Chapter 4: Unconstrained Nonlinear Optimization

Edoardo Amaldi

DEIB – Politecnico di Milano
edoardo.amaldi@polimi.it

Academic year 2023-24

4.1 Examples

1) Statistical estimation

Random variable X with density $f(x, \underline{\theta})$, where $\underline{\theta} \in \mathbb{R}^m$ is parameter vector, and independent observations $x_1, \ldots, x_n$.

Maximum likelihood: Estimates $\hat{\underline{\theta}}$ of $\underline{\theta}$ are derived by maximizing

$$L(\underline{\theta}) = f(x_1, \underline{\theta}) \; f(x_2, \underline{\theta}) \ldots f(x_n, \underline{\theta})$$

product of non linear functions
(well, actually, non linear)
and unconstrained

Assumption: $\exists \; \underline{\theta}$ for which all factors are positive.

Since $\ln(.)$ is monotonically increasing, $\hat{\underline{\theta}}$ also maximizes

$$\ln(L(\underline{\theta})) = \sum_{j=1}^{n} \ln(f(x_j, \underline{\theta}))$$

If f is differentiable w.r.t. $\underline{\theta}$ at $\hat{\underline{\theta}}$, necessary optimality conditions:

$$\sum_{j=1}^{n} \frac{\nabla_{\underline{\theta}} f(x_j, \hat{\underline{\theta}})}{f(x_j, \hat{\underline{\theta}})} = \underline{0}$$

For Guassian density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$$

and $\underline{\theta} = (\mu, \sigma)$, we obtain

$$\ln(L(\underline{\theta})) = \ln((\frac{1}{\sigma\sqrt{2\pi}})^n \prod_{j=1}^{n} \exp -\frac{(x_j-\mu)^2}{2\sigma^2}) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) + -\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2$$

Minimum is achieved in a stationary point:

$$\begin{cases} \dfrac{\partial[\ln(L(\underline{\theta}))]}{\partial\mu} = \dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^{n}(x_j-\mu) \overset{!}{=} 0 \\ \text{and} \\ \dfrac{\partial[\ln(L(\underline{\theta}))]}{\partial\sigma} = -\dfrac{n}{\sigma} + \dfrac{1}{\sigma^3}\displaystyle\sum_{j=1}^{n}(x_j-\mu)^2 \overset{!}{=} 0 \end{cases}$$

Thus

$$\Rightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} x_j, \qquad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j-\hat{\mu})^2}$$

2) Training multilayer neural networks

Supervised learning:

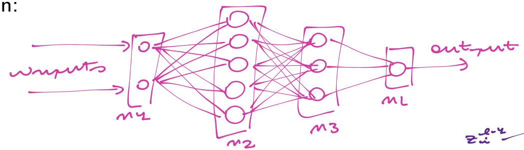
Given a training set $T = \{(\underline{x}^1, \underline{y}^1), \dots, (\underline{x}^p, \underline{y}^p)\}$ where $\underline{y}^k \in [0,1]^{n_{out}}$ desiderd output for $\underline{x}^k \in \mathbb{R}^{n_{in}}$, construct a model that maps $\underline{x}^k$'s into $\underline{y}^k$'s as well as possible.

Multilayer networks:

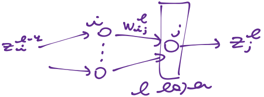
L layers with n_l units in layer l, $n_1 = n_{in}$ and $n_L = n_{out}$.

First layer of inputs $x_1, \dots, x_{n_1}$, other layers with activation units.

Illustration:



Output of unit j of layer l:



$$z_j^l = \phi\left(\sum_{i=1}^{n_{l-1}} w_{ij}^l z_i^{l-1} - w_{0j}^l\right)$$

where weights w_{ij} to be determined and $\phi : \mathbb{R} \to \mathbb{R}$ is sigmoid $\phi(t) = \frac{1}{1+e^{-t}}$.

A multilayer network defnines a mapping $h(\underline{w}, .)$ from $\mathbb{R}^{n_1}$ to $\mathbb{R}^{n_L}$ parametrized by $\underline{w} = \{w_{ij}^l : l = 1, \ldots, L; i = 1, \ldots, n_{l-1}; j = 1, \ldots, n_l\}$.

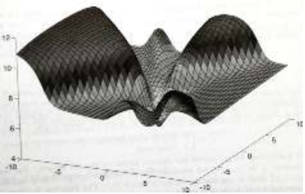
Training problem: Given $T = \{(\underline{x}^1, \underline{y}^1), \ldots, (\underline{x}^p, \underline{y}^p)\}$, determine values of $\underline{w}$ which approximate as well as possible the mapping underlying T.

In general one minimizes

$$\frac{1}{2}\sum_{k=1}^{p}(\|\underline{y}^k - h(\underline{w}, \underline{x}^k)\|_)^2$$

highly non linear and non convex function

challenging (non convex)



Example 1.5.3 of D. Bertsekas, Nonlinear Programming, Athena Scientic 1999.

4.2 Optimality conditions

Generic problem:

$$\min_{\underline{x} \in S} f(\underline{x})$$

simple constraints, eg. the non-negativity

where $S \subseteq \mathbb{R}^n$, $f: S \to \mathbb{R}$ and $f \in \mathcal{C}^1$ or $\mathcal{C}^2$.

Unconstrained case: $S = \mathbb{R}^n$

→ we don't care about "$\underline{d}$", we consider $\underline{d}$ just as the direction we move

Definition: $\underline{d} \in \mathbb{R}^n$ is a *feasible direction* at $\overline{x}$ if

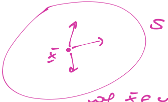
longest step length we could take

$$\exists\ \overline{\alpha} > 0 \text{ such that } \overline{x} + \alpha \underline{d} \in S \qquad \forall \alpha \in [0, \overline{\alpha}] \tag{1}$$

Illustrations:

we ask if we can make strictly positive steps in the direction given by $\underline{d}$ and still remaining in S

set of all the feasible dirs

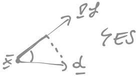


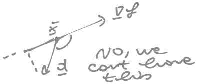
if $\overline{x} \in \text{int}(S)$ then every dir^n is a feasible dir

First order necessary local optimality conditions:

If $f \in \mathcal{C}^1$ on S and $\underline{x}$ is a *local minimum* of f over S, then for any feasible direction $\underline{d} \in \mathbb{R}^n$ at $\underline{x}$

$$\boxed{\nabla^t f(\underline{x})\underline{d} \geq 0,}$$





namely all feasible directions are ascent directions.
no worsening direction

Proof:

Consider $\varphi : [0, \bar{\alpha}] \to \mathbb{R}$ st $\varphi(\alpha) = f(\bar{x} + \alpha \underline{d})$.
Since $\bar{x}$ is a local min (opt) of f on S, then $\alpha = 0$ is a local min of φ on $[0, \bar{\alpha}]$ (so that we remain in $\bar{x}$).

Being $f \in C^1$, also $\varphi \in C^1$, and the Taylor series of φ at $\alpha = 0$ is

$$\varphi(\alpha) = \varphi(0) + \alpha \varphi'(0) + o(\alpha)$$

note: $u(\alpha) = o(\alpha)$ if $u(\alpha) \to 0$ faster than α, when $\alpha \to 0$ itself

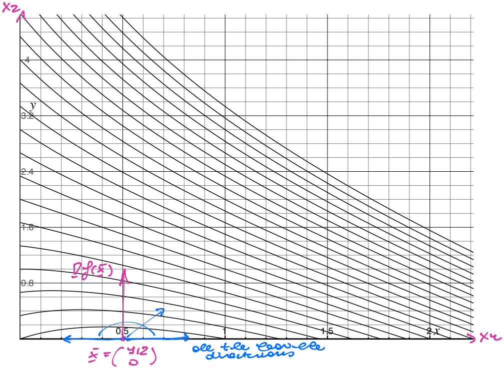
Suppose that $\varphi'(0) < 0$. Then if $\alpha \to 0^+$ we can neglect the remainder $o(\alpha)$ and we have

$$\varphi(\alpha) - \varphi(0) = \alpha \varphi'(0) < 0 \Rightarrow$$

0 would not be a local min of φ, absurd

So we must have $\varphi'(0) \geq 0$, and since $\varphi'(\alpha) = \nabla f(\bar{x} + \alpha \underline{d}) \cdot \underline{d}$ we have the condition of the thesis.

Example: $\min_{x_1,\ x_2 \geq 0}\ f(x_1, x_2) = x_1^2 - x_1 + x_2 + x_1 x_2$



$\Rightarrow$ $\bar{x}$ is a (global) min because $\nabla f(\bar{x}) \cdot \underline{d} \geq 0$ for all feasible directions $\underline{d}$ at $\bar{x}$, even if $\nabla f(\bar{x}) \neq 0$

so the first order "classical" conditions (of derivative = 0) are no more valid here, we have to introduce feasible directions

Second order necessary local optimality conditions:

If $f \in \mathcal{C}^2$ on S and $\underline{\overline{x}}$ is a *local minimum* of f over S then

i) $\nabla^t f(\underline{\overline{x}})\underline{d} \geq 0$ $\quad\quad \forall \underline{d} \in \mathbb{R}^n$ feasible direction at $\underline{\overline{x}}$, the 1st order condition we just saw

ii) if $\nabla^t f(\underline{\overline{x}})\underline{d} = 0$ $\quad$ then $\quad \underline{d}^t \nabla^2 f(\underline{\overline{x}})\underline{d} \geq 0$. $\Leftrightarrow$ $\varphi''(0) \geq 0$

the vectors are orthogonal

$\underline{d} \cdot H \underline{d}$ is ≥ 0, for that $\underline{d}$ at which we have $\perp$.

Proof:

Similarly for (i). Suppose $\nabla^t f(\underline{\overline{x}})\underline{d} = 0$, then we can expand the Taylor series and we get

$$\phi(\alpha) = \phi(0) + \alpha \underbrace{\phi'(0)}_{0} + \frac{1}{2}\alpha^2\phi''(0) + o(\alpha^2).$$

If $\varphi''(0) = \underline{d} \cdot \nabla^2 f(\underline{x})\, \underline{d} < 0$, we can reason as before and we would get

$$\varphi(\alpha) - \varphi(0) = \frac{1}{2}\alpha^2 \varphi''(0) < 0 \Rightarrow$$ for sufficiently small α we would get a better loc min of φ rather than 0, absurd

So we must have $\varphi''(0) \geq 0$ which is the thesis

Corollary: (Unconstrained case)

If $f \in \mathcal{C}^2$ on S and $\underline{x}$ $\boxed{\in int(S)}$ is a *local minimum* of f over S, then

1. $\nabla f(\underline{x}) = \underline{0}$ (stationarity condition)
2. $\nabla^2 f(\underline{x})$ is positive semidefinite.

Proof: Since $\underline{x} \in int(S)$, all the vectors $\underline{d} \in \mathbb{R}^n$ are feasible directions at $\underline{x}$.

So from the previous condition (i) we have $\nabla f(\underline{x}) \cdot \underline{d} \geq 0$ $\forall \underline{d}$ and $-\underline{d}$, so we get (1) here.

While (2) is a consequence of condition (ii) since $\underline{d} \cdot (\nabla^2 f(\underline{x}))\, \underline{d} \geq 0$ $\forall \underline{d} \in \mathbb{R}^n$, so the hessian matrix $\nabla^2 f(\underline{x})$ is positive semidefinite.

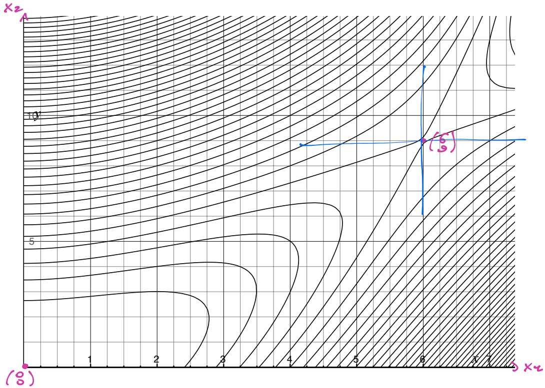
we would satisfy one of the previous conditions

Types of candidate points: local minima, local maxima and saddle points.

Above optimality conditions are not sufficient

eg just take $f(x) = x^3$, for which $x = 0$ is not a local minimum, even if it satisfies the conditions

Example: $\min_{x_1,\ x_2 \geq 0}\ f(x_1, x_2) = x_1^3 - x_1^2 x_2 + 2x_2^2$



Candidate points are $(0,0)$ and $(6,9)$.

- $\binom{6}{9}$ is not a loc min, even though for $x_1=6$ and $x_2=9$ is a loc min wrt x_2 and for $x_2=9$ and $x_1=6$ is a loc min wrt x_1 ($\Rightarrow$ is a saddle) extension (on a general set S) of the classical opt conditions

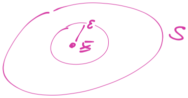
Sufficient local optimality conditions:

stationary point

If $f \in \mathcal{C}^2$ on S and $\underline{\overline{x}} \in int(S)$ such that $\nabla f(\underline{\overline{x}}) = \underline{0}$ and $\nabla^2 f(\underline{\overline{x}})$ is positive definite, then $\underline{\overline{x}}$ is a *strict local minimum* of f over S, namely

$$f(\underline{x}) > f(\underline{\overline{x}}) \qquad \forall \underline{x} \in \mathcal{N}_\epsilon(\underline{\overline{x}}) \cap S.$$

Proof:



Let $\underline{d} \in \mathcal{B}_\epsilon(\underline{0})$ be any feasible direction such that $\underline{\overline{x}} + \underline{d} \in S \cap \mathcal{B}_\epsilon(\underline{\overline{x}})$.

Then

$$f(\underline{\overline{x}} + \underline{d}) = f(\underline{\overline{x}}) + \underbrace{\nabla^t f(\underline{\overline{x}})}_{\underline{0}} \underline{d} + \frac{1}{2} \underline{d}^t \nabla^2 f(\underline{\overline{x}}) \underline{d} + o(\|\underline{d}\|^2)$$

Since $\nabla^2 f(\underline{\overline{x}})$ is pos def, then $\exists a > 0$ st $\underline{d} \cdot (\nabla^2 f(\underline{\overline{x}}))\, \underline{d} \geq a \cdot \|\underline{d}\|^2$ (where a is related to the smallest eigenval of the hessian).

Thus for $\|\underline{d}\|$ sufficiently small, we have that

$$f(\underline{\overline{x}} + \underline{d}) - f(\underline{\overline{x}}) \geq \frac{a}{2} \|\underline{d}\|^2 > 0$$

$\Rightarrow f(\underline{\overline{x}}) < f(\underline{\overline{x}} + \underline{d}) \Rightarrow \underline{\overline{x}}$ is a strict loc min along the direction $\underline{d}$

Since this holds $\forall \underline{d} \in \mathbb{R}^n$ such that $\underline{\overline{x}} + \underline{d} \in S \cap \mathcal{B}_\epsilon(\underline{\overline{x}})$, f is locally strictly convex.

Convex problems

$$\min_{x \in C \subseteq \mathbb{R}^n} f(\underline{x}) \qquad \text{where } C \subseteq \mathbb{R}^n \text{ convex and } f: \ C \to \mathbb{R} \text{ convex}$$

Every local minimum is a global minimum.

a stronger result for convex problems

Necessary and sufficient (NS) conditions:

Let f be convex and $\mathcal{C}^1$ on $C \subseteq \mathbb{R}^n$ convex. $\underline{x}^*$ is a *global minimum* of f on C if and only if

$$\boxed{\nabla^t f(\underline{x}^*)(\underline{y} - \underline{x}^*) \geq 0 \qquad \forall \underline{y} \in C.}$$

Proof:

nec. cond: if $f \in C^1$ and $\underline{x}^*$ is a loc min (or global also) then we have that (all feasible dirs are ascending, ie)

$$\nabla f(\underline{x}^*) \cdot \underline{d} \geq 0 \qquad \forall \underline{d} \text{ feasible dir at } \underline{x}^*$$

ie $\forall \underline{d} = \underline{y} - \underline{x}^*$ where $\underline{y} \in C$

suf. cond: here we use the characterization of convexity:

f is convex $\iff f(\underline{y}) \geq f(\underline{x}^*) + \underbrace{\nabla f(\underline{x}^*) \cdot (\underline{y} - \underline{x}^*)}_{\geq 0 \text{ from nec cond}} \quad \forall \underline{y} \in C$

$\Rightarrow f(\underline{y}) \geq f(\underline{x}^*) \ \forall \underline{y} \in C$

$\Rightarrow \underline{x}^*$ is a glob opt

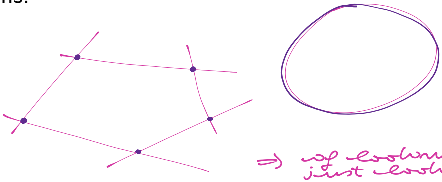
Recall: Given any $C \subseteq R^n$ convex, $\underline{x} \in C$ is an *extreme point* of C if it cannot be expressed as a convex combination of two different points of C.

eg we may want to maximize a linear function (which was convex & concave)

Property: (maximization of convex functions)

Let f be convex defined on C convex, bounded and closed. If f has a (finite) *maximum* over C, then $\exists$ an *optimal extreme point* of C.

Illustrations:



⇒ if looking for max points just look at extreme points

it is basically the main result of linear programming (LP)

Special case: Linear programming

4.3 Iterative methods and convergence

relevant since all methods for constrained or not solvers are iterative in nature

Generic Nonlinear Optimization (N.O.) problem:

new notation (w.l.o.g.) for the constraints

$$\begin{array}{lll} \min & f(\underline{x}) & \\ s.t. & \boxed{g_i(\underline{x}) \leq 0} & 1 \leq i \leq m \\ & \underline{x} \in S \subseteq \mathbb{R}^n & \end{array}$$

If $X = \{\underline{x} \in S \ : \ g_i(\underline{x}) \leq 0, \ 1 \leq i \leq m\} \subset \mathbb{R}^n$ then constrained problem.

Difficulty depends on f and X. Usually f and g_i are at least continuously differentiable.

In some cases (e.g., LP and combinatorial optimization) an optimal solution can be found in a finite number of elementary operations.

Efficiency depends on how this number grows with the instance size (polynomial vs exponential).

(i.e. # variables, # constraints, ecc

Most N.O. methods are **iterative**

- start from $\underline{x}_0 \in X$
- generate a sequence $\{\underline{x}_k\}_{k \geq 0}$ that "converges" to a point of $\Omega = \{$"desired solutions"$\}$.

Different meanings of "**converge**" and "**desired solutions**":

- $\{\underline{x}_k\}_{k \geq 0}$ converges to a point of Ω

  or $\exists$ a limit point of $\{\underline{x}_k\}_{k \geq 0}$ which belongs to Ω

  a good estimate after sufficient number of iterations

  while here we really converge in the classical sense

- $\Omega =$ set of global optima

  or $\Omega =$ set of candidate points satisfying 1st/2nd order necessary optimality conditions

  e.g. if $X = \mathbb{R}^n$ we can have $\Omega = \{\underline{x} \in \mathbb{R}^n : \nabla f(\underline{x}) = \underline{0}\}$

Often but not always descent methods: $f(\underline{x}_{k+1}) < f(\underline{x}_k)$ for each k

Interested in robust and efficient methods.

1) **Robustness** associated to global convergence

i.e. normally it does not matter the x_0 chosen from where we start

Definition: An algorithm is **globally** (**locally**) **convergent** if $\{\underline{x}_k\}_{k\geq 0}$ satisfies one of previous properties for any $\underline{x}_0 \in X$ (only for $\underline{x}_0$ in a neighborhood of an $\underline{x}^* \in \Omega$).

a candidate point

2) **Efficiency** characterized by convergence speed

Assume that $\lim_{k\to\infty} \underline{x}_k = \underline{x}^*$ where $\underline{x}^* \in \Omega$

Definitions: $\{\underline{x}_k\}_{k\geq 0}$ *converges* to $\underline{x}^*$ with *order* $p \geq 1$ if $\exists\ r > 0$ and $k_0 \in \mathbb{N}$ such that

$$\boxed{\|\overbrace{\underline{x}_{k+1} - \underline{x}}^{err(k+1)}{}^*\| \leq r\ \|\overbrace{\underline{x}_k - \underline{x}}^{err(k)}{}^*\|^p \qquad \forall k \geq k_0.}$$

Largest p is the **order of convergence** and smallest $r > 0$ is the **rate**.

If $p = 1$ and $r < 1$ **linear** convergence, if $p = 1$ and $r \geq 1$ **sublinear** convergence.

this one is slower than

N.B.: If $p = 1$ the distance w.r.t. $\underline{x}^*$ decreases at each iteration by a factor r.

Example: (E1) Consider $\underbrace{1+\frac{1}{k}}_{a_k} \xrightarrow{k\to+\infty} 1$

- verify that there is a linear convergence
- and that the rate is 1.

Actually, we need to study the ratio of LHS/RHS:

$$\frac{\|e(k+1)\|}{\|e(k)\|^p} = \frac{|a_{k+1}-1|}{|a_k-1|^p} = \frac{\frac{1}{k+1}}{\left(\frac{1}{k}\right)^p} = \frac{k^p}{k+1} \overset{k\to+\infty}{\sim} k^{p-1} \longrightarrow \begin{cases} 0 & p<1 \\ \boxed{1 \quad p=1} \\ +\infty & p>1 \end{cases}$$

$\Rightarrow$ only for that p we get that $\exists$ an r

(E2) Study similarly $a_k = 1+\frac{1}{2^k}$ (ref: $r=1/2$)

Definition: The **convergence** is **superlinear** if there exists $\{r_k\}_{k\geq 0}$ with $\lim_{k\to\infty} r_k = 0$ such that

$$\|\underline{x}_{k+1} - \underline{x}^*\| \leq r_k \, \|\underline{x}_k - \underline{x}^*\| \qquad \forall k \geq k_0.$$

FASTNESS
⋮
quadratic
super-linear
linear
sub-linear
⋮

Example: $1 + \frac{1}{k^k}$

Definition: If $p = 2$ (and r not necessarily < 1), the **convergence** is **quadratic**.

Example: $1 + \frac{1}{2^{2^k}}$

4.4 Line search methods

Unconstrained optimization problem:

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

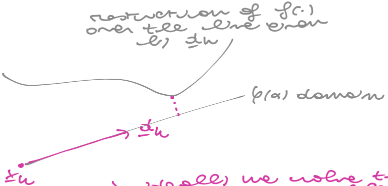
otherwise the prob
is unbounded (and
is not interesting)

with $f : \mathbb{R}^n \to \mathbb{R}$ of class $\mathcal{C}^1$ or $\mathcal{C}^2$ and bounded below.

Iterative methods: start from $\underline{x}_0 \in \mathbb{R}^n$ and generate $\{\underline{x}_k\}_{k \geq 0}$ "converging" to an $\overline{x} \in \Omega$.

See Chap. 3 of J. Noceal, S. Wright, Numerical Optimization, Springer 1999.

1) General scheme



Select $\underline{x}_0$ and $\varepsilon > 0$, set $k := 0$

Repeat

Choose *search direction* $\underline{d}_k \in \mathbb{R}^n$

=> ideally we solve this small min problem to get α_k

Determine *step length* $\alpha_k > 0$ along $\underline{d}_k$ s.t. $\quad \min_{\alpha \geq 0} \ \phi(\alpha) = f(\underline{x}_k + \alpha \underline{d}_k)$

costly problem, but we will see that we don't have to solve it at optimality necessarily

Set $\underline{x}_{k+1} := \underline{x}_k + \alpha_k \underline{d}_k$ and $k := k + 1$

Until termination criterion is satisfied

Termination criterion: $\|\nabla f(\underline{x}_k)\| < \varepsilon$ or $|f(\underline{x}_k) - f(\underline{x}_{k+1})| < \varepsilon$ or $\|\underline{x}_{k+1} - \underline{x}_k\| < \varepsilon$

Often approximate α_k (also $f(\underline{x}_{k+1}) < f(\underline{x}_k) \ \forall k \geq 0$).

Flexibility in choice of $\underline{d}_k$ and α_k, efficiency depends on both.

2) Search directions

In many *line search methods*, i.e., iterative methods based on search directions,

$$\boxed{\underline{d}_k = -D_k \nabla f(\underline{x}_k)}$$

with positive definite $n \times n$ matrix D_k.

$\underline{d}_k$ is a *descent direction* because of the matrix D_k being pos def

Now allow this we can compute

$$\nabla f(\underline{x}_k) \cdot \underline{d}_k = \nabla f(\underline{x}_k) \cdot \left(-D_k \nabla f(\underline{x}_k)\right) =$$

$$= -\underbrace{(\underline{z} \cdot A \underline{z})}_{>0} < 0$$

Example 1: Gradient method

Given $f \in \mathcal{C}^1$, consider linear approximation of $f(\underline{x}_k + \underline{d})$ at $\underline{x}_k$

$$l_k(\underline{d}) := f(\underline{x}_k) + \nabla^t f(\underline{x}_k)\underline{d}$$

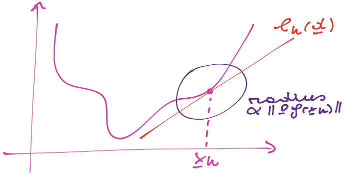
and choose $d_k \in \mathbb{R}^n$ minimizing $l_k(\underline{d})$ over sphere of radius $\|\nabla f(\underline{x}_k)\|$:

then about α_k we are free

region where we look for $\underline{d}$

$$\begin{array}{ll} \min & \nabla^t f(\underline{x}_k)\underline{d} \\ \text{s.t.} & \|\underline{d}\| = \|\nabla f(\underline{x}_k)\|. \end{array} \tag{1}$$

since $\nabla f(\underline{x}_k) \cdot \underline{d} = \|\nabla f(\underline{x}_k)\| \cdot \|\underline{d}\| \cdot \cos\theta$
we minimize (1) when $\cos(\theta) = -1$
that is $\theta = \pi$



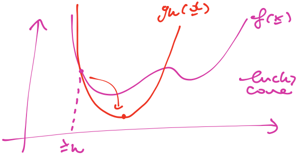
Steepest descent direction:

$$\boxed{\underline{d}_k = -\nabla f(\underline{x}_k)}$$

where $D_k = I_n$.

Clearly $\underline{d}_k$ is a descent direction if $\nabla f(\underline{x}_k) \neq \underline{0}$.

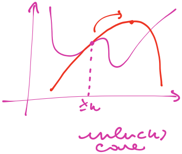
Example 2: Newton method



Given $f \in \mathcal{C}^2$ and $H(\underline{x}_k) = \nabla^2 f(\underline{x}_k)$.

Consider quadratic approximation of $f(\underline{x}_k + \underline{d})$ at $\underline{x}_k$

$$q_k(\underline{d}) := f(\underline{x}_k) + \nabla^t f(\underline{x}_k)\underline{d} + \frac{1}{2}\underline{d}^t H(\underline{x}_k)\underline{d}$$



and choose $d_k \in \mathbb{R}^n$ and α_k leading to a stationary point of $q_k(\underline{d})$.

Since $\nabla_{\underline{d}}\ q_k(\underline{d}) = \underline{0}$ implies $\nabla^t f(\underline{x}_k) + \underline{d}^t H(\underline{x}_k) = \underline{0}$, if $H^{-1}(\underline{x}_k)$ exists then

Newton direction:

$$\boxed{\underline{d}_k = -H^{-1}(\underline{x}_k)\nabla f(\underline{x}_k),}$$

where $D_k = H^{-1}(\underline{x}_k)$.

$\nabla f(\underline{x}_k) \cdot \underline{d}_k = \nabla f \cdot (-H^{-1}\nabla f) =$
$= -\underbrace{(\underline{a} \cdot H^{-1}\underline{a})}_{>0 \text{ if } H \text{ is pos def}} < 0$

If $H(\underline{x}_k)$ is p.d. and $\nabla f(\underline{x}_k) \neq \underline{0}$, $\underline{d}_k$ is a descent direction.

If $H(\underline{x}_k)$ is not p.d., $\underline{d}_k$ may not be defined ($\nexists\ H^{-1}(\underline{x}_k)$) or may be an ascent direction.

3) Step length

To guarantee global convergence, an approximate solution α_k of *line search*:

$$\min_{\alpha \geq 0} \ \phi(\alpha) = f(\underline{x}_k + \alpha \underline{d}_k).$$

is sufficient.

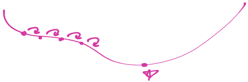
Different methods to generate α_k and stop when appropriate conditions are satisfied (simple, after a few iterations).

$f(\underline{x}_k + \alpha_k \underline{d}_k) < f(\underline{x}_k)$ does not suffice.

Basic principles:

eg if the real opt is far away

- α must not be too small (to avoid premature convergence)



- α must not be too large (to avoid oscillations)

or even we might diverge

Wolfe conditions:

Sufficient reduction (decrease):

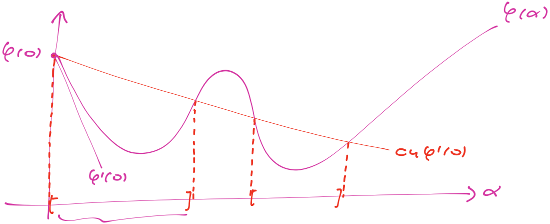
$$\boxed{\phi(\alpha) \leq \phi(0) + c_1 \alpha \phi'(0) \qquad \text{con} \quad c_1 \in [0,1]}$$

which is equivalent to

$$f(\underline{x}_k + \alpha \underline{d}_k) \leq \underbrace{f(\underline{x}_k)}_{\varphi(0)} + c_1 \alpha \underbrace{\nabla^t f(\underline{x}_k) \underline{d}_k}_{\varphi'(0)} \qquad \text{(Armijo criterion)}$$

$\phi'(0) < 0$ since $\underline{d}_k$ is a descent direction, $c_1 \leq 1/2$ so that it is satisfied by the minimum of a quadratic convex $\phi(\alpha)$ (exercise set n.6).

Illustration:



too small α's, so we now introduce another term $c_2 \varphi'(0)$

To avoid too small steps also condition:

$\Rightarrow c_1 < c_2 < 1$

$$\boxed{\phi'(\alpha) \geq c_2 \phi'(0) \qquad \text{con} \quad c_2 \in (c_1, 1)}$$

which is equivalent to

$$\nabla^t f(\underline{x}_k + \alpha \underline{d}_k)\underline{d}_k \geq c_2 \nabla^t f(\underline{x}_k)\underline{d}_k.$$

In general $c_2 = 0.9$ for (quasi)-Newton and $c_2 = 0.1$ for non-linear conjugate gradient.

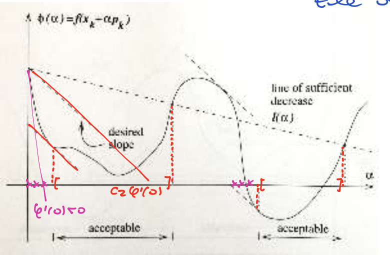
Weak Wolfe conditions:

$$\boxed{\begin{aligned} \phi(\alpha) &\leq \phi(0) + c_1 \alpha \phi'(0) \qquad (2) \\ \phi'(\alpha) &\geq c_2 \phi'(0) \qquad (3) \end{aligned}}$$

with $0 < c_1 < c_2 < 1$.

we can also cut too large values (i.e., since the old value here e.g.)

there are also other conditions but these are the most famous ones



we run through the α axis, check the value there of $\varphi'(\alpha)$, and basing on the conditions we decide to take (accept) or not that α value

See Chap. 3 of J. Noceal, S. Wright, Numerical Optimization, Springer 1999.

Strong Wolfe conditions:

$$\boxed{\begin{aligned} \phi(\alpha) &\leq \phi(0) + c_1 \alpha \phi'(0) \quad &(4) \\ |\phi'(\alpha)| &\leq c_2 |\phi'(0)| \quad &(5) \end{aligned}}$$

with $0 < c_1 < c_2 < 1$.

Exclude values of α with $\phi'(\alpha)$ too positive, far from stationary points of ϕ.

Conditions are invariant w.r.t. affine transformation of the variables.

Proposition:

If $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^1$ and $\underline{d}_k$ descent direction at $\underline{x}_k$ such that f is bounded below along $\{\underline{x}_k + \alpha \underline{d}_k \ : \ \alpha > 0\}$. Then if $0 < c_1 < c_2 < 1$ there exist intervals of step lengths satisfying the Wolfe conditions (weak and strong).

Simple consequence of the mean value theorem.

Still we need to see how can
we easily find such α values.
We will focus on the weak Wolfe case.

Metods for 1-D search

Many methods (with/without derivatives) to determine an approximate solution α_k of

$$\min_{\alpha \geq 0} \; \phi(\alpha) = f(\underline{x}_k + \alpha \underline{d}_k)$$

satisfying appropriate conditions (e.g. Wolfe) which guarantee global convergence.

In general, two phases:

- determine $[\alpha_{min}, \alpha_{max}]$ containing "acceptable" step lengths ("bracketing phase"),
- select a good value α within $[\alpha_{min}, \alpha_{max}]$ via bisection or interpolation.

Bisection

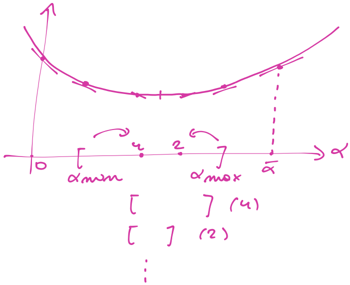
$\phi \in \mathcal{C}^1$, $\phi'(0) < 0$ since $\underline{d}_k$ descent direction and $\exists\, \overline{\alpha}$ such that $\phi'(\alpha) > 0$ for $\alpha \geq \overline{\alpha}$.

Start from $[\alpha_{min}, \alpha_{max}]$ with $\phi'(\alpha_{min}) < 0$ and $\phi'(\alpha_{max}) > 0$ and iteratively reduce it.

Iteration: set $\tilde{\alpha} = \frac{1}{2}(\alpha_{min} + \alpha_{max})$

if $\phi'(\tilde{\alpha}) > 0$ **then** $\alpha_{max} := \tilde{\alpha}$

if $\phi'(\tilde{\alpha}) < 0$ **then** $\alpha_{min} := \tilde{\alpha}$



Linear convergence with rate $1/2$

To find initial $[\alpha_{min}, \alpha_{max}]$:

1) $\alpha_{min} := 0$ e $s := s_0$

2) compute $\phi'(s)$

if $\phi'(s) < 0$ **then** $\alpha_{min} := s$, $s := 2s$, **goto** 2)

if $\phi'(s) > 0$ **then** $\alpha_{max} := s$, **stop**

Adaptation to determine α_k satisfying Wolfe conditions.

<u>Procedure</u>:

i) select $\alpha > 0$ and set $\alpha_{min} = \alpha_{max} = 0$

ii) **if** α satisfies Wolfe (2) **then** goto iii)

else $\alpha_{max} := \alpha$, $\alpha := \frac{\alpha_{min}+\alpha_{max}}{2}$, **goto** ii)

iii) **if** α satisfies Wolfe (3) **then** $\alpha_k = \alpha$, **stop**

else $\alpha_{min} := \alpha$

$$\alpha := \begin{cases} 2\alpha_{min} & \text{if } \alpha_{max} = 0 \\ \frac{\alpha_{min}+\alpha_{max}}{2} & \text{if } \alpha_{max} > 0 \end{cases}$$

goto ii)

Proposition: If $f \in \mathcal{C}^1$ is bounded below along ray $\{\underline{x}_k + \alpha \underline{d}_k \ : \ \alpha \geq 0\}$, the procedure stops after a finite number of iterations and yields α_k satisfying Wolfe conditions.

4) Global convergence of line search methods

Suitable assumptions on $\boxed{\alpha_k \text{ and } \underline{d}_k}$ can guarantee global convergence.

steepest descent (with −) direction

Key aspect: angle θ_k between $\underline{d}_k$ and $-\nabla f(\underline{x}_k)$

$$\cos(\theta_k) = -\frac{\nabla^t f(\underline{x}_k)\underline{d}_k}{\|\nabla f(\underline{x}_k)\|\|\underline{d}_k\|}$$

since $\underline{a} \cdot \underline{b} = \|\underline{a}\| \cdot \|\underline{b}\| \cdot \cos(\theta)$

General result showing how far $\underline{d}_k$ can deviate from $-\nabla f(\underline{x}_k)$ and still give rise to globally convergent iterations.

For a proof assuming weak Wolfe conditions, see J. Noceal, S. Wright, Numerical Optimization, Springer 1999, p. 43-44.

Theorem: (Zoutendijk)

Consider any line search method iteration with descent $\underline{d}_k$ and α_k satisfying Wolfe conditions. Suppose f is bounded below on $\mathbb{R}^n$, $f \in \mathcal{C}^1$ on open set N containing $L_0 = \{\underline{x} \in \mathbb{R}^n \, : \, f(\underline{x}) \leq f(\underline{x}_0)\}$ and $\nabla f(\underline{x})$ is Lipschitz continuous on N, i.e., $\exists \; L > 0$ such that

level set

$$\|\nabla f(\underline{x}) - \nabla f(\overline{x})\| \leq L \|\underline{x} - \overline{x}\| \qquad \forall \underline{x}, \overline{x} \in N.$$

models bounded curvature

Then

$$\boxed{\sum_{k \geq 0} \cos^2(\theta_k) \|\nabla f(\underline{x}_k)\|^2 < +\infty.} \tag{6}$$

we sum over all iterations

if the series does not diverge it means that the argument → 0 for k→∞, that is:

$$\underbrace{\cos(\theta_k)^2}\ \underbrace{\|\nabla f(\underline{x}_k)\|^2} \xrightarrow{k \to +\infty} 0$$

if this is $\geq \delta > 0$ ⇒ this → 0

This holds if the directions d_k are not too close to orthogonal w.r.t. $-\nabla f(\underline{x}_k)$, i.e. d_k are still "close" to the $-\nabla f(\underline{x}_k)$ direction

⇒ the gradient method is globally convergent, i.e.

$$\lim_{k \to +\infty} \|\nabla f(\underline{x}_k)\| = 0 \quad \forall \underline{x}_0$$

Consequence: The gradient method ($\cos \theta_k = 1$) satisfying Wolfe conditions is globally convergent.

If D_k symmetric and p.d. $\forall k \geq 0$ and $\exists$ constant M such that

$$\|D_k\| \|D_k^{-1}\| \leq M \qquad \forall k \geq 0$$

(bounded condition number), it can be verified that

$$\cos\theta_k \geq 1/M.$$

the δ of before is therefore related to the condition number of D_k

In such cases Newton and quasi-Newton methods are globally convergent.

4.5 Gradient method

(+ computationally light
+ globally convergent
− slow convergence)

Given $f : \mathbb{R}^n \to \mathbb{R}$ with $f \in \mathcal{C}^1$, look for a stationary point.

Gradient method with exact 1-D search:

exact since we set/choose the α_k by solving at optimality that 1-D search problem

Choose $\underline{x}_0$, set $k := 0$

Iteration k:

$$\underline{d}_k := -\nabla f(\underline{x}_k)$$

Determine $\alpha_k > 0$ such that $\min_{\alpha \geq 0} \; \phi(\alpha) = f(\underline{x}_k + \alpha \underline{d}_k)$

$f(\cdot)$ restricted on values on the direction $\underline{d}_k$

$$\underline{x}_{k+1} := \underline{x}_k + \alpha_k \underline{d}_k$$

$$k := k + 1$$

Termination criteria: $\|\nabla f(\underline{x}_k)\| < \varepsilon$ or $|f(\underline{x}_k) - f(\underline{x}_{k+1})| < \varepsilon$ or $\|\underline{x}_{k+1} - \underline{x}_k\| < \varepsilon$.

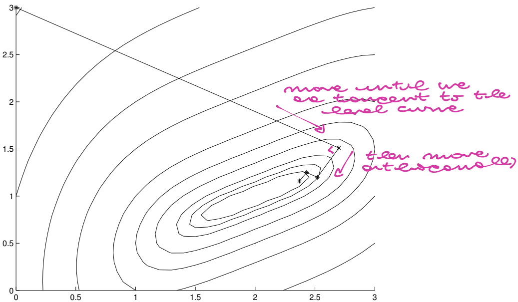
Property: If 1-D search is exact, the successive directions are orthogonal.

since we know that

$\varphi'(\alpha_k) = \nabla f(\underline{x}_k + \alpha_k \underline{d}_k) \cdot \underline{d}_k = \nabla f(\underline{x}_{k+1}) \cdot \underline{d}_k \overset{!}{=} 0$

$\Rightarrow \underline{d}_{k+1} \cdot \underline{d}_k = -\nabla f(\underline{x}_{k+1}) \cdot \underline{d}_k = 0$

this is bad news!



Example: zig-zag trajectory, very slow convergence

We first consider the case of quadratic strictly convex functions.

Any $\mathcal{C}^2$ function can be well approximated around any local/global minimum by such a function.

Quadratic strictly convex functions:

$$f(\underline{x}) = \frac{1}{2}\underline{x}^t Q \underline{x} - \underline{b}^t \underline{x} \quad \text{with } Q \text{ symmetric and p.d.}$$

the Hessian matrix

at Hessian of strictly convex

Global minimum is unique solution of $Q\underline{x} = \underline{b}$ ($\nabla f(\underline{x}) = \underline{0}$) and α_k can be determined explicitly:

$$\phi(\alpha) = f(\overbrace{\underline{x}_k - \alpha \underbrace{\nabla f(\underline{x}_k)}_{d_k}}^{x_{k+1}}) = \frac{1}{2}(\underline{x}_k - \alpha\nabla f(\underline{x}_k))^t Q(\underline{x}_k - \alpha\nabla f(\underline{x}_k)) - \underline{b}^t(\underline{x}_k - \alpha\nabla f(\underline{x}_k))$$

$$\Rightarrow \varphi'(\alpha) = \frac{\partial}{\partial\alpha}\varphi(\alpha) = (-\nabla f(x_k))^T Q (x_k - \alpha \nabla f(x_k)) - \underline{b}^T(-\nabla f(x_k)) \overset{!}{=} 0$$

since $\nabla f(x_k)^T = x_k^T Q - \underline{b}^T$ (derivative of wrt x) we end $\underline{b}^T = x_k^T Q - \nabla f(x_k)^T$ and we plug it in to get

$$\varphi'(\alpha) = -\nabla f(x_k)^T Q x_k + \alpha \nabla f(x_k)^T Q \nabla f(x_k) + (-\nabla f(x_k)^T + x_k^T Q \nabla f(x_k)) \overset{!}{=} 0$$

so isolating α_k we get

$$\boxed{\alpha_k = \frac{\nabla f(x_k)^T \nabla f(x_k)}{\nabla f(x_k)^T Q \nabla f(x_k)} = \frac{d_k^T d_k}{d_k^T Q d_k}}$$

Convergence analysis

Often consider convergence rate of $f(\underline{x}_k) \to f(\underline{x}^*)$ instead of $\|\underline{x}_k - \underline{x}^*\| \to 0$ when $k \to \infty$.

→ so actually we indifferent the choice of what to monitor

Proposition: If $H(\underline{x}^*)$ is p.d., $\underline{x}_k$ converges (super)linearly at $\underline{x}^*$ w.r.t. $|f(\underline{x}_k) - f(\underline{x}^*)|$ if and only if it converges in the same way w.r.t. $\|\underline{x}_k - \underline{x}^*\|$.

Indeed

$$f(\underline{x}) \approx f(\underline{x}^*) + \frac{1}{2}(\underline{x} - \underline{x}^*)^t H(\underline{x}^*)(\underline{x} - \underline{x}^*)$$

$(\ldots) = (\ldots)$

and $\exists$ a neighborhood $N(\underline{x}^*)$ such that

from those bounds we get the convergence relation of the term above

$$\lambda_1^{'} \|\underline{x} - \underline{x}^*\|^2 \le |f(\underline{x}) - f(\underline{x}^*)| \le \lambda_n^{'} \|\underline{x} - \underline{x}^*\|^2 \qquad \forall \underline{x} \in N(\underline{x}^*)$$

with $\lambda_1^{'} = \lambda_1 - \varepsilon > 0$ and $\lambda_n^{'} = \lambda_n + \varepsilon$, where $\varepsilon > 0$ and $0 < \lambda_1 \le \ldots \le \lambda_n$ are the eigenvalues of $H(\underline{x}^*)$. assumed to be positive def (which was a strong assumption)

N.B.: This equivalence does not hold in general (e.g., functions non everywhere $\mathcal{C}^1$)

Quadratic strictly convex functions:

$f(\underline{x}) = \frac{1}{2}\underline{x}^t Q \underline{x} - \underline{b}^t \underline{x}$ and weighted norm $\|\underline{x}\|_Q^2 := \underline{x}^t Q \underline{x}$.

Since $Q\underline{x}^* = \underline{b}$ (?) we have that
$\frac{1}{2}\|\underline{x} - \underline{x}^*\|_Q^2 = \frac{1}{2}[(\underline{x}-\underline{x}^*)^T Q (\underline{x}-\underline{x}^*)] = \ldots = f(\underline{x}) - f(\underline{x}^*)$

while in practice we rarely use the exact version of the method

Theorem: If gradient method with exact 1-D search is applied to any quadratic strictly convex $f \in \mathcal{C}^2$, for any $\underline{x}_0$ we have $\lim_{k\to\infty} \underline{x}_k = \underline{x}^*$ and

the method is globally convergent

$$\boxed{\|\underline{x}_{k+1} - \underline{x}^*\|_Q^2 \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 \|\underline{x}_k - \underline{x}^*\|_Q^2,} \qquad (1)$$

where $0 < \lambda_1 \leq \ldots \leq \lambda_n$ are the eigenvalues of Q.

convergence is linear (due to the square here and everywhere) and the rate it depends on Q

Proof sketch:

Zoutendijk's theorem implies global convergence.

Since exact 1-D search, easy to verify that

$$\|\underline{x}_{k+1} - \underline{x}^*\|_Q^2 = \left(1 - \frac{\underline{g}_k^t \underline{g}_k}{(\underline{g}_k^t Q \underline{g}_k)(\underline{g}_k^t Q^{-1} \underline{g}_k)}\right) \|\underline{x}_k - \underline{x}^*\|_Q^2,$$

where $\underline{g}_k = Q\underline{x}_k - \underline{b}$ $= \nabla f(\underline{x}_k)$

Then just apply Kantorovich inequality:

If Q p.d. (with λ_1 and λ_n smallest and largest eigenvalues), for each $\underline{x} \neq \underline{0}$ we have

$$\frac{(\underline{x}^t \underline{x})^2}{(\underline{x}^t Q \underline{x})(\underline{x}^t Q^{-1} \underline{x})} \geq \frac{4\lambda_n \lambda_1}{(\lambda_n + \lambda_1)^2}.$$

□

If $\lambda_1 = \lambda_n$ ($Q = \gamma I$), method "converges" in one iteration.

Upper bound (1) is reached for some choices of $\underline{x}_0$ (Aikake).

Linear convergence whose rate depends on *condition number* $\kappa = \frac{\lambda_n}{\lambda_1}$ of Q:

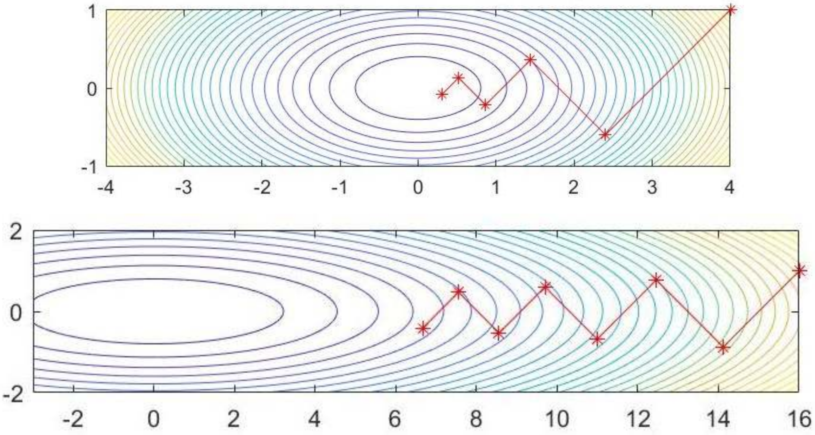
$$\boxed{r = \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right) = \left(\frac{\kappa - 1}{\kappa + 1}\right)}$$

the closer κ to 1 the smaller r; if the spektrum of Q is very wide then $\kappa \gg 1$ and $r \approx 1$.

which mos slow convergence

Example:

$$\min f(x_1, x_2) = \frac{1}{2}x_1^2 + \frac{a}{2}x_2^2 \quad \text{with } a \geq 1 \text{ and eigenvalues } \frac{1}{2} \text{ and } \frac{a}{2}$$



Some points of $\{\underline{x}_k\}$ for $a = 4$ (top) and $a = 16$ (bottom), starting from $\underline{x}_0 = \begin{pmatrix} a \\ 1 \end{pmatrix}$.

Arbitrary nonlinear functions:

Theorem: If $f \in \mathcal{C}^2$ and gradient method with exact 1-D search converges to $\underline{x}^*$ with $H(\underline{x}^*)$ p.d., then

we have a weaker result but involving the f(.) terms

$$\boxed{f(\underline{x}_{k+1}) - f(\underline{x}^*) \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 [f(\underline{x}_k) - f(\underline{x}^*)]}$$

where $0 < \lambda_1 \leq \ldots \leq \lambda_n$ are eigenvalues of $H(\underline{x}^*)$.

We cannot expect better convergence with inexact (approximate) 1-D search.

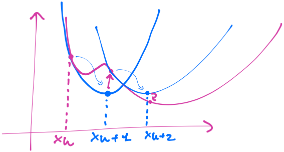
α_k minimizing $\phi(\alpha)$ may not be the best choice, we could try to "extract" 2nd order information about $f(\underline{x})$. which was assumed C^2 so we are good

Example: for $f(\underline{x})$ quadratic strictly convex, $\alpha_k = 1/\lambda_{k+1}$ lead to $\underline{x}^*$ in at most n iterations!

size of the space ($\mathbb{R}^n$) in which we are optimizing

4.6 Newton method

but this is of course a "lucky" case, but with H we could have performance issues



Let $f \in \mathcal{C}^2$ and $H(\underline{x}) = \nabla^2 f(\underline{x})$.

Consider quadratic approximation of $f(\underline{x})$ at $\underline{x}_k$:

$$q_k(\underline{x}) := f(\underline{x}_k) + \nabla^t f(\underline{x}_k)(\underline{x} - \underline{x}_k) + \frac{1}{2}(\underline{x} - \underline{x}_k)^t H(\underline{x}_k)(\underline{x} - \underline{x}_k)$$

and choose as $\underline{x}_{k+1}$ a stationary point of $q_k(\underline{x})$, namely

$$\nabla f(\underline{x}_k) + H(\underline{x}_k)(\underline{x}_{k+1} - \underline{x}_k) = \underline{0}.$$

If $H(\underline{x}_k)$ is not singular, $H^{-1}(\underline{x}_k)$ exists and

due to the hessian evaluation and inversion here

$$\boxed{\underline{x}_{k+1} := \underline{x}_k - \underbrace{H^{-1}(\underline{x}_k)\nabla f(\underline{x}_k)}_{d_k}.}$$

If $H(\underline{x}_k)$ is p.d., $f \in \mathcal{C}^2$ implies that $H^{-1}(\underline{x}_k)$ p.d. over $N(\underline{x}_k)$ and iteration is well defined, otherwise $\underline{d}_k$ may not be a descent direction.

(− expensive computation
− locally convergent
+ fast convergence)

the extreme opposite of the previous method

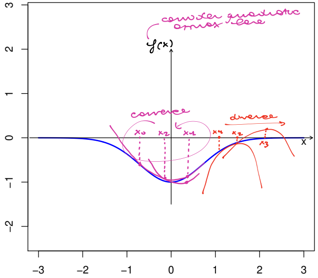
In the "pure" Newton method, $\alpha_k = 1$ for each k.

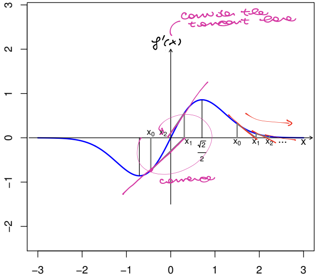
For f quadratic and strictly convex, global minimum in a single iteration.

Property: Newton method is invariant w.r.t. affine and non singular coordinate changes (see exercise set 6).

Observation: Newton method is <u>not globally convergent</u>, but very <u>fast local convergence</u> if $\underline{x}_0$ is sufficiently close to a desired solution.

<u>Example</u>: $\min_{x \in \mathbb{R}} f(x) = -\exp(-x^2)$ with global minimum $x^* = 0$ and $f'(x) = 2x\exp(-x^2)$





If $-0.2 \leq x_0 \leq 0.2$, $\{x_k\}_{k \in \mathbb{N}}$ converges at $x^* = 0$. If $x_0 > 1$, $\{x_k\}_{k \in \mathbb{N}}$ diverges.

Alternative interpretation of Newton method (1-D case):

$f(x) \in \mathcal{C}^2$ and look for x^* such that $f'(x) = 0$.

Method of tangents (Newton-Raphson) to determine the zeros of a 1-D function:

At iteration k, $f'(x)$ is approximated with the tangent at x_k

$$z = f'(x_k) + f''(x_k)(x - x_k)$$

x_{k+1} corresponds to the intersection with the x-axis: $x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$

n-D case: Determine a stationary point of $f(\underline{x})$ by solving non linear system $\nabla f(\underline{x}) = \underline{0}$ with "Newton-Raphson" method.

Theorem: (proof see Nocedal and Wright, 1999 edition, p. 52-53)

stationary point — strict local min

Suppose $f \in \mathcal{C}^2$ and $\underline{x}^*$ such that $\nabla f(\underline{x}^*) = \underline{0}$ and $H(\underline{x}^*)$ p.d. and $\exists\, L > 0$ such that

$$\|H(\underline{x}) - H(\underline{y})\| \leq L\|\underline{x} - \underline{y}\| \qquad \forall \underline{x}, \underline{y} \in N(\underline{x}^*)$$

(Lipschitz) requirement of the Hessian

then, for $\underline{x}_0$ sufficiently close to local minimum $\underline{x}^*$,

i) $\{\underline{x}_k\} \to \underline{x}^*$ with a quadratic convergence order,

ii) $\{\|\nabla f(\underline{x}_k)\|\} \to 0$ quadratically when $k \to \infty$.

very very fast convergence, but x_0 may be required to be also a lot close to x^*

Disadvantages:

⇒ idea: find a trade off method

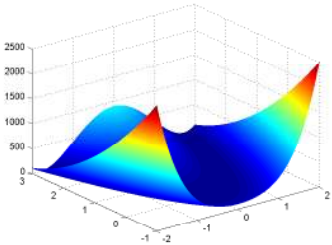
- If $H(\underline{x}_k)$ is singular the step is not defined.
- If $H^{-1}(\underline{x}_k)$ is not p.d., Newton direction may not be descent direction.
- "taking the pure newton step" Even for a descent direction $\alpha_k = 1$ may increase the value of f.
- Computation of $H^{-1}(\underline{x}_k)$ at each iteration ($O(n^3)$ complexity).
- Only *locally convergent*: if $\underline{x}_0$ is not close enough to $\underline{x}^*$, $\{\underline{x}_k\}_{k\geq 0}$ may not converge.
- Since $\{\underline{x}_k\}_{k\geq 0}$ converges from any $\underline{x}_0$ sufficiently close to any stationary point with non singular $\nabla^2 f(\underline{x})$, it may converge to local maxima.

For a comparison between gradient and Newton methods, see Nocedal and Wright, Numerical Optimization, Edition 1999, p. 199.

Rosenbrock function

$$f(\underline{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2.$$

quadratic and nonconvex.



Fourth computer laboratory: explore the considerable difference in convergence speed between various line search methods.

Modifications and extensions

1) If $\alpha_k = 1$ does not satisfy Wolfe (alternative) conditions then inexact 1-D search.

on update only, useful for local convergence

2) To guarantee global convergence

(gradient) I or (newton) H^{-1}

$$\underline{d}_k = -D_k \nabla f(\underline{x}_k)$$

with $D_k \neq [\nabla^2 f(\underline{x}_k)]^{-1}$. If D_k is symmetric and p.d., $\underline{d}_k$ is a descent direction.

Trade-off between steepest descent and Newton directions:

$$\boxed{D_k := (\varepsilon_k I + \nabla^2 f(\underline{x}_k))^{-1}}$$

where $\varepsilon_k > 0$ are smallest values such that eigenvalues of $(\varepsilon_k I + \nabla^2 f(\underline{x}_k))$ are $\geq \delta > 0$. Such ε_k making D_k p.d. always exist.

but we still need to invert the matrix

Coincides with "pure" Newton method when getting closer to a local minimum.

3) Trust region methods

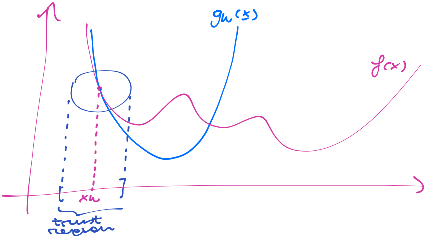
another idea to elaborate the newton method

Idea: simultaneously determine $\underline{d}_k$ and α_k by minimizing local quadratic approximation $q_k(\underline{x})$ at $\underline{x}_k$ over a *trust region* on which $q_k(\underline{x})$ provides a good approximation of $f(\underline{x})$.

→ becomes constrained opt, but ... the constraints

Example: $B_k = \{\underline{x} \in \mathbb{R}^n \, : \, \|\underline{x} - \underline{x}_k\| \leq \Delta_k\}$

Illustration:



min $q_k(x)$
st $x \in B_k$

In general, *trust region subproblem* $\min_{\underline{x} \in B_k} q_k(\underline{x})$ can be solved in closed form or it has low computational requirements.

The trust region size (e.g.Δ_k) is updated adaptively during the iterations based on an estimate of the quality (e.g. $\max |f(\underline{x}) - q_k(\underline{x})|$) of the quadratic approximation over it.

$q_k(\cdot)$ and $f(\cdot)$ similar $\Rightarrow$ larger Δ_k
-- -- different $\Rightarrow$ smaller Δ_k

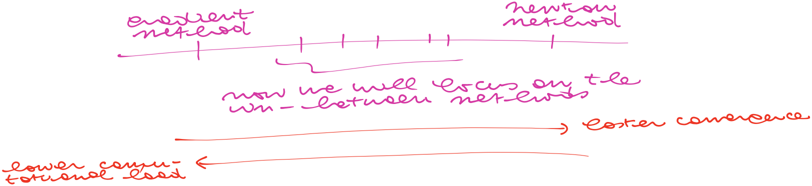
3) Trust region methods

Idea: simultaneously determine $\underline{d}_k$ and α_k by minimizing local quadratic approximation $q_k(\underline{x})$ at $\underline{x}_k$ over a *trust region* on which $q_k(\underline{x})$ provides a good approximation of $f(\underline{x})$.

Example: $B_k = \{\underline{x} \in \mathbb{R}^n : \|\underline{x} - \underline{x}_k\| \leq \Delta_k\}$

Illustration:

anyway, for now we know the two extremes of the methods



In general, *trust region subproblem* $\min_{\underline{x} \in B_k} q_k(\underline{x})$ can be solved in closed form or it has low computational requirements.

The trust region size (e.g.Δ_k) is updated adaptively during the iterations based on an estimate of the quality (e.g. $\max |f(\underline{x}) - q_k(\underline{x})|$) of the quadratic approximation over it.

4.7 Conjugate direction methods

Aim: faster convergence than gradient method and lower computational load than Newton method.

First consider quadratic strictly convex functions

seen since locally also non linear functions will be of that shape

$$\min_{\underline{x}\in\mathbb{R}^n} q(\underline{x}) = \frac{1}{2}\underline{x}^t Q\underline{x} - \underline{b}^t\underline{x}$$

with Q $n \times n$ symmetric and p.d.

Definition: Given $n \times n$ and symmetric Q, two nonzero $\underline{d}_1, \underline{d}_2 \in \mathbb{R}^n$ are Q-**conjugate** if $\boxed{\underline{d}_1^t Q \underline{d}_2 = 0.}$ extension of orthogonality

why are they useful?

Example:

$f(x_1, x_2) = 12x_2 + 4x_1^2 + 4x_2^2 - 4x_1x_2$

$\Rightarrow Q = \begin{pmatrix} 8 & -4 \\ -4 & 8 \end{pmatrix}$ if $\underline{d}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ then $\underline{d}_2 = \begin{pmatrix} a \\ b \end{pmatrix}$ must satisfy

$\underline{d}_1^T Q \underline{d}_2 = 8a - 4b = 0 \Rightarrow b = 2a$

and assume a we can get a conjugate directions

Proposition: If Q p.d. and nonzero $\underline{d}_0, \ldots, \underline{d}_k$ are mutually Q-conjugate, then $\underline{d}_0, \ldots, \underline{d}_k$ are linearly independent.

Proof: If $\exists\ \lambda_0, \ldots, \lambda_k$ st $\sum \underline{d}_i \lambda_i = \underline{0}$ then we can pre/post multiply either sides by $Q \underline{d}_i^T$ we get

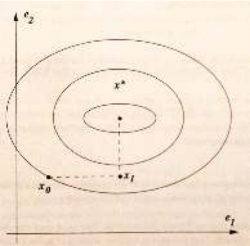
$$0 + \lambda_i \underline{d}_i (Q \underline{d}_i^T) = \underline{0} (Q \underline{d}_i^T) = 0$$

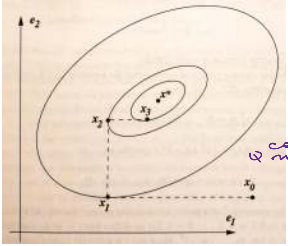
$$\Rightarrow \lambda_i (\underbrace{\underline{d}_i Q \underline{d}_i^T}_{>0 \text{ given } Q \text{ pos def}}) = 0 \iff \lambda_i = 0 \ \forall i$$

Geometric/algebraic interpretation

If Q is diagonal, $q(\underline{x})$ can be minimized via 1-D search along coordinate directions.

easy separable case





case of Q not diagonal

Nocedal and Wright, Numerical Optimization, Edition 1999, p. 104-105.

If Q is not diagonal and $\underline{d}_0, \ldots, \underline{d}_{n-1}$ are n mutually Q-conjugate, linear variable transformation

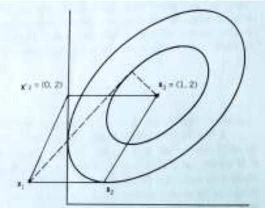
$$\underline{x} = \sum_{i=0}^{n-1} \alpha_i \underline{d}_i$$

leads to $\frac{1}{2}\underline{x}^T Q \underline{x} - \underline{b}^T \underline{x}$ write that $\underline{x}$

$$\tilde{q}(\underline{\alpha}) = \frac{1}{2}(\sum_{i=0}^{n-1} \alpha_i \underline{d}_i)^t Q (\sum_{i=0}^{n-1} \alpha_i \underline{d}_i) - \underline{b}^t (\sum_{i=0}^{n-1} \alpha_i \underline{d}_i) = \sum_{i=0}^{n-1} \left[\frac{1}{2} \alpha_i^2 \underline{d}_i^T Q \underline{d}_i - \alpha_i \underline{b}^T \underline{d}_i \right]$$

$$= \sum_{i=0}^{n-1} \tilde{q}_i(\alpha_i)$$

So the min of $q(\underline{x})$ over $\mathbb{R}^n$ reduces to at most n 1D search $\Rightarrow$ sum of n 1D quadratic terms



Bazaraa, Sherali, Shetty, Nonlinear Programming – Theory and Algorithms, third edition, Wiley Interscience, 2006, p. 316

Theorem: (Conjugate directions)

Let $\{\underline{d}_i\}_{i=0}^{n-1}$ be n nonzero mutually Q-conjugate directions.
For any $\underline{x}_0 \in \mathbb{R}^n$, $\{\underline{x}_k\}_{k\geq 0}$ generated according to the usual line search scheme

$$\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k \qquad (1)$$

but with the same Q conj directions

with

since q(·) is a quadratic form

$$\alpha_k = -\frac{\underline{g}_k^t \underline{d}_k}{\underline{d}_k^t Q \underline{d}_k} \qquad \text{and} \qquad \underline{g}_k := \nabla q(\underline{x}_k) = Q\underline{x}_k - \underline{b}$$

terminates to the (unique) global optimal solution $\underline{x}^*$ of $q(\underline{x})$ in at most n iterations, that is

$$\underline{x}_n = \underline{x}_0 + \sum_{k=0}^{n-1} \alpha_k \underline{d}_k = \underline{x}^*.$$

Proof.:

Since $\underline{d}_k$'s are linearly independent, $\exists\ \alpha_k$'s such that

$$\underline{x}^* - \underline{x}_0 = \alpha_0 \underline{d}_0 + \ldots + \alpha_{n-1} \underline{d}_{n-1}.$$

multiplying by $\underline{d}_k^T Q$ we get

$\underline{d}_k^T Q(\underline{x}^* - \underline{x}_0) = ($ all zeros except for the $\alpha_k \underline{d}_k$ term $)$

$$\Rightarrow \quad \alpha_k = \frac{\underline{d}_k^T Q(\underline{x}^* - \underline{x}_0)}{\underline{d}_k^T Q \underline{d}_k}$$

Now, following the iterative process (1) of the line search scheme we get

$$\underline{x}_k - \underline{x}_0 = \alpha_0 \underline{d}_0 + \ldots + \alpha_{k-1} \underline{d}_{k-1}$$

and hence the mutual Q-orthogonality of the $\underline{d}_k$'s implies that

$$\Rightarrow \quad \underline{d}_k^T Q(\underline{x}_k - \underline{x}_0) = 0$$

$\underline{d}_k^T Q(\ldots) = \underline{d}_k^T Q(\ldots)$

So now we can rewrite

$$\alpha_k = \frac{\underline{d}_k^T Q(\underline{x}^* - \underline{x}_k + \underbrace{\underline{x}_k - \underline{x}_0})}{\underline{d}_k^T Q \underline{d}_k} = \frac{\underline{d}_k^T Q(\underline{x}^* - \underline{x}_k)}{\underline{d}_k^T Q \underline{d}_k} =$$

$$= \frac{\underline{g}_k^T \underline{d}_k}{\underline{d}_k^T Q \underline{d}_k}$$

since $\underline{g}_k = \nabla f(\underline{x}_k) = Q\underline{x}_k - \underline{b}$ and the fact that $Q\underline{x}^* = \underline{b}$

Property: (Expanding subspace)

Let $\underline{d}_0, \ldots, \underline{d}_{n-1}$ be nonzero mutually Q-conjugate vectors. Then, for any $\underline{x}_0 \in \mathbb{R}^n$, $\{\underline{x}_k\}_{k \geq 0}$ generated according to

the sequence

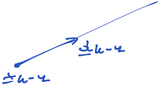
$$\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k \qquad \text{with} \qquad \alpha_k = -\frac{\underline{g}_k^t \underline{d}_k}{\underline{d}_k^t Q \underline{d}_k}$$

is such that

$$\underline{x}_k = \underline{x}_0 + \sum_{j=0}^{k-1} \alpha_j \underline{d}_j$$

all what we just derived before

$\boxed{\text{minimizes}}$ $q(\underline{x}) = \frac{1}{2}\underline{x}^t Q \underline{x} - \underline{b}^t \underline{x}$ not only on the line

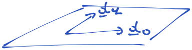


$$\{\ \underline{x} \in \mathbb{R}^n\ :\ \underline{x} = \underline{x}_{k-1} + \alpha \underline{d}_{k-1},\ \alpha \in \mathbb{R}\ \}$$

but also on the affine subspace $V_k = \{\ \underline{x} \in \mathbb{R}^n\ :\ \underline{x} = \underline{x}_0 + span\{\underline{d}_0, \ldots, \underline{d}_{k-1}\}\ \}$.

In particular, $\underline{x}_n$ is the global optimum of $q(\underline{x})$ on $\mathbb{R}^n$.

since all the $\underline{d}$'s are a basis of the entire $\mathbb{R}^n$



Proof:

Since $q(\underline{x})$ is strictly convex, $\underline{x}^*$ is a global min of $q(\underline{x})$ over $C \subseteq \mathbb{R}^n$ $\Longleftrightarrow$

moreover, it works since we optimize one direction at a time (assume the fact that the) are $\perp$)

Property: (Expanding subspace)

Let $\underline{d}_0, \ldots, \underline{d}_{n-1}$ be nonzero mutually Q-conjugate vectors. Then, for any $\underline{x}_0 \in \mathbb{R}^n$, $\{\underline{x}_k\}_{k \geq 0}$ generated according to

the sequence

$$\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k \qquad \text{with} \qquad \alpha_k = -\frac{\underline{g}_k^t \underline{d}_k}{\underline{d}_k^t Q \underline{d}_k}$$

is such that

$$\underline{x}_k = \underline{x}_0 + \sum_{j=0}^{k-1} \alpha_j \underline{d}_j$$

minimizes $q(\underline{x}) = \frac{1}{2}\underline{x}^t Q \underline{x} - \underline{b}^t \underline{x}$ not only on the line

$$\{\ \underline{x} \in \mathbb{R}^n\ :\ \underline{x} = \underline{x}_{k-1} + \alpha \underline{d}_{k-1},\ \alpha \in \mathbb{R}\ \}$$

but also on the affine subspace $V_k = \{\ \underline{x} \in \mathbb{R}^n\ :\ \underline{x} = \underline{x}_0 + span\{\underline{d}_0, \ldots, \underline{d}_{k-1}\}\ \}$.

In particular, $\underline{x}_n$ is the global optimum of $q(\underline{x})$ on $\mathbb{R}^n$.

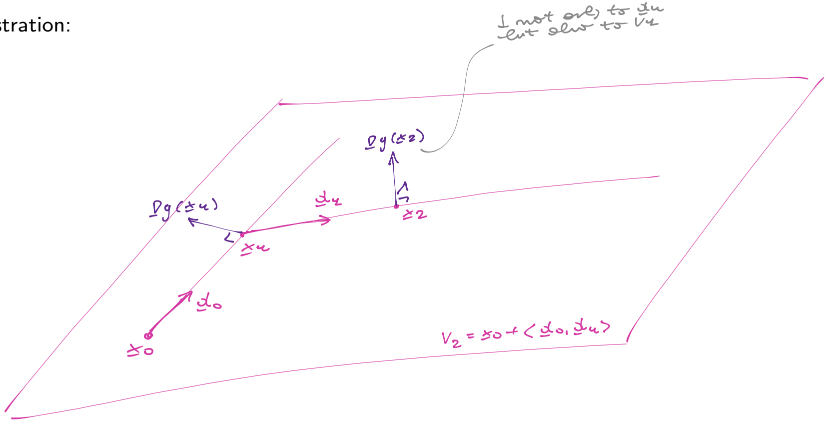
Proof:

Since $q(\underline{x})$ is strictly convex,

$\underline{x}^*$ is a global min of $q(\underline{x})$ over $C \subseteq \mathbb{R}^n$ $\iff$ $\nabla q(\underline{x}^*)(\underline{x}-\underline{x}^*) \geq 0$ for all $\underline{x} \in C$

So we verify that $\underline{\nabla} q(\underline{x}_k) = Q\underline{x}_k - \underline{b}$ is orthogonal ($\perp$) to V_k $\forall k = 1, \ldots, n$

Illustration:



Consequence: In conjugate direction method the gradients $\underline{g}_k$ satisfy $\underline{g}_k^t \underline{d}_i = 0$ for all i with $1 \leq i < k$.

4.7.1 Conjugate gradient method for quadratic convex functions

Initialization: Arbitrary $\underline{x}_0$, $\underline{g}_0 = \nabla q(\underline{x}_0) = Q\underline{x}_0 - \underline{b}$, $\boxed{\underline{d}_0 := -\underline{g}_0}$ and $k = 0$

Iteration: $\underline{x}_{k+1} := \underline{x}_k + \alpha_k \underline{d}_k$ with $\alpha_k = -\frac{\underline{g}_k^t \underline{d}_k}{\underline{d}_k^t Q \underline{d}_k}$ (exact 1-D search)

but there we still matrix Q here

$$\boxed{\underline{d}_{k+1} := \underbrace{-\underline{g}_{k+1}}_{\text{steepest descent dir}} + \beta_k \underline{d}_k} \quad \text{with} \quad \beta_k = \frac{\underline{g}_{k+1}^t Q \underline{d}_k}{\underline{d}_k^t Q \underline{d}_k}.$$

This solves the problem of how to derive the Q-conj directions

<u>Observations</u>:

- $\alpha_k = -\frac{\underline{g}_k^t \underline{d}_k}{\underline{d}_k Q \underline{d}_k}$ minimizes $q(\underline{x})$ along line through $\underline{x}_k$ generated by $\underline{d}_k$

$$\frac{dq(\underline{x}_k + \alpha \underline{d}_k)}{d\alpha} = \underline{d}_k^t Q(\underline{x}_k + \alpha \underline{d}_k) - \underline{b}^t \underline{d}_k \overset{!}{=} 0$$

$\Rightarrow$ (...) we get really that definition of α_k

- Limited computational requirements, no matrix inversions are needed.

To show that global optimal solution is found after at most n iterations, just verify that directions are mutually Q-conjugated.

it's just algebra

Proposition:

At each iteration k in which the optimum solution of $q(\underline{x})$ has not yet been found ($\underline{g}_i \neq \underline{0}$ for $i = 0, \ldots, k$)

i) $\underline{d}_0, \ldots, \underline{d}_{k+1}$ generated are mutually Q-conjugate

ii) $\alpha_k = \frac{\underline{g}_k^t \underline{g}_k}{\underline{d}_k^t Q \underline{d}_k} \neq 0$ (not really relevant, it is just needed for the next iii)

iii) $\beta_k = \frac{\underline{g}_{k+1}^t (\underline{g}_{k+1} - \underline{g}_k)}{\underline{g}_k^t \underline{g}_k} = \underbrace{\frac{\underline{g}_{k+1}^t \underline{g}_{k+1}}{\underline{g}_k^t \underline{g}_k}}_{= \frac{\|g_{k+1}\|^2}{\|g_k\|^2}}$ } we then form is the rewriting for quadratic functions, since it does not need the matrix Q (which is constant only for strictly convex functions)

Advantages: No need for matrix inversions, limited computational requirements.

Disadvantages:

- Exact or at least accurate 1-D search otherwise the directions may loose Q-conjugacy.

- The method is not invariant w.r.t. affine transformations of the coordinates. which is a really bad news

Fourth computer laboratory: compare the convergence speed of gradient, conjugate gradient and Newton methods.

vs conjugate gradient (of below)

4.7.2 Conjugate direction methods

the real method to use in practice for large scale (huge n) problems

For arbitrary functions with large n, approximate α_k and β_k must not depend on Hessian.

Arbitrary $\underline{x}_0$ and $\underline{d}_0 = -\nabla f(\underline{x}_0)$

scheme for the Q-conj dir generation

$$\underline{x}_{k+1} := \underline{x}_k + \alpha_k \underline{d}_k \quad \text{with inexact 1-D search and} \quad \underline{d}_{k+1} = -\nabla f(\underline{x}_{k+1}) + \beta_k \underline{d}_k.$$

Most popular formulae for β_k:

RHS equivalent expr of the previous rule

$$\beta_k^{FR} = \frac{\|\nabla f(\underline{x}_{k+1})\|^2}{\|\nabla f(\underline{x}_k)\|^2} \qquad \text{Fletcher-Reeves}$$

LHS expr of the previous rule

$$\beta_k^{PR} = \frac{\nabla^t f(\underline{x}_{k+1})(\nabla f(\underline{x}_{k+1}) - \nabla f(\underline{x}_k))}{\|\nabla f(\underline{x}_k)\|^2} \qquad \text{Polak-Ribière}$$

Observation: $\underline{d}_k$ is a descent direction if exact 1-D search

$$\nabla^t f(\underline{x}_k)\underline{d}_k = -\|\nabla^t f(\underline{x}_k)\|^2 + \beta_{k-1}\cancel{\nabla^t f(\underline{x}_k)\underline{d}_{k-1}} = -\|\nabla^t f(\underline{x}_k)\|^2 < 0.$$

0 from ⊥ (orthog)

For quadratic functions the method coincides with $\boxed{\text{CG}}$ method.
while we denote as $\boxed{\text{CD}}$ this extension to arbitrary functions

For nonquadratic functions, Polak-Ribière version turns out to be more efficient than Fletcher-Reeves one.

Observations

→ computational load is very low

- At each iteration it suffices to store $\underline{x}_k$, $\nabla f(\underline{x}_k)$, $\nabla f(\underline{x}_{k+1})$ and $\underline{d}_k$.

  where every m-th time we take a steepest descent step

- Version with "restart" in which $\beta_k = 0$ every m iterations ($m << n$) is globally convergent.

  When $\beta_k = 0$, $d_{k+1} = -\nabla f(\underline{x}_{k+1})$ and all previous information is lost.

  For large n, we hope to find a solution way before n iterations!

  → this restart version is just a theoretical comment, useless in practice

4.7.3 Convergence

Convergence for quadratic functions

Let $q(\underline{x}) = \frac{1}{2}\underline{x}^t Q \underline{x} - \underline{b}^t \underline{x}$ be quadratic strictly convex with $\lambda_1 \leq \ldots \leq \lambda_n$ the eigenvalues of Q, then

$$\boxed{\|\underline{x}_{k+1} - \underline{x}^*\|_Q^2 \leq \left(\frac{\boxed{\lambda_{n-k}} - \lambda_1}{\boxed{\lambda_{n-k}} + \lambda_1}\right)^2 \|\underline{x}_0 - \underline{x}^*\|_Q^2}$$

- smaller Q norm
- smaller λ values

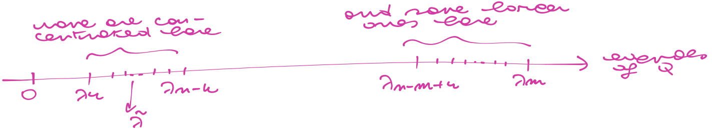
BUT

- err_{k+1} vs err_0
- neglect some of the largest eigenvalues

where $\|\underline{x} - \underline{x}^*\|_Q^2 = (\underline{x}^t - \underline{x}^*)^t Q(\underline{x} - \underline{x}^*) = 2(q(\underline{x}) - q(\underline{x}^*))$.

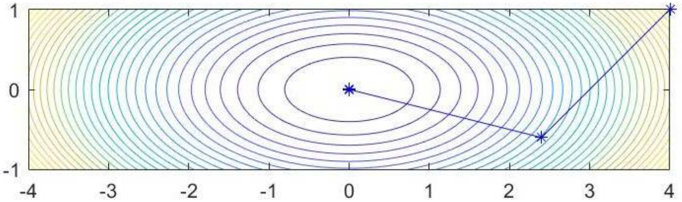
If m large eigenvalues and other $n - m$ "concentrated" around a $\tilde{\lambda}$, after $m+1$ iterations $\|\underline{x}_{m+1} - \underline{x}^*\|_Q \approx \varepsilon \|\underline{x}_0 - \underline{x}^*\|_Q$ with $\varepsilon = (\lambda_{n-m} - \lambda_1)/2\tilde{\lambda}$, that is, we have an accurate estimate of the solution after $m+1$ iterations.

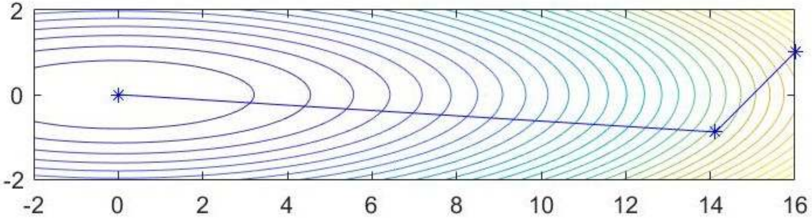
Illustration:



<u>Example</u>:

$$\min f(x_1, x_2) = \frac{1}{2}x_1^2 + \frac{a}{2}x_2^2 \quad \text{with } a \geq 1 \text{ and hence eigenvalues } \frac{1}{2} \text{ and } \frac{a}{2}$$





The sequence $\{\underline{x}_k\}$ for $a = 4$ (top) and $a = 16$ (bottom), starting from $\underline{x}_0 = \begin{pmatrix} a \\ 1 \end{pmatrix}$.

Convergence for arbitrary functions

1) If $f \in \mathcal{C}^2$ and $\{\underline{x}_k\}_{k\geq 0}$ generated by the F-R method with exact 1-D search converges to $\underline{x}^*$ with p.d. $H(\underline{x}^*)$, then

i.e. x^* is a strict local min

$$\boxed{\lim_{k\to\infty} \frac{\|\underline{x}_{k+n} - \underline{x}^*\|}{\|\underline{x}_k - \underline{x}^*\|} = 0,}$$

$\dfrac{\| \text{err}(k+n) \|}{\| \text{err}(k) \|}$

which reaches, $\to 0$ the ratio

namely convergence is superlinear within n iterations.

in n exact

Similar result also for inexact 1-D search.

2) Global convergence of F-R method even without "restart" (for P-R?).

Zoutendijk's theorem implies:

For F-R method with inexact 1-D search satisfying strong Wolfe conditions with $0 < c_1 < c_2 < 1/2$, we have

$$\boxed{\lim_{k\to\infty} \inf \|\nabla f(\underline{x}_k)\| = 0.}$$

A sub-sequence has $\|\nabla f(\underline{x}_k)\|$ that converges to 0.

due to the inf presence

to solve the issue of not being CG invariant to affine transformations

4.7.4 Preconditioned conjugate gradient method

The conjugate gradient method (CG) can be accelerated by a variable change $\boxed{\underline{x} = S\underline{y},}$ where S is $n \times n$ symmetric and non singular.

By applying CG to

issue: find S st SQS becomes well conditioned

$$h(\underline{y}) = q(S\underline{y}) = \frac{1}{2}\underline{y}^t SQS\underline{y} - \underline{b}^t S\underline{y}$$

we obtain

$$\underline{y}_{k+1} = \underline{y}_k + \alpha_k \underline{\tilde{d}}_k$$

directions in the y space

with α_k determined by 1-D search, $\underline{\tilde{d}}_0 = -\nabla h(\underline{y}_0)$ and $\underline{\tilde{d}}_k = -\nabla h(\underline{y}_k) + \beta_{k-1}\underline{\tilde{d}}_{k-1}$ for $k = 1, \ldots, n-1$ where

$$\beta_{k-1} = \frac{\nabla^t h(\underline{y}_k)\nabla h(\underline{y}_k)}{\nabla^t h(\underline{y}_{k-1})\nabla h(\underline{y}_{k-1})}.$$

Setting $\underline{x}_k = S\underline{y}_k$, $\nabla h(\underline{y}_k) = S\underline{g}_k$, $\underline{d}_k = S\underline{\tilde{d}}_k$, we obtain the equivalent **preconditioned conjugate gradient method**:

$$\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k$$

with α_k determined by 1-D search, $\underline{d}_0 = -S\underline{g}_0$ and

$$\underline{d}_k = -S\underline{g}_k + \beta_{k-1}\underline{d}_{k-1} \qquad \text{for } k = 1, \ldots, n-1$$

where

$$\beta_{k-1} = \frac{\underline{g}_k^t S^2 \underline{g}_k}{\underline{g}_{k-1} S^2 \underline{g}_{k-1}}.$$

Clearly when $S = I$ it coincides with the standard CG method.

Since $\nabla^2 h(\underline{y}) = SQS$, $\tilde{\underline{d}}_0, \ldots, \tilde{\underline{d}}_{n-1}$ are (SQS)-conjugate. Moreover $\underline{d}_k = S\tilde{\underline{d}}_k$ implies that $\underline{d}_0, \ldots, \underline{d}_{n-1}$ are Q-conjugate.

To achieve faster convergence, we look for S such that SQS has a smaller condition number than Q or eigenvalues that are distributed into "groups".

Recall: a good approximate solution can be found in a number of iterations not much larger than the number of groups.

4.8 Quasi-Newton methods

Instead of using/inverting $\nabla^2 f(\underline{x}_k)$, second order derivative information is extracted from variations in $\nabla f(\underline{x})$.

easier and interpretable

Generate $\{H_k\}$ of symmetric p.d. approximations of $[\nabla^2 f(\underline{x}_k)]^{-1}$ and take

$$\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k \quad \text{with} \quad \underline{d}_k = -H_k \nabla f(\underline{x}_k),$$

where $\alpha_k > 0$ minimizes $f(\underline{x})$ along $\underline{d}_k$ or satisfies some inexact 1-D search conditions.

Advantages w.r.t. Newton method:

- since H_k's are symmetric and p.d., always well defined and descent direction,
- only involves first order derivatives,
- H_k is constructed iteratively, each iteration is $O(n^2)$. which was good wrt $O(n^3)$ of pure-newton method (for matrix inversion there)

Disadvantages w.r.t. conjugate direction methods: requires storing/handling matrices.

rather than the few vectors of CG

Idea: Second order derivative information is extracted from $\nabla f(\underline{x}_k)$ and $\nabla f(\underline{x}_{k+1})$.

Quadratic approximation of $f(\underline{x})$ around $\underline{x}_k$:

$$f(\underline{x}_k + \underline{\delta}) \approx f(\underline{x}_k) + \underline{\delta}^t \nabla f(\underline{x}_k) + \frac{1}{2}\underline{\delta}^t \nabla^2 f(\underline{x}_k)\underline{\delta}.$$

Differentiating (wrt $\underline{\delta}$) we obtain

$$\nabla f(\underline{x}_k + \underline{\delta}) \approx \nabla f(\underline{x}_k) + \nabla^2 f(\underline{x}_k)\underline{\delta}.$$

Substituting $\underline{\delta}$ with $\underline{\delta}_k := \underline{x}_{k+1} - \underline{x}_k$ we can move the gradients to the LHS one, and so we also define $\underline{\gamma}_k = \nabla f(\underline{x}_{k+1}) - \nabla f(\underline{x}_k)$.

So we have

$$\underline{\gamma}_k \approx \nabla^2 f(\underline{x}_k)\,\underline{\delta}_k \iff [\nabla^2 f(\underline{x}_k)]^{-1}\underline{\gamma}_k \approx \underline{\delta}_k$$

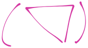
and this was the role we was assuming for H_{k+1} matrix (H_{k+1} since we need $k+1$ iters to compute $\underline{\delta}_k$ and $\underline{\gamma}_k$)

Since $\underline{\delta}_k$ and $\underline{\gamma}_k$ can only be determined after 1-D search, we select H_{k+1} symmetric and p.d. such that

$$\boxed{H_{k+1}\underline{\gamma}_k = \underline{\delta}_k} \qquad \textbf{(secant condition)}. \tag{1}$$

H_{k+1} is not univocally defined: n equations and $n(n+1)/2$ (variables) degrees of freedom.

symmetric part we solve for



how do we update H_k?

Simple way is by successive updates:

a generic expression of a symm rank 1 matrix

$$H_{k+1} = H_k + a_k \underbrace{\underline{u}\,\underline{u}^t}_{} \tag{2}$$

where $\underline{u}\,\underline{u}^t$ symmetric matrix of rank 1 and a_k proportionality coefficient.

To satisfy (1) we must have

$$H_k \underline{\gamma}_k + a_k \underline{u}\,\underline{u}^t \underline{\gamma}_k = \underline{\delta}_k$$

$\underbrace{a_k u}_{\text{scalar}} \underbrace{(u^T \gamma_k)}_{\text{scalar}} = \delta_k - H\gamma_k$

$\Rightarrow u \,//\, \delta_k - H\gamma_k$

and hence $\underline{u}$ and $(\underline{\delta}_k - H_k \underline{\gamma}_k)$ must be collinear.

Since a_k accounts for proportionality, we can set $\underline{u} = \underline{\delta}_k - H_k \underline{\gamma}_k$ and hence $a_k \underline{u}^t \underline{\gamma}_k = 1$.

$\Rightarrow a_k = \frac{1}{u \cdot \gamma_k}$

Rank one update formula:

$$\boxed{H_{k+1} = H_k + \frac{\overbrace{(\underline{\delta}_k - H_k\underline{\gamma}_k)}^{u}\overbrace{(\underline{\delta}_k - H_k\underline{\gamma}_k)^t}^{u^T}}{(\underline{\delta}_k - H_k\underline{\gamma}_k)^t \underline{\gamma}_k}} \tag{3}$$

$\longrightarrow \frac{1}{\text{scalar}} = a_k$

Properties

1. For quadratic strictly convex functions, $H_n = Q^{-1}$ in at most n iterations, even with inexact 1-D search.
   eventually H_k becomes Q^{-1}
2. No guarantee that H_k is p.d.!
   really bad, so we now improve it

Rank two updates

$$H_{k+1} = H_k + a_k \underline{u}\,\underline{u}^t + b_k \underline{v}\,\underline{v}^t \qquad (4)$$

are more interesting.

$$(\,) + a_k \underbrace{(\,)}_{\doteq\, \delta_k} uv + b_k \underbrace{(\,)}_{\doteq\, H_k \gamma_k} vv = (\,)$$

To satisfy (1) we have

$$H_k \underline{\gamma}_k + a_k \underline{u}\,\underline{u}^t \underline{\gamma}_k + b_k \underline{v}\,\underline{v}^t \underline{\gamma}_k = \underline{\delta}_k$$

view the condition as a vector formula with free choices

where $\underline{u}, \underline{v}$ are not determined univocally.

Setting $\boxed{\underline{u} = \underline{\delta}_k}$ and $\boxed{\underline{v} = H_k \underline{\gamma}_k}$, we obtain $a_k \underline{u}^t \gamma_k = 1$ and $b_k \underline{v}^t \gamma_k = -1$ and hence the rank two update formula:

$$\boxed{H_{k+1} = H_k + \frac{\underline{\delta}_k \underline{\delta}_k^t}{\underline{\delta}_k^t \underline{\gamma}_k} - \frac{H_k \underline{\gamma}_k \underline{\gamma}_k^t H_k}{\underline{\gamma}_k^t H_k \underline{\gamma}_k}} \qquad \text{Davidon-Fletcher-Powell (DFP)} \qquad (5)$$

which has nice
and properties

as it involves γ_k we use the variation of the gradient

Proposition: If $\boxed{\underline{\delta}_k^t \underline{\gamma}_k > 0 \quad \forall k}$ (curvature condition),

the DFP method preserves the positive definiteness of H_k, i.e., if H_0 is p.d. then H_k is p.d. for all $k \geq 1$.

Proof: By induction, suppose that H_k is pd (we need to show then) then $\underline{z}^T H_{k+1} \underline{z} > 0 \quad \forall \underline{z} \neq \underline{0}$

If H_k is pd it admits a Cholesky factorization
$$H_k = L_k L_k^T$$

eliminating the k subscript (for clearness) and by setting
$$\underline{a} = L^T \underline{z} \quad \text{and} \quad \underline{b} = L^T \underline{\gamma}$$

we get (rewriting $\underline{z}$ and the H_{k+1} update rule):
$$\underline{z}^T \left(H - \frac{H \underline{\gamma} \underline{\gamma}^T H}{\underline{\gamma}^T H \underline{\gamma}} \right) \underline{z} = \underline{a}^T \underline{a} - \frac{(\underline{a}^T \underline{b})^2}{\underline{b}^T \underline{b}} \geq 0 \quad (\#)$$

neglecting the other term (B)

because of Cauchy-Schwarz $|\underline{a}^T \underline{b}| \leq \|\underline{a}\| \cdot \|\underline{b}\|$

Since $\underline{z} \neq \underline{0}$, the equality holds iff $\underline{a}$ and $\underline{b}$ are collinear, i.e. iff $\underline{z}$ and $\underline{\gamma}$ are.

Since $\delta^T \gamma > 0$ (curvature cond) we have that (rewriting the term (B)):
$$\underline{z} \left(\frac{\underline{\delta} \underline{\delta}^T}{\underline{\delta}^T \underline{\gamma}} \right) \underline{z} \geq 0$$
where equality holds iff $\underline{\delta}$ and $\underline{\gamma}$ are collinear

and so here (#) we get the sum of two ≥ 0 terms

Fact: The curvature condition $\underline{\delta}_k^t \underline{\gamma}_k > 0$ holds for every $k \geq 0$ provided that the 1-D search satisfies (weak or strong) Wolfe conditions.

$$\left\{ \begin{array}{l} g(\underline{x}) = \underline{x}^T Q \underline{x} - \underline{b}^T \underline{x} \\ \nabla g(\underline{x}) = 2Q\underline{x} - \underline{b}^T \end{array} \right\}$$

$$= \nabla g(\underline{x}_k + u) - \nabla g(\underline{x}_k) =$$

$$= (2Q\underline{x} - \underline{b}) \Big|_{\underline{x}_k}^{\underline{x}_k+u} = 2Q(\underline{x}_k + u - \underline{x}_k) = 2Q\underline{\delta}_k$$

Proof*:

For quadratic strictly convex functions, $\underline{\gamma}_k = Q\underline{\delta}_k$ implies $\underline{\delta}_k^t Q \underline{\delta}_k = \underline{\delta}_k^t \underline{\gamma}_k > 0$ because Q is p.d.

For arbitrary functions:

Weak Wolfe conditions:

$$f(\underline{x}_k + \alpha_k \underline{d}_k) \leq f(\underline{x}_k) + c_1 \alpha_k \nabla^t f(\underline{x}_k) \underline{d}_k \quad \text{(Armijo criterion)} \tag{6}$$

$$\nabla^t f(\underline{x}_k + \alpha_k \underline{d}_k) \underline{d}_k \geq c_2 \nabla^t f(\underline{x}_k) \underline{d}_k \tag{7}$$

with $0 < c_1 < c_2 < 1$.

Since $\underline{\delta}_k = \alpha_k \underline{d}_k$, (7) implies

$$\nabla^t f(\underline{x}_{k+1}) \underline{\delta}_k \geq c_2 \nabla^t f(\underline{x}_k) \underline{\delta}_k,$$

which in turn implies

$$\underline{\gamma}_k^t \underline{\delta}_k \geq (c_2 - 1) \alpha_k \nabla^t f(\underline{x}_k) \underline{d}_k$$

with $(c_2 - 1) < 0$, $\alpha_k > 0$, and $\nabla^t f(\underline{x}_k) \underline{d}_k < 0$ because $\underline{d}_k$ is a descent direction. $\square$

Properties

For quadratic strictly convex functions, DFP method with exact 1-D search:

1. terminates in at most n iterations with $H_n = Q^{-1}$,
2. generates Q–conjugate directions (from $H_0 = I$ it generates CG directions),
3. secant condition is hereditary, i.e., $H_i \underline{\gamma}_j = \underline{\delta}_j$ for $j = 0, \ldots, i-1$.

   we not only $H_{k+1} \underline{\gamma}_k = \underline{\delta}_k$, but this was preserved also for past $\underline{\gamma}_k$ and $\underline{\delta}_k$ vectors

For arbitrary functions:

4. if $\underline{\delta}_k^t \underline{\gamma}_k > 0$ (curvature condition), all H_k are p.d. if H_0 is p.d. (hence descent method),
5. each iteration is $O(n^2)$,

   much faster than linear

6. superlinear convergence rate (in general only local),
7. if $f(\underline{x})$ convex, DFP method with exact 1-D search is globally convergent.

BFGS method

so that the idea was to make the approximation matrices easily invertible (analytically)

the hessian ... its inverse

We can construct an approximation of $\nabla^2 f(\underline{x}_k)$ rather than of $[\nabla^2 f(\underline{x}_k)]^{-1}$.

Since we aim at $B_k \approx \nabla^2 f(\underline{x}_k)$, B_k must satisfy $\boxed{B_{k+1}\underline{\delta}_k = \underline{\gamma}_k.}$

Taking $B_{k+1} = B_k + a_k \underline{u}\,\underline{u}^t + b_k \underline{v}\,\underline{v}^t$, with similar manipulations, we have:

again using a rank 2 update

$$B_{k+1} = B_k + \frac{\underline{\gamma}_k \underline{\gamma}_k^t}{\underline{\gamma}_k^t \underline{\delta}_k} - \frac{B_k \underline{\delta}_k \underline{\delta}_k^t B_k}{\underline{\delta}_k^t B_k \underline{\delta}_k} \tag{8}$$

same formula of before with γ and δ exchanged

which should be inverted to obtain $H_{k+1} = B_{k+1}^{-1}$

By applying twice *Sherman–Morrison indentity*

$$(A + \underline{a}\,\underline{b}^t)^{-1} = A^{-1} - \frac{A^{-1}\underline{a}\,\underline{b}^t A^{-1}}{1 + \underline{b}^t A^{-1}\underline{a}}, \quad A \in \mathbb{R}^{n \times n} \text{ non singular, } \underline{a}, \underline{b} \in \mathbb{R}^n, \text{ denominator} \neq 0,$$

we obtain the Broyden Fletcher Goldfarb and Shanno (BFGS) update formula:

$$H_{k+1} = H_k + \left(1 + \frac{\underline{\gamma}_k^t H_k \underline{\gamma}_k}{\underline{\delta}_k^t \underline{\gamma}_k}\right) \frac{\underline{\delta}_k \underline{\delta}_k^t}{\underline{\delta}_k^t \underline{\gamma}_k} - \frac{H_k \underline{\gamma}_k \underline{\delta}_k^t + \underline{\delta}_k \gamma_k^t H_k}{\underline{\delta}_k^t \underline{\gamma}_k} \tag{9}$$

Indeed $B_{k+1} H_{k+1} = I$ if $B_k H_k = I$.

The BFGS method has same properties 1 to 5 as DFP method.

In pratice, it is more robust w.r.t. to rounding errors and inexact 1-D search.

→ BFGS actually was better than DFP

BFGS and DFP are two extreme cases of unique Broyden family of update formulae:

$$H_{k+1} = (1-\phi)H_{k+1}^{\text{DFP}} + \phi H_{k+1}^{\text{BFGS}}$$

convex comb of the DFP and BFGS matrices

with $0 \leq \phi \leq 1$.

Properties: (Broyden family)

- H_{k+1} satisfies secant condition and is p.d. if $\underline{\delta}_k^t \underline{\gamma}_k > 0$.
- Methods invariant w.r.t. affine variable transformations.

  ⇒ we get the speed of CG and the good properties of the newton method
- If $f(\underline{x})$ quadratic strictly convex, methods with exact 1-D search find $\underline{x}^*$ in at most n iterations ($H_n = Q^{-1}$) and the generated directions are Q-conjugate.
- Quasi-Newton methods are much less "sensitive" to inexact 1-D search than CD ones.

  and since this true for arbitrary functions is a very good property

Convergence of quasi-Newton methods

Complex analysis because approximation of Hessian (inverse) is updated at each iteration.

Convergence speed for $\{B_k\}$ or $\{H_k\}$ with inexact 1-D search (Wolfe cond.) where $\alpha_k = 1$ is tried first:

Theorem: (Dennis and Moré)

this is in model, guaranteed always (we take x_0)

Consider $f \in \mathcal{C}^3$ and quasi-Newton method with B_k p.d. and $\alpha_k = 1$ for each k.
If $\lim_{k\to\infty} \underline{x}_k = \underline{x}^*$ with $\nabla f(\underline{x}^*) = \underline{0}$ and $\nabla^2 f(\underline{x}^*)$ is p.d., $\{\underline{x}_k\}$ converges superlinearly if and only if

stationary point

local opt

$$\lim_{k\to\infty} \frac{\|(B_k - \nabla^2 f(\underline{x}^*))\underline{d}_k\|}{\|\underline{d}_k\|} = 0. \tag{10}$$

If quasi-Newton $\underline{d}_k$ approximates Newton direction well enough, $\overset{(x_0)}{\alpha_k} = 1$ satisfies Wolfe cond. when $\underline{x}_k \to \underline{x}^*$.

we don't ask $B_k \to \nabla^2 f(\underline{x}^*)$ but just that the B_k accuracy (in approximating the Hessian) increases along directions $\underline{d}_k$

Observation: No need that $B_k \to \nabla^2 f(\underline{x}^*)$, it suffices that B_k's become increasingly accurate approximations of $\nabla^2 f(\underline{x}^*)$ along $\underline{d}_k$!

The necessary and sufficient condition (10) is satisfied by quasi-Newton methods such as BFGS and DFP.

Comparing the convergence rates of gradient, Newton and BFGS methods:

for Rosenbrock's function, see p. 199 (Chap. 8) of J. Nocedal, S. Wright, Numerical Optimization, Springer, 1999.

Global convergence:

Under some assumptions, can guarantee global convergence for arbitrary functions with inexact 1-D search.

In general "classical" globalization techniques (restart or trust region) are not adopted because no examples of non convergence are known.

Widely used: quasi-Newton methods with BFGS and DFP updates and 1-D search procedures satisfying Wolfe conditions.

Chapter 5: Constrained nonlinear optimization

Edoardo Amaldi

DEIB – Politecnico di Milano
edoardo.amaldi@polimi.it

Course material on WeBeep 2022-23 - Optimization

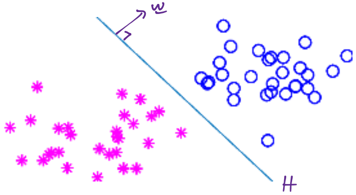
Academic year 2023-24

5.1 Example: Design linear classifiers and train SVMs

Support Vector Machines (SVMs) for binary classification.

Training set $T = \{ (\underline{x}^i, y^i) \, : \, \underline{x}^i \in \mathbb{R}^n, \, y^i \in \{-1, 1\}, \, i = 1, \ldots, p \}$.

Linear classifier: Suppose T is linearly separable

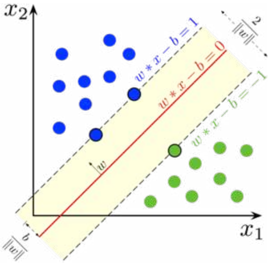


hyperplane $H(\underline{w}, b) = \{\underline{x} \in \mathbb{R}^n \, : \, \underline{w}^t \underline{x} = b\}$ separates the points of the two classes if

$\underline{w}^T \underline{x}_i - b \geq 1$ for $i : y_i = +1$
$\underline{w}^T \underline{x}_i - b \leq -1$ for $i : y_i = -1$

H not unique. the "best" one would be the one with the largest margin

If T is linearly separable, H with **largest margin** (min distance from H to any $\underline{x}^i$) is the most robust w.r.t. noise.



Since width $= \frac{2}{\|\underline{w}\|}$, hard-margin linear SVM training:

we assume there is complete separation

and we look for a linear separation

$$\max \text{ width} = \max \frac{2}{\|\underline{w}\|^2} = \min \frac{\|\underline{w}\|^2}{2}$$

$$\text{st} \quad y_i(\underline{w}^T \underline{x}_i - \underline{b}) - 1 \geq 0 \quad \forall i$$ (we all points are correctly classified)

$\underline{w} \in \mathbb{R}^n$
$b \in \mathbb{R}$

quadratic obj function
linear constraints

Remark: H with maximum margin is completely determined by the *support vectors* (closest $\underline{x}^i$'s to H).

so we also reduce the # of the constr, which for now were the we are constr for each point

Decision function: $h(\underline{w}, b, \underline{x}) = \underline{w}^t \underline{x} - b$.

Extensions:

1) Soft margin for nonlinearly separable T (not convex)

2) Nonlinear classifiers by applying kernels.

See Computer Lab 5.

For other applications see e.g. Chap. 6-8 of S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge Press, 2004.

5.2 Necessary optimality conditions

generic problem structure:

Consider

$$\boxed{\begin{array}{lll} \min & f(x) & \\ \text{s.t.} & g_i(\underline{x}) \leq \underline{0} & i \in I = \{1, ..., m\} \\ & \underline{x} \in \mathbb{R}^n & \end{array}} \qquad (1)$$

wlog since we can always bring us in this form

where $f, g_i \in C^1$.

Assumption: Feasible region $S = \{ \underline{x} \in \mathbb{R}^n : g_i(x) \leq 0, \forall i \in I \} \neq \emptyset$ but its interior can be empty.

Definitions: For each $\underline{x} \in S$ given we a feasible point

- $\mathscr{D}(\underline{x}) = \{\, \underline{d} \in \mathbb{R}^n \;:\; \exists\, \overline{\alpha} > 0 \text{ such that } \underline{x} + \alpha \underline{d} \in S,\ \forall \alpha \in [0, \overline{\alpha}] \,\}$

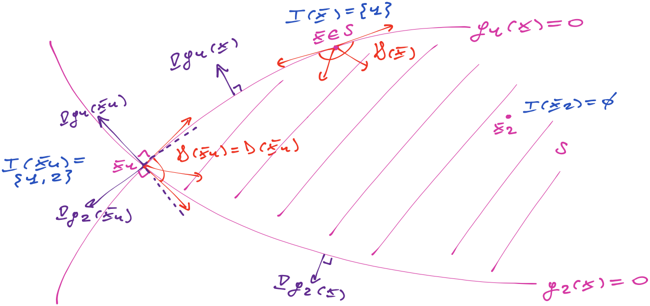
  cone of the **feasible directions**.

  i.e. the constr which are saturated at equality)

- $I(\underline{x}) = \{\, i \in I \;:\; g_i(\underline{x}) = 0 \,\} \subseteq I$ set of **indices** of the **active constraints**.

- $\mathrm{D}(\underline{x}) = \{\, \underline{d} \in \mathbb{R}^n \;:\; \nabla^t g_i(\underline{x}) \underline{d} \leq 0, \forall i \in I(\underline{x}) \,\}$

  cone of the **directions constrained by the gradients of the active constraints**.



Definitions: For each $\underline{x} \in S$ even on a feasible point

- $\mathscr{D}(\underline{x}) = \{\, \underline{d} \in \mathbb{R}^n \,:\, \exists\, \overline{\alpha} > 0 \text{ such that } \underline{x} + \alpha \underline{d} \in S,\ \forall \alpha \in [0, \overline{\alpha}] \,\}$

  cone of the **feasible directions**.

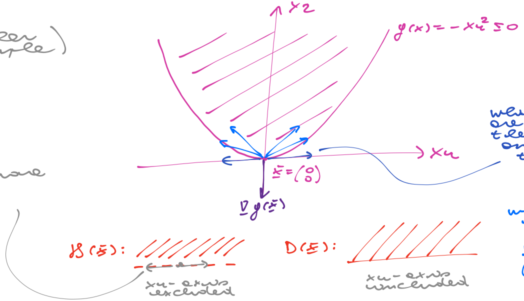
- $I(\underline{x}) = \{\, i \in I \,:\, g_i(\underline{x}) = 0 \,\} \subseteq I$ set of **indices** of the **active constraints**. (i.e. the constr. which are saturated at equality)

- $\mathrm{D}(\underline{x}) = \{\, \underline{d} \in \mathbb{R}^n \,:\, \nabla^t g_i(\underline{x})\underline{d} \le 0, \forall i \in I(\underline{x}) \,\}$

  cone of the **directions constrained by the gradients of the active constraints**.

(another example)

because these directions



while these directions are valid for $\mathrm{D}(\underline{x})$ since their projection is 0 on the direction of the gradient, i.e. $\nabla g(\underline{x}) \cdot (\text{--}\!\to) = 0$

while the above ones are clearly $\nabla g(\underline{x}) \cdot (\text{--}) \le 0$ ($\underline{d}$)

Property: $\overline{\mathscr{D}}(\underline{\overline{x}}) \subseteq \mathrm{D}(\underline{\overline{x}})$ for all $\underline{\overline{x}} \in S$.

Proof:

Given any $\underline{d} \in \mathscr{D}(\underline{\overline{x}})$, for sufficiently small α we have

$$0 \geq g_i(\underline{\overline{x}} + \alpha \underline{d}) = \cancel{g_i(\underline{\overline{x}})} + \alpha \nabla g(\underline{\overline{x}})^T \underline{d} + o(\alpha)$$

$\forall i \in I(\underline{\overline{x}})$ with $g_i(\underline{\overline{x}}) = 0$

$$\Rightarrow \underbrace{\cancel{\alpha}}_{\alpha > 0} \nabla g(\underline{\overline{x}})^T \underline{d} \leq 0$$

$$\Rightarrow \nabla g(\underline{\overline{x}})^T \underline{d} \leq 0 \quad \forall i \in I(\underline{\overline{x}})$$

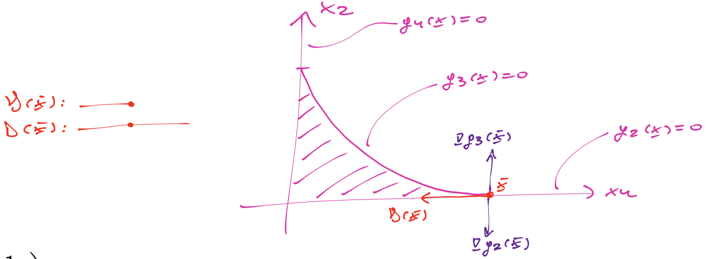
that is $\underline{d} \in D(\underline{\overline{x}})$ and hence $D(\underline{\overline{x}}) \subseteq D(\underline{\overline{x}})$.

But $D(\underline{\overline{x}})$ is a closed set, then we have also that $\overline{D}(\underline{\overline{x}}) \subseteq D(\underline{\overline{x}})$

Not all $\underline{d} \in \mathrm{D}(\overline{x})$ are feasible directions. as we now write the example of $-x_1^2$

Example:

$$g_1(\underline{x}) = -x_1 \leq 0$$
$$g_2(\underline{x}) = -x_2 \leq 0$$
$$g_3(\underline{x}) = -(1 - x_1)^3 + x_2 \leq 0$$



At $\overline{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ we have $\mathscr{D}(\overline{x}) = \{ (\alpha, 0) : \alpha < 0 \}$

Since $\nabla g_2(\bar{x}) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$ and $\nabla g_3(\bar{x}) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ we have

$\overline{\mathscr{D}}(\bar{x}) \subset D(\bar{x}) = \left\{ \begin{pmatrix} \alpha \\ 0 \end{pmatrix} : \alpha \in \mathbb{R} \right\}$

Theorem: (Extension of first order necessary optimality conditions)

If $f \in C^1$ on S and $\underline{\overline{x}} \in S$ is a local minimum of f on S, then

$$\boxed{\nabla f^t(\underline{\overline{x}})\underline{d} \geq 0 \qquad \forall \underline{d} \in \boxed{\overline{\mathscr{D}}(\underline{\overline{x}})},}$$

that is, all feasible directions are ascent directions.

where now we characterize more precisely the "feasible directions" in terms of $\mathscr{D}(\overline{x})$

Proof:

The result holds $\forall \underline{d} \in \mathscr{D}(\underline{\overline{x}})$.

For every $\underline{d} \in \overline{\mathscr{D}}(\underline{\overline{x}})$, $\exists$ a sequence $\{\underline{d}^k\}$ with $\underline{d}^k \in \mathscr{D}(\underline{\overline{x}})$ such that $\lim_{k\to\infty} \underline{d}^k = \underline{d}$.

Since $\nabla f^t(\underline{\overline{x}})\underline{d}^k \geq 0, \forall \underline{k}$, then $\lim_{k\to\infty} \nabla f^t(\underline{\overline{x}})\underline{d}^k = \nabla f^t(\underline{\overline{x}})\underline{d} \geq 0$.

But $\overline{\mathscr{D}}(\underline{\overline{x}})$ is difficult to characterize.

while $\mathscr{D}(\overline{x})$ was easy to characterize, we

Since $\mathrm{D}(\underline{\overline{x}})$ is well characterized, we introduce further conditions.

Definition: (Constraint Qualification CQ – Zangwill)

The constraint qualification assumption holds at $\underline{\overline{x}} \in S$ if $\boxed{\overline{\mathscr{D}}(\underline{\overline{x}}) = \mathrm{D}(\underline{\overline{x}})}$

Theorem: (Karush-Kuhn-Tucker necessary optimality conditions)

Suppose $f, g_i \in C^1$ and CQ assumption holds at $\underline{\overline{x}} \in \{\underline{x} \in \mathbb{R}^n : g_i(\underline{x}) \leq 0, \forall i \in I\}$.

feasible region

If $\underline{\overline{x}}$ is a local minimum of f over S then $\exists\ u_1, \ldots, u_m \geq 0$ (KKT-multipliers) such that:

one for each constr

$$\boxed{\nabla f(\underline{\overline{x}}) + \sum_{i \in I(\underline{\overline{x}})} u_i \nabla g_i(\underline{\overline{x}}) = \underline{0}} \equiv \begin{cases} \nabla f(\underline{\overline{x}}) + \sum_{i=1}^{m} u_i \nabla g_i(\underline{\overline{x}}) = \underline{0} \\ u_i g_i(\underline{\overline{x}}) = 0 \quad \forall i \in I \end{cases}$$

$$\Leftrightarrow \nabla f(\underline{\overline{x}}) = \sum_{i \in I(\underline{\overline{x}})} (-u_i) \nabla g_i(\underline{\overline{x}})$$

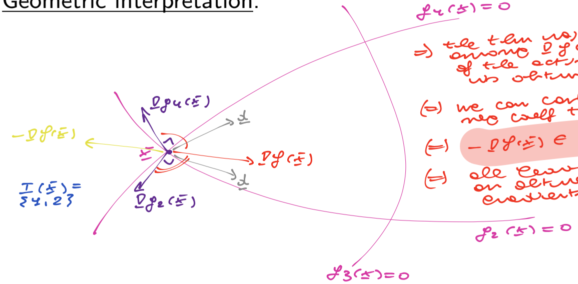
a rewriting where all the indices $i \in I$ (all the constr + that new condition)

not linear conditions since both u_i and $\underline{\overline{x}}$ are unknowns ⇒ we will study cases

$\underline{\overline{x}}$ must also satisfy all the constraints $g_i(\underline{x}) \leq 0, \forall i \in I$.

i.e. $\underline{\overline{x}}$ must be ∈ feasible region

Geometric interpretation:



⇒ the two ways that the angle among $\nabla f(\underline{\overline{x}})$ and the gradients of the active constraints at $\underline{\overline{x}}$ we obtain

⟺ we can combine $\nabla g_i(\underline{\overline{x}})$ with neg coeff to get $\nabla f(\underline{\overline{x}})$

⟺ $-\nabla f(\underline{\overline{x}}) \in \text{cone}(\nabla g_i(\underline{\overline{x}}) : i \in I(\underline{\overline{x}}))$

⟺ all feasible directions $\underline{d}$ form an angle with all the active constr gradients (we are minimizing directions)

Theorem: (Karush-Kuhn-Tucker necessary optimality conditions)

Suppose $f, g_i \in C^1$ and CQ assumption holds at $\underline{\overline{x}} \in \{\underline{x} \in \mathbb{R}^n : g_i(\underline{x}) \leq 0, \forall i \in I\}$.

If $\underline{\overline{x}}$ is a local minimum of f over S then $\exists\, u_1, \ldots, u_m \geq 0$ (KKT-multipliers) such that:

$$\nabla f(\underline{\overline{x}}) + \sum_{i \in I(\underline{\overline{x}})} u_i \nabla g_i(\underline{\overline{x}}) = \underline{0} \;\equiv\; \begin{cases} \nabla f(\underline{\overline{x}}) + \sum_{i=1}^{m} u_i \nabla g_i(\underline{\overline{x}}) = \underline{0} \\ u_i g_i(\underline{\overline{x}}) = 0 \quad \forall i \in I \end{cases}$$

$\underline{\overline{x}}$ must also satisfy all the constraints $g_i(\underline{x}) \leq 0, \forall i \in I$.

$u_i > 0 \Rightarrow$ the corresponding constr must be saturated at equality)

Geometric interpretation:

So based on this we will develop the cases

Proof:

Assuming CQ holds at $\underline{\overline{x}}$, we have $\overline{\mathscr{D}}(\underline{\overline{x}}) = \mathrm{D}(\underline{\overline{x}})$.

nec cond

NC for $\underline{\overline{x}}$ to be a local minimum of f over S is

(thanks to the CQ assumption)

$$\nabla^t f(\underline{\overline{x}})\underline{d} \geq 0, \quad \forall \underline{d} \ \text{ such that } \ \overbrace{\nabla^t g_i(\underline{\overline{x}})\underline{d} \leq 0}^{\text{that will be } \underline{d} \in D(\underline{\overline{x}})} \quad \forall i \in I(\underline{\overline{x}}). \tag{2}$$

Farkas Lemma:

$$\begin{cases} A\underline{u} = \underline{b} \\ \underline{u} \geq 0 \end{cases} \text{ has a solution } \Leftrightarrow \begin{cases} \underline{b}^t \underline{d} \geq 0 \\ \forall \underline{d} \ \text{ such that } \ \underline{d}^t A \geq 0 \end{cases}$$

thinking about the structure of the term before we can take

$$\underbrace{\nabla f(\underline{\overline{x}})}_{\underline{b}} + \sum_{i \in I(\underline{\overline{x}})} u_i \nabla g_i(\underline{\overline{x}}) = 0$$

$$A = \begin{pmatrix} | & & | \\ \nabla g_{i_1}(\underline{\overline{x}}) & \cdots & \nabla g_{i_l}(\underline{\overline{x}}) \\ | & & | \end{pmatrix}$$

which has
- n rows
- $|I| = l$ cols

then (2) is equivalent to
$\underline{b}^T \underline{d} \geq 0 \quad \forall \underline{d} \in \mathbb{R}^n : \underline{d}^T A \geq 0 \quad (2^*)$

But according to Farkas Lemma we have
$(2^*) \iff \exists \underline{u} \geq \underline{0} : A\underline{u} = \underline{b}$
$\iff \exists u_i \geq 0 \ \forall i \in I(\underline{\overline{x}}) : \nabla f(\underline{\overline{x}}) = \sum_{i \in I(\underline{\overline{x}})} (-u_i) \nabla g_i(\underline{\overline{x}})$

Example 1:

$$\begin{array}{ll} \min & f(\underline{x}) = x_1 + x_2 \\ \text{s.t.} & g_1(\underline{x}) = x_1^2 + x_2^2 \leq 2 \\ & g_2(\underline{x}) = -x_2 \leq 0 \end{array}$$

actually $g_1(\underline{x}) = x_1^2 + x_2^2 - 2 \leq 0$ as we must have zero on the RHS

KKT conditions: the candidate points $\bar{x}$ must satisfy the following conditions:

$$\nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} (-u_i) \nabla g_i(\bar{x}) = 0$$

well we use the other formulation (easier for the exercises)

KKT cond

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \left[u_1 \begin{pmatrix} 2\bar{x}_1 \\ 2\bar{x}_2 \end{pmatrix} + u_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} \right] = 0$$

$$u_1 g_1(\bar{x}) = u_1 (\bar{x}_1^2 + \bar{x}_2^2 - 2) = 0$$

$$u_2 g_2(\bar{x}) = u_2 (-x_2) = 0$$

$$u_1 \geq 0$$
$$u_2 \geq 0$$

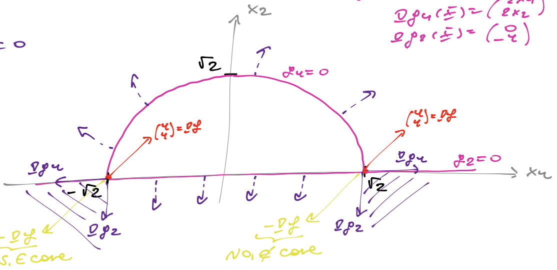
feasibility cond of $\bar{x}$

$$\bar{x}_1^2 + \bar{x}_2^2 - 2 \leq 0$$
$$-\bar{x}_2 \leq 0$$

$\nabla f(\underline{x}) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$\nabla g_1(\bar{x}) = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix}$

$\nabla g_2(\bar{x}) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$



$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} + u_1 \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} + u_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{aligned} & u_1(x_1^2 + x_2^2 - 2) = 0 \\ & u_2(-x_2) = 0 \\ & x_1^2 + x_2^2 \le 2 \\ & -x_2 \le 0 \\ & u_1 \ge 0, u_2 \ge 0 \end{aligned}$$

Four cases:

(C1) $u_1 = 0$, $u_2 = 0$ $\Rightarrow$ $\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \sum 0 \cdots = \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ impossible

(C2) $u_1 = 0$, $u_2 > 0$ $\Rightarrow$ $\begin{pmatrix} 1 \\ 1 \end{pmatrix} + 0 \cdots + u_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ impossible

(C3) $u_1 > 0$, $u_2 = 0$ $\Rightarrow$ $\begin{pmatrix} 1 \\ 1 \end{pmatrix} + u_1 \begin{pmatrix} 2\bar{x}_1 \\ 2\bar{x}_2 \end{pmatrix} + 0 \cdots = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$1 + 2u_1\bar{x}_2 = 0$ impossible

$u_1 = -\frac{1}{2\bar{x}_2} \le 0$

since $x_2 \ge 0$ (i.e. feasibility)

we lie on the circle

(C4) $u_1 > 0$, $u_2 > 0$ $\Rightarrow$ $\begin{cases} \bar{x}_1^2 + \bar{x}_2^2 - 2 = 0 \\ \bar{x}_2 = 0 \end{cases}$ $\Rightarrow$ two pts: $A = \begin{pmatrix} -\sqrt{2} \\ 0 \end{pmatrix}$, $B = \begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix}$

$x_2 = 0$

We now check to have $u_i \ge 0$ or to find which are candidate points

we satisfy KKT conds (all)

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} + u_1 \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} + u_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{aligned} & u_1(x_1^2 + x_2^2 - 2) = 0 \\ & u_2(-x_2) = 0 \\ & x_1^2 + x_2^2 \leq 2 \\ & -x_2 \leq 0 \\ & u_1 \geq 0, u_2 \geq 0 \end{aligned}$$

Four cases:

we are on the circle

(CG) $u_1 > 0$, $u_2 > 0$

$x_2 = 0$

$\Rightarrow \begin{cases} \bar{x}_1^2 + \bar{x}_2^2 - 2 = 0 \\ \bar{x}_2 = 0 \end{cases} \Rightarrow$ two vectors: $B = \begin{pmatrix} -\sqrt{2} \\ 0 \end{pmatrix}$, $A = \begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix}$

We now check to have $u_i \geq 0$ so to they really are candidate points

we using KKT conds (all)

A: $\begin{pmatrix} 1 \\ 1 \end{pmatrix} + u_1 \begin{pmatrix} 2\sqrt{2} \\ 0 \end{pmatrix} + u_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow$ $u_2 = 1$ good, $u_1 = -\frac{1}{2\sqrt{2}} < 0$ bad

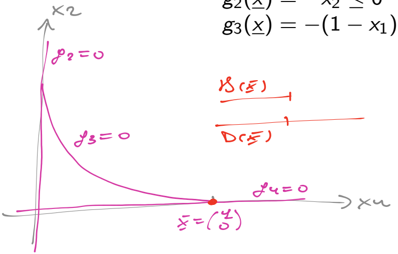
B: $\begin{pmatrix} 1 \\ 1 \end{pmatrix} + u_1 \begin{pmatrix} -2\sqrt{2} \\ 0 \end{pmatrix} + u_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow$ $u_2 = 1$ good, $u_1 = \frac{1}{2\sqrt{2}} > 0$ good

$\Rightarrow$ only B is a candidate point (and indeed it is the local & abs minimum)

If CQ assumption does not hold at $\overline{x}$, KKT conditions need not be necessary for local optimality.

Example 2:

$$\begin{array}{ll} \min & f(\underline{x}) = -x_1 \\ s.t. & g_1(\underline{x}) = -x_1 \leq 0 \\ & g_2(\underline{x}) = -x_2 \leq 0 \\ & g_3(\underline{x}) = -(1-x_1)^3 + x_2 \leq 0 \end{array}$$



we know that $D(\bar{x}) \supset \bar{T}(\bar{x})$ for $\bar{x} = \binom{1}{0}$

now $\bar{x}$ is still the min for $f(x)$, but no multipliers exist to meet KKT conditions.

Since (and $u_2 = 0$) we have

$$\binom{-1}{0} + u_1 \binom{-1}{0} + u_3 \binom{0}{1} \neq \binom{0}{0} \quad \forall u_1, u_3 \geq 0$$

since constr 2 is not active

how do we check to have CQ assumption?

Proposition: (Sufficient conditions for Constraint Qualification)

1) If

- all g_i are linear functions (Karlin)

or

- all g_i are convex and $\exists \underline{a}$ such that $g_i(\underline{a}) < 0, \forall i \in I$, (Slater)

we ∃ or whatever feasible point

CQ assumption holds $\boxed{\text{at every } \underline{x} \in S.}$

2) If $\nabla g_i(\overline{x}),\ i \in I(\overline{x})$, are linearly independent, CQ assumption holds $\boxed{\text{at } \bar{\underline{x}} \in S.}$

the gradients of the
active constr

N.B.: When the gradients of the active constraints are linearly independent,
KKT multiplier vector is unique.

which are our problem structure ensures that feasible region is convex

Theorem: (Necessary and sufficient conditions – convex problems)

Slater cond

If $f \in C^1$, $g_i \in C^1$ $\forall i \in I$ are convex, and $\exists\, \underline{a}$ such that $g_i(\underline{a}) < 0, \forall i \in I$, then $\underline{x}^* \in S$ is a global minimum if and only if $\exists\, u_1, \ldots, u_m \geq 0$ such that

satisfy the KKT conditions

$$\begin{cases} \nabla f\left(\underline{x}^*\right) + \sum_{i=1}^{m} u_i \nabla g_i\left(\underline{x}^*\right) = \underline{0} \\ u_i g_i(\underline{x}^*) = 0 \qquad \forall i \in I. \end{cases}$$

+ (feasibility of $\underline{x}$) & ($\underline{u} \geq \underline{0}$)

Proof:

(⇒) if $\underline{x}^*$ is a (loc) min, then $\underline{x}^*$ satisfies KKT conditions through Slater assumption.

(⇐) if we know the KKT conds then $\nabla f(\underline{x}^*)^T \underline{d} \geq 0 \quad \forall \underline{d} \in \mathcal{D}(\underline{x}^*) = D(\underline{x}^*)$ (from CQ and Farkas lemma)

therefore $\underline{x}^*$ is the (global) min

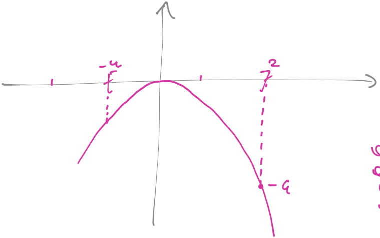
For Linear Programs, it amounts to the complementary slackness theorem.

Remark: Result holds under milder convexity conditions (f pseudoconvex and the g_i's quasiconvex).

If f is not convex, KKT conditions are not sufficient.

Example 3:

$$\min f(x) = -x^2$$
$$\left.\begin{array}{l} g_1(x) = -2 + x \leq 0 \\ g_2(x) = -x - 1 \leq 0 \end{array}\right\} x \in [-1, 2]$$



at $x = -1$ the KKT cond are satisfied but the candidate point is a local min and not a global min

So convexity is really important to get suff cond for global optimality

General case ~ with also equality constr

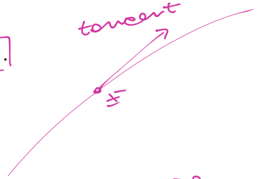
Consider

$$
\boxed{
\begin{array}{lll}
\min & f(x) & \\
\text{s.t.} & g_i(\underline{x}) \leq 0 & i \in I = \{1, ..., m\} \\
 & h_l(\underline{x}) = 0 & l \in L = \{1, ..., p\} \\
 & \underline{x} \in X \subseteq \mathbb{R}^n &
\end{array}
}
$$

where f, g_i, $h_l \in C^1$.

When nonlinear equality constraints, usually $\boxed{\mathscr{D}(\underline{x}) = \{\underline{0}\}}$.

Extend previous results by defining cone of directions.



we now are more difficult to get/see the feasible directions

and the concept of $\mathscr{D}(\bar{x})$ will need to be extended

Definition: *Closed cone of the tangents* at $\underline{\overline{x}}$

where $\{x_k\} \subseteq S$ are feasible pts

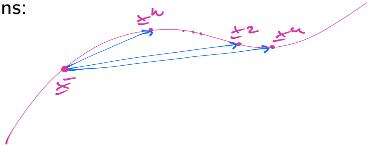
$$\mathscr{T}(\underline{\overline{x}}) = \left\{ \underline{d} \in \mathbb{R}^n : \underline{d} = \lambda \underbrace{\lim_{k\to\infty} \frac{\underline{x}^k - \underline{\overline{x}}}{\|\underline{x}^k - \underline{\overline{x}}\|}}_{\text{limit direction}},\ \lambda \geq 0,\ \left\{\underline{x}^k\right\} \underset{k\to\infty}{\rightarrow} \underline{\overline{x}} \text{ with } \underline{x}^k \neq \underline{\overline{x}} \right\}$$

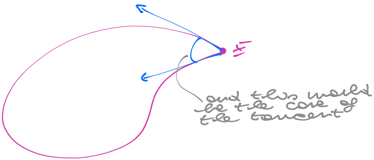
(if $d \in \mathscr{T}(\bar{x})$ and $\exists$ a sequence $(x_k) \to \bar{x}$, then directions of the chords $x_k - \bar{x}$ converge to d)

of the corresponding chords that we get with the sequence

we can also come there through the interior

Illustrations:





Definition: (Constraint Qualification CQ – Abadie)

The CQ assumption holds at $\underline{\overline{x}} \in S$ if $\boxed{\mathscr{T}(\underline{\overline{x}}) = \mathrm{D}(\underline{\overline{x}}) \bigcap \mathrm{H}(\underline{\overline{x}})}$ where

$D(\bar{x}) = \{ d \in \mathbb{R}^n : \nabla g_i(\bar{x}) \cdot d \leq 0 \quad \forall i \in I(\bar{x}) \}$ for all the active constr (convex)

$H(\bar{x}) = \{ d \in \mathbb{R}^n : \nabla h_\ell(\bar{x}) \cdot d = 0 \quad \forall \ell \in L \}$ for all the equality constr (which being at equality are active by definition)

Theorem: (General KKT necessary optimality conditions)

eg the one of above

Suppose $f \in C^1$, $g_i \in C^1\ \forall i$, $h_l \in C^1\ \forall l$ and CQ assumption holds at $\underline{\overline{x}} \in \mathrm{S}$.

If $\underline{\overline{x}}$ is a local minimum of f over S then $\exists\, u_i \geq 0$, $\forall i \in I(\underline{\overline{x}})$ and $v_l \in \mathbb{R}$, $\forall l \in L$ such that

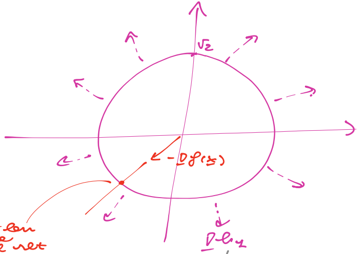
$$\boxed{\nabla f(\underline{\overline{x}}) + \sum_{i \in I(\underline{\overline{x}})} u_i \nabla g_i(\underline{\overline{x}}) + \sum_{l \in L} v_l \nabla h_l(\underline{\overline{x}}) = \underline{0}.}$$

N.B.: If only equalities, KKT conditions coincide with classical Lagrange optimality conditions.

$\Longleftrightarrow$ we can express $-\nabla f(\underline{\overline{x}})$ as a linear combination of all the active constraints

<u>Example 1</u>:

min $\quad f(\underline{x}) = x_1 + x_2$
s.t. $\quad {x^2}_1 + {x^2}_2 = 2$



$-\nabla f(\underline{x}) = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$

$h_1(\underline{x}) = x_1^2 + x_2^2 - 2$

$\underline{\nabla} h_1(\underline{x}) = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix}$

$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$ was then the optimal set

for showing normal wrt to the tangent

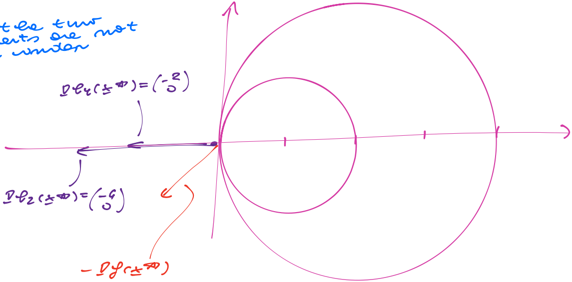
Example 2:

$$\begin{array}{ll} \min & f(\underline{x}) = x_1 + x_2 \\ \text{s.t.} & (x_1-1)^2 + {x^2}_2 - 1 = 0 \\ & (x_1-2)^2 + {x^2}_2 - 4 = 0 \end{array}$$

$R = 1 \quad C = (1,0)$

$R = 2 \quad C = (2,0)$

here the two gradients are not lin indep



we see that we can't express it as lin comb of the two gradients above

but $\underline{x}^* = \binom{0}{0}$ is the optimal sol (being the only feasible point)

$\Rightarrow$ $\nexists\, v_1, v_2$ st the KKT cond are satisfied (since $\underline{x}^*$ here is an irregular point)

Proposition: (Sufficient conditions for CQ)

se esiste

- If g_i convex, h_l linear and $\exists \underline{a} \in X$ such that $g_i(\underline{a}) < 0, \forall i \in I$ and $h_l(\underline{a}) = 0 \ \forall l \in L$, then CQ assumption holds at every $\underline{x} \in S$.
- If $\nabla g_i(\overline{x})$, $\forall i \in I(\overline{x})$, and $\nabla h_l(\overline{x})$, $\forall l \in L$, are linearly independent then CQ assumption holds at $\overline{x} \in S$.

es) extension considering all eq constr & active constr

5.3 Sufficient optimality conditions

Generic NLP

$$(P) \quad \begin{cases} \min & f(\underline{x}) \\ \text{s.t.} & g_i(\underline{x}) \leq 0 \qquad \forall i \in I = \{1, \ldots, m\} \\ & \underline{x} \in X \subseteq \mathbb{R}^n \end{cases}$$

set constr

where X is an arbitrary subset (even discrete).

Definitions

$\begin{cases} > 0 \text{ (we bad) if constr was not saturated} \\ = 0 \text{ (we good) if constr was saturated} \end{cases}$

- The **Lagrange function** associated with (P) is

$$\boxed{L(\underline{x}, \underline{u}) = f(\underline{x}) + \sum_{i \in I} u_i g_i(\underline{x})} \quad \forall \underline{x} \in X \text{ and } \boxed{\underline{u} \geq \underline{0}}$$

set constr remains there

N.B.: $\underline{u} \geq \underline{0}$ since $g_i(\underline{x}) \leq 0$.

- $(\overline{x}, \overline{u})$ with $\overline{x} \in X$ and $\overline{u} \geq \underline{0}$ is a **saddle point** of $L(\underline{x}, \underline{u})$

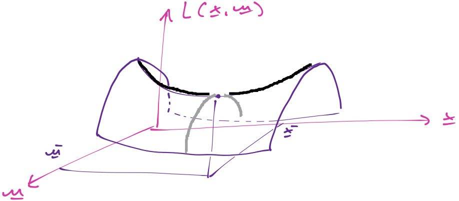
if $\boxed{L(\overline{x}, \overline{u}) \leq L(\underline{x}, \overline{u})}$ $\forall \underline{x} \in X$ and $\boxed{L(\overline{x}, \underline{u}) \leq L(\overline{x}, \overline{u})}$ $\forall \underline{u} \geq 0$,

that is, $\overline{x}$ minimizes $L(\underline{x}, \overline{u})$ over X and $\overline{u}$ maximizes $L(\overline{x}, \underline{u})$ over $\mathbb{R}^m$.

$\bar{x}$ is the min at u fixed

$\bar{u}$ is the max at x fixed

Illustration:



Proposition: (Characterization of saddle points)

$(\overline{x}, \overline{u})$ with $\overline{x} \in X$ and $\overline{u} \geq \underline{0}$ is a saddle point of $L(\underline{x}, \underline{u})$ if and only if

i) $L(\overline{x}, \overline{u}) = \min_{x \in X} L(\underline{x}, \overline{u})$ — this is trivial, was part of the definition

ii) $g_i(\overline{x}) \leq 0 \quad \forall i \in I$ — $\bar{x}$ is a feasible sol (since we ask all the algebraic constraints) also wrt to natural)

iii) $\bar{u}_i g_i(\overline{x}) = 0 \quad \forall i \in I$ — complementary conditions: (if one of them is >0) $\Rightarrow$ (the other has to be $=0$)

Proof*:

Consequence of the definition of the Lagrangian function and a lot of playing with algebra

a very strong result

Theorem: (Sufficient optimality condition)

If $(\overline{x}, \overline{u})$ is a saddle point of $L(\underline{x}, \underline{u})$, then $\overline{x}$ is a *global minimum* of problem (P).

Proof: Condition (i) of the characterization $\Rightarrow L(\overline{x}, \overline{u}) \leq L(x, \overline{u}) \quad \forall x \in X$

Lagrangian function def $\Rightarrow f(\overline{x}) + \underbrace{\cancel{\sum_{i \in I} \overline{u}_i g_i(\overline{x})}}_{\text{this is } = 0 \text{ by characterization (3)}} \leq f(x) + \sum_{i \in I} \overline{u}_i g_i(x) \quad \forall x \in X$

$\Rightarrow f(\overline{x}) \leq f(x) + \sum_{i \in I} \underbrace{\overline{u}_i}_{\geq 0} \underbrace{g_i(x)}_{\text{this will be } \leq 0 \text{ by charact (2)}} \quad \forall x \in X$

and in this way we make appear all the constr of (P)

that's because since $\overline{u} \geq 0$ we have $f(\overline{x}) \leq f(x) \quad \underbrace{\forall x \in X \text{ st } g_i(x) \leq 0 \ \forall i \in I}_{\text{we are feasible for the problem (P)}}$

Observations:

- Result applies to any mathematical program (convex or not, with f and g_i differentiable or not, X continuous or discrete,...).
- For some problems a saddle point may not exist, in general for nonconvex problems.

Example:

$$\begin{array}{ll} \min & f(x) = -x^2 \\ \text{s.t.} & 2x - 1 \leq 0 \\ & 0 \leq x \leq 1 \end{array}$$

where $g(x) = 2x - 1$ and $X = \{x : 0 \leq x \leq 1\}$

$L(\underline{x}, u) = f(\underline{x}) + \sum_{i \in I} u_i g_i(\underline{x}) = \Big|_{|I|=1} (-x^2) + u(2x-1)$

which is concave wrt x (for a fixed $u \geq 0$) $\Rightarrow$ $\min_{x \in [0,1]} L(x, u)$ is reached at an extreme point, i.e. $x = 0$ or $x = 1$

while we can easily find $\underline{x}^* = \frac{1}{2}$ but $\nexists$ saddle points of the type $(\frac{1}{2}, u)$

$\Rightarrow$ it indeed is only a SUF cond, not NEC, but:

Theorem: (saddle point for convex problems)

Suppose f and g_i, $\forall i \in I$ are convex, $X \subseteq \mathbb{R}^n$ is convex and $\exists\, \underline{a} \in X$ such that $\underline{g}(\underline{a}) < \underline{0}$. ($\exists$ an interior feasible point)

If (P) has an optimal solution $\overline{x}$, $\exists\, \overline{u} \geq \underline{0}$ such that $(\overline{x}, \overline{u})$ is a saddle point of $L(\underline{x}, \underline{u})$.

it becomes NEC & SUF for convex problems (before was a concave case)

Connection with KKT conditions for convex problems

If f and $g_i \in C^1$ are convex, $X = \mathbb{R}^n$ and $\exists\, \underline{a} \in X$ such that $\underline{g}(\underline{a}) < \underline{0}$, then $\underline{\overline{x}}$ is an optimal solution if and only if $\underline{\overline{x}}$ satisfies the KKT conditions.

∃ an interior feasible point

Proof: $\underline{\overline{x}}$ is an opt sol $\Longleftrightarrow$ $\exists\, \underline{\overline{u}} \geq \underline{0}$: $(\underline{\overline{x}}, \underline{\overline{u}})$ is a saddle point for $L(\underline{x}, \underline{u})$

($\Leftarrow$) It is the SUF condition for global optimality
($\Rightarrow$) It is the previous prop. thm (about the always existence of a saddle point)

Since $L(\underline{x}, \underline{\overline{u}}) = f(\underline{x}) + \sum \overline{u}_i g_i(\underline{x}) = f(\underline{x}) + \underbrace{\underline{\overline{u}}}_{\geq \underline{0}} \cdot \underline{g}(\underline{x})$ is convex, then we can look (due to condition (u)) for the minimum point of unknown ~ we look for a stationary point

$$\nabla_{\underline{x}} L(\underline{\overline{x}}, \underline{\overline{u}}) = 0$$

which coincides with the KKT conditions

$$\nabla f(\underline{\overline{x}}) + \sum_{i \in I} \overline{u}_i \nabla g_i(\underline{\overline{x}}) = \underline{0}$$

and the CQ assumption holds at every feasible sol

N.B.: 1) Without convexity assumption a stationary point $\underline{\overline{x}}$ may not minimize $L(\underline{x}, \underline{\overline{u}})$.

2) KKT multipliers are then identical to Lagrange multipliers at the saddle point.

5.4 Lagrangian duality

Generic NLP:

$$(P) \quad \begin{cases} \min & f(\underline{x}) \\ \text{s.t.} & g_i(\underline{x}) \leq 0 \qquad \forall i \in I = \{1, \ldots, m\} \\ & \underline{x} \in X \subseteq \mathbb{R}^n \end{cases}$$

To any minimization NLP we can associate a maximization NLP such that, under some assumptions, the objective function values of respective optimal solutions coincide.

Tackle the *primal* problem (P) indirectly, by solving the dual (second) problem.

To try to solve (P), we can look for a saddle point of the Lagrange function.

this can be convenient in some cases

Dual function:

$$w(\underline{u}) = \min_{\underline{x} \in X} L(\underline{x}, \underline{u}) \qquad \forall \underline{u} \geq \underline{0}$$

for every $\underline{u} \geq \underline{0}$ we fix, we get $w(\underline{u})$ by solving that problem over the $\underline{x} \in X$

Well-defined if, for instance, f and the g_i's are continuous and X is compact.

Search for a saddle point (if $\exists$):

$\Rightarrow$ objective function will work for a saddle point:

$$\max_{\underline{u} \geq \underline{0}} \left(\min_{\underline{x} \in X} L(\underline{x}, \underline{u}) \right)$$

Dual problem: $\quad (D) \begin{cases} \max \ w(\underline{u}) \\ \underline{u} \geq \underline{0} \end{cases}$

N.B.: $w(\underline{u})$ and (D) are defined even if no saddle point exists.

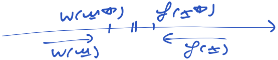
and also works when we have removed the algebraic constraints. But yes of course the function $w(\underline{u})$ may not have good properties, as a trade-off

Observations:

1) Different Lagrangian duals of (P) depending on which $g_i(\underline{x}) \leq 0$ are dualized. Choice affects optimal value of (D) and complexity to evaluate $w(\underline{u})$.

we could decide to keep out none of the constraints

2) Lagrangian dual is useful to solve large-scale LPs and (non)convex/discrete optimization problems.



Theorem: (Weak duality)

For every feasible $\underline{x}$ of (P) and $\underline{u} \geq \underline{0}$ (feasible) of (D), we have $\boxed{w(\underline{u}) \leq f(\underline{x})}$.

Proof: By definition of Lagrangian function:

$$w(\underline{u}) = \min_{\underline{x} \in X} \left(f(\underline{x}) + \underline{u}^T \underline{g}(\underline{x}) \right) \leq f(\underline{x}) + \underbrace{\underline{u}^T}_{\geq 0} \underbrace{\underline{g}(\underline{x})}_{\leq 0 \text{ for } \underline{x} \in X \text{ (we feasible)}} \quad \forall \underline{x} \in X \; (\underline{u} \geq \underline{0})$$

So if $g(\underline{x}) \leq \underline{0}$ then we have $w(\underline{u}) \leq f(\underline{x})$.
(due also to $\underline{u} \geq \underline{0}$)

In particular, for every $\underline{u} \geq 0$ we have $w(\underline{u}) \leq f(\underline{x}^*)$ for an optimal $\underline{x}^*$ of (P).

$\Rightarrow$ the term gives a lower bound on the best opt val of function $(f(\underline{x}^*))$ value

Consequence:

If a feasible solution $\overline{x}$ of (P) and $\overline{u} \geq \underline{0}$ satisfy $\boxed{w(\overline{u}) = f(\overline{x})}$, $\overline{x}$ is optimal for (P) and $\overline{u}$ is optimal for (D).

For Linear Programs the objective function values of optimal solutions of (P) and (D) coincide, for NLPs this is not always the case.

Theorem: (Strong duality)

i) If (P) has a saddle point $(\underline{\overline{x}}, \underline{\overline{u}})$, then – not only $\underline{\overline{x}}$ is optimal for (P)
– but also $\underline{\overline{u}}$ is optimal for (D)

$$\left\{ \begin{array}{l} \max \; w(\underline{u}) \\ \underline{u} \geq \underline{0} \end{array} \right. = \boxed{w(\underline{\overline{u}}) = f(\underline{\overline{x}})} = \min \left\{ f(\underline{x}) \, : \, \underline{g}(\underline{x}) \leq \underline{0}, \underline{x} \in X \right\}.$$

ii) If $\exists$ a feasible $\underline{\bar{x}}$ of (P) and $\underline{\bar{u}} \geq \underline{0}$ such that $w(\underline{\bar{u}}) = f(\underline{\bar{x}})$, then $(\underline{\bar{x}}, \underline{\bar{u}})$ is a saddle point of $L(\underline{x}, \underline{u})$. and also the converse result holds:
$w(\underline{u}^*) = f(\underline{x}^*) \Rightarrow (\underline{x}^*, \underline{u}^*)$ is a saddle point

Proof: (i) Since $(\underline{\bar{x}}, \underline{\bar{u}})$ is a sad. point we have that
$$w(\underline{\bar{u}}) = L(\underline{\bar{x}}, \underline{\bar{u}}) = \min_{\underline{x} \in X} L(\underline{x}, \underline{\bar{u}}) \quad \text{from char (4)}$$

Moreover, by definition of Lagrangian function
$$L(\underline{\bar{x}}, \underline{\bar{u}}) = f(\underline{\bar{x}}) + \underbrace{\cancel{\underline{u}^T \underline{g}(\underline{\bar{x}})}}_{=0 \text{ from char (3)}} = f(\underline{\bar{x}}) = \left\{ \begin{array}{l} \min f(\underline{x}) \\ \text{st } \underline{g}(\underline{x}) \leq 0 \\ \underline{x} \in X \end{array} \right\}$$
SOC (suff opt condition)

$\Rightarrow \; w(\underline{\bar{u}}) = f(\underline{\bar{x}})$

Because of weak duality we have $w(\underline{u}) \leq f(\underline{\bar{x}}) \;\; \forall \underline{u} \geq \underline{0}$
where $w(\underline{\bar{u}}) = \left\{ \begin{array}{l} \max w(\underline{u}) \\ \text{st } \underline{u} \geq \underline{0} \end{array} \right\}$

(iii) Let $\underline{\bar{x}}$ be a feasible sol of (P) and $\underline{\bar{u}} \geq \underline{0}$ st $w(\underline{\bar{u}}) = f(\underline{\bar{x}})$
(i.e. the statement)

By definition of $w(\cdot)$ we have:
$$w(\underline{\bar{u}}) \leq f(\underline{x}) + \underline{\bar{u}}^T \underline{g}(\underline{x}) \quad \forall \underline{x} \in X$$

For $\underline{x} = \underline{\bar{x}}$ we have
$$f(\underline{\bar{x}}) = w(\underline{\bar{u}}) \leq f(\underline{\bar{x}}) + \underline{\bar{u}}^T \underline{g}(\underline{\bar{x}})$$

$$\Rightarrow w(\underline{\bar{u}}) - f(\underline{\bar{x}}) = 0 \boxed{\leq} \underbrace{\underline{u}^T}_{\geq 0} \underbrace{\underline{g}(\underline{\bar{x}})}_{\leq 0 \text{ for } \underline{\bar{x}} \text{ which is feasible}}$$

cond (3)
$\Rightarrow \bar{u}_i g_i(\underline{\bar{x}}) = 0 \quad \forall i \in I$

Since $\underline{\bar{u}} \geq \underline{0}$, $g_i(\underline{\bar{x}}) \leq 0 \;\; \forall i \in I$ (cond (2)) and $w(\underline{\bar{u}}) = L(\underline{\bar{x}}, \underline{\bar{u}}) = \min_{\underline{x} \in X} L(\underline{x}, \underline{u})$ (cond (4))

$\Rightarrow$ we have all the characterization points and so $(\underline{\bar{x}}, \underline{\bar{u}})$ is a saddle point

Consequence:

If f, g_i's and $X \subseteq \mathbb{R}^n$ are convex, $\exists\, \underline{a}$ (a feasible point) such that $\underline{g}(\underline{a}) < \underline{0}$ and (P) has a finite optimal solution, $\exists$ a saddle point $(\underline{\bar{x}}, \underline{\bar{u}})$ and i) holds:

$$\left\{ \begin{array}{l} \max \;\; w(\underline{u}) \\ \underline{u} \geq \underline{0} \end{array} \right. = \min \left\{ f(\underline{x}) \; : \; \underline{g}(\underline{x}) \leq \underline{0}, \underline{x} \in X \right\}.$$

N.B.: *Strong duality*, the optimal values of the two objective functions coincide.

In general, we can have a *duality gap* ($<$ instead of $=$).

we for non-convex problems

1) $\begin{cases} \text{MIN} & x_1^2 + x_2^2 \\ \text{s.c.} & 2x_1 + x_2 \le -4 \end{cases}$

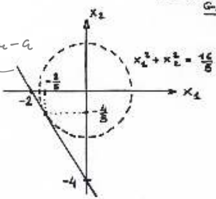
$g_i(x) = 2x_1 + x_2 - 4 \boxed{\le 0}$

f & g CONVEX, $a = \binom{-2}{1}$ $\Rightarrow \exists$ SADDLE-PT $\Rightarrow$ NO DUALITY GAP

line $x_2 = -2x_1 - 4$

$\underline{x}^* = \left(-\frac{8}{5}, -\frac{4}{5}\right)$

$f(\underline{x}^*) = \frac{16}{5}$



$f(x) + \sum_{i \in I} u_i g_i(x)$

$L(\underline{x}, u) = x_1^2 + x_2^2 + 2ux_1 + ux_2 + 4u$ → CONVEX

here we minimized $L(x,u)$ wrt x

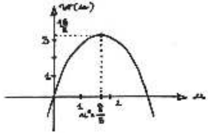
$\begin{cases} \frac{\partial L}{\partial x_1} = 2x_1 + 2u = 0 \\ \frac{\partial L}{\partial x_2} = 2x_2 + u = 0 \end{cases} \Rightarrow \begin{cases} x_1 = -u \\ x_2 = -\frac{u}{2} \end{cases}$ we plug them into $L(x,u)$ $\Rightarrow w(u) = -\frac{5u^2}{4} + 4u$ → CONCAVE

(D) $\max\limits_{u \ge 0} \; w(u)$ $\qquad \frac{\partial w(u)}{\partial u} = -\frac{5u}{2} + 4 \Rightarrow u^* = \frac{8}{5} \ge 0$

Thus $w(u^*) = -\frac{5}{4}\left(\frac{8}{5}\right)^2 + \frac{32}{5} = \frac{16}{5}$

$= f(\underline{x}^*)$

NO DUALITY GAP



2) $\begin{cases} \text{MIN} & -2x_1 + x_2 \\ \text{S.C.} & x_1 + x_2 - 3 \boxed{= 0} \\ & \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in X = \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 4 \end{pmatrix}, \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \begin{pmatrix} 4 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix} \right\} \end{cases}$

INTEGER PROGRAM

(P) : $\underline{x}^* = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, $f(\underline{x}^*) = -3$

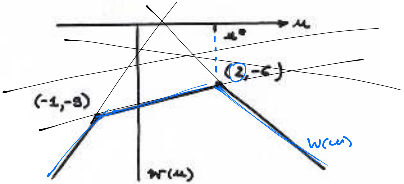
$L(\underline{x}, u) = -2x_1 + x_2 + u\,(x_1 + x_2 - 3)$ u UNRESTRICTED

$$w(u) = \min_{\underline{x} \in X} L(\underline{x}, u) = \begin{cases} -4 + 5u & \text{FOR } u \leq -1 \\ -8 + u & \text{FOR } -1 \leq u \leq 2 \\ -3u & \text{FOR } u \geq 2 \end{cases}$$

For each $\underline{x} \in X$ we get a different linear function in u, we can take the min we take the lower envelope

$u^* = 2$, $w(u^*) = -6 < f(\underline{x}^*)$

$\exists$ DUALITY GAP



Since under certain conditions we can solve (P) indirectly by solving (D)

Property 1: The dual function $w(\underline{u})$ is concave. regardless of any assumption

which is a good news since we are maximizing it
so in general solving the dual (D) is a lot easier than (P)

Proof*:

Observations:

- If $X \subseteq \mathbb{Z}^n$, $w(\underline{u})$ is not everywhere continuously differentiable. Concave piecewise linear function, lower envelope of a (in)finite family of hyperplanes in $\mathbb{R}^{n+1}$.
- In general (D) is easier than (P).
- Since $w(\underline{u})$ is concave local optima are global optima, but need for ad hoc solution method: **subgradient method**. that we solved with also in the case of non-everywhere differentiable functions (as we have seen before)

the set of all the $\underline{x}$s for which the min was reached, where $w(\underline{\tilde{u}}) = \min_{\underline{x} \in X} L(\underline{x}, \underline{\tilde{u}})$

Property 2: For $\underline{\tilde{u}} \in \mathbb{R}^m_+$ let $X(\underline{\tilde{u}}) = \left\{ \underline{x} \in X \; : \; f(\underline{x}) + \underline{\tilde{u}}^t \underline{g}(\underline{x}) = w(\underline{\tilde{u}}) \right\}$ then $\boxed{\underline{g}(\underline{x})}$ is a subgradient of $w(\underline{u})$ at $\underline{\tilde{u}}$ for each $\underline{x} \in X(\underline{\tilde{u}})$.

so it's extremely easy (another good news) getting subgradients

Proof*:

Observations:

- Every subgradient of $w(\underline{u})$ at $\underline{\tilde{u}}$ can be expressed as a convex combination of the subgradients $\underline{g}(\underline{x})$ with $\underline{x} \in X(\underline{\tilde{u}})$.
- If w is continuously differentiable at $\underline{\tilde{u}}$, $X(\underline{\tilde{u}})$ contains a single element $\underline{\tilde{x}}$ and $\underline{g}(\underline{\tilde{x}})$ is the gradient of $w(\underline{u})$ at $\underline{\tilde{u}}$.

Summary

- In general (D) is easier than (P) – even if no saddle point exists.

- <u>If a saddle point exists</u>: we can solve (D) and derive optimal $\underline{x}^*$ of (P) by minimizing $L(\underline{x}, \underline{u}^*)$ over X, ensuring $g_i(\underline{x}^*) \leq 0$ and $u_i^* g_i(\underline{x}^*) = 0 \;\; \forall i \in I$.

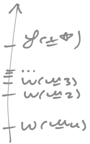
- <u>If no saddle point exists</u>: optimal $\underline{u}^*$ of (D) gives a lower bound $w(\underline{u}^*)$ for $f(\underline{x}^*)$.

  Find $\underline{u}^* \geq \underline{0}$ maximizing $w(\underline{u})$ by using the <u>subgradient method</u> that generates $\{\underline{u}^k\} \rightarrow \underline{u}^*$ when $k \rightarrow \infty$.

  For each $\underline{u}^k$, we have a lower bound $w(\underline{u}^k)$ for $f(\underline{x}^*)$ and we determine $\underline{x}^k$ that minimizes $L(\underline{x}, \underline{u}^k)$ over X.

a sequence of lower bounds

a sequence of problems which are easier to solve



5.5 Second order optimality conditions

Nonlinear program:

$$(P) \qquad \begin{array}{lll} \min & f(\underline{x}) & \\ s.t. & g_i(\underline{x}) \leq 0 & i \in I = \{1, \ldots, m\} \quad u_i \\ & h_l(\underline{x}) = 0 & l \in L = \{1, \ldots, k\} \quad v_l \\ & \underline{x} \in X \subseteq \mathbb{R}^n & \end{array}$$

with f, g_i's and h_l's of class $\mathcal{C}^2$ and X open subset of $\mathbb{R}^n$.

Lagrange function:

$$L(\underline{x}, \underline{u}, \underline{v}) = f(\underline{x}) + \sum_{i=1}^{m} u_i g_i(\underline{x}) + \sum_{l=1}^{k} v_l h_l(\underline{x}) = \boxed{f(\underline{x}) + \underline{u}^t \underline{g}(\underline{x}) + \underline{v}^t \underline{h}(\underline{x})}$$

with $\underline{u} \geq \underline{0}$ and $\underline{v} \in \mathbb{R}^k$.

Hessian submatrix w.r.t. the variables x_j:

$$\underbrace{\nabla^2_{\underline{x}\underline{x}}}_{\text{wrt } \underline{x}} L(\underline{x}, \underline{u}, \underline{v}) = \left(\nabla^2 f(\underline{x})\right) + \left(\sum_{i=1}^{m} u_i \nabla^2 g_i(\underline{x})\right) + \left(\sum_{l=1}^{k} v_l \nabla^2 h_l(\underline{x})\right)$$

Second order KKT necessary conditions:

If $\underline{\overline{x}}$ is a local minimum of (P) and $\nabla g_i(\underline{\overline{x}})$ (at $\underline{\overline{x}}$), with $i \in I(\underline{\overline{x}})$, and $\nabla h_l(\underline{\overline{x}})$ (at $\underline{\overline{x}}$), with $l \in L$, are linearly independent, then $\underline{\overline{x}}$ and some $(\underline{\overline{u}}, \underline{\overline{v}})$ satisfy the KKT conditions:

for the CQ [illegible]

taking the ∇_x of the Lagrangian function and setting it $= \underline{0}$

$$\nabla_{\underline{x}} L(\underline{x}, \underline{u}, \underline{v}) = \nabla f(\underline{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\underline{x}) + \sum_{l=1}^{k} v_l \nabla h_l(\underline{x}) = \underline{0}$$

$$\begin{aligned} g_i(\underline{x}) &\leq 0 && i \in I = \{1, \dots, m\} \\ h_l(\underline{x}) &= 0 && l \in L = \{1, \dots, k\} \\ u_i g_i(\underline{x}) &= 0 && i \in I \\ \underline{u} \geq \underline{0}, \ \underline{v} &\in \mathbb{R}^k. \end{aligned}$$

the 2nd order part

Moreover, every $\underline{d} \in \mathbb{R}^n$ such that

$$\begin{aligned} \nabla^t g_i(\underline{\overline{x}}) \underline{d} &\leq 0 && i \in I(\underline{\overline{x}}) \\ \nabla^t h_l(\underline{\overline{x}}) \underline{d} &= 0 && l \in L \end{aligned}$$

must satisfy

$$\underline{d}^t \nabla^2_{\underline{xx}} L(\underline{\overline{x}}, \underline{\overline{u}}, \underline{\overline{v}}) \underline{d} \geq 0$$

for those $\underline{d}$, the Hessian matrix $\nabla^2_{\underline{xx}} L(\cdot)$ is pos semi-def

Second order KKT sufficient conditions:

Let $\underline{\overline{x}}$ satisfies with $(\underline{\overline{u}}, \underline{\overline{v}})$ the previous KKT conditions.

If

$$\underline{d}^t \nabla^2_{\underline{xx}} L(\underline{\overline{x}}, \underline{\overline{u}}, \underline{\overline{v}}) \underline{d} > 0$$

for each $\underline{d} \neq \underline{0}$ such that

$$\begin{array}{ll} \nabla^t g_i(\underline{\overline{x}}) \underline{d} = 0 & i \in I^+ \\ \nabla^t g_i(\underline{\overline{x}}) \underline{d} \leq 0 & i \in I^0 \\ \nabla^t h_l(\underline{\overline{x}}) \underline{d} = 0 & l = 1, \ldots, k \end{array}$$

makes sense since if $u_i > 0$ then $g_i(\underline{x}) = 0$ by the complementary slackness conditions

where $I^+ = \{i \in I \ : \ u_i > 0\}$ and $I^0 = \{i \in I \ : \ u_i = 0\}$,

then $\underline{\overline{x}}$ is a strict $\boxed{\text{local}}$ minimum of (P).

local as these conditions apply to any problem, also the non-convex ones

See Chap. 12 of J. Nocedal and S. Wright, Numerical Optimization, Springer 1999

5.6 Quadratic programming (QP, estende LP non lineare programmazione)

Optimize a quadratic function subject to linear constraints:

$$(P) \qquad \begin{array}{lll} \min & \boxed{\frac{1}{2}\, \underline{x}^t Q \underline{x} + \underline{c}^t \underline{x}} & \\ s.t. & \underline{a}_i^t \underline{x} \leq b_i & i \in I \\ & \underline{a}_i^t \underline{x} = b_i & i \in E \\ & \underline{x} \in \mathbb{R}^n, & \end{array}$$

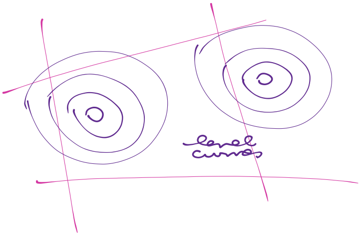
where $Q \in \mathbb{R}^{n \times n}$.

Without loss of generality: Q is symmetric (same function value with $\overline{Q}$ not symmetric and $Q = \frac{1}{2}(\overline{Q} + \overline{Q}^t)$).

Difficulty depends on Q: if Q positive (semi)definite, (P) convex, otherwise can have a large number of local optima.

Example: $\min\{-\overbrace{\underline{x}^t \underline{x}}^{\|\underline{x}\|} : \underbrace{-1 \leq x_i \leq 1, \ i = 1, \dots, n}_{\text{ipercubo in } \mathbb{R}^n}\}$ where all 2^n vertices of $\{-1, 1\}^n$ are local minima.

Illustrations of convex Quadratic Programs (QPs):



optimum points could be on the boundary or also in the middle

$\Rightarrow$ more complex to solve QPs

QPs are the simplest NLP problems besides Linear Programs. Efficient QP algorithms are available.

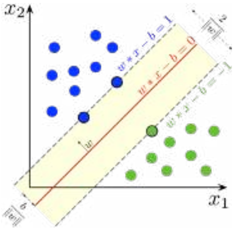
Many direct applications (for portfolio optimization see exercise 9.1).

Example: Training linear Support Vector Machines (SVMs)

Training set $T = \{(\underline{x}^i, y^i) : \underline{x}^i \in \mathbb{R}^n,\ y^i \in \{-1, 1\},\ i = 1, \ldots, p\}$.

Linear decision function: $f(\underline{w}, b, \underline{x}) = \underline{w}^t \underline{x} - b$.

Separating hyperplane with largest margin (width $\frac{2}{\|\underline{w}\|}$) guarantees best generalization.



Hard-margin linear SVM training:

$$\begin{array}{ll} \min\limits_{\underline{w} \in \mathbb{R}^n,\, b \in \mathbb{R}} & \frac{1}{2} \|\underline{w}\|^2 \quad \Rightarrow \text{quadratic problem} \\ \text{s.t.} & y^i(\underline{w}^t \underline{x}^i - b) - 1 \geq 0 \quad i = 1, \ldots, p. \end{array}$$

strictly convex function but possibly huge number of linear constraints.

so the solution would be to move to the dual (to have only one constr)

Reformulated as QP with a single constraint using duality:

$$L(\underline{w}, b, \underline{u}) = \frac{1}{2}\|\underline{w}\|^2 - \sum_{i=1}^{p} u_i(y^i(\underline{w}^t \underline{x}^i - b) - 1)$$

$f(\underline{x}) - \sum_{i \in I} u_i g_i(\underline{x})$ where the only variable now is $\underline{x} = (\underline{w}, b)$

...

Dual: we look for a saddle point, as we saw, so we get

$$\max_{\underline{u} \geq 0} \left(\min_{\underline{w}, b} L(\underline{w}, b, \underline{u}) \right)$$

$$\Longleftrightarrow \quad \max_{\underline{u} \geq 0} L(\underline{w}, b, \underline{u})$$

$$\text{st} \quad \begin{cases} \nabla_{\underline{w}} L(\underline{w}, b, \underline{u}) = \underline{0} \\ \frac{\partial}{\partial b} L(\underline{w}, b, \underline{u}) = 0 \end{cases}$$

we set the gradient $= 0$ so to get a min

$$\underline{w} = \sum_{i=1}^{p} u_i y_i \underline{x}_i$$

$$\sum_{i=1}^{p} u_i y_i = 0$$

which is a simpler case, and being a convex example there was perfect correspondence of solving primal or dual very easy to solve

5.6.1 QP with only equality constraints

Consider

$a_i^T x = b_i \quad \forall i \in E$

$$\min\{\ \frac{1}{2}\ \underline{x}^t Q \underline{x} + \underline{c}^t \underline{x}\ \ :\ \ A\underline{x} = \underline{b}\ \} \qquad (1)$$

this defines a subspace of $\mathbb{R}^n$
to solve these problems
(by construction) is easy

where $A \in \mathbb{R}^{m \times n}$.

Since only linear equations, CQ assumption is satisfied at every feasible point and simple KKT conditions:

gradient of the obj function

gradient of the constr (all of them are active)

$$\left\{\begin{array}{l} Q\underline{x} + \underline{c} + \sum_{i=1}^{m} u_i \underline{a}_i = \underline{0} \\ A\underline{x} = \underline{b}. \end{array}\right.$$

$u_i \in \mathbb{R}$

N.B.: Complementary slackness constraints are automatically satisfied.

More or less direct solution of the linear system:

$$\left(\begin{array}{cc} Q & A^t \\ A & 0 \end{array}\right)\left(\begin{array}{c} \underline{x} \\ \underline{u} \end{array}\right) = \left(\begin{array}{c} -\underline{c} \\ \underline{b} \end{array}\right).$$

If A of full rank and Q is p.d. on subspace $\{\underline{x} \in \mathbb{R}^n\ :\ A\underline{x} = \underline{0}\}$, matrix is non singular.

Null-space method

Determine $Z \in \mathbb{R}^{n \times (n-m)}$ whose columns span the $\overbrace{\text{null space}}^{\ker(A)}$ $\{\underline{x} \in \mathbb{R}^n \ : \ A\underline{x} = \underline{0}\}$ of A.

Z can be computed by (sub)matrix factorization of A (if A sparse by LU factorization).

Given feasible $\underline{x}_0$, any other feasible solution

$$\boxed{\underline{x} = \underline{x}_0 + Z\underline{w}}$$

i.e. we perform a change of variables

for an appropriate $\underline{w} \in \mathbb{R}^{n-m}$.

(1) is equivalent to $\boxed{\text{unconstrained}}$ QP:

$$\frac{1}{2}(\underline{x}_0 + Z\underline{w})^T Q (\underline{x}_0 + Z\underline{w}) + \underline{c}^T(\underline{x}_0 + Z\underline{w})$$

$$\min_{\underline{w} \in \mathbb{R}^{n-m}} \left[\frac{1}{2} \underline{w}^T (Z^T Q Z) \underline{w} + (Q\underline{x}_0 + \underline{c})^T Z \underline{w} \right]$$

if Z^TQZ is pos def $\Rightarrow$ $\exists!$ optimal sol $\underline{w}^*$ obtained by solving

∇(that obj function) $= \underline{0}$ that is

$$(Z^TQZ)\underline{w} = -Z^T(Q\underline{x}_0 + \underline{c})$$

Also other methods but null-space ones are widely used.

5.6.2 QP with equality and inequality constraints

Active-set methods

$$(P) \qquad \begin{array}{lll} \min & q(\underline{x}) = \frac{1}{2}\ \underline{x}^t Q \underline{x} + \underline{c}^t \underline{x} & \\ s.t. & \underline{a}_i^t \underline{x} \leq b_i & i \in I \\ & \underline{a}_i^t \underline{x} = b_i & i \in E \\ & \underline{x} \in \mathbb{R}^n & \end{array}$$

where $Q \in \mathbb{R}^{n \times n}$.

the set of active constr around x^*

Idea: Determine $I(\underline{x}^*) = \{i \in I : \underline{a}_i^t \underline{x}^* = b_i\}$ where $\underline{x}^*$ is an optimal solution, by solving a sequence of QPs with only equality constraints.

idea: if we knew which of the inequality constr are active at the opt set, then we would come back to the previous case of only eq constr

Active-set method for convex QPs

eg through the first part of simplex methods for LPs

Initialization: Find initial feasible $\underline{x}_0$ and

choose $W_0 \subseteq \underbrace{\{i \in I : \underline{a}_i^t \underline{x}_0 = b_i\}}_{I(\underline{x}_0)} \cup E$ of the active constraints at $\underline{x}_0$, with $E \subseteq W_0$.

⇒ W_0 is like the starting working set of constr on which we optimize

Iteration k:

Given current feasible $\underline{x}_k$, determine $\underline{d}_k$ by solving the subproblem:

only consider constr according to W_k

$$\min\{\ q(\underline{x}_k + \underline{d}) : \underline{a}_i^t(\underline{x}_k + \underline{d}) = b_i,\ i \in W_k\ \}, \qquad (2)$$

$\underline{x}_k + \underline{d}$ must satisfy at eq all the W_k constr (and only them)

where W_k is current *working set*, with $W_k \subseteq \underbrace{\{i \in I : \underline{a}_i^t \underline{x}_k = b_i\}}_{\text{a choice (of indexes) among the ineq constr}} \cup E$.

plus the full index set of eq constr

(2) is equivalent to:

$a^T_i x_k = b_i$ already by construction

$$\min\{\ q(\underline{x}_k + \underline{d}) : \underline{a}_i^t \underline{d} = 0,\ i \in W_k\ \}. \qquad (3)$$

so we are just left with these

N.B.: If $Z^t Q Z$ is p.d. (always true if Q is p.d.), (3) has a unique solution $\underline{d}_k$.

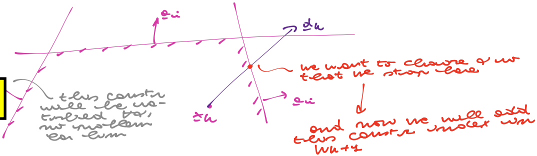
Based on solution $\underline{d}_k$ of (3), we determine α_k, $\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k$ and W_{k+1}.

the ones which for now we left apart

- If $\boxed{\underline{d}_k \neq \underline{0}}$, we determine the largest α satisfying all constraints not in W_k:

$$\alpha_k = \min \left\{ 1, \min_{\substack{i \notin W_k: \\ \underline{a}_i^T \underline{d}_k > 0}} \left(\frac{b_i - \underline{a}_i^T \underline{x}_k}{\underline{a}_i^T \underline{x}_k} \right) \right\}$$

we remove the ones that are ≤ 0 (we already satisfied)



and set $\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k$.

$W_{k+1} = W_k \cup \{i'\}$ where i' is index of one constraint becoming active at $\underline{x}_{k+1}$.

- If $\boxed{\underline{d}_k = \underline{0}}$, $\underline{x}_k$ is a minimum over subspace defined by W_k and we set $\underline{x}_{k+1} = \underline{x}_k$.

but we have to check its optimality, we are not done yet

KKT conditions of (3) imply there are multipliers u_i^k such that:

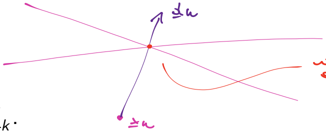
$$(Q\underline{x}_k + \underline{c}) + \sum_{i \in W_k} u_i^k \underline{a}_i = \underline{0}. \tag{4}$$

`If` $u_i^k \geq 0$ for every $i \in W_k \cap I$ `Then` $\underline{x}_k$ is a local optimum of original QP

`Else` $W_{k+1} = W_k \setminus \{i'\}$ where i' is the index with the most negative $u_{i'}^k$.

we if $\exists i' : u_{i'}^k < 0$

- If $\underline{d}_k \neq \underline{0}$, we determine the largest α satisfying all constraints not in W_k:



if more constr become active (at equality) we just add one of them to W_{k+1} (not both)

and set $\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k$.

$W_{k+1} = W_k \cup \{i'\}$ where i' is index of one constraint becoming active at $\underline{x}_{k+1}$.

- If $\underline{d}_k = \underline{0}$, $\underline{x}_k$ is a minimum over subspace defined by W_k and we set $\underline{x}_{k+1} = \underline{x}_k$.

KKT conditions of (3) imply there are multipliers u_i^k such that:

$$Q\underline{x}_k + \underline{c} + \sum_{i \in W_k} u_i^k \underline{a}_i = \underline{0}. \tag{4}$$

`If` $u_i^k \geq 0$ for every $i \in W_k \cap I$ `Then` $\underline{x}_k$ is a local optimum of original QP

`Else` $W_{k+1} = W_k \setminus \{i'\}$ where i' is the index with the most negative $u_{i'}^k$.

Proposition: If Q is p.d. (q is strictly convex), the method (with anti-cycling rule) finds an optimal solution within a finite number of iterations.

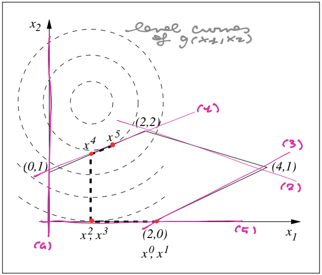
Works because of:

Note: finite number of working sets. and of changes for the Wk

Example:

$$\begin{array}{lll} \min & q(x_1, x_2) = (x_1 - 1)^2 + (x_2 - 2.5)^2 & \\ \text{s.t.} & -x_1 + 2x_2 - 2 \leq 0 & (1) \\ & x_1 + 2x_2 - 6 \leq 0 & (2) \\ & x_1 - 2x_2 - 2 \leq 0 & (3) \\ & -x_1 \leq 0 & (4) \\ & -x_2 \leq 0 & (5) \end{array}$$

Figure:



From J. Nocedal, S. Wright, Numerical Optimization, First Edition, Springer 1999, p. 462-463.

Iteration 0:

$\underline{x}_0 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$ and we take $W_0 = \{3, 5\}$.

since we took x_0 on the intersection of the two constr. and not just a "random inside" feasible set

Since $\underline{x}_0$ is a vertex of the feasible solution polyhedron,

$\underline{x}_0$ minimizes $q(\underline{x})$ w.r.t. W_0 and

$\underline{d}_0 = \underline{0}$ is optimal solution of $\min\{\ q(\underline{x}_0 + \underline{d}) : \underline{a}_i^t \underline{d} = 0,\ i \in W_0\ \}$.

Thus $\underline{x}_1 = \underline{x}_0 + \alpha_0 \underline{d}_0 = \underline{x}_0$.

KKT conditions:

$$\nabla q(\underline{x}_0) = \begin{pmatrix} 2 \\ -5 \end{pmatrix} = u_3 \begin{pmatrix} -1 \\ 2 \end{pmatrix} + u_5 \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

we obtain the multipliers $\begin{pmatrix} u_3 \\ u_5 \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \end{pmatrix}$ for the active constraints.

Since $u_3 < u_5 < 0$, we set $W_1 = W_0 \setminus \{3\} = \{5\}$.

so now we are just continuing moving on the x_1-axis

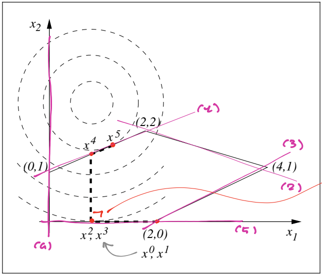
Proposition: If Q is p.d. (q is strictly convex), the method (with anti-cycling rule) finds an optimal solution within a finite number of iterations.

Works because of:

Note: finite number of working sets. and of changes for the W_k

Example:
$$\begin{array}{lll} \min & q(x_1, x_2) = (x_1 - 1)^2 + (x_2 - 2.5)^2 & \\ \text{s.t.} & -x_1 + 2x_2 - 2 \le 0 & (1) \\ & x_1 + 2x_2 - 6 \le 0 & (2) \\ & x_1 - 2x_2 - 2 \le 0 & (3) \\ & -x_1 \le 0 & (4) \\ & -x_2 \le 0 & (5) \end{array}$$

Figure:



x2 becomes optimal as wt' and (is closest also) to the center of the level curves

From J. Nocedal, S. Wright, Numerical Optimization, First Edition, Springer 1999, p. 462-463.

Iteration 1:

Optimal solution of $\min\{\ q(\underline{x}_1 + \underline{d}) : \underline{a}_i^t \underline{d} = 0,\ i \in W_1\ \}$ is $\underline{d}_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$.

Since $\underline{d}_1$ does not violate any constraint with indices not in W_1, $\alpha_1 = 1$ and $\underline{x}_2 = \underline{x}_1 + \alpha_1 \underline{d}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

Since at $\underline{x}_2$ no other constraints are active, we set $W_2 = W_1 = \{5\}$.

Iteration 2:

Optimal solution of $\min\{\ q(\underline{x}_2 + \underline{d}) : \underline{a}_i^t \underline{d} = 0,\ i \in W_2\ \}$ is $\underline{d}_2 = \underline{0}$.

From KKT conditions

$$\nabla q(\underline{x}_2) = \begin{pmatrix} 0 \\ -5 \end{pmatrix} = u_5 \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

we obtain $u_5 = -5$.

Thus $\underline{x}_3 = \underline{x}_2$ and we set $W_3 = W_2 \setminus \{5\} = \emptyset$. so we get to optimize freely, and we moved more vertically towards the level curves center

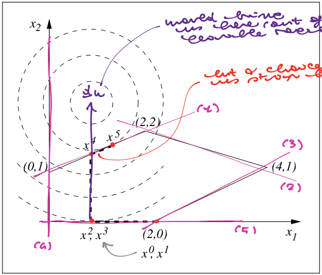
Proposition: If Q is p.d. (q is strictly convex), the method (with anti-cycling rule) finds an optimal solution within a finite number of iterations.

Works because of:

Note: finite number of working sets. and of changes for the W_k

Example:
$$
\begin{array}{lll}
\min & q(x_1, x_2) = (x_1 - 1)^2 + (x_2 - 2.5)^2 & \\
\text{s.t.} & -x_1 + 2x_2 - 2 \leq 0 & (1) \\
& x_1 + 2x_2 - 6 \leq 0 & (2) \\
& x_1 - 2x_2 - 2 \leq 0 & (3) \\
& -x_1 \leq 0 & (4) \\
& -x_2 \leq 0 & (5)
\end{array}
$$

Figure:



From J. Nocedal, S. Wright, Numerical Optimization, First Edition, Springer 1999, p. 462-463.

Iteration 3:

Optimal solution of $\min\{\ q(\underline{x}_3+\underline{d}) : \underline{a}_i^t\underline{d}=0,\ i\in W_3\ \}$ is $\underline{d}_3=\begin{pmatrix} 0 \\ 2.5 \end{pmatrix}$.

Since $\underline{d}_3$ violates constraints (1) and (2) which are not in W_1, $\alpha_3=0.6$ and $\underline{x}_4=\underline{x}_3+\alpha_3\underline{d}_3=\begin{pmatrix} 1 \\ 1.5 \end{pmatrix}$.

Since at $\underline{x}_4$ only constraint (1) becomes active, we set $W_4=\{1\}$.

Iteration 4:

Optimal solution of $\min\{\ q(\underline{x}_4+\underline{d}) : \underline{a}_i^t\underline{d}=0,\ i\in W_4\ \}$ is $\underline{d}_4=\begin{pmatrix} 0.4 \\ 0.2 \end{pmatrix}$.

Since $\underline{x}_4+\underline{d}_4=\begin{pmatrix} 1.4 \\ 1.7 \end{pmatrix}$ satisfies all the constraints with indices not in W_1, we take $\alpha_4=1$, set $\underline{x}_5=\underline{x}_4+\underline{d}_4$ and $W_5=W_4=\{1\}$.

Iteration 5:

Optimal solution of $\min\{\ q(\underline{x}_5 + \underline{d}) : \underline{a}_i^t \underline{d} = 0,\ i \in W_5\ \}$ is $\underline{d}_5 = 0$.

Solving the KKT conditions (4) we obtain $u_1 = 1.25 \geq 0$.

Thus $\underline{x}_5 = \begin{pmatrix} 1.4 \\ 1.7 \end{pmatrix}$ is optimal for the original problem.

5.6.3 Non convex QP and solvers

If Q has some negative eigenvalues, the active-set method for convex QP can be adapted by modifying $\underline{d}_k$ and α_k in certain situations.

See J. Nocedal, S. Wright, Numerical Optimization, First edition, Springer 1999, p. 468-474.

Since W_k may change by just one index at every iteration, efficient QP solvers proceed by successive updates of the factors computed at the previous iterations.

Available active-set-based solvers: LINDO, QPOPT, NAG Library, Matlab,...

5.7 Penalty method and augmented Lagrangian method

Generic NLP:

and changed the function name

now ≥ 0 for uniformity (before was ≤ 0)

$$
\begin{array}{lll}
\min & f(\underline{x}) & \\
s.t. & c_i(\underline{x}) \geq 0 & i \in I \\
 & c_i(\underline{x}) = 0 & i \in E \\
 & \underline{x} \in \mathbb{R}^n &
\end{array}
\tag{1}
$$

where f and c_i's are of class $\mathcal{C}^1$ or $\mathcal{C}^2$.

Notation, examples and proofs: see Chapter 17 of J. Nocedal, S. Wright, Numerical Optimization, Springer, 1999, p. 491-500.

5.7.1 Quadratic penalty method

to reduce to unconstrained opt

in the obj function

Idea: Delete constraints, penalize their violation and solve a sequence of unconstrained optimization problems.

Description for

$$\boxed{\begin{array}{lll} \min & f(\underline{x}) & \\ s.t. & c_i(\underline{x}) = 0 & i \in E = \{1, \dots, m\} \\ & \underline{x} \in \mathbb{R}^n. & \end{array}} \quad (2)$$

Definition: The **quadratic penalty function** problem associated to (2) is

$$\boxed{\min_{\underline{x} \in \mathbb{R}^n} Q(\underline{x}, \mu) = f(\underline{x}) + \frac{1}{2\mu} \sum_{i \in E} c_i^2(\underline{x})} \quad (3)$$

with penalty parameter $\mu > 0$..

to weight the penalization

the square penalizes more the larger violations (violations since in theory $c_i(\underline{x}) = 0$, here eq constr)

We consider $\{\mu_k\}_{k \geq 1}$ with $\lim_{k \to \infty} \mu_k = 0$ and, for each k, we determine an approximate solution $\underline{x}_k$ of (3) using an unconstrained optimization method.

in that the penalization goes to ∞ (we eventually we "erase" the constraints)

Example:

$$\min \quad x_1 + x_2$$
$$s.t. \quad x_1^2 + x_2^2 - 2 = 0$$

with optimal solution $(-1, -1)^t$.

Quadratic penalty problem:

$$\min_{x \in \mathbb{R}^2} Q(x, \mu) = f(x) + \frac{1}{2\mu} \sum_{i \in E} c_i(x)^2 =$$
$$= x_1 + x_2 + \frac{1}{2\mu}(x_1^2 + x_2^2 - 2)$$

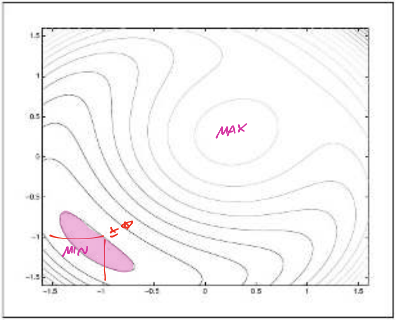


Figure 17.1 Contours of $Q(x;\mu)$ from (17.4) for $\mu = 1$

for $\mu = 1$ the minimizer of $Q(x, \mu)$ is close to $x^* = \binom{-1}{-1}$

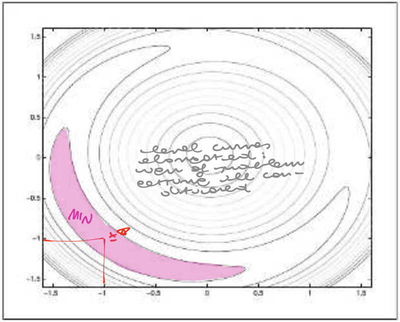


Figure 17.2 Contours of $Q(x;\mu)$ from (17.4) for $\mu = 10$

while here we are farther away from x^*

General scheme

we want to be more and more accurate

0) Select $\varepsilon > 0$, $\mu_0 > 0$, sequence of tolerances $\{\tau_k\}_{k\geq 0}$ with $\tau_k > 0$ and $\lim_{k\to\infty} \tau_k = 0$.
Choose initial $\underline{x}_0^s$ and set $k = 0$.
→ for starting pt

1) Determine an approximate minimizer $\underline{x}_k$ of $Q(\underline{x}, \mu_k)$ starting from $\underline{x}_k^s$ and terminate when $\|\nabla Q(\underline{x}, \mu_k)\| \leq \tau_k$. → we when the required accuracy is reached

terminate this step (2)

2) `If` termination condition is satisfied (e.g, $|f(\underline{x}_{k-1}) - f(\underline{x}_k)| < \varepsilon$)
`Then` return solution $\underline{x}_k$
`Else` choose $\mu_{k+1} \in (0, \mu_k)$ and starting $\underline{x}_{k+1}^s$, set $k = k + 1$ and `Goto 1)`

decrease μ_k we increase penalization
$\dots + \frac{1}{2\mu}\sum$) ⇒ small μ larger weight

Choices:

- For convergence results, it suffices that $\lim_{k\to\infty} \tau_k = 0$.
- $\{\mu_k\}_{k\geq 0}$ generated adaptively starting from μ_0: if minimization of $Q(\underline{x}, \mu_k)$ is "difficult" set e.g. $\mu_{k+1} = 0.7\mu_k$, otherwise $\mu_{k+1} = 0.1\mu_k$. smaller/slower decrease if problem is difficult
- Judicious choice of the starting $\underline{x}_k^s$ when solving unconstrained penalty problem at each iteration: $\underline{x}_{k+1}^s := \underline{x}_k$. → the pt that we get solving the kth unconstrained problem

Convergence

unrealistic, we are assuming to solve at optimality each subproblem ($\tau_k = 0 \ \forall k$)

interesting but unrealistic

Theorem 1: Suppose each $\underline{x}_k$ is a *global minimizer* of $Q(\underline{x}, \mu_k)$ and $\lim_{k\to\infty} \mu_k = 0$, then every limit point $\underline{x}^*$ of $\{\underline{x}_k\}_{k\geq 0}$ generated with above scheme ($\tau_k = 0$, $\forall k \geq 0$) is a *global minimum* of problem (2).

then forces to terminate if $\|\nabla Q(x_k, \mu_k)\| = 0$

Proof:

Let $\overline{x}$ be an optimal solution of (2).

Since x_k is a global min of $Q(x, \mu_k)$ and $\bar{x}$ is feasible for (2, the original problem) then

$$\underbrace{Q(x_k, \mu_k)}_{\text{global min}} \leq \underbrace{Q(\bar{x}, \mu_k)}_{\text{just a feasible pt (for } Q)} \quad \forall k$$

since $\bar{x}$ is feasible and opt for (2)

namely:

$$\underbrace{f(x_k)}_{\to f(x^*)} + \frac{1}{2\mu_k} \sum_{i=1}^{m} c_i(x_k)^2 \leq f(\bar{x}) + \frac{1}{2\mu_k} \underbrace{\sum_{i=1}^{m} c_i(\bar{x})^2}_{0} = f(\bar{x}) \quad (4)$$

$$\iff \sum_{i=1}^{m} c_i(x_k)^2 \leq 2\underbrace{\mu_k}_{\to 0}\left[f(\bar{x}) - f(x_k) \right] \quad \forall k \quad (5)$$

Consider any sub-sequence of $\{x_k\}_{k\geq 0}$ with $k \in K$: $\lim_{k \in K} x_k = x^*$

By assuming $k \to +\infty$ with $k \in K$ we obtain from (5), by taking the limit:

$$\sum_{i=1}^{m} c_i(x^*)^2 = \lim_{k \in K} \sum_{i=1}^{m} c_i(x_k)^2 \leq 0 \Rightarrow c_i(x^*) = 0 \ \forall i \in E$$

(the limit point is feasible for (2))

now we know that $\underline{x}^*$ is not only feasible but also the optimal sol.

B) assume $k \to +\infty$ with $k \in \mathcal{K}$ in (a) we obtain

$$\underline{f(\underline{x}^*)} = f(\underline{x}^*) + \lim_{k \in \mathcal{K}} \frac{1}{2\mu_k} \sum_{i=1}^{m} c_i(\underline{x}_k)^2 = \underline{f(\bar{\underline{x}})}$$

$\Rightarrow$ $\underline{x}^*$ is an opt sol of (2)

we need to solve at opt the unconstrained problems

more restrictive

we demand more and more precision

0 so we would try to drive them to zero

$u_i^* > 0$ and $= 0$ we are not exact

$\Rightarrow -c_i(\underline{x}_k) \approx \mu_k u_i^*$

and so we can tune the μ_k adaptively more precisely the method will diverge, according to u_i^*. This will be the idea of the augmented Lagrangian method

Theorem 2: If

- tolerances $\tau_k > 0$ satisfy $\lim_{k\to\infty} \tau_k = 0$
- $\lim_{k\to\infty} \mu_k = 0$,

then every limit point $\underline{x}^*$ of $\{\underline{x}_k\}_{k\geq 0}$ at which all $\nabla c_i(\underline{x}^*)$, with $i \in E$, are linearly independent is a *KKT point* of problem (2).

For such points, the subsequence defined by $\mathcal{K}$ with $\lim_{k\in\mathcal{K}} \underline{x}_k = \underline{x}^*$ satisfies

$$\boxed{\lim_{k\in\mathcal{K}} -\frac{c_i(\underline{x}_k)}{\mu_k} = u_i^* \qquad \forall i \in E,} \tag{4}$$

where $\underline{u}^*$ satisfies with $\underline{x}^*$ the KKT conditions for problem (2).

the corresponding optimal KKT multipliers

Observation: (4) implies that

i) The minimizer $\underline{x}_k$ of $Q(\underline{x}, \mu_k)$ does not satisfy $c_i(\underline{x}) = 0$ exactly for all $i \in E$ ($c_i(\underline{x}_k) = -\mu_k u_i^*$). To obtain a feasible solution, we must $\mu_k \to 0$.

ii) In some circumstances $-\frac{c_i(\underline{x}_k)}{\mu_k}$ may be used as estimates of u_i^*.

Recall: Lagrange function for problem (2) is

sign irrelevant due to equality constr

$$L(\underline{x}, \underline{u}) = f(\underline{x}) - \sum_{i=1}^{m} u_i c_i(\underline{x}) \tag{5}$$

and KKT conditions require that, apart from $c_i(\underline{x}) = 0$

no complementary slackness since we have equality constr

$$\nabla_{\underline{x}} L(\underline{x}, \underline{u}) = \nabla f(\underline{x}) - \sum_{i=1}^{m} u_i \nabla c_i(\underline{x}) = \underline{0}. \tag{6}$$

By comparing

$Q(\underline{x}, \mu) = f(\underline{x}) + \frac{1}{2\mu} \sum c_i(\underline{x})^2$

$$\nabla_{\underline{x}} Q(\underline{x}, \mu) = \nabla f(\underline{x}) + \frac{1}{\mu} \sum_{i=1}^{m} c_i(\underline{x}) \nabla c_i(\underline{x}) = \underline{0} \tag{7}$$

and (6), it appears that $-\frac{c_i(\underline{x})}{\mu}$ has been substituted with u_i.

It can be proved that if $\tau_k \to 0$ then $\underline{x}_k \to \underline{x}^*$ and $-\frac{c_i(\underline{x}_k)}{\mu_k} \to u_i^*$ $\qquad i = 1, 2, \ldots, m.$

that's the main problem of this method

Observation: When $\mu_k \to 0$ the quadratic penalty problem (3) becomes ill conditioned.

$\min_{\underline{x} \in \mathbb{R}^n} Q(\underline{x}, \mu_k)$

we see this looking at the hessian

$$\nabla^2_{\underline{xx}} Q(\underline{x}, \mu_k) = \nabla^2 f(\underline{x}) + \frac{1}{\mu_k} A^t(\underline{x}) \; A(\underline{x}) + \frac{1}{\mu_k} \sum_{i=1}^{m} c_i(\underline{x}) \nabla^2 c_i(\underline{x}) \tag{8}$$

where $A^t(\underline{x}) = [\nabla c_1(\underline{x}); \ldots; \nabla c_m(\underline{x})]$ and $A \in \mathbb{R}^{m \times n}$ of full rank $m \leq n$, usually $m < n$.

When $\underline{x}$ is close to minimizer of $Q(\underline{x}, \mu_k)$ and assumptions of Theorem 2 are satisfied, (4) implies that

optimal KKT mult

$$\nabla^2_{\underline{xx}} Q(\underline{x}, \mu_k) \approx \nabla^2_{\underline{xx}} L(\underline{x}, \underline{u}^*) + \frac{1}{\mu_k} A^t(\underline{x}) \; A(\underline{x}). \tag{9}$$

Since $\nabla^2_{\underline{xx}} L(\underline{x}, \underline{u}^*)$ does not depend on μ_k and $\frac{1}{\mu_k} A^t(\underline{x}) \; A(\underline{x})$ has $n - m$ eigenvalues of value 0 and m eigenvalues of value $O(1/\mu_k)$, numerical issues arise when $\mu_k \to 0$.

For convenience, this approach can also be extended for inequality constr

Problems with both equality and inequality constraints:

Quadratic penalty problem

$$\min_{\underline{x}\in\mathbb{R}^n} \; Q(\underline{x},\mu) = f(\underline{x}) + \frac{1}{2\mu}\sum_{i\in E} c_i^2(\underline{x}) + \frac{1}{2\mu}\sum_{i\in I}\underbrace{([c_i(\underline{x})]^-)^2}_{= \begin{cases} c_i(\underline{x}) & \text{if } c_i(\underline{x}) \geq 0 \\ 0 & \text{otherwise} \end{cases}} \qquad (10)$$

where $[y]^-$ denotes $\max(-y, 0)$.

also here we may get differentiability problems maybe

Other penalty functions are available.

If only equality constraints, the *exact penalty problem* is

$$\min_{\underline{x}\in\mathbb{R}^n} \; Q(\underline{x},\mu) = f(\underline{x}) + \frac{1}{2\mu}\sum_{i\in E} |c_i(\underline{x})|. \qquad (11)$$

N.B.: Q is not everywhere differentiable.

due to the $|\cdot|$

5.7.2 Augmented Lagrangian method

Idea: Reduce ill-conditioning issues of the unconstrained subproblems (in quadratic penalty method) by introducing explicit estimates of the Lagrange multipliers.

Description for

$$\boxed{\begin{array}{ll} \min & f(\underline{x}) \\ s.t. & c_i(\underline{x}) = 0 \quad i \in E = \{1, \ldots, m\} \\ & \underline{x} \in \mathbb{R}^n. \end{array}} \qquad (12)$$

none of below

Definition: The **augmented Lagrange function** associated to problem (12) is

we add a quadratic penalty term but to the lagrangian function (rather than to the obj function)

$$L_A(\underline{x}, \underbrace{\underline{\lambda}}_{\text{lag mult}}, \underbrace{\mu}_{\text{penalty param}}) = \underbrace{\left[f(\underline{x}) + \sum_{i=1}^{m} \lambda_i c_i(\underline{x}) \right]}_{L(\underline{x}, \underline{\lambda})} + \frac{1}{2\mu} \left[\sum_{i=1}^{m} c_i(\underline{x})^2 \right]$$

we're indifferent as we have equality constr

Since the KKT cond require that $\nabla_x L(\underline{x}^*, \underline{\lambda}^*) = 0$ and $c_i(\underline{x}^*) = 0$ $\forall i$, then at optimality

$L_A(\underline{x}, \underline{\lambda}, \mu) = L(\underline{x}, \underline{\lambda}) \Rightarrow$ no need for assume $\mu \to 0$

Similar approach:

At each iteration: $\mu_k > 0$ and determine an approximate minimizer $\underline{x}_k$ of $L_A(\underline{x}, \underline{u}^k, \mu_k)$ via an unconstrained optimization method, where $\underline{u}^k$ is an updated estimate.

→ we need to understand how to update u^k

Differentiating w.r.t. $\underline{x}$, we obtain

$$\nabla_{\underline{x}} L_A(\underline{x}, \underline{u}, \mu) = \nabla f(\underline{x}) - \sum_{i=1}^{m} \boxed{\left(u_i - \frac{c_i(\underline{x})}{\mu}\right)} \nabla c_i(\underline{x}).$$

$L_A(\underline{x}, \underline{u}, \mu) = [f(x) + \sum_{i=1}^{m} u_i c_i(x)] + \frac{1}{2\mu}[\sum_{i=1}^{m} c_i(\underline{x})^2]$

$\hookrightarrow \nabla L_A = \nabla f \qquad u_i \nabla c_i(x) \qquad 2 c_i(\underline{x}) \nabla c_i(\underline{x})$

Considerations similar to those in proof of Theorem 2 allow to establish that

$$u_i^* \approx u_i^k - \frac{c_i(\underline{x}_k)}{\mu_k} \qquad i \in E, \tag{13}$$

→ relation that would be obtained asymp-totically)

which is equivalent to

$$c_i(\underline{x}_k) \approx \mu_k \left(u_i^k - u_i^*\right) \qquad i \in E. \tag{14}$$

c_i tends to get towards 0 (we to lea-wewent))

> 0

we can keep μ structly moderate (to avoid ill-posed problems)

and instead we drive these → 0

General scheme

0) Choose $\varepsilon > 0$, $\mu_0 > 0$, tolerances $\{\tau_k\}_{k\geq 0}$ with $\tau_k > 0$ and $\lim_{k\to\infty} \tau_k = 0$, $\underline{x}_0^s$ and initial $\underline{u}^0$, set $k := 0$.

1) Determine an approximate minimizer $\underline{x}_k$ of $L_A(\underline{x}, \underline{u}^k, \mu_k)$ starting from $\underline{x}_k^s$ and terminate when $\|\nabla_{\underline{x}} L_A(\underline{x}, \underline{u}^k, \mu_k)\| \leq \tau_k$.

2) `If` overall termination condition is satisfied (e.g, $|f(\underline{x}_{k-1}) - f(\underline{x}_k)| < \varepsilon$)

`Then` Stop

`Else` set

$$u_i^{k+1} = u_i^k - \frac{c_i(\underline{x}_k)}{\mu_k} \quad \text{for } i \in E \tag{15}$$

choose $\mu_{k+1} \in (0, \mu_k)$ and next starting solution $\underline{x}_{k+1}^s$

set $k := k + 1$ and `Goto 1)`

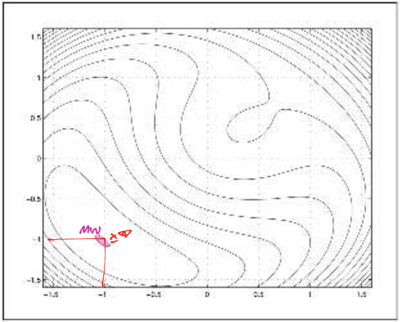
Including in L_A an additional term related to the Lagrange multipliers leads to substantial improvements w.r.t. the quadratic penalty method.

Example:

$$\begin{array}{ll} \min & x_1 + x_2 \\ s.t. & x_1^2 + x_2^2 - 2 = 0 \end{array}$$

with optimal solution $\underline{x}^* = (-1, -1)^t$, optimal multiplier $u^* = -0.5$ and unconstrained optimization subproblem:

From J. Nocedal, S. Wright, Numerical Optimization, Springer, 1999, p. 513-514.



better level curves

Figure 17.5 Contours of $\mathcal{L}_A(x, \lambda; \mu)$ from (17.40) for $\lambda = -0.4$ and $\mu = 1$.

Including in L_A an additional term related to the Lagrange multipliers leads to substantial improvements w.r.t. the quadratic penalty method.

Example:

$$\begin{array}{ll} \min & x_1 + x_2 \\ s.t. & x_1^2 + x_2^2 - 2 = 0 \end{array}$$

with optimal solution $\underline{x}^* = (-1,-1)^t$, optimal multiplier $u^* = -0.5$ and unconstrained optimization subproblem:

From J. Nocedal, S. Wright, Numerical Optimization, Springer, 1999, p. 513-514.

$$\min_{\underline{x} \in \mathbb{R}^2} L_A(\underline{x}, u, \mu) = \left[(x_1 + x_2) - u(x_1^2 + x_2^2 - 2) \right] + \left[\frac{1}{2\mu} (x_1^2 + x_2^2 - 2)^2 \right]$$

Suppose that $\mu_k = 1$ and estimate $u^k = -0.4$.

Contours of $L_A(\underline{x}, -0.4, 1)$ are similar to those of $Q(\underline{x}, 1)$ but the minimizer

$$\underline{x}^k = \begin{pmatrix} -1.02 \\ -1.02 \end{pmatrix} \text{ of } L_A(\cdot)$$

is much closer to $\underline{x}^*$ than the

$$\begin{pmatrix} -1.1 \\ -1.1 \end{pmatrix} \text{ of } Q(\cdot)$$

$\Rightarrow$ performances have improved

Theorem 3:

Let $\underline{x}^*$ be a local minimum of (12) at which the $\nabla c_i(\underline{x}^*)$, $i \in E$, are linearly independent and 2nd order sufficient optimality conditions are satisfied for $\underline{u} = \underline{u}^*$.

Then $\exists\ \overline{\mu} > 0$ such that for all $\mu \in (0, \overline{\mu}]$, $\underline{x}^*$ is a strict local minimum of $L_A(\underline{x}, \underline{u}^*, \mu)$.

N.B.: In general $\underline{u}^*$ is unknown.

The next result

the optimal multiplier

- concerns the more realistic case in which $\underline{u} \neq \underline{u}^*$,
- provides conditions under which $\exists$ a minimizer of L_A close to $\underline{x}^*$ and error bounds on $\underline{x}_k$ and on $\underline{u}^{k+1}$.

Theorem 4:

Suppose the assumptions of Theorem 3 are satisfied at $\underline{x}^*$ and $\underline{u}^*$, and let $\overline{\mu} > 0$ be the corresponding threshold.

Then $\exists$ scalars $\delta > 0$, $\varepsilon > 0$, and M such that

i) For all $\underline{u}^k$ and μ_k satisfying

$$\|\underline{u}^k - \underline{u}^*\| \leq \delta/\mu_k, \qquad \mu_k \leq \overline{\mu}, \tag{16}$$

the problem

$$\min_{\underline{x} \in \mathbb{R}^n \,:\, \|\underline{x} - \underline{x}^*\| \leq \varepsilon} L_A(\underline{x}, \underline{u}^k, \mu_k)$$

=> as well, the problem was well defined

has a unique solution $\underline{x}_k$. Moreover, we have $\|\underline{x}_k - \underline{x}^*\| \leq M\mu_k \|\underline{u}^k - \underline{u}^*\|$.

moreover we have an error bound (which depends on the gradient) of the $\underline{u}^k$ wrt $\underline{u}^*$)

ii) For all $\underline{u}^k$ and μ_k satisfying (16), we have

$$\|\underline{u}^{k+1} - \underline{u}^*\| \leq M\mu_k \|\underline{u}^k - \underline{u}^*\|,$$

where $\underline{u}^{k+1}$ is given by the formula (15). good news about convergence speed

iii) For all $\underline{u}^k$ and μ_k satisfying (16), the matrix $\nabla^2_{\underline{xx}} L_A(\underline{x}_k, \underline{u}^k, \mu_k)$ is positive definite at $\underline{x}_k$ and the $\nabla c_i(\underline{x}_k)$, with $i \in E$, are linearly independent.

we really are well defined

Problems with also inequality constraints:

We can introduce slack variables and substitute $c_i(\underline{x}) \geq 0$, $i \in I$, with

$$c_i(\underline{x}) - s_i = 0, \qquad s_i \geq 0, \qquad i \in I.$$

In LANCELOT solver, the bounds on the variables are explicitly taken into account in the subproblem

$$\min_{\underline{l}_{inf} \leq \underline{x} \leq \underline{l}_{sup}} \; L_A(\underline{x}, \underline{u}^k, \mu_k).$$

5.8 Barrier method

Description for:

$$
\boxed{\begin{array}{lll} \min & f(\underline{x}) & \\ s.t. & c_i(\underline{x}) \geq 0 & i \in I = \{1, \ldots, m\} \\ & \underline{x} \in \mathbb{R}^n. & \end{array}} \tag{1}
$$

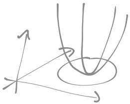
Notation and examples: Chapter 17 of J. Nocedal, S. Wright, Numerical Optimization, Springer, 1999, p. 498-508.

still a "penalty" but works where we try to penalize the approaching close to ∂X

Definition: Let

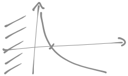
$$X^o = int(\{\underline{x} \in \mathbb{R}^n \, : \, c_i(\underline{x}) \geq 0, i \in I\}) \neq \emptyset,$$

a function defined on $\mathbb{R}^n$ is a **barrier function** if it is continuous over X^o, tends to ∞ when approaching ∂X and has value ∞ on $\mathbb{R}^n \setminus X^o$.



it does not have to go to zero in X^o, it just needs to be continuous

Example: **Logarithmic barrier function** for $c_i(\underline{x}) \geq 0$:



$$-\ln c_i(\underline{x}).$$

Idea: Add to objective function the barrier terms associated to the constraints and solve a sequence of

Definition: The **logarithmic barrier problem** associated to problem (1) is

$$\min_{\underline{x} \in \mathbb{R}^n} \; P(\underline{x}, \mu) = f(\underline{x}) - \mu \sum_{i \in I} \ln c_i(\underline{x}), \tag{2}$$

this constraint need to be ≥ 0

$\rightarrow -\infty$ if we move towards not satisfying it

with the $(-)$ we recover $+\infty$ and hence the penalization

$f(\underline{x}) + \mu \sum [-\ln c_i(\underline{x})]$

with barrier parameter $\mu > 0$.

N.B.: When $\mu \to 0$ the barrier term becomes negligible.

since $a \cdot \ln(b) \xrightarrow[b \to 0]{a \to 0} 0$

We consider $\{\mu_k\}$ with $\lim_{k\to\infty} \mu_k = 0$, start from $\underline{x}_0 \in X^o$ and, for each k, determine an approximate minimizer $\underline{x}_k$ of $P(\underline{x}, \mu_k)$ with an unconstrained optimization method.

Example 1:

$$\begin{aligned} \min \quad & x \\ s.t. \quad & x \geq 0 \\ & 1 - x \geq 0 \end{aligned}$$

with optimal solution $x^* = 0$ and logarithmic barrier problem:

$$\min_{x \in \mathbb{R}} P(x, \mu) = x - \mu \ln x - \mu \ln(1 - x).$$

Compare $P(x, \mu)$ for values of μ from 1 to 0.01.

See J. Nocedal, S. Wright, Numerical Optimization, Springer, 1999, p. 499-500.

Example 2:

$$\begin{array}{ll} \min & (x_1 + 0.5)^2 + (x_2 - 0.5)^2 \\ s.t. & x_1 \in [0, 1] \\ & x_2 \in [0, 1] \end{array}$$

with optimal solution $\underline{x}^* = (0, 0.5)^t$ and logarithmic barrier problem:

Compare contours of $P(\underline{x}, \mu)$ for values of μ from 1 to 0.01.

For $\mu = 0.01$, around $\underline{x}^*$ (more elongated and less elliptical) indicate possible numerical problems.

See J. Nocedal, S. Wright, Numerical Optimization, Springer, 1999, p. 500-502.

General scheme

0) Choose $\varepsilon > 0$, $\mu_0 > 0$, tolerances $\{\tau_k\}_{k \geq 0}$ with $\tau_k > 0$ and $\lim_{k \to \infty} \tau_k = 0$, initial point $\underline{x}_0^s$. Set $k := 0$.

1) Determine an approximate minimizer $\underline{x}_k$ of $P(\underline{x}, \mu_k)$ starting from $\underline{x}_k^s$ and terminate when $\|\nabla P(\underline{x}, \mu_k)\| \leq \tau_k$.

2) `If` overall termination condition is satisfied (e.g, $|f(\underline{x}_{k-1}) - f(\underline{x}_k)| < \varepsilon$)

   `Then` Stop

   `Else` select $\mu_{k+1} \in (0, \mu_k)$ and $\underline{x}_{k+1}^s$, set $k := k + 1$ and `Goto` 1)

Since $\underline{x}_0 \in X^o$ the sequence $\{\underline{x}_k\}$ remains in X^0, the algorithm is an <u>*interior point method*</u>. we we will never get a point on the ∂X, an extreme point

Important connection between a minimum of $P(\underline{x}, \mu)$, denoted $\underline{x}(\mu)$, and a point $(\underline{x}^*, \underline{u}^*)$ satisfying the KKT conditions of problem (1), namely

$$\nabla_{\underline{x}} L(\underline{x}, \underline{u}) = \boxed{\nabla f(\underline{x}) - \sum_{i=1}^{m} u_i \nabla c_i(\underline{x}) = \underline{0}} \tag{3}$$

$$c_i(\underline{x}) \geq 0 \qquad \forall i \in I \tag{4}$$

$$u_i c_i(\underline{x}) = 0 \qquad \forall i \in I \tag{5}$$

$$u_i \geq 0 \qquad \forall i \in I. \tag{6}$$

In a minimizer $\underline{x}(\mu)$ of $P(\underline{x}, \mu)$, we have

KKT multipliers

$$\nabla_{\underline{x}} P(\underline{x}, \mu) = \nabla f(\underline{x}) - \sum_{i=1}^{m} \boxed{\frac{\mu}{c_i(\underline{x})}} \nabla c_i(\underline{x}) = \underline{0}. \tag{7}$$

By defining the estimates of the multipliers

$$u_i(\mu) := \frac{\mu}{c_i(\underline{x}(\mu))} \quad \text{with} \;\; i = 1, \ldots, m, \tag{8}$$

(7) can be rewritten as

$$\boxed{\nabla f(\underline{x}) - \sum_{i=1}^{m} u_i(\mu) \nabla c_i(\underline{x}) = \underline{0}} \tag{9}$$

which is equivalent to (3).

Observation: For $\mu > 0$ the KKT conditions (3)-(6) hold except (5) because

$$u_i(\mu)c_i(\underline{x}(\mu)) = \mu \quad \text{for} \quad i = 1, \ldots, m.$$

When $\mu \to 0$, a minimizer $\underline{x}(\mu)$ of $P(\underline{x}, \mu)$ and

the associated estimate

$$u_i(\mu) := \frac{\mu}{c_i(\underline{x}(\mu))} \quad \text{with} \;\; i = 1, \ldots, m$$

tend to progressively satisfy the KKT conditions of problem (1).

Thus we generate points on the so-called **central path**

$$\{(\underline{x}(\mu), \underline{u}(\mu)) \; : \; \mu > 0\}$$

defined by (8).

Theorem:

Suppose that $X^o \neq \emptyset$ and $\underline{x}^*$ is a local minimum of (1) at which the KKT conditions are satisfied for some $\underline{u}^*$.

Moreover, suppose that

- gradients of the active constraints at $\underline{x}^*$ are linearly independent,
- strict complementarity conditions are satisfied at $\underline{x}^*$ ($\forall i \in I$ exactly one of $c_i(\underline{x}^*)$ or u_i^* is equal to 0),
- 2nd order sufficient conditions are satisfied at $(\underline{x}^*, \underline{u}^*)$.

Then

i) $\exists$ unique continously differentiable vector function $\underline{x}(\mu)$ s.t. $\lim_{\mu \to 0_+} \underline{x}(\mu) = \underline{x}^*$. For all sufficiently small μ, $\underline{x}(\mu)$ is a local minimum of $P(\underline{x}, \mu)$ in some neigborhood of $\underline{x}^*$.

ii) For $\underline{x}(\mu)$ in (i), the Lagrange multiplier estimates $\underline{u}(\mu)$ defined by

$$u_i(\mu) = \mu / c_i(\underline{x}(\mu)) \qquad i = 1, \ldots, m,$$

converge to $\underline{u}^*$ when $\mu \to 0_+$.

iii) $\nabla^2_{\underline{xx}} P(\underline{x}, \mu)$ is positive definite for all sufficiently small μ.

If also equality constraints, one may include quadratic penalty terms (combined log-barrier/quadratic penalty function problem).

Sixth computer lab: application of the logarithmic barrier method to LP.

An **interior point method** for LP

$$\begin{aligned} \min \quad & \underline{c}^t \underline{x} \\ \text{s.t.} \quad & A\underline{x} = \underline{b} \qquad (10) \\ & \underline{x} \geq \underline{0} \qquad (11) \end{aligned}$$

is obtained by applying the barrier method to constraints (11) and by adapting the Newton method to account for (10).

Unlike for Simplex method, such interior point method for LP can be proved to be a polynomial time algorithm.

5.9 Introduction to sequential quadratic programming

Generic NLP:

$$
(P) \qquad \begin{array}{lll}
\min & f(\underline{x}) & \\
s.t. & g_i(\underline{x}) \leq 0 & i \in I = \{1, \ldots, m\} \\
 & h_l(\underline{x}) = 0 & l \in E = \{1, \ldots, p\} \\
 & \underline{x} \in \mathbb{R}^n &
\end{array}
$$

where f, g_i's and h_l's are of class $\mathcal{C}^2$.

Idea: Extend the Netwon method to nonlinearly constrained problems.

Given a current $\underline{x}_k$, we could try to determine an improving direction $\underline{d}_k$ by solving the quadratic approximation of (P):

quadratic approx of $f(\cdot)$ at x_k

$$
(QA_k) \qquad \begin{array}{lll}
\min & \frac{1}{2}\underline{d}^t \nabla^2 f(\underline{x}_k)\underline{d} + \nabla^t f(\underline{x}_k)\underline{d} + f(\underline{x}_k) & \\
s.t. & \frac{1}{2}\underline{d}^t \nabla^2 g_i(\underline{x}_k)\underline{d} + \nabla^t g_i(\underline{x}_k)\underline{d} + g_i(\underline{x}_k) \leq 0 & i \in I = \{1, \ldots, m\} \\
 & \frac{1}{2}\underline{d}^t \nabla^2 h_l(\underline{x}_k)\underline{d} + \nabla^t h_l(\underline{x}_k)\underline{d} + h_l(\underline{x}_k) = 0 & l \in E = \{1, \ldots, p\}
\end{array}
$$

quadratic approx of the $g_i(\cdot)$ and $h_l(\cdot)$

but difficult because of the quadratic constraints.

Observation:

If $(\underline{d}^*, \underline{\eta}^*, \underline{\rho}^*)$ is a stationary point of the Lagrange function associated to (QA_k) it is also a stationary point of the Lagrange function associated to the Quadratic Program:

$$
(QPA_k) \qquad \begin{array}{lll}
\min & \frac{1}{2}\underline{d}^t \nabla^2_{\underline{xx}} L(\underline{x}_k, \underline{\eta}^*, \underline{\rho}^*)\underline{d} + \nabla^t f(\underline{x}_k)\underline{d} + f(\underline{x}_k) & \\
s.t. & \nabla^t g_i(\underline{x}_k)\underline{d} + g_i(\underline{x}_k) \leq 0 & i \in I = \{1, \ldots, m\} \\
 & \nabla^t h_l(\underline{x}_k)\underline{d} + h_l(\underline{x}_k) = 0 & l \in E = \{1, \ldots, p\}
\end{array}
$$

All constraints are linear (approximations).

To obtain a good approximation of (P) via Quadratic Programs, the objective function must include not only a quadratic model of f but also 2nd order information of the g_i's.

General scheme

Let $\underline{x}_0$, $\underline{u}_0$ and $\underline{v}_0$ be estimates of a solution of (P) and of the corresponding multipliers.

Iteration k:

Given $(\underline{x}_k, \underline{u}_k, \underline{v}_k)$ determine $\underline{d}_k$ and the corresponding multipliers $(\underline{\eta}_k, \underline{\rho}_k)$ of the Quadratic Program:

$$
(QP_k) \qquad \begin{array}{lll} \min & \frac{1}{2}\underline{d}^t \nabla^2_{\underline{x}\underline{x}} L(\underline{x}_k, \underline{u}_k, \underline{v}_k)\underline{d} + \nabla^t f(\underline{x}_k)\underline{d} & \\ s.t. & \nabla^t g_i(\underline{x}_k)\underline{d} + g_i(\underline{x}_k) \leq 0 & i \in I = \{1, \ldots, m\} \\ & \nabla^t h_l(\underline{x}_k)\underline{d} + h_l(\underline{x}_k) = 0 & l \in E = \{1, \ldots, p\} \end{array}
$$

Set $\underline{x}_{k+1} := \underline{x}_k + \underline{d}_k$, $\underline{u}_{k+1} := \underline{\eta}_k$ and $\underline{v}_{k+1} := \underline{\rho}_k$

Although (QP_k) derives from (QPA_k) by subsituting the optimal multipliers with the current estimates, it can be proved that:

- feasible region of the subproblem (QP_k) is a *linear approximation* of that of the original problem,
- Lagrange function $L_Q(\underline{d}, \underline{\eta}, \underline{\rho})$ of (QP_k) is a *quadratic approximation* of that of (P).

An iteration of the *Sequential Quadratic Programming method* (SQP) is equivalent to:

- carry out one iteration of the Newton method for the Lagrange function,
- enforce feasibility with respect to the linearization of the feasible region.

The SQP method is well defined:

Proposition:

$(\underline{x}^*, \underline{u}^*, \underline{v}^*)$ is a KKT point of (P) if and only if $(\underline{d}^*, \underline{\eta}^*, \underline{\rho}^*) = (\underline{0}, \underline{u}^*, \underline{v}^*)$ is a KKT point of (QP_k).

Convergence properties similar to those for Newton method:

Quadratic local convergence if

(i) Hessian matrices of the objective function and constraints are Lipschitz continuous,

(ii) constraint qualification assumption is satisfied,

(iii) 2nd order sufficient optimality conditions and strict complementarity conditions are satisfied.

To guarantee global convergence:

- 1-D search that minimizes an appropriate merit function such as

$$M(\underline{x};\mu) = f(\underline{x}) + \frac{1}{2\mu}\Big(\sum_{i=1}^{m} \max\{0, g_i(\underline{x})\} + \sum_{l=1}^{p} |h_l(\underline{x})| \ \Big)$$

- or trust region based approach.

Quasi-Newton versions (without 2nd order derivatives) have also been investigateded.

Several SQP codes are available (SQP, NPSOL, SNOPT, Matlab,...).

Subgradient method

Consider $\min_{x \in \mathbb{R}^n} f(\underline{x})$ with f convex.

Start from an abitrary $\underline{x}_0$.

At k-th iteration: consider $\underline{\gamma}_k \in \partial f(\underline{x}_k)$ and set

$$\underline{x}_{k+1} := \underline{x}_k - \alpha_k \, \underline{\gamma}_k$$

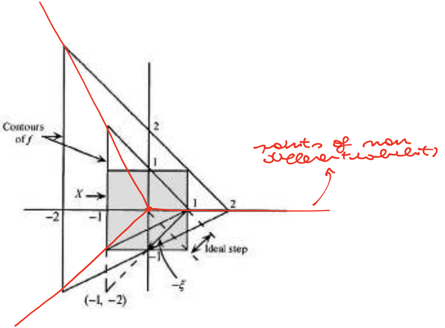
with $\alpha_k > 0$

about the choice of α_k

Observation: No 1-D search (optimization) because for nondifferentiable functions a subgradient $\underline{\gamma} \in \partial f(\underline{x})$ is not necessarily a descent direction!

Example: $\min_{-1 \leq x_1, x_2 \leq 1} \; f(x_1, x_2)$ with $f(x_1, x_2) = \max\{-x_1, x_1 + x_2, x_1 - 2x_2\}$

Level curves in black, points of nondifferentiability $(t, 0)$, $(-t, 2t)$ and $(-t, -t)$ for $t \geq 0$, global minimum $\underline{x}^* = (0, 0)$.



At $\underline{x}_k = (1, 0)^t$ consider $\underline{\gamma}_k = (1, 1) \in \partial f(\underline{x}_k)$, $f(\underline{x})$ increases along $\{\underline{x} \in \mathbb{R}^2 \, : \, \underline{x} = \underline{x}_k - \alpha_k \underline{\gamma}_k, \alpha_k \geq 0\}$ but if α_k is sufficiently small then $\underline{x}_{k+1} = \underline{x}_k - \alpha_k \underline{\gamma}_k$ is closer to $\underline{x}^*$.

good idea to follow

From Chapter 8, Bazaraa et al., Nonlinear Programming, Wiley, 2006, p. 436-437

Theorem:

small steps — but not too small

If f is convex, $\lim_{\|\underline{x}\| \to \infty} f(\underline{x}) = +\infty$, $\lim_{k \to \infty} \alpha_k = 0$ and $\sum_{k=0}^{\infty} \alpha_k = \infty$, the subgradient method terminates after a finite number of iterations with an optimal solution $\underline{x}^*$ or infinite sequence $\{\underline{x}_k\}$ admits a subsequence converging to $\underline{x}^*$.

Stepsize:

In practice $\{\alpha_k\}$ as above (e.g., $\alpha_k = 1/k$) are too slow.

An option: $\alpha_k = \alpha_0 \rho^k$ for a given $\rho < 1$. A more popular one (min problems):

$$\alpha_k = \varepsilon_k \frac{f(\underline{x}_k) - \hat{f}}{\|\underline{\gamma}_k\|^2},$$

where $0 < \varepsilon_k < 2$ and $\hat{f}$ is either the optimal value $f(\underline{x}^*)$ or an estimate.

Stopping criterion: prescribed maximum number of iterations
(even if $\underline{0} \in \partial f(\underline{x}_k)$ it may non be considered at $\underline{x}_k$).

Need to store the best solution $\underline{x}_k$ found. as we may take a "wrong" subgradient

Simple extension for bounds (projections).