Nonlinear Optimization

0. Basics

Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$

$U \subset \mathbb{R}^n$

look for $x_* \in U$: $f(x_*) \leq f(x) \quad \forall x \in U$

0.1 Definition

(i) x_* is called <u>minimizer</u> if $\forall x \in U: f(x_*) \leq f(x)$

(ii) <u>strict minimizer</u> " $<$ "

(iii) the <u>epigraph</u> of f is

$$\operatorname{epi} f := \{ (x, y) \in \mathbb{R}^n \times \mathbb{R} \mid y \geq f(x) \}$$

(iv) the <u>level set</u> L_α for $\alpha \in \mathbb{R}$ is

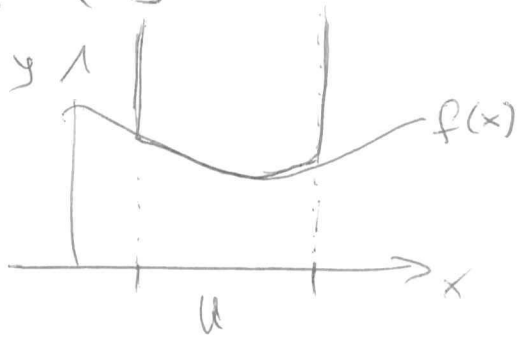
$$L_\alpha f := \{x \in U \mid f(x) \leq \alpha\}$$

(v) the <u>domain</u> of f is

$$\operatorname{dom} f := \bigcup_{\alpha \in \mathbb{R}} L_\alpha f = \{x \in U \mid f(x) < \infty\}$$

(vi) the <u>indicator function</u> of U is

$$\iota_U := \begin{cases} 0, & x \in U \\ \infty, & \text{otherwise} \end{cases}$$



<u>0.2 Lemma</u> Let $f: \mathbb{R}^n \to \mathbb{R}$ be continuous.

Then L_α is closed for all $\alpha \in \mathbb{R}$.

<u>Proof</u> $x_k \to \bar{x}$ with $x_k \in L_\alpha f \Rightarrow f(x_k) \leq \alpha$ f.a. $k \in \mathbb{N}$

$\underset{\text{continuity}}{\Rightarrow}$ $f(\bar{x}) \leq \alpha \Rightarrow \bar{x} \in L_\alpha f$ □

0.3 Theorem (Existence of minimizers)

Let $U \subset \mathbb{R}^n$ be closed, bounded, nonempty,

$f : \mathbb{R}^n \to \mathbb{R}$ be continuous,

Then, a minimizer x_* exists.

Proof: Let $m := \inf_{x \in U} f(x)$. Then there is $(x_k)_k$: $\lim_{k\to\infty} f(x_k) = m$.

Due to $m < \infty$, Thm Heine-Borel $\Rightarrow$ U compact.

$(x_k)_k$ contains a convergent subsequence. Wlog $x_k \to \bar{x}$.

Continuity of f implies $f(\bar{x}) = f(\lim_{k\to\infty} x_k) = \lim_{k\to\infty} f(x_k) = m$.

$\Rightarrow \bar{x}$ is a minimizer. □

0.4 Definition

f is <u>lower semicontinuous</u>,

if $L_\alpha f$ is closed for all $\alpha \in \mathbb{R}$.

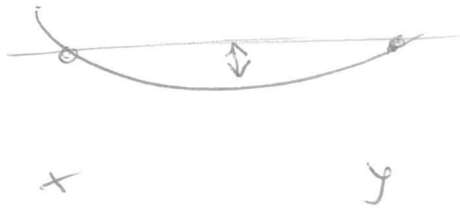
Generalization of 0.3: Assume f is only lower semicont.

A minimizer exists.

Proof: Assume $f(\bar{x}) = m + \varepsilon$, $\varepsilon > 0$. There is a subsequence $(x_{k_j})_j$ with $f(x_{k_j}) \leq m + \frac{\varepsilon}{2} \Rightarrow x_{k_j} \in L_{m+\frac{\varepsilon}{2}} f$.

Closedness of $L_{m+\frac{\varepsilon}{2}}$ implies $\bar{x} \in L_{m+\frac{\varepsilon}{2}} \Rightarrow f(\bar{x}) \leq m + \frac{\varepsilon}{2}$ □

0.5 Definition (Convexity)



(i) $f: \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex, if $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$

for all $x, y \in U$, $\lambda \in [0,1]$

strictly convex " $<$ " for all $x \neq y$, $\lambda \in]0,1[$

(ii) $U \subset \mathbb{R}^n$ is convex if $x, y \in U \Rightarrow \lambda x + (1-\lambda) y \in U \quad \forall \lambda \in [0,1]$

0.6 Theorem (uniqueness)

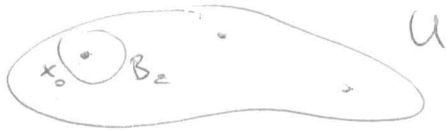
Let $f: U \to \overline{\mathbb{R}}$ be strictly convex and $U \subset \mathbb{R}^n$ convex.
Then, f has at most one minimizer.

Proof: Let x, y be minimizers. Then $\frac{x+y}{2} \in U$ and

$$f\left(\frac{x+y}{2}\right) < \frac{1}{2}\left(f(x) + f(y)\right) = f(x) \Rightarrow x = y \quad \square$$

0.7 Definition (local minimizer)

$x_0 \in U$ is a local minimizer if there is $\varepsilon > 0$ such that x_0 is a (global) minimizer of f in $U \cap B_\varepsilon(x_0)$.



0.8 Theorem (necessary conditions)

Let x_* be a local minimizer of f and $x_* \in \operatorname{int} U$.

Then (i) if $f \in C^1(U)$: $f'(x_*) = 0$

(ii) if $f \in C^2(U)$: $f''(x_*)$ is positive semidefinite

If f and U are convex, $f'(x_*) = 0 \Rightarrow x_*$ is a minimizer.

Proof (i) $0 \le f(x_* + \varepsilon h) - f(x_*) = \varepsilon f'(x_*) h + o(\varepsilon \|h\|)$

$$\Rightarrow f'(x_*) h \ge -\frac{1}{\varepsilon} o(\varepsilon \|h\|) \underset{\varepsilon \to 0}{\to} 0 \Rightarrow f'(x_*) h \ge 0 \quad \forall h \Rightarrow f'(x_*) = 0$$

(ii) $$0 \le f(x_* + \varepsilon h) - f(x_*) = \varepsilon h^T f''(x_*) \varepsilon h + o(\varepsilon^2 \|h\|^2)$$

$$\Rightarrow h^T f''(x_*) h \ge 0.$$

f convex $\Rightarrow (1-\varepsilon) f(x_*) + \varepsilon f(x_* + h) \ge f(x_* + \varepsilon h) = f(x_*) + o(\varepsilon \|h\|)$

$$\Rightarrow f(x_* + h) \ge \frac{1}{\varepsilon} \left(\varepsilon f(x_*) + o(\varepsilon \|h\|) \right) \underset{\varepsilon \to 0}{\to} f(x_*)$$

$\Rightarrow f(x_* + h) \ge f(x_*) \Rightarrow x_*$ is minimizes. $\square$

0.9 Theorem (sufficient conditions)

$x_* \in \operatorname{int} U$, $f \in C^2(U)$, $f'(x_*) = 0$, $f''(x_*)$ spd $\left(v^T f''(x_*) v \geq \alpha \|v\|^2\right)$

$\Rightarrow x^*$ is local minimizer

Proof
$$f(x) - f(x_*) = (x - x_*)^T f''(x_*)(x - x_*) + o(\|x - x_*\|^2)$$
$$\geq \alpha \|x - x_*\|^2 + o(\|x - x_*\|^2)$$

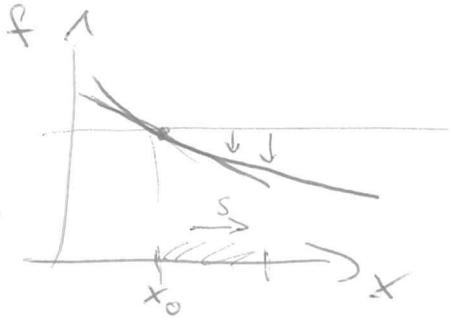
$\to 0$ faster than $\|x - x_*\|^2$

$$\geq 0 \quad \text{for } \|x - x_*\| \text{ sufficiently small} \quad \square$$

I. Unconstrained Optimization

$U = \mathbb{R}^n$, $f \in C^1(\mathbb{R}^n)$, $L_{f(x_0)} f$ bounded $\Rightarrow$ f is bounded from below

I.1 Descent methods



(I.1.1) Algorithm

input: $x_0 \in \mathbb{R}^n$

while „not converged"

choose s_k with $f'(x_k)\, s_k < 0$

choose $\alpha_k > 0$ such that $f(x_k + \alpha_k s_k) < f(x_k)$

$x_{k+1} = x_k + \alpha_k s_k$

(I.1.2) Lemma

(i) $f(x_k)$ is monotonically decreasing

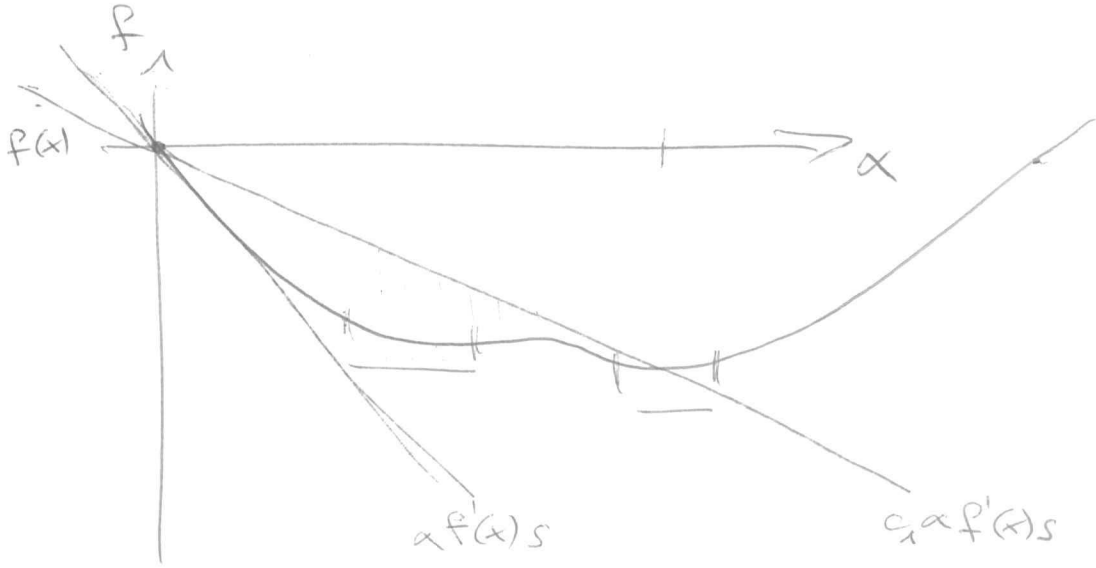
(ii) x_k has (at least) one accumulation point

Proof (i) by construction

(ii) $(x_k)_k \subset L_{f(x_0)} f$ which is compact

Bolzano-Weierstraß $\Rightarrow$ convergent subsequence $\square$

I.2 Line search



Properties of „good" step sizes:

(i) $f(x+\alpha s) \leq f(x) + c_1 \alpha f'(x) s$

$0 < c_1 < 1$

(ii) $|f'(x+\alpha s) s| \leq -c_2 f'(x) s$

$0 < c_2 < 1$

Wolfe-conditions

(I.2.1) **Lemma** (existence of Wolfe step size)

If $0 < c_1 \leq c_2 < 1$, then there is $\alpha > 0$ satisfying conditions (i) and (ii), unless $f'(x) \neq 0$.

Proof: $\varphi(\alpha) := f(x+\alpha s) \in C^1(\mathbb{R})$, $\varphi'(0) < 0$

$L := L_{\varphi(0)}^{\varphi} = \{\alpha \in \mathbb{R}_+ \mid f(x+\alpha s) \leq f(x)\}$ is closed & bounded

$\Rightarrow$ φ assumes a minimum in L. Let α_0 be the minimizer.

If $\alpha \in \partial L \Rightarrow \varphi(\alpha) = \varphi(0)$, but due to $\varphi'(0) < 0$ there are $\alpha > 0$ with $\varphi(\alpha) < \varphi(0) \Rightarrow \alpha_0 \in \operatorname{int} L \Rightarrow \varphi'(\alpha_0) = 0$.

Let $\underline{\alpha} := \min\{\alpha \in \mathbb{R}_+ \mid |\varphi'(\alpha)| \leq -c_2 \varphi'(0)\}$

$\underline{\alpha}$ satisfies (ii).

Then $f(x+\underline{\alpha}s) = f(x) + \int_0^{\underline{\alpha}} f'(x+\alpha s)s \, d\alpha$

$< f(x) + \int_0^{\underline{\alpha}} c_2 f'(x)s \, d\alpha$

$\leq f(x) + \underline{\alpha} c_1 f'(x)s \Rightarrow \underline{\alpha}$ satisfies (i). □

Practical line search

(I.2.2) Algorithm of Armijo linesearch

input $\alpha_0 > 0$

while (i) violated

$\alpha_{k+1} = \frac{\alpha_k}{2}$

output α_k

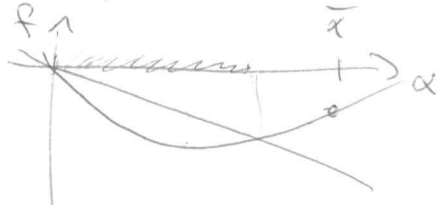
(I.2.3) Wolfe line search

(a) find interval containing a „good" step size

$\underline{\alpha} = 0$, $\bar{\alpha} > 0$

while (i) satisfied

$\bar{\alpha} \leftarrow 2\bar{\alpha}$



(b) find a Wolfe-admissible point

while true

$\beta = \frac{1}{2}(\underline{\alpha} + \bar{\alpha})$

if (i) violated: $\bar{\alpha} \leftarrow \beta$

if $\varphi'(\beta) < c_2 \varphi'(0)$: $\underline{\alpha} \leftarrow \beta$

if $\varphi'(\beta) > -c_2 \varphi'(0)$: $\bar{\alpha} \leftarrow \beta$

else stop

I.3. Steepest descent (Gradient method)

locally optimal choice of s_k:

$$s_k \in \operatorname{argmin}\left\{ \frac{f'(x_k)s_k}{\|s_k\|} \right\} \Rightarrow s_k \text{ - gradient}$$

for Euclidean norm: $s_k = -\nabla f(x_k) = -f'(x)^T$

(I.3.1) Theorem

Let f' be Lipschitz continuous, i.e $\|f'(a) - f'(b)\| \leq L \cdot \|a-b\|$.
Then the gradient method (Alg. I.1.1. with linesearch I.2.1 and $s_k = -\nabla f(x_k)$) creates a sequence $(x_k)_k$. All accumulation points $\bar{x}$ are stationary: $f'(\bar{x}) = 0$.

Proof $\quad (f'(x_{k+1}) - f'(x_k))\, s_{x_k} \geq (c_2 - 1) f'(x_k) s_k \qquad$ (Wolfe (ii))

$\Rightarrow \alpha_k L \|s_k\|^2 \geq (c_2 - 1) f'(x_k) s_k \qquad$ Lipschitz

$$f(x_{k+1}) \leq f(x_k) + c_1 \alpha_k \underbrace{f'(x_k) s_k}_{<0} \qquad \text{(Wolfe (i))}$$

$$\leq f(x_k) + c_1 \frac{c_2 - 1}{L \|s_k\|^2} f'(x_k) s_k \; f'(x_k) s_k$$

$$= f(x_k) - c_1 \frac{1 - c_2}{L} \|\nabla f(x_k)\|^2 \qquad s_k = -\nabla f(x_k)$$

$$= f(x_0) - c_1 \frac{1 - c_2}{L} \sum_{i=0}^{k} \|\nabla f(x_k)\|^2 .$$

f bounded from below $\Rightarrow \sum_{i=0}^{k} \|\nabla f(x_k)\|^2 < \infty$

$$\Rightarrow \lim_{k \to \infty} \|\nabla f(x_k)\| = 0$$

Let ξ be an accumulation point of $(x_k)_k$. Then in any neighborhood of ξ there are points with arbitrarily small gradient. Lipschitz continuity $\Rightarrow f'(\xi) = 0$ □

Example: $f(x) = x^3$ with $c_1 \geq \frac{1}{3}$.

Convergence speed

model problem $f(x) = \frac{1}{2} x^T A x - b^T x$, A spd $\Rightarrow$ unique minimizer $A^{-1}b$

gradient: $\nabla f(x) = Ax - b$

exact linesearch: $f(x_k - \alpha \nabla f(x_k)) = \min$

$$\frac{1}{2} (x - \alpha \nabla f)^T A (x - \alpha \nabla f) - b^T (x - \alpha \nabla f)$$

differentiate with α: $-\nabla f^T A (x - \alpha \nabla f) + b^T \nabla f = 0$

$$\Rightarrow \alpha = \frac{-b^T \nabla f + \nabla f^T A x}{\nabla f^T A \nabla f} = \frac{\nabla f^T \nabla f}{\nabla f^T A \nabla f}$$

(I.32) **Theorem** For the model problem, gradient method with exact linesearch converges linearly:

$$\|x_{k+1} - x_*\|_A \leq \Theta \|x_k - x_*\|_A, \quad \Theta = 1 - \frac{1}{2\kappa}$$

$\|x\|_A^2 := x^T A x$, $\kappa = \|A\| \cdot \|A^{-1}\| = \frac{\lambda_{max}}{\lambda_{min}}$ condition of A

Proof $f(x_{k+1}) = f(x_k + \alpha_k s_k) = f(x_k) + f'(x_k)\alpha_k s_k + \frac{1}{2} f''(x_k)\alpha_k^2 s_k^2$

$= f(x_k) - \alpha_k \nabla f^T \nabla f + \frac{\alpha_k^2}{2} \nabla f^T A \nabla f$

$= f(x_k) - \frac{(\nabla f^T \nabla f)^2}{\nabla f^T A \nabla f} + \frac{1}{2} \frac{(\nabla f^T \nabla f)^2}{\nabla f^T A \nabla f}$

$= f(x_k) - \frac{1}{2} \frac{\|\nabla f\|^4}{\|\nabla f\|_A^2}$

$f(x) = f(x_*) + \underbrace{f'(x_*)}_{=0}(x - x_*) + \frac{1}{2} f''(x_*)(x - x_*)^2$

$= f(x_*) + \frac{1}{2}(x - x_*)^T A (x - x_*)$

$= f(x_*) + \frac{1}{2}\|x - x_*\|_A^2$

$\Rightarrow \|x_{k+1} - x_*\|_A^2 = 2\left(f(x_{k+1}) - f(x_*)\right)$

$= 2\left(f(x_k) - \frac{1}{2} \frac{\|\nabla f\|^4}{\|\nabla f\|_A^2} - f(x_*)\right)$

$= \|x_k - x_*\|_A^2 - \frac{\|\nabla f\|^4}{\|\nabla f\|_A^2}$

Moreover $\nabla f = Ax_k - b = Ax_k - Ax_* = A(x_k - x_*)$

$$\Rightarrow \|x_k - x_*\|_A^2 = (A^{-1}\nabla f)^T A (A^{-1}\nabla f) = \nabla f^T A^{-1} \nabla f$$

$$\Rightarrow \|x_{k+1} - x_*\|_A^2 = \|x_k - x_*\|_A^2 - \frac{\|\nabla f\|^4}{\|\nabla f\|_A^2} = \|x_k - x_*\|_A^2 \underbrace{\left(1 - \frac{\|\nabla f\|^4}{\nabla f^T A \nabla f \; \nabla f^T A^{-1} \nabla f}\right)}_{\Theta_k^2}$$

$$\Theta_k^2 \leq 1 - \frac{1}{\lambda_{max} \lambda_{min}^{-1}} = 1 - \frac{1}{\kappa}$$

$$\Rightarrow \Theta_k \leq 1 - \frac{1}{2\kappa} \quad \text{for } \kappa \gg 1 \qquad \square$$

Remark: Improvement possible: $\Theta = \frac{\kappa - 1}{\kappa + 1}$.

I.4 Termination criteria

$$f(x_k) - f(x_*) \leq TOL^2$$

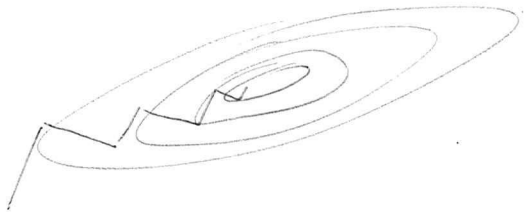
$$\| \| x_k - x_* \|_A^2 \geq \frac{1}{2} \Theta^{-2m} \| x_{k+m} - x_* \|_A^2 = \Theta^{-2m} \left(f(x_{k+m}) - f(x_*) \right)$$

Thus $f(x_k) - f(x_*) = f(x_k) - f(x_{k+m}) + f(x_{k+m}) - f(x_*)$

$$\leq f(x_k) - f(x_{k+m}) + \Theta^{2m} \left(f(x_k) - f(x_*) \right)$$

$$f(x_k) - f(x_*) \leq \frac{f(x_k) - f(x_{k+m})}{1 - \Theta^{2m}} \overset{!}{\leq} TOL^2$$

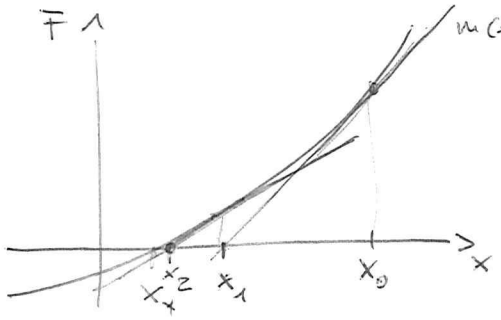
$\kappa(f''(x_*)) \gg 1 \Rightarrow$ steepest descent is slow



$\nabla f(x_*) = 0$

I.5 Newton's method

ordinary Newton method for $F(x) = 0$, $F: \mathbb{R}^n \to \mathbb{R}^n$



$m(x) = F(x_0) + F'(x_0)(x - x_0) = 0$

$m(x_1) = 0 \Rightarrow x_1 = x_0 - \underbrace{F'(x_0)^{-1} F(x_0)}_{\Delta x_0}$

I.5.1 Theorem Let $F \in C^2(\mathbb{R}^n)$, $F(x_*) = 0$, $F'(x)$ nonsingular

Then, Newton's method converges locally quadratically.

Proof

$$
\begin{aligned}
x_{k+1} - x_* &= x_k - F'(x_k)^{-1} F(x_k) - x_* + F'(x_k) \overbrace{F(x_*)}^{=0} \\
&= x_k - x_* - F'(x_k)^{-1} \left(F(x_k) - F(x_*) \right) \\
&= x_k - x_* - F'(x_k)^{-1} \int_{t=0}^{1} F'(x_* + t(x_k - x_*)) (x_k - x_*) \, dt \\
&= x_k - x_* - F'(x_k)^{-1} \int_{t=0}^{1} \left[F'(x_k) + F'(x_* + t(x_k - x_*)) - F'(x_k) \right] (x_k - x_*) \, dt \\
&= - F'(x_k)^{-1} \int_{t=0}^{1} \underbrace{\left(F'(x_* + t(x_k - x_*)) - F'(x_k) \right)}_{= O(\|x_k - x_*\|)} dt \; (x_k - x_*)
\end{aligned}
$$

$$
\Rightarrow \quad \|x_{k+1} - x_*\| = O\left(\|x_k - x_*\|^2 \right) \qquad \square
$$

I.6. Affine-conjugate Newton methods

$f(x) \to \min, \quad F = \nabla f, \quad f \in C^2(\mathbb{R}^n)$ strictly convex

$$\Delta x = -F'(x)^{-1} F(x)$$

descent direction: $-F(x)^T F'(x)^{-1} F(x) = f'(x)\Delta x < 0 \Leftarrow F'(x)^{-1}$ spd $\Leftrightarrow F'(x)$ spd

choice of stepsize α

$$\begin{aligned}
f(x+\alpha\Delta x) &= f(x) + \int_{s=0}^{1} f'(x+s\alpha\Delta x)\,\alpha\Delta x\,ds \\
&= f(x) + \alpha f'(x)\Delta x + \int_{s=0}^{1}\left[f'(x+s\alpha\Delta x) - f'(x)\right]\alpha\Delta x\,ds \\
&= f(x) + \alpha f'(x)\Delta x + \int_{s=0}^{1}\int_{t=0}^{1} f''(x+ts\alpha\Delta x)\,s\alpha\Delta x\,dt\,\alpha\Delta x\,ds \\
&= f(x) + \alpha f'(x)\Delta x + \frac{\alpha^2}{2}\Delta x^T f''(x)\Delta x \\
&\quad + \int_{s=0}^{1}\int_{t=0}^{1}\left(f''(x+ts\alpha\Delta x) - f''(x)\right) s\alpha^2\Delta x^2\,dt\,ds \\
&= f(x) - \alpha\Delta x^T F'(x)\Delta x + \frac{\alpha^2}{2}\Delta x^T F'(x)\Delta x \\
&\quad + \int_{s=0}^{1}\int_{t=0}^{1}\left(F'(x+ts\alpha\Delta x) - F'(x)\right) s\alpha^2\Delta x^2\,ds\,dt
\end{aligned}$$

(T. 6.1) affine conjugate Lipschitz condition

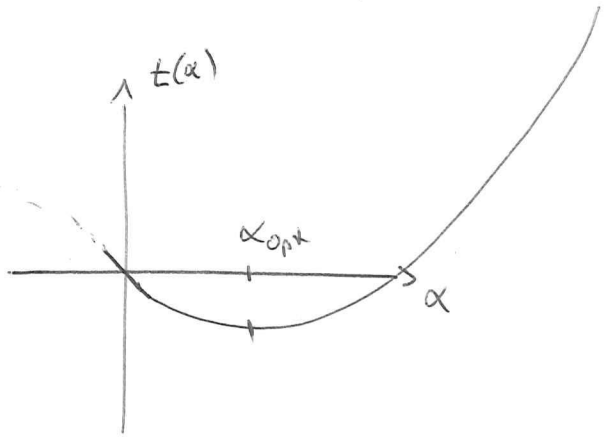
$$\Delta x^T \left(F'(x+\Delta x) - F'(x) \right) \Delta x \leq \omega \left(\Delta x^T F'(x) \Delta x \right)^{3/2} \qquad \text{f.a. } x, \Delta x \in \mathbb{R}^n$$

$$\Rightarrow f(x+\alpha\Delta x) \leq f(x) + \left(-\alpha + \frac{\alpha^2}{2} + \frac{\omega}{6}\alpha^3 \sqrt{\Delta x^T F'(x) \Delta x} \right) \Delta x^T F'(x) \Delta x$$

energy norm $\|y\|_x^2 := y^T F'(x) y$

optimal step size

$$t(\alpha) = -\alpha + \frac{\alpha^2}{2} + \frac{\omega\alpha^3}{6} \|\Delta x\|_x$$



$$\alpha_{opt} = \underset{\alpha > 0}{\arg\min}\ t(\alpha)$$

$$t'(\alpha) = -1 + \alpha + \frac{\omega}{2}\alpha^2 \|\Delta x\|_x \overset{!}{=} 0$$

$$\Rightarrow \alpha_{opt} = \frac{-1 + \sqrt{1 + 2\omega \|\Delta x\|_x}}{\omega \|\Delta x\|_x}$$

$$= \frac{2}{1 + \sqrt{1 + 2\omega \|\Delta x\|_x}} \qquad \in]0, 1]$$

→ „damped" Newton method

I.6.4 Theorem $\alpha_{opt} = \alpha$ satisfies the Armijo rule I.2.(i), i.e.

$$f(x+\alpha\Delta x) - f(x) \le c_1 \alpha f'(x)\Delta x \quad \text{with } c_1 = \frac{1}{6}.$$

Proof: Let $h = \omega \|\Delta x\|_x$, $\alpha = \dfrac{2}{1+\sqrt{1+2h}} \le 1$.

Then $h\alpha^2 = \dfrac{4h}{1+2\sqrt{1+2h}+1+2h} \le \dfrac{4h}{2h} = 2$.

$$-\frac{t(\alpha)}{\alpha} = 1 - \frac{\alpha}{2} - \frac{h}{6}\alpha^2 \ge 1 - \frac{1}{2} - \frac{2}{6} = \frac{1}{6}.$$

Therefore $f(x+\alpha\Delta x) - f(x) \le t(\alpha)\,(-F(x)^T\Delta x) = -\dfrac{t(\alpha)}{\alpha}\alpha f'(x)\Delta x \le \dfrac{1}{6}\alpha f'(x)\Delta x$ □

I.6.5 Theorem If $\omega\|\Delta x\|_x \le \frac{3}{2}$, then α satisfies the curvature condition I.2.(ii), i.e. $|f'(x+\alpha\Delta x)\Delta x| \le -c_2 f'(x)\Delta x$ with $c_2 < \frac{2}{3}$.

Proof:

$$|f'(x+\alpha\Delta x)\Delta x| = \left| F(x)^T\Delta x + \int_{s=0}^{1} F'(x+s\alpha\Delta x)\alpha\Delta x^2\, ds \right|$$

$$= \left| F(x)^T\Delta x + \alpha\Delta x^T F'(x)\Delta x + \int_{s=0}^{1} \left(F'(x+s\alpha\Delta x) - F'(x)\right)\alpha\Delta x^2\, ds \right|$$

$$\leq \left| -\Delta x^T \bar{F}'(x) \Delta x + \alpha \Delta x^T \bar{F}'(x) \Delta x \right| + \int_{s=0}^{1} \omega \alpha s \|\Delta x\|_x^3 \, ds$$

$$= \left(\underbrace{\left| -1 + \alpha \right|}_{\leq 0} + \frac{\alpha}{2} \omega \|\Delta x\|_x \right) \|\Delta x\|_x^2$$

$$= \left| -1 + \alpha - \frac{\alpha}{2} \omega \|\Delta x\|_x \right| \|\Delta x\|_x^2$$

$$= \left| \underbrace{-1 + \alpha + \frac{\alpha}{2} \omega \|\Delta x\|_x}_{= \xi'(\alpha) = 0} - \alpha \omega \|\Delta x\|_x \right| \|\Delta x\|_x^2$$

$$= \underbrace{\alpha \omega \|\Delta x\|_x}_{c_2} \left(-f'(x) \Delta x \right)$$

For $h = \omega \|\Delta x\|_x \leq \frac{3}{2}$ it is easily verified that $\alpha h \leq \frac{2}{3}$ □

I.7 Practical step size control

$$\alpha_{opt} = \frac{2}{1+\sqrt{1+2\omega\|\Delta x\|_x}}$$ depends on (unknown) ω!

From (I.6.2), $f(x+\alpha\Delta x) \le f(x) - (\alpha - \frac{\alpha^2}{2})\|\Delta x\|_x^2 + \frac{\alpha^3}{6}\omega\|\Delta x\|_x^3$

we obtain $\frac{\alpha^3}{6}\omega\|\Delta x\|_x^3 \ge \left| f(x+\alpha\Delta x) - f(x) + (\alpha - \frac{\alpha^2}{2})\|\Delta x\|_x^2 \right| =: |\Xi|$

$$\Rightarrow \quad \omega \ge \frac{6|\Xi|}{\alpha^3\|\Delta x\|_x^3} =: [\omega]$$

$\longrightarrow$ computable step size $[\alpha_{opt}] = \frac{2}{1+\sqrt{1+2[\omega]\|\Delta x\|_x}} \ge \alpha_{opt}$

I.7.1 Lemma (bitcounting lemma)

Let $\omega \le 2[\omega]$. Then

(I.7.2) $f(x+[\alpha_{opt}]\Delta x) \le f(x) - \frac{1}{6}[\alpha_{opt}]\left([\alpha_{opt}]+2\right)\|\Delta x\|_x^2$

Proof Let $\alpha := [\alpha_{opt}]$.

$$f(x+\alpha\Delta x) - f(x) \le \left(-\alpha + \frac{\alpha^2}{2} + \frac{\alpha^3}{6}\omega\|\Delta x\|_x\right)\|\Delta x\|_x^2$$

$$\le \left(-\alpha + \frac{\alpha^2}{2} + \frac{\alpha^3}{6}2[\omega]\|\Delta x\|_x\right)\|\Delta x\|_x^2$$

α is obtained by minimizing $[t] = -\alpha + \frac{\alpha^2}{2} + \frac{\alpha^3}{6}[\omega]\|\Delta x\|_x$

$$\Rightarrow 1 - \alpha - \frac{\alpha^2}{2}[\omega]\|\Delta x\|_x = 0 \Rightarrow \frac{\alpha^2}{2}[\omega]\|\Delta x\|_x = 1-\alpha$$

$$\Rightarrow f(x+\alpha\Delta x) - f(x) \leq \left(-\alpha + \frac{\alpha^2}{2} + \frac{2}{3}\alpha(1-\alpha)\right)\|\Delta x\|_x^2$$
$$\leq \alpha\left(-\frac{1}{3} - \frac{\alpha}{6}\right)\|\Delta x\|_x^2 = -\frac{\alpha}{6}(\alpha+2)\|\Delta x\|_x^2 \qquad \square$$

I.7.3 Damped Newton iteration

choose $[\omega]_{-1} > 0$, $x_0 \in \mathbb{R}^n$

for $k = 0, \ldots$

$\overline{F}'(x_k)\,\Delta x_k = -\overline{F}(x_k)$

$[\omega]_k \leftarrow [\omega]_{k-1}$

while (I.7.2) violated

$[\omega]_k \leftarrow [\omega]_k \left(f(x_k + [\alpha_{opt}]\Delta x_k)\right) > [\omega]_k$

$x_{k+1} \leftarrow x_k + [\alpha_{opt}]\Delta x_k$

$[\omega]_k \leftarrow [\omega]_k \left(f(x_k + [\alpha_{opt}]\Delta x_k)\right)$ // decrease in $[\omega]_k$ possible

Termination criterion

require (i) quadratic model is good approximation, ordinary Newton method converges to a unique minimizer

$$\Leftarrow [\omega]\, \|\Delta x\|_k \leq 1$$

(ii) $f(x) \leq f(x_*) + TOL^2$

$$f(x_k) - f(x_*) = \sum_{i=k}^{\infty} \left(f(x_i) - f(x_{i+1}) \right)$$

$$\leq \sum_{i=k}^{\infty} \left(\alpha_i - \frac{\alpha_i^2}{2} + \frac{\omega}{6} \alpha_i^3 \|\Delta x_i\|_{x_i} \right) \|\Delta x_i\|_{x_i}^2$$

$$\leq \sum_{i=k}^{\infty} \left(\frac{1}{2} + \frac{1}{6} \right) \|\Delta x_i\|_{x_i}^2$$

$$\leq \frac{5}{6} \sum_{i=k}^{\infty} \Theta^{2(i-k)} \|\Delta x_k\|_{x_k}^2$$

$$= \frac{5}{6} \frac{1}{1-\Theta^2} \|\Delta x_k\|_{x_k}^2$$

$$\Theta \approx \frac{\|\Delta x_k\|_{x_k}}{\|\Delta x_{k-1}\|_{x_{k-1}}}$$

II.8 Nonconvex Problems

II.8.1 Hessian-modifications

$\delta x = -B^{-1} F(x)$ with B spd $\Rightarrow$ δx is descent direction

simple choice $B = F'(x) + \eta M$, M spd

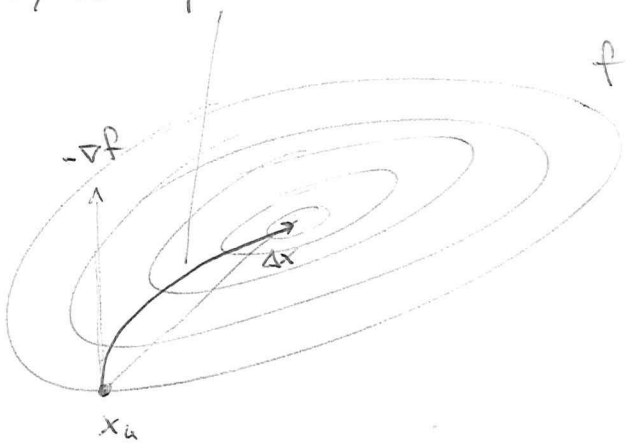
($M = I$: Levenberg-Marquardt method)

Choice of η

$\eta \to 0$: ordinary Newton method

$\eta \to \infty$: gradient method (preconditioned) with small step size

$\eta \mapsto \delta x(\eta)$ for $\eta > \eta_0$ defines a Levenberg-Marquardt path



II.8.2 Theorem Let $\eta_k \geq 0$ be minimal such that

$B \geq \varepsilon I$ for $\varepsilon > 0$, and Armijo condition holds, then all accumulation points of $(x_k)_k$ are stationary.

Proof as for Thm. I.2.1 □

II.8.3 Trust Region methods

quadratic model $f(x) + f'(x)\delta x + \frac{1}{2}\delta x^T f''(x)\delta x$
good for „small" δx, i.e. $\|\delta x\|_M \leq \rho$

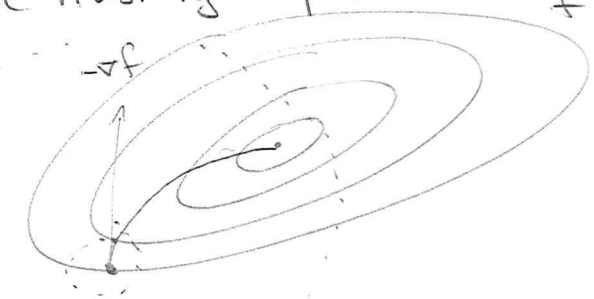
$$\delta x \in \operatorname*{argmin}_{\|\delta x\|_M \leq \rho} f(x) + f'(x)\delta x + \frac{1}{2}\delta x^T f''(x)\delta x$$

Choice of ρ

$\rho \to 0$: preconditioned gradient method with very small step size

$\rho \to \infty$: if $f''(x)$ spd → Newton

$\delta x(\rho)$ „defines" the trust-region path



II 8.4. Theorem Let ρ_k be chosen maximal such that δx_k satisfy the Armijo condition. Then, all accumulation points ξ of $(x_k)_k$ are stationary and satisfy the necessary second order optimality conditions $f''(\xi) \geq 0$.

Proof Stationarity of ξ as in I.3.1.

(i) Due to $f'' \in C(\mathbb{R}^n)$, f'' is bounded in a neighborhood of ξ:
$\|f''(x)\| \leq R$. Then with $\rho = \frac{2}{R}(1-c_1)$, $s = -\nabla f(x_k)$

$$f(x_k + \rho s) \leq f(x_k) + \rho f'(x_k)s + \rho^2 \frac{R}{2}\|s\|^2$$
$$\leq f(x_k) + \rho\left(1 - \rho\frac{R}{2}\right) f'(x_k)s$$
$$= f(x_k) + c_1 \rho f'(x_k)s$$

$\Rightarrow x_{k+1}(\rho)$ satisfies Armijo $\Rightarrow \rho_k \geq \rho$ bounded from below

(ii) WLOG. $x_k \to \xi$. $f(\xi) \leq f(x_k)$.

Assume $\exists p \in \mathbb{R}^n, \|p\|_2 = 1$: $p^T f''(\xi) p = -\varepsilon < 0$.

Then $f(x_{k+1}) \leq f(x_k + \rho p) \leq \underbrace{f(x_k)}_{\to f(\xi)} + \underbrace{f'(x_k)p}_{\to 0}\rho$
$$\leq \frac{\varepsilon}{2}\rho^2$$

$$\xrightarrow[k\to\infty]{} f(\xi) - \frac{\varepsilon}{2}\rho^2 < f(\xi). \quad \lightning \qquad \square$$

II 8.5 Minimization of cubic upper bound

Assumption: $F' = f''$ Lipschitz continuous

$$\| F'(x) - F'(y) \|_M \leq \omega \| x - y \|_M$$

Then $$f(x+\delta x) \leq f(x) + f'(x)\delta x + \frac{1}{2}\delta x^T F'(x)\delta x + \frac{\omega}{6}\|\delta x\|_M^3 =: m(x)$$

(II.8.6) $$\delta x \in \operatorname{argmin} m(x)$$

Choice of ω

$\omega \to 0$: ordinary Newton method

$\omega \to \infty$: preconditioned gradient method, small step size

$\delta x(\omega)$ „defines" a path for $\omega \geq 0$.

II 8.7 Theorem The paths $\delta x(\eta)$, $\delta x(\rho)$, $\delta x(\omega)$ are identical for sufficiently large η, ω, small ρ up to parametrization.

Proof: (i) Let $\delta x = \delta x(\omega)$. Then (II.8.6) implies

$$0 = F(x) + F'(x)\delta x + \frac{\omega}{6}\cdot\frac{3}{2}(\delta x^T \Pi \delta x)^{1/2} \cdot 2\Pi\delta x$$
$$= F(x) + \Big(F'(x) + \underbrace{\frac{\omega}{2}\|\delta x\|_M}_{\eta}\cdot\Pi\Big)\delta x$$

$$\Rightarrow \delta x = \delta x(\eta) \text{ with } \eta = \frac{\omega}{2}\|\delta x\|_M.$$

(ii) a) $\|\delta x\|_4 < \rho \Rightarrow 0 = F(x) + F'(x)\,\delta x$

$\Rightarrow \quad \delta x = \delta x(\eta)$ with $\eta = 0$

$= \delta x(\omega)$ with $\omega = 0$

b) $\|\delta x\|_4 = \rho \Rightarrow \exists\, \lambda \in \mathbb{R}$: $0 = F(x) + F'(x)\,\delta x + \lambda\, \Pi\, \delta x$

$= F(x) + \underbrace{(F'(x) + \lambda \Pi)}_{\eta}\, \delta x$

$\Rightarrow \delta x(\eta)$ with $\eta = \lambda$.

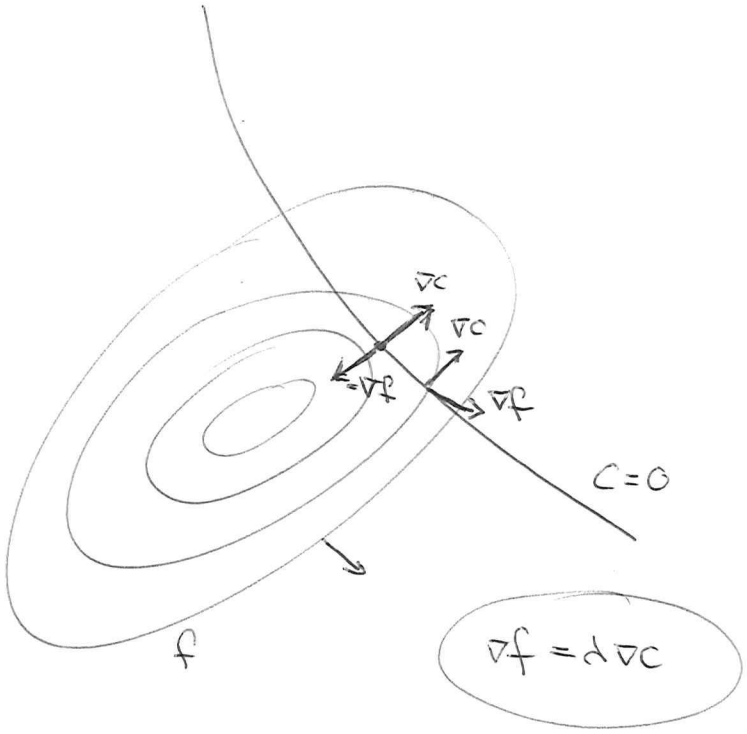
□

III. Equality constrained optimization

(III.1) $\min_{x \in \mathbb{R}^n} f(x)$ s.t. $c(x) = 0$, $\quad f \in C^2(\mathbb{R}^n, \mathbb{R})$, $\quad c \in C^2(\mathbb{R}^n, \mathbb{R}^m)$

→ admissible set $\mathcal{U} = \{x \in \mathbb{R}^n \mid c(x) = 0\}$

III.2 Necessary optimality conditions

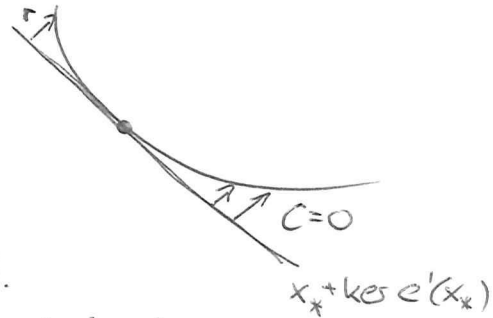


III.2.1 Theorem Let x_* minimizer of (III.1). Then $f'(x_*)\cdot \delta x = 0$ for all $\delta x \in \ker c'(x_*)$.

Proof Let $A = c'(x_*) \in \mathbb{R}^{m\times n}$ and $\mathbb{R}^n = \ker A \oplus (\ker A)^\perp$

$\rightarrow$ unique representation of $x = x^0 + x^\perp$. $c(x) = c(x^0, x^\perp)$.

The implicit function theorem guarantees the existence of $r \in C^1(\ker A, (\ker A)^\perp)$ with $c(x^0, x^\perp) = 0 \Leftrightarrow x^\perp = x^0 + r(x^0)$ in a neighborhood of x_*.



Let $\delta x \in \ker A$ and define

$x(\varepsilon) = x_* + \varepsilon \delta x + r(x_* + \varepsilon \delta x)$.

$\rightarrow$ $x(\varepsilon)$ is admissible ($c(x(\varepsilon)) = 0$).

As x_* is minimizer, $\varepsilon = 0$ is a minimizer of $f(x(\varepsilon))$ and hence

$$0 = \left.\frac{\partial f(x(\varepsilon))}{\partial \varepsilon}\right|_{\varepsilon=0} = f'(x(0))\cdot x'(0) = f'(x_*)\left[\delta x + r'(x_*)\delta x\right]$$

Impl. fct. thm. $r'(x_*) = \frac{\partial c}{\partial x^\perp}(x_*)^{-1} \underbrace{\frac{\partial c}{\partial x_0}(x_*)}_{=0} = 0.$

$$0 = f'(x_*)\delta x.$$

□

III.2.2 Definition Let $U \subseteq X$ be a subspace. Then

$U^\circ = \{v \in X^* \mid \forall u \in U : v \cdot u = 0\}$ is called annihilator of U.

III 2.3 Theorem Let $A \in \mathbb{R}^{m \times n}$ have full rank (linear independence constraint qualification) LICQ

Then $(\ker A)^\circ = \operatorname{im} A^T$ $\quad m \leq n$

Proof: (i) $\delta x \in \ker A$: $(A^T \lambda)^T \delta x = \lambda^T A \delta x = 0 \Rightarrow \operatorname{im} A^T \subset (\ker A)^\circ$

(ii) $\dim (\ker A)^\circ = n - \dim \ker A = n - (n-m) = m$
$\dim \operatorname{im} A^T = m$
$\Big\} \dim \operatorname{im} A^T = \dim (\ker A)^\circ$

$\Rightarrow \operatorname{im} A^T = (\ker A)^\circ$ □

III.2.4
Corollary (Karush-Kuhn-Tucker)

Let x_* be the minimizer of (III.1) and let $c'(x_*)$ have full rank. Then there is a Lagrange multiplier $\lambda \in \mathbb{R}^m$ such that

$$f'(x_*) = \lambda^T c'(x_*) \Leftrightarrow \nabla f(x_*) = c'(x_*)^T \lambda$$
$$c(x_*) = 0$$

III.2.1 Theorem Let x_* be minimizer of (III.1). Then $f'(x_*)\delta x = 0$ for all $\delta x \in \ker c'(x_*)$.

Proof Let $A = c'(x_*) \in \mathbb{R}^{m\times n}$ and $\mathbb{R}^n = \ker A \oplus (\ker A)^\perp$.
Unique representation of $x = x^0 + x^\perp$. Let $c(x) = c(x^0, x^\perp)$
The impl. fct thm guarantees existence of $r \in C^1(\ker A, (\ker A)^\perp)$
with $c(x^0, x^\perp) = 0 \Leftrightarrow x^\perp = x^0 + r(x^0)$ in a neighborhood of x_*.
Let $\delta x \in \ker A$ and define $x(\varepsilon) = x_* + \varepsilon\delta x + r(x_* + \varepsilon\delta x)$
$\Rightarrow x(\varepsilon)$ admissible ($c(x(\varepsilon)) = 0$) for small $|\varepsilon|$.
As x_* is minimizer, $\varepsilon = 0$ is minimizer of $f(x(\varepsilon))$ and
hence $0 = \left.\frac{\partial f(x(\varepsilon))}{\partial \varepsilon}\right|_{\varepsilon=0} = f'(x(0))\cdot x'(0) = f'(x_*)\left[\delta x + r'(x_*)\delta x\right]$.
Impl. fct. thm: $r'(x_*) = \frac{\partial c}{\partial x^\perp}(x_*)^{-1} \cdot \underbrace{\frac{\partial c}{\partial x^0}(x_*)}_{=0} = 0$.
$\Rightarrow 0 = f'(x_*)\delta x$ □

[nice but algorithmically not practical since we'd need a basis of $\ker A$ in every point]
Can we characterize $f'(x_*)$ better?

III.2.2 Definition Let $U \subseteq X$ be a subspace. Then
$U^0 = \{v \in X^* \mid \forall u \in U : v\cdot u = 0\}$ is called the annihilator of U.

E.2.5 Example

$\min_{x \in \mathbb{R}^2} \frac{1}{2} x^T x \quad s.t. \quad x_1 + x_2 = 1 \quad \rightarrow c' = [1 \; 1]$

KKT

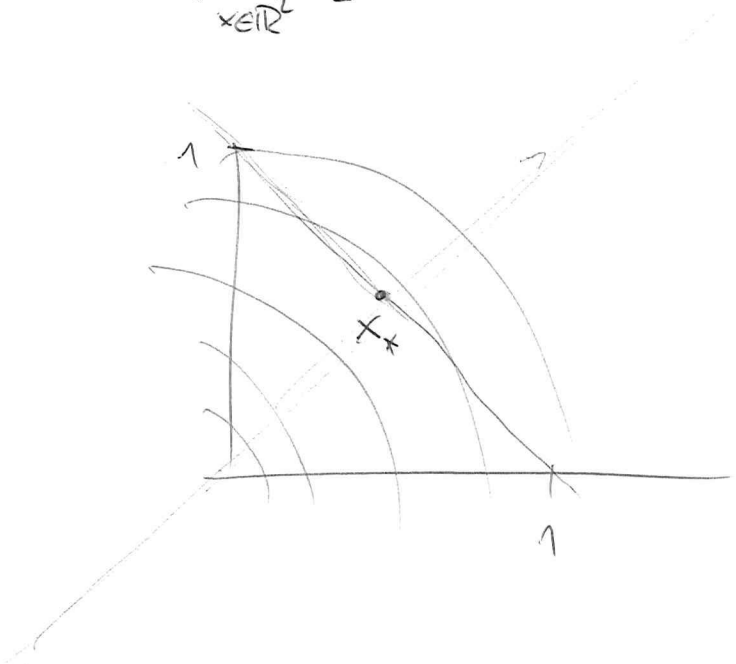
$$x - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \lambda = 0$$

$$[1 \; 1]\, x - 1 = 0$$

$$x = \lambda \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

$$\lambda \underbrace{[1 \; 1] \begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{2} = 1$$

$$\lambda = \frac{1}{2}$$



III. 3. Second order conditions

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad c(x) = 0 \in \mathbb{R}^m, \; m < n$$

$$\exists \lambda \in \mathbb{R}^m: \quad \nabla f(x_*) - c'(x_*)^T \lambda = 0$$
$$- c(x_*) = 0$$

III. 3. 1 Definition

The Lagrange function is

$$L(x, \lambda) = f(x) - c(x)^T \lambda.$$

$$\nabla_x L(x, \lambda) = \nabla f(x) - c'(x)^T \lambda$$
$$\nabla_\lambda L(x, \lambda) = - c(x)$$

$\Rightarrow$ KKT points are stationary points of Lagrangian

2nd derivative

$$L_{xx} = f'' - c''(x)^T \lambda = f'' - \sum_i \lambda_i c_i''$$

III.32 Theorem

Let $c'(x_*) = 0$ have full rank and (x_*, d) be a KKT point.

(i) x_* is local minimizes $\Rightarrow$ $L_{xx}(x_*, d)$ positive semidefinite on $\ker c'(x_*)$

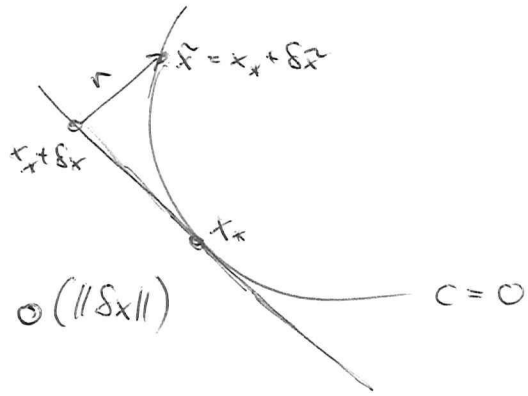
(ii) L_{xx} positive definite on $\ker c'(x_*)$ $\Rightarrow$ x_* is local minimizer

Proof Let $\delta\tilde{x} = \tilde{x} - x_*$ with $c(\tilde{x}) = 0$. Then

$$f(\tilde{x}) = f(\tilde{x}) - \underbrace{d^T c(\tilde{x})}_{=0} = L(\tilde{x}, d) = L(x_*, d) + \underbrace{L_x(x_*, d)\,\delta\tilde{x}}_{=0} + \frac{1}{2}\delta\tilde{x}^T L_{xx}(x_*, d)\delta\tilde{x} + o(\|\delta\tilde{x}\|^2)$$

$$= f(x_*) + \frac{1}{2}\delta\tilde{x}^T L_{xx}\delta\tilde{x} + o(\|\delta\tilde{x}\|^2)$$

Now let $\delta x \in \ker c'(x_*)$ and $\delta\tilde{x} = \delta x + r(x_* + \delta x)$



$\Rightarrow \|r(x_* + \delta x)\| = o(\|\delta x\|)$

Then $\delta\tilde{x}^T L_{xx} \delta\tilde{x} - \delta x^T L_{xx} \delta x$

$$= \underbrace{(\delta\tilde{x} - \delta x)}_{\|r\| = o(\|\delta x\|)} L_{xx} \underbrace{(\delta\tilde{x} + \delta x)}_{= O(\|\delta x\|)} = o(\|\delta x\|^2)$$

Thus, $f(\tilde{x}) - f(x_*) = \frac{1}{2} \delta x^T L_{xx} \delta x + o(\|\delta x\|^2)$.

(i) x_* minimizes $\Rightarrow f(\tilde{x}) - f(x_*) \geq 0 \quad \forall \tilde{x} = x_* + \delta\tilde{x}(\delta x), \ \delta x \in \ker c'(x_*) \cap B_\varepsilon(0)$

$\Rightarrow \delta x^T L_{xx} \delta x \geq -o(\|\delta x\|^2) \Rightarrow \delta x^T L_{xx} \delta x \geq 0.$

(ii) L_{xx} pos. def. on $\ker c'(x_*)$

$\Rightarrow \exists \alpha > 0: \quad f(\tilde{x}) - f(x_*) \geq \frac{1}{2} \delta x^T L_{xx} \delta x - o(\|\delta x\|^2)$

$\geq \frac{\alpha}{2} \|\delta x\|^2 - o(\|\delta x\|^2)$

$\geq \frac{\alpha}{4} \|\delta x\|^2$ for $\|\delta x\|$ suff. small $\square$.

III.4 Newton-KKT

$L'(x,\lambda) = 0 \rightarrow$ Newton's method

$$F\begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{bmatrix} \nabla f(x) - c'(x)^T \lambda \\ -c(x) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$F'\begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{bmatrix} L_{xx} & L_{x\lambda} \\ L_{\lambda x} & L_{\lambda\lambda} \end{bmatrix} = \begin{bmatrix} H & C^T \\ C & \end{bmatrix} \quad \text{with } H = f''(x) - c''(x)^T\lambda, \; C = -c'(x)$$

III.4.1 Theorem Assume $f, c \in C^{2,1}(\mathbb{R}^n)$ and (x_*, λ) satisfy sufficient 2nd order conditions. Then, Newton's method converges locally quadratically to (x_*, λ).

Proof: Thm I.5.1: need to show F' is Lipschitz continuous and F' is invertible

Let $F'(x_*,\lambda)\begin{bmatrix} \delta x \\ \delta\lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \rightarrow C\delta x = 0 \Rightarrow \delta x \in \ker C$

$$0 = [\delta x^T, \delta\lambda^T] \underbrace{F'(x_*,\lambda)\begin{bmatrix} \delta x \\ \delta\lambda \end{bmatrix}}_{=0} = \delta x^T H \delta x \geq \alpha \|\delta x\|^2$$

$\Rightarrow \delta x = 0 \; | \; C^T\delta\lambda = 0 \Rightarrow \delta\lambda = 0$ ∎

III.5. Penalty method

(III.5.1) $\min_{x \in \mathbb{R}^n} f(x)$ s.t. $c(x) = 0 \in \mathbb{R}^m$, $c, f \in C^2(\mathbb{R}^n)$

Idea
- transform (approximately) to unconstrained problem
- apply methods for unconst. problems

III.5.2 quadratic penalty

$$f_\mu(x) = f(x) + \frac{1}{\mu} \|c(x)\|^2, \quad \mu > 0$$

$x_\mu \in \operatorname{argmin} f_\mu(x)$

$x_\mu \to ?$

III.5.3 Theorem Let $\mu_k \to 0$ and x_k a global minimizer of f_{μ_k}. Then, every accumulation point ξ of $(x_k)_k$ is a global minimizers of (III.5.1), and $c(x_k) = O(\sqrt{\mu_k})$.

Proof Let x_0 be feasible, i.e. $c(x_0) = 0$. W.l.o.g. $x_k \to \xi$. Then

$$f(x_0) = f_{\mu_k}(x_0) \geq f_{\mu_k}(x_k) = f(x_k) + \frac{1}{\mu_k} \|c(x_k)\|^2$$

(i) $\Rightarrow \|c(x_k)\|^2 \le \mu_k \left(f(x_0) - f(x_k)\right) \le \mu_k \cdot \text{const} \Rightarrow c(x_k) = O(\sqrt{\mu_k})$

$\Rightarrow c(\xi) = 0$ feasibility

(ii) $f(x_0) \ge f(\xi)$ optimality

□

III. S. 4. Theorem Let (x_*, λ) satisfy sufficient 2nd order cond.

Then $x_\mu - x_* = O(\sqrt{\mu})$

$$\frac{2}{\mu} c(x_\mu) - \lambda = O(\sqrt{\mu})$$

Proof $$L'\left(x_\mu, \tfrac{2}{\mu} c(x_\mu)\right) = \underbrace{L'(x_*, \lambda)}_{=0} + L''(x_*, \lambda) \begin{bmatrix} x_\mu - x_* \\ \frac{2}{\mu} c(x_\mu) - \lambda \end{bmatrix} + o\left(\|x_\mu - x_*\| + \|\tfrac{2}{\mu} c(x_\mu) - \lambda\|\right)$$

$$\| \quad \begin{bmatrix} \nabla f(x_\mu) - \frac{2}{\mu} c'(x_\mu)^T c(x_\mu) \\ -c(x_\mu) \end{bmatrix} = \begin{bmatrix} 0 \\ -c(x_\mu) \end{bmatrix} = O\left(\|c(x_\mu)\|\right).$$

$= O(\sqrt{\mu})$.

□

$$f_\mu(x) = f(x) + \frac{1}{\mu}\|c(x)\|^2 \quad \to x_\mu \qquad \|x_* - x_\mu\| = O(\sqrt{\mu})$$

III.5.5 Corollary Assumptions as in III.5.4. Then, $\|x_\mu - x_*\| = O(\mu)$.

Proof: $\|\frac{2}{\mu} c(x_\mu) - \lambda\| = O(\sqrt{\mu})$

$$\leq \frac{2}{\mu}\|c(x_\mu)\| - \|\lambda\| \quad \Rightarrow \quad \|c(x_\mu)\| = \frac{\mu}{2}\left(\|\lambda\| + O(\sqrt{\mu})\right) = O(\mu) \qquad \square$$

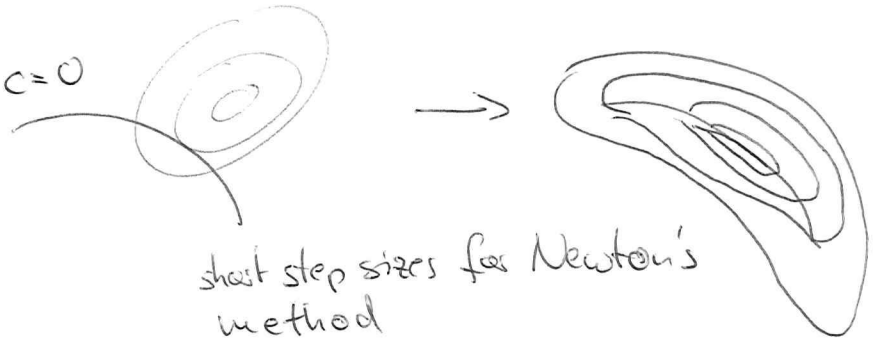
Drawbacks of penalty

(i) Hessian is ill-conditioned

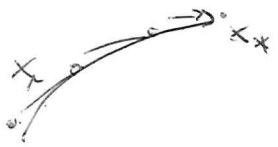
$$f_\mu''(x) = \underbrace{f''(x)}_{=O(1)} + \underbrace{\frac{2}{\mu} c'(x)^T c'(x)}_{O(\frac{1}{\mu})} + \underbrace{\frac{2}{\mu} c(x)^T c''(x)}_{O(1)}$$

$\to \frac{2}{\mu} c'(x)^T c'(x)$ not regular due to rank $\leq m \Rightarrow$ ker $c'(x)$ of dim $n-m > 0$

(ii) f_μ develops steep curved valleys (banana shaped)



partial remedy: continuation



III. 5.6 Algorithm

x_0, μ_0 given

choose $\sigma < 1$

for $k = 0, \ldots$

$$\mu_{k+1} = \sigma \mu_k$$

$$x_{k+1} = x_k - f''_{\mu_{k+1}}(x_k)^{-1} \nabla f_{\mu_{k+1}}(x_k)$$

if σ sufficiently large $\rightarrow$ Newton converges quickly

III.6 Augmented Lagrangian method

observations: (i) $L(x,\lambda) = f(x)$ for admissible x

(ii) $x_* = \arg\min_x L(x,\lambda_*)$ if f convex

Idea: $L_\mu(x,\lambda) = L(x,\lambda) + \frac{1}{\mu}\|c(x)\|^2$

III.6.1 Theorem Let x_*,λ_* satisfy necessary & sufficient optimality conditions.
Assume LICQ holds. Then $\exists \bar\mu > 0$: $\forall \mu \le \bar\mu$:
x_* is locally unique minimizer of $L_\mu(\xi,\lambda_*)$

Proof: (i) $L'_\mu(x,\lambda) = \underbrace{f'(x) + c'(x)^T\lambda}_{=0} + \frac{2}{\mu}c'(x)^T\underbrace{c(x)}_{=0} = 0$

(ii) $L''_\mu(x,\lambda) = \underbrace{L_{xx}(x,\lambda)}_{H} + \frac{2}{\mu}c'(x)^Tc'(x) + \frac{2}{\mu}\underbrace{c(x)^T}_{=0}c''(x)$

$= H + \frac{2}{\mu}C^TC$

Let $u \in \mathbb{R}^n$ arbitrary. Then $\exists w \in \ker C,\ v \in \mathbb{R}^m$:

$u = w + C^Tv$

$u^TL''_\mu u = u^THu + \frac{2}{\mu}\|Cu\|^2$

$= w^THw + 2w^THC^Tv + v^TCHC^Tv + \frac{2}{\mu}\|CC^Tv\|^2$

$$\geq \alpha \|w\|^2 - 2\|HC^T\| \cdot \|w\| \cdot \|v\| - \|CHC^T\| \cdot \|v\|^2 + \frac{2}{\mu}\|CC^T v\|^2$$

$$= \underbrace{\left(\sqrt{\alpha}\|w\| - \frac{\|HC^T\|}{\sqrt{\alpha}}\|v\|\right)^2}_{\geq 0} + \underbrace{\left(\frac{2}{\mu}\beta - \|CHC^T\| - \frac{\|HC^T\|^2}{\alpha}\right)}_{>0 \text{ for suff. small } \mu.}\|v\|^2 \geq \gamma\|u\|^2 \quad \square$$

III.6.2 Theorem Let x_*, λ_* satisfy necessary & sufficient conditions & LICQ.

Then there is a neighborhood of (x_*, λ_*) such that for $\mu < \bar{\mu}$:

$x_\mu := \arg\min_x L_\mu(x, \lambda) \in U$ satisfies

$$\|x_\mu - x_*\| + \|\lambda - \tfrac{2}{\mu} c(x_\mu) - \lambda_*\| \le M \mu \|\lambda - \lambda_*\|$$

Proof $x_\mu(\lambda)$ exists as implicit function for $L'_\mu(x, \lambda) = 0$ in U.

Since L''_μ is spd (see proof III.6.1), x_μ is a minimizer.

$$\begin{bmatrix} 0 \\ c(x_\mu) \end{bmatrix} = L'(x_\mu, \lambda - \tfrac{2}{\mu} c(x_\mu)) = \underbrace{L'(x_*, \lambda_*)}_{=0} + \underbrace{L''(x_*, \lambda_*)}_{\text{invertible}} \begin{bmatrix} x_\mu - x_* \\ \lambda - \tfrac{2}{\mu} c(x_\mu) - \lambda_* \end{bmatrix} + o\left(\|x_\mu - x_*\| + \|\lambda - \tfrac{2}{\mu} c(x_\mu) - \lambda_*\|\right)$$

$$\Rightarrow \|x_\mu - x_*\| + \|\lambda - \tfrac{2}{\mu} c(x_\mu) - \lambda_*\| \le \|L''(x_*, \lambda_*)^{-1}\| \cdot \|c(x_\mu)\| + o\left(\|x_\mu - x_*\| + \|\cdot\|\right)$$

$$\Rightarrow \|x_\mu - x_*\| + \|\lambda - \tfrac{2}{\mu} c - \lambda_*\| \le 2\gamma \|c(x_\mu)\|$$

$$\Rightarrow \tfrac{2}{\mu} \|c(x_\mu)\| - \|\lambda - \lambda_*\| \le 2\gamma \|c(x_\mu)\|$$

$$\Rightarrow \|c(x_\mu)\| \le \tfrac{\mu}{2} (1 - \mu\gamma)^{-1} \|\lambda - \lambda_*\|$$

$$\Rightarrow \|x_\mu - x_*\| + \|\lambda - \tfrac{2}{\mu} c - \lambda_*\| \le \underbrace{\mu\gamma (1 - \mu\gamma)^{-1}}_{M} \|\lambda - \lambda_*\| \qquad \square$$

Apparently, $\lambda - \frac{2}{\mu} c(x_\mu)$ is a better estimate of λ_* than λ itself (by a factor μM): first order multiplier update.

III. 6.3 Algorithm (augmented Lagrangian method)

for $i = 1, \dots$

$$x_{i+1} = \arg\min_x L_\mu(x, \lambda_i)$$

$$\lambda_{i+1} = \lambda_i - \frac{2}{\mu} c(x_{i+1})$$

III. 6.4 Corollary AL method converges locally linearly if μ is sufficiently small.

Proof: $\|\lambda_{i+1} - \lambda_*\| \le \mu M \|\lambda_i - \lambda_*\|$ converges linearly

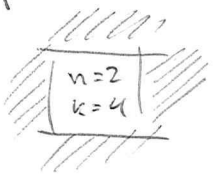
$\|x_{i+1} - x_*\| \le \mu M \|\lambda_i - \lambda_*\|$

□

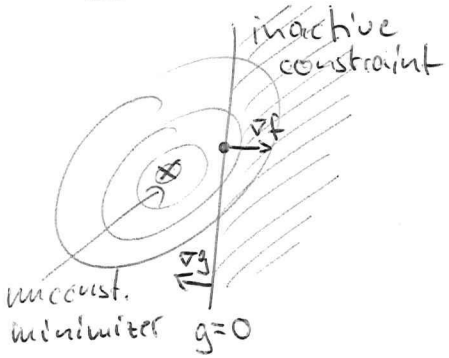
IV Inequality Constraints

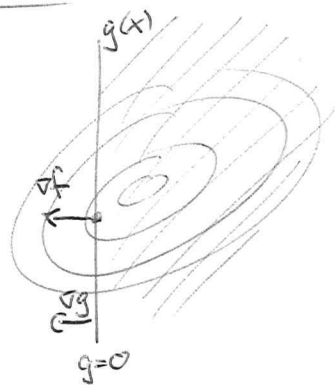
$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t. } g(x) \geq 0 \text{ component-wise in } \mathbb{R}^k$$

observation: $k > n$ possible, e.g.



IV.1 Optimality conditions





Necessary conditions: $\exists \lambda \in \mathbb{R}^k$:

$$\nabla f(x_*) = g'(x_*)^T \lambda \qquad \text{adjoint equation}$$

$$g(x_*) \geq 0, \quad \lambda \geq 0$$

$$\lambda^T g(x_*) = 0 \qquad \text{complementarity}$$

if constraint qualifications are satisfied
e.g. LICQ (linear independence of active constraint gradients).

IV. 1.1 Definition

active set $A(x) = \{ i \in \{1, \dots, k\} \mid g_i(x) \leq 0 \}$

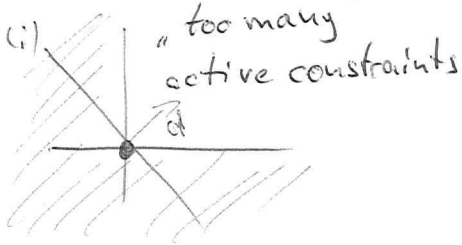
inactive set $I(x) = \{ i \in \{1, \dots, k\} \mid g_i(x) > 0 \}$

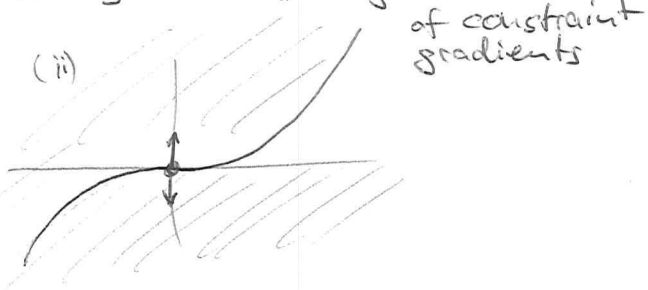
Sometimes, LICQ is too restrictive.

Example

(i) $\min x+y$ s.t. $x \geq 0$, $y \geq 0$, $x+y \geq 0$

(ii) $\min x$ s.t. $y \geq 0$, $x^3 - y \geq 0$





Mangasarian - Fromovitz - CQ (MFCQ):

$\exists\, d \in \mathbb{R}^n : g_A'(x_*)\, d > 0$ ("all constraints point in similar directions")

(i) MFCQ ✓

(ii) MFCQ ×

IV.2 Active set methods

Idea: ignore inactive constraints with $g_i(x_*) > 0$.
then $\min_x f(x)$ s.t. $g_A(x) = 0$ has solution x_*.

strategy: (i) find correct active set
(ii) solve equality constrained problem

IV.2.1 Simple algorithm

$A_0 \leftarrow \emptyset$

for $k = 0, \ldots$

$x_k \leftarrow \arg\min f(x)$ s.t. $g_{A_k}(x) = 0$

if KKT satisfied

break

$A_{k+1} \leftarrow A(x_k) \setminus \{i \mid \lambda_i \leq 0\}$

nasty details:

- solvability of equality constrained subproblems
- cycling: choose an active set A_k encountered before

efficiency:

- in practice, often good
- in theory, exponential run time possible
(combinatorial number of possible active sets)

IV.3 Penalty & barrier methods

Idea: penalize infeasibility

$$\min f(x) + \frac{1}{\mu} \sum_i \min(0, g_i(x))^2$$

consider $\mu \to 0$

- same approach as for equality constraints
- similar performance, same issues (infeasible iterates, ill-conditioned Hessians, banana-shaped valleys)

augmented Lagrangian approach possible as well

Idea: penalize approach to infeasibility

$$\min f(x) - \mu \sum_i \log g_i(x)$$

consider $\mu \to 0$.

- feasible iterates, but starting points are difficult to define
- ill-conditioned Hessians
- path-following (homotopy) for $\mu \to 0$ (central path) is efficient, both practically and theoretical: polynomial run time for linear problems.