Definition

Level set of $f\colon \mathbb{R}^d \supset D \to \mathbb{R}$

Nonlinear optimisation

Theorem

Compact level sets and minima

Nonlinear optimisation

Example

Linear regression

Nonlinear optimisation

Theorem

For convex problems, local minima are global

Nonlinear optimisation

Theorem

Strictly convex problems are uniquely solvable

Nonlinear optimisation

Theorem

Variational inequality for the directional derivative

Nonlinear optimisation

Theorem

Second order necessary condition

Nonlinear optimisation

Theorem

Sufficient second order condition

Nonlinear optimisation

Theorem

Quadratic growth condition

Nonlinear optimisation

Theorem

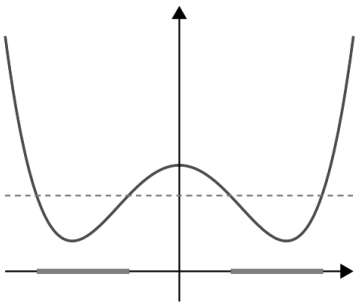
Tangents of convex functions

Nonlinear optimisation

Let $f\colon \mathbb{R}^n \supset D \to \mathbb{R}$ by a continuous function and Ω **closed**. If there exists a $\omega \in \Omega$, such that $\boldsymbol{\mathcal{N}(f, f(\omega))}$ **is compact** there exists a global minimum of f on Ω.

Let $a := \inf_{x\in\Omega} f(x) \leqslant f(\omega)$. As Ω is closed, $N := \Omega \cap \mathcal{N}(f, f(\omega))$ is compact and we have $a = \inf_{x\in N} f(x)$. By the theorem of Weierstrass there exists a $\hat{x} \in \Omega$ with $\inf_{x\in N} f(x) = f(\hat{x})$.

For $a \in \mathbb{R}$, $\mathcal{N}(f, a) := \{x \in D : f(x) \leqslant a\}$ is the **level set** of f with respect to a.



Let f be convex, $D \neq \varnothing$ open and $\Omega \subset D$ **convex**. Any local minimum of f is global. The set of solutions is convex.

Let $x \in \Omega$ be a local minima. Then $\exists r > 0$ with $f(x) \leqslant f(y)$ for all $y \in \Omega \cap B(x, r)$. Let $y \in \Omega$ and $t > 0$ so small, that $x_t := x + t(y - x) \in B(x, r)$. Since Ω is convex, $x_t \in \Omega$ for all $t \in [0, 1]$. Since f is convex, $f(x) \leqslant f(x_t) = f((1-t)x+ty) \leqslant (1-t)f(x) + tf(y)$, which yields $f(x) \leqslant f(y)$.

If $x, y \in \Omega$ are solutions, for all $z \in \Omega$ $f((1-t)x + ty) \leqslant (1-t)f(x)+tf(y) \leqslant (1-t)f(z)+tf(z) = f(z)$, so $(1-t)x+ty$ is a minimum, too.

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

$$H := 2\begin{pmatrix} \sum_{k=1}^m x_k^2 & \sum_{k=1}^m x_k \\ \sum_{k=1}^m x_k^2 & m \end{pmatrix}, b := -2\begin{pmatrix} \sum_{k=1}^m x_k y_k \\ \sum_{k=1}^m y_k \end{pmatrix}$$

$c := \sum_{k=1}^m y_k^2$. If two x_k are different, H is positive definite, as all principal minors are positive (CS). Thus F is strictly convex and has a unique minimum.

If $x \in D$ is a local minimum f and f is directionally differentiable in x, $f'(x; h) \geqslant 0$ for all $h \in \mathbb{R}^n$.

As D is open, $\exists r > 0$ with $f(y) \geqslant f(x)$ for all $y \in B(x, r)$. For $h \in \mathbb{R}^n$ and small t we have $x + th \in B(x, r)$ and thus $f(x + th) - f(x) \geqslant 0$, i.e. $\frac{f(x+th)-f(x)}{t} \geqslant 0$.

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

The absolute has a minimum in 0, but is not differentiable there, but we have $|\cdot|'(x; h) = |h| \geqslant 0$ for all h.

If $f \in C^1$ and x is a local min, $f'(x; h) = \nabla f(x)^\mathsf{T} h \geqslant 0\ \forall h \in \mathbb{R}^n$ (var. ineq.). Taking $h = -\nabla f(x)^\mathsf{T} h$, we get $\nabla f(x) = 0$.

Let f by strictly convex. If x is a minimum of f, it is unique and thus strict.

Let $x \neq y \in \Omega$ be two (by the previous theorem, global) minima of f and $a := \min_{x\in\Omega} f(x)$. Then $f\left(\frac{x+y}{2}\right) < \frac{f(x)+f(y)}{2} = a$, which that only x and y are minima of f.

Examples. The exponential function is strictly convex (AM-GM), but has no minimum. If H is positive definite, $\frac{1}{2}x^\mathsf{T} H x + b^\mathsf{T} x$ is strictly convex.

Let f be $\mathcal{C}^2$ in a neighbourhood of $x \in D$, $\nabla f(x) = 0$ and $f''(z)$ be **positive semidefinite for all** $z \in B(x, \delta)$ with some $\delta > 0$. Then x is a local minimum of f.

For $y \in B(x, \delta)$ and $\theta \in [0, 1]$

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

by Taylor's theorem.

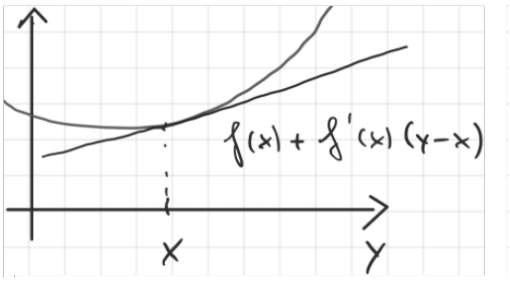
$x = 0$ is a local minimum of $f(x) := x^{2p}$, where $p \in \mathbb{N}_{\geqslant 2}$. We have $f'(0) = f''(0) = 0$, which is not positive definite.

Let f be $\mathcal{C}^2$ in a neighbourhood of $x \in D$ and x a local minimum of f. Then we have $\nabla f(x) = 0$ and that $f''(x)$ is **positive semidefinite**.

For $h \in \mathbb{R}^n$ let $g(t) := f(x + th)$. Then $g \in \mathcal{C}^2$ has a local minimum in $t = 0$. By Taylor $\exists\theta \in [0, 1]$ with $g(t) = g(0) + g'(0)t+\frac{t^2}{2}g''(\theta t)$. As x is a local minimum of g, $0 \leqslant \frac{g(t)-g(0)}{t^2} = \frac{1}{2}g''(\theta t)$. The continuity of g'' yields $g''(0) = h^\mathsf{T} f''(x)h \geqslant 0$ for $t \searrow 0$.

$f(x) = x^4$ has a global minimum in $\tilde{x} = 0$, but $f''(\tilde{x}) = 0$.

Let f be differentiable on D. Then f is (strictly) convex on Ω iff $f(y) \overset{(>)}{\geqslant} f(x) + \nabla f(x)^\mathsf{T}(y - x)$ for all $x, y \in \Omega$.



Let f be $\mathcal{C}^2$ in a neighbourhood of $x \in D$, $\nabla f(x) = 0$ and $f''(x)$ **positive definite**. Then $\exists r, a > 0$ such that $f(y) \geqslant f(x) + a\|y - x\|^2$ for all $x \in B(x, r)$, so x is a strict minimum.

Taylor: $f(y) = f(x) + \frac{1}{2}(y - x)f''(x + \theta(y - x))(y - x)$ and

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

Theorem

Convex variational inequality

Nonlinear optimisation

Definition

Descent direction

Nonlinear optimisation

Examples + Lemma

Descent direction

Nonlinear optimisation

Algorithm

General descent algorithm

Nonlinear optimisation

Assumptions

ALC and AFD

Nonlinear optimisation

Definition

Efficient step size

Nonlinear optimisation

Definition

(strictly) Gradient-related descent direction

Nonlinear optimisation

Assumptions

(ALG) and (AHP)

Nonlinear optimisation

Lemma

Convergence results for general descent methods

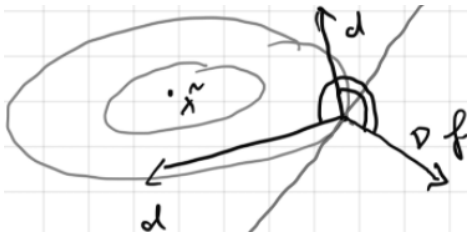
Nonlinear optimisation

Theorem

Convergence of descent algorithms

Nonlinear optimisation

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



If $\tilde{x}$ is a local minimum, we have $\nabla f(\tilde{x})^\mathsf{T}(x - \tilde{x}) \geqslant 0$, so a necessary condition is that there **exists no descent direction**.

Let f be differentiable in D and convex in $\Omega \subset D$. Then $x \in \Omega$ is a minimiser of f if and only if $\nabla f(x)^\mathsf{T}(y - x) \geqslant 0 \ \forall y \in \Omega$.
" $\Longrightarrow$ ": If x be a local solution, then $x + t(y - x) = (1 - t)x + ty \in \Omega$ for all $t \in [0, 1]$, $y \in \Omega$. For small $t > 0$ $\frac{f(x+t(y-x))-f(x)}{t} \geqslant 0$. Take $t \searrow 0$. (convexity of f not needed)
" $\Longleftarrow$ ": As is f convex and all tangents lie below the graph, we have $f(y) - f(x) \geqslant \nabla f(x)^\mathsf{T}(y - x) \geqslant 0$ and by a previous theorem x is a global minimum.
For $x \in \operatorname{int}(\Omega)$, we have $\nabla f(x)^\mathsf{T} d \geqslant 0$ for all directions $d \in \mathbb{R}^n$ and thus $\nabla f(x) = 0$.

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.

Step ② is only of academic nature, e.g. use $|\nabla f(x)| < \varepsilon$ instead.

For a descent direction d $\exists c > 0$ with $f(x + ad) < f(x)$ for all $a \in (0, c]$: 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 $a \in (0, c]$.

The reverse direction of this lemma doesn't hold, take $x \mapsto -x^2$, $\tilde{x} = 0$, $d := 1$.

The **antigradient / steepest descent** $d = -\nabla f(x) \neq 0$ and $-A^{-1}\nabla f(x)$ for positive definite A are descent directions.

Assume (ALG). A step size with

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

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

(ALC): for $x^{(0)} \in \mathbb{R}^d$ the level set $\mathcal{N}(f, f(x^{(0)}))$ is compact.
(AFD): We have $f \in \mathcal{C}^1$ on an open, convex set $D_0 \supset \mathcal{N}(f, f(x^{(0)}))$.

In descent methods, $f(x^{k+1}) < f(x^{(k)})$ and thus $x^{(k)} \in \mathcal{N}\left(f, f(x^{(0)})\right)$. If (ALC) holds, $(x^{(k)})_{k \in \mathbb{N}}$ and $(f(x^{(k)}))_{k \in \mathbb{N}}$ are bounded.

(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).

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

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

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

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

The antigradient (and assuming (AHP), the Newton descent direction) is strictly gradient related ($c_3 = c_4 = 1$).

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. 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** minimiser of f. 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

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

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.

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

② F has a unique minimiser $\tilde{x}$, which is the **only stationary point** of f,

③ $\frac{a}{2}|x - \tilde{x}|^2 \leqslant f(x) - f(\tilde{x}) \leqslant \frac{1}{2a}|\nabla f(x)|^2 \ \forall x \in \mathcal{N}(f, f(x^{(0)}))$.

Definition

Exact step size

Definition

Armijo step size and algorithm

Definition

Powell step size

Assumptions

R1 and R2: Sufficiently fast decay

Algorithm

Powell Algorithm

Lemma

Origin of the term steepest descent

Algorithm + remark

The gradient method (steepest descent)

Algorithm + interpretation

Damped Newton method

Example

Steepest descent direction vs Newton direction for a quadratic function.

Algorithm

Variable metric method

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

We consider $\varphi(\sigma) := f(x+\sigma d)$. The exact step size $\sigma_E > 0$ is such that $\varphi'(\sigma_E) = 0$ and $\varphi'(s) < 0$ for $s \in [0, \sigma_E)$. The exact step size is the "first" local minimum of φ.
If ∇f is L-cts (AGL), we have $\sigma_E \geqslant -\frac{\nabla f(x)^\mathsf{T} d}{L|d|^2}$ and $f(x+\sigma_E d) \leqslant f(x) - \frac{1}{2L}\left(\frac{\nabla f(x)^\mathsf{T} d}{|d|}\right)^2$, so σ_E is **efficient**. If $f(x) = \frac{1}{2}x^\mathsf{T} H x + b^\mathsf{T} x$ with positive definite H, $\sigma_E = -\frac{\nabla f(x)^\mathsf{T} d}{d^\mathsf{T} H d}$

(R1): There exists a constant $c_1 > 0$ independent of k, such that $f(x^{(k)} + \sigma_k d^{(k)}) - f(x^{(k)}) \leqslant c_1\sigma_k \nabla f(x^{(k)})^\mathsf{T} d^{(k)} < 0$.

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

(R2): There exists a constant $c_2 > 0$ 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 σ_k satisfies the **sufficient decrease condition**: $f(x^k + \sigma_k d^k) \leqslant f(x^k) - c_1 c_2 \frac{(\nabla f(x^k)^\mathsf{T} d^k)^2}{|d^k|^2}$.

σ_P should fulfil (R1) and $\nabla f(x+\sigma d)^\mathsf{T} d \geqslant \beta \nabla f(x)^\mathsf{T} d$ with $0 < \delta < \beta < 1$. 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 $I_3 := (s_2, \infty)$.

$G_1(\sigma) := \begin{cases} \frac{f(x+\sigma d) - f(x)}{\sigma \nabla f(x)^\mathsf{T} d}, & \sigma > 0, \\ 1, & \sigma = 0, \end{cases}$ (cts.), $G_2(\sigma) := \frac{\nabla f(x+\sigma d)^\mathsf{T} d}{\nabla f(x)^\mathsf{T} d}$. 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. (ALC), (AGL) imply that σ_P is an efficient step size.

$\arg\min_{d\in\mathbb{R}^n, |d|=1} \nabla f(x)^\mathsf{T} d = -\frac{\nabla f(x)}{|\nabla f(x)|}$.

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

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

(a) If $G_1(\sigma) \geqslant \delta$ and $G_2(\sigma) \leqslant \beta$, stop and let $\sigma_P := \sigma_0$.

(b) 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.

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

(a) If $\sigma_j \in I_2$, stop and set $\sigma_P := \sigma_j$.

(b) If $\sigma_j \in I_1$, set $a_{j+1} := \sigma_j$ and $b_{j+1} := b_j$.

(c) 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.

Like gradient method but instead $d^k = -f''(x^k)^{-1}\nabla f(x^k)$.

Let $A := f''(x)$ be SPD and $\langle x, y\rangle_A := x^\mathsf{T} A y$. We have $\tilde{d} := -\frac{A^{-1}\nabla f(x)}{|A^{-1}\nabla f(x)|_A} = \arg\min_{|d|_A=1} \nabla f(x)^\mathsf{T} d$.

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

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

We have $\nabla f(x)^\mathsf{T}\tilde{d} = -|A^{-1}\nabla f(x)|_A$.

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. Define $x^{k+1} = x^k + \sigma_k d^k$, $k \to k+1$, return to ②.

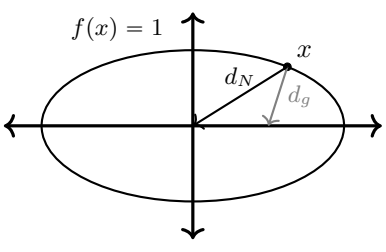
After initially fast decrease, one observes slow convergence especially for functions with e.g. ellipse-shaped isolines. We e.g. have $0 = \varphi'(\sigma_E) = \nabla f(\underbrace{x^k + \sigma_E d^k}_{=x^{k+1}})^\mathsf{T} d^k = d^{k+1} d^k$, i.e. $d^{k+1} \perp d^k$, which leads to the slow convergence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

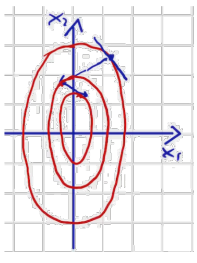
We want to **account for curvature information** (f'') **without having to compute the second derivative**.

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.



vs.


The grey anti-gradient direction $d_g = -\nabla f(x)$, (orthogonal to isolines of f), is not optimal. The NEWTON direction is better: $d_N = -f''(x)^{-1}\nabla f(x) = -H^{-1}Hx = -x$.

Algorithm

BFGS-method

Nonlinear optimisation

Definition

H-Orthogonality

Nonlinear optimisation

Theorem

BFGS method for quadratic problems

Nonlinear optimisation

Theorem

Properties of the CG method

Nonlinear optimisation

Algorithm

CG method

Nonlinear optimisation

Algorithm

Trust region Newton method

Nonlinear optimisation

Idea

Trust region method

Nonlinear optimisation

Definition

Active / inactive inequality constraints and the active set

Nonlinear optimisation

Definition

Tangent cone

Nonlinear optimisation

Definition

Admissible approximation

Nonlinear optimisation

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

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 the rank-2-update $A^{(k+1)} = A^{(k)} - \frac{A^{(k)}s^{(k)}(A^{(k)}s^{(k)})^\mathsf{T})}{(s^{(k)})^\mathsf{T}A^{(k)}s^{(k)}} - \frac{y^{(k)}(y^{(k)})^\mathsf{T}}{(y^{(k)})^\mathsf{T}s^{(k)}}$. Set $k \to k+1$ and go to step ②.

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

① $d^{(k-1)} \neq 0$ and $d^{(0)}, \dots, d^{(k)}$ are H-orthogonal,

②

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

③ $f(x^{(k)}) = \min_{z\in V_k} f(x^{(0)} + z)$.

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 $A^{(n)} = H$.

Given: $0 < \delta_1 < \delta_2 < 1$, $\sigma_1 \in (0,1)$, $\sigma_2 > 1$, $\sigma_0 > 0$, $x^{(0)} \in \mathbb{R}^n$.

① $d^{(k)}$ = solution of $\min_{|d|\leqslant\rho_k} f_k(d)$. If $f(x^{(k)}) = f_k(d^{(k)})$, then stop.

② $r_k := \frac{f(x^{(k)}) - f(x^{(k)}+d^{(k)})}{f(x^{(k)}) - f_k(x^{(k)}+d^{(k)})}$. 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.

③ If $r_k < \delta_1$ (unsuccessful), choose $\rho_{k+1} \in (0, \delta_1\rho_k)$, $x^{(k+1)} = x^{(k)}$, $\nabla f(x^{(k+1)}) = \nabla f(x^{(k)})$, $f''(x^{(k+1)}) = f''(x^{(k)})$. $k = k+1$, go to ②.

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 $d^{(k+1)} = -\nabla f(x^{(k+1)} + \beta_k d^{(k)}$. Set $k \to k+1$ and return to ②.

We consider $\min_{x\in\mathbb{R}^n} f(x)$ subject to $\begin{cases} c_i(x) = 0, & i \in E, \\ c_i(x) \geqslant 0, & i \in I \end{cases}$ 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 $\Omega = \{x \in \mathbb{R}^n : c_i(x) = 0, i \in E, c_i(x) \geqslant 0, i \in I\}$,

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

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 $\min_{|d|\leqslant\rho_k} f_k(d)$,
- update $x^{(k+1)} = x^{(k)} + d^{(k)}$.

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 $n \in \mathbb{N}$.

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

The tangent cone is a cone ($\tilde{t_k} := \frac{1}{a} t_k$).

If $x \in \operatorname{int}(\Omega)$, then $T_\Omega(x) = \mathbb{R}^n$.

Theorem

Variational inequality - General case

Nonlinear optimisation

Definition

Linearised cone

Nonlinear optimisation

Assumptions

ACQ and LICQ

Nonlinear optimisation

Theorem

KKT conditions

Nonlinear optimisation

Theorem

$T_\Omega(\tilde{x}) \subset L_\Omega(\tilde{x})$

Nonlinear optimisation

Proof

of the KKT theorem

Nonlinear optimisation

Lemma

Farkas

Nonlinear optimisation

Definition

Critical cone

Nonlinear optimisation

Theorem

Second order necessary condition for constraint problems

Nonlinear optimisation

Theorem

Second order sufficient optimality condition for constraint problems

Nonlinear optimisation

For $x \in \Omega$, the **linearised cone of Ω in $x \in \Omega$** is

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

Thus $L_\Omega(x)$ for $\Omega := \{x \in \mathbb{R}^n : h(x) = 0, g(x) \leqslant 0\}$ depends on g and h, whereas $N_\Omega(x)$ and $T\Omega(x)$ don't.

Let $\hat{x} \in \Omega$ be a solution of the constrained problem and $f \in \mathcal{C}^1$. Then $\nabla f(\hat{x})^\mathsf{T} d \geqslant 0$ holds for all $d \in T_\Omega(\hat{x})$.
For $d \in T_\Omega(\hat{x})$ we have by TAYLORs theorem,

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

Let $\hat{x}$ be solution to the constrained problem, 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 $\tilde{\lambda}_i c_i(\hat{x}) = 0$ for all $i \in E \cup I$ (complementarity).

Let $x \in \Omega$.
ABADIE constraint qualification (ACQ): $T_\Omega(x) = L_\Omega(x)$.
Linear independence constraint qualification (LICQ): $\{\nabla c_i(x) : i \in \mathcal{A}(x)\}$ is **linearly independent**.
LICQ $\implies$ ACQ.

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 FARKAS lemma 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 the variational inequality, 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.

W.l.o.g. assume $c_i(x)$, $i \in \{1, \dots, m\}$ be the **active** constraints in $\tilde{x}$. Let $d \in T_\Omega(\tilde{x})$. For k **sufficiently large** and c_i **is an equality constraint**, by TAYLOR expansion, $\exists \alpha \in [0, 1]$

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

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 $d \in L_\Omega(\tilde{x})$.

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

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

We have $w \in C(\tilde{x}, \tilde{\lambda})$ if and only if $\nabla c_i(\tilde{x})^\mathsf{T} w = 0 \ \forall i \in E$, $\forall i \in \mathcal{A}(\tilde{x}) \cap I$ s.th. $\tilde{\lambda}_i > 0$ and $\nabla c_i(\tilde{x})^\mathsf{T} w = 0 \ \forall i \in \mathcal{A}(\tilde{x}) \cap I$ s.th. $\tilde{\lambda}_i = 0$.
For $d \in C(\tilde{x}, \tilde{\lambda})$ we have $\nabla f(\tilde{x})^\mathsf{T} d = \sum_{i \in \mathcal{A}(\tilde{x})} \tilde{\lambda}_i \nabla c_i(\tilde{x})^\mathsf{T} d = 0$. **Thus $C(\tilde{x}, \tilde{\lambda})$ contains all directions where, based on first order information, we cannot decide if f decreases or increases.**

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 $C^\mathsf{T} d = 0$.

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 the constrained problem.

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

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

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

Lemma

(ACQ) holds for affine linear constraints $c_i(x) = a_i^\mathsf{T} x + b_i$ for $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$.

Nonlinear optimisation

Assumption

Mangasarian-Fromovitz

Nonlinear optimisation

Lemma

LICQ implies MFCQ

Nonlinear optimisation

Without proof

Implications between constraint qualifications

Nonlinear optimisation

Definition

Normal cone

Nonlinear optimisation

Theorem

Convergence of the Newton method

Nonlinear optimisation

example

Step size decreases to fast

Nonlinear optimisation

Theorem

Taylor in $\mathbb{R}$ with remainder

Nonlinear optimisation

Definition

Stationary points

Nonlinear optimisation

Convergence analyisis steps for descent methods

Nonlinear optimisation

(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*.

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 $i \in I \setminus \mathcal{A}(\tilde{x})$, then $c_i(\tilde{x}) > 0$. Then $\exists t_0 > 0$ such that $c_i(\tilde{x} + tw) > 0$ $\forall t \in [0, t_0]$, 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 $c_i(x^{(k)}) = c_i(x^{(k)}) - c_i(\tilde{x}) = a_i^\mathsf{T}(x^{(k)} - \tilde{x}) = \frac{t_0}{k} a_i^\mathsf{T} w \geqslant 0$ since $c_i(\tilde{x}) = 0$ and $w \in L_\Omega(\tilde{x})$, 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 $w \in T_\Omega(\tilde{x})$.

(LICQ) $\implies$ (MFCQ) $\implies$ (ACQ).
If (SQC) holds in $\tilde{x} \in \Omega$, then (MFCQ) holds.

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.

Let f'' **be Lipschitz continuous** in a neighbourhood of a local minimum $\hat{x}$ of f and let $f''(\hat{x})$ be **positive definite**. Then the NEWTON method

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

converges locally quadratically to $\hat{x}$.

$d^{(k)} := f''(x^{(k)})^{-1} \cdot (-\nabla f(x^{(k)}))$ is the **Newton direction**.

The **damped Newton method** is $x^{(k+1)} = x^{(k)} - \sigma_k f''(x^{(k)})^{-1} \nabla f(x^{(k)})$ with $\sigma_k < 1$.

For $x \in \Omega$, $N_\Omega(x) := \{v \in \mathbb{R}^n : v^\mathsf{T} w \leqslant 0 \ \forall w \in T_\Omega(x)\}$ is the **normal cone** to $T_\Omega(x)$. The elements of $N_\Omega(x)$ are **normal vectors**.

Let $\tilde{x}$ be a local solution to the constraint problem. Then $-\nabla f(\tilde{x}) \in N_\Omega(\tilde{x})$.
By the variational inequality $\nabla f(\tilde{x}) d \geqslant 0$, i.e. $-\nabla f(\tilde{x}) d \leqslant 0$ holds for all $d \in T_\Omega(\tilde{x})$.

Let $I \subset \mathbb{R}$ be an interval and $f \colon I \to \mathbb{R}$ in $\mathcal{C}^{n+1}(I)$. Then there exits a $\theta \in [0, 1]$ such that

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!} (x-a)^k + \frac{f^{(n+1)}(a + \theta(x-a))}{(n+1)!} (x-a)^{n+1}.$$

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 $\frac{1}{2}$:

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

We want to show $\nabla f(x^{(k)}) \to 0$. We first show that $\frac{\nabla f(x^{(k)})^\mathsf{T} d^{(k)}}{|d^{(k)}|} \xrightarrow{k \to \infty} 0$. If σ_k is efficient, we have

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

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

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

so to ensure that $\nabla f(x^{(k)}) \to 0$ we have to avoid $\nabla f(x^{(k)}) \perp d^{(k)}$ for large k.

If $\nabla f(x) = 0$ holds, x is a **stationary point** of f.
Stationary points need not be extrema, consider z.B. $f(x) := x^3$ and $x = 0$.

Requirements for the search directions

Definition

Strict complementarity

Problems with box constraints

Definition

Slater constraint qualification

Algorithm

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

A LAGRANGE multiplier λ satisfies **strict complementarity** if $\lambda_i > 0$ for all $i \in I \cap \mathcal{A}(\tilde{x})$.
Then $C(\tilde{x}, \tilde{\lambda}) = \{d \in \mathbb{R}^n : \nabla c_i(\tilde{x})^\mathsf{T} d = 0 \ \forall i \in \mathcal{A}(\tilde{x})\} = \ker(G(\tilde{x}))$ for $G(\tilde{x}) \coloneqq (\nabla c_i(\tilde{x})^\mathsf{T})_{i \in \mathcal{A}(\tilde{x})}$. Let $(s_k)_{k=1}^{\ell}$ be a basis of $\ker(G(\tilde{x}))$. The second order optimality conditions reduce to $Z^\mathsf{T} \nabla^2_{xx} L(\tilde{x}, \tilde{\lambda}) Z$ being positive definite on $\mathbb{R}^\ell$.

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

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

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

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 $c_i(x) \coloneqq a_i^\mathsf{T} x + b_i$ 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 $i \in I$.
One can show that if (SQC) holds in $\tilde{x} \in \Omega$, then (MFCQ) holds.

Let $\Omega \coloneqq \{x \in \mathbb{R}^n : v_i \leqslant x_i \leqslant w_i \ \forall i \in \{1, \dots, n\}\}$ and for simplicity assume $v < w$ componentwise. Then $x \in \Omega$ can be rewritten as $Gx \geqslant r$, where $G = (I, -I)^\mathsf{T}$, $r = (v, -w)$.
At most one constraint can be active, so $G(x) \coloneqq (\nabla c_i(x))_{i \in \mathcal{A}(x)} = \operatorname{diag}((\pm 1)_{i=1}^n)$ and thus $\{\nabla c_i : i \in \mathcal{A}(\tilde{x})\}$ is linearly independent and thus (ACQ) holds.
$L(x, \lambda) = f(x) - \sum_{j=1}^n \lambda_j^{(\ell)}(x_j - v_j) - \sum_{j=1}^n \lambda_j^{(u)}(-x_j + w_j)$, by KKT: $\lambda_i^{(\ell)} = \left[\frac{\partial f(x)}{\partial x_i}\right]_+$ and $\lambda_i^{(u)} = \left[\frac{\partial f(x)}{\partial x_i}\right]_-$ are unique.
$C(\tilde{x}, \tilde{y}) = \left\{d \in L_\Omega(\tilde{x}) : d_i = 0 \text{ if } \frac{\partial f(\tilde{x})}{\partial x_i} \neq 0\right\}$.

$Q \in \mathbb{R}^{n \times n}$ sym., PD on $\ker(A)$, $A \in \mathbb{R}^{m \times n}$, $\operatorname{rang}(A) = m \leqslant n$.

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 $h \coloneqq -Hq = \binom{h_1}{h_2}$ and $B \coloneqq HQH^\mathsf{T} =: \left(\begin{smallmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{smallmatrix}\right)$, 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 $B_{22} \tilde{x}_z = h_2 - B_{21} \tilde{x}_y$. $\tilde{x} \coloneqq H^\mathsf{T} \binom{\tilde{x}_y}{\tilde{x}_z}$.

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