infty} f(x_k) = \inf_{x\in\Omega} f(x). \qquad \square$$

Example 1.2.4 Die Funktion

$$f\colon \mathbb{R}^2 \to \mathbb{R},\ e^{(x-y)^2}(x^4 + y^4 - \sin(y^3))$$

besitzt ein globales Minimum, da sie stetig und koerzitiv ist. ◇

Proof. Für $x, y \in \mathbb{R}$ gilt $(x-y)^2 \geqslant 0$ und $\sin(y^3) \in [-1, 1]$ $(\star)$, also $e^{(x-y)^2} \geqslant 1$. Somit folgt

$$f(x,y) \geqslant 1\cdot(x^4+y^4-\sin(y^3)) \overset{(\star)}{\geqslant} x^4+y^4-1 = \|(x,y)\|_4^4-1 \xrightarrow{\|(x,y)\|_4\to\infty} \infty,$$

wobei $\|(x,y)\|_4 := (x^4+y^4)^{1/4}$ die Minkowski-4-Norm ist. Da auf $\mathbb{R}^n$ alle Normen äquivalent sind, folgt die Aussage. $\square$

Alternative: Es gilt $x^4 + y^4 - 1 \geqslant x^2 + y^2 - 2$ (da $x^4 + y^4 - 1 - (x^2 + y^2 - 2) = (x^2 - 1/2)^2 + (y^2 - 1/2)^2 + 1/2 \geqslant 0$), wobei $\sqrt{x^2 + y^2}$ die "‘Standard"’ euklidische Norm auf $\mathbb{R}^2$ ist.

Remark 1.2.5 In Korollar 1.2.2 können wir die Stetigkeit von f durch die schwächere Voraussetzung ersetzen, dass der Epigraph $\operatorname{epi}(f) := \{(x,a) \in \mathbb{R}^{n+1} : f(x) \leqslant a\}$ von f abgeschlossen ist.

Epigraph

Proof. (Skizze) Ist $\operatorname{epi}(f)$ abschlossen, so ist f unterhalbstetig und das genügt für den Beweis von Satz 1.2.1. $\square$

Lemma 1.2.6 (Positive Definitheit)
Eine Matrix $A \in \mathbb{R}^{n\times n}$ ist genau dann positiv definit, wenn ein $C > 0$ existiert, sodass $x^\mathsf{T} Hx \geqslant C\|x\|^2$ für alle $x \in \mathbb{R}^n$ gilt.

Proof. "‘ $\Longleftarrow$ "’ is clear.

"‘ $\Longrightarrow$ "’: We can assume that A is symmetric as $2\langle Ax, x\rangle = \langle (A + A^\mathsf{T})x, x\rangle$ holds for all $A \in \mathbb{R}^{n\times n}$ and $x \in \mathbb{R}^n$ as $\langle Ax, x\rangle = \langle x, A^\mathsf{T} x\rangle$ holds and the scalar product on $\mathbb{R}^n$ is symmetric (and $A + A^\mathsf{T}$ is symmetric). Thus A is positive definite if and only $A + A^T$ is positive definite.

For $x \in \mathbb{R}^n$ define the Rayleigh quotient $\mathcal{R}_A(x) := \frac{\langle Ax,x\rangle}{\langle x,x\rangle}$. For $v_{\min}$, the eigenvector of the smallest eigenvalue $\lambda_{\min}$, $\mathcal{R}_A(v_{\min}) = \frac{\langle \lambda_{\min} v_{\min}, v_{\min}\rangle}{\langle v_{\min}, v_{\min}\rangle} = \lambda_{\min}$ holds. By the theorem of Courant-Fischer we know that $\lambda_{\min}$ is the minimum of $\mathcal{R}_A$. As $\mathcal{R}_A(cx) = \mathcal{R}_A(x)$ for all $c \neq 0$, it suffices restrict ourselves to the case $\|x\| = 1$. Then $\mathcal{R}_A(x) \geqslant \lambda_{\min}$ is equivalent to $\langle Ax, x\rangle \geqslant \lambda_{\min}\|x\|^2$. As A is symmetric and positive definite as argued above, $\lambda_{\min} > 0$. $\square$

Example 1.2.7 (Unrestringierte quadratische Aufgabe)
Für positiv definites $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n$ ist

$$\min_{x\in\mathbb{R}^n} \frac{1}{2} x^\mathsf{T} Ax + b^\mathsf{T} x \tag{QU}$$

ein quadratische Optimierungsproblem. Die Zielfunktion ist koerzitiv und besitzt somit nach Korollar 1.2.2 ein globales Minimum. ◇

Proof. Nach Lemma 1.2.6 existiert ein $C > 0$, sodass $\frac{1}{2}\langle Ax, x\rangle \geqslant C\|x\|^2$ für alle $x \in \mathbb{R}^n$ gilt. Mit der Cauchy-Schwartz-Ungleichung (CS) folgt $\langle b, x\rangle \geqslant -\|b\|\|x\|$ und somit $f(x) \geqslant (C\|x\| - \|b\|)\|x\| \xrightarrow{\|x\|\to\infty} \infty$, da $C\|x\| - \|b\| \geqslant 0$ für $\|x\| \geqslant \frac{\|b\|}{C}$ gilt. □

Example 1.2.8 (Lineare Regression II)
Ausmultiplizieren der Zielfunktion (1) ergibt

$$f(a,b) = \sum_{k=1}^m y_k^2 - 2\sum_{k=1}^m y_k(ax_k + b) + \sum_{k=1}^m (ax_k + b)^2 = \frac{1}{2}z^\mathsf{T}Hz + c^\mathsf{T}z + d$$

mit $z := (a,b)^\mathsf{T}$ sowie

$$H := 2\begin{pmatrix} \sum_{k=1}^m x_k^2 & \sum_{k=1}^m x_k \\ \sum_{k=1}^m x_k & m \end{pmatrix}, c := -2\begin{pmatrix} \sum_{k=1}^m x_k y_k \\ \sum_{k=1}^m y_k \end{pmatrix} \text{ und } d := \sum_{k=1}^m y_k^2.$$

Sind alle x_k gleich, ist die Aufgabe nicht sinnvoll gestellt. Sind mindestens zwei der x_k verschieden, ist H positiv definit, da alle Hauptminoren positiv sind. Die Hauptminoren von H (wir können den Vorfaktor vernachlässigen) sind $\sum_{k=1}^m x_k^2$ und

$$\det(H) = m \cdot \sum_{k=1}^m x_k^2 - \left(\sum_{k=1}^m x_k\right)^2 \overset{(\star)}{>} 0.$$

Sind die x_k verschieden, existiert ein $x_k \neq 0$ und somit gilt $\sum_{k=1}^m x_k^2 > 0$.◇

Proof. (von ($\star$)) Aus $\left(\sum_{k=1}^m x_k y_k\right)^2 \overset{\text{CS}}{\leqslant} \left(\sum_{k=1}^m x_k^2\right)\left(\sum_{k=1}^m y_k^2\right)$ folgt mit $y_k = 1$ für alle $k \in \{1, \dots, m\}$, $\left(\sum_{k=1}^m x_k\right)^2 \leqslant m \cdot \sum_{k=1}^m x_k^2$ und Gleichheit genau dann wenn $x := (x_k)_{k=1}^m$ und $(y_k)_{k=1}^m$ linear abhängig sind, das heißt $x = \lambda y = (\lambda, \dots, \lambda)$ für ein $\lambda \in \mathbb{R}$. Sind also zwei x_k verschieden ist die Ungleichung strikt. □

1.3 Konvexe Optimierungsaufgaben

Definition 1.3.1 (Konvexe Menge)
Eine Menge $C \subset \mathbb{R}^n$ ist konvex wenn für alle $x, y \in C$ und $t \in (0,1)$ gilt: $(1-t)x + ty \in C$.

Definition 1.3.2 ((Streng) konvexe Funktion)
Sei $C \subset D$ konvex und nichtleer. Eine Funktion $f\colon D \to \mathbb{R}$ heißt konvex auf C wenn

$$f((1-t)x + ty) \leqslant (1-t)f(x) + tf(y)$$

für alle $x, y \in C$ und $t \in (0,1)$ gilt.

Für strenge Konvexität ersetzt man $\leqslant$ durch $<$ und schließt $x = y$ aus. Eine Funktion f heißt konkav, wenn $-f$ konvex ist.

Example 1.3.3 Affin lineare Funktionen sind konvex, die Abbildungen $f(x) := x^4$ und $g(x) := e^x$ sind streng konvex. ◇

Im Weiteren betrachten wir $f\colon \mathbb{R}^n \supset D \to \mathbb{R}$, wobei D offen und nichtleer und $\Omega \subset D$ konvex ist. Ist f konvex auf Ω, so ist das Problem

$$\min_{x\in\Omega} f(x) \tag{C}$$

eine konvexe Optimierungsaufgabe.

Theorem 1.3.1: Lokale Minima von (C) sind global

Jede lokale Minimum von (C) ist global. Die Menge aller Lösungen von (C) ist konvex.

Proof. ① Sei $x \in \Omega$ eine lokale Lösung von (C). Dann existiert ein $r > 0$, sodass $f(x) \leqslant f(y)$ für alle $y \in \Omega \cap B(x, r)$ gilt. Seien $y \in \Omega$ beliebig und $t > 0$ so klein, dass $x_t := x + t(y - x) \in B(x, r)$ gilt. Da Ω konvex ist, gilt $x_t \in \Omega$ für alle $t \in [0, 1]$.

Da f konvex ist, folgt

$$f(x) \leqslant f(x_t) = f((1-t)x + ty) \leqslant (1-t)f(x) + tf(y).$$

Umstellen und teilen durch $t > 0$ ergibt $f(x) \leqslant f(y)$.

② Seien $x, y \in \Omega$ Lösungen von (C). Dann gilt für alle $z \in \Omega$

$$f((1-t)x+ty) \leqslant (1-t)f(x)+tf(y) \leqslant (1-t)f(z)+tf(z) = f(z),$$

somit ist $(1-t)x + ty$ auch ein Minimum. □

Corollary 1.3.4
There does not exist a convex function with exactly two minima.

Theorem 1.3.2: Streng konvex $\Longrightarrow$ Eindeutigkeit

Sei f streng konvex. Ist x eine Lösung von (C), so ist sie eindeutig und ein strenges Minimum.

Proof. Seien $x \neq y \in \Omega$ zwei Lösungen von (C), also nach Satz 1.3.1 globale Minima von f und $a := \min_{x\in\Omega} f(x)$. Dann gilt

$$f\left(\frac{x+y}{2}\right) < \frac{f(x)+f(y)}{2} = a,$$

was ein Widerspruch dazu darstellt, dass x und y Lösungen von (C) sind. Somit ist $x = y$ ein strenges Minimum. □

Counterexample 1.3.5 (Streng konvexe Funktion ohne Minimum)
Die Exponentialfunktion ist streng konvex (Ungleichung vom gewichteten arithmetischen und geometrischen Mittel) jedoch wird ihr Infimum, Null, nicht angenommen. ◇

Example 1.3.6 (Positiv definite quadratische Aufgabe)
Ist H positiv definit, ist $f(x) = \frac{1}{2}x^\mathsf{T} Hx + b^\mathsf{T} x$ streng konvex. ◇

Proof. Für $x, y \in \mathbb{R}^n$ und $t \in (0,1)$ gilt $t^2 < t$ und $(1-t)^2 < 1-t$. Es folgt

$$\begin{aligned} f((1-t)x+ty) &= \underbrace{(1-t)^2}_{<1-t} \underbrace{x^\mathsf{T} Hx}_{\geqslant 0} + (1-t)b^\mathsf{T} x + \underbrace{t^2}_{<t} \underbrace{y^\mathsf{T} Hy}_{\geqslant 0} + tb^\mathsf{T} y \\ &< (1-t)x^\mathsf{T} Hx + (1-t)b^\mathsf{T} x + ty^\mathsf{T} Hy + tb^\mathsf{T} y \\ &= (1-t)f(x) + tf(y). \end{aligned}$$ □

Corollary 1.3.7 (Lineare Regression III)
Sind zwei x_k verschieden, ist H nach Beispiel 1.3.6 positiv definit und somit strikt konvex. Somit ist die Lösungsgerade eindeutig bestimmt.

27.04.2020

1.4 Further optimisation problem examples

Example 1.4.1 (Nonlinear regression)
Again consider measurements $((\xi_i, \eta_i))_{i=1}^m$ and assume a nonlinear model $\eta(\xi) = g(x_1, \dots, x_n; \xi)$ and minimise

$$f(x) := \sum_{i=1}^m (\eta_i - g(x, \xi_i))^2 .$$ ◇

Example 1.4.2 (Newton cooling)
Consider $T(t)$, which is the coffee temperature at time t, the outside temperature T_a and measurements $((t_i, T_i))_{i=1}^m$ and a initial temperature $T_0 > T_a$. We consider the Newton cooling model, given by

$$\hat{T} := \frac{\mathrm{d}T}{\mathrm{d}t} = -K(T - T_a), \tag{2}$$

for some constant $K \geqslant 0$. By solving (2), we would like to identify T_0, T_a, K and t_0. We will use separation of variables:

$$\int_{T_0}^{T} \frac{\mathrm{d}\tilde{T}}{\tilde{T} - T_a} = \ln\left(\frac{T - T_a}{T_0 - T_a}\right) = -\int_{t_0}^{t} K \,\mathrm{d}t = -k(t - t_0),$$

and obtain

$$T(t) = T_a + (T_0 - T_a)e^{-K(t-t_0)}.$$

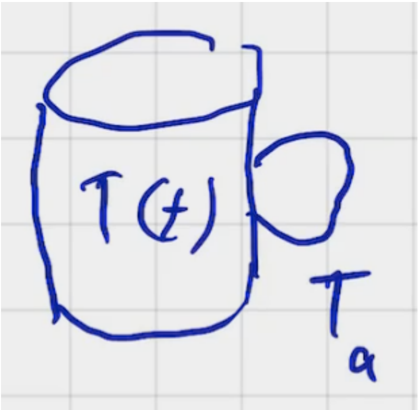


Figure 3: TODO

In the language of nonlinear regression from above, our objective is

$$g(x, \xi) = x_1 + (x_2 - x_1)e^{-x_3(\xi - x_4)}$$

and the admissable set is

$$\Omega := \{x \in \mathbb{R}^4 : x_2 - x_1 \geqslant 0, x_3 \geqslant 0, x_4 \geqslant \xi_1\},$$

so we have linear inequality constraints. ◇

Example 1.4.3 (Rosenbrock function (Banana shaped valley))
Consider the fourth-order polynomial

$$f(x,y) := 100 \underbrace{(y - x^2)^2}_{\text{parabola valley}} + \underbrace{(1-x)^2}_{\text{inclination}}$$

Clearly, $f(x,y) \geqslant 0$ for all $x, y \in \mathbb{R}$ and $f(x,y) = 0$ if $y - x^2 = 0$ and $x = 1$, so $(x,y) = (1,1)$ is a global minimum. ◇

Another example is the Himmelblau function

$$f(x,y) \coloneqq (x^2 + y - 11)^2 + (x + y^2 - 7)^2,$$

which has four local minima, which are also global with function value zero, four saddle points and a local maximum.

There are more examples, i.e. Bazaraa-Shetty, Dixon.

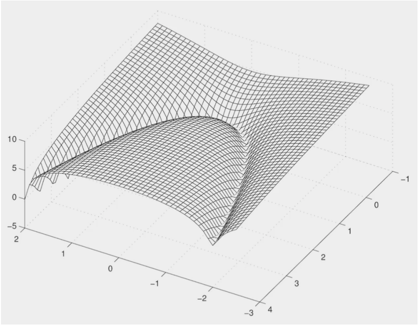


Figure 4: The Rosenbrock function (but scaled??).

2 Problem ohne Restriktion - Theorie

2.1 Optimalitätsbedingungen

Notwendige Bedingungen erster Ordnung

Sei im Weiteren $D \subset \mathbb{R}^n$ offen und nicht leer. Für hinreichende differenzierbare Funktionen $f\colon D \to \mathbb{R}$ betrachten wir

$$\min_{x \in D} f(x). \tag{PU}$$

Der Satz von Fermat besagt, dass für in $\hat{x} \in D$ differenzierbare f mit lokalem Minimum $\hat{x}$, $\nabla f(\hat{x}) = 0$ folgt. Wir nennen dies eine notwendige Bedingung erster Ordnung.

Remark 2.1.1 Wir benutzen die Konvention $f'(x)^\mathsf{T} = \nabla f(x)$.

Definition 2.1.2 (stationärer Punkt)
Gilt $\nabla f(x) = 0$, so ist x ein stationärer Punkt von f.

stationärer Punkt

Example 2.1.3 (Rosenbrock function)
We have

$$\nabla f(x,y) = \begin{pmatrix} -2 \cdot 100(y - x^2) \cdot 2x - 2(1-x) \\ 200(y - x^2) \end{pmatrix}$$

and thus $\nabla f(1,1) = 0$. ◇

Example 2.1.4 (Quadratische Optimierungsaufgabe II)
Für *symmetrische* $H \in \mathbb{R}^{n \times n}$ und $b \in \mathbb{R}^n$ sei $f(x) := \frac{1}{2} x^\mathsf{T} H x + b^\mathsf{T} x$. Es ist $\nabla f(x) = Hx + b$:

$$\begin{aligned} \frac{\partial f}{\partial x_k} &= \frac{1}{2} \frac{\partial}{\partial x_k} \left(\sum_{i=1}^n x_i [Hx]_i \right) + \frac{\partial}{\partial x_k} \sum_{i=1}^n b_i x_i \\ &= \frac{1}{2} [Hx]_k + \frac{1}{2} \sum_{i=1}^n x_i \frac{\partial}{\partial x_k} \left(\sum_{i=1}^n h_{ij} x_i \right) + b_k \\ &= \frac{1}{2} [Hx]_k + \frac{1}{2} \sum_{i=1}^n \underbrace{h_{ik}}_{=h_{ki}} x_i + b_k = [Hx]_k + b_k. \end{aligned}$$

Somit muss jede Lösung von $\min_{x \in \mathbb{R}^n} f(x)$ die Gleichung $Hx = -b$ erfüllen. Ist H positiv definit (also insbesondere invertierbar), ist die eindeutige Lösung $\hat{x} = -H^{-1} b$. ◇

30.04.2020

Example 2.1.5 Stationäre Punkte müssen keine Extrema sein, betrachte z.B. $f(x) := x^3$ und $x = 0$. ◇

Definition 2.1.6 (Richtungsableitung)
Eine Funktion f ist in $x \in D$ in Richtung $h \in \mathbb{R}^n$ richtungsdifferenzierbar, wenn die Richtungsableitung

Richtungsableitung

$$f'(x; h) := \lim_{t \searrow 0} \frac{f(x + th) - f(x)}{t}$$

existiert. Ist f in alle Richtungen h richtungsdifferenzierbar, so heißt f richtungsdifferenzierbar in x.

Example 2.1.7 Die Betragsfunktion f besitzt ein lokales Minimum in 0, ist aber dort nicht differenzierbar. Jedoch ist f dort richtungsdifferenzierbar: für $t > 0$ und $h \in \mathbb{R}$ gilt

$$\frac{f(th) - f(0)}{t} = |h|$$

und somit $f'(0; h) \geqslant 0$. ◇

Theorem 2.1.1: Variationsungleichung

Ist $x \in D$ ein lokales Minimum von (PU) und f in x richtungsdifferenzierbar, so gilt

$$f'(x; h) \geqslant 0 \quad \forall h \in \mathbb{R}^n . \qquad \text{(Variationsungleichung)}$$

Proof. Da D offen ist, existiert ein $r > 0$, sodass $f(y) \geqslant f(x)$ für alle $y \in B(x, r)$ gilt. Sei $h \in \mathbb{R}^n$. Für betragsmäßig kleine t gilt $x+th \in B(x, r)$ und somit

$$f(x + th) - f(x) \geqslant 0 \implies \frac{f(x + th) - f(x)}{t} \geqslant 0.$$ □

Die Intuition hinter der Variationsungleichung ist, dass es keine Richtung gibt, in der f fällt.

Remark 2.1.8 (Fermat's rule in the $\mathcal{C}^1$ case) If $f \in \mathcal{C}^1$, then

$$f'(x; h) = \lim_{t \searrow 0} \frac{f(x + th) - f(x)}{t} = \nabla f(x)^\mathsf{T} h.$$

If x is a local minimum, by the above theorem, we have $\nabla f(x)^\mathsf{T} h \geqslant 0$ for all $h \in \mathbb{R}^n$. Taking $h = -\nabla f(x)^\mathsf{T}$, we get $-|\nabla f(x)|^2 \geqslant 0$, and thus $\nabla f(x) = 0$.

Lemma 2.1.9 (Convex variational inequality (HW 6.2))
Let $f \colon \mathbb{R}^n \to \mathbb{R}$ be differentiable and (strictly) convex. We have

$$\left\langle \nabla f(y) - \nabla f(x), y - x \right\rangle \overset{(>)}{\geqslant} 0$$

for all x, y (with $x \neq y$).

Proof. For all x, y (with $x \neq y$)

$$\begin{aligned}
\left\langle \nabla f(y), y - x \right\rangle &= -\left\langle \nabla f(y), x - y \right\rangle \\
&= -\lim_{h \to 0} \frac{1}{h} \left(f(y + h(x - y)) - f(y)\right) \\
&= -\lim_{h \searrow 0} \frac{1}{h} \left(f(y + h(x - y)) - f(y)\right) \\
&\overset{(>)}{\geqslant} -\lim_{h \searrow 0} \frac{1}{h} \left((1 - h) f(y) + h(f(x)) - f(y)\right) \\
&= -\lim_{h \searrow 0} (f(x) - f(y)) = f(y) - f(x).
\end{aligned}$$

Thus

$$\begin{aligned}\langle \nabla f(y) - \nabla f(x), y - x \rangle &= \langle \nabla f(y), y - x \rangle + \langle \nabla f(x), x - y \rangle \\ &\geqslant f(y) - f(x) + f(x) - f(y) = 0.\end{aligned}$$ □

Notwendige Bedingungen zweiter Ordnung

Theorem 2.1.2: Notwendige Bedingung 2. Ordnung

Ist f eine $\mathcal{C}^2$-Funktion in einer Umgebung von $x \in D$ und x eine lokales Minimum von (PU), so muss neben $\nabla f(x) = 0$ auch $f''(x)$ positiv semidefinit sein.

Counterexample 2.1.10 ($f''(x) > 0$ for local minimum x)
Even for $f\colon \mathbb{R} \to \mathbb{R}$ with $f \in \mathcal{C}^2$ and a local minimum x we need not have $f''(x) > 0$: consider the zero function or $f(x) = x^4$, which has a global minimum in $\tilde{x} = 0$. We have $f''(x) = 12x^2$ and $f''(\tilde{x}) = 0$. ◇

Proof. Für $h \in \mathbb{R}^n$ sei $g(t) := f(x+th)$. Dann hat g ein lokales Minimum bei $t = 0$ und es gilt $g \in \mathcal{C}^2(\mathbb{R})$. Nach dem Satz von Taylor existiert ein $\theta \in [0,1]$, sodass

$$g(t) = g(0) + \underbrace{g'(0)}_{=h^\mathsf{T}\nabla f(x) = 0} t + \frac{t^2}{2} g''(\theta t)$$

gilt. Da x ein lokales Minimum von g ist, folgt $0 \leqslant \frac{g(t)-g(0)}{t^2} = \frac{1}{2} g''(\theta t)$. Da g'' stetig ist, folgt $g''(0) = h^\mathsf{T} f''(x) h \geqslant 0$ für $t \searrow 0$. □

Example 2.1.11 (Quadratic function) For $f(x) = \frac{1}{2} x^\mathsf{T} H x + b^\mathsf{T} x$ we have $f''(x) = H$, so if $\tilde{x}$ is a solution to (PU), H must be positive semidefinite. ◇

Example 2.1.12 For the Rosenbrock function g we have

$$g''(x,y) = \begin{pmatrix} -400(y - x^2) + 800x^2 + 2 & -400x \\ -400x & 200 \end{pmatrix}$$

and thus $g''(1,1)$ is positive definite by Silvester's theorem, because the principal minors 802 and $\det(g''(1,1))$ are positive. ◇

Hinreichende Bedingungen zweiter Ordnung

Aus der Gültigkeit von notwendigen Bedingungen erster und zweiter Ordnung kann man nicht auf lokale Extrema schließen (vgl. Beispiel 2.1.5)

Theorem 2.1.3: Hinreichende Bedingung 2. Ordnung

Seien f eine $\mathcal{C}^2$-Funktion in einer Umgebung von $x \in D$, $\nabla f(x) = 0$ sowie $f''(z)$ positiv semidefinit für alle $z \in B(x, \delta)$ für ein $\delta > 0$. Dann ist x ein lokales Minimum von (PU).

Proof. Für $y \in B(x, \delta)$ und ein $\theta \in [0, 1]$ gilt

$$f(y) - f(x) = \underbrace{f'(x)}_{=0}(y - x) + \frac{1}{2}\underbrace{(y - x)}_{h}^{\mathsf{T}} f''(\underbrace{x + \theta(y - x)}_{=z \in B_\delta(x)})(y - x) \geqslant 0$$

nach dem Satz von Taylor. □

Example 2.1.13 Consider $f_k(x, y) := x^2 + kxy + y^2$ for $k \in \{1, 2, 3\}$. They can be written as $f_k(x, y) = \frac{1}{2}\left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right)^{\mathsf{T}} A_k \left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right)$ with $A_k := \left(\begin{smallmatrix} 2 & k \\ k & 2 \end{smallmatrix}\right)$.

We have $\nabla f_k(x, y) = A_k \left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right)$ and $f_k''(x, y) = A_k$. Thus the stationary points of f_k are exactly $\ker(A_k)$ and $(0, 0)^{\mathsf{T}}$ is always a stationary point.

For $k = 1$, $(0, 0)^{\mathsf{T}}$ is a local minimiser of f_1 by theorem TODO, as $f_1''(0, 0) = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ is positive definite. Hence $(0, 0)^{\mathsf{T}}$ fulfills a quadratic growth condition theorem 2.1.4.

For $k = 2$, $(0, 0)^{\mathsf{T}}$ is a local minimiser by theorem 2.1.3, even though $f_2''(0, 0) = \left(\begin{smallmatrix} 2 & 2 \\ 2 & 2 \end{smallmatrix}\right)$ is not positive definite. But since $f_2''(x, y)$ is positive semidefinite everywhere and thus in particular in a neighbourhood of $(0, 0)^{\mathsf{T}}$, $(0, 0)^{\mathsf{T}}$ is a local minimiser.

For $k = 3$, $(0, 0)^{\mathsf{T}}$ is not a local minimiser, as $f_3''(0, 0) = \left(\begin{smallmatrix} 2 & 3 \\ 3 & 2 \end{smallmatrix}\right)$ is not positive semidefinite.

We have thus seen that the definiteness of a matrix $A \in \mathbb{R}^{n \times n}$ can also be viewed as the optimality conditions of $(0, 0)^{\mathsf{T}}$ of the quadratic $f(x) := x^{\mathsf{T}} Ax$. ◇

Example 2.1.14 (Quadratic function) Notice than if H is only invertible, a stationary point must not be a minimum: consider $f(x, y) := x^2 - y^2$, which has a stationary point in $(0, 0)$, but this is not a local minimum. ◇

Example 2.1.15
For a constant function, every $x \in \mathbb{R}^n$ is a local minimiser.

Consider

$$f(x) = \begin{cases} 0, & x \in [-1, 1], \\ (x - 1)^4, & x > 1, \\ (x + 1)^4, & x < -1 \end{cases}$$

with

$$f''(x) = \begin{cases} 0, & x \in [-1, 1], \\ 12(x - 1)^2, & x > 1, \\ 12(x + 1)^2, & x < -1. \end{cases}$$

We have $f''(x) \geqslant 0$ for all $x \in \mathbb{R}$ and $f'(x) = 0$ for all $x \in [-1, 1]$ and thus all $x \in [-1, 1]$ are local minimisers by theorem 2.1.3. ◇

Example 2.1.16 (Lineare Regression V)
Es hängt $f''(x) = H$ nicht von x ab. Ist H positiv semidefinit und gilt $Hx + b = 0$, so ist x ein lokales Minimum. Ist H positiv definit, ist x sogar eindeutig. ◇

Theorem 2.1.4: Quadratische Wachstumsbedingung

Seien f $\mathcal{C}^2$ in einer Umgebung von $x \in D$, $\nabla f(x) = 0$ und $f''(x)$ positiv definit. Dann existieren $r, a > 0$, sodass

$$f(y) \geqslant f(x) + a\|y - x\|^2 \quad \forall x \in B(x, r) \quad \text{(quadr. Wachstumsbdg.)}$$

gilt, das heißt x ist ein strenges lokales Minimum von (PU).

Proof. (Skizze) Mit Taylor-Entwicklung folgt für ein $\theta \in (0, 1)$

$$f(y) = f(x) + \underbrace{\nabla f(x)^\mathsf{T}}_{=0}(y - x) + \frac{1}{2}(y - x)f''(x + \theta(y - x))(y - x)$$

und mit Lemma 1.2.6

$$\begin{aligned}
&(y - x)f''(x + \theta(y - x))(y - x) \\
= \; &\underbrace{(y - x)f''(x)(y - x)}_{\geqslant a\|y-x\|^2} + \underbrace{(y - x)\left[f''(x + \theta(y - x)) - f''(x)\right](y - x)}_{|\cdot| \leqslant \frac{a}{2}\|y-x\|^2 \text{ für kleine } \|y-x\|^2, \text{ da } f \in \mathcal{C}^2} \\
\geqslant \; &\frac{a}{2}\|y - x\|^2.
\end{aligned}$$

□

Lemma 2.1.17
If quadratic growth condition holds in x, then $f''(x)$ is positive definite.

Proof. By the theorem of Taylor

$$\cancel{f(x)} + \frac{1}{2}f''(x)[y - x, y - x] + R(y) \geqslant \cancel{f(x)} + a\|y - x\|^2,$$

where $\lim_{y \to x} \frac{R(y)}{\|y-x\|^2} = 0$ and which can be rearranged to

$$\frac{1}{2}f''(x)[y - x, y - x] \geqslant a\|y - x\|^2 - R(y).$$

Now let $v \in \mathbb{R}^n \setminus \{0\}$. For $y := x + tv$ we have

$$\frac{1}{2}f''(x)[tv, tv] \geqslant at^2\|v\|^2 - R(x + tv),$$

and thus

$$\frac{1}{2}f''(x)[v, v] \geqslant a\|v\|^2 - \frac{R(x + tv)}{t^2} \xrightarrow{t \searrow 0} a\|v\|^2.$$

The claim follows with lemma 1.2.6. □

04.05.2020

Example 2.1.18
Betrachte $f(x) := x^{2p}$ für $p \in \mathbb{N}$ mit lokalem Minimum $x = 0$. Dass $f'(0) = 0$ und $f''(0) > 0$ implizieren, dass ein lokales Minimum vorliegt, ist nur für $p = 1$ richtig: es gilt $f'(x) = 2px^{2p-1}$ und somit $f'(0) = 0$ und $f''(x) = (2p-1)(2p)x^{2p-2}$ und somit ist $f''(0) = 0$ nur für $p > 1$ und somit nicht positiv definit. ◇

2.2 Konvexe Optimierungsaufgaben

Theorem 2.2.1: Tangenten

Sei f differenzierbar auf D. Dann ist f (strikt) konvex in Ω genau dann wenn $f(y) \overset{(>)}{\geqslant} f(x) + \nabla f(x)^\mathsf{T}(y-x)$ für alle $x, y \in \Omega$ gilt.

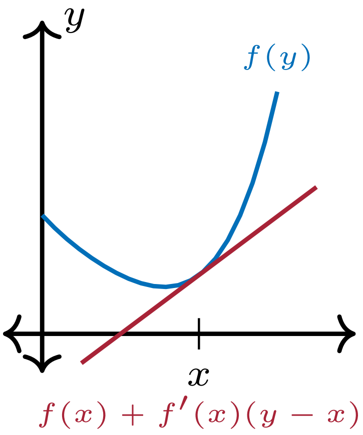


Figure 5: Tangents of convex functions always are below their graph.

Proof. Follows from the proof of lemma 2.1.9. □

Corollary 2.2.1
Let f be differentiable (strictly) convex function. If $\nabla f(x) = 0$ for some $x \in \Omega$, then x is a (strict) global minimiser of f.

Proof. For $y \in \Omega$ we have $f(y) \overset{(>)}{\geqslant} f(x) + \nabla f(x)^\mathsf{T}(y-x) = f(x)$, so x is a (strict) global minimiser. □

Theorem 2.2.2: Konvexe Variationsungleichung

Sei f differenzierbar in D, und konvex auf Ω. Dann löst $x \in \Omega$ das Problem (C) genau dann wenn

$$\nabla f(x)^\mathsf{T}(y-x) \geqslant 0 \quad \forall y \in \Omega \quad \text{(konvexe Variationsungleichung)}$$

Proof. "‘ $\Longrightarrow$ "’: Ist x ein lokale Lösung, gilt $x + t(y-x) = (1-t)x + ty \in \Omega$ für alle $t \in [0,1]$ und $y \in \Omega$. Für kleine $t > 0$ gilt also

$$\frac{f(x + t(y-x)) - f(x)}{t} \geqslant 0.$$

Mit $t \searrow 0$ folgt die Aussage.

"‘ $\Longleftarrow$ "’: Da f konvex ist gilt nach Satz 2.2.1 für alle $y \in \Omega$

$$f(y) - f(x) \geqslant \nabla f(x)^\mathsf{T}(y-x) \geqslant 0$$

und nach Satz 1.3.1 ist x sogar ein globales Minimum. □

Remark 2.2.2 (zum Beweis) Für die Hinrichtung des obigen Beweis ist die Konvexität von f nicht notwendig.

Summa summarum gilt also: ist $f\colon D \to \mathbb{R}$ differenzierbar auf eine konvexen Menge $D \subset \mathbb{R}^n$ und x ein lokale Minimerer von f auf D, so erfüllt x die notwendige Optimalitätsbedingung $\nabla f(x)^\mathsf{T}(y - x) \geqslant 0$ für alle $y \in \Omega$. Im Allgemeinen ist die Variationsungleichung jedoch nicht hinreichend. Ist f jedoch konvex auf D, so ist ein $x \in D$, welcher die Variationsungleichung erfüllt, ein globaler Minimerer von f auf D. Daher werden bei konvexen Optimierungsaufgaben keine Bedingungen zweiter Ordnung verwendet: die Bedingung erster Ordnung ist bereits hinreichend.

Corollary 2.2.3 ($\boldsymbol{x} \in \mathbf{int}(\boldsymbol{\Omega})$)
*Für $x \in \operatorname{int}(\Omega)$ gilt $\nabla f(x)^\mathsf{T} d \geqslant 0$ für jede Richtung $d \in \mathbb{R}^n$ und somit $\nabla f(x) = 0$. **(braucht man konvexität hier?)***

3 Problem ohne Restriktionen - Verfahren

07.05.2020

Um

$$\min_{x \in \mathbb{R}^n} f(x) \tag{PU}$$

numerisch zu lösen, muss $\nabla f(\hat{x}) = 0$ für den Minimierer $\tilde{x} \in \mathbb{R}^n$ gelten. Diese Gleichung kann man numerisch lösen, etwa mit dem Newton-Verfahren. Man interessiert sich für Verfahren, welche $\nabla f(\hat{x}) = 0$ lösen und gleichzeitig die Minimierung in (PU) berücksichtigen. Dazu gehören Abstiegsverfahren; iterative Verfahren, welche schrittweise den Funktionswert von f verkleinern. Man sucht also eine Folge $(x^k)_{k\in\mathbb{N}}$ mit $x^k \to \hat{x}$ und $f(x^{k+1}) < f(x^k)$ für alle $k \in \mathbb{N}$.

Sei im Folgenden $f\colon \mathbb{R}^n \to \mathbb{R}$ differenzierbar an der Stelle x.

Definition 3.0.1 (Descent direction)
A descent direction of f in x is $d \in \mathbb{R}^n$ with $\nabla f(x)^\mathsf{T} d < 0$.

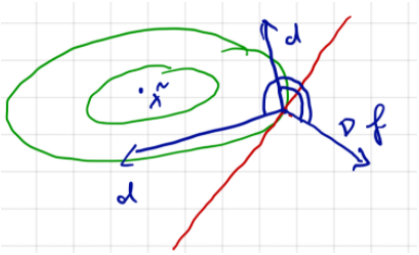


Figure 6: Because $a^\mathsf{T} b = |a||b|\cos(\sphericalangle(a,b))$, the requirement $a^\mathsf{T} b < 0$ is equivalent to $\sphericalangle(a,b) \in (\frac{\pi}{2}, \frac{3\pi}{2})$.

Remark 3.0.2 If $\tilde{x}$ is a local minimum, we have $\nabla f(\tilde{x})^\mathsf{T}(x - \tilde{x}) \geqslant 0$ for all $x \in \mathbb{R}^n$ by theorem 2.2.2, so a necessary condition is that there exists no descent direction.

Lemma 3.0.3 (Existenz einer positiven Schrittweite)
Für eine Abstiegsrichtung d existiert ein $c > 0$, sodass $f(x + ad) < f(x)$ für alle $a \in (0, c]$ gilt.

Proof. We have

$$\nabla f(x)^\mathsf{T} d = \lim_{a \searrow 0} \frac{f(x + ad) - f(x)}{a} < 0$$

and thus there exists a $c > 0$ such that $\frac{f(x+ad)-f(x)}{a} < 0$ for all $a \in (0, c]$.□

Counterexample 3.0.4 (HW 4.2)
The reverse direction of the above lemma is not correct: consider $f(x) := -x^2$ with $x = 0$ and $d = 1$. Then $f(x + td) = -t^2 < 0 = f(x)$ for all $t \in (0, a]$ for any $a \in \mathbb{R}_{>0}$. It follows $\nabla f(x)^\mathsf{T} d = -2 \cdot 0 \cdot 1 = 0 \nless 0$. ◇

Example 3.0.5 (Abstiegsrichtungen)
Sei $\nabla f(x) \neq 0$.

1. Der Antigradient $-\nabla f(x)$ ist eine Abstiegsrichtung (genannt *steepest descent*):

   Antigradient

   $$\nabla f(x)^\mathsf{T} \cdot (-\nabla f(x)) = -\|\nabla f(x)\| < 0.$$

2. Ist $A \in \mathbb{R}^{n\times n}$ symmetrich und positiv definit, so ist $-A^{-1}\nabla f(x)$ eine Abstiegsrichtung: es ist auch A^{-1} positiv definit und somit gilt $z^\mathsf{T} A^{-1} z > 0$ für alle $z \in \mathbb{R}^d \backslash\{0\}$ und somit $-\nabla f(x)^\mathsf{T} A^{-1} \nabla f(x) < 0$.

   ist λ ein Eigenwert von A, ist λ^{-1} ein Eigenvektor von A^{-1}

◇

The following is an general descent algorithm.

① Choose $x^0 \in \mathbb{R}^n$ and set $k := 0$.

② If $\nabla f(x^k) = 0$ holds, stop.

③ Compute a descent direction d^k and a step size σ_k such that

$$f(x^k + \sigma_k d^k) < f(x^k).$$

Define $x^{k+1} = x^k + \sigma_k d^k$.

④ Set $k \to k+1$ and return to step ②.

Remark 3.0.6 (Stopping criteria) Step ② is only of academic nature, in reality we choose e.g. $|\nabla f(x^k)| < \varepsilon$ or $|f(x^{k+1}) - f(x^k)| + |x^{k+1} - x^k| < \varepsilon$ or $f(x^{k+1}) - f(x^k) \approx \sigma_k \nabla f(x^k)^\mathsf{T} d^k < \varepsilon$ and $|x^{k+1} - x^k| = \sigma_k |d^k| < \rho$.

3.1 Das Newton-Verfahren

07.05.2020

Setzen wir $F\colon \mathbb{R}^n \to \mathbb{R}^n$, $x \mapsto \nabla f(x)$, so müssen wir $F(x) = 0$ lösen.

Ist $x^{(k)}$ bestimmt, so verhält sich $F(x)$ nahe bei $x^{(k)}$ in der ersten Näherung wie $F(x^{(k)} + F'(x^{(k)}) \cdot (x - x^{(k)})$. Wir lösen also das lineare Gleichungssystem

$$F(x^{(k)}) + F'(x^{(k)}) \cdot (x - x^{(k)}) = 0.$$

Ist $F'(x^{(k)})$ invertierbar, so ist unsere neue Näherung

$$x^{(k+1)} = x^{(k)} - F'(x^{(k)})^{-1} F(x^{(k)}).$$

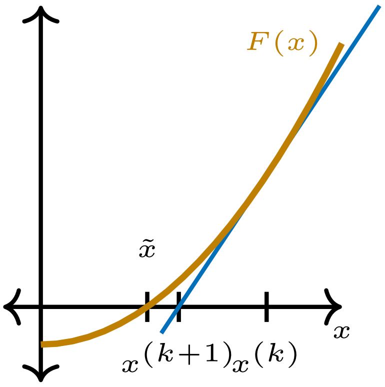


Figure 7: The new Newton iterate is the zero of the tangent to F at x^k.

Für die Konvergenzanalyse benötigen wir folgenden Voraussetzungen

① F ist auf einer offenen Teilmenge $D \subset \mathbb{R}^n$ differenzierbar und hat eine Nullstelle $\hat{x} \in D$.

② F' ist Lipschitz-stetig in D.

③ Es existiert $F'(\hat{x})^{-1}$.

Der Konvergenzbeweis des Newton-Verfahrens beruht auf den folgenden Aussagen.

Lemma 3.1.1
Für alle $x, y \in D$ gilt $\|F(x) - F(y) - F'(y)(x - y)\| \leqslant \frac{L}{2}\|x - y\|^2$.

Proof. Definiere

$$g\colon [0,1] \to \mathbb{R}^n,\ t \mapsto F(y + t(x - y))$$

Die linke Seite ist dann gleich

$$\begin{aligned}
\|g(1) - g(0) - F'(y)(x-y)\| &= \left\| \int_0^1 g'(t)\,\mathrm{d}t - F'(y)(x-y) \right\| \\
&= \left\| \int_0^1 \left[F'(y + t(x-y)) - F'(y) \right] (x-y)\,\mathrm{d}t \right\| \\
&\leqslant \|x-y\| \int_0^1 \|F'(y+t(x-y)) - F'(y)\|\,\mathrm{d}t \\
&\overset{\textcircled{2}}{\leqslant} \|x-y\| \int_0^1 L\|y + t(x-y) - y\|\,\mathrm{d}t \\
&= L\|x-y\|^2 \int_0^1 t\,\mathrm{d}t = \frac{L}{2}\|x-y\|^2,
\end{aligned}$$

wobei wir die Lipschitz-Stetigkeit von F' nutzen. □

Lemma 3.1.2

Seien $A \in \mathbb{R}^{n\times n}$ invertierbar und $S \in \mathbb{R}^{n\times n}$ mit $\|A^{-1}\| < \|S\|^{-1}$, so existiert $(A+S)^{-1}$ und es gilt

$$\|(A+S)^{-1}\| \leqslant \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|S\|}.$$

Proof. Es gilt $\|A^{-1}(-S)\| \leqslant \|A^{-1}\|\|S\| < 1$ und somit

$$(I - A^{-1}\cdot(-S))^{-1} = \sum_{k=0}^{\infty} A^{-k}(-S)^k,$$

was zu

$$(A+S)^{-1} = \sum_{k=0}^{\infty} A^{-k-1}(-S)^k$$

äquivalent ist. Es folgt

$$\|(A+S)^{-1}\| \overset{\triangle\neq}{\leqslant} \|A^{-1}\| \sum_{k=0}^{\infty} \|A^{-1}\|^k \|S\|^k = \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|S\|}.$$ □

Corollary 3.1.3 (des Banachschen Fixpunktsatzes)

Sei $G\colon \overline{B}(\hat{x}, r) \to \mathbb{R}^n$ ein Kontraktion, das heißt L-Lipschitz stetig mit $L < 1$. Ist $\hat{x}$ ein Fixpunkt von F, so ist es der einzige in $\overline{B}(\hat{x}, r)$. Ausgehend von jedem $x^{(0)} \in \overline{B}(\hat{x}, r)$ konvergiert die Folge $x^{(k+1)} := G(x^{(k)})$ gegen $\hat{x}$ und es gilt $\|x^{(k)} - \hat{x}\| \leqslant L^k\|x^{(0)} - \hat{x}\|$.

Theorem 3.1.1: Konvergenz des Newton-Verfahrens

Es existieren $\delta, c > 0$, sodass das Newton-Verfahren für jeden Startpunkt $x^{(0)} \in B(\hat{x}, \delta)$ eine gegen $\hat{x}$ konvergente Folge $(x^{(k)})_{k\in\mathbb{N}}$ definiert, welche

$$\|x^{(k+1)} - \hat{x}\| \leqslant c\|x^{(k)} - \hat{x}\|^2 \qquad \text{(quadratische Konvergenz)}$$

erfüllt.

Proof. (Skizze) ① Mit Lemma 3.1.2 folgt

$$\|F'(x)^{-1}\| \leqslant 2\|F'(\hat{x})^{-1}\|$$

für alle $x \in B(\hat{x}, \delta_1)$: Zunächst gilt

$$\|F'(x) - F'(y)\| \overset{②}{\leqslant} L\|x - y\| \overset{???}{\leqslant} \frac{1}{2\|F'(\hat{x})^{-1}\|}.$$

Wähle $A := F'(\hat{x})$ und $S := F'(x) - F'(\hat{x})$, so folgt

$$\|A^{-1}\|\|S\| = \|F'(\hat{x})^{-1}\|\|F'(x) - F'(\hat{x})\| \leqslant \frac{1}{2} < 1$$

Mit Lemma 3.1.2 folgt, dass $(A + S)^{-1} = F'(x)^{-1}$ existiert und

$$\|F(x)^{-1}\| \leqslant \frac{\|F'(\hat{x})^{-1}\|}{1 - \|F'(\hat{x})^{-1}\| \cdot \|F'(x) - F'(\hat{x})\|} \leqslant \frac{\|A^{-1}\|}{1 - \frac{1}{2}} = 2\|A^{-1}\|$$

gilt.

② Mit ① folgt für $x, y \in B(\hat{x}, \delta_1)$

$$\begin{aligned}\|F'(x)^{-1} - F'(y)^{-1}\| &= \|F'(x)^{-1}(F'(y) - F'(x))F'(y)^{-1}\| \\ &\leqslant \|F'(x)^{-1}\|\|F'(y) - F'(x)\|\|F'(y)^{-1}\| \\ &\leqslant 2\|F'(\hat{x})^{-1}\| \cdot L\|x - y\| \cdot 2\|F'(\hat{x})^{-1}\| \\ &= 4L\|F'(\hat{x})^{-1}\|^2 \cdot \|x - y\|.\end{aligned}$$

③ Das Newton-Verfahren ist Fixpunktiteration für $G(x) := x - F'(x)^{-1}F(x)$. Ferner ist G eine Kontraktion in $B(\hat{x}, \delta)$, wobei wir $\delta \leqslant \delta_1$ klein genug wählen. Für $x, y \in B(\hat{x}, \delta)$ gilt

$$\begin{aligned}G(x) - G(y) &= x - y - F'(x)^{-1}F(x) + F'(y)^{-1}F(y) \\ &= F'(x)^{-1}F'(x) \cdot (x - y) - F'(x)^{-1}F(x) \\ &\quad\ + F'(y)^{-1}F(y) \\ &= F'(x)^{-1}(F'(x) \cdot (x - y) - F(x) + F(y)) \\ &\quad + (F'(y)^{-1} - F'(x)^{-1})F(y)\end{aligned}$$

und somit mit Lemma 3.1.1

$$\begin{aligned}\|G(x) - G(y)\| &\overset{\triangle\neq}{\leqslant} \|F'(x)^{-1}\| \cdot \|F'(x) \cdot (x - y) - F(x) + F(y)\| \\ &\quad + \|F'(y)^{-1} - F'(x)^{-1}\| \cdot \|F(y)\| \\ &\leqslant 2\|F'(\hat{x})^{-1}\| \cdot \frac{L}{2}\|x - y\|^2 \\ &\quad + 4L\|F'(\hat{x})^{-1}\|^2\|x - y\| \cdot \|F(y)\| \\ &\leqslant \|x - y\| \left(L\|A^{-1}\| \cdot \|x - y\| + \|A^{-1}\|^2 \cdot \|F(y)\|\right) \\ &\leqslant \frac{1}{2}\|x - y\|,\end{aligned}$$

④ Mit Lemma 3.1.1 folgt die quadratische Konvergenz:

$$\begin{aligned}\|x^{(k+1)} - \hat{x}\| &= \|x^{(k)} - F'(x^{(k)})^{-1}F(x^{(k)}) - \hat{x}\| \\ &\overset{\triangle\neq}{\leqslant} \|F'(x^{(k)})^{-1}\| \\ &\qquad \cdot \|\underbrace{F(\hat{x})}_{=0} - F(x^{(k)}) - F'(x^{(k)})(x^{(k)} - \hat{x})\| \\ &\leqslant 2\|F'(\hat{x})^{-1}\| \cdot \frac{L}{2}\|x^{(k)} - \hat{x}\|,\end{aligned}$$

also $c = L \cdot \|F'(\hat{x})^{-1}\|$. □

11.05.2020

Unsere Forderung an f sind die Voraussetzungen des folgenden Satzes

Theorem 3.1.2: Konvergenz des Newton-Verfahrens

Seien f'' Lipschitz-stetig in einer Umgebung eines lokalen Minimums $\hat{x}$ von f und $f''(\hat{x})$ positiv definit (und somit invertierbar). Dann konvergiert das Newton-Verfahren

$$x^{(k+1)} = x^{(k)} - f''(x^{(k)})^{-1} \nabla f(x^{(k)}) \tag{3}$$

lokal quadratisch gegen $\hat{x}$.

Remark 3.1.4 (Numerische Umsetzung)
In der numerischen Umsetzung invertiert man $f''(x^{(k)})$ nicht, sondern löst stattdessen das Gleichungssystem

$$f''(x^{(k)}) \cdot (x^{(k+1)} - x^{(k)}) = -\nabla f(x^{(k)}),$$

das heißt man findet eine Richtung $d^{(k)} \in \mathbb{R}^n$, sodass

$$f''(x^{(k)}) d^{(k)} = -\nabla f(x^{(k)})$$

gilt und setzt

$$x^{(k+1)} := x^{(k)} + d^{(k)}$$

Definition 3.1.5 (Newton-Richtung)
Der Vektor $d^{(k)} := -f''(x^{(k)})^{-1} \cdot \nabla f(x^{(k)})$ heißt Newton-Richtung.

Newton-Richtung

Da $f''(x^k)$ und somit $(f''(x^k))^{-1}$ positiv definit sind, ist die Newton-Richtung eine Abstiegsrichtung nach Beispiel 3.0.5 (2).

Remark 3.1.6 (gedämpftes Newton-Verfahren)
Damit wird das Newton-Verfahren aber nicht automatisch ein Abstiegsverfahren, denn es wählt immer die Schrittweite $\sigma = 1$, und die kann zu groß sein! Deshalb wendet man das gedämpfte Newton-Verfahren

$$x^{(k+1)} = x^{(k)} - \sigma_k f''(x^{(k)})^{-1} \nabla f(x^{(k)})$$

mit $\sigma_k < 1$ an.

Remark 3.1.7 (SQP-Verfahren)
Wir können das Newton-Verfahren als

$$f''(x^{(k)}) \cdot (x^{(k+1)} - x^{(k)}) + \nabla f(x^{(k)}) = 0$$

schreiben, was die notwendige Optimalitätsbedingung für Lösungen der quadratischen Optimierungsaufgabe

$$\min_{x \in \mathbb{R}^n} \nabla f(x^{(k)})^\mathsf{T} \cdot (x - x^{(k)}) + \frac{1}{2}(x - x^{(k)})^\mathsf{T} f''(x^{(k)}) \cdot (x - x^{(k)}) \qquad (Q_k)$$

ist.

Ist $f''(x^{(k)})$ positiv definit, so ist die eindeutige Lösung $x^{(k+1)} = x^{(k)} - f''(x^{(k)})^{-1}\nabla f(x^{(k)})$. Damit ist das Newton-Verfahren äquivalent zu Lösung einer Folge quadratischen Optimierungsaufgaben (wenn $f''(\hat{x})$ positiv definit ist) und damit das sequentiell-quadratisches Optimierungsverfahren-Verfahren (*sequential quadratic programming*-Verfahren)

$$\min_z \nabla f(x^{(k)})^\mathsf{T} z + \frac{1}{2} z^\mathsf{T} f''(x^{(k)}) z \quad \text{and} \quad x^{(k+1)} := x^{(k)} + z^{(k)}.$$

3.2 Descent methods - general properties

By lemma 3.0.3 we have $f(x^{(k)} + \sigma_k d^{(k)}) < f(x^{(k)})$ for some descent direction $d^{(k)}$ and small $\sigma_k > 0$. Our goal now is convergence analysis, i.e. showing that $\nabla f(x^{(k)}) \xrightarrow{k\to\infty} 0$.

Definition 3.2.1 (Assumptions)
Assumption level set compact (ALC): for $x^{(0)} \in \mathbb{R}^d$, $\mathcal{N}(f, f(x^{(0)}))$ is compact.

(AFD): We have $f \in \mathcal{C}^1$ on a open convex set $D_0 \supset \mathcal{N}(f, f(x^{(0)}))$.

Remark 3.2.2 In descent methods, $f(x^{k+1}) < f(x^{(k)})$, i.e. $x^{(k)} \in \mathcal{N}\left(f, f(x^{(0)})\right)$. If (ALC) holds, $(x^{(k)})_{k\in\mathbb{N}}$ and $(f(x^{(k)}))_{k\in\mathbb{N}}$ are bounded and by theorem 1.2.1, a minimiser $\hat{x}$ exists.

In order to get convergence, we have to formulate requirements for the descent direction and the step size.

Requirements for the step size

Counterexample 3.2.3 (Step size decreases to fast)
Consider $f(x) := x^2$, $d^{(k)} := -1$ and $\sigma_k := 2^{-k-2}$ for all $k \geqslant 0$.

The sequence $(x^{(k)})_{k\in\mathbb{N}}$ defined by $x^{(k+1)} = x^{(k)} + \sigma_k d^{(k)} = x^{(k)} - \frac{1}{2^{k+2}}$ and $x^{(0)} = 1$ converge to $\frac{1}{2}$:

$$x^{(k+1)} = x^{(0)} - \sum_{j=0}^{k} \frac{1}{2^{k+2}} = 1 - \frac{1}{4}\frac{1 - \frac{1}{2^{k+1}}}{1 - \frac{1}{2}} = \frac{1}{2} + \frac{1}{2^{k+2}} \xrightarrow{k\to\infty} \frac{1}{2}. \qquad \diamond$$

We aim to show $\nabla f(x^{(k)}) \to 0$. The first step will be to show that

$$\frac{\nabla f(x^{(k)})^\mathsf{T} d^{(k)}}{|d^{(k)}|} \xrightarrow{k\to\infty} 0.$$

Consider a first order approximation

$$f(x + \sigma d) - f(x) \approx \sigma \nabla f(x)^\mathsf{T} d.$$

Definition 3.2.4 (Sufficiently fast descent)
(R1): There exists a constant $c_1 > 0$ independent of k, such that

$$f(x^{(k)} + \sigma_k d^{(k)}) - f(x^{(k)}) \leqslant c_1 \sigma_k \nabla f(x^{(k)})^\mathsf{T} d^{(k)} < 0.$$

Let us consider the sequence $(f(x^{(k)}))_{k\in\mathbb{N}}$, which is bounded by (ALC) and monotone (by design of descent algorithm) and thus convergent. Then $\sigma_k \nabla f(x^{(k)}) d^{(k)} \to 0$.

Definition 3.2.5 (Lower step size bound)
(R2): There exists a constant $c_2 > 0$ independent of k such that

$$\sigma_k \geqslant -c_2 \frac{\nabla f(x^{(k)})^\mathsf{T} d^{(k)}}{|d^{(k)}|^2}.$$

If (R1) and (R2) hold, then

$$\begin{aligned} f(x^{(k)} + \sigma_k d^{(k)}) &\overset{(R1)}{\leqslant} f(x^{(k)}) + c_1\sigma_k \underbrace{\nabla f(x^{(k)})^\mathsf{T} d^{(k)}}_{<0} \\ &= f(x^{(k)}) - c_1\sigma_k |\nabla f(x^{(k)})^\mathsf{T} d^{(k)}| \\ &\overset{(R2)}{\leqslant} f(x^{(k)}) - \underbrace{c_1 c_2}_{=:c} \frac{(\nabla f(x^{(k)})^\mathsf{T} d^{(k)})^2}{|d^{(k)}|^2}. \end{aligned}$$

Definition 3.2.6 (Efficient step size condition)
Assume (ALC). A step size σ_k with (R1) and (R2) satisfies the sufficient decrease condition. A step size with

$$f(x^{(k)} + \sigma_k d^{(k)}) \leqslant f(x^{(k)}) - c\left(\frac{\nabla f(x^{(k)})^\mathsf{T} d^{(k)}}{|d^{(k)}|}\right)^2 \tag{ES}$$

with a constant $c > 0$ independent of k, is called efficient.

Assume σ_k is efficient, i.e.

$$0 \xleftarrow{k\to\infty} f(x^{(k+1)}) - f(x^{(k)}) \leqslant -c\left(\frac{\nabla f(x^{(k)})^\mathsf{T} d^{(k)}}{|d^{(k)}|}\right)^2 < 0.$$

Then $\frac{\nabla f(x^{(k)})^\mathsf{T} d^{(k)}}{|d^{(k)}|} \xrightarrow{k\to\infty} 0$, as we wanted. We have

$$\frac{\nabla f(x^{(k)})^\mathsf{T} d^{(k)}}{|d^{(k)}|} = |\nabla f(x^{(k)})| \cos\left(\sphericalangle(\nabla f(x^{(k)}), d^{(k)})\right).$$

Remark 3.2.7 (HW 4.3 (a)) We have seen that if σ_k is efficient, $(f(x^{(k)}))_{k\in\mathbb{N}}$ is strictly monotonically decreasing. If $(f(x^{(k)}))_{k\in\mathbb{N}}$ diverges, it can only diverge to $-\infty$, and then f does not possess a minimiser.

Remark 3.2.8 (Constant step size, HW 4.3(c)) A constant step size σ_k is not efficient and we can't even guarantee $\|f(x^{(k+1)})\| \leqslant \|f(x^{(k)})\|$: Consider $f(x) := x^2$ and $x^{(0)} = 1$ and $d^{(k)} = -\nabla f(x^{(k)})$. Then $x^{(k)} = (-1)^k$, which is not convergent.

If $f(x) = x^2$ for $x \geqslant 0$ and x^4 for $x < 0$, it is continuously differentiable, but we have $\|f'(x^{(1)})\| = \|4 \cdot (-1)\| > \|2 \cdot 1\| = \|f'(x^{(0)})\|$.

Requirements for the search directions

Thus to ensure that $\nabla f(x^{(k)}) \to 0$ we have to avoid $\nabla f(x^{(k)}) \perp d^{(k)}$ for large k (this is slow convergence). We have

$$\cos(\sphericalangle(\nabla f(x^{(k)}), d^{(k)})) = \frac{\nabla f(x^{(k)})^\mathsf{T} d^{(k)}}{|d^{(k)}||\nabla f(x^{(k)})|} =: \beta_k \tag{4}$$

Then $\beta_k |\nabla f(x^{(k)})| = \frac{\nabla f(x^{(k)})^\mathsf{T} d^{(k)}}{|d^{(k)}|} \to 0$ for efficient step sizes, as shown above. We can infer from this that $\nabla f(x^{(k)}) \to 0$ only if $-\beta_k \geqslant c > 0$ is bounded away from zero for all $k \in \mathbb{N}$.

This constraint limits the angle of the argument of the cosine to something strictly between $\frac{\pi}{2}$ and $\frac{3\pi}{2}$, and gives rise to a cone of possible directions:

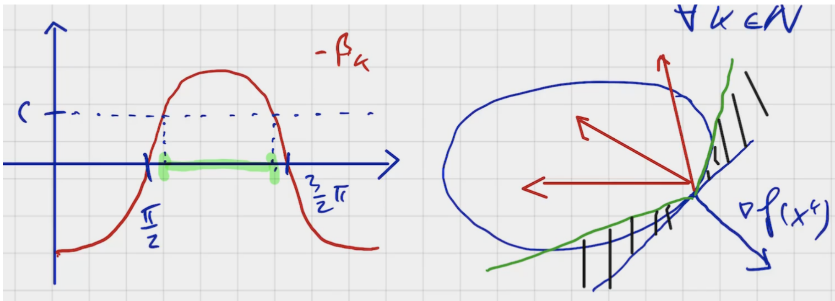


Figure 8: Directions in the black dashed area are descent directions, which are not gradient related, because they could asymptotically be perpendicular to the gradient.

Definition 3.2.9 ((strictly) Gradient-related)
Let $x \in \mathcal{N}\left(f, f(x^{(0)})\right)$. Then $d \in \mathbb{R}^n$ is (strictly) gradient-related if there exists a $c_3 > 0$ such that

gradient-related

$$-\nabla f(x)^\mathsf{T} d \geqslant c_3 |\nabla f(x)||d|$$

holds (and there exists a $c_4 > 0$ independent of x and d such that $c_4|\nabla f(x)| \geqslant |d| \geqslant \frac{1}{c_4}|\nabla f(x)|$).

Note that by Cauchy-Schwartz, $-\nabla f(x)^\mathsf{T} d \leqslant |\nabla f(x)||d|$, which is the opposite of the above inequality. If d is strictly gradient related, it "grows similar to" $\nabla f(x)$.

Example 3.2.10 (Steepest descent is strictly gradient related)
The direction $d := -\nabla f(x)$ is strictly gradient related ($c_3 = c_4 = 1$). ◇

14.05.2020

Definition 3.2.11 ((ALG) and (AHP))
(AGL): ∇f is Lipschitz continuous.

(AHP): (uniformly positive definite) for $f \in \mathcal{C}^2$ and $a > 0$ there holds that $h^\mathsf{T} f''(x) h \geqslant a|h|^2$ for all $h \in \mathbb{R}^n$ and for all $x \in D \subset \mathbb{R}^n$ (which is an open set).

Remark 3.2.12 The function $x \mapsto e^x$ is not uniformly positive definite for $D = \mathbb{R}$.

Lemma 3.2.13

Let $f\colon \mathbb{R}^n \supset D \to \mathbb{R}$ be a $\mathcal{C}^2$ function and D be an open convex subset containing $N\left(f, f(x^{(0)})\right)$ and (AHP) be fulfilled. Then

① $\mathcal{N}\left(f, f(x^{(0)})\right)$ *is convex and compact,*

② (PU) *has a unique solution $\tilde{x}$, which is the only stationary point of f,*

③ *for all $x \in \mathcal{N}(f, f(x^{(0)}))$ we have*

$$\frac{a}{2}|x-\tilde{x}|^2 \leqslant f(x) - f(\tilde{x}) \leqslant \frac{1}{2a}|\nabla f(x)|^2.$$

Proof. Alt, Lecture Notes □

Theorem 3.2.1: Convergence of descent algorithms

Under the same assumptions as in the lemma above and if $d^{(k)}$ is gradient related in $x^{(k)}$ and $(\sigma_k)_{k\in\mathbb{N}}$ are efficient, then $x^{(k)} \to \tilde{x}$, which is the unique solution to (PU). There exists a $q \in (0,1)$ such that

$$f(x^{(k)}) - f(\tilde{x}) \leqslant q^k \left(f(x^{(0)}) - f(\tilde{x})\right)$$

and

$$|x^{(k)} - \tilde{x}|^2 \leqslant \frac{2}{a} q^k \left(f(x^{(0)}) - f(\tilde{x})\right).$$

Remark 3.2.14 The a from the theorem is the a from the lemma. **(is it???)**

Remark 3.2.15 (Linear convergence)
Taking the square root yields

$$|x^{(k)} - \tilde{x}| \leqslant C\hat{q}^k$$

with $\hat{q} := \sqrt{q}$. Hence $(x^{(k)})_{k\in\mathbb{N}}$ is linearly convergent.

Proof. As σ_k is efficient and by gradient relatedness, we have

$$\begin{aligned} f(x^{(k)}) - f(x^{(k+1)}) &\geqslant c\left(\frac{\nabla f(x^{(k)})^\top d^{(k)}}{|d^{(k)}|}\right)^2 \overset{(4)}{=} c\beta_k^2 |\nabla f(x^k)|^2 \\ &\overset{(\star)}{\geqslant} 2\alpha\beta_k^2 c\left(f(x^k) - f(\tilde{x})\right) \geqslant \tilde{c}\left(f(x^k) - f(\tilde{x})\right) \end{aligned}$$

by lemma 3.2.13 $(\star)$ for some $\tilde{c} \in (0,1)$. We have thus shown

$$f(x^{k+1}) - f(x^k) \leqslant -\tilde{c}\left(f(x^k) - f(\tilde{x})\right). \tag{5}$$

We thus have

$$\begin{aligned} 0 &\leqslant f(x^{k+1}) - f(\tilde{x}) = f(x^{k+1}) - f(x^k) + f(x^k) - f(\tilde{x}) \\ &\overset{(5)}{\leqslant} \underbrace{(1-\tilde{c})}_{=:q\in(0,1)} \left(f(x^k) - f(\tilde{x})\right) \leqslant q^k \left(f(x^0) - f(\tilde{x})\right) \end{aligned}$$

for all $k \in \mathbb{N}$, yielding the first estimate. The second claim follows directly from the first one and lemma 3.2.13 ③. □

3.3 Line search methods

We consider $\varphi(\sigma) := f(x + \sigma d)$ with $\varphi'(\sigma) = \nabla f(x)^\mathsf{T} d$.

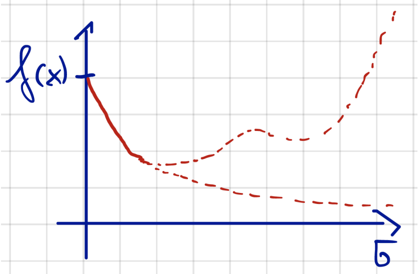


Figure 9: Let d be a descent direction. Then $\varphi'(0) = \nabla f(x)^\mathsf{T} d < 0$. We don't what φ to decrease so much.

Exact step size

We take $\sigma > 0$ such that

$$\min_{s\geqslant 0} \varphi(s) = \varphi(\sigma).$$

But this is not always possible, if e.g. f has multiple minima or $\varphi(s) = e^{-s}$. But if $x \in \mathcal{N}(f, f(x^0))$ is compact (ALC), there exists a $s > 0$ such that $\varphi(s) > \varphi(0)$.

Definition 3.3.1 (Exact step size)
The exact step size $\sigma_E > 0$ is such that

$$\varphi'(s) = \begin{cases} < 0, & \text{if } s \in [0, \sigma_E), \\ = 0, & \text{if } s = \sigma_E. \end{cases}$$

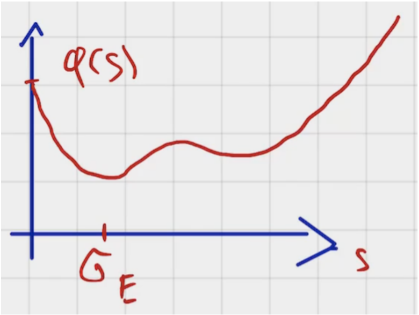


Figure 10: The exact step size is the "first" local minimum of φ.

If ∇f is Lipschitz-continuous (AGL), we have

$$\begin{aligned} \varphi'(\sigma_E) = 0 &= \nabla f(x + \sigma_E d)^\mathsf{T} d = \nabla f(x)^\mathsf{T} d + \left[\nabla f(x + \sigma_E d) - \nabla f(x)\right]^\mathsf{T} d \\ &\overset{\text{CS}}{\leqslant} \nabla f(x)^\mathsf{T} d + |\nabla f(x + \sigma_E d) - \nabla f(x)|\,|d| \leqslant \nabla f(x)^\mathsf{T} d + \sigma_E L |d|^2. \end{aligned}$$

Thus a lower bound is

$$\sigma_E \geqslant \tilde{\sigma} := -\frac{\nabla f(x)^\mathsf{T} d}{L|d|^2}.$$

Moreover, one can show that

$$f(x + \sigma_E d) \leqslant f(x) + \frac{1}{2}\tilde{\sigma}\nabla f(x)^\mathsf{T} d,$$

and thus σ_E is an efficient step size.

Example 3.3.2 (Quadratic function)
Usually, σ_E is difficult to compute. But if $f(x) = \frac{1}{2}x^\mathsf{T} H x + b^\mathsf{T} x$ and $f''(x) = H$ is positive definite, then

$$\sigma_E = -\frac{\nabla f(x)^\mathsf{T} d}{d^\mathsf{T} H d}. \qquad \diamond$$

Proof. We have

$$\varphi(t) = f(x+td) = \frac{1}{2}x^{\mathsf{T}}Hx + td^{\mathsf{T}}Hx + \frac{1}{2}t^2 d^{\mathsf{T}}Hd + b^{\mathsf{T}}x + \sigma b^{\mathsf{T}}x$$

and thus

$$\varphi'(t) = d^{\mathsf{T}}Hx + \sigma d^{\mathsf{T}}Hd + b^{\mathsf{T}}d \overset{!}{=} 0 \iff t = -\frac{(Hx+b)^{\mathsf{T}}d}{d^{\mathsf{T}}Hd}.$$

One easily sees that the above result also holds if H is not symmetric (HW 4.4). □

Armijo step size

Let $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$ be a direction. For the Armijo step size σ_A, we require want that (R1) and (R2) hold because then σ_A is efficient.

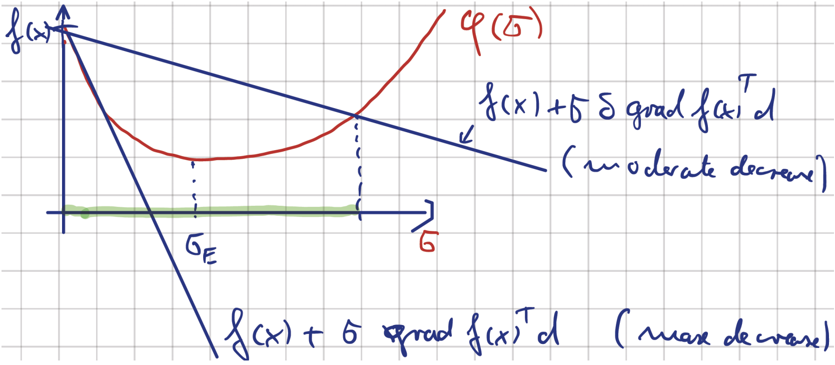


Figure 11: The requirements for σ_A visualised.

The algorithm (by Armijo and Goldstein) is

① Choose the flattening parameter $\delta \in (0,1)$, efficiency parameter $\gamma > 0$ and $0 < \beta_1 \leqslant \beta_2 < 1$.

② Initial step size. Take $\sigma_0 \geqslant -\gamma \frac{\nabla f(x)^{\mathsf{T}} d}{|d|^2}$.

③ If $f(x+\sigma_j d) \leqslant f(x) + \delta\sigma_j \nabla f(x)^{\mathsf{T}} d$, then $\sigma_A = \sigma_j$.

④ Else: reduce σ_j such that $\tilde{\sigma}_j \in [\beta_1\sigma_j, \beta_2\sigma_j]$ and iterate $j \to j+1$ and return to step ③.

Assuming (ALC), one can show that after finitely many steps, (R1) and (R2) are satisfied, so σ_A is efficient (proof in lecture notes).

Powell step size

We demand σ_P to fulfil (R1) and $\nabla f(x+\sigma d)^{\mathsf{T}} d \geqslant \beta \nabla f(x)^{\mathsf{T}} d$ with $0 < \delta < \beta < 1$.

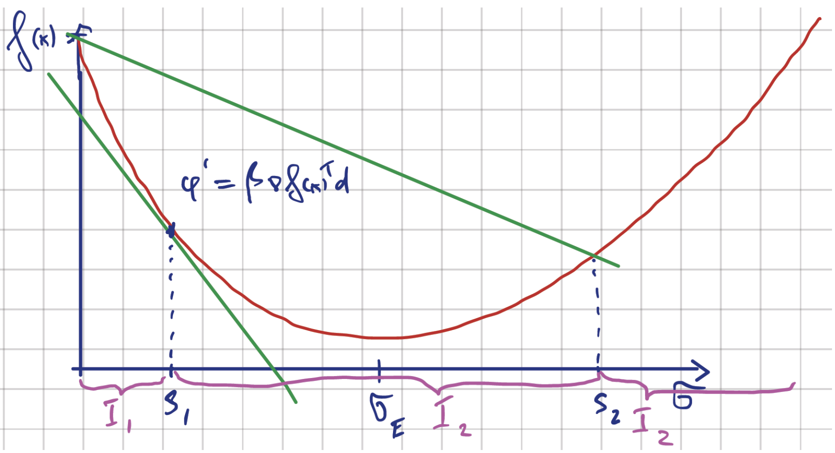


Figure 12: The intersections s_1 and s_2 divide $[0, \infty)$ into three intervals $I_1 := [0, s_1)$, $I_2 := [s_1, s_2]$ and $I_3 := (s_2, \infty)$.

Thus the requirements boil down to $\sigma_P \in I_2$.

We will set up an iteration scheme with nested intervals. Define

$$G_1(\sigma) := \begin{cases} \frac{f(x+\sigma d)-f(x)}{\sigma \nabla f(x)^\mathsf{T} d}, & \text{if } \sigma > 0, \\ 1, & \text{if } \sigma = 0, \end{cases}$$

which is continuous, and

$$G_2(\sigma) := \frac{\nabla f(x + \sigma d)^\mathsf{T} d}{\nabla f(x)^\mathsf{T} d}$$

From (R1) we get $G_1(\sigma) \geqslant \delta$ and from the second condition we get $G_2(\sigma) \leqslant \beta$. Moreover, $G_1(\sigma) \geqslant \delta$ and $G_2(\sigma) > \beta$ holds only in I_1 and $G_1(\sigma) \geqslant \delta$ and $G_2(\sigma) \leqslant \beta$ holds only in I_2 and $G_1(\sigma) < \delta$ and $G_2(\sigma) \leqslant \beta$ only in I_3.

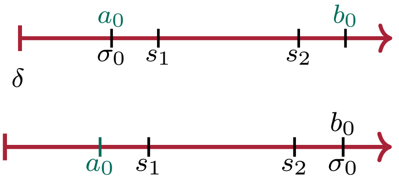


Figure 13: Visualisation of steps 1b and 1c.

The POWELL algorithm is

① Initialisation. Choose $\sigma_0 > 0$ and set $j := 0$.

- ⓐ If $G_1(\sigma) \geqslant \delta$ and $G_2(\sigma) \leqslant \beta$, stop and let $\sigma_P := \sigma_0$.
- ⓑ If $\sigma_0 \in I_1$, define $a_0 := \sigma_0$ and $b_0 := 2^\ell \sigma_9$, where ℓ is chosen minimally, such that $G_1(b_0) < \delta$. Go to step ②.
- ⓒ If $\sigma_0 \in I_3$, define $b_0 := \sigma_0$ and $a_0 = 2^{-\ell}\sigma_0$, where ℓ is chosen minimally, such that $G_2(a_0) > \beta$.

② Compute $\sigma_j := \frac{1}{2}(a_j + b_j)$.

- ⓐ If $\sigma_j \in I_2$, stop and set $\sigma_P := \sigma_j$.
- ⓑ If $\sigma_j \in I_1$, set $a_{j+1} := \sigma_j$ and $b_{j+1} := b_j$.
- ⓒ If $\sigma_j \in I_3$, set $a_{j+1} := a_j$ and $b_{j+1} := \sigma_j$.

③ Set $j \to j + 1$ and go to step ②.

Remark 3.3.3 A typical choice is $\delta := 0.1$ and $\beta := 0.9$. The step size σ_P is computed in finitely many steps.

Assuming (ALC) and (AGL), one can show that σ_E, σ_A and σ_P are efficient.

3.4 The gradient method (steepest descent)

Consider the following algorithm.

1. Initialise. Choose $x^0 \in \mathbb{R}^n$, $\varepsilon > 0$ and set $k := 0$.
2. If $|\nabla f(x^k)| < \varepsilon$, then stop.
3. Compute $d^k := -\nabla f(x^k)$ and choose an efficient step size σ_k (e.g. with Powell or Armijo). Define $x^{k+1} = x^k + \sigma_k d^k$ and set $k \to k+1$ and return to 2.

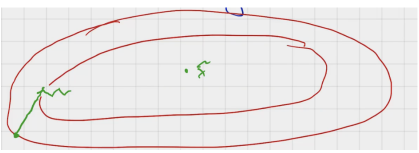


Figure 14: **TODO**

Remark 3.4.1 After initially fast decrease, one observes slow convergence especially for functions with stretched-out valleys. We e.g. have $0 = \varphi'(\sigma_E) = \nabla f(\underbrace{x^k + \sigma_E d^k}_{=x^{k+1}})^\mathsf{T} d^k = -(d^{k+1})^\mathsf{T} d^k$, i.e. $d^{k+1} \perp d^k$, which leads to the slow convergence detailed above.

They way out is to respect level sets, i.e. to account for the curvature of f, which is encoded in f''.

Lemma 3.4.2 (Origin of the term steepest descent)
We have $-\frac{\nabla f(x)}{|\nabla f(x)|} = \arg\min_{|d|=1} \nabla f(x)^\mathsf{T} d$.

Proof. For $d \in \mathbb{R}^n$ with $|d| = 1$, we have

$$\nabla f(x)^\mathsf{T} d \overset{\text{CS}}{\geqslant} -|\nabla f(x)||d| = -|\nabla f(x)|.$$

For $d = -\frac{\nabla f(x)}{|\nabla f(x)|}$ we get

$$\nabla f(x)^\mathsf{T} d = -|\nabla f(x)|^2. \qquad \square$$

3.5 The damped Newton method

1. Choose $x^0 \in \mathbb{R}^n$, $\varepsilon > 0$ and set $k := 0$.
2. If $|\nabla f(x^k)| < \varepsilon$, then stop.
3. Compute d^k as the solution of

$$f''(x^k) d^k = -\nabla f(x^k)$$

and an efficient step size σ_k. Define $x^{k+1} = x^k + \sigma_k d^k$ and set $k \to k+1$ and return to 2.

We compare the Newton direction to the direction of steepest descent.

Interpretation of the Newton direction

Define $A := f''(x)$ and assume that A is symmetric and positive definite. Then $\langle x, y \rangle_A := x^\mathsf{T} A y$ is a scalar product, which induces the norm $|x|_A = \sqrt{x^\mathsf{T} A x}$.

We can now prove a result analogous to lemma 3.4.2, which shows that the Newton direction is the steepest descent in the norm induced by A.

Lemma 3.5.1 (Newton direction is optimal in induced norm)
We have $\bar{d} := -\frac{A^{-1}\nabla f(x)}{|A^{-1}\nabla f(x)|_A} = \arg\min_{|d|_A=1} \nabla f(x)^\mathsf{T} d$.

Proof. For $d \in \mathbb{R}^n$ with $|d|_A = 1$ we have

$$\nabla f(x)^\mathsf{T} d = \langle A^{-1}\nabla f(x), d \rangle_A \overset{\text{CS}}{\geqslant} -|A^{-1}\nabla f(x)|_A |d|_A = -|A^{-1}\nabla f(x)|_A.$$

Finally, we have

$$\begin{aligned}\nabla f(x)^\mathsf{T}\bar{d} &= -\frac{\nabla f(x)^\mathsf{T} A^{-1}\nabla f(x)}{|A^{-1}\nabla f(x)|_A} \\ &= -\frac{\nabla f(x)^\mathsf{T} A^{-1}\nabla f(x)}{\sqrt{\nabla f(x)^\mathsf{T} A^{-1}\cancel{AA^{-1}}\nabla f(x)}} = -|A^{-1}\nabla f(x)|_A. \qquad \square\end{aligned}$$

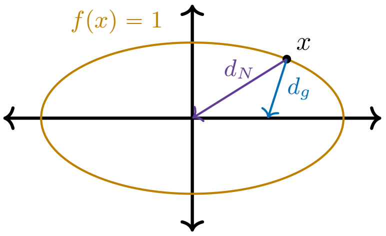


Figure 15: If we do gradient descent from point x on, the anti-gradient search direction $d_g = -\nabla f(x)$ (blue), which is orthogonal to the yellow isolines of f, is not optimal when trying to reach the minimum in the origin. In contrast, the purple Newton direction is $d_N = -f''(x)^{-1}\nabla f(x) = -H^{-1}Hx = -x$, which is a better search direction.

Example 3.5.2 (Quadratic function)
Consider $f(x) := \frac{1}{2}x^\mathsf{T} Hx$ with e.g. $H = \left(\begin{smallmatrix} b & 0 \\ 0 & 1 \end{smallmatrix}\right)$ for $b \in (0, 1]$, whose unique minimum (H positive definite, so f is strictly convex) is the origin. Then the condition of H is equal to $\frac{1}{b}$, so the problem becomes ill-conditioned small b. The level sets of f are ellipses, whose half axes are parallel to the axes.

From the figure, where $b = \frac{1}{4}$, we can see that with an exact step size, the Newton method for quadratic problems converges in one step.

If we instead consider $H = \operatorname{diag}(1, b)$ for $b \in (0, 1]$ and start in $x^0 = (b, 1)^\mathsf{T}$, the iterates are $x_1^{(k)} = (-1)^k b \left(\frac{1-b}{1+b}\right)^k$ and $x_2^{(k)} = \left(\frac{1-b}{1+b}\right)^k$, which converges very slowly. ◇

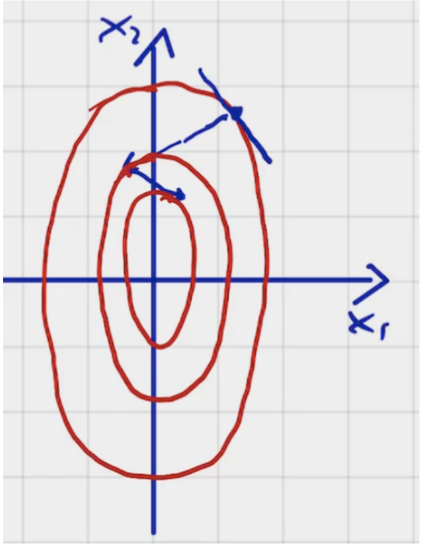


Figure 16: The second case from above.

Remark 3.5.3 If (AHP) holds, one can show that after finitely many steps $\bar{k}$, $\sigma_k = 1$ for all $k \geqslant \bar{k}$, so from $k = \bar{k}$, the convergence is quadratic if f'' is Lipschitz continuous, else it is super linear.

If we replace $f''(x^k)$ with $f''(x^0)$ for all k, we still get linear convergence. If we recompute $f''(x^k)$ after every n steps, we get super linear convergence.

(If we have an complicated objective function, we can replace derivatives with difference quotients and still achieve super linear convergence.)

25.05.2020

3.6 Variable metric and quasi-Newton methods

We want to account for curvature information (f'') without having to compute the second derivative, because it might be very expensive.

Consider the following general algorithm.

1. Choose $x^{(0)} \in \mathbb{R}^n$, $\varepsilon > 0$ and set $k := 0$.
2. If $|\nabla f(x)| < \varepsilon$, stop.

③ Compute the positive definite matrix $A^{(k)}$ and the search direction $d^{(k)} := -(A^{(k)})^{-1}\nabla f(x^{(k)})$ and an efficient step size σ_k. Set $x^{(k+1)} = x^{(k)} + \sigma_k d^{(k)}$, $k = k+1$ and go to step ②.

Special cases of this algorithm are the steepest descent ($A^{(k)} = I$) and the damped Newton method ($A^{(k)} = f''(x^{(k)})$).

If $A^{(k)}$ is uniformly positive definite, then $d^{(k)}$ are strictly gradient related, so the convergence result applies.

Quasi-Newton methods

We would like to construct $(A^{(k)})_{k\in\mathbb{N}}$ such that $A^{(k)}$ are positive definite for all $k \in \mathbb{N}$, the transition $A^{(k)} \to A^{(k+1)}$ is computationally cheap and that $A^{(k)} \approx f''(x^{(k)})$.

Example 3.6.1 (Motivation: Quadratic function)
Consider $f(x) = \frac{1}{2}x^\mathsf{T}Hx + b^\mathsf{T}x$, where H is positive definite. The associated quadratic minimisation problem has a unique solution $\tilde{x} = -H^{-1}b$.

We have

$$\begin{aligned} f''(x^{(k+1)})(x^{(k+1)} - x^{(k)}) &= H(x^{(k+1)} - x^{(k)}) + b - b \\ &= \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}). \end{aligned} \tag{6}$$

Assume that $H \in \mathbb{R}^{n\times n}$ is unknown but we know $\nabla f(x^{(k)})$, $x^{(k)}$ and assume that $x^{(k+1)} - x^{(k)}$ are linearly independent for $k \in \{0, \dots, n-1\}$.

Define the shortcuts $s^{(k)} := x^{(k+1)} - x^{(k)}$ and $y^{(k)} := \nabla f(x^{(k+1)}) - \nabla f(x^{(k)})$ and let $S = (s^{(0)}, \dots, s^{(n-1)})$, which is regular, as $x^{(k+1)} - x^{(k)}$ are linearly independent.

Thus (6) becomes $HS = (y^{(0)}, \dots, y^{(n-1)})$, i.e. $H = (y^{(0)}, \dots, y^{(n-1)})S^{-1}$, which is uniquely defined. ◇

To compute $A^{(k+1)}$, we demand that

$$A^{(k+1)}s^{(k)} = y^{(k)}. \tag{Quasi-Newton equation}$$

It gets the name from being the Newton equation if f is a quadratic function.

We now compute $A^{(k+1)}$ from $A^{(k)}$ in two steps:

① Let

$$\tilde{A}^{(k)} := A^{(k)} - \frac{(A^{(k)}s^{(k)})(A^{(k)}s^{(k)})^\mathsf{T}}{(s^{(k)})^\mathsf{T}A^{(k)}s^{(k)}}.$$

This is called a symmetric rank-1 update.

For $x, y \in \mathbb{R}^n$, the matrix $xy^\mathsf{T} = (y_1x, \dots, y_nx)$ has rank one. If $A^{(k)}$ is positive definite and symmetric, $\tilde{A}^{(k)}$ is symmetric and positive semidefinite: symmetry is clear and for $x \in \mathbb{R}^n$ we have

$$\begin{aligned} x^\mathsf{T}\tilde{A}x &= x^\mathsf{T}Ax - \frac{x^\mathsf{T}(As)(As)^\mathsf{T}x}{s^\mathsf{T}As} = |x|_A^2 - \frac{\langle x, s\rangle_A^2}{|s|_A^2} \\ &\overset{\text{CS}}{\geqslant} = |x|_A^2 - \frac{|x|_A^2|s|_A^2}{|s|_A^2} = 0. \end{aligned}$$

Finally, we have $\tilde{A}^{(k)}s^{(k)} = 0$:

$$\tilde{A}s = As - \frac{(As)(As)^\mathsf{T} s}{s^\mathsf{T} As} = As - \frac{(As)\cancel{s^\mathsf{T} As}}{\cancel{s^\mathsf{T} As}} = 0.$$

② $A^{(k+1)}$ is now obtained from $\tilde{A}^{(k)}$ by another rank one update:

$$A^{(k+1)} = \tilde{A}^{(k)} + \gamma_k w^{(k)}(w^{(k)})^\mathsf{T},$$

where $\gamma^{(k)}$ and $w^{(k)}$ are chosen such that $A^{(k+1)}$ is positive definite and satisfies (Quasi-Newton equation).

We have

$$A^{(k+1)}s^{(k)} = \tilde{A}^{(k)}s^{(k)} + \gamma_k w^{(k)}(w^{(k)})^\mathsf{T} s^k = \gamma_k w^{(k)}(w^{(k)})^\mathsf{T} s^k \overset{!}{=} y^{(k)}$$

We choose $w^{(k)} = y^{(k)}$ and $\gamma_k = \frac{1}{(w^{(k)})^\mathsf{T} s^k}$.

Remark 3.6.2 One can show that $A^{(k+1)}$ is positive definite if $A^{(k)}$ is positive definite and $(s^{(k)})^\mathsf{T} A^{(k+1)} s^{(k)} = (s^{(k)})^\mathsf{T} y^{(k)} > 0$ holds (cf. HA6.3).

Example 3.6.3 (Quadratic function)
Consider $f(x) := \frac{1}{2}x^\mathsf{T} Hx$. Then

$$\begin{aligned}(s^{(k)})^\mathsf{T} y^{(k)} &= (x^{(k+1)} - x^{(k)})^\mathsf{T}(\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) \\ &= (x^{(k+1)} - x^{(k)})^\mathsf{T} H(x^{(k+1)} - x^{(k)}) > 0\end{aligned}$$

if $s^{(k)} \neq 0$. ◇

Altogether, we obtain the quasi-Newton (or rank-2 or BFGS) update

$$A^{(k+1)} = A^{(k)} - \frac{A^{(k)}s^{(k)}(A^{(k)}s^{(k)})^\mathsf{T})}{(s^{(k)})^\mathsf{T} A^{(k)}s^{(k)}} + \frac{y^{(k)}(y^{(k)})^\mathsf{T}}{(y^{(k)})^\mathsf{T} s^{(k)}} \tag{7}$$

named after Broyden, Fletcher, Goldfarb and Shanno, which can be also written as

$$A^{(k+1)} := \left(I - \frac{y^{(k)}(s^{(k)})^\mathsf{T}}{\langle y^{(k)}, s^{(k)}\rangle}\right) A^{(k)} \left(I - \frac{s^{(k)}(y^{(k)})^\mathsf{T}}{\langle y^{(k)}, s^{(k)}\rangle}\right) + \frac{y^{(k)}(y^{(k)})^\mathsf{T}}{\langle y^{(k)}, s^{(k)}\rangle}.$$

It fulfills (Quasi-Newton equation) (HA 6.3):

$$\begin{aligned}A^{(k+1)}s^{(k)} &= \left(I - \frac{y^{(k)}(s^{(k)})^\mathsf{T}}{\langle y^{(k)}, s^{(k)}\rangle}\right) A^{(k)} \left(s^{(k)} - \frac{s^{(k)}\cancel{(y^{(k)})^\mathsf{T} s^{(k)}}}{\cancel{\langle y^{(k)}, s^{(k)}\rangle}}\right) \\ &\quad + \frac{y^{(k)}\cancel{(y^{(k)})^\mathsf{T} s^{(k)}}}{\cancel{\langle y^{(k)}, s^{(k)}\rangle}} \\ &= \left(I - \frac{y^{(k)}(s^{(k)})^\mathsf{T}}{\langle y^{(k)}, s^{(k)}\rangle}\right) A^{(k)} \underbrace{(s^{(k)} - s^{(k)})}_{=0} + y^{(k)} = y^{(k)}.\end{aligned}$$

If f is strictly convex and A_0 symmetric positive definite, then $A^{(k)}$ is symmetric positive definite for all $k \in \mathbb{N}$ (HA 6.3): By lemma 2.1.9, we have $\langle y^{(k)}, s^{(k)} \rangle > 0$ for $s^{(k)} \neq 0$ and for $z \neq 0$

$$\begin{aligned}\langle z, A^{(k+1)} z \rangle &= \left\langle z, \left(I - \frac{y^{(k)}(s^{(k)})^\top}{\langle y^{(k)}, s^{(k)} \rangle}\right) A^{(k)} \left(I - \frac{s^{(k)}(y^{(k)})^\top}{\langle y^{(k)}, s^{(k)} \rangle}\right) z \right\rangle \\ &\quad + \left\langle z, \frac{y^{(k)}(y^{(k)})^\top}{\langle y^{(k)}, s^{(k)} \rangle} z \right\rangle \\ &= \left\langle \left(I - \frac{s^{(k)}(y^{(k)})^\top}{\langle y^{(k)}, s^{(k)} \rangle}\right) z, A^{(k)} \left(I - \frac{s^{(k)}(y^{(k)})^\top}{\langle y^{(k)}, s^{(k)} \rangle}\right) z \right\rangle \\ &\quad + \frac{\|(y^{(k)})^\top z\|^2}{\langle y^{(k)}, s^{(k)} \rangle} \\ &= \left\| \left(I - \frac{s^{(k)}(y^{(k)})^\top}{\langle y^{(k)}, s^{(k)} \rangle}\right) z \right\|_{A^{(k)}}^2 + \frac{\|(y^{(k)})^\top z\|^2}{\langle y^{(k)}, s^{(k)} \rangle}.\end{aligned}$$

The symmetry is clear.

BFGS-method for quadratic problems

Consider $f(x) := \frac{1}{2} x^\top H x + b^\top x$, where H is symmetric and positive definite.

The BFGS algorithm for this unconstrained quadratic problem is

1. Choose $x^{(0)} \in \mathbb{R}^n$, $A^{(0)} \in \mathbb{R}^{n \times n}$ positive definite, $\varepsilon > 0$ and set $k := 0$.
2. If $|\nabla f(x)| < \varepsilon$, stop.
3. Compute $d^{(k)} = -(A^{(k)})^{-1} \nabla f(x^{(k)})$, an exact step size σ_k and set $x^{(k+1)} = x^{(k)} + \sigma_k d^k$, $s^{(k)} = x^{(k+1)} - x^{(k)}$ and $y^{(k+1)} = \nabla f(x^{(k+1)}) - f(x^{(k)})$ and preform (7). Set $k \to k+1$ and go to step 2.

Definition 3.6.4 (H-Orthogonality)
Let $H \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix. Then directions $d^{(0)}, \ldots, d^{(k)}$ for $k < n$ are conjugate or H-orthogonal if $d^{(i)} \neq 0$ and $(d^{(i)})^\top H d^{(j)} = 0$ for all $0 \leqslant i < j \leqslant k$.

H-orthogonal

Theorem 3.6.1: BFGS method for quadratic problems

Let $H \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix. Then, the BFGS-method generated H-orthogonal search directions $d^{(k)}$. The minimum is found in $m \leqslant n$ steps. If $m = n$, then $A^{(n)} = H$.

Remark 3.6.5 The BFGS method is not suitable for quadratic problems, as it is much faster to solve $H\tilde{x} + b = 0$ by a Cholesky decomposition.

When using BFGS for nonlinear problems, it is numerically more efficient to use a update for $(A^{(k)})^{-1}$. We obtain super linear convergence if f'' is Lipschitz continuous. In general we get

$$\frac{|(A^{(k)} - f''(\tilde{x}) d^{(k)}|}{|d^{(k)}|} \xrightarrow{k \to \infty} 0.$$

Remark 3.6.6 (SHERMAN-MORRISON-WOODBURY formula) For the A^{-1} update we use

$$(A + uv^\mathsf{T})^{-1} = A^{-1} - \frac{A^{-1}uv^\mathsf{T}A^{-1}}{1 + \langle A^{-1}u, v \rangle}.$$

We first show that for $u, v \in \mathbb{R}^n$ and an invertible matrix $A \in \mathbb{R}^{n\times n}$, the matrix $A + uv^\mathsf{T}$ is invertible if and only if $1 + \langle A^{-1}u, v \rangle \neq 0$.

Proof. For $u = 0$ or $v = 0$ the statement is clear. As $A + uv^\mathsf{T} = A(I + A^{-1}uv^\mathsf{T})$ holds, $A + uv^\mathsf{T}$ is invertible if and only if $I + A^{-1}uv^\mathsf{T}$ is, which is the case if and only if $\det(I + A^{-1}uv^\mathsf{T}) \neq 0$ holds. For $\tilde{u} := A^{-1}u$ we show

$$\det(I + \tilde{u}v^\mathsf{T}) = 1 + \langle \tilde{u}, v \rangle$$

Let $(\lambda_j)_{j=1}^n$ be the eigenvalues of $\tilde{u}v^\mathsf{T}$. Then $(1+\lambda_j)_{j=1}^n$ are the eigenvalues of $I + \tilde{u}v^\mathsf{T}$. The matrix $\tilde{u}v^\mathsf{T}$ has rank one and thus has the eigenvalue zero with multiplicity $n - 1$: Thus

$$\det(I + \tilde{u}v^\mathsf{T}) = \prod_{i=1}^n (1 + \lambda_i) = 1 + \sum_{i=1}^n \tilde{u}_i v_i = 1 + \langle \tilde{u}, v \rangle. \qquad \square$$

We can now show the desired formula.

Proof. We have $(A + uv^\mathsf{T})^{-1} = (I + A^{-1}uv^\mathsf{T})^{-1}A^{-1}$ and thus it suffices to show

$$(I + \tilde{u}v^\mathsf{T})^{-1} = I - \frac{\tilde{u}v^\mathsf{T}}{1 + \langle \tilde{u}, v \rangle} \tag{8}$$

with $\tilde{u} := A^{-1}u$. Multiplying both sides of (8) by $I + \tilde{u}v^\mathsf{T}$ yields the identity matrix on the left and on the right

$$\begin{aligned}
\left(I - \frac{\tilde{u}v^\mathsf{T}}{1 + \langle \tilde{u}, v \rangle}\right)(I + \tilde{u}v^\mathsf{T}) &= I + \tilde{u}v^\mathsf{T} - \frac{\tilde{u}v^\mathsf{T}(I + \tilde{u}v^\mathsf{T})}{1 + \langle \tilde{u}, v \rangle} \\
&= I + \tilde{u}v^\mathsf{T} - \frac{\tilde{u}v^\mathsf{T} + \tilde{u}v^\mathsf{T}\tilde{u}v^\mathsf{T}}{1 + \langle \tilde{u}, v \rangle} \\
&= I + \frac{\cancel{\tilde{u}v^\mathsf{T}} + \tilde{u}v^\mathsf{T}\langle \tilde{u}, v \rangle \cancel{- \tilde{u}v^\mathsf{T}} - \tilde{u}v^\mathsf{T}\tilde{u}v^\mathsf{T}}{1 + \langle \tilde{u}, v \rangle} \\
&= I + \frac{\tilde{u}v^\mathsf{T}\langle \tilde{u}, v \rangle - \tilde{u}v^\mathsf{T}\tilde{u}v^\mathsf{T}}{1 + \langle \tilde{u}, v \rangle} \\
&= I + \frac{\cancel{\tilde{u}v^\mathsf{T}\langle \tilde{u}, v \rangle - \langle \tilde{u}, v \rangle \tilde{u}v^\mathsf{T}}}{1 + \langle \tilde{u}, v \rangle} = I. \qquad \square
\end{aligned}$$

3.7 Conjugate gradient (CG) methods

04.06.2020

In applications, $A^{(k)}$ might be too large to store efficiently. We thus want to generate H-orthogonal directions without such a quasi-NEWTON A^k update.

The following lemma shows that if we have k H-orthogonal direction and the standard descent algorithm with an exact step size for a quadratic problem, then we not only have that the iterates x_k minimise along the line of the search direction, but also in the subspace spanned by the previous search directions.

Lemma 3.7.1 (Structural properties for quadratic problems)
Let $d^{(0)}, \dots, d^{(n-1)}$ by H-orthogonal directions. For any $x^{(0)} \in \mathbb{R}^n$, the iteration $x^{(k+1)} + \sigma_k d^k$, where $\sigma_k = -\frac{\nabla f(x^{(k)})^\intercal d^k}{(d^{(k)})^\intercal H d^{(k)}}$ is the exact step size, compute the solution $x^{(n)} = -H^{-1}b$ of (QU) in at most n steps.

Proof. Clearly, $d^{(0)}, \dots, d^{(n-1)}$ are linearly independent. Then choose σ_k such that

$$\tilde{x} - x^{(0)} = \sum_{i=0}^{n-1} \sigma_i d^{(i)}. \tag{9}$$

As this is exactly the iteration from above, it suffices to show that σ_i is the exact step size from above.

Multiplying (9) with $(d^{(k)})^\intercal H$ yields

$$(d^{(k)})^\intercal H(\tilde{x} - x^{(0)}) = \sum_{i=0}^{n-1} \sigma_i (d^{(k)})^\intercal H d^{(i)} = \sigma_k (d^{(k)})^\intercal H d^{(k)}$$

by the H-orthogonality of $d^{(0)}, \dots, d^{(n-1)}$, yielding

$$\sigma_k = \frac{(d^{(k)})^\intercal H(\tilde{x} - x^{(0)})}{(d^{(k)})^\intercal H d^{(k)}}.$$

As $H\tilde{x} + b = 0$, we have

$$\begin{aligned}\sigma_k &= -\frac{(d^{(k)})^\intercal (Hx^{(0)} + b)}{(d^{(k)})^\intercal H d^{(k)}} \\ &= -\frac{(d^{(k)})^\intercal (H(x^{(1)} - \sigma_0 d^{(0)}) + b)}{(d^{(k)})^\intercal H d^{(k)}} = -\frac{(d^{(k)})^\intercal (Hx^{(1)} + b)}{(d^{(k)})^\intercal H d^{(k)}}\end{aligned}$$

again by the H-orthogonality, as $k \neq 0$. Continuing in this fashion we obtain

$$\sigma_k = -\frac{(d^{(k)})^\intercal (Hx^{(k)} + b)}{(d^{(k)})^\intercal H d^{(k)}},$$

which was to show. □

Corollary 3.7.2
We have $x^{(k)} = \arg\min_{\sigma \in \mathbb{R}} f(x^{k-1} + \sigma d^{(k-1)})$ and on $x_0 + V_k$, with $V_k := \operatorname{span}(d^{(0)}, \dots, d^{(k-1)}$ and $(d^{(i)})^\intercal \perp \nabla f(x^{(k)})$ for $i < k$.

Proof. It suffices to show $(d^{(i)})^\intercal \perp \nabla f(x^{(k)})$ for $i < k$, which we do inductively.

First, let $k = i + 1$. Then

$$\begin{aligned}(d^{(i)})^\intercal \nabla f(x^{(i+1)}) &= (d^{(i)})^\intercal (Hx^{(i+1)} + b) \\ &= (d^{(i)})^\intercal (Hx^{(i)} + b) + \sigma_i (d^{(i)})^\intercal H d^{(i)} \\ &= (d^{(i)})^\intercal (Hx^{(i)} + b) - (d^{(i)})^\intercal \nabla f(x^{(i)} + b) = 0.\end{aligned}$$

For $k \geqslant i + 1$ not that

$$\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) = H(x^{(k+1)} - x^{(k)}) = \sigma_k H d^{(k)}. \qquad \square$$

Consider the following CG-algorithm.

① choose $x^{(0)} \in \mathbb{R}^n, \varepsilon > 0$ and set $k := 0$ and $d^{(0)} = -H(x^{(0)} + b)$.

② If $|\nabla f(x^{(k)})| \leqslant \varepsilon$, stop.

③ Compute $\sigma_k = \frac{|\nabla f(x^{(k)}|^2}{|d^{(k)}|_H^2}$ and set $x^{(k+1)} = x^{(k)} + \sigma_k d^{(k)}$. We have

$$\nabla f(x^{(k+1)}) = Hx^{(k+1)} + b = \nabla f(x^{(k)}) + \sigma_k H d^k.$$

Compute $\beta_k := \frac{|\nabla f(x^{(k+1)})|^2}{|\nabla f(x^{(k)})|^2}$ and set $d^{(k+1)} = -\nabla f(x^{(k+1)}) + \beta_k d^{(k)}$.

④ Set $k \to k+1$ and return to step ②.

Remark 3.7.3 Note that $\sigma_k = \frac{|\nabla f(x^{(k)}|^2}{|d^{(k)}|_H^2}$ is exact, as

$$\begin{aligned} \sigma_E &= -\frac{\nabla f(x^{(k)})^\mathsf{T} d^{(k)}}{|d^{(k)}|_H^2} \\ &= -\frac{\nabla f(x^{(k)})^\mathsf{T}(-\nabla f(x^{(k)}) + \beta_{k-1} d^{(k-1)})}{|d^{(k)}|_H^2} \\ &= -\frac{\nabla f(x^{(k)})^\mathsf{T}(-\nabla f(x^{(k)})}{|d^{(k)}|_H^2}, \end{aligned}$$

as $\nabla f(x^{(k)}) \perp d^{(k-1)}$ by corollary 3.7.2.

Theorem 3.7.1: Properties of the CG method

As long as $\nabla f(x^{(k-1)}) \neq 0$, we have

① $d^{(k-1)} \neq 0$,

② We have

$$\begin{aligned} V_k &= \operatorname{span}(\nabla f(x^{(0)}), H\nabla f(x^{(0)}), \dots, H^{k-1}\nabla f(x^{(0)}) \\ &= \operatorname{span}(\nabla f(x^{(0)}), \dots, \nabla f(x^{(k-1)})) \\ &= \operatorname{span}(d^{(0)}, \dots, d^{(k-1)}), \end{aligned}$$

③ The directions $d^{(0)}, \dots, d^{(k)}$ are H-orthogonal,

④ $f(x^{(k)}) = \min_{z \in V_k} f(x^{(0)} + z)$.

Proof. In the Script □

Convergence analysis for (QU)

Recall that $\kappa(H) = \|H\|\|H^{-1}\|$ is the condition of the matrix H. If H is symmetric positive definite with eigenvalues $0 < \lambda_1 < \dots, \lambda_n$, we have $\kappa(H) = \frac{\lambda_n}{\lambda_1} \geqslant 1$. Thus if λ_1 is very small, the matrix will become ill-conditioned (cf. above).

condition

For steepest descent with exact step size, i.e. $\sigma = \sigma_E$, for (QU) one can show that

$$|x^{(k+1)} - \tilde{x}|_H \leqslant \Bigg(\underbrace{\frac{\kappa(H) - 1}{\kappa(H) + 1}}_{<1}\Bigg)^k |x^{(0)} - \tilde{x}|_H$$

Theorem 3.7.2: CG for (QU)

For the CG method applied to (QU) there holds

$$|x^{(k)} - \tilde{x}|_H \leqslant 2 \left(\frac{\sqrt{\kappa(H)} - 1}{\sqrt{\kappa(H)} + 1} \right)^k |x^{(0)} - \tilde{x}|_H$$

Remark 3.7.4 For $H = \operatorname{diag}(1, b)$, we have

$$\frac{\sqrt{\kappa(H)} - 1}{\sqrt{\kappa(H)}} = \frac{\sqrt{b} - 1}{\sqrt{b}},$$

so for $b = \frac{1}{100}$, we have $\frac{1-b}{1+b} = 0.\overline{9801}$ but $\frac{1-\sqrt{b}}{1+\sqrt{b}} = 0.\overline{81}$

Preconditioning of CG-algorithms

The speed of convergence depends on the condition of the matrix. The closer the condition number is to one, the faster the convergence. Is there a way to alter the system matrix such that the condition is improved without changing the problem such that it becomes to expensive to solve?

Yes, and the idea is to modify H such that the isolines of the modified $\tilde{f}$ become close to circles. Let $B \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix and replace $Hx = -b$ with $\overline{H}\overline{x} = -b$, where $\overline{H} := H \cdot B$ and $\overline{x} := B^{-1}x$. Then

$$\langle x, HBy \rangle_B = x^\mathsf{T} BHBy = (HBx)^\mathsf{T} By = \langle HBx, y \rangle_B,$$

so HB is self-adjoint with respect to $\langle \cdot, \cdot \rangle_B$.

The main idea is to replace $\langle \cdot, \cdot \rangle$ with $\langle \cdot, \cdot \rangle_B$ in the CG algorithm, such that one obtains

$$|\tilde{x} - x^{(k)}|_H \leqslant 2 \left(\frac{\sqrt{\kappa(HB)} - 1}{\sqrt{\kappa(HB)} + 1} \right)^k |\tilde{x} - x^{(0)}|_H.$$

Thus the main task of preconditioning is to find a B such that evaluation of By is cheap and $\kappa(HB)$ is small. Typical choices are $B = D^{-1}$, where $D = \operatorname{diag}(H)$ or an incomplete Cholesky decomposition of H.

3.8 Trust region method

08.06.2020

Up to now, we have computed a search direction d^k and a step size σ_k (line search) and we used the update $x^{(k+1)} = x^{(k)} + \sigma_k d^{(k)}$.

The new idea is now to

- use a local model f_k of f, e.g. $f_k = f(x^{(k)}) + \nabla f(x^{(k)})^\mathsf{T} d$ or $f_k = f(x^{(k)}) + \nabla f(x^{(k)})^\mathsf{T} d + \frac{1}{2} d^\mathsf{T} f''(x^{(k)}) d$,
- take radius $\rho_k > 0$ and consider the trust region $B_{\rho_k}(x^{(k)})$,
- compute $d^{(k)}$ as a global solution to

trust region

$$\min_{|d| \leqslant \rho_k} f_k(d), \tag{10}$$

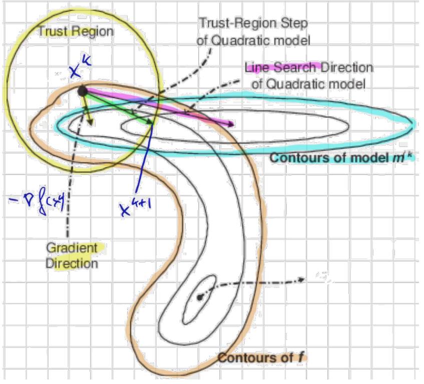


Figure 17: Basic idea of the trust region method.

- update $x^{(k+1)} = x^{(k)} + d^{(k)}$.

Remark 3.8.1 If we choose a linear model, then the solution of (10) is $d^{(k)} = -\rho_k \frac{\nabla f(x^{(k)})}{|\nabla f(x^{(k)})|}$, with is standard gradient descent, which can suffer from slow convergence.

For the model f_k (where $x^{(k)} \in \mathbb{R}^n$ and $\rho_k > 0$ are given) we require that

- $f_k(0) = f(x^{(k)})$
- for a solution $d^{(k)}$ of (10) we have that $f_k(d^{(k)}) = f(x^{(k)})$ implies that $\nabla f(x^{(k)}) = 0$.

Example 3.8.2
Assume that $f \in \mathcal{C}^2$ and choose the quadratic model $f_k(d) = f(x^{(k)}) + \nabla f(x^{(k)})^\mathsf{T} d + \frac{1}{2} d^\mathsf{T} f''(x^{(k)}) d$. Then $f_k(0) = f(x^{(k)})$. If $f_k(d^{(k)}) = f(x^{(k)})$, then

$$f_k(d^{(k)}) = f(x^{(k)}) \leqslant f_k(d)$$

for all $d \in B_{\rho_k}(0)$ and thus $\tilde{d} = 0$ is a local solution to

$$\min_{|d| \leqslant \rho_k} \underbrace{\nabla f(x^{(k)})^\mathsf{T} d + \frac{1}{2} d^\mathsf{T} f''(x^{(k)}) d}_{=:F(d)}$$

in the interior of the admissable set $B_{\rho_k}(0)$, so it is a solution of $0 = \nabla F(\tilde{d})$. We have

$$\nabla F(\tilde{d}) = \nabla f(x^{(k)}) + f''(x^{(k)}) \tilde{d} = \nabla f(x^{(k)})$$

and thus $\nabla f(x^{(k)}) = 0$, as required. ◇

Remark 3.8.3 (How do we choose the radius?)
We compute

$$r_k := \frac{f(x^{(k)}) - f(x^{(k)} + d^{(k)})}{f(x^{(k)}) - f_k(x^{(k)} + d^{(k)})}, \tag{11}$$

which is the ratio of the real descent and the model descent. Now choose $0 < \delta_1 < \delta_2 < 1$. We set $x^{(k+1)} + x^{(k)} + d^{(k)}$ and if

$$r_k \begin{cases} \in (\delta_1, \delta_2), & \text{we keep } \rho_k, \\ \geqslant \delta_2, & \text{we increase } \rho_k, \\ \leqslant \delta_1, & \text{we decrease } \rho_k \end{cases}$$

Trust region Newton methods

Consider the quadratic model

$$f_k(d) = f(x^{(k)}) + \nabla f(x^{(k)})^\mathsf{T} d + \frac{1}{2} d^\mathsf{T} f''(x^{(k)}) d$$

and

$$\min_{|d| \leqslant \rho_k} f_k(d). \tag{12}$$

If this were an unconstrained problem, it would coincide with the Newton method, however that problem only has a solution if $f''(x^{(k)})$ is positive definite, while (12) always has a solution.

In the following we consider

$$\min_{|d|\leqslant\rho} \underbrace{c + b^\mathsf{T} d + \frac{1}{2} d^\mathsf{T} A d}_{=:\Phi(d)} \tag{13}$$

and let $g(d) := \frac{1}{2}(|d|^2 - |\rho|^2)$, so $|d| \leqslant \rho$ if and only if $g(d) \leqslant 0$. We also have $\nabla g = d$.

We introduce the Lagrangian

$$\Phi_\lambda(d) := \Phi(d) + \lambda g(d) = -\frac{1}{2}\rho^2 + c + b^\mathsf{T} d + \frac{1}{2} d^\mathsf{T}(A + \lambda I)d.$$

Lemma 3.8.4 (see the next chapter)
Let $\tilde{d} \in \overline{B}_\rho(0)$ and $\lambda \geqslant 0$ such that $\tilde{d}$ is a (strict) global minimum of Φ_λ. If the complementarity condition $\lambda g(\tilde{d}) = 0$ holds, then $\tilde{d}$ is a (strict) global minimum of (13).

Proof. For all $d \in B_\rho(0)$, we have

$$\Phi(\tilde{d}) = \Phi_\lambda(\tilde{d}) \leqslant \Phi_\lambda(d) = \Phi(d) + \underbrace{\lambda}_{\geqslant 0} \underbrace{g(d)}_{\leqslant 0} \leqslant \Phi(d).$$ □

Lemma 3.8.5 (See KKT conditions in next chapter)
Let $\tilde{d} \in \mathbb{R}^n$ be a global minimum of (13). Then there exists exactly one Lagrange multiplier $\lambda \geqslant 0$ such that $\nabla\Phi_\lambda(\tilde{d}) = \nabla\Phi(\tilde{d}) + \lambda \underbrace{\tilde{d}}_{=\nabla g} = 0$, $d^\mathsf{T}\Phi_\lambda''(\tilde{d})d = d^\mathsf{T}(A + \lambda I)d \geqslant 0$ for all $d \in \mathbb{R}^n$ and $\lambda g(\tilde{d}) = 0$.

Theorem 3.8.1: KKT conditions for the trust region method

A $\tilde{d} \in B_\rho(0)$ is a global solution to (13) if and only if there exists a $\lambda \geqslant 0$ such that

① $(A + \lambda I)\tilde{d} = -b$,

② $\lambda(|\tilde{d}| - \rho) = 0$, i.e. $|d| = \rho$ if $\lambda > 0$,

③ $A + \lambda I$ is positive semidefinite.

Proof. By the two lemmas above. □

For a stopping criterion, we show for $\Phi(d)$:

Lemma 3.8.6
Let $\tilde{d} \in B_\rho(0)$ by a global solution to (13). Then $\Phi(\tilde{d}) = c$ if and only if $b = 0$ and A is positive semidefinite.

Proof. " $\implies$ ": We have $\Phi(0) = c = \Phi(\tilde{d}) \leqslant \Phi(d)$, so 0 is a solution of (13). By theorem 3.8.1 ① we get $b = 0$ and by ② we get that $\lambda = 0$ and thus by ③, A is positive semidefinite.

" $\impliedby$ ": Then $\Phi(d) = c + \frac{1}{2}d^\mathsf{T}Ad$. As this is a convex objective, $\tilde{d} = 0$, which satisfies $\nabla\Phi(\tilde{d}) = 0$, and $\tilde{d}$ is a global minimum and we have $\Phi(\tilde{d}) = c$. □

Applying this to the quadratic model yields the following: if $d^{(k)}$ is a global solution of (12), then $f_k(d^{(k)}) = f^{x^{(k)}}$ if and only if $\nabla f(x^{(k)}) = 0$ and $f''(x^{(k)})$ is positive semidefinite.

Indeed, $f_k(d^{(k)}) = f(x^{(k)})$ is a useful stopping criterion.

Lemma 3.8.7 (Estimation of descent)
If $\tilde{d} \in \overline{B}_\rho(0)$ is the global solution of (13)*, then*

$$c - \Phi(d^{(k)}) \geqslant \frac{1}{2}|b| \min\left(\rho, \frac{|b|}{|A|}\right).$$

Proof. n.a. □

We can now formulate the trust region Newton algorithm: Given $0 < \delta_1 < \delta_2 < 1$, $\sigma_1 \in (0,1)$, $\sigma_2 > 1$ and $\sigma_0 > 0$,

1. Choose $x^{(0)} \in \mathbb{R}^n$,
2. Compute global solution $d^{(k)}$ of $\min_{|d| \leqslant \rho_k} f_k(d)$. If $f(x^{(k)}) = f_k(d^{(k)})$, then stop.
3. Compute r_k from (11). If $r_k \geqslant \delta_1$ (successful step), set $x^{(k+1)} = x^{(k)} + d^{(k)}$, compute $\nabla f(x^{(k+1)})$, $f''(x^{(k+1)})$ and update ρ_k:

   $$\text{if } r_k \begin{cases} \in [\delta_1, \delta_2), & \text{choose } \rho_{k+1} \in [\delta_1 \rho_k, \rho_k], \\ \geqslant \delta_2, & \text{choose } \rho_{k+1} \in [\rho_k, \delta_2 \rho_k], \end{cases}$$

   set $k \to k+1$ and go to 2.
4. If $r_k < \delta_1$ (not successful step), choose $\rho_{k+1} \in (0, \delta_1 \rho_k)$ and set $x^{(k+1)} = x^{(k)}$, $\nabla f(x^{(k+1)}) = \nabla f(x^{(k)})$, $f''(x^{(k+1)}) = f''(x^{(k)})$ and $k \to k+1$ and go to step 2.

4 Problems with constraints - theory

15.06.2020

We consider

$$\min_{x\in\mathbb{R}^n} f(x), \qquad \text{subject to} \begin{cases} c_i(x) = 0, & i \in E, \\ c_i(x) \geqslant 0, & i \in I \end{cases} \tag{14}$$

where $I, E \subset \mathbb{N}$ are disjoint index sets. The constraints $c_i(x) \overset{(\geqslant)}{=} 0$ are called (in)equality constraints. The admissable set is

admissable set

$$\Omega = \{x \in \mathbb{R}^n : c_i(x) = 0, i \in E, c_i(x) \geqslant 0, i \in I\},$$

so we can rewrite (14) as $\min_{x\in\Omega} f(x)$.

Remark 4.0.1 If the c_i are continuous, Ω is closed.

Definition 4.0.2 ((In)Active constraints, active set) Let $x \in \Omega$, i.e. admissable. Then $c_i(x)$, $i \in I$ is called active if $c_i(x) = 0$ and inactive if $c_i(x) > 0$. The active set is

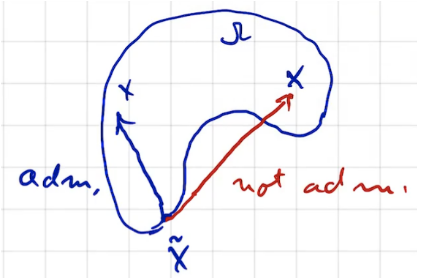
$$\mathcal{A}(x) := E \cup \{i \in I : c_i(x) = 0\}.$$

4.1 Tangent cone, linearised cone and regularity

Remark 4.1.1 (What we know already) Assume the admissible set Ω is convex and that $\hat{x}$ is a solution of (14). Take $x \in \Omega$, then $\hat{x}+t(x-\hat{x}) \in \Omega$ if $t \in [0,1]$. For small $t > 0$, we have

$$0 \leqslant \frac{f(\hat{x}+t(x-\hat{x})) - f(\hat{x})}{t} \xrightarrow{t\to 0} \nabla f(\hat{x})^\mathsf{T}(x-\hat{x})$$

for all $x \in \Omega$.



cone

Figure 18: If Ω is not convex, remark 4.1.1 no longer holds.

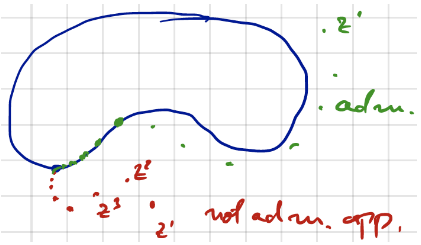


Figure 19: An example of a admissable approximation and an inadmissable approximation.

Definition 4.1.2 (Cone) A set $K \subset \mathbb{R}^n$ is a cone if $K \subset aK$ for all $a > 0$.

Definition 4.1.3 (Admissible approximation) Let $x \in \Omega$. Then the sequence $(x^{(n)})_{n\in\mathbb{N}}$ is called admissable approximation of x if $x^{(n)} \to x$ and $x^{(n)} \in \Omega$ for almost all $n \in \mathbb{N}$.

Definition 4.1.4 (tangent cone) A direction $d \in \mathbb{R}^n$ is a tangent to Ω in $x \in \Omega$ if there exists an admissable approximation $(x^{(k)})_{k\in\mathbb{N}}$ of x and a sequence $(t_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ converging to zero such that

$$\lim_{k\to\infty} \frac{x^{(k)} - x}{t_k} = d.$$

The tangent cone of Ω in x is

$$T_\Omega(x) := \{d \in \mathbb{R}^n : d \text{ is tangent to } \Omega \text{ in } x\}.$$

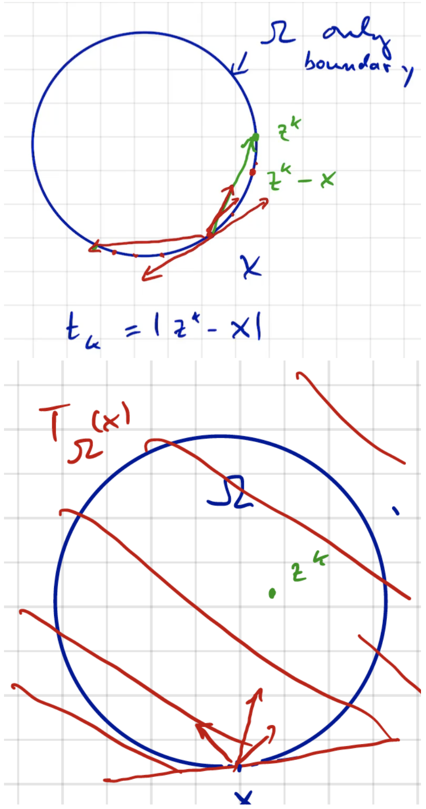


Figure 20: If Ω is the boundary of a circle, then the tangent cone at a point x is exactly what one would imagine. If Ω is disk, then the tangent cone is the whole space.

Example 4.1.5 (The tangent cone is a cone) Let $d \in T_\Omega(x)$. For $a > 0$ define $\tilde{t}_k := \frac{t_k}{a} > 0$. Then $\tilde{t}_k \to 0$. Then

$$\frac{z^{(k)} - x}{\tilde{t}_k} = a\frac{z^{(k)} - x}{t_k} \to ad,$$

so $ad \in T_\Omega(x)$. ◇

Remark 4.1.6 The tangent cone is closed and nonempty and it is convex if Ω is, too.

Example 4.1.7 Let $x \in \operatorname{int}(\Omega)$. Then there exists a $\varepsilon > 0$ such that $B_\varepsilon(x) \subset \Omega$. Any $x^{(k)} \in \{y \in \mathbb{R}^n : \|x - y\| = \varepsilon\}$ defines a tangent, so $T_\Omega(x) = \mathbb{R}^n$. ◇

The following theorem is a generalisation of 2.2.2 for non-convex Ω.

Theorem 4.1.1: Variational inequality - General case

Let $\hat{x} \in \Omega$ be a solution of (14) and $f \in \mathcal{C}^1$. Then $\nabla f(\hat{x})^\mathsf{T} d \geqslant 0$ holds for all $d \in T_\Omega(\hat{x})$.

Proof. Let $d \in T_\Omega(\hat{x})$. Then there exists a approximating sequence $(x^{(k)})_{k\in\mathbb{N}}$ and $(t_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ such that $x^{(k)} \to \hat{x}$ and $x^{(k)} \in \Omega$ for all $k \geqslant k_0$ and

$$d = \lim_{k\to\infty} \frac{x^{(k)} - \hat{x}}{t_k}.$$

By Taylors theorem, we have

$$\begin{aligned} 0 \leqslant \frac{f(x^{(k)}) - f(\hat{x})}{t_k} &= \frac{1}{t_k}\Big(f(\hat{x} + (x^{(k)} - \hat{x})) - f(\hat{x})\Big) \\ &= \frac{1}{t_k}\Big(\cancel{f(\hat{x})} + \nabla f(\hat{x} + \xi(x^{(k)} - \hat{x}))^\mathsf{T}(x^{(k)} - \hat{x})\cancel{- f(\hat{x})}\Big) \\ &= \underbrace{\nabla f(\hat{x} + \xi(x^{(k)} - \hat{x}))}_{\to \nabla f(\hat{x})}{}^\mathsf{T} \underbrace{\frac{x^{(k)} - \hat{x}}{t_k}}_{\to d} \to \nabla f(\hat{x})^\mathsf{T} d. \end{aligned}$$

□

Example 4.1.8 Let $f(x) := x_1 + x_2$ and consider

$$\min_{x\in\mathbb{R}^2} f(x) \qquad \text{such that} \qquad x_1^2 + x_2^2 \leqslant 2.$$

We have $E := \varnothing$ and $I := \{1\}$. The admissable set Ω is a circle with radius $\sqrt{2}$ centered in the origin. We have $\nabla f(x) = (1, 1)^\mathsf{T}$, $c_1(x) := 2 - x_1^2 - x_2^2$ and thus $\nabla c_1(x) = -2x$.

Conditions for descent. As before, consider a small perturbation s. We want $0 > f(x+s) - f(x) \approx \nabla f(x)^\mathsf{T} s$, i.e.

$$\nabla f(x)^\mathsf{T} s < 0. \tag{15}$$

Conditions for admissibility

$$0 \leqslant c_1(x+s) \approx c_1(x) + \nabla c_1(x)^\mathsf{T} s,$$

i.e. of first order, we demand

$$c_1(x) + \nabla c_1(x)^\mathsf{T} s \geqslant 0. \tag{16}$$

Consider two cases.

1. If $c_1(x)$ is inactive, i.e. $c_1(x) > 0$, then (16) holds for all $s \in \mathbb{R}^n$ if $|s|$ is small enough. In our example, this corresponds to the fact that if we are inside the circle, all directions are admissable. For admissable descent take $s = -\alpha \nabla f(x)$ with α sufficiently small.
2. If $c_1(x)$ is active, i.e. $c_1(x) = 0$, then (15) and (16) must hold. Consider the Lagrangian

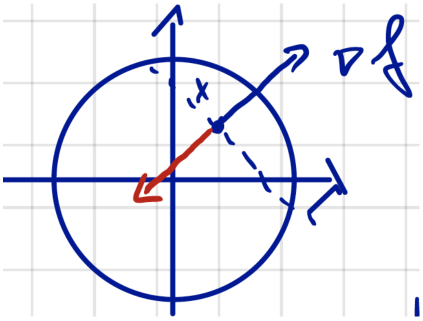
$$L(x, \lambda_1) := f(x) - \lambda_1 c_1(x),$$

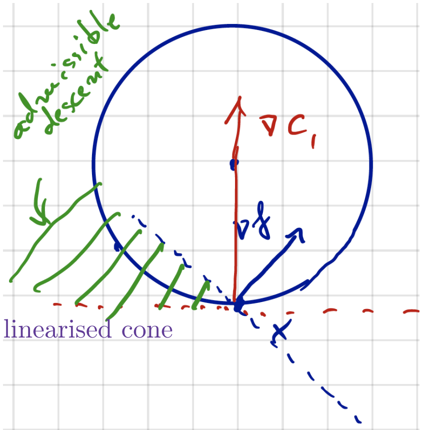
then $\nabla_x L(\hat{x}, \lambda_1) = 0$ and $\lambda_1 c_1(\hat{x}) = 0$ because $\hat{x}$ is an active point, i.e. $c_1(\hat{x}) = 0$. ◇

We have seen that the linearisation of the constraints is an important tool in order to investigate optimality conditions.

Definition 4.1.9 (Linearised cone)
For $x \in \Omega$, the linearised cone of Ω in $x \in \Omega$ is

$$L_\Omega(x) := \left\{ d \in \mathbb{R}^n : \begin{array}{l} d^\mathsf{T} \nabla c_i(x) = 0 \ \forall i \in E, \\ d^\mathsf{T} \nabla c_i(x) \geqslant 0 \ \forall i \in I \cap \mathcal{A}(x) \end{array} \right\}.$$





linearised cone

Figure 21: All descent directions have to lie below the dotted line by (15). By (16), the admissable directions have to lie above the dotted red line.

18.06.2020

Example 4.1.10 ($T_\Omega(x)$ is independent of definition of Ω)
Consider a circle centered at zero with radius $\sqrt{2}$, $\Omega := \{x \in \mathbb{R}^n : c_1(x) = 0\}$, and $c_1(x) := x_1^2 + x_2^2 - 2$. Consider $x = (-\sqrt{2}, 0)^\mathsf{T} \in \Omega$. For $k \in \mathbb{N}$, define

$$x^{(k)} := \begin{pmatrix} -\sqrt{2 - \frac{1}{k^2}} \\ -\frac{1}{k} \end{pmatrix}.$$

Then $\|x^{(k)}\| = \sqrt{2}$, i.e. $(x_k)_{k \in \mathbb{N}} \subset \Omega$. Define $t_k := \|x^{(k)} - x\|$, then $\frac{x^{(k)} - x}{t_k} \to (0, -1)^\mathsf{T}$. Thus $T_\Omega(x) = \operatorname{span}((0,1))$.

For $d \in L_\Omega(x)$ we need $\nabla c_1(x)^\mathsf{T} d = 0$. We have $\nabla c_1(x) = -2x$ and thus $\nabla c_1(x)^\mathsf{T} d = -2\sqrt{2} d$. Thus $L_\Omega(x) = T_\Omega(x)$.

If we redefine Ω as $\{x \in \mathbb{R}^2 : (x_1^2 + x_2^2 - 2)^2 = 0\}$, $T_\Omega(x)$ has changed: As $\nabla c_1(x) = 4(x_1^2 + x_2^2 - 2) \cdot x$, we require $\nabla c_1(x)^\mathsf{T} d$. As $x \in \Omega$, we have $(x_1^2 + x_2^2 - 2) = 0$, i.e. $\nabla c_1(x) = 0$ and thus $L_\Omega(x) = \mathbb{R}^2$. ◇

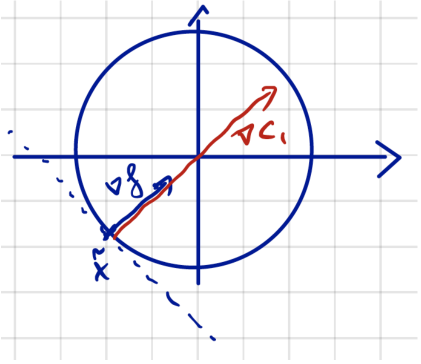


Figure 22: In $x = \hat{x}$, there are not descent directions. We notice that $\nabla f(\hat{x}) = \lambda_1 \cdot \nabla c_1(\hat{x})$ for some $\lambda_1 > 0$.

Example 4.1.11 ($L_\Omega(x) \neq T_\Omega(x)$) Consider Ω defined by $c_1(x) := 1 - x_1^2 - (x_2-1)^2 \geqslant 0$, which is a closed disk centered around $(0,1)^\mathsf{T}$ with radius 1 and the half plane $c_2 := -x_2 \geqslant 0$. Thus $\Omega = \{(0,0)^\mathsf{T}\}$.

Let $x^{(k)}$ be an admissable approximating sequence. Then $x^{(k)} = 0$ for all $k \geqslant k_0$. Thus $T_\Omega(x) = \Omega$.

We have $\nabla c_1(x) = (-2x_1, 2-2x_2)^\mathsf{T}$ and thus $\nabla c_1(0,0) = (0,2)^\mathsf{T}$. We have $\nabla c_2(0) = (0,-1)$.

Thus $d \in L_\Omega(x)$ if and only if $(0,2)^\mathsf{T} d \geqslant 0$ and $(0,-1)^\mathsf{T} d \geqslant 0$, i.e. $2d_2 \geqslant 0$ and $-d_2 \geqslant 0$, i.e. $d_2 = 0$ and thus $L_\Omega(x) = \operatorname{span}((1,0)^\mathsf{T}) \neq T_\Omega(x)$. ◇

Remark 4.1.12 To derive optimality conditions, it will be important to assure $T_\Omega(\hat{x}) = L_\Omega(\hat{x})$, where $\hat{x}$ solves (14).

Definition 4.1.13 (ACQ, LICQ)
Let $x \in \Omega$.

① Abadie constraint qualification (ACQ): $T_\Omega(x) = L_\Omega(x)$.

② Linear independence constraint qualification (LICQ): $\{\nabla c_i(x) : i \in \mathcal{A}(x)\}$ is linearly independent.

4.2 First order necessary optimality conditions

The Lagrangian is

Lagrangian

$$L(x,\lambda) := f(x) - \sum_{i \in E \cup I} \lambda_i c_i(x).$$

Theorem 4.2.1: KKT conditions (1939 / 1951)

Let $\tilde{x}$ be solution to (14), f and $(c_i)_{i \in I \cup E}$ be $\mathcal{C}^1$ functions such that (ACQ) is satisfied. Then there exists a vector $(\tilde{\lambda}_i)_{i \in I \cup E}$ of Lagrange multipliers such that

① $\nabla_x L(\tilde{x}, \tilde{\lambda}) = 0$,
② $c_i(\tilde{x}) = 0$ for all $i \in E$,
③ $c_i(\tilde{x}) \geqslant 0$ for all $i \in I$,
④ $\tilde{\lambda}_i \geqslant 0$ for all $i \in I$,
⑤ $\tilde{\lambda}_i c_i(\tilde{x}) = 0$ for all $i \in E \cup I$.

① - ⑤ are the Karush-Kuhn-Tucker (KKT) conditions.

Remark 4.2.1 ⑤ is called complementarity condition; we infer $\tilde{\lambda}_i > 0 \implies c_i(\tilde{x}) = 0$ and that $c_i(\tilde{x}) > 0$ (i.e. c_i is inactive) implies that $\tilde{\lambda}_i = 0$. Thus ① can be written as

complementarity condition

$$0 = \nabla f(\hat{x}) - \sum_{i \in \mathcal{A}(\hat{x})} \tilde{\lambda}_i \nabla c_i(\hat{x}).$$

Assume that (LICQ) holds, then the Lagrange multipliers $\tilde{\lambda}_i$ are uniquely defined (above).

Example 4.2.2 ((LICQ) being fulfilled) Consider

$$\min_{x\in\mathbb{R}^2} x_1 + x_2 \quad \text{such that} \quad 2 - x_1^2 - x_2^2 \geqslant 0,\ x_2 \geqslant 0,$$

whose solution is $\hat{x} = (-\sqrt{2}, 0)^\mathsf{T}$. Let $c_1(x) := 2 - x_1^2 - x_2^2$ and $c_2(x) := x_2$. We have $\nabla c_1(x) = -2x$ and thus $\nabla c_1(\hat{x}) = (2\sqrt{2}, 0)^\mathsf{T}$. We have $\nabla c_2(x) = (0,1)^\mathsf{T}$ and $\nabla f(x) = (1,1)^\mathsf{T}$. We indeed have a unique decomposition $\nabla f(\hat{x}) = \lambda_1 \nabla c_1(\hat{x}) + \lambda_2 \nabla c_2(\hat{x})$ with $\lambda_1, \lambda_2 > 0$. ◇

Example 4.2.3 Consider

$$\min_{x\in\mathbb{R}^2} \left(x_1 - \frac{3}{2}\right)^2 + \left(x_2 - \frac{1}{2}\right)^4 \quad \text{such that} \quad \begin{cases} 1 - x_1 - x_2 \geqslant 0, \\ 1 - x_1 + x_2 \geqslant 0, \\ 1 + x_1 - x_2 \geqslant 0, \\ 1 + x_1 + x_2 \geqslant 0, \end{cases}$$

whose solution is $\hat{x} := (1,0)^\mathsf{T}$, where c_1 and c_2 are active. We have $\nabla f(x) = (2x_1 - 3, 4(x_2 - 0.5)^3)$ and thus $\nabla f(\hat{x}) = (-1, -0.5)^\mathsf{T}$. We have $\nabla c_1(x) = (-1,-1)$, $\nabla c_2(x) = (-1,1)$ and

$$\nabla f(\hat{x}) = \begin{pmatrix} -1 \\ -\frac{1}{2} \end{pmatrix} = \frac{3}{4}\begin{pmatrix} -1 \\ -1 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} -1 \\ 1 \end{pmatrix} = \lambda_1 \nabla c_1(\tilde{x}) + \lambda_2 c_2(\tilde{x}),$$

i.e. $\tilde{\lambda} := \left(\frac{3}{4}, \frac{1}{4}, 0, 0\right)^\mathsf{T}$, then all KKT conditions are fulfilled. ◇

22.06

Definition 4.2.4 (Strict complementarity)
A Lagrange multiplier $\tilde{\lambda}$ satisfies strict complementarity if $\tilde{\lambda}_i > 0$ for all $i \in I \cap \mathcal{A}(\tilde{x})$.

strict complementarity

Example 4.2.5 (Linear equality constraints)
For $m \leqslant n$ and $a^{(i)} \in \mathbb{R}^n$ consider

$$\min f(x) \quad \text{such that} \quad {a^{(i)}}^\mathsf{T} x = b_i,\ i \in \{1, \dots, m\}.$$

Define $A := (a^{(1)} \dots, a^{(m)})^\mathsf{T} \in \mathbb{R}^{m\times n}$. We thus have $c_i(x) := {a^{(i)}}^\mathsf{T} x - b_i$ and thus $\nabla c_i(x) = a^{(i)}$. Now, (LICQ) is equivalent to $\operatorname{rank}(A) = m$.

We have

$$L(x,\lambda) = f(x) - \sum_{i=1}^{m} \lambda_i c_i(x) = f(x) - \lambda^\mathsf{T}(Ax - b) = f(x) - (A^\mathsf{T}\lambda)^\mathsf{T} x + \lambda^\mathsf{T} b$$

and thus

$$\nabla_x L(x,\lambda) = \nabla f(x) - A^\mathsf{T}\lambda \quad \text{and} \quad \nabla_\lambda L(x,\lambda) = -(Ax - b)$$

Thus if $(\tilde{x}, \tilde{\lambda})$ is a solution of (14), we have

$$\nabla f(\tilde{x}) - A^\mathsf{T}\tilde{\lambda} = 0 \quad \text{and} \quad b - A\tilde{x} = 0,$$

i.e. $\nabla L(\tilde{x}, \tilde{\lambda}) = 0$ and thus $(\tilde{x}, \tilde{\lambda})$ is a stationary point of L. ◇

4.3 Proof of the KKT conditions

We introduce the following notation: $A^\mathsf{T}\tilde{x} = [\nabla c_i(\tilde{x})]_{i\in\mathcal{A}(\tilde{x})}$.

Lemma 4.3.1
Let $\tilde{x} \in \Omega$.

1. $T_\Omega(\tilde{x}) \subset L_\Omega(\tilde{x})$.
2. *(LICQ) implies (ACQ).*

Proof. 1 Without loss of generality assume $c_i(x)$, $i \in \{1, \dots, m\}$ be the active constraints in $\tilde{x}$. Let $d \in T_\Omega(\tilde{x})$. Then there exists an admissable approximation $(x^{(k)})_{k\in\mathbb{N}}$ and a sequence $(t_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ such that $\frac{x^{(k)}-\tilde{x}}{t_k} \xrightarrow{n\to\infty} d$. For k sufficiently large and c_i is an equality constraint, we have that

$$\begin{aligned} 0 &= \frac{1}{t_k} c_i(x^{(k)}) = \frac{1}{t_k} c_i(\tilde{x} + (x^{(k)} - \tilde{x})) \\ &= \left[\underbrace{c_i(\tilde{x})}_{=0} + \underbrace{\nabla c_i(\tilde{x} + \alpha(x^{(k)} - \tilde{x}))^\mathsf{T}}_{\to \nabla c_i(\tilde{x})} \right] \underbrace{\frac{x^{(k)} - \tilde{x}}{t_k}}_{\to d} \xrightarrow{k\to\infty} \nabla c_i(\tilde{x})^\mathsf{T} d \end{aligned}$$

by Taylor expansion for $i \in E$. Similarly we can show that $\nabla c_i(\tilde{x})^\mathsf{T} d \geqslant 0$ for $i \in I \cap \mathcal{A}(\tilde{x})$. Thus $d \in L_\Omega(\tilde{x})$.

2 see Lecture notes or Nocedal + Wright, uses the implicit function theorem. □

We will write $x \geqslant 0$ if $x_i \geqslant 0$ for all i.

Lemma 4.3.2 (Farkas, 1902)
Let $K := \{By + Cw : y \in \mathbb{R}^m, y \geqslant 0, w \in \mathbb{R}^p\}$ with $B \in \mathbb{R}^{n\times m}$ and $C \in \mathbb{R}^{n\times p}$. For each $g \in \mathbb{R}^n$ either

- *$g \in K$ OR*
- *there exists $d \in \mathbb{R}^n$ such that $g^\mathsf{T} d < 0$, $B^\mathsf{T} d \geqslant 0$ and $C^\mathsf{T} d = 0$.*

Proof. see Nocedal + Wright.

Sketch for $n = 2$, $m = 3$, $p = 3$ and $C = 0$, i.e. $K = \{By : y \geqslant 0\}$. Let $B := (b_1, b_2, b_3) \in \mathbb{R}^{2\times 3}$.

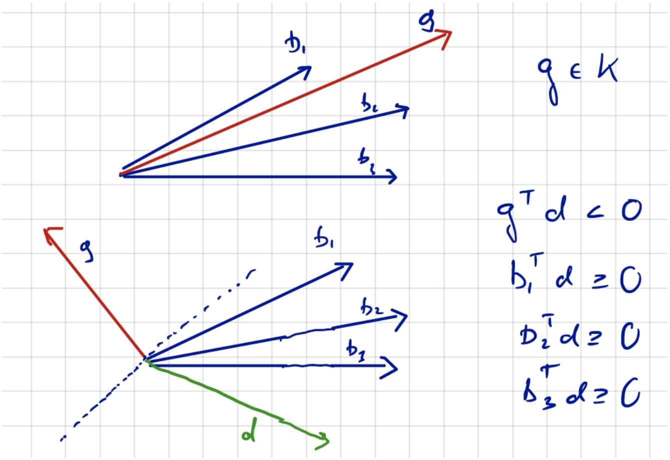


Figure 23: The vector d has to lie below the dotted line.

Proof. (Of theorem 4.2.1) Let

$$N := \left\{ \sum_{i \in \mathcal{A}(\tilde{x})} \lambda_i \nabla c_i : \lambda \geqslant 0 \right\}$$

and $g := \nabla f(\tilde{x})$. By lemma 4.3.2 either

$$\nabla f(\tilde{x}) = \sum_{i \in \mathcal{A}(\tilde{x})} \lambda_i A^{\mathsf{T}}(\tilde{x}) \tilde{\lambda}$$

with $\tilde{\lambda}_i \geqslant 0$ for $i \in \mathcal{A}(\tilde{x}) \cap I$ or there exists a $d \in \mathbb{R}^n$ such that $\nabla f(\tilde{x})^{\mathsf{T}} d < 0$, $\nabla c_i^{\mathsf{T}} d = 0$ for $i \in E$ and $\nabla c_i^{\mathsf{T}} d \geqslant 0$ for $i \in \mathcal{A}(\tilde{x}) \cap I$.

We can rewrite those three conditions as $\nabla f(\tilde{x})^{\mathsf{T}} d < 0$ and $d \in L_\Omega(\tilde{x})$. By assumption $\tilde{x} \in \Omega$ and (ACQ) hold. Thus we have $\nabla f(\tilde{x})^{\mathsf{T}} d < 0$ for a $d \in T_\Omega(\tilde{x})$, which is a contradiction to theorem 4.1.1, so the first option has to hold.

Define $\tilde{\lambda}_i = 0$ for $i \notin \mathcal{A}(\tilde{x})$, so the last condition (complementarity condition) holds. □

4.4 Second order optimality conditions

In the following, we assume that f and the c_i are $\mathcal{C}^2$ functions.

Definition 4.4.1 (Critical cone)
If $(\tilde{x}, \tilde{\lambda})$ satisfy the KKT conditions, the critical cone is

critical cone

$$C(\tilde{x}, \tilde{\lambda}) = \{ w \in L_\Omega(\tilde{x}) : \nabla c_i(\tilde{x})^{\mathsf{T}} w = 0 \ \forall i \in \mathcal{A}(\tilde{x}) \cap I \text{ s.th. } \tilde{\lambda}_i > 0 \}.$$

We have $w \in C(\tilde{x}, \tilde{\lambda})$ if and only if

$$\nabla c_i(\tilde{x})^{\mathsf{T}} w \begin{cases} = 0 \ \forall i \in E \text{ and } \forall i \in \mathcal{A}(\tilde{x}) \cap I \text{ s.th. } \tilde{\lambda}_i > 0 \\ \geqslant 0 \ \forall i \in \mathcal{A}(\tilde{x}) \cap I \text{ s.th. } \tilde{\lambda}_i = 0. \end{cases}$$

Remark 4.4.2 For $d \in C(\tilde{x}, \tilde{\lambda})$ we have

$$\nabla f(\tilde{x})^\top d = \sum_{i \in \mathcal{A}(\tilde{x})} \tilde{\lambda}_i \nabla c_i(\tilde{x})^\top d = 0.$$

Thus $C(\tilde{x}, \tilde{\lambda})$ contains all directions where, based on first order information, we cannot decide if f decreases or increases.

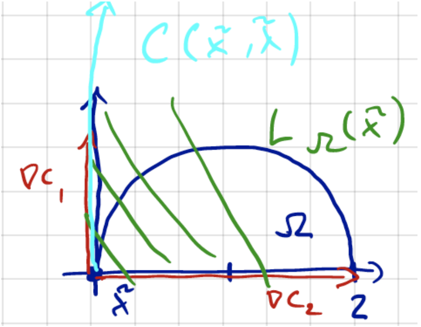


Figure 24: The admissable set Ω and the critical and linearised cones.

Example 4.4.3 Consider

$$\min_{x=(x_1,x_2)} x_1 \quad \text{such that} \quad x_2 \geqslant 0,\ 1-(x_1-1)^2 - x_2^2 \geqslant 0,$$

whose solution is $\tilde{x} := (0,0)$.

We have $\mathcal{A}(\tilde{x}) = \{1,2\}$, $\nabla c_1(x) = (0,1)^\top$ and $\nabla c_2(x) = -2(x_1-1, x_2)^\top$ and thus $\nabla c_2(\tilde{x}) = (2,0)^\top$. Then (ACQ) holds, as $(0,1)^\top$ and $(2,0)^\top$ are linearly independent.

We have $\nabla f(x) = (1,0)^\top = 0 \cdot \nabla c_1(\tilde{x}) + \frac{1}{2}\nabla c_2(\tilde{x})$. Thus

$$\begin{aligned} C(\tilde{x}, \tilde{\lambda}) &= \{d \in L_\Omega(\tilde{x}) : \nabla c_2(\tilde{x}) d = 0\} = \{d \in L_\Omega(\tilde{x}) : d_1 = 0\} \\ &= \{d \in \mathbb{R}^2 : d_1 = 0, d_2 \geqslant 0\}, \end{aligned}$$

as $T_\Omega(\tilde{x}) = \{d \in \mathbb{R}^2 : d_1, d_2 \geqslant 0\}$. ◇

25.06.2020

Theorem 4.4.1: 2nd order necessary condition

Let $\tilde{x}$ be a local solution to (14), assume that (LICQ) holds and let $\tilde{\lambda}$ be such that the KKT conditions are satisfied. Then

$$w^\top \nabla^2_{xx} L(\tilde{x}, \tilde{\lambda}) w \geqslant 0$$

holds for all $w \in C(\tilde{x}, \tilde{\lambda})$.

Proof. Let $w \in C(\tilde{x}, \tilde{\lambda}) \subset L_\Omega(\tilde{x}) = T_\Omega(\tilde{x})$. By the proof of lemma 4.3.1 (2) we know that there exists a admissable approximation $(x^{(k)})_{k\in\mathbb{N}}$ and $(t_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ such that

$$c_i(x^{(k)}) = t_k \nabla c_i(\tilde{x})^\top w. \tag{17}$$

We have

$$\begin{aligned} L(x^{(k)}, \tilde{\lambda}) &= f(x^{(k)}) - \sum_{i \in E \cup I} \tilde{\lambda}_i c_i(x^{(k)}) \\ &\overset{(17)}{=} f(x^{(k)}) - t_k \sum_{i \in E \cup I} \tilde{\lambda}_i \underbrace{\nabla c_i(\tilde{x})^\top w}_{=0} = f(x^{(k)}) \end{aligned}$$

as $w \in C(\tilde{x}, \tilde{\lambda})$.

Taylor-expansion of $L(x^{(k)}, \tilde{\lambda})$ around $\tilde{x}$ yields

$$\begin{aligned} f(x^{(k)}) &= L(x^{(k)}, \tilde{\lambda}) \\ &= \underbrace{L(\tilde{x}, \tilde{\lambda})}_{=f(\tilde{x})} + (x^{(k)} - \tilde{x})^{\mathsf{T}} \underbrace{\nabla_x L(\tilde{x}, \tilde{\lambda})}_{=0 \quad \text{(KKT)}} \\ &\quad + \frac{1}{2}(x^{(k)} - \tilde{x})^{\mathsf{T}} \nabla^2_{xx} L\big(\tilde{x} + \xi(x^{(k)} - \tilde{x}), \tilde{\lambda}\big)(x^{(k)} - \tilde{x}) \\ &= f(\tilde{x}) + \frac{1}{2}(x^{(k)} - \tilde{x})^{\mathsf{T}} \nabla^2_{xx} L\big(\tilde{x} + \xi(x^{(k)} - \tilde{x}), \tilde{\lambda}\big)(x^{(k)} - \tilde{x}) \end{aligned}$$

and thus dividing by t_k^2 yields

$$\begin{aligned} 0 \leqslant \frac{f(x^{(k)}) - f(\tilde{x})}{t_k^2} &= \frac{1}{2} \frac{(x^{(k)} - \tilde{x})^{\mathsf{T}}}{t_k} \nabla^2_{xx} L\big(\tilde{x} + \xi(x^{(k)} - \tilde{x}), \tilde{\lambda}\big) \frac{(x^{(k)} - \tilde{x})}{t_k} \\ &\xrightarrow{k \to \infty} \frac{1}{2} w^{\mathsf{T}} \nabla^2_{xx} L(\tilde{x}, \tilde{\lambda}) w, \end{aligned}$$

where the inequality comes from the fact that $(x^{(k)})_{k \in \mathbb{N}}$ is an admissable approximation, so the inequality holds for all $k \geqslant k_0$ for some $k_0 \in \mathbb{N}$. □

Theorem 4.4.2: 2$^{\text{nd}}$ order sufficient condition

Let $\tilde{x} \in \Omega$ and $\tilde{\lambda}$ such that $(\tilde{x}, \tilde{\lambda})$ satisfies the KKT conditions. If there exists a $\sigma > 0$ such that

$$w^{\mathsf{T}} \nabla^2_{xx} L(\tilde{x}, \tilde{\lambda}) w \geqslant \sigma |w|^2$$

holds for all $w \in C(\tilde{x}, \tilde{\lambda})$, then $\tilde{x}$ is a strict local solution to (14).

Proof. Lecture Notes or Nocedal and Wright. □

Remark 4.4.4 The proof shows a locally quadratic behaviour analogous to theorem 2.1.4:

$$f(x^{(k)}) \geqslant f(\tilde{x}) + \frac{\sigma}{4} |x^{(k)} - \tilde{x}|^2$$

holds for all admissable approximations $(x^{(k)})$.

Remark 4.4.5 Assume strict complementarity, i.e. $\tilde{\lambda}_i > 0$ for all $i \in \mathcal{A}(\tilde{x}) \cap I$. Then

$$C(\tilde{x}, \tilde{\lambda}) = \{d \in \mathbb{R}^n : \nabla c_i(\tilde{x})^{\mathsf{T}} d = 0 \ \forall i \in \mathcal{A}(\tilde{x})\}.$$

Otherwise we had the distinction for the inequality constraints with corresponding Lagrange parameters being nonnegative, but we assume strict complementarity, so the only chance for a parameter to belong to the active set is that it is positive.

If $G(\tilde{x}) := (\nabla c_i(\tilde{x})^{\mathsf{T}})_{i \in \mathcal{A}(\tilde{x})}$, then

$$C(\tilde{x}, \tilde{\lambda}) = \ker(G(\tilde{x})),$$

which is a subspace of $\mathbb{R}^n$. Then the second order optimality conditions becomes

$$w \nabla^2_{xx} L(\tilde{x}, \tilde{\lambda}) w \geqslant \sigma |w|^2 \qquad \forall w \in \ker(G(\tilde{x})).$$

Let $\ell := \dim(\ker(G(\tilde{x})))$ and $(s_k)_{k=1}^{\ell}$ a basis of $\ker(G(\tilde{x}))$. Then $Z := (s_1 \dots s_\ell) \in \mathbb{R}^{n\times \ell}$ is a null-space matrix, i.e.

$$\ker(G(\tilde{x})) = \{Za : a \in \mathbb{R}^{\ell}\}$$

and the second order optimality conditions reduce to $Z^{\mathsf{T}}\nabla^2_{xx}L(\tilde{x},\tilde{\lambda})Z$ being positive definite on $\mathbb{R}^{\ell}$.

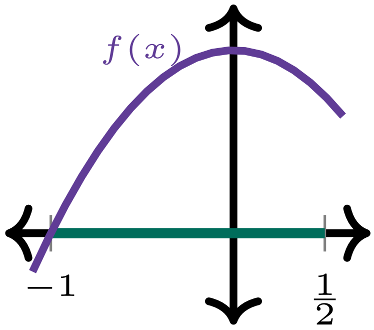


Figure 25: The objective function from example 4.4.6

Example 4.4.6 Consider

$$\min_x 1 - x^2 \quad \text{such that} \quad x \geqslant -1,\ x \leqslant \frac{1}{2}.$$

We have a local minimum in $\tilde{x}_2 = \frac{1}{2}$ and a global minimum in $\tilde{x}_1 = -1$. The Lagrangian is

$$L(\tilde{x}, \lambda_1, \lambda_2) := 1 - x^2 - \lambda_1(x+1) - \lambda_2\left(\frac{1}{2} - x\right).$$

In $\tilde{x}_2 = \frac{1}{2}$, c_1 is inactive so $\tilde{\lambda}_1 = 0$.

By the KKT conditions

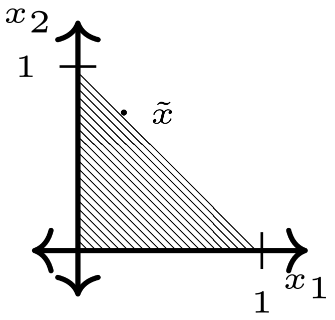
$$0 \overset{!}{=} L_x(\tilde{x},\tilde{\lambda}) = -2\tilde{x} - \tilde{\lambda}_1 + \tilde{\lambda}_2 = -2\tilde{x} + \tilde{\lambda}_2$$

and thus $\lambda_2 = 1$.

Thus

$$C(\tilde{x},\tilde{\lambda}) = \{d \in \mathbb{R} : c_2'(\tilde{x}_2)d = 0\} = \{0\}$$

and so the second order condition is satisfied for any $\sigma > 0$. ◇



Example 4.4.7 Consider

$$\min -\frac{1}{2}\sqrt{x_1} - \frac{1}{2}x_2 \quad \text{such that} \quad x_1 \geqslant 0,\ x_2 \geqslant 0,\ 1 - x_1 - x_2 \geqslant 0.$$

The Lagrangian is

$$L(x,\lambda) = -\frac{1}{2}\sqrt{x_1} - \frac{1}{2}x_2 - \lambda_1 x_1 - \lambda_2 x_2 - \lambda_3(1 - x_1 - x_2)$$

The KKT conditions yield

$$\nabla_x L(x,\lambda) = \begin{pmatrix} -\frac{1}{4\sqrt{x_1}} - \lambda_1 + \lambda_3 \\ -\frac{1}{2} - \lambda_2 + \lambda_3 \end{pmatrix} \overset{!}{=} 0,$$

where the complementarity condition is

$$\lambda_1 x_1 = 0, \quad \lambda_2 x_2 = 0, \quad \lambda_3(1 - x_1 - x_2) = 0$$

Plugging $\lambda_1 = \lambda_3 - \frac{1}{4\sqrt{x_1}}$ and $\lambda_2 = \lambda_3 - \frac{1}{2}$ and assuming $\lambda_3 \neq 0$ (otherwise $\lambda_2 < 0$), we get $\lambda_3 x_1 = \frac{1}{4}\sqrt{x_1}$ and $\lambda_3 x_2 = \frac{1}{2}x_2$, so $\lambda_3 = \frac{1}{2}$ and thus $\lambda_2 = 0$ and $x_1 = \frac{1}{4}$ and thus $\frac{1}{2}(1 - \frac{1}{4} - x_2) = 0$, so $x_2 = \frac{3}{4}$, i.e. $\tilde{x} = \frac{1}{4}(1,3)^{\mathsf{T}}$. We have $\lambda_1 = \lambda_2 = 0$ and $\lambda_3 = \frac{1}{2}$, and thus strict complementarity.

We have

$$\nabla^2 L_{xx}(\tilde{x},\tilde{\lambda}) = \operatorname{diag}\left(\frac{1}{8}\tilde{x}_1^{-\frac{3}{2}}, 0\right) = \operatorname{diag}(1,0),$$

which is a matrix with rank one and is not positive definite on $\mathbb{R}^2$. But this is not necessary for the second order condition.

Consider

$$C(\tilde{x}, \tilde{\lambda}) = \{d \in \mathbb{R}^2 : \nabla c_3^\mathsf{T} d = 0\} = \operatorname{span}((1, -1)).$$

Thus for $w \in C(\tilde{x}, \tilde{\lambda})$ we get $w_1^2 = w_2^2$ and thus

$$w^\mathsf{T} L_{xx}(\tilde{x}, \tilde{\lambda}) w = w_1^2 = \frac{w_1^2 + w_2^2}{2} = \frac{1}{2}|w|^2,$$

so the second order sufficient condition is satisfied. ◇

4.5 Problems with box constraints

Let

$$\Omega := \{x \in \mathbb{R}^n : v_i \leqslant x_i \leqslant w_i \ \forall i \in \{1, \dots, n\}\}$$

be a box in $\mathbb{R}^n$. Let $v := (v_i)_{i=1}^n$, $w := (w_i)_{i=1}^n$ and for simplicity assume $v < w$ componentwise. Consider

$$\min_{v \leqslant x \leqslant w} f(x), \qquad f \colon \mathbb{R}^n \to \mathbb{R} \text{ sufficiently smooth.} \tag{18}$$

We reformulate (18) into standard form: we can rewrite $v \leqslant x \leqslant w$ as $x - v \geqslant 0$ and $w - x \geqslant 0$, which is equivalent to

$$\begin{pmatrix} I \\ -I \end{pmatrix} x + \begin{pmatrix} -v \\ w \end{pmatrix} \geqslant 0$$

or $Gx + r \geqslant 0$ with $G := (I, -I)^\mathsf{T}$ and $r := (-v, w)^\mathsf{T}$.

At most one of $v_i \leqslant x_i \leqslant w_i$ can be active (if $x_i = v_i < w_i$, then $w_i - x_i > 0$), so $G(x) := (\nabla c_i(x))_{i \in \mathcal{A}(x)} = \operatorname{diag}((\pm 1)_{i=1}^n)$ and thus $\{\nabla c_i : i \in \mathcal{A}(\tilde{x})\}$ is linearly independent and thus (LICQ) holds. Thus the λ_i are uniquely defined.

We have

$$L(x, \lambda) = f(x) - \sum_{j=1}^n \lambda_j^{(\ell)}(x_j - v_j) - \sum_{j=1}^n \lambda_j^{(u)}(-x_j + w_j)$$

By the KKT conditions we have

$$0 \overset{!}{=} \frac{\partial}{\partial x_i} L(x, \lambda) = \frac{\partial f(x)}{\partial x_i} - \lambda_i^\ell + \lambda_i^{(n)}$$

and $\lambda_i^\ell, \lambda_i^{(n)} \geqslant 0$. Only one of them can be non-zero as at most one of the inequalities can be active.

Case 1: $\frac{\partial f(x)}{\partial x_i} = 0$. Then $\lambda_i^{(\ell)} = \lambda_i^{(u)} = 0$.

Case 2: $\frac{\partial f(x)}{\partial x_i} > 0$. Then $\lambda_i^{(\ell)} = \frac{\partial f(x)}{\partial x_i}$ and $\lambda_i^{(u)} = 0$.

Case 3: $\frac{\partial f(x)}{\partial x_i} < 0$. Then $\lambda_i^{(u)} = -\frac{\partial f(x)}{\partial x_i}$ and $\lambda_i^{(\ell)} = 0$.

In a more compact way: $\lambda_i^{(\ell)} = \left[\frac{\partial f(x)}{\partial x_i}\right]_+$ and $\lambda_i^{(u)} = \left[\frac{\partial f(x)}{\partial x_i}\right]_-$.

Let use investigate second order conditions. We have

$$C(\tilde{x},\tilde{y}) = \left\{ d \in L_\Omega(\tilde{x}) : d_i = 0 \text{ if } \frac{\partial f(\tilde{x})}{\partial x_i} \neq 0 \right\}.$$

4.6 Further regularity conditions

29.06.2020

Lemma 4.6.1 (affine linear c_i imply (ACQ))
Assume all constraints are affine linear, i.e. $c_i(x) = a_i^\mathsf{T} x + b_i$ for $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$. Then (ACQ) holds.

Proof. We show $L_\Omega(\tilde{x}) \subset T_\Omega(\tilde{x})$. Let $w \in L_\Omega(\tilde{x})$. Then $a_i^\mathsf{T} w = 0$ for $i \in E$ and $a_i^\mathsf{T} w \geqslant 0$ for $i \in \mathcal{A}(\tilde{x}) \cap I$, as $\nabla c_i = a_i$.

If $i \in I \backslash \mathcal{A}(\tilde{x})$, then $c_i(\tilde{x}) > 0$. Then there exists a $t_0 > 0$ such that $c_i(\tilde{x} + tw) > 0$ for all $t \in [0, t_0]$, so c_i "‘stays"’ inactive.

Let $(x^{(k)} := \tilde{x} + \frac{t_0}{k} w)_{k\in\mathbb{N}}$. For $i \in \mathcal{A}(\tilde{x}) \cap I$ we have (as $c_i(\tilde{x}) = 0$)

$$c_i(x^{(k)}) = c_i(x^{(k)}) - c_i(\tilde{x}) = a_i^\mathsf{T}(x^{(k)} - \tilde{x}) = \frac{t_0}{k} a_i^\mathsf{T} w \geqslant 0$$

since $w \in L_\Omega(\tilde{x})$, so $(x^{(k)})_{k\in\mathbb{N}}$ is an admissable approximation.

For $i \in E$ we have

$$c_i(x^{(k)}) = c_i(x^{(k)}) - c_i(\tilde{x}) = \frac{t_0}{k} a_i^\mathsf{T} w \geqslant 0,$$

by the same reasoning as above, so $(x^{(k)})_{k\in\mathbb{N}}$ is an admissable approximation. Moreover,

$$\lim_{k\to\infty} \frac{x^{(k)} - \tilde{x}}{\frac{t_0}{k}} = \lim_{k\to\infty} \frac{\frac{t_0}{k} w}{\frac{t_0}{k}} = w,$$

so $w \in T_\Omega(\tilde{x})$. □

Definition 4.6.2 (Mangasarian-Fromovitz)
(MFCQ) holds if there exists a $w \in \mathbb{R}^n$ such that

$$\nabla c_i(\tilde{x})^\mathsf{T} w \begin{cases} > 0, & \forall i \in \mathcal{A}(\tilde{x}) \cap I \\ = 0, & \forall i \in E. \end{cases}$$

and $\{\nabla c_i\}_{i\in E}$ is linearly independent.

Lemma 4.6.3
We have (LICQ) $\implies$ (MFCQ) $\implies$ (ACQ).

We only prove the first implication.

Proof. Let $G(\tilde{x}) := (\nabla c_i(\tilde{x})^\mathsf{T})_{i\in\mathcal{A}(\tilde{x})}$. By (LICQ) it has maximal rank. Then there exists a $w \in \mathbb{R}^n$ such that

$$\nabla c_i(\tilde{x})^\mathsf{T} w = \begin{cases} 1, & \forall i \in \mathcal{A}(\tilde{x}) \cap I, \\ 0, & \forall i \in E. \end{cases}$$

This is because as $G(\tilde{x})$ has maximal rank, adding an additional column doesn't change the rank. A linear system $Ax = b$ is solvable if the rank of A is equal to the rank of the extended matrix $A|b$. The system is solvable as A as maximal rank and thus we can append any b, in particular one with ones in first components for the active inequality constraints and zeros for all the equality constraints. □

Remark 4.6.4 (Uniqueness of Lagrange multiplier)
If (MFCQ) holds, the Lagrange multipliers need not be unique. Recall that when (ACQ) holds, we can write, thanks to the KKT conditions, the gradient as a linear combination of the active constraint gradients. If these are linearly independent, there is only one unique combination of scalars, so in case of (LICQ) the Lagrange multipliers are unique.

Definition 4.6.5 (Slater constraint qualification)
Let $D \subset \mathbb{R}^n$ be an open and convex subset such that $-c_i$ is a convex $\mathcal{C}^1$ function on D for $i \in I$ and $c_i(x) := a_i^\top x + b_i$ is an affine linear function for $i \in E$.

Then the global Slater condition holds if the set $(a_i)_{i \in E}$ is linearly independent and there exists a $v \in \mathbb{R}^n$ such that $c_i(v) = 0$ for all $i \in E$ and $c_i(v) \geqslant 0$ for $i \in I$.

Remark 4.6.6 (SQC $\implies$ MFCQ) One can show that if (SQC) holds in $\tilde{x} \in \Omega$, then (MFCQ) holds.

4.7 Geometric interpretation of necessary optimality conditions

Definition 4.7.1 (Normal cone)
For $x \in \Omega$

$$N_\Omega(x) := \{v \in \mathbb{R}^n : v^\top w \leqslant 0 \ \forall w \in T_\Omega(x)\}$$

is the normal cone to $T_\Omega(x)$. The elements of $N_\Omega(x)$ are normal vectors.

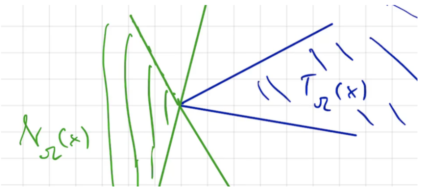


Figure 26: The two green lines are perpendicular to the limiting rays of the cone $T_\Omega(x)$.

Theorem 4.7.1: Normal cone and (14)

Let $\tilde{x}$ be a local solution to (14). Then $-\nabla f(\tilde{x}) \in N_\Omega(\tilde{x})$.

Proof. By theorem 4.1.1 $\nabla f(\tilde{x})d \geqslant 0$, i.e. $-\nabla f(\tilde{x})d \leqslant 0$ holds for all $d \in T_\Omega(\tilde{x})$. Thus $-\nabla f(x) \in N_\Omega(\tilde{x})$. □

Remark 4.7.2 If $\tilde{x} \in \operatorname{int}(\Omega)$, then $T_\Omega(\tilde{x}) = \mathbb{R}^n$ (shown before). Then $N_\Omega(\tilde{x}) = \{0\}$, so $\nabla f(\tilde{x}) = 0$ if $\tilde{x}$ solves (14).

4.8 Duality

Consider

$$\min f(x) \quad \text{such that} \quad c_i(x) \geqslant 0 \ \forall i \in \{1, \dots, m\} \tag{19}$$

and define $c(x) := (c_i(x))_{i=1}^{m}{}^{\mathsf{T}}$. The Lagrangian is

$$L(x, \lambda) := f(x) - \lambda^{\mathsf{T}} c(x).$$

for $\lambda \in \mathbb{R}^m$. The dual objective function is

dual objective function

$$q \colon \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}, \ \lambda \mapsto \inf_x L(x, \lambda).$$

We define $D := \{\lambda \in \mathbb{R}^m : q(\lambda) > -\infty\}$. The dual problem is

$$\max q(\lambda) \quad \text{such that} \quad \lambda \geqslant 0. \tag{P_D}$$

We will see that there's a relationship between the problems and its dual. Under certain conditions, the Lagrange multiplier corresponding to the solution of the problem is a solution of the dual problem. Under further restrictions, the Lagrange multiplier corresponding to the solution of the dual problem is a solution of the original problem. Sometimes we can circumvent solving the original problem by solving the dual problem.

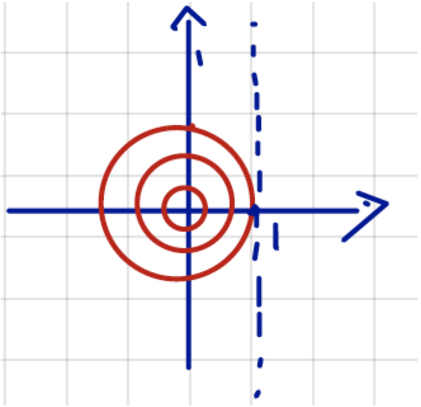


Figure 27: Isolines of f.

Example 4.8.1 Consider

$$\min \frac{1}{2}\left(x_1^2 + x_2^2\right) \quad \text{such that} \quad x_1 - 1 \geqslant 0,$$

whose solution is $\tilde{x} := (1, 0)^{\mathsf{T}}$ with $f(\tilde{x}) = \frac{1}{2}$.

The Lagrangian is

$$L(x, \lambda_1) := \frac{1}{2}\left(x_1^2 + x_2^2\right) - \lambda_1(x_1 - 1)$$

The map $x \mapsto L(x, \lambda_1)$ is convex, so the global solution satisfies $\nabla_x L(x, \lambda_1) = 0$. We have (for $\lambda_1 \neq 0$ otherwise the solution is $\tilde{x} = (0, 0)$, which is not admissable)

$$\nabla_x L(x, \lambda_1) = \begin{pmatrix} x_1 - \lambda_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

so $(x_1, x_2) = (\lambda_1, 0)$. We have

$$q(\lambda_1) = \frac{1}{2}\lambda_1^2 - \lambda_1(\lambda_1 - 1) = \lambda_1\left(1 - \frac{1}{2}\lambda_1\right),$$

which has the two zeros 0 and 2 and achieves its global maximum in $\lambda = 1$.

The dual problem is

$$\max_{\lambda_1 \geqslant 0} \lambda_1 \left(1 - \frac{1}{2}\lambda_1\right) = 1$$

with $q(1) = \frac{1}{2}$. ◇

Lemma 4.8.2
The function q is concave and D is convex.

Proof. Lecture Notes. □

Lemma 4.8.3
For each $\bar{x} \in \Omega$ for (19) and every $\bar{\lambda} \geqslant 0$ we have

$$q(\bar{\lambda}) \leqslant f(\bar{x}).$$

Proof. We have

$$q(\bar{\lambda}) = \inf_x \left(f(x) - \bar{\lambda}^{\mathsf{T}} c(x)\right) \leqslant f(\bar{x}) - \underbrace{\bar{\lambda}^{\mathsf{T}} c(\bar{x})}_{\geqslant 0} \leqslant f(\bar{x}).$$ □

Let us consider the KKT conditions for (19):

$$\begin{aligned} \nabla f(\bar{x}) - \nabla c(\bar{x})\bar{\lambda} &= 0, \\ c(\bar{x}) &= 0, \\ \bar{\lambda} &\geqslant 0, \\ \bar{\lambda} c_i(\bar{x}) &= 0 \ \forall i \in \{1, \dots, m\}, \end{aligned}$$

where $\nabla c(x) = (\nabla c_j(x))_{j=1}^m$ is the Jacobian of $c(\bar{x})$, transposed.

Theorem 4.8.1

Let $\bar{x}$ be a solution to (19) and f and each $-c_i$ be convex functions on $\mathbb{R}^n$. Then every $\bar{\lambda}$ such that $(\bar{x}, \bar{\lambda})$ satisfies the KKT conditions is a solution to the dual problem (P_D).

5 Algorithms for problems with linear constraints

06.07.2020

Our general strategy for quadratic problems will be to start with null space methods to handle linear equality constraints, then use the active set method to handle linear constraints. In order to treat general problems, we will use successive quadratic approximation, which approximates the objective by a quadratic function. Recall that locally, in a neighbourhood of an optimum, the (sufficiently smooth) objective function behaves like a quadratic function. We then use the null space and active set method to solve the next step of the approximating quadratic problem with linear constraints.

In the final step, we will deal with general nonlinear constraints, which will be done by linearising the constraints, building again on what we have done algorithmically before. We then need another control-function, which assures that we stay in the admissable region.

5.1 Quadratic objective functions

Equality constraints

Consider

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^\mathsf{T} Q x + q^\mathsf{T} x \quad \text{subject to} \quad Ax = b, \tag{QE}$$

where $Q \in \mathbb{R}^{n \times n}$, $A \in \mathbb{R}^{m \times n}$ and $m \leqslant n$. We assume (for uniqueness of the solution) that $\operatorname{rank}(A) = m$, Q is symmetric and $d^\mathsf{T} Q d \geqslant \alpha |d|^2$ for all $d \in \ker(A)$.

We know that (QE) has a unique solution $\hat{x}$ and the corresponding Lagrange multiplier $\tilde{\lambda}$ is uniquely defined, as because A has maximal rank, the gradients of the linear constraints are linearly independent and thus (ACQ) holds. Now consult remark 4.6.4.

The solution satisfies

$$\begin{pmatrix} Q & -A^\mathsf{T} \\ A & 0 \end{pmatrix} \begin{pmatrix} \tilde{x} \\ \lambda \end{pmatrix} = \begin{pmatrix} -q \\ b \end{pmatrix}, \tag{20}$$

where the second row represents the constraint $Ax = b$ and the first row can be rewritten as $A\tilde{\lambda} = Q\tilde{x} + q = \frac{\mathrm{d}}{\mathrm{d}x} x^\mathsf{T} Q x + q^\mathsf{T} x$, which is the KKT condition.

Applying the transformations $\tilde{x} = x + p$, $h := Ax - b$ and $g := q + Qx$, (20) becomes

$$\underbrace{\begin{pmatrix} Q & A^\mathsf{T} \\ A & 0 \end{pmatrix}}_{=:K} \begin{pmatrix} -p \\ \lambda \end{pmatrix} = \begin{pmatrix} g \\ h \end{pmatrix},$$

where the KKT-Matrix K is invertible, so we can apply standard results such as the Cholesky-decomposition.

Instead, we solve the second system with the nullspace method, where we eliminate the constraints and then consider a free optimisation problem of lower dimension.

nullspace method

① Consider the QR-decomposition of $A^\mathsf{T} \in \mathbb{R}^{n\times m}$, which finds an unitary matrix $H \in \mathbb{R}^{n\times n}$ and an upper triangular matrix $R \in \mathbb{R}^{m\times m}$ such that

$$HA^\mathsf{T} = \begin{pmatrix} R \\ 0 \end{pmatrix}$$

To structure H, we write

$$H = \begin{pmatrix} Y^\mathsf{T} \\ Z^\mathsf{T} \end{pmatrix},$$

where $Y \in \mathbb{R}^{n\times m}$ and $Z \in \mathbb{R}^{n\times n-m}$. As H is unitary, $\operatorname{rank}(H) = n$ and thus $\operatorname{rank}(Y) = m$ and $\operatorname{rank}(Z) = n - m$, i.e. the columns of Y and Z span $\mathbb{R}^n$: for $x \in \mathbb{R}^n$, there exists a unique decomposition

$$x = Yx_y + Zx_z = H^\mathsf{T} \begin{pmatrix} x_y \\ x_z \end{pmatrix},$$

where $x_y \in \mathbb{R}^m$ and $x_z \in \mathbb{R}^{n-m}$.

For $d \in \ker(A)$ we have

$$0 = Ad = A(Yd_y + Zd_z) = AH^\mathsf{T} \begin{pmatrix} d_y \\ d_z \end{pmatrix} = (R^\mathsf{T}, 0) \begin{pmatrix} d_y \\ d_z \end{pmatrix} = R^\mathsf{T} d_y.$$

As R is invertible, we have thus shown

$$d \in \ker(A) \iff d_y = 0,$$

i.e. $d = Zd_z$. Z **is the nullspace matrix** because it generates $\ker(A)$: $\ker(A) = \{Zd_z : d_z \in \mathbb{R}^{n-m}\}$. For a decomposition $x = Yx_y + Zx_z$ we have found $Zx_z \in \ker(A)$ and thus $Yx_y \in (\ker(A))^\perp$.

② We want to solve the inhomogeneous linear system $Ax = b$, by finding a special solution of the inhomogeneous equation and all solutions to the homogeneous equation, the latter of which are generated by the nullspace matrix.

Recall that $\begin{pmatrix} Y^\mathsf{T} \\ Z^\mathsf{T} \end{pmatrix} A^\mathsf{T} = \left(\begin{smallmatrix} R \\ 0 \end{smallmatrix}\right)$ holds, which, by transposition, is equivalent to $(AY, AZ) = (R^\mathsf{T}, 0)$, i.e. $AY = R^\mathsf{T}$ and $AZ = 0$.

Using the decomposition $\tilde{x} = Y\tilde{x}_y + Z\tilde{x}_z$, we obtain

$$b = A\tilde{x} = AY\tilde{x}_y + 0 = R^\mathsf{T}\tilde{x}_y.$$

As R^T is a lower triangular matrix (and thus invertible), we can solve $b = R^\mathsf{T}\tilde{x}_y$ by forward substitution. The solution is given by $\tilde{x}_y = R^{-\mathsf{T}}b$.

To find a special solution of $Ax = b$, we take $Y\tilde{x}_y = w$ and obtain $Aw = b$, so we can write the admissable set $\Omega := \{x \in \mathbb{R}^n : Ax = b\}$, as $\{w + Zz : z \in \mathbb{R}^{n-m}\}$. We have found a reduction of the problem:

$$\min_{z \in \mathbb{R}^{n-m}} f(w + Zz).$$

By the definition of f we get

$$\begin{aligned} f(w + Zz) &= \frac{1}{2}(w + Zz)^\mathsf{T} Q(w + Zz) + q^\mathsf{T}(Zz + w) \\ &= \frac{1}{2} z^\mathsf{T}(Z^\mathsf{T}QZ)z + \langle Z^\mathsf{T}Qw, z\rangle + \langle Z^\mathsf{T}q, z\rangle + C, \end{aligned}$$

where $C = \frac{1}{2}w^\mathsf{T} Q w \in \mathbb{R}$ is a constant with respect to z. We can thus rewrite the reduced problem as

$$\min_{z \in \mathbb{R}^{n-m}} \underbrace{\frac{1}{2} z^\mathsf{T} \tilde{Q} z + \tilde{q}^\mathsf{T} z}_{=:F(z)}$$

with $\tilde{Q} := Z^\mathsf{T} Q Z$ and $\tilde{q} := Z^\mathsf{T}(Qw + q)$.

By our assumptions on Q, $\tilde{Q}$ is positive definite, so above the problem has a unique solution. The optimality condition is $\nabla F(\tilde{z}) = 0$, i.e. $\tilde{Q}\tilde{z} = -\tilde{q}$. We have

$$\tilde{Q}\tilde{z} = Z^\mathsf{T} Q Z \tilde{x}_z = -Z^\mathsf{T} q - Z^\mathsf{T} Q w = -Z^\mathsf{T} q - Z^\mathsf{T} Q Y \tilde{x}_y. \qquad (21)$$

We rewrite the approach: we have

$$-Hq = -\begin{pmatrix} Y^\mathsf{T} \\ Z^\mathsf{T} \end{pmatrix} q = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} =: h$$

for $h_1 \in \mathbb{R}^m$ and $h_2 \in \mathbb{R}^{n-m}$. Define

$$\begin{aligned} B := HQH^\mathsf{T} &= \begin{pmatrix} Y^\mathsf{T} \\ Z^\mathsf{T} \end{pmatrix} Q(Y, Z) \\ &= \begin{pmatrix} Y^\mathsf{T} Q Y & Y^\mathsf{T} Q Z \\ Z^\mathsf{T} Q Y & Z^\mathsf{T} Q Z \end{pmatrix} =: \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}. \end{aligned}$$

With this new notation, we have

$$\underbrace{Z^\mathsf{T} Q Z}_{=B_{22}} \tilde{x}_z = \underbrace{-Z^\mathsf{T} q}_{=h_2} - \underbrace{Z^\mathsf{T} Q Y}_{B_{21}} \tilde{x}_y,$$

so

$$B_{22}\tilde{x}_z = h_2 - B_{21}\tilde{x}_y.$$

As B_{22} is symmetric and positive definite, we can solve this with the Cholesky decomposition.

$$\tilde{x} = Y\tilde{x}_y + \tilde{Z}x_z$$

We have now found

$$\tilde{x} = Y\tilde{x}_y + Z\tilde{x}_z.$$

3 We now compute the Lagrange multiplier $\tilde{\lambda}$.

Why do we care about the Lagrange multiplier even if, for inequality constraints, we don't have to checks its sign? However be aware that we will use this as a building block for the inequality constraints, where the KKT conditions state that the Lagrange multiplier must be nonnegative. This will be very important for the "bookkeeping" in the active set method.

From the KKT conditions we know that $Q\tilde{x} - A^\mathsf{T}\lambda = -q$. Inserting $\tilde{x} = H^\mathsf{T}\left(\begin{smallmatrix} \tilde{x}_y \\ \tilde{x}_z \end{smallmatrix}\right)$ yields

$$QH^\mathsf{T} \begin{pmatrix} \tilde{x}_y \\ \tilde{x}_z \end{pmatrix} - A^\mathsf{T}\lambda = -q.$$

Multiplying with H from the left yields

$$\underbrace{HQH^\mathsf{T}}_{=B} \begin{pmatrix} \tilde{x}_y \\ \tilde{x}_z \end{pmatrix} - \underbrace{HA^\mathsf{T}}_{=\begin{pmatrix} R \\ 0 \end{pmatrix}} \lambda = \underbrace{-Hq}_{=h}.$$

Because of the term $\binom{R}{0}$, only the first m components of λ can be recovered and we obtain the simplified equation

$$\boxed{R\lambda = -h_1 + B_{11}\tilde{x}_y + B_{12}\tilde{x}_z.}$$

As R is upper triangular, this can be easily solved (see above).

In summary, the algorithm for (QE) is

1. Compute the QR-decomposition of A^T: compute $H \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$, such that $HA^\mathsf{T} = \binom{R}{0}$.

   Define $h := -Hq = \binom{h_1}{h_2}$ and $B := HQH^\mathsf{T} =: \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22}\end{pmatrix}$, where $h_1 \in \mathbb{R}^m$ and $B_{11} \in \mathbb{R}^{m\times m}$.

2. Solve $R^\mathsf{T}\tilde{x}_y = b$ and $B_{22}\tilde{x}_z = h_2 - B_{21}\tilde{x}_y$ to obtain $\tilde{x} = H^\mathsf{T}\begin{pmatrix}\tilde{x}_y \\ \tilde{x}_z\end{pmatrix}$.

3. Solve $R\lambda = B_{11}\tilde{x}_y + B_{12}\tilde{x}_Z - h_1$ via forward substitution.

Problems with inequality constraints - the active set method

We consider

$$\min_{x\in\mathbb{R}^n} x^\mathsf{T}Qx + q^\mathsf{T}x \quad \text{subject to} \quad Ax = b, \quad hx \geqslant r, \tag{22}$$

where $Q \in \mathbb{R}^{n\times n}$ is symmetric and $A \in \mathbb{R}^{m\times n}$ with $m \leqslant n$ and $h \in \mathbb{R}^{p\times n}$.

The admissable set is $\Omega := \{x \in \mathbb{R}^n : Ax = b, hx - r \geqslant 0\}$. For convenience we introduce a new index set $J(x) = \mathcal{A}(x) \cap I$, which is comprised of the active indices. We have

$$L_\Omega(x) = \{d \in \mathbb{R}^n : Ad = 0, \ \langle g^j, d\rangle \geqslant 0 \ \forall j \in J(x)\}.$$

Define $G(x) := ((g^j)^\mathsf{T})_{j\in J(x)}$. Then $d \in L_\Omega(x)$ if and only if $Ad = 0$ and $G(x)d \geqslant 0$.

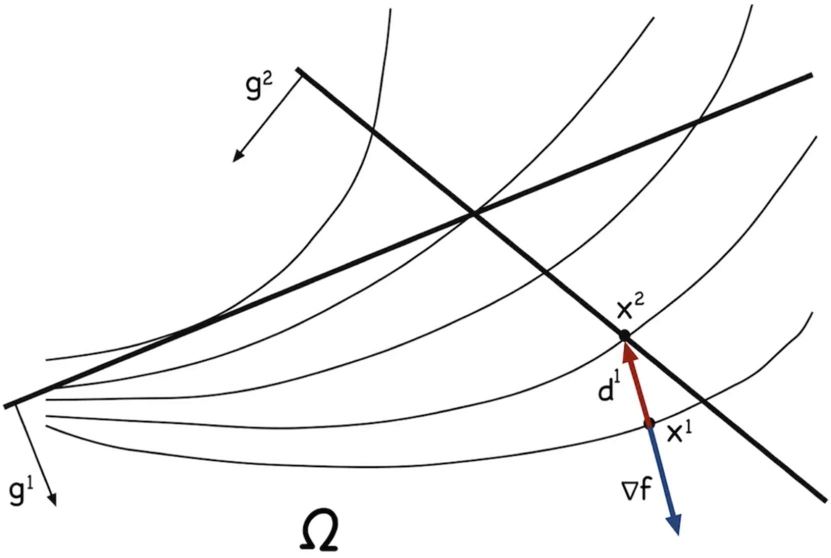


Figure 28: The thin black lines are the isolines of f and the thick black lines are two lines defining the admissable set which have normals g^1 and g^2.

Example 5.1.1 (Active set method)

① Suppose we start with $x^1 \in \operatorname{int}(\Omega)$. We do free optimisation, i.e. gradient descent.

② We reach $\partial\Omega$ in $x^{(2)}$, where g^2 gets active. We check if we are in a minimum, i.e. $\nabla f = \mu \nabla g^2$ for some $\mu > 0$, which is not the case in our picture.

We now search in the affine subspace defined by $\langle g^2, d \rangle = 0$, i.e. $x^2 + \{d \in \mathbb{R}^2 : \langle g^2, d \rangle = 0\}$, as we can't leave Ω but can "go along its boundary", as one can see in figure Figure 29.

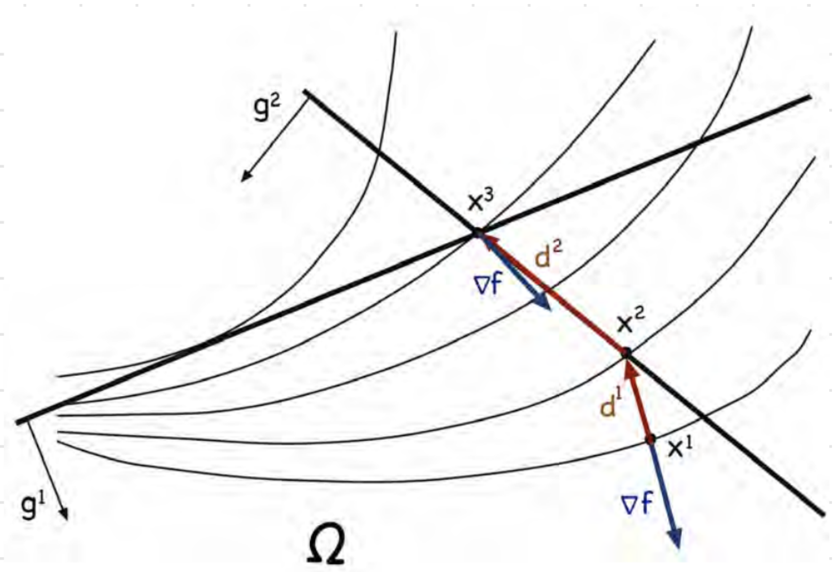


Figure 29: TODO

③ We reach x^3. In x^3, the inequality g^1 is activated. We check if the KKT conditions are satisfied, i.e. if $\nabla f = \mu_1 g^1 + \mu_2 g^2$. From figure Figure 29 we can see that $\mu_1 > 0$ and thus $\mu_2 < 0$. We know that such μ_1 and μ_2 must exist as g^1 and g^2 are linearly independent. Thus the KKT conditions are *not* satisfied.

How can we find a *new* search direction? We know that d has to satisfy $\langle g^1, d \rangle \geqslant 0$ and $\langle g^2, d \rangle \geqslant 0$ for admissibility and $\langle \nabla f, d \rangle < 0$ for descent. We thus obtain

$$0 > \langle \nabla f, d \rangle = \underbrace{\mu_1}_{>0} \underbrace{\langle g^1, d \rangle}_{\geqslant 0} + \underbrace{\mu_2}_{<0} \underbrace{\langle g_2, d \rangle}_{\geqslant 0}. \qquad \diamond$$

We would like to have a high value of descent. We see that this first terms is positive, which is not what we want. To minimise this term, we choose d such that $\langle g^1, d \rangle = 0$ and $\langle g^1, d \rangle > 0$, i.e. we deactivate g^2 and activate g^1.

④ We thus perform a subspace-search in g^1 and find a point x^4, which intersects an isoline of f: $\nabla f = \mu g^1$ with $\mu > 0$, i.e. the KKT conditions are satisfied.

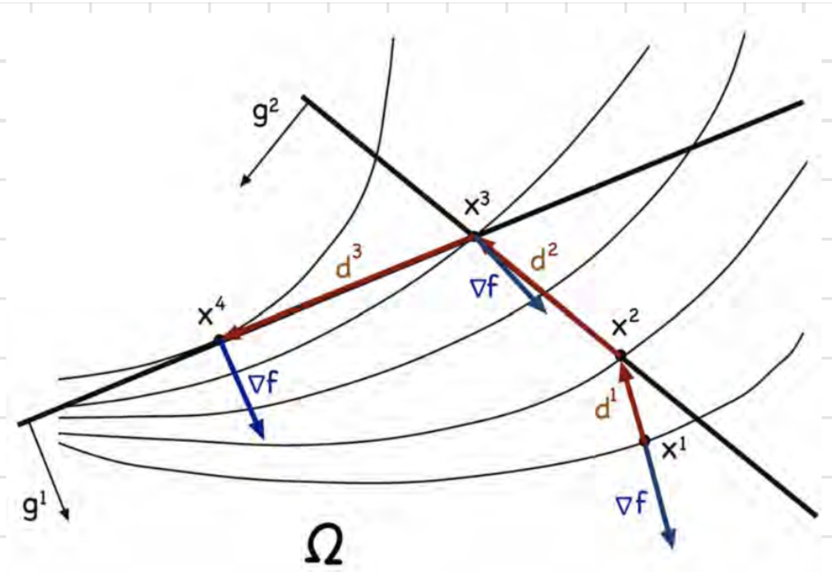


Figure 30: TODO

We will now precisely formula the active set method. Let x^k be the current iterate, $J_k := J(x^k) = \mathcal{A}(x^k) \cap I$ and $P_k := |J_k|$ the number of active inequalities. Furthermore, let $G_k := G(x^k) = ((g^i)^\mathsf{T})_{i\in J_k}$ and $B_k := \begin{pmatrix} A \\ G_k \end{pmatrix}$ the active linear system at the k-th iterate. 09.07

Note that $\ker(B_k) \subset \ker(A)$.

In x^k we solve the following problem

$$\min_{d\in\mathbb{R}^n} \underbrace{\frac{1}{2}\langle Qd, d\rangle + \langle Qx^k + q, q\rangle}_{=f(x^k+d)+c} \quad \text{subject to} \quad B_k d = 0. \tag{Q_k}$$

The solution of (Q_k) is a direction d^k and multipliers μ_j^k, where $j \in J_k$ for the inequality constraints and the Lagrange multipliers λ_i^k for the equality constraints, where $i \in \{1, \dots, m\}$. We add $\mu_j^k = 0$ if $j \notin J_k$. In summary we obtain the vectors $\lambda^k \in \mathbb{R}^m$ and $\mu^k \in \mathbb{R}^{P_k}$ or $\tilde{\mu} \in \mathbb{R}^p$ (filled with zeros).

Our assumptions are that B_k has maximal rank, i.e. $(a^i)_{i=1}^m$ and $(g^j)_{j\in J_k}$ are linearly independent. (Then (ACQ) is satisfied, so we can apply the nd we obtain unique multipliers.) To get a unique solution of the problem, we require that "system-matrix" Q is positive definite on $\ker(B_k)$, or, stronger, that Q is positive definite on $\ker(A)$.

In summary: (Q_k) has a unique solution and λ^k and μ^k are unique. The admissable set for (Q_k) is

$$\Omega_k := \{d \in \mathbb{R}^n : Ad = 0, G_k d = 0\} \subset L_\omega(x^k).$$

We see $\Omega_k \subset \Omega$ for all k, i.e. all directions from Ω_k are admissable for the original problem.

We now evaluate the KKT conditions for (Q_k), which are $\nabla f(x^k + d^k) - A^\mathsf{T}\lambda^k - G_k^\mathsf{T}\mu^k = 0$, as we only have equality constraints. This can be rewritten as

$$Q(d^k + x^k) + q - A^\mathsf{T}\Lambda^k - G_k^\mathsf{T}\mu^k = 0. \tag{23}$$

Thanks to the linear independence of the active gradients of the constraints, we can recover λ^k and μ^k uniquely.

We consider three cases.

① Assume $d^k = 0$ and $\mu^k \geqslant 0$. Then (23) becomes

$$\nabla f(x^k) - A^\mathsf{T}\Lambda^k - G_k^\mathsf{T}\mu^k = 0,$$

implying that x^k satisfies the KKT conditions. We have a quadratic problem, which is positive definite of $\ker(A)$, so we know that x^k is the unique solution by convexity of the problem.

② Consider $d^k = 0$ and $\mu^k \ngeqslant 0$, i.e. there exists a $j \in J_k$ such that $\mu_j^k < 0$. We choose $j \in J_k$ such that $\mu_j^k = \min_{i \in J_k} \mu_i^k < 0$ and "deactivate it", i.e. the new active index set is $\hat{J}_k := J_k \backslash \{j\}$. We solve the new corresponding problem $\hat{Q}_k$, yielding the solution $\hat{d}^k$ and multipliers $\hat{\mu}^k$ and $\hat{\lambda}^k$. We show that $\hat{d}^k \neq 0$.

If $d^k = 0$, we know by (23) that

$$Qx^k + q - A^\mathsf{T}\lambda^k - G_k^\mathsf{T}\mu^k = 0$$

and the KKT conditions for $(\hat{Q}_k)$ are

$$Q\hat{d}^k + Qx^k + q - A^\mathsf{T}\hat{\lambda}^k - G_k\hat{\mu}^k = 0.$$

If $\hat{d}^k = 0$, we obtain $A^\mathsf{T}\lambda^k + G_k^\mathsf{T}\mu^k = A^\mathsf{T}\hat{\lambda}^k + G_k^\mathsf{T}\hat{\mu}^k$. As μ^k has one component more than $\hat{\mu}^k$. In that component j we get that $\mu_j g^j$ is a linear combination of all other a^i and g^i. As $\mu_j \neq 0$, the same holds for g^j, which is a contradiction to the assumption that B_k has maximal rank. Thus $\hat{d}^k \neq 0$.

③ Assume $d^k \neq 0$. We show that d^k is a descent direction. By (23) we have

$$\nabla f(x^k) = Q^k x^k + q = -Qd^k + A^\mathsf{T}\lambda^k + G_k^\mathsf{T}\mu^k.$$

Multiplying by $d^k \neq 0$ yields

$$\langle \nabla f(x^k), d^k \rangle = -\underbrace{\langle Qd^k, d^k \rangle}_{<0} + \left\langle \begin{pmatrix} \lambda^k \\ \mu^k \end{pmatrix}, \underbrace{B_k d^k}_{=0} \right\rangle < 0. \tag{24}$$

This yields the following algorithm.

① Choose $x^0 \in \Omega$ and set $k := 0$

② Setup (Q_k) and compute d^k, λ^k, μ^k with the null space method.

③ If $d^k = 0$, $\mu^k \geqslant 0$, then stop, as x^k is a solution.

④ If $d^k = 0$ and $\mu^k \ngeqslant 0$.

ⓐ Let $\mu_j^k = \min_{i \in J_k} \mu_i^k$ and $J_k := J_k \backslash \{j\}$.

ⓑ Delete corresponding row in G_k and solve the new problem, which we call Q_k.

⑤ If $d^k \neq 0$, compute a step size σ_k and define $x^{k+1} := x^k + \sigma_k d^k$.

How can we compute the step size? Consider $x^{k+1} = x^k + \tau d^k$ for $\tau > 0$. Our desiderata for the step size τ are

① that τ yields the maximal descent, i.e. τ should be the $t > 0$ for which

$$\begin{aligned} f(x^k + td^k) &= \frac{1}{2}(x^k + td^k)^\mathsf{T} Q(x^k + td^k) + q(x^k + td^k) \\ &= f(x^k) + t\,\underbrace{(Qx^k + q)^\mathsf{T} d^k}_{=\langle \nabla f(x^k), d^k \rangle} + \frac{1}{2}t^2 (d^k)^\mathsf{T} Q d^k \\ &\overset{(24)}{=} f(x^k) + (-d^k)^\mathsf{T} Q d^k + \frac{1}{2}t^2 (d^k)^\mathsf{T} Q d^k \\ &= f(x^k) + \left(\frac{1}{2}t^2 - t\right)(d^k)^\mathsf{T} Q d^k \end{aligned}$$

is maximised. We have $\arg\min \frac{1}{2}t^2 - t = 1$, so we take $\tau = 1$.

② that τ is admissable; we require that

$$A(x^k + \tau d^k) = b \tag{25}$$

$$(g^j)^\mathsf{T}(x^k + \tau d^k) \geqslant r_j \quad \forall j \in \{1, \dots, p\}. \tag{26}$$

Since $Ax^k = b$ and d^k is an admissable direction, we have $A(x^k + \tau d^k) = Ax^k + \tau \cdot 0 = Ax^k$, so (25) is satisfied for all $\tau \geqslant 0$.

If $Gx^k \geqslant r$, i.e. $(g_j)^\mathsf{T} d^k \geqslant 0$, (26) holds for all $\tau \geqslant 0$. This is true especially for all $j \in J_k$.

Let $j \in \{1, \dots, p\} \backslash J_k$ and $(g^j)^\mathsf{T} d^k < 0$. Then $(g^j)^\mathsf{T}(x^k + \tau d^k) \geqslant r_j$. Thus $\tau (g^j)^\mathsf{T} d^k \geqslant r_j - (g^j)^\mathsf{T} x^k$, so we have to require that

$$\tau \leqslant \frac{r_j - (g^j)^\mathsf{T} x^k}{(g^j)^\mathsf{T} d^k}.$$

We thus define

$$I_k := \{ i \in \{1, \dots, p\} : \langle g_i, d^k \rangle \leqslant 0 \}$$

and

$$\tau_k = \min_{j \in I_k} \frac{r_j - (g^j)^\mathsf{T} x^k}{(g^j)^\mathsf{T} d^k}$$

if $I_k \neq \varnothing$ and ∞ else. Thus choose $\sigma_k := \min(\tau_k, 1)$.

We summarise our findings in the following theorem.

Theorem 5.1.1: Active set method

Let Q be positive definite on $\ker(A)$ and for all $x \in \Omega$, $B(x) = \begin{pmatrix} A \\ G(x) \end{pmatrix}$ have maximal rank. The the algorithm computes the solutions of (QLI) in finitely many steps.

Proof. See script. □

Remark 5.1.2 (Computation of an admissable initial value)
We solve the auxiliary problem

$$\begin{cases} \min \tilde{f}(x,y,z) = \frac{1}{2}\sum_{i\in E} y_i^2 + \frac{1}{2}\sum_{i\in I} z_i^2 \quad \text{subject to} \\ (a^i)^\mathsf{T} x + y_i = b_i \ \forall i \in E, \quad (g^i)^\mathsf{T} x + z_i \geqslant r_i \ \forall i \in I. \end{cases} \tag{AP}$$

where we have introduced slack variables y and z. Note that $(x^0, y^0, z^0) = (0, b, r)$ is admissable. If the original problem has an admissable point, then (AP) has solution in $(x, 0, 0)$. **(WHYY?)**

5.2 Nonlinear objective, linear equality constraints

13.07

Consider

$$\min_x f(x) \quad \text{subject to} \quad Ax = b, \tag{27}$$

where $A \in \mathbb{R}^{n\times m}$ has rank m. Then $\ell := \dim(\ker(A)) = n - m$. The admissable set is

$$\Omega := \{x \in \mathbb{R}^n : Ax = b\} = w + \ker(A),$$

where w is a special solution, i.e. $Aw = b$.

Recall that to find w and $\ker(A)$, we preformed at QR decomposition of A^T: $HA^\mathsf{T} = \left(\begin{smallmatrix} R \\ 0 \end{smallmatrix}\right)$, with $H = (Y^\mathsf{T}, Z^\mathsf{T})^\mathsf{T}$, with $Y \in \mathbb{R}^m$ and $Z \colon \mathbb{R}^\ell \to \ker(A)$, thus Z is called nullspace matrix. We then considered the unconstrained problem

$$\min_{z\in\mathbb{R}^\ell} \underbrace{f(w + Zz)}_{=:F(z)}$$

Consider $Z := (\zeta^1, \dots, \zeta^\ell)$, where $\zeta^i \in \mathbb{R}^n$. We have

$$\frac{\partial F}{\partial z_i} = \frac{\partial F}{\partial z_i} f\left(w + \sum_{k=1}^{\ell} \zeta^k z_k\right) = \nabla f^\mathsf{T} \zeta^i = (\zeta^i)^\mathsf{T} \nabla f$$

and thus $\nabla_z F(z) = Z^\mathsf{T} \nabla x f$. Similarly, we get

$$\nabla_{zz} F(z) = Z^\mathsf{T} f_{xx} Z.$$

Consider a descent method

$$z^{k+1} = z^k + \sigma_k v^k,$$

where v^k is a descent direction in z^k. Then for $d^k = Zv^k$ we have

$$\nabla f(x^k)^\mathsf{T} d^k = \nabla f(x^k)^\mathsf{T} Z v^k = \left(Z^\mathsf{T} \nabla f(x^k)\right)^\mathsf{T} v^k = \nabla F(z^k)^\mathsf{T} v^k < 0,$$

i.e. d^k is a descent direction for f in x^k. Thus we can instead of preforming the descent method in the reduced setting in $\mathbb{R}^\ell$, we can preform in in $\mathbb{R}^n$.

This yields the following (reduced descent method) algorithm.

① Choose $x^0 \in \Omega$, compute Z and define $k := 0$.

② If $Z^\mathsf{T} \nabla f(x^k) = 0$, stop.

③ Compute a descent direction $d^k = Zv^k$ and an efficient step-size σ_k and set $x^{k+1} := x^k + \sigma_k d^k$ and $k \to k+1$ and go to step ②.

Remark 5.2.1 (Reduced gradient method)
We have

$$v^k = -\nabla F(z^k) = -Z^\mathsf{T} \nabla f(x^k)$$

and thus

$$d^k = Zv^k = -ZZ^\mathsf{T} \nabla f(x^k),$$

so we don't need v^k explicitly and can instead only rely on the null-space matrix Z.

Example 5.2.2 (Nonlinear regression with cubic splines)
Consider measurements ξ_i, η_i with $i \in \{1, \dots, m\}$. Our model function is $\eta(\xi) = g(x, \xi)$, where g is a natural cubic spline. Without loss of generality let $\xi_1 < \xi_2 < \ldots < \xi_m$ and cover the interval $[\xi_1, \xi_m]$ by nodes τ_k with $k \in \{1, \dots, N\}$ with $\tau_0 \leqslant \xi$ and $\tau_N \geqslant \xi_m$. We define $\Delta\tau_i = \tau_{i+1} - \tau_i$. We demand that on $[\tau_i, \tau_{i+1}]$, g is a third order polynomial and that $g(x, \cdot) \in \mathcal{C}^2([\tau_0, \tau_N])$.

On $[\tau_i, \tau_{i+1}]$, we define

$$g_i(x, \tau) := \frac{1}{\Delta\tau_i} \left(\gamma_{i+1}(\tau - \tau_i)^3 + \gamma_i(\tau_{i+1} - \tau)^3\right) + \beta_i(\tau - \tau_i) + \alpha_i$$

with $\gamma_0 = \gamma_N = 0$. Since $g \in \mathcal{C}^2([\tau_0, \tau_N])$, we know that g, g' and g'' have to be continuous in nodes τ_i. This gives conditions for the coefficients. All in all we get

$$\min f(x) = \sum_{i=1}^{m} (\eta_i - g(x, \xi_i))^2 \quad \text{subject to}$$
$$\beta_i = \beta_{i-1} + 3\gamma_i(\Delta\tau_i - \Delta\tau_{i-1}) \quad \text{and}$$
$$\alpha_i = \alpha_{i-1} + \beta_{i-1}\Delta\tau_{i-1} + \gamma_i(\Delta\tau_{i-1} - \Delta\tau_i) \quad \forall i \in \{1, \dots, N-1\}.$$

We end up with a free optimisation problem with a parameter vector $x = (\alpha_0, \beta_0, \gamma_1, \dots, \gamma_{N-1})^\mathsf{T}$. ◇

5.3 Nonlinear objective and inequality constraints

Consider

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad Ax = b \quad \text{and} \quad Gx \geqslant r,$$

where $A \in \mathbb{R}^{m \times n}$, $G \in \mathbb{R}^{p \times m}$, $b \in \mathbb{R}^m$ and $r \in \mathbb{R}^p$.

Motivation. Consider

$$\min_{x \in \mathbb{R}^n} f(x).$$

The necessary optimality condition is

$$\nabla f(\tilde{x}) = 0,$$

which we solve with Newton's method: consider the linearisation

$$\nabla f(x^k) + f''(x^k)(x - x^k) = 0. \tag{28}$$

The solution of (28) **is the next iterate.** We now interpret (28) as the necessary optimality condition of the quadratic problem

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}(x-x^k)^\mathsf{T} f''(x^k)(x-x^k) + \nabla f(x^k)(x-x^k).$$

This formulation allows us to include constraints, by instead of $\mathbb{R}^n$ minimising over some set $\Omega \subset \mathbb{R}^n$, whose solution is $x = x^{k+1}$, where

$$\Omega := \{x \in \mathbb{R}^n : Ax = b, Gx \geqslant r\}.$$

Defining $d := x - x^k$, we have $x = x^k + d$. As $Ax^k = b$, we demand that $Ad = 0$. We obtain

$$\min \langle f(x^k), d \rangle + \frac{1}{2}\langle d, f''(x^k)d \rangle \quad \text{subject to} \quad Ad = 0,\ Gx^k + bd \geqslant r, \tag{QP_k}$$

which we can solve with the technique learned in the previous sections.

The necessary assumptions are that $f''(x^k)$ is positive definite on $\ker(A)$, so then there exists a unique solution (QP_k). We further assume that $B(x) := \begin{pmatrix} A \\ G(x) \end{pmatrix}$ has maximal rank, the multipliers λ and μ are unique.

The optimality conditions for (QP_k) are that

$$f''(x^k)d^k + \nabla f(x^k) - A^\mathsf{T}\lambda^{k+1} - G^\mathsf{T}\mu^{k+1} = 0$$

and $\mu^{k+1} \geqslant 0$ and (complementarity) $\langle \mu^{k+1}, G(x^k + d^k) - r \rangle = 0$.

We consider two cases for the solution d^k.

(1) If $d^k = 0$, then the optimality conditions above reduce to the KKT conditions and thus x^k is the solution of (PLI).

(2) If $d^k \neq 0$, then d^k is a descent direction: by the KKT conditions

$$\nabla f(x^k) = -f''(x^k)d^k + A^\mathsf{T}\lambda^{k+1} + G^\mathsf{T}\mu^{k+1}.$$

Multiplying by d^k yields

$$\langle \nabla f(x^k), d^k \rangle = -\underbrace{\langle f''(x^k)d^k, d^k \rangle}_{>0} + \langle \lambda^{k+1}, \underbrace{Ad^k}_{=0} \rangle + \underbrace{\langle \mu^{k+1}, Gd^k \rangle}_{\leqslant 0} < 0,$$

where the last inequality is by complementarity: we have

$$0 = \underbrace{\langle \underbrace{\mu^{k+1}}_{\geqslant 0}, \underbrace{Gx^k - r}_{\geqslant 0} \rangle}_{\geqslant 0} + \langle \mu^{k+1}, Gd^k \rangle.$$

As d^k is admissable, $x^{k+1} = x^k + d^k$.

One also does a further step size computation with $\tau \geqslant 1$. Often we take $\tau_k = 1$ and then this is called SQP method.

SQP method

The assumptions for $\tilde{x}$ being a local minimum of (QLI) is that

(1) $f \in \mathcal{C}^2$ in $B_\delta(\tilde{x})$

(2) f'' is Lipschitz on $B_\delta(\tilde{x})$ (comes from Newton's method)

(3) $B(x)$ has full rank

(4) $d^\mathsf{T} f''(\tilde{x})d \geqslant \alpha|d|^2$ for all d such that $Ad = 0$ and $G(\tilde{x})d = 0$.

(5) (strict complementarity) if $\langle g^i, \tilde{x} \rangle = r_i$, then $\tilde{\mu}_i > 0$.

Theorem 5.3.1

Assuming ① - ⑤, then the SQP method converges locally quadratically, i.e.

$$|x^{k+1}-\tilde{x}|+|\lambda^{k+1}-\tilde{\lambda}|+|\mu^{k+1}-\tilde{\mu}| \leqslant c\big(|x^k-\tilde{x}|+|\lambda^k-\tilde{\lambda}|+|\mu^k-\tilde{\mu}|\big)$$

for all $x \in B_{\delta_1}(\tilde{x})$.

Here we need a globalisation strategy; we would need a damped approach to first go into the region of quadratic convergence, so we might do a globalisation with so gradient steps, for example.

Index